データベース・システム 60－6
（１９８７．７．２０）

大規模ＤＢ／ＤＣシステムにおける履歴情報取得／システム回復高速化の一方式

金居　貞三郎、大町　一彦
（日立　システム開発研究所）

金融オンラインシステムをはじめとして、オンラインデータベースシステムの大規模化が進んでいる。これらのＤＢ／ＤＣシステムでは、大規模化にともない、要求される処理能力が増加し、システムが社会に与える影響も増大している。このため、大規模ＤＢ／ＤＣシステムに対する処理能力向上および信頼性向上の要求が益々強まっている。これらの要求に対し、システムの性能上のボトルネックとなりやすいジャーナル取得の高速化、および障害時のシステム回復時間の短縮が重要な技術課題となる。本報告では、不揮発半導体メモリである半導体ディスクを活用した稼動履歴情報取得処理および障害発生時のシステム回復処理の高速化方式について述べる。

"A High-Speed Logging and Recovery in Very Large Scale Database/Data Communication Systems"
(in Japanese)

by Sadasaburo Kanai and Kazuhiko Oomachi(Systems Development Laboratory, Hitachi,Ltd.,1099 Ohzenji,Asao-ku,Kawasaki-shi,215,Japan)

On-line database systems are widely accepted for a great variety of applications. Among them, banking on-line systems have tended to become increasingly larger in scale in recent years. As becoming larger, they need very high performance and reliability. So it is necessary to get high performance in logging which is one of the bottleneck in system performance and shorten the time of system recovery. This paper summarizes high-speed logging and recovery technique in large scale DB/DC(database/data communication) systems.

1．はじめに

通信ネットワークを介してオンライン処理でデータベースを利用するオンラインデータベースシステムは、銀行・証券等の金融オンラインシステム、航空・列車の座席予約システムをはじめとして、現在広く適用されている。これらのＤＢ／ＤＣ（Data Base/Data Communication）システムは、データ量や端末数の増加、ネットワーク規模の拡大、適用業務の多様化等により益々大規模化しつつある。そして、大規模化に伴い、システムの利用者が増加し、システムが社会に与える影響も増大している。このため、これらの大規模ＤＢ／ＤＣシステムに対する処理能力向上および信頼性向上の要求が益々増大している。代表的な銀行オンラインシステムでは、データ量は１００ＧＢを超え、ピーク時の１時間当たりのシステム全体での処理件数は１００万件に達しようとしている。したがって、ハードウェアならびにソフトウェアの両面からの大幅な処理能力向上が要求されている。ＤＢ／ＤＣシステムでは、データ保証の観点から、ハードウェアやソフトウェアの各種障害に備えてシステムの稼動履歴情報であるジャーナルを磁気ディスクや磁気テープ等の外部記憶装置に取得している。データベースアクセスが並行処理可能であるのに対し、ジャーナルは取得の時系列的な順序がその目的上重要であるため、逐次的に処理せざるをえない。したがって、ジャーナル取得がシステムの性能上のボトルネックになりやすく、処理能力を向上する上でジャーナル取得の高速化が重要な技術課題になっている。また、信頼性に関しては、利用者の増加や適用分野・業務の広がりによりＤＢ／ＤＣシステムは現代社会にとって不可欠の存在になっており、高信頼なシステムの実現が要求されている。例えば、銀行・証券等の金融オンラインシステムに障害が発生した場合、処理中断が長引くとシステムを運用している企業に莫大な損害を与えるのみでなく、社会的に重大な影響を及ぼす恐れがある。このため、各種障害に備え障害によるシステムの停止を防止するとともに、障害発生時には短時間でシステムを正常な状態に回復する必要がある。

一方、近年のハードウェア技術、特に半導体技術の進歩により、半導体ディスク、ディスクキャッシュ、バッファ付磁気テープ等の新しい記憶装置が開発・実用化されつつある。上記のジャーナル取得や障害回復に関する技術課題は、主として計算機システムの入出力系の問題であり、これらの新記憶装置の活用が重要な鍵となる。すなわち、これらの新記憶装置に使用されている半導体メモリの特徴である高速ランダムアクセス性を活かして、ジャーナル取得および障害回復の高速化を実現するアプローチが考えられる。本稿では、上記の新記憶装置のうちの半導体ディスクを主に活用したジャーナル取得／障害回復の高速化方式について述べる。

