ディスクアレイシステムにおけるSSTF方式の適用および評価

2N－8

山本康友　山本　彰
(株)日立製作所システム開発研究所

1．はじめに

ディスクアレイでは，ホストからのデータ更新に対する，冗長データの作成およびライトによるディスクアクセス増が性能悪化の原因となる。

これを回避する技術として，ホストとディスク装置の間にキャッシュメモリを設け，ディスク装置へのアクセスデータを格納しておき，これをもってホストからの入出力要求に応答する方式が広く適用されている。この技術では，ホストからのデータ更新に対してキャッシュメモリに更新データを格納した時点でライトの完了を報告し，冗長データの作成および，該更新データ／冗長データのディスク装置へのライトを，ホストアクセスとは非同期に実行する。この処理方式をライトアフタと呼ぶ。

ライトアフタの採用で，ホストへの応答時間の悪化は回避できる。しかし，ライトアフタ処理によるディスクアクセス自体はなくならないため，ディスク装置の利用率は改善されない。

一方，ディスクアレイシステムにおける大容量化，低価格化を目指した大容量ディスク装置の採用により，ディスク利用率は今後ますます増加する傾向にある。

この状況をふまえ，ライトアフタ処理における平均シーク時間削減を目的として，更新データのスケジュールにSSTF(Shortest Seek Time First)アルゴリズムを適用する。本講演では，SSTFの実装方式と性能解析結果について述べる。

2．ディスクアレイシステムの構成

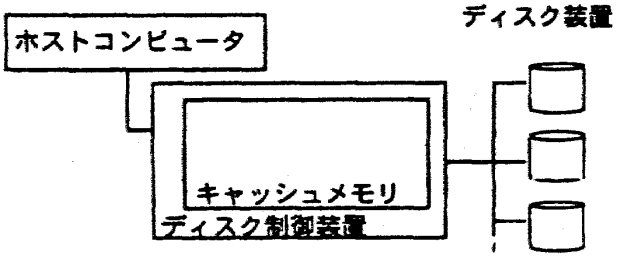

図1　ディスクアレイシステム構成

SSTFを実装したディスクアレイシステムを図1に示す。本システムでは，ホストからの入出力処理を，必ずキャッシュメモリを介して実行し，更新データをライトアフタ処理によってディスク装置へ反映する。

RAID5のディスクアレイでは冗長データはパリティと呼ばれ，対応するデータの排他的論理和で求められる。パリティは，対応する全てのデータ値からだけでなく，更新されたデータの更新値と該データとパリティの更新前値からも生成可能である。ランダムアクセスのパリティ生成は，後者の方法で行われることが多い。

従って，更新データに対するライトアフタ処理では，次の一連のディスクアクセスが発生する。

・データおよび冗長データの更新前値リード
・　〃　の更新値ライト

3．SSTFとその実装

3．1　SSTFについて

SSTFは，ディスク装置に対する入出力要求のスケジュールアルゴリズムの一つで，複数の入出力要求のうち，現ヘッド位置からのシーク距離が最短となるものを処理対象として，平均シーク時間を削減することを目的とする。

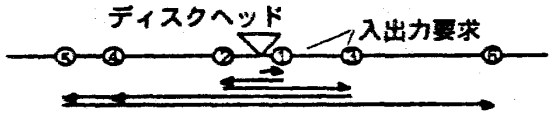

図2　SSTFの概念

SSTFには次に挙げる二つの問題がある。

(P1)　処理完了時間が保証できない
(P2)　スケジュールの度に，現ヘッド位置から最短シークの入出力要求を求めるオーバヘッドが大きい

3．2　SSTFの実装

各更新データは，ホストからキャッシュメモリに格納された時点でライトの完了をホストに報告するため，以後のライトアフタ処理完了時にホスト報告の必要はない。よって，各更新データに対するライトアフタ処理は処理完了までの時間に制約を受けない。したがって，更新データのライトアフタ処理ス

Application and Evaluation of SSTF for Disk Array Systems
Yasutomo YAMAMOTO,Akira YAMAMOTO
System Development Laboratory,Hitachi,Ltd.

