ﾏﾙﾁﾒﾃﾞｨｱ通信と分散処理 86－24
グループウェア 26－24
（1998. 1. 30）

マルチメディア情報検索のための階層蓄積管理方式の検討

－蓄積システムの構築と磁気テープ装置の評価－

渋谷美継* 裏野博明* 柳田崇* 脇英世**
*通信放送機構 **東京電機大学教授

大容量性、高速応答性及び経済性に優れたビデオサーバを構築するためにハードディスク装置と光磁気ディスク装置及び磁気テープ装置を用いた階層蓄積管理方式の検討を行う。

本報告ではハードディスク装置のみで構成されるビデオサーバに対して、前記各装置を用いたビデオサーバのそれぞれの装置にデータを適切に配置し、また蓄積されたデータをアクセス頻度に応じて再配置する方法と、読み出しの低速な装置を組み合わせても読み出しレスポンスを劣化させないためにデータの分割による蓄積方法を提案し実験によりその有効性を証明する。

階層蓄積管理方式の検証のために、蓄積システムの構築と磁気テープ装置の新規開発及び評価を行ったので合せて報告する。

Examination of a hierarchical storage management system for multi-media information retrieval

－A construction of storage system and evaluation of the magnetic tape devices.

Mitsugu Shibuya* Hiroaki Urano* Takasi Yanagita* Hideyo Waki**
*Telecommunication Advancement Organization of Japan **Tokyo Denki University

Here we discuss the method of hierarchical storage management to construct the video server system which has large capacity, high response speed and also cost effectiveness.

The system consists of hard discs, opti-magnetic disc archives and magnetic tape archives.

Appropriate data allocation method, re-allocation method accessing frequency, improvement of the response time in combination with the hard disc and magnetic tape archive are proposed and examined through the experiment. These methods are very much different from the conventional single storage media system such as hard disc system.

Here we also report the new magnetic tape archive which has the big data capacity, and their experimental evaluation on the system.

1 はじめに

近年デジタル圧縮技術並びにデータ転送技術の進展が目覚しく、ビデオデータを蓄積したビデオサーバからデータを複数の端末に提供するビデオオンデマンド(VOD)の開発が盛んに行われている。インフラの整備と共にホテル、公共施設、美術館、博物館、学校内教育システムなどと用途も多岐に広がっている。

情報の増加と共にこれらも静止画や文字から、動画映像へと比率が増している。動画映像はテキストデータなどに比べて情報量も多く、これらに

対応するためには更に大規模な蓄積と経済性に優れたビデオサーバが必要になっている。

従来このようなビデオサーバを実現するために階層蓄積技術に関して、ハードディスクのような高速装置と低速で大容量を持つ装置を組み合わせたシステムにデータを分散する方法が提案されている[1][2][3]。またアクセス頻度の高いデータはハードディスク装置へその他のデータは大容量の光磁気ディスク装置へ蓄積すると共に、データの分割を行って先頭データをハードディスク装置へ後続のデータを光磁気ディスク装置へ蓄積して頭出し遅延を少なくする方法や、ドライブ装置やロボットハンドの競合による遅延を考慮した高速読み出し技術の提案がされている[3]。

データへのアクセス頻度の算出方法として映像ファイルの参照確立分布の解析[4]も行われている。しかし対象となる映像データが1分程度と小さく、また蓄積装置間での長時間データ交換に関する検討はなされていなかった。

階層蓄積管理方式の検討のために、ハードディスク装置に加えて光磁気ディスク装置及び磁気テープ装置を組み合わせたビデオサーバを構築し、扱うデータとして30分以上の長時間AVデータを使うことにした。これらを使って各蓄積装置へのデータの分散配置法と経時変化やアクセス頻度に応じて再配置を行うマイグレーション方式、ハードディスク装置と磁気テープ装置を使った場合の見かけ上の読み出しレスポンスを速める方式の検討を進めている。

ビデオサーバの構築には、大容量の情報蓄積媒体として磁気テープ装置が経済的に有効である。本システムでは世界で初めてのコンパクさと性能、機能を持ったテープストリーマライブラリ(以降テープストリーマ)の開発も同時に行ったので合せて報告する。

