(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第4378335号
(P4378335)

(45)発行日　平成21年12月2日(2009.12.2)　(24)登録日　平成21年9月18日(2009.9.18)

(51)Int.Cl.		F I		
G06F 3/06	(2006.01)	G06F	3/06	301Z
G06F 12/00	(2006.01)	G06F	12/00	518M
G06F 13/10	(2006.01)	G06F	13/10	340A
		G06F	12/00	514A

請求項の数 18　(全 14 頁)

(21)出願番号　特願2005-262155 (P2005-262155)
(22)出願日　平成17年9月9日 (2005.9.9)
(65)公開番号　特開2007-72975 (P2007-72975A)
(43)公開日　平成19年3月22日 (2007.3.22)
審査請求日　平成20年1月21日 (2008.1.21)

早期審査対象出願

(73)特許権者　390009531
インターナショナル・ビジネス・マシーンズ・コーポレーション
ＩＮＴＥＲＮＡＴＩＯＮＡＬ　ＢＵＳＩＮＥＳＳ　ＭＡＳＣＨＩＮＥＳ　ＣＯＲＰＯＲＡＴＩＯＮ
アメリカ合衆国１０５０４　ニューヨーク州　アーモンク　ニュー　オーチャード　ロード
(74)代理人　100086243
弁理士　坂口　博
(74)代理人　100091568
弁理士　市位　嘉宏
(74)代理人　100108501
弁理士　上野　剛史

最終頁に続く

(54)【発明の名称】ディスクへのトランザクション・データ書き込みの方式を動的に切り替える装置、切り替える方法、及び切り替えるプログラム

(57)【特許請求の範囲】
【請求項１】
トランザクション・データをディスクに書き込む装置であって、
前記トランザクション・データを格納するディスク書き込み用キューとメモリ書き込み用キューに関連づける情報を管理するためのキュー管理テーブルを有するメモリであって、前記ディスク書き込み用キューと前記メモリ書き込み用キューがデータの耐久性保証方法を切り替える単位であるリージョンごとに設けられているものと、
先行トランザクションのディスクへの書き込み処理が終了する前に次のトランザクションが発生する頻度として定義されるディスクへの書き込み負荷が所定の閾値を超えたことを条件として前記キュー管理テーブルに前記メモリ書き込み用キューに関連づける情報を登録し、前記負荷が閾値を下回ることを条件として前記キュー管理テーブルから前記メモリ書き込み用キューに関連づける情報を削除する手段と、
前記キュー管理テーブルに前記ディスク書き込み用キューに関連づける情報のみが登録されている場合、前記ディスク書き込み用キューに格納された前記トランザクション・データを受け取り、冗長化メモリを使用せずにディスクに書き込む手段と、
前記キュー管理テーブルに前記メモリ書き込み用キューに関連づける情報が登録されている場合、前記メモリ書き込み用キューに格納された前記トランザクション・データを受け取り冗長化メモリに書き込む手段と、
を有するトランザクション・データをディスクに書き込む装置。
【請求項２】

前記キュー管理テーブルに前記メモリ書き込み用キューに関連づける情報が登録されている場合、前記トランザクション・データを前記冗長化メモリに書き込むために、前記メモリに書き込む手段にスレッドを割り当てる手段をさらに含む請求項１に記載のトランザクション・データをディスクに書き込む装置。

【請求項３】

前記メモリに書き込む手段が、前記メモリ書き込み用キューに格納された前記トランザクション・データをマルチキャストで複数のメモリに書き込むことを特徴とする請求項１または２に記載のトランザクション・データをディスクに書き込む装置。

【請求項４】

前記メモリに書き込む手段が、マルチキャストの宛先からの応答により所定の数のメモリへのコピーが完了したことが判明したら、データ書き込みの完了通知を返す手段をさらに含む、請求項３に記載のトランザクション・データをディスクに書き込む装置。

【請求項５】

アプリケーションのプロセスの障害が発生した場合にも、前記負荷が閾値を超えたとする、請求項１に記載のトランザクション・データをディスクに書き込む装置。

【請求項６】

データベースの一時的な応答遅延が発生した場合にも、前記負荷が閾値を超えたものとする、請求項１に記載のトランザクション・データをディスクに書き込む装置。

【請求項７】

データベースに障害が発生した場合にも、前記負荷が閾値を超えたものとする、請求項１に記載のディスクへの書き込み方式を変更する装置。

【請求項８】

ディスクの障害が発生した場合にも、前記負荷が閾値を超えたものとする、請求項１に記載のトランザクション・データをディスクに書き込む装置。

【請求項９】

前記管理する手段が、さらに、データの耐久性保証方法を切り替える場合、前記ディスク書き込み用キューに排他ロックをかける手段を含む請求項１に記載のトランザクション・データをディスクに書き込む装置。

【請求項１０】

ライト・スルー方式とライト・ビハインド方式でトランザクション・データのディスクへの書き込み方式を変更する方法であって、

先行トランザクションのディスクへの書き込み処理が終了する前に次のトランザクションが発生する頻度として定義されるディスクへの書き込み負荷が所定の閾値より下から上に変わったことに応答して、キュー・マネージャがディスク書き込み用キューとメモリ書き込み用キューに関連づける情報をキュー管理テーブルに登録するステップであって、前記ディスク書き込み用キューと前記メモリ書き込み用キューがデータの耐久性保証方法を切り替える単位であるリージョンごとに設けられているものと、

前記負荷が前記閾値より上から下に変わったことに応答して、前記キュー・マネージャが前記キュー管理テーブルからメモリ書き込み用キューに関連づける情報を削除するステップと、

前記キュー管理テーブルに前記ディスク書き込み用キューに関連づける情報のみが登録されている場合、ディスク用データ記録手段が前記ディスク書き込み用キューに格納された前記トランザクション・データを受け取り、冗長化メモリを使用せずにディスクに書き込むステップと、

前記キュー管理テーブルに前記メモリ書き込み用キューに関連づける情報が登録されている場合、メモリ用データ記録手段が前記メモリ書き込み用キューに格納された前記トランザクション・データを受け取り冗長化メモリに書き込むステップと、

