アーカイブシステムにおけるバックアップ系列のデータ管理方式に関する検討

高橋　秀和† 　大森　匡† 　星　守†

† 電気通信大学大学院情報システム学研究科　〒 182–8585 東京都調布市調布ヶ丘 1–5–1
E-mail: †{hide,omori,hoshi}@hol.is.uec.ac.jp

あらまし　近年，多数の PC のストレージバックアップを一ヶ所のアーカイブシステムに集めて管理する形態が注目されている。このような状況では，蓄積されたバックアップ系列のデータ量は 1 ペタバイト級にまでなると予想される。そのため、集められたバックアップ系列を RAID, MAID(低消費電力ディスクファーム), テープロボットの 3 つからなる階層記憶上で適切に分割配置する必要がある．一方，PC の利用環境では，利用目的や保存されているデータは多様であり、同時に、各 PC からのデータのリストア要求や複数の PC に分散した共有データのリストア，リストアしたい時間帯、リストア時間最適化等、多様な要求に応える必要がある。本稿では，以上の要求を扱うことを動機として、リストアの要求分布に応じたデータクラスタリングを使ってバックアップ系列のデータ管理を行う方式を検討し、シミュレーションによる試行結果を述べる．

キーワード　バックアップ管理, ストレージシステム, データライフサイクル管理, アーカイブ

An Approach of Backup-Data Management in Mass Archive Systems

Hidekazu TAKAHASHI†, Tadashi OHMORI†, and Mamoru HOSHI†

† 1–5–1 Chofugaoka, Chofu-shi, Tokyo, 182–8585 Japan
E-mail: †{hide,omori,hoshi}@hol.is.uec.ac.jp

Abstract　Recently an upcoming trend of enterprise data management is a centralized management of many PC's backup-data in archival storage systems. The size of these backup data can grow up to a petabyte scale. Thus such backup datasets need to be divided and be stored appropriately onto a system complex of RAIDs, MAIDs (farms of low power-consumption disks) and a tape robot. On the other hand, an archival storage system must satisfy many requests, including restoring a single PC as of a given time-point or time-interval, restoring a shared data-space of multiple PC's, or finishing the restoring itself within a given time-period. Motivated by this situaion, this paper discusses an idea of applying a data-clustering technique to manage sequences of huge backup datasets in an archival storage system. A simulation test of its effects on an archival storage system is described.

Key words　Backup, Storage System, Data Life Cycle Management, Archive System

1. 本稿の背景と目的

今日，散在するクライアント PC のストレージを一ヶ所のバックアップサーバやストレージサーバに集中管理する事例が増えている [1] [2] [3]。 このような背景に立って、本稿は、複数の PC のバックアップ系列を 1 つのアーカイブシステムに集め，ストレージ階層の中でリカバリ時間要求を考慮したデータ配置管理の方式を検討する。

図 1 に、本稿におけるバックアップ系列の集中管理を行う状況を示した。図中、クライアント PC はバックアップの取り方とリストアの要求について、適当なポリシーを持っているとする。バックアップサーバは、このポリシーに沿って、バックアップ系列のデータをとり、1 次ストレージとして常時稼働しているアレイディスクに集める． 蓄積されるバックアップ系列のデータ量は総合的には数ペタバイト級になるため，時間の経過とともに、古くなったデータは、1 次ストレージから低消費電力型ディスク装置 [1] [4] [6] (図中の 2 次ストレージ) やテープ装置階層 (図中の 3 次ストレージ) に移

