jh220058

次世代学術情報基盤に向けた基盤ソフトウェアの実践的な研究・開発・評価

杉木 章義（北海道大学）

概要

2021 年度から本格稼働した NII Research Data Cloud (RDC) と高性能仮想化基盤 mdx，情報基盤センターや HPCI の大型計算機群との連携を深化させ，次世代の計算科学・データ科学・データ駆動科学・データ利活用分野を支える学術情報基盤に向けた基盤ソフトウェアの研究・開発・評価を行う. 以下の 3 点を中心に実環境へのデプロイ・活用を目指した実践的な研究開発を行う．(1) 仮想化・コンテナによる高再現性を有する計算環境と高性能計算環境との融合，(2) セキュアデータ流通環境の構築とプライバシー保護，(3) RDC・mdx・大型計算機間でのデータ連携と情報の共有．本報告書では，初年度の成果について報告する．

1 共同研究に関する情報

1.1 共同研究を実施した拠点名

- 北海道大学 情報基盤センター
- 東京大学 情報基盤センター
- 大阪大学 サイバーメディアセンター
- mdx

1.2 課題分野

- データ科学・データ利活用課題分野

1.3 共同研究分野 (HPCI 資源利用課題のみ)

- 超大規模情報システム関連研究分野

1.4 参加研究者の役割分担

- 杉木 章義（北海道大学，研究代表者）：コンテナ基盤に関する研究開発
- 空閑 洋平（東京大学，副代表者）：仮想化基盤・ネットワークに関する研究開発
- 田浦 健次朗（東京大学）：mdx 計算機との調整（全体管理）
- 塙 敏博（東京大学）：mdx 計算機との調整（アーキテクチャ周辺）
- 鈴村 豊太郎（東京大学）：mdx 計算機との調整（アプリ領域や計算機運用）
- 中村 遼（東京大学）：ネットワークに関する研究開発
- 宮本 大輔（東京大学）：セキュリティに関する研究開発
- 合田 憲人（NII）：コンテナ基盤・ストリーム環境に関する研究開発
- 竹房 あつ子（NII）：コンテナ基盤・ストリーム環境に関する研究開発
- 藤原 一毅（NII）：研究データ管理に関する研究開発
- 伊達 進（大阪大学）：他拠点との接続
- 建部 修見（筑波大学）：ストレージに関する研究開発

2 研究の目的と意義

本研究課題の前年度となる 2021 年度から本格稼働した NII Research Data Cloud (RDC) と高性能仮想化基盤 mdx，HPCI/JHPCN の大型計算機群との連携を深化させ，次世代の計算科学・データ科学・データ駆動科学・データ利活用分野を支える学術情報基盤に向けた基盤ソフトウェアの研究，開発，評価を行う．各機関の基盤ソフトウェア研究者が集まり，大きく以下の 3 点を中心に，学術基盤の実環境へのデプロイ・活用を目指した実践的な研究開発を行う．

- **テーマ 1：仮想化，コンテナによる高再現性計算環境と高性能計算環境とのコンバージェンス**
  仮想化やコンテナ技術による高い計算再現性，柔軟性，利便性を利用者に届ける技術の開発（環境の自動構築）と性能評価を実施する．後者の性能評価によって, 仮想化, コンテナを中心とするクラウド環境と従来のスーパーコンピュータ（以下, スパコン）による高性能計算環境のコンバージェンスの可能性を模索する．
- **テーマ 2：セキュアデータ環境の構築とプライバシー保護**
  安全性やプライバシー保護を向上させ，ユーザ間及びグループ間で安心してデータ共有できる環境の設計と構築を実施する．
- **テーマ 3：RDC，mdx，スパコン間のデータ連携，情報共有**
  RDC 上のデータを mdx 及びスパコンの高性能な計算環境でシームレスに処理できる次世代学術計算基盤の実現へ向けた研究開発を行う．また，次期学術情報ネットワーク SINET6 も活用し，mdx とスパコン間，スパコン・スパコン間のデータ転送や共有も実現する．

3 当拠点公募型研究として実施した意義

今年度から JHPCN に新設された「データ科学・データ利活用分野」の研究を推進するためには，各拠点の研究者が連携し，その基盤となるソフトウェアの整備が必要である．各基盤ソフトウェア研究者が個別に開発してきた成果を持ち寄り，mdx を中心とした実環境にデプロイすることで，研究環境の整備を進めるとともに，現状の課題や問題点を明らかにする．

