7B－4

高速UNIXファイルシステムの開発における多重アクセス制御方式の実現

秋沢 充　山下 洋史*1　加藤 寛次*1　鬼頭 昭*2　牧 敏行*3　山田 秀則*3
(株)日立製作所コンピュータ事業本部　*1(株)日立製作所 中央研究所
*2 (株)日立製作所 ソフトウェア開発本部　*3 日立コンピュータエンジニアリング(株)

1．はじめに

近年，UNIXワークステーション(WS)においては急激なCPU性能向上に対しファイルアクセス性能が追従しきれていない状況にある。ファイルアクセス性能はシステム性能を決定する重要な要因であるため，様々な高速化のアプローチが行われている[1][2]。報告者らは汎用で高コスト・パフォーマンスな高速UNIXファイルシステムとして，ストライピングを用いたバーチャルアレイ・ファイルシステム(Virtual Array File System :VAFS)を提案した[3]。本稿ではVAFSの高速化を達成するために必要となる多重アクセス制御方式の実現方法について報告する。

2．多重アクセス制御方式の課題

VAFSは特殊なハードウェアを必要とせず，SCSIバスに標準接続された既存のディスク装置を高速・大容量のディスク装置として扱えるようにする高コスト・パフォーマンスなファイルシステムである。複数のディスク装置に分割格納したファイルを高速にアクセスするために，ファイルシステムにおける非同期I/O制御方式およびデバイスドライバにおけるディスク装置の多重アクセス制御方式を開発する。なお，デバイスドライバを特にバーチャルアレイ・デバイスドライバ(VADD)と呼び，ファイルを格納する複数のディスク装置をバーチャルアレイ・ディスク(VAHD)と呼ぶ。

多重アクセス制御方式はVAFSのアクセス性能の向上を図るため，ディスク装置の並列動作，SCSIバス上の高速データ転送およびSCSIバスの利用率向上を実現することを目的とする。

そのためには，READY状態のディスク装置に直ちにアクセス要求を発行できるキューイング方式，ロックの粒度を上げてオーバヘッドを最小化する排他制御方式，SCSIプロトコルのオーバヘッドを低減するアクセス要求のマージ方式および並列動作するディスク装置の並列度を高めるアクセス・スケジューリング方式の開発が課題となる。

An Implementation of Multiple Disk Access Method in the Design of Performance Improved UNIX File System
Mitsuru AKIZAWA,Hirofumi YAMASHITA*1, Kanji KATO*1, Akira KITO*2, Toshiyuki MAKI*3, Hidenori YAMADA*3
Computer Group,Hitachi,Ltd.
*1 Central Research Laboratory,Hitachi,Ltd.
*2 Software Development Center,Hitachi,Ltd.
*3 Hitachi Computer Engineering Co.,Ltd.

3．多重アクセス制御方式の実現

多重アクセス制御方式では，上記課題をVADDの内部で下記方式により解決する。

3.1　アクセス要求キューイング方式

アクセス要求をキューイングするI/Oキューが1本しかないと，各ディスク装置へのアクセス要求がシリアライズされてしまう。その結果，READY状態のディスク装置が他にあっても新たにアクセス要求を発行することができなくなる。

VAFSでは図1のようにI/Oキューおよびその管理に必要な情報をディスク装置ごとに持たせる。各I/Oキューには，そのディスク装置に対するアクセス要求だけをキューイングし，READY状態であれば常にキュー内のアクセス要求を発行できるようにする。

3.2　ディスク装置の排他制御方式

READY状態にあるディスク装置へ常にアクセス要求を発行できるようにするためには，ディスク装置単位に排他制御を行う必要がある。VAFSではVAHDを構成するディスク装置単位に排他制御を行う。これにより，アクセス要求処理中は同一のディスク装置にアクセス要求を発行せず，他のREADY状態にあるディスク装置へは全く独立にアクセス要求を発行できるようにする。

