二重化内部データバスを持つRAIDシステムの制御

6U－1

村田 浩樹　高橋 伸彰　新島 秀人　宗藤 誠治
日本アイ・ビー・エム株式会社 東京基礎研究所

1 はじめに

近年,RAIDシステムの普及が進んでいる. RAID5はSCSI-2の規格にあるコマンド並列処理(Tagged Command queueing)の機能によりRAID3と同等程度の高い転送速度を実現する事が可能となるが[2], 一つの書き込み要求に対して4回のI/Oが必要となるため, 書き込み時のオーバーヘッドが大きく, 高速なライトアクセスを要求するようなアプリケーションには適さない. 現在, 書き込み時間が実効的に見えなくなるようにディスクキャッシュやライトアシストディスクを搭載することが検討され, 実用化されている. また, ディスク本体への書き込み速度の向上も検討されている[3]. 本稿では, 文献[3]において述べている二重化内部データバスを持つRAIDシステムの制御ソフトウェアの構成と, キャッシュ・アルゴリズムについて述べる.

2 従来技術

普及の進んでいる小型RAIDシステムには, ハードディスクを複数実装したシステムに導入するソフトウェア製品,ATAドライブをシングルバスで接続した製品, およびSCSIドライブを独立のコントローラに接続した製品などが出荷されている. ソフトウェアによる製品は, パリティ生成やデータ書き込み時の4回のデータ転送のためにホストのリソースを使用するため, サーバへの導入にはパフォーマンスに対する考慮が必要である. これに対し, ハードウェアで実現されたRAID製品群は高価とはなるが, ホストに影響を与えない. ハードウェア製品の中でもATAドライブをシングルバスに接続した製品は, ハードウェア量が比較的少なく, ハードウェアRAID製品群の中では最も安価な部類に入る. しかし, データ書き込みに際して4回のI/Oを順々に行なわなければならず,RAIDとしての性能はソフトウェアRAIDと殆んど変らない. SCSIドライブを用いたRAID製品は比較的高価となるが, 基本的には内部データバスをドライブ台数分持つため全てのドライブの並列動作が可能であり, データの書き込み動作を, ‘旧データ, 旧パリティの読みだし’と‘新データ, 新パリティの書き込み’の2ステップで完了することができる.

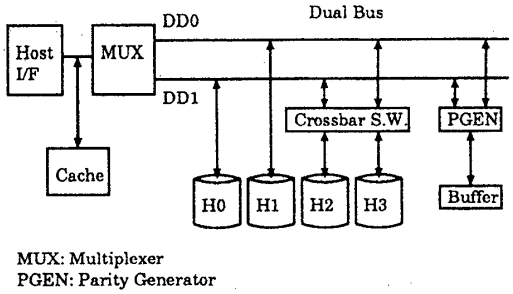

図1: 二重化内部データバスを持つRAIDシステム

3 二重化内部データバス構成

ATAドライブを用いたシングル・データバス構成の安価さと, SCSIドライブを用いた構成の高速さを合わせ持つシステムとして, 文献[3]において我々は二重化内部データバスを持つRAIDシステムを提案した. 構成を図1に示す. RAID5システムでは書き込み動作中, 一時期にデータ転送を行なっているドライブは高々二台であり, データバスが少なくとも二本あれば論理的な最小時間でデータの書き込みを完了できるという事実を用いている.

4 制御プログラムの概要

図2に制御プログラムの構成の概要を示す. 制御プログラムは並行プログラミングの手法を用いてμITRON上に構築されている. SCSI Interrupt Handlerはホストからのコマンド受取り, メッセージの送受, データの送受をおこなう.Command Queue taskはTagged Commandを含むコマンドをqueueする.SCSI Command Handlerはコマンドのチェックと, ハードディスクもしくはキャッシュへの操作を含まないコマンドの処理をおこなう. Cache Managerはキャッシュの管理をおこなう. 構成については次節にて述べる. Read/Write DistributerはRAID5の方式に従って,Read/Writeデータの各ハードディスクへの割り振りを行なう. Supervisor TaskはCode Pageや,Data Reconstruction Task,Cache Flush Taskを制

Control for RAID System with Dual Data Bus Architecture
Hiroki Murata, Nobuaki Takahashi, Hideto Nijima, Seiji Munetoh
IBM Research, Tokyo Research Laboratory

