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

特許第4473612号
(P4473612)

(45)発行日　平成22年6月2日(2010.6.2)　(24)登録日　平成22年3月12日(2010.3.12)

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

請求項の数 16　外国語出願　(全 20 頁)

(21)出願番号　特願2004-81902(P2004-81902)
(22)出願日　平成16年3月22日(2004.3.22)
(65)公開番号　特開2005-18736(P2005-18736A)
(43)公開日　平成17年1月20日(2005.1.20)
審査請求日　平成19年2月26日(2007.2.26)
(31)優先権主張番号　10/602223
(32)優先日　平成15年6月23日(2003.6.23)
(33)優先権主張国　米国(US)

(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　山神　憲司
アメリカ合衆国カリフォルニア州ロスガトス　カイルニベル　１０８

審査官　桜井　茂行

最終頁に続く

(54)【発明の名称】リモートコピーシステム

(57)【特許請求の範囲】
【請求項１】
複数の第１のデータボリュームと、関連する第１のホストから書き込みリクエストを受け取って、前記書き込みリクエストに関連する書き込みデータを前記複数の第１のデータボリュームに記憶するように構成されている第１のストレージコントローラと、を有する第１のストレージシステムと、
第２のデータボリュームと、前記第１のデータボリュームを前記第２のデータボリュームにミラーリングするための前記書き込みデータを含む第１のデータを受信する第２のストレージコントローラと、ジャーナルボリュームと、を有する第２のストレージシステムと、
複数の第３のデータボリュームと、前記複数の第１のデータボリュームを前記複数の第３のデータボリュームに各々ミラーリングするジャーナルを含む第２のデータを受信する第３のストレージコントローラと、を有する第３のストレージシステムと、
を備え、
前記ジャーナルボリュームは、前記複数の第１のデータボリューム及び複数の第３のデータボリュームと対応しており、
前記ジャーナルは、前記書き込みデータと、前記第１のデータボリュームに書き込み順序を提供するシーケンス番号を有し、
前記第１のホストから前記第１のストレージシステムへの書き込みリクエストは、前記第１のデータが前記第２のストレージシステムによって受信された後、完了し、

前記第１のホストから前記第１のストレージシステムへの書き込みリクエストは、前記第３のストレージシステムによって受信された第２のデータとは独立して完了し、
前記第３のデータボリュームに記憶される書き込みデータは、前記ジャーナルのシーケンス番号によって提供される書き込み順序にしたがって生成され、
前記ジャーナルは、前記ジャーナルボリュームに先入れ先出しで積み重ねられ、
当該ジャーナルは、更新の順序情報及び前記第１のデータボリュームの識別情報を含み、当該更新の順序情報及び前記第１のデータボリュームの識別情報に基づいて前記第３のデータボリュームに非同期で送信されることを特徴とするリモートコピーシステム。

【請求項２】
前記第２のストレージシステムは、前記第１のストレージシステムの比較的近くに位置し、前記第３のストレージシステムは、前記第１のストレージシステムの比較的遠くに位置していることを特徴とする請求項１に記載のリモートコピーシステム。

【請求項３】
前記第２のストレージシステムは、前記第１のストレージシステムから１００マイル以内に位置し、前記第３のストレージシステムは、前記第１のストレージシステムから１００マイルを超えて位置していることを特徴とする請求項２に記載のリモートコピーシステム。

【請求項４】
前記第２のストレージシステムに結合される第２のホストをさらに備え、
前記第２のストレージシステムは、もし前記第１のストレージシステムに障害が発生した場合には、プライマリストレージシステムとして機能するように構成されていることを特徴とする請求項１に記載のリモートコピーシステム。

【請求項５】
前記第３のストレージシステムに結合される第３のホストをさらに備え、
前記第３のストレージシステムは、もし前記第１のストレージシステムに障害が発生した場合には、前記第１のストレージシステムに代ってプライマリストレージシステムとして機能するように構成されていることを特徴とする請求項１に記載のリモートコピーシステム。

【請求項６】
前記ジャーナルは、前記第１のホストからの書き込みリクエストに基づいて前記書き込みデータが前記第１のデータボリュームに記憶された時刻に関する情報をさらに含むことを特徴とする請求項１に記載のリモートコピーシステム。

【請求項７】
前記第１のストレージシステムに障害が発生した場合、前記ジャーナルは、前記書き込みデータを確保するため、前記書き込みデータが前記第２のストレージシステムにコピーされた後、前記第３のデータボリュームに受信されることを特徴とする請求項１に記載のリモートコピーシステム。

【請求項８】
前記ジャーナルボリュームが、コントロールデータ領域と、ジャーナルデータ領域と、を含み、当該コントロールデータ領域は、コントロールデータのみを記憶する構成であり、当該ジャーナルデータ領域は、ジャーナルデータのみを記憶する構成であることを特徴とする請求項１に記載のリモートコピーシステム。

【請求項９】
複数の第１のデータボリュームと、関連する第１のホストから書き込みリクエストを受け取って、前記書き込みリクエストに関連する書き込みデータを前記複数の第１のデータボリュームに記憶するように構成されている第１のストレージコントローラと、を有する第１のストレージシステムと、
前記第１のストレージシステムに結合され、且つ、第２のデータボリュームと、前記第１のデータボリュームに記憶されたデータを前記第２のデータボリュームにコピーするための書き込みデータを含む第１のデータを同期して受信するよう構成された第２のストレ

ージコントローラと、ジャーナルボリュームと、を有する第２のストレージシステムと、
前記第２のストレージシステムに結合され、且つ、複数の第３のデータボリュームと、前記ジャーナルボリュームからの第２のデータを非同期で受信するよう構成された第３のストレージコントローラと、を有する第３のストレージシステムと、
を備え、
前記ジャーナルボリュームは、前記複数の第１のデータボリューム及び複数の第３のデータボリュームと対応しており、
前記第の２データは、前記第１のデータボリュームに記憶されたデータを、前記第２のストレージシステムを介して、前記第３のデータボリュームにコピーするジャーナルを有し、
前記ジャーナルは、前記書き込みデータと、前記第１のデータボリュームに書き込み順序を提供するシーケンス番号を有し、
前記第３のデータボリュームに記憶される書き込みデータは、前記シーケンス番号によって提供される書き込み順序にしたがって生成され、
前記ジャーナルは、前記ジャーナルボリュームに先入れ先出しで積み重ねられ、
当該ジャーナルは、更新の順序情報及び前記第１のデータボリュームの識別情報を含み、当該更新の順序情報及び前記第１のデータボリュームの識別情報に基づいて前記第３のデータボリュームに非同期で送信されることを特徴とするリモートコピーシステム。
【請求項１０】
前記第２のストレージシステムは、前記第１のストレージシステムの比較的近くに位置し、前記第３のストレージシステムは、前記第１のストレージシステムの比較的遠くに位置していることを特徴とする請求項９に記載のリモートコピーシステム。
【請求項１１】
前記第２のストレージシステムは、前記第１のストレージシステムから１００マイル以内に位置し、前記第３のストレージシステムは、前記第１のストレージシステムから１００マイルを超えて位置していることを特徴とする請求項１０に記載のリモートコピーシステム。
【請求項１２】
前記第２のストレージシステムに結合される第２のホストをさらに備え、
前記第２のストレージシステムは、もし前記第１のストレージシステムに障害が発生した場合には、プライマリストレージシステムとして機能するように構成されていることを特徴とする請求項９に記載のリモートコピーシステム。
【請求項１３】
前記第３のストレージシステムに結合される第３のホストをさらに備え、
前記第３のストレージシステムは、もし前記第１のストレージシステムに障害が発生した場合には、前記第１のストレージシステムに代ってプライマリストレージシステムとして機能するように構成されていることを特徴とする請求項９に記載のリモートコピーシステム。
【請求項１４】
前記ジャーナルは、前記第１のホストからの書き込みリクエストに基づいて前記書き込みデータが前記第１のデータボリュームに記憶された時刻に関する情報をさらに含むことを特徴とする請求項９に記載のリモートコピーシステム。
【請求項１５】
前記第１のストレージシステムに障害が発生した場合、前記ジャーナルは、前記書き込みデータを確保するため、前記書き込みデータが前記第２のストレージシステムにコピーされた後、前記第３のデータボリュームに受信されることを特徴とする請求項９に記載のリモートコピーシステム。
【請求項１６】
前記ジャーナルボリュームが、コントロールデータ領域と、ジャーナルデータ領域と、を含み、当該コントロールデータ領域は、コントロールデータのみを記憶する構成であり

