

【公報種別】特許法第 17 条の 2 の規定による補正の掲載

【部門区分】第 6 部門第 3 区分

【発行日】平成 18 年 12 月 14 日 (2006.12.14)

【公開番号】特開 2004-272884 (P2004-272884A)

【公開日】平成 16 年 9 月 30 日 (2004.9.30)

【年通号数】公開・登録公報 2004-038

【出願番号】特願 2003-428627 (P2003-428627)

【国際特許分類】

G 0 6 F 3/06 (2006.01)

G 0 6 F 12/00 (2006.01)

【F I】

G 0 6 F 3/06 3 0 4 F

G 0 6 F 3/06 3 0 1 Z

G 0 6 F 12/00 5 3 1 D

【手続補正書】

【提出日】平成 18 年 10 月 27 日 (2006.10.27)

【手続補正 1】

【補正対象書類名】特許請求の範囲

【補正対象項目名】全文

【補正方法】変更

【補正の内容】

【特許請求の範囲】

【請求項 1】

データを保存するストレージボリュームを備えるローカルストレージシステムと、
前記ローカルストレージシステム内のストレージボリュームのミラー動作が可能であり、
データを保存するストレージボリュームを備える少なくとも 2 台のリモートストレージ
システムと、

前記ローカルストレージシステムと前記少なくとも 2 台のリモートストレージシステム
間の通信の途絶を検出する検出部と、

前記検出部の通信の途絶の検出に応答して、前記通信の途絶までの前記ローカルストレ
ージシステム内のストレージボリューム及び前記通信の途絶を検出したリモートストレ
ージシステム内のストレージボリュームになされた変更記録を生成し、その変更記録のコピ
ーを前記ローカルストレージシステム及び前記少なくとも 2 台のリモートストレージシス
テム内の他のストレージボリュームに提供する制御部と、

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】

前記制御部により提供される変更記録のコピーは、前記他のストレージボリュームに生
成されることを特徴とする請求項 1 記載のストレージシステム。

【請求項 3】

前記通信の途絶を検出したリモートストレージシステム内のストレージボリュームを、
前記他のストレージボリュームに同期化させるデータ同期化部をさらに備える

ことを特徴とする請求項 2 記載のストレージシステム。

【請求項 4】

前記制御部は、前記変更記録を、他のローカルストレージシステム内のストレージボリ
ュームにコピーできる場合は前記他のローカルストレージシステム内のストレージボリ
ュームにコピーし、前記他のローカルストレージシステム内のストレージボリュームにコピ
ーできない場合は前記リモートストレージシステム内のストレージボリュームにコピーす
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】

第一のストレージボリュームが第二のストレージボリュームでミラーされる直列接続されたストレージボリュームと、

前記第一のストレージボリュームと前記第二のストレージボリューム間の通信の途絶を検出する検出部と、

前記検出部による前記通信の途絶の検出に応答して、前記第一のストレージボリューム及び前記第二のストレージボリュームになされた変更記録を生成し、その変更記録のコピーを前記第一のストレージボリューム及び前記第二のストレージボリュームにそれぞれ直列接続されたストレージボリュームにそれぞれ提供する制御部と、

を備えることを特徴とするストレージシステム。

【請求項 6】

前記変更記録は、前記ストレージボリュームのうちで新たな直列接続となる新ペアを同期させるのに用いられることを特徴とする請求項 5 に記載のストレージシステム。

【請求項 7】

第一のストレージボリュームを備えるプライマリストレージシステムと、第二のストレージボリュームと第三のストレージボリュームを備えるセカンダリストレージシステムとがデータ通信可能に接続されたストレージシステムにおけるストレージシステム内に保存されたデータを同期させるデータ同期化方法において、

前記プライマリストレージシステム内の第一のストレージボリュームから前記セカンダリストレージ内の第二のストレージボリュームへの通信の途絶を検出するステップと、

前記途絶を検出すると、前記プライマリストレージシステムでは前記第一のストレージボリュームに保存されたデータの記録を提供するとともに前記セカンダリストレージシステムでは前記第二のストレージボリュームに保存されたデータの記録を提供するステップと、

