(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5286212号
(P5286212)
(45)発行日　平成25年9月11日(2013.9.11)　(24)登録日　平成25年6月7日(2013.6.7)

(51)Int.Cl.　F I
G06F 3/06 (2006.01)　G06F 3/06 304F

請求項の数 2 (全 21 頁)

(21)出願番号　特願2009-223621(P2009-223621)
(22)出願日　平成21年9月29日(2009.9.29)
(65)公開番号　特開2011-76130(P2011-76130A)
(43)公開日　平成23年4月14日(2011.4.14)
審査請求日　平成24年2月29日(2012.2.29)

(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
(74)代理人　100098660
弁理士　戸田　裕二
(72)発明者　平岩　友理
神奈川県川崎市麻生区王禅寺１０９９番地　株式会社日立製作所　システム開発研究所内
(72)発明者　西　好行
神奈川県横浜市戸塚区戸塚町５０３０番地　株式会社日立製作所　ソフトウェア事業部内

最終頁に続く

(54)【発明の名称】ストレージクラスタ環境でのリモートコピー制御方法及びシステム

(57)【特許請求の範囲】
【請求項１】
アプリケーションプログラムと、ストレージクラスタリング制御プログラムと、前記ストレージクラスタリング制御プログラムとは独立して動作するコピー制御プログラムと、を実行し、前記アプリケーションプログラムの実行に従って書き込みデータを送信する第一のホスト計算機と、
前記第一のホスト計算機と接続し、第一のボリュームを提供する第一のストレージ装置と、
前記第一のホスト計算機及び前記第一のストレージ装置と接続し、第二のボリュームを提供する第二のストレージ装置と、
前記第一のストレージ装置及び前記第二のストレージ装置と接続し、第三のボリュームを提供する第三のストレージ装置と、
を有する計算機システムであって、
前記第一のストレージ装置は、前記第一のボリュームに前記書き込みデータを格納し、
前記第一のストレージ装置は、前記第一のボリュームをコピー元とし、前記第二のボリュームをコピー先とする第一のコピーペアについての同期リモートコピー処理によって、前記書き込みデータを前記第二のストレージ装置へ送信し、
前記第一のストレージ装置は、前記書き込みデータに基づき、前記書き込みデータが格納された記憶領域のアドレスを含む情報と書き込みデータとの組である第一のジャーナルエントリーを作成し、

前記第一のストレージ装置は、前記第一のボリュームをコピー元とし、前記第三のボリュームをコピー先とする第二のコピーペアについての非同期リモートコピー処理によって、前記第一のジャーナルエントリーを前記第三のストレージ装置へ送信し、
前記第三のストレージ装置は、前記第一のジャーナルエントリーとして送信された前記書き込みデータを前記第三のボリュームに格納し、
前記第二のストレージ装置は、前記第一のストレージ装置より送信された前記書き込みデータを前記第二のボリュームに格納し、前記書き込みデータに基づき第二のジャーナルエントリーを作成し、
前記ストレージクラスタリング制御プログラムを実行することで、前記第一のホスト計算機は、前記第一のボリュームへの前記書き込みデータの書き込みが異常終了したことを検知し、前記第一のコピーペアのペア状態を変更することで検知後の前記書き込みデータを前記第二のボリュームに対して書き込み可能とするストレージ装置切り替え指示を前記第二のストレージ装置に送信し、前記ストレージ装置切り替え指示に基づく切り替え処理の完了後に、前記切り替え処理を完了したことを示す情報を出力し、
前記コピー制御プログラムを実行することで、前記第一のホスト計算機は、前記切り替え処理を完了したことを示す情報の出力の有無を監視し、前記出力を検知することにより、前記ストレージクラスタリング制御プログラムによって異常終了を検知し前記切り替え処理を完了したことを、前記ストレージ装置切り替え指示の送信より後に検知し、前記第二のボリュームをコピー元とし、前記第三のボリュームをコピー先とする第三のコピーペアについての別な非同期リモートコピー処理によるコピー処理を開始するための差分リシンク指示を前記第二のストレージ装置に送信し、
前記第二のストレージ装置は、前記差分リシンク指示を受信すると、前記切り替え処理が行われる前より作成した前記第二のジャーナルエントリーを、前記第三のストレージ装置に送信することにより、差分リシンクを実行し、
前記計算機システムは、前記第三のストレージ装置と接続し、前記アプリケーションプログラムを実行することで前記第三のボリュームに格納したデータを読み込む第二のホスト計算機を有し、
前記第一のホスト計算機は、前記第一のコピーペアが前記ストレージクラスタリング制御プログラムの制御対象であることを示すコピー定義情報を有し、
前記コピー制御プログラムを実行することで、前記第一のホスト計算機は、前記第一のコピーペア及び前記第二のコピーペアを対象とするペア状態指示を受付け、前記第一のコピーペアのペア状態又は前記第二のコピーペアのペア状態を表示し、
前記コピー制御プログラムを実行することで、前記第一のホスト計算機は、前記コピー定義情報に基いて、前記第二のコピーペア及び前記第三のコピーペアを対象とする状態変更指示は処理し、前記第一のコピーペアを対象とする状態変更指示の処理は抑止し、
前記コピー制御プログラムを実行することで、前記第一のホスト計算機は、前記差分リシンク指示送信に基く第三のコピーペアを対象とする副ボリューム間差分リシンクが差分コピーで処理できないことを示す情報を前記第二のストレージ装置から取得し、前記ストレージクラスタリング制御プログラムによる前記第一のコピーペアの状態変化を要因とした前記第三のコピーペアの副ボリューム間差分リシンクが失敗したメッセージを表示する、
ことを特徴とする計算機システム。

【請求項２】

アプリケーションプログラムと、ストレージクラスタリング制御プログラムと、前記ストレージクラスタリング制御プログラムとは独立して動作するコピー制御プログラムと、を実行し、前記アプリケーションプログラムの実行に従って書き込みデータを送信する第一のホスト計算機と、
前記第一のホスト計算機と接続し、第一のボリュームを提供する第一のストレージ装置と、
前記第一のホスト計算機及び前記第一のストレージ装置と接続し、第二のボリュームを

提供する第二のストレージ装置と、
前記第一のストレージ装置及び前記第二のストレージ装置と接続し、第三のボリュームを提供する第三のストレージ装置と、
を有する計算機システムであって、
前記第一のストレージ装置は、前記第一のボリュームに前記書き込みデータを格納し、
前記第一のストレージ装置は、前記第一のボリュームをコピー元とし、前記第二のボリュームをコピー先とする第一のコピーペアについての同期リモートコピー処理によって、前記書き込みデータを前記第二のストレージ装置へ送信し、
前記第一のストレージ装置は、前記書き込みデータに基づき、前記書き込みデータが格納された記憶領域のアドレスを含む情報と書き込みデータとの組である第一のジャーナルエントリーを作成し、
前記第一のストレージ装置は、前記第一のボリュームをコピー元とし、前記第三のボリュームをコピー先とする第二のコピーペアについての非同期リモートコピー処理によって、前記第一のジャーナルエントリーを前記第三のストレージ装置へ送信し、
前記第三のストレージ装置は、前記第一のジャーナルエントリーとして送信された前記書き込みデータを前記第三のボリュームに格納し、
前記第二のストレージ装置は、前記第一のストレージ装置より送信された前記書き込みデータを前記第二のボリュームに格納し、前記書き込みデータに基づき第二のジャーナルエントリーを作成し、
前記ストレージクラスタリング制御プログラムを実行することで、前記第一のホスト計算機は、前記第一のボリュームへの前記書き込みデータの書き込みが異常終了したことを検知し、前記第一のコピーペアのペア状態を変更することで検知後の前記書き込みデータを前記第二のボリュームに対して書き込み可能とするストレージ装置切り替え指示を前記第二のストレージ装置に送信し、前記ストレージ装置切り替え指示に基づく切り替え処理の完了後に、前記切り替え処理を完了したことを示す情報を出力し、
前記コピー制御プログラムを実行することで、前記第一のホスト計算機は、前記切り替え処理を完了したことを示す情報の出力の有無を監視し、前記出力を検知することにより、前記ストレージクラスタリング制御プログラムによって異常終了を検知し前記切り替え処理を完了したことを、前記ストレージ装置切り替え指示の送信より後に検知し、前記第二のボリュームをコピー元とし、前記第三のボリュームをコピー先とする第三のコピーペアについての別な非同期リモートコピー処理によるコピー処理を開始するための差分リシンク指示を前記第二のストレージ装置に送信し、
前記第二のストレージ装置は、前記差分リシンク指示を受信すると、前記切り替え処理が行われる前より作成した前記第二のジャーナルエントリーを、前記第三のストレージ装置に送信することにより、差分リシンクを実行し、
前記計算機システムは、前記第三のストレージ装置と接続し、前記アプリケーションプログラムを実行することで前記第三のボリュームに格納したデータを読み込む第二のホスト計算機を有し、
前記第一のホスト計算機は、前記第一のコピーペアが前記ストレージクラスタリング制御プログラムの制御対象であることを示すコピー定義情報を有し、
前記コピー制御プログラムを実行することで、前記第一のホスト計算機は、前記第一のコピーペア及び前記第二のコピーペアを対象とするペア状態指示を受付け、前記第一のコピーペアのペア状態又は前記第二のコピーペアのペア状態を表示し、
前記コピー制御プログラムを実行することで、前記第一のホスト計算機は、前記コピー定義情報に基いて、前記第二のコピーペア及び前記第三のコピーペアを対象とする状態変更指示は処理し、前記第一のコピーペアを対象とする状態変更指示の処理は抑止し、
前記コピー制御プログラムを実行することで、前記第一のホスト計算機は、前記差分リシンク指示送信に基く第三のコピーペアを対象とする副ボリューム間差分リシンクが差分コピーで処理できないことを示す情報を前記第二のストレージ装置から取得し、前記第二のボリュームからの全コピーによって前記第三のコピーペアのコピーを再開する第二の指

