データ駆動型研究を加速する東北大学研究データレイク IZUMI

野﨑 真治 [1)], 木村 優太 [1)], 佐藤 信夫 [2)], 宗形 聡 [3)], 中村 隆喜 [3)]

1) 東北大学　情報部デジタル基盤整備課
2) 東北大学　データシナジー創生機構
3) 東北大学　サイバーサイエンスセンター
shinji.nozaki.d5@tohoku.ac.jp

Tohoku University Research Data Lake IZUMI Powering Data-driven Research

Shinji Nozaki[1)], Yuta Kimura[1)], Nobuo Sato[2)], Satoshi Munakata[3)], Takaki Nakamura[3)]

1) Digital infrastructure Division, Tohoku Univ.
2) Organization for Innovations in Data Synergy, Tohoku Univ.
3) Cyberscience Center, Tohoku Univ.

概要

国内外で広がりを見せているデータ駆動型研究やオープンサイエンスを推進するため，東北大学では 2025 年 6 月から東北大学研究データレイク IZUMI の運用を開始した。本稿では，IZUMI の構築に際して検討した設計およびアーキテクチャの要点を紹介する。また，研究データ管理に関連する他システムとの連携や実環境で実施したシステム性能評価の結果も報告する。

1. はじめに

近年のデジタルトランスフォーメーション（以下「DX」という）の進展に伴い，国内外の学術研究機関では，従来の仮説検証型の研究開発に加えて，大規模データの解析により新たな科学的知見を生み出すデータ駆動型研究の展開が急速に進んでいる。また，G7 仙台科技大臣会合（Communique）での共同声明[1]や国の統合イノベーション戦略を受け，研究活動で生成・収集したデータやデータに基づいて得られた成果を広く公開し，新たなイノベーション創出につなげるオープンサイエンスの推進も求められている。

こうした状況のもと，東北大学（以下「本学」という）では，東北大学ビジョン 2030 の中でコネクテッドユニバーシティ戦略[2]を掲げ，研究 DX によるデータ駆動型研究とオープンサイエンスを推進している。研究 DX では，教職員が高い生産性をもって研究や周辺業務に従事できるよう，本学内外の様々なシステムが連携する研究データ基盤の開発を進めている。具体的には，研究者データベースと連携し，公的資金に紐づく研究のデータマネジメントプラン作成や成果論文との関連付けを管理する研究データ管理アプリ，学務データおよび研究データを情報資産としてカタログ化する統合データカタログシステムなどのシステムを開発し，運用を始めている。しかし，これらのシステムを有機的に連携させ，データ駆動型研究とオープンサイエンスのさらなる加速化を実現するためには，研究データ基盤の中核となる研究データレイクの整備が課題となっていた。

そこで本学では，2024 年度に東北大学研究データレイク IZUMI（以下「IZUMI」という）を構築し，試験運用期間を経て 2025 年 6 月から正式運用を開始した[3]。IZUMI の構築では，ユーザである教職員が利用に際して発生する作業負担を可能な限り低減しつつ，研究データの一元管理による利活用促進と即時オープンアクセス（以下「OA」という）をともに実現できるよう，設計やアーキテクチャを検討した。本稿では，IZUMI で採用した設計やアーキテクチャの概要について述べるとともに，実運用システムの性能評価として試験運用期間中に実施したスループット測定の結果について報告する。また，今後の展開として研究データ基盤における他システムとの連携を自動化する取り組みについても紹介する。

2. IZUMI の構築

2.1　システム全体像

IZUMI は，本学のサイバーサイエンスセンター計算機室内に構築したオンプレミスのストレージシステムである。本学の教職員約 3,500 名が利用対象者となっている。

IZUMI のシステム構成図を図 1 に示す。IZUMI は実効容量で 5PB のストレージ装置を有している。また，当該ストレージ上で Nextcloud，Amazon S3 互換ストレージ（以下「S3 という」），および

