jh231008

Tomo-e Gozen データプラットフォームの開発

瀧田 怜（東京大学）

概要

東京大学木曽観測所では、時間軸天文学の未開拓領域である「秒」の時間スケールの変動現象を捉えるために 105 cm シュミット望遠鏡に 84 枚の CMOS センサからなる Tomo-e Gozen カメラを搭載し、毎晩広視野動画サーベイを行っている。これは世界でも類を見ないユニークなデータセットであるが、データセンターを持たない観測所で日々増え続ける膨大なデータをアーカイブし続けることは困難である。この観測データ等を持続的にアーカイブし、また迅速に世界へ発信するデータプラットフォームを mdx 上に構築することが目標である。このために、我々は木曽観測所を SINET に接続することで、リアルタイムに観測データを mdx に転送できることを確認した。また転送されたデータを使い、簡易的なデータアーカイブも作成した。今後はこのアーカイブ機能の強化や、「秒」スケールの動画データを生かしたデータベースの開発を引き続き行う予定である。

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

- 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)

1.4 参加研究者の役割分担

- 瀧田 怜: データプラットフォームの開発
- 酒向 重行: 木曽観測所、および Tomo-e Gozen プロジェクトとの連携、研究統括

2 研究の目的と意義

東京大学木曽観測所が進める Tomo-e Gozen 計画では、時間軸天文学の未開拓領域である「秒」の時間スケールの変動現象を捉えるために、広視野動画サーベイを行っている。さらにはデータをリアルタイムで解析し、その中に潜む突発現象や高速移動現象を検出して、情報を迅速に世界へ発信することが目標である。

本研究の意義は、我が国が生成するユニークな天文学データセットである Tomo-e Gozen の観測データを持続的にアーカイブし、その観測データおよび解析済みデータを迅速に世界へ発信するデータプラットフォームを構築することにある。このような機能は山岳部に位置する木曽観測所のようなオンサイトの施設では運用が困難であるが、SINET6 と mdx の組み合わせにより打開できる。さらに mdx の豊富な計算機リソースを利用することで、Tomo-e Gozen の膨大なデータをリアルタイムに解析可能となる。海外の ZTF, Pan-STARRS, LSST などの代表的な時間軸サーベイ計画では、各国のデータセンターが、取得する観測データを管理、解析、配信する役割を担っている。最も短い「秒」の時間スケールをカバーす

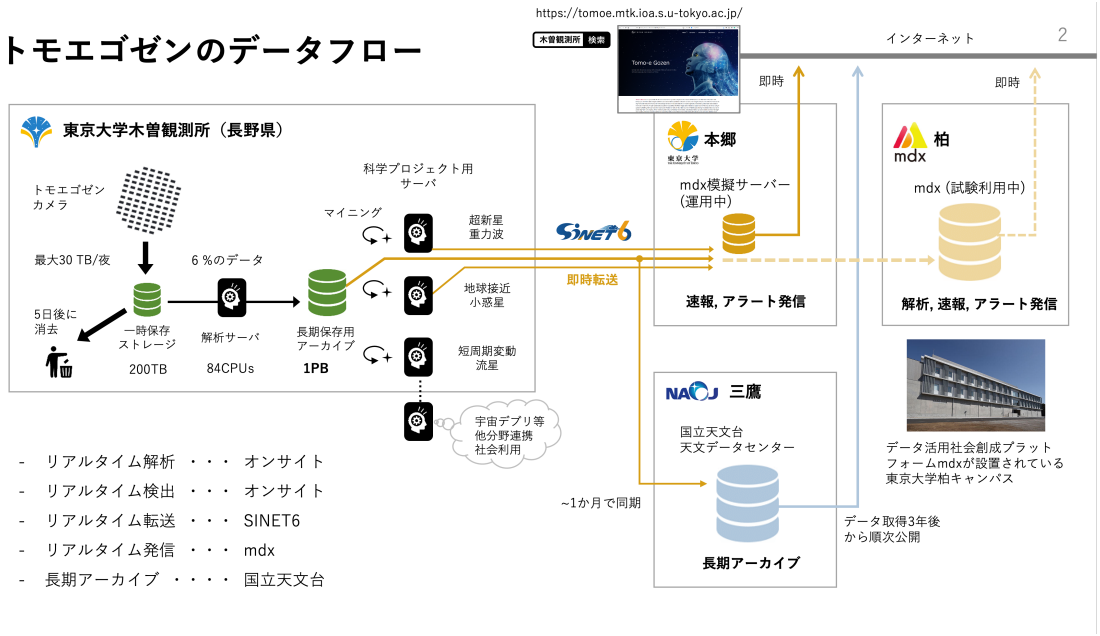

図 1　トモエゴゼンのデータフロー。(PC クラスタワークショップ in 大阪 2023「ビッグデータと HPC」, 2023.6.21, 酒向の発表資料より抜粋)

