(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4289056号
(P4289056)
(45)発行日　平成21年7月1日(2009.7.1)　(24)登録日　平成21年4月10日(2009.4.10)

(51)Int.Cl.　　F I
G06F 12/00 (2006.01)　　G06F 12/00 531D
G06F 3/06 (2006.01)　　G06F 12/00 533J
　　G06F 3/06 304F

請求項の数 9　(全 21 頁)

(21)出願番号　特願2003-207782(P2003-207782)
(22)出願日　平成15年8月19日(2003.8.19)
(65)公開番号　特開2004-348701(P2004-348701A)
(43)公開日　平成16年12月9日(2004.12.9)
審査請求日　平成18年3月20日(2006.3.20)
(31)優先権主張番号　特願2003-86920(P2003-86920)
(32)優先日　平成15年3月27日(2003.3.27)
(33)優先権主張国　日本国(JP)

(73)特許権者　000005108
株式会社日立製作所
東京都千代田区丸の内一丁目６番６号
(74)代理人　100100310
弁理士　井上　学
(72)発明者　須藤　敦之
東京都国分寺市東恋ケ窪一丁目２８０番地
株式会社日立製作所中央研究所内
(72)発明者　馬場　恒彦
東京都国分寺市東恋ケ窪一丁目２８０番地
株式会社日立製作所中央研究所内

審査官　田川　泰宏

最終頁に続く

(54)【発明の名称】計算機システム間のデータ二重化制御方法

(57)【特許請求の範囲】
【請求項１】
第１の計算機システムと、該第１の計算機システムに接続される第１のストレージシステムとを有する正システムと、第２の計算機システムと、該第２の計算機システムに接続された第２のストレージシステムとを有する副システムとを備え、かつ少なくとも前記第１、第２のストレージシステムの間が相互接続されているシステムにおけるデータ二重化制御方法であって、
前記第１の計算機システムの処理によって前記第１のストレージシステムが保持するデータの更新を再現可能なログを該第１のストレージシステム内の特定記憶装置に登録するステップと、
前記ログの複製のために設定された第２のストレージシステム内の特定記憶装置に前記第１のストレージシステムの特定記憶装置に登録されたログをコピーするステップと、
前記第１計算機システムの処理によって前記第１のストレージシステムが保持するデータを更新するステップと、
前記コピーステップにより前記第２のストレージシステム内の特定記憶装置の保持内容に変更が生じたことを前記第２の計算機システムが検知するステップと、
前記第２の計算機システムが前記第２のストレージシステム内の特定記憶装置の保持内容の変更を読み込むステップと、
前記第２の計算機システムがログを読み込み該ログにしたがって前記第２のストレージに保持する前記データの複製を更新するステップ、

を有することを特徴とするデータ二重化制御方法。
【請求項２】
前記第１の計算機システムの処理によって前記第１のストレージシステムが保持するデータの更新を再現可能なログは、１つまたは複数のトランザクションからなり該トランザクションの開始と終了を区分する情報を含むことを特徴とする請求項１のデータ二重化制御方法。
【請求項３】
前記第１の計算機システムの処理による前記第１のストレージシステムが保持するデータの更新を再現可能なログを該第１のストレージシステム内の特定記憶装置に登録するステップは、第１の計算機システムが指定したログ入出力単位で行うことを特徴とする請求項１のデータ二重化制御方法。
【請求項４】
前記第１の計算機システムの処理によって前記第１のストレージシステムが保持するデータの更新を可能なログを該第１ストレージシステム内の特定記憶装置に登録するステップと
前記ログの複製のために設定された第２のストレージシステム内の特定記憶装置に前記第１のストレージシステムの特定記憶装置のログの変更をコピーするステップとを同期で行うことを特徴とする請求項１のデータ二重化方法。
【請求項５】
請求項１記載のデータ二重化制御方法において、前記第１の計算機システムが停止したことを検知するステップと、前記第２の計算機システムが前記第１の計算機システムから業務を引き継ぐステップを更に有することを特徴とするデータ二重化制御方法。
【請求項６】
請求項１記載のデータ二重化制御方法において、前記第１の計算機システムが停止したことを検知するステップと、前記第２の計算機システムが前記第１の計算機システムから業務を引き継ぐステップと、前記第２の計算機システムが前記第２のストレージシステム内の特定記憶装置のログを読み込み該ログにしたがってデータを更新するステップをさらに有することを特徴とするデータ二重化制御方法。
【請求項７】
データベースを構築する第１のデータベースサーバに接続され、前記データベースのデータを格納する記憶領域及び前記データベースの更新履歴を示す更新ログを格納する記憶領域を備える第１のストレージシステムと、第２のデータベースサーバと、に接続される第二のストレージシステムであって、
第１の記憶領域と、前記第二のデータベースサーバが用いるデータを格納する第二の記憶領域と、
前記第１の記憶領域及び前記第二の記憶領域に対してデータの入出力を制御するストレージ制御部と、を備え、
前記ストレージ制御部は、
前記更新ログを前記第１のデータベースサーバから前記第一のストレージシステムを介して、受領する受領部と、
前記更新ログを、前記第１の記憶領域に格納し、
前記第１の記憶領域に格納されたことを前記第二のデータベースサーバに通知し、
前記第二のデータベースサーバから、前記第１の記憶領域に格納される更新ログを前記第２の記憶領域に反映する反映要求を受け、
前記第２の記憶領域に前記更新ログを反映し、
前記反映が完了した場合、完了したことを示す完了報告を前記第２のデータベースサーバに通知する、ことを特徴とする第２のストレージシステム。
【請求項８】
データベースを構築する第１のデータベースサーバに接続され、前記データベースのデータを格納する記憶領域及び前記データベースの更新履歴を示す更新ログを格納する記憶

領域を備える第１のストレージシステムと通信可能な第１の記憶領域及び第２の記憶領域を備える第２のストレージシステムに接続される第２のデータベースサーバであって、
前記第１のストレージシステムから送信される前記１の記憶領域に新たな更新ログが格納されたか否かを検出する検出手段と、
前記検出手段により検出した場合は、前記第１の記憶領域から更新ログを読み出す読み出し手段と、
前記読み出された更新ログを前記第２の記憶領域に更新する更新要求を前記第２のストレージシステムに送信する更新要求送信手段と、
前記第二のストレージシステムから前記更新要求に対応する更新処理の完了を示す前記更新完了を受信する手段と、を有することを特徴とする第２のデータベースサーバ。

【請求項９】
請求項８記載の第２のデータベースサーバであって、
第一のデータベースサーバの障害を検出した場合、
前記更新ログの前記第１の記憶領域への格納を停止するよう前記第２のストレージシステムに指示する停止指示手段と、
前記第１の記憶領域から更新ログを読み出し、前記第２の記憶領域に未反映の更新ログの有無を確認する確認手段と、
前記未反映の更新ログがあった場合は、前記第２のストレージシステムに前記第２の記憶領域に反映するよう指示する更新指示手段と、
前記更新指示手段に対し、前記第２のストレージシステムから反映完了報告を受信した場合、前記第２のデータベースサーバで、前記第１のデータベースサーバの処理を引き継ぐことを特徴とする第２のデータベースサーバ。

【発明の詳細な説明】
【０００１】
【発明の属する技術分野】
本発明は計算機およびストレージ装置からなる業務システムに関し、特に複数のシステム間でデータを複製するデータ二重化制御方法、およびデータを複製したシステムへの高速な切り替えを実現する方法に関する。
【０００２】
【従来の技術】
データベースサーバおよびストレージ装置からなる業務システムが複数ある場合のデータ複製方法として、データベースサーバ上で動作するＤＢＭＳが実行する方法がある。ＤＢＭＳがデータ複製する方法については、例えば非特許文献１に記述がある。複数のシステムのデータベースサーバ同士を接続し、一方のシステム上で動作するＤＢＭＳの更新情報を別のシステムに転送することでデータ複製する技術である。
【０００３】
また、同様なシステムのデータ複製方法として、ストレージ装置間のデータコピー機能を使用する方法がある。ストレージ装置間のデータコピー機能については、例えば非特許文献２に記述がある。複数システムのストレージ同士をファイバーチャネルで接続し、一方のストレージ装置のディスクドライブに更新があると、別のストレージ装置のディスクドライブにもデータの更新を反映する技術である。
【非特許文献１】
Oracle9i製品カタログ（http://www.oracle.co.jp/products/catalog/pdf/9iDBr2J07266-01.pdf）、第６頁。
【非特許文献２】
日立統合ストレージソリューション「Storeplaza」カタログ(hppt://www.hitachi.co.jp/Prod/comp/storeplaza/data/stpzclg.pdf)、第５頁。
【０００４】
【発明が解決しようとする課題】
従来のデータ複製方法を実行する場合、通常のデータベース業務を行う以上のコストが必