、当該ジャーナルデータ領域は、ジャーナルデータのみを記憶する構成であることを特徴とする請求項９に記載のリモートコピーシステム。

【発明の詳細な説明】
【技術分野】
【０００１】
本発明はストレージシステムに係わり、さらに詳細にはリモートコピー機能を行うように構成されたストレージシステムに係わる。
【背景技術】
【０００２】
データは全ての計算プロセスの基本となる基礎リソースである。近年のインターネットやｅ－ビジネスの爆発的な発展に伴い、データ記憶システムの需要は急速に増加している。一般に、ストレージネットワークは二つののアプリケーションや構成を包含するもので、ネットワーク接続ストレージ（ＮＡＳ）またはストレージ・エリア・ネットワーク（ＳＡＮ）などである。
【０００３】
ＮＡＳはストレージサーバとそのクライアントの間でファイルフォーマットのデータを転送するためにイーサーネット上でＩＰを用いる。ＮＡＳにおいては、ディスクアレイやテープアレイなどの総合ストレージシステムは、ＴＣＰ／ＩＰなどのメッセージコミュニケーションプロトコルを用いて、ローカルエリアネットワーク（ＬＡＮ）を通してメッセージネットワークに直接接続される。クライアントサーバシステムでは、ストレージシステムはサーバとして機能する。
【０００４】
一般に、ＳＡＮは異種のサーバとストレージリソース間でデータを移動させる専用高性能ネットワークである。従来のメッセージング・ネットワークでは、ＮＡＳと異なり、クライアントとサーバ間のトラフィックの輻輳を避けるためにそれぞれに専用のネットワークが用いられる。ＳＡＮでは、ストレージリソースとプロセッサーあるいはサーバ間に直接の接続を設定することが可能である。ＳＡＮはサーバの間でシェアすることも、また特定のサーバに専用にすることも可能である。それは一カ所に集中させることも、地理的に離れた場所に拡張することも可能である。ＳＡＮインターフェイスは、ファイバーチャネル（ＦＣ）やエンタプライズシステムコネクション（ＥＳＣＯＮ）や小型コンピュータシステムインターフェイス（ＳＣＳＩ）やシリアルストレージアーキテクチャ（ＳＳＡ）や高性能パラレルインターフェイス（ＨＩＰＰＩ）あるいは将来出現するであろう他のプロトコルのような各種のインターフェイスによることが可能である。
【０００５】
たとえば、インターネットエンジニアリングタスクフォース（ＩＥＴＦ）は、ＴＣＰ／ＩＰ上でブロックストレージを可能にする新しいプロトコルあるいは標準ｉＳＣＳＩを開発中であり、一方幾つかの会社はｉＳＣＳＩをＳＡＮの主たる標準にするためにホストプロセッサーからｉＳＣＳＩ－ＴＣＰ／ＩＰプロトコルスタックを外す努力をしている。用いられているストレージシステムの形式がどうであろうとも、データストレージシステムのユーザは、ストレージユニット（あるいはストレージサブシステム）の障害で貴重なデータが失われることを防ぐためのデータバックアップに強い関心を持っている。
【０００６】
したがって、データストレージシステムは一般に、プライマリユニットに障害が発生した時に、エマージェンシーリカバリーのためのデータを記憶するバックアップユニットを含んでいる。しかしながら、障害は、ユニット自身の故障による場合も、たとえばユニットが設置されている場所での地震とか台風による自然災害の発生による場合もある。もしバックアップユニットがプライマリユニットの近くに設置されていると、自然災害が発生した場合に両方とも破壊される可能性がある。したがって、多くのストレージシステムのユーザは、たとえば１００マイル以上離れた遠い距離を離してプライマリユニットとバッ

クアップユニットを設置する傾向がある。実際に、複数のユーザはプライマリユニットとバックアップユニットを別の大陸に設置することまでやっている。

【発明の開示】

【発明が解決しようとする課題】

【０００７】

現在、ストレージシステムでは、データをバックアップあるいはセカンダリサイトにコピーするために二つの動作モード、すなわち同期モードと非同期モードが用いられている。

【０００８】

同期モードでは、ホストからプライマリストレージシステムへの書き込みリクエストは、書き込みデータがセカンダリストレージシステムにコピーされ、それが確認された後でのみ完了する。したがって、セカンダリストレージシステムから確認を受け取るまで、ホストからの書き込みデータはプライマリシステムのキャッシュに保存されているので、このモードではセカンダリシステムでデータの消失がないことが保証される。

【０００９】

さらに、プライマリストレージシステムにおけるプライマリボリューム（ＰＶＯＬ）とセカンダリストレージシステムにおけるセカンダリボリューム（ＳＶＯＬ）は同一に保持されているので、もしＰＶＯＬに障害が発生しても、速やかにＳＶＯＬを用いてＰＶＯＬを再生することができる。しかしながら、このモードの下では、プライマリストレージシステムとセカンダリストレージシステムを、たとえば１００マイル以上離して設置することは出来ない。そのようなことがなければ、ストレージシステムは効率よくホストからの書き込みリクエストを実行できる。

【００１０】

非同期モードでは、ホストからプライマリストレージシステムへの書き込みリクエストは、プライマリシステムのみへ書き込みデータを記憶することで完了する。書き込みデータは、続いてセカンダリストレージシステムにコピーされる。すなわち、プライマリストレージシステムへデータを書き込むことは、セカンダリストレージシステムにデータをコピーすることとは独立した処理である。したがって、プライマリシステムとセカンダリシステムは、たとえば１００マイルあるいはそれ以上ずっと離して別々に設置されてもよい。

【００１１】

しかしながら、ＰＶＯＬとＳＶＯＬは同一に維持されているので、もしプライマリシステムがダウンするとデータが失われる可能性がある。したがって、同期モードと非同期モードの利点を提供する、言い換えればデータ喪失がないことを保証していながらプライマリシステムとセカンダリシステムをずっと離して設置できる、データストレージシステムあるいはリモートコピーシステムを提供することが望まれている。

【課題を解決するための手段】

【００１２】

本発明の実施例は、リモートコピー機能を行うべく構成されたストレージシステムに係わる。一実施例においては、プライマリストレージシステムは、仲介ストレージシステムに書き込みデータをコントロールデータとともに同期して送る。仲介ストレージシステムは書き込みデータとコントロールデータを、たとえばジャーナルボリューム（ＪＮＬ）などのボリュームに記憶する。仲介ストレージシステムはデータの到着の順序を維持し、コントロールデータの各々に順序情報を割り当てあるいは連結させる。

【００１３】

つづいて、仲介ストレージシステムは書き込みデータとその対応するコントロールデータを非同期で、あるいはプライマリストレージシステムにおける書き込みデータとコントロールデータの記憶とは関連無しにセカンダリストレージシステムに送る。セカンダリストレージシステムは、コントロールデータとコントロールデータに関連した順序情報に従い、書き込みデータをセカンダリボリューム（ＳＶＯＬ）に記憶する。ここに用いられて

いる、「ストレージシステム」という用語は、データを記憶するように構成され、たとえばディスクアレイユニットなどの１台ないしはそれ以上のストレージユニットあるいはストレージサブシステムを持つコンピュータシステムを意味する。

【００１４】

したがって、ストレージシステムは、１台ないしはそれ以上のホストと、１ないしはそれ以上のストレージサブシステムを含むコンピュータシステム、あるいは単にストレージサブシステムまたはユニット、あるいはコミュニケーションリンクを介して相互に結合した複数のストレージサブシステムまたはユニットを意味することもある。ここに用いられている、「ストレージサブシステム」という用語は、データを記憶するように構成され、ストレージ領域と１台ないしはそれ以上のホストからのリクエストを処理するストレージコントローラを含むコンピュータシステムを意味する。ストレージサブシステムの例はディスクアレイユニットである。

【００１５】

ここに用いられている、「ホスト」という用語は、１台ないしはそれ以上のストレージシステムあるいはストレージサブシステムに結合され、ストレージシステムあるいはストレージサブシステムにリクエストを送るように構成されたコンピュータシステムを意味する。ホストはサーバかあるいはクライアントであるかもしれない。ここに用いられている、「リモートコピーシステム」という用語は、リモートコピー機能を行うように構成されたコンピュータシステムを意味する。リモートコピーシステムは単一のストレージシステム、ストレージサブシステムまたはストレージユニット、あるいはネットワークまたはコミュニケーションリンクで結合された複数のストレージユニット、ストレージシステム、ストレージサブシステムを意味する。

