テープマイグレーション機能を有する三次記憶装置におけるファイル管理手法とその衛星画像データベースシステムへの適用に関する研究

根本 利弘

目次

図目次

表目次

第1章　序論

1.1　背景

近年，プロセッサや通信回線の速度向上に伴い，音声や動画像を扱うマルチメディアデータベースや衛星画像や数値モデルグリッドデータを扱う地球環境情報システムなど，大規模なデータを扱うアプリケーションが構築されるようになってきた．米国では，NASAを中心としてEarth Observing System Data and Information System（EOSDIS）の構築が進められている．このシステムは，現在稼働中，あるいは今後打ち上げが予定されている複数の地球観測衛星のデータのアーカイブを目的としており，米国内の8つのデータアーカイブセンタにより，1日に1TB以上のデータを格納を行う予定である．また，東京大学生産技術研究所でも，1983年より受信を開始した気象衛星NOAAによる地表面の観測画像，および1995年より受信を開始した気象衛星GMS（ひまわり）による観測画像のデータベースを構築中であるが，これらのデータの容量は原画像のみでも約9TBになる．これらの原データに対して幾何補正や放射量補正を施した一次処理画像，さらには海面温度分布，植生指数などに変換したデータなどを加えると数十～数百TBにも及ぶ．

二次記憶装置は急激にその容量を増し，価格を下げつつあるものの，地球環境情報のような膨大なデータをアーカイブするには未だ十分ではない．二次記憶装置のみで膨大なデータをアーカイブすることは極めて大きなコストを要することとなり，現状ではまだ非現実的である．数十TB以上の膨大なデータをアーカイブする記憶システムとして，大規模三次記憶システムは必要不可欠である．

現在，三次記憶システムに用いられるメディア/ドライブ装置としては，記録したデータの消去，更新が可能なものに限ると，光/光磁気ディスクと磁気テープに二分される．光磁気（MO : Magneto Optical）ディスクに関しては，3.5インチの直径を持つディスクが一般に広く用いられ，その容量は最大約1.3GBである．また，最近入手可能となり，今後

表 1.1: 三次記憶ドライブ装置の例（ドライブが有する圧縮機能を用いない場合）

ドライブ装置名	メディア種類	容量	読込み速度	書込み速度
富士通 MCK3130SS	MO（3.5inch）	1.3GB	2.98~5.09MB/s	0.99~1.7MB/s
松下 LF-D201JD	DVD-RAM	9.4GB	2.77MB/s	←
HP SureStore DAT40	4mm テープ	20GB	3MB/s	←
Exabyte Mammoth	8mm テープ	20GB	3MB/s	←
Quantum DLT8000	1/2inch テープ	40GB	6MB/s	←
STK Redwood SD-3	1/2inch テープ	50GB	11MB/s	←
Quantum SDLT220	1/2inch テープ	110GB	11MB/s	←
IBM 3580 Ultrium	1/2inch テープ	100GB	15MB/s	←
Sony DIR-1000H	19mm テープ	96GB	64MB/s	←
Ampex DST312	19mm テープ	330GB	15MB/s	←

の普及が期待されるメディア/ドライブとしてDVD（Digital Versatile Disk）-RAMがある．DVD-RAMは，転送速度2.77MB/s，容量は4.7GB（片面），9.4GB（両面）であり，さらに，青色レーザをヘッドに用い10GBを超える容量を持つDVDの開発も行われている．

一方，磁気テープメディア/ドライブ装置は，現在までに様々なテープ/ドライブ装置が開発，発売されており，そのデータ転送速度，データ容量とも多岐にわたっている．幾つか例を挙げると，パーソナルコンピュータやワークステーションなどでは，4mmテープ（DAT : Digital Audio Tape）や8mmテープ（Exabyte）が広く用いられている．4mm DATは，転送速度は250KB/s~3MB/s，容量は1.3~20GB，Exabyte 8mmテープ/ドライブは転送速度250KB/s~3MB/s，容量は2.3~20GBである．これらに加え，ワークステーションなどでは，DLTが広く用いられている．DLTは1/2インチのサーペンタインフォーマットのテープを用い，容量15~40GB，転送速度1.25~6MBである．また，最近利用可能となったテープドライブ装置としてDLTを発展させたSuper DLT，およびIBM，HP，Seagateにより規格化が行われたLTO（Linear Tape-Open）がある．いずれも，約100GBの容量，10~15MB/sの転送速度が実現されている．さらに，ハイエンドでは，100GB以上の容量，10MB/s以上の転送速度を持つテープ/ドライブも数多く商用化されている．SonyのDIR-1000Hでは転送速度は64MB/sに，Ampex DST312では1メディア当たりの容量は330GBにも達している．

表 1.2: 三次記憶ライブラリシステムの例

システム名	メディア数	最大容量	ドライブ数
Hitachi GT0DVDE0-02D1300	150	705GB	2（DVD-RAM 片面）
Hitachi GT0DVDH0-05D1300	350	3.29TB	4（DVD-RAM 両面）
HP DAT40x6	6	120GB	1（HP DAT40）
Exabyte 220	20	400GB	2（Exabyte Mammoth）
Exabyte 480	80	1.6TB	2 or 4（Exabyte Mammoth）
STK 9740	326∼494	20.6TB	∼10（DLT7000）
STK 9310	∼6000	300TB	2∼16（STK SD-3）

三次記憶システムのライブラリシステムもまた，そのサイズは多岐にわたっている．現在，数本程度を格納するものから数千本のテープを格納するものに至るまで，様々なサイズのライブラリシステムが商用となっている．テープ 10 本程度を格納する小規模なライブラリシステムは，一般的に 1 台のドライブ装置とロボットアーム，テープラックにより構成され，テープを格納するラックとドライブの間を水平または垂直に移動し，テープのロード/アンロードを行う．大規模なライブラリシステムでは，1000 本以上のテープが格納でき，複数のドライブを備える．

光/光磁気ディスクもその容量を増しつつあるが，磁気テープと比較するとその容量は小さく，単位容量当たりのコストは大きい．磁気テープに対し，光/ 光磁気ディスクはランダムアクセスが可能である，メディアのロード後のシークが不要であるなどの利点を持つものの，テープドライブ装置に比べ，ドライブ自体の転送速度は低く，メディア当たりの容量が小さく，単位容量当りのコストた高いため，現状では，地球環境情報などの大規模三次記憶システムとして用いるには適しているとは言いがたく，多くの場合，磁気テープによるアーカイバが用いられている．

大規模アーカイバでは高速なアクセスを実現するために，データ転送速度の高いテープドライブ装置，高速動作可能なロボットアームが用いられており，非常に高価なシステムとなっている．その一方で，小規模アーカイバは近年急速にその価格を下げつつあり，近い将来コモディティ化される可能性は極めて高い．小規模アーカイバに用いられているミドルレンジのテープドライブ装置自体は既にコモディティ化が進んでおり，ワークステー

ションでは必須な周辺機器となりつつあり，またパーソナルコンピュータでの利用に対する需要も高まってきている．今後，ワークステーションやパーソナルコンピュータで扱うデータ量はますます増大すると考えられることより，このようなテープドライブ装置を用いた小規模アーカイバもコモディティ化する可能性は非常に高い．

本研究では，コモディティ化された小規模アーカイバを複数台接続することで，安価に大規模アーカイバが実現可能であることを示すとともに，三次記憶システム上のファイルに対するアクセス特性を利用した高性能化手法の確立を目的とする．プロセッサの分野においては，スーパーコンピュータ，メインフレーム，ワークステーションサーバなどにおいてはマルチプロセッサは当然のものとなっている．各ワークステーションメーカは，クライアントワークステーションで用いているものと同じプロセッサを複数利用し高性能なサーバを実現している．二次記憶システムにおいても，単一の大型磁気ディスク装置が用いられることはなく，PCや小型ワークステーションなどで用いられている小型磁気ディスク装置を複数利用し，大容量化，I/Oの高速化を実現するとともに，冗長ディスクにより信頼性を向上させるディスクアレイにより，大規模高性能二次記憶装置が実現されている．

一方，三次記憶装置の需要が高まり，注目されるようになってから日は浅く，その研究自体はまだ黎明期にあると言っても過言ではない．大規模三次記憶システム上のファイルへのアクセスの高速化に関する研究はまだ少なく，大規模三次記憶システムの構築法に関する研究はほとんど無いのが現状である．

1.2 研究の概要

本研究では，並列計算機やディスクアレイのごとく，小規模アーカイバをコンポーネントとし，これらを複数台接続して大容量な三次記憶システムの構築法を提案するとともに，この三次記憶システムにおいて高性能を実現する手法を示すことを目的とする．小規模アーカイバを利用することで大規模アーカイバを安価に実現することが可能となり，また，接続する小規模アーカイバの数を変更することで必要に応じて容量を容易に変更することが可能となる．この複数台の小規模アーカイバにより実現された大規模三次記憶シス

テムにおけるスケーラブルな性能を実現するため，大規模三次記憶システム上のファイルに対するアクセス特性を利用した高速化手法の提案が本研究の主題である．

三次記憶システムのファイルへのアクセスの特徴としては，

- ファイル内の部分参照性
- ファイル間の参照局所性

が挙げられる．ファイル内の部分参照性とは，1つのファイル全体ではなくそのファイルの一部分のみが参照されるということである．東京大学生産技術研究所で受信されている衛星画像を例にとると，衛星NOAAによる画像1シーンは約100~120MBであり，約5000km×5000kmの範囲が観測される．NOAA衛星により観測される範囲はシーン毎に異なるが，例えば衛星が東京上空を通過した場合には，東経110度から東経170度，北緯70度から北緯10度の範囲が観測される．GMSでは，画像のサイズは約100MBであり，緯度0度，東経140度を中心とし，東経70度から西経150，北緯70度から南緯70度までの範囲が観測されている．しかしながら，利用者が興味を持ち，実際に必要とする範囲は，日本や日本のごく一部など，全画像に比べて狭い範囲であり，必要とするデータ量は全ファイルの数%から数十%程度である．にもかかわらず，例え全ファイルの数%しか必要としていなくても，現状の階層ファイルシステムにおいては，マイグレーション単位をファイルとしているため，全ファイルを一度テープからディスクへマイグレートしなければならない．衛星画像をはじめ，一般的に三次記憶システムにアーカイブされるファイルのサイズは大きく，従って，ファイル全体を三次記憶装置から二次記憶装置へマイグレートするためには長い時間を要する．すなわち，実際には必要としないデータをマイグレートするために長時間を要することとなる．

この問題を解決するための手法として，ファイルをブロックに分割して管理するとともに，必要な部分のみをマイグレートすることでマイグレーションに要する時間を短縮する部分マイグレーション機能を提案した．アプリケーションプログラム自体がファイル内の必要な部分を判断し，その部分のみのマイグレートを実行する方法も考えられるが，アプリケーションプログラムがマイグレートの対象となる全てのファイルの管理をする必要がある，アプリケーション毎に変更を加えなければならないなど，マイグレーションを強

く意識したアプリケーションプログラムを作成する必要があり，現実的であるとは言えない．階層ファイルシステム自体に部分マイグレーション機能をもたせることで，部分マイグレーション機能について意識することなく，従来のファイルシステム上のファイルを扱うアプリケーションとして作成されたプログラムが，部分マイグレーション機能による恩恵を享受することが可能となる．

一方，ファイル間の参照局所性とは，各ファイルのアクセス頻度に偏りがある，すなわち，頻繁にアクセスされるファイルとほとんどアクセスされないファイルが存在するということである．生産技術研究所における気象衛星NOAAおよび気象衛星GMSによる観測画像データベースシステムを例にとると，受信システムの故障などによる欠測を除き受信可能な画像は全て受信され，受信された全画像がアーカイブされる．すなわち，観測範囲や日時，観測領域の天候等によらず，全ての画像がアーカイブされている．一方，生産技術研究所における衛星画像データベースの利用者の多くは日本国内の研究者であり，従って，受信毎に観測領域が異なるNOAA画像においては，日本が中心に観測されている画像にアクセスが多くなる．植生に興味を持つ利用者は，可視センサにより観測される太陽光の反射率によって植生を観測するため，昼間の観測データのみを解析に利用する．海洋あるいは陸地など地表面に興味をもつ研究者は，対象とする陸地あるいは海洋が雲で覆われていない画像のみを使用する．また，研究者は常に同じ地域や期間を研究対象とする訳ではない．このような理由により，アーカイブされる画像に対するアクセス頻度は画像により異なり，また，そのアクセス頻度の偏りは時々刻々変化することとなる．

複数のアーカイバが存在する状況，すなわち，小規模アーカイバを複数用いて大規模三次記憶システムを構築した場合には，アクセス頻度の偏りは応答性能の劣化を招く重大な問題となる．例えば，アクセス頻度が高いファイルが一つのアーカイバ内に集中していると，そのアーカイバは他のアーカイバよりも多くのリクエストを受けることになる．総アクセス数が少ない場合には問題ないが，アクセス数が多くなると，このアーカイバ内のドライブ装置は常に使用されることとなる．このとき，新たにこのアーカイブ内のファイルに対してアクセスリクエストが発行された場合，他のアーカイバ内のドライブ装置が使用されていないのにもかかわらず，アクセスリクエストはブロックされてしまう．すなわち，アーカイバ全体がもつリクエスト処理能力を生かせず，処理能力は飽和してしまう

こととなる．

この問題を解決する手法として，筐体間でテープの移送を行うための移送装置を用いて小規模アーカイバを1つのエレメントとし，これらを複数接続して構成されるスケーラブルアーカイバを提案するとともに，筐体間のアクセス頻度の偏りを平坦化するためのメディアマイグレーション方式を提案した．使用可能ドライブ装置を持たないエレメントアーカイバ内のファイルに対して新たなリクエストが生じた際に，そのファイルが存在するメディアを他のエレメントアーカイバ内の空きドライブ装置まで移動させてサービスを行うフォアグランドマイグレーション，および各エレメントアーカイバの過去一定期間のアクセス数を均衡化することで負荷の分散を図るとともに，フォアグランドマイグレーションをスムーズに実行可能とするために各エレメントアーカイバ間の空きスロット数を平衡化するバックグランドマイグレーションを用いることにより，全ドライブ装置が持つ処理能力を有効に利用することが可能となる．

また，三次記憶装置においては，各メディアのアクセス頻度の偏り，メディア内の各データのアクセス頻度の偏りも性能劣化を招く要因となる．テープドライブ装置では，一本のテープ上の異なるデータに同時にアクセスすることはできず，逐次的にアクセスする必要がある．つまり，あるテープ上のデータがアクセスされている場合にはそれが終了するまで，そのテープ上の他のデータへのリクエストはブロックされることとなる．少数のテープ上にアクセス頻度の高いデータが複数存在している場合には，それらを逐次的にアクセスしなければならないため，例えライブラリ内に使用されていないドライブがあってもそれを利用できず，ライブラリの持つ性能を十分に活用することができない．また，磁気テープにおいては，テープ上のアクセス頻度の高いデータの位置も問題となる．今日のメディアでは，テープ長が100mを越えるものも少なくない．このため，ヘッドをテープ上のある位置から他の離れた位置までシークさせるために要する時間は極めて長くなる．従って，例えば，頻繁にアクセスされるファイルがテープの先頭付近と終端付近に記録されている場合には，これらのファイルの間を何度もヘッドを往復させる必要が生じ，そのために長時間を要するため，応答性能が悪化することとなる．

これらの問題を解決するためには，高アクセス頻度データをなるべく異なるテープ上に配置するとともに，各テープ上においては高アクセス頻度データをクラスタリングする

必要がある．これにより，異なるテープ上のデータを同時にアクセスでき，また，一本のテープにおいては，高アクセス頻度データを連続してアクセスする場合のシーク時間を短縮することが可能となり，応答時間を短縮することができる．一般的にはファイルが作成された時点でそのアクセス頻度を予測するのは困難であり，テープ上の高アクセス頻度データ間のシーク時間を短縮するため，このような配置を実現するためには，アクセス頻度が判明した後でのデータの再配置が必要となる．

アクセス頻度が判明した後，テープ上のデータを動的に再配置する手法として，テープを巻戻すことなく，テープの途中でロード/イジェクトできるドライブを用いるという条件の下，高アクセス頻度データの複製を予め用意しておいた空き領域に作成する手法提案した．高アクセス頻度データをあらかじめ用意されていた領域へクラスタリングするために，高アクセス頻度データが連続してアクセスされるときのシーク長が短縮されるとともに，オリジナルデータと複製データを異なるテープ上に配置することにより高アクセス頻度データのアクセシビリティが向上し，応答時間の短縮が図られる．

1.3 本論文の構成

本論文は以下の構成をとる．

まず，第2章において，三次記憶システムに関する研究について触れる．特に三次記憶システムの応答性能を向上させる手法に関する研究を中心に取り上げる．

第3章では，ファイルの部分参照性を利用し，ファイルの必要とされる部分のみをテープとディスク間でマイグレートする部分マイグレーション機能を備えたファイルシステムについて説明する．ファイルシステムの構築法を示すとともに，試作システムにより基本性能を測定し，その結果を述べる．さらに，放射量補正/幾何補正プログラムと植生指数生成プログラムという衛星画像処理において実際に用いられる2つのアプリケーションプログラムを用い，実アプリケーションに対する部分マイグレーション機能の有効性を明らかにする．

第4章においては，複数の小規模アーカイバと隣接アーカイバ間でメディアの移送を可能とするメディアマイグレーション装置により構成されるスケーラブルアーカイバについ

て説明するとともに，スケーラブルアーカイバにおける負荷分散手法であるホットデクラスタリングの有効性について述べる．シミュレーションによりスケーラブルアーカイバの基本性能を示し，ホットデクラスタリングの有効性を明らかにするとともに，ドライブ故障時，およびファイルストライピング環境下のシミュレーション結果を示し，このような状況においてもホットデクラスタリングが有効であることを示す．

第5章では，高アクセス頻度ファイルをクラスタリングすることでシーク時間を短縮するとともに，ファイルの複製による多重アクセスによる応答時間の短縮を図り，応答性能を向上させるホットレプリケーションについて説明する．高アクセス頻度データの複製作成法について述べるとともにシミュレーションによりその効果を明らかにする．

第6章では，東京大学生産技術研究所において運用されている衛星画像データベースシステムに対するアクセス履歴を用い，ホットデクラスタリング，ホットレプリケーションの有効性を示す．まず，衛星画像データベースシステムについて簡単に述べ，衛星画像データベースシステムへのインターネットを通じたアクセスの履歴について説明をする．その後，このアクセス情報に基づきアーカイブシステムのシミュレーションを行うことで，ホットデクラスタリング，ホットレプリケーションが実システムに対しても有効であることを示す．

最後に第7章において，本論文についてまとめ，今後の課題について述べる．

第2章　関連研究

2.1　従来の三次記憶システムの高性能化手法

三次記憶システムに対する関心は高まりつつあるものの，三次記憶システムや三次記憶システムに用いられるドライブ装置に関する研究は未だ少ない．しかしながら，これまでにいくつかの三次記憶装置の高性能化に関する研究が発表されている．本章では，これらの研究について説明する．

2.1.1　三次記憶システム一般の高性能化手法

I/O スケジューリング

三次記憶システム上のデータへのアクセスのために要する時間では，メディアの交換やテープ上のデータ先頭位置までのシークに要する時間が大きな割合を占める．そこで，メディア交換やシーク時間が削減されるようにアクセス順序を変更することにより，応答時間の短縮を図ることが可能である．I/O スケジューリングは，三次記憶システムに限らず，ディスクシステムにおいても用いられている高性能化手法である．

B. K. Hillyer，A. Silberschatz は論文 [22] においてサーペンタインテープドライブ Quantum DLT4000 の詳細なモデルを構築し，さらに論文 [23] にてこのモデルを用いてシミュレーションを行い，いくつかの I/O スケジューリング法の比較を行っている．最適なスケジューリングは非対称巡回セールスマン問題であり，NP 完全問題であるが，LOSS と呼ばれる巡回セールスマン問題の貪欲近似解法を用いたスケジューリング法により，計算コストを抑えつつ最適値に極めて近い応答性能が得られ，I/O リクエストの実行順序を変更しない場合に対して応答時間が短縮可能であることが示されている．さらに，論文 [25] では，サーペンタインテープドライブ IBM 3570 Magstar MP に関してシミュレーションを行い，同様の結果を得ている．

