5Q-1

高速データベースマシン HDMの障害回復方式

峯村 治実　板倉 國司*　下平 純一*　中村 俊一郎　石田 喬也
三菱電機㈱ 情報電子研究所
*三菱電機東部コンピュータシステム㈱

1. はじめに

近年、高まりつつあるリレーショナル・データベース・システムの高速化需要に答えるため、我々は、高速データベースマシンHDMの開発を行っている。HDMは、並列処理や最適化したディスクアクセス方法などによって高速処理を実現している[2-4]。本論文では、HDMの障害回復方式について述べる。本方式は、SQL文による論理REDOログなどの手法により、ログ量を大幅に減らし、実処理時のオーバーヘッドを少なくできるという特徴を有する。

2. データベースシステムの障害回復

一般にデータベースシステムの障害回復方法としては、UNDO・REDOログ、チェックポイント、およびダンプ・ロードといった技法が知られている[1]。更新処理の途中でシステムがダウンした場合に、データベースの首尾一貫性を保つため、その処理によって変更されたデータを元に戻すのに使用されるのがUNDOログであり、また、システム再立ち上げ時に、障害発生前に正常終了した更新処理を再実行するのに用いられるのがREDOログである。一般にUNDO・REDOログは、それぞれ更新前、および更新後の値をディスクなどに記録したものである。チェックポイントは、REDO、およびUNDOの処理量（ログ量）を減らすために、バッファ中の更新された値を定期的にディスクに書き出す技法である。また、ダンプ・ロードは、ディスククラッシュなどの媒体障害からの回復を行なうための技法で、予め他の媒体（MTなど）にコピー（ダンプ）しておいた全データを、障害発生時にディスクにロードすることによって回復処理を行なうものである。ロード後にREDOを行うことにより、障害発生直前まで回復することができる。

HDMは、これらの技法（UNDO・REDOログ、チェックポイント、ダンプ・ロード）を、以下に述べる手法で実現している。以下、HDMにおける各技法の実現方式、および特徴について述べる。

3. UNDOログ

HDMは、1台のマスター・プロセッサと複数台のスレーブ・プロセッサからなるマルチプロセッサのデータベースマシンであり[2]、それらの並列処理と、各プロセッサに搭載した大容量メモリのキャッシュとしての使用により、高速処理を実現している。これらの点を考慮して、HDMでは以下のようにしてUNDOログ機能を実現する。

- 各プロセッサで個別に取得する。
- トランザクション単位に、更新前の物理ページ・イメージとして、ディスク上のUNDOログ領域に取得する。
- ログ取得効率を向上させるため、各プロセッサのメモリ上に設けたUNDOログバッファによりバッファリングを行う（図1参照）。あるページが、ログバッファからディスク上のログ領域へ書き出されるのは、以下の2つの場合である。
  ①ログバッファ満杯時
  ②キャッシュ上の対応するページがディスクに掃き出される直前
  従って、更新されたページがキャッシュ上にあるうちは、そのページのログをディスクに書き出す必要はないため、UNDOログの量を減らすことができる。
- UNDOは、ログバッファ、およびログ領域から読み込んだ更新前イメージを逆順に（すなわち、新しいものから順に）、該当するページに物理的に重ね書きすることによって行う。

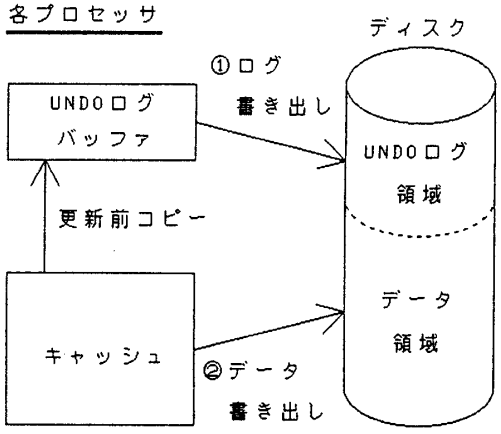

図1　UNDOログの管理方法

4. REDOログ

従来の障害回復手法では、REDOログとして、更新後の物理ページイメージ、あるいは更新後のレコードの値そのものを用いていた。しかし、リレーショナル・データベースでは、1つのSQL文で大量データの更新が簡単に行えてしまうため、従来方式では、ログの量が膨大なものになってしまうという問題点があった。

Recovery Method of the High Speed Database Machine, HDM
Harumi MINEMURA[1], Kuniji ITAKURA[2], Jun-ichi SHIMODAIRA[2], Shun-ichiro NAKAMURA[1], Takaya ISHIDA[1]
1 Mitsubishi Electric Corporation, 2 Mitsubishi Electric Computer System (Tokyo) Corporation

そこでＨＤＭでは、更新後の値をREDOログとして記録しておかなくても、どのような更新処理を行ったかさえ記録しておけば、その処理を再実行することにより回復を行うことができるという点に着目し、REDOログとして、更新処理のＳＱＬ文（実際は、そのＨＤＭの中間言語による表現）を取得することとした。これにより、回復に要する時間は少し長くなるものの、REDOログの量を大幅に減らすことができる。以下に具体的な処理方法を示す。

