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

特許第4452494号
(P4452494)

(45)発行日　平成22年4月21日(2010.4.21)　(24)登録日　平成22年2月5日(2010.2.5)

(51)Int.Cl.　F I
G06F 3/06 (2006.01)　G06F 3/06 304F
G06F 12/00 (2006.01)　G06F 3/06 301Z
G06F 12/00 531D

請求項の数 8　外国語出願　(全 26 頁)

(21)出願番号　特願2003-428627 (P2003-428627)
(22)出願日　平成15年12月25日 (2003.12.25)
(65)公開番号　特開2004-272884 (P2004-272884A)
(43)公開日　平成16年9月30日 (2004.9.30)
審査請求日　平成18年10月27日 (2006.10.27)
(31)優先権主張番号　10/386277
(32)優先日　平成15年3月11日 (2003.3.11)
(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　渡辺　直企
アメリカ合衆国カリフォルニア州サニーベイル　イーストエルカミノリアル９６５

審査官　高瀬　勤

最終頁に続く

(54)【発明の名称】複数リモートストレージでのリモートコピー停止後のデータ同期化方式

(57)【特許請求の範囲】

【請求項１】

第一、第二、第三、第四のストレージボリュームが直列接続され、各ストレージボリュームに保存されたデータを同期化させるストレージシステムにおいて、

前記第二のストレージボリュームと前記第三のストレージボリュームとの間のリモートコピー動作の停止を検出する検出部と、

前記検出部による前記リモートコピー動作の停止が検出されたことを契機に、前記第二のストレージボリュームになされた前記リモートコピー動作の停止までの第一の変更記録を生成するとともに前記第三のストレージボリュームになされた前記リモートコピー動作の停止までの第二の変更記録を生成し、前記第一の変更記録の複写情報を前記第一のストレージボリュームに保存するとともに前記第二の変更記録の複写情報を前記第四のストレージボリュームに保存する制御部と、

前記第一の変更記録の複写情報を用いて、前記第四のストレージボリュームを前記第二のストレージボリュームに同期化させ、又は、前記第二の変更記録の複写情報を用いて、前記第三のストレージボリュームを前記第一のストレージボリュームに同期させるデータ同期化部と、

を備えることを特徴とするストレージシステム。

【請求項２】

前記第一のストレージボリュームを有するプライマリストレージシステムと、

前記第二、第三及び第四のストレージボリュームを有する少なくとも２台のセカンダリ

ストレージシステムと、
を備え、
前記プライマリストレージシステムは、
ホストコンピュータ又は前記少なくとも２台のセカンダリストレージシステムが発行するＩ／Ｏ要求に基づいて生成されるコマンドエントリを格納するための第一のキューを有する第一のディスクコントローラを備え、
前記少なくとも２台のセカンダリストレージシステムは、それぞれ、
前記ホストコンピュータ又は前記プライマリストレージシステムが発行するＩ／Ｏ要求に基づいて生成されるコマンドエントリを格納するための第二のキューを有する第二のディスクコントローラを備える
ことを特徴とする請求項１に記載のストレージシステム。
【請求項３】
前記第一及び第二のディスクコントローラは、それぞれ、前記第一及び第二のキューとして、ワーキングキュー、中間キュー及びライトヒストリキューを備える
ことを特徴とする請求項２に記載のストレージシステム。
【請求項４】
前記第一及び第二のディスクコントローラは、それぞれ、前記第一及び第二のキューとして、コールバックキュー及びライトヒストリキューを備える
ことを特徴とする請求項２に記載のストレージシステム。
【請求項５】
第一、第二、第三、第四のストレージボリュームが直列接続されたストレージシステム内のストレージボリュームに保存されたデータを同期化させるデータ同期化方法において、
前記第二のストレージボリュームと前記第三のストレージボリュームとの間のリモートコピー動作の停止を検出する検出ステップと、
前記リモートコピー動作の停止を検出したことを契機に、前記第二のストレージボリュームにおいて前記第二のストレージボリュームになされた前記リモートコピー動作の停止までの第一の変更記録を生成するとともに前記第三のストレージボリュームにおいて前記第三のストレージボリュームになされた前記リモートコピー動作の停止までの第二の変更記録を生成し、前記第一の変更記録の複写情報を前記第一のストレージボリュームに保存するとともに前記第二の変更記録の複写情報を前記第四のストレージボリュームに保存する制御ステップと、
前記第一の変更記録の複写情報を用いて、前記第四のストレージボリュームを前記第二のストレージボリュームに同期化させ、又は、前記第二の変更記録の複写情報を用いて、前記第三のストレージボリュームを前記第一のストレージボリュームに同期させる同期ステップと
を備えることを特徴とするデータ同期化方法。
【請求項６】
前記第一のストレージボリュームを有するプライマリストレージシステムでは、第一のディスクコントローラが、第一のキューに、ホストコンピュータ又は前記少なくとも２台のセカンダリストレージシステムが発行するＩ／Ｏ要求に基づいて生成されるコマンドエントリを格納させ、
前記第二、第三及び第四のストレージボリュームを有する少なくとも２台のセカンダリストレージシステムでは、それぞれ、第二のディスクコントローラが、第二のキューに、前記ホストコンピュータ又は前記プライマリストレージシステムが発行するＩ／Ｏ要求に基づいて生成されるコマンドエントリを格納させる
ことを特徴とする請求項５に記載のデータ同期化方法。
【請求項７】
前記第一及び第二のディスクコントローラは、それぞれ、前記第一及び第二のキューとして、ワーキングキュー、中間キュー及びライトヒストリキューを用いる

ことを特徴とする請求項６に記載のデータ同期化方法。
【請求項８】
前記第一及び第二のディスクコントローラは、それぞれ、前記第一及び第二のキューとして、コールバックキュー及びライトヒストリキューを用いる
ことを特徴とする請求項６に記載のデータ同期化方法。

【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、プライマリ（ローカル）ストレージシステムと、セカンダリ（リモート）ストレージシステムで構成されるデータ処理ストレージシステムに関する。
【背景技術】
【０００２】
商業、政府、その他の企業体によるデータ処理の使用拡大により、夥しい量のデータが保存されており、この多くのものはそれら企業体の日々のオペレーションに対して極めて重要なものである。
【０００３】
例えば、莫大な数の財務上のトランザクションは現在、完全に電子的に処理されている。
【０００４】
業務上、例えば、航空会社は将来のチケット予約に関する情報が失われると大混乱になるというリスクを背負っている。信頼できるデータへのニーズの結果として、元データが損傷又は損失してもデータが使用できるように、保持されているデータの一つ以上のコピーを、ローカルサイトから離れた位置に、バックアップすることが普通である。データが重要である程、バックアップ方法は精巧なものでなければならない。
【０００５】
例えば、重要な又は貴重なデータを守るための一つの方法は、そのデータのコピーをローカルストレージシステムから地理的に離れたサイトにバックアップ保存することである。
【０００６】
各リモートストレージシステムは、ローカルストレージシステムで保持されるデータのミラーイメージを維持し、ローカルストレージシステムのローカルデータが更新される度に、その保存されたデータをローカルストレージシステムのローカルデータイメージの“ミラー”変更に更新する。
【０００７】
ローカルストレージシステムのデータをミラーするためのリモートストレージシステムの一例は、“Method and Appratus for Mirroring Data in a Remote Storage System”と題される米国特許No. 5,933,653に記されている。
【０００８】
【特許文献１】米国特許５９３３６５３号公報
【発明の開示】
【発明が解決しようとする課題】
【０００９】
リモートストレージシステムに転送される更新データは、遠隔コピー操作のオーバヘッドを低減するために、しばしばキューされグループ化されて、インターネットの如き、ネットワーク転送媒体を通して送信される。この結果、リモートサイトにミラーされるデータイメージはローカルサイトのものとは必ずしも同じではない。
【００１０】
もし、ローカルデータのミラーの為に、複数のリモートストレージが使用される場合は、少なくとも更新が完了するまでの間は、リモートストレージのデータイメージは互いに

異なっている事が有り得る。

【００１１】

これらの異なったデータイメージの存在は、ローカルシステムが損傷した場合には問題となる。ローカルストレージシステムの損傷は、あるリモートストレージシステムに損傷前のローカルストレージシステムに完全ではなくても比較的近いデータイメージを残し、他方、他のシステムでは最終更新操作では決して更新されないより古い“ステール”データイメージが存在する。

【００１２】

かくして、ローカルストレージシステムの損傷時には、リモートストレージシステムのデータをすべて最新データに一致させることを保障する再同期化処理が、システムの再起動以前に必要である。

【００１３】

更に解決を要するもう一つの問題は、リモートコピー動作中に“停止（suspension）”が発生した場合のシステム回復である。予期せぬ事故による途絶、例えば、キャッシュオーバフロー、コピー中のストレージシステム障害、ネットワーク障害等が発生してリモートコピー動作が停止した場合は、再同期化処理が必要になる。