示を送信する、
ことを特徴とする計算機システム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、コンピュータシステムにおけるデータの保護技術に関する。
【背景技術】
【０００２】
社会基盤を担う公共あるいは企業の基盤システムにおいては、当然ながら高い可用性が求められている。そのため、アプリケーションの継続性を実現したストレージクラスタリング技術として特許文献１記載の技術がある。これは通常の業務に使用される第１のストレージ装置と第２のストレージ装置の間のデータの同一性は、同期リモートコピーによって確保される。業務を行うホストにより第１のストレージ装置の障害を検出した際に、ホストアプリケーションの入力／出力が第２のストレージ装置に対して実行できるように、ホストの制御ブロック内の装置アドレス情報を再構成し、ポインタの切り替えを行う。また、同期リモートコピーの制御もあわせて行う。この技術により、ホストアプリケーションを静止させる必要なく、第１のストレージ装置を、第２のストレージ装置に切り替えることができ、業務の可用性を向上させることができる。
【０００３】
上記に加えて、データストレージ市場では、大量のデータを格納したストレージ装置が災害等で破壊されてもデータを喪失しない、いわゆるディザスタリカバリシステムが要求されている。このような市場の要求に応えるべく、リモートコピー技術を利用してデータをバックアップする計算機システムが提供されている。これは、十分に離れた二つの地点に設置されたストレージ装置に同一のデータを格納するものである。一方のストレージ装置のデータが更新されると、その更新は、リモートコピーによってもう一方のストレージ装置に同期的もしくは非同期的に反映される。このため、二つのストレージ装置のデータの同一性が確保される。
【０００４】
さらにデータの安全性を高めるため、相互に十分に離れた三つの地点にストレージ装置を設置する計算機システムが特許文献２に開示されている。この計算機システムでは、通常の業務に使用される第１のストレージ装置と、遠隔地の第２のストレージ装置との間のデータの同一性は、同期リモートコピーによって確保される。一方、第１のストレージ装置と、遠隔地の第３のストレージ装置との間のデータの同一性は、非同期リモートコピーによって確保される。
【０００５】
災害等に起因する障害によって第１のストレージ装置を業務に使用することができなくなった場合、第２のストレージ装置が第１のストレージ装置の業務を引き継ぐ。このとき、第２のストレージ装置も使用することができない場合は、第３のストレージ装置が第１のストレージ装置の業務を引き継ぐ。その結果、深刻な災害が発生した場合でも、データの喪失を防ぐことができる。
【０００６】
このように三つの地点にストレージ装置を設置する計算機システムでは、通常の運用時には、第２のストレージ装置と第３のストレージ装置との間でデータが複製されない。このため、第２のストレージ装置と第３のストレージ装置との間では、データの同一性が保証されない。したがって、第１のストレージ装置の業務を第２のストレージ装置が引き継いだ後で、さらに第２のストレージ装置に障害が発生した場合、第３のストレージ装置が第２のストレージ装置の業務を引き継ぐことができない。
【０００７】
このため、第１のストレージ装置の業務を引き継いだ第２のストレージ装置の運用が開始される前に、第２のストレージ装置と第３のストレージ装置との間のデータの同一性を

確保する。第２のストレージ装置の運用が開始された後は、第２のストレージ装置のデータの更新をリモートコピーによって第３のストレージ装置に反映させる。その結果、第２のストレージ装置に障害が発生したときに、第３のストレージ装置が第２のストレージ装置の業務を引き継ぐことができる。

【０００８】

この引き継ぎの際に、第２のストレージ装置のデータを全て第３のストレージ装置に複製すれば、これらのストレージ装置のデータの同一性が確保される。しかし、このように全てのデータを複製することは、長い時間を要する。特に、近年の大容量のストレージ装置では、数時間以上を要する場合がある。全てのデータの複製が終了するまで、第２のストレージ装置を業務に使用することができないとすれば、長時間のシステム停止によって深刻な経済的損失が発生するおそれがある。この第２のストレージ装置と第３のストレージ装置とのデータの同一性の確保を行う時間を短縮する手段として、特許文献３記載のデータの更新方法がある。この技術は、第２のストレージ装置と第３のストレージ装置の同一性を確保する際に、相互の差異のデータをもう片方に反映することで、データの複製容量を削減し、結果として時間短縮を実現するものである。本明細書では、この特許文献３に示されるようなデータの同一性の確保技術を「副ボリューム間差分リシンク」と呼ぶ。

【先行技術文献】

【特許文献】

【０００９】

【特許文献１】米国特許第７０４３６６５号明細書

【特許文献２】米国特許第７１６７９６２号明細書

【特許文献３】米国特許第７４４７８５５号明細書

【発明の概要】

【発明が解決しようとする課題】

【００１０】

特許文献２及び特許文献３の非同期リモートコピーを特許文献１のストレージクラスタリング技術へ適用する場合、管理者が指定した、クラスタを構成する片方のストレージ装置をコピー元として非同期リモートコピーを行うことが考えられ、クラスタ制御とディザスタリカバリ用途の非同期リモートコピーとが連携してより高性能または高信頼なシステムとして動作することができない。

【００１１】

本発明は、ストレージクラスタリング技術と連携した非同期リモートコピーを提供することを目的とする。

【課題を解決するための手段】

【００１２】

本発明では、非同期リモートコピーを制御するホスト計算機上のプログラムは、ストレージクラスタリング環境でのホスト書き込み先ボリュームの切り替えを行うストレージクラスタリング制御プログラムの切り替え指示と非同期に連携して、非同期リモートコピーのペア操作を行う。

