高速UNIXファイルシステムの基本構想

7B－1

鬼頭昭　秋沢充*1　山下洋史*2　加藤寛次*2　牧敏行*3　山田秀則*3
(株)日立製作所 ソフトウェア開発本部　*1(株)日立製作所コンピュータ事業本部
*2(株)日立製作所 中央研究所　*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，プロセッサの高性能化によりワークステーションの性能が著しく向上したが，これに対応するファイルアクセス性能の向上は十分でない。このため，ファイルアクセス性能向上を目的としてディスクアレイやプログラムによるデータストライピング技術の開発が試みられている[1][2]。筆者らは，ファイルシステムレベルのデータストライピングを行うことで複数台のディスク装置にファイルを分割格納し，並列アクセスによる高速化を実現する高速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)”を提案した[3]。本稿では，VAFSの基本構想及び，WSへの実装・評価の結果について報告する。

2．ファイルアクセス高速化の技術課題

ディスク装置上にあるファイルをいかに効率よく高速に読み出してくるかという観点で，ファイルアクセス方式を考えた場合，CPU処理性能に比べて何倍も遅いディスク装置を複数台接続し1ファイルを複数ディスク装置に分割して並列に読み出すことによるアクセス高速化が有効である。並列多重読み出しを行うためには，ファイル分割(データストライピング)

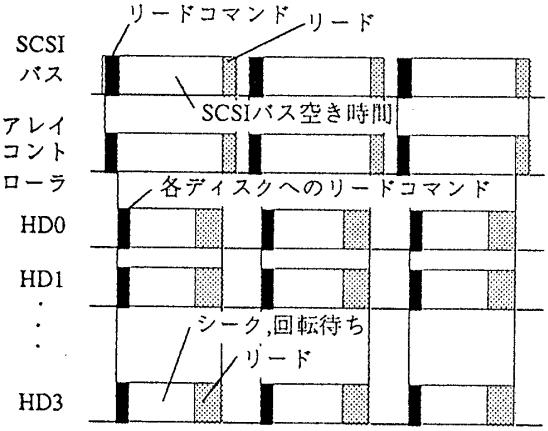

図1　SCSIバス使用状況(UFS/ディスクアレイ)

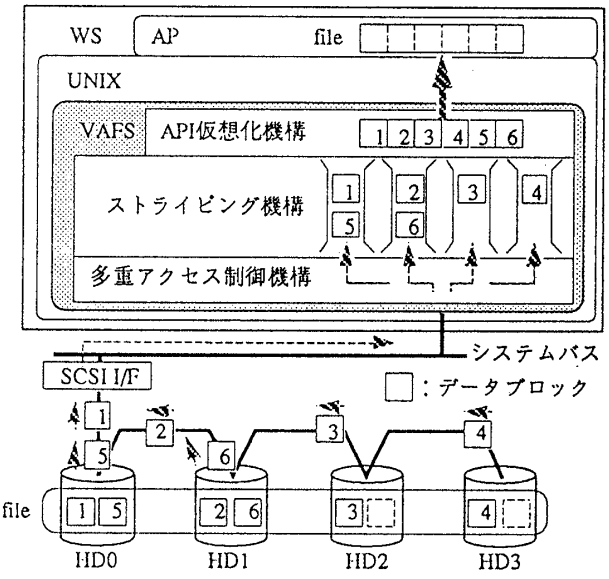

図2　バーチャルアレイ・ファイルシステムの構成

の実現が課題である。ワークステーション(WS)においてデータストライピングを行う場合の実現部位は下記に分類できる。

・ディスク装置
・デバイスドライバ
・ファイルシステム

ディスクアレイに代表されるディスク装置内でのストライピングでは，アレイを構成する複数のディスク装置が，WSからは論理的に1台のディスク装置として扱えるため，OSや上位ソフトウェアはデータストライピングを意識する必要がない。ところが，UNIXファイルシステムは，ディスク装置に対して同期でかつ8kbyte程度の小さな単位でI/Oを要求するため図1に示すようにSCSIバスの利用効率が上がらないことによりディスクアレイ本来の性能が得られない。

また，ディスクアレイ内でストライピングを独立して行うため，ディスク装置内に特別のハードウェアが必要となり，コスト面での問題もある。

これに比べて，デバイスドライバ，ファイルシステムレベルでのストライピングは複数のディスク装置をソフトウェアで制御するものであり，特別のハードウェアを必要としない。

反面，OS内の各部位の改造が必要となり，従来の

The Design Concept of Performance Improved UNIX File System
Akira KITO, Mitsuru AKIZAWA*1, Hirofumi YAMASHITA*2, Kanji KATO*2, Toshiyuki MAKI*3, Hidenori YAMADA*3
Software Development Center,Hitachi,Ltd.
*1 Computer Group,Hitachi,Ltd.
*2 Central Research Laboratory,Hitachi,Ltd.
*3 Hitachi Computer Engineering Co.,Ltd.