要である。また、複数システム間で同期したデータ複製を行うと業務処理の遅延が発生する。

【０００５】

ＤＢＭＳによるデータ複製を行うためには、ＤＢＭＳの動作するデータベースサーバが業務処理を行う負荷に加えて、データ複製処理を行う負荷が加わるためにより高性能なデータベースサーバが必要でありコストが増加するという課題がある。また、複製したデータが一致するためには、ＤＢＭＳの更新処理を実行するたびにデータベースサーバ間で同期通信を行う必要がある。同期通信中は、ＤＢＭＳが次の更新処理を実行できないため、業務が遅延することが課題である。

【０００６】

ストレージ装置でデータ複製を行うためには、ＤＢＭＳが扱うデータの更新を全てコピーするため、ストレージ装置間の接続に広帯域の回線を使う必要がある。広帯域の回線を使用することでコストが増大する課題がある。また、複製したデータが一致するためには、ディスクドライブ上のデータが更新されるたびにストレージ装置間で同期通信を行う必要がある。同期通信中はディスクドライブへの次の更新処理が実行できないため、業務が遅延することが課題である。

【０００７】

また、同期通信による遅延を防ぐため、ＤＢＭＳやストレージ装置間の通信を非同期で実行する方法が存在するが、障害や災害でデータ複製先のシステムに切り替える場合に、未転送分のデータを複製先で再構築する必要が生じ、システムの切り替えが遅延することが課題である。

【０００８】

【課題を解決するための手段】

サーバが外部から受け付けた要求に応じて業務を実行すると、ストレージ装置に保存されたデータへの更新や追加が必要となる。このストレージ装置のデータ複製を行うために全てのデータを複製するのではなく、複製先としてサーバとストレージ装置を用意し、複製元のサーバで実行された業務を復元可能なログをストレージ装置の特定のディスクドライブに保存し、このディスクドライブが更新されるたびに複製先のストレージ装置にディスクドライブのコピーを行う。複製先のストレージ装置へのディスクドライブのコピーが完了したら、コピーされたストレージ装置からログを保存したディスクドライブが更新されたことを複製先のサーバに通知する。複製先のサーバは、ストレージ装置からログを保存したディスクドライブの変更通知を受信できるようにしておき、通知を受けた後でログをディスクドライブから読み取り、複製元のサーバで行われたのと同じ業務処理を実行する。このログを基にした業務処理の実行後、その結果をストレージ装置に反映することでデータの複製が完了する。

【０００９】

このデータ複製方法を実行しているシステムにおいて、複製元のサーバとストレージ装置が障害や保守操作により停止した場合、複製先のストレージ装置に保存された業務データが最新の状態にあるため、複製元のサーバが受信していた業務を複製先で受信するように変更することで、業務の処理を中止せずにサーバとストレージ装置の切り替えを実行する。

【００１０】

サーバとストレージ装置の切り替え実行後に、複製元と複製先双方のサーバとストレージ装置とがデータ複製のために実行していた処理を交替することで、業務を受信し処理しているシステムが停止した場合、再びサーバとストレージ装置の切り替えを実行する。

【００１１】

【発明の実施の形態】

以下で説明する実施形態では、例として計算機上で動作する業務としてデータベースサーバを取り上げるが、計算機上で実行される業務はデータベースに限定するものではない。計算機上で動作する業務は、複製するデータが、正システムで行った業務によって更新さ

れるもので、かつ、そのデータの更新を副システムで再現できるログを生成するものであればよい。例えば、ファイルシステムなどでも実施可能である。

［第１実施形態］

図１は、本発明が適用されたデータベースサーバとストレージ装置を用いたデータ複製システムの一実施例である。

【００１２】

正システムを構成するのは、データベースサーバ２とストレージ装置８である。これらはデータベースサーバ２に内蔵されたストレージ接続装置３とストレージ装置８のディスク制御装置５とがサーバ・ストレージ間接続インタフェース４によって接続される。ストレージ装置８はディスク制御装置５によって読み込み書き込みを行うデータを保存するディスクドライブ６，７を内蔵しており、データベースサーバ２が業務ネットワーク１を通じて業務要求を受け取って処理したデータや、その処理に必要なデータおよびデータベースサーバ２内部で実行された業務データを保持する。

【００１３】

データベースサーバ２とストレージ装置８とは、サーバ・ストレージ間接続インタフェース４を通してデータの読み込み・書き込みを行うだけでなく、データベースサーバ２が要求したディスクドライブ６，７の変更があった場合、ストレージ装置８からデータベースサーバ２に通知を行う方法を有している。

【００１４】

副システムを構成するのは、データベースサーバ１２とストレージ装置１８である。これらはデータベースサーバ１２に内蔵されたストレージ接続装置１３とストレージ装置１８のディスク制御装置１５とがサーバ・ストレージ間接続インタフェース１４によって接続される。ストレージ装置１８はディスク制御装置１５によって読み込み書き込みを行うデータを保存するディスクドライブ１６，１７を内蔵しており、データベースサーバ１２が業務ネットワーク１を通じて業務要求を受け取って処理したデータや、その処理に必要なデータおよびデータベースサーバ１２内部で実行された業務データを保持する。

１８とは、サーバ・ストレージ間接続インタフェース１４を通してデータの読み込み・書き込みを行うだけでなく、データベースサーバ１２が要求したディスクドライブ１６，１７の変更があった場合、ストレージ装置１８からデータベースサーバ１２に通知を行う方法を有している。

【００１５】

ディスク制御装置５とディスク制御装置１５とはストレージ装置間接続インタフェース２０により接続される。これにより、正システムのストレージ装置８と副システムのストレージ装置１８は互いに接続される。ストレージ装置８とストレージ装置１８は、一方のディスクドライブの一つを複製元に、他方のディスクドライブの一つを複製先にあらかじめ設定しておくことで、ストレージ装置間接続インタフェース２０を通して内容を複製する方法を有している。

【００１６】

以下、本実施形態のデータ複製方法およびシステム切り替え方法の動作を説明する。本実施形態では業務を通常実行している正システムと、正システムが何らかの理由で稼働不可能になった時に業務を引き継ぐ副システムとの間でデータ複製を行うものとする。

【００１７】

まず、データ複製方法を実現するための初期設定を正システム、副システム双方について行う。

【００１８】

正システムの初期設定は、業務システムに応じたデータベースを構築することから始める。ストレージ装置８のディスク制御装置５で、データベースサーバ２が使用可能なディスクドライブ６，７を割り当てる。データベースサーバ２はデータベースのデータを保持するディスクドライブ６とデータベースのログを保持するディスクドライブ７とを設定する。ここで言うログとは、データベースの更新作業を逐一表すもので、ログを再実行するこ

とでデータベースの再構築が可能なものである。例えば、データベースが実行したトランザクションログやデータベースサーバが受け取った業務要求全てのＳＱＬコマンドである。トランザクションは、データベースの処理を複数個まとめた処理単位で、その処理が全て成功するか失敗するかのいずれかになる。そのため、多数の業務要求を処理しなければならない業務システムは、トランザクション単位で処理を行うことでデータベースに不整合を発生させないために使用する。

【００１９】

副システムにも正システムと同様のデータベースを構築する。ストレージ装置１８において、ストレージ装置８でデータベースサーバ２が使用可能としたディスクドライブ６と同様なディスクドライブ１６と、ディスクドライブ７と同様なディスクドライブ１７をディスク制御装置１５でデータベースサーバ１２が使用可能となるように割り当てる。データベースサーバ１２は、データベースサーバ２同様に、データベースのデータを保持するディスクドライブ１６とデータベースのログを保持するディスクドライブ１７とを設定する。

