jh250033

SINET を介したデータベース基盤と HPC 基盤の連携による医療画像解析基盤実現に関する研究

課題代表者　村尾 晃平（国立情報学研究所）

概要

本研究の目的は、SINET を介してデータベース基盤と HPC 基盤の連携による医療画像解析に必要なインフラを構築し、実際の医療画像を用いた AI／機械学習の開発を実現することである。JHPCN において重要とされる、HPC 基盤、データベース基盤、ネットワーク基盤の 3 つの基盤を連結してその実現を目指す。

本研究は 2023 年度からの継続であり、医療画像データを収集・格納している国立情報学研究所（NII）の医療画像データベース基盤と、豊富な計算資源を保有する名古屋大学の情報基盤センターの HPC 基盤「不老」を SINET6 経由で連携し、データセキュリティを確保したシステム設計・構築を行ってきている。この基盤連携システム上で AI／機械学習の処理や医学的見地からの大量データの分析を進めながら、2024 年度はさらに HPC 基盤を追加できるように、安全性を考慮したネットワーク構成を見直し、HPC 基盤の連携対象を mdx に広げた。2025 年度はデータベース基盤を拡張し、データベース基盤群と HPC 基盤群の間のみ多対多で連携可能とする環境を構築し、遠隔による遅延効果の測定を行った。また AI 基盤モデルの開発や画像データのクラスタリングなどに活用した。

1. 共同研究に関する情報

共同利用・共同研究を実施している拠点名

名古屋大学 情報基盤センター

mdx I

(1) 課題分野

データ科学・データ利活用課題分野

(2) 参加研究者一覧と役割分担

村尾 晃平：代表、全体統括

森 健策：副代表、高性能計算環境設定、AI 開発

合田 憲人：インフラ設計・セキュリティ

佐藤 真一：画像系 AI 開発

大江 和一：インフラ設計・性能検証

大竹 義人：医療系データ整備・AI 開発

明石 敏昭：医療系データ整備・AI 評価

黒瀬 優介：医療系 AI 開発

崇風 まあぜん：医療系 AI 開発

二宮 洋一郎：医療系データ整備

2. 研究の目的と意義

本研究の目的は、SINET を介してデータベース基盤と HPC 基盤の連携による医療画像解析に必要なインフラを構築し、実際の医療画像を用いた AI／機械学習の開発を実現することである。JHPCN において重要とされる、HPC 基盤、データベース基盤、ネットワーク基盤の 3 つの基盤を連結し、AI/機械学習のための医療画像解析基盤の実現を目指す。

本研究の意義は、使用者や使用目的が限定されるような機微データを保持するデータ基盤に対し、遠隔の計算資源をセキュアかつ計算効率を損なうことなく連携できる実例

となることである。

3. 当拠点の公募型共同研究として実施した意義

本研究では、CT 画像や MRI 画像といった医療現場で撮影され診断治療等に利用される画像データを収集して格納を進めている国立情報学研究所（単に NII と記す）医療ビッグデータ研究センターの医療画像データベース基盤（以下 MIDB 基盤と記す）と外部の HPC 基盤とを連携方法について研究を行うものである。HPC 基盤とデータ基盤がそれぞれ異なる機関に整備されている状況下において、これらの基盤を SINET 経由で連携して運用するここで、医療画像を対象とした人工知能(AI)の学習処理を行うための諸問題の解決を目指す。

NII 医療ビッグデータ研究センターMIDB 基盤には約 9 億枚の医療画像が格納されており、その多くは CT 画像や MR 画像である。画像のみならず、画像に付随する所見文や疾患名などの情報も保存されており、悉皆的な大規模データ基盤となっている。現在も日々データ収集と蓄積を続けており、毎日 20 万枚から 30 万枚の画像データを受信している。これは本 MIDB 基盤の特徴ともいえる。

これらのデータの扱いについては、医療系学会の倫理審査委員会で、使用者や使用目的が定められている。利用者による誤操作などによるデータ漏洩なども当然防ぐ必要がある。そのためには、データ基盤側と HPC 基盤側との間でセキュリティ対策などおいて強い連携が必要となる。名古屋大学の情報基盤センターは、昨年度までの共同研究の実績で基盤連携の準備が整っている。また、医療画像データの扱いと解析に対して経験豊富であることも共同研究で実施する理由である。

