3R－5

トランザクション処理によるディスクアクセスのトレースを用いたHot mirroringの性能評価

茂木 和彦　喜連川 優
東京大学 生産技術研究所

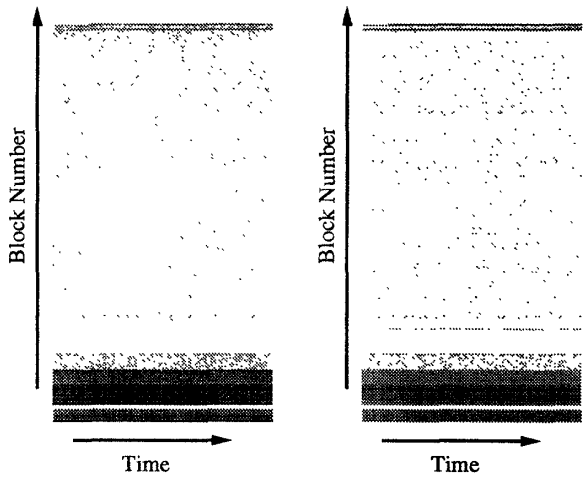
1 はじめに

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)[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では、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法を提案した [2]。本方式は、参照局所性を利用したミラーとRAID5の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を実行した時のディスクアクセスのトレースを採取した。本稿では、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を評価した結果について述べる。

2 トランザクション処理によるディスクアクセスの特徴

RAIDの用途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を挙げることができる。そこで、TPC-Cベンチマークを基に商用ミドルウェアを用いてSPARCstation 20/502 (OS: Solaris2.3) 上に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。データベースのテーブルの構成を表1に示す。これら9つのテーブルと2つのノンクラスタードインデックスをそれぞれ分離・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、テーブルとインデックスの記憶に総計11,420MB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。チェックポイントは実行されない。

100万回のアクセス (読み出し685,296回、書き込み314,704回) におけるアクセス位置の分布の図1に示す。アクセス位置に関しては、アクセス頻度が高い部分と