【００１４】

リモートコピー動作での再同期化の一例は、“Method and Appratus for Independent Operation of a Remote Data Facility.”と題される米国特許No.6,092,066、に記されている。

【００１５】

然しながら、この特許での技術では、ある限定された環境での再同期化処理が可能になるのみである。

【００１６】

ある種のより複雑なシステム停止、例えば、リンク障害、キャッシュオーバフロー、及び／またはドライブ障害等の２つの障害が組み合わさって発生した場合、システムの再初期化をしないで済む再同期化の方策は用意されていない。

【００１７】

このような状況では、この技術では、少なくとも２つのコピーが存在する構成を保障していない為、再同期化処理のためにはボリューム全体のコピーが通常は必要になる。

【課題を解決するための手段】

【００１８】

本発明は、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間でデータのミラー処理を実行する改善されたデータ処理ストレージシステムを提供する。典型的に、各ストレージシステムは、データ記憶の為にボリュームを備え、ボリュームはミラー状態に維持される。

【００１９】

もし、たとえば何れかのストレージボリュームの障害又はネットワーク結合障害により、ボリューム間でのデータ転送が不能になったら、タイムスタンプ付きビットマップがプライマリストレージシステムで生成され、セカンダリストレージシステムの一つに記憶される。これらの記録は、結合リンク確立後にペアを再同期化するために用いられる。

【００２０】

もし、ペアの一つのメンバ又は他方で障害が発生した場合には、障害発生時に、記録が障害発生ストレージボリュームとは別のライトオペレーションステータスのストレージボリュームに書き込まれる。この記録は後でストレージボリュームを再同期化するために使用される。

【００２１】

各プライマリ及びセカンダリストレージシステムは追加のボリュームを備え、ミラーリモートペアの動作停止時は、ビットマップがペアの各メンバから追加のストレージボリュームの一つに保存される事が好ましい。

【００２２】
既に述べたように、これらのビットマップは、以前のペアの一メンバで保持されていた情報が失われても、新しいペアの再同期化のために使用することが出来る。再同期化処理は、新しいペア間でビットマップを交換して、新しいペアの再同期化に必要なライトオペレーションのセットを決定することにより実行される。
【００２３】
次いで、これらのライトオペレーションが実行され、新しく同期化されたペアが誕生する。
【００２４】
プライマリストレージシステムの第一ストレージボリュームと、第二ストレージボリュームをも備えるセカンダリストレージシステムの第一ストレージボリュームとの間でのデータ転送が途絶した後の、ストレージシステムに保存されたデータの同期化方法は、下記のステップから成り立つ事が好ましい。
【００２５】
第一に、プライマリ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のデータ転送の途絶を検出し、次に、プライマリストレージシステムにプライマリストレージシステムの第一ストレージボリュームに書き込まれたデータの記録を提供し、セカンダリストレージボリュームにセカンダリストレージシステムの第一ストレージボリュームに書き込まれたデータの記録を提供する。
【００２６】
次のステップでは、プライマリストレージシステムの第一ストレージボリュームに書き込まれたデータの記録の少なくとも一部のコピーを第二ストレージボリュームに生成する。
【００２７】
次に、第二ストレージボリュームの少なくとも一部のコピーとセカンダリストレージシステムの第一ストレージボリュームに書き込まれたデータの記録を用いて、セカンダリストレージシステムの第一ストレージボリュームがセカンダリストレージシステムの第二ストレージボリュームに同期化される。
【００２８】
他の実施例として、直列的に結合されたストレージシステムに保存されたデータの同期化方法が提供される。
【００２９】
ストレージシステムでは、第一、第二、第三、及び第四のストレージボリュームが直列的に結合されている。第二ストレージボリュームと第三ストレージボリューム間のデータ転送が途絶された後に、第二ストレージボリュームに第二ストレージボリュームに書き込まれたデータの第一の記録を提供し、第三ストレージボリュームに第三ストレージボリュームに書き込まれたデータの第二の記録を提供することにより、システムが再同期化される。
【００３０】
次に、システムは、第一の記録の少なくとも一部を第一ストレージボリュームにコピーし、第二の記録の少なくとも一部を第四ストレージボリュームにコピーする。コピー終了後のステップでは、コピーされた記録の少なくとも一部を用いて、第二、第三ストレージボリュームの少なくとも一つを、第一、第四ストレージボリュームの少なくとも一つに同期化する
【発明の効果】
【００３１】
何れかのストレージボリュームの障害又はネットワーク結合障害により、ストレージボリューム間でのデータ転送が不能になったら、タイムスタンプ付きビットマップがプライマリストレージシステムで生成され、セカンダリストレージシステムの一つに記憶される。

【００３２】
ペアの一つのメンバ又は他方で障害が発生した場合には、障害発生時に、記録が障害発生ストレージボリュームとは別のライトオペレーションステータスのストレージボリュームに書き込まれる。この記録は後でストレージボリュームを再同期化するために使用される。
【発明を実施するための最良の形態】
【００３３】
図１は、プライマリ又はローカルサイト１２と複数のセカンダリ又はリモートサイト１４がデータ通信ネットワーク１０７で通信的に結合された、データ処理システム１０を構成するストレージシステムを示す。
【００３４】
ローカルサイト１２はホストプロセッサ１０１とローカルストレージシステム１０４を備えている。
【００３５】
ホストプロセッサ１０１とストレージシステム１０４は、ローカルサイト１２で保持されるデータイメージの更新データをリモートサイト１４に通信する為に、リモートサイト１４にデータ通信ネットワーク１０７を通して結合されている。
【００３６】
かくして、リモートサイト１４はローカルサイト１２のミラーデータイメージを保持する。
【００３７】
リモートサイト１４は、各々リモートストレージシステム（セカンダリストレージ１，セカンダリストレージ２）１０５、１０６を備えている。各サイト１４は対応するホストプロセッサ（リモートホスト）１０２、１０３を含んでも含まなくても良い。
【００３８】
リモートストレージシステム１０５、１０６は、ローカルストレージシステム１０４のミラーデータイメージを保持する為に、ローカルストレージシステム１０４のストレージボリューム１４２と極めて類似したストレージボリュームを備えている事が好ましい。
【００３９】
リモートサイト１４は、ローカルストレージシステム１０４が計画的又は非計画的に停まった場合には、当該データにアクセスされるものの、ローカルサイト１２のデータが破壊されるような災害に見舞われた場合でも、サイトとデータを維持するためには、リモートサイト１４はローカルストレージシステム１０４から地理的に離れた場所に設置した方がより好ましいと言える。
【００４０】
リモートストレージシステム１０５，１０６はローカルストレージシステム１０４と実質的に同じである為に、ローカルストレージシステム１０４に関する議論はそのまま等しくリモートストレージシステム１０５，１０６に当てはまる。
【００４１】
ローカルサイト１２では、ホストプロセッサ１０１は、ネットワークインターフェース（Ｉ／Ｆ）１１１を経由してデータ通信ネットワーク１０７に結合し、入出力（Ｉ／Ｏ）バス１０８と入出力インターフェース１１０を経由してローカルストレージシステム１０４に結合している。
【００４２】
ローカルストレージシステム１０４はディスクコントローラ１４１を備え、ディスクコントローラ１４１は、入出力（Ｉ／Ｏ）バス１０８に結合する為の入出力インターフェース１３０とデータ通信ネットワーク１０７に結合する為のネットワークインターフェース１３１とを有する。
【００４３】
ローカルストレージシステム１０４は更に、ディスクユニット１４０で構成されるスト

レージボリューム１４２を備え、ディスクコントローラ１４１が入出力インターフェース１３７，１３８を経由してデータパス１３９によってストレージボリューム１４２に結合する。

【００４４】

ディスクコントローラ１４１自身は、中央処理ユニット（ＣＰＵ）１３３を有し、ＣＰＵ１３３は内部バス１３２経由でメモリ１３４に結合し、更に本デイスクコントローラ１４１の各種インターフェース（即ち、入出力インターフェース１３０，１３８及びネットワークインターフェース１３１等）に結合している。

【００４５】

メモリ１３４は、ホストプロセッサ１０１からのＩ／Ｏ要求に対応して、ストレージボリューム１４２へのリード、ライトデータを格納する為のキャッシュメモリ１３５を有する。メモリは更に、コントロール情報１３６の如きある種のデータ構造や情報を維持する。

【００４６】

ＣＰＵ１３３は、ストレージボリューム１４２に対する（例えばホストプロセッサ１０１により齎される）あらゆるデータイメージの変更をリモートストレージシステム１０５，１０６に送信する、リモートコピー処理を通常通りに実行する事が好ましい。