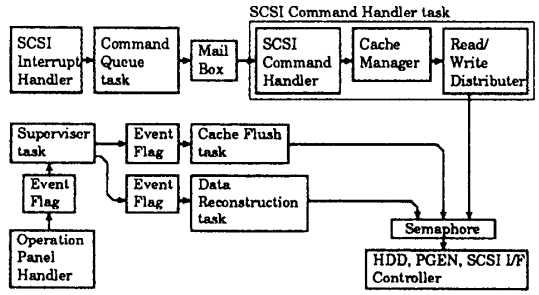

図 2: 制御プログラムの概要

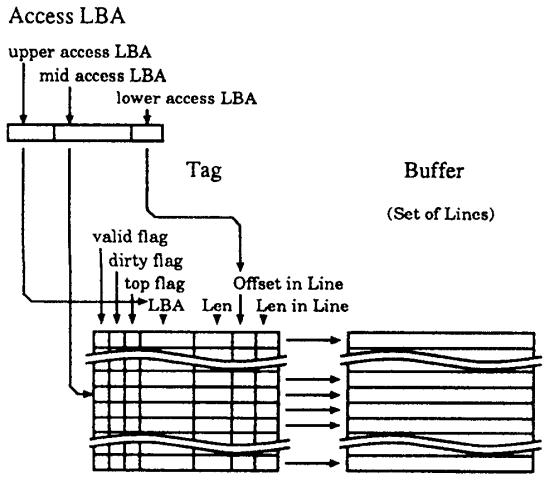

図 3: アクセス LBA と Tag,Buffer の構成

御する. Data Reconstruction Task はデータの再構成を通常のアクセスの裏側で行なう.Cache Flush Task は, キャッシュ内のダーティ・データを, 通常のアクセスの裏側で対応するハードディスクに書き込み, クリーン・データとする. Operation Panel Handler はオペレーション・パネルを制御する. HDD, PGEN, SCSI I/F Controller はハードディスク, パリティ・ジェネレータ, SCSI インターフェイスを制御し, 実際のデータ転送を制御する.

5 キャッシュの構成

キャッシュを構成するバッファの利用効率を向上し, リードもしくはライトされてキャッシュされたデータの読みだしを高速化するため, キャッシュを構成するバッファを, 数セクタから十数セクタ程度のラインで分割して管理する. 各ラインにアドレスを付け, アクセス・データのアドレスと関連付けて, アクセス・データのアドレスによって, どのラインにキャッシュするかを決める (図 3).

各々のラインに対してタグを 1 本づつ用意する. タグは, 対応するラインのデータが有効なものであるか, 無効なものであるかを示す valid フラグ, ライン内のデータが HDD にあるものと同じか, ライトによってライン内でのみ更新されたデータかを示す dirty フラグ. ライン内にデータの先頭があるかどうかを示す Top フラグ, ラインにキャッシュされているデータの先頭 LBA の上位ビットを示す LBA, 長さを示す Len, ライン内でのデータの先頭位置を示す Offset in Line と, ライン内の有効なデータの長さを示す Len in Line からなる.

また, 複数の Line にまたがるデータを一括して扱うためにタグの中の Len に二重の意味を持たせている. データの先頭を含む Line に対応するタグの Len は, そのデータの長さを示すが, それ以外のタグの Len は, データの先頭のバッファ内での絶対位置を示す. これによって, 複数のラインにまたがるデータの中間辺りのデータがアクセスされても, 有効なデータの範囲が 2 回のタグ・アクセスでわかる. 新たなデータがキャッシュされる場合にも,LBA,Len をはじめとするタグの内容は, データの先頭に対応するタグ以外すべて同じなので, 複数のタグの更新は単に同じデータを書き込むだけで済む. このように,Len に二重の意味を持たせることで, 複数の Line にまたがるデータを一括して扱うことが, 実用的な速度で可能となる.

6 おわりに

二重化内部データバスを持つ RAID システムの制御プログラムの構成と, キャッシュの構成の概要について述べた. 現在, キャッシュ以外の部分についてはほぼ実装が終了している. 今後はキャッシュを実装し性能の評価を行なう.

参考文献

[1] D.A.Patterson, G.Gibson, R.H.Katz: A Case for Redundant Arrays of Inexpensive Disks (RAID), Report no. UCB/CSD 87/391, Computer Science Div. University of California, Berkeley, (1987).

[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, 1993 年 4 月 26 日号, no.579, pp.78-103, (1993).

[3] 新島秀人, 宗藤誠治, 村田浩樹, 高橋伸彰：二重化内部データバスを持つ RAID システム, 第 49 回情報処理学会全国大会論文集, 7K-3, (1994).