2009－MPS－73 (1)
2009／3／5

グリッドデータファームによる太陽地球系物理分野における分散データ型データインテンシブ処理モデルの構築と評価

山本 和憲[1], 木村 映善[2], 村田 健史[3], 建部 修見[4], 松岡 大祐[5], 宮地 英生[6]
[1]愛媛大学大学院理工学研究科 [2]愛媛大学大学院医学系研究科
[3]独立行政法人 情報通信研究機構電磁波計測研究センター
[4]筑波大学大学院システム情報工学研究科
[5]独立行政法人 海洋研究開発機構地球シミュレータセンター [6]株式会社ケイ・ジー・ティー

太陽地球系物理データは，近年の増大化・大規模化に伴い，大規模データ処理が期待されている．大規模データ処理にはデータインテンシブ処理が行われる場合が多いが，分散データ管理下においてはデータ管理・共有の問題が生じるため，有効なデータインテンシブ処理モデルが求められている．本研究では，メタデータ利活用システム(STARS)とグリッドデータファーム(Gfarm)による分散データ型データインテンシブ処理モデルを提案し，8台の計算ノードによる実装を行った．さらに，構築したシステムで衛星観測データと計算機シミュレーションデータの並列分散処理の性能評価を行った．その結果，ファイルサイズが小さい場合はメタデータの階層化が有効であり，ファイルごとにデータ処理量が異なる場合はFIFO型スケジューリングが有効であることを検証した．

Development and Performance Evaluation of Distributed Parallel Processing System for Solar-Terrestrial Physics Observation Data and 3-D Computer Simulation Data based on Grid Datafarm Architecture

Kazunori YAMAMOTO[1], Eizen KIMURA[2], Ken T. MURATA[3], Osamu TATEBE[4], Daisuke MATSUOKA[5] and Hideo MIYACHI[6]
[1]Graduate School of Science and Engineering, Ehime University
[2]Ehime University School of Medicine
[3]Applied Electromagnetic Research Center, NICT
[4]Graduate School of Systems and Information Engineering, University of Tsukuba
[5] The Earth Simulator Center, JAMSTEC [6]KGT Inc.

In the Solar-Terrestrial Physics field, there has been tremendous increase of satellite observation data and computer simulation data. Since most of data files and computer resources are distributed over the Internet, analysis environments for data intensive processing are required. In this study, we propose a data intensive processing model of distributed data based on meta-data system and Grid Datafarm. A testing system is constructed with 8 filesystem nodes. As a result of small-data processing of observation data on the system, parallel processing is found effective using meta-data file at local disk and hierarchical Gfarm file. As for parallel visualizations of simulation data, it was achieved high parallelization efficiency of 97.6% when using FIFO-type scheduling.

1. はじめに

太陽地球系物理(STP: Solar-Terrestrial Physics)分野の衛星観測データおよび計算機シミュレーションデータは近年，増大化・大規模化している．これに伴い，複数衛星による多地点長期観測データの統計解析処理や長時間ステップの3次元可視化処理などの大規模データ処理が期待されている．

STP分野の大規模データ処理では，大量のデータファイルに同一処理を施すデータインテンシブ処理を行う場合が多い．しかし，データファイルが分散管理されており，有効なデータファイルの管理・共有方法やファイルI/Oの負荷分散が可能なデータインテンシブ

