Japan Geoscience Union Meeting 2014
(28 April - 02 May 2014 at Pacifico YOKOHAMA, Kanagawa, Japan)

MGI37-P01 会場:3 階ポスター会場 時間:4 月 29 日 18:15-19:30

NICT サイエンスクラウドを用いたシミュレーションデータと衛星観測データの大規模統計解析環境の構築
Construction of the Large-Scale Statistical Analysis Environment of the STP field data based on the NICT Science Cloud

山本 和憲[1*] ; 長妻 努[1] ; 久保田 康文[1] ; 村田 健史[1] ; 亘 慎一[1] ; 建部 修見[2] ; 田中 昌宏[2] ; 木村 映善[3]
YAMAMOTO, Kazunori[1*] ; NAGATSUMA, Tsutomu[1] ; KUBOTA, Yasubumi[1] ; MURATA, Ken T.[1] ; WATARI, Shinichi[1] ; TATEBE, Osamu[2] ; TANAKA, Masahiro[2] ; KIMURA, Eizen[3]

[1] 情報通信研究機構, [2] 筑波大学, [3] 愛媛大学
[1]National Institute of Information and Communications Technology, [2]University of Tsukuba, [3]Ehime University

太陽地球系物理分野では、地球磁気圏ダイナミクスを解明するための主な研究手法として、衛星観測と計算機シミュレーションが確立されている。衛星観測では国際太陽地球系物理観測（ISTP）計画により、複数衛星による多地点観測データが蓄積されてきた。計算機シミュレーションでは、計算機技術の向上により、データの３次元化、大規模化、高精度化が顕著である。両者のデータは年々増大化している。

これに対して、これまでのデータ解析・可視化は、イベント期間の特定の観測データを使用したり、シミュレーションデータの３次元メッシュ構造の特定断面のみを可視化したりするなど、蓄積・生成されたデータ資源を十分に活用できているとは言えない。

両者のデータの性質は相補的な関係であり、データ同化に向けた相関関係を調べる取り組みは、データの信頼性と現象解明の可能性を増す。このためには、過去に蓄積されたデータを活用し、衛星観測データでは長期間の複数衛星を用いることで空間領域に対して、シミュレーションデータでは高時間分解能で計算することで時間軸に対して、信頼性を高める必要がある。

一方で、両データのアーカイブ、解析・可視化は独立した環境で行われてきた。そのため、データフォーマットおよび時間軸・座標系が統一されておらず、過去に蓄積した大量のデータファイルに対して統合的に可視化・解析することは容易ではない。

本研究では、NICT サイエンスクラウド [1]（以下、サイエンスクラウド）が提供している計算機リソースおよびデータ処理システムをマッシュアップし、両データのデータフォーマットを統一したデータセットを作成する環境を構築した。データセットを作成するプロセスは、データアーカイブ、時間軸・座標系の統一、物理量の抽出、両データのマージの４つからなる。

データアーカイブ

分散管理された衛星観測データの収集には、NiCTy+Download Agent を利用することにより、分散管理情報をメタデータベース化した STARSDB を参照して定常的にデータを収集できる。また、サイエンスクラウドではスパコンをハウジングできる環境を提供しており、スパコンから出力された大規模データは、基幹ネットワークが 10GbE で接続された分散ストレージ（Gfarm[2]）に出力することができる。

時間軸・座標系の統一

衛星観測データの時間分解能と座標系をシミュレーションデータに合わせる。サイエンスクラウドが提供する SEDOC を利用することで、衛星観測データのデータフォーマットの差異や時間単位で出力されたファイル群を意識することなく、指定した時間分解能にサンプリングして配列に格納される。また、衛星軌道データについては、主な座標系に変換することができる。

物理量の抽出

座標系が統一された衛星軌道の任意の座標値に該当する、シミュレーションデータの物理量を抽出する。サイエンスクラウドが提供する V_Aurora は、衛星観測データとシミュレーションデータの融合表示および指定した座標値の物理量の抽出が可能である。

両データのマージ

時間軸・座標系が統一された両データを 1 つのファイルにマージする。

データアーカイブされた大量のデータファイルに対して、上記の一連の処理を行った場合、データ読み込み時のディスク I/O およびデータ補間時の CPU の処理性能がボトルネックになる。本研究では、この問題を解決するため Pwrake[3] によるワークフローを作成した。Pwrake は分散配置されたファイルがあるノード上でプロセスを起動するようにジョブ管理するため、ローカルディスク I/O を活用したデータ処理が行える。

本発表では、過去に蓄積した LANL シリーズの衛星データと Global MHD シミュレーションデータを用いて、両データのデータセットの作成および相関図を作成した結果について発表する。

参考文献

[1] 村田健史, サイエンスクラウドは第四の研究基盤となるか？ , 学術の動向, Vol. 17, No. 6, pp. 42-47, 2012.

[2] 建部 修見，森田 洋平，松岡 聡，関口 智嗣，曽田哲之，ペタバイトスケールデータインテンシブコンピューティングのための Grid Datafarm アーキテクチャ，情報処理学会論文誌：ハイパフォーマンスコンピューティングシステム, Vol.43, No. SIG6 (HPS 5), pp.184-195 , 2002.

[3] Masahiro Tanaka, Osamu Tatebe, Workflow Scheduling to Minimize Data Movement using Multi-constraint Graph Partitioning, The 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid2012), pp.65-72, Ottawa, Canada, May 13-16, 2012.

キーワード: 計算機シミュレーションデータ, 衛星観測データ, 並列分散処理, Gfarm, Pwrake, NICT サイエンスクラウド
Keywords: computer simulation data, satellite observation data, parallel distributed processing, Gfarm, Pwrake, NICT Science Cloud