データベースシステム 113－23
（１９９７．７．１５）

トランザクション処理環境におけるアクセストレースを用いたHot Mirroringの性能解析

茂木 和彦 喜連川 優

東京大学 生産技術研究所
〒106 東京都港区六本木 7-22-1

あらまし

２次記憶装置の高性能化・高信頼化を目的とした冗長情報を記録するディスクアレイ (RAID) の開発が進められている。サイズは小さいが多数のアクセス要求があるような負荷に適しているRAID5やミラーにおいては、それぞれ性能やデータ容量に関する欠点を保持している。RAID5とミラーの双方の利点を有効に利用するために「hot mirroring」と名付けた記憶管理法を提案した。本手法について、より現実的な負荷での性能評価を行うため、TPC-Cベンチマークを基にしたトランザクション処理環境を構築し、ディスクアクセスのトレースを採取した。本稿では、この処理環境におけるディスクアクセスの特徴を簡単に述べ、その後、このアクセス負荷に対するhot mirroringを用いたディスクアレイの性能評価の結果を述べる。

キーワード： TPC-C, ディスクアクセストレース, ディスクアレイ, RAID5, ミラー, 2次記憶装置

Performance Analysis of Hot Mirroring Using Disk Access Traces in a Transaction Processing Environment

Kazuhiko MOGI Masaru KITSUREGAWA

Institute of Industrial Science, University of Tokyo
7-22-1, Roppongi, Minato-ku, Tokyo 106, JAPAN.

Abstract

Recently RAID has attracted strong attention as a high performance and high reliable secondary storage system. For the loads which consist of a large number of small accesses, RAID5 disk arrays and mirrored disk arrays are the best suited among several RAID levels. The major drawback of RAID5 disk arrays is, however, in the large overhead incurred for small writes and the significant performance degradation on disk failure. Mirrored disk arrays have considerably smaller storage capacity than that of RAID5 disk arrays. In order to get not only higher performance but also larger capacity, we propose yet another storage scheme named "hot mirroring". For the performance evaluation of hot mirroring in a realistic environment, we collected disk access traces in a transaction processing environment based on TPC-C. In this paper, we describe the characteristics of these traces and analyze the performance of hot mirroring using these traces.

Key Words : TPC-C, Disk access trace, Disk array, RAID5, Mirror, Secondary storage

1 はじめに

2次記憶装置の高性能化・高信頼化のため、冗長情報を記録するディスクアレイ (RAID[1]) の開発が進められている。その中で、サイズは小さいが多数のアクセス要求があるような負荷に対してはミラーや RAID5 が良いと考えられている。RAID5 ではパリティを用いた冗長化を行っており、データ書き込み時のパリティ更新ためのオーバヘッドやディスク故障時のデータ復旧作業の影響による性能の低下が問題となっている。この点に関して優れているミラーでは、データのコピーを保持することによる冗長化を行っており、データ容量が少ないという問題点が存在する。これらの問題を解決するために「Hot mirroring」と名付けた記憶管理手法を提案した [2]。本方式は、アクセス頻度が高いもの (ホットブロック) をミラー、アクセス頻度が低いもの (コールドブロック) を RAID5 と、参照局所性を利用した階層構成により高性能性と高記憶効率性の両立を目指すものである。本手法について、より現実的な負荷での性能評価を行うため、TPC-C ベンチマーク [3] を基にしたトランザクション処理環境を構築し、そこでのディスクアクセスのトレースを採取し、それを用いた性能評価を行った。本稿では、この処理環境におけるディスクアクセスの特徴を簡単に述べ、その後、このアクセス負荷に対する Hot mirroring を用いたディスクアレイの性能評価の結果を述べる。