【００２０】

次に、正システムのストレージ装置８と副システムのストレージ装置１８との間で、ストレージ装置間接続インタフェース２０を通じてデータベースのログを保持するディスクドライブ７をディスクドライブ１７にコピーするように設定する。このディスクドライブコピーは、同期コピー、非同期コピーいずれとも可能である。例えば、同期コピーとは、データベースサーバ２からのディスクドライブ７への情報書き込み要求に対して、ディスクドライブ７への情報書き込みとディスクドライブ１７への情報書き込みとの両方が終了してから、データベースサーバ２に書き込み完了報告を行うものを指す。一方、非同期コピーは同期コピーと異なり、ディスクドライブ７への情報書き込みが完了した時点でデータベースサーバ２に書き込み完了報告を行う。非同期コピーを行う場合、正システムと副システムのログディスクが常に一致するとは限らず、システム切り替え時にデータが欠損することがある。

【００２１】

そして、副システムのデータベースサーバ１２からストレージ装置１８のディスク制御装置１５に対して、ディスクドライブ１７の更新が行われたらデータベースサーバ１２に通知を行うように設定する。

【００２２】

正システムに障害・災害などが発生した場合に、副システムに切り替えるため正システムの停止を迅速に検知する必要がある。そのため、正システムのデータベースサーバ２と副システムのデータベースサーバ１２の間で正システムが稼働していることを通知するための通信設定を行う。例えば、正システムのデータベースサーバ２から副システムのデータベースサーバ１２に業務ネットワーク１を経由して一定時間間隔で通知を行う方法がある。また、正システムの稼働状態を監視する外部のサーバから副システムへの切り替えを指示する方法や、副システムから一定時間間隔で正システムに稼働状態を問い合わせる方法もある。

【００２３】

以上のような設定がデータベースサーバ２，１２とストレージ装置８，１８で完了した後、正システムのデータベースサーバ２で業務処理を開始する。以下では、データ複製の手順について説明する。

【００２４】

データ複製第１ステップ１０１：業務処理要求は、業務ネットワーク１を通じてデータベースサーバ２に到着する。業務処理要求は業務ネットワーク上のプロトコルに応じて送付され、データベースサーバ２の管理するデータ内容を参照するものや更新するものからなる。例えば、ＴＣＰ／ＩＰプロトコルによって送付される、ＳＱＬコマンドの組み合わせからなる。

【００２５】

業務処理要求を受信したデータベースサーバ２は、ネットワークプロトコル層の解析を行い、データベースへの業務処理内容を取り出し、業務処理内容の解析を行った後、業務処理を実行する。例えば、ＴＣＰ／ＩＰプロトコルの解析を行い、ＳＱＬコマンドを取り出し、その処理をデータベースで実行する処理がある。

【００２６】

データ複製第２ステップ１０２：業務処理の内容が、データベースの更新処理を伴う場合にはストレージ装置内に保持しているデータを更新する必要がある。その場合には、ストレージ装置接続装置３からサーバ・ストレージ接続インタフェース４を通じて、ディスク制御装置５に対してディスクドライブ７への更新ログの書き込みをストレージ装置８に指示する。例えば、データベースサーバ２にホストバスアダプタを装着し、ファイバーチャネルケーブルを通じてＳＣＳＩコマンドをディスクコントローラに送信することに当たる。また、この更新ログの書き込みはデータ更新の度に常に実行するだけでなく、トランザクション単位で実行が完了したときに行う方法、データベースサーバ２で用意したログ用バッファの空きが無くなった場合や一定時間が経過した後などのようにデータベースサーバ２が指定する入出力の単位で条件に応じて実行する方法なども用いることができる。また、本実施形態では簡単のため１回の書き込み要求のように図示したが、通常は別のディスクドライブへの書き込み要求は複数の要求に分けて送信される。

【００２７】

データ複製第３ステップ１０３：更新ログの書き込み要求を受けたディスク制御装置５は、ディスクドライブ７へと情報の書き込みを行う。

【００２８】

データ複製第４ステップ１０４：ディスクドライブ７はその情報の書き込みが終了すると副システムのストレージ装置１８内のディスクドライブ１７にコピーするように設定されているため、ディスク制御装置５はストレージ装置間接続インタフェース２０を通じて副システムのストレージ装置１８にあるディスク制御装置１５にディスクドライブ７の更新内容を送信し、ディスクドライブ１７に書き込むよう指示し、ディスク制御装置１５はディスクドライブ１７へと書き込みを行う。この書き込みが完了したら、ディスク制御装置５からデータベースサーバ２に更新ログ書き込み完了報告を行う。例えば、ストレージ装置間接続インタフェース２０としてファイバーチャネルケーブルを用い、ストレージ装置の管理ソフトウェアでディスクドライブ７，１７のコピーを設定することで実現できる。また、本実施形態では、ディスクドライブ７の更新直後にディスクドライブ１７へのコピーを行う同期コピー方法としているが、一定時間間隔でコピーを実行する非同期コピー方法を用いることも可能である。ただし、非同期コピー方法を用いた場合、ディスクドライブ１７への書き込みが完了しなくてもストレージ装置８内のディスクドライブ７への書き込みが完了したらデータベースサーバ２に完了報告を行うため、システム切り替え時にディスクドライブのデータがコピーされていない事態も発生しうる。

【００２９】

データ複製第５ステップ１０５：更新ログの書き込み完了報告を受けたデータベースサーバ２は、ディスク制御装置５にディスクドライブ６へ更新データの書き込みを指示し、ディスク制御装置５はディスクドライブ６への情報の書き込み処理を行う。

【００３０】

データ複製第６ステップ１０６：ディスクドライブ１７への更新を実行後、ディスク制御装置１５はあらかじめデータベースサーバ１２から更新を通知するように指定されているため、更新が発生したことをデータベースサーバ１２に通知する。この更新通知要求と更新通知のインタフェースは、例えば、データベースサーバ１２からストレージ装置１８内の特殊なディスクドライブへの読み込み要求の応答としてディスク制御装置１５が通知する方法や、データベースサーバ１２から更新通知を要求するディスクドライブ１７への専用コマンドに対する応答としてディスク制御装置１５が通知する方法、また、ディスク制御装置１５からディスクの更新を通知する専用の割り込みインタフェースをデータベースサーバ１２内のストレージ接続装置１３に設ける方法などがある。また、ディスク制御装

置１５からデータベースサーバ１２への通知は、更新が発生する度に実行する方法に限定するわけではなく、一定時間間隔ごとに通知する方法や、データベースサーバ２から指示をストレージ装置８に発行したものをストレージ装置１８に伝えてデータベースサーバ１２への通知を実行させる方法などがある。

【００３１】

データ複製第７ステップ１０７：ディスクドライブ１７の更新通知を受けたデータベースサーバ１２は、ディスクドライブ１７の更新分の情報を読み込み、その更新ログにしたがって、ディスクドライブ１６上のデータを更新するようにストレージ接続装置１３からサーバ・ストレージ間接続インタフェース１４を通じてディスク制御装置１５に情報書き込み要求を通知する。例えば、データベースサーバ１２にホストバスアダプタを装着し、ファイバーチャネルケーブルを通じてＳＣＳＩコマンドをディスクコントローラに送信する方法がある。

【００３２】

正システムのデータベースサーバ２が業務処理要求を実行するたびに、このように正システムのデータを副システムに複製することで、正システムのデータベースサーバ１２にデータ複製のための負荷をかけることなく、また業務ネットワーク１上にデータ複製のためのデータ送信を行うことなく、ストレージ装置８，１８間でのデータ転送量を小さくして、データ複製のコストを抑え、業務の遅延を小さくすることができる。

【００３３】

正システムが災害や機器の障害などにより停止した場合、副システムに業務処理を切り替える。正システムの保守作業を行うため必要がある場合でも、正システムを停止させ、副システムに業務処理を切り替えることがある。図２は正システム停止後に副システムが業務処理を引き継ぐ処理を行う手順を示した。装置構成は図１と同様であるため詳細は省略する。以下では、業務処理引き継ぎの手順を説明する。

【００３４】

正システムが停止すると、システムを切り替えて副システムのデータベースサーバ１２とストレージ装置１８で業務処理を引き継ぐ。正システムの停止は、例えばデータベースサーバ２とデータベースサーバ１２との間で一定時間間隔で通信を行うハートビート通信や、データベースサーバ２，１２以外の監視サーバを業務ネットワーク１に接続してハートビート通信を行う方法で検出可能である。業務処理の引き継ぎは、データベースサーバ２で受け取っていた業務処理要求をデータベースサーバ１２が受け取れるように設定を変更することで可能である。例えば、データベースサーバ２が業務要求受信に用いていたネットワークアドレスを引き継ぐ方法がある。