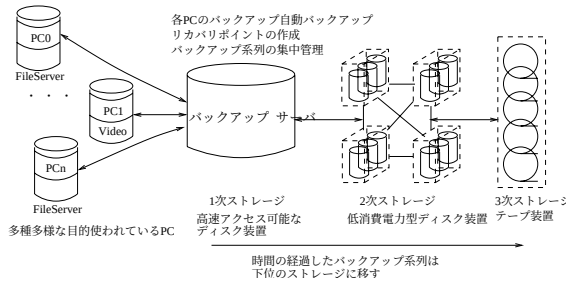


図 1　バックアップ系列の集中管理モデル

すと仮定する。

一般に，1 次ストレージから下位のストレージに移されたバックアップ系列は，時間軸に沿って一定サイズごとにまとめて管理される．一方で，PC の利用特性は多様であるため，発生するリストア要求も多様である．例えば，ビデオサーバのような PC では,「過去 3 日間に録画されたデータをリストアせよ」といったような，時間軸に沿って長いリストア要求が多く発生すると考えられる．普通のファイルサーバ用途が主であれば、複数 PC にまたがった共有フォルダの指定時点へのリストアなどが考えられる。このような多様なリストア要求が混在したとき、既存の管理手法では，リストアに必要なバックアップ系列が複数のディスクやテープに無条件に分割管理されるため，リストア処理時間は下げられない。このことから，下位のストレージに移されたバックアップ系列の管理において，バックアップ対象である PC の利用特性や発生するリストア要求の性質を考慮したデータ管理手法が，リストア要求の処理時間制約や 2 次ストレージの駆動ドライブ数削減に対して効果的であると期待できる。

以上の動機にたって、本稿では，2 次/3 次ストレージにおけるバックアップ系列の分割配置方法として、文献 [7] で提案した GR 木によるアーカイブ管理技法を適用した時の効果をシミュレーションによって示す.

2. 問題設定とアプローチ

以下，本稿の仮定をあげる.

1）クライアント PC のリストア要求について

一般に，PC の使用目的は多様であり，様々なリストア要求が発生する．これらのリストア要求には，ある種のデータ空間のリストアにかかる時間への制約 (RTO) や，リストア可能な時点の範囲への制約 (RPO) がある.

クライアント PC は，バックアップサーバに対して，自分のデータ空間についてのリストア要求の型を，2 次ストレージにいる場合と 3 次ストレージに移した後の場合について記述すると仮定する．注意すべきは，観測記録データ列か通常のファイル空間かなどのコンテンツ情報自体は教えない．そのようなデータモデル的な情報 [9] の利用は本稿では考えない。

リストア要求の型は，1 つの PC について時系列的に取り出したい場合と，複数の PC にまたがった共有データ空間を取り出したい場合 (データグリッドなどの共有データプール) を考える．1 つのデータ空間に複数のリストア要求が混合することは許す．このようなリストア要求のことを、以下では、(リストアの) ポリシーと呼ぶ。

2）2 次ストレージのモデル

2 次ストレージとしては，最近登場している一時停止型ディスクの集まり (MAID) [6] を想定する．MAID は，一定期間アクセスのない時は一時的に停止状態になる低消費電力仕様ディスク装置のクラスタである．3 次ストレージは従来のテープアーカイバとする.

MAID は，単純に，強制的な駆動とサスペンドをバックアップサーバから指示できるようなディスク装置 N 台とする．(本稿では $N = 16$ 台程度とする). MAID は，頻繁にアクセスされないようなデータ群の記憶用途であり、N 台のうち同時に稼働するドライブを少なく保って消費電力を抑制する [6] [1]。主に、ディスク to ディスク型のアーカイブ機構の一つと考えられる(注1)。

以上の仮定の下で，本稿は，クライアント PC のリストア要求が多様に与えられたとき，適切にバックアップ系列をクラスタリングして 2 次/3 次ストレージの記憶格納単位へとクラスタを配置することで，低消費電力，かつ，リストア要求の制約を満たしたデータ管理方式を示す.