【００４７】

かくして、リモートストレージシステム１０５，１０６は、ローカルストレージシステム１０４で維持されるデータをミラーする。

【００４８】

要約すると、リモートコピー処理は以下の如く動作する。

【００４９】

ディスクコントローラ１４１は、ストレージボリューム１４２のデータ（データイメージ）に対するＩ／Ｏライト要求、即ち、追加、変更、削除、あるいはその他の更新を受け取ると、そのデータをストレージボリューム１４２に書き込む。ストレージボリューム１４２へのこの様な各書き込み又は少なくともデータイメージの一部分はリモートサイト１４にミラーされる為に、データメッセージが生成され、リモートコピーキュー（ここでは示されていない）に格納される。

【００５０】

リモートコピーキューは、ＣＰＵ１３３上で稼動しているリモートコピープロセスにより周期的に監視されている。一つ以上のデータメッセージを有するキューが検出されると、データメッセージは取り出されて、リモートストレージシステム１０５，１０６の各々に送信され、データメッセージ中のデータが書き込まれ、結果として、ローカルサイト１２のデータに対するデータイメージが更新されることになる。

【００５１】

データメッセージに関する情報は、ストレージシステム１０４、１０５、１０６でヒストリカル情報として保持され、ローカルストレージシステム１０４から送信済み、送信途上、又はリモートストレージシステム１０５，１０６での受信完了を示している。各ストレージシステムは、ヒストリカル情報をキュー構造で管理する事が好ましい。

【００５２】

図２はキュー構造を示している。図２で示す如く、デイスクコントローラ１４１は、ワーキングキュー１１０、ロールバックキュー１１１、中間キュー１１２、およびライトヒストリキュー１１３を有する。

【００５３】

ワーキングキュー、ロールバックキュー、中間キュー、及びライトヒストリキュー１１０～１１３は、リモートストレージシステム１０５（各々キュー１２０、１２１、１２２及び１２３）、１０６（各々キュー１３０、１３１、１３２及び１３３）でミラーされている。これらのキューは、先入れ先出し（ＦＩＦＯ）構成で実装される。

【００５４】

Ｉ／Ｏリードライト要求は典型的に、コマンドエントリとこれに伴う又は継続するデータ（ライト要求の場合）により構成される。

【００５５】

コマンドエントリは、データが書き込まれるストレージボリューム１４２上（ライト要求の場合）、又はデータが読み出される（リード要求の場合）ボリューム上のロケーションを特定し、その他実装によって必要とされる情報を特定する。

【００５６】

ホストプロセッサ１０１より、リモートストレージシステム１０５，１０６でミラーされているデータイメージの変更を伴うＩ／Ｏ要求を受けとった場合には、コマンドエントリには、シーケンス番号が割り当てされる。コマンドエントリはコマンドエントリと割り当てられたシーケンス番号で形成される事になる。

【００５７】

コマンドエントリは次いで、ワーキングキュー１１０に格納される。キューは、リモートストレージシステム１０５，１０６にデータを送信するに先立つ、データとデータメッセージに対するヒストリカル情報を構成する。

【００５８】

コマンドエントリがワーキングキュー１１０に滞在している間に、対応するライト要求が処理される。処理には、ライト要求に対応するデータ（要求と共に受信するか、要求に継続して受信するかは、ホストプロセッサ１０１とローカルストレージシステム１０４に用いられる通信プロトコルによる）を受信し、データの為にキャッシュメモリ１３５のエリアを割り当て、当該エリアに受信データを書き込む事が含まれる。

【００５９】

キャッシュメモリ１３５内のデータへのポインターは、対応するコマンドエントリに関連付けられる。Ｉ／Ｏ要求データが正常に受信されたか、エラーが検出されたかのステータスを示すステータスメッセージが、Ｉ／Ｏ要求の発行元に返される。

【００６０】

図２は、ディスクコントローラ１４１がシーケンス番号“１５”を割り当てられたライト要求を受領中の状態を示す。コマンドエントリは割り当てられたシーケンス番号とライト要求のコマンドエントリで形成される。

【００６１】

コマンドエントリは次いで、上述されたように更なる処理の為にワーキングキュー１１０に格納される。コマンドエントリは、Ｉ／Ｏ要求が正常に処理された場合には、ワーキングキュー１１０からロールバックキュー１１１に移される。

【００６２】

ロールバックキュー１１１は、後に十分に説明されるように、ローカルストレージシステム１０４とリモートストレージシステム１０５，１０６でのロールバック同期化の為に用いられる一時的な保存エリアである。

【００６３】

リモートストレージシステム１０５，１０６も基本的には同じ目的の為に、ロールバックキュー１１１を含む同一のキュー構造を有する。

【００６４】

ロールバックキュー１１１に保持されるコマンドエントリに対応するデータは、一つのストレージシステムで障害が検出された場合に廃棄処理をする場合、又は生存ストレージシステムの間での巡回処理をする場合を含めて、本発明での同期化処理に於いて使用される。

【００６５】

コマンドエントリは、ロールバックキュー１１１から関連データがストレージボリューム１４２への書き込み待ちか書き込み中であるエントリを保存する中間キュー１１２に移動される。

【００６６】

ローカルストレージボリューム１４２への書き込みが完了すると、コマンドエントリは、プライマリストレージシステム１０４をミラーするリモートストレージシステムの保存するデータイメージの更新のためにリモートストレージシステム１０５，１０６に送信される、データメッセージ形成のためのリモートコピー要求を形成するために使用される。

【００６７】

次いで、コマンドエントリからポインターが抽出され、エントリはライトヒストリキュー１１３に格納される。

【００６８】

上述したように、図２は、ホストプロセッサ１０１より受信しシーケンス番号１５が割り当てられ、従ってワーキングキュー１１０に格納される、Ｉ／Ｏ要求からのコマンドエントリを示す。

【００６９】

シーケンス番号１３と１４のコマンドエントリはこの時点でロールバックキュー１１１に保持され、関連データのストレージボリューム１４２への書き込み待ちの状態になっている。

【００７０】

中間キュー１１２は、シーケンス番号１０，１１、及び１２を持つコマンドエントリを保持している。シーケンス番号１０のコマンドエントリは書き込み中か、次の書き込み待ちの何れかである。シーケンス番号７，８及び９のコマンドエントリの関連データはストレージボリューム１４２に書き込み済みで、従ってライトヒストリキュー１１３に格納されている。

【００７１】

リモートストレージシステム１０５，１０６は実質的に同様なキュー構造を保持し、同様の方法で動作する。かくして、例えば、リモートストレージ１０５はシーケンス番号１０を割り当てられたコマンドエントリを持つデータメッセージを受信中であり、全データが受信中の間は、ワーキングキューに格納されている。

【００７２】

シーケンス番号１０はローカルストレージシステム１０４によって割り当てられたものである。ひとたび、データメッセージが受信されると、コマンドエントリは、ワーキングキューから、シーケンス番号６～９のコマンドエントリを現在保持していることを図２が示しているロールバックキュー１２１に移動される。

【００７３】

中間キュー１２２はシーケンス番号５が割り当てられたデータメッセージのコマンドエントリを有し、エントリはリモートストレージシステム１０５のストレージボリューム１４０に書き込み中である。

【００７４】

書き込みが完了すると、エントリはシーケンス番号１～４のデータメッセージのコマンドエントリと共に、ライトヒストリキュー１２３に移動される。ライトヒストリキュー１２３の深さによっては、最も早いエントリ、例えばシーケンス番号１のエントリは、シーケンス番号５のコマンドエントリを格納させるために、排出されても良い。

【００７５】

他のリモートストレージシステム１０６も同様なデータキュー（１３０、１３１、１３２及び１３３）を有する。図２では、リモートストレージ１０６は現在シーケンス番号１２に関するデータメッセージを受信中であり、コマンドエントリはワーキングキュー１３０に格納されている。

【００７６】

ロールバックキュー１３１は現在シーケンス番号８～１１の制御情報を保有している。ストレージシステムがヒストリカル情報を追跡するためのキューは、メモリ１３４のみならず、ストレージボリューム１４２にも保持される事が好ましい。

【００７７】

ローカルストレージシステム１０４は更に、メモリ１３４中にリモートコピーステータステーブル１１４を保持し、送信済みのシーケンス番号、リモートストレージシステム１０５、１０６にて受信され応答が得られたシーケンス番号を特定できるようになっている。

【００７８】

例えば、リモートストレージシステム１０５（テーブル１１４では“Ｓ１”で表示）で受信済みの最新のデータメッセージはシーケンス番号９で、リモートストレージシステム１０６（“Ｓ２”）で受信済みの最新のデータメッセージはシーケンス番号１１である事が判る。リモートコピーステータステーブル１１４は更に、リモートストレージシステムのロールバックキュー１２１，１３１とライトヒストリキュー１２３，１３３に関する情報も保持している。

【００７９】

かくして、リモートコピーステータステーブル１１４では、リモートストレージシステム１０５，１０６のロールバックキュー１２１，１３１は各々４データメッセージエントリ分の“長さ”を有し、１０ＭＢのデータを保持できる事がわかる。