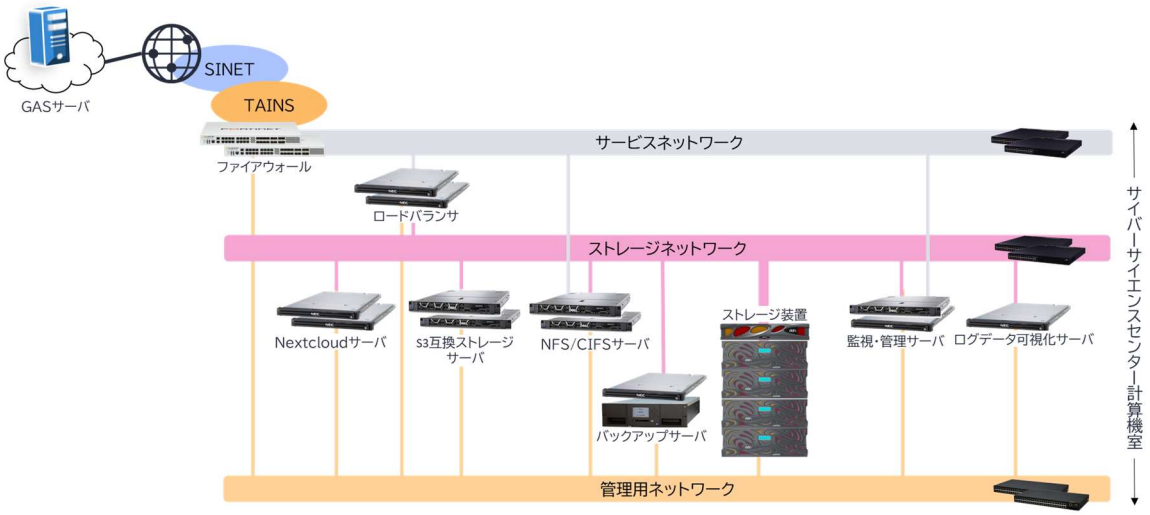

図 1　IZUMI のシステム構成図

NFS の各サービスを提供するサーバがそれぞれ配備されている。ストレージ装置やサーバは，本学の学内ネットワーク（以下「TAINS」という）の中に構築した IZUMI 専用のネットワークに接続する。専用ネットワークでは，TAINS との接続ポイントでファイアウォールによるセキュリティ対策を実施する。Nextcloud と S3 の両サービスでは，各サーバの前段でロードバランサによる負荷分散を行い，学内外の端末からアクセス可能としている。一方，NFS サービスではインターネットからのアクセスを遮断し，学内からのみ利用可能としている。なお，NFS は提供準備中であり，NFS クライアントのグローバル IP アドレスを運用サイドで把握する方法などの課題について現在検討を進めている。また，各機器や専用ネットワーク回線を含めた IZUMI の最大スループットは 20Gpbs となっている。

2.2　IZUMI の設計

IZUMI は本学で初めて本格的に提供されるオンプレミスの研究データレイクシステムである。初めて提供されるシステムにおいて重要なのは，まずユーザに利用してもらうことである。ユーザに利用してもらうことができれば，そこからネガティブなまたはポジティブなフィードバックを得て，より良いシステムに改良していくことができる。本学では，組織メンバを対象ユーザとするシステムにおいて，ユーザの利用を阻害しうる要素には以下の 3 つがあると考えた。

- 申請しないと利用できない／利用の申請に手間がかかる
- 直感的にシステムを操作できない／システムの学習コストが高い
- 既に使っているシステムからの移行が手間である／システム同士が連携しない

IZUMI の設計に際し，これらの阻害要素を可能な限り取り除き，使い始めようとするユーザの抵抗を払拭できるよう，次の 3 つの設計方針を立ててシステムを構築した。