ファイルアクセスインタフェースを保った形でのサポートが課題となる。

3. VAFSの基本構想

図2に示すように, VAFSはストライピング機構, 多重アクセス制御機構, API仮想化機構から構成される。 ストライピング機構はファイルを分割し各ディスク装置に振り分ける機能であり, 多重アクセス制御機構はSCSIバスの使用効率を上げるために各ディスク装置に対し非同期にI/O要求を出せるようにする機能である。 また, API仮想化機構では, アプリケーションプログラムに対しデータストライピングを意識させないためにインタフェースの仮想化を行う。

VAFSでは, ストライピング機構をファイルシステムレベルで実現する。これは, ディスク装置のフォーマットをファイルシステム形式とすることで, デバイスドライバで実現する場合に比べて, 先読み, シリンダグループを使った最適配置による高速読みだし等の効果を考慮したことによる。iノードを各ディスク装置上に必要とするため, iノードサーチ時間が増えるが, 先読み, 最適配置による効果を優先した。

ファイル読み出し処理を同期で行うUNIXでは, 多重アクセスによる高速化を最大限に引き出すために, 上記各機構に加えて非同期ファイルアクセス機能の提供を行う。非同期ファイルアクセス機能の提供においては下記に示す2種類のアクセス法を検討した。

(1) 従来のUNIXファイルアクセスインタフェースをそのまま使用し, ファイルシステム内部で非同期I/Oに変換することでアプリケーションプログラムの書き換えなしにファイルアクセス性能の向上を行う。
(2) アプリケーションプログラムレベルでのファイルアクセス最適化を目的として非同期R/Wシステムコールの新規追加を行い, 1プロセスから, 複数ディスク装置に対しI/O終了を待たず次々にI/O要求を発行できるようにする。

上記各機構により従来のファイルアクセスインタフェースを保ったままで, 図3のようにSCSIバスの使用効率が上がり単位時間当たりのスループットが向上する。非同期システムコールを使用すれば, さらにきめ細かなファイルアクセス制御が行えるため性能向上が期待できる。

4. VAFSの実装と評価

VAFSの実装に際しては従来のファイルシステムとの共存を実現するため, vnodeインタフェースを用いた。これによりVAFSでマウントされたディスク装置をNFSで共有することができる。また, デバイスドライバは, 標準UNIXファイルシステムであるUFSと共通

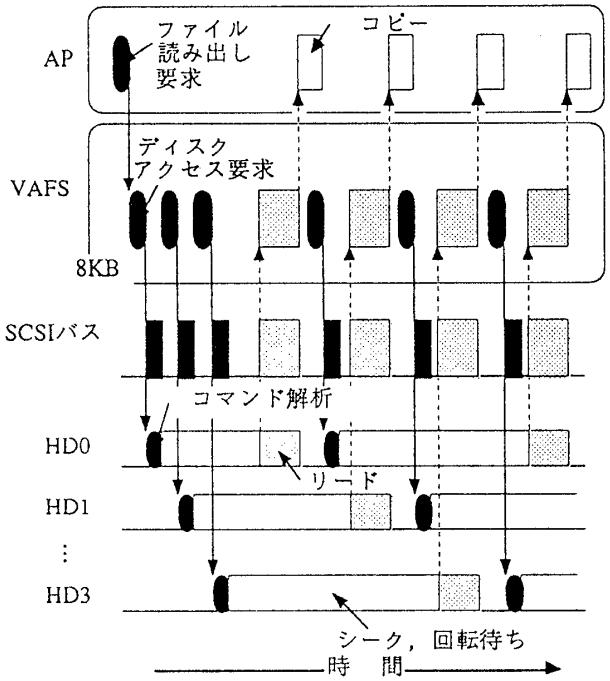

図3 多重アクセス制御方式

とし, 多重アクセスサポートのためにSCSIバスのディスコネクト・リコネクト機能を組み込む。また, ディスクブロック上で連続するI/O要求が来た場合のブロックをまとめてディスク装置に対し一括要求を出すブロックマージ機能を用いて単体のディスクアクセスの高速化を図る。 これを日立製のWS(3050RX)に実装したところ, 10MB/sのSCSIバス上に3台のディスク装置を接続した場合にシーケンシャルREAD性能8.1MB/sを得た。

5. おわりに

非同期ファイルアクセス機能, 多重アクセス機能を持つ高速UNIXファイルシステム, VAFSを提案し, これをWS上に実装した。 これにより従来のファイルクセスインタフェースを崩さずにファイルアクセス高速化が行えることを示した。今後は冗長データ追加によるVAFSの信頼性向上および, 複数I/Oバスの適用による高速化を行う予定である。

参考文献

[1]D.A.Patterson,P.Chen,G.Gibson,and R.H.Kats, "Introduction to Redundant Arrays of Inexpensive Disks (RAID)",spring COMPCON '89, pp.112-117, Feb.1989
[2]Andy Debaets他7, "High Performance PA-RISC Snake Motherboard I/O" ,spring COMPCON '93, pp.433-440
[3]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」,情報処理全国大会講演論文集4-61, 平4-年10月

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。