3）提案方式の概要

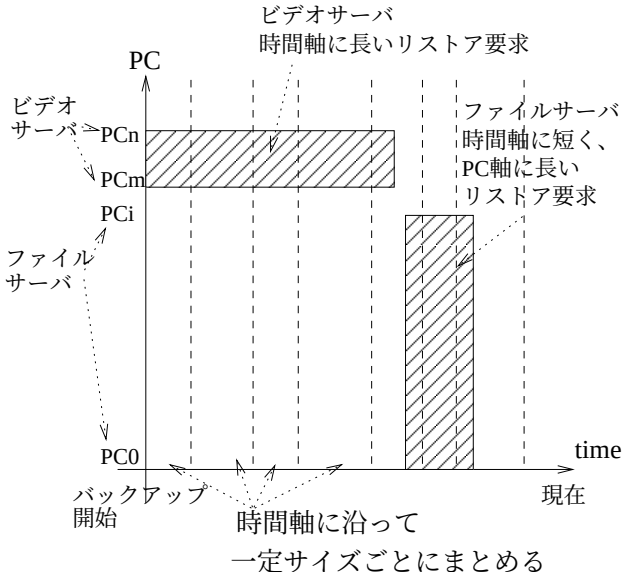
ここで，データクラスタリングの手法として、GR 木と呼ぶ手法 [7] を用いる。GR 木とは，多次元データのうちいくつかの部分次元で問い合わせが出ることを考慮して適切に空間分割を行う R 木の一種である．著者らは，文献 [8] で GR 木を使ったアーカイブデータの管理技法を検討したことがある．そこで本稿では、図 2 のように，時間軸と PC 軸の 2 次元にバックアップデータを置いて，リストアの問い合わせ要求に応じてこのデータ群をクラスタへ分割する。すなわち，2 次/3 次ストレージのディスクやテープ 1 台を葉ノード 1 つと考え，葉ノードに対応するバックアップデータの集合を、そのノードに対応した記憶領域に格納する.

図 2 に、本稿で検討する分割配置方式の概要を示す。同図-(a) は、既存の分割方式、つまり、時間軸に沿って全 PC のデータを一定サイズの記憶単位に分割し、そのまま単位記憶領域へと配置する場合を表す。ここで、ビデオサーバ用途の PC では時間軸に沿って長い幅、つまり一定の時区間に入るデータ列を取り出したいはずであるし、ファイルサーバなら、指定一時点へのスナップショットをリストアする要求が多いと考えられる。しかし、図 2 では、こうしたリストアしたいデータの集りを無視して分割する。 一方、図 2-(b) は、

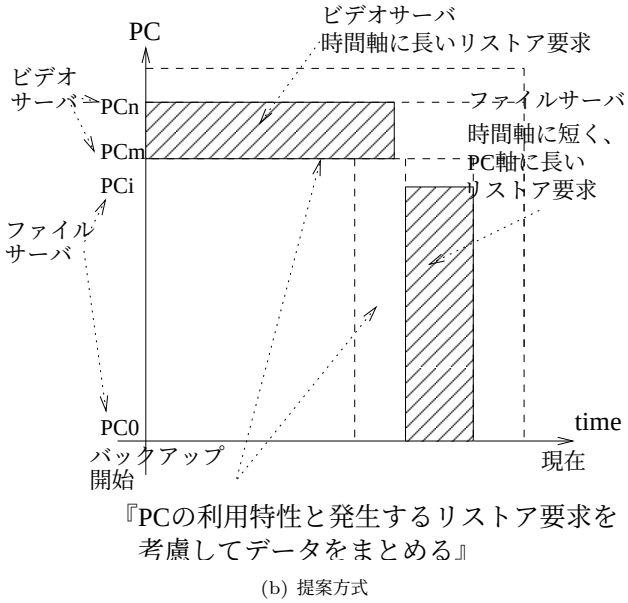
(注1)：ただし、頻繁な駆動の on/off があると信頼性が下りやすいという課題もある.

GR 木により、PC 軸と時間軸の 2 次元空間上のデータ列をクラスタリングして葉ノードへ分割した場合である。この方が、アクセスする葉ノードに対応する MAID のドライブや 3 次ストレージのテープへのアクセス数は減るはずである。

以下、この考えに基づいて、具体的な事例を使って、適切なバックアップデータ系列の管理技法ができるか，2 次ストレージの駆動ドライブ数を低く保てるか，等を評価する．



(a) 既存方式



(b) 提案方式

図 2　バックアップ系列の管理方式とリストア要求

3. バックアップ系列管理用アーカイブシステムのモデル

本節では，本稿で設定するアーカイブシステムにおけるバックアップ系列管理のモデルを示す．

このモデルは，バックアップ対象である 10 台の PC，1 次ストレージの高速アクセス可能なディスク装置，2 次ストレージの低消費電力型ディスク装置，そして，3 次ストレージのテープロボットから成る．