各大学の情報基盤センター群や HPCI の第 2 階層機関が導入・運用しているスパコンは我が国の計算基盤の重要な一角を占めており，それらは日本の全都道府県を 100Gbps 以上の帯域幅で接続した SINET 経由で，広く利用可能となっている．しかし，それらが短期間で柔軟に連携し，有機的な計算環境として十分に機能しているとはいい難い状態にある．一方，計算基盤に対する需要は, 従来からの計算能力に対する需要とともに，(1) データの収集・蓄積・管理・活用のための基盤に対する需要，(2) データのセキュリティの重要性が著しく増している．また，(3) ソフトウェアの進化やその普及によって，高性能計算環境の使い方に対する期待値やノルムも変化している．2021 年度の NII RDC と mdx の稼働開始はこの状態からの脱却の第一歩であるが，ソフトウェアや支援体制の面で，欠けている要素や，解決すべき課題は数多くある．本研究により，次世代の学術情報基盤を形成する大きな力となることを目標とする．

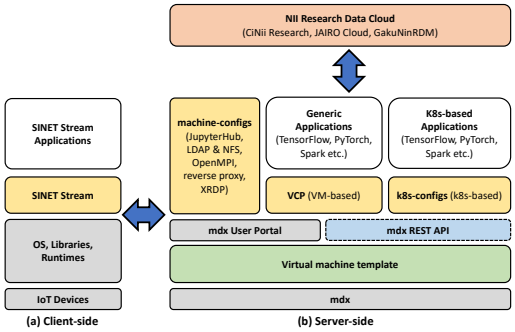

図 1　次世代学術情報基盤の全体構成

4　前年度までに得られた研究成果の概要

2022 年度（当該年度）新規採択課題のため，該当しない．

5　今年度の研究成果の詳細

研究期間の前半では，各共同研究者が個別に研究開発を続けてきた基盤ソフトウェアの成果を mdx に移植する作業を行なった．初期の全体的な研究成果（図 1）については，大学 ICT 推進協議会（AXIES）の研究論文誌「学術情報処理研究」に掲載されている [1].

研究期間の後半では，上記の成果を mdx 上でより発展させるとともに，拠点間でセキュアに高速転送を行う Multi-threaded scp などの新規開発も行なった．

5.1　仮想化・コンテナによる高再現性を有する計算環境と高性能計算環境との融合

テーマ 1 の研究の再現性と高性能計算を両立する計算環境に関しては，machine-configs，VCP (Virtual Cloud Provider), k8s-configs [8] の三つで研究を進めた．本研究課題では，提供する計算環境を一つに絞り込まず，複数の計算環境を許容し，相互に情報交換することで発展させるアプローチをとっている．また，アプリケーション研究者に複数の選択肢を提供し，その都度，最適な研究環境を構築できるようにしている．

5.1.1　machine-configs と仮想マシンテンプレート

machine-configs は，データ利活用に必要な Jupyter クラスタ環境一式をセットアップする Ansible スクリプト集であり，GitHub を通して公開している．machine-configs では，データの共有やユーザアカウント管理のための NFS と LDAP サーバ，JupyterLab 環境，Open-MPI による並列計算環境，JupyterLab に接続するためのリバースプロキシ，デスクトップ画面接続のための xrdp などを一括して自動で導入する．

仮想マシンテンプレートは，mdx の利用に必要なデバイスドライバや設定などを導入したインストール済み OS イメージである．本テンプレートは Amazon EC2 の AMI（Amazon Machine Image）などに相当する．仮想マシンテンプレートは，ベンダが導入当初，作成したものも提供されているが，研究分担者らが継続的にメンテナンスを行なっているテンプレートが広く利用されている．現在，Ubuntu 20.04 と 22.04 が提供されており，デスクトップ版とサーバ版，CPU 対応版と GPU 対応版で 4 種類の組み合わせが存在する．今後，RHEL クローンの Rocky Linux の対応を予定している．

machine-configs 及び仮想マシンテンプレートは，本研究課題以前から開発されているが，継続的に大きな改良やメンテナンスが行われている．

5.1.2　VCP

VCP は，コンテナ技術を利用して共通の API から主要なクラウド上で研究，教育のためのアプリケーション環境の構築を可能にするシステムである．本研究では，VCP を用いて mdx 上でも同様にアプリケーション環境の構

