RAID型ファイルシステムVAFS/HRのファイルアクセス制御方式

6U－6

高良 亜紀子　山下 洋史*　高橋 英男*　畠山 敦*　裏谷 郁夫*　山田 秀則　城田 浩二
日立コンピュータエンジニアリング（株）　*（株）日立製作所

1．はじめに

近年，ネットワークの普及に伴い複数のワークステーション間でファイルの共有化が進んでいる。共有化されたファイルには複数のユーザがアクセスするため，ファイルアクセスの高速化とファイルデータの高信頼化が要求される。筆者らは既に，複数のディスク装置にファイルを分割して格納する”バーチャルアレイ・ファイルシステム(VAFS)”を開発し，ファイルアクセスの高速化を実現している[1][2][3]。今回，更にパリティデータを付与して，ディスク装置の故障に対するデータ保証を行うRAID型ファイルシステムVAFS/HR(High Reliability)を提案し，ファイルシステムの高性能化，高信頼化に取り組む。本稿では，VAFS/HR上でのファイルアクセス制御方式について報告する。

2．VAFS/HRのファイルアクセス制御方式

2.1　ファイルアクセス制御の課題

ユーザが指定したファイルに対してアクセスを行うためには，ディレクトリ検索機能と該当するファイルのデータアクセス機能の二つの機能が必要である。これら二つの機能をサポートするにあたって，VAFS/HRではそのファイルの格納形態に付随し，性能上解決するべき次の二つの課題がある。

（1）ディスク装置増加時のディレクトリ検索時間の短縮

図1に示すように，VAFS/HRを構成する各ディスク装置はUFS(UNIX File System)として管理されており，同じディレクトリ構造となっている。分割されたデータはサブファイルとして所定のディレクトリ下に配置される。そのため，VAFS/HR上でファイルアクセスするためには，各ディスク装置ごとにユーザが指定したパス名にしたがってディレクトリを検索していけばよい。しかし，VAFS/HRを構成するディスク装置の台数が増加していくと線形的にディレクトリ検索の時間が増加してしまうため，これを防ぐ必要がある。

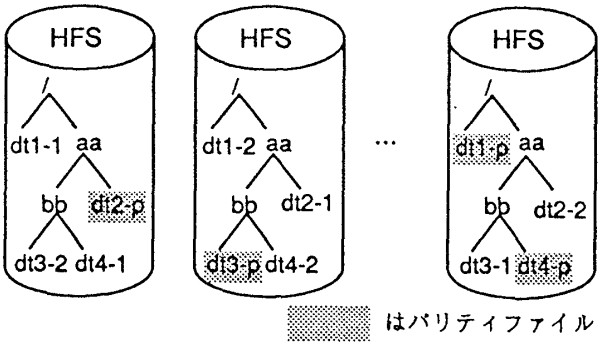

図1　VAFS/HRのディレクトリ構造

（2）ディスク装置故障時のシステムバッファの有効利用

UNIXでは，ディスク装置から読み出されたファイルのデータは，ディスク装置番号とディスクの物理的なブロック番号で管理されたシステムバッファでキャッシングされ，ファイルアクセス性能の高速化が行われている。しかし，ディスク装置故障時に他のサブファイルから復元したサブファイルのデータはそのままではシステムバッファにキャッシングできない。復元したデータについては故障ディスク装置に格納されたそのファイルのi-nodeにアクセスできず，システムバッファにキャッシングするために必要なディスクの物理的なブロック番号がわからないからである。そこで，物理ブロック番号がわからなくてもシステムバッファを管理し，他のサブファイルから復元したサブファイルのデータについてもシステムバッファを利用できるようにする必要がある。

上述したこれらの課題を解決するために表1のアプローチを採用する。

The file access method of VAFS/HR - a software RAID file system
Akiko Kora, Hirofumi Yamashita*, Hideo Takahashi*, Atsushi Hatakeyama*, Ikuo Uratani*, Hidenori Yamada and Koji Shirota

表1　ファイルアクセス制御の課題とアプローチ

No.	課　題	アプローチ
1	ディレクトリ 検索時間の短縮	サブファイル リンク方式
2	縮退時のシステム バッファの有効利用	縮退時対応システム バッファ管理方式