【００１６】

したがって、リモートコピーシステムは、プライマリストレージシステム、セカンダリストレージシステム、仲介システムあるいはそれらの組み合わせを意味する。リモートコピーシステムはまた、１台ないしはそれ以上のホストを含むこともある。一実施例においては、プライマリストレージシステム１１０ａのボリュームは、仲介ストレージシステム１１０ｃを用いてセカンダリストレージシステム１１０ｂにミラーリングされている。仲介ストレージシステム１１０ｃは一般にプライマリストレージシステム１１０ａの比較的近くに位置し、一方セカンダリストレージシステム１１０ｂは、仲介ストレージシステム１１０ｃと／あるいはプライマリストレージシステムから比較的遠方に位置している。

【００１７】

プライマリストレージシステムと関連しているホストから書き込みリクエストを受け取ると、書き込みデータはプライマリストレージシステム１１０ａから仲介ストレージシステム１１０ｃに同期的にコピーされる。仲介ストレージシステムにコピーされた書き込みデータは、コントロールデータとジャーナルデータを含むジャーナル情報の形式である。書き込みデータはジャーナルデータに対応し、コントロールデータはジャーナルデータについての管理情報を提供する。

【００１８】

次に、ホストの書き込みリクエストに関連して、ジャーナルは仲介ストレージシステム１１０ｃからセカンダリストレージシステム１１０ｂに非同期的にコピーされる。書き込みデータは、上記のコピー処理の間にプライマリストレージシステムと仲介ストレージシステムが損傷を受けない限り信頼できる。一実施例においては、仲介ストレージシステムは１ないしはそれ以上のジャーナルボリュームを含むがデータボリュームは含まないので、装置コストは低くなる。一実施例では、リモートコピーシステムは第１のストレージコントローラと第１のデータボリュームを含む第１のストレージシステムを含む。第１のストレージコントローラは第１のデータボリュームに対するデータアクセスリクエストを制御するように構成されている。

【００１９】

第１のストレージシステムは第１のストレージシステムに関連する第１のホストからの

書き込みリクエストを受け取ると第１のデータボリュームに書き込みデータを記憶し、コントロールデータとジャーナルデータを含むジャーナルを生成するように構成されている。第２のストレージシステムはジャーナルボリュームを含み、第１のストレージシステムにより生成されたジャーナルをジャーナルボリュームに受信し記憶するように構成されている。第３のストレージシステムは第２のデータボリュームを含み、第２のストレージシステムからジャーナルを受け取り、コントロールデータにより提供される情報に従ってジャーナルのジャーナルデータを第２のストレージシステムに記憶する。

【００２０】

一実施例においては、ストレージシステムは、第１のホストからデータアクセスリクエストを受け取る第１のストレージコントローラと、第１のストレージコントローラと結合し第１のストレージコントローラの制御により情報を記憶するように構成され、プライマリボリュームを含んでいる第１のストレージ領域と、第１のストレージコントローラから少なくとも１００マイル離れて設置されている第２のストレージコントローラと、第２のストレージコントローラと結合し第２のストレージコントローラの制御に従い情報を記憶するように構成され、セカンダリボリュームを含んでいる第２のストレージ領域と、を含んでいる。セカンダリボリュームはプライマリボリュームをミラーリングする。

【００２１】

第１のストレージコントローラはプライマリボリュームに第１のホストからの書き込みリクエストに関連した書き込みデータを記憶し、書き込みリクエストに応じてコントロールデータとジャーナルデータを含むジャーナルを生成するように構成されている。ジャーナルデータは書き込みデータに対応しており、ジャーナルは第１のストレージ領域の外に設けられたジャーナルボリュームに同期的に転送される。

【００２２】

別の実施例においては、リモートコピーシステム動作は、プライマリストレージシステムに関連したプライマリホストから書き込みリクエストを受け取った後にプライマリストレージシステムでコントロールデータとジャーナルデータを生成し、プライマリストレージシステムとセカンダリストレージシステムの間でデータのミラーリングを行うために仲介ストレージシステムにジャーナルを転送することを含んでいる。セカンダリストレージシステムは仲介ストレージシステムから遠く離れて配置されている。

【発明の効果】

【００２３】

一実施例においては、仲介ストレージシステムは１ないしはそれ以上のジャーナルボリュームを含むがデータボリュームは含まないので、装置コストは低くなる。

【発明を実施するための最良の形態】

【００２４】

図１Ａは、本発明の一実施例によるリモートコピーシステム５０は複数のストレージシステムを含むことを示している。リモートコピーシステムはプライマリストレージシステム１１０ａとセカンダリストレージシステム１１０ｂと仲介ストレージシステム１１０ｃを含んでいる。ストレージシステムは相互にコミュニケーションリンク１２０ａと１２０ｂを介し結合されている。一実施例においては、以下に説明するようにリンク１２０ａはリンク１２０ｂよりかなり短いので、リンク１２０ａはファイバー・チャネルであり、リンク１２０ｂは公衆コミュニケーションリンクである。本実施例では、ストレージシステム１１０ａと１１０ｂと１１０ｃはディスクアレイユニットか、あるいはストレージサブシステムである。

【００２５】

図１Bはデータの読み出し／書き込みリクエストを処理するように構成されたストレージコントローラ６２と、書き込みリクエストに応じてデータを記憶する記録媒体を含むストレージユニット６３を含む典型的なストレージサブシステム６０（たとえばプライマリシステム１１０ａ）の例を示す。コントローラ６２はホストコンピュータ（たとえばホスト１００ａ）と結合したホストチャネルアダプター６４と他のサブシステム（たとえばス

トレージシステム１１０ｃあるいは１１０ｂ）と結合したサブシステムチャネルアダプター６６とストレージサブシステム６０におけるストレージユニット６３に結合したディスクアダプター６８を含んでいる。

【００２６】

本発明の実施例では、これらのアダプターの各々はデータを送受信するポート（図示せず）とポートを経由したデータ転送を制御するマイクロプロセッサー（図示せず）を含んでいる。コントローラ６２はストレージユニット６３から読み出したあるいは書き込むデータを一時的に記憶するために用いるキャッシュメモリー７０も含んでいる。一実施例では、ストレージユニットは複数の磁気ディスクドライブ（図示せず）である。サブシステムはホストコンピュータに複数の論理ボリュームをストレージ領域として提供する。ホストコンピュータはこれらの論理ボリュームの識別子をストレージサブシステムからデータを読み出すあるいはそこにデータを書き込むために用いる。論理ボリュームの識別子は論理ユニット番号（ＬＵＮ）と呼ばれる。論理ボリュームは単一の物理ストレージデバイスあるいは複数のストレージデバイスに含むことができる。同様に、複数の論理ボリュームを単一の物理ストレージデバイスに関連させることが出来る。

【００２７】

より詳細なストレージサブシステムの説明が、参考として援用されている、現在の特許権者が権利を有する、２００２年６月５日に出願された日本国特許出願２００２-１６３７０５により優先権が主張されており、２００３年５月２１日に出願された「データストレージサブシステム」という名称の米国特許出願＿＿＿＿＿＿により提供されている。

【００２８】

図１Ａに戻ると、本発明の実施例においては、書き込みデータはプライマリシステム１１０ａと仲介システム１１０ｃとの間では同期して、仲介システム１１０ｃとセカンダリシステム１１０ｂとの間では非同期で送られる。したがって、仲介システム１１０ｃは一般にセカンダリシステム１１０ｂからの距離に比較してプライマリシステム１１０ａの比較的近くに位置している。たとえば、仲介システムはプライマリシステムから約５マイル以下、あるいは１０マイル以下もしくは２０マイル以下に位置している。一実施例においては、仲介システムはプライマリシステムから１００マイルを超えて離れることは無い。比較してみると、本発明の一実施例においては、仲介システムはセカンダリシステムから５０マイル以上あるいは１００マイル以上あるいは２００マイル以上あるいは異なった大陸に位置している。