O. Sandstå, R. Midtstraum は，論文 [45] において，サーペンタインテープドライブを対象としたMPScan*と呼ばれるスケジューリング法を提案した．リクエストされているデータを，記録されているトラックの読み込み方向で二分し，それぞれを物理的な位置でソートして正方向，逆方向の順に一往復（パス）で全てをアクセスするSCANアルゴリズムを元に，トラック間の移動コストを避けるために複数のパスで全てをアクセスするようにしたMPScan（Multi Pass Scan）アルゴリズムがあるが，MPScan*はさらにMPScanに改良を加えたアルゴリズムである．MPScan*は，MPScanによりスケジューリングされた各パスのリクエストされているデータをトラック移動のコストに基づき他のパスへ挿入してパス数を減らし，性能の向上を図るものである．シミュレーションおよびTandberg MLR1ドライブ装置を用いた実験を行い，LOSSと同程度，あるいは実験条件によってLOSSよりも良い結果が得られることを示している．

B. K. Hillyer, R. Rastogi, A. Silberschatz は論文 [21] において，Envelope と呼ばれる，ヘリカルスキャンテープドライブ装置をはじめとするロジカルアドレスとテープ上の物理位置が線形に対応するテープドライブ装置を有するテープジュークボックスにおいて，レプリカが存在する場合のスケジューリング法を提案した．Envelope では，まず，レプリカを持たないデータに対するリクエストをスケジューリングし，その後，レプリカを持つデータに対するリクエストを，サービスに要するコストに基づきスケジュールに加えていくアルゴリズムである．シミュレーションにより有効性を示すとともに，提案したスケジューリング法を用いてレプリカの配置についても検討し，高アクセス頻度データを1本のテープにまとめて，各テープの後方にレプリカを配する配置法が良いという結論を得ている．

データストライピング

データを分割して複数のメディアに分散配置し，複数のドライブ装置により同時にI/O処理を実行することでスループットを高める手法である．データストライピングもまた，三次記憶システム固有の手法ではなく，ディスクシステムを含むI/Oシステム一般において用いられる手法である．

A. L. Drapeau, R. H. Katz は論文 [16] において，ドライブ4台を持つExabyte社製EXB-

120 ライブラリ装置およびドライブ 8 台を持つ Ampex 社製 DST600 ライブラリ装置上においてデータストライピングを用いた場合の性能評価を行っている．大きなファイルへのアクセスにはストライピングは有効であるが，多くの小さいファイルに対してランダムにアクセスを行う場合には，ストライピングを行わない場合と比較してより多くのメディア交換が必要となるために，性能劣化が見られることを示している．

L. Golubchik，R. Muntz は論文 [20] において，複数のサイズのデータが混在する環境下においてストライプ幅を変化させるなど，論文 [16] と比べより一般的な環境下でのデータストライピングの評価を行っている．各リクエストに対する最適なストライプ幅はワークロードに強く依存すること，また，全てのデータを同じ幅でストライピングした場合に最も性能が向上することが示されている．

データ配置の最適化

三次記憶システムにおいては，メディアを交換する時間，またテープドライブ装置の場合にはシークに要する時間は，全応答時間に対して大きな割合を占める．従って，三次記憶システム上の高アクセス頻度データを少数のメディアへクラスタリングすることでデータへアクセス時のメディアの交換を削減し平均応答時間を短縮する，あるいはテープドライブの場合に，テープ先頭へ高アクセスデータを配置し，平均シーク距離を短縮することで平均応答時間を短縮することが可能である．

S. Christodoulakis，P. Traintafillou，F. A. Zioga は論文 [4] において，ディスクライブラリおよびテープライブラリにおけるデータの最適配置を解析的に求めている．ディスクライブラリではアクセス頻度の高いデータから順に 1 つのメディアへ配置し，これを繰り返し行い全データを配置するとともに，全ドライブ数-1 台のドライブにアクセス頻度の高いメディアをマウントしたままとし，他の 1 台のドライブで残りのメディアを交換してアクセスリクエストに応じることが最適であるとしている．テープライブラリでは，テープのシークコストが大きいため，テープのロード/イジェクト時に巻戻す必要のあるドライブでは各テープの先頭，テープ途中でイジェクト可能なドライブの場合には各テープの中央に高アクセス頻度データを格納する配置が最適であるという結論を導いている．

Log-structured File System（LFS）

LFS[43] はディスクシステムの書き込み性能を向上させる目的で開発された手法であるが，書き込みがアペンドで行われるという特徴を利用し，三次記憶への拡張が行われている．

J. T. Kohl，C. Staelin，M. Stonebraker は LFS を三次記憶へ拡張し HighLight と呼ばれるファイルシステムを開発した [28]．HighLight では二次記憶と三次記憶を用い，三次記憶は二次記憶上から溢れたデータのバックアップとして用いられる．マイグレートは二次記憶上のセグメントを単位として行われる．読み出しの際，データが二次記憶上にない場合には三次記憶から所望のデータを含むセグメントが一度二次記憶上にマイグレートされ，アクセスが行われる．三次記憶上のガベージコレクションは実装されていない．

D. A. Ford，J. Myllymaki による三次記憶を用いた LFS では，データは三次記憶のみに記録される [18]．データの書き込みのためにセグメントバッファが用いられ，このセグメントバッファが満たされると書き込みドライブ内のメディアへデータがマイグレートされる．読み出しは未使用ドライブを用い，データは直接三次記憶から読み出される．ガベージコレクションは対象となるメディア全体を解放することにより行われ，具体的にはガベージコレクション対象メディア上の有効データ全てを書き込みドライブ内のメディアへコピーすることにより行われる．

2.1.2 特定データ，アプリケーションに対する高速化手法

結合演算

現在の商用のデータベース管理システムでは，リレーションはディスク上に存在すると仮定しており，三次記憶は単にバックアップ用デバイスとして位置づけられている．このため，三次記憶上にリレーションが存在する場合には，その処理は極めて低速なものとなる．

S. Sarawagi，M. Stonebraker は論文 [49] において，2 つのテープ上のリレーション間の結合演算について検討を行っている．ドライブ装置が 1 台，2 つのリレーションは異なるメディア上に存在し，リレーションのサイズよりディスク容量が小さい場合を想定してい

る．複数の結合演算方式について，三次記憶の遅延による影響を減らすためにキャッシュに可能な限りデータを読み込むとともに，タプルIDリストを用いてI/Oリクエストを前処理することで，ランダムI/Oをシーケンシャル I/Oに変換し，実行時間を短縮している．

J. Myllymaki，M. Livnyは論文[32]，[33]において，それぞれディスク上のリレーションとテープ上のリレーション，およびテープ上のリレーション同士の結合演算について，メインメモリ量を変化させて様々なアルゴリズムの性能や必要ディスク量の検討を行い，また，ディスクI/Oとテープ I/Oの並列実行の効果についても検討している．低速デバイス上に大きなリレーションがある場合には，それを外部リレーションとすることで実行時間が短縮される．また，データの読み込みをダブルバッファ化し，ディスクアクセスとテープアクセスをオーバラップすることにより，実行時間が短縮可能であることも示した．

データベース管理システム

三次記憶システムを考慮していないデータベース管理システムでは，問合せ実行，ディスクキャッシュ管理，三次記憶デバイス管理などのプロセスは独立したコンポーネントとして実装され，また，問合せはその発行順に処理が行われるものが多い．このため，データが三次記憶上のデータを読み込む際に，多数の三次記憶システムメディア交換やシークが必要となってしまい，性能が劣化してしまう．

S. Sarawagi，M. Stonebrakerは，問合せ処理，ディスクキャッシュ，三次記憶の状態についての情報を有する統合的スケジューラを持つデータベース管理システムのアーキテクチャを提案し，三次記憶システム上のデータに対する問合せの高速化を行っている[47]，[46]，[50]．スケジューラは，三次記憶システムから読み込むべきデータ，ディスクが溢れたときに三次記憶へ書き込むべきデータ，次に実行すべき問合せを決定する．これにより三次記憶システムのI/Oを減らし，高速化を図った．

多次元配列データ

科学データなどの多次元データは一般的に，格子をある方向から直線的に読み込んだ順に三次記憶装置に記録される．一方，多次元データが参照されるときには，より小さい次元の全データ内の一部の領域が参照されることが多い．すなわち，参照される領域は三次

記憶上において広く散在することとなり，シーク距離やメディア交換の増大を招き，性能が劣化することとなる．

この問題を解決するため，S. Sarawagi，M. Stonebraker は論文 [48] において，エンドユーザおよびDBMS から収集した利用パターンに対して最適化した多次元配列データ格納法を提案した．三次記憶装置としては光ディスクを対象としている．まず，各読み出しに対して読み込まれるブロック数が最小となるような単位（多次元タイル）に分割し，それらをシークが最小となるように並べ換える．さらに異なるパターンの読み出しに対応するためにコピーを作成し，メディア交換回数が最小となるように配置することで，応答時間の短縮を図っている．

一方，L. T. Chen，D. Rotem，Arie Shoshani，B. Drach，M. Keating，S. Louis は論文 [2]，[3]，[1] においてテープドライブを対象とし，多次元配列データの格納法を述べている．まず同時にアクセスされる変数をグループ化し，グループ内の変数をデータがアクセスされる最小基本単位に分割し，各問合せにおいて同時にアクセスされる基本単位間の全距離が最小となるように並べ，隣り合う基本単位を接続してファイルする．その後，マウント時間やシーク時間などのハードウェアのパラメータより計算される各問合せの重要度により重み付けされた平均応答時間が最小となるように各ファイルを三次記憶上の各メディアに配置することで，応答時間を短縮している．

文書アーカイブ

Web ページなどのハイパーテキスト化された文書では，各ページ間にリンクが張られており，多くの場合，利用者はこのリンクに従って文書を参照する．すなわち，ある文書がリクエストされた場合，次にリクエストされうる文書の予測が可能である．

A. Kraiss，G. Weikum は論文 [29] において，ディスクをキャッシュとしたテープアーカイブシステムにおける文書データのプリロード（プリフェッチ），キャッシュリプレースアルゴリズムについて述べている．まず，アクセス履歴よりアーカイブされている各文書間の遷移をマルコフモデルに基づいてモデル化する．このモデルに基づき，ある文書がリクエストされた場合に，各文書がその後にリクエストされる確率を求め，データサイズ，転送時間などの情報と統合し，三次記憶からキャッシュへプリロードすべき文書，および

キャッシュ上のリプレースすべき文書を決定し，キャッシュ性能を高めて応答時間の短縮を図っている．

2.2 本論文提案手法との関連

本論文では，ファイル内の部分参照性を考慮した高速化手法として，部分マイグレーション手法，ファイル間の局所参照性を考慮した高速化手法としてホットデクラスタリング手法，およびホットレプリケーション手法の提案を行っている．本節では，これらの手法と従来の三次記憶システムの高速化に関する研究との関連について述べる．

ファイルの部分参照を考慮した三次記憶システムの高速化に関する研究としては，多次元配列データの格納法 [48][2][3][1] があげられるのみである．これらは，予測されるアクセスパターンを元にあらかじめ原ファイルを複数のサブファイルに分割してアーカイブする方法を導入しているが，データのマイグレートはやはりこの分割されたファイルを単位としており，ファイル内の任意の部分のみをマイグレートすることはできない．このため，予測されたアクセスと異なる場合には，やはりユーザが必要としない部分もマイグレートされることになる．現在の商用の階層ファイルシステム [19][41] に目を向けてみても，ファイル内の部分参照性は全く考慮されておらず，ファイル内の一部分を読み書きするためにもファイル全体をテープからディスクへマイグレートする必要がある．また，米国の UC Berkeley による SEQUOIA 2000 プロジェクト [51][52] や，NASA などのいくつかの研究機関による EOS（Earth Observing System）プロジェクト [55] の一環として研究・開発されている衛星画像データアーカイブのための大規模ファイルシステムでもデータのマイグレーションの単位はファイルであり，ファイル内の部分参照性は考慮されていない．

一方，ファイル間の参照局所性に基づく動的なファイルの再配置に関連する研究としては，S. Christodoulakis らによる最適データ配置 [4]，B. K. Hillyer らが検討を行ったレプリカの配置 [21] があげられるが，示された配置とするために，全データを新たに記録し直す必要があり，そのためには膨大な時間，作業用空間を必要とする．一方，本論文で扱う手法は，テープ上に記録されているデータに対して，高アクセス頻度データの複製のみをあらかじめ確保しておいた領域に書き込むため，全データを読み込み，再び書き戻す必要

はない．地球環境データやマルチメディアデータなどの大規模三次記憶システムに格納されるファイルのアクセス頻度は，アクセスが行われてはじめてその頻度が明らかになり，また，時々刻々と変化する場合も少なくないため，アーカイブ時に最適な配置にすることは困難である．

テープの移送という点に関しては，いくつかの商用のテープアーカイバでは複数の筐体間でテープの移送が可能となっており，Storage Tecknology 社の NEARLINE シリーズでは，パススルーとよばれる装置を筐体間に設けることで，筐体間のテープの移送が可能となっている [15]．また，ソニーの PetaSite では，カセットの出し入れを行うコンソール，ドライブを搭載するコンソール，カセットラックのみのコンソールをコンポーネントとして直線上に組み合わせて設置することで容量，ドライブ数の拡張が可能であり，さらに縦横のカセット搬送路間でのカセットの受渡しを可能とするジャンクションコンソールによって，垂直方向に分岐，延長が可能となっている [14]．しかしながら，これらの機構は，単にテープ移動の省力化や，設置場所に柔軟に対応することに主眼が置かれており，これらのアーカイバを利用するためのソフトウェアにおいては，その負荷状況に応じて自動的にテープを他の筐体に移送するというような機能はサポートされていない [13][19][41][17]．

ディスクアレイにおいては，参照局所性を利用した高速化手法は数多く研究されており，本論文において用いた熱と温度の概念も G. Copeland らにより，ディスクアレイ上においてファイルを格納するディスクを決定するために導入されたものである [6]．また，G. Weikum らは，さらにこの研究を発展させ，データが作成された場合や拡張された場合のディスクアレイ上のファイルの動的なデータ配置法を提案した [54]．しかしながら，これらの研究はディスクアレイ上でのデータの配置法に関するものであり，テープアーカイバとディスクアレイでは，アクセス方法や速度など異なる点が多く，ディスクアレイでの結果をテープアーカイバにそのまま適用することはできない．テープアーカイバにおけるテープの配置手法についての研究はなく，商用のアーカイバにおいてもテープを移送するための機構が備わっているものが存在するにもかかわらず，それにより負荷分散を図るためのアルゴリズムは未だ確立していない．

また，第 2.1 節において説明し，これまで本研究との関連を述べていない高速化手法に関しては，本論文で述べる部分マイグレーション，ホットデクラスタリング，ホットレプ

リケーションとは基本的に独立した高速化手法であるため，本論文で提案する手法と同時に適用することが可能であり，それによりさらなる高速化が図れると考えられる．

第3章　部分マイグレーションによる三次記憶システムの高速化

3.1　はじめに

本章では，衛星画像データを対象に8ミリテープ装置を用いて，利用者の必要とする部分のみをマイグレートすることを可能とする三次記憶ファイルシステム（PFS : Partially Migratable File System）について述べる．一般に，大規模なデータを扱う処理においては，利用者が処理の対象とするのはファイル全体ではなく，ファイルの一部分である場合が多い．そのため，ファイルシステムが必要な部分のみをテープからディスクへマイグレートすることが可能であれば，そのI/O性能は大きく改善すると期待される．さらに，必要な部分のみのマイグレーションが実現されれば，不要な部分がディスクにキャッシュされることもなくなるため，ディスクの利用効率も向上する．まず，3.2節においてPFSの特徴的な機能を述べ，その後，3.3節でPFSの構成，動作例を示す．さらに，3.4節において，PFSのテープのシーク時間，ファイル読み込み速度などの基本性能を示した後，衛星画像処理においてしばしば利用される2種類のアプリケーション，放射量/幾何補正処理，NDVI作成処理をPFS上で実行し，従来のファイル単位のマイグレーションに対して大幅な処理時間の短縮が達成されることを示す．

3.2　PFSの特長

以下，三次記憶ファイルシステムPFSにおける特徴的な機能を以下に述べる．

3.2.1 部分マイグレーション機能

現在の階層ファイルシステムではマイグレーションの単位はファイルであり，ファイル全体をマイグレーションするために長時間待たねばならず，低い性能しか得られていない．一般に，大きなファイルを扱うアプリケーションでは必ずしもファイル全体をアクセスする必要はなく，その一部のみをアクセスするだけで十分であることが多い．

動画像や衛星画像，地球環境の数値モデルのグリッドデータなどの大規模なファイルでは，データ構造は単純であることが多く，ある決まった構造の繰り返しで構成される傾向にある．例えば，衛星によるリモートセンシング画像は，固定数あるいは任意数のラインによって構成されている．ラインは衛星に搭載されたこのセンサによる観測データであり，決まった構造を有している．従って，ファイル内の任意のラインの位置を容易に決定することが可能であるために，ファイルシステムがテープからディスク，あるいはディスクからテープへ，あるラインを選択的に転送することを可能とすれば，ファイル内の一部のみを処理するアプリケーションプログラムの処理時間は大幅に短縮される．いいかえれば，ファイルの先頭からの相対位置によって指定されたファイルの一部をマイグレートすることが可能なファイルシステムにより，より高速に既存のアプリケーションを実行することが可能となる．

上述のような，ファイルシステムがファイルの一部のみをテープからディスク，ディスクからテープへマイグレートする機能を部分マイグレーション機能と呼ぶ．部分マイグレーション機能を実装することを目的としてPFSを開発した．

3.2.2 高速シーク機構

SunOS 4.1.3で提供されている磁気テープのデバイスドライバによるシークでは，目的のファイルまで，その前にあるファイルマーク毎に一つずつ順にシークを行う．すなわち，ファイルマーク毎にテープが停止することになり，テープ上の目的のファイルまで直接シークすることができない．このため，シークに非常に長い時間を要することとなる．また，ファイルの先頭位置へのシークしかサポートされておらず，ファイル内の任意の位置へのシークはできないため，ファイルの先頭部のデータが不要な場合でも，ファイルの先

頭よりデータを読み込まなければならない．Solaris 2.5 などのいくつかの最新のオペレーティングシステムで提供されているデバイスドライバでは，任意のファイルの先頭位置まで停止することなくシークさせることは可能となっているものの，やはりファイル内の任意の位置までシークをさせることは不可能である．

この問題を解決するため，独自に 8 ミリテープドライブ用のデバイスドライバを開発した．Exabyte 8500 は 75 倍速のポジショニングをサポートしており，この機能を利用し，高速に任意のファイル内の任意の位置にまで直接シークを行うことを可能とするデバイスドライバを開発した．

3.2.3 透過的なアクセス手段の提供

PFS では，利用者への負担を軽減するために，従来のファイルシステム（UFS : Unix File System）と同様のインタフェースが提供されており，また，マイグレーションは完全に利用者から隠蔽されている．すなわち，利用者はデバイスやファイルシステムの内部動作を意識せずに PFS 上のデータにアクセスできることができる．

3.2.4 プリフェッチ

多くの大規模ファイルを処理するアプリケーションプログラムは，ファイルの一部分にアクセスするのみであるが，その一部分をアクセスする際にも，全てを一度に読み込むわけではなく，ある決まった単位でシーケンシャルにアクセスし，逐次的に処理を行う場合が多い．従って，ユーザプログラムがあるブロックを処理している間，ファイルシステムが次のブロックをテープからディスクへマイグレートするプリフェッチ機能は非常に効果的であると考えらる．PFS でも，このプリフェッチ機能を採用し，アプリケーションプログラムの実行と三次記憶装置に対する I/O 処理をオーバーラップさせ，アプリケーションの実行時間の短縮を図っている．