前記第一のストレージボリュームに保存されたデータの記録の少なくとも一部のコピーを前記第三のストレージボリュームに生成するステップと、

前記第三のストレージボリュームに生成された前記少なくとも一部のコピーと前記第二のストレージボリュームに保存されたデータとを用いて、前記セカンダリストレージシステム内の第二のストレージボリュームを前記第三のストレージボリュームに同期させるステップと、

を備えることを特徴とするデータ同期化方法。

【請求項 8】

前記少なくとも一部のコピーは、前記プライマリストレージシステム内の第一のストレージボリュームに保存されたデータの記録のコピーであることを特徴とする請求項 7 に記載のデータ同期化方法。

【請求項 9】

前記第一のストレージボリューム及び前記第二のストレージボリュームに保存されるデータの記録は、ビットマップを構成することを特徴とする請求項 7 に記載のデータ同期化方法。

【請求項 10】

前記データを同期化させるステップは、

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間でビットマップを交換するステップと、

前記交換されたビットマップを用いて、前記セカンダリストレージシステム内の第二のストレージボリュームを前記第三のストレージボリュームに同期させるライト動作のセットを決定するステップと、

前記セットされたライト動作を実行するステップと、

を備えることを特徴とする請求項 7 に記載のデータ同期化方法。

【請求項 11】

前記記録を提供するステップは、

前記記録の対象となるストレージボリュームがプライマリストレージシステム内のストレージボリュームであるかを決定するステップと、

前記決定するステップでプライマリストレージシステム内のストレージボリュームであると決定された場合に、前記プライマリストレージシステム内の他のストレージボリュームが存在するときには前記他のストレージボリュームに前記記録を生成するステップと、

前記決定するステップでプライマリストレージシステム内のストレージボリュームであると決定された場合に、前記プライマリストレージシステム内の他のストレージボリュームは存在しないが前記セカンダリストレージシステム内の他のストレージボリュームが存在するときには前記セカンダリストレージシステム内の他のストレージボリュームに前記記録を生成するステップと、

前記決定するステップでプライマリストレージシステム内のストレージボリュームであると決定されなかった場合に、前記セカンダリストレージシステム内の他のストレージボリュームに前記記録を生成するステップと、

を備えることを特徴とする請求項 7 に記載のデータ同期化方法。

【請求項 12】

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から前記第三のストレージボリューム間のデータ通信の途絶を検出するステップと、

前記途絶を検出すると、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保存されたデータの第一の記録を提供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保存されたデータの第二の記録を提供するステップと、

前記第一の記録の少なくとも一部を前記第一のストレージボリュームにコピーするとともに前記第二の記録の少なくとも一部を前記第四のストレージボリュームにコピーするステップと、

前記コピーされた少なくとも一部の記録を用いて、前記第二と前記第三のストレージボリュームの少なくとも一つを前記第一と前記第四のストレージボリュームの少なくとも一つに同期させるステップと、

を備えることを特徴とするデータ同期化方法。

【請求項 13】

前記第一のストレージボリュームは、前記第三のストレージボリュームに同期化されることを特徴とする請求項 12 に記載のデータ同期化方法。

【請求項 14】

前記第一のストレージボリュームは、前記第四のストレージボリュームに同期化されることを特徴とする請求項 12 に記載のデータ同期化方法。

【請求項 15】

第一のストレージボリュームを備えるプライマリストレージシステムと、第二のストレージボリュームと第三のストレージボリュームを備えるセカンダリストレージシステムとがデータ通信可能に接続されたストレージシステムにおけるストレージシステム内に保存されたデータを同期させるデータ同期方法において、

前記プライマリストレージシステム内の第一のストレージボリュームから前記セカンダリストレージ内の第二のストレージボリュームへの通信の途絶を検出するステップと、

前記途絶を検出すると、前記プライマリストレージシステムでは前記第一のストレージボリュームに保存されたデータの記録を提供するとともに前記セカンダリストレージシステムでは前記第二のストレージボリュームに保存されたデータの記録を提供するステップと、

前記第一のストレージボリュームに保存されたデータの記録のコピーを前記第三のストレージボリュームに生成するステップと、

前記第三のストレージボリュームに生成された前記記録のコピーを用いて、前記プライマリストレージシステム内の第一のストレージボリュームを前記第三のストレージボリュームに同期させるステップと、
を備えることを特徴とするデータ同期化方法。