【発明の効果】

【００１３】

本発明によれば、クラスタ制御と連携した高性能または高信頼なディザスタリカバリ用途の非同期リモートコピーを実現することができる。

【図面の簡単な説明】

【００１４】

【図１】図１は、本実施形態の計算機システムの構成を表している。

【図２】図２は、ストレージ装置の構成の詳細を表している。

【図３】図３は、コピー制御情報の構成の詳細を表している。

【図４】図４は、コピーグループ情報の構成の詳細を表している。

【図５】図５は、ペア情報の構成の詳細を表している。

【図６】図６は、ストレージクラスタリング及び副ボリューム間差分リシンクを適用したコピーグループの関係を示した模式図である。
【図７】図７は、ストレージクラスタリング制御プログラムの動作を示したフローである。
【図８】図８は、コピー制御プログラムのストレージクラスタリング制御に対応するための動作を示したフローである。
【図９】図９は、コピー制御プログラムの実際にコマンド発行処理を行う部分の動作を示したフローである。
【図１０】図１０は、コマンドラインインタフェースの説明図である。
【発明を実施するための形態】
【００１５】
本発明の実施例を、図１乃至図１０を用いて説明する。
【００１６】
計算機からの入出力制御や、ストレージ装置内でのボリュームに基づく記憶領域使用方法は、従来技術に示したものと同様である。
【実施例１】
【００１７】
図１は、本実施形態の計算機システムの構成を表している。本システムは、ホスト計算機１０００と、ホスト計算機１０００からアクセスするストレージ装置２０００、ホスト計算機１０００とストレージ装置２０００をつなぐデータネットワーク１００９、管理計算機１１００、ホスト計算機１０００とストレージ装置２０００と管理計算機をつなぐ管理ネットワーク１０１０で構成される。ホスト計算機１０００には、ホスト計算機１０００の設定など、運用を管理するためのホスト計算機入出力装置１００３が付属している。管理計算機には、管理計算機の設定など、運用を管理するための管理計算機入出力装置が付属している。
【００１８】
データネットワーク１００９は、データ通信用のネットワークであって、本実施形態では、ＳＡＮ（Ｓｔｏｒａｇｅ　Ａｒｅａ　Ｎｅｔｗｏｒｋ）である。なお、データネットワーク１００９は、データ通信用のネットワークであればＳＡＮ以外のネットワークでもよく、例えばＩＰネットワークでもよい。
【００１９】
また、管理ネットワーク１０１０は、データ通信用のネットワークであって、本実施形態では、ＩＰネットワークである。なお、管理ネットワーク１０１０は、データ通信用のネットワークであればＩＰネットワーク以外のネットワークでもよく、例えばＳＡＮ（Ｓｔｏｒａｇｅ　Ａｒｅａ　Ｎｅｔｗｏｒｋ）でもよい。
【００２０】
また、データネットワーク１００９及び管理ネットワーク１０１０は、同一のネットワークであってもよい。また、管理計算機とホスト計算機１０００とは、一つの計算機によって実現されてもよい。
【００２１】
なお、説明の都合上、本実施形態では、ストレージ装置２０００を３台、ホスト計算機１０００を１台としたが、これらは複数あってもよい。
【００２２】
ストレージ装置２０００は、データを格納する領域であるボリューム２３０１を有する。また、ボリューム２３０１も、データをホスト計算機１０００からの書き込みにより格納するデータボリューム（「Ａ０００１」「Ａ０００２」と図示）と、リモートコピーを行う際にコピー中のデータを格納するジャーナルボリューム（「ＪＮＬ－Ａ」と図示）という用途がある。本実施形態のリモートコピーは、「Ａ０００１」から「Ｃ０００１」がペア構成でデータのコピーが行われる場合、「Ａ０００１」の書き込みのデータが「ＪＮＬ－Ａ」にも格納され、「ＪＮＬ－Ｃ」に転送され、「ＪＮＬ－Ｃ」をもとに「Ｃ０００

１」に反映されるものとする。ここではストレージ装置をボリュームについてのみ記載している。ストレージ装置の詳細は後述する。

【００２３】

なお、以後の説明ではクラスタリング対象の二つのストレージ装置を指す場合はストレージ装置Ａやストレージ装置Ｂと呼び、遠隔地のストレージ装置を指す場合はストレージ装置Ｃと呼び、特定のストレージ装置を指さない場合はストレージ装置またはストレージ装置２０００と呼ぶ。

【００２４】

ボリューム２３０１は、コピーペアを構成できる。なお、コピーペアは、複製元のボリューム２３０１と、当該複製元のボリューム２３０１に記憶されたデータの複製を記憶する複製先のボリューム２３０１とから構成される。

【００２５】

ホスト計算機１０００は、ＣＰＵ１００１、メモリ１００２、キーボードやマウスやディスプレイ等の入出力装置１００３、ストレージＩ／Ｆ１００４、管理Ｉ／Ｆ１００５を備え、各々が接続されている。

【００２６】

ストレージＩ／Ｆ１００４は、ホスト計算機１０００をデータネットワーク１００９に接続するネットワークインタフェースである。ストレージＩ／Ｆ１００４は、データネットワーク１００９を介してストレージ装置２０００とデータ及び制御指示を送受信する。

【００２７】

管理Ｉ／Ｆ１００５は、ホスト計算機１０００を管理ネットワーク１０１０に接続するネットワークインタフェースである。管理Ｉ／Ｆ１００５は、管理ネットワーク１０１０を介してストレージ装置２０００及び管理計算機とデータ及び制御指示を送受信する。

【００２８】

メモリ１００２には、アプリケーション１００６、ストレージクラスタリング制御プログラム３０００、コピー制御プログラム４０００、ＯＳ１００７、入出力構成制御情報１００８、及び、コピー制御情報５０００が格納される。

【００２９】

アプリケーション１００６は、ＯＳ１００７に要求を出すことにより、ストレージ装置２０００に備わるボリューム２３０１にデータを読み書きする。例えば、アプリケーション１００６は、ＤＢＭＳ（Ｄａｔａ　Ｂａｓｅ　Ｍａｎａｇｅｍｅｎｔ　Ｓｙｓｅｔｅｍ）又はファイルシステム等である。

【００３０】

ＯＳ１００７は、このシステムを管理するオペレーティングシステムである。アプリケーション１００６などからの要求に従い、入出力の対象となる機器を入出力構成制御情報１００８から特定し、ストレージ装置とのデータの入力／出力を行う機能をその一部として有する。

【００３１】

ストレージクラスタリング制御プログラム３０００は、従来技術記載のストレージクラスタリングを制御する。すなわち、ストレージクラスタリングの対象となるボリューム間の同期コピーの制御、及び、入出力対象となっているボリュームの障害が発生した場合に、入出力を一旦停止させ、同期コピーのコピー方向を逆転させ、入出力構成制御情報内のボリューム情報を書き換えることで入出力が元の同期コピー先に切り替わるようにした上で入出力を再開させるという処理を行う。

【００３２】

コピー制御プログラム４０００は、コピー運用の手順に基づきコピー制御を行う。また、コピー制御のため、ストレージマイクロプログラム２２０６にコピーペアを制御させる要求及びコピーペアの状態を取得させる要求を送信する。

【００３３】

コピー制御情報５０００は、コピー制御プログラム４０００がコピー制御を行うための

情報を格納した領域である。詳細は後述する。
アプリケーション１００６、ストレージクラスタリング制御プログラム３０００、コピー制御プログラム４０００、ＯＳ１００７はＣＰＵ１００１によって実行される。
【００３４】
なお、説明の都合上、図１では、アプリケーション１００６を一つとしたが、複数あってもよい。
【００３５】
なお、ストレージ装置Ｃは非同期リモートコピーによって複製されたデータを用い、ホスト計算機１０００が実行していた業務処理を実行するホスト計算機１０００Ｂが接続される。
【００３６】
なお、典型的にはホスト計算機１０００とストレージ装置Ａとストレージ装置Ｂとは一つのサイト（ローカルサイト）に存在し、ホスト計算機１０００Ｂとストレージ装置Ｃとは別なサイト（リモートサイト）に属することで、ローカルサイトでの災害を想定したディザスタリカバリ環境を構築する。しかし、サイトと各装置・計算機の関係はこれ以外の場合もありうる。
【００３７】
図２は、図１におけるストレージ装置の構成の詳細を表している。
【００３８】
ストレージ装置２０００は、ディスク装置２１００及びディスクコントローラ２２００を備える。ディスク装置２１００は、ホスト計算機１０００に書き込み要求されたデータを格納し、ホスト計算機１０００から読み込み要求された格納データをホスト計算機に送信する。ディスクコントローラ２２００は、ストレージ装置２０００の処理を制御する。
【００３９】
ディスク装置２１００は、複数のボリューム２３０１を備える。ボリューム２３０１は、物理的な記憶領域であるハードディスクドライブ（ＨＤＤ）、フラッシュドライブ（ＳＳＤ）、又は、論理的な記憶領域である論理デバイス（Ｌｏｇｉｃａｌ　Ｄｅｖｉｃｅ）のいずれであってもよく、本発明では、ボリュームの種類を問わない。なお、説明の都合上、図２では、３つのボリューム２３０１を示したが、ボリューム２３０１の数はいくつであってもよい。
【００４０】
ディスクコントローラ２２００は、ホストＩ／Ｆ２２０１、管理Ｉ／Ｆ２２０２、ディスクＩ／Ｆ２２０３、メモリ２２０４、及び、ＣＰＵ２２０５を備える。
【００４１】
メモリ２２０４には、ストレージマイクロプログラム２２０６及びコピーペア情報２２０７が格納される。
【００４２】
ストレージマイクロプログラム２２０６は、ＣＰＵ２２０５によって実行される。ストレージマイクロプログラム２２０６は、ホスト計算機１０００からの要求に応じて、コピーペアを制御し、コピーペアの状態を取得し報告する。
【００４３】
ストレージマイクロプログラム２２０６によるコピーペアの制御については後ほど説明する。
【００４４】
コピーペア情報２２０７は、ストレージ装置２０００に備わるボリューム２３０１のうち、コピーペアを構成するボリューム２３０１の情報が格納される。
【００４５】
なお、本実施形態では、ストレージマイクロプログラム２２０６及びコピーペア情報２２０７は、ディスクコントローラ２２００のメモリ２２０４に格納されるとしたが、本発明はこれに限定されない。例えば、ストレージマイクロプログラム２２０６及びコピーペ