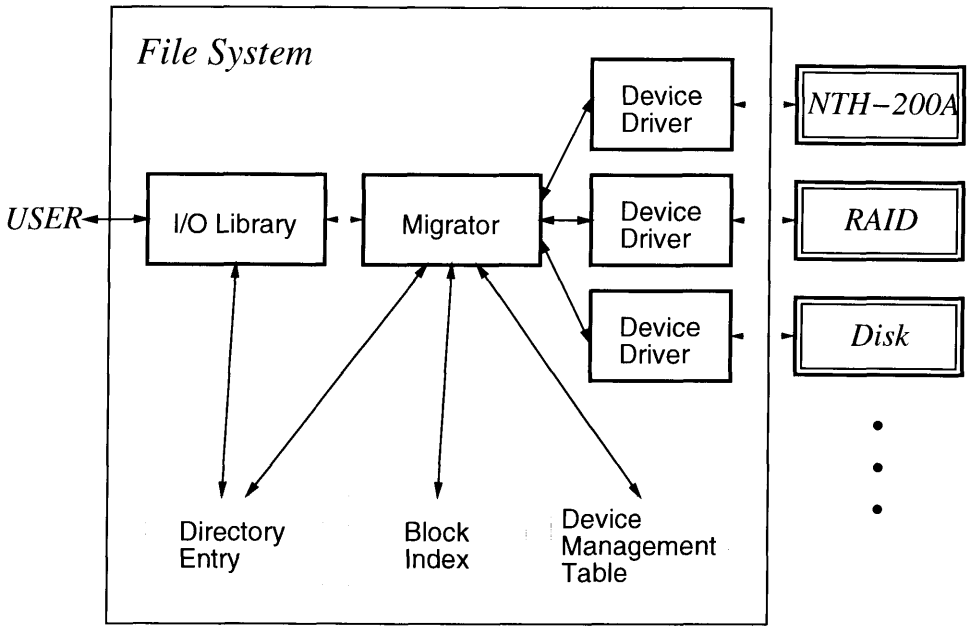

図 3.1: PFS のソフトウェア構成

3.2.5 ストライピング

各ジュークボックスは2つのドライブをもつことを利用し，データ転送速度を上げるためにデータを2つに分割して異なるテープに記録する2ウェイストライピングを採用している．ロボットアームは1つしかないため，テープのロードは逐次的に行わなければならないものの，データの転送は並列に行うことが可能である．

3.3 部分マイグレーションファイルシステム PFS の構成

3.3.1 ソフトウェア構成

PFS のソフトウェア構成を図 3.1 に示す．PFS は，大きく I/O ライブラリ，マイグレータ，デバイスドライバの3つにより構成される．

I/O ライブラリは，PFS 上のデータへの透過的なアクセスを提供する．ユーザはこの PFS 専用の I/O ライブラリを用いている限りにおいて，UFS 上のファイルと同様なアクセスが可能となる．OS のレベルで PFS を完全に UFS と見せるように実装することも可能ではあるが，OS に強く依存した実装となるために移植性が悪く，I/O ライブラリによる手法を

採用している．I/Oライブラリで提供されるPFSへのアクセスのための関数はマイグレータとソケット通信を行うことでI/O処理を行うが，その内部動作は利用者には隠蔽されており，インタフェースはUFSにおけるopen，read，writeなどのシステムコールと同様になっている．すなわち，利用者はこれらのシステムコールをPFS用のI/Oライブラリの関数に変更するだけで，PFS上に存在するをファイルであることを意識せず，UFS上のファイルと同様にアクセスすることが可能となる．

マイグレータはサーバプロセスとして動作し，PFS上の全てのデバイスを管理し，また，I/Oライブラリを使用しているクライアントプログラムの要求に応じて部分マイグレーションを実行する．このために，マイグレータはディレクトリエントリ，ブロックインデックス，デバイス管理テーブルを使用する．

PFSにおいて，UFSと同様の階層ディレクトリ構造，ファイル名を実現するため，PFSに対するディレクトリエントリとしてUFS上に同名のファイルが作成される．ファイル名，ユーザID，グループID，ファイルモード，ファイル更新時刻等は，それぞれ，このUFS上のディレクトリエントリのものが使用される．また，PFSにおける各ファイルのディレクトリは，UFSにおけるディレクトリエントリファイルのディレクトリがそのまま対応する．従って，UFSにおける階層ディレクトリ構造がそのままPFS上で実現される．各エントリファイル内には，エントリファイルであることを示す識別子，ファイルインデックス番号，ファイルサイズが記述されている．ファイルインデックス番号はUFSにおけるiノードに対応し，PFSにおけるファイルのポインタを示す．ファイルサイズはPFS上のファイルのサイズである．以上の手法によりPFSでは，UFSと同様の階層ディレクトリ，ファイルが提供され，利用者はUFSと同様にディレクトリ名，ファイル名によって容易にファイルを扱うことができる．

PFSでは，全てのファイルはブロックに分割されて管理され，部分マイグレーションはこのブロックを単位として実行される．ブロックインデックスはPFS上の全ブロックに関する情報をもち，部分マイグレーション実行時に，要求されたブロックがどのデバイス上に存在するのかを判断するために用いられる．

デバイス管理テーブルは各デバイス毎に1つ存在し，それぞれのデバイスの状態を格納している．マイグレータはデバイス管理テーブルにより，各デバイスの空き領域をチェッ

クし，二次記憶から三次記憶へデータをマイグレートするかを決定する．

デバイスドライバは各デバイスに対する透過的なアクセス法を提供する．利用者はI/Oライブラリを通し，これらのデバイスにアクセスすることになる．

3.3.2 PFS上のファイルへのアクセス時の動作例

以下に，アプリケーションプログラムがPFS上へのファイルの読み込み/書き込み時のPFSの動作の例を示す．

- 読み込み時の動作例

  1. I/Oライブラリをリンクしたクライアントプログラムが，読み込み要求メッセージをマイグレータへ送る．
  2. マイグレータはディレクトリエントリから，クライアントプログラムから受け取った要求に対応するブロックのインデックス番号を取り出し，ブロックインデックスを参照する．
  3. 要求されたブロックが二次記憶上に存在する場合は，そのブロックの情報をI/Oライブラリへ返す．
  4. 要求されたブロックが二次記憶上に存在しない場合は，マイグレータはまず，要求されたブロックを三次記憶から二次記憶上へマイグレートし，その後，二次記憶上のそのブロックに関する情報をI/Oライブラリへ返す．
  5. クライアントプログラムはマイグレータより返されたブロック情報をもとにI/Oライブラリを通してデータを読み込む．同時にマイグレータは読み込みが行われているブロックに続く次のブロックを三次記憶から二次記憶へマイグレートする．

- 書き込み時の動作例

  1. I/Oライブラリをリンクしたクライアントプログラムが，書き込み要求メッセージをマイグレータへ送る．

2. マイグレータはディレクトリエントリから，クライアントプログラムから受け取った要求に対応するブロックのインデックス番号を取り出し，ブロックインデックスを参照する．

3. 要求されたブロックがPFS上に存在している場合，すなわち，利用者が上書きを要求した場合，マイグレータは必要ならばそのブロックを三次記憶から二次記憶へマイグレートし，二次記憶上のブロックの情報をI/Oライブラリへ返す．

4. 要求されたブロックがPFS上に存在しない場合，マイグレータはデバイス管理テーブルを参照して二次記憶上の空き領域をさがし，そのブロックの情報をI/Oライブラリへ送る．

5. クライアントプログラムはマイグレータより返されたブロック情報をもとにI/Oライブラリを通してデータを書き込む．

3.4 PFS試作システムによる性能評価

以下に，PFS試作システムの環境について述べる．その後，テープのシーク，シーケンシャルリードなどの基本性能を測定する．

3.4.1 実験環境

図3.2に実験環境の構成を示す．三次記憶装置として2台のExabyte 8500 8ミリテープドライブをもち，200巻のテープを格納可能なテープジュークボックスを用いている．各テープドライブはSCSIインタフェースによりメインメモリ64MBのSUN SPARCstation 10/41に接続される．NOAA衛星画像データは1シーンが約100MBであるので，容量が約5GBであるテープ1巻に50シーンが格納可能である．従って，8ミリテープジュークボックス1台の容量は1TBになり，約10000シーンのNOAA衛星画像データを蓄積できる．最新のExabyte 8ミリテープドライブは圧縮機能を備えているが，NOAA画像に対して圧縮は効果が少なく，本実験においては圧縮機能を用いていない．

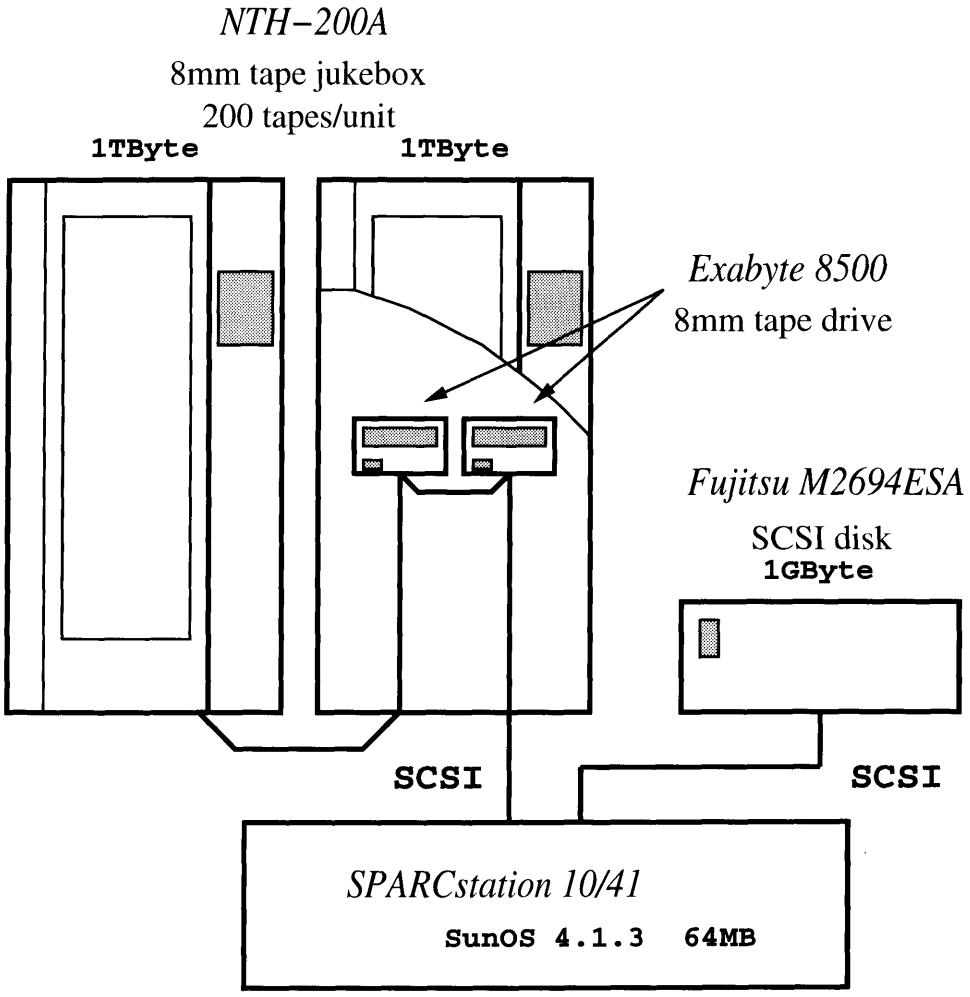
NTH-200A
8mm tape jukebox
200 tapes/unit
1TByte
1TByte
Exabyte 8500
8mm tape drive
Fujitsu M2694ESA
SCSI disk
1GByte
SCSI
SCSI
SPARCstation 10/41
SunOS 4.1.3 64MB

図 3.2: 実験環境

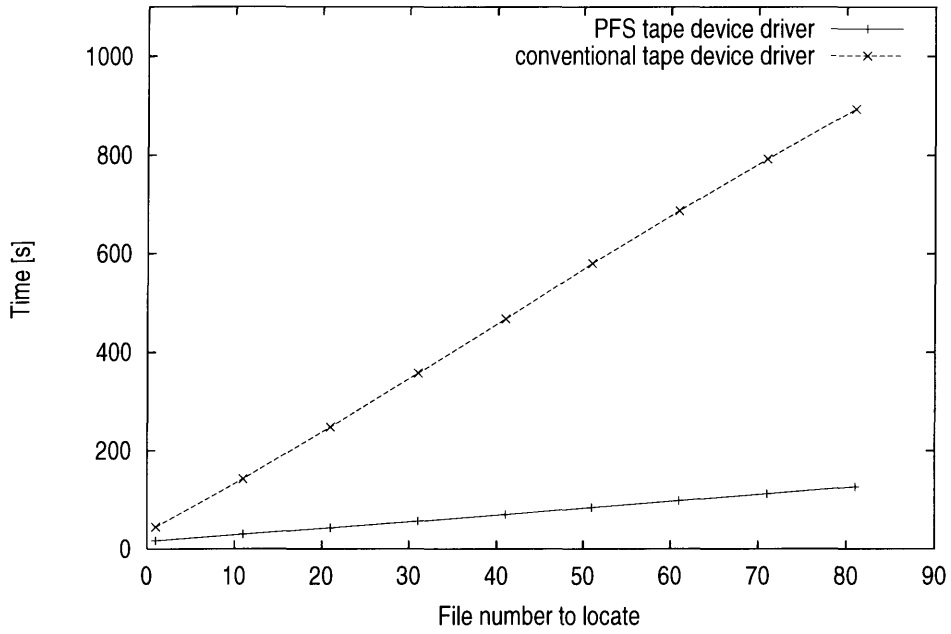

図 3.3: シーク時間の比較

3.4.2 基本性能測定

シーク操作

図 3.3 に，我々が開発した SCSI テープデバイスドライバを用いた場合と，SunOS 4.1.3 標準の SCSI テープデバイスドライバで mt コマンドを用いた場合の，テープの先頭から目的のファイルまでのシークに要する時間を示す．この実験では，テープ上に 50MB のファイルが順に繰り返し記録されている．我々のシステムでは 2 ウェイストライピングを行っているため，100MB の画像は 1 本のテープの上では 50MB となる．横軸はテープの先頭からのファイル数，縦軸はテープ先頭からそのファイル先頭までシークさせた場合に要する時間である．SunOS 4.1.3 のデバイスドライバでは各ファイルマーク毎に一度停止するため，シークは非常に低速である．一方，我々が開発したデバイスドライバでは，目的の位置まで停止することがないため，高速なシークが行えることが判る．

表 3.1: 100MB のファイルの読み込み時間

ファイルシステム	時間（秒）	転送速度
シングルドライブ	220	465 KB/sec
2 ウェイストライピング	116	883 KB/sec

ストライピング

本システムで用いた 8 ミリテープジュークボックス内の 2 台の Exabyte テープドライブを用いて 2 ウェイストライピングを行った場合と，2 ウェイストライピングを行わずにドライブ 1 台のみを使用した場合の，100MB のファイルの読み込み時間を表 3.1 に示す．ストライピングを行うことで読み込み時間が約半分になることが判り，大きなファイルを扱うアプリケーションでは例えロボットが 1 台でもストライピングは有効であることが判る．

3.5 衛星画像処理アプリケーションを用いた PFS の性能評価

試作ファイルシステム上において 2 つの代表的な衛星画像処理アプリケーションプログラムを実行させ，部分マイグレーション機能の効果を明らかにする．一方は，指定された領域を切り出し，放射量補正/幾何補正を行うものであり，他方は指定領域を切り出し，植生指数（NDVI:Normalized Differential Vagitation Index）を求めるものである．この 2 種類のアプリケーションを用いたのは，これら 2 つの処理はいずれも衛星画像処理においては頻繁に行われる処理であり，また，前者は I/O 処理に対して計算処理の負荷が大きく，これに対して後者はその逆の特性を有し，対称的なアプリケーション特性を有している．

表 3.2 に，実行時のパラメータを示す．NOAA 衛星画像の 1 ラインは 2048 画素であり，この他に各ライン毎に観測時刻や較正用の情報などが含まれているため，1 ラインのデータ量は 22180 バイトになる．1 シーンのライン数は画像によって異なるが，実験に用いた画像は 4297 ラインであり，このライン数は我々がアーカイブしている画像の平均的な値である．

PFS ブロックサイズは，PFS でのファイルを管理する際のブロックサイズで，1MB である．部分マイグレーションもこの PFS ブロックサイズを単位として実行される．ディス

表 3.2: 実験パラメータ

NOAA 衛星画像	
ファイルサイズ	95307460 bytes
1 ラインのサイズ	22180 bytes
ライン数	4297 lines

ファイルシステム	
PFS ブロックサイズ	1 MB
ディスクブロックサイズ	256 KB
8 ミリテープブロックサイズ	256 KB

クブロックサイズ，8 ミリテープブロックサイズは，それぞれディスク，8 ミリテープに対してデータを読み書きする際のブロックサイズであり，256KB とした．

3.5.1 放射量・幾何補正

放射量・幾何補正プログラムでは，まず，10bit のセンサー出力値を画素毎に温度に変換し，その後，幾何補正処理ルーチンによって歪みを含んだ画像を，ポーラステレオ図法やメルカトール図法などの指定された地図画像に変換する．図 3.4 はこのプログラムの実行時間を示している．横軸は抽出領域を緯度・経度で表している．例えば，5 度は東京を中心とし，緯度，経度ともに ±5 度の領域の地図画像を作るということを示す．出力地図画像のサイズは，抽出領域のサイズによらず，常に 512×512 画素である．縦軸は実行時間を表す．

従来のファイルシステムでは，まずファイル全体をテープからディスク上に移す必要があり，その後，処理プログラムがディスク上の画像データを処理することになる．図 3.4 の “conventional file system” がこのときの時間を表す．一方，ファイルが既にディスク上にある場合には，テープからディスクへのファイルのマイグレーションは不要であり，この場合の実行時間は “file on disk” に示される．“PFS with prefetch” に示される線は，PFS において 1 画像ブロックをプリフェッチした場合の実行時間である．この場合の処理時間はファイルがディスク上にある場合の処理時間とほぼ同じであることが判る．放射量・幾何

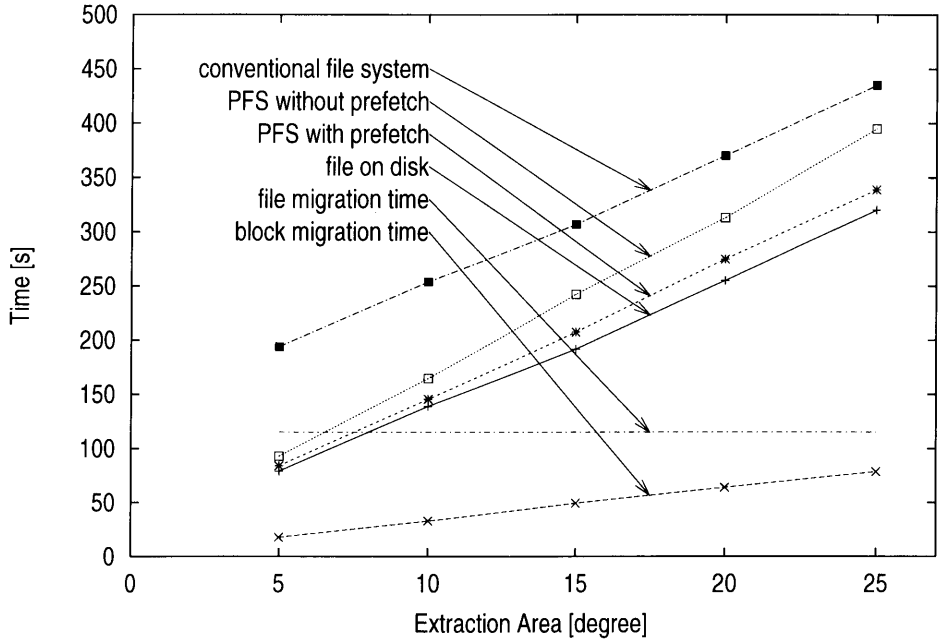

図 3.4: 放射量・幾何補正プログラムの実行時間

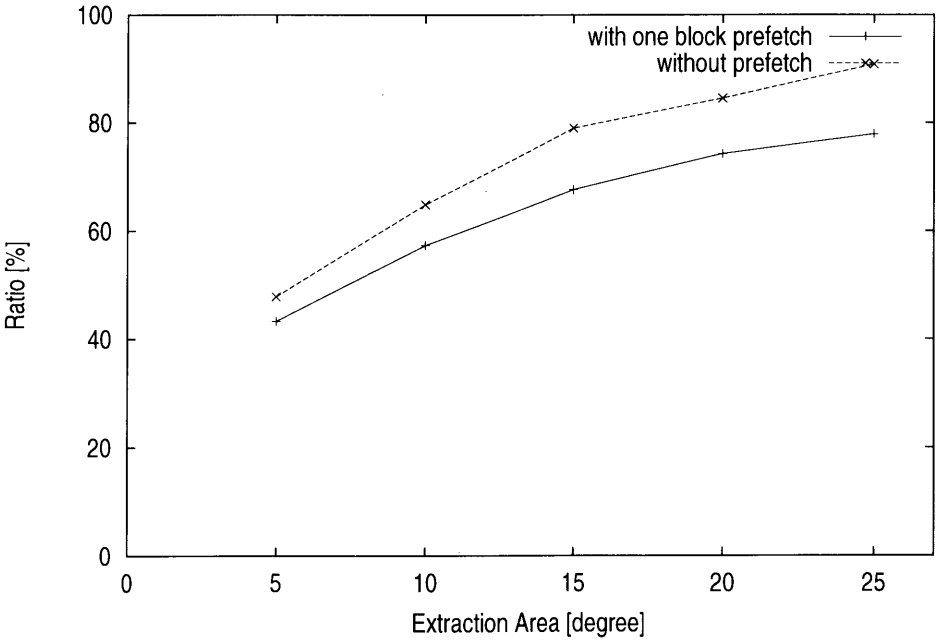

図 3.5: 放射量・幾何補正プログラムの相対実行時間

補正処理はI/O処理に比べて計算処理の比重が大きく，プリフェッチが有効である．“PFS without prefetch”にPFSにおいてプリフェッチを用いなかった場合の処理時間が示されているが，この結果と比較すると，プリフェッチが非常に効果的に働いていることが判る．