【００８０】

リモートストレージシステム１０５，１０６各々のライトヒストリキュー１２３，１３３は５個のデータメッセージに５個のエントリを有する。ライトヒストリキュー１２３，１３３のサイズはバイト単位の容量又は他の形でリモートコピーステータステーブル１１４に表示されても良い。

【００８１】

図３及び図４は、本発明が解決しようとしている問題点を説明するダイアグラムである。

【００８２】

図３はプライマリストレージボリューム２０に対応して、少なくとも２つのセカンダリストレージボリューム２１，２２を備える、多重リモートコピー動作の為の構成を説明する。

【００８３】

２３，２４の矢印は、プライマリストレージボリューム２０とセカンダリストレージボリューム２１，２２との間のリモートコピーリンクを示す。

【００８４】

矢印２５が示す図の次の部分では、リモートコピーリンクの一つ２３が停止（suspend）している。リンクはネットワーク障害又はストレージボリューム障害で停止され得る。ここで用いられる“停止（suspend）”の用語は、原因の如何によらず、二つのユニットの間でのデータ転送が途絶していることを示す。

【００８５】

停止が発生すると、停止中のペアの各ストレージボリューム（プライマリストレージボリューム２０とセカンダリストレージボリューム２１）は自らのビットマップ情報を生成する。

【００８６】

これらのビットマップ情報は、図３ではＢｐ，Ｂｓで表示されている。このビットマップ情報は、停止時のストレージボリュームの状態を反映している。ビットマップ情報生成のための方法については、後に記すが、関連米国特許No. 10/042,376に記されている。

【００８７】

図の次の部分は、停止に加えて発生する三つの可能なシナリオを図で表示したものである。

【００８８】

第一のシナリオは矢印２７により図の左側に示され、第二のシナリオは矢印２８で示され、第三のシナリオは矢印２９で示されている。矢印２７で示され、Ａ－３のラベルで示される第一のシナリオでは、セカンダリストレージボリューム２１が障害になった場合で

ある。この場合は、矢印３０で示されるように、プライマリストレージボリューム２０とセカンダリストレージボリューム２２のペアが残っている。

【００８９】

図３の矢印２８で示される第二のシナリオでは、セカンダリストレージ２２が障害になった場合である。この状況に対する対応は、矢印３１とＡ－６のラベルで示される。この状況では、プライマリストレージ２０と稼動可能なセカンダリストレージ２１間のリンクが回復されるのを待つ。

【００９０】

リンクが回復すると、プライマリストレージボリューム２０とセカンダリストレージボリューム２１は２つのビットマップＢｐとＢｓの情報を用いて再同期化を行う事が出来る。

【００９１】

図３の右側部分に、矢印２９で第三の状態が示される。この状態では、プライマリストレージボリューム２０が障害になり、ビットマップＢｐはもはや使用不能である。この状態では、図３の右側下のＡ－８でラベルされる部分で示されるように、データの全コピー動作がセカンダリストレージボリューム（Ｓ１）２０とセカンダリストレージボリューム（Ｓ２）２１を再同期化するための唯一の手段である。

【００９２】

図４は、直列接続多重コピーを実現するプライマリとセカンダリストレージの異なった構成を示す図である。

【００９３】

図の最上段に示すとおり、プライマリストレージボリューム（Ｐ）４０は、セカンダリストレージボリューム（Ｓ１）４２，（Ｓ２）４４，（Ｓ３）４６との間でリモートコピー機能を実現する為に、直列的に構成されている。

【００９４】

矢印４１は、セカンダリストレージボリューム４２とセカンダリストレージボリューム４４と間のリモートコピーリンクが停止している状況を示している。この状態では、更に図示されているように、ビットマップＢｐとＢｓがセカンダリストレージボリューム４２と４４によって生成されている。この状況において、追加して発生する４種類の障害が存在する。これらは、図の下側の部分に矢印１，２，３及び４で示されている。

【００９５】

矢印１では、プライマリストレージボリューム４０が障害になっている場合を示している。この状況では、ケースＢ－４に示すように、構成は変更され、リモートコピーペアＳ１とＳ２及びＳ２とＳ３は同じままである。

【００９６】

ケース２では、停止ペアのメンバではないセカンダリストレージボリュームが障害になっている。この状況では、ケースＢ－６に示されるように、リモートコピーペアは不変のままで、構成のみが変更されている。

【００９７】

図示されているように、プライマリストレージボリューム４０はセカンダリストレージボリューム４２及び４４との関連して動作を継続する。

【００９８】

矢印３は停止ペアのメンバであるセカンダリストレージボリュームが障害になった場合を反映している。このケースでは、ビットマップＢｐからの変更情報は失われる。

【００９９】

この結果、Ｂ－８に示されているように、プライマリストレージボリューム４０とセカンダリストレージボリューム４４は再同期化不能である。プライマリストレージボリューム４０とセカンダリストレージボリュームのペアは、新しいデータのコピーを全て転送することによってのみ再同期が可能になる。

【０１００】

矢印４は最後のケースで、セカンダリストレージボリュームが障害になっている。前述した状況では、セカンダリストレージボリュームの障害でビットマップＢｓが損失して、変更情報が失われる結果となる。この状況では、ケースＢ－１０に示されるように、例えばセカンダリストレージボリューム４６をターゲットに全データの再コピーを実施することが唯一の解である。

【０１０１】

図３、図４に示したビットマップ喪失の状況は、本発明の技術により克服する事が出来る。このことについては次で説明する。

【０１０２】

図３，図４の幾つかの状況で説明したように、図３のＡ－３とＡ－７及び図４のＢ－７とＢ－９の場合には、ミラーペアのビットマップの一つが失われている事に注意が必要である。

【０１０３】

本発明の実施例によれば、ペアの停止を検出した場合、停止されたリモートコピーペアのメンバ外のストレージボリューム上でビットマップのコピーを生成する事により、この問題は克服される。離れたロケーションにビットマップをコピーすることにより、ビットマップの一つ（Ｂｐ又はＢｓ）が失われた場合でも、容易に再同期を実施することが出来る。図５に本発明の好ましい実施例を説明する。

【０１０４】

図５の上部で示す通り、プライマリストレージボリューム５０はセカンダリストレージボリューム５２，５４とペア状態になっている。矢印５６で示す図の次の部分では、プライマリストレージボリューム５０とセカンダリストレージボリューム５２の間での停止の発生を示している。

【０１０５】

ケースＡ－２’で示す通り、この停止の発生を契機に、プライマリストレージボリューム５０は異なるセカンダリストレージボリューム５４上にビットマップのコピーを生成する。セカンダリストレージボリューム５４は、プライマリストレージボリューム５０をミラーするボリュームとは異なるボリュームである。ビットマップＢｐ’は、ビットマップＢｐと必ずしも同一でなくても良い。セカンダリストレージボリューム５４でのビットマップコピーに対する最小限の要求は、その開始時間がオリジナルコピーＢｐの開始時間に比べて同じかより古い事である。

【０１０６】

ビットマップの開始時刻がより古い場合は、ビットマップコピーが示す全ての変更は、単に同じデータを同じアドレスに上書きする事になるだけである。本発明の親出願では、ストレージが停止発生時と同時刻からのビットマップコピーを生成する技術を開示しているが、本発明では、ストレージが同時刻のビットマップコピーを生成するのが不可能な場合でも、代わりにより古いビットマップコピーを用いる事により十分に満足できる解を得られることを述べている。

【０１０７】

図５のケースＡ－７で示す通り、プライマリビットマップＢｐが損失しても、セカンダリストレージボリューム５２と５４はプライマリビットマップＢｐのコピーＢｐ’を用いて再同期することが可能である。