2．従来のジャーナル取得／障害回復方式の問題点

ジャーナル取得／障害回復の高速化方式について述べる前に、従来のジャーナル取得／障害回復方式の概要および問題点をのべる。

2.1 従来方式の概要

ＤＢ／ＤＣシステムにおいては、システムに発生するハードウェアおよびソフトウェアの各種障害に対しても、データ保全、システムの状態の回復、正常処理の再開を保証する必要がある。すなわち、データベースや主メモリ（本稿では、主記憶装置の他に仮想記憶装置を含めて主メモリと総称する）上の常駐情報を障害発生直前の論理的に一貫した状態に回復する必要がある。このため、ＤＢ／ＤＣシステムでは、システム稼動中の各種稼動履歴情報を磁気ディスクや磁気テープ等の外部記憶装置に取得し、障害発生時にはこれらの稼動履歴情報を用いて上記回復処理を行う。このとき、業務処理の論理的にまとまった単位であるトランザクションは原子的、すなわち、各トランザクションは完了した状態か、全く実行されなかった状態の何れかでなければならない。上記稼動履歴情報のうち、主としてトランザクション処理に関する稼動履歴情報をジャーナルという。ジャーナルは各種障害に備える目的の他に、各種バッチ処理や業務処理内容記録の目的でも取得される。ジャーナルには通常以下の種類がある。

①データベース更新ジャーナル
②主メモリ常駐情報更新ジャーナル
③入力メッセージに関するジャーナル
④出力メッセージを回復するためのジャーナル
⑤トランザクション開始／終了を示すジャーナル

上記の①～⑤のジャーナル以外に、各システム毎に必要に応じて各種のジャーナルが取得される。

ジャーナル以外の稼動履歴情報としては、ジャーナル取得媒体などに関する管理情報等、システム障害時の回復処理で用いるシステム回復管理情報がある。また、システム回復処理に要する時間を短縮するために、システム稼動中に一定間隔でチェックポイントを設け、システム障害時には最新のチェックポイントを回復処理の起点とする。特に、主メモリ常駐情報については、その内容をチェックポイント

時に磁気ディスク等に取得することにより、システム回復時間を短縮させる手法がよく使われる。障害回復の観点からの標準的なＤＢ／ＤＣシステムの構成を図2.1 に示す。

上記のジャーナルの取得方法として、性能上の理由により通常バッファリング技法が用いられる。すなわち、ジャーナルを一旦主メモリ上のバッファに書込み、バッファが満杯になった時点でまとめて取得媒体に書込むことにより、ジャーナル取得効率を向上させている。しかし、ジャーナルの取得目的から、バッファ満杯時以外にトランザクション終了時やデータベース実更新時にも関連するジャーナルを媒体に取得する必要がある。したがって、単位時間当たりのトランザクション件数が増加するにつれ、高速にジャーナルを取得することが要求される。このため、銀行オンラインシステム等の高トラフィックなシステムでは、何らかの方法によりジャーナル取得を高速化する必要がある。

システム障害時には、ホットスタンバイ機等の別系あるいは回復後のＣＰＵ／主メモリにおいて、システムを再開始する。このとき、データベースや主メモリ常駐情報等の各種データを障害発生前の状態に回復する必要がある。回復処理の概要は以下のとおりである（図2.2 ）。