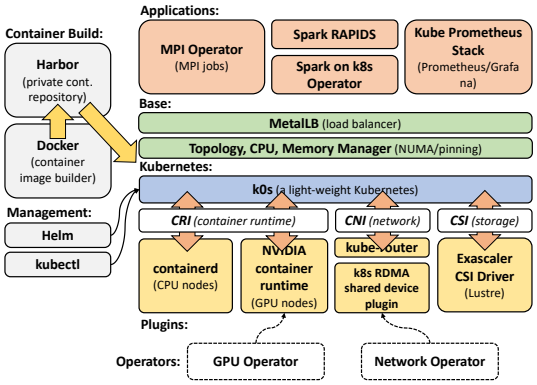

図 2　k8s-configs

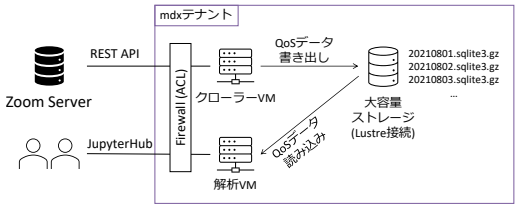

図 3　mdx 上に構築した長期データ保存環境

築ができるようにするため，mdx REST I/F の開発，mdx REST I/F の利用を容易にする Python ライブラリの開発，および VCP ポータブル版の開発を行った [3][6]．VCP の mdx 対応により，VCP が現在提供している公開アプリケーションテンプレートを用いて，講義演習環境，HPC クラスタ環境，JupyterHub が構築できることを確認した．また，群馬大学では VCP と講義演習環境テンプレートを用いて Jupyter ベースの演習環境を構築して，授業で利用していただくことができた．さらに，Open OnDemand の構築や，スケールアウト／インが可能な HPC クラスタテンプレートを行っており，発表準備を進めている．

5.1.3　k8s-configs

k8s-configs [2] は mdx に最適化された Kubernetes を自動展開する Ansible スクリプトである（図 2）．mdx は高性能な CPU 及び GPU，ROCEv2 による RDMA 通信可能なネットワーク，Lustre 並列ファイルシステムなどスパコンと同等のハードウェアを備えており，Harbor コンテナレジストリ，MetalLB ロードバランサーなどともに Kubernetes 用の最適なプラグインを選択し，同時に導入する．また，内部管理データベースのパスワードも暗号化して管理するなど，セキュリティ面にも配慮している．k8s-configs が存在しない場合には，利用者が Kubernetes 用の個別コンポーネントを順に設定や検証しながら導入していく必要があるが，その負担が大幅に低減されている．

mdx 上に Kubernetes を自動展開するところまでは，今年度前半までに作業を終えており，それ以降はさらに Kubernetes 上にアプリケーションを自動展開する作業を進めている．JHPCN の他の研究課題（jh220051 等）とも連携しつつ，JupyterHub（Zero to JupterHub），Spark，MySQL または PostgreSQL などの展開の検証を進めている．今年度は個別のアプリケーションの導入までの検証を終えており，次年度以降はさらに特定の研究課題の要望に特化した環境構築作業を進めていく予定である．また，機械学習の推論ワークロードを対象に Kubernetes のスケジューリングに関する研究も進めている [5]．

5.2　セキュアデータ流通環境の構築とプライバシー保護

テーマ 2 のセキュアデータ流通環境とプライバシー保護に関しては，mdx の成功にも関わるような非常に大きなテーマを含んでおり，またテーマ 1 の計算環境にも依存している．

今年度は，データ蓄積・解析基盤の事前検討を進めるとともに，実用的かつ非常に効果の高い手法として Multi-threaded scp の研究開発を実施した．その成果について報告する．

また，IoT センサ等からデータをセキュアかつ容易に収集する SINET Stream を利用者が mdx で容易に活用できるようにするための環境整備を進めている．仮想マシンやコンテナイメージの検証環境などのセキュア環境及び仮想ネットワーク環境の構築も検討している．mdx では一部ノードで Intel SGX の利用も試験的に可能となっており，これらの活用についても検討している．

