4D－10　RDBMSへのディスクキャッシュ装置の適用による高速化手法

保田　浩之 、住田　利幸
沖電気工業(株)　コンピュータシステム開発本部

１．はじめに

リレーショナル型データベース管理システム（ＲＤＢＭＳ）のトランザクション性能を向上させるには、ディスクアクセスタイム、特にライトアクセスタイムの短縮と、アクセス回数の削減が不可欠である。磁気ディスク装置のアクセスタイムはディスクキャッシュ装置を付加することによって短縮することが可能であり、ＲＤＢＭＳ自身が管理するデータベースバッファと組み合わせることにより高速化を図ることができる。

筆者らは小容量のライトデータ専用のバッファ（ライトバッファ）を付加したディスクキャッシュ装置を試作し、ミッドレンジコンピュータ　OKITAC 8300 に適用した。本稿では、そのバッファ管理アルゴリズム、実機評価結果について述べる。

２．ディスクキャッシュ装置の概要

２．１　構成

試作したディスクキャッシュ装置は、最大３２ＭＢのディスクキャッシュメモリを内蔵し、さらにバッテリバックアップされたＳＲＡＭからなる６４ｋＢのライトバッファを持つ。ＳＣＳＩをインターフェースとする磁気ディスク装置を４台まで接続でき、内蔵のファームウェアでディスクへのアクセス制御、ＤＭＡ制御、キャッシュ／バッファ管理を行う。

２．２　書き込み方式

ライトデータは一旦ライトバッファへ格納してＣＰＵにライト終了を通知した後に、適当なタイミングでディスクに書き出す（ライトバック）。（図１）　このため、見かけ上のライトアクセスタイムは大幅に短縮される。また、バッファへのライト動作とライトバック動作は並列に実行可能である。

バッファ内のデータは近隣のブロックをまとめたグループを単位としてＬＲＵで管理し、連続したブロックは一括してライトバックする。

３．評価システム

３．１　データベース管理システムＲＥＡＭ

ＲＥＡＭは、ＪＩＳに準拠したＳＱＬをサポートするＲＤＢＭＳであり、トランザクション終了毎にジャーナルを書き出すコミットメント制御を行う。また、主記憶上で独自のデータベースバッファの管理を行う。

３．２　評価モデル

評価には、Debit Credit ベンチマークの１０ＴＰＳモデル（総ファイル容量　約２５０ＭＢ）を使用し、トランザクション実行性能を測定した。また、ディスクキャッシュ装置の内蔵ファームウェアに内部での各種処理時間の集計機能を組み込んで、動作特性を解析した。

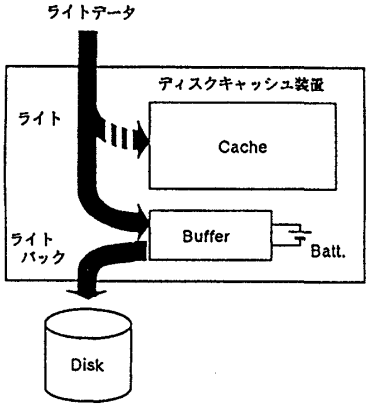

図１．ディスクキャッシュ装置の書き込み方式

４．特性評価

４．１　ディスクＩＯ特性

トランザクション実行時のディスクＩＯ特性を測定した結果、レコードの更新、ジャーナルライトが主体となるためにライト比率（ライト回数／総アクセス回数）は、０．３～０．６と高い。また、ディスクキャッシュ装置上でのリードヒット率は２０～７０％（キャッシュ３２ＭＢ時）であり、一般のアプリケーションに比べて低い。ＲＥＡＭのデータベースバッファの容量を大きくするほどライト比率が高くなり、リードヒット率は下がる。

アクセスの頻度およびライト比率が高いため、ライトバック動作を隠蔽しきれずに競合による待ち時間が発生するが、ライトバッファの溢れは発生しなかった。

４．２　リードヒット率とスループット

リードミス時にディスクをアクセスしようとする時にライトバック動作中であるとディスクキャッシュ装置の内部で待ち時間が発生する。内部ログによれば、リード時のミスペナルティの大部分をこの待ち時間が占めている。これは、リードヒット率が上がれば競合の確率も下がり、ミスペナルティも減少してスループットが向上するということを意味する。図２をみても、

Performance Improvement Methods of RDBMS with Hardware Disk Cache
Hiroyuki HOTA , Toshiyuki SUMIDA
Oki Electric Industry Co., Ltd.