【００２９】

リモートコピーシステムはプライマリストレージシステム１１０ａにコミュニケーションリンク１３０ａを経由して結合しているプライマリホスト１００ａとコミュニケーションリンク１３０ｂを経由してセカンダリストレージシステム１１０ｂに結合しているセカンダリホスト１００ｂを含んでいる。プライマリホストはプライマリストレージシステム１１０ａのストレージ領域あるいはボリュームにアクセス（読み出し及び書き込み）するためのアプリケーションプログラム（ＡＰＰ）１０２を含んでいる。一実施例においては、ＡＰＰ１０２はもしプライマリホスト１００ａあるいは／およびプライマリストレージシステム１１０ａが使えなくなった時には（言い換えれば障害が発生した時には）、特に企業などのビジネスユーザに支障なくデータにアクセス出来るようにセカンダリホスト１００ｂにフェイルオーバを動作させることが出来る。

【００３０】

リモートコピーマネージメントソフトウエア（ＲＣＭ）１０１は、リモートコピーシステムを制御するためのインターフェイスをユーザおよび／あるいはアプリケーションに提供するためにホスト１００ａと１００ｂの双方のうえで走る。システム５０は三つの形式のボリュームを含んでいる。プライマリストレージシステムは１ないしはそれ以上のプライマリボリューム（ＰＶＯＬ）１１１ａを含んでいる。ＰＶＯＬ１１１ａはＡＰＰ１０１が読み出しあるいは書き込むプロダクションデータを含んでいる。セカンダリストレージシステムは１ないしはそれ以上のセカンダリボリューム（ＳＶＯＬ）１１１ｂを含んでい

る。ＳＶＯＬ１１１ｂはＰＶＯＬ１１１aのコピーデータを含んでいる。仲介ストレージシステムは１ないしはそれ以上のジャーナルボリューム（ＪＶＯＬあるいはＪＮＬ）１１２を含んでいる。ＪＶＯＬ１１２はＰＶＯＬ１１１aに書き込まれた書き込みデータとそれに対応するコントロールデータを記憶する。

【００３１】

図２は本発明の一実施例によるＪＮＬボリュームまたはＪＶＯＬ１１２の例を示す。ＪＶＯＬはリモートコピーを実行するに際し用いられるジャーナルあるいはジャーナル情報を記憶するように構成されている。ジャーナルはジャーナルデータとそのコントロールデータを対として持っている。ジャーナルデータはＰＶＯＬ１１１aに記憶されたデータあるいはホスト１００からＰＶＯＬに書き込まれたデータに対応している。コントロールデータは対応するジャーナルデータに関係した管理情報を含む。

【００３２】

本発明の実施例においては、コントロールデータとジャーナルデータは両方ともシーケンシャルに、言い換えれば受け取ったジャーナルはＦＩＦＯメモリーに順番に積み重ねられるように、同じジャーナルボリュームに記憶される。仲介ストレージシステムは複数のこのようなＦＩＦＯメモリーを含むことがある。一実施例では、最初のコントロールデータはボリューム１１２において最初に記憶され、それに対応する最初のジャーナルデータは同じボリュームに記憶される。

【００３３】

次に第２のコントロールデータは最初のジャーナルデータの隣に記憶され、第２のコントロールデータに対応する第２のジャーナルデータはその隣に記憶され、以下同様になる。コントロールデータは、ジャーナルデータが得られるＰＶＯＬ１１１aに対する識別子であるインデックス（ＩＤＸ）２１１、たとえばプライマリストレージシステム１１０ａあるいはジャーナルグループ２００（図４）におけるＰＶＯＬに割り当てられた固有の番号を、含んでいる。アドレス２１２は書き込みデータがそこから書かれるＰＶＯＬのオフセットアドレス、たとえば書き込みデータのスターティング論理ブロックアドレス（ＬＢＡ）、を提供する。長さ２１３は書き込みデータの長さ、たとえば論理ブロックの数あるいは書き込みデータの総バイト数、を提供する。

【００３４】

時刻２１４はホストがＰＶＯＬ１１１ａにデータを書き込んだ時を示す。シーケンス番号（ＳＥＱ＃）２１５は書き込みのシーケンス情報を提供する。すなわち、シーケンス番号はプライマリストレージシステム１１０ａ内の書き込み順序を提供する。ＪＶＯＬアイデンティフィケーション（ＪＶＯＬ＿ＩＤ）２１６は対応するジャーナルデータ、たとえばプライマリストレージシステムあるいはジャーナルグループ２００におけるジャーナルボリュームに割り当てられた固有の番号、を含むジャーナルボリュームを識別する。ジャーナルオフセット（ＪＦＳ）２１７は、ジャーナルデータの記憶を開始する、あるいはジャーナルデータのスターティングアドレスである、ジャーナルボリュームにおけるオフセットアドレスを提供する。あるいは、コントロールデータが対応するジャーナルデータに近接して記憶されているので、コントロールデータはＪＶＯＬ＿ＩＤ２１６とＪＯＦＳ２１７を含まないかもしれない。

【００３５】

仲介ストレージシステム１１０ｃは、第１のポインター（ＪＯＰＴＲ）２１８と第２のポインター（ＪＩＰＴＲ）２１９を保持している。ＪＯＰＴＲ２１８はリモートストレージシステム１１０ｂに送られるジャーナルを指す。ＪＩＰＴＲ２１９はプライマリシステムから受け取った次のジャーナルを記憶するアドレスを指す。したがって、ＪＩＰＴＲ２１９は、リモートストレージシステム１１０ｂに未だ送られていないジャーナルを新しいジャーナルが重ね書きしないように、ＪＯＰＴＲ２１８より前に進んではならない。

【００３６】

図８に関連して説明されるが、仲介ストレージシステム１１０cはジャーナルボリューム上で空間を割り付ける。この空間の割付はＪＩＰＴＲ８２０にコントロールデータとジ

ャーナルデータの長さを加える、言い換えれば次のＪＩＰＴＲ＝ＪＩＰＴＲ＋（コントロールデータとジャーナルデータの長さ）にすることで実行される。前述のように、仲介ストレージシステム１１０cがプライマリストレージシステム１１０ａからＪＮＬＷＲコマンドを受け取ると、仲介ストレージシステム１１０ｃはコントロールデータとジャーナルデータを記憶するためにＪＮＬボリューム１１２上に、キャッシュメモリーと併行してスペースを割り付ける。キャッシュメモリーにジャーナルが記憶されると、仲介ストレージシステム１１０ｃはＪＮＬＷＲコマンドの完了を送る。続いてジャーナルはＪＮＬボリューム１１２‘に記憶される。

【００３７】

図３は本発明の他の実施例におけるＪＮＬボリュームあるいはＪＶＯＬ１１２'を示す。上記にて説明したように、ＪＶＯＬはリモートコピーを行うために用いられるジャーナルあるいはジャーナル情報を記憶するように構成されている。ジャーナルはジャーナルデータとそのコントロールデータの一対を含んでいる。コントロールデータはＪＶＯＬ１１２'のコントロールデータ領域２３０に記憶され、ジャーナルデータはＪＶＯＬ１１２’のジャーナルデータ領域２４０に記憶される。一実施例においては、ＪＶＯＬは先入れ先出し（ＦＩＦＯ）ストレージであって、コントロールデータとジャーナルデータは受け取った順序で読み出される。一実施例では、コントロールデータは、ジャーナルデータが引き出されるＰＶＯＬ１１１aの識別子であるインデックス（ＩＤＸ）２１１'を含んでおり、たとえば固有の数値がプライマリストレージシステム１１０ａあるいはジャーナルグループ２００（図４）のＰＶＯＬに割り当てられる。

【００３８】

アドレス２１２'は書き込みデータがそこから書かれるＰＶＯＬのオフセットアドレス、たとえば書き込みデータのスターティング論理ブロックアドレス（ＬＢＡ）、を提供する。長さ２１３'は書き込みデータの長さ、たとえば論理ブロックの数あるいは書き込みデータの総バイト数、を提供する。時刻２１４'はホストがＰＶＯＬ１１１ａにデータを書き込む時を示す。シーケンス番号（ＳＥＱ＃）２１５’は書き込みのシーケンス情報を提供する。すなわち、シーケンス番号は、プライマリストレージシステム１１０ａ内における書き込み順序を与える。ＪＶＯＬアイデンティフィケーション（ＪＶＯＬ＿ＩＤ）２１６'は対応するジャーナルデータを含むジャーナルボリュームを識別し、たとえばプライマリストレージシステムあるいはジャーナルグループ２００において固有の数値がジャーナルボリュームに割り当てられる。ジャーナルオフセット（ＪＯＦＳ）２１７'はジャーナルデータの記憶が始まるジャーナルボリュームのオフセットアドレスあるいはジャーナルデータのスターティングアドレスを提供する。コントロールは、本発明の実施例ではコントロールデータとジャーナルデータは異なった領域に記憶されているので、さらにＪＯＦＳ２１７'とＪＶＯＬ＿ＩＤ２１６’を含んでいる。