4. 前年度までに得られた研究成果の概要

2023 年度には、データ基盤として NII の医療画像ビッグデータ MIDB 基盤、HPC 基盤として名古屋大学情報基盤センターの不老を設定し、これらをセキュアに連携するインフラを構築した（図 1 の mdx 連携以外の部分）。そのポイントとしては、連携に SINET の L2VPN を用いたこと、不老から専用の GPU を 2 ノード切り出したこと、専用の GW サーバを設け、計算時のみに sshfs 接続するようオンデマンド sshfs の枠組みを作ったことが挙げられる。GPU の切り出しにあたっては、ジョブ投入できるユーザを限定し、専用の GW につながるネットワークを該当 GPU ノードに限定した。この環境で実験した結果、遠隔に起因する遅延はあるものの、AI 学習においてはデータがバッチ単位で読み込まれるため通信トラフィックは間欠的であり、並列計算に対しての耐性もあることがわかった。

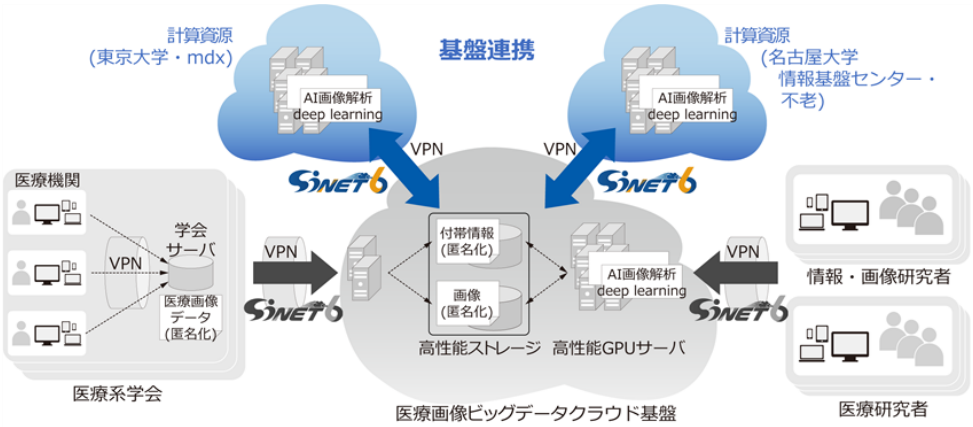

図 1 データ基盤と HPC 基盤の連携概念図

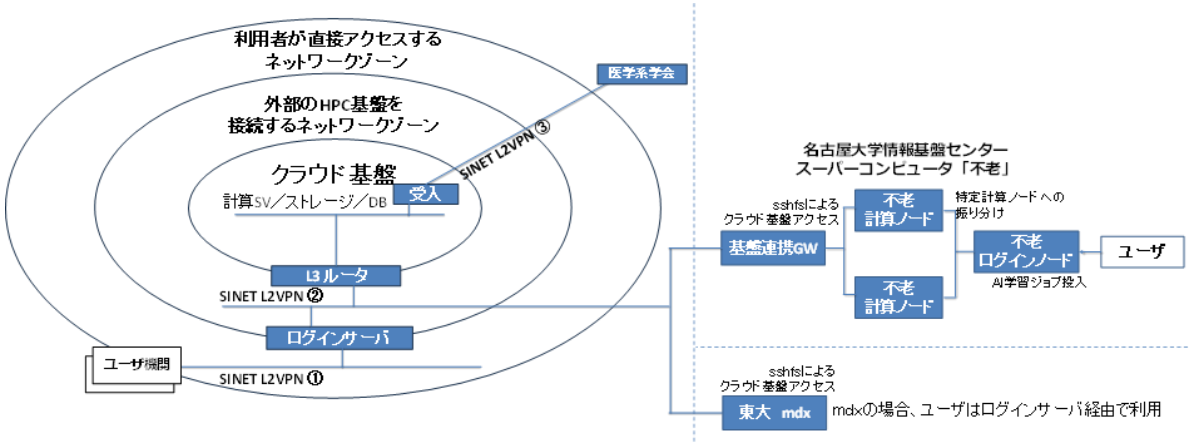

図 2　基盤連携におけるネットワーク概念図

2024 年度は、外部 GPU 連携による強化を行った。ここでのポイントは、MIDB 基盤のログインサーバの別立てと SINET L2VPN によるネットワーク緩衝地帯を設けることによるセキュリティ強化であった(図 2)。これにより、ログインサーバやバッチ投入の仕組みのない mdx などからも利用が可能となった。

