5L－4

# 部分マイグレーション機能を有する大規模階層ファイルシステムの試作

根本 利弘† 迫 和彦‡ 喜連川 優† 高木 幹雄†
†東京大学 生産技術研究所
‡株式会社 日立製作所

## 1 はじめに

近年、地球環境問題に対する関心が高まり、現在では、変動する地球環境を把握するためには、同時刻に広い範囲を繰り返し観測できる人工衛星による観測データは必要不可欠なものである。このような背景のもと、我々は10年以上もの間、気象衛星NOAAによる画像の直接受信・アーカイブを続けてきており、さらに本年4月よりGMS(ひまわり)によるデータの受信・アーカイブも開始した。

しかしながら、衛星画像は広域性・反復性ゆえ、1シーンでもデータ量は膨大であるが、研究者が地球環境の把握のためには何シーンものデータを利用する必要があり、このため、必要な画像データを高速、かつ容易に取得できるようなシステムが望まれている。

そこで、現在、我々は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を構築中である。本ファイルシステムは8mmテープアーカイバ、D1テープアーカイバ、ディスクアレイを階層的に組み合わせることで、デバイスを意識せずに、高速なアクセスを実現することを目的としている。本稿では、この階層ファイルシステムについて報告を行う。

## 2 システム構成

### 2.1 概要

図1に本ファイルシステムの構成を示す。ソフトウェアはマイグレーションサーバとクライアントプログラムの2つに大別され、クライアントプログラムは、専用のI/Oライブラリを用いて階層ファイルシステムのファイルにアクセスする。I/Oライブラリの各関数は、マイグレーションサーバと通信を行い、必要なデータの入出力等を実現する。マイグレーションサーバはデバイスドライバを通して各デバイスを管理する。

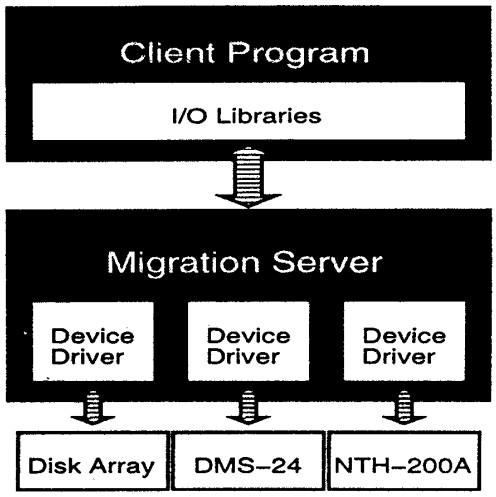


図1: 階層ファイルシステムの構成

### 2.2 階層構造の実現

各デバイスドライバは、デバイスに対するI/O処理の他に、そのデバイスに存在するブロックの情報、デバイスの利用状況を管理する。マイグレーションサーバは、デバイスを管理する際、デバイス上にデータがない場合にマイグレートするもとになるデバイスを示すリスト、デバイスの容量が少なくなった際のデータのマイグレート先のデバイスを示すリストをともに管理し、全てのデバイスを木構造で管理することで階層構造の構築を実現し、柔軟な変更を可能とする。

#### 2.2.1 データの書き込み

階層ファイルシステムにデータを書き込む際には、最下層のデバイスにも同時に書き込む。これは、対象としている衛星データは、最下層のテープデバイスのアクセス速度に比べ、衛星からの送信速度が低速であり、また、一度書き込まれたデータが変更されることは非常に少ないためである。

---

Very Large Hierarchical File System Supporting Partial Migration
Toshihiro NEMOTO, Kazuhiko SAKO, Masaru KITSUREGAWA and Mikio TAKAGI
Institute of Industrial Science, University of Tokyo
Roppongi 7-22-1, Minato-ku, Tokyo, Japan