【０１０８】

本再同期化については、図５の下側部分Ａ－８に示されている。

【０１０９】

図６は本発明のもう一つの好ましい実施例を示す。図６は図４と同様に、図の上部での障害と同じ状態を示す。

【０１１０】

Ｂ－２でリモートコピー動作が停止する。本状態の検出を契機に、ストレージシステムはビットマップＢｐのコピーをＢ－２’に示されるように他のボリュームに作る。

【０１１１】
図示されているケースでは、プライマリストレージボリューム６０はプライマリビットマップＢｐのコピーＢｐ’を受け取り、セカンダリストレージボリューム６６はビットマップＢｓのコピーＢｓ’を受け取る。
【０１１２】
かくして、Ｂ－７で示されるセカンダリストレージボリューム６２の障害では、ビットマップＢｐが失われても、ビットマップを用いて、再同期化が不能になることを避ける事が出来る。Ｂ－８に示す通り、セカンダリストレージボリューム６４はプライマリストレージボリューム６０に再同期化する事が可能である。
【０１１３】
もし停止がセカンダリストレージボリューム６２の代わりにセカンダリストレージボリューム６４の障害で発生しても、セカンダリストレージボリューム６６はセカンダリストレージボリューム６２に再同期化可能である。
【０１１４】
図７は本発明の好ましい一つの実現方法を説明するためのフローチャートである。同期型リモートコピーでは、各ホストＩ／Ｏ動作がリモートＩ／Ｏ動作を引き起こす。
【０１１５】
かくして、図７に記す処理は、ホストＩ／Ｏ毎に実行される。ローカルストレージは、図７の処理が完了するまでは、ホストに対するステータス報告を行わない。この準備動作を図７ａに示す。
【０１１６】
図７に、ローカルストレージシステム１０４（図１）がホストプロセッサ１０１（図１）から受け取った更新データをリモートストレージシステム１０５，１０６にコピーする為のリモートコピー動作の基本的ステップを説明する。
【０１１７】
ホストプロセッサ１０１からのローカルストレージシステム１０４で保存されているデータイメージを変更するＩ／Ｏライト要求は、リモートストレージシステム１０５，１０６で保存されているミラーデータイメージに対しても同様の変更を必要とする。
【０１１８】
Ｉ／Ｏライト要求は、割り当てられたシーケンス番号と要求に対するデータのポインターを備えたコマンドエントリを生成する。コマンドエントリは、全データが受信され、受領応答がホストプロセッサ１０１に返送されるまでの間は、ワーキングキュー１１０に格納されている。
【０１１９】
次いで、コマンドエントリはロールバックキュー１１１に移動される。ロールバックキュー１１１が満杯になるか、フラッシュコマンドが受信されると、コマンドエントリは中間キュー１１２に移動される。中間キューに滞在する間に、要求された対応データはストレージボリューム１４２に書き込まれる。
【０１２０】
ＣＰＵ１３３上で実行しているリモートコピー処理は周期的に中間キューの内容を調査して、各リモートストレージシステム１０５，１０６にコピーすべき更新データコマンドのエントリが存在しないかを判定する。
【０１２１】
次いで図７を参照すると、ローカルストレージシステム１０４は、各リモートストレージ１０５，１０６が次のデータメッセージを受信できるかをチェックする。チェックはＲＣステータステーブル１１４を参照して行う。ＲＣステータステーブル１１４により、各リモートストレージシステム１０５，１０６がどのデータメッセージを受信して応答済みか、又リモートストレージシステム１０５，１０６が保持する各種キューのサイズ情報等がローカルストレージシステム１０４に伝わる。
【０１２２】

これにより、ローカルストレージシステム１０４は、特定のリモートストレージシステムが別のデータメッセージや関連データを受信できる余裕があるか否かを判定する。もしなければ、ステップ５０１から抜け、プロセス６００１にて停止が発生しているかの判定が開始される。（あとで議論される。）

更に、リモートストレージシステム自身が、例えばＳＣＳＩの“ビジー（０ｘ０８）”又は“キュー満杯（０ｘ２８）”の何れかを用いて、更なるデータメッセージ受信の不可を返答する事も可能である。同期型リモートコピー動作の場合は、リモート（セカンダリ）ストレージシステムのキュー状態のチェックは不要である。

【０１２３】

もし対象のリモートストレージシステムにデータメッセージを受信するのに十分な余裕がある場合には、ローカルストレージシステム１０４はリモートコピー（ＲＣ）コマンドを、ステップ５０３にてデータが後続するデータメッセージの形で、リモートストレージシステム（たとえばリモートストレージシステム１０５）に発行する。

【０１２４】

次いで、ローカルストレージシステム１０４は、データメッセージが受信されたか否かを示すステータス報告の受信を待つ（ステップ５０４）。

【０１２５】

受信後には、ローカルストレージシステム１０４は、ステップ５０５にて、全てのリモートストレージシステムがデータメッセージでの更新を完了したかをチェックする。もし完了していなければ、処理はステップ５０７に移動し、ＲＣステータステーブル１１４を更新して、対象リモートストレージシステムはデータメッセージを受信した事を反映させて、次のリモートストレージシステムがデータメッセージを受信するためにステップ５０１に戻る。

【０１２６】

しかし、もしステップ５０５で全てのリモートストレージシステムがデータメッセージを受信した事が判明したら、データメッセージ（コマンドエントリ）は、ステップ５０６にてライトヒストリキューに移動され、ＲＣステータステーブル１１４はステップ５０７で更新され、このデータメッセージに対する処理は５０８で完了する。

【０１２７】

ステップ５０１の判定にて、リモートストレージがデータメッセージを受信できない場合には、処理フローはステップ６００１に移動し、停止状態が発生しているかを判定する。この動作での最初の判定は、リモートコピーペアのステータスについてである。リモートコピーペアの継続を妨げる何らかの異常事態が存在すれば、リモートコピーペアは停止しなければならないと判定される。

【０１２８】

本判定の結果として、停止が発生し、主要な処理はステップ６００２に示すビットマップのコピーを生成する事である。ビットマップを生成する方法は、以下に議論される。

【０１２９】

しかしながら、以下でも議論されるように、ひとたびビットマップが生成されると、制御フローはステップ５０８に移動し、多重リモートコピー処理を完了し、再同期処理へ移行する。本発明の親出願の技術を用いることにより、プライマリストレージは停止ポイントを検出し、本ポイントをシーケンス番号で示し、ビットマップのコピーを有する他のストレージに提供することができる。

【０１３０】

図８は、リモートストレージシステムがリモートコピー処理のデータメッセージを受信するためのステップを広範囲に説明するフローダイアグラムである。最初のステップ７００１は停止が発生しているかを判定する。この判定は、図７に関連して説明したものと同じ技術を用いて実行される。

【０１３１】

もし停止が発生していたら、図９に関連して以下に説明する技術を用いて、ビットマッ

プのコピーがステップ７００２で生成される。

【０１３２】

もし停止が発生していなければ、本処理フローはステップ６０１に移動する。ステップ６０１にて、リモートストレージシステムはデータメッセージを受信して、ステップ６０２にて、このデータメッセージのためのキュー資源の使用状態をチェックする。

【０１３３】

言い換えると、利用できる余裕があるかを判定する。ステップ６０２での判定は、他のリモートストレージシステムのキュー内容に比較して、各キュー（ロールバック、中間、及びライトヒストリキュー）の整合状態も考慮して判定される。

【０１３４】

もしステップ６０２での判定の結果、リモートストレージシステムがデータをこの時点では受領できない場合は、ステップ６０２を抜け、処理はステップ６０６に移動して、リモートストレージシステムは“ビジー”ステータスメッセージをローカルストレージシステムに返送し、受信処理より抜ける。

【０１３５】

次いで、ローカルストレージシステム１０４は、後で当該トランザクションを再度試行しなければならないと判定する。反対に、ステップ６０２にて、データが受信可能と判定された場合には、データはステップ６０３にて受信される。

【０１３６】

ステップ６０４にてリモートストレージシステムはデータ転送完了のステータスを返送して、ステップ６０５にて、データメッセージはメッセージ受信のために用いられたワーキングキューよりロールバックキューに移動される。ステップ６０７にてこの処理は完了する。

【０１３７】

図９は、例えば図７のステップ６００２又は図８のステップ７００２で実行される、ビットマップのコピーを生成する手順の好ましい実施例を説明するフローチャートである。

【０１３８】

図１０は、図９のフローチャートの説明で用いられる構成例を説明するダイアグラムである。

【０１３９】

図１０のストレージ７２は、ボリューム７２と７４で構成される停止中のリモートコピーペアのプライマリストレージである。

【０１４０】

停止ポイントからのビットマップを生成するのに用いることができる３つの異なった技術が存在する。停止ポイントからのビットマップを生成する一つの技術は、本発明の親出願で述べられている技術を用いることである。この技術は図９に要約されている。

【０１４１】

ストレージは、変更情報のすべてを反映する、ロールバックキューとライトヒストリキューを持っているため、多重リモートコピー環境での各種ストレージボリュームは同一時刻又は同一シーケンス番号からのビットマップを生成する事が出来る。

【０１４２】

図９のステップ８００１で示すとおり、ストレージは自分自身が停止中のリモートコピーペアのプライマリストレージボリュームかセカンダリストレージボリュームかをチェックする。

【０１４３】

図１０で示されるケースでは、ストレージボリューム７２が本ペアのプライマリストレージボリュームで、ストレージボリューム７４がセカンダリストレージボリュームである。言い換えれば、“プライマリ”及び“セカンダリ”の用語は、相対的な用語である。プライマリストレージボリューム７２が実行する処理は、ステップ８００２にて示す。

【０１４４】

ステップ８００３にて、ストレージボリューム７２は、時刻又はリモートコピーコマンドからのシーケンス番号で表示される停止ポイントを検出する。これを契機に、ストレージボリューム７２は、“ビットマップコピー生成コマンド”をプライマリストレージボリューム７０
に対して発行する。

