衛星画像の格納を目的とした大規模階層ファイルシステムの設計

根本 利弘　迫 和彦　喜連川 優　高木 幹雄
{nemoto,sako,kitsure,takagi}@tkl.iis.u-tokyo.ac.jp
東京大学生産技術研究所
〒106 東京都港区六本木 7-22-1

近年、地球環境問題に対する関心が高まり、衛星データは、ダイナミックに変動する地球環境の解明には必要不可欠なものとなった。しかしながら、衛星データは膨大であり、現状の二次記憶装置の容量では不十分であり、また、磁気テープ装置などからデータをマイグレーションさせるためには、長時間を要する。

衛星データの高速アクセスを実現するため、現在、我々は二次記憶装置に比べて安価で大容量である磁気テープ装置等の三次記憶装置を組合せたファイルシステムの構築を行っている。衛星データに対するアクセス傾向に着目し、本階層ファイルシステムでは、部分マイグレーション機構、画像ブロックのプリフェッチ機能を採用し、高速化を図っている。

本稿では、衛星画像データの特徴に触れた後、この衛星画像データをアーカイブするための階層ファイルシステムについて述べ、さらに試作システムによる実験結果を報告をする。

Design of Very Large Hierarchical File System for Archiving Satellite Images

Toshihiro NEMOTO, Kazuhiko SAKO, Masaru KITSUREGAWA and Mikio TAKAGI
Institute of Industrial Science, University of Tokyo
7-22-1, Roppongi, Minato-ku, Tokyo, 106 Japan

Recent attention on global environmental changes has stimulated the development of large scale global information systems. Satellite images play a very important role for understanding these global changes. However, the data size is very large and current file systems are not efficient enough to handle such huge data files. Migration of a whole file from tape to disk takes a very long time. Usually users are not interested in a whole image but in only a small portion of it. Thus, there is no need for full migration.

We are now implementing a partially migratable file system based on 8mm tape robotics. The file system migrates only the necessary portion of a file onto the disk. Two real application programs: radiometric/geometric correction and NDVI (Normalized Difference Vegetation Index) generation, were chosen and executed using our experimental file system. Large performance improvements were achieved compared to the conventional file level migration scheme.

1 はじめに

近年、地球環境問題に対する関心が高まり、地球環境を把握することは急務となってきた。同時に広範囲を繰り返し観測できるという特長をもつ衛星データは、ダイナミックに変動する地球環境の解明には必要不可欠なものとなった。このような背景のもと、東大生産技術研究所では10年以上もの間、気象衛星NOAA (National Oceanic and Atomospheric Administration) によって観測された地表面画像データの受信・蓄積を続けてきており、また、1995年4月よりNOAAに加えて新たにGMS (Geostationary Meteorological Satellite) によるデータの受信、蓄積を開始した。

しかしながら、衛星データはその特徴ゆえ、必然的に膨大なものとなる。例えば、NOAAおよびGMSによって観測された画像データは1シーンあたり約100MB、NOAAでは1日あたり6~8シーン、GMSでは1日に24シーンのデータが送られてくる。さらに、これらのデータを使用するために必要な放射量補正や幾何補正を行った後の画像はより大きなものとなる。このため、衛星画像をデータベース化し、効率的に利用するためには、現状の二次記憶装置の容量では不十分である。

この問題を解決する方法の一つとして、現在、我々は二次記憶装置と、二次記憶装置に比べて安価で大容量である磁気テープ装置等の三次記憶装置を組合せたファイルシステムの構築を行っている。このファイルシステムでは、二次記憶装置として高速でランダムアクセス可能な磁気ディスクアレイ、三次記憶装置として安価で大容量な8mmテープジュークボックス、およびD1テープジュークボックスを階層化し、両者の利点を兼ね備えたファイルシステムを実現する。衛星データにアクセス対するアクセス傾向に着目し、本階層ファイルシステムでは、部分マイグレーション機構、画像ブロックのプリフェッチ機能を採用し、高速化を図る。

本稿では、衛星画像データの特徴に触れた後、この衛星画像データをアーカイブするための階層ファイルシステムについて述べ、さらに試作システムによる実験結果を報告をする。