①最終のチェックポイント時に磁気ディスク等の媒体に取得した主メモリ常駐情報を媒体から読出して主メモリ内に展開する。

②最終のチェックポイント時以降のジャーナルをジャーナル取得媒体から読出し、データベースおよび主メモリ常駐情報を回復する。また、出力メッ

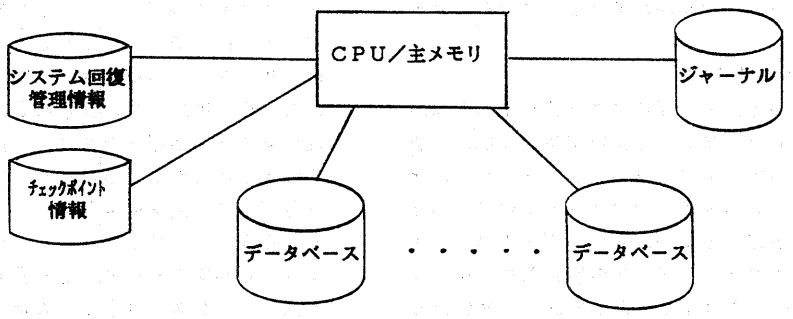

図2.1 従来方式でのＤＢ／ＤＣシステムの構成

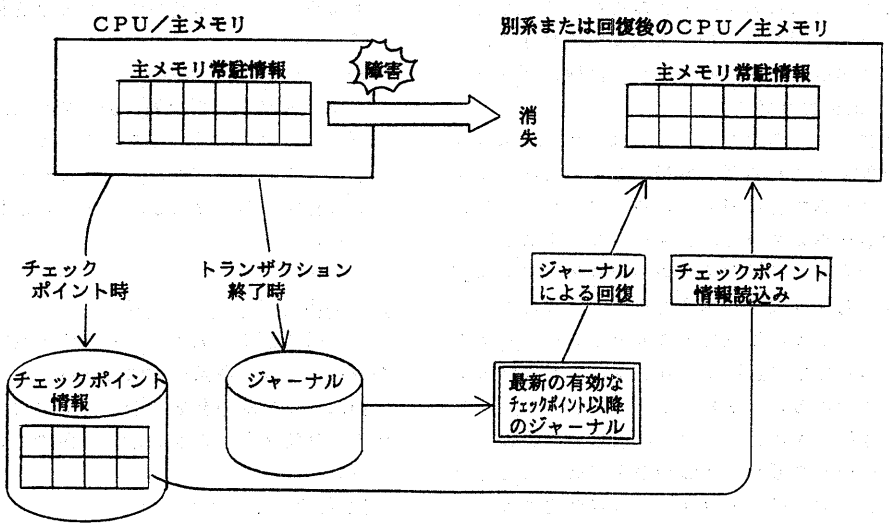

図2.2 従来のシステム回復方法

セージに関するジャーナルを用いて、出力メッセージおよびメッセージキューを回復する。

2.2 従来方式の問題点

上記の方式には以下の問題点がある。

(1)ジャーナル取得能力

全てのジャーナルを同一の媒体に取得し、しかもトランザクションの終了と同期して取得する。このため、磁気ディスク等の場合には機械的動作に要する時間が占める割合が大きく、ジャーナル取得能力の向上に限界がある。この問題点に対し、例えばジャーナルを複数台の磁気ディスクに並行に分散して取得する等の方法により性能向上を図ることが考えられる。しかし、この方法ではトラフィックに比例して磁気ディスクの台数を増やす必要があり、システム構成上の限界がある。また、次に述べるジャーナル蓄積量に関しては改善されていない。

(2)ジャーナル蓄積量

ジャーナルは、その種別により保存すべき期間が異なる。トランザクション障害時およびシステム障害時の回復処理でのみ必要なジャーナルは比較的短期間（基本的には次のチェックポイントまで）保存するだけでよい。一方、データベース更新ジャーナルはデータベースの新たなバックアップコピーが作成されるまで必要であり、通常数日ないしは数週間の保存が必要である。業務処理内容記録の目的で取得するジャーナルはさらに長く、数年以上の保存が必要とされる場合がある。従来方式では、これらのジャーナルをひっくるめて取得するため、ジャーナル蓄積量が膨大になる、ジャーナルの編集処理が必要、等の問題点がある。

(3)システム回復時間

大規模ＤＢ／ＤＣシステムでは、ホットスタンバイ機能による待機系計算機への即時切替等により、システム障害に備えている。このため、システム回復処理において、ジャーナルによるデータベースや主メモリ常駐情報等の各種データの回復に要する時間が大きな割合を占めるようになっている。システム回復処理で必要とするジャーナルは、基本的には最新のチェックポイント以降のジャーナルである。したがって、チェックポイントの間隔を短くすることにより回復処理で読むジャーナルの量を削減し、回復時間を短縮することができる。一方、近年の主メモリ大容量化により主メモリ常駐データの量を大幅に増加させることが可能になり、システムの全体性能向上を目的として、データベースのアクセス頻度の高い部分を主メモリに常駐化させる傾向にある。しかし、主メモリ常駐情報の増加はチェックポイント時の処理オーバヘッド増大を招き、チェックポイントの間隔を長くする必要がある。したがって、主メモリ常駐データ量の増大は、障害時のシステム回復処理に要する時間を長引かせる結果を招く。