2 トランザクション処理環境におけるディスクアクセスの特徴

2.1 トランザクション処理環境

RAID の利用先として重要なものの 1 つにトランザクション処理システムがある。トランザクション処理システムにおけるより現実に近いディスクのアクセス負荷を得るために、Sybase SQL Server™ 10 を用いて SPARCstation 20/502 (Solaris2.3) 上に TPC-C ベンチマークを基にしたトランザクション処理環境を構築し、データ領域に対するディスクアクセスのトレースを修正を加えた sd ドライバを用いて収集した。

本処理環境におけるデータベースのテーブル間の参照関係を図 1 に、また、各テーブルの構成を表 1 に示す。9 つのテーブルと 2 つのノンクラスタード索引[1] をそれぞれ分離・独立した領域に記録する。データベースの規模は 14 ウエアハウスとし、テーブルと索引の記憶のために総計 11,420MB ディスク容量を割り当てた。システム上のデータ用バッファキャッシュは 56MB とした。今回は、データベースサーバ上の未書き込みデータの書き込み処理 (以下、チェックポイント[2] と呼ぶ) を実行した場合と、実行しない場合の 2 種類の条件についてアクセストレースを採取した。そのときの Order テーブルの初期タプル数を表に示す[3]。

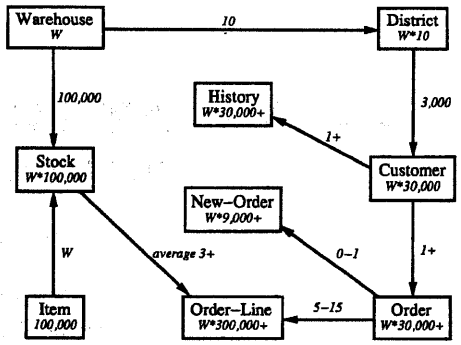


図 1: テーブル間の参照関係

	Order テーブル 初期タプル数
チェックポイント未実行	824101.9 × 14
チェックポイント実行	844402.9 × 14

表 2: Order テーブル 初期タプル数

トランザクションの処理内容は TPC-C と同様に以下の 5 種類を実行する。

New-Order 新規の商品注文処理を行う。

Payment 顧客からの支払いの処理を行う。

Order-Status 顧客の最新の注文に関する内容や配送状態を調べる。

Delivery 倉庫毎にその配下にある地区販売店毎に最も古い未配送状態にある注文の配送処理を行う。

Stock-Level ある地区販売店における最新 20 個の注文に含まれる商品品目について、それらの在庫量の確認処理を行う。

また、これらの実行比率とアクセスを行うテーブルを表 3 に示す。

処理名	比率	アクセステーブル
New-Order	10/23	W(r), D(rw), C(r), I(r), S(rw), O(w), OL(w), NO(w)
Payment	10/23	W(rw), D(rw), C(rw), H(w)
Order-Status	1/23	C(r), O(r), OL(r)
Delivery	1/23	C(rw), O(rw), OL(rw), NO(rw)
Stock-Level	1/23	D(r), S(r), OL(r)

