5AA-5

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性解析

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。RAIDからその性能を最大限に引き出すためには、アプリケーションの特性をいかにして記憶管理に用いるかが重要な鍵となる。Informed prefetching and caching[2] のようなアプリケーションから直接アクセス情報を受け取る手法が存在するが、これらではアプリケーションやOSに手を加える必要があるという問題点が存在する。RAIDコントローラ側のみで採取・利用可能なアプリケーション特性のみを用いることも考えられ、この方針の有効性について検討する必要がある。RAIDの用途で重要なものとしてトランザクション処理システムが存在する。本稿では、TPC-Cベンチマーク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から読み取れる特性を示し、それをRAIDの高性能化に活用する手法について検討する。

2 アクセス特性とRAIDの高性能化

文献 [3] で用いたトレースデータのうち、周期的にホスト上の未書き込みデータをディスクに書き込む処理を実行したものを基に検討する。データベースの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を分離・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、ホスト上のデータバッファは56MBとした。アクセスは2KBの固定長である。

領域毎のアクセス特性　全体と各領域毎のブロックのアクセス頻度分布を図1に示す。このように、各記憶領域毎にそのアクセス特性は大きく異なる。これらを大きく分類すると、文献 [3] に示されたように、データ領域の大半を占めアクセスの局所性が存在する挿入が実行されるテーブルのデータを記憶する領域と、基本的には多数のアクセスがほぼ一様に行われるそれ以外の領域に2分することができる。例えば後述するような手法を用いて、領域毎のアクセス特性をRAIDの高性能化に役立てることが可能である。

テーブル名	タプル数	容量 (Mbytes)
Warehouse	14	1
District	10×14	1
Customer	30,000×14	320 + 14
History	840,000×14 以上	734
New-Order	約 9,000×14	4
Order	840,000×14 以上	447 + 408
Order-Line	8,400,000×14 以上	8984
Item	100,000	10
Stock	100,000×14	497

表 1: テーブルの構成

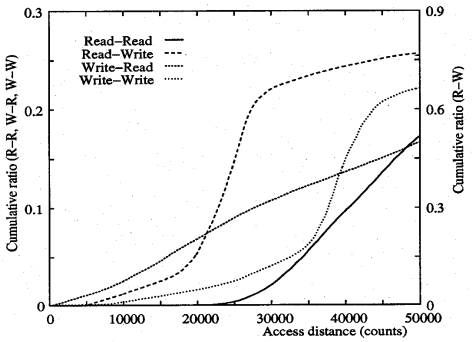

図 2: 同一ブロックへのアクセス間隔 (100万アクセス)

ブロックの参照局所性　データの参照局所性により、アクセスが実行されるブロック間にはアクセス頻度差が存在する。しかし、ホスト上にはデータバッファが存在し、本当にアクセス頻度が高いデータに関してはディスクアクセスが必ずしも多数行われる訳ではない。本トレースではデータバッファはLRUアルゴリズムで管理されていると推測され、基本的に同一ブロックへのアクセス間隔が長くなる特徴を示す (図2)。(アクセス間隔は、データバッファのページ数と同程度以上である。) ただし、データバッファに存在しきれないデータと時間的な局所性を持つ処理により、書き込み-読み出しのアクセス間隔は短いものも存在する。図示はしないが、アクセス間隔についても領域毎に特性が異なる。

ブロックのグループ化と参照局所性　負荷分散や階層構成を利用した高性能化手法には、データの記録位置を動的を変更するものがある。このとき、どの程度の単位で記録位置の変更を行うかは重要なパラメータとなる。適当なサイズでブロックをグループ化したものを単位として考えた時の記憶領域量とアクセス割合を

An Analysis of Characteristics of Disk Accesses in a Transaction Processing Environment
Kazuhiko Mogi and Masaru Kitsuregawa
Institute of Industrial Science, University of Tokyo

