C-028

ATAディスク適用ストレージ向け高信頼化技術の開発(2)
～データインテグリティ向上技術～
Development of High-Reliable Controls for ATA-based Storage System (2)
~ Improvement of Data Integrity ~

中川 豊† 新井 政弘† 松並 直人† 八木沢 育哉‡
Yutaka Nakagawa Masahiro Arai Naoto Matsunami Ikuya Yagisawa

1. はじめに

近年、エンタープライズ用途のストレージとして、価格／性能比と信頼性に優れる RAID[1]型ストレージが一般的に用いられている。非ミッションクリティカルデータの増加や、証拠性情報の長期間保管義務化を背景に、長期アーカイブなどのコスト重視の用途で、高価なサーバ向け FC ディスクに代わり、安価で大容量な PC 向け ATA ディスクを活用しようとする動きが広まりつつある。

本稿では、このような用途に適用するストレージのデータインテグリティに関する高信頼化技術に関し、文献[2]に示されるストレージ要件に基づき、活用上の課題について述べる。そして、課題を解決する信頼性向上方式として、高速ライト&コンペア方式および、ATA 対応データ保証コード方式について提案する。

2. 目的

文献[2]に示されるように、FC ディスクと比べてエンタープライズ用途における使用実績の短い ATA ディスクをストレージに用いる際、信頼性への配慮が特に重要となる。データの信頼性の観点で、ストレージには以下の 4 点が必要となる：

(1) 記録：正しく記録できたことを保証すること
(2) 読出：読み出したデータが誤っていないかを検査できること
(3) 復元：データが誤っていた際、正しいデータを復元できること
(4) 予防：復元を妨げる障害を予防保守すること

本稿では、データインテグリティに関する(1)記録、(2)読出、を実現する高信頼化技術について説明する。

3. 課題

3.1 記録（ライト保証）

PC 向けの ATA ディスクはサーバ向けの FC ディスクに比べ Non-Recoverable エラーレートが高く、セクタ障害が発生する可能性が相対的に高くなる。そのため、実績の少ない ATA ディスクを使いこなすためには、RAID 型ストレージにおいても、書き込み不良に起因するリード時のエラーを未然に防ぎ、二重障害を予防することが信頼性確保のために重要である。ライト保証の確実な方法の 1 つは、ストレージコントローラがライト後にデータを読み返し、比較チェックを行うことである。しかしながら、単純にこのプロセスを実行すると、ライト、リードの 2 回の I/O アクセスが発生するため、性能が著しく低下してしまう。このため、性能低下の防止が課題となる。

なお、長期アーカイブ用途では、ストレージへのライト処理はシーケンシャルアクセスとなる場合が多い。本稿では、ライト処理に関しては、シーケンシャルアクセスを中心に説明することにする。

3.2 読出（リード保証）

データを正しくライトしても、長期保管中の突発的・後発的な事象により、データを正しく読み出せない可能性が考えられる。このため、リード時にデータの正当性を確認することが重要である。ディスク装置は、セクタのデータに対して ECC コードを付加しているが、バースト誤り等の障害モードによっては検出能力を超えてしまうことが考えられる。

そのため、一部のストレージでは、エラー検出率を高めるために、ECC コードとは別の方法で、データ誤りを検出可能なデータ保証コードをストレージの機能により付加している。特に ATA ディスクでは、FC ディスクに比べて Non-Recoverable エラーレートが高いため、データ保証コードの付加が一層重要である。

このデータ保証コード方式は、ストレージコントローラが 512 バイトのホストデータに、例えば 8 バイトの保証コードを付加してディスクに記録する方式である。なお、「データ＋保証コード」で構成したブロックのことを、以下本稿では「論理ブロック」と呼ぶことにする。FC ディスクの場合は、セクタサイズを変更できるので、「512 バイト+8 バイト」の 520 バイトのセクタを構築し、論理ブロックを格納できる。しかし、ATA ディスクのセクタサイズは、512 バイトに固定され変更できないため、ATA ディスク専用の保証コード付加方式の開発が必要となる。

4. 方式の提案

4.1 高速ライト&コンペア方式

性能低下を防止しつつライトを保証する方式として、高速ライト&コンペア方式を提案する。

はじめに本方式の着眼点について述べる。

ATA ディスクの利用が見込まれる長期アーカイブ用途で

† （株）日立製作所 システム開発研究所
Systems Development Laboratory, Hitachi, Ltd.
‡ （株）日立製作所 RAIDシステム事業部
Disk Array Systems Division, Hitachi, Ltd.