ア情報２２０７は、ディスクコントローラ２２００に接続されるフラッシュメモリに格納されてもよいし、ディスク装置２１００に備わるボリューム２３０１に格納されてもよい。

【００４６】

ホストＩ／Ｆ２２０１は、ストレージ装置２０００をデータネットワーク１００９に接続するネットワークインタフェースである。ホストＩ／Ｆ２２０１は、データネットワーク１００９を介してホスト計算機１０００とデータ及び制御指示を送受信する。

【００４７】

管理Ｉ／Ｆ２２０２は、ストレージ装置２０００を管理ネットワーク１０１０に接続するネットワークインタフェースである。管理Ｉ／Ｆ２２０２は、管理ネットワーク１０１０を介してホスト計算機１０００及び管理計算機１２００とデータ及び制御指示を送受信する。

【００４８】

ディスクＩ／Ｆ２２０３は、ディスクコントローラ２２００をディスク装置２１００に接続するインタフェースである。

【００４９】

次に、ストレージ装置によるリモートコピーについて説明する。

【００５０】

同期リモートコピーは、主ボリュームに対してホストから書き込み要求を受け付けた場合、当該書き込みデータを副ボリュームにデータコピーした後に、ホストに対して書き込み完了を返すコピー方式である。

【００５１】

同期リモートコピーが実行される際、正と副のボリュームのデータおよびコピーの状況を示したり操作するために、ディスクコントローラはペア状態（Ｓｉｍｐｌｅｘ、Ｉｎｉｔｉａｌ－Ｃｏｐｙｉｎｇ、Ｄｕｐｌｅｘ、Ｓｕｓｐｅｎｄ及びＤｕｐｌｅｘ－Ｐｅｎｄｉｎｇ）と呼ばれる情報を管理する。以下、各ペア状態毎に同期リモートコピーの処理内容について説明する。

【００５２】

＜Ｓｉｍｐｌｅｘ状態＞

Ｓｉｍｐｌｅｘ状態は、正と副のボリューム間でコピーが開始されていない状態である。

【００５３】

＜Ｄｕｐｌｅｘ状態＞

Ｄｕｐｌｅｘ状態は、同期リモートコピーが開始され、後述する初期化コピーも完了して正副のボリュームのデータ内容が同一となった状態である。同期リモートコピーの場合、主ボリュームに対して行われた書き込みの内容が副ボリュームに対してコピーされた後に、書き込みを行ったホストに対して正常完了のメッセージが返される。従って、書き込み途中の領域を除けば、主ボリュームのデータと副ボリュームのデータの内容は同じとなる。

【００５４】

＜Ｉｎｉｔｉａｌ－Ｃｏｐｙｉｎｇ状態＞

Ｉｎｉｔｉａｌ－Ｃｏｐｙｉｎｇ状態は、Ｓｉｍｐｌｅｘ状態からＤｕｐｌｅｘ状態へ遷移するまでの中間状態であり、この期間中に、必要ならば主ボリュームから副ボリュームへの初期化コピー（主ボリュームに既に格納されていたデータのコピー）が行われる。初期化コピーが完了し、Ｄｕｐｌｅｘ状態へ遷移するために必要な処理が終わったら、ペア状態はＤｕｐｌｅｘとなる。なお、当該状態にはＳｉｍｐｌｅｘ状態にてホスト計算機から「作成」指示を受信した契機で遷移する。

【００５５】

＜Ｓｕｓｐｅｎｄ状態＞

Ｓｕｓｐｅｎｄ状態は、主ボリュームに対する書き込みの内容を副ボリュームに反映さ

せない状態である。この状態では、正副のボリュームのデータは同じでない。オペレータやホストからの「停止」指示を契機に、ペア状態は他の状態からＳｕｓｐｅｎｄ状態へ遷移する。

【００５６】

それ以外に、同期リモートコピーを行うことが出来なくなった場合に自動的にペア状態がＳｕｓｐｅｎｄ状態に遷移することが考えられる。

【００５７】

以下の説明では、後者の場合を障害Ｓｕｓｐｅｎｄ状態と呼ぶことにする。障害Ｓｕｓｐｅｎｄ状態となる代表的な原因としては、正副のボリュームの障害、正副のディスクコントローラの障害、正副間の通信障害が考えられる。Ｓｕｓｐｅｎｄ状態となった正副のディスクコントローラは、当該状態となった以降の正と副のボリュームに対する書き込み位置を記憶する。

【００５８】

＜Ｄｕｐｌｅｘ－Ｐｅｎｄｉｎｇ状態＞

Ｄｕｐｌｅｘ－Ｐｅｎｄｉｎｇ状態は、Ｓｕｓｐｅｎｄ状態からＤｕｐｌｅｘ状態に遷移するまでの中間状態である。この状態では、主ボリュームと副ボリュームのデータの内容を一致させるために、主ボリュームから副ボリュームへのデータのコピーが実行される。正と副のボリュームのデータが同一になった後、ペア状態はＤｕｐｌｅｘとなる。なお、Ｄｕｐｌｅｘ－Ｐｅｎｄｉｎｇ状態におけるデータのコピーは、Ｓｕｓｐｅｎｄ状態で正と副のディスクコントローラが記録した書き込み位置を利用して更新が必要な部分だけをコピーする差分コピーが用いられる。なお、当該状態にはＳｕｓｐｅｎｄ状態（障害Ｓｕｓｐｅｎｄ状態を含む）にてホスト計算機から「リシンク」指示を受信した契機で遷移する。

【００５９】

なお、以上の説明ではＩｎｉｔｉａｌ－Ｃｏｐｙｉｎｇ状態とＤｕｐｌｅｘ－Ｐｅｎｄｉｎｇ状態は別々な状態としたが、これらをまとめて一つの状態として管理装置の画面に表示したり、状態を遷移させても良い。

【００６０】

次に非同期リモートコピーについて説明する。

【００６１】

非同期リモートコピーの場合、Ｄｕｐｌｅｘ状態の副記憶領域に対する書き込みデータの反映は、記憶装置１２０のホスト１１０に対する書き込みの正常完了メッセージの送信とは無関係（非同期）に行われる。

【００６２】

非同期リモートコピーの場合、正記憶領域から副記憶領域へのデータコピーの方法として、以下の方法がある。

【００６３】

例えば、正記憶装置が、データが書き込まれた記憶領域のアドレスを含んだ制御情報及び書き込まれたデータの組（以下「ジャーナルエントリ」と呼ぶ）を、データの書き込みの度に作成し、これを副記憶装置へ転送し、副記憶領域へ反映させる方法がある（記憶先はキャッシュメモリ又は／及びジャーナルボリュームである）。さらにこの発展形として、ジャーナルエントリの制御情報に書き込みの時間順序を示す情報を含め、副記憶領域へジャーナルエントリを反映させる際にはこの時間順序を示す情報を利用して時間順序どおりに反映する方法がある。