【００３９】

ジャーナルには二つの形式があり、更新ジャーナルとベースジャーナルである。更新ジャーナルはホストから書き込まれたデータに対するジャーナルである。ジャーナルはホストがデータをＰＶＯＬ１１１ａに書き込む際に取られる。ベースジャーナルは、組み合わせを行う前にＰＶＯＬ１１１ａに存在している、先に存在しているデータに対するジャーナルである。ベースジャーナルはＰＶＯＬの新しいコピーが作成された時、あるいは再同期化が必要な際に取られる。

【００４０】

図３は本発明の一実施例によるジャーナルグループを示す。ジャーナルグループはジャーナルが生成されるべきボリューム（１ないしはそれ以上のボリューム）のセットである。ボリュームは只一つのジャーナルグループ３００の一部であることも可能である。ジャーナルグループ３００は１ないしはそれ以上のデータボリューム１１１あるいは１ないしはそれ以上のジャーナルボリュームを含んでいる。ジャーナルグループ３００はマスターあるいはリストアの何れかの属性を持っている。マスタージャーナルグループ３００ａはジャーナルを生成し、一方リストアジャーナルグループ３００ｂはジャーナルをＳＶＯＬ

１１１ｂに更新する。

【００４１】

マスタージャーナルグループはプライマリストレージシステム１１０ａに関連し、１ないしはそれ以上のＰＶＯＬ１１１ａを含み、随意に１ないしはそれ以上のジャーナルボリュームを含む。リストアジャーナルグループはセカンダリストレージシステム１１０ｂに関連し、１ないしはそれ以上のＰＶＯＬ１１１ｂを含み、随意に１ないしはそれ以上のジャーナルボリュームを含む。ジャーナルグループはたとえばＪＶＯＬ１１２のような仲介的な属性（図示せず）を持つこともある。本発明の実施例では、仲介的なジャーナルグループはマスタージャーナルグループやリストアグループとグループ化している。このような仲介的なジャーナルグループ（図示せず）は仲介ストレージシステム１１０ｃと連携し、１ないしはそれ以上ＪＶＯＬを含み、随意に１ないしはそれ以上のＳＶＯＬ１１１ｂを含む。

【００４２】

図５は本発明の一実施例によるジャーナルグループ（ＪＮＬＧ）テーブル４００を示している。ストレージシステムは、対応するジャーナルグループが生成されると、ＪＮＬＧテーブル４００を生成し保持する。

【００４３】

図４はプライマリストレージシステム１１０ａとセカンダリストレージシステム１１０ｂと仲介ストレージシステム１１０ｃにより保持されているＪＮＬＧテーブル４００の内容を示す。ジャーナルグループ番号（ＧＲＮＵＭ）４１０はストレージシステム１１０内でジャーナルグループに個別に割り当てられた番号を示す。ジャーナルグループ名（ＧＲＮＡＭＥ）４２０は一般にユーザが割り当てるものであるが、ジャーナルグループに与えられた名称を示す。もし、２ないしはそれ以上のグループが同じ名称ＧＲＮＡＭＥ４２０を持つ場合は、それらはリモートミラーリング関係にある。ジャーナルグループの属性（ＧＲＡＴＴＲ）４３０はジャーナルグループに割り当てられた属性、たとえばマスター（ＭＡＳＴＥＲ）や仲介（ＩＮＴＥＲＭＥＤＩＡＲＹ）やリストア（ＲＥＳＴＯＲＥ）などを示す。

【００４４】

上記で説明したように、マスタージャーナルグループはジャーナルグループにおけるデータボリューム（ＰＶＯＬ）からジャーナルを生成する。仲介ジャーナルグループはマスタージャーナルグループとリストアジャーナルグループ間に於いて過渡的なジャーナルグループである。リストアジャーナルグループはジャーナルボリュームからデータボリューム（ＳＶＯＬ）に対してジャーナルを再生する。グループステータス（ＧＲＳＴＳ）４４０はジャーナルグループのステータスを示す。ジャーナルグループは以下のＣＯＰＹやＰＡＩＲやＳＵＳＰやＳＭＰＬなどのステータスをもつこともある。

【００４５】

ＣＯＰＹステータスはベースジャーナルを取り出す先であるジャーナルグループにデータボリュームがあることを示す。すなわち、組み合わせ（言い換えればベースジャーナルに記憶されている事前に存在するデータ）前にＰＶＯＬに記憶されているデータがＳＶＯＬにコピーするために検索されることである。ＰＡＩＲステータスは全ての予め存在したデータはＳＶＯＬにコピーされ、ＳＶＯＬへコピーするためにジャーナルグループは更新ジャーナルから更新されたデータを検索されあるいは検索してきた。ＳＵＳＰステータスすなわちサスペンドステータスは、ジャーナルグループが、更新ジャーナルから更新したデータを取ることあるいは検索することを中断していることを示す。ＳＭＰＬステータスは、ジャーナルグループのどのボリュームもベースジャーナルから予め存在するデータを取得することを開始していない、言い換えればリモートコピーを開始しようとすることを示している。

【００４６】

アレイ（ＤＶＯＬ＿ＩＮＦＯ）４５０はジャーナルグループの全てのデータボリュームについての情報を記憶する。アレイの各エントリーは次の情報を含んでいる。ストレージ

システム１１０においてボリュームに個別に指定されるボリューム識別（ＶＯＬＩＤ）４５１（ＶＯＬＩＤ４５１は一般に整数の値を持つ）と、ユーザによりボリュームに指定されるボリューム名称（ＶＯＬＮＡＭＥ）４５２と、たとえばＣＯＰＹやＰＡＩＲやＳＵＳＰやＳＭＰＬやその他のボリュームのステータスを示すボリュームステータス(ＶＯＬＳＴＳ)４５３を含んでおり、ポインター（ＢＪＰｔｒ）４５４はベースジャーナルを取得する経過を保持する。ジャーナルグループにおけるデータボリュームにはグループにおいてＤＶＯＬ＿ＩＮＦＯ４５０の索引のために個別のインデックスが指定される。

【００４７】

アレイ（ＪＶＯＬ＿ＩＮＦＯ）４６０はジャーナルグループにおける全てのジャーナルボリュームに関連する情報を記憶する。ＪＶＯＬ＿ＩＮＦＯ４６０はジャーナルボリューム１１２の識別子であるＶＯＬＩＤ４６１を含んでいる。ジャーナルグループにおけるジャーナルボリュームにはグループにおいてＪＶＯＬ＿ＩＮＦＯ４５０アレイの索引のために個別のインデックスが指定される。

【００４８】

図６は本発明の一実施例によるベースジャーナル生成のためのプロセス５００を示す。プロセス５００はまた初期コピープロセスとも呼ばれる。ベースジャーナルはユーザがＰＡＩＲ＿ＣＲＥＡＴＥまたはＰＡＩＲ＿ＳＹＮＣコマンドを発行すると取得される。ＰＡＩＲ＿ＣＲＥＡＴＥコマンドによりプライマリストレージシステムの第１のボリュームがセカンダリストレージシステムの第２のボリュームと組み合わされる。本発明の実施例では、もし以下の条件、すなわち、（１）プライマリストレージシステム１１０ａとセカンダリストレージシステム１１０ｃに保持されている二つのＪＮＬグループ３００が同じＧＲＮＡＭＥ４２０を持っている、（２）プライマリストレージシステム１１０ａの二つのＪＮＬグループ３００の一つがＧＲＡＴＴＲ４３０においてＭＡＳＴＥＲ属性を持ち、セカンダリストレージシステム１１０ｃにおける他方がＲＥＳＴＯＲＥ属性を持っている、（３）二つのＪＮＬグループ３００からの二つのデータボリュームが同じインデックス番号を持っている、が満たされる場合には、二つのデータボリュームはペアの関係にある。

