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

特許第4546629号
(P4546629)

(45)発行日　平成22年9月15日(2010.9.15)　(24)登録日　平成22年7月9日(2010.7.9)

(51)Int.Cl.		F I		
G06F 12/00	(2006.01)	G06F	12/00	533J
G06F 3/06	(2006.01)	G06F	12/00	518A
G06F 12/08	(2006.01)	G06F	12/00	531D
G06F 12/16	(2006.01)	G06F	3/06	304B
		G06F	3/06	540

請求項の数 8　(全 21 頁)　最終頁に続く

(21)出願番号　特願2000-285237(P2000-285237)
(22)出願日　平成12年9月14日(2000.9.14)
(65)公開番号　特開2002-49517(P2002-49517A)
(43)公開日　平成14年2月15日(2002.2.15)
審査請求日　平成19年8月6日(2007.8.6)
(31)優先権主張番号　特願2000-159547(P2000-159547)
(32)優先日　平成12年5月25日(2000.5.25)
(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　100093861
弁理士　大賀　眞司
(72)発明者　占部　喜一郎
神奈川県小田原市国府津２８８０番地　株式会社日立製作所　ストレージシステム事業部内
(72)発明者　裏谷　郁夫
神奈川県小田原市国府津２８８０番地　株式会社日立製作所　ストレージシステム事業部内

審査官　桜井　茂行

最終頁に続く

(54)【発明の名称】記憶システム、記憶システムの応答方法及び記録媒体

(57)【特許請求の範囲】
【請求項１】
計算機と、第二のストレージ装置と、に接続された第一のストレージ装置であって、
プロセッサと、
前記プロセッサから参照される制御プログラムやデータが保存されたメモリと、を有し、
前記プロセッサは、
前記計算機から送信された複数の書き込み対象のデータに、前記計算機からの書き込み要求が発行された順に識別子を付与し、
前記書き込み対象のデータを、前記書き込み要求とは非同期に、前記第二のストレージ装置に送信し、
前記計算機によって実行されるアプリケーションプログラムのトランザクションごとに対するコミットに基づいて、前記計算機が発行した、前記第二のストレージ装置が前記書込み対象のデータを格納したかを確認するデータ格納確認コマンドを前記計算機から受信して、前記第二のストレージ装置のボリュームに書き込まれるデータの前記識別子を前記第二のストレージ装置に問い合わせ、
前記データ格納確認コマンドを受信するまでに、前記計算機から書き込まれたデータに付加した第１の識別子と、前記第二のストレージ装置から得た第２の識別子とを比較し、この比較結果に基づいて、前記データ格納確認コマンドに対する完了応答を、前記計算機へ送信する、ストレージ装置。

【請求項２】
前記識別子は、複数の前記書込み対象のデータの順序性を示すシーケンス番号である、請求項１記載のストレージ装置。
【請求項３】
前記第２の識別子は、前記第二のストレージ装置への前記問い合わせの直前に、前記第二のストレージ装置に格納されたデータに付与された識別子であり、
前記第１の識別子は、前記データ格納確認コマンドを受信する直前に、前記計算機から書き込まれた最新データに対する識別子である、請求項２記載のストレージ装置。
【請求項４】
前記アプリケーションプログラムが前記トランザクションのコミットポイントでコミットコマンドを送信した直後に、前記計算機から、前記データ格納確認コマンドを受信する、請求項１記載のストレージ装置。
【請求項５】
前記計算機は前記アプリケーションプログラムを複数備え、
前記複数のアプリケーションプログラムはグループ分けされており、
前記識別子は、前記複数のアプリケーションプログラムのグループを識別するための情報を含む、請求項２記載のストレージ装置。
【請求項６】
計算機に接続する第一のストレージ装置と、第二のストレージ装置と、の間のデータコピー制御方法であって、
前記第一のストレージ装置は、
前記計算機から送信された複数の書き込み対象のデータに、前記計算機からの書き込み要求が発行された順に識別子を付与し、
前記書き込み対象のデータを、前記書き込み要求とは非同期に、前記第二のストレージ装置に送信し、
前記計算機によって実行されるアプリケーションプログラムのトランザクションごとに対するコミットに基づいて、前記計算機が発行した、前記第二のストレージ装置が前記書込み対象のデータを格納したかを確認するデータ格納確認コマンドを前記計算機から受信して、前記第二のストレージ装置のボリュームに書き込まれるデータの前記識別子を前記第二のストレージ装置に問い合わせ、
前記データ格納確認コマンドを受信するまでに、前記計算機から書き込まれたデータに付加した第１の識別子と、前記第二のストレージ装置から得た第２の識別子とを比較し、この比較結果に基づいて、前記データ格納確認コマンドに対する完了応答を、前記計算機へ送信する、リモートコピー制御方法。
【請求項７】
計算機に接続する第一のストレージ装置に、第二のストレージ装置へのデータコピーを実行させるためのプログラムが記録されたコンピュータ読取可能な媒体であって、
前記プログラムは、
前記計算機から送信された複数の書き込み対象のデータに、前記計算機からの書き込み要求が発行された順に識別子を付与するステップと、
前記書き込み対象のデータを、前記書き込み要求とは非同期に、前記第二のストレージ装置に送信するステップと、
前記計算機によって実行されるアプリケーションプログラムのトランザクションごとに対するコミットに基づいて、前記計算機が発行した、前記第二のストレージ装置が前記書込み対象のデータを格納したかを確認するデータ格納確認コマンドを前記計算機から受信して、前記第二のストレージ装置のボリュームに書き込まれるデータの前記識別子を前記第二のストレージ装置に問い合わせるステップと、
前記データ格納確認コマンドを受信するまでに、前記計算機から書き込まれたデータに付加した第１の識別子と、前記第二のストレージ装置から得た第２の識別子とを比較するステップと、

この比較結果に基づいて、前記データ格納確認コマンドに対する完了応答を、前記計算機へ送信するステップと、を前記第一のストレージ装置に実行させるものである、媒体。

【請求項８】

計算機と、

前記計算機に接続する第一のストレージ装置と、

前記第一のストレージ装置に接続する第二のストレージ装置と、

を備える、データコピー制御システムであって、

前記第一のストレージ装置は、

プロセッサと、

前記プロセッサから参照されるプログラムやデータが保存されたメモリと、を有し、

前記プロセッサは、

前記計算機から送信された複数の書き込み対象のデータに、前記計算機からの書き込み要求が発行された順に識別子を付与し、

前記書き込み対象のデータを、前記書き込み要求とは非同期に、前記第二のストレージ装置に送信し、

前記計算機によって実行されるアプリケーションプログラムのトランザクションごとに対するコミットに基づいて、前記計算機が発行した、前記第二のストレージ装置が前記書込み対象のデータを格納したかを確認するデータ格納確認コマンドを前記計算機から受信して、前記第二のストレージ装置のボリュームに書き込まれるデータの前記識別子を前記第二のストレージ装置に問い合わせ、

前記データ格納確認コマンドを受信するまでに、前記計算機から書き込まれたデータに付加した第１の識別子と、前記第二のストレージ装置から得た第２の識別子とを比較し、この比較結果に基づいて、前記データ格納確認コマンドに対する完了応答を、前記計算機へ送信する、データコピー制御システム。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、ディスク制御装置間の非同期リモートコピーにおけるデータ同期制御に関するものである。

【０００２】

【従来の技術】

近年のコンピュータシステムは、銀行及び証券業務の基幹業務を、大型コンピュータによる一括管理から、クライアント・サーバシステムを中心とする分散システムへ移行している。このような分散システム環境では、クライアントからの要求を複数のサーバとディスクアレイ装置を用いてデータを処理するＨＡ（Ｈｉｇｈ　Ａｖａｉｌａｂｉｌｉｔｙ:高可用性）クラスタ構成が採られている。このようなＨＡクラスタ構成では、地震などの災害に備えて遠隔地にあるデータセンター間でデータを二重化する方法が採用されてきている。二重化は、通常、２台のディスクアレイ装置（記憶システム）を公衆回線や専用回線等を介してリモート接続し、ホストコンピュータ装置からローカルのディスクアレイ装置への書込みデータを、リモートのディスクアレイ装置へコピーする方法が採られている。

【０００３】

ディスクアレイ装置間で二重化する方法は、大別すると同期方式と非同期方式の２種類がある。

【０００４】

同期方式では、まず、ローカル側のホスト装置からの書込み要求を、ローカル側のディスクアレイ装置のキャッシュに書込む。続いて、ローカル側のディスクアレイ装置は、キャッシュに書込まれたデータを、リモートのディスクアレイ装置に転送する。更に、ローカル側のディスクアレイ装置は、書込み要求のデータがリモート側のディスクアレイ装置により受信されたこと示す応答信号を受信すると、この後、ホストに対して、前記書込み要求に対する応答を返す。つまり、ローカル側のディスクアレイ装置は、リモート側のディ

スクアレイ装置にデータが届けられたことを確認して、ホストへ応答を返す。この応答は、ホスト装置に対しリモート側にデータが渡されたことを保証する。（同期とは、ホストからの書込み要求とリモート側のディスクアレイ装置へのコピーが同期して行われるという意味で用いられる。）この同期方式は、リモート側からの応答信号を待つ遅延が発生するため、ローカルとリモート間でデータ伝送の伝播遅延が少ない比較的近距離（１００Ｋｍ以内）に適しているが、例えば公衆回線網等を使用する遠距離転送には適さない。なお、ローカル側およびリモート側のディスクアレイ装置に記録されたデータは、それぞれのドライブ制御回路を介してそれぞれの物理ディスクへ書込まれる。