表 3: トランザクションの実行比率

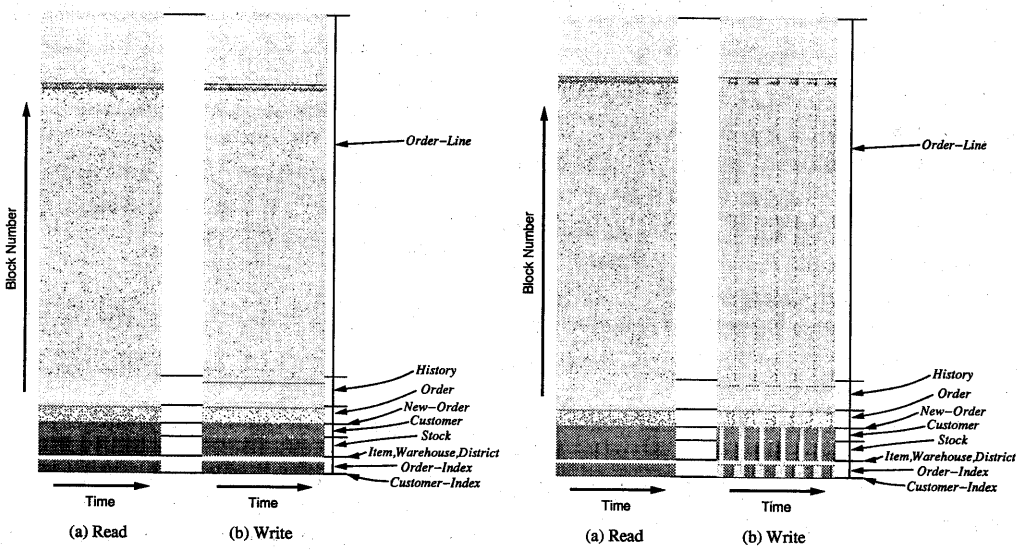
[1] Sybase SQL Server 10 においては、クラスタード索引とノンクラスタード索引の 2 種類の索引が存在する。クラスタード索引は、タプルの物理的な順序を利用した索引であり、常にデータと同じ領域内に記録される。一方のノンクラスタード索引はデータと異なる領域に記録可能である。

[2] Sybase SQL Server 10 における呼称に従う。実行頻度は「回復間隔」パラメータで決定され、実行する場合には 12 分に設定した。

[3] History tbl. の初期タプル数は Order tbl. のそれとほぼ等しく、Order-Line tbl. の初期タプル数は Order tbl. の約 10 倍である。

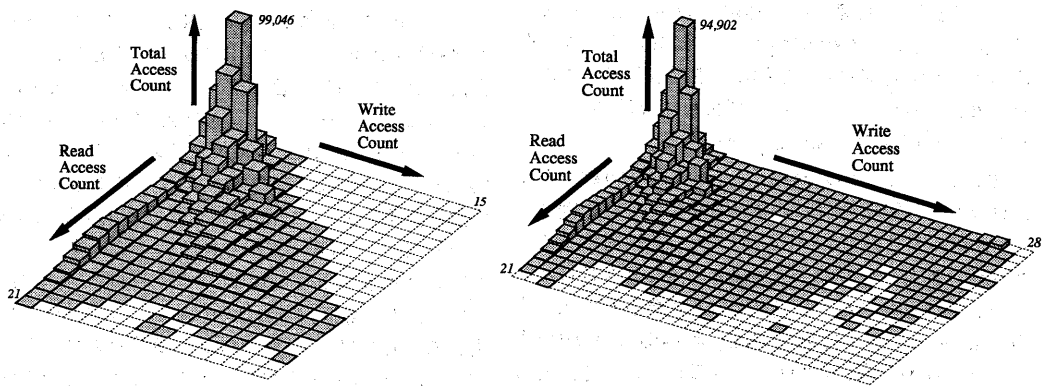
テーブル名	最大タプル長	容量 (MB)	クラスタード 索引	ノンクラスタード 索引
Warehouse	92 bytes	1	W_ID	—
District	97 bytes	1	D_ID, D_W_ID	—
Customer	665 bytes	320 + 14	C_W_ID, C_D_ID, C_LAST	C_W_ID, C_D_ID, C_ID
History	46 bytes	734	—	—
New-Order	7 bytes	4	NO_W_ID, NO_D_ID, NO_O_ID	—
Order	23 bytes	447 + 408	O_W_ID, O_D_ID, O_ID	O_C_ID, O_D_ID, O_W_ID, O_ID
Order-Line	52 bytes	8984	OL_W_ID, OL_D_ID, OL_O_ID, OL_NUMBER	—
Item	86 bytes	10	I_ID	—
Stock	310 bytes	497	S_I_ID, S_W_ID	—

表 1: テーブルの構成



(a) チェックポイント 未実行　　(b) チェックポイント 実行

図 2: アクセス位置の分布



(a) チェックポイント 未実行　　(b) チェックポイント 実行

図 3: ブロックのアクセス頻度の分布 (100 万アクセス)