図3.5は，全ファイルをマイグレートした場合に対する相対的な実行時間を示している．抽出領域が全ファイルに対して小さい場合には，部分マイグレーション機能により大幅な実行時間の短縮が実現されていることが判る．

3.5.2 植生指数生成

正規化差植生指数（NDVI:Normalized Differential Vegetation Index）は，光合成活動の程度を示す指標であり，陸上における植生の変化を観測するために広く用いられている．NDVIは，可視センサの出力と，近赤外センサの出力より以下の式で求められる．

$$\mathrm{NDVI} = \frac{\mathrm{NIR} - \mathrm{VIS}}{\mathrm{NIR} + \mathrm{VIS}} \tag{3.1}$$

VIS ： 可視センサ（チャネル1）の出力値

NIR ： 近赤外センサ（チャネル2）の出力値

図3.6は，NDVI生成プログラムの実行時間を示している．横軸はNDVIを生成するライン数を表し，縦軸は実行時間を表す．“file migration time”に示される約120秒の水平な線はファイル全体を8ミリテープからディスク上にマイグレートするのに要する時間であり，“file on disk”はファイルがディスク上にある場合のNDVI生成に要する時間である．従来のファイルシステムでは，ファイルがテープ上にある場合には，全てのファイルをマイグレートし，その後，ディスク上のデータの処理を行うため，これらの時間を合計した“conventional file system”に表される時間となる．“read file from the disk”はディスクから指定された抽出領域を読み込むのに要する時間を示しているが，“file on disk”と“read file from the disk”はほぼ同じ値であることより，I/O処理がこのNDVI生成プログラムの全処理の大部分をしめていることが判る．“PFS with prefetch”に示される線は，PFSを用いて部分マイグレーション機構を使用した場合の実行時間である．“read file on 8mm tape”にファイルをテープから読み込む処理のみを行った場合の実行時間を示しているが，この値

はPFSを用いた場合の値との差が極めて小さく，理想的性能を達成していることが判る．また，ブロックプリフェッチが効果があることも判る．

図3.7は全ファイルをマイグレートした時の実行時間に対する，部分マイグレーション機能を用いた場合の実行時間を示している．相対実行時間は処理ライン数にほぼ比例しており，PFSを用いることで処理時間の大きな短縮を図れることが判る．

3.6 まとめ

大規模データを対象とし，8ミリテープロボティクスを用いた三次記憶ファイルシステムの設計，試作を行った．部分マイグレーション機構を提案し，試作システム上で2種類のアプリケーションプログラムを実行することによりその性能評価を行った．放射量・幾何補正プログラム，NDVI生成プログラムいずれにおいても，実行時間が大幅に短縮されることを示し，部分マイグレーションの有効性を示した．加えて，プリフェッチ機能，2ウェイストライピング，高性能テープデバイスドライバが，三次記憶ファイルシステムの性能向上に貢献することを示した．利用者の観点では，従来のファイルシステムと同様のインタフェースを有する専用I/Oライブラリをリンクするだけでで PFSを用いることができ，その機構を考慮する必要はない．

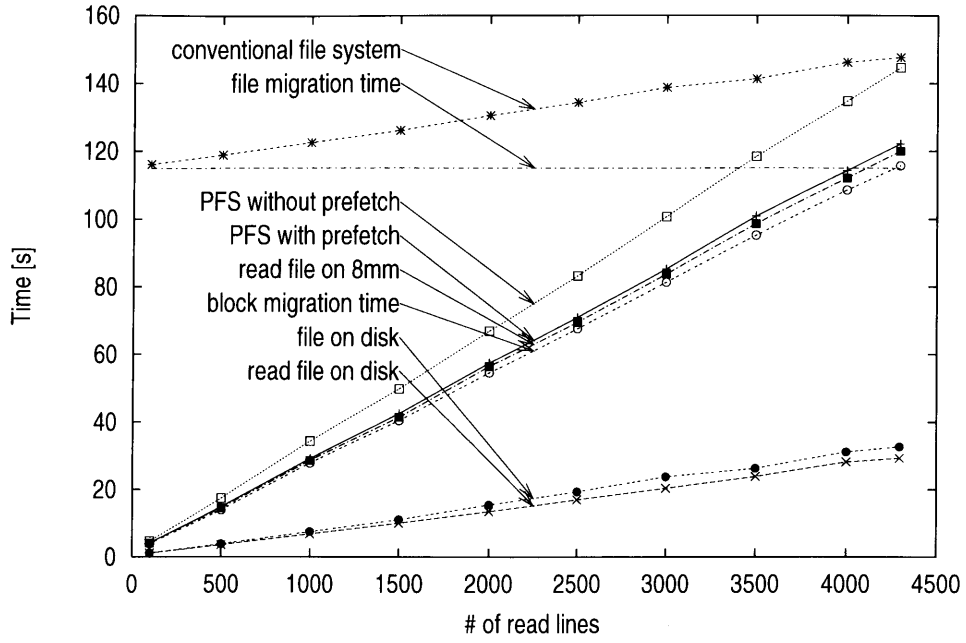

図 3.6: NDVI 生成プログラム実行時間

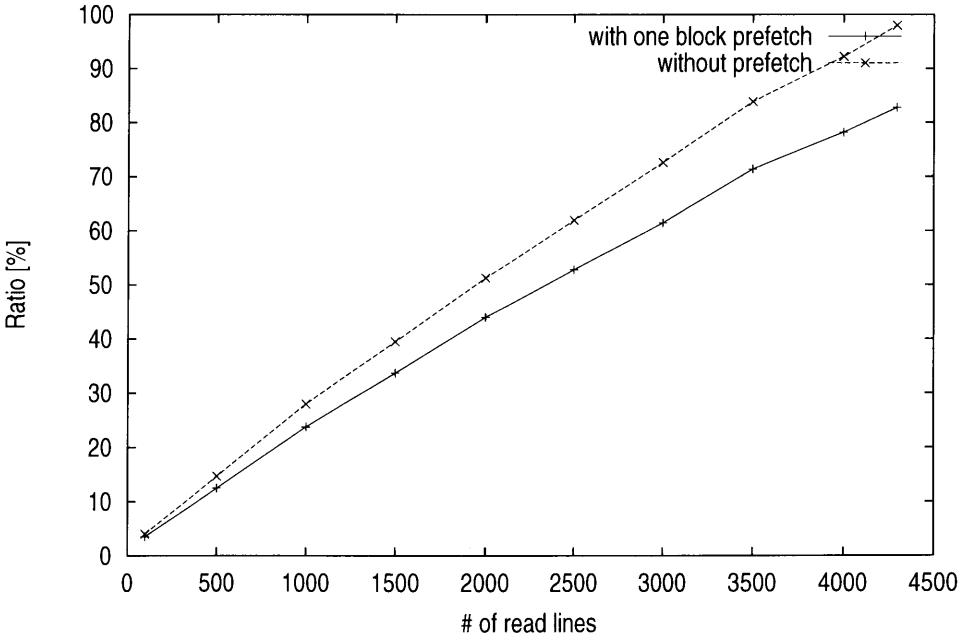

図 3.7: NDVI 生成プログラム実行時間

第4章　ホットデクラスタリング：テープマイグレーションを用いた負荷分散による三次記憶システムの高速化

4.1　はじめに

本節では多数台の小規模アーカイバにより構成されるスケーラブルテープアーカイバについて述べるとともに，スケーラブルテープアーカイバ上で負荷分散を行うホットデクラスタリングについて説明を行い，シミュレーションによりその有効性を明らかにする．まず，4.2節でスケーラブルテープアーカイバについて説明を行い，4.3節においてホットデクラスタリングが用いる熱と温度の概念について述べ，4.4節でホットデクラスタリングによる負荷分散手法について詳細に説明する．その後，4.5節においてシミュレーションによりホットデクラスタリングの基本性能を明らかにし，4.6節にてテープドライブが故障した場合のホットデクラスタリングの効果，4.7節でファイルストライピング環境下でのホットデクラスタリングの効果について示す．

4.2　スケーラブルテープアーカイバの構造

複数のエレメントアーカイバと呼ばれる小規模テープアーカイバにより構成され，それらのエレメントアーカイバ間で物理的にテープの移送が可能であるテープアーカイブシステムをスケーラブルテープアーカイバと呼ぶ．スケーラブルテープアーカイバは，エレメントアーカイバ数を変更することにより，任意の規模のテープアーカイバを構成できるという特徴を有する．

4.2.1 スケーラブルテープアーカイバ実装システム

システム構成

図4.1にエレメントアーカイバとしてNCLコミュニケーション社製のNTH-200Bを用いて構成したスケーラブルテープアーカイバを示す．エレメントアーカイバNTH-200Bは2台のExabyte社製EXB-8505 8mmテープドライブ，1台のテープハンドラロボット，200スロットのテープラックをもつ．更にテープハンドラロボットとテープマイグレーション装置を制御するためのコントローラをもち，RS-232Cによりホストコンピュータと接続される．コントローラはホストコンピュータからのコマンドに応じてテープハンドラロボット，テープマイグレーション装置を制御する．テープドライブとホストコンピュータとはSCSIバスにより接続され，通常の8mmテープドライブと同様にテープドライブ用のコマンドによって制御される．

テープの移送方法

テープマイグレーション装置はテープを乗せるためのワゴンをもち，このワゴンが隣り合うエレメントアーカイバ間を移動することでテープを移送する．具体的なテープの移送手順を以下に示す．

1. テープマイグレーション装置のワゴンが移送元のエレメントアーカイバ内に存在しない場合，ワゴンを移送元へ移動させる．

2. 移送元のテープハンドラロボットが移送すべきテープをラックあるいはテープドライブからテープマイグレーション装置のワゴンへ移す．

3. 移送元エレメントアーカイバから隣接するエレメントアーカイバへテープマイグレーション装置のワゴンを移動させる．

4. テープが最終移送先エレメントアーカイバへ到達していない場合は，テープハンドラロボットが移送されたテープを次のマイグレーション装置のワゴンへ移し，テープが最終移送先エレメントアーカイバへ到達するまで3~4を繰り返す．

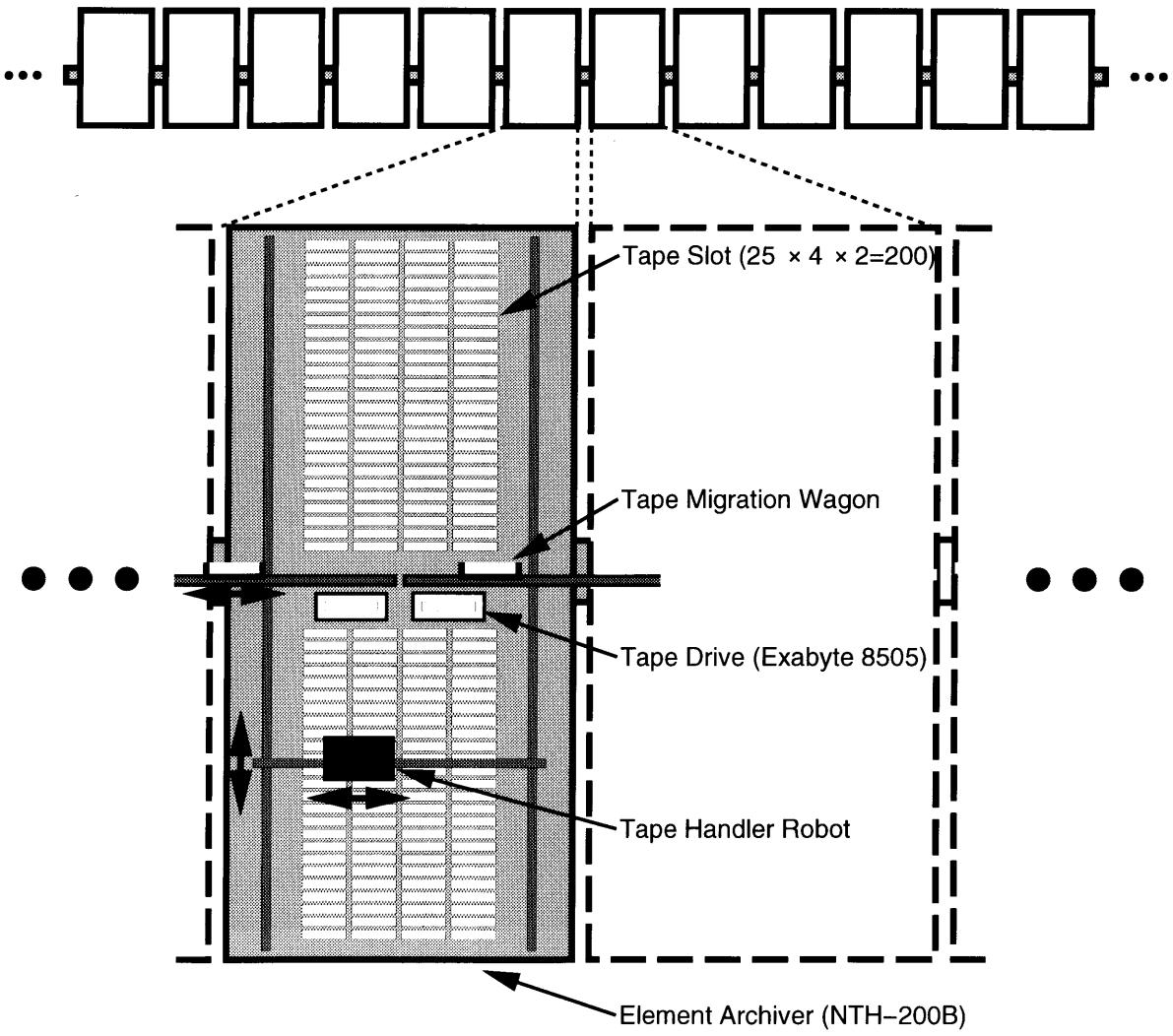

図 4.1: スケーラブルテープアーカイバの構成

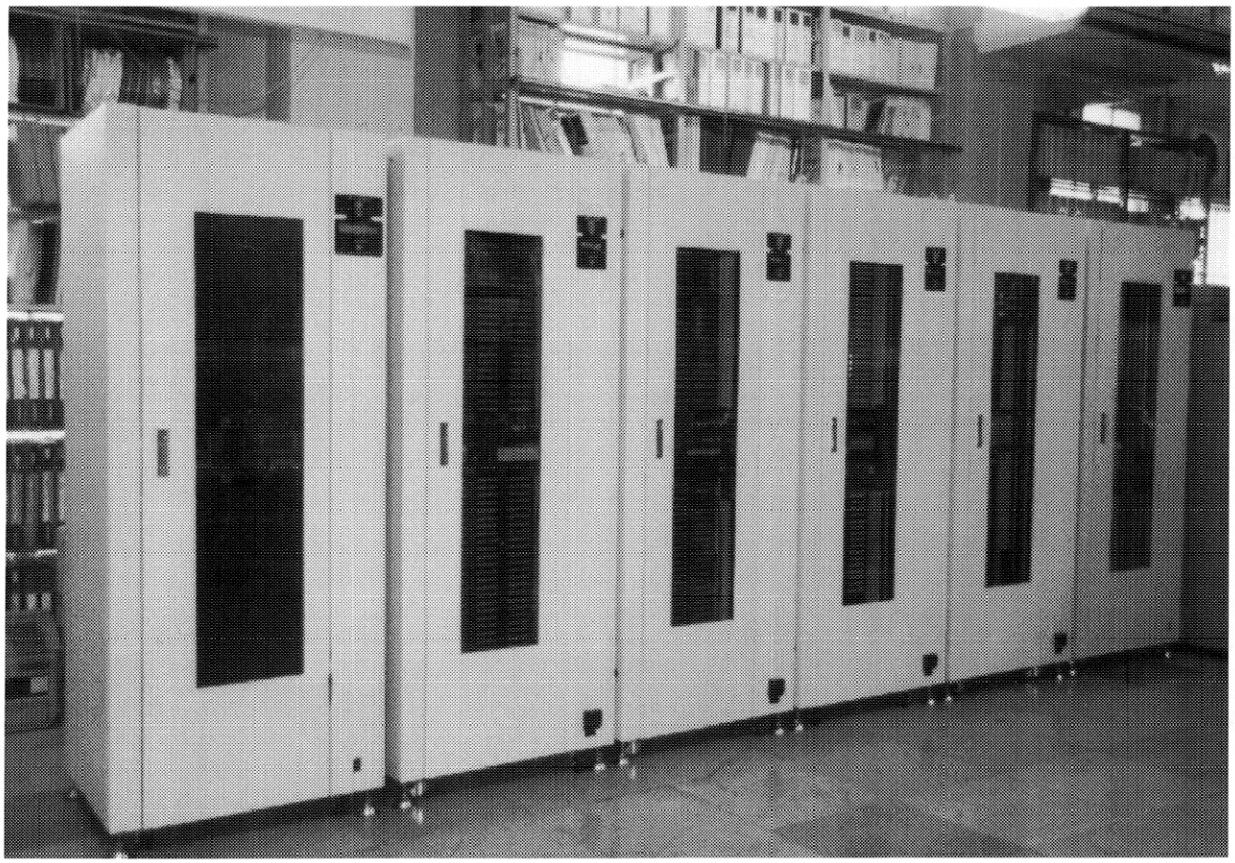

図 4.2: 試作スケーラブルテープアーカイバ

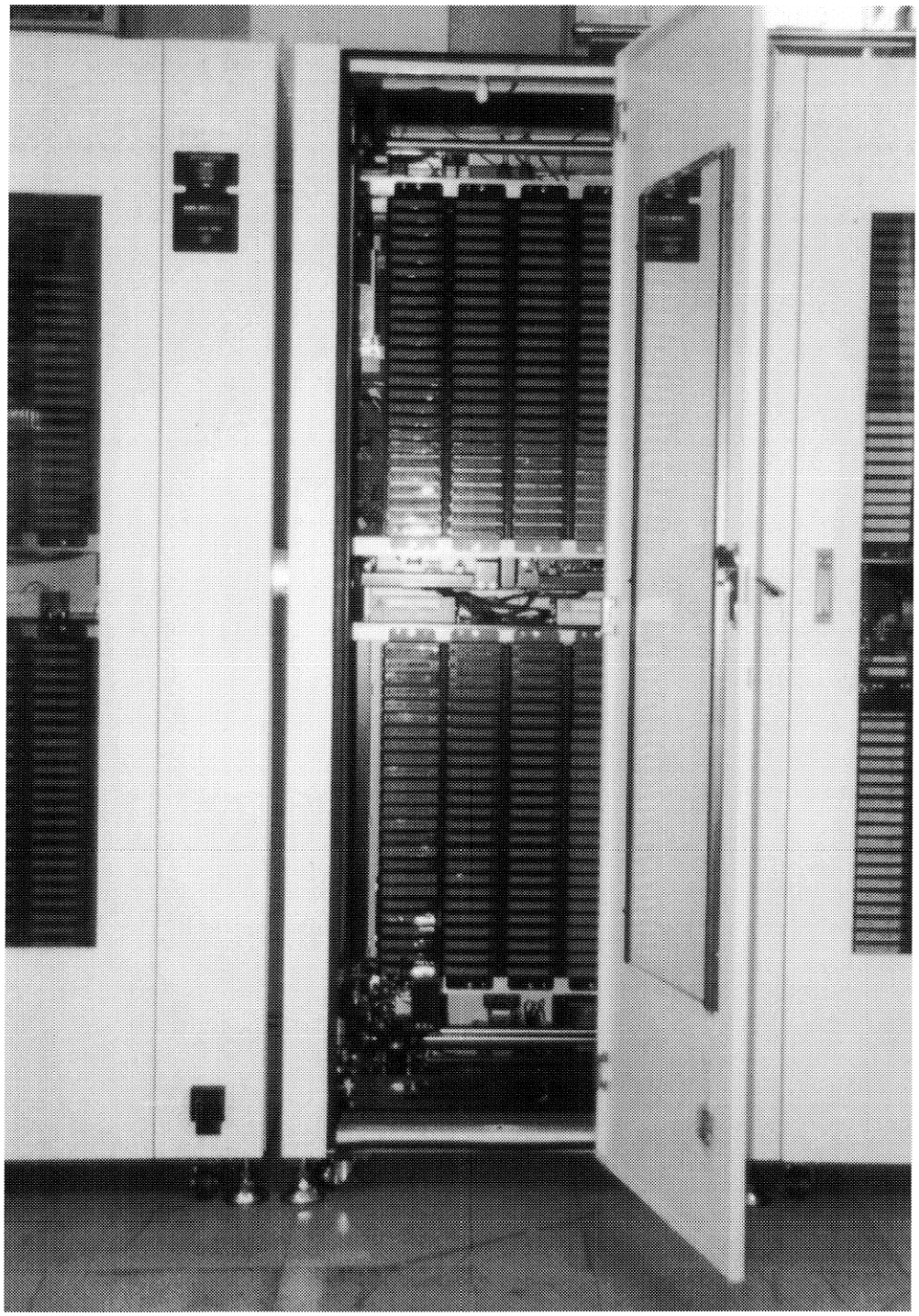

図 4.3: 試作エレメントアーカイバ（NTH-200B）

5. 移送先のテープハンドラロボットがテープマイグレーション装置のワゴンから，ラック，テープドライブへテープを移す．

NTH-200B では，テープマイグレーション装置とテープハンドラロボットは物理的に干渉する場合があり同時に動作させることができないため，上述の操作は逐次的に実行する必要がある．