【０００５】

一方、非同期方式は、遠距離転送に向いており、ローカル側のホスト装置からの書込み要求に対するホスト装置への応答（書込み完了）は、ローカル側のキャッシュに前記書込み要求のデータが書かれた時点で、書込み完了をホスト装置へ返される。キャッシュに書かれたデータは、ホスト装置への応答後に、別のタイミングで（この意味で非同期）、リモート側のディスクアレイ装置へコピー（転送）される。この非同期方式では、リモート側のディスクアレイ装置へデータを転送するタイミングとは関係なく、ホスト装置へ上記書込み要求に対する応答を行うので、応答タイミングが同期方式と比べて速くなり、ホストは次の処理に早く移ることができる。

【０００６】

なお、リモートコピーに関しては、特表平８－５０９５６５号に記載がある。

【０００７】

【発明が解決しようとする課題】

上記非同期にリモート側にデータを転送する方法では、ローカル側のディスクアレイ装置が、当該ディスクアレイ装置内にデータを格納した時点で、リモート側にデータがストアされたか否かに関係なく、ホストに対し、書込み要求完了を報告する。このため、ローカル側のホストは、ホスト書込み要求のリモート側への同期完了確認（ホスト書込み要求によるデータがリモート側のディスクアレイ装置へ確実に転送されたかの確認）が困難であった。このホスト書込み要求のリモート側での同期完了確認は、特に、データベースの履歴ログファイル等でのデータベースのトランザクション単位のコミット（データが確実にストレージに記憶されたという保証）で必要とされる。なお、コミットとは、１つのトランザクションに関する複数のデータベースの更新結果を実際の記憶システム上にログファイルと共に書込む一連の処理のことである。

【０００８】

また、災害時のデータリカバリの観点から正サイト障害によってディスクアレイ装置内に残存していた未転送のデータは失われ、副サイトに切替り副サイトで運用を開始した時に保証されるデータが分からない問題があった。

【０００９】

しかしながら、上記従来技術の非同期転送方式は、非同期転送の特性上から、ホストＩＯに対する同期確定方法を備えていなかった。即ち、データベース(ＤＢ)運用上必要となる、ＡＰＰ（アプリケーションプログラム）からのトランザクションに対するコミットポイントでの書込み要求がリモートサイト（副サイト）に確実に書かれたか否かを確認する方法が提供されていなかった。

【００１０】

以下、課題を具体的に説明する。まず、コンピュータが１つの記憶システムと接続される場合を説明し、その後、記憶システムが非同期データ転送（非同期リモートコピー）を行っている場合につき、その課題を明確に説明する。

【００１１】

初めに、コンピュータに１つの記憶システムが接続されている場合を説明する。コンピュータのアプリケーションが、書込みコマンド（要求）を実行すると、通常、コミット（commit）コマンド無しの状態では、その書込みコマンドのデータはコンピュータ内のデータバッファ上に書込まれるだけで、データバッファ上のデータと、記憶システム内のデー

タは一致しない。その後、アプリケーションがコミットコマンドを発行すると、データバッファ上のデータが記憶システムへ実際に書込みコマンドによって書込まれる。その後、記憶システムは、書込みデータをキャッシュメモリに記憶すると（この時点で記憶システム内のデータとコンピュータ内のデータは一致する）、書込みコマンドを発行したコンピュータに対し、書込み要求完了で応答する。コンピュータは、その書込み要求完了を確認すると、前記コミットコマンドに対するリターンをアプリケーションへ発行する。アプリケーションは、このリターンにより、記憶システム内のデータとコンピュータ内のデータが一致していることを知る。

【００１２】

次に、非同期リモートコピーを行っている場合について説明する。コンピュータのアプリケーションが、コミットコマンドを発行すると、データバッファ内のデータが書込みコマンドでローカル側の記憶システムのキャッシュへ書込まれる。ローカル側の記憶システムは、その応答として、書込み完了をコンピュータへ返す。コンピュータは、書込み完了を受けると、アプリケーションへコミットに対するリターンを返す。しかしながら、このリターンは、単にローカル側の記憶システムのデータがデータバッファ内のデータと一致したことを示すのみであり、リモート側の記憶システム内のデータとデータバッファ内のデータが一致したことを示すものではない。従って、ローカル側の記憶システムが書込み要求完了を返した後で、かつ、リモート側へのデータコピーが終了する前にローカル側の記憶システム内のデータが消失した場合に、アプリケーションがリモート側のデータを用いて処理を継続しようとすると、コミットのリターンを受け、記憶システム上でデータが確定したと確認したにもかかわらず、誤ったデータを用いて処理を継続することになる。即ち、非同期リモートコピー中に、もし、障害等が発生した場合には、従来からのコミット機能ではアプリケーションは満足な結果を得られない場合が発生することになる。

【００１３】

このように、従来技術の非同期リモートコピーは、非同期転送の特性上から、ホストＩＯに対するデータ同期確定方法を備えていない。その結果、従来技術の非同期転送方式では、データベース(ＤＢ)運用上必要となる、ＡＰＰ（アプリケーションプログラム）からのトランザクションに対するコミットポイントでの書込み要求がリモートサイト（副サイト）に確実に書かれたか否かを確認できないという課題がある。

【００１４】

本発明の目的は、ホストＩＯ（書込み要求）に対するリモートサイトへのデータ同期確定を、ホストアプリケーションのコミット単位または任意な時点での同期確定を可能にすることにある。

【００１５】

【課題を解決するための手段】

かかる課題を解決するため本発明においては、ローカル側の上位装置とリモート側の記憶システムとに接続され、前記上位装置から与えられる書込み対象のデータを記憶すると共に、当該書込み対象のデータを前記リモート側の記憶システムに送信する記憶システムであって、前記リモート側の記憶システムが前記書込み対象のデータを受信したか否かを問い合わせる問合せコマンドを前記上位装置から受信する手段と、前記上位装置へ、前記問合せコマンドに対する応答を送信する手段とを備えることを特徴とする。

また本発明においては、データを記憶するディスク装置と、外部上位装置と前記ディスク装置との間で前記データを転送するディスク制御装置とを有する記憶システムであって、前記ディスク制御装置は、前記上位装置から与えられた前記ディスク装置に書き込むべき書込み対象のデータを他の記憶システムへ送信する回路と、前記上位装置から入力されるコマンドを実行する回路及びソフトウェアとを備え、前記コマンドは、前記他の記憶システムが前記書込み対象のデータを受信したか否かに関する情報を問合せるものであり、かつ、自記憶システムに対して、前記外部上位装置にその問合せに対する結果を報告することを要求するものであることを特徴とする。

さらに本発明においては、上位装置に対する記憶システムの応答方法であって、前記上

位装置から送信される書込みコマンドを受信する第１のステップと、前記書込みコマンドに対する応答を前記上位装置に返す第２のステップと、前記書込みコマンド内のデータを他記憶システムへ送信する第３のステップと、前記上位装置から送信される前記他の記憶システムが前記データを受信したか否かを問う問合せコマンドを受信する第４のステップと、前記上位装置へ、前記問合せコマンドに対する応答を送信する第５のステップとを備えることを特徴とする。

さらに本発明においては、ローカル側の上位装置とリモート側の記憶システムとに接続され、前記上位装置から与えられる書込み対象のデータを対応する正ボリュームに格納すると共に、当該書込み対象のデータを副ボリュームが設けられた前記リモート側の記憶システムに送信する記憶システムであって、上位装置との間で情報を送受信するための第１のインタフェース回路と、他の記憶システムとの間で情報を送受信するための第２のインタフェース回路とを備え、前記上位装置から送信される書込み対象のデータを前記第１のインタフェース回路を介して受信し、当該書込み対象のデータを、当該書込みデータに付与した識別情報と共に前記第２のインタフェース回路を介して前記他の記憶システムに送信し、前記上位装置からの前記書込み対象のデータに関する問合せコマンドを前記第１のインタフェース回路を介して受信し、前記問合せコマンドにおいて指定された前記書込み対象データに付与した前記識別情報を前記第１のインタフェース回路を介して前記上位装置に送信すると共に、最後に前記副ボリュームに格納された前記書込み対象のデータに付与された前記識別情報を、前記第２のインタフェース回路を介して前記他の記憶システムに問い合わせ、前記他の記憶システムから送信される、最後に前記副ボリュームに格納された前記書込み対象のデータに付与された前記識別情報を前記第２のインタフェース回路を介して受信し、当該識別情報を、前記第１のインタフェース回路を介して前記上位装置に送信することを特徴とする。

【００１６】

さらに本発明においては、コンピュータ上で動作するアプリケーションから複数の書込み要求を受けた際に、前記複数の書込み要求内のデータを他記憶システムに非同期にコピーする記憶システムと接続される前記コンピュータに、インストールされるプログラムが記録された記録媒体であって、前記プログラムは、前記アプリケーションから問合せコマンドを受ける第１ステップと、前記記憶システムから、前記記憶システムに記憶された前記アプリケーションに関する書込み対象のデータを識別する識別情報を得る第２ステップと、前記記憶システムから、前記他の記憶システムに記憶された前記アプリケーションに関する書込み対象のデータを識別する識別情報を得る第３ステップと、前記第２ステップで得た識別情報と前記第３ステップで得た識別情報を基に、前記問合せに応答する第４のステップとを備える処理を前記コンピュータに実行させることを特徴とする。

さらに本発明においては、ローカル側ホストからの書込み対象のデータを、他の記憶システムへリモート非同期データ転送を行う記憶システムであって、前記記憶システムは、前記書込み対象のデータが前記他の記憶システムへ記憶されたか否かを前記ローカル側ホストへ報告することを特徴とする。