5.2.1 データ蓄積・解析基盤の事前検討

本項目の事前検討として，実際にプライバシ情報を含むデータの蓄積・解析基盤を mdx 上に構築し運用を開始した [4]．本プロジェクトは，東京大学が運用している遠隔会議システム Zoom で計測されたユーザの回線品質 (QoS) データを取得し，東京大学のリモートワーク・遠隔講義の実態を調査する目的で実施されている．本計測対象の Zoom QoS データは，1 日あたり 1 から 2GB 程度のログデータである．また，QoS データは，プライバシ情報を含むことから，匿名化処理を実施した後にデータを蓄積する必要がある．そこで，本プロジェクトでは，プロジェクトごとにネットワークが隔離された mdx 上に，データ取得のためのクローラ仮想マシンと，データ解析マシンをデプロイし，データ取得・処理・蓄積・解析を自動で実施する環境を構築した．本 QoS データは，時系列データであることからファイルによるパーティションが容易であり，そこで，本プロジェクトでは, 日毎で QoS データを分割し, SQLite 形式に変換し保存する方式を採用した．データ蓄積をデータベースサーバではなく，SQLite ファイルによるファイル保存を選択したことにより，データ解析の際にファイルを毎回読み出す必要がある一方で，データ蓄積自体にはサーバが必要ないため，S3 や Lustre ストレージ上での長期データ蓄積が可能である．現在は，外部ストレージと SQLite ファイルの組み合わせで運用しているが，今後は Apache Parquet のような列指向のフォーマットや，圧縮形式に対応したファイルフォーマットについても利用を検討する．本プロジェクトより，mdx 上でプライバシデータを含むデータの蓄積・解析基盤のリファレンス環境を構築した経験を基に，今後 mdx 自体の基盤環境にフィードバックを実施していく．

5.2.2 Multi-threaded scp

mdx は SINET からの L2VPN 接続に対応しており，外部から閉域網経由で既存のネットワークを仮想マシンに接続し，セキュアなデータ流通を実現することができる．一方で，mdx をはじめ IaaS 型クラウド上に構築される計算環境は，ユーザが仮想マシンに直接ログインして利用するケースだけでなく，前節で触れた JupyterHub を前提とした Web からの利用や，Kubernetes によるコンテナ環境など，さまざまな構成がありうる．そのため計算環境の構成によっては，常に SINET L2VPN を用いたセキュアなデータの流通が可能であるとは限らない．そこでネットワークの構成に依らず，アプリケーションのレイヤでセキュアに，高速かつ手軽にデータをやり取りするためのソフトウェアとして，Multi-threaded scp (`mscp`) を開発した [7].

計算機間でファイルを転送する際，最も広く利用されるツールのひとつが `scp` である．`scp` はデータの転送に SSH コネクションを用いるため，マシンへネットワーク越しに SSH ログインさえできればファイルを転送でき，またデータはネットワーク中では暗号化されるため，高い利便性とセキュリティを誇る．しかし `scp` は，暗号化のオーバーヘッドや，TCP であるため高遅延ネットワークではスループットが出ないといった問題が性能上の課題として知

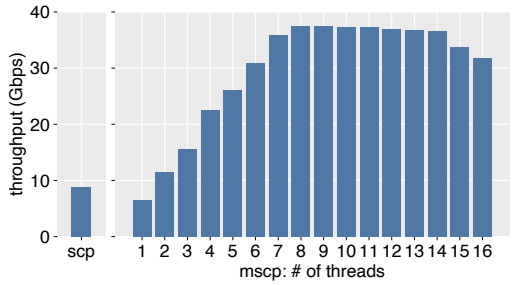

図 4　scp と mscp のスループットの比較

られている．mdx や HPCI の大型計算機群は 10Gbps を超える高速リンクで相互に接続されており，scp ではこうした広帯域ネットワークを十分に利用することができない．

mscp は，SSH 越しにファイルを転送する Secure File Transfer Protocol (SFTP) を活用し，複数の SSH コネクション越しに複数の CPU コアを用いて並列にファイルを転送する．mscp は，SSH できればすぐに使えるという scp の利点はそのままに，広帯域ネットワークでの高速なファイル転送を実現する．そのため他の高速ファイル転送ツールのように，追加のサーバソフトウェアを計算機にインストールして設定する必要はない．つまり mscp は，mdx のようなクラウド環境でも，仮想マシンをデプロイしてすぐさま利用可能な高速かつセキュアなファイル転送ツールである．