バックアップ対象である 10 台の PC（PC0〜PC9）は，利用特性が異なる．10 台の PC の中で PC8, 9 は，HDD ビデオサーバとして主に利用され，残りの PC0〜PC7 に関しては，ファイルサーバとして主に利用されるとした．

1 次ストレージ上では，バックアップサーバが稼働する．各 PC のバックアップは，このバックアップサーバによって行われる．バックアップサーバは，バックアップイメージを集めて管理する．各 PC，1 日 1 回，PC 全体のバックアップイメージ（以下，フルイメージ）を採る．さらに，1 日 5 回，差分バックアップを行い，そのバックアップイメージ（以下，差分イメージ）を採る．ただし，各 PC のフルイメージの取得は，1 週間に 1 回は PC 全体のフルバックアップによって取得する．それ以外のフルイメージ取得については，1 次ストレージ上のバックアップサーバによって，以前に採ったフルイメージと差分イメージから新たなフルイメージを生成していると仮定する。これらのバックアップイメージは，1 次ストレージ上に，1 週間（7 日間）集め続けられる．

図 3 に、このときの PC1 台あたりのバックアップデータ列を示す。

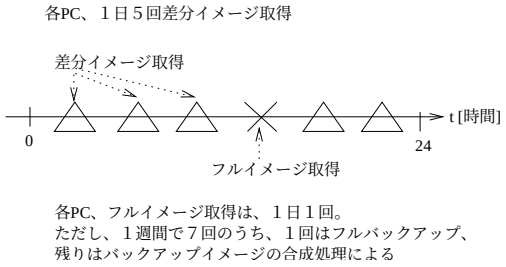


図 3　1 次及び 2 次ストレージにおける PC 1 日分のバックアップ系列

集められた各 PC のバックアップイメージは，時間の経過とともに使用頻度が下がる．そこで，1 次ストレージに保存されたバックアップイメージは，1 週間経過する毎に，2 次，3 次と，より下位のストレージへと移して行く．

先ず，バックアップ開始から 1 週間経過後，1 次ストレージに集められたバックアップイメージは，2 次ストレージへ移す．この時，バックアップイメージの集合に対して，GR 木を利用したデータのクラスタリングを行う．GR 木の葉ノードには，バックアップイメージ自体が収められる．そして，1 つの葉ノードに収められたバックアップイメージの集合は，2 次ストレージの 1 つのディスクドライブに保存される．すなわち，バックアップイメージの集合は，このクラスタリン

グによって適切なパッキングが行われ，2 次ストレージに保存される．

さらに 1 週間経過すると，2 次ストレージに移されたバックアップイメージを、3 次ストレージにあたるテープロボットに移す．テープロボットに移されたバックアップイメージへのアクセス頻度は，極端に減少していると考えられる．そこで、テープにバックアップ系列を移す際，PC の利用特性に従ってバックアップ系列を圧縮する (図 4 参照)．ファイルサーバとして主に利用される PC のバックアップイメージについては，差分イメージは全て残す．フルイメージは，各 PC，1 週間のうち最初の 1 回のみ残し，残りは削除する．HDD ビデオサーバとして主に利用される PC に関しては，フルイメージのみを残し，差分イメージは全て削除する．バックアップ系列の圧縮完了後，GR 木によるクラスタリングを再度行い，データのパッキングを変更する．この圧縮により，リストア処理の時間が犠牲になるが，アクセス頻度を考えて問題視しない．