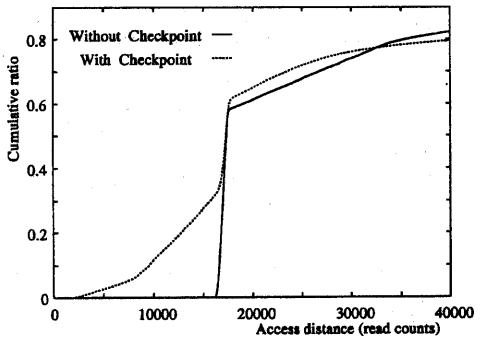


図 4: 読み出し-書き込みアクセス間隔の分布

2.2 ディスクアクセスの特徴

上述のトランザクション処理環境におけるディスクアクセスの特徴を簡単にまとめる。25 万回のアクセスにおけるアクセス位置の分布の図 2 に示す。図の縦軸はブロックの論理番号を示し、横軸は経過時間を示す。アクセス頻度が高い領域とアクセス頻度が低い領域へとかなり明確に分離される。アクセス頻度が低い領域は、Histroy, Order, Order-Line の各テーブルのデータを記録する領域であり、全データ領域の大半をこれらのテーブルで占めている。これらは、New-Order, Payment のトランザクションにより挿入が実行されるテーブルである。これらのテーブルは、180 日分以上の処理データを記録することができるだけの容量が割り当てられているため、実際に使われる部分は狭い範囲に限られ、このような特徴を示している。

チェックポイント実行時には、書き込みアクセスに周期性が見られる。チェックポイント未実行の時には、ホストのデータバッファ上でデータを置き直す時に古い更新されたデータの書き込み処理を行う必要があり、書き込みが時間的に一様に実行される。一方のチェックポイント実行時には、その処理により未書き込みデータが周期的に強制的に書き込まれる。そのため、書き込み無しにデータの置換可能な時期が有り、図のような周期性が現れることになる。

100 万アクセス中の読み出し/書き込み回数によりブロックを分類した時のアクセス頻度の分布を図 3 に示す。読み書き双方 1 回のみのブロックに対するアクセスが一番多いものの、数回程度の読み書きが行われているブロックに対するアクセスも多数存在する。索引や Item テーブルに関しては、読み出しのみの (またはその比率が高い) ブロックが多数存在する。その他のものに関しては、リード・モディファイ・ライトが行われているものが多い。チェックポイント実行時には、多数の書き込みが実行されるものが増加する。これは、データバッファ上で再利用性が高く、チェックポイントを実行しなければ書き込まれないものが存在することを示している。

50 万アクセス中の書き込みに関し、何アクセス前に同一ブロックへの読み出しが行われたかを調べた結果を図 4 に示す。図の横軸は読み出し-書き込み間の読み出しアクセス数を、図の縦軸は横軸以下の割合を示す。データバッファは LRU で管理されていると考えられ、データバッファで再利用されない更新データはほぼ同じアクセス間隔でリード・モディファイ・ライトが実行されると考えられる。再利用されないブロックが多数存在するため、チェックポイント未実行時にはある読み出し-書き込みアクセス間隔を持つものの割合が高くなっている。チェックポイントを実行する場合には、チェックポイント処理による書き込みが実行されるため、実行しない場合の立上り点前で書き込まれるものがかなり存在する。