5.　今年度の研究成果の詳細

①多対多の分散型を想定した実験

これまで、一か所のデータ基盤に対して複数の HPC 基盤を接続する実験をおこなってきたが、今後のマルチモーダルな AI 開発にあたっては、複数のデータ基盤からデータを得て計算する場合も考えられる。そこで、基盤どうしが互いに多対多の状況で AI 開発の実験を行い、インフラとして対策すべき課題を見出す。

このような分散型のインフラの設計にあたって、まず、ネットワークとしては、セキュリティと運用性の観点から、データ基盤毎に独立した VLAN で接続することになる。その上で、大きく次の２つの方式が考えられる。

１）HPC 基盤の側に新たなゲートウエイ・サーバ（GW）を導入し、そこに接続先ごとに同一アドレスで異なるポート番号を割当てる（ポート方式）

２）HPC 基盤のサーバ側に VLAN ごとに異なる IP を付与する（IP 方式）

図 3 に示すように、ポート方式では共通の IP アドレスを使ってポート番号で振り分ける。IP 方式では HPC 基盤そのもののネットワーク接続を DB 基盤を増やすたびに加える。

この 2 つの方式に対してセキュリティ（監査性）と運用面での差異を表 1 のように比較した。

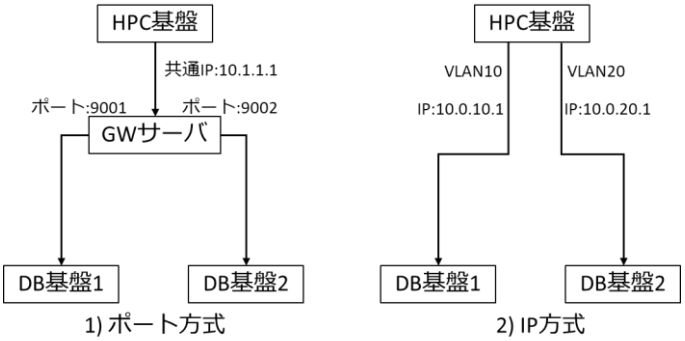

図 3　データ基盤への接続方式比較

表 1　複数のデータ基盤への接続方式の比較

観点	ポート方式	IP 方式
監査性	IP とポート番号の組合せを見る必要あり。	IP で明確に追跡可能。設定のヒューマンエラーを検知しやすい。
運用性	HPC 基盤の運用が大規模でセンターで行われていると、GW 設定が外出しできるので NAT 設定を実装・運用しやすい。	サーバに VLAN サブインターフェースを追加する必要があり、ACL の見直しなどネットワークチームとの調整が必要で固定運用向き。

名古屋大学の不老の場合は、GPU ノードは情報基盤センターとしての運用があるため、独立した GW を導入してポート方式にすることで、素早く実装できた。一方、東大の mdx の場合はプロジェクトの管理者が仮想マシンをデプロイすることができるので、監査性の良い IP 方式を実装することが容易にできた。

②データ基盤のフラッシュディスク化による効果実験

画像データの場合は数百キロバイトのサイズの小さいものから数ギガバイトの大きなものと種々のファイルサイズのものまである。データ基盤側でのファイルストレージの物理的な違いが、様々なサイズのデータへのアクセスが必要となる AI 学習の速度にどのように影響するかを調査する。具体的には図 4 のような実験環境を使った。

データ基盤としては、NIIMIDB 基盤と Dell のサーバ PowerEdgeT640 に Synology のフラッシュディスク FS3410 を接続した環境を第 2 の仮想的基盤とした。以降では前者を単に MIDB 基盤、後者を Synology と略して表記する。

MIDB 基盤のディスクは 7,200rpm のニアラインディスクであるが、RAID6 で構成されており、読込が並列に行われるので読込速度性能を重視している。Synology は SSD を使っているのでニアラインディスクよりも速い読込が期待できるが、耐障害性と寿命を重視して RAID-F1 で構成している。この結果、MIDB 基盤と Synology でどちらが読み書き性能を出せるかは測定してみないとわからない。