本提案の階層蓄積管理方式では新たに開発したテープストリーマを大容量データ蓄積の重要装置として位置付けている。

2 階層蓄積方式

一般的に大容量データの階層蓄積の考え方は図1に示すように、読み出し速度の高速、中速及び低速の蓄積装置を組み合わせ、アクセスの高いデータをハードディスクなどの高速な装置に蓄積し、アクセスが中くらいだと中速の光磁気ディスク装置、アクセスが非常に少ないデータは磁気テープ装置に蓄積する方法が取られている。またデータ容量が小さいほど高速装置へ大きいデータは低速装置に蓄積される傾向にある。

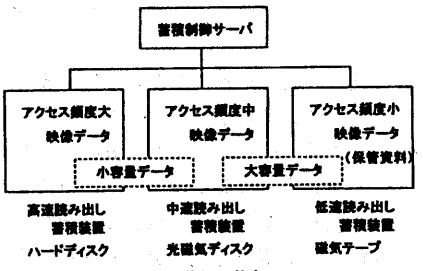

図1 階層蓄積の基本

VODにおけるビデオサーバも、高速低容量の装置と低速大容量の装置を組み合わせた混在システムになることは十分考えられる。大容量の蓄積装置を単なるバックアップ用のアーカイブ装置として使うのでなく、積極的に高速装置を補助するシステム装置として位置づけるべきである。

映像情報の場合、データ容量が大きいためそれに合せてデータの転送時間も比例して長くなる。視聴者側から見れば、早いデータ提供と情報が途切れずリアルタイムに視聴できる事は当然の要望である。

我々の取り組む階層蓄積管理の課題と検討内容は次のものである。

① データごとの蓄積装置の選択

多数の被蓄積データを蓄積する時に、どのデータをどの蓄積装置に入れるか選択する必要が出てくる。ハードディスク装置主体のシステムならデータを順次蓄積していけばいいが、それぞれ特性の異なる装置の場合は、ファイル容量、アクセス頻度（初期時は分からないので予測）、データの持つ情報の新規性など、また各媒体装置の蓄積容量、運用コストなどの要点を考慮して優先順位を定めて選択する方法を求める。現行 VOD 環境の AV データ提供の実状を調査し分析する。

② データのマイグレーション

1 項の選択によって各々蓄積装置に蓄積されているデータは運用の経時によってアクセス頻度が変化してくる。またデータも新規に追加されてくるのでアクセス頻度の変化や更新に対応してデータの入れ替えを行う。アクセス頻度の高くなったデータをできるだけ高速な装置へ、頻度の低くなったデータを低速な装置へと再配置するマイグレーションは、限られた蓄積装置を有効に使いレスポンスや蓄積効率を考える場合に重要な技術である。

③ 低速装置の読み出し速度向上

大容量低速な蓄積装置に蓄積されたデータを要求に応じていかに早く読み出すかは、低速蓄積装置の最大の課題である。装置の機能アップによる高速転送には限界があり、装置の組み合わせ技術を工夫する必要がある。今回低速装置にテープストリーマを使い、ハードディスクとの組み合わせで見かけ上の読み出しレスポンスを速める実験を行う。特にデータ容量を数百 MB から 1GB 以上の大容量に設定する事によって低速装置の有効性を検証できる。

3 実験システム

3.1 システム構成

情報蓄積システムは、映像情報を MPEG-2 データにエンコードする入力系とデータの蓄積、書誌情報の管理を行なう蓄積系、ここには情報蓄積サーバと光磁気ディスク・テープ各蓄積装置が含まれる。読み出した MPEG-2 データをデコードする出力系で構成されている(図 2)。ネットワークは ATMLAN、Ethernet, SCSI でそれぞれ接続されている。

3.2 アーキテクチャ

入力系と出力系はそれぞれ MPEG-2 データの書込み及び読み出しを操作するクライアントにもなっている。この操作で情報蓄積サーバを作動し、蓄積サーバの指令制御で光磁気ディスクライブラリ(以降光磁気ディスク)とテープストリーマが実行動作する。入力系と出力系へのデータは ATM LAN 上で ftp 転送を行なっている。

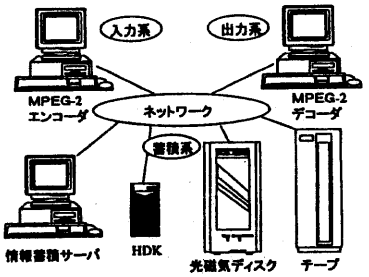

図2 情報蓄積システムの構成

情報蓄積サーバとテープストリーマ間及びテープストリーマ内の制御情報は Ethernet 上で TCP/IP ソケット通信を行っている。これは制御データの相互通信が頻繁に行なわれるために実験設備として ATM LAN 上の障害などに影響されないように考慮した。蓄積系内の MPEG-2 データ転送は Fast SCSI-2 を使用している。システムネットワークを図 3 に示す。