本年度は mscp の実装を開始し，最初の性能評価としてローカル環境に用意した 2 台のマシン間でファイルを転送する簡単な計測実験を行った．図 4 は，AMD Ryzen 9 7950X CPU を搭載した 2 台のマシンを 100 Gbps で接続し，メモリ上に配置した 30 GB のファイルを転送して計測したスループットを示している．図 4 が示すように，scp は 8.7 Gbps と 10 Gbps に到達できなかった一方，mscp はコネクション数に応じて 8 コネクションまでスループットが増加し，最終的に 37.5 Gbps と，scp の 4.29 倍のスループットを達成した．次年度移行も mscp の実装を継続し，さらなる性能の向上や広帯域高遅延ネットワークでの評価を実施していく予定である．

mscp のソースコードはオープンソースで公開されており，macOS といくつかの Linux ディストリビューションで利用可能である [10].

5.3　RDC・mdx・大型計算機間でのデータ連携と情報の共有

5.3.1　NII RDC

NII Research Data Cloud（RDC）は，データ検索基盤 CiNii Research，データ公開基盤 JAIRO Cloud，データ管理基盤 GakuNin RDM の三本柱からなる次世代研究データ基盤である．mdx では，NII RDC と関連して（1）GakuNin RDM における mdx ストレージの利用，（2）GakuNin RDM と mdx 計算環境の連携，および（3）JAIRO Cloud ベースのデータ公開リポジトリシステムの開発を行っている．（1）は，mdx が提供する S3 互換ストレージを GakuNin RDM の外部ストレージとしてマウントする機能である．研究者は GakuNin RDM の UI を通じて mdx ストレージにあるファイルを管理・共有できる．（2）は，mdx テナント VM に利用者が構築した JupyterHub を用いて，GakuNin RDM からの操作で mdx 計算環境に Jupyter コンテナを自動構築する機能である．Jupyter コンテナは，GakuNin RDM 側で定義された Python や R のパッケージがあらかじめインストールされ，かつ，GakuNin RDM の標準ストレージにあるファイルがコピーされた状態で起動する．Jupyter 側で得られた解析結果などの出力ファイルはワンクリックで GakuNin RDM へ書き戻すことができる．これにより，GakuNin RDM のプロジェクトに参加する研究者同士がデータ解析プログラムとその実行環境を容易に共有・再利用し，互いの

実験結果を再現しながらスムーズに共同研究を進めることができる．(3) は，mdx をデータ公開リポジトリとして利用するにあたり，データ利用者が使いやすい Web UI を提供する機能である．JAIRO Cloud と同じ基盤ソフトウェア「WEKO3」を採用することで NII RDC との相互運用性を確保しつつ，一般ユーザーがデータを登録・閲覧しやすい直感的な UI を整備する予定である．前述の三本柱を軸として，研究活動を支援する様々な機能が今後 NII RDC に搭載される予定である．mdx と NII RDC をできるだけシームレスに統合することで，研究者がストレスなく mdx の性能を活用できる環境づくりを進めていきたい．

NII RDC に関しては，mdx が提供する S3 互換ストレージを GakuNin RDM の外部ストレージとして利用する機能，GakuNin RDM から mdx 計算機上に Jupyter コンテナを自動構築する機能，WEKO3 を基礎としてデータ公開リポジトリとして利用する三つの機能の研究開発を進めた．

5.3.2 mdx とスパコン連携

mdx とそれ以外の大型計算機システムとの連携に関しては，東京大学，北海道大学，大阪大学の計算機システムの利用を今年度申請していたことから，実計算機での動作検証を実施した．東京大学 Wisteria/BDEC-01 は「計算・データ・学習」融合スーパーコンピュータシステムであり，大阪大学の ONION はスパコン SQUID と連携するデータ集約基盤である．北海道大学ハイパフォーマンスインタークラウドの仮想サーバは，mdx と同様に仮想マシン（KVM）を提供しているため，検証環境に含めた．