さらに本発明においては、他の記憶システムへデータの非同期リモートコピーを行う記憶システムに接続されるコンピュータにインストールされるプログラムを記録する記録媒体であって、前記プログラムは、前記コンピュータ内で動作するアプリケーションからコミットコマンドを受けると前記コンピュータ内のバッファにある前記アプリケーションが発行した書込み対象のデータを実際に前記記憶システムへ書込む書込み要求を発行する第１のステップと、前記記憶システムから前記書込み要求に対する書込み完了を受信する第２のステップと、前記書込み対象のデータが前記他の記憶システムにコピーされたか否かを問い合わせる問合せコマンドを前記記憶システムへ発行する第３のステップと、前記記憶システムから前記問合わせコマンドに対する応答を受ける第４のステップと、前記書込み完了を受信し、かつ、前記応答を受けた後で、前記コミットコマンドに対する応答を返す第５のステップとを備える処理を前記コンピュータに実行させることを特徴とする。

【００１７】

【発明の実施の形態】
（実施例）
以下、本発明の一実施例を図１～図３により詳細に説明する。
【００１８】
最初に、図３を用いて本発明を適用するシステム構成例を示す。図３は、一般的なＨＡクラスタの全体構成を示したシステムブロック図である。以下、ディスクアレイ装置１３ａをローカル（又は正サイト）、ディスクアレイ装置１３ｂをリモート（又は副サイト）とし説明する。また、参照番号に「ａ」を添付した方をローカル側、参照番号に「ｂ」を添付した方をリモート側として説明する。
【００１９】
ＨＡクラスタを構成するホストコンピュータ装置１ａ、１ｂはデータベース等のＡＰＰ（アプリケーション・プログラム）２ａ、２ｂ、ミドルソフト３ａ、３ｂ、ＡＰＰやミドルソフトを制御するＯＳ（オペレーティング・システム）４ａ、４ｂ、及び、ディスクアレイ装置との間でコマンド等の入出力を制御する“ＩＯ Ｉ／Ｆ”（入出力インタフェース）２３ａ、２３ｂで構成される。
【００２０】
ミドルソフト３ａ、３ｂは、リモート側（副サイト）では、正サイト障害での副サイト運用時に、副サイト側のホスト装置に対し、副サイトのディスクアレイ装置への書込み禁止状態を解除し、一方、正サイト側では、正サイトと副サイト間でデータを一致させる初期ペア状態の生成、ペア状態のサスペンド（論理的なペアを切断した状態）等のペア制御指令を大型ディスクアレイ装置１３ａ、１３ｂに送るコマンド制御ソフトである。本発明の一実施例では、このミドルソフト３ａにより、本願特有のＳｙｎｃコマンドを提供する（詳細は後述する）。
【００２１】
ホストコンピュータ装置１ａ、１ｂからのＩＯ要求は“ＩＯ Ｉ／Ｆ”（２３ａ、２３ｂ）を介して大型ディスクアレイ装置１３ａ、１３ｂに送られる。
【００２２】
大型ディスクアレイ装置１３ａ、１３ｂは、ＨＯＳＴ Ｉ／Ｆ制御回路１７ａ、１７ｂでホストからのＩＯ要求を受け、キャッシュ２１ａ、２１ｂに、書込み要求のデータ（ＷＲＩＴＥデータ）を書込む。このキャッシュ２１ａ、２１ｂに書かれたデータはドライブ制御２２ａ、２２ｂによって実際の物理ディスク９ａ、９ｂに書込まれる。ここで、キャッシュに書込まれたデータに関しては、大型ディスクアレイ装置は、そのデータの記憶を保証するものとする。また、大型ディスクアレイ装置は、複数の物理ディスク９をＲＡＩＤ（ＲｅｄｕｎｄａｎｃｙＡｒｒａｙ ｏｆ Ｉｎｅｘｐｅｎｓｉｖｅ Ｄｒｉｖｅｓ）１、ＲＡＩＤ５等で保護し物理ディスクの障害に備える。なお、ＲＡＩＤに関しては、"A Casefor Redundant Arrays of Inexpensive Disks (RAID)"by David A. Patterson, et al., Dec 1987, U. C. Berkleyに記載がある。
【００２３】
大型ディスクアレイ装置間でのデータの二重化では、まず、ローカル大型ディスクアレイ装置１３ａのキャッシュ２１ａに書かれたデータが、リモートＩ／Ｆ制御回路１８ａ、１８ｂおよび公衆回線や専用線等のデータ転送路２０ａ、２０ｂを介してリモート大型ディスクアレイ装置１３ｂに送られる。リモート側の大型ディスクアレイ装置１３ｂは、キャッシュ２１ｂに、この送られたデータを書き込む。この後、リモート側の大型ディスクアレイ装置１３ｂは、ローカル側での物理ディスク装置９ａへの書込みと同様に、物理ディスク９ｂへ受信したデータを書込む。
【００２４】
この二重化された状態では、リモート側のＨＯＳＴ Ｉ／Ｆ制御１７ｂはリモートホスト１ｂからの書込みを禁止する。
【００２５】
ここで、リモート側への未転送データは、ＢＩＴＭＡＰメモリ１１ａ、１１ｂによって差

分管理される。このメモリ内の差分情報によって管理される差分データ（未転送データ）が、ホスト装置１ａからのＩＯ要求に同期しないで非同期にリモート側に転送される。この差分データを管理する差分ＢＩＴＭＡＰメモリについては後述する。

【００２６】

なお、図３において、本願発明で定義するＳｙｎｃコマンド等のディスクアレイ装置内での処理は、ＨＯＳＴ／ＩＦ制御回路やリモートＩ／Ｆ制御回路等に、マイクロプロセッサおよびマイクロプロセッサから参照されるプログラムやデータが保持される一般的なメモリ（ＤＲＡＭ等）を設け、マイクロプロセッサがプログラムを実行することによって実現される。また、制御回路内に専用回路を設けることで実現しても良い。一方、当該装置を制御するマイクロプロセッサ、プログラムを記憶するメモリ等を、ＨＯＳＴ／ＩＦ制御回路やリモートＩ／Ｆ制御回路等以外の場所に設け、これにより、ＨＯＳＴ／ＩＦ制御回路やリモートＩ／Ｆ制御回路を制御しながら実行しても良い。

【００２７】

図１は、図３のシステム構成において本願発明を実現する論理ブロックを示す。

【００２８】

ローカル側のホスト装置１ａの論理ブロックは、アプリケーションソフトであるＡＰＰ（アプリケーションプログラム）２ａ、大型ディスクアレイ装置１３ａ、１３ｂ内のボリュームの初期ペア状態の生成、ペア状態のサスペンド等のペア制御を実行するミドルソフト３ａ、及びＯＳ４ａで構成される。

【００２９】

ローカルサイトである大型ディスクアレイ装置１３ａと、リモートサイトである大型ディスクアレイ装置１３ｂは、図示しない光ファイバ、広域回線網等のデータ転送路（図３の２０ａ、２０ｂ）で接続される。大型ディスクアレイ装置１３ａ、１３ｂ内のボリュームは、ＰＶＯＬ（正ボリューム）１０ａとＳＶＯＬ（副ボリューム）１０ｂで構成され、ホストからのデータはＰＶＯＬ１０ａからＳＶＯＬ１０ｂにコピーされ二重化される。なお、これらボリュームは物理ディスク（図３の９ａ、９ｂ）の中に保持される。

【００３０】

ＢＩＴＭＡＰ１１０（図３のＢＩＴＭＡＰメモリ１３ａに保持される）、ＢＩＴＭＡＰ１２０（図３のＢＩＴＭＡＰメモリ１３ｂに保持される）は、それぞれＰＶＯＬ（１０ａ）とＳＶＯＬ（１０ｂ）間のデータ差分管理テーブル（ボリューム単位に管理可能）であり、ＰＶＯＬとＳＶＯＬの全データブロックを数１０ＫＢ（キロバイト）単位でビットマップ化したものである。ＢＩＴＭＡＰ１１０は、ＰＶＯＬ（正ボリューム）に格納されているがリモート側へ未転送であるデータを表し、ＢＩＴＭＡＰ１２０はリモート側へ転送されたがＳＶＯＬ（副ボリューム）へまだ格納されていないデータを表す。通常、ペア（ＰＡＩＲ）状態（二重化状態）がサスペンド(Ｐ ＳＵＳ)した状態となると、ホストからの新たなデータはＰＶＯＬ１０ａにのみ書込まれるので、このデータ分がＰＶＯＬ（１０ａ）とＳＶＯＬ（１０ｂ）の不一致として、このＢＩＴＭＡＰ１１０、１２０によって差分管理されることになる。

【００３１】

ローカル側のＦＩＦＯ７、リモート側のＦＩＦＯ８は、それぞれ、ローカル側の大型ディスクアレイ装置１３ａ、リモート側の大型ディスクアレイ装置１３ｂ間の非同期転送用のバッファキューであり、ペア(ＰＡＩＲ)状態の時に使用される。

【００３２】

ローカル側のホスト１ａからのＩＯ（入出力）の書込みデータは、大型ディスクアレイ装置１３ａのＰＶＯＬ（１０ａ）対応のキャッシュに置かれる。その後、データは、物理ディスク９ａに書込まれると同時に、そのＩＯ単位に、そのＩＯを識別するための識別子であるシーケンス番号が付加され、即ち、ホストＩＯデータは順序化され、ホストコンピュータからの書込み要求の発行順に、一旦ＦＩＦＯ７にキューされる。なお、識別子は番号の他、タイムスタンプ等、データをホストコンピュータからの書込み要求の発行順に一意に識別できる識別子ならその種類は問わない。また、ＦＩＦＯバッファ７にキャッシュ上