4.3 熱，温度

G. Copeland らによりディスクアレイ上のデータ配置に対して導入された熱と温度の概念 [6] を拡張し，スケーラブルテープアーカイバにおける熱と温度を定義する．論文 [6] ではデータの熱はある一定期間のそのデータに対するアクセス数の合計と定義され，データの温度はそのデータの熱をデータサイズで割ったものと定義されている．この定義をそのままスケーラブルテープアーカイバに適用するならば，テープ，エレメントアーカイバの熱はそれぞれ，そのテープあるいはエレメントアーカイバ内の全データの熱の総和となる．また，テープの温度はテープの熱をそのサイズで割ったものとなる．

スケーラブルテープアーカイバにおいては，各エレメントアーカイバ内のテープドライブ数が異なる場合には，各エレメントアーカイバの熱をそのエレメントアーカイバ内のドライブ数で割り，熱を正規化する．エレメントアーカイバの熱はそのエレメントアーカイバの負荷を抽象化したものであり，従って，エレメントアーカイバの処理能力が異なる場合にはその処理能力で正規化する必要がある．試作スケーラブルテープアーカイバにおいては，各エレメントアーカイバはそれぞれ 2 台のテープドライブをもつが，そのうちの 1 台が故障した場合，その処理能力は 1/2 となるため，このエレメントアーカイバの熱を 2 倍に再定義することでこのような状況に対処することが可能となる．以降では，エレメントアーカイバの熱とは，この正規化された熱のことを言う．ディスクアレイでは一般的に各コンポーネントディスクは同種のものが用いられておりその処理能力は等しく，また，ヘッドは 1 つしかないため，熱の正規化の必要はない．

また，論文 [6] では温度はディスクアレイにおけるデータマイグレーションのコストパフォーマンスを評価するための尺度として導入されている．即ちデータのサイズをその

データを移動させるためのコストとみなし，小さいコストでより大きな熱の移動が可能となるデータの選択に用いている．しかしながら，スケーラブルテープアーカイバではテープを移動させるコストはテープによらず常に一定であり，従って，スケーラブルテープアーカイバではテープの温度は熱と同義とする．

4.4 ホットデクラスタリング

高温のテープが一部のエレメントアーカイバに集中すると，それらのエレメントアーカイバはより多くのアクセス要求を受けることになり，低温のエレメントアーカイバ内にアイドル状態のテープドライブが多数存在するにも関わらず，スケーラブルテープアーカイバ全体の応答時間は劣化する．アクセスの集中を減らし，テープドライブを効率的に利用するために，フォアグランドマイグレーションとバックグランドマイグレーションの2種類のテープマイグレーション機構をスケーラブルテープアーカイバに導入して負荷を分散し，応答性能の向上を図る．

4.4.1 フォアグランドマイグレーション

あるエレメントアーカイバ内のテープドライブがすべて使用されているときに，そのエレメントアーカイバ内のテープに対し新たなアクセス要求が生じた場合，アイドル状態のテープドライブをもつエレメントアーカイバへテープを移送し，そのテープドライブを使用することにより応答時間を短縮することが可能である．このようなテープマイグレーション方式をフォアグランドマイグレーションと呼ぶ．

フォアグランドマイグレーションを実行するためには，移送先エレメントアーカイバ内のテープドライブが使用されていないこと，移送先エレメントアーカイバに空きスロットが存在すること，移送元，移送先，中継エレメントアーカイバのテープハンドラロボットおよびテープマイグレーション装置が使用されていないことが必要である．移送先として複数の候補が存在する場合の選択方針としては，

Heat Balancing 熱が最小のエレメントアーカイバを選択する

Space Balancing ラックの空きスロット数の最大のエレメントアーカイバを選択する

Distance Minimizing 移動距離が最小となるエレメントアーカイバを選択する

Random ランダムに選択する

が考えられる．

4.4.2 バックグランドマイグレーション

ある2つのエレメントアーカイバおよびその間に存在するエレメントアーカイバのテープハンドラロボットがすべて使用されておらず，また，それらの間のテープマイグレーション装置も使用されていない場合，その2つのエレメントアーカイバ間でテープを移送することによりエレメントアーカイバのテープ数や熱を平衡化することができる．このようなマイグレーションをバックグランドマイグレーションと呼ぶ．一般に三次記憶装置にアーカイブされるファイルのサイズは大きく，例えば我々が現在格納している衛星画像データベースシステムでは各画像ファイルのサイズは約100MBであり，テープハンドラロボットによるテープの操作に比べテープドライブによるデータの読み書きには長時間を要する．この間，テープドライブはビジーであるものの，テープハンドラロボット，テープマイグレーション装置はアイドル状態にあり，テープを移送することが可能である．ある2つのエレメントアーカイバ間でバックグランドマイグレーションを実行する場合，常にテープ数の多いエレメントアーカイバからテープ数の少ないエレメントアーカイバへテープを移送する．また，移送されるテープに関しては，2つのエレメントアーカイバ間の熱の差を最小化するテープを移送する．例えば，移送元のエレメントアーカイバの熱が大きい場合，温度の高いテープを移送し，逆に移送元の熱が小さい場合，低温のテープを移送する．

バックグランドマイグレーションは2つのエレメントアーカイバのテープハンドラロボットおよびその間のテープマイグレーション装置が使用されておらず，それらのエレメントアーカイバ間の空きスロット数または熱に偏りがある場合に実行される．しかしながら，この偏りに敏感にしすぎると必要以上にバックグランドマイグレーションを実行することとなるため，熱および空きスロット数の差にしきい値を設け，必要以上にバックグランドマイグレーションが実行されるのを防ぐ．また，実行可能なバックグランドマイグ

表4.1: シミュレーションパラメータ

エレメントアーカイバ	
全エレメントアーカイバ数	16台
最大テープ数	200本/台
テープドライブ数	2台/台
テープドライブ	
ロード時間	35秒
シーク速度	25MB/秒
リード/ライト速度	0.5MB/秒
イジェクト時間	20秒
テープハンドラロボット	
移動時間（テープの操作なし）	2秒
移動時間（テープの操作あり）	14秒
テープマイグレーション装置	
ワゴンの移動時間	9秒

レーションの候補が複数存在する場合，実行するバックグランドマイグレーションの選択方針として，移送元，移送先の空きスロット数の差が最も大きいものを優先する方法，移送元，移送先の熱の差が最も大きいものを優先する方法が考えられる．

4.5 基本性能評価

4.5.1 シミュレーション条件

スケーラブルテープアーカイバにおけるマイグレーション方式の基本性能を評価するためシミュレーションを行い，リクエストが発行されてから読み込みが終了するまでの時間の平均をとった平均応答時間およびフォラグランドマイグレーション数，バックグランドマイグレーション数を測定する．シミュレーションパラメータを表4.1に示す．これらの値は4.2節において説明した試作エレメントアーカイバNTH-200Bに基づいて決定した．また，各データのサイズはNOAA，GMSによる画像の平均的なサイズである100MBとし，1本のテープには48のデータが記録されているものとしている．1データの読み込み

表 4.2: カセットテープの初期分布

エレメントアーカイバ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
高アクセス頻度テープ	8	8	8	8	8	88	88	88	88	88	88	8	8	8	8	8
低アクセス頻度テープ	182	182	182	182	182	102	102	102	102	102	102	182	182	182	182	182
合計	190	190	190	190	190	190	190	190	190	190	190	190	190	190	190	190

時間は 200 秒であり，従って平均サイクル時間は 487 秒である[1]．また，カセットテープを移送するためには移送先エレメントアーカイバに空きスロットが必要となるため，ロードファクタを 95%に設定している．アクセス要求の到着時間間隔は負の指数分布に従うとし，アクセスローカリティについては，アクセス頻度の高いテープとアクセス頻度の低いテープの 2 種類が存在し，スケーラブルアーカイバ全体で 80/20 則，すなわち全テープの 20%のテープに対して 80%のアクセス要求が生じるものとした．シミュレーション開始時のテープの初期分布は表 4.2 に示すように，中央の 6 台エレメントアーカイバにアクセス頻度の高いテープを多く配置した．

また，4.5.3 節においてフォアグランドマイグレーション移動候補先の選択方式による影響，4.5.4 節においてマイグレーション移動距離の影響，4.5.5 節においてバックグランドマイグレーション起動条件に対するしきい値の影響について評価を行っているが，この結果によるとこれらのパラメータが結果に与える影響は小さいため，他のシミュレーションにおいては，フォアグランドマイグレーションにおける移送先候補の選択方針としては，熱が最も小さいエレメントアーカイバを選択する手法を用い，また，フォアグランドマイグレーションにおける最大移送距離を 5 とし，バックグランドマイグレーションに関しては，起動条件のしきい値として空きスロット数の差を 3，熱の比を 1.2 に設定し，バックグランドマイグレーションにおけるテープの移送距離を常に 1 としている．

4.5.2 テープマイグレーションの効果

図 4.4 はシミュレーション開始後から 50000 アクセスまでの平均応答時間を示している．フォアグランドマイグレーション，バックグランドマイグレーションのいずれも実行しな

[1] 移動時間（テープ操作なし）＋移動時間（テープ操作あり）＋ロード時間＋シーク時間＋リード/ライト時間＋シーク時間（巻戻し）＋イジェクト時間＋移動時間（テープ操作なし）＋移動時間（テープ操作あり）

い場合と比較し，フォアグランドマイグレーションを実行することにより平均応答時間は大きく短縮される．フォアグランドマイグレーションに加えてバックグランドマイグレーションを導入することにより平均応答時間は更に短縮される．フォアグランドマイグレーションのみを用いた場合には，フォアグランドマイグレーション，バックグランドマイグレーションの両者を用いた場合に比べ，リクエスト到着率が高まるにつれて急速に応答時間が悪化するが，これはフォアグランドマイグレーションのみでは移送されるテープの受け入れ先エレメントアーカイバの空きスロット数が不足した状態の解消が困難であることによる．低温エレメントアーカイバはアイドル状態である時間が長く，フォアグランドマイグレーションによるテープ移送先となりやすいために容易に空きスロットがなくなり，また，リクエストを多く受けないために移送元エレメントアーカイバとなりにくく，テープ数が減少しないことが性能劣化の原因と考えられる．バックグランドマイグレーションを用いることで，空きスロット数の偏りが生じた場合にはテープが移送されることによりこの様な状況が解消され，応答性能が改善される．

図4.5は，図4.4のシミュレーションにおけるマイグレーション数を表している．リクエスト到着率が高くなるにつれてマイグレーション数は増加しているが，ある到着率より高くなるとフォアグランドマイグレーション数は減少する．これは，リクエスト到着率が高いときには各エレメントアーカイバはそのエレメントアーカイバ内のテープに対して発せられた大量のリクエストに対応するために，他のエレメントアーカイバからテープを受け入れる余裕がなくなることによる．

図4.6はリクエスト到着率が126リクエスト/時のときの2000アクセスごとの平均応答時間を示し，同様に図4.7は2000アクセスごとのマイグレーション数を示している．バックグランドマイグレーションを用いることにより収束時間が大きく短縮され，アクセスローカリティの変化に素早く追従可能なことがわかる．フォアグランドマイグレーション，バックグランドマイグレーションを共に使用した場合，シミュレーション開始直後に頻繁にバックグランドマイグレーションが実行され，初期状態の集中した熱が分散されていくことが図4.7からわかる．一方，フォアグランドマイグレーションのみを用いた場合には，シミュレーション開始後，一度平均応答時間が増加した後，530秒付近へ緩やかに収束している．これは，シミュレーション開始時には各エレメントアーカイバは空きスロットを