【００４９】

ＰＡＩＲ＿ＳＹＮＣコマンドは、ペアの関係にあるボリュームに対して、双方が同じデータを保有するように同期をさせるかあるいはミラーリングをさせる。プライマリストレージシステムはこれらの二つののコマンドの一つを受け取ると、プロセス５００を実行する。ステップ５１０においては、ベースジャーナルポインター（ＢＪＰｔｒ）４５４は初期化されて、データボリュームにおける第１のデータ（たとえばデータボリューム上の第１のブロック、トラック、ブロックの一塊やアドレス指定可能な任意のデータなど）からベースジャーナルの取得を開始する。

【００５０】

次のターゲットが検査される（ステップ５１５）。次のターゲットはＢＪＰｔｒ４５４の値により得られる。たとえばもしＢＪＰｔｒがｊを持っていると、次のターゲットはｊ番目のブロックである。一実施例においては、より効率の良い処理のために、ジャーナルについて一度に数ブロックのデータが取得される。したがって、本例ではｊ番目のブロックの後、次のターゲットはｊ番目のブロックからｎブロックである。プロセスは追加のターゲットが有るか否かを決める（ステップ５２０）。プロセス５００はもうそれ以上のターゲットが無い、言い換えれば全てのベースジャーナルが取得された場合には終了する。しかしながら、もしさらにターゲットが存在する場合には、ターゲットブロックに関するコントロールデータが生成される（ステップ５２５）。生成されたコントロールデータはプライマリストレージシステム１１０ａのキャッシュメモリーに記憶される。ターゲットデータはＰＶＯＬからキャッシュメモリーに読み込まれる（ステップ５３０）。ターゲットデータを読んだ後で、コントロールデータは確認される。

【００５１】

コントロールデータは、以下の情報、すなわち、ＩＤＸ２１１やアドレス２１２や長さ２１３を含んでいる。他の情報も同様に含まれている可能性がある。ジャーナルデータと

コントロールデータは仲介ストレージシステム１１０ｃに送られる（ステップ５３５）。一般に、ステップ５３０で確認されたジャーナル（コントロールデータとジャーナルデータの一対）のみが仲介システムに転送される。一実施例では、複数の確認されたジャーナルが単一のコマンドで共に送られる。ジャーナルを無事に転送し終えるとＢＪＰｔｒ４５４は次のターゲットへとインクリメントされる。すなわち、ＢＪＰｔｒ４５４はｊ＋ｎにインクリメントされる。ステップ５１５と５４０はターゲットが無くなるまで繰り返される。

【００５２】

図７は本発明の一実施例による更新ジャーナルを生成するためのプロセス６００を示す。プロセス６００はまた更新コピープロセスとも呼ばれる。ユーザによりＰＡＩＲ＿ＣＲＥＡＴＥコマンドあるいはＰＡＩＲ＿ＳＹＮＣコマンドが発行された後でプロセス６００は開始される（ステップ６０２）。すなわち、プライマリストレージシステムが更新ジャーナルの取得を開始する。ステップ６０２はもしＰＶＯＬが予め存在するデータを持っている場合は、プロセス５００の後で実行される。プライマリストレージシステム１１０ａは書き込みコマンドを受け取ったか否かを判定する（ステップ６１０）。この時には全ての読み取りコマンドが無視される。

【００５３】

ストレージはまたＰＶＯＬのペアステータスがＣＯＰＹであるかＰＡＩＲであるかを判定する。もしこれらの条件が満足されると、プライマリストレージシステムはペアステータスがＣＯＰＹであるかを判定する（ステップ６１５）。もしそうであるならば、ベースジャーナルが書き込みターゲットアドレスに対して既に取得されているかを判定するためにステータスがチェックされる（ステップ６２０）。これはポインターＢＪＰｔｒ４５４を判定することで行われる。すなわち、もし（書き込みターゲットアドレス）＜＝ＢＪＰｔｒ４５４であるならば、プロセス６００はステップ６２５に進む。もしステップ６２０が真であるかステップ６１５が偽である、言い換えれば、ベースジャーナルを取得するプロセスが完了している場合は、更新ジャーナルは書き込みのために取得される。この目的のために、コントロールデータは最初に生成される。

【００５４】

コントロールデータはＩＤＸ２１１とアドレス２１２と長さ２１３を含んでいる。書き込みコマンドはアドレス２１２と長さ２１３を含んでいる。コントロールデータには他の情報が含まれている可能性がある。書き込みデータはホストから受け取り、キャッシュメモリーに記憶される（ステップ６３０）。書き込みデータはステップ６２５で生成されたコントロールデータに係わるジャーナルデータに対応する。コントロールデータとジャーナルデータは仲介ストレージシステム１１０ｃに転送される（ステップ６３５）。プロセス６００は仲介ストレージシステム１１０ｃからの受け取り通知を待つ（ステップ６４０）。受け取り通知を受け取ると、書き込み完了がホストに送られる（ステップ６４５）。仲介ストレージシステムからの受け取り通知が受けとられるまでは書き込み完了がホストに通知されないので、書き込みデータのプライマリシステムと仲介システムへの記憶は保証される。

【００５５】

図８は本発明の一実施例による、コントロールデータとジャーナルデータを含むジャーナル情報を仲介ストレージシステム１１０ｃに転送するプロセス７００を示している。プライマリストレージシステム１１０ａは、仲介ストレージシステム１１０ｃにジャーナルデータを送るためのＪＮＬ書き込みコマンド（ＪＮＬＷＲコマンド）を発行する（ステップ７０２）一実施例においては、コマンドは１ないしはそれ以上のパラメータ、たとえばジャーナルデータの長さなどを含んでいる。コントロールデータの長さは、固定データ長、たとえば６４バイト、が本発明の実施例のコントロールデータに用いられているのでコマンドパラメータには含まれていない。

【００５６】

もう一つの方法としては、可変長のコントロールデータが用いられることもあり、その

場合はその長さについての情報がパラメータに含まれる必要がある。仲介ストレージシステム１１０ｃは、コマンドパラメータにおいて提供される情報に従いＪＮＬボリューム上に記憶空間を割り付ける（ステップ７１０）。以下でより詳細に説明するように、空間の割り付けは、書き込み性能を向上させるために、キャッシュメモリーバッファー上でも実行される。割り付けられたバッファーはボリュームにおいて割り付けられたストレージ空間に関係している。

【００５７】

一旦ストレージ割り付けが実行されると、転送の準備が出来たパケットあるいはメッセージがプライマリストレージシステム１１０ａへ送られる。プライマリシステムは、転送の準備の出来たパケットを受け取ると、ジャーナル情報を仲介ストレージシステムに転送する（ステップ７２０）。一実施例においては、コントロールデータが最初に送られ、次いでジャーナルデータが送られる。仲介ストレージシステムはキャッシュメモリー上のバッファーにコントロールデータとジャーナルデータを記憶させる（ステップ７３０）。ジャーナルは最終的には、ステップ７２０におけるバッファーストレージ結合によりプライマリストレージシステム１１０ｂがアイドル状態である時に、仲介システムにおける割り付けられたＪＮＬボリュームに記憶される。

【００５８】

さらに、シーケンス番号および／あるいはその時のタイムスタンプがジャーナルに割り付けられ、言い換えればコントロールデータに付加される。シーケンス番号は順番に（プライマリシステムから）受け取られたジャーナルに割り当てられ、ＪＮＬボリュームに記憶される。シーケンス番号によりプライマリシステムから受け取ったジャーナルの順番が分かるので、データリカバリープロセスに役立つ。本発明の実施例では、仲介システム１１０ｃはジャーナルにシーケンス番号を付加し、さもなければ、シーケンス番号を管理する。仲介システムにおいてはカウンター１５２は、プライマリシステム１１０ａ（図１を参照）により転送されたジャーナルにシーケンス番号を付加するために用いられる。タイムスタンプも、仲介ストレージシステム１１０ｃがジャーナルを受け取った時刻を示すために、ジャーナルに付加される。

【００５９】

他の実施例においては、シーケンス情報は、ジャーナルが仲介システムに転送される前に、プライマリシステムにおいてジャーナルに付加される。同様に、タイムスタンプは、それが仲介システムへ転送された時刻を示すために、プライマリシステムによってジャーナルに付加することもある。一旦ジャーナルが正常に受信され記憶されると、仲介システムはジャーナルを問題なく受信したことの認証をプライマリシステムに送る。その後は、プライマリストレージシステム１１０ａは書き込みジャーナルコマンドの完了を発行する。