3.3 蓄積システムの主機能

本システムは、入力系で作られた MPEG-2 デ

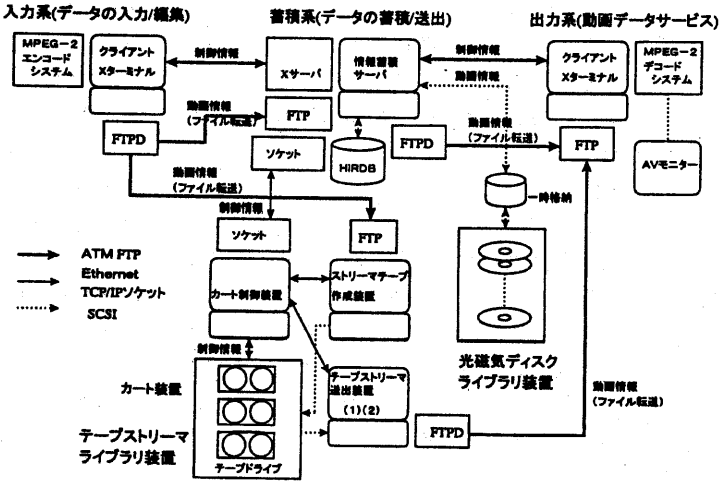

図3. 情報蓄積システムネットワーク図

ータとそのデータに関係する情報を蓄積装置に入力蓄積し管理する事が目的である。主に次の機能に分けられる。

・データの蓄積と提供

蓄積装置へのMPEG-2データの書込みと読み出し。データの入出力はクライアント端末においてコンテンツ格納先情報一覧を使って入力実行する。

・データの書誌情報管理

クライアント端末からデータファイル名、コンテンツ名、製作者名、日付、容量、格納媒体、日付などの映像データに関係する情報を登録する。

・ライブラリ管理

データファイルの存在が蓄積装置内の個々の媒体上のどこにあるのか詳細なディレクトリを管理する。

・蓄積媒体管理

蓄積装置内に使用目的の媒体が存在するか、又その媒体が現在使用可能であるかを管理する。特にテープストリーマの場合は、上書きをさせないためのプロテクトが必要となる。

3.4 サーバ及びクライアントシステム

情報蓄積システムは実験操作性を考慮して、情報蓄積システムをサーバシステムとクライアントシステムに分離した。サーバとクライアントは前述の機能の制御操作を行う。

・サーバシステム

サーバシステムは情報蓄積システム系の中心を成しており、光磁気ディスク並びにテープストリーマの動作制御と媒体情報や書誌情報類のデータ管理を行なう。

・クライアントシステム

クライアントシステムは、情報蓄積システムを作動させるための操作部である。サーバシステムに対してデータファイルや書誌情報の書き込み／読み出しと管理を実行させる。またクライアントシステムには入力系のMPEG-2エンコーダと出力系のMPEG-2デコーダが各々接続されておりAVデータの加工と確認を行う。

4 蓄積装置

4.1 光磁気ディスクライブラリ

両画面で2.6GBの記録容量を持つ5.25インチの光磁気ディスクとキャッシュ用のハードディスクで構成され,25枚のディスクを1記録ファイルとして取り扱っている。4台のドライブでパラレル運転が可能である。光磁気ディスクをパーティションに割り当て、大容量ハードディスクの性能に近づけるマルチボリューム機能や、ハードディスクとの一体により同一ファイルの読み出し要求に対してはメモリしてあるハードディスクから読み出す事によって光磁気ディスクへのアクセス速度を高める機能を持たせている。主な性能を表1に示す。

表1. 光磁気ディスク性能仕様

項目	性能
平均回転待ち時間	約10ms
平均シーク時間	約39ms
転送速度	2.3~4.6MB/s
平均ロード時間	約5秒
平均アンロード時間	約2.5秒
アクセス時間	約3秒(マウント・デマウント時)
ディスク1枚の記録容量	2,607MB／枚
光磁気ディスクカートリッジ収納枚	180枚
全記録容量(1024B/セクタ使用時)	469GB
光ディスクドライブ台数	4台