3.3　アクセス要求マージ方式

SCSIプロトコルのオーバヘッドを低減して高速化を

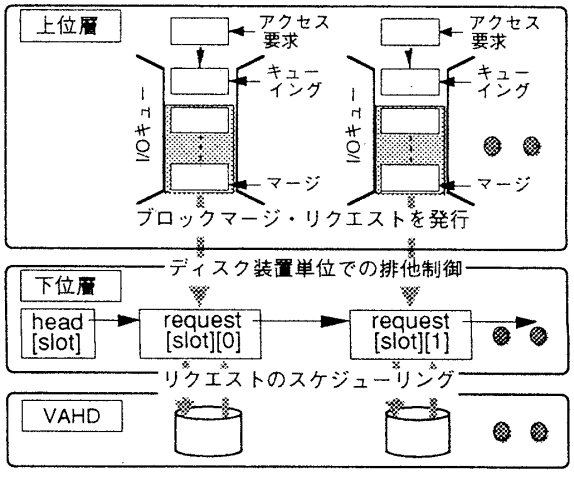

図1．VADDの構成

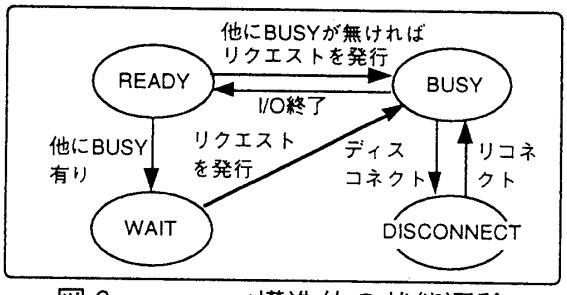

図２．request構造体の状態遷移

図るために，複数のアクセス要求を一括して１回のリクエストにまとめて発行するアクセス要求のマージ機能をVADDに導入する。

すなわち，ディスク装置へアクセス要求を発行する際に，I/Oキュー内にキューイングされている複数のアクセス要求のアクセス対象ブロックがディスク媒体上で連続する場合には，これらを一括して一つのアクセス要求として発行する。

3.4　アクセス・スケジューリング方式

SCSIバス仕様として定義されているディスコネクト／リコネクト機能を用い，VADDの下位層でアクセス・スケジューリングを行う。

上位層からアクセス要求が渡されると，ディスク装置に対応したリクエスト構造体request[slot][i]を確保して，ヘッダhead[slot]のリクエストチェインにつなぐ。リクエスト構造体は図２のように以下の４状態を遷移する。

●READY：ディスク装置がアクセス要求待ちであり，リクエスト構造体が使用可能である状態。

●BUSY：ディスク装置とホストがコネクト中でありSCSIバスを占有している状態。

●DISCONNECT：ディスク装置は動作中であるが，ホストとは切り離されバスを空け渡している状態。

●WAIT：SCSIバスが他のディスク装置に使用されているため，バスの空きを待っている状態。

リクエスト構造体をリクエストチェインにつないだ後，ヘッダにリンクされている他のすべてのリクエスト構造体の状態をチェックする。BUSY状態のものがなければこれをBUSY状態としてリクエストを発行する。BUSY状態のものがあればリクエストを発行せずにWAIT状態とする。

VADDはSCSIバスにバスフリーの時間ができると，直ちにこのリンクをたどりWAIT状態またはDISCONNECT状態のリクエスト構造体を探し，該当するディスク装置へアクセス要求を発行する。

このようなディスコネクト／リコネクト制御を行なうことにより，ディスク装置内の媒体から先読みバッファへのデータの読み出しと，先読みバッファからバッファキャッシュへのデータ転送をディスク装置間で全く独立に行うことができる。これにより，ディスク装置の動作の並列度を高めることが可能になりI/Oスループットを向上できる。また，ディスク装置からバッファキャッシュへのデータ読み出しは，ディスク装置内の先読みバッファに読み込まれたデータに対して行うことができるようになるため，媒体読み出し以上の高速転送が実現できることになる。VAHDに格納されたVAFSファイルをアクセスする動作を図３に示す。

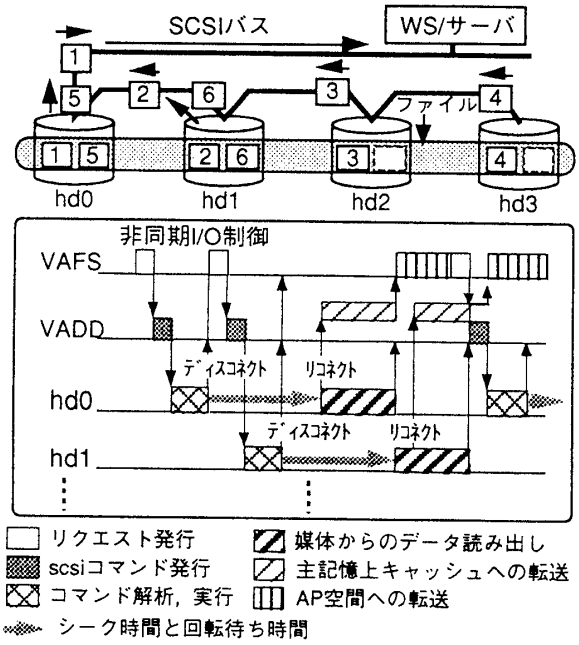

図３．VAFSのファイルアクセス動作

以上の各機能を統合して多重アクセス制御方式をVADDに実現し，そのフィージビリティを確認した。

4．おわりに

アクセス要求キューイング方式，ディスク装置の排他制御方式，アクセス要求マージ方式およびアクセス・スケジューリング方式を開発することにより，VAFSの高速アクセスを可能にする多重アクセス制御方式を実現した。

参考文献

[1]手塚,「ﾜｰｸｽﾃｰｼｮﾝ用Unixｶｰﾈﾙの高速化の試み」,情処研報，Vol.92，No.22(ｵﾍﾟﾚｰﾃｨﾝｸﾞｼｽﾃﾑ54-4)，pp.25-32，1992.3.13

[2]Andy DeBaets, et. al, "High Performance PA-RISC Snakes Motherboard I/O" , Digest of Papers, COMPCON Spring '93, pp.433-440

[3]秋沢他5,「ﾊﾞｰﾁｬﾙｱﾚｲ・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,4-62,1992.10

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。