を含むディスクへの書き込み方式を変更する方法。

【請求項１１】

前記キュー管理テーブルに前記メモリ書き込み用キューに関連づける情報が登録されて

いる場合、前記トランザクション・データを冗長化メモリに書き込むために、書き込みタイミング制御手段が前記メモリに書き込む手段にスレッドを割り当てるステップをさらに含む請求項１０に記載のディスクへの書き込み方式を変更する方法。
【請求項１２】
前記登録するステップの前に、登録する前記キューに対応するリージョンのディスク書き込み用キューに前記キュー・マネージャが排他ロックをかけるステップを含む請求項１０に記載のディスクへの書き込み方式を変更する方法。
【請求項１３】
前記削除するステップの前に、削除する前記キューに対応するリージョンのディスク書き込み用キューに前記キュー・マネージャが排他ロックをかけるステップを含む請求項１０に記載のディスクへの書き込み方式を変更する方法。
【請求項１４】
アプリケーションのプロセスのフェイルオーバーが発生した場合にも、前記キュー・マネージャは、前記負荷が閾値を超えたとみなして処理する、請求項１０に記載のディスクへの書き込み方式を変更する方法。
【請求項１５】
データベースの一時的な応答遅延が発生した場合にも、前記キュー・マネージャは、前記負荷が閾値を超えたものとみなして処理する、請求項１０記載のディスクへの書き込み方式を変更する方法。
【請求項１６】
データベースに障害が発生した場合にも、前記キュー・マネージャは、前記負荷が閾値を超えたものとみなして処理する、請求項１０に記載のディスクへの書き込み方式を変更する方法。
【請求項１７】
ディスクの障害が発生した場合にも、前記キュー・マネージャは、前記負荷が閾値を超えたものとみなして処理する、請求項１０に記載のディスクへの書き込み方式を変更する方法。
【請求項１８】
請求項１０から１７のいずれかに記載の方法を実現するステップを、コンピュータに実行させるためのプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、オンライン・トランザクション処理（ＯＬＴＰ）等において、動的な負荷の変動、プロセスやハードウェアの障害に応じて、ディスクへのトランザクション・データの書き込みを、ライト・スルー方式とライト・ビハインド方式で動的に切り替えることが可能なトランザクション・データをディスクに書き込む装置、書き込む方法、及び書き込むプログラムに関する。
【背景技術】
【０００２】
証券取引システムのようなオンライン・トランザクション処理アプリケーションでは、コミット済みのデータが障害によって失われることは許されない。通常、障害に対するデータの耐久性を保証する方法として、コミット時にディスクに直接データを書き込むライト・スルー方式が用いられる。
【０００３】
従来は、データのディスク書き込みは一般的にレイテンシが大きくスループットも低いため、それを改善する手法としてライト・ビハインド方式が使われてきた。図１は、ライト・ビハインド方式によるデータの書き込み処理を簡単に説明している。ライト・ビハインド方式では、まず、コミット時に異なるノードやプロセス上のメモリ（遠隔メモリ）にデータを一時的にコピーして冗長化しておく（１０１）。コピー先の遠隔メモリの場所や

構成などはどのような障害に耐えたいかによって変わる。コミット完了後、ローカルプロセスのメモリ（ローカルメモリ）上のデータを非同期にディスクに書き込む（１０２）。もし、ディスクにデータが書き込まれる前に障害が発生しても、遠隔メモリ上からデータを復旧できる（１０３）。ディスクに書き込みが完了したらローカルメモリおよび遠隔メモリからデータを削除する（１０４）。メモリはディスクに比べてデータの書き込み速度が高速なため、アプリケーションはレイテンシを小さくしスループットを高められる。

【０００４】

しかし、ライト・スルー方式とライト・ビハインド方式を切り替えるというデータの耐久性保証はアプリケーションの起動前に静的に設定するのが一般的であり、実行時の負荷の変化、プロセスやハードウェアの障害に対応して方式を動的に変更することはできなかった。

【０００５】

また、ライト・スルー方式とライト・ビハインド方式を選択する特許文献１では、マルチプロセッサシステムにおいて、複数のプロセッサが共有する領域へは、ＣＰＵのライトアクセスは競合を防止するためライト・スルー方式でアクセスし、１つのＣＰＵに固有のアドレスへのアクセスには、ライト・ビハインド方式でアクセスするのみで、アプリケーションの状況に応じて動的に変更することはできなかった。

【０００６】

【特許文献１】特開平９－９３２６５号公報

【発明の開示】

【発明が解決しようとする課題】

【０００７】

コミット時のデータの保存先を、ディスクではなく遠隔メモリにすることによって、トランザクション・データの耐久性を保証しつつレイテンシを向上できる。しかし、ディスクに比べてメモリは高価であり、結果として容量が小さい。そのため、ライト・ビハインド方式では、ある一定量以上のコミット済みデータが一定時間以内にメモリに蓄積すると、メモリの使用容量が限界に達してしまい、それ以上にコミット済みデータを保存することができなくなる。その結果、メモリが溢れてしまう。一方で、これを防ぐためにトランザクション・データのディスク書き込みを待つと、ディスク書き込み処理がボトルネックとなってしまう。

【課題を解決するための手段】

【０００８】

上記課題を解決するため、本発明においては、ライト・スルー方式とライト・ビハインド方式でトランザクション・データのディスクへの書き込み方式を動的に変更することが可能な、トランザクション・データをディスクに書き込む装置を提供する。該装置は、前記トランザクション・データを格納するディスク書き込み用キューとメモリ書き込み用キューとを管理するためのキュー管理テーブルを有するメモリを備える。また、該装置は、負荷が所定の閾値を超えたことを条件として前記キュー管理テーブルに前記メモリ書き込み用キューを登録し、前記負荷が閾値を下回ることを条件として前記キュー管理テーブルから前記メモリ書き込み用キューを削除する手段を有する。そして、該装置は、前記ディスク書き込み用キューに格納された前記トランザクション・データを受け取りディスクに書き込む手段と、前記メモリ書き込み用キューに格納された前記トランザクション・データを受け取り冗長化メモリに書き込む手段とを有する。該装置により、負荷に応じて、ディスクへのデータ書き込みをライト・スルー方式とライト・ビハインド方式で動的に切り替えることが可能になり、システムのリソースを有効に活用することができる。

