RAID型ファイルシステムVAFS/HRの構想

6U－3

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

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。筆者等は既に，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（Higih 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の基本構想について説明する。

2．VAFS/HRの概要

2.1　VAFS/HRの基本仕様

RAID型ファイルシステムVAFS/HRの基本仕様を表1に示す。

VAFS同様にファイルアクセス時に複数のディスク装置を並列に制御することにより，VAFSと同程度の性能を達成する。信頼性の面では，RAIDと同様にパリティデータを格納しておくことでディスク装置故障時のデータ保証を実現する。可用性の面では，ディスク装置が故障しても継続運用を可能にする。

2.2　VAFS/HRの実装部位

VAFS/HRを実装する部位として，ファイルシステム層とデバイスドライバ層の二つが考えられる。表2にそれぞれの比較を示す。

表1　VAFS/HRの基本仕様

No.	項　目	仕　様
1	性　能	VAFSと同程度。シーケンシャル読み出し性能8MB/s (SCSIバス10MB/s時)
2	信頼性	故障ディスク装置のデータ保証
3	可用性	ディスク装置故障時の継続運転

表2　VAFS/HRの実装部位による比較

No.	比較項目	ﾌｧｲﾙｼｽﾃﾑ層	ﾃﾞﾊﾞｲｽﾄﾞﾗｲﾊﾞ層
1	性　能	高速	中速
2	移植性	困難	容易
3	開発規模	大	小

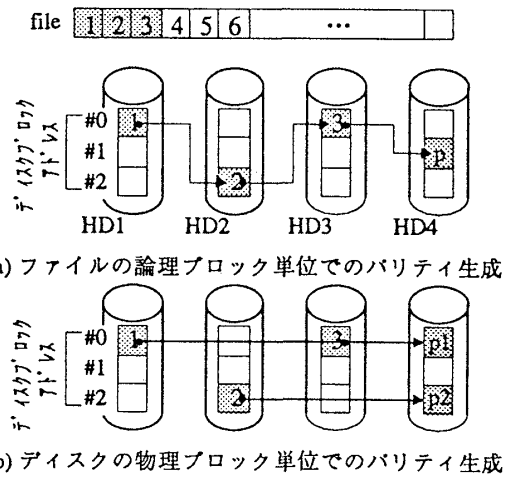
（1）性能比較

デバイスドライバ層に実装するよりファイルシステム層に実装した方が性能は良い。その理由は，次の三つである。

第一に，UNIXファイルシステムは同期ファイルアクセス方法を採用しているが，ファイルシステム層に実装する場合にはこれに手を加え非同期化することが可能である。

第二に，図1に示すようにパリティ生成単位をファイルの論理ブロックとすることができ，たとえファイルがディスク装置上の物理的に不連続に配置されるような場合でもパリティ生成に伴うI/O数を最小にすることができる。

第三に，ファイルシステム層で管理しているシステムバッファを使用してパリティ生成を行うことができるため，パリティ生成用の特別なバッファを設けなくてよくバッファ管理のオーバヘッドを最小にすることができる。



(a) ファイルの論理ブロック単位でのパリティ生成

(b) ディスクの物理ブロック単位でのパリティ生成

図1　パリティ生成の仕方

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

（2）移植性

ファイルシステム層に実装するよりデバイスドライバ層に実装した方が，他のOSへの移植やOSのバージョンアップに伴う移植を行いやすい。その理由は，デバイスドライバ層は入出力インタフェースが各OSとも公開されていることと，インタフェースの変化がほとんど無いことからである。

（3）実装規模

ファイルシステム層に実装するよりデバイスドライバ層に実装した方が，実装規模が小さくなる。その理由は，ファイルシステム層に実装する場合にはopen(), close(), write(), read()などのv-nodeオペレーションと呼ばれる関数全てを実装する必要があるが，全てのv-nodeオペレーションはデバイスドライバ層ではREAD処理とWRITE処理に帰着されるからである。

ファイルシステム層に実装する場合とデバイスドライバ層に実装する場合を比較するとどちらも一長一短であるが，実装規模が大きくなり移植性が悪くなっても性能的に有利なファイルシステム層での実装を選ぶことにした。

2.2 VAFS/HRの全体構成

VAFS/HRの全体構成を図1に示す。VAFS/HRは，(1)ファイル管理モジュールと(2)ファイルアクセス制御モジュールと(3)パリティ生成モジュールおよび(4)障害回復モジュールから構成される。(1)のファイル管理モジュールと(2)のファイルアクセス制御モジュールはVAFSの既存のモジュールに追加変更を加えた。また，(3)のパリティ生成モジュールと(4)の障害回復モジュールは新規に開発した。

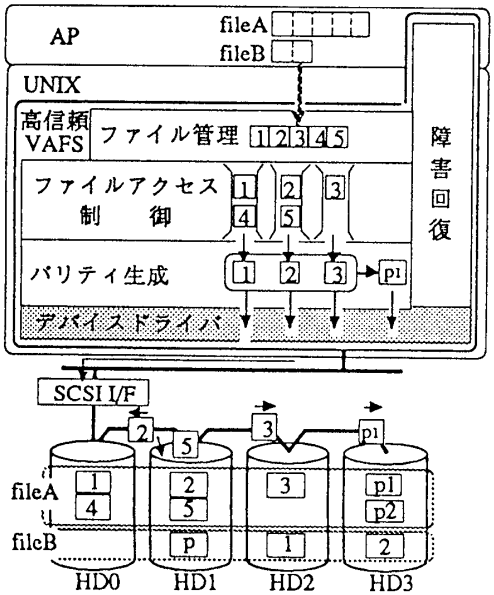


図2 VAFS/HRの全体構成

（1）ファイル管理モジュール

従来のVAFSで行っていたファイル分割（ストライピング）処理に加えてパリティ生成単位であるパリティグループの管理を行う。

（2）ファイルアクセス制御モジュール

VAFSで行っていたディスク装置の並列アクセス制御に加えて高速ディレクトリ検索とディスク装置故障時に対応したシステムバッファの管理を行う。

（3）パリティ生成モジュール

システムバッファ上でパリティ生成処理を行う。

（4）障害回復モジュール

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va_recoveryコマンドと，ファイル管理情報の整合性チェックとパリティデータ整合性チェックを行うVAFS/HR用のfsckコマンドの二つのユーティリティ・コマンドにより障害回復処理を行う。

2.3 性能評価

VAFS/HRを日立製WSである3050RX(PA-RISC 80MHz, SCSI-bus10MB/s)上でプロトタイピングし，性能を測定した。その結果，最大で，正常時に最大8.3 MB/s，ディスク装置故障時に8.0MB/sのファイルアクセス性能が達成でき，VAFS/HRのフィージビリティを確認することができた。

3. おわりに

ファイルを複数のディスク装置に分割格納しさらにパリティデータを付加したVAFS/HRを開発した結果，ディスク装置が故障してもデータが保証できる高性能UNIXファイルシステムが実現できた。

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。