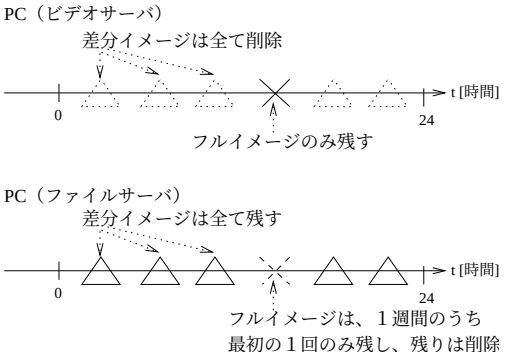


図 4　2 次から 3 次ストレージへの各 PC のバックアップ系列の圧縮

一方，バックアップサーバでは，集められたバックアップイメージの集合に対してディレクトリ情報管理を行う．これにより，リストア要求が発生すると，まず，このディレクトリ情報から，リストアのために必要なバックアップイメージのメタ情報，つまり，どの PC の何時の時点のバックアップイメージを取得する必要があるかに関する情報を取得する．そして，この情報に基づき，バックアップサーバは，バックアップイメージを取得しリストア処理を行う．

4. シミュレーション結果

前節で示したバックアップ系列管理のモデルに基づき，リストア要求を与えてシミュレーションを行った．そして，リストア要求発生時の，2 次ストレージにおけるディスクドライブの平均駆動数についての評価を行った．

10 台の PC から 1 週間集めたバックアップイメージの集合は，時間軸 (t 軸) と PC 軸による 2 次元空間上で管理する．この 2 次元空間における t 軸の値域は，$0 \leqq t \leqq 168$，単位は [時間]．PC 軸の値域は，$0 \leqq PC \leqq 9$ で，全 PC10 台を表す．各 PC のバックアップイメージのサイズは，固定とし，フルイメージは，50GB，差分イメージは，1GB とした． 2 次，3 次ストレージにおけるディスク，またはテープ 1 本の容量は，250GB とした．これに合わせて GR 木の葉ノードのサイズは，250GB とした．また、GR 木でデータ分割を行うに際し、各バックアップデータは、1GB のチャンクの集まりとして表現した。

GR 木設計時の問い合わせ分布は，具体的なリストア要求を想定して用意した Q1〜Q3 の 3 種類の混合とした．各問い合わせ分布の，問い合わせ発生領域，問い合わせ幅，発生確率を表 1 にまとめる．全て時間軸と PC 軸に対する 2 次元問い合わせである。このうち、Q1 は，HDD ビデオサーバとして利用された PC において見られる，時間軸に沿った長いリストア要求を想定している．Q2 に関しては，ファイルサーバとして利用されている PC 数台を，指定された時点の状態をリストアする要求を想定している．Q3 については，指定された 1 台の PC について，指定された時点の状態へのリストアを想定する．この Q3 が想定するリストア要求は，基本的要求であると考え，問い合わせ発生領域を全域とした．

表 1　GR 木設計時の問い合わせパターン

	問い合わせ発生領域	問い合わせ幅	発生確率
Q1	$[0,168] * [8,9]$	$t = 72, pc = 1$	40%
Q2	$[0,168] * [0,7]$	$t = 24, pc = 3$	50%
Q3	$[0,168] * [0,9]$	$t = 24, pc = 0$	10%

この問い合わせ分布に従って，2 次ストレージ上に保存されたバックアップ系列に対して，リストア要求を与えた．与えたリストア要求は、以下の $q1 \in Q1, q2 \in Q2$ の 2 種類である：

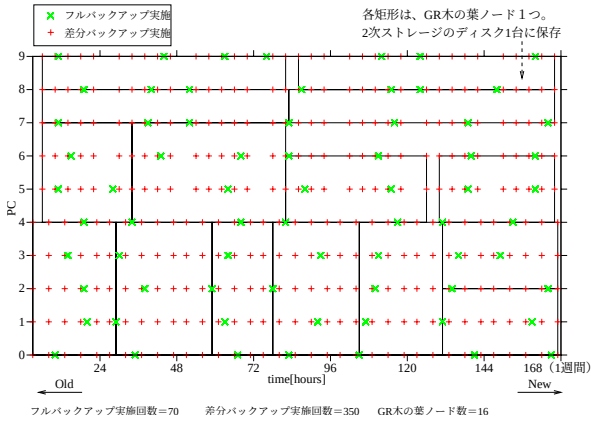
* $q1 \in Q1$ は，PC9 について、1 時点 t_1(=48,72,96,120,144 のどれか 1 つ) を指定されて起動し、t_1 から見て 24 時間前から t_1 までの時区間における PC9 のデータの変分をリストアする。つまり、1 週間のうち第 x 日目に記録したデータ列を PC9 について取り出す要求である。$t_1 - 24$ の状態のリストアと、それ移行 t_1 までの差分イメージとフルイメージ全てを取り出すことになる。

