Grid Datafarmによる太陽地球系物理観測データの大規模統計解析システム

A large-scale statistical analysis system for satellite and ground-based observations via Grid Datafarm

山本 和憲 [1]; # 村田 健史 [2]; 木村 映善 [3]; 建部 修見 [4]; 海老原 祐輔 [5]; 上野 玄太 [6]; 北本 朝展 [7]
Kazunori Yamamoto[1]; # Takeshi Murata[2]; Eizen Kimura[3]; Osamu Tatebe[4]; Yusuke Ebihara[5]; Genta Ueno[6]; Asanobu Kitamoto[7]

[1] 愛媛大・理工; [2] 愛大・メディアセンター; [3] 愛媛大 CITE; [4] 筑波大・シス情・コンピュータサイエンス; [5] 名大高等研究院; [6] 統数研; [7] 国情研
[1] Ehime Univ; [2] CITE, Ehime University; [3] CITE, Ehime Univ.; [4] Computer Science, Univ. of Tsukuba; [5] Nagoua Univ., IAR; [6] ISM; [7] NII

http://www.infonet.cite.ehime-u.ac.jp/

宇宙科学研究は，科学衛星による宇宙空間探査を行い，観測データ処理により未だ人類が到達していない空間についての知見を得ることが研究の主目的のひとつである．近年，打ち上げられる衛星数は増大し，また観測データは複雑になっている．それらの観測データにより遠い宇宙空間の物理現象を理解するためには，観測される多種多様なデータを効率的に解析する環境が必要とされる．科学衛星観測データが，特定の組織だけではなく国内外の大学や研究機関で独立に公開されていることが，わが国の科学観測データベースの特徴のひとつである．そのために，解析者は，データの検索，収集，各データファイルからのデータの読み出し，異なるデータの比較といった数段階に渡るプロセスを経て，初めてデータから情報や知識を取得できる．これらの作業量の負担が，これまでの宇宙科学研究の妨げとなっていた．

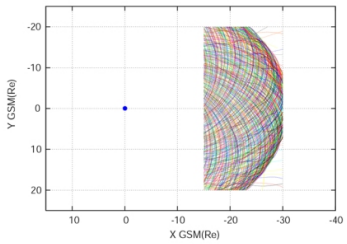
本研究では，分散管理・公開されている複数の科学衛星観測データをシームレスに検索・収集し，さらにこれらの観測データから所望する宇宙空間の物理パラメータを抽出する環境を構築する．クライアントは，まず，分散管理・公開されている観測データファイルの情報をメタデータベースから取得する．次に，メタ情報を使って必要とするデータファイルをダウンロードする．本研究で用いるメタデータベースは，我々の研究グループが開発してきたSTARS（科学衛星地上解析参照システム）メタデータベース [1] を用いた．システムでは，さらに，これらの取得データをGRIDのミドルウェアのひとつであるGfarm[2] システム上で解析する．Gfarmは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を実現するためのアーキテクチャを提供している．科学衛星観測データは多岐にわたり複雑であるが，個々のデータサイズはテレメトリの制限により比較的小さい．したがって，本研究では，Gfarmの全ノードにダウンロードデータファイルをリプリケートする．レプリケーションは，Gfarmの機能の一つであるgfrepコマンドを用いる．データ処理を各ノードで分散することで，観測データ統計処理が可能となる．

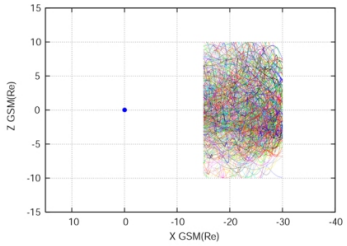
構築した試験システムを使った観測データの統計処理例を図に示す．図は，GEOTAIL衛星が地球磁気圏尾部（-30Re ≦ x ≦ -15 Re, -20 Re ≦ y ≦ 20 Re, -10 Re ≦ z ≦ 10 Re）に位置する軌道パスを抽出している．従来のイベント検索ではx，y，z，Bx（同衛星が観測する磁場のx成分）の条件に合う観測日時を独立して求め，これらのandをとる方法が主流であるが，この方法では負荷分散が容易ではない．本研究では，まずx，y，zの条件を満たす観測日時をあらかじめ抽出する．この処理は100%の負荷分散が可能である．本研究では，366個の期間が抽出されたが，これらの期間は最短で3秒，最長で709171秒とばらついている．次に，これらのイベントをGfarmの並列プログラム実行コマンドにより各ノードで処理（Bxの条件で検索）する．処理が終わったノードは次の期間を処理する．十分に粒度が小さい場合には，この方法でも十分に高い負荷分散が期待できる．本システムによる負荷分散の結果を調べたところ、高い負荷分散効果を得ることができた。なお、本研究ではGfarmに4ノードを用いたが、発表時にはさらに大規模なシステムでの報告を行う予定である．

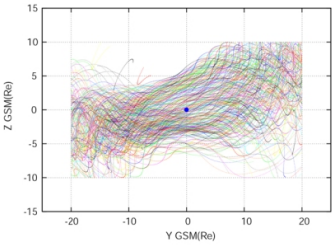
参考文献：

[1] 村田健史: 国際太陽地球系物理観測の広域分散メタデータベース，電子情報通信学会 (B)，vol.J86-B，no.7，pp.1331-1343，Jul. 2003.

[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).


Y GSM(Re)
X GSM(Re)


Z GSM(Re)
X GSM(Re)


Z GSM(Re)
Y GSM(Re)