る Tomo-e Gozen のデータを、日本を代表するデータセンター (SINET+mdx) が世界へ配信する意義は大きい (図 1)。

Tomo-e Gozen 計画では「トモエゴゼン スカイアトラス」による教育普及活動や、スペースデブリ等の民間活用計画も進めており、これらのプラットフォームとしての役割も期待されている。

3　当拠点公募型研究として実施した意義

Tomo-e Gozen が生成する年間 2 PB に及ぶ膨大な観測データの全てを、データセンター設備を持たない東京大学木曽観測所のオンサイトに保存することは非現実的である。そのため、全観測データに対しワンスキャンで時間変動現象を探索した後、観測データを 10 % 程度のサイズに非可逆圧縮 (動画を時間方向に圧縮、天体の周辺のみをトリミングなど) してオンサイトでアーカイブしている。それでも年間で 200 TB 弱のペースでデータは増え続けており、今後数年で限界を迎える見込みである。また、解析が不十分なためワンスキャンでは見逃される変動現象も少なくない。これらの問題に対して我々は、木曽観測所を SINET に接続し、Tomo-e Gozen のデータアーカイブシステムを mdx 上に移行する計画を 2019 年より進めてきた。2022 年に SINET6 が近郊の都市 (松本市) に延長されたこと、官学連携の枠組みで地元自治体 (木曽広域連合) が管理する 10Gbps ファイバー網を専用線として利用できることになったこと、そして外部資金 (科研費) が得られたことにより、2023 年 には木曽観測所が無事に SINET6 に接続された。これにより、Tomo-e Gozen の観測データをほぼリアルタイムで mdx へ転送することが可能となった。さらに mdx 上に Tomo-e Gozen のデータ公開のプラットフォームを構築することで、

宇宙の変動現象を持続的に世界へ発信でき、このユニークなデータセットを基に科学的成果の創出が期待される。

4 前年度までに得られた研究成果の概要

該当しない。

5 今年度の研究成果の詳細

本研究課題の目的は (1) Tomo-e Gozen のデータプラットフォームを mdx に構築することと、(2) Tomo-e Gozen による動画観測を利用した「秒」スケールの測光データベースの開発である。2023 年度は主に木曽観測所と SINET (mdx) との接続に関する部分の開発・検討を行った。

5.1 木曽観測所と SINET の接続

2023 年 4 月に木曽観測所と SINET (松本 DC) との接続工事は完了した。しかし接続直後から上り回線のみにパケットロスが発生し、数 10 Mbps 程度の速度しか達成できていなかった。5 月に塩尻–松本間のファイバー研磨を行ったものの目立った効果は得られなかった。しかしその数日後から、一時的に 8 Gbps 程度の転送速度を達成するようになった。ただし安定的な運用には至らず、通信が高速な期間と低速な期間が数週間単位で切り替わる状態であった (6 ヶ月のうち安定して 1 Gbps 以上で通信できたのは 2 ヶ月程度)。東大情報基盤センターや理学系情報システムチームとも相談しつつ調査をしたところ、9 月に通信速度が著しく低下したタイミングで松本 DC の SINET ルータ側でエラーを検知していることが分かった。さらに、過去の速度低下のタイミングとエラーの出現が同期していることから、SINET への接続工事区間である塩尻–松本間の機器に問題があることが特定された。その後 10 月に行われた調査で、直近の工事箇所である、松本 DC 内に設置した機器の光モジュールに問題があることが確認され、これを交換した結果通信エラーが解消された。問題解消後は安定して数 Gbps で通信ができている。

5.2 木曽観測所から mdx へのデータ転送

上記の問題解決に向けた作業と並行して、木曽観測所から mdx へのデータ転送の枠組みも作り始めた。まず木曽観測所内で Tomo-e Gozen 用に整備されたローカルの 10 Gbps ネットワークを SINET に載せ変えた。SINET L2VPN により木曽観測所と本郷キャンパスとを接続し、さらに理学系のホスティングを利用することで東大のネットワークとして運用を行っている。観測所の計算機群が SINET に繋がったことで、mdx からは東大のファイアウォールを通して接続が可能となった (図 2)。

