

RAID型ファイルシステムVAFS/HRの開発

山下 洋史, 高橋 英男, 畠山 敦

(株)日立製作所システム開発研究所

UNIX OSのファイルシステム層でRAIDアーキテクチャを実現したVirtual Array File System / High Reliability を開発した。VAFS/HRではディスク装置を並列に動作させるため、UNIX標準ファイルシステム(UFS)に比べて高速である。また、ファイル単位でパリティグループを構成するため、パリティ生成のために発生する読み出しが従来のRAIDに比べて少ない。10MB/sのSCSIバスにディスク装置5台を接続した構成で、読み出しで8MB/s (UFS比2.3倍) , 書き込みで5.3MB/s (同1.8倍) のスループットを達成した。

DEVELOPMENT OF RAID FILESYSTEM VAFS/HR

Hirofumi Yamashita, Hideo Takahashi, Atsushi Hatakeyama

Hitachi,Ltd. Systems Development Laboratory

1099 Ohzenji Asao, Kawasaki, 215 Japan

The Virtual Array File System / High Reliability (VAFS/HR) has been developed. VAFS/HR is an implementation of RAID as a UNIX file system. Disk parallelism gives performance advantage over the standard UNIX file system UFS. Fewer disk blocks are read at parity generation than conventional RAID because parity groups consist of blocks from the same file. A read performance of 8MB/s (2.3 times faster than UFS) and a write performance of 5.3MB/s (1.8 times faster than UFS) was achieved on a SCSI-bus 10MB/s with 5 disks.

1. はじめに

近年、クライアント・サーバ・システム(CSS)アーキテクチャに基づいてシステム構築が行なわれるようになってきている。CSS環境下では、ファイルの共用と一元管理、並びに大容量の二次記憶空間の提供を目的としたファイルサーバが重要な位置を占める。ファイルサーバに対しては、多数のユーザが同時にファイルアクセスを行うため、ファイルアクセスの高性能化とファイルデータの高信頼化が求められる。そこで、各社とも高性能化や高信頼化のニーズに答えるためにRAIDアーキテクチャ[1][2]の採用を行っている。

RAIDアーキテクチャとは、(1)複数のディスク装置にファイルを分割して格納しディスク装置を並列にアクセスすることと、(2)パリティデータと呼ばれる冗長データを付加することによってディスク装置が故障してもデータが失われないようにすることを特長とするアーキテクチャである。RAIDアーキテクチャの実現手段としては、ディスク・コントローラ内でRAID制御を行うハードRAIDと、WSのオペレーティングシステム内でRAID制御を行うソフトRAIDの二つがある。従来は、ディスクアレイ装置に代表されるようにハードRAIDが主流であったが、最近では、Windows/NT Advanced ServerやCorel RAIDといったソフトRAID製品が登場し始めている[3]。

報告者等も、高速高信頼ファイルシステム技術の開発の一環としてソフトRAIDの開発を行っている。すでに、RAID 0、すなわちファイルを複数のディスク装置に分割格納し、これを並列に読み出すことによりアクセス性能を高める“バーチャルアレイ・ファイルシステム(VAFS)”を開発した[4][5][6]。VAFSは、SCSIバスの転送速度が10MB/sの時に最大8MB/sのファイルアクセス性能を達成したが、RAID 0のためパリティデータの付加は行っておらず、信頼性は低い。そこで、さらにVAFSでRAID 4、すなわちパリティデータの付加を実現し、信頼性を向上させるVAFS/HR(HR:High Reliability)を開発した。本稿では、VAFS/HRの概要について報告する。

2. VAFS/HRの構想

2.1 VAFS/HRの基本仕様

VAFS/HRの基本仕様を表2.1に示す。

これらの仕様を満たすために、VAFS/HRはRAID(Redundant Arrays of Inexpensive Disks)アーキ

表 2.1 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

テクチャをベースとする。すなわち、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、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 ソフトウェアパリティ・オーバーヘッド

VAFS/HRは、UNIXオペレーティングシステム内に実装することを想定しており、ソフトウェアRAIDとなる。そのため、RAID実現のためのCPU負荷、特にパリティ生成のオーバーヘッドの影響の見積もりを行う必要がある。そこで、動作周波数80MHzのCPUを用いてパリティ生成時間と他のファイルアクセス処理時間の実測比較を行った。その結果を、図2.1に示す。