1）Agility…すぐに利用できること

当初の構想では，学内ユーザは申請不要で利用可能とし，ポータル機能を提供してストレージ領域の確保や容量変更，共有ユーザの追加・削除，データ公開（即時 OA 化）などほぼすべての設定をユーザのセルフサービスで行えるようにする設計であった。しかし，予算やスクラッチ開発のリスクを考慮し，最終的には IZUMI はストレージサービスの提供に特化し，利用申請やポータルで実現しようとしていた各種設定は，クラウド上の Google Apps Script（以下「GAS」という）サーバで実行する設計とした（図 1 参照）。利用申請については，教員は基本的に申請不要で好きなときにいつでも利用開始できるようにし，職員は Google フォームに必要事項を入力して GAS サーバ経由で申請する仕組みとした。ポータルに相当する各種設定も同様に，設定ごとに用意した Google フォームにユーザが記入した上で送信することで，GAS サーバ経由で IZUMI に反映させる設計とした。このようにして，申請手続きの簡素化かつ自動化を実現した。

2）Usability…簡単に利用できること

IZUMI では，直感的に操作可能な Web ブラウザベースのデータサービスとしてオープンソースのストレージシステムである Nextcloud を採用した。Nextcloud は，Google Drive や Box，OneDrive など主要なクラウドストレージと類似した UI でデータの格納や取得ができる。データ共有に関連する操作もマウスクリックで可能であり，非公開から OA 化（公開）までのデータ公開状態を研究段階に応じて柔軟に設定できる。研究データ管理で使用する機能に限れば，普段から研究や業務で PC を利用するユーザであれば習得は難しくない。また Nextcloud は，教育・研究用途のストレージインフラや複数のストレージ基盤と連携するデータ

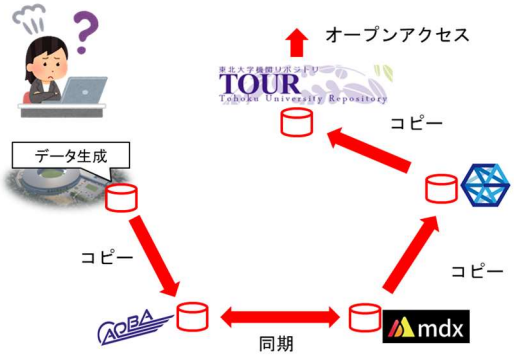

図 2　コンピュータ中心アーキテクチャ

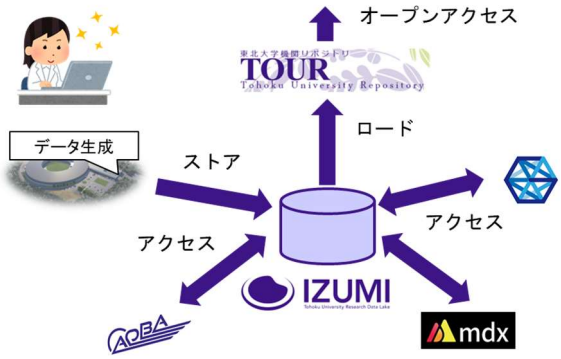

図 3　データ中心アーキテクチャ

集約基盤として，多くの大学や公共機関で採用されており [4, 5]，既に学生や教職員など多様なユーザに使われている実績がある。このことからもユーザビリティの高さが伺える。

3）Interoperability…既存システムと繋がること

本学では，国立情報学研究所（以下「NII」という）との密接な協力関係のもと，NII が提供する研究データ管理基盤 GakuNin RDM を学内向けに展開している。また，5 年以上にわたって研究データ管理に関して Proof-of-Concept などの様々な施策を一緒に推進してきた。GakuNin RDM のユーザにとって，IZUMI との間でシームレスにデータをやり取りできることは，IZUMI の利用を決断する上で必須の機能である。そこで IZUMI では，S3 サービスで提供するバケットを Nextcloud の外部ストレージとしてユーザが利用できるようにし，かつ当該バケットのキー情報をユーザに提供して GakuNin RDM の外部ストレージとしても登録できる設計とした。これにより，Nextcloud と GakuNin RDM が共通の S3 バケットを経由して直接つながることになり，ユーザは容易に相互のデータのやり取りが可能となる。