また、図示はしないが、アクセスデータに時間的局所性を持つものがあり、書き込み-読み出しのアクセス間隔が短いブロックが存在する。

3 アクセストレースを用いたシミュレーションによる性能評価

3.1 Hot mirroring の実装方式

ミラー ホット領域の管理テーブル Hot mirroring においては、アクセス頻度によりミラー ホット領域と RAID5 コールド領域間でデータブロックの記録位置を移動させる。論文 [2] 中では、性能を重視した、データの記録位置を 1 ブロック毎に両領域で自由に移動する方式を採用した。このとき、記録位置管理のためのテーブルの大きさが問題となる。今回収集したトレースの 1 回のアクセスサイズは 2Kbytes であり、ブロック記録位置の管理テーブルに全データ容量の 0.2%[4] の容量が必要である。管理テーブルの大きさを削減するために、ミラー ホット領域を RAID5 コールド領域のキャッシュ的に用いる実装を考える。本評価では、RAID5 コールド領域に記録されているときには、その記録位置をブロック番号により固定的に決定する。ミラー ホット領域に記録されているデータブロックの管理は図 5 のような管理テーブルを用いてを行う。ブロックの記録位置の管理のために 2 つのテーブルを用いる。片方は、ミラー ホット領域のブロックの識別子 (ミラー ID) とそのブロックに関する情報を管理するテーブルであり、もう一つはブロック番号から記録されているミラー ID を調べる時に用いるハッシュ

[4]データ ID からブロック記録位置への変換テーブルに 1 エントリあたり 4bytes 必要であると仮定した。(4/2048 = 0.2)

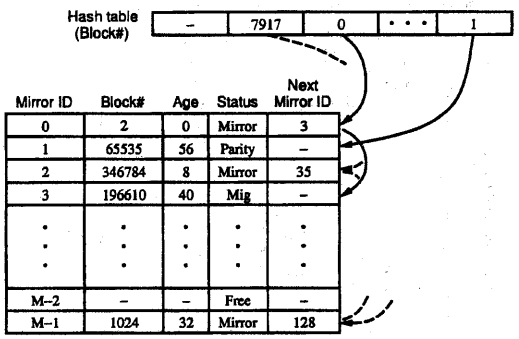


図 5: ミラー ホット領域の管理テーブル

テーブルである。

片方のテーブルでは、ミラー ホット領域内のブロックに関する情報として以下の 3 つのものを記録する。まず、あるミラー ID の領域に記録されているデータブロックのブロック番号を記録する。2 つめはブロック状態を記録する。最後の 1 つは、そのブロックが最後にアクセスされてからの経過時間の指標となる「Age」である。更に、ミラー ID 間のリンクを示すエントリを保持する。本評価では、ブロック番号からそれが記録されているミラー ID を検索するためにハッシュテーブルとこのリンクを用いた検索法を採用する。あるブロック番号を持つエントリが存在しない時には、そのブロックはミラーホット領域に記録されていない。本評価における Hot mirroring の構成では、6.7%のデータブロックがミラーホット領域に記録可能であるとすると、管理テーブルの大きさは全データ容量の 0.05%[5] 程度あれば良い。

書き込みとマイグレーションの実行方式 本評価においては、不揮発性書き込みバッファを利用し、アクセス要求が存在しないディスクとディスク ペア (ミラー ペアを保持するディスクの組みを指す) のもう片方のディスクに対して書き込みを実行する負荷分散を行う。ただし、書き込みバッファが満杯になったときにはこのような負荷分散を行うことができない。このときには、(書き込みが可能なトラック数による) 空き領域量とアクセスキュー長を利用した負荷分散を行う。本評価における実装では、書き込み時にその記録位置が変更されることを利用し、書き込み実行時に 7 ブロック分のデータを同一トラック内に存在するミラー ブロックに割り付け、その書き込みを一括化することにより効率的な書き込みを実行する。