パリティ生成オーバーヘッドは、ファイルアクセス時のOSのオーバーヘッド(ファイルアクセス処理オーバーヘッド)やシステムバッファからユーザ空間へのデータコピー(データコピー)時間に比べて小さく、パリティ生成処理をCPUで行ってもCPU処理時間は約10%程度しか増加しない。また、パリティ生成処理を加えたCPU処理時間は、理想的なディスク装置への書き込み時間の8ms

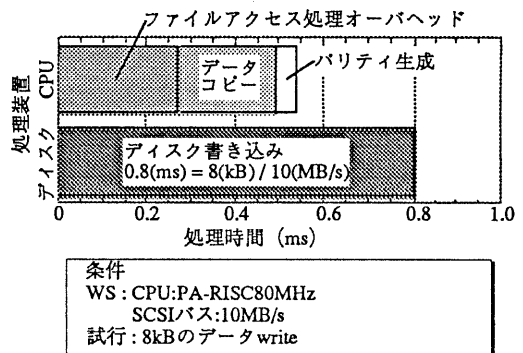


図 2.1 パリティ生成オーバーヘッド

(SCSIバスの転送速度が10MB/sの時)より短い。したがって、パリティ生成のオーバーヘッドの影響はほとんどなく、VAFS/HRをUNIX内に実装可能であると言える。

2.3 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2.2にそれぞれの比較を示す。

表2.2 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	デバイスドライバ層
1	性能	高速	中速
2	移植性	難	容易
3	開発規模	大	小

(1) 性能

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。その理由は次の二つである。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、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ーヘッドを最小にすることができる。

(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、インタフェースの変化がほとんど無いことである。

(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、性能的に有利なファイルシステム層での実装を採用する。

2.4 VAFS/HRのパリティグループ管理方式

2.4.1 RAIDの性能上の課題

— ライトペナルティの低減

RAID構成の場合には各ディスク装置の同一物理ブロック番号のブロック間でパリティ生成を行うが、図2.2に示すようにライトペナルティが発生し、書き込み性能が低下する。特にパリティグループ中の1ブロックのデータだけを書き込む場合には計4ブロックのI/Oが発生し、単体ディスク装置の1/4まで性能が低下する。

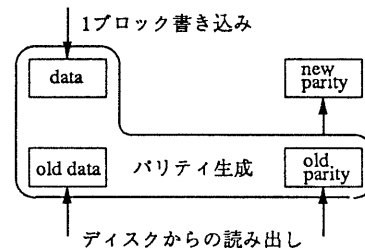


図2.2 ライトペナルティ

1ブロックだけのデータ書き込みはランダム書き込みだけでなく、図2.3に示すようにファイルが連続的にディスク装置に格納できず断片化される場合にはシーケンシャル書き込みでも発生する。ファイルの断片化はファイルの作成や削除を繰り返していくことにより簡単に生じる。すなわち、ライトペナルティの低減を検討する際には断片化されたファイルのシーケンシャル書き込みについても考慮する必要がある。

ライトペナルティの低減については仮想ストライピング方式[7]が提案されているが、ガーページ

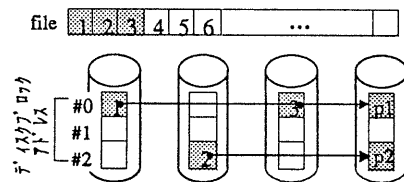


図2.3 断片化されたファイルのライトペナルティ

コレクティングを行う必要がある。そのため、性能が不安定になることや、処理が複雑になるという問題がある。

2.4.2 ファイル内パリティグループ管理方式

ライトペナルティを低減し特に断片化されたファイルのシーケンシャル書き込み性能の向上を目的として、図2.4に示すファイル内パリティグループ管理方式を提案する。ファイル内パリティグループ管理方式は、ファイルの論理ブロック単位でのパリティ生成を行う方式である。これにより、図2.4の断片化されたファイルでファイルの1ブロック目から3ブロック目まで書き込むような場合でもディスク装置からの読み出しは発生せず、1ブロックのパリティ書き込み済み済み、ライトペナルティを低減できる。

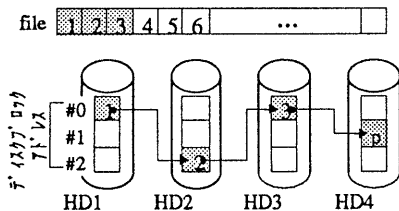


図2.4 ファイル内パリティ管理方式

3. VAFS/HRの実装

3.1 VAFS/HRのシステム構成

VAFS/HRのシステム構成を図3.1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。各モジュールの機能概要を表3.1に示す。