【００３５】

システム切り替え第１ステップ２０１：まず、データベースサーバ１２からディスクドライブ７からディスクドライブ１７への情報の書き込みを停止するようにストレージ装置１８に指示を出してから、データベースサーバ１２がログの書き込まれたディスクドライブ１７を参照し、未実行の業務がディスクドライブ１７に存在するか確認する。ディスクドライブ７からディスクドライブ１７への書き込みを停止する理由は、データベースサーバ２が正常に稼動しているか不明でありディスクドライブ７への書き込みが正常に行われるかも明らかでないため、業務を引き継ぐ副システムにそのような書き込みを反映させないためである。

【００３６】

システム切り替え第２ステップ２０２：未実行の業務処理があればその業務を実行してディスクドライブ１６のデータ更新を行うようストレージ接続装置１３からサーバ・ストレージ間接続インタフェース１４を通じてディスク制御装置１５に情報の書き込み要求を通知する。ただし、正システムのデータベースサーバ２からディスクドライブ７へのログ書き込みが、トランザクション単位で行われなかった場合、ディスクドライブ１７上のログがトランザクションの途中で途切れていることも起こりうる。そのような場合は、データベースサーバ１２が途切れたトランザクションによるデータ更新を取り消すためのデータ

更新が必要になる。
【００３７】
システム切り替え第３ステップ２０３：ディスク制御装置１５は、要求を受けた情報の書き込みをディスクドライブ１６に行う。
【００３８】
システム切り替え第４ステップ２０４：ディスクドライブ１６の情報書き込みが完了したら、データベースサーバ１２で業務要求を受け付けて業務処理を開始する。
【００３９】
さらに、正システムが障害・災害から回復し再び動作するようになった場合や保守作業完了で正システムが動作可能になった場合、本実施形態で説明してきたデータ複製方法を、副システムから正システムに複製する方向に適用することで、正システムが停止中に副システムで実行した業務処理によるデータやログの更新を正システムに反映させることができる。
【００４０】
例えば、ストレージ装置８，１８の間でディスクドライブ１７の更新部分をディスクドライブ７にコピーする設定を行い、データベースサーバ２でディスクドライブ７の更新ログにしたがってディスクドライブ６のデータを更新することで、副システムのデータ複製を実行可能である。このように、正システムと副システムが同時に停止することがなければ、交互に本発明のデータ複製方法を適用することで業務停止時間を小さくできる。
【００４１】
また、本実施形態では正システムと副システムが一対一の形態を説明したが、正システムから複数の副システムへのデータ複製を行う方法や、正システムから副システムへ複製したデータをさらに別の副システムへデータ複製を行なう方法も容易に構築可能である。
【００４２】
図３から図６に、本実施形態の主な構成要素であるデータベースサーバおよびストレージ装置の処理手順をフローチャートで示した。以下で各図のフローチャートについて説明する。
【００４３】
図３に示した正システムデータベースサーバ処理手順のフローチャートについて説明する。
【００４４】
まず、データベースサーバの初期設定を行なう（３０１）。例えば、初期設定には、データベースの構築やディスクドライブの割り当てなどがある。
【００４５】
次に、データ複製システムを構築するまで、ストレージ装置と副システムの初期設定完了を待つ（３０２）。例えば、ストレージ装置間のディスクコピーの設定や副システムのデータベース構築の完了を待つことになる。
【００４６】
データ複製システムの初期設定が一通り完了すると、業務処理要求受付を開始する（３０３）。例えば、インターネット経由で行う商取引の商品管理などが業務処理にあたる。
【００４７】
業務処理要求の受付開始後が、業務処理要求が到着したか（３０４）判定する。
業務処理要求が到着していれば、業務処理を実行する（３０５）。業務処理要求が到着していなければ業務処理要求を待ち、業務処理要求の到着を判定を続ける。業務処理要求が到着し、業務処理がデータ更新を伴う場合、業務処理のログをストレージ装置に書き込む（３０６）。そして、ログの書き込み完了報告を受信する（３０７）。その後でデータの更新要求をストレージ装置に送る（３０８）。
【００４８】
ストレージ装置に対して行なったデータの更新書き込み要求について、ストレージ装置からの書き込み完了報告を受信する（３１０）ことで業務処理要求が完了する。

【００４９】
正システムは一度稼動すると、ここで説明したように、業務処理要求の実行とそれに伴うログの更新およびデータの更新を繰り返し行なう。
【００５０】
図４に示した副システムデータベースサーバ処理手順のフローチャートについて説明する。
【００５１】
正システム停止後に交替して業務処理を実行するのがデータ複製の目的であるため、副システムのデータベースサーバには正システムの設定にあわせた初期設定を行う（４０１）。例えば、コピーするログディスクやデータディスクの用意などがデータ複製のためには必要となる。
【００５２】
次に、正システムからコピーされたログを参照するため、ストレージ装置のログディスク更新を検出する設定をする（４０２）。例えば、副システムのデータベースサーバから副システムのストレージ装置にログディスクの更新を通知するように指定する方法や、副システムのデータベースサーバから定期的にストレージ装置内のログディスクを読み込んで更新が行われたかを判定する方法などが考えられる。これらにより、ログの更新を検出できる。
【００５３】
そして、ログが更新されるとその内容を副システムで実行してデータベースのデータも更新する、データ複製処理を開始する（４０３）。
【００５４】
システム切り替えが必要かを判定するため、正システムが正常に稼働しているか（４０４）判定する。例えば、正システムから副システムに対して10秒間稼働状態の通知がない場合にシステムを切り替えるとか、正システムと副システムのデータベースサーバ以外の稼働状態監視サーバを業務ネットワークに接続して状態監視をさせることによって正システムが稼働していないと判断した場合は業務処理を副システムが引き継ぐというように方針決めておき、判定を実行することになる。なお、このように正システムの稼働状態を判定する方法は複数考えられ、ここに示した方法に限定されるものではない。
【００５５】
もし正システムが正常に稼働していないと判定した場合には、システム切り替えの処理を行なう。まず、正システムの業務引継処理を実行する（４１０）ことで、業務ネットワークと接続可能としてから、ログディスクの更新分で未実行の業務処理を実行する（４１１）。そして実行の結果、データの更新をストレージ装置に送る（４１２）。ストレージ装置のデータ更新完了報告を受信する（４１３）ことで、正システムのデータ複製が完了したとみなす。そして、ログディスクの未実行業務処理のデータがすべてストレージ装置に反映されたら、業務処理要求受付を開始する（４１４）。
【００５６】
もし稼動状態が正常であると判定した場合は、ストレージ装置からログディスクの更新を検出したか（４０５）判定する。ログディスクの更新を検出する方法は、ストレージ装置からデータベースサーバに割り込みを発生させる方法や、データベースサーバからストレージ装置へ発行する特殊なＩ／Ｏコマンドへの応答を返す方法、あるいはデータベースサーバが一定時間間隔でログディスクの内容を読み込み、その情報を解析する方法などがある。ログディスクの更新を検出したら、ログディスクの更新分を読み込む（４０６）。そして、そのログ更新分を適用してデータ更新を実行する（４０７）。さらに、データの更新実行によって発生するデータの更新をストレージ装置に送る（４０８）ことで、データが正システムの最新のものと一致するようにする。ストレージ装置のデータ更新完了報告を受信する（４０９）と再び状態通知の受信やストレージ装置の更新通知待ちの処理を繰り返してデータ複製をしながら、システム切り替えの準備をする。
【００５７】