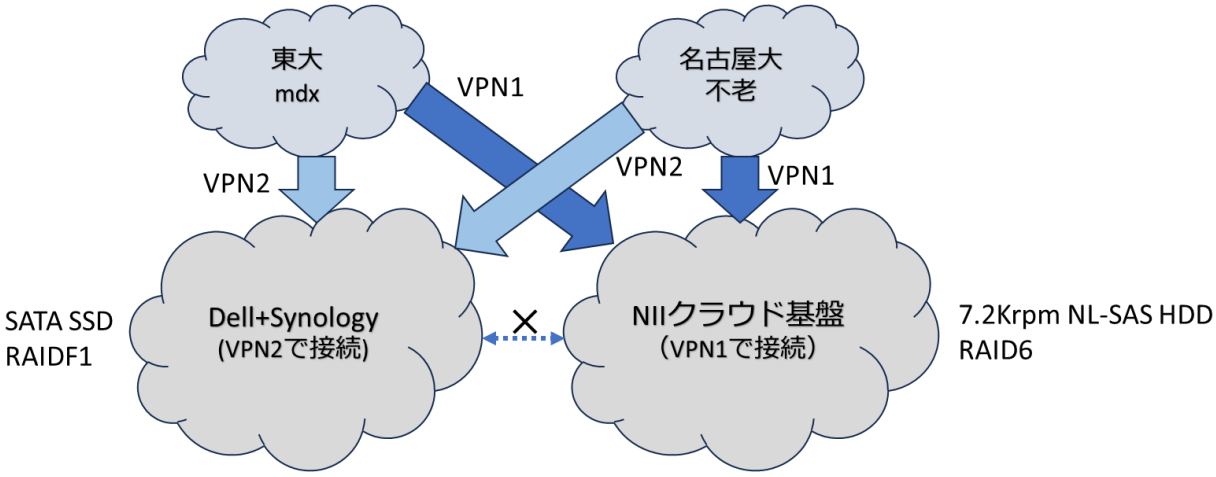

図 4　実験環境の概要

そこで、fio コマンドを使ってネットワーク越しの読み書き性能を測定した。その際の設定ポイントは次の 7 つである。

1. シーケンシャル読み書きで 4 種類のブロックサイズ（4k, 64k, 1M, 4M bytes）
2. ランダム読み書き（ブロックサイズは 4k bytes 固定）
3. テスト用ファイルサイズ＝10G bytes
4. 並列数(numjobs)＝4
5. Linux の非同期 I/O を使う （ioengine = libio), (iodepth を使用しない)
6. OS キャッシュをバイパスして直接 I/O を使う (direct = 1)
7. 複数ジョブの結果をまとめて表示する (group_reporting)

名古屋大学の不老から基盤に対して読み書きのスループットを計測した結果を図 5 に示す。シーケンシャルの読み書きについて、読み込み（図中の SEQ_READ）については Synology と MIDB 基盤でほとんど差が無いが、書き込み（図中の SEQ_WRITE）については Synology の方が僅かに速い。ランダムの読み書き（図中の RAND_READ と RAND_WRITE）については Synology への書き込みが MIDB 基盤より 2 倍以上速い。SSD の効果は書き込みで効果が出ることがわかった。またブロックサイズは 1MB 以上で性能が発揮できることがわかる。

同様の実験を東大の mdx から行った結果を図 6 に示す。図 5 と比較して、mdx の方が不老よりも物理距離が圧倒的に近いのでスループットが全体に高く出ていることがわかる。シーケンシャルの読み書きでは MIDB 基盤の方が Synology よりもスループットが良好な結果となった。ランダムアクセスでは図 5 と同様に書込みで Synology の方が 2 倍以上速いスループットが出た。ここでもブロックサイズが 1MB 以上で性能発揮できることがわかった。