（1）ファイル管理モジュール

ファイル管理モジュールは、ファイル配置の管理とデータストライピング及びファイル内パリティグループ管理を行うモジュールである。データストライピングとディレクトリ構造の管理機能については、従来のVAFSの機能にパリティファイルの管理機能を追加する。ファイル内パリティグループ管理機能については新規に作成する。

（2）パリティ生成モジュール

パリティ生成モジュールは、システムバッファ上でパリティ生成処理を実行するモジュールである。また、パリティ生成の単位であるパリティグ

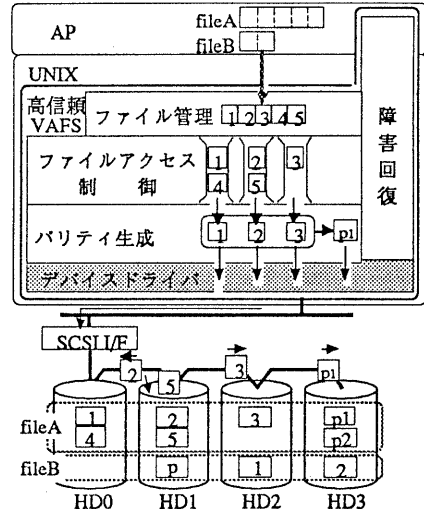


図3.1 VAFS/HRの全体構成

表3.1 VAFS/HRの機能概要

No.	モジュール	機能概要
1	ファイル管理	○ディレクトリ構造管理 ○データストライピング ◎ファイル内パリティグループ管理
2	パリティ生成	◎パリティデータ生成 ◎システムバッファ上でのパリティグループ管理
3	ファイルアクセス制御	○ユーザ指定パス名からのファイル検索 ●非同期I/O制御 ◎故障ディスク装置内データのアクセス（縮退運転）
4	障害回復	◎故障ディスク装置上のデータの回復

- ：従来VAFSの機能
- ：従来VAFSの機能に追加変更
- ◎：新規作成機能

ループとなるシステムバッファの組管理も行う。本モジュールは、新規に開発する。

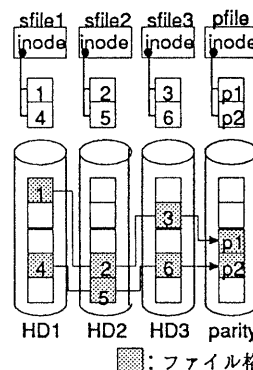
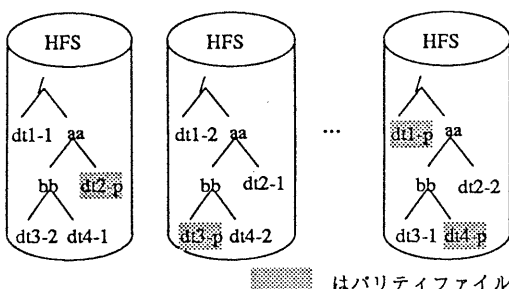
VAFS/HRではメモリコピーのオーバヘッドを最小限にするために、パリティ生成用の特別なバッファを設けずシステムバッファ上でパリティ生成を行う。パリティ生成処理は、システムバッファからディスク装置へ書き込みを行う延長で起動される。

（3）ファイルアクセス制御モジュール

ファイルアクセス制御モジュールは、ユーザが指定したパス名のファイルを検索し、そのファイルへのアクセスを行うモジュールである。ファイ

パス名をキーとしたファイル検索とファイルへのアクセスについては、ディスク装置が故障しても継続して実行できるように従来のVAFSの機能に追加変更を行う。また、非同期I/O制御については、従来のVAFSの機能を用いる。

障害回復モジュールは、ディスク装置の故障が発生した場合に、VAFS/HRを正常状態に回復するためのモジュールである。そのため、故障したディスク装置上に構築されていたデータを、交換したディスク装置上に復元するva_recoveryコマンドを新規に開発する。



に格納する。パリティ・データは各サブファイルの同一論理ブロック番号のブロック間で生成し、パリティファイルとしてHDnに格納する。これにより、各サブファイルとパリティファイルの同一論理ブロック番号のブロックがパリティグループとして管理される。

3.3 パリティ生成モジュール

パリティ生成処理の高速化のために、パリティ生成に必要なブロックの読み出し処理を多重に行う。そのために、図3.5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した。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。これにより、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。