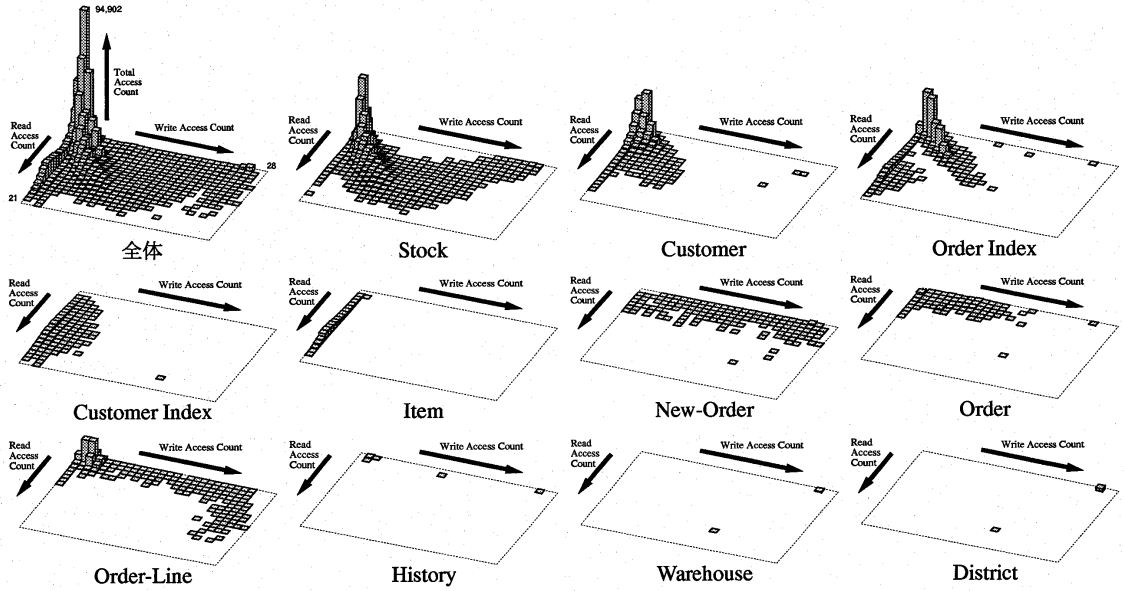

図 1: 領域毎のブロックのアクセス頻度の分布 (100 万アクセス)

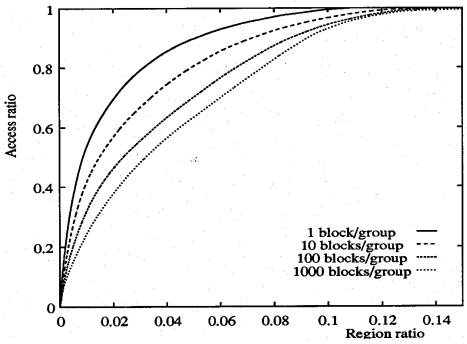

図 3: グループ化による参照局所性 (1000 万アクセス)

図 3 に示す。文献 [3] で示されたように、テーブル毎に異なった領域を割り当てることによりアクセスされる領域は集約化される。そのため、ある程度大きな単位でグループ化してもアクセスの大半を占めるグループの容量はさほど変わらず、図 3 のような特性を示す。

読み出しのプリフェッチ　各領域において 1 ブロックの読み出しが多い。ただ、幾つかの候補を読み出して、その中から適切なものを選択する処理も行われている。その候補が多数ある場合には、複数のブロックを読み出す必要がある。それらを連続的に配置しておくことにより、単純なプリフェッチによりアクセス数を削減することができる。本トレースではこの効果を利用可能であり、特に、Customer テーブルでその効果が大きい (表 2)。

連続的な書き込み　各領域において 1 ブロックの書き込みが多い。ただし、挿入が実行されるテーブルに関

サイズ	5×1	5×5	5×10	5×20	5×100
全体	3.8%	8.8%	10.5%	11.4%	11.9%
Cust. のみ	22.3%	34.8%	35.7%	35.8%	36.6%

(5 ブロック先読み時のヒット率, 10 万アクセス)

表 2: プリフェッチの効果

しては、挿入されたデータが連続した領域に割り当てられることが期待でき、この場合には書き込みを一括化により効率的に実行することができる。

3 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環境におけるディスクアクセスのトレースを基に、アプリケーション特性の RAID のデータ記憶管理への活用法について検討した。キャッシュは、十分な大きさがない場合にはその効果は小さい。テーブルを記憶する領域毎にそのアクセス特性は異なり、領域毎に適切な管理を行うことにより RAID の高性能化に寄与できると考えられる。

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*, pp. 109–116, Jun. 1988.

[2] R.H. Patterson, G.A. Gibson, E. Ginting, D. Stodolsky, and J. Zelenka. Informed Prefetching and Caching. In *Proc. of 15th ACM SOSP*, pp. 79–95, Dec. 1995.

[3] 茂木和彦, 喜連川優. 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. 信学技報 DE97–28, 電子情報通信学会 データ工学研究会, Jul. 1997.