【００６０】

図９は本発明の一実施例による、仲介ストレージシステム１１０ｃからセカンダリストレージシステム１１０ｂへジャーナルを送るプロセス９００を示す。本発明の実施例においては、仲介ストレージシステム１１０ｃのＪＮＬボリューム１１２に記憶されたジャーナルは、プライマリストレージシステム１１０ａの書き込みコマンドとは非同期で、言い換えればプライマリシステムからの書き込みコマンドの合間に、セカンダリストレージシステム１１０ｂへ送られる。プロセス９００はプライマリシステムから仲介システムへジャーナルを転送することに関してはプロセス７００と同様である。パラメータが付いたＪＮＬ ＷＲコマンドが仲介システムからセカンダリストレージシステムへ発行される（ステップ９０２）。セカンダリシステムはパラメータにより規定されたデータ長に従いキャッシュメモリー上にバッファー空間を割り付け、転送準備完の通知を仲介システムに送り返す（ステップ９１０）。仲介システムはコントロールデータとそれに対応するジャーナルデータを含んでいるジャーナルを送る（ステップ９２０）。一実施例においては、コントロールデータは最初に転送され、続いてジャーナルデータが転送される。セカンダリシステムは割り当てられたバッファー空間にジャーナルを記憶し、仲介システムにジャーナ

ル受信の受け取り通知を送る（ステップ９３０）。

【００６１】

ジャーナルデータはプロセス７００のステップ７３０において割り当てられたシーケンス番号とタイムスタンプに基づきＳＶＯＬに記憶される。たとえば、下位のシーケンス番号を持つジャーナルは上位のシーケンス番号を持つジャーナルの前に再記憶される。受け取り通知を受信すると、仲介システムはデータ書き込みの完了を示すＷＲＪＮＬコマンドを発行する（ステップ９４０）。仲介システムのジャーナルボリュームに関連したポインター、たとえばＪＯＰＴＲ８１０は、次のバッチのデータまで前進しセカンダリシステムにコピーされる。

【００６２】

図１０は、本発明の一実施例にしたがって、セカンダリストレージシステム１１０ｂのＳＶＯＬへ、その対応するコントロールデータを用いてジャーナルデータを記憶するプロセス１０００を示している。セカンダリストレージシステム１１０ｂは、ＲＥＳＴＯＲＥの属性を持つＪＮＬグループ３００上で定期的にプロセス１０００（ステップ１００２）を実行する。一実施例では、プロセス１０００は１０秒ごとに実行される。ＳＶＯＬに記憶する予定の、コントロールデータとジャーナルデータを含む次のジャーナルは、シーケンス番号を用いて選択される（ステップ１００５）。この目的のために、セカンダリストレージシステム１１０ｂは再記憶した、言い換えればそのジャーナルをＳＶＯＬに記憶した、ジャーナルのシーケンス番号の追跡をしている。セカンダリストレージシステムは、最も最近に再記憶されたジャーナルのシーケンス番号とキャッシュメモリーに一時的に記憶されたジャーナルに係わるシーケンス番号とを比較し、再記憶する次のジャーナルを決定する。

【００６３】

ステップ１００５で選択されたジャーナルのコントロールデータは、ジャーナルデータの記憶領域、言い換えれば特定のＳＶＯＬとその中の場所、を決めるために用いられる（ステップ１０１０）。たとえば、コントロールデータの以下の場所、すなわちＩＤＸ２１１、アドレス２１２、長さ２１３、が調べられる。ＩＤＸ２１１はＭＡＳＴＥＲ ＪＮＬグループのＰＶＯＬ１１１ａ、言い換えればプライマリシステムにおけるプライマリストレージボリュームのインデックスを示す。ジャーナルデータは同じインデックスを持つＳＶＯＬに記憶される（ステップ１０１５）。すなわち、ジャーナルデータはＩＤＸ２１１で特定され、アドレス２１２で示されるアドレスで、長さ２１３に対応する長さによりＳＶＯＬに記憶される。本発明の実施例では、コントロールデータは、ＰＶＯＬとＳＶＯＬにおけるジャーナルデータの記憶場所はミラーリングされているので、ＳＶＯＬには記憶されない。

【００６４】

図１１は本発明の一実施例の場合に於いて、プライマリシステム１１０ａに障害が発生した場合に、リモートコピーシステム５０がフェイルオーバを実行するところを示している。フェイルオーバはプライマリストレージシステム１１０ａまたはホスト１００ａまたはその両者が停止し、セカンダリホスト１００ｂが適当なアプリケーションを走らせる状態あるいはプロセスを言い、ここでセカンダリストレージシステムは新しい「プライマリ」ストレージシステムとして機能する。もしプライマリストレージシステム１１０ａが未だ動作可能であるか、あるいは障害が発生した後で再起動した場合は、セカンダリストレージシステム１１０ｂのデータボリューム１１１ｂをＰＶＯＬと設定し、二つのサイトの間でミラーリングを継続することが求められる。

【００６５】

仲介ストレージシステム１１０ｃは従来と同様に仲介ストレージとして用いられる。しかしながら、仲介ストレージシステム１１０ｃはセカンダリストレージシステム１１０ｂから遠く離れている可能性があるので、新しいプライマリシステム（言い換えればセカンダリシステム１１０ｂ）はホスト１００ｂの書き込みリクエストに対して非同期でジャーナルを送信する。この目的のために、プロセス６００はステップ６３５と６４０無しに実

行される。すなわち、ジャーナルは同期して生成されるが、ホストの書き込みリクエストに対しては非同期で送られる。コントロールデータには、仲介システム１１０ｃへ送信する前に、新しいプライマリシステム１１０ｂでタイムスタンプとシーケンス番号が付加される。したがって、フェイルオーバの際には、仲介システムでは、このようなステップを実行する必要はない。

【００６６】

図１２は本発明の他の実施例における、リモートコピーシステム５０’を示す。システム５０’はプライマリストレージシステム１１０ａ’とセカンダリストレージシステム１１０ｂ’と仲介ストレージシステム１１０ｃ’を含む。プライマリストレージシステムは複数のボリューム１１１ａ’を含み、コミュニケーションリンク１３０ａ’を経由してプライマリホスト１００ａ’と結合されている。プライマリホスト１００ａ’はアプリケーション１０２’とＲＣＭ１０１ａ’を含む。セカンダリシステムは複数のボリューム１１１ｂ’を含み、コミュニケーションリンク１３０ｂ’を経由してセカンダリホスト１００ｂ’と結合されている。セカンダリシステムはＲＣＭ１０１ｂ’を含んでいる。仲介システムはジャーナルボリューム１１２’と複数のデータボリューム１１１ｃ’を含んでいる。一実施例では、データボリューム１１１ｃ’は、たとえば他のストレージシステムにあるジャーナルボリューム１１２’から遠く離れた場所に位置しているかもしれない。

【００６７】

仲介ホスト１００ｃ’はコミュニケーションリンク１３０ｃ’を経由して仲介システム１１０ｃ’に結合している。仲介ホストはＲＣＭ１０１ｃ’を含んでいる。システム５０’では、仲介システムとセカンダリシステムは共にプライマリシステムとミラーリングを行っている。仲介システムにおけるデータミラーリングは、上述のプロセス７００、９００および１０００を用いて行われる。システム５０’は、プライマリストレージシステム１１０ａ’が停止した場合は、仲介ストレージシステム１１０ｃ’あるいはセカンダリストレージシステム１１０ｂ’にフェイルオーバすることができる。このような構成に於いては、仲介ストレージシステムは、セカンダリストレージシステムよりもプライマリシステムやユーザに近い可能性があるので、セカンダリシステムよりも効率の良いストレージセンターの役を果たすものである。フェイルオーバの際には、仲介ホスト１００ｃ’は単独であるいはホスト１００ａ’と協調してプライマリホストとして機能する。

【００６８】

上記の詳細説明は本発明の特定の実施例を説明するために提供されたものであって、制約を目的としたものではない。多くの改造や変更が本発明の範囲内に於いて可能である。したがって、本発明は特許請求の範囲により定義されるものである。