のデータをキューイングする際は、このＦＩＦＯバッファに、キャッシュ上のデータの複製を作り実際にキューイングする方法でも良いが、データ量の増加を防ぐため、キャッシュ上のデータを管理するポインタ・アドレス等を管理するキューであることが望ましい。この場合、実際のデータをリモート側へ転送する場合は、キャッシュから実際のデータが転送されることになる。

【００３３】

このシーケンス番号を付加したデータは、ホストへのＩＯ完了応答報告と非同期に、ローカル側の大型ディスクアレイ装置１３ａからリモート側の大型ディスクアレイ装置１３ｂへ転送される。リモート側の大型ディスクアレイ装置１３ｂは、このデータを受信すると、このデータをシーケンス番号順にＦＩＦＯ８にキューイングする。このＦＩＦＯ８にキューされたデータは、シーケンス番号順にＳＶＯＬ（副ボリューム）１０ｂ対応のキャッシュに置かれ、その後、物理ディスク９ｂに書込まれる。なお、受信データを一旦ＦＩＦＯ８にキューイングした後にシーケンス番号順にキャッシュに置く理由は、ローカル側の大型ディスクアレイ装置１３ａの制御の都合や転送路の通信状態の関係で、リモート側の大型ディスクアレイ装置１３ｂは、書込み要求のデータを必ずしも、ホストコンピュータからの書込み要求の発行順序で、受信しないからである。本実施例では、前述のように、ＦＩＦＯ８に受信データをキューイングし、その後、受信データに割り付けられているシーケンス番号が番号順に揃った段階で、こんどは、ＦＩＦＯからデータをシーケンス番号順に読み出してＳＶＯＬ対応のキャッシュへ正式に書込むことで、その順序性を保証している。当然のことながら、リモート側の大型ディスクアレイ装置１３ｂがデータを受信した順序がそのまま、ホストコンピュータから発行される書込み要求の発行順序と一致することが保証されるなら、大型ディスクアレイ装置１３ｂは、受信データをキャッシュへＳＶＯＬデータとして直接書込める。

【００３４】

大型ディスクアレイ装置１３ａ、１３ｂ間の伝送路等の転送障害等によって非同期転送が出来ない場合、大型ディスクアレイ装置１３ａ、１３ｂは、ＦＩＦＯ７、８にキューされた未転送データをＢＩＴＭＡＰ１１０、１２０に差分データとして、ボリューム単位に、管理し、二重化を障害サスペンド(ＰＳＵＥ)状態とする。

【００３５】

ローカル側からリモート側へのデータの転送が進むと、ＦＩＦＯ７内にはデータが無くなる。一方、ＦＩＦＯ８には、データが受信され、例えば、ＦＩＦＯ８内にシーケンス番号「１」～「５」を持つデータが記憶される。その後、ＦＩＦＯ８からＳＶＯＬ対応のキャッシュへ受信データが格納されると、ＦＩＦＯ８内のデータは減る。この際、ＳＶＯＬへ新たに書込まれたデータに割り付けられていたシーケンス番号が新たなＳＶＯＬ書込みシーケンス番号となる。従って、ＦＩＦＯ８から実際にＳＶＯＬ対応のキャッシュへ書込まれたデータに割り当てられているシーケンス番号をリモート側の大型記憶システム内に設けられるメモリにＳＶＯＬ書込みシーケンス番号として記憶しておき、ローカル側から問合せがあった場合に、このメモリ内のシーケンス番号をＳＶＯＬ書込みシーケンス番号として報告すれば良い。そして、リモート側がＳＶＯＬ書込みシーケンス番号として「５」を報告すると、ミドルウェア３ａは、（ＰＶＯＬシーケンス番号「５」）＜＝（ＳＶＯＬ書込みシーケンス番号「５」）が成立したことから、データ同期が確立したと判断する。

【００３６】

次に、本願発明である、ホスト１ａのＡＰＰ２ａとミドルソフト３ａと大型ディスクアレイ装置１３ａの連携によってＡＰＰ２ａのＣｏｍｍｉｔ時のリモートサイトへの同期確認がどのように実行されるか説明する。

【００３７】

これは、ホスト１ａのミドルソフト３ａが、ＡＰＰ２ａから同期確認の要求を受け、大型ディスクアレイ装置１３ａ、１３ｂの状態とＦＩＦＯ７、８キュー等から得られる、ＰＶＯＬとＳＶＯＬに格納されるデータに関するシーケンス番号を比較し、コミットコマンド

発行直前の“ＷＴＩＴＥ Ｉ／Ｏ”（書込み要求）のデータがリモートサイトに同期したかを確認し、ＡＰＰ２ａへリターンを返すことで、可能になる。

【００３８】

以下、ホスト１ａのＡＰＰ２ａとミドルソフト３ａと大型ディスクアレイ装置１３ａの連携によってＡＰＰ２ａのＣｏｍｍｉｔのリモートサイトへの同期要求がどのように行われるか具体的に説明する。

【００３９】

ホスト１ａのＡＰＰ２ａからのＣｏｍｍｉｔが発行されＡＰＰ２ａのＣｏｍｍｉｔをリモートサイトへ同期要求する場合、まず、ＡＰＰ２ａはトランザクションとしてデータベースへＷＲＩＴＥし、最後にＣｏｍｍｉｔを発行する。通常はこのＣｏｍｍｉｔで完了である。

【００４０】

本願発明では、大型ディスクアレイ装置間でデータが二重化構成の場合、更に、Ｃｏｍｍｉｔの直後に、本発明の特徴である“Ｓｙｎｃコマンド”を発行する。Ｓｙｎｃコマンドは、ホスト装置上で動作するライブラリ及びホストコマンドあり、ソフトウエア製品としてＣＤＲＯＭ、フロッピーディスク等のソフトウェア記憶媒体を介して提供される。このＳｙｎｃコマンドはミドルソフト３によってアプリケーションに提供される。また、パラメータとしてｇｒｏｕｐ（後で説明する）と最大の同期完了待ち時間を指定するｔｉｍｅｏｕｔによって定義される。

【００４１】

なお、このＳｙｎｃコマンドを発行するタイミングは、Ｃｏｍｍｉｔ直後に限定されず、アプリケーションソフトが、リモート側とローカル側でデータの同期が必要と判断した時点に任意に発行可能である。言い換えれば、同期を取る／取らないの選択が、Ｓｙｎｃコマンドの発行の有／無で任意にできるので、Ｓｙｎｃコマンドを使用するアプリケーションソフトに対する自由度がある。但し、非同期リモートコピーにおいて、トランザクション単位でコミットを行う場合、リモート側の大型ディスクアレイ装置内に書込みデータが確実に記憶されているか否かを確認するという意味では、アプリケーションにおける、コミットコマンドの発行、引き続いての本発明のＳｙｎｃコマンドの発行という２つのコマンドの流れは必然である。

【００４２】

このミドルソフト３ａは、Ｓｙｎｃコマンドを受けると、大型ディスクアレイ装置のペア状態をチェックし、ＰＡＩＲ状態であればＦＩＦＯ７の最新のＰＶＯＬシーケンス番号（最も新しくホストから受領したデータであって、ローカル側のキャッシュ２１ａに書かれており、リモートへ未コピーのデータに付与されているシーケンス番号）をローカル側のディスクアレイ装置１３ａから取得し、このＰＶＯＬの最新シーケンス番号をＣｏｍｍｉｔ直後のシーケンス番号として保持する。次に、ミドルソフト３ａは、ローカル側の大型ディスクアレイ装置１３ａを介して、リモート側の大型ディスクアレイ装置１３ｂ内のＳＶＯＬ書込みシーケンス番号（最も新しくＳＶＯＬ（リモート側のキャッシュ２１ｂ）へ書き込まれたデータに付与されているシーケンス番号）を取得する。更に、ＰＶＯＬシーケンス番号とＳＶＯＬ書込みシーケンス番号を比較し、

（ＰＶＯＬシーケンス番号）

＜＝（ＳＶＯＬ書込みシーケンス番号）

が成立するまでＳＶＯＬ書込みシーケンス番号を繰返し取得し、テストを繰り返す。条件が成立すると、ミドルソフト３ａは、同期完了の応答としてこのＳｙｎｃコマンドの呼び出しもとであるＡＰＰ２ａにリターンを返す。ＡＰＰ２ａは、このリターンを受けると、つまり、このＳｙｎｃコマンドが完了したことで同期完了とみなす。

【００４３】

ここで、このＰＶＯＬシーケンス番号とＳＶＯＬ書込みシーケンス番号の取得およびデータ同期について、以下、詳細に説明する。ＰＶＯＬシーケンス番号は、最も新しくローカル側のキャッシュに書込まれたデータに付与されたシーケンス番号であるから、例えば、

図１においては、ＦＩＦＯ７から最新のＰＶＯＬシーケンス番号「５」が獲得され、これがミドルソフト３ａへ報告される。一方、リモート側のＳＶＯＬ書込みシーケンス番号に関しては、ＦＩＦＯ８にはデータ「１」「２」がキューイングされた状態であることから、ＳＶＯＬには「０」のシーケンス番号を持つデータまでが書込まれていることが分かる。つまり、ＦＩＦＯ８を用いて、番号「０」が、ＳＶＯＬ書込みシーケンス番号であることが分かり、このＳＶＯＬ書込みシーケンス番号「０」がミドルソフト３ａへ報告される。この状態では、（ＰＶＯＬシーケンス番号「５」）＞（ＳＶＯＬ書込みシーケンス番号「０」）であるから、ミドルソフト３ａは、同期がとれていないと判断する。