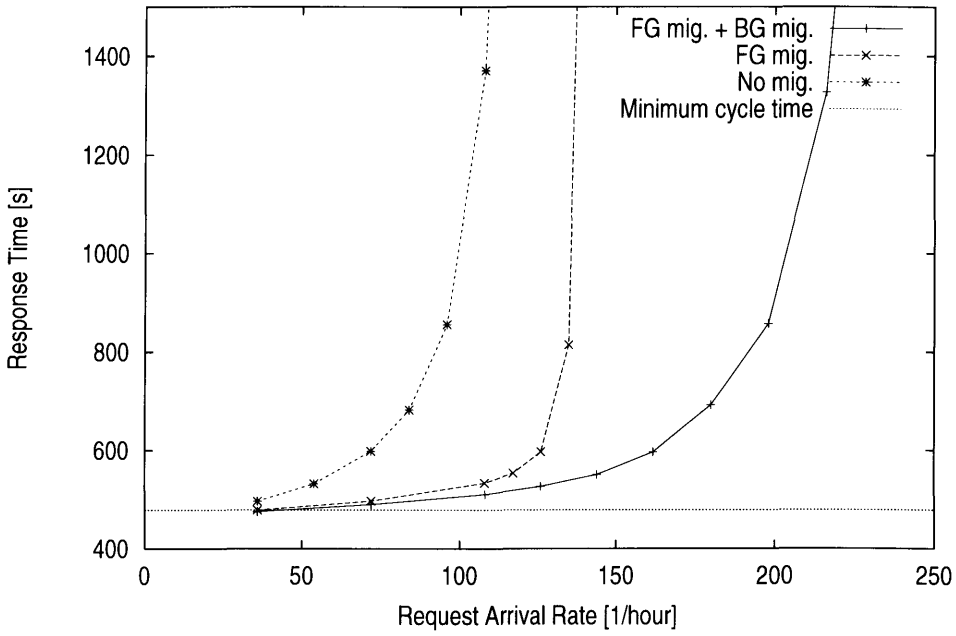

図 4.4: 50000 アクセスの平均応答時間

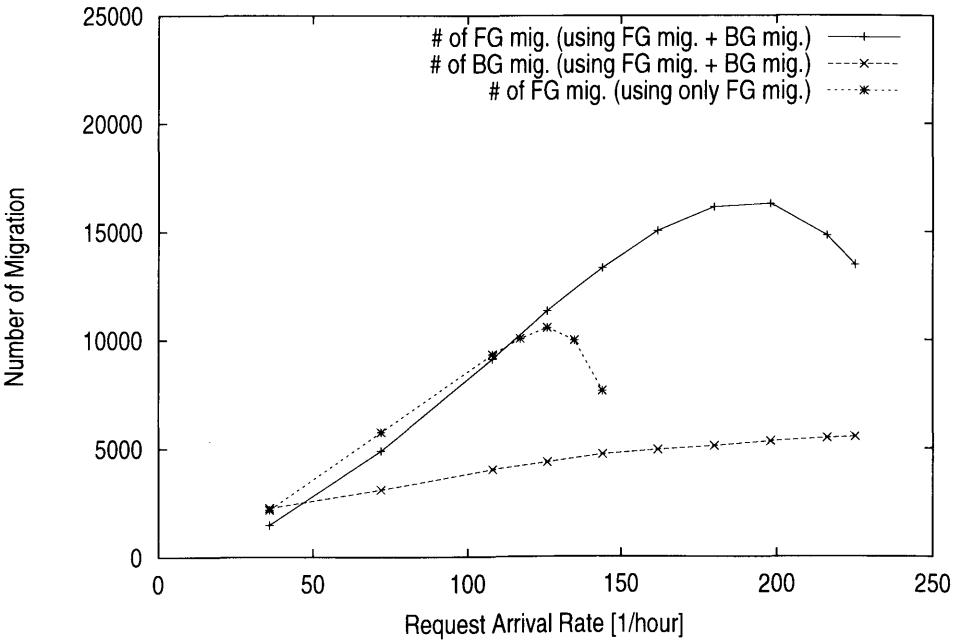

図 4.5: 50000 アクセスのマイグレーション数

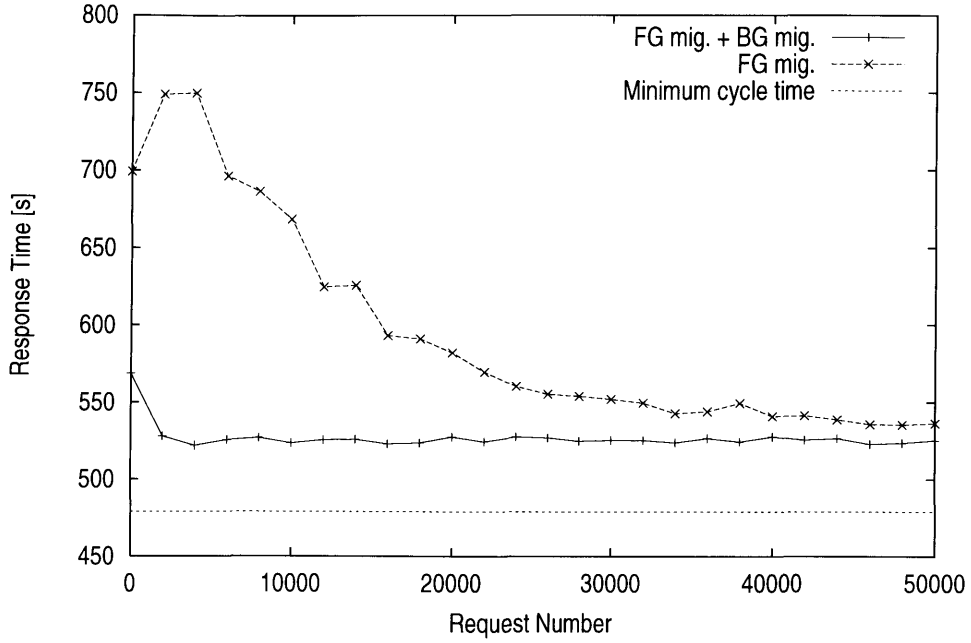

図 4.6: 2000 アクセス毎の平均応答時間

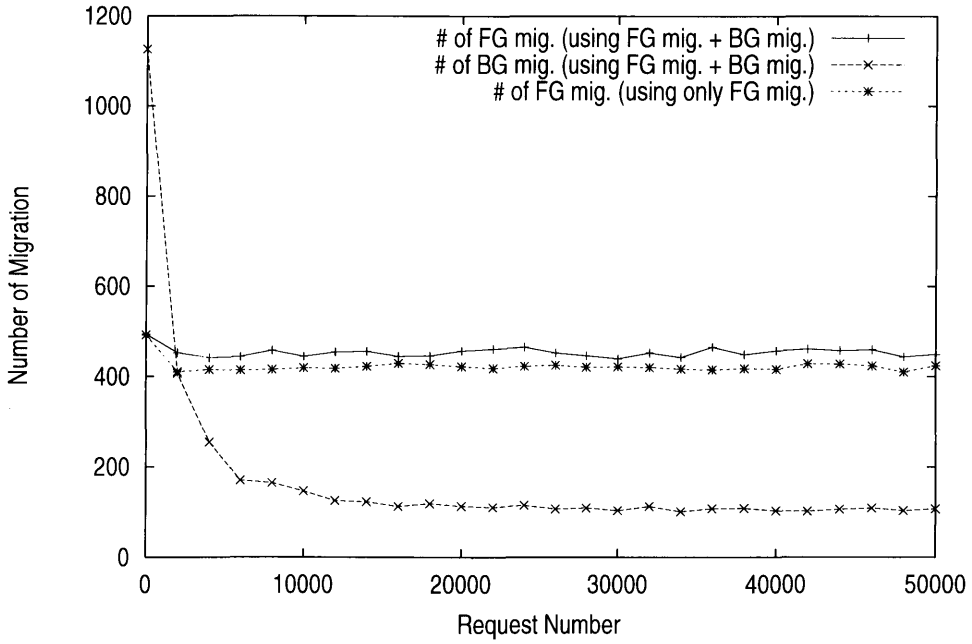

図 4.7: 2,000 アクセス毎のマイグレーション数

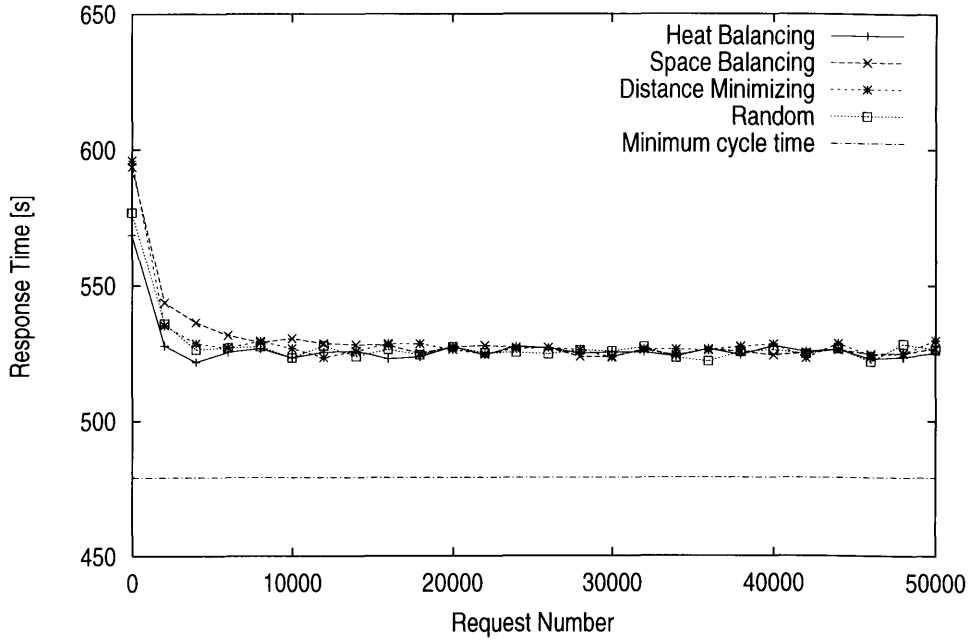

図 4.8: フォアグランドマイグレーションの移動先候補の選択方針による影響（応答時間変化）

もつため，フォアグランドマイグレーションはブロックされることなく実行されるが，短期間に高温エレメントアーカイバの近隣に存在する低温エレメントアーカイバの空きスロットは移送された高温のテープで埋められることとなり，その後はテープを受け入れることができず，フォアグランドマイグレーションが有効に機能しなくなるためである．このために平均応答時間が悪化する．その後，徐々にスケーラブルテープアーカイバは定常状態へ収束するが，これは低温アーカイバから高温アーカイバへのマイグレーションが生じ，空きスロットが生成されることによる．このようなマイグレーションの発生率は低く，そのため，収束の速度は緩やかなものとなる．一方，フォアグランドマイグレーションに加えてバックグランドマイグレーションを用いることにより，空きスロット数に偏りがある場合にはバックグランドマイグレーションによりテープが移送され低温アーカイバに空きスロットが生成されるために，収束速度は大きく向上する．

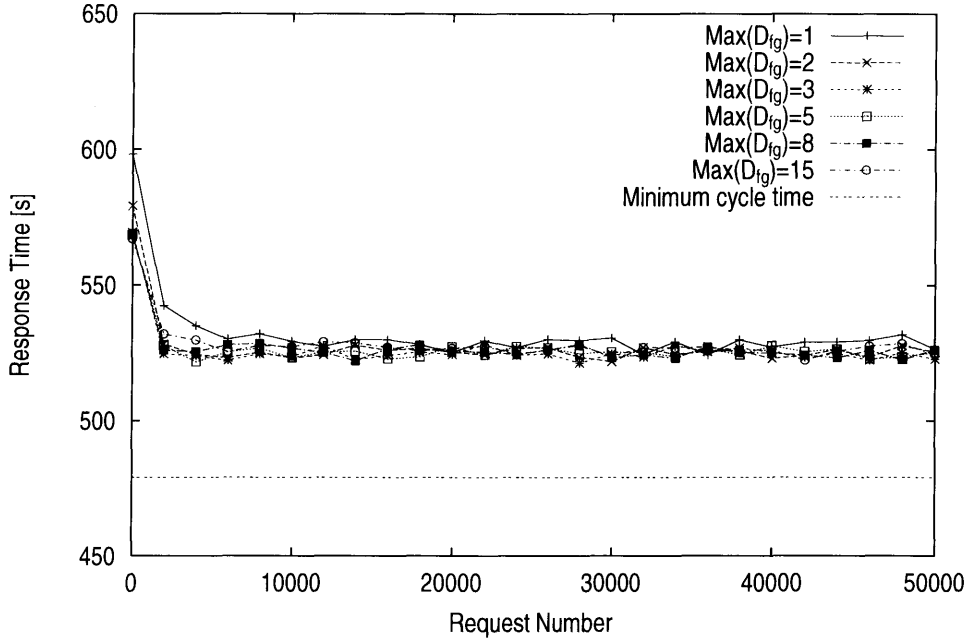

図 4.9: フォアグランドマイグレーション移動距離の影響（応答時間変化）

4.5.3 フォアマイグレーション移動先候補の選択方式による影響

図 4.8 は，フォアグランドマイグレーションにおける移動先エレメントアーカイバの選択方針を変えた場合の，リクエスト到着率が 126 リクエスト/時の場合の 2000 リクエスト毎の平均応答時間である．平均応答時間に関しては，わずかながら Heat Balancing を用いた場合が収束時間が短い．これは初期状態に存在するエレメントアーカイバ間の熱の不均衡の解消に Heat Balancing が最も貢献するためである．しかしながら，各方針の差は極めて小さく，収束後の平均応答時間においては方針による差は見られない．フォアグランドマイグレーションの違いによるテープの移送先の違いはバックグランドマイグレーションが解消するため，いずれの方針でもフォアグランドマイグレーションは有効に動作する．

4.5.4 マイグレーション移動距離の影響

図 4.9 はフォアグランドマイグレーションによるテープの最大移動距離を 1,2,3,5,8,15 エレメントアーカイバと変化させた場合の，リクエスト到着率が 126 リクエスト/時のときの 2000 アクセス毎の応答時間の変化を示している．フォアグランドマイグレーションの最

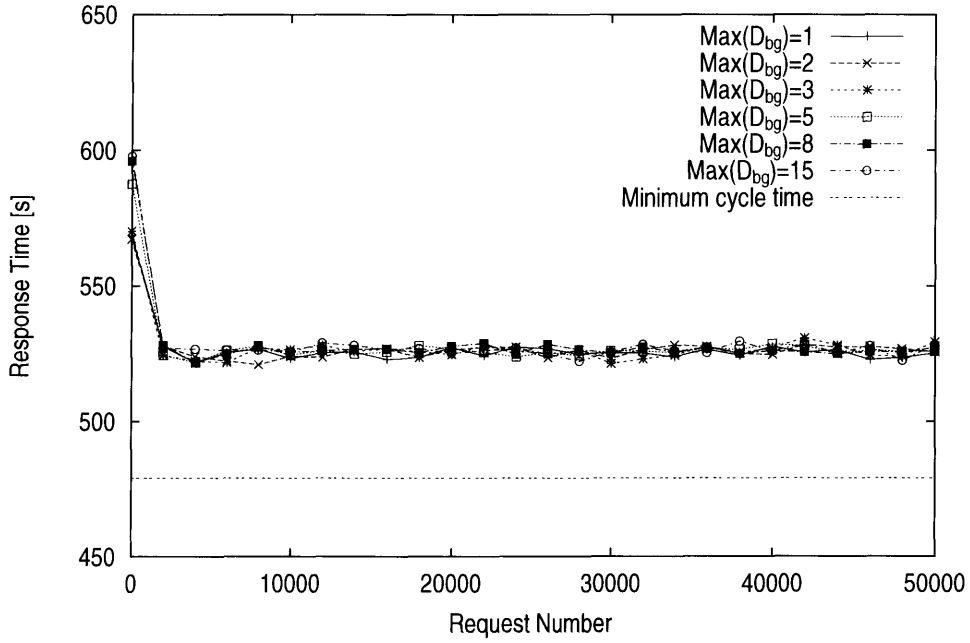

図 4.10: バックグランドマイグレーション移動距離の影響（応答時間変化）

大移動距離を大きくすることは，より多くのエレメントアーカイバを移送先の候補とすることになり，空きドライブが存在する確率が増すことになる．しかしながら，マイグレーション中は，その間のエレメントアーカイバのテープハンドラーロボット，テープ移送装置をロックするために，他のリクエストがブロックされることになる．シミュレーションの結果によると，最大移動距離が 1 エレメントアーカイバの場合には，他の場合と比較して僅かながら収束時間が長い．これは最大移動距離が大きい場合には，フォアグランドマイグレーションにより，直接，より遠くの熱の低いエレメントアーカイバへ熱いテープを移送できるために収束が早まるためである．一方，最大移送距離を 3 以上としても収束時間は変化しない．これは長い距離のフォアグランドマイグレーションは，その間に存在するより多くのエレメントアーカイバのロボットハンド，移送装置が使用されていないことが必要となるために，実行回数が少ないことによる．

図 4.10 はバックグランドマイグレーションによるテープの最大移動距離を 1,2,3,5,8,15 エレメントアーカイバと変化させた場合の，リクエスト到着率が 126 リクエスト/時のときの 2000 アクセス毎の応答時間の変化を表している．バックグランドマイグレーションにおけるテープの移送距離に関しては，長い距離のバックグランドマイグレーションは，

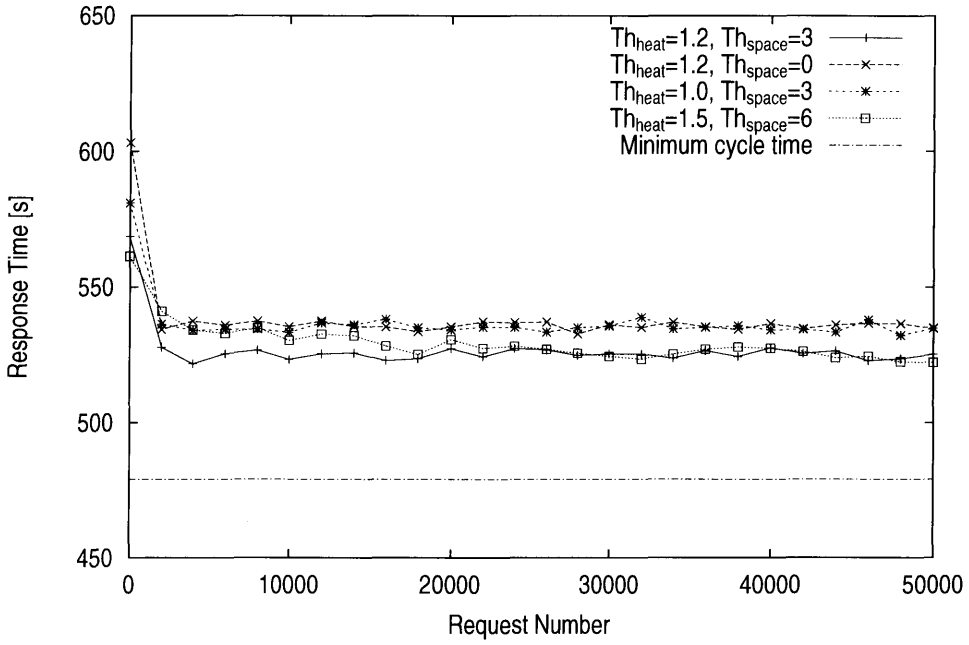

図 4.11: バックグランドマイグレーション起動条件に対するしきい値の効果（平均応答時間変化）

より短い距離の複数のバックグランドマイグレーションに分割可能であり，分割されたバックグランドマイグレーションを並列に実行することによりテープの移送に要する時間が短縮でき，また，並列に実行できない場合でも時間的なオーバーヘッドは小さいため，最大移動距離が短い場合の方が有利である．しかしながら，シミュレーション結果によるとその差は見られない．これは，フォアグランドマイグレーション同様，バックグランドマイグレーションもその間のエレメントアーカイバのハンドラロボット，移送装置が使用されていないことを必要とするために，長い距離のバックグランドマイグレーションが行われる頻度が低いためである．

4.5.5 バックグランドマイグレーション起動条件に対するしきい値の効果

バックグランドマイグレーションは 2 つのエレメントアーカイバ間の熱および空きスロット数に偏りがある場合に実行される．しかしながら，この偏りに敏感にしすぎると必要以上にバックグランドマイグレーションが実行されることとなるため，熱および空きスロット数の差にしきい値を設け，必要以上にバックグランドマイグレーションが実行され

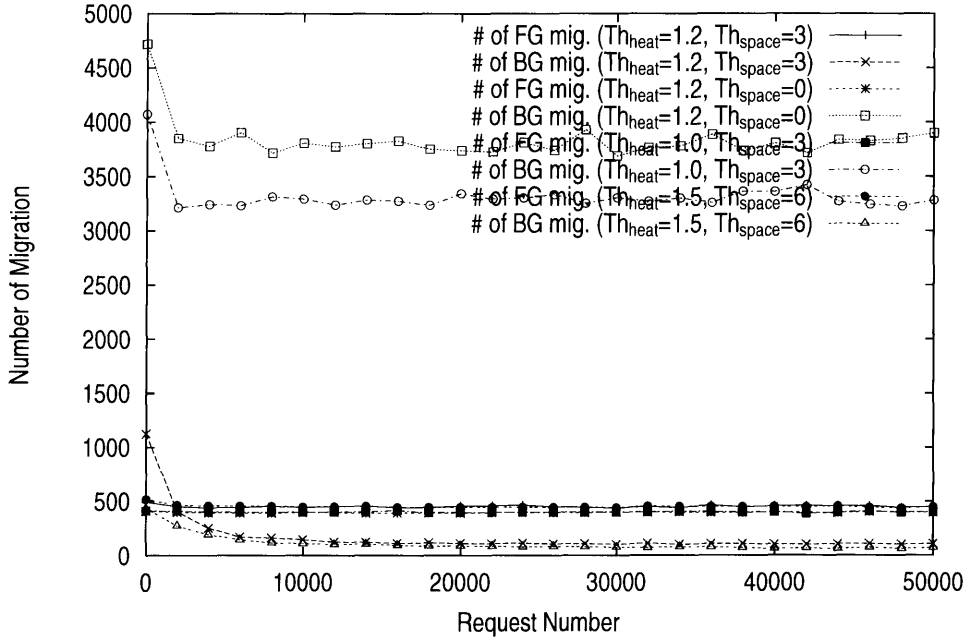

図 4.12: バックグランドマイグレーション起動条件に対するしきい値の効果（マイグレーション数の変化）

るのを防ぐ必要がある．図4.11は，バックグランドマイグレーションを実行する条件として，空きスロット数の差が3より大きく，かつ熱の比が1.2より大きい場合，空きスロット数の差が0より大きく，かつ熱の比が1.2より大きい場合，空きスロット数の差が3より大きく，かつ熱の比が1.0より大きい場合，空きスロット数の差が6より大きく，かつ熱の比が1.5より大きい場合の2000リクエスト毎の応答時間を変化を，図4.12はそのときのマイグレーション数を表している．熱，空きスロット数の両者に関し，その感度を敏感にすると非常に多くのバックグランドマイグレーションが実行され，そのためにフォアグランドマイグレーションの実行が妨げられるためにかえって応答時間が悪化していることが分かる．また，起動条件を緩くすると，バックグランドマイグレーションが減少するために，スケーラブルテープアーカイバの収束時間はやや長くなるが，その差は小さい．