2.3　データ中心アーキテクチャ

研究データを格納するデータレイクがデータ駆動型研究やオープンサイエンスの推進に資するためには，研究者がデータを利活用したいときにいつでも所望のデータにアクセスできる状況をデータレイクで提供する必要がある。そのためには，研究過程で生じるあらゆる研究データをデータレイクに集約し，一元的に管理できるようにすることが重要である。

従来のコンピュータ中心アーキテクチャでは，図 2 に示すようにデータ生成元をはじめスーパーコンピュータ[6]，仮想計算機基盤[7]，データ管理基盤，機関リポジトリなどの各コンピュータがそれぞれ研究データを一時的あるいは永続的に格納するストレージを持つことになる。この場合，研究者は利用したいコンピュータのストレージに必要な研究データをコピーしたり，データ更新によりコンピュータ間でストレージを同期したりする必要がある。その結果，同じ内容のデータが複数のコンピュータに格納された状態が頻発する，どこにあるのが公開可能な最新バージョンのデータなのか把握できなくなる，などデータ管理上の問題が発生して円滑な研究データの利活用や即時 OA 化が困難となる。

そこで IZUMI では，データ中心アーキテクチャを指向したデータレイクとなるようにシステムを設計した。データ中心アーキテクチャのイメージを図 3 に示す。データ中心アーキテクチャでは，全ての研究データは IZUMI に集約され，一元管理される。各コンピュータは必要なときに IZUMI にアクセスして，そこから必要なデータを取得または格納（更新）する。このようなアーキテクチャとすることで，研究者はコンピュータ間でのデータコピーや同期処理から解放され，コンピュータ中心アーキテクチャで生じるサイロ化や管理の煩雑さなどのデータ管理上の問題も解消される。こうして，研究成果（論文）やその基礎データ（実験データやソースコードなど）の検索・再利用・公開の促進が可能となる。

3. 提供サービスと申請プロセス

3.1　IZUMI の提供サービス

IZUMI で提供する利用用途別のサービス一覧を表 1 に示す。IZUMI では，利用用途別に個人利用，グループ利用，一般公開の 3 つのサービスを提供する。

個人利用サービスは，利用者個人に 100GB の専用ストレージ領域を提供するサービスである。利用者は必要に応じて拡張申請することにより，ストレージ領域の拡張が可能である。本サービスは Nextcloud でのみ提供する。利用者は Nextcloud の URL 共有機能を用いて，個人領域にあるデータを学内外の第三者と共有することができる。

グループ利用サービスは，利用者グループに対して 100GB のストレージ領域を提供するサービスであり，データ共有が必要な共同研究などでの利用を想定している。本サービスも，必要に応じてグループ管理者が拡張申請することにより，ストレージ領域の拡張が可能である。本サービスは

表 1　利用用途別サービス

	個人	グループ	一般公開
利用申請	不要※1	要	要
基本容量	100GB	100GB	制限なし
拡張容量	利用者指定	利用者指定	-
費用負担※2	基本容量：大学 拡張容量：利用者	利用者（グループ管理者）	大学
利用者退職時の対応	データ削除	データ削除	データ維持
提供方法	Nextcloud	Nextcloud，S3，NFS（調整中）， GakuNin RDM 拡張ストレージ	Nextcloud
学外者との共有	URL 共有（パスワードや有効期限を設定可）	・Nextcloud：URL 共有 ・GakuNin RDM：拡張ストレージ	-

※1 教員のみ。職員は利用申請が必要。※2 当面利用者の費用負担はなし。

S3 互換オブジェクトストレージのバケットとして提供される。グループメンバ内でのデータ共有を目的として REST API や Nextcloud 外部ストレージを利用でき，学認参加機関の研究者との共有も可能な GakuNin RDM の拡張ストレージとしても利用できる。なお，Nextcloud 外部ストレージでは URL 共有機能による学外者とのデータ共有も可能である。