3．ジャーナル取得／障害回復の高速化方式

3.1 基本的な考え方

(1)半導体ディスクの活用

半導体ディスクを用い、その特性である高速ランダムアクセス性を活かす。半導体ディスクは、磁気ディスクの駆動装置部を半導体メモリに置き換えた外部記憶装置であり、電源断に備えて内臓ディスクおよびバッテリにより不揮発化を実現している。磁気ディスクに対して、平均アクセス時間は順アクセスの場合は約1/10、ランダムアクセスの場合には約1/30、転送速度も現状では２倍程度の高速転送が可能であり、大幅に入出力時間を短縮することができる。ただ、記憶容量は百ＭＢ～数百ＭＢと近年の大型ディスクと比べて少なく、しかもかなり高価である。したがって、一時的に保存が必要なデータや特定の重要データで、しかも高速アクセスが要求される場合に適している。

(2)ジャーナルの用途別取得

ジャーナルをシステム回復用、記録用、データベース障害回復用等の用途別に分類し、システム回復用ジャーナルの媒体として半導体ディスクを用いる。

(3)データベースのキャッシング

データベースの比較的最近の更新結果を半導体ディスクに一時的に保存する。また、主メモリ常駐データに対しては、そのコピーを半導体ディスクに保持することによりシステム回復時間を短縮する。

3.2 機能別ジャーナル方式

本方式の主な目的は、ジャーナル取得高速化およびジャーナル蓄積量の削減である。

図3.1 に方式の概要を示す。本方式では、ジャーナルを回復用および記録用の２種類に分類して取得する。

(1)回復用ジャーナル

回復用ジャーナルは、システム障害およびトランザクション障害からの回復処理で用いるジャー

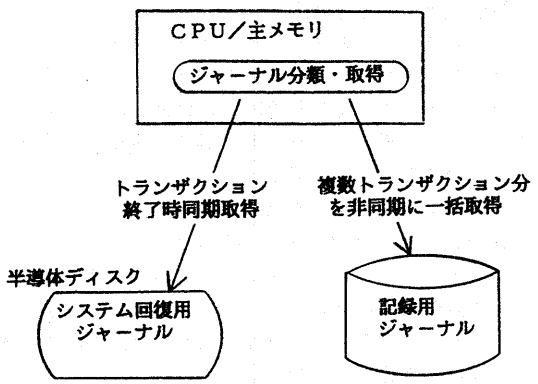

図3.1 機能別ジャーナル方式の概要

ナルである。回復用ジャーナルは、トランザクションの終了と同期して取得する必要がある。したがって、トラフィックに比例して取得回数が増加するため高速に取得する必要があり、取得媒体として半導体ディスクを用いる。半導体ディスクは高価格かつ比較的小容量であるが、システム回復用のジャーナルは基本的には最新のチェックポイント時以降の情報があればよいので、ラップアラウンド方式で繰返し使用することができる。また、回復用ジャーナルは、媒体障害の発生に備え多重化することも信頼性確保に有効である。

(2)記録用ジャーナル

データベース障害の回復処理で用いる情報、各種バッチ処理で用いる情報、業務上の処理内容を記録するための情報、等をまとめて記録用ジャーナルとして取得する。記録用ジャーナルは、トランザクション終了と同期して取得する必要がないため、磁気ディスクや磁気テープ等に複数トランザクション分をまとめて取得する。これらの装置は、磁気ディスクの回転待ち時間等機械的動作に要する時間が大きいという問題点が存在するが、まとめ書きにより取得効率を向上させる。記録用ジャーナルの媒体は複数用意し、１つの媒体が満杯になると、次の媒体を使用する。記録用ジャーナルは、最終的にはオフライン業務で使用できるようにする。