【００６４】

また、本方法の効率的な方法として、正記憶領域の同一領域に対する書き込みが連続して発生した場合には、正記憶装置は、途中の書き込みに対するジャーナルエントリは副記憶装置へ転送せず、最後の書き込みに対するジャーナルエントリのみを転送する方法がある。

【００６５】

さらに、もう一つの例としては、ある一定時間に主ボリュームに対して書き込まれたデータを差分データとして保持し、副ボリュームへコピーする方法がある。この方法の場合、差分データを副ディスクコントローラへ全て転送してからボリュームへコピーする。

【００６６】

非同期リモートコピーも、ペア状態（Ｓｉｍｐｌｅｘ、Ｉｎｉｔｉａｌ－Ｃｏｐｙｉｎｇ、Ｄｕｐｌｅｘ、Ｓｕｓｐｅｎｄ、Ｄｕｐｌｅｘ－Ｐｅｎｄｉｎｇ及びＳｕｓｐｅｎｄｉｎｇ）を用いて操作と管理を行う。Ｓｉｍｐｌｅｘ、Ｉｎｉｔｉａｌ－Ｃｏｐｙｉｎｇ、Ｓｕｓｐｅｎｄ及びＤｕｐｌｅｘ－Ｐｅｎｄｉｎｇ状態については同期リモートコピーと同様である。

【００６７】

＜Ｄｕｐｌｅｘ状態＞

Ｄｕｐｌｅｘ状態も基本的には同期リモートコピーの場合と同じであるが、書き込みデータの副ボリュームへのコピーが非同期に行われるため、副ボリュームは主ボリュームより少し遅れながら追従する。

【００６８】

＜Ｓｕｓｐｅｎｄｉｎｇ状態＞

Ｓｕｓｐｅｎｄｉｎｇ状態とは、Ｄｕｐｌｅｘ状態からＳｕｓｐｅｎｄ状態へ遷移するまでの中間状態である。非同期リモートコピーの場合は、Ｓｕｓｐｅｎｄｉｎｇ状態を経由してＳｕｓｐｅｎｄ状態へ遷移する。同期リモートコピーのＳｕｓｐｅｎｄ状態で述べた正副のボリュームに対する書き込み位置の記録には、コピーできなかったジャーナルエントリを書き込み位置の記録に加える。

【００６９】

図３は、図１におけるコピー制御情報の構成の詳細を表している。

【００７０】

コピー制御情報５０００は、コピーグループ情報５１００、ペア情報５２００、ペア状態情報５３００、コピー環境情報５４００から構成される。

【００７１】

コピーグループ情報５１００は、複数のコピーペアをグループ化したコピーグループに関する情報を示す。なお、コピーグループ情報５１００については、図４で詳細を説明する。

【００７２】

ペア情報５２００は、コピーグループ毎に、グループとしてまとめて操作するコピーペアの情報を有する。なお、ペア情報５２００については、図５で詳細を説明する。

【００７３】

ペア状態情報５３００は、ペア情報５２００に記載されたペア各々のペア状態の情報を保持する領域であり、コピー制御プログラム４０００がストレージ装置２０００に発行したペアの状態取得指示の結果得られた状態が格納される。

【００７４】

コピー環境情報５４００は、コピー制御を行うパラメタを格納する領域である。本実施の形態では、オプション５４０１を有し、ここには副ボリューム間差分リシンクを実行する場合に、差分リシンクができなかった場合はリシンク処理を中断する「中断」と、差分リシンクができなかった場合は差分リシンクのかわりにボリューム全体のコピーを行ってリシンク処理を行う「続行」の、２つのいずれかの値が設定される。

ペア状態情報５３００に格納される指示の結果得られた情報以外のこれらの値は、入出力装置１００３もしくは管理計算機を介してユーザが設定する。

【００７５】

図４は、図３におけるコピーグループ情報５１００の構成の詳細を表している。

【００７６】

コピーグループ情報５１００は、グループ識別子５１０１、第一属性５１０２、第二属性５１０３のセットからなる。グループ識別子５１０１は、コピーグループの識別子を格

納する領域である。第一属性は、同期コピーや非同期コピーなど、コピー種別を格納する領域である。第二属性５１０３は、ストレージクラスタリング機能の適用対象であるか否かという、第一属性５１０２とは別の機能での属性を格納する領域である。

【００７７】

図５は、図３におけるペア情報５２００の構成の詳細を表している。

【００７８】

ペア情報５２００は、グループ識別子５２０１、プライマリデバイス識別子５２０２、セカンダリデバイス識別子５２０３を含む。図３はコピーグループがＧ１からＧ３まで３つある場合の例である。例えばＧ１は、デバイス識別子が「Ａ０００１」を主ボリューム、「Ｂ０００１」を副ボリュームとするペアと、「Ａ０００２」を主ボリューム、「Ｂ０００２」を副ボリュームとするペアの２つのペアから構成されている。このデバイス識別子により、対象となる装置、及び、その装置内のいずれのボリュームかが特定できる。

【００７９】

図６は、図４、及び、図５における情報の例示を元にした、本実施形態におけるストレージクラスタリング及び副ボリューム間差分リシンクを適用したコピーグループの関係を示した模式図である。

【００８０】

Ｇ１は、ストレージクラスタリングが適用中であるコピーグループで、ストレージ装置上は同期コピーが設定され、「Ａ０００１」ボリュームへの書き込みは、同期的に「Ｂ０００１」ボリュームに反映される。同様に「Ａ０００２」ボリュームへの書き込みは、同期的に「Ｂ０００２」ボリュームに反映される。すなわちＧ１は、「Ａ０００１」と「Ｂ０００１」、及び、「Ａ０００２」と「Ｂ０００２」の２つのコピーペアから構成されている。

【００８１】

更に、「Ａ０００１」ボリュームと、「Ａ０００２」ボリュームは、Ｇ２として非同期コピーが設定されている。すなわちＧ２は、「Ａ０００１」と「Ｃ０００１」、及び、「Ａ０００２」と「Ｃ０００２」の２つのコピーペアから構成され、書き込みが非同期的にコピーされている。

【００８２】

このＧ１とＧ２で示された構成は、マルチターゲット構成と呼ばれる。

【００８３】

また、Ｇ１とＧ２のコピーグループを対象に、副ボリューム間差分リシンクを適用するためのＧ３が設定されている。Ｇ１の副ボリュームとＧ２の副ボリュームを対象に、「Ｂ０００１」と「Ｃ０００１」、及び、「Ｂ０００２」と「Ｃ０００２」の２つのコピーペアから構成される。通常Ｇ３ではボリュームへの書き込みの反映はおこなわれない。これを「差分リシンク待機」と呼び、図６ではこれを破線で示している。Ｇ３は、副ボリューム間差分リシンク指示がされた時に、差分のデータの調整を行い、Ｇ２のかわりにペアとして非同期コピーを開始し、ペア状態になる。その際、Ｇ２が「差分リシンク待機」に変わる。指示の仕方の詳細は後述する。

【００８４】

なお、差分リシンク待機状態のペアの正ボリュームを提供するストレージ装置は、クラスタを構成するもう一方のストレージ装置から同期リモートコピーを介して受信した書き込みデータをジャーナルエントリ化して、ジャーナルボリュームに格納している。

【００８５】

図７は、ストレージクラスタリング制御プログラム３０００の動作を示したフローである。

【００８６】

ストレージクラスタリング制御プログラムは、クラスタリング適用中のボリュームへの読み込み又は／及び書き込みの成否を監視する（ステップ３００１）。

【００８７】

次に読み込み又は／及び書き込みが失敗してスワップが必要になったか、もしくは、ユーザからの指示があってスワップが必要になったか、いずれかの状態を検知する（ステップ３００２）。すなわち非計画切り替えと計画切り替えのいずれかによる処理の開始を判断する。なお、「スワップ」とは読み込み又は書き込み要求を出すストレージ装置を切り替えることを指す。