【００４４】

ローカル側からリモート側へのデータの転送が進むと、ＦＩＦＯ７内にはデータが無くなる。一方、ＦＩＦＯ８には、データが受信され、例えば、ＦＩＦＯ８内にシーケンス番号「１」～「５」を持つデータが記憶される。その後、ＦＩＦＯ８からＳＶＯＬ対応のキャッシュへ受信データが格納されると、ＦＩＦＯ８内のデータは減る。この際、ＳＶＯＬへ新たに書込まれたデータに割り付けられていたシーケンス番号が新たなＳＶＯＬ書込みシーケンス番号となる。従って、ＦＩＦＯ８から実際にＳＶＯＬ対応のキャッシュへ書込まれたデータに割り当てられているシーケンス番号をローカル側の大型記憶システム内に設けられるメモリにＳＶＯＬ書込みシーケンス番号として記憶しておき、ローカル側から問合せがあった場合に、このメモリ内のシーケンス番号をＳＶＯＬ書込みシーケンス番号として報告すれば良い。そして、リモート側がＳＶＯＬ書込みシーケンス番号として「５」を報告すると、ミドルウェア３ａは、（ＰＶＯＬシーケンス番号「５」）＜＝（ＳＶＯＬ書込みシーケンス番号「５」）が成立したことから、データ同期が確立したと判断する。

【００４５】

なお、上記において、ＰＶＯＬシーケンス番号、ＳＶＯＬ書込みシーケンス番号をＦＩＦＯから求めるのでなく、ローカル側およびリモート側のＲＡＭ等の記憶手段を別途設け、２つの番号を随時これらに記憶するようにしておき、必要な時は、これを読むようにしても良い。

【００４６】

また、上記説明では、ホスト１ａ内のＡＰＰ２ａはＯＳ４ａからは１つのプロセスであるが、複数のプロセスとしてＡＰＰ２ａは存在することもできる。この場合、この複数のプロセスからそれぞれＳｙｎｃコマンドが発行されるが、それぞれのＳｙｎｃコマンド受領時点でのＰＶＯＬ最新シーケンス番号をそれぞれのＳｙｎｃコマンド対応に取得し、これをＳＶＯＬ最新シーケンス番号と比較することで、複数のプロセス単位でも同期確認することが可能である。

【００４７】

また、ホスト１ａ内のアプリケーションソフト単位にグループ化して、このグループ単位に同期確認をとることも可能である。この場合、このグループ単位に非同期転送のシーケンス番号を維持する。大型ディスクアレイ装置内に複数のグループを定義しておき、アプリケーションソフトが同期確認を行う場合、そのアプリケーションソフトが同期完了要求にグループ名を指定することでグループ単位に独立して同期確認が可能になる。

【００４８】

図２は、図１を用いて説明した事項の全体制御フローを示す図である。以下制御フローに基づき詳細に説明する。ここで、図２の説明で用いるＳＶＯＬシーケンス番号は、図１の説明のＳＶＯＬ書込みシーケンス番号と同意味で使用する。

【００４９】

まず、制御フローは、ホスト１ａのアプリケーションソフトであるＡＰＰ２ａと、ペア制御及びＳｙｎｃコマンドを実行するミドルソフト３ａと、大型ディスクアレイ装置１３ａホスト１ａと物理ディスク９ａ間でデータの転送制御を行うＤＫＣ１３０ａ（図３参照）の制御を示す。

【００５０】

ＡＰＰ２ａは、コミットポイントになると、コミットコマンドを発行し、ホストコンピュ

ータ１ａのデータバッファ内に記憶していたデータを大型ディスクアレイ装置１３ａへ書込む（図２には示していない）。

【００５１】

次に、ＡＰＰ２ａは、データベースへのコミット完了直後にＳｙｎｃコマンドを発行する。ここで、Ｓｙｎｃコマンドは、Ｓｙｎｃコマンドであることを示すコマンドコード（図示せず）の他、２つの引数を有する。第１引数であるｇｒｏｕｐは、前述したグループ名を指定する。第２引数であるｔｉｍｅｏｕｔは、最大の同期完了待ち時間を指定する。ＡＰＰ２ａは、Ｓｙｎｃコマンドを発行した後、同期完了の報告をミドルソフト３ａから受けるか、若しくは同期失敗の報告を受けるまで待つ。

【００５２】

ミドルソフト３ａは、Ｓｙｎｃ(ｇｒｏｕｐ、ｔｉｍｅｏｕｔ)コマンド３１を実行する。Ｓｙｎｃ(ｇｒｏｕｐ、ｔｉｍｅｏｕｔ）コマンド３１は、このコマンド内でまずローカル側の大型ディスクアレイ装置１３ａ内のＰＶＯＬ（正ボリューム）１０ａのペア状態を調べるために、ＰＶＯＬ状態取得コマンド２０１を大型ディスクアレイ装置１３ａに発行する（ＰＶＯＬ状態取得ステップ３２）。大型ディスクアレイ装置１３ａの制御部分であるディスク制御部（ＤＫＣ）１３０ａは、このコマンド応答としてＰＶＯＬペア状態２０２を返す（ＰＶＯＬ状態応答ステップ３９）。なお、ｇｒｏｕｐが指定されていた場合は、このｇｒｏｕｐ単位にＰＶＯＬのシーケンス番号を管理することになる。以下、ｇｒｏｕｐがある特定グループに限定されていると考え、ｇｒｏｕｐの違いを省いて説明する。

【００５３】

ミドルソフト３ａは、ＰＶＯＬペア状態２０２を大型ディスクアレイ装置１３ａから受けると、ＰＶＯＬ状態のチェックを行い、状態がＰＡＩＲ以外(ＰＳＵＳ,ＰＳＵＥ)であれば二重化はサスペンドであるとして同期失敗をＡＰＰ２ａへ返す（ＰＶＯＬ状態のチェックステップ３３）。なお、大型ディスクアレイ装置はこれらの状態を管理する機能を有するものとする。

【００５４】

一方、ＰＶＯＬ状態のチェックステップ３３は、状態がＰＡＩＲであれば二重化状態と判断し、ＦＩＦＯ７にキューされている書き込みデータの最新のＰＶＯＬシーケンス番号を調べるため、ＰＶＯＬシーケンス番号取得コマンド２０３を、大型ディスクアレイ装置１３ａに発行する（ＰＶＯＬシーケンス番号を取得するステップ３４）。

【００５５】

大型ディスクアレイ装置１３ａは、このコマンド応答として、ＦＩＦＯ７上にキューされている最新のデータに付加されている最新のＰＶＯＬシーケンス番号をＰＶＯＬ最新シーケンス番号２０４として、ミドルソフト３ａへ返す（ＰＶＯＬ最新シーケンス応答ステップ４０）。

【００５６】

本発明でのミドルソフト３ａは、この取得したＰＶＯＬ（正ボリューム）シーケンス番号を同期確認の間保持し、後で取得するＳＶＯＬ（副ボリューム）シーケンス番号との比較に使用する。

【００５７】

次に、ミドルソフト３ａは、リモートサイトのＳＶＯＬシーケンス番号を入手するために、ＳＶＯＬシーケンス番号取得コマンド２０５を大型ディスクアレイ装置１３ａに発行する（ＳＶＯＬシーケンス番号を取得するステップ３５）。大型ディスクアレイ装置１３ａは、リモートサイトの大型ディスクアレイ装置１３ｂからＳＶＯＬ１０ｂに書き込まれたデータに対応する最新のシーケンス番号を取得し、このＳＶＯＬシーケンス番号取得コマンドの応答として、最新のＳＶＯＬ書込みシーケンス番号２０６をミドルソフト３ａへ返す（ＳＶＯＬ書込みシーケンス応答ステップ４１）。

【００５８】

なお、リモートサイトの大型ディスクアレイ装置１３ｂからＳＶＯＬ１０ｂに書き込まれたデータに対応する最新のシーケンス番号を取得するには、まず、ローカル側の大型ディ

スクアレイ装置１３ａが、リモート側の大型ディスク装置１３ｂに対して、ＳＶＯＬへ書込まれた最新のデータに関する最新のＳＶＯＬ書込みシーケンス番号を問い合わせるコマンドを発行する。本実施例では、このコマンドは、通常のコピーデータを転送する通信路を用いて、リモートＩ／Ｆ制御部１８ａ、１８ｂを介して問い合わせられるものとする。これを受信したリモートの大型ディスクアレイ装置は、この問い合わせコマンドを解析し、ＳＶＯＬへ書込まれた最新のデータに関する最新のＳＶＯＬ書込みシーケンス番号を取得する。

【００５９】

ここで、最新のＳＶＯＬ書込みシーケンス番号は、大型ディスクアレイ装置１３ｂが有する特定のメモリに記録され、リモート側がコピーデータを確実に保証できる形で保持した段階で、更新されているものとする。本実施例では、先に説明したように、リモート側でのデータ受信順序がホストコンピュータが発行したデータ順序と必ずしも一致していない場合を考慮し、ＦＩＦＯ８を使用する場合を示している。そのため、リモート側がコピーデータを確実に保証できる形とは、本実施例では、受信データを、シーケンス番号に抜けが無い状態で、ＦＩＦＯ８からＳＶＯＬに対応するキャッシュメモリに管理を移した状態である（なお、本システムは、キャッシュに記憶されたデータは物理ディスクに書き込むことを保証するものとする）。そして、この状態のデータに付与されたシーケンス番号がＳＶＯＬ書込みシーケンス番号となる。例えば、ＦＩＦＯ８内にシーケンス番号「２」、「３」、「５」を有する受信データがあるとすると、シーケンス番号「２」「３」を有する受信データはＳＶＯＬに対応するキャッシュへ書込まれるがシーケンス番号「５」を有する受信データはＳＶＯＬ対応のキャッシュへは書込まれない。この段階でのＳＶＯＬ書込みシーケンス番号は「３」である。その後、シーケンス番号「４」を有するデータを受信すると、シーケンス番号「４」および「５」の受信データが順次ＳＶＯＬ対応のキャッシュに格納される。ここで初めて、リモート側がコピーデータを確実に保証できる形で保持した段階の最新のＳＶＯＬ書込みシーケンス番号は、「５」ということになる。一方、ローカル側の記憶システムや、ローカル側とリモート側の間の通信路が、リモート側での受信データの順序性を保証できるものであれば、リモート側の記憶システムは受信データを受信した時点で、例えば、リモートＩ／Ｆ制御回路がデータを受信した時点で、その受信したデータに付与されているシーケンス番号を、本実施例でいう最新のＳＶＯＬ書込みシーケンス番号として用いることができる。

