RAID型ファイルシステムVAFS/HRのパリティ生成方式

6U－5

裏谷郁夫　山下洋史　高橋英男　畠山敦　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ともない複数のワークステーション(WS)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，複数のファイル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。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(HighReliability)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上でのパリティデータの生成方式について報告する。

2．パリティデータの生成方式

2.1　パリティ生成時の課題

VAFS/HRではメモリコピーのオーバヘッドを最小限にするために，パリティ生成用の特別なバッファを設けずシステムバッファ上でパリティ生成を行う。パリティ生成処理は，システムバッファからディスク装置へ書き込みを行う延長で起動される。パリティデータは通常のRAIDと異なりディスク装置の物理ブロック単位で生成するのではなく，ファイルの論理ブロック単位で生成する。そのため，パリティ生成方式を実装するにあたっては，次の二つのことが課題となる。

（1）システムバッファ上でのパリティグループ管理

システムバッファのヘッダには，ファイルの論理ブロック番号は登録されていない。これは，システムバッファはディスク装置のキャッシュであり，ディスク装置の物理ブロック番号で管理されているからである。そこで，ファイルの論理ブロック単位でパリティデータを生成するためには，システムバッファからファイルの論理ブロックを一意に求められるようにする必要がある。

（2）パリティ生成処理の高速化

sync時など一度に多数のシステムバッファに対して逐次的にパリティ生成処理を行う場合では，パリティ生成処理がボトルネックとなりファイルアクセス性能が低下する。パリティ生成処理はパリティ生成に必要なブロックの読み出し処理とパリティ計算処理からなり，パリティ生成に必要なブロックの読み出し処理がそのうちの大部分の時間を占める。そこで，パリティ生成の高速化を図るために，パリティ生成に必要な読み出し処理の高速化を行う必要がある。

これらの課題に対して表1に示すアプローチを行った。

表1　パリティ生成の課題とアプローチ

No.	課　題	アプローチ
1	システムバッファ上でのパリティグループ管理	論理ブロック物理ブロック対応管理方式
2	パリティ生成処理の高速化	パリティ時I/O多重化方式

2.3　論理ブロック物理ブロック対応管理方式

本方式は，同一パリティグループのシステムバッファを束ねて管理する方式である。これによりシステムバッファからファイルの論理ブロックを一意に求めることができ，パリティグループ内の他のブロックをアクセスすることが可能になる。

本方式を実現するために，図1に示すblinksテーブルを新規に導入する。blinksテーブルはパリティグループごとに割り当てられる。blinksテーブルには，図2に示すように(1)ファイルの論理ブロック番号と(2)パリティグループ内のシステムバッファへのポインタが登録されている。また，システムバッファからblinksテーブルを参照できるように，システムバッ

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

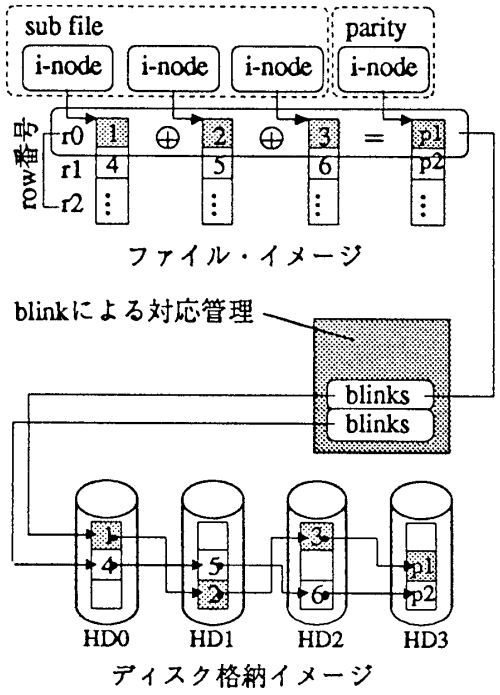


図1　論理ブロック物理ブロック対応管理方式

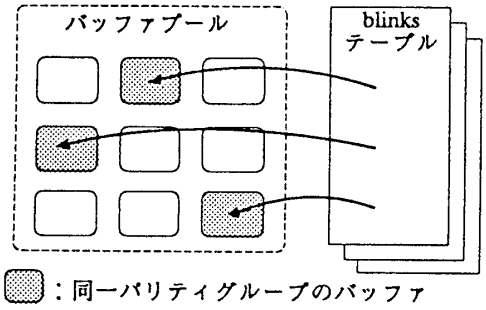


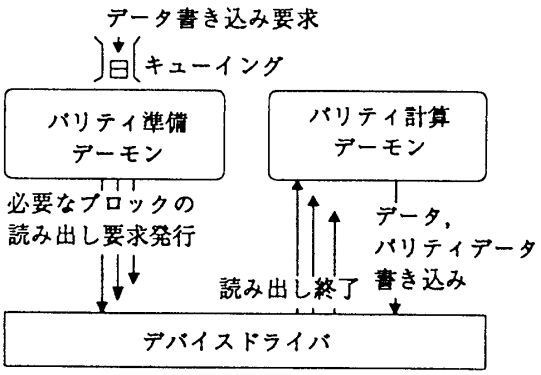
図2　blinksテーブル

ファのヘッダにもblinksテーブルへのポインタを登録しておく。パリティ生成を行うシステムバッファは，自ヘッダのblinksテーブルへのポインタをもとにblinksテーブルを参照し，同一パリティグループのシステムバッファを検索する。

2.4　パリティ時I/O多重化方式

本方式は，パリティ生成に必要なブロックの読み出し及び書き込み処理を多重に行う方式である。本方式により，sync時のように一度に多数のシステムバッファに対してパリティ生成を行う場合でもパリティ生成処理がボトルネックとなることを防ぐことができるようになる。

本方式を実現するために，図2に示すようにそれぞれパリティ準備デーモンとパリティ計算デーモン



と呼ぶ二つの専用プロセスを用意する。パリティ準備デーモンでは，パリティ生成に必要なブロックの読み出し要求を各ディスク装置のデバイスドライバに並列に発行する。この読み出し要求の終了はパリティ準備デーモンでは行わずパリティ計算デーモンで行う。すなわち，パリティ計算デーモンでは，パリティ生成に必要なブロックの読み出し処理が全て終了した時点で起動され，パリティデータを計算する。本方式により，パリティ生成に必要なブロックの読み出し処理を各ディスク装置に並列して行うことができ，複数のシステムバッファに対してパリティ生成処理を行う場合には，それらを多重に行うことができるようになる。

3．おわりに

VAFS/HRでは，論理ブロック物理ブロック対応管理方式とパリティ時I/O多重化方式を開発したことにより，UNIXのシステムバッファを用いてファイルの論理ブロック単位でのパリティ生成を高速に行うことができるようになった。

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。