4.2 テープストリーマライブラリ

本装置を開発するにあたっては、大容量化と装置全体のコンパクト化の他に、今後の方向を考慮した装置自体の展開性、汎用性にも十分対応性のあることなどを条件にした。データ容量、ドライブ数、転送レート等実験への適合性と他の機器との整合性もとれるようにしている。

放送用デジタルVTRとして画質などの性能、機能に優れ、非常に小型であるDVCPROフォーマットをベースにして開発を行なった。テープ幅6.35mm、カセットサイズ125mm＊78mmの小型セットを使っているためストリーマドライバ本体を非常に小型に出来ている。データの書込みを行なうストリーマドライバ1台とデータの読み出しを行なうストリーマドライバが2台、更に小型カセット120巻を収納する3列40段の収納棚が小型のラックに組み込まれている。

ラック内の前面部をロボット機構1台が移動し、ロボットハンド機構を使ってストリーマドライバと収納棚へのカセットの着脱を行なう。ストリーマドライバは各々個別のワークステーションに接

続されており蓄積サーバと TCP/IP ソケット通信を行っている。

書込み、読み出しのデータの入力及び出力は ATM LAN を通じてクライアントシステムの MPEG-2 エンコーダ及びデコーダに接続されている。

小型のカセットは 1 巻で 20GB の容量が有り、MPEG-2 圧縮で 6Mbps のデータなら約 7 時間分の AV データが書込める。120 巻のカセット合計で 2.4T バイト約 800 時間分の AV データが蓄積できる。転送レート 25Mbps での書込み及び読み出しが実現できた。テープストリーマの性能を表 2 に示す。

表2. テープストリーマ性能仕様

項目	性能
データ転送速度	3.1MB/s
インターフェース	Fast SCSI−2
エラーレート	10^{-17}以下(ECC, Retry)
バッファメモリ	8.2GB(1ドライブ)
リワインド時間	110秒 max
リポジショニング時間	2秒 max
データ容量	20GB
最大カセット収納本数	120本
全データ容量	2, 400GB
ストリーマドライブ台数	3台(書込み1台、読み出し2台)

4.2.1 テープストリーマの動作

テープストリーマの運用制御はテープストリーマ制御装置で行なっている。

テープストリーマ制御装置はカセット及び書込まれるデータのファイル名、ファイルサイズの管理と蓄積サーバからの指令であるデータの書込みまたは読み出しの要求に対する通信を行ないながらテープストリーマを作動させる。

カセットの識別はカセット固有の番号をバーコード情報で管理し、カセットの背面にバーコードラベルを貼って格納棚に格納した後、バーコードリードを実行させるとバーコードを読み取った以降カセットは自動管理される。使用したいカセットの必要に応じてブランク指定するとデータの書込みが可能となる初期状態になる。カセットの移動や入れ替えを行なった場合は、其のたびにバーコードリードの実行によって格納棚内の配置とカセットの存在が認識される。これらのカセットの情報は常に情報蓄積サーバに送られ登録管理される。

データの書込みの管理もカセット毎に行なっており、データが書き込まれる毎にカセット内の残容量を計算し、残量が常に表示される。

データの不用意な消去を防ぐために書込まれているデータの上書き更新はできない。

また新規の書込みデータのファイルサイズより残容量が少ない場合は事前にプロテクト表示し、書込み途切れを防ぐようにしている。蓄積サーバとの通信内容は、他に書込み、読み出しの実行指令、テープストリーマの動作状態、カセットデータの検索、データ削除などがある。

5 蓄積システムの特性と評価

5.1 ハードディスク装置の遅延特性

seek.exe プログラムによりハードディスクをアクセスする際のオーバヘッド時間の測定結果を図 4 に示す。内周から外周までのシーク距離を変えた場合のオーバーヘッドを示す。

シーク距離には単純比例せず小さくなるとばらつきが大きくなっている。またシーク距離が小さい場合でも 6〜7ms のオーバヘッドが常に存在しシークを行うペナルティが大きいことが分かる。

内周側でも外周側でも、シーク距離が同じであればオーバヘッド時間も同じであった。オーバーヘッドの大半はヘッドシークと回転待ち時間で占められている。

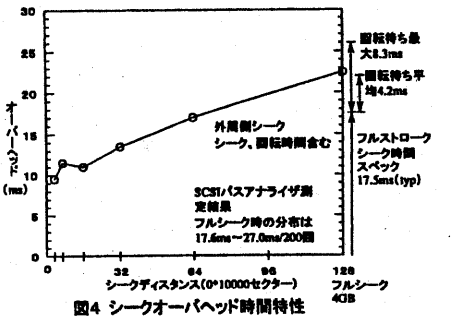

図4 シークオーバヘッド時間特性

5.2 テープストリーマの遅延特性

テープストリーマの記録媒体は長尺の磁気テープであり磁気テープはカセットに収まった構造をしている。カセットは通常格納棚に保管されているのでロボットハンドを使ってドライバに着脱する機構が必要になる。これらの事からデータアク