【００８８】

上記ステップ３００２での判定が偽であった場合、ステップ３００１に戻り読み込み又は／及び書き込み要求の監視を継続する。

【００８９】

上記ステップ３００２での判定が真であった場合、以下の切り替え処理を行う。

【００９０】

まず、対象となるボリュームへの読み込み又は／及び書き込み要求の処理を一時的に停止する（ステップ３００３）。

【００９１】

ストレージクラスタリング対象のボリュームを有するペア状態の変更を行う（ステップ３００４）。すなわち、図６の例で述べると、「Ａ０００１」のデータが「Ｂ０００１」に向けて同期的にコピーされていたものを、「停止」指示をストレージ装置へ送り、「Ｂ０００１」への書き込みができる状態に変更する。「Ａ０００２」と「Ｂ０００２」のペアも同様に処理する。

【００９２】

次に、入出力構成制御情報１００８の交換書き換えを行う（ステップ３００５）。通常、アプリケーションは、読み込み又は／及び書き込み要求の発行先として指定する情報とともに読み込み又は／及び書き込み要求がＯＳ１００７に渡る。ＯＳ１００７は渡された読み込み又は／及び書き込み要求を処理する時に、入出力構成制御情報１００８に含まれる読み込み又は／及び書き込み要求発行先の情報に対応するデバイス識別子を求め、そのデバイス識別子を対象にした読み込み又は／及び書き込み要求処理を行う。本ステップでは、クラスタリング対象のボリューム情報と、そのペアとされているボリュームの情報を交換する。「Ａ０００１」と「Ｂ０００１」の例で述べると、交換前に「Ａ０００１」を指し示していた情報は、交換後に「Ｂ０００１」を指し示すようにし、逆に交換前に「Ｂ０００１」を指し示していた情報は、交換後に「Ａ０００１」を指し示すようになる。

【００９３】

それから、上記ステップ３００３で停止させた読み込み又は／及び書き込み要求処理を再開する（ステップ３００６）。

【００９４】

最後に、交換処理を完了したことを示すメッセージを出力し（ステップ３００７）、ステップ３００１の読み込み又は／及び書き込み要求監視に戻る。

【００９５】

なお、ステップ３００５ではコピー方向の反転を伴った「リシンク」指示をストレージ装置に送信してもよい。仮に当該処理を行う要因が切り替え元ストレージ装置２０００とホスト計算機１０００との間のネットワーク障害の場合は、逆方向の同期リモートコピーが開始され、データの冗長性が確保される。

【００９６】

図８は、コピー制御プログラム４０００の、ストレージクラスタリング制御に対応するための動作を示したフローである。

【００９７】

コピー制御プログラム４０００は、ストレージクラスタリング制御プログラム３０００が行う交換の完了メッセージの出力の有無を監視する（ステップ４１０１）。

【００９８】

メッセージが検知されたかを判定する（ステップ４１０２）。

【００９９】

上記ステップ４１０２での判定が偽であった場合、ステップ４１０１に戻りメッセージの監視を継続する。

【０１００】

上記ステップ４１０２での判定が真であった場合、以下の処理を行う。

【０１０１】

まず、Ｇ１の状態を確認する指示を発行し、ストレージ装置２０００からＧ１の状態情報を取得し、ストレージクラスタリング制御プログラム３０００による交換処理後の状態であるか否かを確認する（ステップ４１０３）。

【０１０２】

次に、Ｇ２及びＧ３の状態を確認する指示を発行し、ストレージ装置２０００からＧ２及びＧ３の状態情報を取得し、障害でペア状態が停止しているなど差分リシンクができない状態あるか否かを確認する（ステップ４１０４）。

【０１０３】

上記ステップ４１０３、及び、ステップ４１０４の結果から、Ｇ１が交換処理後の状態であり、かつ、Ｇ２及びＧ３の状態に異常がない、副ボリューム間差分リシンクをすべき状態か否かを判定する（ステップ４１０５）。

【０１０４】

上記ステップ４１０５での判定が真であった場合、以下の処理を行う。

【０１０５】

差分リシンク指示を発行する（ステップ４１０６）。ここでＧ２がペア状態で、Ｇ３が差分リシンク待機状態であるのを、Ｇ３がペア状態で、Ｇ２が差分リシンク待機状態に切り替える指示である。この切り替えを、ボリュームの情報を全てコピーするのではなく、Ｇ２の副ボリューム（すなわちＧ３の副ボリューム）と、Ｇ３の主ボリュームの差分だけをコピーするように設定しなおすことで、全てコピーするよりも短期間にコピーできる副ボリューム間差分リシンクを行う。

【０１０６】

次に、ストレージ装置２０００から情報を取得し、ステップでの副ボリューム間差分リシンクが実行できたか否かを判定する（ステップ４１０７）。ストレージクラスタリング制御プログラムへの作りこみを抑止するため、本フローでの差分リシンクがストレージクラスタリングのボリューム交換と非同期的に実行される。そのため、ボリューム交換実行と差分リシンク実行の間に時間差が生じることがある。通常その差は非同期コピーのジャーナルボリュームの容量を大きくとることで解決されるが、ジャーナルボリュームの容量を少なく運用していた場合など、差分リシンクが失敗に終わることがある。この事象を本ステップで検知する。

【０１０７】

上記ステップ４１０７での判定が真であった場合、差分リシンクが成功したため本フローの処理を終了する。

【０１０８】

上記ステップ４１０７での判定が偽であった場合、差分リシンクできなかったため、以下の処理を行う。

【０１０９】

コピー環境情報５４００に設定されたオプション情報が、「続行」であるか「中断」であるかを判定する（ステップ４１０８）。

【０１１０】

上記ステップ４１０８での判定が「続行」であった場合、Ｇ３の主ボリュームのデータをＧ３の副ボリュームに全てコピーすることでペア状態を確立させる全コピーをすることで、Ｇ２を待機状態に、Ｇ３をペア状態に切り替える指示を行い、処理を終了する（ステップ４１０９）。

【０１１１】

上記ステップ４１０８での判定が「中断」であった場合、又は、上記ステップ４１０５

での判定が偽であった場合、切り替えができなかったことを示すメッセージを出力し、処理を終了する（ステップ４１１０）。

【０１１２】

図９は、コピー制御プログラム４０００の実際にコマンド発行処理を行う部分の動作を示したフローである。

【０１１３】

コピー制御プログラム４０００は、「作成」、「停止」、「リシンク」指示やペア状態取得指示など、コピー制御に関する指示を受け取る（ステップ４２０１）。

【０１１４】

指示で指定されているコピーグループに対応するコピーグループ情報５１００の第二属性５１０３を参照し、ストレージクラスタリングの対象となっているか否かを判定する（ステップ４２０２）。

【０１１５】

上記ステップ４２０２での判定が真であった場合、以下の処理を行う。

【０１１６】

指示で指定された処理内容が、停止やリシンクなどのペアの状態を変える制御系の指示であるか否かを判定する（ステップ４２０３）。

【０１１７】

上記ステップ４２０３での判定が真であった場合、その指示は許可されていないものとしてエラーとして処理し、終了する（ステップ４２０４）。

【０１１８】

上記ステップ４２０３での判定が偽であり指示で指定された処理内容がペアの状態情報取得などの参照系の指示であった場合、又は、上記ステップ４２０２での判定が偽であった場合、要求通りコマンド発行を行い、処理を終了する（ステップ４２０５）。

【０１１９】

図１０は、コピー制御プログラム４０００がユーザもしくはスクリプトプログラム等に提供するコマンドラインインタフェースの説明図である。

【０１２０】

コマンドラインインタフェースは、コマンド４３０１、対象４３０２、パラメタ４３０３からなる。

【０１２１】

コマンド４３０１は、ペア作成や状態取得などの指示を記述する。

【０１２２】

対象４３０２は、対象となるグループやペアを記述する。

【０１２３】

パラメタ４３０３は、指示に必要なパラメタがあればそれを記述する。

【０１２４】

コピーグループＧ１に対して状態を取得する場合は「状態取得，Ｇ１」と指示する。コピーグループＧ３を差分リシンク待機状態に設定する場合は、「ペア作成，Ｇ３，差分リシンク待機状態」と指示する。

