階層ストレージの制御アーキテクチャ

3B－4

杉本　欽一　　辻澤　隆彦
NEC　機能エレクトロニクス研究所

1 はじめに

現在小型コンピュータシステム用のディスクサブシステムとして、光ディスクあるいは磁気ディスクなどを複合した階層ストレージシステムの試作を行っている。これは、現在普及しつつあるディスクアレイシステム[2]が同じスペックのドライブの並列運転により性能と信頼性を獲得することを目的としているのに対して、階層ストレージは異なった種類のデバイスを階層状に組み合わせる事により、性能と信頼性を獲得することを目的としているものである。[3]

本報告では本階層ストレージ装置における、制御ソフトウェアの特徴、基本アルゴリズムに関して述べる。特にその制御方式の特徴である、ホストからの入出力要求を待ち行列により管理するスケジューラと、マルチチャネルＩ／Ｏ環境やシーケンシャルアクセスの高効率の管理が可能なハードウェアおよびその制御方式について述べる。

2 システムの特徴

階層ストレージシステムは、様々な形態の記憶デバイスを複合することにより、システムのトタールパフォーマンスの向上を目指したものであり、次のような特徴を持ったファイル装置を目指している。

- サポートの容易性
- システムの拡張性
- シーケンシャルデータへの対応

このような特徴を持たせるために、ハードウェア的には図１のような構成をとり、ホストコンピュータからは物理的に単体のディスク装置として使用可能である。この構成によりホストコンピュータのハードウェア／ソフトウェアアーキテクチャの変更への対応や保守が容易となる。

一方、最近注目されているビデオオンディマンドにおけるディジタル画像サーバーのようなシーケンシャル入出力性能の確保も比較的容易に実現可能である。特に、２チャネルの内部Ｉ／Ｏバス及び複数の記憶デバイスの使用に対してスケジューリングを行うことにより、最適なアクセス方式を実現できる。

本階層ストレージシステムでは、これらの特徴を持たせるために、ホストコンピュータからの入出力要求を待ち行列で管理してスケジューリングを行い、そのスケジューラ上にキャッシュ管理を実現している。

3 待ち行列Ｉ／Ｏ管理方式

本階層ストレージシステムのＩ／Ｏ管理はコントロールボード上のファームウェアとして実装されており、その管理は待ち行列を使用し効率化を図っている。

階層ストレージシステムのＩ／Ｏ管理はホストコンピュータのＩ／Ｏ要求を短時間に完結し、ホストと接続しているＩ／Ｏバスを短時間に開放することを目的としている。よって、ホストコンピュータのＩ／Ｏ要求を優先度を高く設定し、優先的にＩ／Ｏ処理を実行することにより、Ｉ／Ｏバス占有時間の短縮を行う。

このＩ／Ｏ管理アルゴリズムは次の各ステップて実行される。（図２参照）

①ホストコンピュータからの処理要求コマンドをホストインタフェースで受け、コマンドキューに保持する。
②スケージユーラがコマンドキューより処理要求を引き取る。
③引き取ったコマンドの実行に必要なバッファあるいはディスク領域をバッファ／ディスクマネージャに要求する。
④必要なデバイスが確保出来た場合はコントロールブロックと呼ぶタグがＩ／Ｏスケジューラに返される。
⑤バッファ／ディスクデータの同期が必要な場合は、所要の処理を優先度の高いコマンド実行キューに追加する。
⑥Ｉ／Ｏスケジューラは処理の重要度に応じて３つのコマンド実行キューに追加する。
⑦コマンドに対する処理の負荷が低い時に優先度の高いコマ

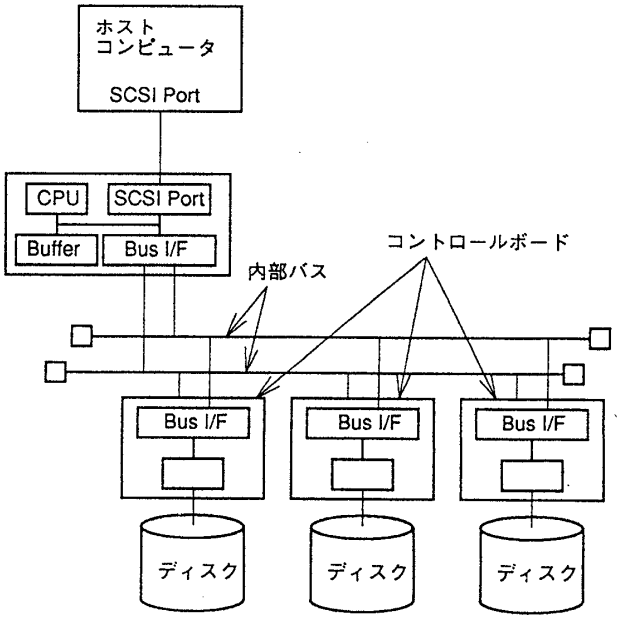

図１　階層ストレージ試験装置の概略

A Hierarchical Storage System Architecture.
Kinichi Sugimoto, Takahiko Tsujisawa
Functional Devices Research Laboratories, NEC Corporation
4-1-1 Miyzaki,Miyamae-ku,Kanagawa 216,Japan