ライトバッファを持たない場合よりヒット率の性能に対する影響が大きいことがわかる。

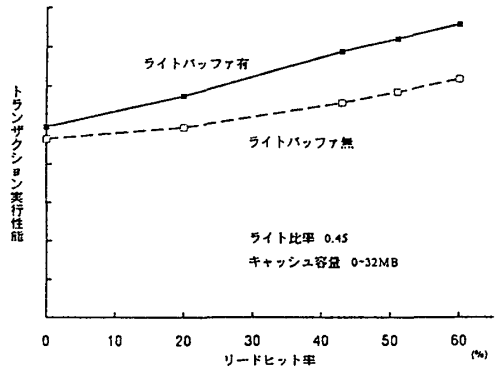

図2．リードヒット率とトランザクション性能

4．3　バッファライトとライトバックの並列動作

ライトバッファへのデータライトとライトバックの並列動作ができない場合、競合による待ち時間が発生する。そこで、これを可能にしたことによる効果を確認するため、並列動作を禁止した場合との比較を行った。その結果、並列動作を可能にした場合、ライトアクセスタイムを平均で約4割短縮できることを確認した。

4．4　ライトバックタイミング

ライトバックを行うタイミングを決定するアルゴリズムとして、次の2つの方式を選択し、比較を行った。

一方は、ライトバッファが空の状態からライトデータを格納して空でなくなった時点から、一定時間（1秒）後にその時点にバッファに格納されている全てのデータを連続して書き出す方式である。（図3－a 一斉フラッシュ方式）　そして他方は、ライトバックを連続して行わずに、バッファ内に格納されているデータの量に応じて、データ量が少ないときは長く、多いときは短くした間隔（0.1～0.8秒）で一度ずつ書き出す方式である。（図3－b 可変間隔方式）

この二つの方式で、他は同一の条件として、まとめ書きの効率を表すライトバック比率（ライトバック回数／ライトアクセス回数）を比較したところ、一斉フラッシュ方式で0．6～0．8、可変間隔方式で0．5～0．7であり、平均して2割ほど後者の方が値が小さく、トランザクション実行性能も1割近く向上することを確認した。ライトアクセス回数に対してライトバック回数が少なくなってスループットが向上すると共に、ミスリードアクセスと競合する確率も減少するためである。

これは、一度に全てのデータを書き出さずに、ＬＲＵ上位のブロックがバッファ内に長く留まるようにしたことで、シーケンシャルなジャーナルライトや小容量の BRANCHテーブルの更新などを、より多くまとめて書き出すようになるためであると思われる。

5．システム性能評価

ディスクＩＯネックとなっている場合にディスクキャッシュ装置を適用した場合の効果を図4に示す。この図で、平均アクセスタイムはファームウェア、ＳＣＳＩ上でのオーバーヘッド、データ転送時間を含むものであり、実行性能は相対値である。平均アクセスタイムは1／3～2／3に短縮される。そしてＣＰＵを追加しても性能向上の効果がなかったのが、ディスクキャッシュ装置を付加することでほぼ2倍の性能を達成している。これはディスクの台数を2倍以上に増設したことに相当する。

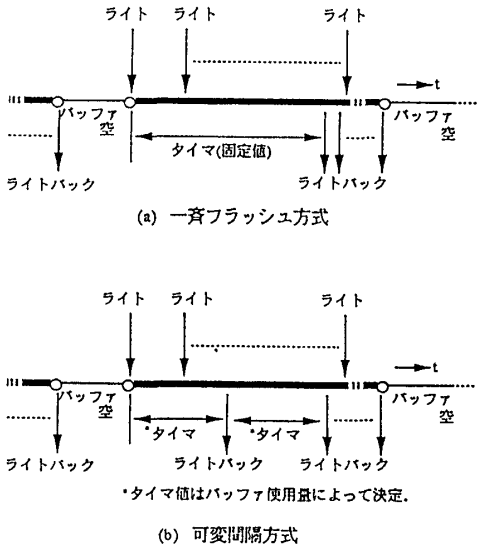

図3．ライトバックタイミング

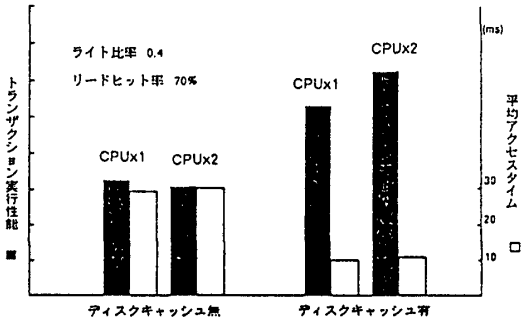

図4．ディスクキャッシュ装置の効果

6．まとめ

ＲＤＢＭＳのトランザクション処理を高速化するためのディスクキャッシュ装置の処理方式と、その効果について述べた。今回試作したディスクキャッシュ装置は、ライトバック方式を実現する機構としてはライトバッファとして小容量のＳＲＡＭとバッテリを付加したのみであるのでコストパフォーマンスに優れている。

【参考文献】

[1] Anon. et al.: "A Measure of Transaction Processing Power", Datamation, pp. 112-118, April 1, 1985.

[2] 喜連川他：データベース処理におけるベンチマーク、情報処理、Vol.31, No.3, 1990