【０１４５】

次のステップ８００４で示す通り、停止中のリモートコピーペアのプライマリストレージボリューム７２が例えば図１０のストレージボリューム７８のような、もう一つのセカンダリストレージを持っている場合も、同様の生成ステップが実行される。

【０１４６】

かくして、ステップ８００５で示す通り、ビットマップはセカンダリストレージボリューム７８にも生成される。

【０１４７】

ストレージボリューム７４で実行される処理は、既に述べられたものと同様である。停止中のリモートコピーペアのセカンダリストレージが、ストレージボリューム７６、８０のように、一つ以上のセカンダリストレージ自身を有する場合には、ストレージ７４上のセカンドビットマップのコピーは、ストレージボリューム７６、８０の何れかに作成される。この処理はステップ８００７に示される。

【０１４８】

図１１は本発明の他の実施例を示す。図１１に於いて、行は各種ストレージを示し、図の番号のついたブロックは各種データ転送動作を示す。

【０１４９】

例えば、図１１の最上段は１４個のデータがホストシステムからリモートシステムに転送された事を示す。このデータが受け取られると、セカンダリストレージボリュームＳ１，Ｓ２，及びＳ３に転送される。

【０１５０】

データブロック１の開始時刻の相違は、データを一つのロケーションから他のロケーションに転送するのに必然的に発生する処理及び転送遅延を反映している。

【０１５１】

図１１では、動作１０００で、データブロック４をセカンダリストレージボリュームＳ１からＳ２に送信する間に、停止が発生した事を想定している。これを時刻ｔ０とする。実際、第四番目のリモートコピー要求は失敗している。

【０１５２】

結果として、少なくともビットマップＢｐとＢｐ’は変更情報を含まなければならない。

【０１５３】

時刻ｔ１でセカンダリストレージボリュームＳ１は動作１００１に示すように、ビットマップのコピーを持っているプライマリストレージボリュームＰに“停止検出”メッセージを発行する。

【０１５４】

時刻ｔ２にて、プライマリストレージボリュームＰは“停止検出”メッセージを受信し、リモートコピー動作をストップする。

【０１５５】

次いで、プライマリストレージボリュームＰは、応答１００２を要求の送信者に送信する。プライマリストレージボリュームＰが停止発生以後の変更を記録できれば、リモートコピー動作の凍結中にも、プライマリストレージボリュームＰはホスト結合を維持する。

【０１５６】

もしプライマリストレージボリュームＰがこの情報の維持が不可能なら、ビットマップのコピーが生成される迄、プライマリストレージボリュームＰはホストからディスコネクトしなければならない。説明の実施例では、この変更情報はキャッシュメモリ又はディス

クに格納される。

【０１５７】

時刻ｔ３で、セカンダリストレージボリュームＳ１はプライマリストレージボリュームＰからの応答を受信する。セカンダリストレージボリュームＳ１は変更情報と共に“ビットマップ生成”メッセージを送信する（動作１００３で示す）。変更情報は、停止ポイントｔ０から“応答”受信ポイント（時刻ｔ３）までカバーする。

【０１５８】

時刻ｔ４で、“ビットマップ生成”メッセージが変更情報と共に受信される。これを契機に、プライマリストレージボリュームＰは、ビットマップＢｐに対応するビットマップのコピーＢｐ’の生成を開始する。

【０１５９】

この処理が終了次第、プライマリストレージボリュームＰとセカンダリストレージボリュームＳ１との間のリモートコピー動作を再開する。これは動作１００４として示される。

【０１６０】

時刻ｔ５にて、セカンダリストレージボリュームＳ１は同様にリモートコピー動作を開始する。

【０１６１】

ビットマップＢｓのコピー生成については、動作１００５に関連して、後に示す。停止が１００５で（又はより早く）発生したとすると、セカンダリストレージボリュームＳ１とＳ２のみが、停止を認識している。リモートコピー要求Ｎｏ．４は失敗し、セカンダリストレージボリュームＳ２には届かない。

【０１６２】

かくして、少なくともビットマップＢｓとＢｓ’はリモートコピー要求Ｎｏ．３の処理後の変更情報を含まなければならない。従って、セカンダリストレージボリュームＳ２は停止中のペアのビットマップＢｓの生成を開始する。

【０１６３】

時刻ｔ２にて、セカンダリストレージボリュームＳ３は“停止検出”メッセージを受信して、これを契機に、ビットマップのコピーＢｓ’の生成を行う。セカンダリストレージボリュームＳ３は、その後、要求の送信者（セカンダリストレージボリュームＳ２）に対して、“応答”を送信する。これは、図１１の動作１００６に示されている。

【０１６４】

セカンダリストレージボリュームＳ２が“応答”を受信すると、セカンダリストレージボリュームＳ２とＳ３との間のリモートコピーリンクを再開する。

【０１６５】

この結果、セカンダリストレージボリュームＳ２は、例えば図１１に示すような要求２０～２５の実行等の他の用途に用いられる事が可能となる。

【０１６６】

図１１に関連して述べられた動作を要約すると、その動作の共通の手順は；

（１）第一のストレージボリュームで停止を検出して、ビットマップを生成する。

【０１６７】

（２）本ビットマップのコピーを有するべき第ニのストレージボリュームを選択する。

【０１６８】

（３）第一と第二のストレージボリュームとの間のリモートコピー動作を凍結する。

【０１６９】

もしこれら２つのストレージボリュームの間のリモートコピー動作が第二のストレージボリュームから第一の方向であれば、第二のストレージボリュームが、第一と第二の間のリモートコピー動作を凍結するために、第一ストレージボリュームに“凍結”メッセージを送信しなければならない。

【０１７０】

（４）第一のストレージボリュームでの変更情報（停止ポイントからの変更を含む）を第二のストレージボリュームに送信する。

【０１７１】

（５）第二のストレージボリュームでビットマップのコピーを生成する。

【０１７２】

（６）第一と第二のストレージボリューム間のリモートコピー動作を再開する。

【０１７３】

図１２では、ローカルストレージシステム（又はデージーチェーン構成での中間的サイトになっている場合のリモートストレージシステム）の障害が検出された場合の同期化処理で行われるステップを示す。ステップ７０１で障害が検出される。障害検出は多数の方法でなされる。

【０１７４】

例えば、ストレージシステム同士はハートビートメッセージを交換しあって、もし未検出なら障害を示す。又は、ストレージシステムは自ら障害を検し、しかもその障害をシステムの他の要素に対して報告出来ることもある。又は他の全ての慣用的障害検出技術が適用されても良い。

【０１７５】

障害が検出されると、残りのストレージシステム、即ち図２の構成に示すストレージシステム１０５，１０６などは、互いに交信しあって、新しいリモートコピーマネージャを互選する。

【０１７６】

選出されたストレージシステムは、残り全てのストレージシステムのデータイメージが同一状態に同期化されるように、リモートコピー動作を制御する。あるいは、リモートコピーマネージャは、システム管理者により事前に指定されても良い。

【０１７７】

しかしながら、リモートコピーマネージャが決まると、選択されたマネージャはローカルストレージシステムとなり、スッテプ７０３にて、残りのストレージシステムに維持される各種キューの内容と構造に関する情報収集を行う。この情報には、ロールバックキューとライトヒストリキューのエントリ範囲（エントリ数）即ち、データメッセージを含むエントリ数等、が含まれる。

【０１７８】

例えば、図２において、ローカルストレージシステム１０４が障害になり、ステップ７０２にて、リモートストレージシステム１０５が選択又は他の判定によりリモートコピーマネージャになると、リモートストレージシステム１０６は、ロールバックキュー１３１内にデータメッセージ８、９、１０、及び１１がありライトヒストリキュー１３３内にデータメッセージ３、４、５、及び６があることを報告する。

【０１７９】

障害検出時には、リモートストレージシステム１０５は、好都合にも、中間キューの内容を空にしてストレージボリュームに対するデータ書き込みを維持できるため、リモートストレージシステム１０６の中間キューの内容は、短時間の間にライトヒストリキューに追加される。

【０１８０】

ステップ１１０１にて、次にペアは停止中か否か判定される。もし、停止中でなければ、以下に示す通り、処理はステップ７０４に移る。

【０１８１】

これに対して、停止中の場合は、停止中のペアは再同期化されなければならない。これはステップ１１０２に示され、処理は図１３で示され、そこで説明される。

【０１８２】

残りのストレージシステムについてのリモートコピー環境に関する情報を受領し、その後一つに障害が発生した場合、選択されたリモートコピーマネージャは、ロールバック処

理を行うかロールフォワード処理を行うかの決定を行う。

【０１８３】

典型的には、この決定は、リモートコピー環境の管理者かユーザによって、障害時での同期化処理を最善に行う方法として、フラグをセットしておくことにより、事前に決定されている。

