スケーラブルテープアーカイバを用いた大規模ファイルシステムにおけるファイル編成方式の検討

1R－6

根本 利弘　喜連川 優　高木 幹雄
東京大学 生産技術研究所

1 はじめに

現在、我々は衛星画像データを対象とした階層ファイルシステムを開発している [1]。この階層ファイルシステムでは三次記憶システムとして、コモディティ化されたテープアーカイバを複数接続し、隣接するアーカイバ間で直接テープを移送可能とすることで、大容量かつ高性能を安価に実現することを目指したスケーラブルアーカイバを用いている [2]。アクセスローカリティをもつデータに対しても十分な性能を持つ階層ファイルシステムを実現するためには、三次記憶におけるファイル編成が極めて重要となる。本稿では、この階層ファイルシステムにおけるファイル編成について検討を行う。

2 スケーラブルテープアーカイバ

2.1 ハードウェア構成

スケーラブルテープアーカイバは複数台のエレメントアーカイバと、その間でテープの移送を可能とする移送装置により構成される。図 1はエレメントアーカイバとしてNTH-200B を用いた試作スケーラブルアーカイバであり、各エレメントアーカイバはテープを操作するための1台のロボティクス、2台の8mm テープドライブ、200巻のカセットラックを持つ。また、移送機構はカセットを乗せるワゴンを持ち、このワゴンがエレメントアーカイバ間を行き来することでエレメントアーカイバ間のカセットの移送が行われる。

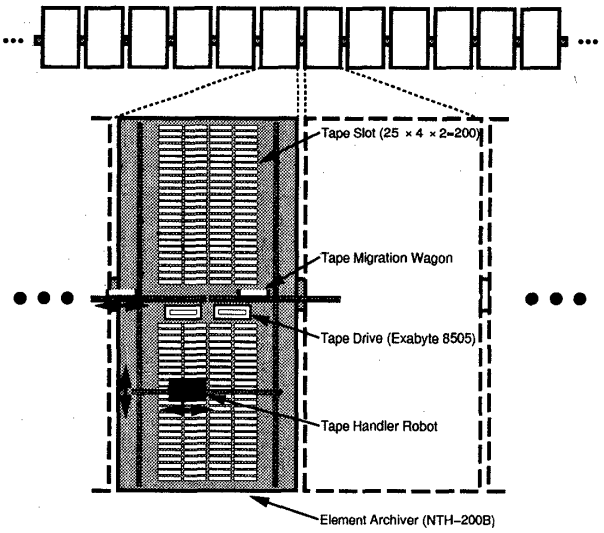

図 1: NTH-200B によるスケーラブルテープアーカイバの構成

2.2 カセットマイグレーション機構

スケーラブルアーカイバでは負荷分散の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種類のカセットマイグレーション機構を取り入れている。

フォアグランドマイグレーションはあるエレメントアーカイバ内のカセットに対してアクセス要求が生じた時に、そのエレメントアーカイバ内のテープドライブが全て使用中である場合に、他のテープドライブが空いているエレメントアーカイバへカセットを移動させるものである。一方、バックグランドマイグレーションは、エレメントアーカイバのロボティクス、移送機構が使用されていない時に、エレメントアーカイバ間でカセットを移動させることで、エレメントアーカイバ間のアクセス頻度 (熱) やテープの数の偏りを減少させ、負荷の均衡化を図るものである。

2.3 ファイル編成

スケーラブルテープアーカイバ上のデータのアクセス頻度 (熱) の偏りを考慮し、以下の2つのファイル編成方式の検討を行う。

熱集中方式 アクセス頻度の高いデータを特定のカセットテープに集中させる。

降順熱分散方式 アクセス頻度の高いデータをテープ間に分散し、テープ内でアクセス頻度の高いデータから順にカセットテープの先頭に配置する。

集中方式は、アクセス頻度の高いデータを集中させることで、ドライブにおけるテープの交換回数を削減し、応答速度の向上を目指すものである。一方、分散ソートは、シーク時間を短縮し、応答速度の向上を目指すものである。

Consideration about File Organization for Large Scale Tertiary File System using Scalable Tape Archiver
Toshihiro NEMOTO, Masaru KITSUREGAWA and Mikio TAKAGI
Institute of Industrial Science, University of Tokyo
Roppongi 7-22-1, Minato-ku, Tokyo, Japan

表 1: シミュレーションパラメータ

スケーラブルアーカイバ	
エレメントアーカイバ	16 台
カセット数	190 本
テープ容量	5GByte (50 ファイル)
ファイルサイズ	100 MByte
テープドライブ	
テープロード時間	35 秒
シーク時間	4 秒/ファイル
リード/ライト時間	200 秒/ファイル
テープ排出時間	20 秒
ロボティクス	
移動時間	2 秒
カセット操作時間	12 秒
移送装置	
移動時間	9 秒