ミラー ホット領域への書き込みを実行するため、ミラー ホット領域に存在するコールドブロックのマイグレーション (RAID5 コールド領域への書き戻し) を実行する必要がある。本評価における実装では、マイグレーションにおける読み出し処理を 1 トラック内の 7 ブロックに対して一括実行し、そのコストを削減する。また、読み出された 7 ブロックのうち、RAID5 コールド領域内で同一のトラックに記録されているものについては、それらの書き込みを一括化することによりそのコストの削減を図る。この効果を高めるために、書き込みバッファからミラーホット領域へ 7 ブロックを書き込むときに、可能な限り RAID5 コールド領域における記録位置が同一のトラックに存在するものを割り当てる[6]。コールドブロックの発見は、LRU アルゴリズムの近似となる Age を用いたアルゴリズムを用いて行う。書き込みが実行される度に、書き込みが実行されるディスク ペアの上に配置されている 4 トラックに対して Age の値に 1 を加える。アクセスが行われたブロックの Age の値は 0 にリセットされる。マイグレーションの読み出しアクセスは、ディスクペア毎独立に行う。各ディスクペア毎にトラックを順次チェックし、書き込み可能ブロックが 7 つ未満で、かつ、ディスクペア毎に定まる閾値以上の Age を持つブロックが 7 つ以上ある場合、Age の値が大きなものから順に 7 ブロックを選択しマイグレーションを行う。

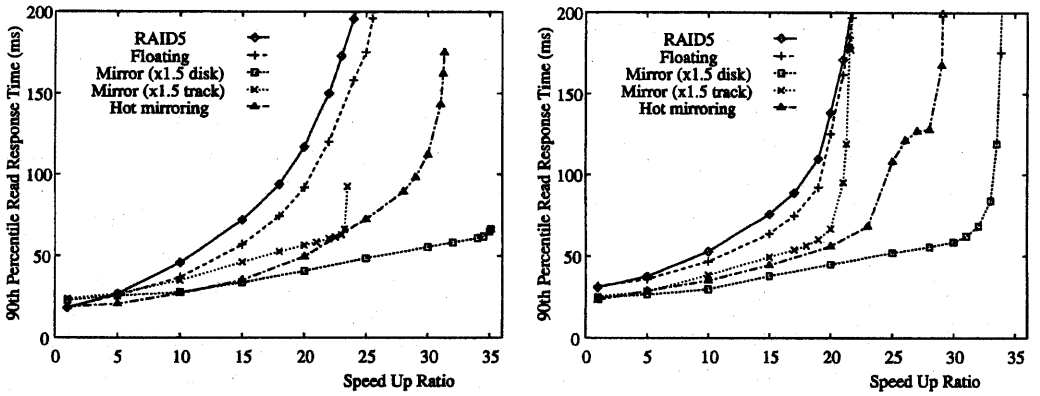
本評価においては、マイグレーションの読み出し処理を他のアクセス要求が存在しない時に (書き込みより低優先度で) 実行する。Hot mirroring においてはチェインド デクラスタリング [4] を用いたミラー ペアの配置により、あるディスクに存在するミラー ブロックはディスク ペアにより 2 種類に分類されるため、書き込みが実行可能なトラックが少ないディスク ペアに対してマイグレーションを実行する。書き込み可能なトラック数があるディスクペアに 80 以上存在する場合には十分な空き領域が存在すると判断し、そのディスクペアに対するマイグレーションの実行を控える。書き込み可能トラック数が少ない時にはマイグレーションの読み出しを書き込みと同時に実行する。マイグレーションにより読み出されたデータの RAID5 コールド領域への書き込みは即時に実行する。

3.2 シミュレーションによる性能評価

アクセストレースを用いて Hot mirroring の性能評価を行った。表 4 にディスクのパラメータを示す。5 台

[5] ミラー ID に関しては、記録位置で管理する。ブロック番号とミラー ID 間のリンクは各々4bytes、ブロック状態と Age の記録には各々1byte、ハッシュテーブルに 1 ミラーブロックあたり 4bytes 必要であるとすると、$(4+4+1+1+4)/2048 \times 0.067 = 4.6 \times 10^{-4}$。

[6] この割当を行うためには、RAID5 コールド領域内の各トラック毎に、幾つのブロックが書き込みバッファに存在しているかを管理する必要がある。本評価では、50 ブロック/トラック と仮定した。1 トラックあたり 1byte 必要であるから、テーブル量の全データ容量に対する比率は、$\frac{1/50}{2048} = 9.7 \times 10^{-6}$ となる。