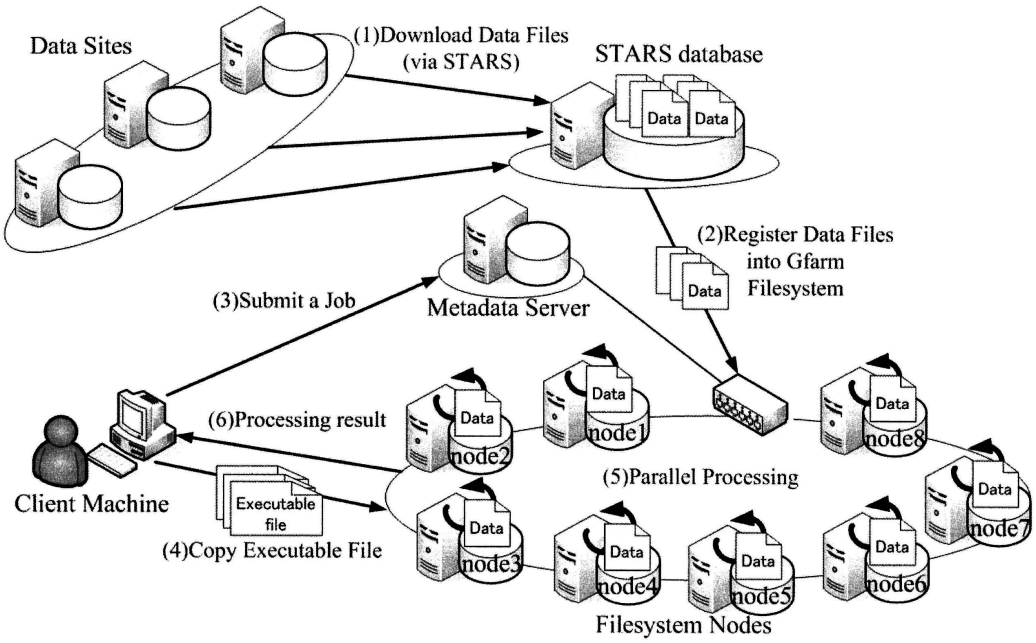


図 1　STARS と Gfarm による分散データ型データインテンシブ処理システム

表 1　システム構築に用いた計算機スペック

	Filesystem Node x8	Metadata Server x1 Client Machine x1
CPU	Athlon 64 x2 Dual Core 2GHz	Dual Core AMD Opteron 1.81GHz
Memory	2GB	1GB
Disk	1.2TB(using RAID0)	232GB
OS	Fedora Core5	Fedora Core5

処理モデルが求められている．

本稿では，太陽地球系物理データのメタデータ利活用システムである STARS(Solar-Terrestrial data Analysis and Reference System[1]とグリッドデータファームアーキテクチャの参照実装である Gfarm(Grid Datafarm) [2] を用いて，分散データ型データインテンシブ処理モデルを構築する．また，衛星観測データと計算機シミュレーションデータを用いて，ファイル共有方法と負荷分散の有効性・実用性の評価を行う．

2. 分散データ型データインテンシブ処理モデルの提案

2.1 モデルの概要

本研究では，分散管理されているデータを一度集約することで複数データ処理の多様性を確保し，集約後に並列分散処理を行い負荷分散することで計算機資源の性能を確保するデータインテンシブ処理モデルを提案する．このモデルでは，データインテンシブ処理に要する時間 *Ttot* を数式(1)のように，データファイルを集約して並列分散処理環境を整えるまでの前処理時間 *Tpre* と，集約後の並列分散処理時間 *Tpar* に分ける．この内，前処理時間 *Tpre* は，数式(2)のようにデータ検索時間 *Tsea*，データ取得時間 *Tget*，並列分散処理のためのデータファイル分配時間 *Tdis* に分けられる．

$$Ttot = Tpre + Tpar \quad (1)$$

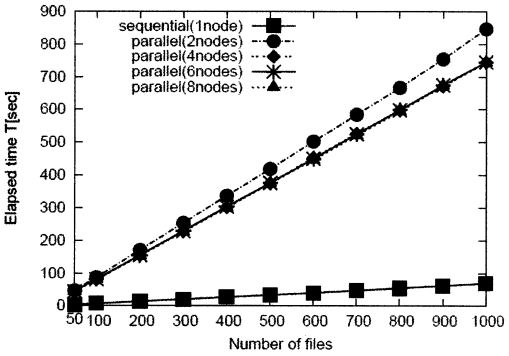
$$Tpre = Tsea + Tget + Tdis \quad (2)$$

提案するデータインテンシブ処理モデルでは，分散データ処理に付随する前処理のユーザの負担量が，大量のデータファイル処理を伴うために増大し，合計処理時間に対する前処理時間 *Tpre* の占める割合が無視できなくなる点に着目する．そこで，各プロセスの仮想化・統合化・自動化を行うことでユーザの負担を軽減し，前処理時間 *Tpre* の短縮を図る．