【００６０】

リモート側のディスクアレイ装置１３ｂは、取得した最新のＳＶＯＬ書込みシーケンス番号を、前記問い合わせコマンドの応答として、ローカル側の大型ディスクアレイ装置１３ａへ転送する。

【００６１】

また、リモート側の最新のシーケンス番号を取得する別の方法として、リモート側の大型ディスクアレイ自体が、定期的に、又は、ＳＶＯＬへのデータ書込み更新が有った際に、最新のＳＶＯＬ書込みシーケンス番号をローカル側へ通知し、これをローカル側で記録するようにし、Ｓｙｎｃコマンドが発生した時点で、ローカル側に記録したこの最新のＳＶＯＬシーケンス番号を読み出して用いるようにしても良い。

【００６２】

ミドルソフト３ａは、次のステップ３６で、保持していたＰＶＯＬ（最新）シーケンス番号とＳＶＯＬ（書込み）シーケンス番号を比較し、

ＰＶＯＬシーケンス番号＜＝ＳＶＯＬ書込みシーケンス番号

であれば当該ＰＶＯＬシーケンス番号に対応するデータはＳＶＯＬ側に書き込み済みであるとして、ＡＰＰ２ａに同期完了を返す（ＰＶＯＬとＳＶＯＬシーケンス比較ステップ３６）。

【００６３】

一方、

ＰＶＯＬシーケンス番号＞ＳＶＯＬシーケンス番号

であれば同期が完了していない（リモート側へデータが書き込まれていない）として、次のステップに進みタイムアウトのチェックを行い指定ｔｉｍｅｏｕｔ値を超えていれば同期完了タイムアウトとして同期失敗をＡＰＰ２ａへ返す（タイムアウトのチェックステップ３７）。ここで、指定ｔｉｍｅｏｕｔ値を超えていなければ一定時間待ち（ＷＡＩＴステップ３８）、その後、ステップ３５からステップ３８を同期が完了するまで繰り返す。繰り返しの中で、ステップ３６で“ＰＶＯＬシーケンス番号＜＝ＳＶＯＬシーケンス番号”の条件が成立した時点で、同期完了し、Ｓｙｎｃコマンドを発行した呼出し元であるＡＰＰ２に制御が戻る。ＡＰＰ２ａは、Ｓｙｎｃコマンドの制御が戻った時点で、戻り値をチェックし同期完了を確認する。

【００６４】

以上説明したように、ホスト１ａのＡＰＰ２ａは、ミドルソフト３ａと連携して、データベース等のトランザクションのコミットを行った直後に本願で示したＳｙｎｃコマンドを実行するので、本Ｓｙｎｃコマンド発行前にディスクアレイ装置へ書込まれたデータが、リモート側のディスクアレイ装置にデータが確実に格納されているか否かをＡＰＰ２は、知ることが出来る。見方を変えれば、本実施例を使用するホストコンピュータ内のアプリケーションからは、非同期コピーが、Ｓｙｎｃコマンドを発行することによってアプリケーション毎に同期処理（データがコンピュータ内とリモートの記憶装置内で一致する処理）ができるリモートコピーとして見えることになる。

【００６５】

（他の実施例）

図４および図５は、ホスト装置１ａのミドルソフト制御の負荷を軽減する目的で、ミドルソフト３ａの制御をローカル側の大型ディスクアレイ装置１３ａ内で実施した本発明の他の実施例を示したものである。

【００６６】

図４は、図１、図２のところで説明したミドルソフト制御を、大型ディスクアレイ装置内で実施した場合の、本発明の論理的なブロック図を示す。図４において、図１のミドルソフト３ａで実現していた機能を図４のＳｙｎｃ制御３００（後述）で置換えた以外は、論理的構成はほぼ同じである。Ｓｙｎｃ制御３００の機能は、“ＨＯＳＴ　Ｉ／Ｆ制御回路（１７ａ）”または“リモートＩ／Ｆ制御回路（１８ａ）”上で本機能を実現するプログラムを実行することで実現される。

【００６７】

ホスト１ａ内のＡＰＰ２ａは、前記実施例と同様にトランザクションとしてデータベースへＷＲＩＴＥし、最後にＣｏｍｍｉｔを発行し、その後、本発明の特徴コマンドであるＳｙｎｃコマンドを発行する。このＳｙｎｃコマンドは、ＯＳを介してディスクアレイ装置１３ａに入力される。なお、ここでは、Ｃｏｍｍｉｔ動作を確認するという意味でＳｙｎｃコマンドをＣｏｍｍｉｔの発行の後に入れているが、Ｓｙｎｃコマンドを発行するタイミングは、この時点に限定されるものでなく、アプリケーションプログラムが必要と認識した時点で発行できる。

また、以降説明するディスクアレイ装置側でのＳｙｎｃコマンドの処理も、当該Ｓｙｎｃコマンドの発行時点に左右されるものではない。

【００６８】

このＳｙｎｃコマンドは、大型ディスクアレイ装置によって、これに接続されるホストに対して提供され、大型ディスクアレイ装置に内蔵されたマイクロプログラム等の機能（後述のＳｙｎｃ制御３００が相当する）である。例えば、図１のＨＯＳＴ　Ｉ／Ｆ制御回路内のマイクロプロセッサ等が、リモートＩ／Ｆ制御回路等と連絡しながら実行する。

【００６９】

ここで、Ｓｙｎｃ(ａｐｐＩＤ、　ｇｒｏｕｐ、ｔｉｍｅｏｕｔ)コマンド５１は、パラメータとして、ローカル側の大型ディスクアレイ装置１３ａのＳｙｎｃ制御３００がＳｙｎｃ要求プロセスを識別するためのａｐｐＩＤと、前記のｇｒｏｕｐ（前述したグループ名）と、および、最大の同期完了待ち時間を指定するＴｉｍｅｏｕｔを有する。

【００７０】
このＳｙｎｃコマンド５１は、ＯＳ４ａを介し大型ディスクアレイ装置１３ａのＳｙｎｃ制御３００に渡される。このＳｙｎｃ制御３００はＳｙｎｃコマンドを受けると、前記実施例と同様にボリュームのペア状態をチェックし、ＰＡＩＲ状態であれば、ＦＩＦＯ７の“最新のＰＶＯＬシーケンス番号”とＳｙｎｃコマンドパラメータとして渡される“ａｐｐＩＤ”をディスクアレイ装置内のメモリに登録し、このＰＶＯＬ最新シーケンス番号をＣｏｍｍｉｔの直後のシーケンス番号として保持する。
【００７１】
次にＳｙｎｃ制御３００は、リモート側のＳＶＯＬ書込みシーケンス番号を取得し、
“ＰＶＯＬシーケンス番号”
＜＝“ＳＶＯＬ書込みシーケンス番号”
が成立するまでＳＶＯＬ書込みシーケンス番号を繰返しテストし、このＳｙｎｃ制御３００内で待つ。条件が成立するとこのＳｙｎｃ制御３００は、同期完了の応答として、ＯＳ４ａを介し“ａｐｐＩＤ”を、呼び出しもとであるＡＰＰ２ａに戻す。ＡＰＰ２ａは、この“ａｐｐＩＤ”を受けると、Ｓｙｎｃコマンドが完了したことで同期完了とみなす。なお、ＰＶＯＬシーケンス番号およびＳＶＯＬ書込みシーケンス番号の具体的な取得方法は、先の実施例と同様である。
【００７２】
上記では、ホスト１ａ内のＡＰＰ２ａは、ＯＳ４ａからは1つのプロセスとして説明したが、複数のプロセスとしてＡＰＰ２ａは存在することもできる。この場合、複数のプロセスからＳｙｎｃコマンドが発行されると、Ｓｙｎｃ制御３００は、ａｐｐＩＤで識別し、プロセス対応のそれぞれのＳｙｎｃコマンド受領時点でのａｐｐＩＤとＰＶＯＬ最新シーケンス番号を登録して、後で得るＳＶＯＬシーケンス番号と比較するので複数のプロセス単位で同期確認することが可能である。結果は、登録したａｐｐＩＤを有するアプリケーションへ報告される。
【００７３】
また、ホスト１ａ内のアプリケーションソフトを複数合わせてグループ化して、このグループ単位に同期確認をとることも可能である。この場合、このグループ単位に非同期転送のシーケンス番号を維持する。大型ディスクアレイ装置内に複数のグループを定義しおき、アプリケーションソフトが同期確認を行う場合、そのアプリケーションソフトが同期完了要求にグループ名を指定することでグループ単位に独立して同期確認が可能になる。
【００７４】
図５は、図４で説明した事項の全体制御フローを示す図である。以下制御フローに基づき詳細に説明する。ここで、図５の説明で用いるＳＶＯＬシーケンス番号は、図４の説明のＳＶＯＬ書込みシーケンス番号と同意味で使用する。制御フローは、ホスト１ａのアプリケーションソフトであるＡＰＰ２ａと、ペア制御及びＳｙｎｃコマンドを実行するローカル側の大型ディスクアレイ装置１３ａ内のＳｙｎｃ制御３００と、及び、リモート側大型ディスクアレイ装置１３ｂのディスク制御ＤＫＣ１３０ｂの制御を示す。
【００７５】
ＡＰＰ２ａは、データベースへのコミット完了直後にＳｙｎｃ(ａｐｐＩＤ、ｇｒｏｕｐ、ｔｉｍｅｏｕｔ)コマンド５１を発行する。Ｓｙｎｃコマンド５１の第１引数であるａｐｐＩＤは、大型ディスクアレイ装置１３ａのＳｙｎｃ制御３００が、ホスト１ａ内のＳｙｎｃ要求プロセス（アプリケーション）を識別するための引数である。これにより、Ｓｙｎｃ要求を発行するプロセス（アプリケーション）毎に同期確定を行うことが可能となる。第２引数であるｇｒｏｕｐは、前述したグループ名を指定する場合に用いる。第３引数であるｔｉｍｅｏｕｔは、最大の同期完了待ち時間を指定する。
【００７６】
大型ディスクアレイ装置１３ａのＳｙｎｃ制御３００は、Ｓｙｎｃ(ａｐｐＩＤ、ｇｒｏｕｐ、ｔｉｍｅｏｕｔ)コマンド５１を実行する。まずＳｙｎｃ制御３００は、大型ディスクアレイ装置１３ａ内のＰＶＯＬ９のペア状態を調べる（ＰＶＯＬ状態のチェックステ