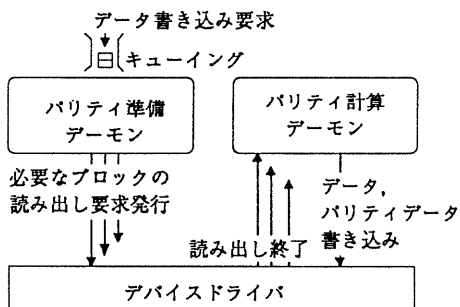


図3.5 パリティ生成

3.4 ファイルアクセス制御モジュール

3.4.1 バス名検索

図3.2で示したように、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バス名にしたがってディレクトリを検索する。

3.4.2 縮退運転

故障したディスク装置内のサブファイルのデータは、残りのサブファイルのデータを読み出し、パリティ計算を行って復元する。縮退運転時には、パリティ生成モジュールのパリティ準備デーモンとパリティ計算デーモンを用いてこれらの処理を行う。

3.4.3 非同期I/O制御

非同期I/O制御機能は、前に発行したアクセス要求の終了を待たずに、次のアクセス要求をファイルシステムからデバイスドライバへ発行して、複数のHDを並列動作させる機能である。

read()とwrite()システムコールを非同期化したシステムコールとしてasread()とaswrite()を導入する。さらに非同期システムコールを同期化するaswait()システムコールを導入する。

また、従来のread()およびwrite()システムコールを使用するAPがVAFS/HR上のファイルをアクセスする場合には、これらをカーネル内部で非同期システムコールに自動変換させる。

3.5 障害回復モジュール

va_recoveryコマンドは、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プログラムと(2)ファイルデータ復元プログラムおよび(3)ユーザ操作ミス防止プログラムの三つのプログラムから構成される。

(1) ディレクトリ構造復元プログラム

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プログラム

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ファイルデータを復元する。

(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防止する。

4. VAFS/HRの性能評価

4.1 測定条件

VAFS/HRの読み出し性能と書き込み性能を、正常運転とディスク1台が故障した縮退運転について測定する。比較のためにUNIX標準ファイルシステムのUFSと、ストライピングは行いパリティデータは格納しないVAFSの性能も測定する。表4.1に測定に用いる各ファイルシステムの構成を示す。

表4.1 測定するファイルシステムの構成

No	ファイルシステム	ディスク台数	運転モード	備考
1	VAFS/HR	5	正常	4データ+1パリティ
2	VAFS/HR	5	縮退	データディスク1台故障
3	VAFS	4	正常	4データ, パリティなし
4	UFS	1	正常	1データ

測定は表4.2に示すようにシーケンシャルおよびランダムアクセスに対して行う。シーケンシャルアクセスでは8MBのファイルの読み出しと書き込みを行い、ランダムアクセスでは8MBのファイルに対して8KB単位で2MB分の読み出しと書き込みを行う。なお、書き込みは新規ファイルに対して行う。ファイルアクセスにはシステムコールasread(),aswrite(),aswait()を用いる。

表4.2 アクセスパターン

	ファイルサイズ	アクセス総量	アクセス単位
シーケンシャル	8MB	8MB	32KB
ランダム		2MB	8KB

表4.3 測定システムの機器構成

名称	形式	台数	仕様
WS	日立 3050RX	1	CPU: PA-RISC 80MHz, 主記憶容量: 64MB, SCSI: 10MB/s
HD	HP C2244	5	直径: 3.5", 容量: 566MB, 回転数: 5400rpm

測定に用いるシステムの構成を表4.3に示す。WSには日立製WSである3050RXを使用し、ディスク装置にはHP社製のディスク装置であるHPC2244を用いた。WSとディスク装置は1本のSCSIバスでディジーチェーン接続されている。

4.2 測定結果

測定結果を図4.1, 図4.2に示す。シーケンシャルアクセスでは、VAFS/HRの読み出しで正常時に8.3MB/s (UFSの2.4倍), 縮退時に8.0MB/s (UFSの2.3倍)となった。これは正常, 縮退共にVAFSと同等である。書き込みでは正常時に5.3MB/s (UFSの1.8倍), 縮退時に6.5MB/s (UFSの2.2倍)となった。VAFSと比較すると、正常時でVAFSの約8割, 縮退時ではほぼ同等である。

ランダムアクセスでは、VAFS/HRの読み出しで正常時に1.7MB/s (UFSの1.7倍), 縮退時に1.0MB/s (UFSと同等)となった。書き込みでは正常時に0.75MB/s (UFSと同等), 縮退時に0.88MB/s (UFSの1.2倍)となった。これは、正常時の読み出しでVAFSと同等, それ以外ではVAFSの約5割の性能である。