3.2 システム構成

本研究では太陽地球系観測データのメタデータ利活用システム(STARS)[1]とグリッドデータファーム(Gfarm)[2]を用いて，提案する分散データ型データインテンシブ処理モデルのシステム構築を行う．本シス

(a)



(b)

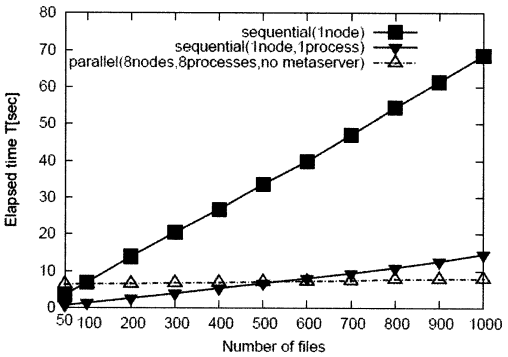


図 2　衛星観測データの逐次処理と並列分散処理の比較（■は逐次処理，●は 2 並列，◆は 4 並列，＊は 6 並列，▲は 8 並列，△は 8 並列（改良手法を適用），▼は逐次処理（1 プロセスで複数ファイルを処理）である）：(a) 改良手法を用いない場合，(b) メタデータのローカルキャッシュとメタデータの階層化を用いた場合

テムでは，データを集約するまでのデータ検索，データ取得，データフォーマットの差異の吸収を STARS が行い，集約後の並列分散処理を Gfarm が行う．

8 台のファイルシステムノードで構成したシステムを図 1 に，計算機スペックを表 1 に示す．ユーザはまず，STARS 経由でデータサイトからデータファイルを取得し（図 1-(1)），Gfarm ファイルシステムに登録する（図 1-(2)）．処理内容によっては効率的な負荷分散のために，登録と同時に各ファイルシステムノードにファイルを複製することもある．続いて，並列分散処理のジョブを投入すると（図 1-(3)），クライアントマシンから実行プログラムが各ファイルシステムノードにコピーされ（図 1-(4)），並列分散処理が行われる（図 1-(5)）．最後に，処理結果の表示の整合が取られ，ユーザ端末に結果が返される（図 1-(6)）．

4.　長期間衛星観測データの並列分散処理

本節では図 1 のシステムを用いて，衛星観測データの逐次処理と Gfarm による並列分散処理の比較実験を行う．逐次処理はファイルシステムノードを 1 台使用し，Gfarm によるオーバーヘッドは生じないものとする．並列分散処理は並列数 2～8 で行う．実験は 1,000 ファイルのデータインテンシブ処理を行い，データファイルの全数値レコードを出力するファイル I/O 処理を行う．データファイルは 1 ファイルあたり約 50KB のデータを用いて，ファイルサイズが小さい場合の Gfarm による並列分散処理の有効性を検証する．

実験結果を図 2 に示す．図 2(a)より逐次処理の方が Gfarm による並列分散処理よりも，実行時間が短いことが分かる．これは，Gfarm ファイルシステムに登録されたファイル名をメタデータサーバに参照する通信時間が，並列化により短縮される時間を上回っているためである．本研究では，メタデータの階層化[3]により，1 プロセスで複数ファイルの処理を行うことで，データファイルが小さい場合でも並列分散処理が有効であることを検証した（図 2(b)）．

5.　計算機シミュレーションデータの並列 3 次元可視化処理

本節では，図 1 のシステムを計算機シミュレーションデータの 3 次元可視化に応用し，並列可視化の有効性を調べる．実験に使用したデータは 1 ファイル約 80MB であり，150 ステップの可視化を行った

実験結果を図 3 に示す．シミュレーションの可視化は，物理現象の変化によりファイルごとにデータ処理時間が異なるため，ファイル数を等分割する手法では負荷分散が最適化されない（図 3(a)）．本研究では Gfarm に FIFO 型スケジューリング[4]を適用して有効性を検証した結果，97.6%の並列化効率を得ることができた（図 3(b)）．

6.　考察

Gfarm の並列分散処理はデータインテンシブ処理で各プロセスが独立しており，プロセス間通信が行われないため，並列数に依存しない高い並列化効率が期待される．しかし，並列分散処理時に Gfarm ファイル