(a) Normal mode (b) Rebuild mode

図 6: アクセストレースを用いた性能評価 (チェックポイント 未実行)

capacity	850 Mbytes
cylinders/disk	4350
tracks/cylinder	2 (3)
sectors/track	50
sector size	2048 bytes
revolution time	13.33 ms
seek time model	$2.867 + 0.0029 \cdot d + 0.13 \cdot \sqrt{d}$
track skew	6 sectors

表 4: ディスクモデル

のデータディスクに対して 1 台のパリティディスクを持つグループが 3 つある構成 (3*(5D+P)) を仮定する。各ディスクの 10%をミラー ホット領域に割り当てる[7]。チェックポイント未実行におけるトレースデータを用いて評価した性能を図 6 に示す。図の横軸はトレースデータに基づく到着シーケンスの加速率 (加速率 x では到着時間間隔は $1/x$ になる) を意味する。縦軸は、(a) 50 万アクセス中、(b) 復旧開始から終了までの復旧動作中、の 90%読み出しレスポンスタイムを示す。1000 万回の初期化アクセスの後に計測した。比較のため RAID5、フローティング、ミラーの性能も示した。データ容量を等しくするため、ミラーではディスク台数を 1.5 倍にしたものと、シリンダ内のトラック数を 1.5 倍にしたものを用いた。RAID5 では 0.15%、その他では 0.1%の容量の不揮発性書き込みバッファの存在を仮定した。Hot mirroring は RAID5 やフローティングより高い性能を示す。ディスク台数を増やしたミラーよりは性能が低いものの、そのコストを考えると Hot mirroring は十分に良い性能であるということができる。復旧動作時に Hot mirroring の性能が単調に悪化しないのは、書き込みのための空き作成処理の関係で書き込み時の動作状態が幾つかの状態に分けられるためである。通常動作時で×1.5 トラック ミラーの性能が低めなのは、以下の理由による。書き込みバッファが大きな RAID5 では、それを利用した読み出しアクセスの削減効果が大きくなる。RAID5 やフローティングでは、書き込みバッファを用いた書き込み一括化によるコスト削減効果がより大きく現れる。更に、本評価でのアクセスはある狭い領域に集中しているため RAID5 等では平均シーク時間が小さくなる。一方のミラーではデータをミラーペアにより完全に分離しているため、書き込み時にロングシークを必要し、性能の低下を招いている。

[7]全ディスク容量の 75%のデータを記録可能であり、ホット領域に 6.6%のデータブロックが記録可能である。

4 まとめ

TPC-C を基にしたトランザクション処理の実行時のテーブル記憶領域へのディスクアクセスのトレースを採取し、その特徴を調べた。このトレースを用いて Hot mirroring の性能評価を行った。性能とシステムコストの双方を考えると Hot mirroring はかなり良い構成であると言うことができる。今後は、アクセス特性の更なる利用について検討する予定である。

参考文献

[1] D. A. Patterson, G. Gibson, and R. H. Katz, "A Case for Redundant Arrays of Inexpensive Disks (RAID)," In *Proc. of ACM SIGMOD*, pp. 109–116, Jun. 1988.

[2] K. Mogi and M. Kitsuregawa, "Hot mirroring : A method of hiding parity update penalty and degradation during rebuilds for RAID5," In *Proc. of ACM SIGMOD*, pp. 183–194, Jun. 1996.

[3] Transaction Processing Performance Council (TPC), "*TPC BENCHMARK(TM) C Standard Specification*," revision 3.1, Jun. 1996.

[4] H. Hsiao and D. DeWitt, "Chained Declustering: A New Availability Strategy for Multiprocessor Database Machines," In *Proc. of IEEE Data Engineering*, pp. 456–465, Feb. 1990.