図５に示した正システムストレージ装置処理手順のフローチャートについて説明する。
【００５８】
まず、ストレージ装置内のディスクドライブをデータベースサーバに割り当てるなどの、初期設定を行う（５０１）。
【００５９】
そして、本実施形態のデータ複製方法を行なうため、正システムのログディスクを副システムのログディスクに対応付けしコピーの設定をする（５０２）。この設定を行なう前に、副システムのデータベースサーバとストレージ装置の初期設定を完了しておく必要がある。
【００６０】
設定が完了したら、読み込み・書き込み処理を開始（５０３）し、データベースサーバからのデータ更新要求などを受け付ける状態になる。
【００６１】
処理要求受信（５０４）を待つ状態から要求を受信すると、まず書き込み要求か（５０５）判定する。書き込み要求でなければ、読み込み要求された情報をデータベースサーバに転送（５１１）し、データベースサーバに情報読み込み完了報告を送信する（５１２）。実際はディスクドライブのコントロールなどの要求も受信するが、ここでは読み込み要求と同じものとみなしている。書き込み要求を受信した場合には、要求された情報をディスクに書き込む（５０６）処理を行い、そのディスクがコピーを設定したディスクか（５０７）判定する。コピー設定されていなければ、データベースサーバにデータ書き込み完了報告を送信する（５１０）。コピーを設定したディスクであれば、副システムのストレージ装置に書き込み要求と情報を転送（５０８）し、副システムのストレージ装置から書き込み完了報告を待つ（５０９）。副システムから完了報告を受け取るとデータベースサーバに情報書き込み完了報告を送信する（５１０）。ここでは、正システムと副システムのストレージ装置間で同期コピーを行なう方法としている。
【００６２】
このように、正システムのストレージ装置はデータベースサーバからの処理要求を待ち、ディスクの情報読み込み・書き込み処理と副システムへのディスクコピー処理を繰り返す。
【００６３】
図６に示した副システムストレージ装置処理手順のフローチャートについて説明する。
【００６４】
まず、ストレージ装置内のディスクドライブをデータベースサーバに割り当てや外部ストレージ装置からのディスクコピー設定などの、初期設定を行う（６０１）。
【００６５】
そして、読み込み・書き込み処理開始（６０２）後、データベースサーバからの要求を受信可能な状態となる。さらに、データベースサーバから更新通知するディスクを指定される（６０３）ことでデータ複製の準備が整う。
【００６６】
処理要求受信（６０４）を開始し、処理要求を受けるとその要求が書き込み要求か（６０５）判定する。書き込み要求でなければ、読み込み要求された情報を要求元に転送（６１０）し、要求元に情報転送完了報告を送信する（６１１）ことで処理要求の実行が完了する。一方、書き込み要求であった場合、まず、要求された情報をディスクに書き込む（６０６）。そして、要求元に情報書き込み完了報告を送信する（６０７）。通常の処理要求はこれで処理が完了するが、データ複製方法を実施するためにディスクへの書き込みが発生した場合は、そのディスクがデータベースサーバに更新通知を指定されたディスクか（６０８）判定し、指定されたディスクであれば、データベースサーバに更新を通知する（６０９）。指定されていなければ、通知せずに処理を完了する。
【００６７】
このように、副システムのストレージ装置は正システムからのログディスク書き込みと副

システムのデータベースサーバからの読み込み・書き込み要求を処理し、更新を通知するディスクであればその通知を送信する処理を繰り返して、データ複製処理を実現する。

【００６８】

本実施形態のように正システムから副システムに切り替えるが、上記のようなデータ複製方法を用い、特にストレージ装置８，１８間のディスクコピーが同期コピー方法で行うとデータ複製に必要なログを漏れなくコピーできるため、システム切り替えに伴う遅延を小さくすることができる。

［第２実施形態］

第１実施形態では、ディスクドライブの更新をデータベースサーバに通知するために通常のディスクドライブ読み込み書き込み以外のインタフェースを必要としたが、以下で説明する第２実施形態においてはディスクドライブの読み込み書き込みインタフェースのみでデータ複製を実現する。データ複製システムの構成は、図１に示す第１実施形態と同様に構成される。

【００６９】

第１実施形態では、副システムにおいてディスクドライブ１７の更新をデータベースサーバ１２に通知するための設定をした。これに対し、本実施形態ではデータベースサーバ１２からディスクドライブ１７をポーリングで監視し、更新を検知する。

【００７０】

更新の検知は以下のような手順で可能である。データベースサーバ１２で、ディスクドライブ１７にログが書き込まれる位置を保持し、その位置のデータを定期的に読み込んで更新されたかを判定する。更新されていれば、ログに従って業務処理をデータベースサーバ１２で行う。処理が完了したら、ログが書き込まれる位置の更新を行い、再び定期的に読み込んで更新されたかの判定処理を繰り返す。

【００７１】

データベースの更新ログは、通常一定の領域に順次上書きされないようにディスクに書き込む。そして、領域の終端まで書き込むと再び領域の先頭から順次書き込む。そのため、更新ログの書き込みが上書きされる前にデータベースサーバで内容を読み込むことができることと、更新ログを１つずつ区別することができるようになっていることが保証される場合、ポーリングで監視してデータのディスクドライブを更新することでデータの複製が可能である。

【００７２】

このように、データベースサーバからディスクドライブの更新をポーリングによって監視する方法でデータ複製する場合、ポーリング間隔を十分小さくすることによって、システム切り替えによる遅延を小さくすることができる。また、第１実施形態と同様のシステム構成であり、データ複製にかかるコストを小さくできる。さらに、ストレージ装置８、１８間のディスクコピーを同期コピー方法で行うとデータ複製に必要なログを漏れなくコピーできるため、システム切替に伴う遅延を小さくすることができる。

［第３実施形態］

第１実施形態、第２実施形態では、データベースサーバ２，１２とストレージ装置８，１８が直接接続された場合やストレージエリアネットワークで接続されたことを前提としていたが、本実施形態では、ストレージ装置としてネットワークアタッチトストレージ（ＮＡＳ）装置を使用して実現する。本実施形態のデータ複製システムの構成は、図１に示す第１実施形態と同様の構成である。

【００７３】

本実施形態では、ストレージ装置間のディスクコピー方法が第１実施形態、第２実施形態と異なる。ＮＡＳ装置はファイルシステムでのアクセス要求を受信する。そのため、ディスク制御装置５，１５間のストレージ間接続インタフェースもファイル単位でのアクセスを実行する。そのため、ディスク制御装置５内でディスクドライブ７の変更を検知するのではなく、データベースサーバ２が操作するログファイルの更新を検知する必要がある。更新の検知には、ディスク制御装置５でログファイルの更新を定期的に監視するデーモン

を実行しておき、更新が起こったらストレージ間接続インタフェースを通じてファイルのコピーをディスクドライブ１７に書き込む。また、副システムでログの更新を通知するインタフェースも、前述の第１実施例、第２実施例のものと異なる。ディスク制御装置１５にはログファイル更新を検知するデーモンを備える。データベースサーバ１２にはこのデーモンと通信を行うプロセスを生成しておき、つまり、ログファイルの更新があったら通知される機構を構築する。あるいは、第２実施形態で示したデータベースサーバ１２からポーリングで監視する方法をログファイルに適用してもよい。

【００７４】

また、ログデータのコピー方法としてディスク制御装置５と別のディスクドライブ単位で変更を検知可能なディスク制御装置をストレージ装置８，１８に設け、その間でディスクの更新を行う方法も可能である。この場合、別途設けたディスク制御装置により、第１実施形態のようにディスクドライブ７からディスクドライブ１７へのデータコピーを実行する。ログデータの更新をデータベースサーバ１２に通知する方法は、上記のようなディスク制御装置１５のデーモンとデータベースサーバ１２のプロセスで通信を行う方法やデータベースサーバ１２からポーリングで監視する方法によって可能である。

【００７５】

また、ＮＡＳ装置ではログファイルをコピーする際に、ファイルを全てストレージ装置８，１８間で転送する必要がある。データ転送量を削減するため、ログをおくためのディレクトリを作成し、更新ログを１個ずつファイルとしてそのディレクトリに置いていくことで更新ログの転送量を削減可能である。データベースは更新ログを作成した日時をファイル名に使用して書き込む。これによりログの一意の識別か可能になる。また、一定以上の時間が経過したログを削除することで、ログの複製を保証することができ、ディスクドライブを使い尽くすことがなくなる。ログデータの更新のデータベースサーバ１２への通知は、上記同様であるが、ディスク制御装置１５で実行するログ更新を監視するデーモンは、ログディレクトリの下にあるファイルの監視を行い、新たなログファイルが作成された場合にデータベースサーバ１２のプロセスに通知を行う。データベースサーバ１２でファイル更新をポーリングで監視するデーモンを実行する場合も、同様にログディレクトリの下に新たなファイルが作成されたかを監視する。

【００７６】

このようにシステムを構築することで、データ複製が実現される。システム構成は第１実施形態と同様であり、コストを削減することが可能である。また、システム切り替え時の遅延も小さくすることが可能である。