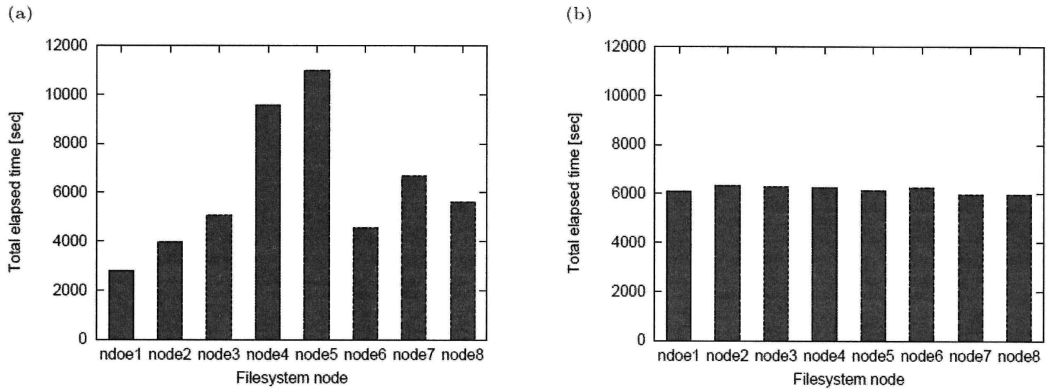


図 3　各ファイルシステムノードの合計可視化処理時間：(a) 各ノードにファイル数を均等に分配，(b)FIFO 型スケジューリングによるファイル分配

システムに登録された Gfarm ファイル名を参照するために 1 台のメタデータサーバを共有するため，図 2(a) のようにファイルサイズが小さい処理においてはオーバーヘッドの占める割合が大きくなり逐次処理の方が有利となる．並列分散処理の主なオーバヘッドであるメタデータベースアクセス時間を短縮した場合においても，ジョブ投入時のオーバーヘッドがあるため，図 2(b) のようにファイル数が一定数よりも小さいときには 1 プロセスで複数ファイルを逐次処理する方が有利となる．これは，Gfarm がライトアットワンスなデータサイズが大きいデータを対象としている理由の 1 つである．

図 3(b)の FIFO 型スケジューリングでは負荷分散が最適化されるため，並列数に依存せず高い並列化効率が得られる．ただし，全データファイルを各ノードに複製する必要があるため，並列化効率とデータファイル複製時間がトレードオフの関係になる．本実験では，12GB の全データファイルを全ノードに複製するのに要した時間は約 77 分であり，1 回目の可視化では両スケジューリング手法に処理時間の差が見られなかった．なお，2 回目以降の可視化では複製が不要となり，提案手法が有利となる．

7. 今後の課題

衛星観測データ処理では，データファイルサイズやデータ処理粒度がヘテロなデータセットを組み合わせた融合型データインテンシブ処理により，分野横断的な多目的データ処理環境の実現が期待される．また，シミュレーションデータの並列可視化では可視化処理とファイルシステムノードへのデータファイル転送のパイプライン処理によるリアルタイム可視化が期待される．

謝辞　本研究にご協力して下さいました宇宙航空研究開発機構・篠原育准教授に感謝致します．本研究は文部科学省の科学研究費補助金・学術創成研究費「宇宙天気予報の基礎研究」(17GS0208，代表者：柴田一成)の助成を受けて行いました．本研究では，宇宙航空研究開発機構科学衛星運用・データ利用センター及び京都大学生存圏研究所により公開されている衛星観測データを利用致しました．また，NICT リアルタイム地球磁気圏シミュレーションデータは，情報通信研究機構の SX-8R で計算致しました．

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，" 信学論(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，"ペタスケール広域分散データ解析のための Grid Datafarm アーキテクチャ，" ハイパフォーマンスコンピューティングと計算科学シンポジウム HPCS2002 論文集, pp.89-96, Jan. 2002.
3) Song Jiang　Xiaodong Zhang, Efficient distributed disk caching in data grid management, in: Cluster Computing, 2003. Proceedings. 2003 IEEE International Conference, pp. 446- 451, 2003.
4) Joseph Y-T. Leung, "Handbook of Scheduling: Algorithms, Models, and Performance Analysis," Chapman & Hall/CRC, 2004.