3 性能評価

3.1 シミュレーションパラメータ

表 1はシミュレーションに用いたパラメータであり、試作スケーラブルテープアーカイバのデータを元にしている。アクセス到着時間は負の指数分布に従うものとし、スケーラブルテープアーカイバではフォアグランド・バックグランドマイグレーションをともに行っている。

3.2 シミュレーション結果

図 2は、アクセス頻度の高いデータとアクセス頻度の低いデータの 2 種類が存在するとし、初期状態として高アクセス頻度のデータを集中させてテープ上に配置した場合と、分散させてソートし配置した場合の、50000 アクセスの平均応答時間である。アクセス頻度の高いデータ数が 3200 個と 800 個の 2 通りの測定を行い、3200 個の場合は熱集中方式ではカセット 64 本と全ドライブ数の 2 倍、800 個では 16 本と全ドライブ数の 1/2 となる。この結果によると、いずれの場合でも、熱集中方式よりも降順熱分散方式の方よい性能を示し、高頻度テープの数がドライブ数の 1/2 となっても、性能は降順熱分散方式の方が優れている。これは、テープドライブのシーク時間がロボティクスのテープ操作時間+テープロード・排出時間よりも大きいため、シーク時間の削減が結果として全体としての応答性能の向上に大きく貢献するためである。

図 3は、アクセスローカリティが 70/30 の Zipf 分布をなす 15200 個のデータを 3040 本のテープにランダムに配置した場合と、これらのテープのうち、アクセス頻度の高いものより 1%、2%、5%のテープに関してテープ内のデータをアクセス頻度順にソートして配置した場合の 50000 アクセスの平均応答時間を表している。全カセットのデータを降順熱分散方式で配置するのには大きなコストを要するが、一部分のカセットをソートするだけでも十分に性能向上が得られることが分かる。

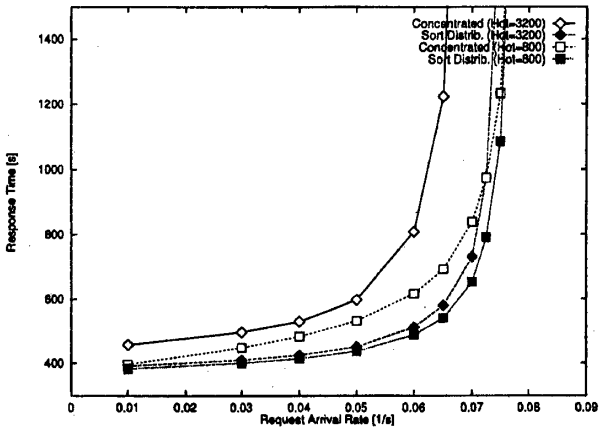

図 2: 熱集中方式と降順熱分散方式の比較

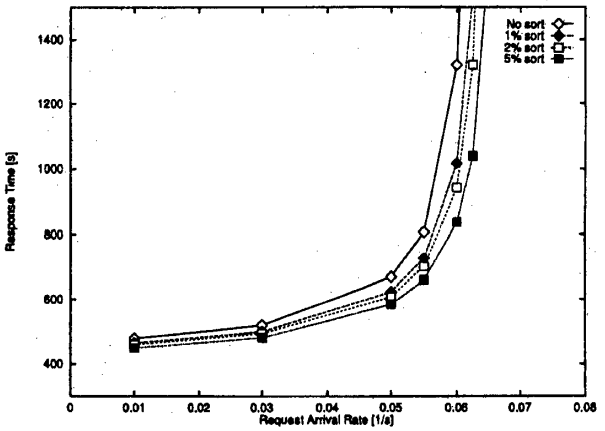

図 3: 一部のカセットのデータにソートしたときの効果

4 おわりに

スケーラブルテープアーカイバのファイル編成に関してアクセス頻度の高いデータを特定テープに集中させる方式と、テープ間に分散させてテープ内でソートする方式の検討を行った。シミュレーションによると分散しソートすることで応答性能の向上が得られることが示された。今後は、動的にファイル編成を行った際の性能評価を行う予定である。

参考文献

[1] K. Sako, T. Nemoto, M. Kitsuregawa, and M. Takagi. "Partial migration in an 8mm tape based tertiary storage file system and its performance evaluation through satellite image processing applications". In *Proceedings of 6th International Conference on Information Systems and Management of Data*, 1995.

[2] T. Nemoto, Y. Sato, K. Mogi, K. Ayukawa, M. Kitsuregawa, and M. Takagi. "Performance evaluation of cassette migration mechanism for scalable tape archiver". In *"Digital Image Storage and Archiving System"*, volume 2606, pp. 48–58. SPIE, 1995.