一般公開サービスは，利用者が OA 化のため個人領域で URL 共有したデータを本学が管理する領域へ委譲することで，個人によるデータ修正を不可としつつ，異動や退職に伴うアカウント削除後も発行済み URL を永続的に有効化することができる。これにより，機関リポジトリに登録した研究成果データへの URL リンクを維持することができる。

3.2　申請プロセス

IZUMI では，Google フォームを活用した利用申請受付および GAS サーバを用いた自動設定システムを実装している。以下，本プロセスの詳細について述べる。

1) 申請受付
   フォーム送信により利用申請や領域拡張申請，グループ利用申請などの申請を受理する。
2) データ処理
   フォーム送信後，本学の認証システムおよびマスタデータとの照合が行われ，申請内容を検証する。不備が検出された場合には申請者にメールで自動通知される。内容に問題がなければ，GAS サーバで申請内容を記載した CSV ファイルを生成し，IZUMI の監視・管理サーバに送信する。ファイル送信は，監視・管理サーバ上で稼働する Web アプリケーションを用いて行われる。
3) システム設定の自動化
   監視・管理サーバに送信された申請データは，申請種別ごとに用意したスクリプトによって定期的に処理され，LDAP や Nextcloud，S3 バケットなどの設定が自動的に反映される。

これらの仕組みにより，利用者は迅速かつ効率的に IZUMI の各種サービスを利用可能となっている。

4. 研究データ基盤における IZUMI

本学では，研究 DX によりデータ駆動型研究とオープンサイエンスの推進を図っている。高いレベルの研究 DX を実現し，教職員が研究活動やサポート業務に専念できる環境整備を目的として，研究データ基盤を開発している。研究データ基盤の全体像を図 4 に示す。

研究データ基盤には，従来から業務で使用してきた学務・人事・財務の各システムやバックエンドのデータベース，図書館が運用する機関リポジトリに加え，IZUMI に先行する形で開発された研究データ管理アプリと統合データカタログシステムが含まれている。研究データ管理アプリは，研究者によるデータマネジメントプラン（以下「DMP」という）作成の支援，外部資金や成果論文と作成した DMP の紐づけ，機関リポジトリと連携した即時 OA 化支援などの機能を持つ。統合データカタログシステムは，経営戦略データベースや研究データ管理アプリと連携して，業務・学務データおよび研究データをカタログ化し，情報資産として管理する機能を有する。

IZUMI は，研究データ基盤の中核として位置付けられており，研究データを一元管理するデータプラットフォームとして他システムと連携して動作する。研究データ管理アプリでは，即時 OA 化支援として研究データの公開 URL を登録することができる。IZUMI の一般公開サービスで発行される研究データの共有用 URL をそこに登録すれば，アプリを起点にして IZUMI で取得した公開 URL の機関リポジトリへの登録を連動させることができる。しかし，現時点では研究者が手動で公開 URL をアプリに登録する運用となっており，データ連携のために研究者の作業負担が生じている。そこで，Nextcloud が保有する外部連携用 Web API（OCS API）を活用し，研究データ管理