2 衛星画像データの概要

2.1 気象衛星NOAA

気象衛星NOAAは、米国海洋大気庁NOAA (National Oceanic and Admospheric Administration) による気象衛星である。軌道は、高度約800km、周期約102分の極軌道であり、衛星が日本付近上空を通過する十数分の間、データを受信することが可能である。現在、NOAA-12号、NOAA-14号が運用中であり、2つの衛星によって1日に6~8回、データを受信することができる。

衛星に搭載されているAVHRR (Advanced Very High Resolusion Radiometer) センサは、可視1チャネル、近赤外1チャネル、赤外3チャネルの計5つのチャネルより構成され、進行方向と垂直に直下点より±55.4°の範囲を1秒間に6回スキャンし、1回の受信で得られる画像は、衛星の進行方向に約5000km (約4000ライン)、垂直方向に約3000km (2048ピクセル) の範囲をカバーする。

2.2 気象衛星GMS (ひまわり)

気象衛星GMS (Geostationary Meteorological Satellite) は、日本の気象庁による静止衛星であり、現在、GMS-5号が運用されている。衛星は自転しながら北から南の方向へセンサを動かし、1時間毎に地表面の観測を行う。

衛星には、可視1チャネル、赤外3チャネルより構成されるVISSR (Visible and Infrared Spin Scan Radiometer) が搭載されており、可視チャネルによる画像は1画素6bit、サイズ

9164×10000、赤外チャネルによる画像は1画素8bit、サイズ2291×2500である。

2.3 衛星画像データへのアクセスの特徴

衛星画像データを対象とした階層ファイルシステムの構築に際して考慮すべき衛星画像データの特徴、アクセス傾向を以下にあげる。

- 膨大なファイルサイズ
  NOAA、GMSともに、1シーンあたりの画像ファイルサイズは約100MBと膨大であり、NOAAでは1日に約0.6~0.8GB、GMSでは1日あたり約2.4GB、1年では合計で約1TBにもおよぶ。
- 一定のファイルサイズ
  NOAA、GMSとも原画像1シーンあたりのサイズはほぼ一定であり、ファイル毎のばらつきは少ない。また、アクセスされる単位も1ラインが単位とされることが多く、ほぼ一定である。
- 偏りのあるアクセス
  衛星画像は最新の画像、晴れていて地表面が観測されている画像、特別な気象現象のある画像などに対するアクセスが多いなど、アクセスに偏りがある。
- 限られた領域へのアクセス
  ユーザが衛星原画像に対して処理を行う場合には、ある地点を中心とする領域を対象とする場合が多い。すなわち、ファイル中の連続する一部の領域のみをアクセスすることが多い。

3 階層ファイルシステム

衛星画像データを格納するための階層ファイルシステムを構築するにあたり、衛星データの特徴、アクセス傾向を考慮し、以下のことを設計方針とする。

- 通常のファイルシステム(いわゆるUFS)上のファイルにアクセスするのと同様なアクセス手法の提供
- データの存在するデバイスを意識する必要がないトランスペアレントなアクセス手法の提供
- 二次記憶装置の効率的な利用
- 三次記憶装置に存在するデータのマイグレーション時間の短縮

本ファイルシステムでは、階層記憶管理、部分マイグレーション機構、画像ブロックプリフェッチ機能を用いることによりこれらの方針の実現を試みる。

3.1 階層記憶管理

三次記憶まで階層記憶方式を拡張することで、大量のデータを効率良く、かつ扱いやすく格納することを実現する。

ユーザがあるデータにアクセスしようとすると、ファイルシステムはそのデータがどのデバイス上にあるかを調べ、二次記憶装置上にない場合には三次記憶装置から二次記憶装置に必要なデータを転送し、ユーザにデータを提供する。また、二次記憶装置の容量が少なくなった場合は、ファイルシステムはアクセス頻度の低いデータを三次記憶装置に戻したり、あるいは不要なデータの破棄を行ったりする。このような方法で、ユーザに対しデバイスを意識させることないデータへのオンラインアクセスを提供する。

3.2 部分マイグレーション

部分マイグレーションとは、ファイルシステムにおいて衛星データを幾つかのブロックに分割して管理し、ユーザのアクセス要求に対して、ファイル全体ではなく、ユーザの要求している部分を含むブロックのみを三次

