5P−7 フラッシュメモリを利用する トランスペアレントバックアップシステムの設計と性能評価

高倉弘喜　上林彌彦
京都大学工学部

1. はじめに

データ更新時にディスクアクセスを必要としない主記憶データベースは、数千から数万tpsのような高速なデータ処理が行えると期待されている。しかし、主記憶を無停電電源などで不揮発化してもその信頼性はディスクに比べ低いと考えられる。このため、主記憶データを定期的に二次記憶にバックアップする必要がある。高速なデータベースシステムはデータ処理だけでなく障害後の回復処理も効率的に行わなければならない。回復処理に要する時間は検査点間隔に比例すると考えられるので、できるだけ短い間隔でバックアップを行う必要がある。検査点間隔を短くするには、システムにほとんど影響を及ぼさないバックアップ方式が必要となる。本研究では、システムの動作中に自動的にバックアップを行なうトラスペアレントバックアップ方式を提案した[3]。

本稿では、電源が切れてもデータを保存でき、かつ、高速なランダムアクセスを行なうフラッシュメモリをバックアップ媒体として利用するトランスペアレントバックアップシステムについて述べる。フラッシュメモリを利用することにより、1000tpsのシステムで30分間隔のバックアップが1分弱で終了し、回復処理も最悪の場合で10秒以下というきわめて高速な検査点処理および回復処理が行える。

2. 基本的事項

• フラッシュメモリ

フラッシュメモリは電源を切ってもそのデータを保持し、ページ単位にランダムアクセスが行えるといった性質を持つ。また、アクセスサイクルは数十から百ns程度である。データにアクセスする前に、メモリ内部でメモリセルとデータレジスタ間のデータ転送を行う。また、データ書き込みの前にチップ全体かブロック単位の消去を行なう必要がある。これらの時間はディスクのアクセス待ち時間に相当すると考えられる。データ転送時間は数十μs程度と極めて短く、また、消去時間は10ms程度である。これらの時間はシーケンシャルアクセスやランダムアクセスの区別なく同じである。従って、従来のディスクへのバックアップ[2][4]のように回復処理の効率化のためアクセス頻度順にページをシリンダに割り付けるような複雑な計算をせずに、主記憶のイメージデータをそのままバックアップすることができる。

• バッテリーバックアップの問題点

主記憶の周辺回路に異常が起こった場合、通常の手段ではバッテリーバックアップされた主記憶データを読み出すことは出来ないし、修理でデータを失う可能性が高い。特に、DRAMコントローラなどの制御回路が故障するとデータは完全に失われる。これに対して、ディスク装置の周辺回路や機械部分が故障してもディスク面さえ無事ならデータは失われない。主記憶の周辺回路が故障する確率はかなり低いが、高信頼性システムはこのような障害にも対応すべきであると考えられる。

Design and Performance Evaluation of a Transparent Backup System utilizing Flash Memory
Hiroki TAKAKURA　Yahiko KAMBAYASHI
Faculty of Engineering, Kyoto University

3. トラスペアレントバックアップシステム

3.1 システム構成の概要

本研究では主記憶にデュアルポートDRAMを利用する方式や主記憶を二重にした方式(2-プレーンバックアップ方式)を提案した[3]。どちらの方式もデータ処理に影響をほとんど及ぼすことなく主記憶データのバックアップを取ることができる。例として2-プレーンバックアップ方式を図1に示す。本方式はデータベース管理部(DM)、バックアップ回復処理部(BRM)、主記憶プレーン(MMP)、バックアップメモリプレーン(BMP)およびバッファで構成される。通常(図1(a))は2枚のメモリを一枚の主記憶として扱う。バックアップ中(図1(b))はMMPが通常の処理を受け、BMP中のデータをある検査点時刻のイメージデータとしてBRMへ読み出す。バックアップ中にMMPとBMPは内容が異ってしまうので、バックアップ後に補正を行なう(図1(c))。本方式は、シンクロナスDRAMのようなキャッシュ内蔵型DRAMなどを利用するシステムでもキャッシュのヒット率に影響を及ぼさずにバックアップを行なうことができる。

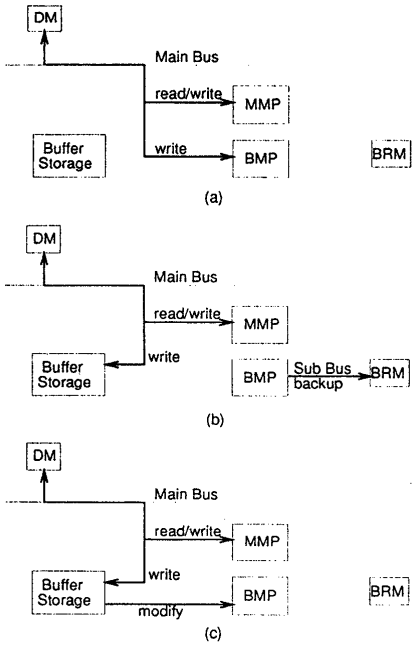

図1: 2-プレーンバックアップ方式

3.2 バックアップ回復処理部の構成

BRMはフラッシュメモリとその制御回路で構成される。フラッシュメモリには書き変え回数に制限があるため、検査点の度に主記憶全体をバックアップするのは望ましくない。このため、各ページごとに1ビットのカウンタを用意する。検査点時