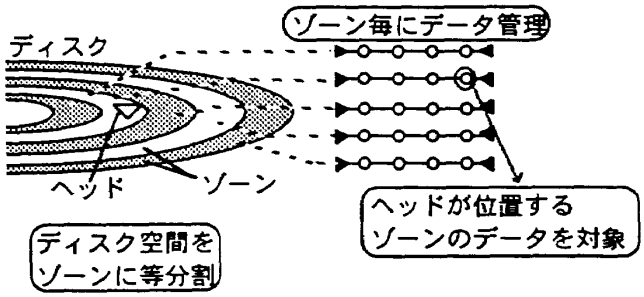

図３　SSTF実装方式

ケジュールにのみ，SSTFを適用する場合には(P1)は問題とならない。

問題(P2)のスケジュールオーバヘッドを解消し，SSTFを疑似的に実装する方式を図３に示す。

まず，ディスク装置の記憶空間をいくつかの小領域（以下，ゾーン）に分割する。そして，各更新データおよび冗長データを，各々が属するディスク装置およびゾーン別にグループ管理する。スケジュール時には，現在ヘッドが属するゾーンのデータグループから任意の一つを処理対象とすることで，ある程度シーク距離の短い更新データを少ないオーバヘッドで選択できる。

４．実装方式のシーク時間削減効果

SSTF実装方式について，平均シーク時間削減効果をシミュレーションプログラムを用いて見積もった。シミュレーションの前提条件となるディスク装置の仕様は以下の通りである。

平均シーク時間	11.5ms
シリンダ数	4553
セクタ数/CYL	(外周)132　(内周)107
ヘッド数	28

シミュレーションでは，ゾーン分割数が16，32，128のSSTF実装方式について，一台のディスク装置に対してキャッシュメモリに保持している更新データ数を変化させ，平均シーク時間の変化を調べた。同時に，オーバヘッドを無視してシーク最短の更新データを検索する本来のSSTF（真性SSTF）についても調べた。シミュレーション結果を図４に示す。

更新データが十分多い場合，SSTF実装方式の平均シーク時間は「一つのゾーン内での平均シーク時間」に収束すると予測できる。実際，図４によると，各ゾーン分割数のSSTF実装方式について，対象データ数の増加に従い，点線で示した「ゾーン内平均シーク時間」に収束する傾向が見られる。

また，各ディスク装置への更新データが少量（数

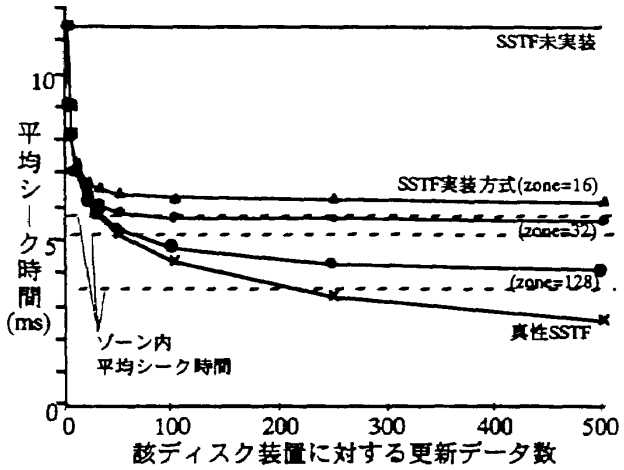

図４　シーク時間削減効果

個）でもSSTFスケジュールによる削減効果は認められ，ゾーン分割数の２倍程度の更新データがキャッシュメモリに保持されていれば，ほぼ収束値程度の削減効果が得られることがわかる。

この結果より，SSTFを実装して，目標性能を得るために必要なキャッシュメモリの容量を決定できる。例えば，ディスク装置100台を接続するディスクアレイシステムのキャッシュ管理をデータ保持単位４KB，更新データのキャッシュ占有率50%とすると，ライトアフタ処理の平均シーク時間を4.5msにするには，ゾーン分割数を128，実装キャッシュメモリを206MB以上にすればよい。

５．おわりに

ライトアフタ処理における平均シーク時間の削減，および実装時のオーバヘッド増加抑制が可能なSSTF実装方式について述べた。

また，実装時に平均シーク時間を目標値まで削減するのに必要なゾーン分割数およびキャッシュメモリ容量の見積もり方法も示した。

参考文献

(1) D.A.Patterson et.al,"A Case for Redundant Arrays of Inexpensive Disks(RAID)", ACM SIGMOD conference proceedings, Chicago, IL., pp.103-116, June 1-3 1988.
(2) A.S.TANENBAUM, "MINIX Operating Systems : Design and Implementation," Prentice-Hall,Inc.