セスに対するレスポンスが遅い欠点がある。一般にバックアップやアーカイブのような大量保管装置として多く使われることが多い。

今回実験に用いたテープストリーマの遅延特性を図5､6に示す。図5はMPEG-2圧縮で

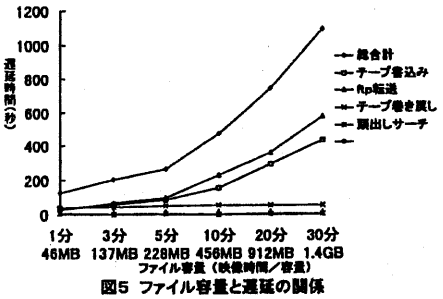

図5 ファイル容量と遅延の関係

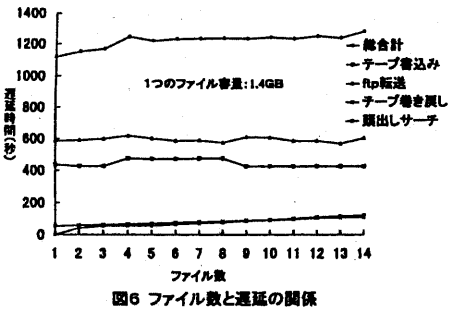

図6 ファイル数と遅延の関係

46MBから1.4GBまで(実映像で1分から30分に相当)6種の映像データを無記録のカセットにそれぞれ別々に書込みを行った場合のデータ容量と遅延の特性図。転送時間のばらつきはあるがほぼデータ容量に比例する事が分かる。図6は1.4GB(30分相当)の映像データを1つのカセットに別ファイル名で順次書込みを行った場合のデータファイル数と遅延の特性図である。最大容量20GB内に14のデータが記録されている。データ転送速度は図5、6のテープ書込み特性から平均3MB/秒あるがばらつきが多く一定していない。これはテープ上の傷によるエラーなどの発生で再書込みがあった事が考えられる。読み出し速度もほぼ同じでばらつきがある。ハードディスク装置のシークオーバーヘッドに相当するテープストリーマのアクセス時間(カセットを捜してからドライバに装着後データの書き込み或いは検索状態になるまで)は図6の頭出しサーチ特性から2～130秒かかる。目的とするデータ位置がテープ上の後ろにあるほどサーチ時間を要する。他にも遅延を発生するものとしてカセット取り出し/格納のロボットハンド動作、テープ巻き戻しがある。低速読み出し装置の速度向上実験では、データの分割と容量はこれらの遅延時間によって決められる。

6 結び

ハードディスク、光磁気ディスク、テープストリーマを組み合わせたビデオサーバシステムを構築し、長時間のAVデータを効率よく管理し送出する手法として三媒体併存型の階層蓄積管理法式を提案した。システム並びに蓄積管理システムの有効性･実用性を実証するために

①データごとの蓄積装置の選択

②蓄積装置間のデータマイグレーション

③テープストリーマとハードディスクの組み合わせによる低速読み出し装置の見かけ上の速度向上について実証していくための基礎データとしてハードディスクのアクセスとオーバーヘッド時間またテープストリーマのファイル容量、ファイル数と遅延の関係などの明確化を図った。今後上記②,③について実験を行い最適マイグレーション方式の決定、読み出し速度の向上などを図っていく予定である。

参考文献

[1] 藤井実、浅井清、“階層的ファイル自動管理システムの設計”情報処理学会論文誌 Nov, 1980 vol21 no, 6

[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 ﾃﾞｰﾀﾍﾞｰｽｼｽﾃﾑ研究会 104-9 1995,7

[3] 鈴木偉元、西村一敏、阪本秀樹、“階層化蓄積ビデオサーバの性能解析”電子情報通信学会論文誌 D-I volJ80-D-I no, 3 1997,3

[4] 鈴木偉元、石橋豊、西村一敏、“映像ファイルの参照確立分布の解析”電子情報通信学会秋大会 1992