【０１８４】

ステップ７０４にて、ロールバック処理が用いられることを決定した場合は、ステップ７０５にて、リモートコピーマネージャは、全てのストレージシステムが共有している番号中で、最大のシーケンス番号を持つデータメッセージを決定する。

【０１８５】

例えば、図２の例では、シーケンス番号９を持つデータメッセージが条件に該当する。それ故に、ステップ７０６にて、選択されたリモートコピーマネージャは、ロールバックコマンドを他のすべてのストレージシステムに対して発行し、受領時にシーケンス番号９を超えるデータメッセージを破棄させる。

【０１８６】

かくして、再び図２に戻って、リモートストレージシステム１０６は、ロールバックコマンドを受領すると、シーケンス番号１０と１１のデータメッセージを破棄して、本処理はステップ７１５で完結する。

【０１８７】

もしこれに対して、ステップ７０４にて、ロールフォワードが決定されたなら、ステップ７０７にて、ロールフォワードシーケンス番号が決定される。

【０１８８】

本処理は、リモートコピー処理が各種のロールバックキューとライトヒストリキューの内容を比較して、最新のデータメッセージ番号、もしある場合、を有するストレージシステムを決定する事よりなされる。

【０１８９】

かくして、図２にて、選出されたリモートコピーマネージャがストレージシステム１０５の場合は、リモートストレージシステム１０６がシーケンス番号１０と１１を有するデータメッセージを持っており、自分自身は持っていない事を悟る。したがって、データメッセージ１０と１１をリモートストレージシステム１０６からコピーさせることによって、各システムで維持されていたデータイメージが同期化される。

【０１９０】

かくして、ステップ７０８にて、選出されたリモートコピーマネージャが最新のデータストレージシステムを決定すれば、ステップ７０８より７０９に移り、選択されたＲＣマネージャは、最新データメッセージを有するストレージシステムから、更新データを受け取る。これは選択されたＲＣマネージャ自身であることも、又他のストレージシステムのひとつである場合もある。

【０１９１】

いずれにしても、選択されたＲＣマネージャは更新データを受け取り、ステップ７１０にて、更新を必要としている他のストレージシステムに更新データを選択的に又は部分的に転送して、全てのリモートストレージシステムのデータイメージの同期化処理を行う。処理はステップ７１５で完結する。

【０１９２】

反対に、ステップ７０８にて、同期化処理は最新の更新データメッセージを持つストレージシステムが実施すべきと判定され、そのストレージシステムが選択されたリモートコピーマネージャではないなら（又は他のストレージシステムが同期化システムと前もって決定されたら）、ＲＣマネージャは、ストレージシステムが必要としている更新情報を比較して、最新の更新データを持つストレージシステムに対して、更新情報を送付する。

【０１９３】

次いで、ステップ７１２にて、データメッセージ形式の更新データが、データイメージ

の同期化を必要としている全てのストレージシステムに送られ、本処理はステップ７１５で完結する。

【０１９４】

もしロールバック処理が選択される場合は、本発明の親出願で議論されたように、例えば、リモートストレージシステム１０５，１０６（図２）のロールバックキュー１２１，１３１は整合されている事が好ましい。

【０１９５】

整合処理は、ローカルストレージシステム１０４にて、ＣＰＵ１３３上で走行するリモートコピー処理とリモートコピーステータステーブル１１４（リモートストレージシステム１０５，１０６が維持している各種キューの内容の情報が提供されている）を用いて実行される。

【０１９６】

例えば、あるリモートストレージシステムはローカルストレージシステムからのリモートコピーデータメッセージを受信不能で、他のリモートストレージシステムは可能と言う事態が発生する。

【０１９７】

この様な場合は、同期処理が必要なときには確実に実行されるように、他のリモートストレージシステムと比較時に、ひとつのリモートストレージシステムのキューの中のデータメッセージ間に少なくとも一個の共通データメッセージが存在することが必要であることを、同期化の目的のためには、考慮しなければならない。

【０１９８】

図１３は再同期化処理の詳細を説明するフローダイアグラムである。最初に、ビットマップのコピーを有する各ストレージボリュームは、リモートコピーペアのステータスをチェックし、ステータスに関する情報を交換する。これはステップ１２０１に示す。ビットマップのコピーを用いて再同期化を必要とする環境では、再同期化処理は停止状態にあるペア間で必要であるのみならず、ビットマップコピー即ち、Ｂｐ，Ｂｐ’，Ｂｓ，Ｂｓ’を有するストレージボリュームのペアの組合せでも必要である。

【０１９９】

ステップ１２０２にて、ビットマップコピーを用いた再同期化が不必要と判定されたなら、処理は終了する。一方、再同期化処理が必要なら、次いでビットマップが交換され（１２０３）、マージされ（１２０４）、ステップ１２０５で示されるように、マージしたビットマップを用いた再同期化処理が実行される。

【０２００】

以上で説明した方法と装置により、多重リモートコピー環境でのコピーを用いた再同期化が可能になる。例えば、リモートコピーの停止処理を採用したシステムのケースでは、特に、再同期化が必要になり次第、システム管理者の介入なしに、自動的に実行される。

【０２０１】

本発明の好ましい実施例について説明してきたが、本発明の範囲を逸脱する事なく、多数の改変が可能である事に注意願いたい。

【図面の簡単な説明】

【０２０２】

【図１】は、本発明の一実施例を実装するローカル及び複数のリモートストレージシステムを含むデータ処理システムを示すブロックダイアグラムである。

【図２】は、ローカル及びリモートストレージシステムの各々で実装され、ローカルストレージシステムによってリモートストレージシステムへ転送されるデータ更新の履歴情報を保持する為のキューの構造を示すダイアグラムである。

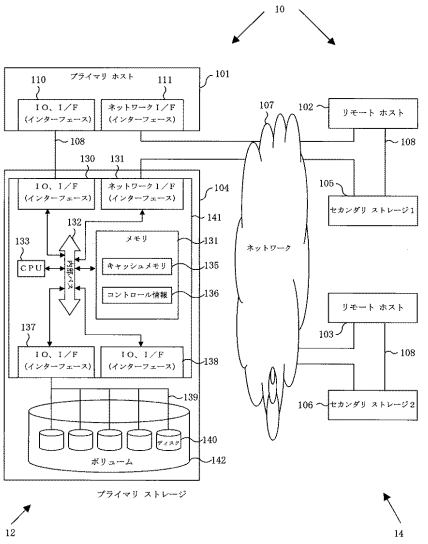
【図３】は、多重リモートコピー環境で停止を引き起こす各種の障害形態を示すダイアグラムである。

【図４】は、異なった多重リモートコピーアーキテクチャにおいて停止を引き起こす各種の障害形態を示すダイアグラムである。

【図５】は、本発明の方法の一実地例を示すダイアグラムである。
【図６】は、本発明の他の実地例を示すダイアグラムである。
【図７】は、停止発生時の多重リモートコピープロセスを示すフローチャートである。
【図７ａ】は、図７に関連する準備用ステップを示すフローチャートである。
【図８】は、停止発生時の他の多重リモートコピープロセスを示すフローチャートである。
【図９】は、停止発生時のビットマップのコピーを生成する為の好ましい方法を示すフローチャートである。
【図１０】は、図９の説明に用いられるダイアグラムである。
【図１１】は、ビットマップのコピーを生成するための動作の詳細シーケンスを示す。
【図１２】は、再同期化の一方法を示すフローチャートである。
【図１３】は、再同期化の他の方法を示すフローチャートである。
【符号の説明】
【０２０３】
１０１・・・プライマリ ホスト（ホストプロセッサ）
１０２・・・リモートホスト、
１０３・・・リモートホスト
１０４・・・ローカルストレージシステム
１０５・・・リモートストレージシステム
１０６・・・リモートストレージシステム
１０７・・・データ通信ネットワーク
１１０，１３０，１３７，１３８・・・ＩＯ　Ｉ／Ｆ（インターフェース）
１１１，１３１・・・ネットワークＩ／Ｆ（インターフェース）
１３２・・・内部バス
１３３・・・ＣＰＵ
１３４・・・メモリ
１３５・・・キャッシュメモリ
１３６・・・コントロール情報
１４０・・・ディスクコントローラ
１４２・・・ストレージボリューム