①REDOログは、マスター・プロセッサのディスク上に設けたREDOログ領域に、ＨＤＭの中間言語形式で取る。中間言語には、トランザクションIDや実行したユーザのIDなど、再実行に必要な情報が記されている。

②トランザクション管理プログラムは、１つの更新処理単位（action、すなわち、１ＳＱＬ文）が正常終了したとき、その中間言語をログ領域に書き出し、ユーザに処理結果を返す。

③ログ領域が満杯に近くなったとき（80%を越えたとき）、フロントエンドの計算機に接続されたカートリッジＭＴなどにREDOログをコピーする（現在、ＨＤＭは、図２に示すように、フロントエンドの計算機として当社のエンジニアリング・ワークステーションME1000シリーズをＳＣＳＩで接続するようになっている）。

④REDOは、コミットされたトランザクションの各actionを時間順に再実行することによって行う。

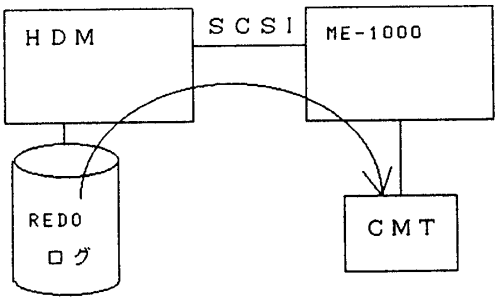

図２　REDOログのコピー

5. チェックポイント

ＨＤＭでは、UNDOログの量を減らすためと、システム回復時間を短縮するために、チェックポイントを設けている。チェックポイントは、Transaction Consistent Checkpointとする。すなわち、トランザクションの切れ目でチェックポイントを取る。また、チェックポイントでは、キャッシュがディスクに書き出されるため、ディスク上のデータベースは首尾一貫した状態になる（UNDOが不要になる）。そのため、チェックポイントでのUNDOログの消去が可能になる。チェックポイントは、以下の場合に取られる。

・任意のプロセッサのディスク上のログ領域が満杯に近くなったとき（80%を越えたとき）。
・前回のチェックポイントから一定時間（例えば、５分）経過したとき。

また、チェックポイントでは以下の処理を行う。

・キャッシュのディスクへの書き出し
・UNDOログの消去
・REDOログ領域が満杯に近くなっている場合は、フロントエンドのＣＭＴ等へのREDOログのコピー

6. ダンプ・ロード

ダンプは、各テーブル、インデックス、およびレコードを生成するＳＱＬ文の中間言語を、フロントエンドのＣＭＴ等にセーブすることによって行う。ロードは、セーブされたダンプファイル（中間言語）をそのまま再実行することによって行われる。この論理ダンプ・ロードによって、ディスクの特定部分が破壊された場合でも、データを正しくロードでき、また、特定のテーブルのみのダンプ・ロードも容易に行うことができる。なお、ＳＱＬの<insert>文では、同時に１つのレコードしか追加できず効率が悪いため、複数のタプルを一括してストアできる<multi-insert>機能を設けた。これにより、ロードに要する時間を短縮することができる。

7. おわりに

ＨＤＭで実現しようとしている障害回復方式は、以下に示すように非常に特徴のあるものである。

・更新されたページをキャッシュからディスクに書き出す直前にUNDOログをディスクに書き出すため、更新されたデータがキャッシュ上にあるうちはUNDOログを書き出さなくてもよく、ログ量を減らすことができる。
・REDOログを、ＳＱＬ文で取得することにより、ログ量を大幅に減らすことができる。
・チェックポイント時にUNDOログを消去できるため、ログ量を減らすことができる。
・データをＳＱＬ文に変換してダンプ（論理ダンプ）するため、ディスク上の特定の（物理的な）位置が破壊されて使用できなくなった場合でも、データを正しくロードできる。

現在、ダンプ・ロード機能が完成し、REDOログ機能を開発中である。今後は、UNDOログ機能を実装するとともに、ログ取得のオーバーヘッドや、障害からの回復時間などの測定・評価を行っていく予定である。

［参考文献］

(1) Gray, J. N., "Notes on Data Base Operating Systems," published in R. Bayer, et al.(eds.), "Operating Systems: An Advanced-Course," Springer-Verlag (1978).

(2) Nakamura, S. et al., "A High Speed Database Machine - HDM," Proc. 5th Int. Workshop on Database Machines (1987).

(3) 中村他「高速ﾃﾞｰﾀﾍﾞｰｽﾏｼﾝHDMのｱｰｷﾃｸﾁｬ」情報処理学会第35回全国大会、4Cc-6(1987)

(4) 峯村他「高速ﾃﾞｰﾀﾍﾞｰｽﾏｼﾝHDMの性能評価」情報処理学会第35回全国大会、4Cc-7(1987)