【０１２５】

これらの指示やパラメタなどは、図１０の記載以外にもコピーの機能に従い様々な形態がありうる。

【０１２６】

以上、本実施形態での図８を用いたコピー制御プログラム４０００の、ストレージクラスタリング制御に対応するための動作において、副ボリューム間差分リシンク指示の前に、差分リシンクが可能な状態かを予め確認した上で差分リシンクを実施する形態を説明したが、これをストレージクラスタリングの動作を検知した後に、差分リシンクを行い、その結果成功しなかった場合に、改めて関連するコピーグループの状態を確認し、その結果ペアの状態異常などストレージによる差分リシンクが可能なエラーだった場合に、ペア状

態の確認などの回復処理を行い、再度、副ボリューム間差分リシンクを指示する形態としてもよい。この場合、Ｇ１などのコピーグループの状態確認を完了する時間を待たずに差分リシンクを行うことで、ディザスタリカバリ可能な状態により早く戻すことができる。

【０１２７】

また、本実施形態での図８を用いたコピー制御プログラム４０００の、ストレージクラスタリング制御に対応するための動作において、ステップ４１０４において、関連するコピーグループＧ２及びＧ３の状態の取得を行った。これを、状態取得の前に、副ボリューム間差分リシンクを実行できる状態に変更するためのＧ２に対するペアの停止指示を実施する形態としてもよい。ストレージクラスタリングによるボリュームの切り替えは、ユーザによるメンテナンスの準備として計画的に実施、もしくは、障害から非計画的に実施される。この場合、ストレージ障害の場合は、Ｇ２の状態が停止状態になるが、データネットワーク１００９は正常に動作しているがホスト計算機１０００とストレージ装置２０００の間のなんらかの回線障害の場合は、Ｇ２の状態はペア状態のままである。すなわちストレージクラスタリングの動作に伴って差分リシンクを実施すべき場合でもＧ２の状態が複数とりうる可能性がある。差分リシンクができるようにペア状態を確定させるステップを行うことで、計画的もしくは非計画的なボリュームの切り替えに対しても、コピー制御プログラム４０００は一つの手順で実施できる。それによりエラー時のユーザ操作の必要性を削減することができる。

【０１２８】

以上の説明により、

アプリケーションプログラムと、ストレージクラスタリング制御プログラムと、コピー制御プログラムと、を実行する第一のホスト計算機と、

前記第一のホスト計算機と接続し、第一のボリュームを提供する第一のストレージ装置と、

前記第一のホスト計算機及び前記第一のストレージ装置と接続し、第二のボリュームを提供する第二のストレージ装置と、

前記第一のストレージ装置及び前記第二のストレージ装置と接続し、第三のボリュームを提供する第三のストレージ装置と、

を有する計算機システムについて説明した。

【０１２９】

そして、

前記第一のストレージ装置は、前記第一のボリュームをコピー元とし、前記第二のボリュームをコピー先とする第一のコピーペアについての同期リモートコピー処理によって、前記アプリケーションプログラムの実行に従ってホスト計算機が送信した書き込みデータを前記第二のストレージ装置へ送信し、

前記第一のストレージ装置は、前記第一のボリュームをコピー元とし、前記第三のボリュームをコピー先とする第二のコピーペアについての非同期リモートコピー処理によって、前記アプリケーションプログラムの実行に従ってホスト計算機が送信した書き込みデータを前記第三のストレージ装置へ送信し、

前記ストレージクラスタリング制御プログラムを実行することで、前記第一のホスト計算機は、前記書き込みデータの書き込みが異常終了したことを検知し、前記第一のコピーペアのペア状態を変更することで検知後の前記書き込みデータを前記第二のボリュームに対して書き込み可能とする第一の指示を前記第二のストレージ装置に送信し、

前記コピー制御プログラムを実行することで、前記第一のホスト計算機は、前記ストレージクラスタリング制御プログラムによって異常終了を検知したことを、前記第一の指示の送信より後に検知し、前記第二のボリュームをコピー元とし、前記第三のボリュームをコピー先とする第三のコピーペアについての別な非同期リモートコピー処理によるコピー処理を開始するための差分リシンク指示を前記第二のストレージ装置に送信することを説明した。

【０１３０】

また、前記計算機システムは、前記第三のストレージ装置と接続し、前記アプリケーションプログラムを実行することで前記第三のボリュームに格納したデータを読み込む第二のホスト計算機を有してもよいことを説明した。

【０１３１】

また、

前記第一のホスト計算機は、前記第一のコピーペアが前記ストレージクラスタリング制御プログラムの制御対象であることを示すコピー定義情報を有し、

前記コピー制御プログラムを実行することで、前記第一のホスト計算機は、前記第一のコピーペア及び前記第二のコピーペアを対象とするペア状態指示を受付け、前記第一のコピーペアのペア状態又は前記第二のコピーペアのペア状態を表示し、

前記コピー制御プログラムを実行することで、前記第一のホスト計算機は、前記コピー定義情報に基いて、前記第二のコピーペア及び前記第三のコピーペアを対象とする状態変更指示は処理し、前記第一のコピーペアを対象とする状態変更指示の処理は抑止してもよいことも説明した。

【０１３２】

また、前記コピー制御プログラムを実行することで、前記第一のホスト計算機は、前記差分リシンク指示送信に基く第三のコピーペアを対象とする副ボリューム間差分リシンクが差分コピーで処理できないことを示す情報を前記第二のストレージ装置から取得し、前記ストレージクラスタリング制御プログラムによる前記第一のコピーペアの状態変化を要因とした前記第三のコピーペアの副ボリューム間差分リシンクが失敗したメッセージを表示しても良いことを説明した。

【０１３３】

また、前記コピー制御プログラムを実行することで、前記第一のホスト計算機は、前記差分リシンク指示送信に基く第三のコピーペアを対象とする副ボリューム間差分リシンクが差分コピーで処理できないことを示す情報を前記第二のストレージ装置から取得し、前記第二のボリュームからの全コピーによって前記第三のコピーペアのコピーを再開する第二の指示を送信してもよいことを説明した。

【０１３４】

このようなクラスタリング制御と非同期リモートコピー制御との連携を取ることで、仮に第一のストレージ装置が機能停止した場合でも非同期ＲＣを継続できることから計算機システムの信頼性が向上する。また、コピー制御プログラムは、障害検知のための高度な処理を含むストレージクラスタリング制御プログラムとは別なプログラムであるため、処理のモジュール性が向上し、安定した動作が可能となる他、計算機システムが非同期リモートコピーを実施しない情況から実施する情況にシステム構成を変更する場合（または逆の場合）にも柔軟な設定変更が可能となる。

【０１３５】

また、非同期リモートコピーの場合、対象となるストレージ装置の通信距離が遠く、そして一時的な通信不良が発生する可能性があり、また同期リモートコピーと比較して複雑な処理を必要とするため、ペア操作指示に時間を必要とすることがある。そのため、ストレージクラスタ制御プログラムからコピー制御プログラムへのストレージ装置切り替えの通知を、切り替え指示よりも後に行う点は、アプリケーションプログラムからの読み込み又は／及び書き込み要求の停止時間短縮、又はストレージクラスタリング制御プログラムの処理をプログラム設計者が想定した以上に停止させないことによる安定動作につながる。