* $q2 \in Q2$ は，PC0〜PC4 のうちどれか一つの PC について、指定された時点 t_2 における状態を全てリストアする。t_2 は、24, 48, 72, 96, 120 のうちから 1 つ指定する。t_2 から見て直近のフルイメージとそれ以後 t_2 までの差分イメージを取り出す要求である。

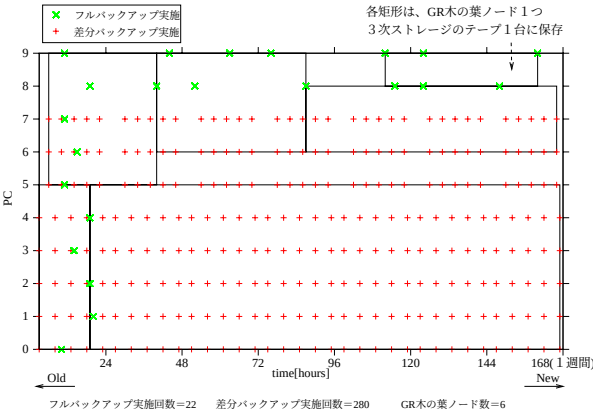
上記のリストア要求を、GR 木に基づいてデータ群を分割して配置管理を行った場合 (図 2-(b)) と，既存手法による場合 (図 2-(a)) について適用した。

図 5 に、2 次ストレージと 3 次ストレージにおいて GR 木を利用してデータ管理を行った際のバックアップデータ系列のパッキングの様子を示す．同図-(a) が 2 次ストレージにお

ける分割を、同図-(b) が 3 次ストレージのそれを示す。(図中の×印はフルバックアップイメージ、＋印は差分イメージである). 比較として、既存手法によるデータ管理を行った様子を、2 次ストレージの場合について図 6 に示した。いずれの場合も、分割のサイズは 250GB 単位である。



(a) 2 次ストレージ



(b) 3 次ストレージ

図 5　GR 木を利用したバックアップ系列管理

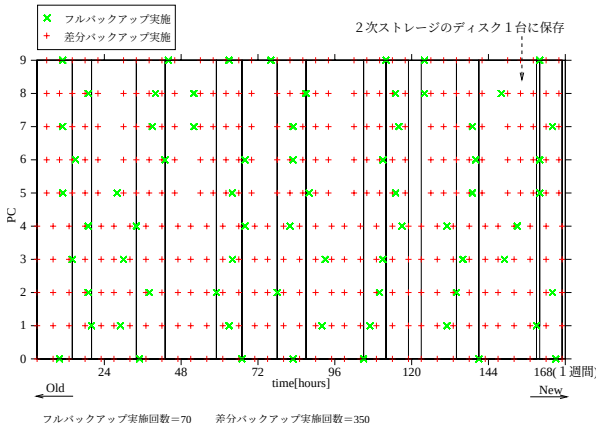


図 6　既存手法によるバックアップ系列管理（2 次ストレージ）

表 2 に、リストア要求 $q1$(5 通り) における提案方式と既存方式の 2 つについて示す。同じく、表 3 には、$q2$(5x5=25 通り) についての 2 次ストレージ上の駆動ドライブ数の平均値を示す。また、$q1, q2$ 各々の実行に必要なフルイメージと差分イメージの数の平均値を、表 4 に示しておく。$q1, q2$ どちらのリストア要求を与えた場合でも，GR 木を利用した場合の方が，既存手法による場合に比べ，駆動ディスクドライブ数が少ない．3 次ストレージにおいても、ほぼ同様の結果が得られている。

表 2　平均駆動ドライブ数 (2 次ストレージ). $q1$ の場合.