Tomo-e Gozen で取得されたデータは、直結された計算機群によって即座に簡易解析が行われる。この計算機群の間では Redis によって解析の進捗が管理されている。そこでこの Redis の通知を mdx 側でも受信することにし、解析完了の通知を受けて該当データを取得することにした。当初は前述した通信障害があったことで、一晩の解析済みデータの転送が完了するのにほぼ一日を費やす状況であった。その後 10 月に通信障害が解消されたことで、リアルタイムにデータの転送ができることを確認した。

5.3 Tomo-e Gozen データプラットフォームの開発

木曽観測所内では、観測システムに連動する形でデータアーカイブが稼働している。今回 mdx 上にデータプラットフォームを開発するにあたって、アーカイブするデータファイルのパス構造が変わることも有り、新たにデータベースを立ち上げることにした。検索用の観測

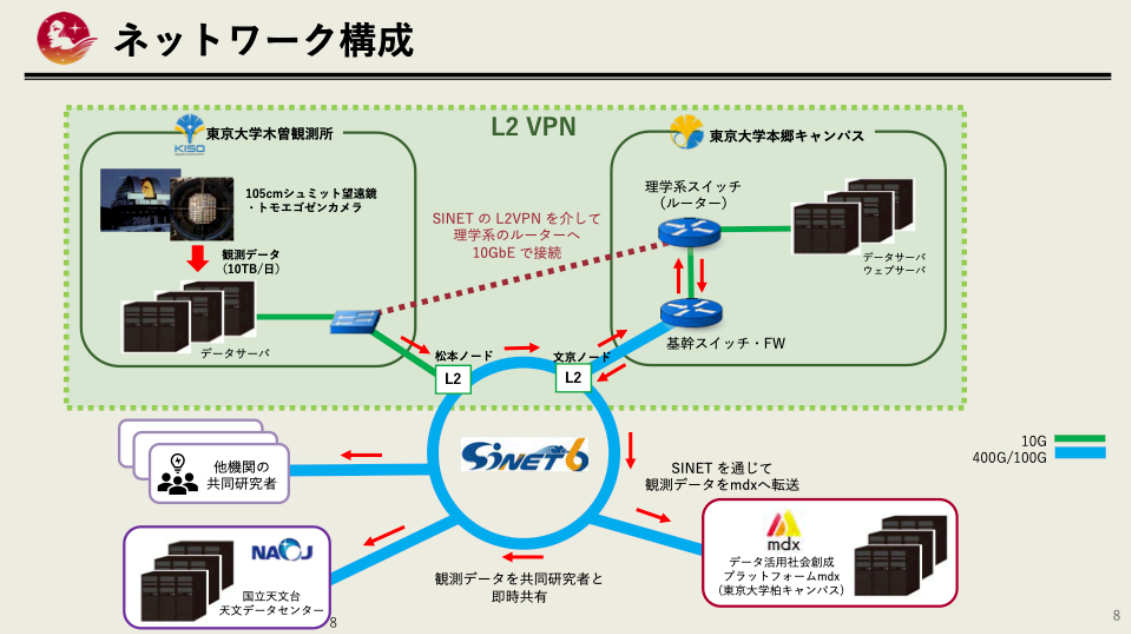

図 2　木曽観測所と SINET のネットワーク構成図。(2024 年天文学会春季年会「Tomo-e Gozen 高速データ転送のための木曽観測所からの SINET 接続」, 2024.3.14, 森の発表資料より抜粋)

ログとしては毎日一度だけ追記する形で良く、また検索性能を高めるためには列指向型のデータベースが適していると考えた。さらに、将来的に他の計算機へ移植する際の利便性を考慮して Apache Arrow 形式のデータベースを利用することにした。観測データから、時刻や望遠鏡の向いている方角、実際に得られた画像の座標等の情報を取得し、それらを観測日ごとの parquet ファイルとして出力した。次にこの parquet ファイルを元にデータアーカイブ機能を作成した。木曽観測所内では、データベースとして PostgreSQL を利用していたので、この部分を pyarrow (duckdb) を利用する形に置き換えている。

5.4　測光データベースの開発

Tomo-e Gozen では他の大型望遠鏡では得られない、「秒」の時間スケールでの測光データを取得できる (図 3)。このユニークなデータセットを活用するために測光データベースを新たに作成する。このためには、既存の高速移動天体検出システムを流用する。このシステムでは 2 fps で取得された各フレームに対して点源検出や測光を行うが、そのデータは in-memory で利用され、移動天体に関する情報のみが残される。この in-memory にしか存在しない情報を出力することが出発点である。