公開鍵の登録やログイン方法，コマンドなど各計算機の利用方法が異なるため作業に時間を要したが，検証の結果，学術研究におけるデータの利活用という点に関していずれのシステムも目標が一致しているが，細部の実装方針が異なり，それぞれに高度によく実現されているが，相互システムの連携のためには，アプリケーション研究者による多大な調整や作業が必要であることが分かった．また，今年度は具体的な検証アプリケーションまでは想定していなかったため，各拠点の最低資源量のみを申請していたが，それでもスパコン利用を想定した計算時間トークンを十分に使い切ることができなかった．具体的なアプリケーションを想定した本格的な連携については，次年度以降の課題とする．

6 今年度の進捗状況と今後の展望

既存の基盤ソフトウェアに関する研究成果を mdx に移植する，または mdx 用に新規に研究開発するという初期の目標は達成した．今年度の成果により，データ利活用を行う学際的なアプリケーション研究者が mdx 上で直ちに研究を進められる，あるいは若干の作業で研究を開始できる状態になったと信じている．また，mdx 以外の各拠点の計算システムや研究者の環境とのセキュアな高速転送にも対応した．

本課題は次年度（2023 年度）も継続課題として採択されている．研究期間の半ばから，研究代表者の家庭の事情等により，全体調整が不足し，研究がやや遅れたが，次年度は研究代表者を東京大学の空閑先生に交代し，更なる発展が見込まれる予定である．

7 研究業績一覧（発表予定も含む）

学術論文 （査読あり）

[1] 杉木 章義, 空閑 洋平, 竹房 あつ子, 藤原 一毅, 合田 憲人, 中村 遼, 塙 敏博, 鈴村 豊太郎, 宮本 大輔, 田浦 健次朗, 伊達 進, 建部 修見, “データ活用社会創成

に向けた基盤ソフトウェア環境の構築", 大学 ICT 推進協議会「学術情報処理研究」, 第 26 巻 1 号, pp 1–9, 2022 年 12 月. https://doi.org/10.24669/jacn.26.1_1

国際会議プロシーディングス （査読あり）

国際会議発表 （査読なし）

国内会議発表 （査読なし）

[2] 杉木 章義, “データ利活用に向けた高性能 Kubernetes 環境構築の検討”, SWoPP 2022, 情報処理学会研究報告 (2022-HPC-185(18)), pp. 1–7, 2022 年 7 月.

[3] 大江 和一, 竹房 あつ子, 丹生 智也, 塙 敏博, 工藤 知宏, 合田 憲人, “クラウド環境構築システム VCP の mdx への適用と OSS 化に向けた試作”, 情報処理学会研究報告 2022-OS-156 No. 10, pp. 1–7, 2022 年 7 月.

[4] 空閑 洋平, 中村 遼, “遠隔会議システムの計測データを用いた広域ネットワーク品質計測”, インターネットと運用技術シンポジウム論文集 (IOTS2022), 2022 年 12 月.

[5] 三井 郁央, 杉木 章義, “マルチインスタンス GPU を用いた推論ワークロードのクラスタスケジューリング”, 情報処理学会第 85 回全国大会, 2023 年 3 月.

[6] 大江 和一, 丹生 智也, 竹房 あつ子, 合田 憲人, “クラウド環境構築システム VCP ポータブル版の開発と活用事例の紹介”, AXIES2022 ポスター発表, 2022 年 12 月.

[7] 中村 遼, 空閑 洋平, ”複数コネクションを用いる高速な scp の実装”, 情報処理学会研究報告 (2023-OS-158(19)), pp1–7, 2023 年 2 月.

公開したライブラリ等

[8] A. Sugiki, “k8s-configs: Ansible playbook for building a containerized platform on mdx”. https://github.com/a-sugiki/k8s-configs

[9] 大江 和一, 丹生 智也, 竹房 あつ子, 合田 憲人, “mdx REST Client for Python の GitHub での公開”, https://github.com/nii-gakunin-cloud/mdx-rest-client-python, 2022 年 8 月.

[10] 中村 遼, “mscp: transfer files over multiple ssh (SFTP) connections”, https://github.com/upa/mscp

その他（特許，プレス発表，著書等）