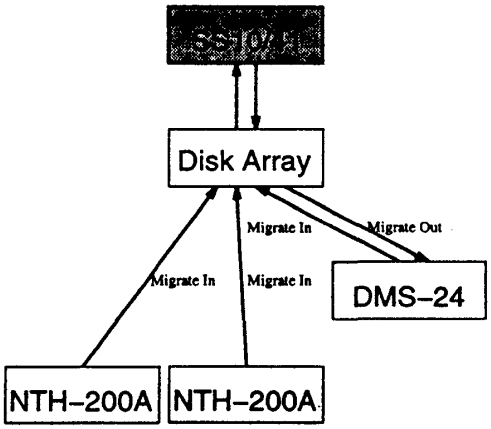


図 2: ファイルシステムの階層構造

### 2.2.2 マイグレートイン

クライアントプログラムからアクセス要求を受けると、マイグレーションサーバは最上位のデバイスにアクセスする。最上位デバイスにデータがない場合には、そのデバイスのマイグレートもとのデバイスへアクセスしマイグレートを行う (マイグレートイン)。そのデバイスにもデータが存在しない場合には、さらにそのデバイスのマイグレートもとへアクセスしマイグレートする、ということを繰り返してデータのマイグレーションを行う。

### 2.2.3 マイグレートアウト

あるデバイスへデータが書き込まれた際には、マイグレーションサーバはそのデバイスの残り容量をチェックする。容量が少ない場合には、古いデータをマイグレート先へマイグレートする (マイグレートアウト)。最下層デバイスを除き、マイグレート先のデバイスリストにデバイスがなにもない場合にはそのデータを廃棄する。

## 3 試作システム

ファイルサーバとして SPARCstation 10/41 (Solaris 2.4)、記憶装置として、8mm テープジュークボックス NTH-200A を 2 台、D1 テープジュークボックス DMS-24 を 1 台、17GB のディスクアレイを 1 台用い、試作システムを構築した。8mm テープドライブは標準 SCSI、D1 テープドライブ、ディスクアレイは Fast Wide SCSI で接続されている。各デバイスのデバイスドライバは OS 標準のものを用い、その上に階層ファイルシステムのデバイスドライバを実装している。また、ディスクアレイは、UFS としてマウントして使用することも、マウントせずに raw デバイスとして使用することもできる。

表 1: データ読み込み時間

| デバイス | 時間 (秒) |
|---|---|
| DMS-24 (UFS Disk) | 75.8 |
| NTH-200A (UFS Disk) | 159.0 |
| DMS-24 (raw Disk) | 60.2 |
| NTH-200A (raw Disk) | 155.5 |

階層構造は図 2のように、最下層に 2 台の NTH-200A を配置して全てのデータを 8mm テープでアーカイブし、これらのデータをディスクアレイでキャッシュし、ディスクアレイ上の古くなったデータを D1 テープにマイグレートし、D1 テープ上の古くなったデータは破棄する、という構成にしている。

この階層ファイルシステムを用いて、テープの入れ換え・シークを伴わない場合の、ディスク上には存在しない NOAA データ 1 シーン (70554580 byte) を D1 テープジュークボックス (DMS-24)、および 8mm テープジュークボックス (NTH-200A) から読み込む時間を表 1に示す。なお、8mm テープのブロックサイズは 10Kbyte、D1 テープ、raw デバイス時のディスクアレイのブロックサイズは 1Mbyte である。

ディスクをマウントせずに raw デバイスとして使用することで、ディスクアレイへのデータの読み込み・書き込み時間が短縮され、その結果、階層ファイルシステムにおける読み込み時間が短縮された。しかしながら、D1 テープドライブ、ディスクアレイの転送速度はそれぞれ、16MByte/s、20MByte/s であり、これらのデバイスの性能を十分に引き出せていないことがわかる。

## 4 おわりに

8mm テープジュークボックス、D1 テープジュークボックス、ディスクアレイを用い、これらを階層的に組み合わせたファイルシステムの試作を行った。今後は、各デバイスに対する専用のドライバを開発し、アクセス速度の向上を図る予定である。

## 参考文献

[1] 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, 1995. 2F-9.

[2] 根本利弘, 迫和彦, 喜連川優, 高木幹雄. "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". 情報処理学会研究報告, Vol. 95, No. 65, pp. 65–71, 1995.