［第４実施形態］

図７は本発明が適用されたデータベースサーバとストレージ装置およびそれらの上で動作するソフトウェアを用いたデータ複製システムの実施例である。本実施例では、サーバ上で動作するアプリケーションプログラムをＤＢＭＳとしているが、データの更新をログとして出力するプログラムであれば本発明は適用可能である。例えばトランザクションモニタであってもよい。

【００７７】

データ複製システムの構成は、図１に示す第１実施形態と同様に構成されるが、図７においてはデータベースサーバ２、１２上のストレージ接続装置３、１３は省略した。データベースサーバ２、１２上では、それぞれＯＳ７２、７５とＤＢＭＳ７１、７４とが動作する。ＯＳ７２、７５はデータベース２、１２のハードウェア制御や他のアプリケーションプログラムの動作環境として実行される。また、ＤＢＭＳ７１、７４はシステムの業務を実行するアプリケーションプログラムである。さらに、ストレージ装置８、１８上のディスク制御装置５、１５上ではデータベースサーバ２、１２からの要求を受付けてディスクドライブ６、７、１６、１７の更新を行う制御ソフトウェア７３、７６が動作する。第１実施形態でストレージ接続装置３、１３を通して行っていた処理は、ＤＢＭＳ７１、７４からはＯＳ７２、７５の機能によって内部で処理されているものとする。

【００７８】

本実施形態では第２実施形態と同様のシステム構成で、各装置とそれら装置の上で動作するソフトウェアによってデータ複製を実現する。つまり、ストレージ装置１８からデータベースサーバ１２へと、ディスクドライブ１６、１７の情報が変更されたことを通知するインタフェースは持たない。

【００７９】

正システムを構成するデータベースサーバ２上で動作するＯＳ７２はストレージ装置８内のディスクドライブ６、７に情報を書き込んだり、あるいはその情報を読み込んだりする操作を実行できる。制御ソフトウェア７３はサーバ・ストレージ間接続インタフェース４を通じて送られるＯＳ７２からの要求に応じて情報のディスクドライブ６、７への反映や情報のＯＳ７２への転送などの処理をする。そして、ＤＢＭＳ７１は、ＯＳ７２の上で動作し、システムの業務を行う。副システムはデータベースサーバ１２とストレージ装置１８、およびこれらの上で動作するＯＳ７５、ＤＢＭＳ７４そして制御ソフトウェア７６で正システムと同様の処理を行うことが可能な構成とする。

【００８０】

制御ソフトウェア７３、７６の間では、ストレージ間接続インタフェース２０を通じて互いの保持するディスクドライブの内容を転送し、指定されたディスクドライブに情報を反映することが出来る。例えば、ストレージ装置間のリモートコピーや同期リモートコピーを用いることで実現できる。

【００８１】

まず、正システムと副システムの初期設定を行う。正システムでＯＳ７２が認識したストレージ装置８内のディスクドライブ６、７をＤＢＭＳ７１のデータディスクとログディスクに割当てる。副システムでは、正システムのディスク構成にあわせて、ＯＳ７５が認識したストレージ装置１８内のディスクドライブ１６をデータディスクに、ディスクドライブ１７をログディスクにＤＢＭＳ７４が割当てる。また、ＤＢＭＳ７１とＤＢＭＳ７４との間で互いの稼動状態をチェックする。正システムが停止したときに、迅速に副システム業務を引き継ぐためである。

【００８２】

ストレージ装置８、１８の間の初期設定は、ディスクドライブ７の情報の変更を、ディスクドライブ１７にコピーするようにする。これによって、ＤＢＭＳ７１で実行した処理のログを副システムにコピーし、ＤＢＭＳ７４から参照できるようにする。このコピーの方式は同期、非同期いずれとも可能であるが、非同期の場合、正システムが停止した時点までの全てのログを転送できる保証はない。一方、同期とした場合、正システムが停止するまでのログが完全に副システムで参照可能となり、高速で正確なデータ複製が可能である。

【００８３】

ここまでの設定が終わったら、以下で説明する手順を実行することで正システムから副システムへのデータ複製が可能である。そこで、正システムで業務を開始する。ＤＢＭＳ７１は業務ネットワーク１からの業務処理要求を受付けて処理する。このときの処理の内容を逐一再生可能なものがログにあたり、ＤＢＭＳ７１からＯＳ７２の機能を経由して制御ソフトウェア７３にディスクドライブ７に書き込みように要求する（７０２）。ただし、業務処理要求は業務ネットワークから受け取るだけではなく、データベースの構築や保守のためにデータベースサーバ２上で実行した処理についても業務処理要求に含まれる。そして、完全なデータ複製を行うためには、これら全ての業務処理要求についてログを作成しディスクドライブ７に書き込みを行う必要がある。

【００８４】

ＤＢＭＳ７１がディスクドライブ７にログの書き込みを要求する単位には、例えばトランザクションをコミットした単位とする方法がある。あるいは、ＤＢＭＳ７１が持つログ用のバッファが一杯になった場合や一定時間経過した場合にＤＢＭＳ７１が持つログ用のバッファの内容をディスクドライブ７に書き込み要求する方法などがある。いずれの場合にも、ＤＢＭＳ７１が実行した業務のログが副システムのＤＢＭＳ７４に欠損なく、またＤ

ＢＭＳ７１、７４それぞれが管理するデータの間で不整合が生じることなく複製するためには、ＤＢＭＳ７１がディスクドライブ７に要求した単位でディスクドライブ７とディスクドライブ１７間の情報が同期していることが必要である。以下では、ディスクドライブ７の情報とディスクドライブ１７の情報とが同期しているものとする。しかし、必ずしも同期している必要は無く、非同期でディスクドライブ７からディスクドライブ１７へのコピーを行ってもよい。但し、非同期コピーを行った場合、正システムと副システムのログが一致することは保証されず、正システムから副システムへ切り替える時にＤＢＭＳ７１とＤＢＭＳ７４が扱うデータに不整合があり得るため、正システムと副システム間でデータ内容を確認したり修正したりする作業が必要になることがある。

【００８５】

要求を受けた制御ソフトウェア７３は、要求をログ用のディスクドライブ７に書き込む（７０３）。このディスクドライブ７は初期設定でストレージ装置１８のディスクドライブ１７にコピーすると指定しているため、制御ソフトウェア７３はストレージ装置間接続インタフェース２０を通して制御ソフトウェア７６に変更分の書き込み要求を転送し、これを受けた制御ソフトウェア７６がディスクドライブ１７に要求された情報を書き込む（７０４）。ここで、ディスクドライブ７とディスクドライブ１７の間で同期コピーを行うためには、ディスクドライブ１７への情報書き込みが完了してからデータベースサーバ２に完了報告を行い、その報告を受けた後でＤＢＭＳ７１が業務処理の続きを行うようにすることが必要である。こうすると、ＤＢＭＳ７１の実行した業務のログと副システムに複製しているログとが一致する。一方、ディスクドライブ７への書き込みが完了した時点でデータベースサーバ２に書き込み完了報告を行い、ディスクドライブ１７へのコピーをＤＢＭＳ７１からの要求と無関係の任意の時点で行うと、非同期コピーとなり副システムで参照するログが正システムのものと一致しないことが起こりうる。

【００８６】

さらに、ＤＢＭＳ７１は業務処理によって生じるデータの変更をＯＳ７２の機能を通じて制御ソフトウェア７３に送る（７０４）。そして制御ソフトウェア７３は要求された情報書き込み処理や読み込み処理をディスクドライブ６に対して行う（７０５）。ただし、ＤＢＭＳ７１からのディスクドライブ７へのログ書き込み要求とディスクドライブ６へのデータ書き込み要求は順序が決められているわけではない。データ書き込みはＯＳ７２上のデータバッファに蓄積して定期的に要求する方法や一定時間以上業務処理で使用しなかったら書き込む方法などがある。正システムの実行した業務を追うためには、実行した処理のログは副システムに必ず転送しなければならないが、データについては正システムが業務に使用する範囲で整合性が保たれていればよい。

【００８７】

一方、副システムのＤＢＭＳ７４はＯＳ７５の機能を使用して制御ソフトウェア７６に要求を出し（７０６）、ディスクドライブ１７の情報が変更されていないか情報を読み込む（７０７）。この結果、ディスクドライブ１７上のログ内容に変更があったら、そのログの業務を実行し、副システムのディスクドライブ１６のデータを更新するため、制御ソフトウェア７６に情報更新要求を発行する（７０８）。この要求を受けた制御ソフトウェア７６はディスクドライブ１６に情報変更の内容を反映する（７０９）。これらの処理を一定時間ごとに行い、ログの処理を実行してデータを更新して正システムのデータを副システムに複製する。あるいは、実施例１のように、サーバ１２からストレージ制御装置１５あるいは制御ソフトウェア７６に指示しておき、ディスクドライブ１７の情報が更新されたらそれをサーバ１２に通知するインタフェースによりログの変更を認識する方法でも可能である。