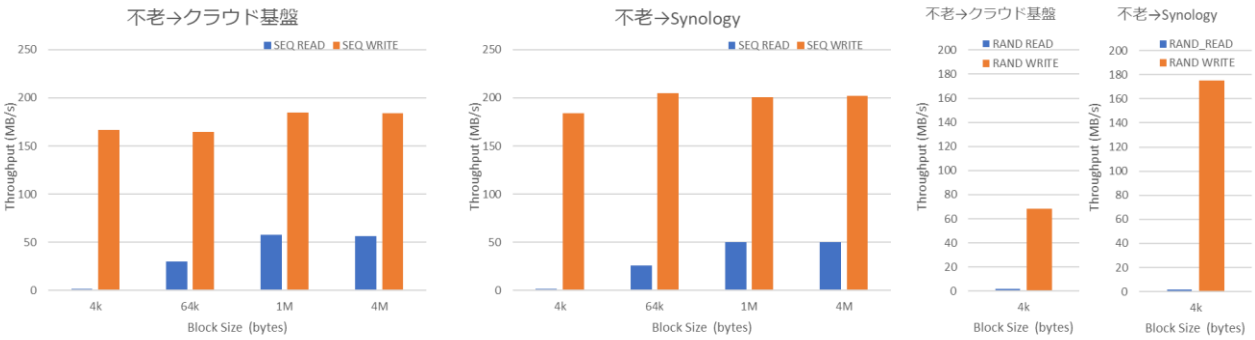

図 5　不老から 2 つのデータ基盤へのスループット計測

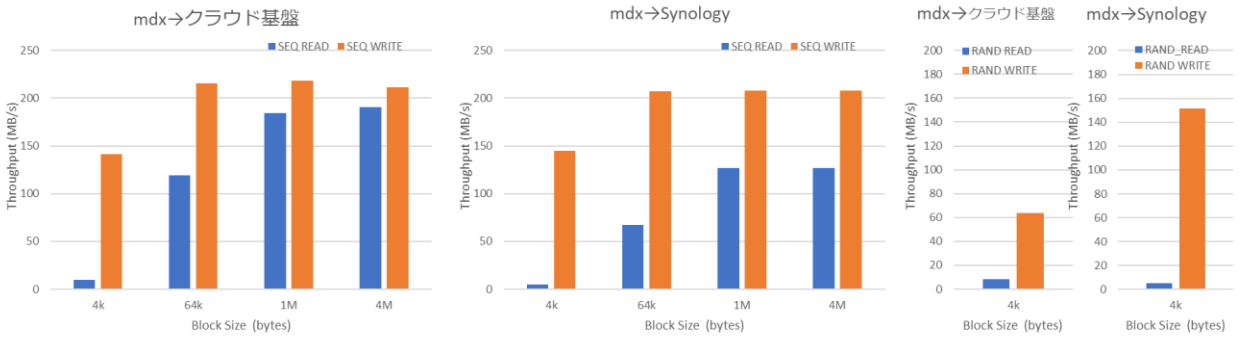

図 6　mdx から 2 つのデータ基盤へのスループット計測

③物理的距離による遅延効果の定量化

図 5 と図 6 を比較すると、シーケンシャルの READ において不老の方が mdx よりも物理的距離による遅延効果が大きく表れている。これは、HPC 基盤側からの一連の READ リクエスト・ディスク読み出し・受信確認(ACK)が直列に行われるためで、WRITE では次の ACK を待たずにデータを送り出せるから（パイプラインが有効）であると考えられる。

ここで、距離による遅延を理論的に考えてみる。不老と MIDB 基盤の回線距離を約 300km と仮定する。光の速度は約 30 万 km/s なので、片道の情報伝達には 1ms、往復で 2ms となるはずである。しかし、光ファイバーの屈折率は波長 1.5μm の光に対して約 1.44 なので、光の伝わる速度は 1/1.44=0.69 で約 70%に落ちる。これを考慮に入れると片道 1.4ms、往復で 2.8ms となる。途中にルータがいくつかあるが、その遅延は数十μ秒程度で無視できると思われる。シーケンシャル READ の場合は、これに加えて OS の処理待ち分が上乗せされるので追加の遅延が発生し、シーケンシャル WRITE の場合は、パイプライン効果で 1 回の I/O 当たりの見かけの往復時間が 2.8ms より少なくなることが考えられる。

そこで、mdx 上で、仮想サーバを占有していることを活用し、tc コマンドを利用して遅延を疑似的に作り出して遅延効果を見ることにした。tc コマンドで送信側（Exgress）に徐々に遅延を入れていくと、遅延 5ms の時に不老における SEQ_READ に近くなった。受信側（Ingress）に遅延を入れるには仮想的なポートを設定してそこに転送して遅延を発生させるテクニックが必要ではあるが、2ms の遅延を加えた。これにより、図 7 に示すように、送信側遅延 5ms、受信側遅延 2ms で不老の状況が再現できた。