記憶装置から二次記憶装置に転送する機構である。

この部分マイグレーション機構を導入することで、ユーザが必要とする部分の転送のみで済ませ、不要な部分の転送時間を削減することが可能となり、必要なデータが二次記憶上にない場合の三次記憶装置から二次記憶装置へのマイグレーション時間の短縮が期待できる。また、部分マイグレーションを行うことで、不必要なデータによって二次記憶装置が占有される可能性が低くなり、アクセス頻度の高いデータをより多く二次記憶装置上に保持することが可能となる。これは、アクセス頻度に偏りが大きいと考えられる衛星データを管理する上で効果的であると考えられる。

3.3 画像ブロックのプリフェッチ機能

衛星画像データに対しては、ある領域から連続してアクセスが行われる場合が多い。すなわち、ある画像ブロックにユーザがアクセスした後の次のアクセス要求は、直前にアクセスした画像ブロックの次のブロックに対するアクセス要求である確率が高いことを意味する。

このような衛星画像データのアクセス傾向を考慮し、画像ブロックのプリフェッチ機能を導入する。画像ブロックのプリフェッチ機能とは、ある画像ブロックにユーザがアクセス要求を出した場合、ファイルシステムは必要であればユーザの要求するブロックを二次記憶装置にマイグレーションするとともに、その次の画像ブロックも二次記憶装置にマイグレーションをする。すなわち、ユーザがはじめに要求したブロックにアクセスしている間に、ファイルシステムは次のブロックのマイグレーションを行うことでファイルシステムの高速化を図る。

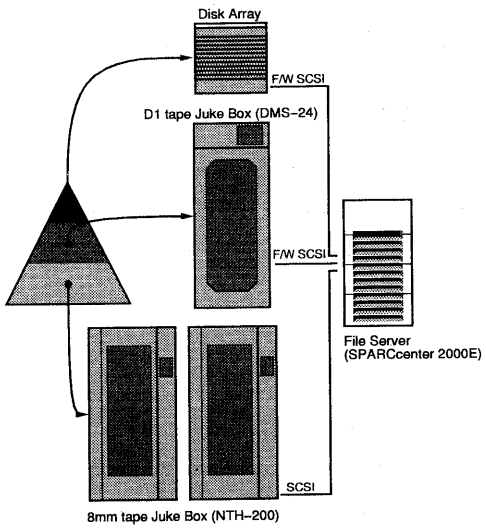

図 1: ハードウェア構成

4 システム構成

4.1 ハードウェア

構築を行っているファイルシステムのハードウェア構成を図1に示す。ファイルサーバとしてサン・マイクロシステムズ社製の SPARCcenter 2000E を使用し、これに接続される三次記憶装置として NTH-200 8mm テープジュークボックス、および DMS-24 D1 テープジュークボックスを、二次記憶装置として約 100GB のディスクアレイを用いる。

NTH-200 は 1 ユニットあたり 200 本のテープを格納し、1 ユニットに 2 台の Exabyte ドライブを搭載する。DMS-24 は 1 本あたり約 100GB の D1 テープを 24 本格納し、1 台のドライブを搭載する。

8mm テープは安価であるがアクセス速度が低く、D1 テープは高価であるがアクセス速度が高いので、DMS24 を NTH-200 よりも上位に配置して階層ファイルシステムを構築する。

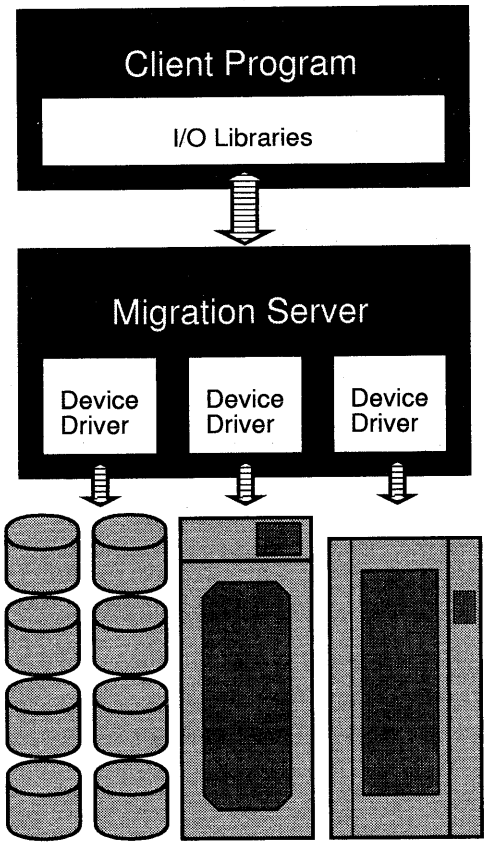

図 2: ソフトウェア構成

4.2 ソフトウェア

ソフトウェアの観点からみたファイルシステムの構成を図 2 に示す。本ファイルシステムは大きくマイグレーションサーバと I/O ライブラリを含むクライアントプログラムによって構成される。