まずは試験的に一日分のデータを作成し、この測光結果をもとに、約 9 秒間の時系列の測光データ (ライトカーブ) のデータベースを作成した。各フレームについて、点源検出と測光が行われているため、フレーム間のデータのクロスマッチには kd-tree を用いてライトカーブを作成する。次にこの Tomo-e Gozen のライトカーブを、既知の天体カタログとをクロスマッチさせることで天体情報を取得する。ここでは現在最も位置精度が高い Gaia カタログ (Data

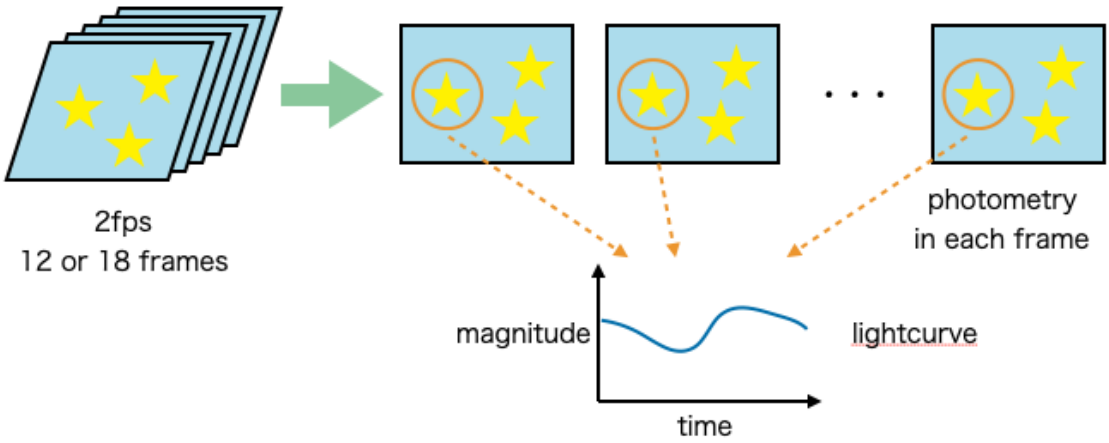

図 3　Tomo-e Gozen の動画データから作成するライトカーブ。

Release 3) を利用した。また高速移動天体検出パイプラインが出力する測光データはキャリブレーションがされていないので、これを補正するために Gaia カタログの測光値との比較を行った。しかし現状の測光値では正しく補正することができないことが判明し、これに対応するために測光方法を変えたデータセットを作成する必要があることが判明した。これは高速移動天体検出チームと連携して行う。

6　今年度の進捗状況と今後の展望

2023 年度は木曽観測所の SINET への接続がなされたものの、十分な速度が出ないという問題を抱えていたため、その解消を待つ必要があった。この問題が解決してからは、Tomo-e Gozen で取得したデータの転送は順調に進んでおり、SINET による高速通信を十分にいかすことが出来ている。一方で現在のペースでデータの蓄積が進むと、mdx ポイントの消費に伴い早々にデータが溢れることが確実である。この問題については HPCI 共有ストレージ (共有型) 利用研究課題を活用し、取得から一定期間が経過したデータを HPCI 共有ストレージに移動することを検討している。

今年度はこれらのデータ転送とアーカイブ機能の充実を目指し、まずは Tomo-e Gozen チーム内へのデータ公開を始める。特に mdx 上のストレージと HPCI 共有ストレージをシームレスに繋いで利用出来るようにする必要がある。

また「スカイアトラス」を利用した教育利用については 2024 年度より科研費も獲得しており (代表: 名古屋市科学館学芸員 毛利氏)、名古屋市科学館との協力の元に進める予定である。

7　研究業績一覧（発表予定も含む）

学術論文　（査読あり）
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）

- “Tomo-e Gozen と mdx”, 瀧田 他, 木曽シュミットシンポジウム 2023, 2023/5/30–31
- “激動の宇宙の姿をもとめて – 木曽トモエゴゼンとビッグデータ天文学”, 酒向, PC クラスタワークショップ in 大阪 2023,

2023/6/21–22

- “Tomo-e Gozen による広域動画サーベイ”, 瀧田 他, プラネタリウムで俯瞰する多波長全天/広域サーベイ, 2023/7/19–20
- “Tomo-e Gozen 高速データ転送のための木曽観測所からの SINET 接続”, 森, 酒向, 瀧田 他, 日本天文学会春季年会, 2024/3/11–15

公開したライブラリ等

その他（特許，プレス発表，著書等）