【発明を実施するための最良の形態】

【０００９】

以下、発明の実施の形態を通じて本発明を説明するが、以下の実施形態は特許請求の範囲にかかる発明を限定するものではなく、また実施形態の中で説明されている特徴の組み合わせは、発明の内容を理解しやすくするためのものもあり、その全てが発明の解決手段

に必須であるとは限らない。

【００１０】

以下の実施の形態では、主に方法またはシステムについて説明するが、当業者であれば明らかなとおり、本発明はコンピュータで使用可能なプログラムとしても実施できる。したがって、本発明は、ハードウェアとしての実施形態、ソフトウェアとしての実施形態またはソフトウェアとハードウェアとの組合せの実施形態をとることができる。プログラムは、ハードディスク、ＣＤ－ＲＯＭ、光記憶装置または磁気記憶装置等の任意のコンピュータ可読媒体に記録できる。

【００１１】

図２は、負荷の変動、プロセスやハードウェアの障害に応じて、ディスクへのトランザクション・データの書き込みを、ライト・スルー方式とライト・ビハインド方式で動的に切り替えるシステムが作動するハードウェア構成２００の概略を示している。中央演算処理装置であるＣＰＵ２０１は各種オペレーティング・システムの制御下で様々なプログラムを実行する。ＣＰＵ２０１はバス２０２を介して、メモリ２０３、ディスク２０４、ディスプレイ・アダプタ２０５、ユーザ・インタフェース２０６およびネットワーク・インタフェース２０７と相互接続される。ディスク２０４には、コンピュータを本発明を実現するためのシステムとして機能させるためのソフトウェア、オペレーティング・システムおよびデータベース・システム等が含まれる。

【００１２】

ユーザ・インタフェース２０６を通して、キーボード２０８およびマウス２０９に接続され、ディスプレイ・アダプタ２０５を通してディスプレイ装置２０８に接続され、ネットワーク・インタフェース２０７を通してネットワーク２１１に接続される。ネットワーク２１１からさらに冗長化用ノード／プロセス２１２に接続される。冗長化用ノード／プロセスは、遠隔メモリ２１３を有する。この遠隔メモリ２１３は、ライト・ビハインド方式でトランザクション・データの耐久性を保証するため、コミット後でディスク２０４に反映されていないメモリ２０３のトランザクション・データを冗長化して保持する。このハード構成２００は、コンピュータ・システム、バス配置およびネットワーク接続の一実施形態の例にすぎず、本発明の特徴は、さまざまなシステム構成で、同一の構成要素を複数有する形態で、または、ネットワーク上にさらに分散された形態でも実現することができる。

【００１３】

図３は、ディスクへのトランザクション・データの書き込みを、ライト・スルー方式とライト・ビハインド方式で動的に切り替える装置の機能および処理されるデータを、概念的に表したものである。ここで表現される機能ブロックおよびデータの構成は、さらに細かい単位で表現可能であり、また、もっと大きい単位の機能ブロックでも表現可能である。これら機能ブロックの表現単位の違いは、発明を実施する際の本質的問題ではないことは、当業者には明らかである。ディスク３０１は、ハード・ディスクなどの一般にデータ書き込み速度が遅い装置で、データベース等のデータが最終的に記録される装置である。メモリ３０２は、ディスク３０１に比べて、書き込み速度が高速であり、ライト・ビハインド方式での書き込みの場合に、一時的にデータを冗長化するのに使用される。メモリ３０２は、障害に耐えるために使用されるので、通常は異なるノードやプロセス上に配置される。リージョン３０３は、データの耐久性保証方法を切り替える単位であり、あるトランザクションでアクセスするデータの集合は、必ず特定のリージョンに含まれる。リージョンの定義は、データ全体を一つのリージョンとすることもできるし、また、証券取引システムなどでは、会社毎の取引データをそれぞれのリージョンとして定義することもできる。ちなみにトランザクションはリージョンの境界を越えてデータにアクセスすることはできない。

【００１４】

キュー管理テーブル３０４は、トランザクションのコミット毎のデータが入ったキューを管理するテーブルである。キューはリージョンとデータ記録手段の組み合わせごとに用

意される。図３においては、リージョンａおよびリージョンｂと、ディスク用データ記録手段３０５およびメモリ用データ記録手段３０６の組み合わせとなるので、キューはディスク用キュー｛Ｄ、ａ｝と｛Ｄ、ｂ｝、およびメモリ用キュー｛Ｍ、ａ｝と｛Ｍ、ｂ｝の４つとなる。ディスク用キュー｛Ｄ、ａ｝と｛Ｄ、ｂ｝は、ディスク３０１にデータを書き込むためのトランザクション・データを管理し、メモリ用キュー｛Ｍ、ａ｝と｛Ｍ、ｂ｝は、メモリ３０２にデータを書き込むためのトランザクション・データを管理する。キュー・マネージャ３０７が、キュー管理テーブル３０４に対し、キューの登録、削除および変更等を行う。キュー・マネージャ３０７の機能は、書き込みタイミング制御手段３０８などの一部として取り込まれていてもいい。

【００１５】

ディスク用データ記録手段３０５は、キュー｛Ｄ、ａ｝およびキュー｛Ｄ、ｂ｝からトランザクション・データを受け取り、ディスク３０１に書き込む。メモリ用データ記録手段３０６は、キュー｛Ｍ、ａ｝およびキュー｛Ｍ、ｂ｝からトランザクション・データを受け取り、メモリ３０２に書き込む。キューにはトランザクション・データそのもの、またはトランザクション・データが記録されている場所を示すポインタが入っていて、これらに従って、データ記録手段３０５および３０６は、データの書き込み処理を行う。キュー｛Ｍ、ａ｝およびキュー｛Ｍ、ｂ｝は、通常はシステムのローカルのメモリ上（図２のメモリ２０３）にあるが、他のシステムに存在することも可能である。書き込みタイミング制御手段３０８は、ディスク用データ記録手段３０５およびメモリ用データ記録手段３０６にスレッドを割り当て、データ記録手段３０５およびメモリ用データ記録手段３０６がトランザクション・データを受取って書き込み処理を行う。また、書き込みタイミング制御手段３０８は、スレッドのプールを管理し、このプールから必要に応じてデータ記録手段３０５およびメモリ用データ記録手段３０６にスレッドを割り当て、不要になったスレッドを再びプールに戻す。そして、キューに対しては、それぞれのプロセスがデータの読み取りと書き込みを別々に行うので、データの順番を間違えて処理するのを防止するために、それぞれのプロセスは、キューに対して排他ロックをかける。