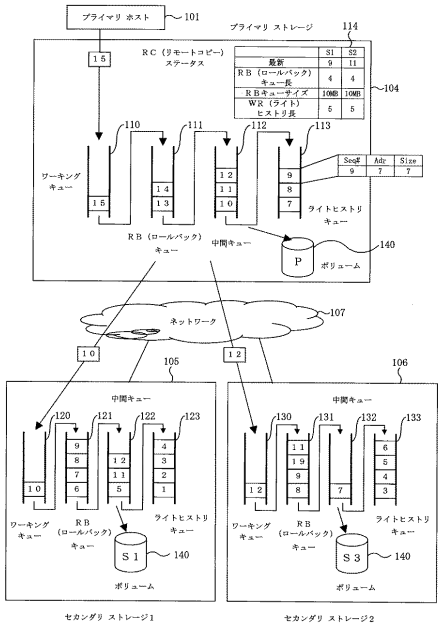
【図１】

図１



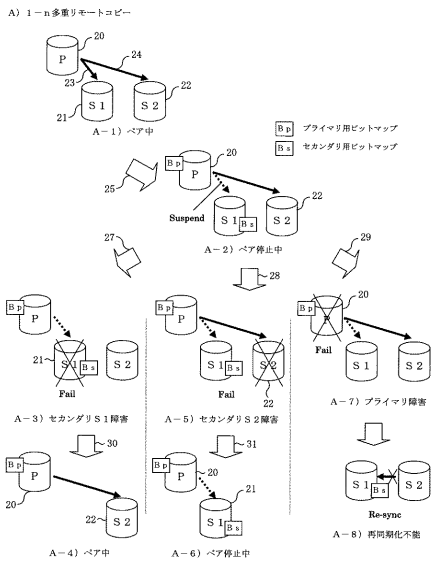
【図２】

図２



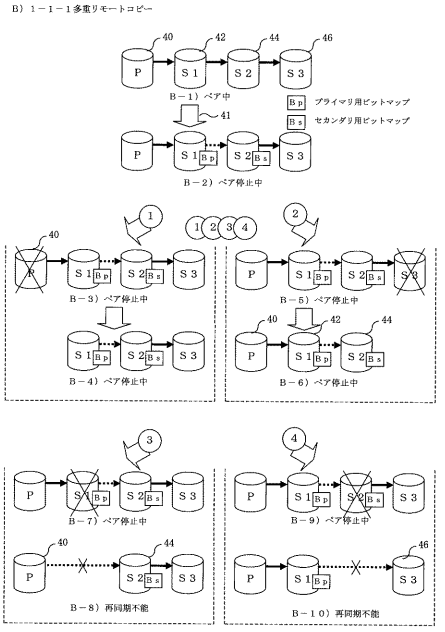
【図３】

図３



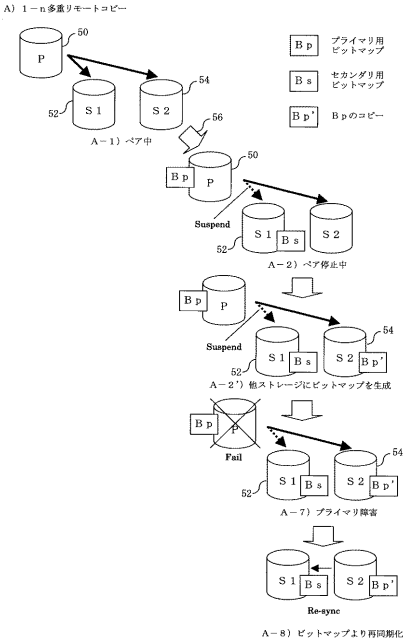
【図４】

図４



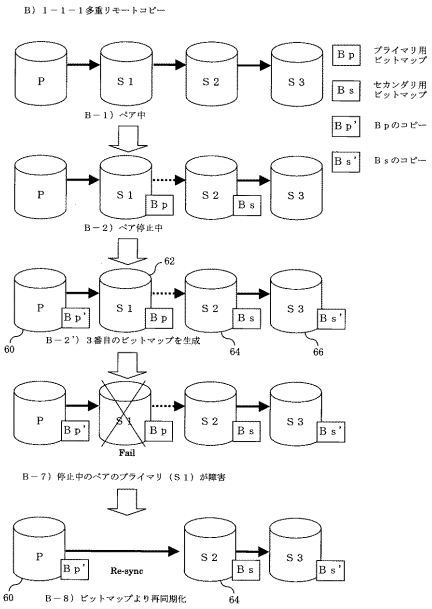
【図５】

図５



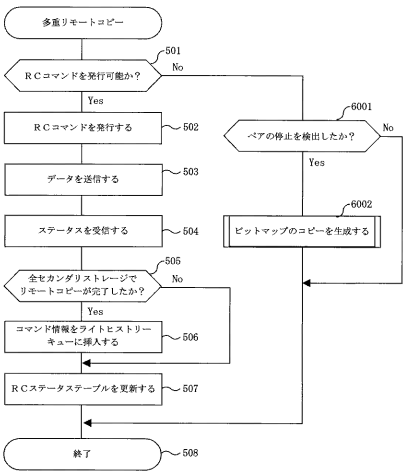
【図６】

図６



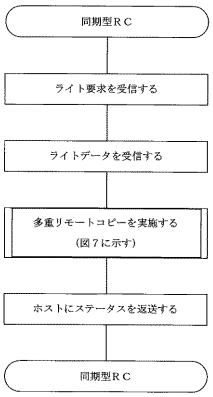
【図７】

図７



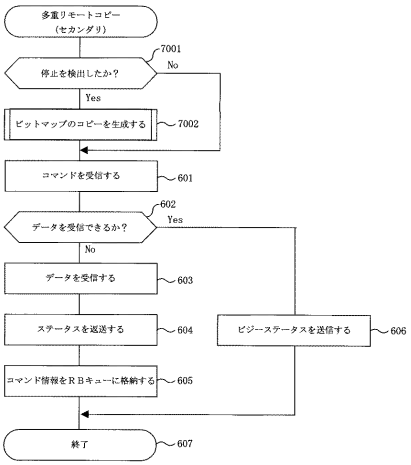
【図７ａ】

図７ａ



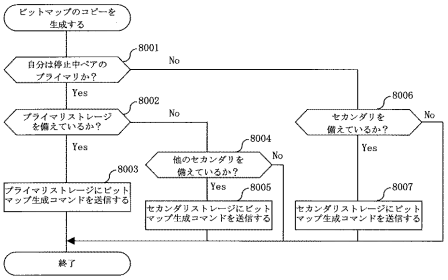
【図８】

図８



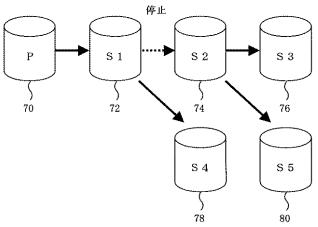
【図９】

図９



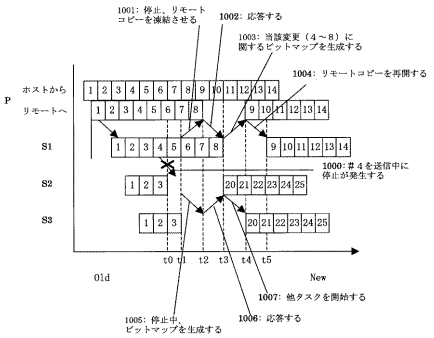
【図１０】

図１０



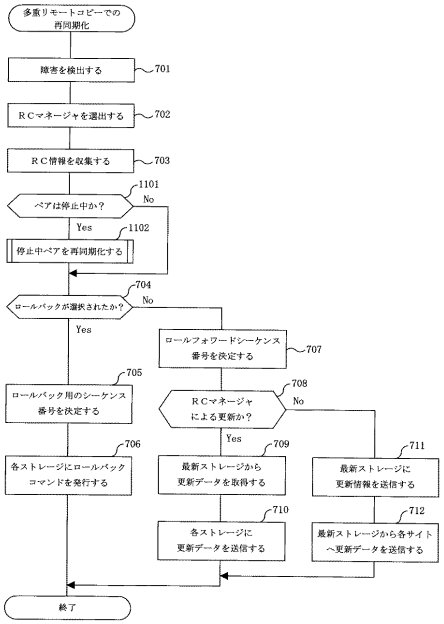
【図１１】

図１１



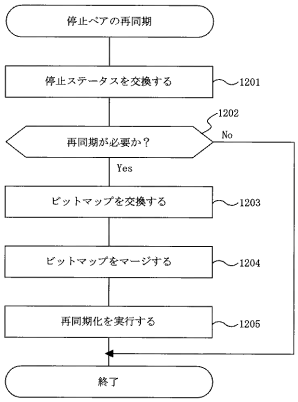
【図１２】

図１２



【図１３】

図１３


停止ペアの再同期
停止ステータスを交換する
1201
1202
再同期が必要か？
No
Yes
ビットマップを交換する
1203
ビットマップをマージする
1204
再同期化を実行する
1205
終了

フロントページの続き

(56)参考文献 特開平１０－１３３９２６（ＪＰ，Ａ）
特開２００１－２９０６８７（ＪＰ，Ａ）
特開２０００－３０５８５６（ＪＰ，Ａ）
特開平１１－３３８６４７（ＪＰ，Ａ）
特開２００３－２６３２８０（ＪＰ，Ａ）
特開２００３－１２２５０９（ＪＰ，Ａ）
特開２００３－１３１９１７（ＪＰ，Ａ）

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