2.2　サブファイル・リンク方式

本方式は，図2に示すように対応するディレクトリやサブファイルのi-node内に互いのi-node番号を登録することにより，1台のディスク装置に対してディレクトリ検索を行うだけで他のディスク装置内のサブファイルのi-nodeを獲得できる方式である。以下に本方式におけるディレクトリ検索アルゴリズムを示す。

(step1) 1台目のディスク装置に対してディレクトリ検索を行う。
(step2) 1台目のディスク装置が故障している場合には，2台目のディスク装置に対してディレクトリ検索を行う。
(step3) ユーザが指定したパス名のサブファイルやディレクトリのi-nodeをディレクトリ検索を行った1台目（2台目）のディスク装置から獲得する。
(step4) step3で獲得したi-node内に登録されている他のサブファイルのi-node番号をもとに，次のディスク装置のサブファイルのi-nodeを獲得する。次のディスク装置が故障している場合には，その次のディスク装置のサブファイルのi-nodeを獲得する。
(step5) step4をディスク装置台数分，繰り返す。
(step6) ユーザの指定したパス名要素分，step1からstep5の動作を繰り返す。

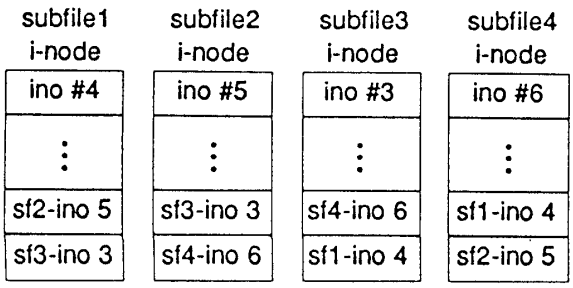

図2　サブファイル・リンク方式

本方式により，ディスク装置台数が増加しても1台目（2台目）でだけディレクトリ検索を行えばよく，ディレクトリ検索時間が短縮できる。

2.3　縮退時対応システムバッファ管理方式

本方式は，ファイルの論理ブロックをキーにシステムバッファの管理や検索を行えるようにした方式である。これによって，物理ブロック番号がわからない故障ディスク装置上のサブファイルのデータについてもシステムバッファを利用できる。

本方式を実現するために，図3に示すようにパリティデータの生成や管理を行うためのテーブルであるblinksテーブルを用いる。blinksテーブルは，各サブファイルの同一の論理ブロック番号のブロックで構成されるパリティグループごとに割り当てられている。物理ブロック番号がわからないデータに関しては，サブファイルの論理ブロック番号をもとにblinksテーブルからシステムバッファを検索する。

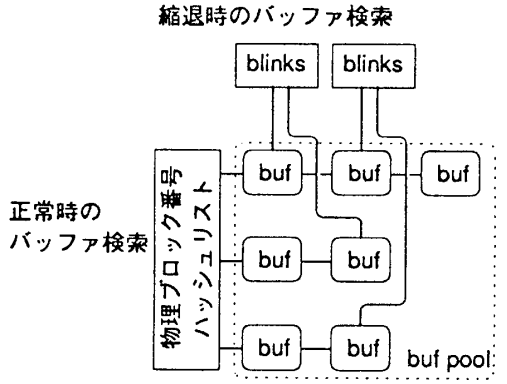

図3　縮退時対応システムバッファ管理方式

3.　おわりに

VAFS/HRでは，サブファイル・リンク方式と縮退時対応システムバッファ管理方式を開発した結果，ディスク装置が故障した場合でも正常時と同様のファイルアクセスが可能となった。

参考文献

[1]秋沢他5，「バーチャルアレイ・ファイルシステム(vafs)の基本構想」，情報処理学会第45回全国大会講演論文集4-62，(平4-10)
[2]秋沢他6，「ストライプド高速UNIXファイルシステムの開発」，情報処理学会システムソフトウェアとオペレーティングシステム研究会61-2，(平5-8)
[3]鬼頭他6，「高速UNIXファイルシステムの構想」他4件，情報処理学会第47回全国大会講演論文集，7B-1~5，(平5-10)

注)UNIXオペレーティングシステムはUNIX System Laboratories, Inc.が開発し，ライセンスしています。