【００１６】

本発明では、メモリのリソースを最大限に活用するため、ディスクへのトランザクション・データの書き込み処理では、通常の負荷の場合はメモリを使わないようにしておき、低負荷状態から高負荷状態に移行する場合にメモリを効率的に使って一度に大量のデータを処理できるようにする。すなわち、通常の負荷の場合は、メモリへのデータ蓄積を行なわないライト・スルー方式でデータ書き込みを処理し、メモリの空き容量を最大に保つことで、高負荷時にライト・ビハインド方式に移行した際のメモリに蓄積できるデータ量を多くする。また、高負荷状態から平均的な負荷に戻る場合は、次に来る高負荷状態に耐えるため、ライト・ビハインド方式からライト・スルー方式に戻してメモリの空き容量を多くし、次回のライト・ビハインド方式でメモリに蓄積できるデータ量を多くできるようにする。

【００１７】

また、ライト・スルー方式でディスクに障害が発生した場合、アプリケーションを停止させないようにするために、一時的にライト・ビハインド方式に切り替えたい場合がある。逆に、ライト・ビハインド方式で、遠隔メモリに障害が発生し、同じくアプリケーションを停止させないようにするために、ライト・スルー方式に切り替えなければならない場合もある。例えば、ディスク、それを管理するＯＳ（オペレーティング・システム）やデータベースが何らかの原因で一時的に応答が遅くなる応答遅延が発生する場合がある。データベース・プロセスに障害が発生した場合、データベースが一時的に応答できない状況もある。同様に、ライト・ビハインド方式で、遠隔メモリを管理するＯＳやプロセスの負荷が高くなり応答しなくなる場合がある。これらの障害は短時間で回復することが多いので、ある一定時間以上遠隔メモリやディスクが応答しない場合は、ライト・スルー方式とライト・ビハインド方式を相互に切り替えられると、アプリケーションに対して応答時間を保証できる。

【００１８】
さらに、ライト・ビハインド方式でアプリケーションが動作中にプロセスに障害が発生し、別プロセスが引き継ぐフェイルオーバーの場合、遠隔メモリとディスク上のデータの差分をディスクに反映する必要がある。しかし、引継ぎ後のプロセスが、差分の反映が完了するのを待ってからでないと新しいトランザクションを実行できないとすると、アプリケーションの可用性が低下してしまう。そこで、アプリケーションの可用性を重視する場合、差分を反映している間はライト・ビハインド方式でディスクに書き込みつつ、同時にアプリケーションの処理を即時に開始できることが望ましい。この場合、差分を反映し終え、さらにフェイルオーバー後のトランザクションによる新たなデータも反映し終えるまでは、ライト・ビハインド方式を使うことになる。さらに、これらの処理が完了した後は、必要に応じてライト・スルー方式に切り替えられることが望ましい。なお、フェイルオーバーも、一時的に応答が遅くなる応答遅延と考えることもできる。

【００１９】
さらに、負荷はある特定の種類のデータに集中する場合がある。例えば、証券取引アプリケーションでは、特定の株銘柄の売買が短期間に集中する場合がある。そのため、ライト・スルー方式とライト・ビハインド方式の切り替えは、株取引の銘柄などのデータの種類ごと（例えば、リージョンごと）に行なえるようにしておく必要がある。

【００２０】
図４に、図３の構成を用いて、ライト・スルー方式を実現した状態の一例を示す。ここでは、リージョンａとリージョンｂ（４０３）のデータは、ディスク用データ記録手段４０５によって書き込みが行われる。ライト・スルー方式でデータ書き込みをするため、メモリ用データ記録手段４０６は使用されない。キュー管理テーブル４０４には、ディスク用データ記録手段４０５が利用するリージョンａとリージョンｂについてのキュー｛Ｄ、ａ｝およびキュー｛Ｄ、ｂ｝が登録される。ここでは、リージョンａのトランザクションがコミットされる場合を例として説明する。データは関連するキュー｛Ｄ、ａ｝に格納される。ライト・スルー方式では、コミット完了前にディスクにデータが書き込まれる必要があるため、書き込みタイミング制御手段４０８は、ディスク用データ記録手段４０５にスレッドを割り当て、データを渡し、ディスク用データ記録手段４０５が書き込みを行なう。

【００２１】
ディスク用データ記録手段４０５によるデータの書き込みが完了するとコミットが完了する。リージョンの異なるトランザクションは互いに関連がないため、書き込みタイミング制御手段４０８は、データ記録手段に対してキューごとにスレッドを割り当てる。なお、図４ではライト・スルー方式でデータ書き込みを行うので、キュー管理テーブル４０４を経由しないでディスク用データ記録手段４０５がデータを受取ることも可能である。しかし、ライト・スルー方式でのデータ書き込みの際にも、データがキュー管理テーブル４０４を経由するようにすれば、書き込みタイミング制御手段４０８や、ディスク用データ記録手段４０５の機能は、ライト・ビハインド方式の際の機能と共通化できるため、機能構成を単純なものとすることが可能となる。また、必ずキューを経由してデータ書き込み処理をすれば、ライト・ビハインド方式からライト・スルー方式に切り替えた場合にも、データ順にキューの残存データの書き込み処理をしやすくなる。