3.3 ＤＢキャッシング方式

本方式は、システム障害時の回復時間を短縮することを主な目的とする。

ＤＢキャッシング方式は、機能別ジャーナル方式を前提とする。概要を以下に示す（図3.2 ）。

(1)ＤＢキャッシュ

データベース全体あるいはその一部に対して、トランザクションによって更新されたページを一時的に格納するための領域（ＤＢキャッシュ）を半導体ディスク上に設ける。

(2)更新ページ書込み

トランザクション終了時、ＤＢキャッシング対象の部分についてはデータベース更新ジャーナルを取得するかわりに半導体ディスク上のＤＢキャッシュに更新ページを書込む。このとき、ページ更新中のシステム障害に備えて、シャドウ方式[4]で更新する。シャドウ方式におけるページマップはＤＢキャッシュ同様半導体ディスク上に設けるとともにその更新履歴情報を回復用ジャーナルの

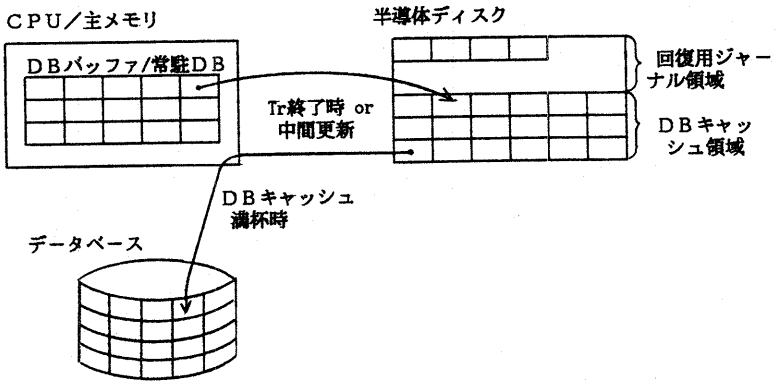

図3.2 ＤＢキャッシュ方式の概要

一種のページマップ更新ジャーナルとして半導体ディスクに取得し、実際の更新はチェックポイント時に行う。なお、トランザクション途中でのデータベース更新の場合も同様に行う（ただし、更新ページはトランザクション終了まで有効化しない）。

(3)データベース実更新

ＤＢキャッシュ領域は、ＬＲＵ等のアルゴリズムにより管理し、あふれたページに対してはデータベースを実更新する。このとき、データベースの該ページを格納するブロックを直接更新する。

(4)主メモリ常駐データベース

主メモリ常駐データベースは、必ずＤＢキャッシングの対象とする。主メモリ常駐データベースは、システム開始時に全ページを磁気ディスクからＤＢキャッシュにロードし、主メモリからのアンロード時にＤＢキャッシュから磁気ディスクにアンロードする。また、各ページはＤＢキャッシュからの追い出し対象外とする。

(5)システム回復

システム障害発生時には、ＤＢキャッシング対象のデータベースに対し、以下の手順で回復処理を行う。

①ページマップの回復

回復用ジャーナル中のページマップ更新ジャーナルを用いてページマップを回復する、

②主メモリ常駐データベースの回復

主メモリ常駐データベースの各ページの内容を、①で回復したページマップを用いてＤＢキャッシュの該当する箇所から読込み、主メモリ内に展開する。

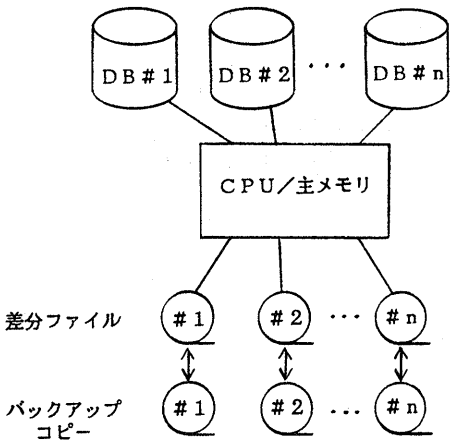

図3.3 差分ファイル方式の概要

3.4 差分ファイル方式

差分ファイル方式はＤＢキャッシング方式を用いた時のデータベース障害対策であり、主な目的はデータベース障害時の回復時間の短縮である。差分ファイルは、データベースの更新後情報をページ単位に取得するファイルである。

