RAID型ファイルシステムVAFS/HRの障害回復方式

6U－7

山田 秀則　山下 洋史*　高橋 英男*　畠山 敦*　裏谷 郁夫*　城田 浩二　高良 亜紀子
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加して，ディスク装置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，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運用面で重要となるVAFS/HRの障害回復方式について報告する。

2．VAFS/HRの障害回復方式

2.1　障害回復の課題

VAFS/HR使用中に発生する障害としては，ディスク装置が故障する場合と停電やOSのパニックによるシステム・ダウンの場合の二つに大別できる。これらの障害が発生した場合にはすみやかに回復処理を行えるようにする必要がある。以下にこれら二つの障害に対する回復処理の課題を示す。

（1）故障ディスク装置の交換とデータの復元

VAFS/HRはパリティデータをファイルに付加することにより，ディスク装置１台の故障までならデータの復元を保証する。しかし，ディスク装置が２台以上故障するとデータの復元は不可能となるため，ディスク装置の故障が発生した場合は次の故障が発生する前にそのディスク装置を交換しデータを復元しておく必要がある。

（2）システム・ダウン発生時のVAFS/HRの整合性のチェック

VAFS/HRでは遅延書き込みをサポートしているため，ファイル更新中にシステム・ダウンが発生するとファイル管理情報の不整合や図１に示すようなファイルデータとパリティデータの不整合が発生する可能性がある。このような不整合は，ファイルのデータ不正やファイルシステム破壊につながる恐れがある。特にパリティデータの不整合を放置したままの状態でディスク装置故障が発生しデータの復元を行うと，復元したデータは不正なものとなってしまう。したがって，システム・ダウンが発生した場合はファイルにこのような不整合が生じていないかをチェックする必要がある。

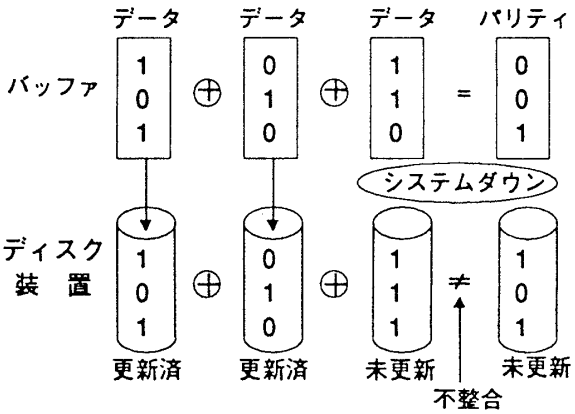

図１　システムダウン発生時の課題

上述したこれらの課題に対して表１のアプローチを採用する。

表１　障害回復の課題とアプローチ

No.	課　題	アプローチ
1	故障ディスク装置の交換とデータの復元	va_recoveryコマンド
2	システムダウン時のVAFS/HR整合性チェック	VAFS/HR用 fsckコマンド

System failure recovery in VAFS/HR - a software RAID file system
Hidenori Yamada, Hirofumi Yamashita*, Hideo Takahashi*, Atsushi Hatakeyama*, Ikuo Uratani*, Koji Shirota and Akiko Kora
Hitachi Computer Engineering Co.,Ltd.
*Hitachi, Ltd.

2.2 va_recoveryコマンド

va_recoveryコマンドは，図2に示すように故障したディスク装置上に構築されていたデータを復元するコマンドである。VAFS/HRではVAFS/HRを構成する各ディスク装置はUFS(UNIX File System)として管理し，分割したファイルデータやパリティデータはUFSのファイルとして管理している。そのため，va_recoveryコマンドは(1)ディレクトリ構造復元モジュールと(2)ファイルデータ復元モジュールおよび(3)ユーザ操作ミス防止モジュールの三つのモジュールから構成される。

（1）ディレクトリ構造復元モジュール

VAFS/HRを構成する各ディスク装置のディレクトリ構造は同一であるため，他の正常なディスク装置のディレクトリ構造を交換ディスク装置にコピーする。

（2）ファイルデータ復元モジュール

ファイルデータは，他の正常なディスク装置に格納されているパリティファイルを含めたサブファイルの排他的論理和を計算し復元する。

（3）ユーザ操作ミス防止モジュール

ディスク装置が故障した場合には故障ディスク装置番号を他の正常なディスク装置のシステムエリアに書き込む。ユーザがディスク装置を交換しデータ復旧を行う場合には，交換したディスク装置が故障ディスク装置であるかを他の正常ディスク装置のシステムエリアに書かれている故障ディスク装置番号からチェックし，ディスク装置交換時のミスを防止する。

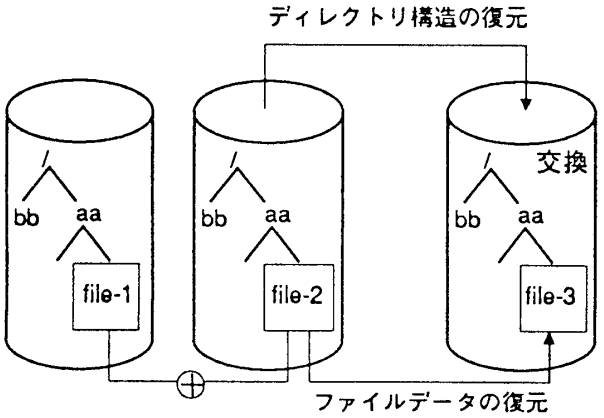

図2　交換ディスク装置のデータ復元方式

2.3　fsckコマンド

VAFS/HR用のfsck コマンドは，(1)ファイル管理情報の整合性チェックモジュールと(2)パリティデータ整合性チェックモジュールから構成される。

（1）ファイル管理情報整合性チェックモジュール

まず，各ディスク装置がUFSとして不整合がないかチェックし，次にディスク装置間で不整合がないかチェックする。ディスク装置間での主なチェック項目は，ディレクトリ構造の同一性チェックである。ここまでのチェックで不整合が検出されたファイルは削除する。そして，残ったファイルに対して(2)のファイルデータとパリティデータの整合性チェックモジュールで整合性のチェックを行う。

（2）パリティデータ整合性チェックモジュール

各ディスク装置に格納されているファイルデータからパリティデータを計算し，それがディスク装置に格納されているパリティデータと一致しているかチェックする。このチェックの高速化を図るために，i-nodeにファイル更新中フラグを追加しチェック対象となるファイルを絞り込むようにする。このフラグは，ファイル更新開始時にオンにし，終了時にオフにする。fsckコマンド実行時にこのフラグがオンになっているファイルはパリティデータの不整合の可能性があるファイルであり，パリティデータ整合性チェックの対象となる。

3．おわりに

VAFS/HRでは，va_recovery コマンドと VAFS/HR用fsck コマンドを開発した結果，ディスク故障やシステム・ダウンのような障害が発生した場合でも，それを回復して継続運用することが可能となった。

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。