ップ５３３）。
【００７７】
Ｓｙｎｃ制御３００は、ＰＶＯＬ状態のチェック３３によって状態がＰＡＩＲ以外(ＰＳＵＳ、ＰＳＵＥ)であれば二重化はサスペンドであるとして同期失敗をＡＰＰ２ａに返す。状態がＰＡＩＲであれば二重化状態であり、書き込みデータはＦＩＦＯ７にキューされているので、最新のＰＶＯＬシーケンス番号とａｐｐＩＤをディスクアレイ装置内のメモリに登録する（ａｐｐＩＤとＰＶＯＬシーケンス番号を登録するステップ５３４）。この登録したＰＶＯＬシーケンス番号とａｐｐＩＤは、同期確認の間保持され、後で、ＳＶＯＬシーケンス番号との比較に使用される。
【００７８】
次に、Ｓｙｎｃ制御３００は、ＰＶＯＬシーケンス番号とリモートサイトのＳＶＯＬシーケンス番号の比較のために、ＳＶＯＬシーケンス番号取得コマンド５０１をリモート側の大型ディスクアレイ装置１３ｂのディスク制御部１３０ｂ（図３のリモート側のディスクアレイ装置参照）に発行する（ＳＶＯＬシーケンス番号を取得するステップ５３５）。
【００７９】
リモート側の大型ディスクアレイ装置１３ｂは、このコマンド５０１の応答としてＳＶＯＬ書込みシーケンス応答として最新のＳＶＯＬ書込みシーケンス番号５０２を、ローカル側のディスクアレイ装置１３ａへ返す（ＳＶＯＬ書込みシーケンス応答ステップ５４１）。なお、ＳＶＯＬ書込みシーケンス番号の具体的な取得方法は先の実施例と同様である。
【００８０】
ローカル側のディスクアレイ装置１３ａのＳｙｎｃ制御３００は、次のステップ３６で、保持していたａｐｐＩＤに対応するＰＶＯＬシーケンス番号とＳＶＯＬシーケンス番号を比較し、
“ＰＶＯＬシーケンス番号”＜＝“ＳＶＯＬシーケンス番号”
であれば当該ＰＶＯＬシーケンス番号はＳＶＯＬ側に書き込み済みであるとして、該当ａｐｐＩＤと合せて同期完了を、アプリケーションソフトＡＰＰ２ａを返す。一方、
“ＰＶＯＬシーケンス番号”＞“ＳＶＯＬシーケンス番号”
であれば同期が完了していないので次の待ちステップ５３７に進む（ａｐｐＩＤのＰＶＯＬとＳＶＯＬシーケンスを比較するステップ５３６）。
【００８１】
次のタイムアウトをチェックするステップ５３７では、タイムアウトのチェック５３７を行い、指定ｔｉｍｅｏｕｔ値を超えていれば同期完了タイムアウトとして同期失敗をＡＰＰ２ａへ返す。一方、指定ｔｉｍｅｏｕｔ値を超えていなければ一定時間ＷＡＩＴ５３８する（ＷＡＩＴステップ５３８）。
【００８２】
その後、ステップ５３５から同期が完了するまでステップ５３６を繰り返し、
“ＰＶＯＬシーケンス番号”＜＝“ＳＶＯＬシーケンス番号”
の条件が成立した時点で、同期が完了したとして、Ｓｙｎｃコマンド５１の呼び出し元であるＡＰＰ２ａに、Ｓｙｎｃ制御から、制御を返す。ＡＰＰ２ａは、Ｓｙｎｃコマンドの制御が戻った時点で、ａｐｐＩＤをチェックすると共に戻り値（同期が完了したか否かを示す値）をチェックし同期完了を確認する。
【００８３】
以上説明したようにディスクアレイ装置が、Ｓｙｎｃコマンドを受信すると、ホストからの書込み要求のデータが確実にリモート側のディスクアレイ装置にコピーされたか否か確認した上で、ホストに対し報告することが可能となる。従って、ホストはディスクアレイ装置が非同期リモートコピーを行っている場合でも、Ｓｙｎｃコマンドを発行することによって、Ｓｙｎｃコマンドの発行前にディスクアレイ装置へ書込まれているデータがリモート側へ転送されているかを正確に知ることができ、データベース等で必要となるＣｏｍｍｉｔ制御を確実なものとすることができる。見方を変えれば、本実施例を使用するホストコンピュータ内のアプリケーションからは、非同期コピーが、Ｓｙｎｃコマンドを発

行することによってアプリケーション毎に同期処理（データがコンピュータ内とリモートの記憶装置内で一致する処理）ができるリモートコピーとして見えることになる。

【００８４】

上記実施例では、コミットコマンドの発行に続けてＳｙｎｃコマンドの発行という形で発明を説明したが、次のような変形も可能である。

【００８５】

一つ目は、最初の実施例においてミドルソフト３ａを機能拡張し、これにより、非同期データ転送を行っている場合は、アプリケーションから受けた従来のコミットコマンドを「従来のコミットコマンド＋Ｓｙｎｃコマンド」と認識して、処理する方法である。具体的には、機能拡張されたミドルソフト３ａは、コミットコマンドを受けると、まず、コンピュータ内のデータバッファ内にあるデータを記憶システムへ書込む。それと合わせて、コミットの対象となるデータが記憶されるボリュームが非同期コピー中か否かを判断する。対象ボリュームが非同期コピー中の場合は、ミドルソフト３ａは本発明の最初の実施例に対応した問合せコマンドをローカル側の記憶システムへ発行する。その後、ミドルソフト３ａは、ローカル側の記憶システムから書込み完了を受けるが、この時点ではアプリケーションに対してコミットコマンドに対するリターンを返さず、先の問合せコマンドに対する応答が有るのを待つ。そして、ミドルソフト３ａは、記憶システムから問合せコマンドに対する応答を受け、リモート側の記憶システムで書込みデータの確定が取れた時点で初めて、前記アプリケーションへ、コミットコマンドに対するリターンを返すようにする。この方法では、アプリケーションはＳｙｎｃコマンドを発行する必要が無いので、非同期コピー中であることを意識しなくてすむというメリットがある。

【００８６】

２つ目は、ローカル側の記憶システムが非同期リモートコピーを行っている際、書込みコマンドを発行するホストコンピュータに対して、リモート側の記憶システムが書込みコマンドのデータをどの時点まで記憶したかを、定期的に、報告する方法である。例えば、ローカル側の記憶システムは、定期的に、ホストからのある書込み要求が発行された時刻を記憶しておき、この書込み要求以前に発行された全ての書込み要求に対応するデータがリモート側の記憶システムへ確実に記憶された段階で、先に記録した時刻をホストコンピュータへ返すようにする。

このようにすれば、ホストコンピュータは定期的にどの時刻までの書込み要求が確実にリモート側の記憶システムへ記憶されたかが分かるようになる。一方、ホスト側のアプリケーションは、従来と同様にコミットコマンドを発行する。このコミットコマンドを受けた時点でミドルソフトは、ローカル側の記憶システムへデータバッファ内のデータを書込むため、書込み要求を発行するが、この際、書込み要求を発行した最終時刻を記憶するようにする。そして、ミドルソフトは、この書込み要求に対する全ての書込み完了を受け、かつ、当該最終時刻に等しいか又は遅い、先に述べたローカル側の記憶システムが定期的に報告する時刻を受けた場合に、アプリケーションに対し、先のコミットコマンドに対するリターンを返すようにする。この２つ目の方法では、ホスト側のアプリケーションはＳｙｎｃコマンドを発行する必要が無いので、非同期コピー中であることを意識しなくてすむというメリットがある。また、記憶システム側は自発的にリモート側の状況をホストコンピュータに通知するだけなので、ホストコンピュータと記憶システム間のインタフェースが簡単なものとなる。