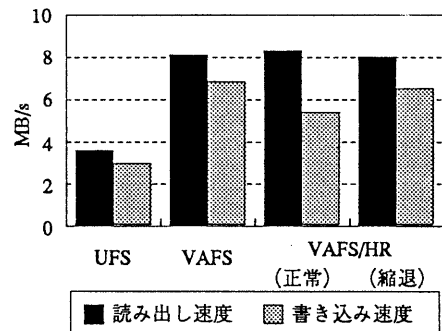


図4.1 シーケンシャルアクセス性能

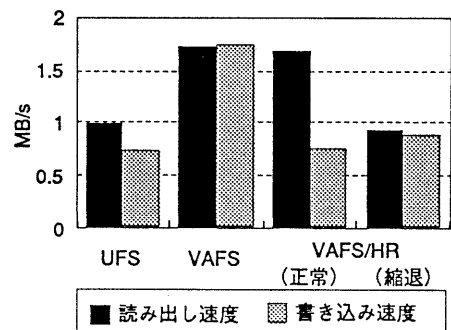


図4.2 ランダムアクセス性能

4.3 考 察

4.3.1 シーケンシャル読み出し性能

正常時の読み出し性能は、SCSIバスの転送速度(10MB/s)がボトルネックとなるため、8.3MB/sであった。また、正常時の読み出し性能がVAFSと同等なのは、正常時にはパリティデータを読まないでVAFSとVAFS/HRで処理内容に差がないためである。

縮退時に僅かな性能低下が見られるのはデータ修復のためのxor演算処理のためである。読み出しデータ量については、各 row につき正常時にはデータブロックを4つ読み出すのに対し、縮退時にはデータブロックを3つとパリティブロックを1つ読み出すため読み出しブロック数としては変わらず、性能に影響しない。

4.3.2 シーケンシャル書き込み性能

正常時の性能がVAFSの約8割になっているのは、VAFS/HRではデータブロック4ブロックにつきパリティブロックを1ブロック多く書き込むためである。縮退時の性能が正常時に比べて向上するのは故障ディスクに対する書き込みを行わない分だけ正常時よりも書き込みブロック数が減るからである。

4.3.3 ランダム読み出し性能

縮退時の性能が正常時の約半分になるのは、故障ディスク上のデータブロックを復元する度に正常な全てのディスクから読み出しが生ずることにより、読み出しブロック数がほぼ2倍に増えるからである。

4.3.4 ランダム書き込み性能

一般的なRAID4のランダム書き込みでは1ブロックの書き込み処理の度にパリティブロックの読み出しと書き込みが起こる。このため、パリティブロックを格納したディスクがボトルネックとなり、書き込み処理性能は単体ディスク性能の約半分になる。しかし、VAFS/HRでは、ファイルとして使用していないブロックは読み出さないため新規ファイルの書き込み処理では単体ディスク性能と同等の性能、すなわちUFSと同等の性能が得られる。

5. おわりに

高速UNIXファイルシステム“バーチャルアレイ・ファイルシステム(VAFS)”にパリティデータを付加したVAFS/HRの検討とプロトタイピングを行った。その結果、以下の結論を得た。

(1) 性 能

VAFS/HRのファイルアクセス性能は、最大でSCSIバス(10MB/s)の80%である8MB/sとなり、VAFS/HRの基本仕様が実現できた。

(2) 信頼性

va_recoveryコマンドの開発により、故障ディスク装置内のデータを新しいディスク装置に復元することができた。

(3) 可用性

縮退運転機能により、ディスク装置が故障しても継続してシステムを運用することができるようになった。

参 考 文 献

- [1] D.A. Patterson, et al. "Introduction to Redundant Arrays of Inexpensive Disks(RAID)", spring COMPCON'89, pp.112-117, Feb.1989
- [2] 「ディスクアレイ装置 性能向上で分散システムの要に」, 日経エレクトロニクスno.579, pp.77-103, (平5-4)
- [3] 「ハードRAIDに遜色ない性能と信頼性」, 日経オープンシステムズno.9, pp.155-161, (平5-12)
- [4] 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」, 情報処理学会全国大会講演論文集4-62, (平4-10)
- [5] 秋沢他4, 「ストライプド高速UNIXファイルシステムの開発-バーチャルアレイ・ファイルシステム(VAFS)-」, 情報処理学会OS研究会61-2, (平5-8)
- [6] 鬼頭他6, 「高速UNIXファイルシステムの構想」他4, 情報処理学会全国大会講演論文集7B 1-5, (平5-9)
- [7] 茂木他1, 「動的パリティストライプの再編成によるRAID5型ディスクアレイの高性能化手法(仮想ストライピング)に関する基本検討」, 信学技報CPSY93-32, (平5-8)

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。