【００２２】
図５に、図３の構成を用いて、ライト・ビハインド方式を実現した状態の一例を示す。ここでは、リージョンａとリージョンｂ（５０３）のデータは、それぞれディスク用データ記録手段５０５によりディスク５０１に書き込まれ、メモリ用データ記録手段５０６によりメモリ５０２に書き込まれる。従って、キュー管理テーブルには、ディスク用データ記録手段５０５に渡すデータを管理するためのキュー｛Ｄ、ａ｝およびキュー｛Ｄ、ｂ｝と、メモリ用データ記録手段５０６に渡すデータを管理するためのキュー｛Ｍ、ａ｝およびキュー｛Ｍ、ｂ｝の４つが登録される。ライト・ビハインド方式でのデータ書き込みでは、コミット処理の前に冗長化用の遠隔メモリ（メモリ５０２）に書き込む必要がある。

書き込みタイミング制御手段５０８は、メモリ用データ記録手段５０６にスレッドを割り当てる。

【００２３】

アプリケーションからのトランザクション・データはキュー｛Ｍ、ａ｝および｛Ｍ、ｂ｝に格納される。メモリ用データ記録手段５０６により、メモリ５０２へのデータの書き込みがなされると、アプリケーションのコミット処理が完了する。その後、書き込みタイミング制御手段５０８は、ディスク用データ記録手段５０５にスレッドを割り当て、ディスク用データ記録手段５０５はデータを受取って書き込み処理を行う。なお、ディスク用データ記録手段５０５が、ディスク５０１にデータの書き込みを完了すると、冗長化したデータは不要となるので、書き込みタイミング制御手段５０８は、メモリ５０２のデータの削除を指示する。ライト・ビハインド方式の場合、アプリケーション側のコミット処理は、メモリ５０２への書き込み処理が完了していればいいので、ディスク用データ記録手段５０５に対するスレッドの割り当てはコミット完了前であってもよい。また、メモリからのデータの削除処理は、個々のトランザクション毎に行なってもいいし、性能向上のために複数トランザクション分の削除をまとめてバッチ処理してもよい。

【００２４】

図６に、リージョンａについて、ディスクへのデータ書き込み方式を、ライト・スルー方式からライト・ビハインド方式へ変更する処理のフロー６００を例示する。ステップ６０１で処理が開始される。ステップ６０２で負荷の検出を行う。ステップ６０３で負荷が所定の閾値を超えたか否か判断する。例えば、所定の閾値は、先行トランザクションの処理のディスク書き込みが終了する前に次のトランザクションが発生する程度のトランザクション量および頻度を設定する。これは、アプリケーションのスループットやレイテンシに対する要求を満たす範囲内で負荷に対する閾値を設定することになる。ステップ６０３で、負荷が閾値を超えていないと判断される場合（Ｎｏ）は、ステップ６０６に進んで、ディスクへのデータ書き込み方式を変更することなく処理を終了する。

【００２５】

一方、ステップ６０３で、負荷が閾値を超えていると判断された場合（Ｙｅｓ）は、ステップ６０４に進む。ステップ６０４では、キュー｛Ｄ、ａ｝について、リージョンａの識別子を鍵とした排他ロックを取得する。排他ロックの鍵はリージョンの識別子以外のものでも可能である。この排他ロックの取得は、データ書き込みをライト・スルー方式からライト・ビハインド方式に変更している間に、アプリケーションやデータ書き込み手段がキューにアクセスしてエラーとなることを防ぐためである。ステップ６０５では、キュー管理テーブルにキュー｛Ｍ、ａ｝を登録する。ステップ６０６で排他ロックを解除する。ステップ６０７で、キュー｛Ｍ、ａ｝について、メモリ用データ記録手段スレッドを割り当てる。但し、ステップ６０７では、当該スレッドのインスタンスが残っている場合は、スレッドの割り当ては行う必要はない。この処理は不要となる。ステップ６０８で変更処理を終了する。ステップ６０６が完了した後、メモリ用データ記録手段にスレッドが割り当てられ、ライト・ビハインド方式によるデータの書き込み処理が始まる。

【００２６】

ちなみに、ステップ６０２では、アプリケーションの動作中にプロセスに障害が発生し、別のプロセスに引き継ぐフェイルオーバー、データベースの一時的な応答遅延や、データベースのフェイルオーバー、また、ディスクのフェイルオーバーを検出した場合にも、負荷が閾値を超えたとみなして、ステップ６０３からステップ６０４に進んで、ディスクへのデータ書き込みをライト・スルー方式からライト・ビハインド方式に変更するようにしてもよい。

【００２７】

ちなみに、ステップ６０２で負荷の検出できる例として、アプリケーションのフェイルオーバー時の差分反映が挙げられる。ライト・ビハインド方式でアプリケーションが動作中にプロセスに障害が発生し、別プロセスが引き継ぐフェイルオーバーの場合、遠隔メモリとディスク上のデータの差分をディスクに反映する必要がある。しかし、これが完了す

るのを待っていたのでは、アプリケーションの可用性が低下してしまう。そこで、本発明を用いて、差分を反映している間はライト・ビハインド方式でディスクに書き込みつつ、同時にアプリケーションの処理を即時に開始することができる。その後、差分を反映し終え、さらにフェイルオーバー後のトランザクションによる新たなデータも反映し終えた時点で、必要ならライト・スルー方式に切り替えてもよい。

【００２８】

また、ステップ６０２で負荷の検出できる例として、データベースの一時的な応答遅延への対応が挙げられる。ライト・スルー方式でデータベースに書き込み中、データベースが何らかの原因で一時的に応答が遅くなる場合がある。これに備えて、ある一定時間以上データベースが応答しない場合は、ライト・ビハインド方式に切り替えることによって、アプリケーションに対して応答時間を保証できる。その後データベースが応答したら、ライト・スルー方式に戻してもよい。動的な切り替えにより、耐えられる応答遅延時間を長くできる。

【００２９】

さらに、ステップ６０２で負荷の検出できる例として、データベースのフェイルオーバー時間の隠蔽する場合が挙げられる。ライト・スルー方式でデータベースに書き込み中、データベースに障害が発生しデータベース・プロセスがフェイルオーバーする場合がある。通常、データベースのフェイルオーバーには時間がかかり、その間アプリケーションへの応答を返せなくなるが、これを応答遅延の1つととらえ、同様にライト・ビハインド方式に切り替えれば、データベースのフェイルオーバー時間をアプリケーションに対して隠蔽することができる。これによりアプリケーションの可用性を上げられる。動的な切り替えにより、耐えられるデータベースのフェイルオーバー時間を長くできる。