以下に方式の概要を述べる。

(1)差分ファイルの取得（図3.3 ）

データベースのサブグループ毎に差分ファイルを定義する。そして、ＤＢキャッシング対象のデータベースのページ単位に以下の契機で順次更新内容を差分ファイルに書出す。

(a)一般のデータベース

ＤＢキャッシュからあふれたページをデータベースにデステージングする際、同一内容を差分ファイルに出力する。

(b)主メモリ常駐データベース

アンロード時、ＤＢキャッシュ内の当該データベースの各ページの内容を差分ファイルに出力する。

(2)データベース障害回復

データベース障害時には、障害の発生したデータベースのサブグループに対応するバックアップコピーをロードし、差分ファイル中の更新結果をかぶせる。

(3)バックアップコピーのアップデート（図3.4 ）

満杯になった差分ファイルはオフラインとなる。これをページ番号でソートし、バックアップコピーとマージする。これにより、データベースへのアクセスを中断することなく、バックアップコピーをアップデートすることができる。また、差分ファイル媒体を再使用可能とすることができる。

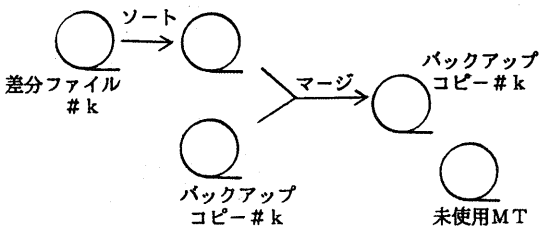

図3.4 バックアップコピーのアップデート

3.5 方式の評価

ジャーナル取得能力、ジャーナル蓄積量、およびシステム回復時間について、方式を評価する。

(1)ジャーナル取得能力

機能別ジャーナル方式では、ジャーナル取得に関しては記録用ジャーナルがボトルネックとなる。何トランザクション分のジャーナルを記録用ジャーナルにまとめて取得するかにより上記方式の効果は異なるが、１０～２０トランザクション分をまとめて処理すれば、ジャーナル取得量削減による効果を無視しても従来方式と比較して５～１０倍の性能向上が期待できる。ＤＢキャッシュ方式および差分ファイル方式と組合せることにより、さらに３０～５０％程度の性能向上が期待できる。

(2)ジャーナル蓄積量

機能別ジャーナルによる記録用ジャーナル削減効果、ＤＢキャッシュおよび差分ファイル方式によるデータベース更新ジャーナルの不要化により、記録用ジャーナルの蓄積量を約３０～５０％削減することができる。

(3)システム回復時間の短縮

機能別ジャーナル方式においては、システム回復用ジャーナルの媒体が半導体ディスクとなるので、従来方式に対して１/２～１/３程度にシステム回復時間を短縮することができる。さらに、ＤＢキャッシュ方式および差分ファイル方式と組合せることにより、約１/３～１/１０にシステム回復時間を短縮することができる。

４．おわりに

大規模ＤＢ／ＤＣシステムにおけるジャーナル取得／障害回復の高速化方式について報告した。方式の特徴は、半導体ディスクの高速ランダムアクセス性を活かし、①ジャーナルを用途別にシステム回復用と記録用に分類し、各々に適した取得媒体、取得方式を用いること、②データベース（特に主メモリ常駐データベース）の更新内容を一時的に格納する領域を半導体ディスク上に設け、データベース更新ジャーナルの不要化およびシステム回復時の主メモリ常駐データベースの高速回復を可能とすること、である。本方式により、①ジャーナル取得高速化、②ジャーナル蓄積量削減、③システム回復時間短縮、④データベース回復時間短縮、を実現することができる。

５．参考文献

(1)J.N.Gray.“Notes on Data Base Operating Systems.” R.Bayers,R.M.Graham,and G.Seegmuller (eds.),Operating Systems:An Advanced Course. Springer-Verlag(1978).

(2)J.S.M.Verhofstad.“Recovery Techniques for Database Systems.” ACM Comp.Surv.Vol.10,No.2 (June 1978).

(3)K.Elhardt,and R.Bayer,“A Database Cache for High Performance and Fast Restart in Database Systems.” ACM TODS Vol.9,No.4(December 1984).

(4)R.A.Lorie.“Physical Integrity in a Large Segmented Database.” ACM TODS Vol.2, No.1 (March 1977).