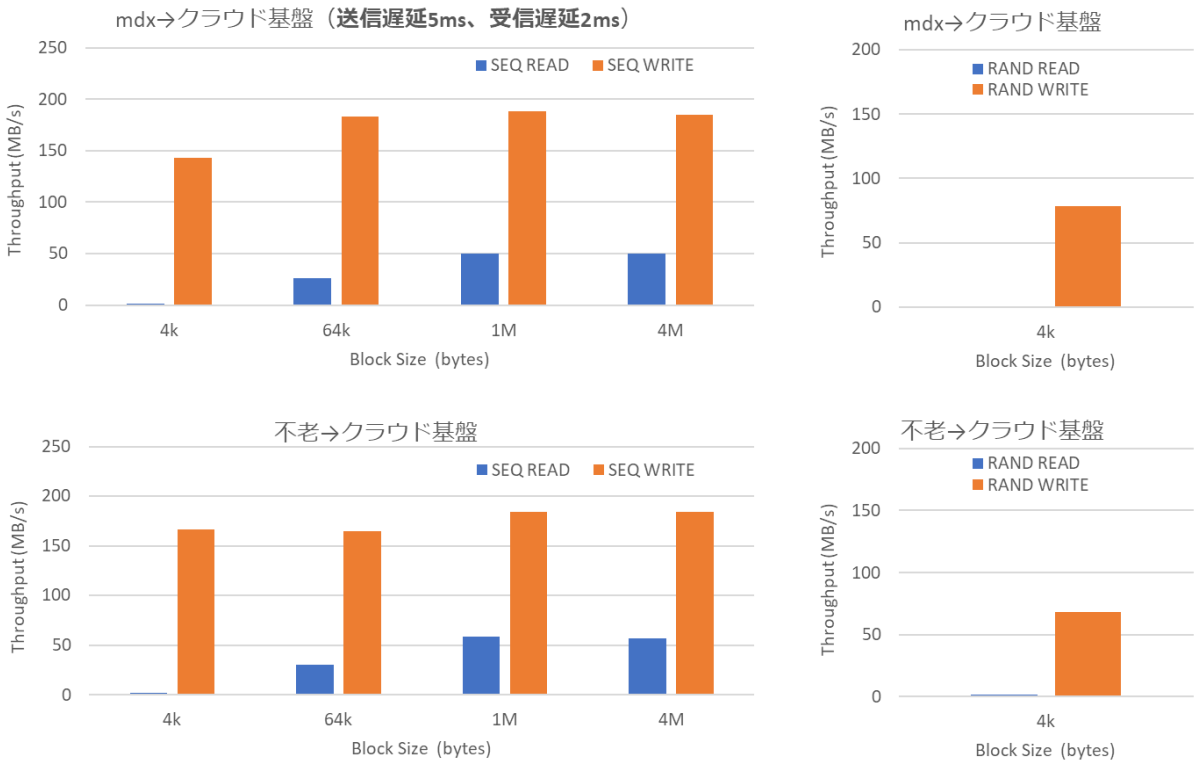

図 7　mdx 上で遅延を入れて不老上のスループットを再現

④AI 開発におけるスループット実験

不老および mdx を使って同じ AI 画像処理を行って性能比較をした。日本眼科学会が収集した約 150 万枚（容量約 3.5TB）の画像には、標準的な眼底画像やステレオ画像、画像を貼り合わせたパノラマ画像など様々な種類の画像が混在している。ここから標準的な眼底画像を選出するために、AI モデル DINO-ViT を使って特徴量を抽出し、次元削減 UMAP とクラスタリング K-means を利用し、クラスタを選出していく方法をとった。最初の 1 万件で不老では 17 分 46 秒、mdx では 11 分 20 秒であり、その差は約 6 分であった。これを 150 万件に広げるとその差は約 15 時間（42 時間 vs. 27 時間）であった。このようにファイル数が多い場合は、ランダムアクセスになるので遠距離による遅延が大きくなる。初年度に報告したように CT 画像の場合はシリーズで 1 ファイルにするような対策がとれたが、今回のように 1 ファイルずつが別の意味を持つ場合には、何等かのデータのパック化、並列処理化を導入して、計算効率を高める可能性が示唆された。

6. 進捗状況の自己評価と今後の展望

この基盤連携のプロジェクトは JHPCN としては今年度で終了するが、ここで作った連携の仕組みの一部は残り、基盤モデル開発やマルチモーダル AI 開発、時系列解析などで利用されると見込んでいる。

また、基盤連携全体のセキュリティの考え方、実現方法についてはノウハウとして残って将来の基盤連携開発に生かされるであろう。

※7. 研究業績はウェブ入力です