【００３０】

そして、ステップ６０２で負荷の検出できる例として、ディスクのフェイルオーバーも挙げられる。ライト・スルー方式またはライト・ビハインド方式で使用中のファイルやデータベースまたはディスクそのものに障害が発生した場合、書き込み先を他のファイルやデータベースに切り替える必要がある場合もある。これは単にディスク用データ記録部の書き込み先を切り替えるだけで実現できる。

【００３１】

また、ステップ６０２の負荷の検出は、リージョン毎に行うことも可能である。例えば、リージョンｂについて負荷が閾値を超えたと判断した場合は、ステップ６０４で、リージョンｂの識別子を鍵とした排他ロックを取得し、ステップ６０５でキュー管理テーブルにキュー｛M、ｂ｝を登録する。ステップ

【００３２】

図７に、リージョンａについて、ディスクへのデータ書き込み方式を、ライト・ビハインド方式からライト・スルー方式へ変更する処理のフロー７００を例示する。ステップ７０１で処理が開始される。ステップ７０２で負荷の検出を行う。ステップ７０３で負荷が所定の閾値を下回るか否か判断する。ステップ７０３で、負荷が所定の閾値を下回ると判断される場合（Ｎｏ）は、ステップ７０７に進んで、ディスクへのデータ書き込み方式を変更することなく処理を終了する。一方、ステップ７０３で、負荷が閾値を下回ると判断された場合（Ｙｅｓ）は、ステップ７０４に進む。ステップ７０４では、キュー｛D、ａ｝およびキュー｛M、ａ｝につき、リージョンａの識別子を鍵とした排他ロックを取得する。ステップ７０５では、キュー管理表からキュー｛M、ａ｝を削除する。ステップ７０６で排他ロックを解除する。ステップ７０７で変更処理を終了する。

【００３３】

また、ステップ７０３では、アプリケーションの動作中のフェイルオーバーの完了、データベースの一時的な応答遅延の回復や、データベースのフェイルオーバーの完了、またはディスクのフェイルオーバーの完了を検出した場合にも、ステップ７０４に進んで、ディスクへのデータ書き込みをライト・ビハインド方式からライト・スルー方式に変更するようにしてもよい。なお、ステップ７０５では、キュー｛M、ａ｝のインスタンスそのも

のを削除してしまうのではなく、図８のように単にキュー管理テーブルから登録を抹消しておく。キュー｛M、ａ｝のインスタンスが残っていれば、キュー管理テーブルから抹消された後も、メモリ用データ記録手段に割り当てられたスレッドは、キュー｛M、ａ｝から残存するライト・ビハインド方式のデータを取得し書き込み処理を継続できる。但し、キュー｛M、ａ｝のデータ書き込みが終われば、リソースに応じて、インスタンスを削除することが望ましい。

【００３４】

図９は、図６と図７のそれぞれの処理を、同一フローで行う場合の一例を示したフロー９００である。ステップ９０２の負荷の検出はステップ６０２およびステップ７０２と共通化できるので、二つの処理フローは図９のように融合することができる。ステップ９０３とステップ９１３は負荷が閾値よりも上か下かを判断する処理である。ステップ９０３では、負荷が閾値より下から上に変化があったか否かを判断する。ステップ９１３では、負荷が閾値より上から下に変化があったか否かを判断する。ステップ９０４からステップ９０６は、図６のステップ６０４からステップ６０７と同じである。ステップ９１４からステップ９１６は、図７のステップ７０４からステップ７０６と同じである。ステップ９０８では処理は終了となるが、負荷の変化に対応するため、定期的に処理は再開され、ステップ９０２に戻って再び負荷の検出をし、ライト・スルー方式とライト・ビハインド方式の切り替え処理を始める。

【００３５】

本発明は、様々な形態で実施可能である。例えば図１０に示すように、データベースとマルチキャスト通信を用いたライト・ビハインド方式に適用できる。ディスクをデータベース１００１として、ディスク用データ記録手段１００５がデータを書き込み、また、マルチキャスト通信１０１０を用いて、ディスク用データ記録手段１００５が複数の遠隔メモリ１００２にコピーすることで冗長度を上げられる。メモリ用データ記録手段１００６がデータ書き込み完了を通知するタイミングを変えることで遠隔メモリの障害に対する耐久度を変えることもできる。例えば、コピー先の遠隔メモリが３つある場合、そのうち所定の数（例えば、最初の２つ）の遠隔メモリからのコピー完了応答が届いた時点で書き込み完了を通知すれば、最大１つの遠隔メモリの障害に耐えることができる。

【００３６】

また、本発明はリージョンが１つしかない場合にも適用可能である。リージョンがシステムに１つしかない場合も許される。この場合はシステム全体でライト・スルー方式またはライト・ビハインド方式の耐久性保証方法を統一的に切り替えて利用することになる。

【００３７】

本手法により、データの耐久性保証方法であるライト・スルー方式とライト・ビハインド方式を動的な負荷の変動に応じて切り替えることが可能になる。これにより、ディスク・ボトルネックになるような高負荷時に閾値の範囲でメモリ使用量を節約し、閾値を超えるような高負荷に備えることが可能になる。また、ライト・スルー方式とライト・ビハインド方式の両方における対故障性を向上でき、フェイルオーバーにかかる時間を小さくできて、データの耐久性保証方法の切り替えの範囲を柔軟に定義できる。

【００３８】

以上、本発明を実施の形態を用いて説明したが、本発明の技術的範囲は上記実施の形態に記載の範囲には限定されない。上記実施の形態に、多様な変更または改良を加えることが可能であることが当業者に明らかである。その様な変更または改良を加えた形態も本発明の技術的範囲に含まれ得ることが、特許請求の範囲の記載から明らかである。