	平均駆動ドライブ数
GR 木適用手法	1.40
既存手法	4.40

表 3　平均駆動ドライブ数（2 次ストレージ). $q2$ の場合.

	平均駆動ドライブ数
GR 木適用手法	1.44
既存手法	2.16

表 4　各リストア要求が必要とするバックアップイメージの平均数

	フルイメージ	差分イメージ
q1	2.00	7.60
q2	1.00	2.24

5. ま と め

本稿では，複数の PC のバックアップ系列を一カ所のアーカイブシステムに集中管理する状況において，PC の利用目的とリストアポリシーに基いたバックアップデータのクラスタリングを行うことによって、ストレージ階層の中でリカバリ時間要求を考慮したデータの配置管理が可能であるかを検討した。

一般に、アーカイブ環境においては、時空間データのようにコンテンツの属性情報のようなデータモデルが明らかな場合には、それに特化したデータ管理が上位層で行われている [9] [10]．これに対し、本稿では、クライアント PC のリストアポリシーだけを情報として用いた。そして、複数 PC のフルバックアップイメージと差分イメージの系列を集中管理するとき、バックアップデータ列のクラスタリングを所与のリストア要求にあわせて行うことで、2 次ストレージ層で使う MAID 型ディスクの平均駆動ドライブ数を低減でき、結果的に、ストレージ階層におけるデータ集合の分割配置として適切になりうることを示した。

まとめとしては、リストアポリシーしかわからない条件でアーカイブストレージを構成するとき、従来からあるブロック相関度やアクセス頻度情報に加え、本稿の手法も一つの設計の選択肢であるとは言える。逆に、これ以上の効率化の

ためには、対象データの属性情報等に基づいたインデックスづけなど、上位層のデータ管理機能との連係が必要である. MAID 自体の信頼性や低消費電力用途の制御方法として妥当な範囲など、考慮すべき課題も多い。

文　献

[1] Enterprise Watch. NTT-IT、MAID 装置を用いた低消費電力型オンラインストレージ http://enterprise.watch.impress.co.jp/cda/hardware/2006/01/11/6981.html.

[2] Enterprise Watch. シマンテックの Windows 向けデータ保護ソリューション「Backup Exec 10d」. http://enterprise.watch.impress.co.jp/cda/storage/2005/12/26/6927.html.

[3] Enterprise Watch. マイクロソフトのデータ保護ソリューション「Microsoft System Center Data Protection Manager 2006」 http://enterprise.watch.impress.co.jp/cda/storage/2005/10/31/6498.html.

[4] 戸田誠二, 江尻革, 太田光彦, 野口泰生, 武理一郎. オーガニックストレージシステムの大規模ハードウェアへの実装. 電子情報通信学会 信学技報, CPSY2005-16, pp. 7–12, Aug. 2005.

[5] 武理一郎, 野口康生, 土屋芳浩, 荻原一隆, 田村雅寿, 丸山哲太郎. オーガニックストレージシステム -自律し成長するストレージシステム-. 電子情報通信学会 信学技報, CPSY2004-57, pp. 55–60, Dec. 2004.

[6] Dennis Colarelli et al.. Massive arrays of idle disks for storage archives. IEEE Supercomputing 2002, pp. 1–11, 2002.

[7] 大森，佐藤，星. 問い合わせ分布を考慮した R 木における領域分割方式. 電子情報通信学会論文誌 D-I, J86-D-I No.10, pp. 746–761, Oct. 2003.

[8] 高橋，大森，星，高塚. 問い合わせ分布に適応した多次元ファイル編成法 GR 木のアーカイブ環境への適用. FIT2002, LD-5, 2002.

[9] Alexander S. Szalay, Jim Gray et al.. The SDSS skyserver: public access to the sloan digital sky server data.. ACM SIGMOD2002, pp. 570–581, 2002

[10] K.Holtman et al.. Towards Mass Storage Systems with Object Granularity. IEEE Mass Storage Systems(MSS) 2000, 2000.