RAID: Redundant Arrays of Inexpensive Disks
FC: Fibre Channel
ATA: Advanced Technology Attachment
ECC: Error Correcting Code

はシーケンシャル動作が主流となる。シーケンシャルライト動作ではディスク上でのアドレスが連続となるような複数のコマンドが発行されるが、ATA ディスクは FC ディスクと異なり、多重コマンド処理ができず逐次的なコマンド処理となる。そのため、コマンドとコマンドとの間で毎回ディスクの回転待ちが発生する。

そこで、ディスクのライトキャッシュを有効化することを考える。この場合、ディスク装置が複数のコマンドデータをディスクキャッシュで受領し、メディアへの書込みをまとめて一回の処理で行うため、各コマンド間の回転待ちが生じず、大幅な高速化が可能となる。本方式はコンペアのためのリードのオーバヘッド増加を、上記の回転待ち削減によるライト処理高速化により相殺する。

しかし、ディスクキャッシュを有効化した場合、停電など予期せぬ電力切断によってデータを消失する新たな課題が生じる。従来の FC ディスクを採用したストレージでは、一般に多重コマンドによる高速化を図ることができることに加え、上記の問題対処へのデメリットから積極的にディスクキャッシュを利用することは少なかった。

本方式では、この課題を解決するため以下のような制御を行う。図 1 に示すように、ストレージコントローラは(1)ディスクへいくつかのライトコマンドを発行すると、(2)ディスクキャッシュ上のデータをメディアへ書き込むよう指示する。(3)書込みが完了した後、当該データをリードし、(4)ストレージコントローラ上でオリジナルのデータと比較を行う。比較によって記録データが正しいと判断された場合、オリジナルデータをコントローラのキャッシュから解放する。

以上の方式によって、電源切断によるデータ消失の危険を回避し、通常の ATA ディスクのライト性能と遜色のない性能を維持しながら、データのライト保証を実現する。

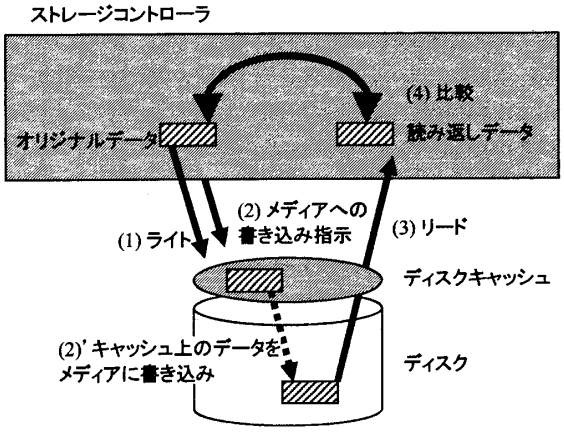

図 1　高速ライト＆コンペア方式

4.2　ATA 対応データ保証コード方式

ディスクのセクタサイズが 512 バイトに固定される ATA ディスクに保証コードを付加する方式として、ATA 対応データ保証コード方式を提案する。

ストレージに搭載するディスクが ATA の場合、データに保証コードを付加すると、ストレージコントローラ上の論理ブロック（データ＋保証コード）のサイズが ATA ディスクのセクタサイズを超えてしまう。そこで、論理ブロックを分割して複数のセクタに格納することを考える。さらに、対応のさせ方は複数考えられるが、記憶領域の利用効率を上げるために、分割した複数の論理ブロックのデータを一つのセクタ内に混在格納させる。ある論理ブロックを更新する際には、その論理ブロックを格納する複数のセクタのデータをいったんリードし、ストレージコントローラ上で当該論理ブロック部分のみ更新した上でディスクに書き戻す、いわゆるリードモディファイライト処理が発生する。このため、このリード処理に伴って性能が低下するという新たな課題が発生する。

そこで、図 2 に示すように、論理ブロックとセクタの対応関係をストレージコントローラ上の管理テーブルで管理することにより、ATA ディスクにおいてデータと保証コードの両者をディスクに書き込む。

これにより、論理ブロックとセクタとの対応関係情報を参照して、ディスク I/O 数が最小になるパターンのアクセスを見出し、実行する。例えば、同一セクタを共有し断片を格納する複数の論理ブロックのリード I/O をまとめるなどの処理を行う。

以上により、性能低下の影響を最小としつつ、リードデータの正当性を保証する。

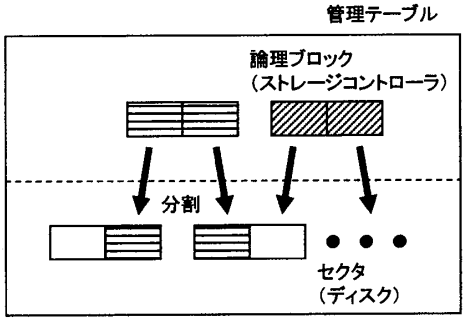

図 2　ATA 対応データ保証コード方式

5. まとめ

ストレージシステムにおいて、PC 向けの ATA ディスクを使いこなすために、データインテグリティを向上し、かつ性能との両立を可能とする手法として、高速ライト＆コンペア方式および ATA 対応データ保証コード方式を提案した。

参考文献

[1] David A. Patterson, et al.: "A Case for Redundant Arrays of Inexpensive Disks (RAID) ", Report no. UCB / CSD 87 / 391, Computer Science Division Department of Electrical Engineering and Computer Science, University of California, Berkeley, 1987.

[2] 新井政弘 他: " ATA ディスク適用ストレージ向け高信頼化技術の開発(1) ～ ATA ディスク高信頼化対策 ～", FIT2005 第 4 回情報科学技術フォーラム講演予稿集