【図面の簡単な説明】

【００６９】

【図１Ａ】は、本発明の一実施例による３箇所のデータセンタを有するリモートコピーシステムを示す。

【図１Ｂ】は、本発明の一実施例による典型的なストレージシステムの例を示す。

【図２】は、本発明の一実施例による仲介ストレージシステムにおいて提供されるジャーナルボリュームを示す。

【図３】は、本発明の他の実施例による仲介ストレージシステムにおいて提供されるジャーナルボリュームを示す。

【図４】は、本発明の一実施例によるマスターとリストアの属性を持つジャーナルグループを示す。

【図５】は、図１のリモートコピーシステムにおいてストレージシステムにより維持されているジャーナルグループテーブルを示す。

【図６】は、本発明の一実施例によるベースジャーナルを生成するプロセスを示す。

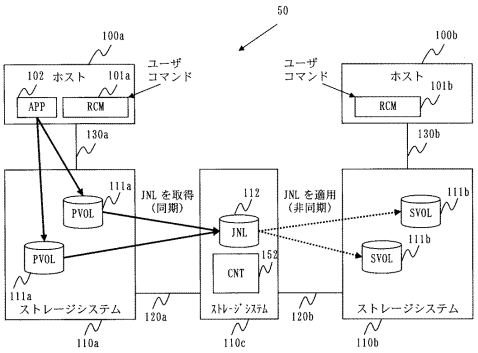
【図７】は、本発明の一実施例による更新ジャーナルを生成するプロセスを示す。

【図８】は、発明の一実施例によるジャーナルをプライマリストレージシステムから仲介ストレージシステムに転送するプロセスを示す。

【図９】は、本発明の一実施例によるジャーナルを仲介ストレージシステムからセカンダリストレージシステムへ送るプロセスを示す。
【図１０】は、本発明の一実施例によるセカンダリストレージシステムにおいてジャーナルを復元するプロセスを示す。
【図１１】は、本発明の一実施例によるリモートコピーシステムにおいて実行されるフェイルオーバあるいはフォールバックプロセスを示す。
【図１２】は、本発明の他の実施例によるリモートコピーシステムを示す。
【符号の説明】
【００７０】
１００ａ・・・ホスト
１０１ａ・・・ＲＣＭ、ユーザコマンド
１０２・・・ＡＰＰ
１１１ａ・・・ＰＶＯＬ
１１１ｂ・・・ＳＶＯＬ
１１０ａ・・・ストレージシステム
１２０ａ・・・ＪＮＬを取得（同期）
１２０ｂ・・・ＪＮＬを適用（非同期）
２３０・・・コントロールデータ領域、コントロールデータ
２４０・・・ジャーナルデータ領域、ジャーナルデータ

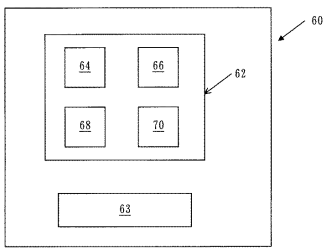
【図１Ａ】

図１Ａ



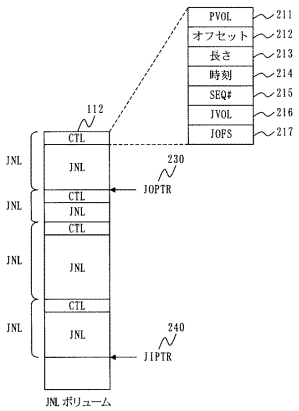
【図１Ｂ】

図１Ｂ



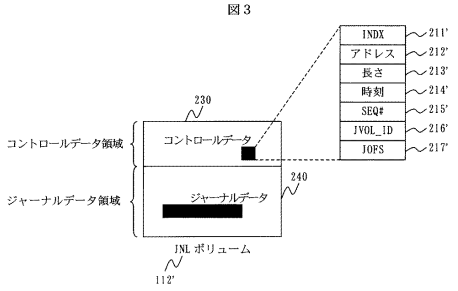
【図２】

図２



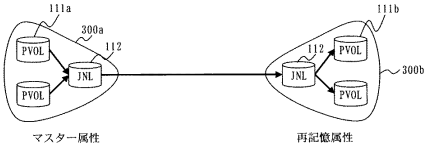
【図３】

図３



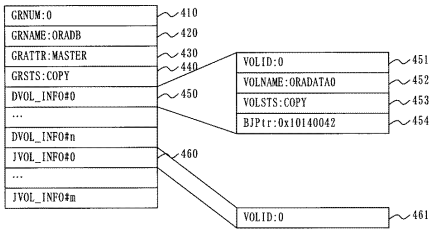
【図４】

図４



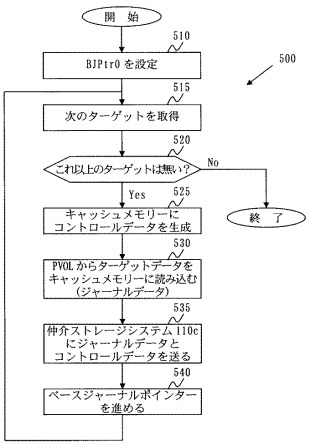
【図５】

図５



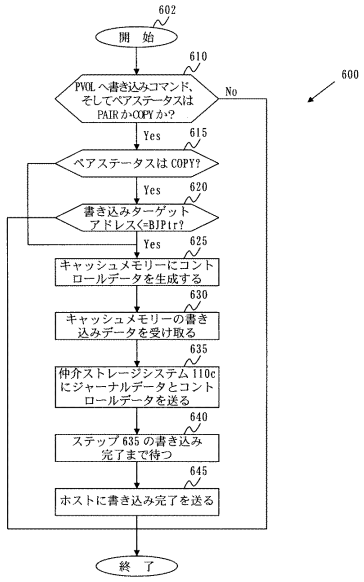
【図６】

図６



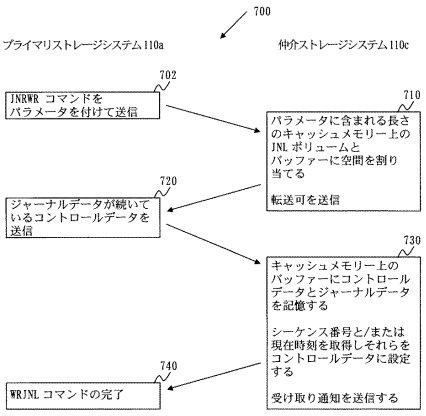
【図７】

図７



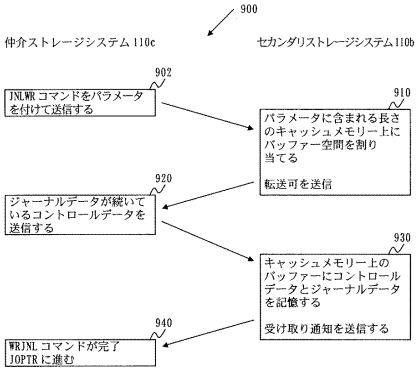
【図８】

図８



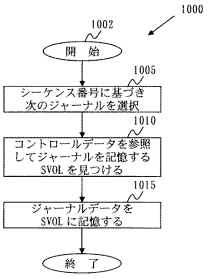
【図９】

図９



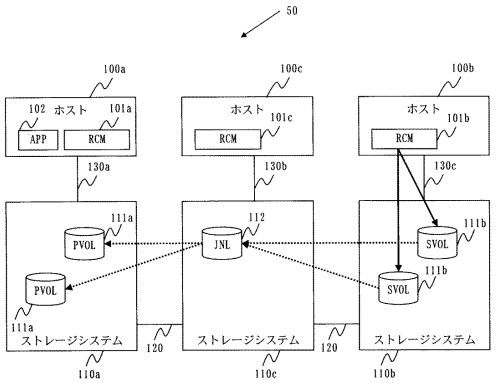
【図１０】

図１０



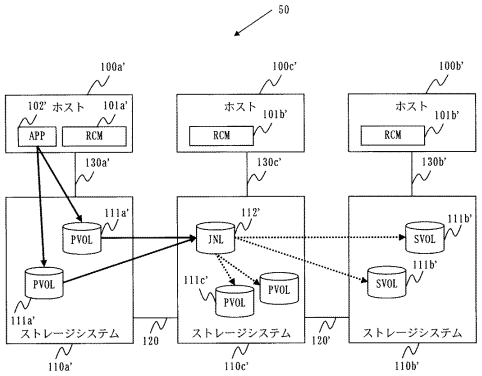
【図１１】

図１１



【図１２】

図１２



フロントページの続き

(56)参考文献 特開２００３－１２２５０９（ＪＰ，Ａ）
特開平１０－０４９４１８（ＪＰ，Ａ）
特開２０００－１８１６３４（ＪＰ，Ａ）

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