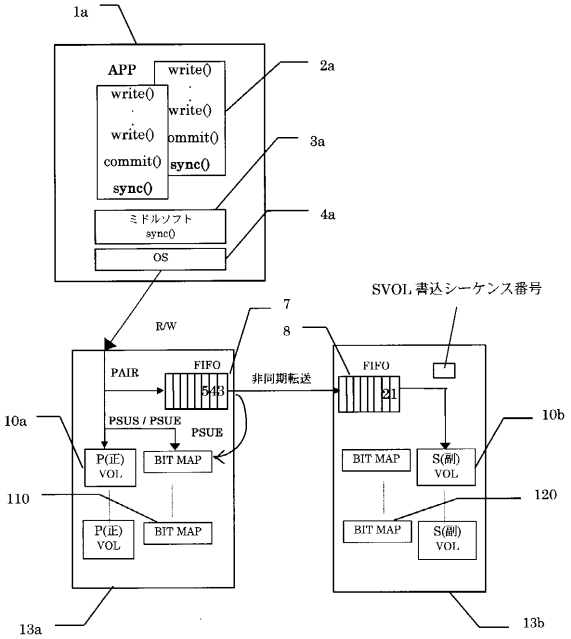
【００８７】

なお、この２つ目の方法において、ローカル側の記憶システムが定期的にどの時刻までの書込み要求をリモート側の記憶システムに確実に書込んだかをホストコンピュータに報告する例をあげている。このような機能を有する記憶システムは、本実施例で説明するようなアプリケーションのためのデータ同期確認にもちろん利用可能であるが、これ以外にも、ホストコンピュータ側から記憶システム状態を知る上で重要な情報を提供しているわけであり、本記憶システムはホスト側から管理しやすい、又は、制御しやすい、又は、使いやすい記憶システムということができる。

【００８８】
【発明の効果】
本発明では、非同期リモートコピーを行っている際、ホストに対し、書込み要求のデータがリモート側に記憶されたか否かを判断するのに必要なリモート側の記憶システムに関する情報をローカル側の記憶システム経由でホストに提供するので、アプリケーションは、ホスト内データとリモート側記憶システム内のデータとの一致を確認できる。
【００８９】
また、本発明では、非同期リモートコピーを行う記憶システムが、ホストに対し、書込み要求のデータがリモート側に記憶されたか否かを通知するので、ホスト内のアプリケーションは、ホスト内データとリモート側記憶システム内のデータとの一致を確認できる。
【図面の簡単な説明】
【図１】本発明の論理的なブロック図を示す。
【図２】本発明の全体制御フローを示す。
【図３】本発明を適用する一般的なＨＡクラスタの全体構成図である。
【図４】本発明の第２の実施例の論理的なブロック図を示す。
【図５】本発明の第２の実施例に相当する全体フローを示す。
【符号の説明】
１ａ、１ｂ　ホスト装置
２ａ、２ｂ　アプリケーションソフト（ＡＰＰ）
３ａ、３ｂ　ミドルソフト
３００　Ｓｙｎｃ制御
４ａ、４ｂ　オペレーティングシステム（ＯＳ）
５　ＰＶＯＬ最新シーケンス番号
６　ＳＶＯＬ書込みシーケンス番号
７　ＦＩＦＯ（ＰＶＯＬ対応）
８　ＦＩＦＯ（ＳＶＯＬ対応）
１０ａ　Ｐ（正）ＶＯＬ
１０ｂ　Ｓ（副）ＶＯＬ
１１０　ＢＩＴＭＡＰ（ＰＶＯＬ対応）
１２０　ＢＩＴＭＡＰ（ＳＶＯＬ対応）
１３ａ　ローカル側大型ディスクアレイ装置
１３ｂ　リモート側大型ディスクアレイ装置
３１　実施例のＳｙｎｃコマンド
５１　他の実施例のＳｙｎｃコマンド

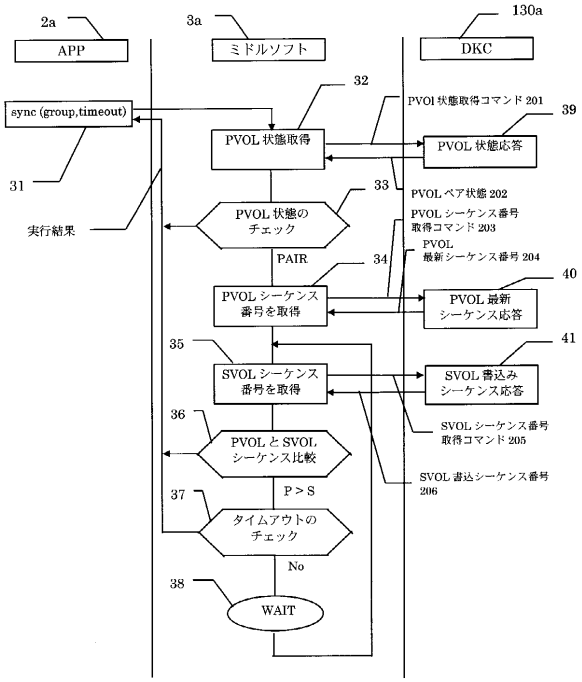
【図１】

図1



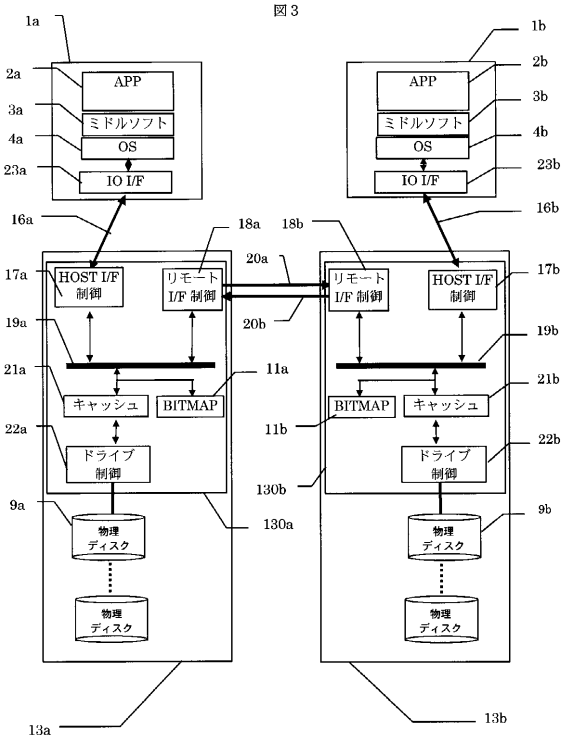
【図２】

図2



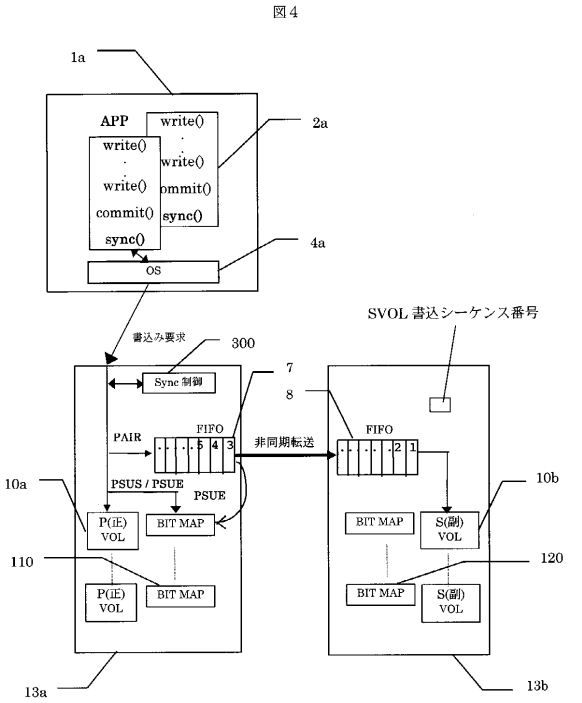
【図３】

図3



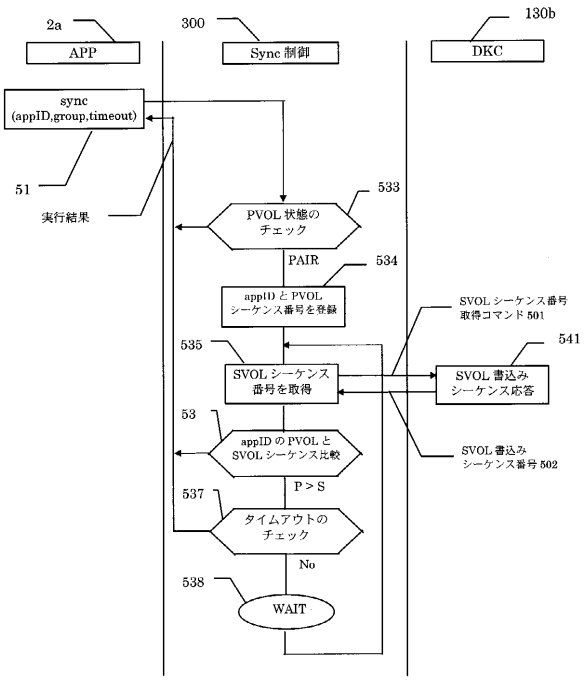
【図４】

図4



【図５】

図5


2a
APP
300
Sync 制御
130b
DKC
sync
(appID,group,timeout)
51
実行結果
533
PVOL 状態の
チェック
PAIR
534
appID と PVOL
シーケンス番号を登録
SVOL シーケンス番号
取得コマンド 501
541
535
SVOL シーケンス
番号を取得
SVOL 書込み
シーケンス応答
53
appID の PVOL と
SVOL シーケンス比較
SVOL 書込み
シーケンス番号 502
P > S
537
タイムアウトの
チェック
No
538
WAIT

フロントページの続き

(51)Int.Cl. FI
G06F 12/08 541
G06F 12/08 557
G06F 12/16 310J

(56)参考文献 特開平07－234811（JP, A）
特開平10－074157（JP, A）
特開平08－263359（JP, A）
特表平08－509565（JP, A）

(58)調査した分野(Int.Cl., DB名)
G06F 12/00
G06F 13/00
G06F 3/06
G06F 12/08
G06F 12/16