以下、各構成要素について簡単に述べる。

4.2.1 マイグレーションサーバ

マイグレーションサーバは、階層記憶制御/管理を行うモジュールである。後述の I/O ライブラリをリンクしたクライアントプログラム間とで、リクエストや結果等の通信を行う。

階層ファイルシステム上で UFS と同様のディレクトリ構造を実現するため、階層ファイルシステム上のファイルのディレクトリエントリを UFS 上のファイルとして用意し、ディレクトリは UFS 上のディレクトリをそのまま用いる。また、同様にディレクトリエントリとして存在するファイルのファイル名、所有者、パーミッション、作成時刻等を階層システム上のファイルに適用する。ディレクトリエントリファイルには、階層ファイルシステムのディレクトリエントリであることを示す Magic Number、UFS における i-node に対応する、階層ファイルシステム上のブロックへのポインタ、階層ファイルシステムにおけるファイルのサイズが記述されている。

また、マイグレーションサーバは階層ファイルシステム上のファイルを管理し、部分マイグレーションを実現するために、インデックスと呼ばれるファイルを用意し、分割された衛星画像の各ブロックの情報を保持する。さらに、各デバイスの利用状況を管理するために、各デバイスに対して存在するデバイス管理テーブルを用意する。

4.2.2 I/O ライブラリ (クライアントプログラム)

I/O ライブラリは、ユーザが階層ファイルシステム上のファイルにアクセスするためのライブラリ群である。I/O ライブラリはマイグレーションサーバとのソケット通信を行うことにより、そのオペレーションを実行する。

I/O ライブラリ関数は、UFS の open、read、write 等のシステムコールと使用法をほぼ同一にすることで、ユーザに対して UFS 上のファイルへのアクセスするとの違和感なく階層システム上のファイルへアクセスすることを可能とする。

表 1: 使用画像

衛星	NOAA
原画像サイズ	95307460 byte
ライン数	4297 line

表 2: ファイルシステムの設定

画像ブロックサイズ	1M byte
ディスクブロックサイズ	256K byte
8mm テープブロックサイズ	256K byte

5 試作システムによる評価

本格的にシステムを構築するにあたり、ファイルサーバとして SPARCstation 10/41、二次記憶装置として SCSI ディスク、三次記憶装置として NTH-200 を用いた試作システム上に階層ファイルシステムを構築し、アプリケーションを実行させて、部分マイグレーション、画像ブロックのプリフェッチによる性能向上を評価した。試作システムでは、2 台の Exabyte ドライブを同時に動作させ、転送速度の向上も行っている。

性能評価に用いたアプリケーションプログラムは、

- 画像の切り出し/放射量・幾何補正プログラム
- NDVI(植生指数) 算出プログラム

を使用した。両アプリケーションによる評価を行う際の使用画像、ファイルシステムの設定はそれぞれ表 1、表 2の通りである。また、いずれのアプリケーションにおいても出力画像は UFS 上のファイルとして出力している。

5.1 画像の切り出し/放射量・幾何補正プログラム

本アプリケーションは、指定された領域を原画像より切り出し、放射量補正・幾何補正を行うプログラムである。図 3に処理領域と実行時間の関係を示す。横軸 x は、東京を中心とした $x° \times x°$の領域の補正画像を出力させることを意味する。また、図 4 に、部分マイグレーションを行わず、ファイル単位のマイグレーションを行った場合の実行時間を 100%とした場合の実行時間の比率を示す。