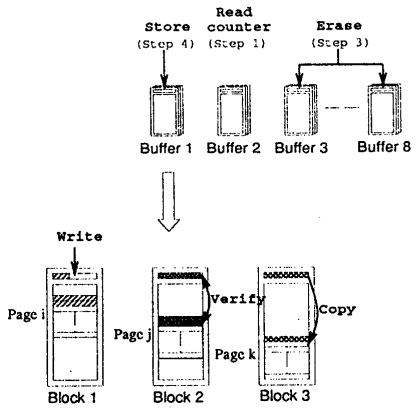

図 2: マルチバッファ

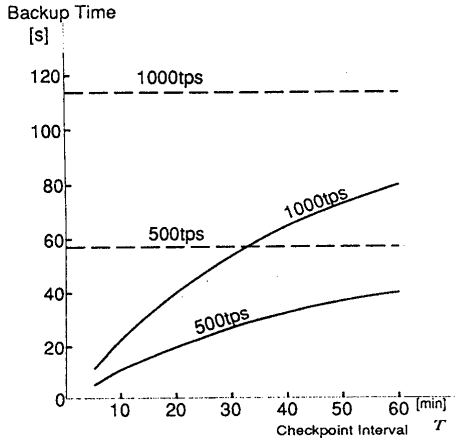

図 3: バックアップに要する時間

刻になると、このカウンタを見て前回の検査点時刻以降に更新を受けたページのみをバックアップする。

フラッシュメモリの消去は 10ms かかり、ブロックの書き込み時間に比べかなり長い。このため本稿のシステムでは、図 2 に示すようなマルチバッファを利用することにより、ブロック消去、データレジスタからメモリセルへのデータ転送および転送後の書き込みデータの確認を全て他のフラッシュメモリの書き込み操作のバックグランドで行なう。従って、バックアップに要する時間はフラッシュメモリの書き込み時間の総計で表されることになる。

4. 性能評価

本稿では、TPC ベンチマーク [1] を元にしたシミュレーションを行なった。

[バックアップ]

検査点間隔に対するバックアップに要する時間を図 3 に示す。図中の破線は主記憶全体をバックアップするのに要する時間を表している。このように、バックアップは検査点間隔が 1 時間の場合でも 1 分程度で終了する。

[回復処理]

回復処理に要する時間は、検査点データを主記憶に戻し、その主記憶データに対してログを適用する時間で表される。従って、検査点時刻直前に障害が発生した場合に最長の回復処理時間がかかることになる。本稿のシステムでは、良く使われるホットスポットデータは直ちに回復し、そうでないデータは必要になった時に回復するインクリメンタル回復を行なう。本稿ではホットスポットデータの回復に要する時間についてのみ考慮し、その他のデータの回復に要する時間は一般に不定であるのでここでは述べない。TPC ベンチマークの場合 Branch と Teller のデータベースがホットスポットであると考えられる。この 2 つのデータベースに関して検査点間隔に対する最長の回復処理時間を図 4 に示す。検査点時刻直後に障害が発生し、BMP のデータが無事であった場合、回復処理に要する時間は、0 であると考えられる。例えば、30 分おきにバックアップを行なう 1000tps のシステムの場合、最長で 9.2 秒、平均すると 4.6 秒で回復処理が終了する。

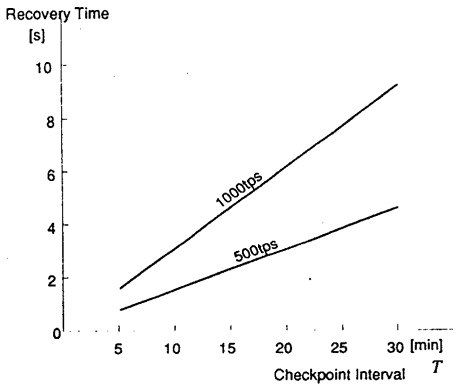

図 4: 回復処理に要する時間

本稿のシステムの特徴はバックアップと回復処理がハードウェア的に制御されることである。このため、トランザクションスループット (tps) が大きくなってデータベースが巨大になった場合、これらの処理を並列に行なうことができる。理想的には、図 3、4 と同じ時間で処理が終了する。

5. まとめ

本稿では、フラッシュメモリを利用するトラスペアレントバックアップシステムの概略について述べた。また、TPC ベンチマークに基づいたシミュレーション結果を述べた。フラッシュメモリを利用することで、非常に簡単な制御で効率的なバックアップと回復処理が実現できる。

参考文献

[1] Jim Gray, "The Benchmark Handbook," Morgan Kaufmann Publishers, 1991, pp.19-117.

[2] L. Gruenwalld, M.H. Eich, "MMDB Reload Algorithms," Proc. of ACM SIGMOD International Conf. on Management of Data, 1991, pp.397-405.

[3] Y. Kambayashi, H. Takakura, "Realization of Continuously Backed-up RAMs for High-Speed Database Recovery," The 2nd International Symposium on DASFAA, 1991.

[4] V. Kumar, A. Burger, "Performace Measurement of Some Main Memory Database Recovery Algorithms," Proc. 7th Int. Conf. Data Engineering, 1991, pp.436-443.