ンド実行キューからＩ／Ｏ処理を実行する。

4 大容量キャッシュ管理機構

階層ストレージシステムでは前述の優先度の異なる複数の待ち行列によりスケジューリングを行うが、その際の各待ち行列への処理の割り振りかたにより、キャッシュ管理を行うことが可能である。基本的にはライトバックキャッシュ管理[1]を行う。現在光磁気ディスクのような比較的書き込みが低速のデバイスに対して、高速の磁気ディスクあるいはシリコンディスクをキャッシュデバイスとして用いることを検討している。

このライトバックキャッシュ機構は次のような方式で実装している。例えばホストコンピュータからの読みだし要求は①ホストコンピュータと高速のキャッシュデバイスとの間の転送と②キャッシュデバイスと低速のデバイスとの転送とに分割され、それぞれ優先度の高い待ち行列と中優先度の待ち行列にそれぞれ追加することにより、スケジュール管理がなされる。

また、キャッシュデバイスとして使用している大容量のディスクにおいてキャッシュヒットの判定と、書き込み時の不要なＩ／Ｏ処理の削除を行うために、ハッシュ法を使用したデータ検索アルゴリズムを用いている。その結果、大量のデータを使用したキャッシュ管理を実現している。

5 評価結果

現在試作中の試験装置においてその制御方式を適用した場合の効果について検討結果を述べる。ここでは磁気ディスク装置および光磁気ディスクで２階層に構成し、磁気ディスクをキャッシュデバイスとして使用した場合の性能の改善に関して予測値を示す。動作条件はディスクに対してランダムなアクセス要求が生じた場合の平均のＩ／Ｏ処理時間から、階層ストレージの処理性能の向上比を割り出したものである。図３には読みだし、書き込みにおいてキャッシュミス率 5% 及び 40% の場合の値を示す。これより、次の点が明らかとなった。

- キャッシュの高ヒット率が不可欠
- 光磁気ディスクのように書き込み処理などに時間を要するデバイスには有効（最大3.5倍）
- 小さなブロックサイズでの性能低下を押さえる為にはシリコンによるキャッシュの併用が必要

6 終わりに

市販されている磁気ディスク装置および光磁気ディスクにより２階層のストレージを構成した場合には、磁気ディスクによるキャッシュ機構に伴う性能の向上は著しいといえるものではない。しかし、光磁気ディスクの低速の書き込み動作に対しては十分有効と言える。また、大容量のファイル管理において集合型の光磁気ディスクファイル装置などに対して適用した場合には大きな効果が期待される。

しかしビデオオンデマンド[4]などの用途では大きなブロック単位での読みだしが主要となり、高いヒット率は期待できない。よって、今後これらの用途に対しては、大きなブロックサイズによる先読み管理などを併用し、Ｉ／Ｏ性能を生かした管理方式の検討が課題と思われる。

参考資料

[1]J.L.Hennessy D.A.Patterson, Computer Architecture A Quantitative Approach, Morgan Kaufmann Publishers Inc., 1990.

[2]D.A.Patterson et al., "Introduction to Redundant Arrays of Inexpensive Disks (RAID)," COMPCON '89 Spring, pp.109-117, 1989.

[3]杉本 他, "階層ストレージの試作　第１報", 信学技報,DE93-38,pp11-18,1993.

[4]石橋 他,"ビデオオンデマンドサービスのための多重特種再生技術の検討", 信学技報,CS92-74,pp101-106, 1992.

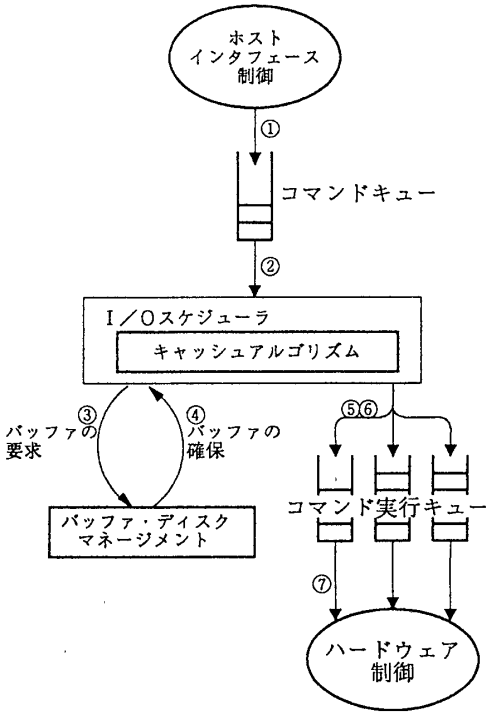

図２　階層ストレージの制御アルゴリズム模式図

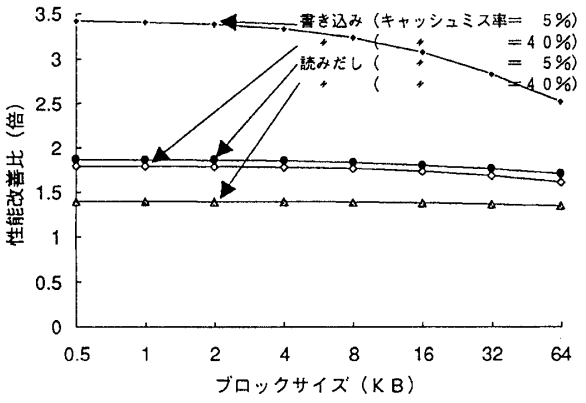

図３　性能改善比