【図面の簡単な説明】

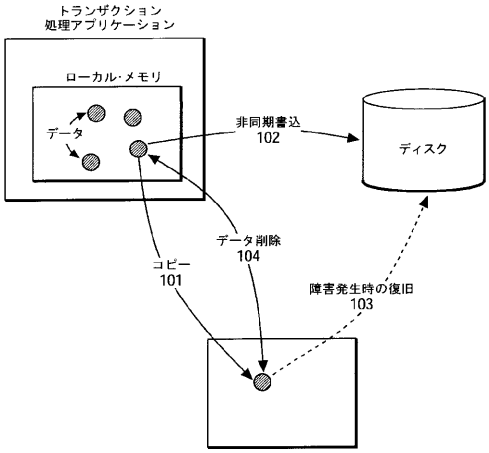
【００３９】

【図１】ライト・ビハインド方式によるデータの書き込み処理を簡単に示す。

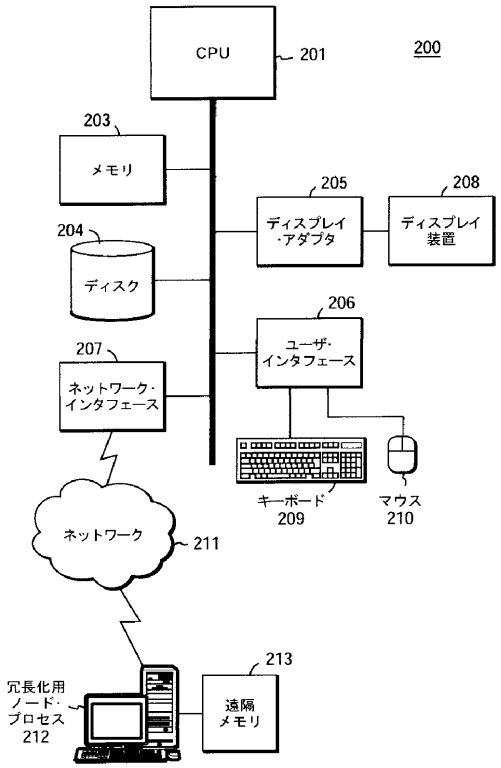
【図２】ディスクへのトランザクション・データの書き込みを、ライト・スルー方式とライト・ビハインド方式で動的に切り替えるシステムが作動するハードウェア構成図の一例を示す。

【図３】ディスクへのトランザクション・データの書き込みを、ライト・スルー方式とライト・ビハインド方式で動的に切り替える装置の機能および処理されるデータの一例を示す。
【図４】ライト・スルー方式を実現した状態の一例を示す。
【図５】ライト・ビハインド方式を実現した状態の一例を示す。
【図６】ディスクへのデータ書き込み方式を、ライト・スルー方式からライト・ビハインド方式へ変更する処理のフローを例示する。
【図７】ディスクへのデータ書き込み方式を、ライト・ビハインド方式からライト・スルー方式へ変更する処理のフロー７００を例示する。
【図８】単にキュー管理テーブルから登録を抹消しておく状態を示す。
【図９】図６と図７のそれぞれの処理を、同一フローで行う場合のフローの一例を示す。
【図１０】データベースとマルチキャスト通信をライト・ビハインド方式に適用した一例を示す。

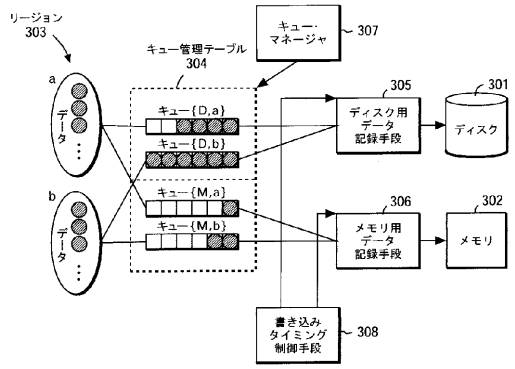
【図１】



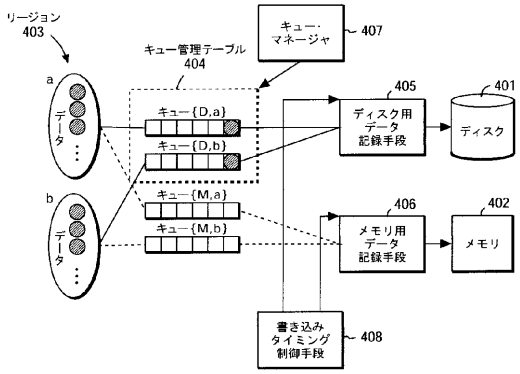
【図２】



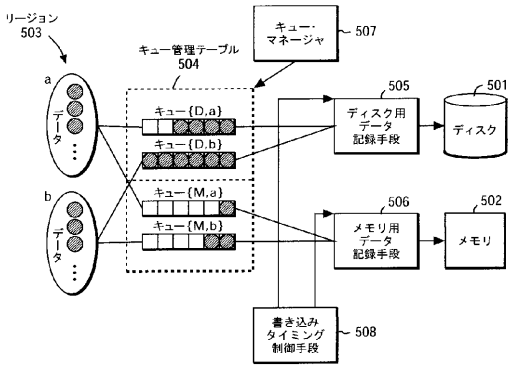
【図３】


リージョン
303
キュー・
マネージャ
307
キュー管理テーブル
304
a
データ
b
データ
キュー{D,a}
キュー{D,b}
キュー{M,a}
キュー{M,b}
305
ディスク用
データ
記録手段
301
ディスク
306
メモリ用
データ
記録手段
302
メモリ
書き込み
タイミング
制御手段
308