4.5.6 テープマイグレーション装置のワゴンの移動速度による影響

本節では，テープマイグレーション装置のワゴン移動速度の性能に対する影響について評価する．試作スケーラブルテープアーカイバではテープマイグレーション装置のワゴン

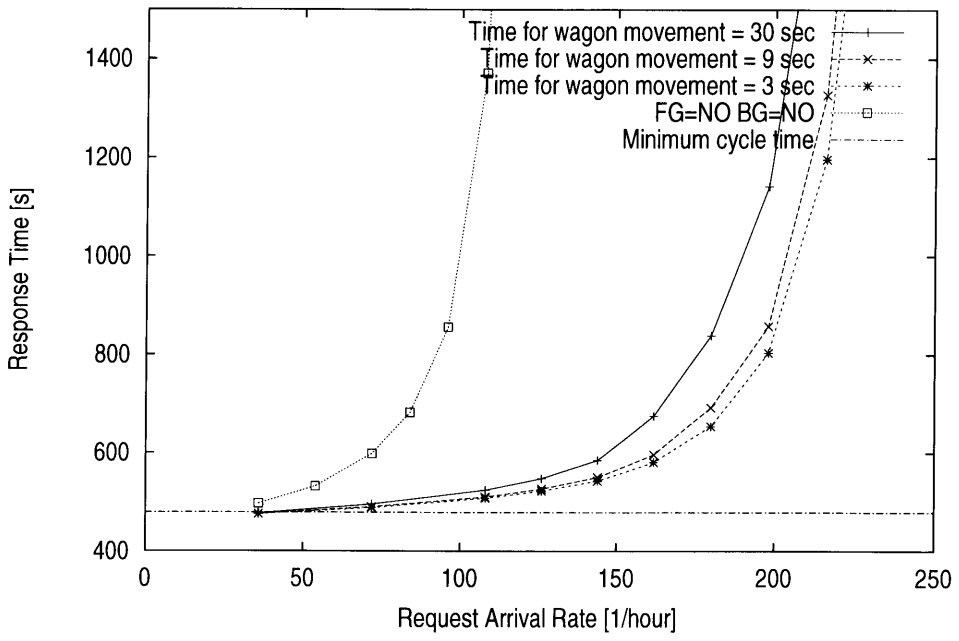

図 4.13: 高速，低速テープマイグレーション装置を用いた際の平均応答時間

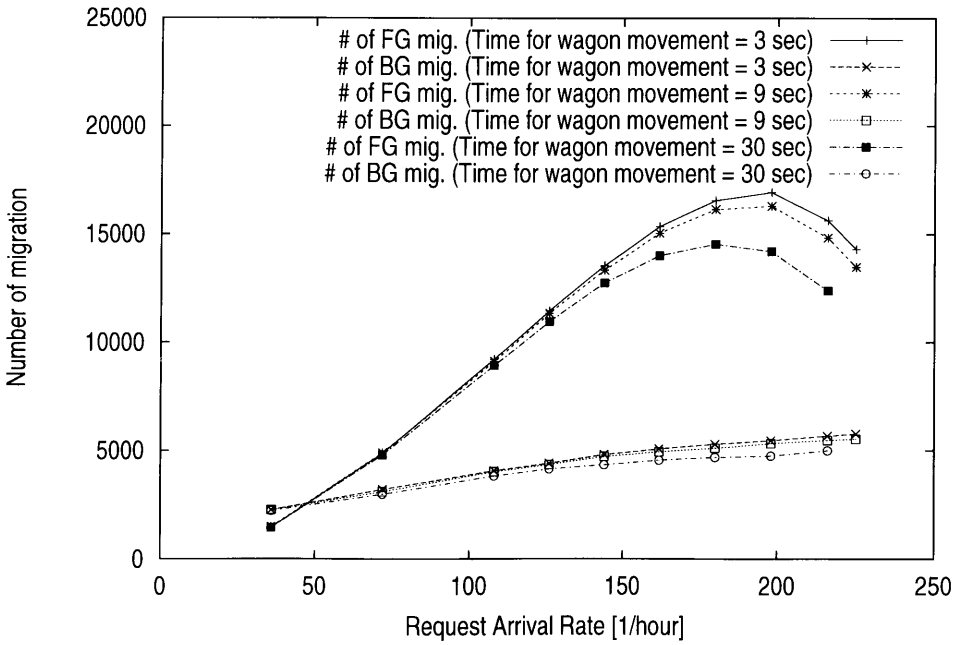

図 4.14: 高速，低速テープマイグレーション装置を用いた際のマイグレーション数

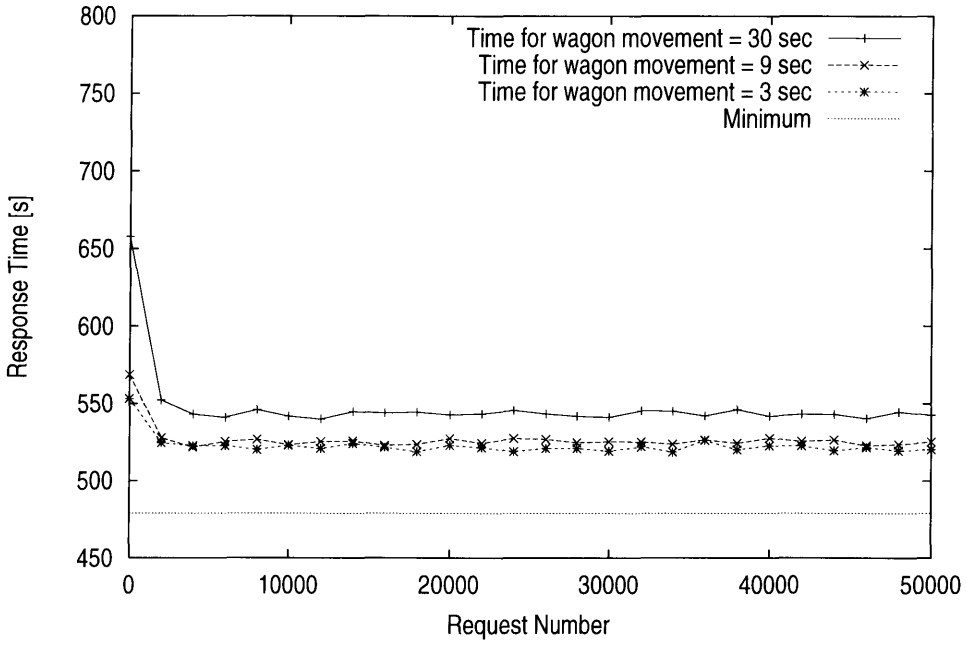

図 4.15: 高速，低速テープマイグレーション装置を用いた際の 2000 アクセス毎の平均応答時間

の移動時間は 9 秒であるが，これに加え，移動に 3 秒および 30 秒を要する場合のシミュレーションを実行した．図 4.13 はテープマイグレーション装置のワゴンの移動速度を変えた場合のシミュレーション開始から 50000 アクセスまでの平均応答時間を示している．テープマイグレーション装置のワゴンの移動速度が遅い場合でも，マイグレーションを用いない場合に比べ，大きく応答時間が短縮されている．図 4.15 はリクエスト到着率が 126 リクエスト/時のときの 2000 アクセスごとの平均応答時間である．ワゴンの移動時間が 30 秒の場合と 9 秒の場合では，平均応答速度にやや差がみられるが，3 秒と 9 秒の間の差は極めて小さく，試作スケーラブルテープアーカイバのテープマイグレーション装置のワゴンを更に高速化しても，これ以上の応答性能の改善は得られないことがわかる．図 4.14 は 2000 アクセスごとのマイグレーション数を表しているが，この図からもワゴンの移動速度が 3 秒と 9 秒の場合のマイグレーション数の違いが非常に小さいことがわかる．マイグレーション装置のワゴン移動距離は約 40cm であるが，この距離の移動時間が 9 秒，即ち移動速度は約 270cm/分であり低速と言える．即ち高価な高速のテープマイグレーション装置を作成する必要はなく，安価な低速のテープマイグレーション装置で十分である．

従来のテープ移送機構をもたないそれぞれ独立したテープアーカイバにわずかなコストのテープマイグレーション装置を追加するだけで極めて高性能な大規模アーカイブ装置を構築することが可能であると言える．

4.5.7 スケーラビリティ

本節では，スケーラブルテープアーカイバを構成するエレメントアーカイバの台数とその応答性能，すなわちスケーラビリティについて評価する．図4.16は，スケーラブルテープアーカイバを構成するエレメントアーカイバが1台，2台，4台，8台，16台，32台，64台の場合の，シミュレーション開始後20000アクセス目から40000アクセス目までの計20000アクセスの平均応答時間を表している．本シミュレーションにおいては，全てのエレメントアーカイバに同数の高アクセス頻度テープが配置される状態を初期配置としており，そのアクセスローカリティは80/20則に従うものとしている．リクエスト到着率は1エレメントアーカイバ当り6リクエスト/時，10リクエスト/時，13リクエスト/時としている．テープマイグレーションを用いない場合においては，スケーラブルテープアーカイバを構成するエレメントアーカイバ数によらず，各エレメントアーカイバ内の高アクセス頻度テープ数は常に等しく，また，各エレメントアーカイバに対する平均リクエスト到着率も変化がないため，各エレメントアーカイバにおける平均応答時間は等しく，従ってエレメントアーカイバ台数を増加させてもスケーラブルテープアーカイバ全体としての平均応答時間は変化しない．一方，テープマイグレーションを用いると，台数を増加させた場合には平均応答時間が短縮される．これは，長時間の平均をとると各エレメントアーカイバに対する平均リクエスト到着率は等しいが，ある短い時間においては，各エレメントアーカイバに対して発行されるリクエストには偏りが存在する．例えば，ある短期間においては，エレメントアーカイバに対してはドライブ装置数より多いリクエストが発行されているが，その隣接エレメントアーカイバでは，ドライブ装置より少い数のリクエストしか発行されていないという状況がある．このような状況に対してフォアグランドマイグレーションは，アクセス要求を受けているテープを他のエレメントアーカイバへ移送することによりあるエレメントアーカイバに集中しているリクエストを分散させることが可能であり，応答時間が短縮される．即ち，テープマイグレーションを用いることに

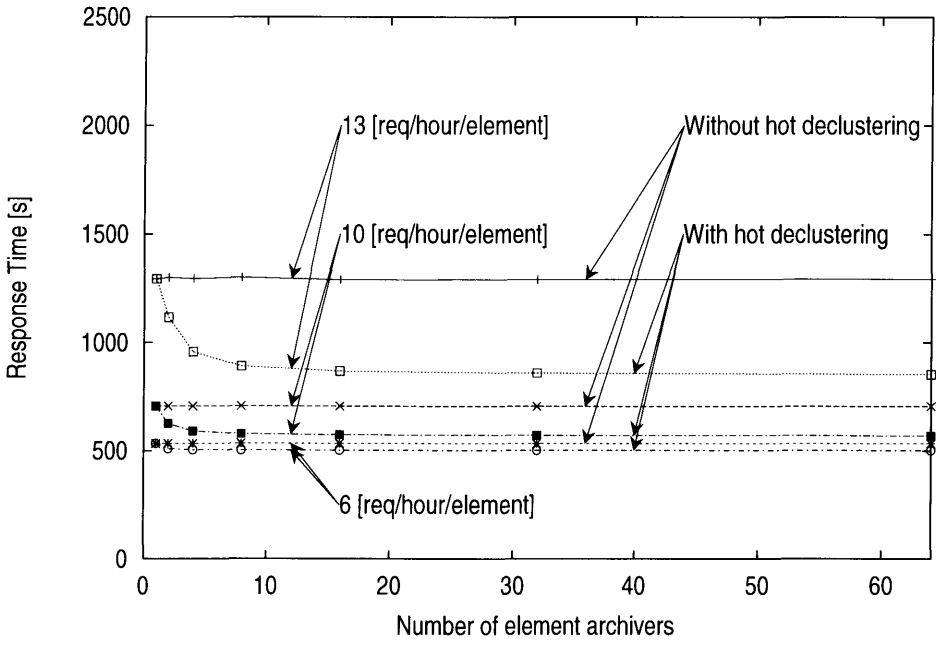

図 4.16: 高アクセス頻度テープを均等に分布させた場合の平均応答時間とエレメントアーカイバ数の関係

より，短い時間内のリクエストの偏りが吸収されることとなり，平均応答時間が短縮される．特に，エレメントアーカイバの数が増加すると，フォアグランドマイグレーション時により多くのエレメントアーカイバを移送先候補とすることができるため，応答性能が向上する．

図 4.17 は，同様にスケーラブルテープアーカイバを構成するエレメントアーカイバが 1 台，2 台，4 台，8 台，16 台，32 台，64 台の場合の，シミュレーション開始後 20000 アクセス目から 40000 アクセス目までの計 20000 アクセスの平均応答時間を表している．本シミュレーションでは，初期状態として高アクセス頻度テープはスケーラブルテープアーカイバ全体にランダムに配置されている．即ち，リクエストの偏りばかりでなく，各エレメントアーカイバ内の高アクセス頻度テープの数には偏りが存在することとなり，そのためテープマイグレーション機構を用いない場合には，最も高アクセス頻度テープが多く存在するエレメントアーカイバに対して発行されたリクエストの応答時間が支配的となり，平均応答時間は悪化する．特にエレメントアーカイバ数が多くなると，各エレメントアーカイバ内の高アクセス頻度テープ数の最大値が大きくなる確率が上昇するため，エレメン

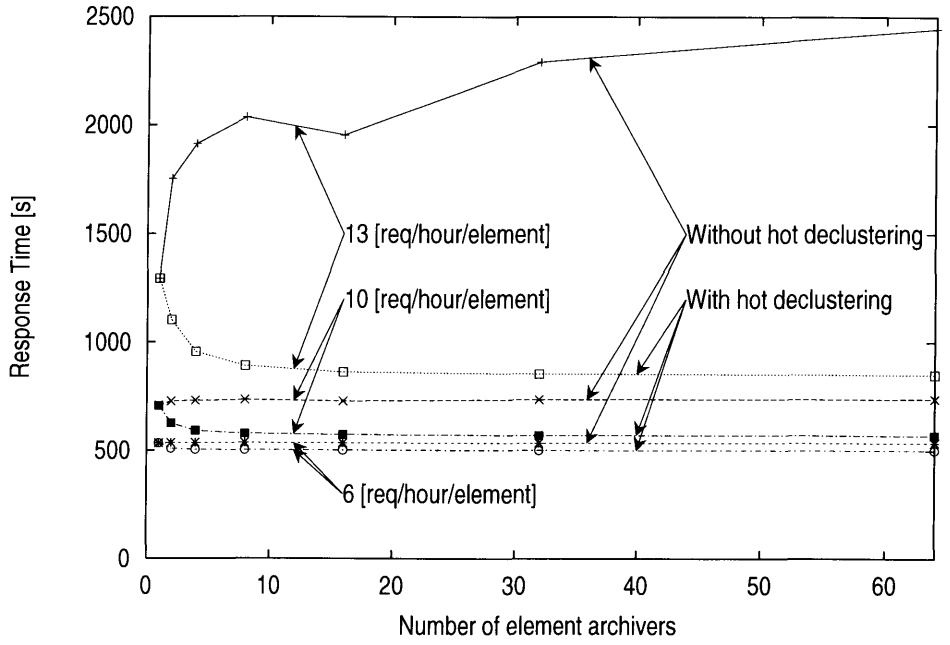

図 4.17: 高アクセス頻度テープをランダムに分布させた場合の平均応答時間とエレメントアーカイバ数の関係

トアーカイバ数の増加にともない平均応答時間も上昇する．一方，バックグランドマイグレーションは各エレメントアーカイバ内の高アクセス頻度テープを再配置し，各エレメントアーカイバのアクセス頻度を平衡化するため，平均応答時間が短縮されることとなる．

4.6 ドライブ故障時の性能

本節では，テープドライブ故障時のシミュレーションを実行し，テープドライブの故障に対するホットデクラスタリングの効果について評価する．シミュレーションパラメータ，テープの初期分布は 4.5 章と同じ値を用いる．

図 4.18 はリクエスト到着率が 144 リクエスト/時，180 リクエスト/時，198 リクエスト/時のときのシミュレーション開始後から 2000 アクセスごとの平均応答時間を表している．このシミュレーションにおいては，スケーラブルテープアーカイバが 30000 リクエストを受けた時点で 8 番目のエレメントアーカイバ内のテープドライブが 1 台故障し，その後更に 30000 リクエストを受けたときにそのテープドライブが回復するものとしている．図 4.18 より 1 台のテープドライブの故障では平均応答時間はほとんど影響を受けないことが

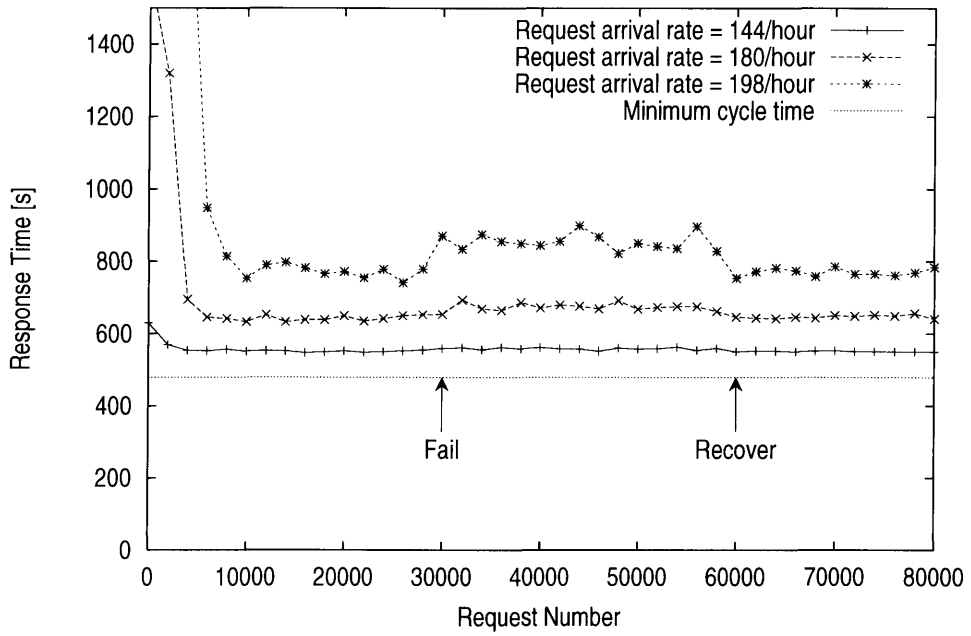

図 4.18: 1 ドライブ故障時の 2000 アクセス毎の平均応答時間

わかる．

図 4.19 はスケーラブルテープアーカイバが 30000 リクエストを受けた時点で第 8 エレメントアーカイバ内のテープドライブが 2 台同時に故障し，その後更に 30000 リクエストを受けたときにそれらのテープドライブが回復するものとしたときの 2000 アクセスごとの平均応答時間である．テープドライブが故障している間は第 8 エレメントアーカイバには使用可能なテープドライブは存在しないが，このような状況においてもスケーラブルテープアーカイバは第 8 エレメントアーカイバから他のエレメントアーカイバへテープを移送し，アクセス要求に応えることができる．リクエスト到着率が 180 リクエスト/時，198 リクエスト/時の場合，2 台のテープドライブが故障すると平均応答時間が悪化するが，これはドライブ故障によりスケーラブルテープアーカイバ自体の処理能力が低下したことによる．

図 4.20 は図 4.18，図 4.19 でのシミュレーションにおいてリクエスト到着率が 180 リクエスト/時のときのスケーラブルテープアーカイバ全体の 2000 アクセスごとのマイグレーション数を表し，図 4.21 は，図 4.18，図 4.19 でのシミュレーションにおいてリクエスト到着率が 180 リクエスト/時のときの故障したテープドライブをもつ第 8 エレメントアー