【００８８】

ＤＢＭＳ７４がディスクドライブ１７からログの変更を読み込んでディスクドライブ１６にあるデータに適用し、データを複製する方法は複数ある。例えば、ＤＢＭＳ７１からディスクドライブ７への書き込みをトランザクションのコミットした単位で行っていれば、ＤＢＭＳ７４はディスクドライブ１７で新たに書き込まれたログをそのまま実行すればよ

い。また、ＤＢＭＳ７１からディスクドライブ７への書き込みがトランザクションのコミットと無関係に実行されている場合は、書き込まれたログをそのまま実行する方法と、ＤＢＭＳ７４でトランザクションの管理テーブル上に確保しておき、そこに読み込んだログを保持しながらトランザクションのコミットを受け取ってから管理テーブルにあるログをデータに反映する方法などがある。

【００８９】

もしも、ＤＢＭＳ７１とＤＢＭＳ７４との間で行っている稼動状態監視で、ＤＢＭＳ７４がＤＢＭＳ７１の停止を検出した時には、ＤＢＭＳ７１が行っていた業務を引き継ぐように設定を変更する（７１０）。ただし、ＤＢＭＳ７４が業務処理を引き継ぐ前に、それまでにＤＢＭＳ７１で実行された正システムの業務データと副システムのデータとの間で整合性を保証するための処理、つまりリカバリを行う必要がある。ディスクドライブ７とディスクドライブ１７の間のログ同期がトランザクションのコミット単位で行われていれば、ディスクドライブ１７に副システムで未実行のログがあるかを確認し、あればそのログの処理を実行して制御ソフトウェア７６に情報更新要求を発行する。（７０８）そして、制御ソフトウェア７６は要求にディスクドライブ１６上の情報を更新する（７０９）。ディスクドライブ１７の未実行のログ全てについてディスクドライブ１６上のデータ更新を実行したらリカバリが完了し、業務ネットワーク１からの業務処理要求をサーバ１２で受けてＤＢＭＳ７４が業務処理を開始する。

【００９０】

上記の、システム間で業務処理を引き継ぐ前に業務データの整合性を保証するために実行する処理、つまり、リカバリはディスクドライブ１７へのログ書き込みがコミット単位であることが保証される場合と、保証されない場合とで異なる。ここで、コミット単位での書き込みが保証されない場合とは、例えば、チェックポイント処理によりコミット前の情報を書き込む場合や、ＤＢＭＳ７１の持つバッファがあふれてコミット前の情報が書き出される場合などがある。以下、コミット単位での書き込みが保証される場合とされない場合、それぞれについてリカバリの方法を説明する。

【００９１】

まず、コミット単位でのログ書き込みが保証される場合のリカバリ処理について説明する。コミット単位での書き込みが保証される場合は、ディスクドライブ１７に書き込まれたログを全てデータに反映すればよい。したがって、ディスクドライブ１７で未実行であるログを認識し、該ログをロールフォワードすることによりリカバリが実現する。同期コピーを用いた場合は、正システムでコミットした情報は副システムに全てコピーされたことが保証され、かつ、該リカバリ処理によりコピーされたログを漏れなく適用できるため、本システムにおいては、トランザクション欠損が無いことが保証される。

【００９２】

コミット単位での書き込みが保証されない場合は、ディスクドライブ７にはコミット済みのトランザクションログとコミット未済みのトランザクションログとが含まれる。ここで、整合性を保証すべきなのは、コミット済みトランザクションだけである。なぜなら、コミット未済みの処理による変更は、正システムのデータに反映されていないからである。そのため、リカバリはコミット済みのトランザクションについてのみ実行し、未済みのトランザクションは実行しない。このようなリカバリ方法として、以下の３つの方法が考えられる。

【００９３】

例えば、ディスクドライブ７に書かれたログを全てロールフォワードした後、未コミット分を検出し、検出した未済みのトランザクションのみをロールバックする方法が考えられる。例えば、ログを終端から先頭に向かってスキャンし未コミットのトランザクションを検出する方法がある。あるいは、あらかじめトランザクションの状態を管理するテーブル（以降、トランザクション管理テーブルとする）をＤＢＭＳ７４が用意し、正システムのＤＢＭＳ７１の停止を検出する前からディスクドライブ１７へのログ書き込みが行われてロールフォワードをするたびにその状態を変更しておく方法がある。後者の場合、リカバ

リするときにトランザクション管理テーブルに登録されている未コミットのトランザクションをロールバックすればよい。

【００９４】

また、正システムのＤＢＭＳ７１が停止したことを検出した時点以降にログの変更を検出した場合については、ロールフォワードを行う前に、トランザクションの状態を調査してからロールフォワードを行う方法も考えられる。コミット済みのトランザクションのロールフォワードが終了したらＤＢＭＳ７１が停止したことを検出した時点より以前について、コミット未済みのトランザクションを調査して、コミット未済みのトランザクションが存在したらロールバックを行う。コミット未済みのトランザクションの調査はＤＢＭＳ７１が停止したことを検出した時点からログの終端までスキャンしたり、管理テーブルの情報を用いるなどして行えばよい。

【００９５】

また、コミットが確定したトランザクションのみロールフォワードする方法も考えられる。例えば、上記管理テーブルにおいてコミット済み状態になった時点で、該トランザクションのロールフォワードをする方法が考えられる。

【００９６】

以上、コミット単位でのログ書き込みが保証されない場合の３つのリカバリ方法を同期コピーとともに用いると、正システムでディスクドライブに書き込まれた情報は副システムにコピーされたことが保証され、かつ、該リカバリ方法により、コミット済みのトランザクションはもれなくロールフォワードされるため、本システムにおいてはトランザクション欠損がないことが保証される。

【００９７】

このように、正システムで実行した業務のログを副システムにコピーして、そのログから業務データを復元する処理を行うことで、正システムのデータベースサーバの処理負荷を上げることなく、また正システムと副システムの間で情報を転送するための通信帯域を抑制して低コストのデータ複製が可能となる。また、正システムから副システムへログを同期コピーで転送すると、業務処理の欠損を起こすことなくデータ複製を完了し、正システムの障害発生時や保守により正システムを停止しなければならない時、迅速に副システムに業務の実行を切り替えることが可能である。

【００９８】

【発明の効果】

本発明によれば、データベースサーバとストレージ装置からなる複数のシステム間で、低コストかつ通常業務の遅延が小さいデータ複製を実現する。また、システム切り替え時の遅延を小さくすることができる。

【図面の簡単な説明】

【図１】正システムと副システムとからなるデータ複製システムおよび正・副システム間のデータ複製方法の概念図である。

【図２】正システムと副システムとからなるデータ複製システムおよび正・副システム間のシステム切り替え方法の概念図である。

【図３】正システムのデータベースサーバが行う処理手順のフローチャートである。

【図４】副システムのデータベースサーバが行う処理手順のフローチャートである。

【図５】正システムのストレージ装置が行う処理手順のフローチャートである。

【図６】副システムのストレージ装置が行う処理手順のフローチャートである。

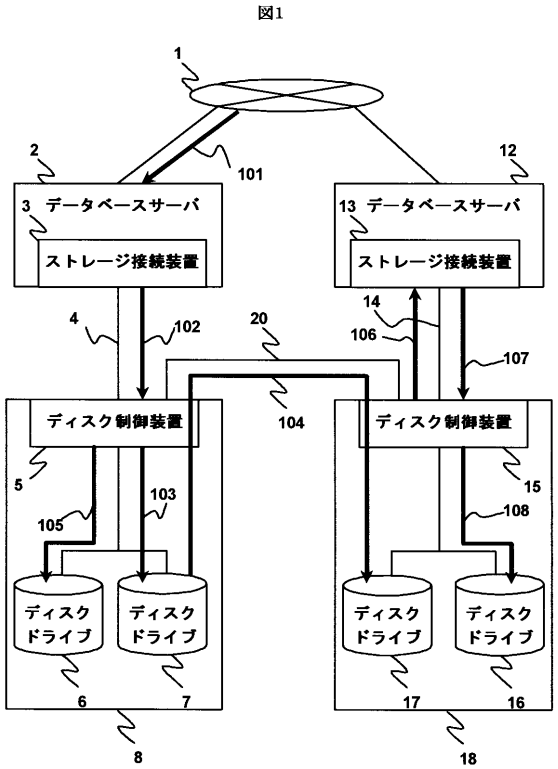
【図７】正システムと副システムとからなるデータ複製システムおよび正・副システム間のシステム切り替え方法の概念図である。