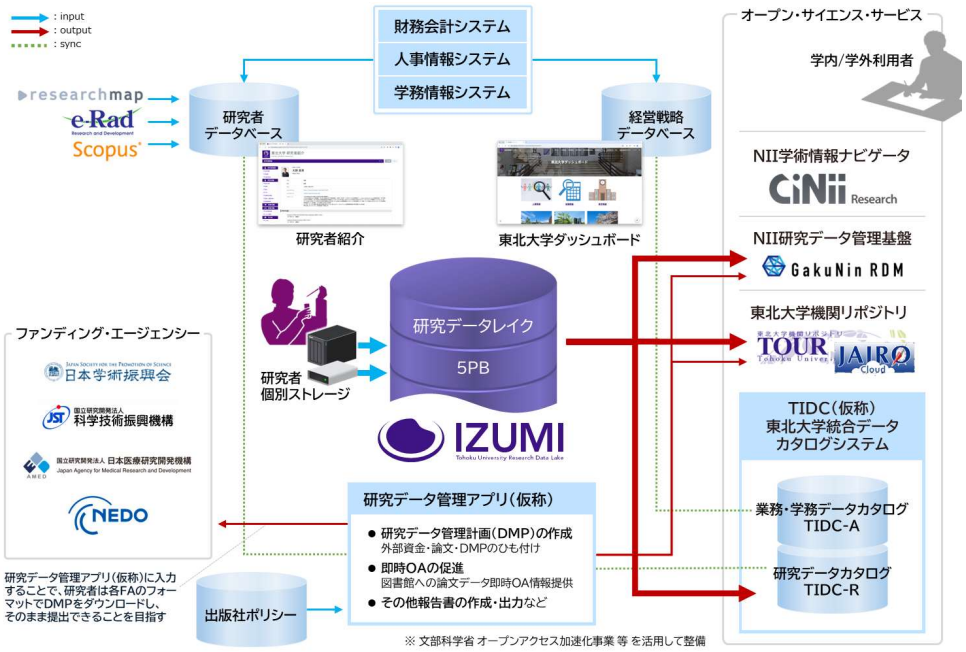

図 4　研究データ基盤の全体像

アプリで事前に公開データの URL 一覧を IZUMI から取得しておき，機関リポジトリに登録する URL をアプリ上で研究者が選択できるようにする機能の開発を構想している。

以上のように，IZUMI の Web API を活用して研究データ管理アプリを含め他のシステムともデータ連携を進めることで，研究データ管理に要する研究者の手間を最小化し，研究データ基盤の利用を促進していくことが今後求められていく。

5. 性能測定

研究データレイクの運用では，クライアントからデータを送受信するときのシステムパフォーマンスが利用者にとって重要な指標となる[5]。本節では，IZUMI の提供サービスごとにクライアントから 1 ファイルを IZUMI へアップロードまたはダウンロードしたときのスループット性能を評価した結果を述べる。スループットの測定は，典型的な利用形態と想定されるキャンパス内のクライアント端末から IZUMI にデータを送受信する方法で実施した。スループットの測定環境を図 5 に，測定に用いたマシンのスペックおよび前提ソフトウェアの詳細を表 2 にそれぞれ示す。

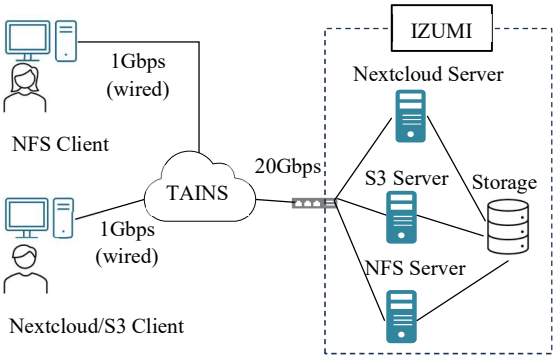

図 5　性能測定環境

本測定では，Nextcloud，S3 および NFS の各ストレージサービスに対し，ファイル送受信時のスループットを算出した。Nextcloud および S3 については，Python プログラムを用いてファイルアップロードまたはダウンロード時の応答時間を計測し，ファイルサイズを応答時間で除算することでスループット値を求めた。ファイル送信では，両サービスともにマルチパートアップロードを実行した。パートサイズはそれぞれ 10MB と 64MB[6] に設定した。マルチパートアップロード時に各パートは並列に送信される。一方，ファイル受信で

表 2　マシンスペックと前提ソフトウェア

	Nextcloud／S3 Client	NFS Client	Nextcloud Server	S3 Server
CPU	13th Gen Intel™ Core™ i9-13900KF 16 コア	Intel™ Xeon™ E5-2420 12 コア	Intel™ Xeon™Silver 4410Y 12 コア	Intel™ Xeon™ Silver 4514Y 16 コア
RAM	64GB	12GB	32GB	512GB
前提ソフト	Python v3.12.0	fio v3.28	-	-