(a) Read　(b) Write

図 1: アクセス位置の分布

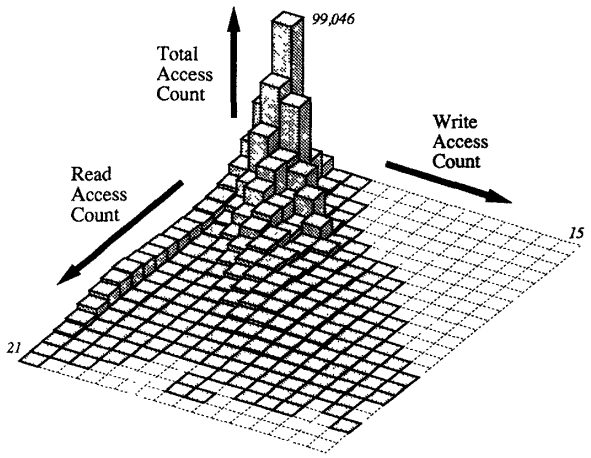


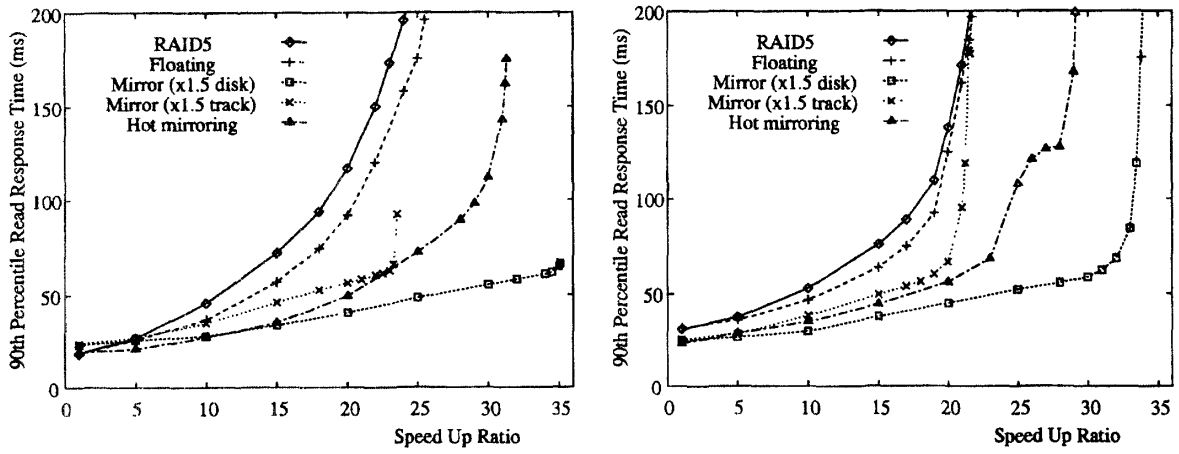
図 2: ブロックのアクセス頻度の分布

アクセス頻度が低い部分へとかなり明確に分離されることがわかる。また、このときの読み出し / 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2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。リード・モディファイ・ライトが行われているものが多い。問い合わせには最新のデータに関する (時間的局所性をもつ) ものがあり、図には明示されないものの、書き込み直後に読み出しが実行されるブロックが存在する。

Performance evaluation of Hot mirroring using a disk access trace by transaction processing
Kazuhiko Mogi and Masaru Kitsuregawa
Institute of Industrial Science, University of Tokyo

Table name	Clustered index	Non clustered index
Warehouse	W_ID	—
District	D_ID, D_W_ID	—
Customer	C_W_ID, C_D_ID, C_LAST	C_W_ID, C_D_ID, C_ID
History	—	—
New-Order	NO_W_ID, NO_D_ID, NO_O_ID	—
Order	O_W_ID, O_D_ID, O_ID	O_C_ID, O_D_ID, O_W_ID, O_ID
Order-Line	OL_W_ID, OL_D_ID, OL_O_ID, OL_NUMBER	—
Item	I_ID	—
Stock	S_I_ID, S_W_ID	—

表 1: テーブルの構成



(a) Normal mode (b) Rebuild mode

図 3: アクセストレースを用いた性能評価

3 アクセストレースを用いた性能評価

アクセストレースを用いた Hot mirroring の性能評価を行った。5 台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 つある構成を仮定する。ディスク容量は 850MB とし、各ディスクの 10% をミラー ホット領域に割り当て、RAID5 コールド領域のキャッシュ的な動作をさせる。この構成においては、全ディスク容量の 75% のデータが記録可能であり、ホット領域に 6.6% のデータブロックを記録可能である。この構成におけるトレースデータを用いて評価した性能を図 3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を、つまり、加速率 x は到着時間間隔を $1/x$ にすることを意味する。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。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 倍にしたものを載せた。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が存在すると仮定した。通常動作時で ×1.5 トラックのミラーの性能が低めなのは、本評価でのアクセスパターンはある狭い領域に集中しているために平均シーク時間が短いが、ミラーでは書き込み時にロングシークを必要とするためである。復旧動作時に Hot mirroring の性能が単調に悪化しないのは、書き込みバッファと空き領域作成動作により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。Hot mirroring は RAID5 とフローティングより高い性能を示す。また、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 は十分に良い性能を出しているということができる。

4 まとめ

TPC-C ベンチマーク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、そのアクセスの特徴を調べた。このトレースを用いて Hot mirroring の性能評価を行った。Hot mirroring は同様な構成の RAID5 やフローティングよりも高い性能を示す。ディスク台数を増やした同容量のミラーよりは性能は低いものの、システムのコストも考えると Hot mirroring が最も良い構成であると言えるであろう。

参考文献

[1] D. A. Patterson, G. A. Gibson, and R. H. Katz. "A Case for Redundant Arrays of Inexpensive Disks (RAID)." In *Proc. of ACM SIGMOD Conf.*, pp. 109–116, June 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 Conf.*, pp. 183–194, June 1996.

[3] Transaction Processing Performance Council(TPC). TPC BENCHMARK(TM) C Standard Specification, Revision 3.1, June 1996.