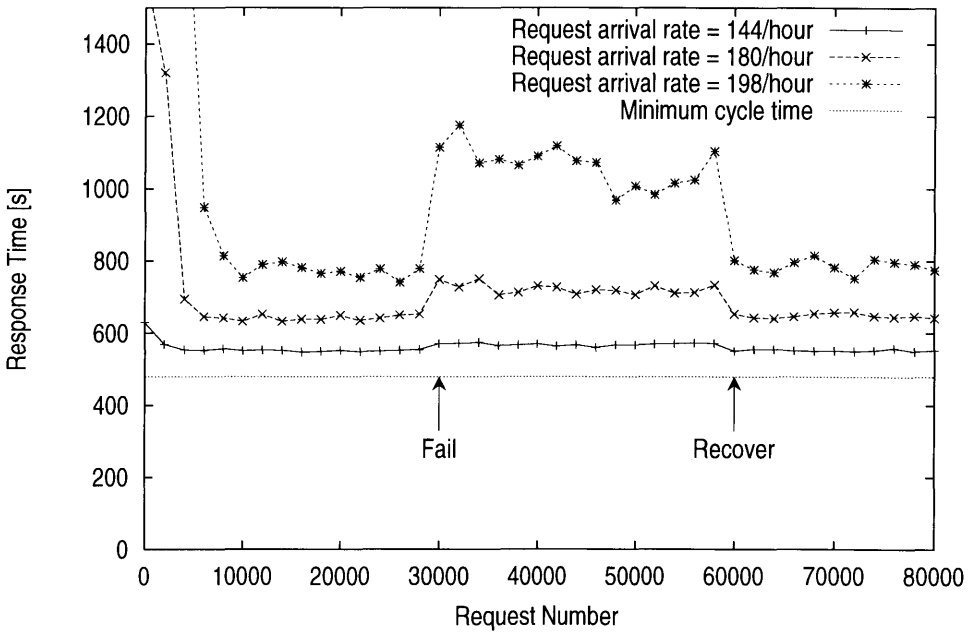

図 4.19: 2 ドライブ故障時の 2000 アクセス毎の平均応答時間

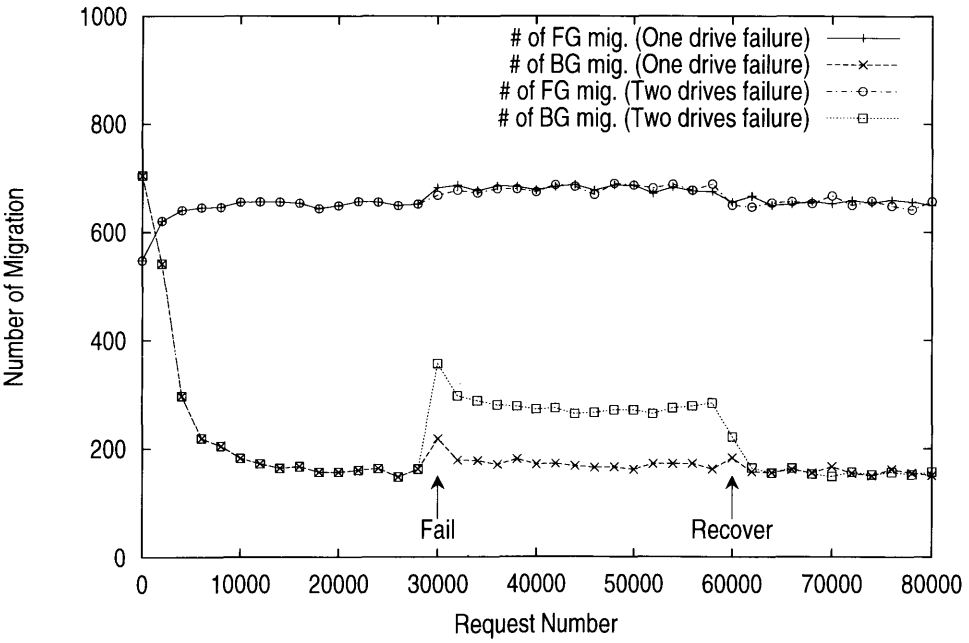

図 4.20: ドライブ故障時の 2000 アクセス毎のマイグレーション数

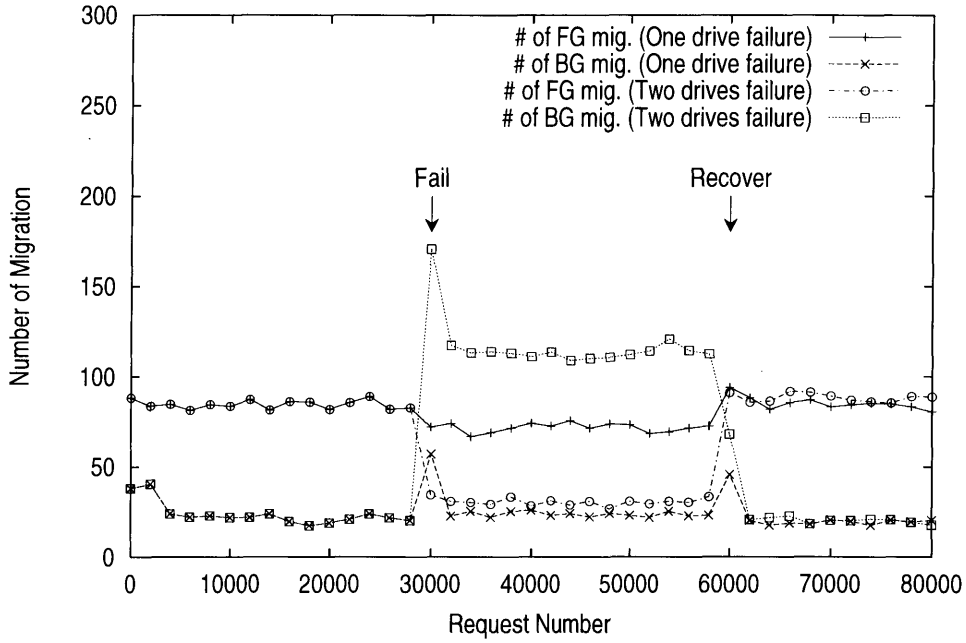

図 4.21: ドライブ故障時の第 8 エレメントアーカイバの 2000 アクセス毎のマイグレーション数

カイバの 2000 アクセスごとのマイグレーション数を表している．エレメントアーカイバの熱はドライブ数によって正規化されているため，ドライブが故障するとそのエレメントアーカイバの熱は大きく変化し，熱の再分散が行われるため，図 4.20，図 4.21 に示されるようにドライブ故障時にバックグランドマイグレーションが増加する．この熱の再分散により，故障したドライブをもつ第 8 エレメントアーカイバには高温テープが減少し，低温のテープが増加することとなり，従って，第 8 エレメントアーカイバが受けるリクエスト数が減少するために第 8 エレメントアーカイバのフォアグランドマイグレーションは減少する．特に 2 台のドライブが故障した場合には，第 8 エレメントアーカイバ内のテープに対するリクエストへのアクセスのためには必ずフォアグランドマイグレーションを必要とするにも関わらず，第 8 エレメントアーカイバのフォアグランドマイグレーション数は大幅に減少していることがわかる．逆に第 8 エレメントアーカイバ以外のエレメントアーカイバでは高温テープが増加するために単位時間あたりに受けるリクエスト数が増え，そのために全体としてフォアグランドマイグレーション数は増加するが，その差は僅かである．スケーラブルテープアーカイバはテープドライブが故障すると自動的に故障したテー

表 4.3: ファイルストライピング時の初期カセットテープ分布

エレメントアーカイバ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
高アクセス頻度テープ（非ストライプデータ）	7	7	7	7	7	39	39	39	39	39	39	7	7	7	7	7
低アクセス頻度テープ（非ストライプデータ）	88	88	88	88	88	56	56	56	56	56	56	88	88	88	88	88
高アクセス頻度テープ（ストライプデータ）	19	19	19	19	19	19	19	19	19	19	19	19	19	19	19	19
低アクセス頻度テープ（ストライプデータ）	76	76	76	76	76	76	76	76	76	76	76	76	76	76	76	76
合計（非ストライプデータ）	95	95	95	95	95	95	95	95	95	95	95	95	95	95	95	95
合計（ストライプデータ）	95	95	95	95	95	95	95	95	95	95	95	95	95	95	95	95
合計	190	190	190	190	190	190	190	190	190	190	190	190	190	190	190	190

プドライブをもつエレメントアーカイバ内の高温テープを再分散させるため，テープドライブの故障時においてもホットデクラスタリングは非常に効果的である．

4.7 ファイルストライピング時の性能評価

4.7.1 シミュレーション条件

本節ではファイルストライピングを用いたときのホットデクラスタリングの効果について評価する．テープアーカイバにおいてはファイルストライピングはデータの転送速度を高める手法として有効であるが [16][20]，複数のテープドライブを同時に使用することとなるため，リクエストがブロックされる確率が高くなる．従って，空きドライブの効率的な使用を可能とするテープマイグレーションはファイルストライピングに対して有効であると期待される．本シミュレーションではアーカイバには 2 種類のサイズのファイルが混在するものとした．一方はファイルサイズが 100MB のストライプされないファイルであり，もう一方はストライプされる 1.6GB のファイルである．ストライプ幅を 1（ストライピングなし），4（4 ウェイストライピング），8（8 ウェイストライピング），16（16 ウェイストライピング）とした 4 つの場合についてシミュレーションを実行した．1.6GB のファイルは 4 ウェイストライプでは 400MB，8 ウェイストライプでは 200MB，16 ウェイストライプでは 100MB にそれぞれ分割される．テープの初期分布を表 4.3 に示す．ストライプされたファイルは初期状態では等間隔でエレメントアーカイバに配置されている．例えば，4 ウェイストライピングでは 1 番目，5 番目，9 番目，13 番目のエレメントアーカイバに配置され，16 ウェイストライピングではすべてのエレメントアーカイバに

配置される．100MB と 1.6GB のファイルのファイル数の比は 16:1，容量比では 1:1 である．スケーラブルテープアーカイバに関する他のパラメータは表 4.1 に従う．

4.7.2 シミュレーション結果

図 4.22 は，シミュレーション開始後 50000 アクセスまでのテープマイグレーションを用いない場合の 100MB の非ストライプファイルに対するリクエストの平均応答時間であり，図 4.23 は，フォアグランドマイグレーション，バックグランドマイグレーションを共に用いた場合の 100MB の非ストライプファイルに対するリクエストの平均応答時間である．横軸は 100MB のファイルへのリクエストと 1.6GB のファイルへのリクエストを合わせた全リクエストの到着率である．図中のストライプ幅は 1.6GB のファイルに対するものである．これらの図より，ホットデクラスタリングを用いることにより応答時間が大きく短縮されることがわかる．図 4.23 よりストライプ幅を大きくした場合に平均応答時間が悪化することがわかるが，これはストライプ幅を増やすとテープハンドラロボットの操作やテープのロード/イジェクトの回数が増加するためである．即ち，ストライプ幅を大きくすることは非ストライプデータに対しては不利に働くことになる．図 4.24，図 4.25 は 1.6GB のファイルに対するリクエストの平均応答時間である．それぞれ図 4.24 はマイグレーションを用いない場合のものであり，図 4.25 はフォアグランドマイグレーション，バックグランドマイグレーションを共に用いた場合のものである．リクエスト到着率が小さい場合には，ストライプ幅を大きくすることにより応答時間が短縮され，また，ホットデクラスタリングを導入することにより性能向上が得られることがわかる．ストライプ幅が大きい場合にはより低いリクエスト到着率で応答性能が悪化してしまうが，これは図 4.23 の場合と同様，テープの操作回数やロード/イジェクト回数の増加によるものである．以上の結果より，ホットデクラスタリングは非ストライプデータおよびストライプデータに対するリクエスト応答時間のいずれも短縮することがわかる．これは，可能な限りテープドライブを利用し，ストライプされたデータへのアクセスの並列性をより高めていることによる．

図 4.26 は，ファイルをストライピングしたときのフォアグランドマイグレーション数およびバックグランドマイグレーション数を表している．ストライプ幅が増えるとフォア

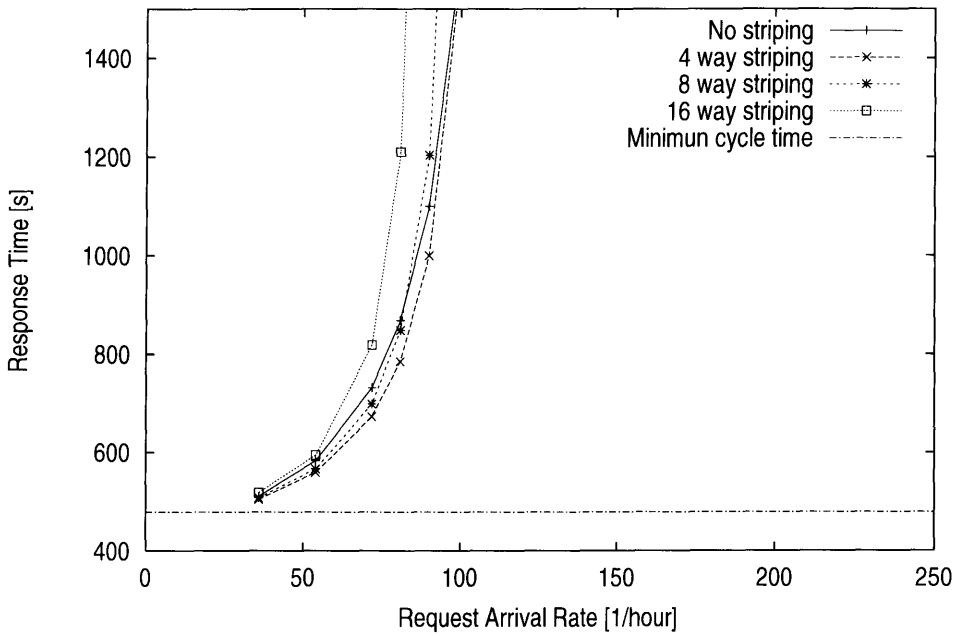

図 4.22: ホットデクラスタリングを用いない場合の非ストライプデータ（100MB）に対するリクエストの平均応答時間

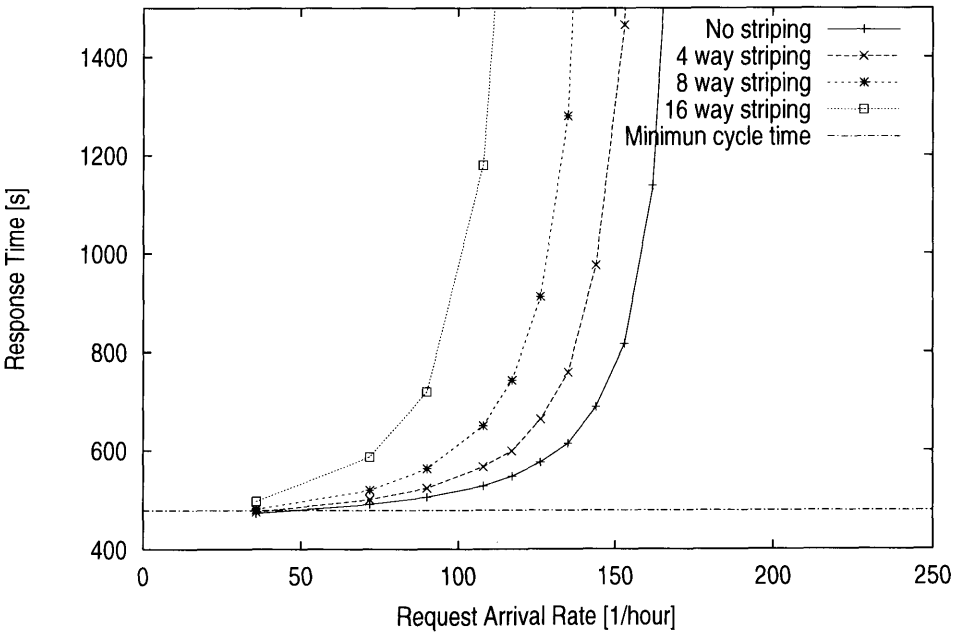

図 4.23: ホットデクラスタリングを用いた場合の非ストライプデータ（100MB）に対するリクエストの平均応答時間

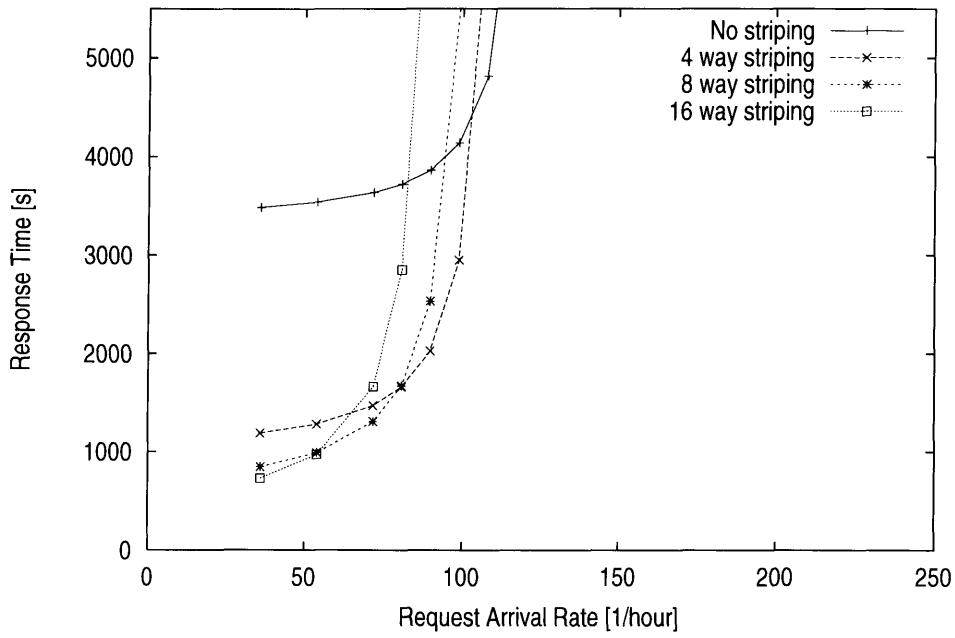

図 4.24: ホットデクラスタリングを用いない場合のストライプデータ（1.6GB）に対するリクエストの平均応答時間

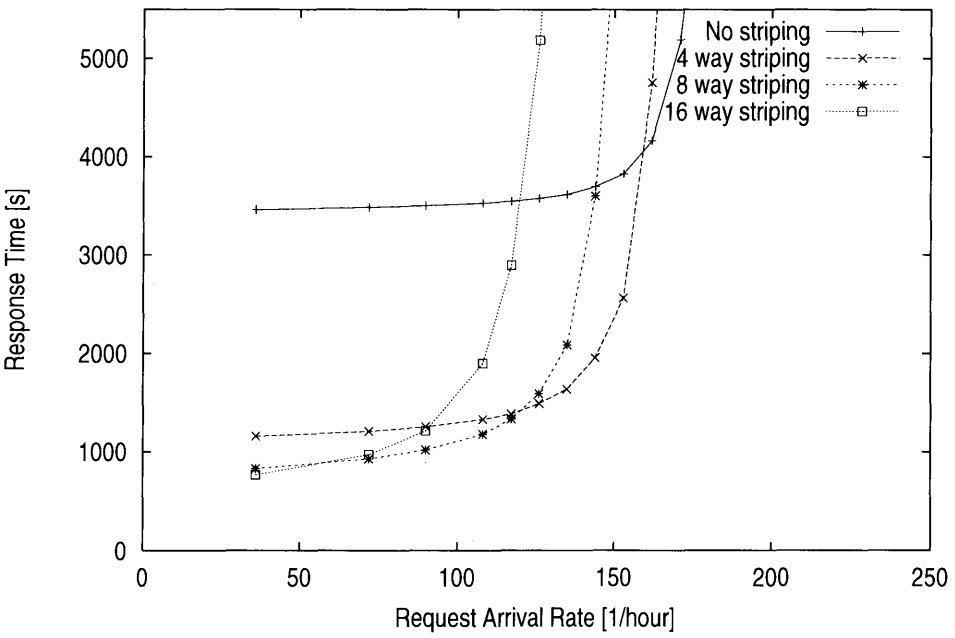

図 4.25: ホットデクラスタリングを用いた場合のストライプデータ（1.6GB）に対するリクエストの平均応答時間

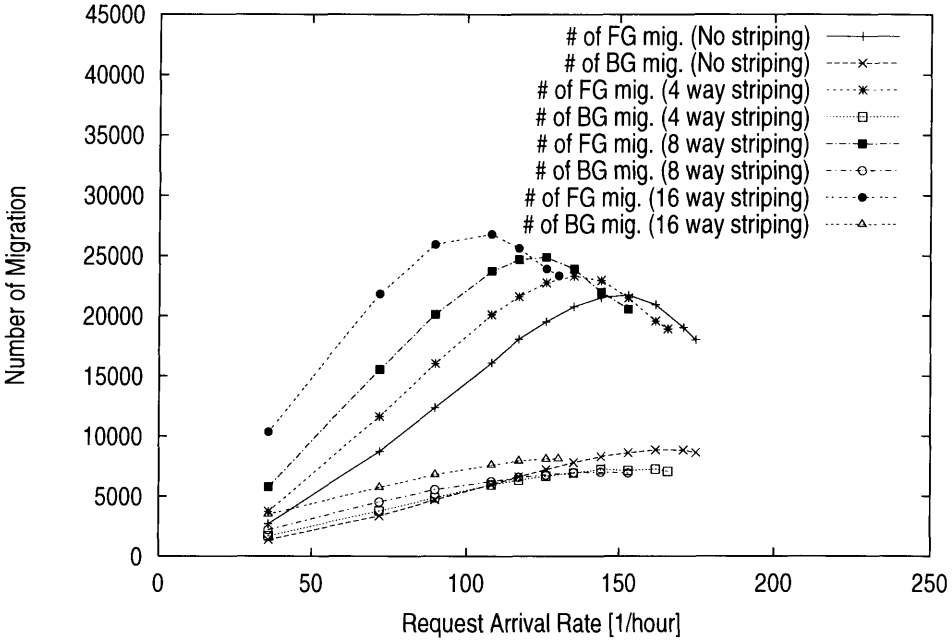

図 4.26: ストライプ時のマイグレーション数

グランドマイグレーション数が増加しているが，これはストライプ幅が大きい場合にはそれだけ多くのストライプされたファイルを操作しなければならないためである．ストライプ幅を増やすとより多くのマイグレーションが実行されるようになり，テープドライブの利用効率は悪化する．よって，図 4.23 に示されるように非ストライプデータの応答時間は劣化するが，図 4.25 に示されるようにストライプされたデータの応答時間は改善されることとなる．つまり，リクエスト到着率が低い場合にはストライプ幅を大きくすることは有効であるが，リクエスト到着率が高い場合にはストライプ幅が小さい場合の方が応答時間は短くなる．

4.8 まとめ

本節では，シミュレーションによりスケーラブルテープアーカイバにおけるホットデクラスタリングの効果について検討した．ホットデクラスタリングは，ホットデクラスタリングを用いない場合に比べ大きく性能を向上させる．フォアグランドマイグレーションは応答性能を向上させるのに対し，バックグランドマイグレーションはスケーラブルテープ

アーカイバの定常状態への収束時間の短縮に有効である．また，ホットデクラスタリングはテープドライブの故障に対しても有効であり．例え使用可能なテープドライブを持たないエレメントアーカイバが存在しても，スケーラブルテープアーカイバはリクエストに応えることが可能であることを示した．さらに，スケーラブルテープアーカイバにおいてファイルストライピングを用いた場合についても，ストライプ数を変化させてシミュレーションを行った．ファイルストライピングを用いたときにおいても，ホットデクラスタリングはテープドライブの利用率を高め，応答時間を短縮することを明らかにした．