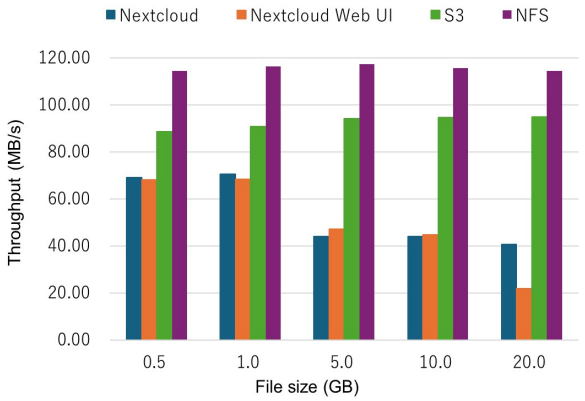

図 6　1 ファイル送信時のスループット

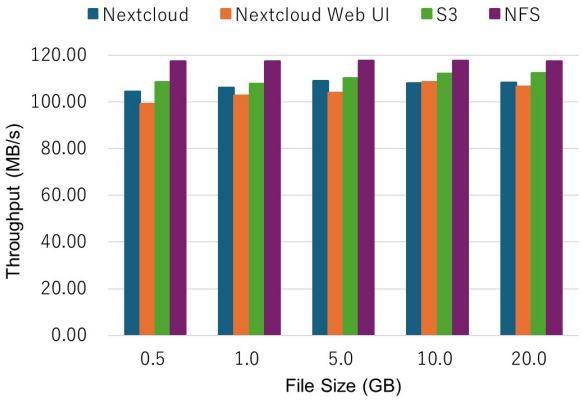

図 7　1 ファイル受信時のスループット

はパート分割を実施していない。また NFS では，fio ツールを用いてマウントポイント上でシーケンシャルライトまたはシーケンシャルリード処理の実行時間を計測し，データサイズを実行時間で除算してスループット値を求めた。ファイルサイズごとに 10 回ずつ測定を行い，その平均値を最終的なスループット評価値とした。なお，Python プログラムおよび fio ツールはともにクライアント端末上で実行した。

図 6 に各サービスの単一ファイル送信時のスループット値を示す。Nextcloud の標準的な利用方法である，クライアント端末で Web ブラウザを用いた手動ファイルアップロード時のスループット計測値（5 回計測の平均値）を Nextcloud Web UI として参考までに含めた。なお，Nextcloud の仕様で手動アップロードの場合も 10MB のパート分割によるマルチパートアップロードが実行される。

スループット性能の評価においては，S3 および NFS がデータサイズに依存することなく一貫して高い値を示した。NFS は S3 よりも優れた性能を観測したが，これは NFS が S3 で使う HTTP よりも下位層で動作し，かつオーバヘッドの少ない RPC 通信を用いてバイナリ化したデータを送受信するためと推察される。ただし，IZUMI の NFS サービスは，研究室のようなクローズドな組織内でのファイル共有を目的としているため，学内外からアクセスする S3 と比べて利用場面が限定される。また，5GB 超のファイルサイズの場合には，Nextcloud でスループットの低下が認められた。Web UI でも同様の傾向であるため，スループット低下の原因は Python プログラムではなく，Nextcloud のマルチパートアップロード機能にあると考えられる。マルチパートアップロードの処理工程ごとに応答時間を計測した結果，全パート送信後にサーバ側で実行される結合工程が総処理時間の 30%以上を占めていることが分かった。従って，結合工程に性能改善の余地があると判断される。以上のことから，5GB を超える大容量の観測データや AI・機械学習用の大規模データ格納には，Nextcloud よりも S3 の利用が推奨される。

図 7 に各サービスの単一ファイル受信時のスループット値を示す。本評価では，ファイル送信時と同様に Web ブラウザから手動でファイルをダウンロードしたときの，スループット計測値（5 回計測の平均値）を Nextcloud Web UI としてプロットした。ダウンロード時のスループット性能については，各サービス間で顕著な差異は認められず，データサイズに依存せずに 100MB/s 以上の高いスループットが安定して得られている。従って，利用シナリオ（例：手動によるデータダウンロード，実験プログラムや連携システムからの直接のデータ取得など）に応じて，Nextcloud，S3，NFS のどのサービスでも選択可能である。