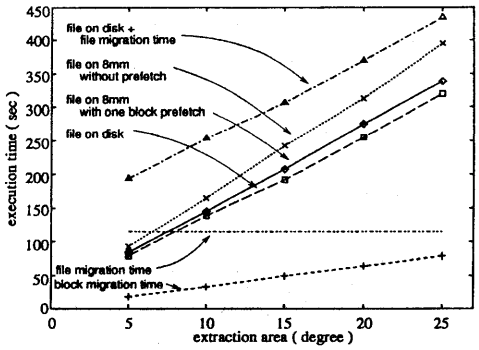

図 3: 画像切り出し・補正プログラム実行時間

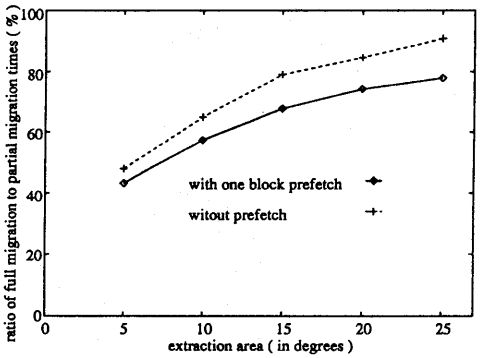

図 4: 画像切り出し・補正プログラム実行時間比

部分マイグレーションを行うことで、アプリケーションの実行時間が大幅に短縮されるということが示されている。また、先行マイグレーションを用いることでさらに高速化が可能であることが示された。

5.2 NDVI(植生指数) 算出プログラム

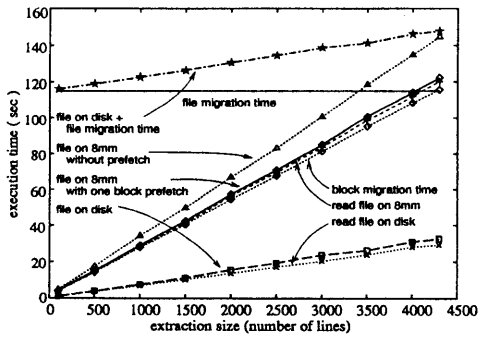

図 5: NDVI 値算出プログラム実行時間

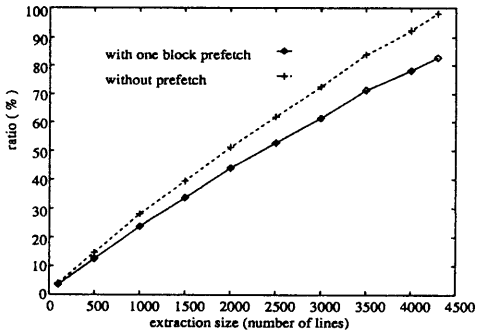

図 6: NDVI 値算出プログラム実行時間比

本アプリケーションは

$$NDVI = \frac{NIR - VIS}{NIR + VIS}$$

VIS : 可視チャネルのディジタルカウント値
NIR : 近赤外チャネルのディジタルカウント値

に基づき、NDVI (Normalized Difference Vegetation Index) をライン毎に求めるアプリケーションである。幾何補正は行わない。このアプリケーションによる実行結果を図 5、図 6 に示す。

NDVI 算出プログラムは、実行時間の大部分がファイル I/O に費されるため、先行マイグレーションによるシステムの性能向上は、画像の切り出し、補正プログラムほど顕著ではない。しかしながら、部分マイグレーションによる実行時間の短縮効果が大きいことが示された。

6 おわりに

本稿では、衛星画像を対象とした階層ファイルシステムについて報告した。衛星画像を格納するために適した階層記憶管理方法として、部分マイグレーション、および先行マイグレーションを採用することとした。試作システムにより、これらの実装方法を示すとともに、その有効性を明らかにした。

参考文献

[1] Stonebraker, M., Frew, J., Dozier, J.,"The Sequoia 2000 Architecture and Implementation Strategy", Sequoia 2000 Technical Report 93/23, University of California, Berkeley, CA, April 1993.

[2] Stonebraker, M., and Olson, M., "Large Object Support in POSTGRES, Proc. 9th Int'l Conf. on Data Engineering, Vienna, Austria, April 1993.

[3] Thomas S. Woodrow, "Hierarchical Storage Management System Evaluation," Third NASA Goddard Conference on Mass Storage Systems and Technologies, NASA Ames Research Center, 1993.

[4] 高橋一夫, 喜連川優, 高木幹雄. "衛星画像の格納を目的とした超大容量アーカイブデータベースシステム". 第 47 回全国大会講演論文集, 3C-3, 情報処理学会, 1993.

[5] 迫和彦, 高橋一夫, 喜連川優, 高木幹雄. "衛星画像データを対象とした階層ファイルシステムの実装". 第 50 回全国大会講演論文集, 2F-9, 情報処理学会, 1995.