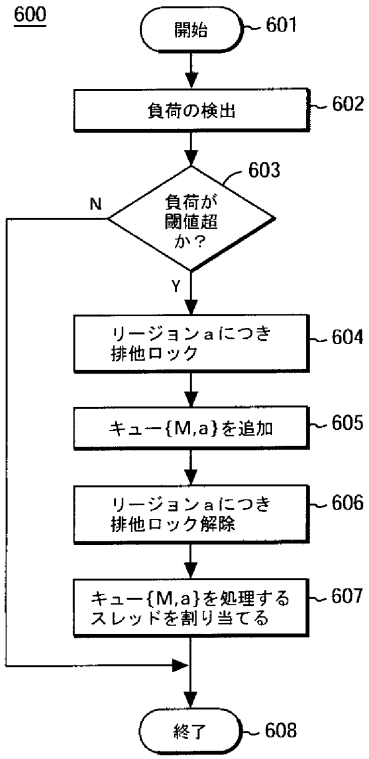
【図４】


リージョン
403
キュー・
マネージャ
407
キュー管理テーブル
404
a
データ
b
データ
キュー{D,a}
キュー{D,b}
キュー{M,a}
キュー{M,b}
405
ディスク用
データ
記録手段
401
ディスク
406
メモリ用
データ
記録手段
402
メモリ
書き込み
タイミング
制御手段
408

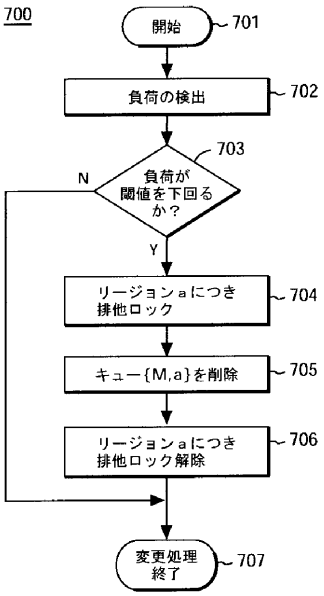
【図５】


リージョン
503
キュー・
マネージャ
507
キュー管理テーブル
504
a
データ
b
データ
キュー{D,a}
キュー{D,b}
キュー{M,a}
キュー{M,b}
505
ディスク用
データ
記録手段
501
ディスク
506
メモリ用
データ
記録手段
502
メモリ
書き込み
タイミング
制御手段
508

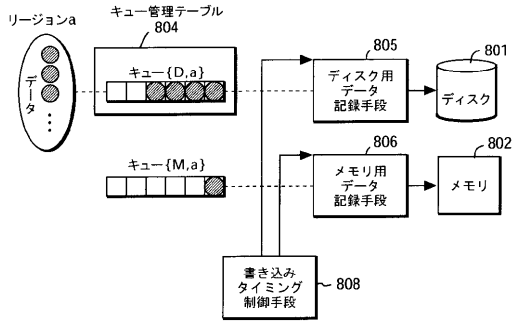
【図６】


600
開始
601
負荷の検出
602
負荷が
閾値超
か？
603
N
Y
リージョンａにつき
排他ロック
604
キュー{M,a}を追加
605
リージョンａにつき
排他ロック解除
606
キュー{M,a}を処理する
スレッドを割り当てる
607
終了
608

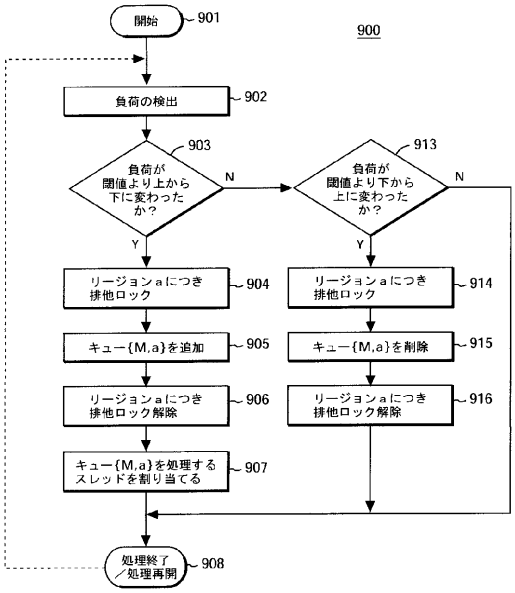
【図７】


700
開始
701
負荷の検出
702
703
負荷が
閾値を下回る
か？
N
Y
リージョンａにつき
排他ロック
704
キュー{M,a}を削除
705
リージョンａにつき
排他ロック解除
706
変更処理
終了
707

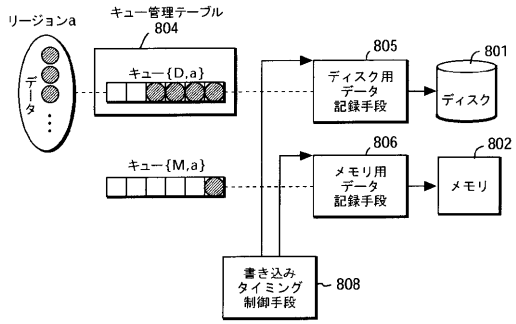
【図８】


リージョンa
キュー管理テーブル
804
データ
キュー{D,a}
805
ディスク用
データ
記録手段
801
ディスク
キュー{M,a}
806
メモリ用
データ
記録手段
802
メモリ
書き込み
タイミング
制御手段
808

【図９】


開始
901
900
負荷の検出
902
903
負荷が
閾値より上から
下に変わった
か？
N
Y
913
負荷が
閾値より下から
上に変わった
か？
N
Y
リージョンaにつき
排他ロック
904
リージョンaにつき
排他ロック
914
キュー{M,a}を追加
905
キュー{M,a}を削除
915
リージョンaにつき
排他ロック解除
906
リージョンaにつき
排他ロック解除
916
キュー{M,a}を処理する
スレッドを割り当てる
907
処理終了
／処理再開
908

【図１０】


リージョンa
キュー管理テーブル
804
データ
キュー{D,a}
805
ディスク用
データ
記録手段
801
ディスク
キュー{M,a}
806
メモリ用
データ
記録手段
802
メモリ
書き込み
タイミング
制御手段
808

フロントページの続き

(74)代理人　100112690
弁理士　太佐　種一
(72)発明者　根山　亮
神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　東京基礎研究所内
(72)発明者　山本　学
神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　東京基礎研究所内
(72)発明者　小澤　陽介
神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　東京基礎研究所内

審査官　木村　雅也

(56)参考文献　特開平０７－０７２９８０（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　　３／０６
Ｇ０６Ｆ　１２／００
Ｇ０６Ｆ　１３／１０