【０１３６】

しかし、本実施例はこれら以外の内容についても開示している。

【符号の説明】

【０１３７】

１０００…ホスト計算機　１１００…管理計算機　２０００…ストレージ装置　３０００…ストレージクラスタリング制御プログラム　４０００…コピー制御プログラム

【図１】

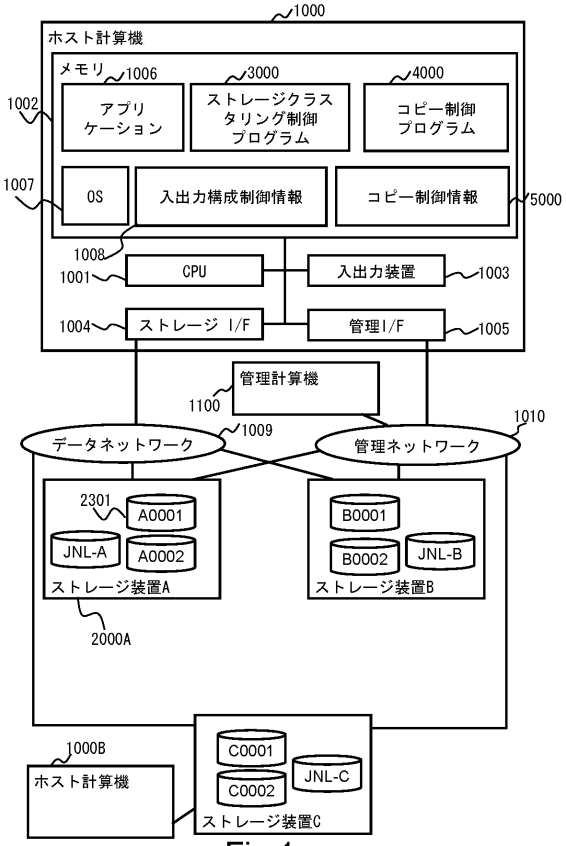

Fig.1

【図２】

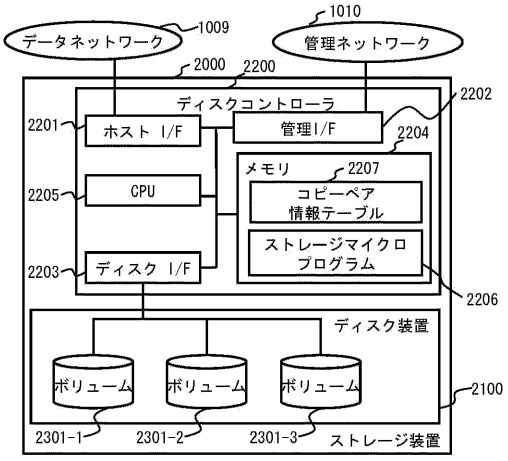

Fig.2

【図３】

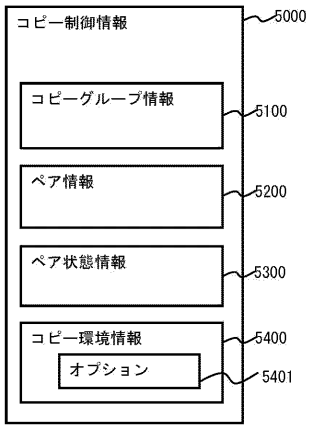

Fig.3

【図４】

5100

5101 グループ識別子	5102 第一属性	5103 第二属性
G1	同期コピー	クラスタリング機能適用中
G2	非同期コピー	ー
G3	非同期コピー	ー

Fig.4

【図５】

5200

グループ識別子	G1	5201
プライマリデバイス識別子	セカンダリデバイス識別子	5203
A0001	B0001	
A0002	B0002	

グループ識別子	G2
プライマリデバイス識別子	セカンダリデバイス識別子
A0001	C0001
A0002	C0002

グループ識別子	G3
プライマリデバイス識別子	セカンダリデバイス識別子
B0001	C0001
B0002	C0002

5202

Fig.5

【図６】

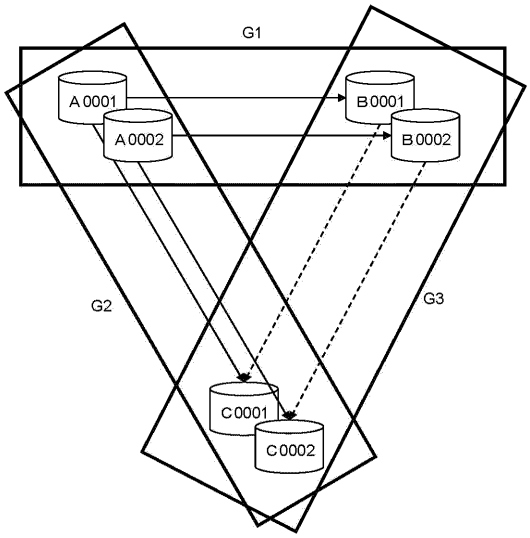

Fig.6

【図７】

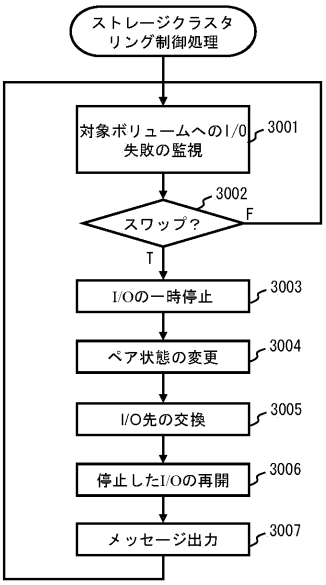

Fig.7

【図８】

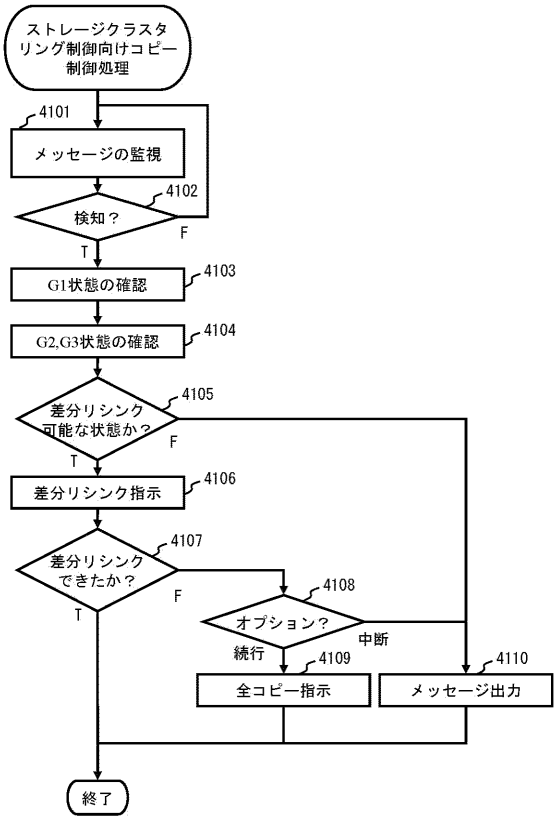

Fig.8

【図９】

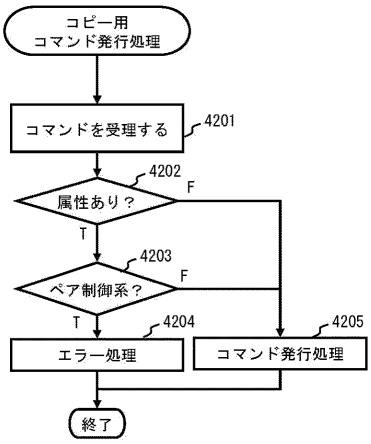

Fig.9

【図１０】

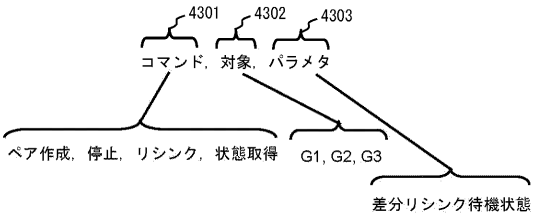

Fig.10

フロントページの続き

(72)発明者　小山田　健一
神奈川県横浜市戸塚区戸塚町５０３０番地　株式会社日立製作所　ソフトウェア事業部内
(72)発明者　松井　義典
神奈川県横浜市戸塚区戸塚町５０３０番地　株式会社日立製作所　ソフトウェア事業部内

審査官　坂東　博司

(56)参考文献　特開２００７－０６６１６２（ＪＰ，Ａ）
特開２００８－０６５４２５（ＪＰ，Ａ）
特開２００７－２４９４４７（ＪＰ，Ａ）
特開２００７－２７２５１０（ＪＰ，Ａ）
特開２００６－０５３９１２（ＪＰ，Ａ）
特開２００４－３１９０１３（ＪＰ，Ａ）
特開２００４－２２７５２８（ＪＰ，Ａ）
特表２００６－５２７８７５（ＪＰ，Ａ）
特開２００８－１３４９８６（ＪＰ，Ａ）
米国特許出願公開第２００４／０２６０８９９（ＵＳ，Ａ１）
米国特許出願公開第２００８／０１０４４４３（ＵＳ，Ａ１）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　　３／０６