6. まとめ

本学では，研究データ基盤の中核として 2025 年 6 月から東北大学研究データレイク IZUMI の運用を開始した。IZUMI の導入により，研究過程で発生・生成したデータを集約して一元管理することが可能となり，それをもとに研究データ基盤を構成するシステム同士を有機的に連携させた，高度な研究 DX を実現した。

本稿では，IZUMI の構築に際して検討した設計およびアーキテクチャの要点を中心に紹介し，研究データ基盤における IZUMI の位置づけや，実環境で測定したシステム性能の評価についても述べた。設計では，Agility，Usability，Interoperability の 3 点を考慮して教職員が IZUMI を利用する際の障壁を可能な限り排除した。また，データ中心アーキテクチャを採用してシステム間のデータコピーや同期を不要とし，研究データ管理を容易化するとともに，利活用や公開を円滑化する仕組みを構築した。性能評価では，IZUMI で提供する Nextcloud，S3，NFS の各ストレージサービスに対し，クライアントから所定サイズの単一ファイルを送受信するときのスループットを計測した。その結果，ファイル送信では 5GB 超のサイズで Nextcloud のスループットは低下するため，大規

模データ送信時には S3 や NFS の使用する方がよいこと，ファイル受信ではデータサイズに依らずどのストレージサービスでも高いスループットが得られることが分かった。

今後は，IZUMI の外部連携用の Web API を用いて，研究データ基盤を構成する他システムとのデータ連携の自動化に取り組む予定である。自動連係の実現を通して，IZUMI 並びに研究データ基盤の利用促進を図り，データ駆動型研究およびオープンサイエンスのさらなる推進につなげていく。

参考文献

[1] G7 仙台科学技術大臣会合：G7 Science and Technology Ministers’ Communiqué，https://www8.cao.go.jp/cstp/kokusaiteki/g7_2023/230513_g7_communique.pdf（2025 年 9 月 16 日参照）

[2] 東北大学ビジョン 2030：東北大学ビジョン 2030（アップデート版）「コネクテッドユニバーシティ戦略」，https://www.tohoku.ac.jp/japanese/profile/vision/01/vision04/（2025 年 9 月 10 日参照）．

[3] 東北大学プレスリリース：AI 時代の研究 DX を加速する研究データ基盤を構築 中核となる東北大学研究データレイク「IZUMI」の運用開始，https://www.tohoku.ac.jp/japanese/2025/05/press20250520-02-izumi.html（2025 年 9 月 10 日参照）．

[4] 伊達 進，寺前 勇希，勝浦 裕貴，ほか：ONION：大阪大学のデータ集約基盤，学術情報処理研究，Vol. 26，pp. 87-96，2022.

[5] 九州大学研究データ管理支援：九州大学研究データ管理用ストレージシステム（QRDM），https://rds.dx.kyushu-u.ac.jp/qrdm（2025 年 9 月 10 日参照）

[6] H. Takizawa, K. Takahashi, Y. Shimomura et al.: AOBA: The Most Powerful Vector Supercomputer in the World, in Sustained Simulation Performance 2022, pp. 71–81. 2024.

[7] T. Suzumura, A. Sugiki, H. Takizawa, et al.: mdx: A Cloud Platform for Supporting Data Science and Cross-Disciplinary Research Collaborations, 2022 IEEE Intl Conf on DASC/PiCom/CBDCom/CyberSciTech, pp. 1–7, 2022.

[8] 宗形 聡，中村 隆喜：東北大学研究データレイク IZUMI の応答時間評価，学術情報処理研究，Vol. 29，2025.（採録決定）

[9] Yi Ting Chung, Takaki Nakamura: Optimizing Object Storage Performance for Large File Uploading, 情報処理学会第 202 回マルチメディア通信と分散処理研究会, 2025.