【符号の説明】

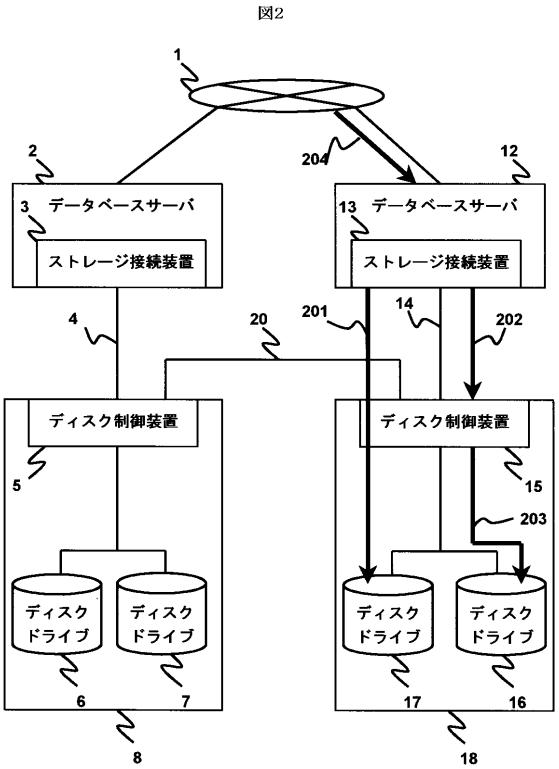
１：業務ネットワーク、２：正システムのデータベースサーバ、３：ストレージ接続装置、４：サーバ・ストレージ間接続インタフェース、５：ディスク制御装置、６：データディスクドライブ、７：ログディスクドライブ、８：正システムのストレージ装置、１２：副システムのデータベースサーバ、１３：ストレージ接続装置、１４：サーバ・ストレー

ジ間接続インタフェース、１５：ディスク制御装置、１６：データディスクドライブ、１７：ログディスクドライブ、１８：副システムのストレージ装置、２０：ストレージ装置間接続インタフェース。

【図１】



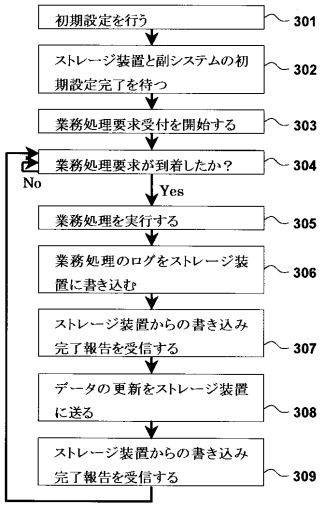
【図２】



【図３】

図3

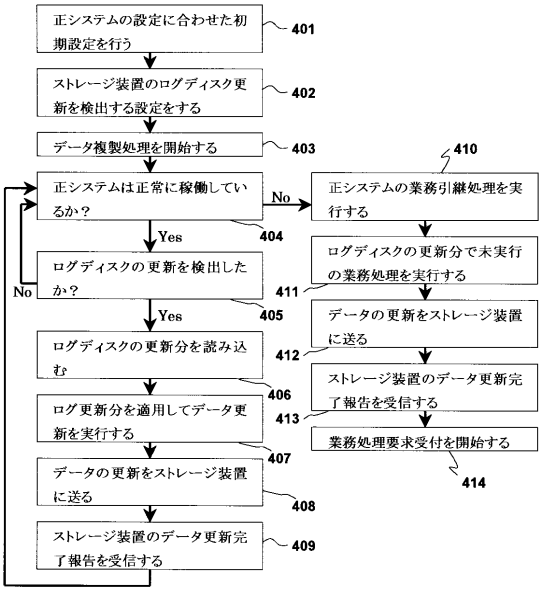
正システムデータベースサーバ処理手順



【図４】

図4

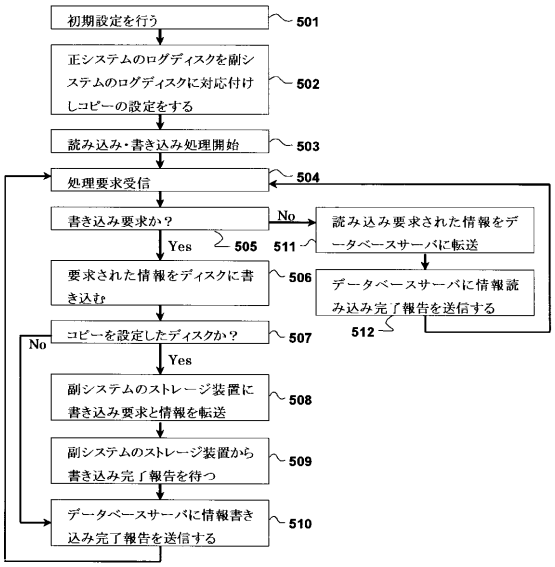
副システムデータベースサーバ処理手順



【図５】

図5

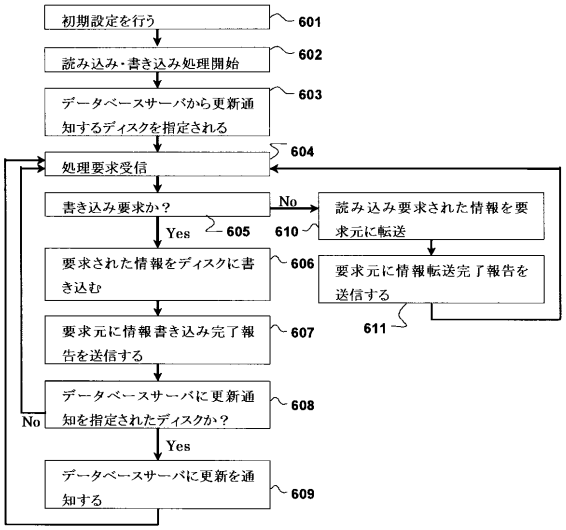
正システムストレージ装置処理手順



【図６】

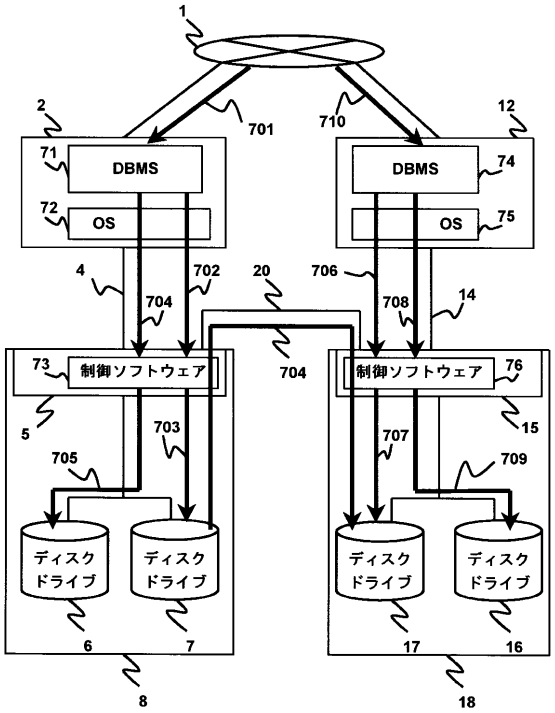
図6

副システムストレージ装置処理手順



【図７】

図7


1
2
701
710
12
71
DBMS
74
DBMS
72
OS
75
OS
4
702
20
706
704
708
14
73
制御ソフトウェア
704
制御ソフトウェア
76
5
703
707
15
705
709
ディスク
ドライブ
ディスク
ドライブ
ディスク
ドライブ
ディスク
ドライブ
6
7
17
16
8
18

フロントページの続き

(56)参考文献　森側　真一，[動向]多様化する災害対策，日経オープンシステム　第１０５号，日本，日経ＢＰ社，２００１年１２月１５日，p.104-107

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