【請求項 16】

第一ストレージシステム内の第一ストレージボリュームと第二ストレージシステム内の第二ストレージボリュームと第三ストレージシステム内の第三ストレージボリュームとを用いるとともに前記第一ストレージボリュームと前記第二ストレージボリュームで第一のミラーペアを構成し、前記第一ストレージボリュームと前記第三ストレージボリュームで第二のミラーペアを構成してリモートコピーを行うリモートコピー方法において、

前記第一のミラーペアがリモートコピー停止中の変更記録を前記第一ストレージシステムが前記第一ストレージボリュームに生成、保存するステップと、

前記第一ストレージシステムが前記変更記録を前記第三ストレージボリュームに転送するステップと、

前記第三ストレージシステムが前記変更記録を保存するステップと、
を備えることを特徴とするリモートコピー方法。

【請求項 17】

前記第一ストレージシステムが前記第一のミラーペアのリモートコピー停止中に動作不能の場合、前記第三ストレージシステムに保存された変更記録を用いて、前記第二ストレージボリュームを前記第三ストレージボリュームに再同期化するステップをさらに備えることを特徴とする請求項 16 に記載のリモートコピー方法。

【手続補正 2】

【補正対象書類名】明細書

【補正対象項目名】全文

【補正方法】変更

【補正の内容】

【発明の詳細な説明】

【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

【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。

【 0 0 0 6 】

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。

【 0 0 0 7 】

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 Apparatus 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。

【 0 0 0 8 】

【特許文献1】米国特許5 9 3 3 6 5 3号公報

【発明の開示】

【発明が解決しようとする課題】

【 0 0 0 9 】

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ー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。

【 0 0 1 0 】

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。

【 0 0 1 1 】

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。

【 0 0 1 2 】

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なものである。

【 0 0 1 3 】

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、キャッシュオーバーフロー、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。

【 0 0 1 4 】

リモートコピー動作での再同期化の一例は、“Method and Apparatus 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。

【 0 0 1 5 】

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。

【 0 0 1 6 】

ある種のより複雑なシステム停止、例えば、リンク障害、キャッシュオーバーフロー、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、システムの再同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。

【 0 0 1 7 】

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。

【 0 0 2 9 】

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。

【 0 0 3 0 】

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 】

【 0 0 3 1 】

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。

【 0 0 3 2 】

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 】

【 0 0 3 3 】

図 1 は、プライマリ又はローカルサイト 1 2 と複数のセカンダリ又はリモートサイト 1 4 がデータ通信ネットワーク 1 0 7 で通信的に結合された、データ処理システム 1 0 を構成するストレージシステムを示す。

【 0 0 3 4 】

ローカルサイト 1 2 はホストプロセッサ 1 0 1 とローカルストレージシステム 1 0 4 を備えている。

【 0 0 3 5 】

ホストプロセッサ 1 0 1 とストレージシステム 1 0 4 は、ローカルサイト 1 2 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 1 4 に通信する為に、リモートサイト 1 4 にデータ通信ネットワーク 1 0 7 を通して結合されている。

【 0 0 3 6 】

かくして、リモートサイト 1 4 はローカルサイト 1 2 のミラーデータイメージを保持する。

【 0 0 3 7 】

リモートサイト 1 4 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) 1 0 5、1 0 6 を備えている。各サイト 1 4 は対応するホストプロセッサ (リモートホスト) 1 0 2、1 0 3 を含んでも含まなくても良い。

【 0 0 3 8 】

リモートストレージシステム 1 0 5、1 0 6 は、ローカルストレージシステム 1 0 4 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 1 0 4 のストレージボリューム 1 4 2 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。

【 0 0 3 9 】

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 を経由してデータバス 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、CPU 133 は内部バス 132 経由でメモリ 134 に結合し、更に本ディスクコントローラ 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】

CPU 133 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。

【 0 0 5 0 】

リモートコピーキューは、CPU 133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。

【 0 0 5 1 】

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。

【 0 0 5 2 】

図2はキュー構造を示している。図2で示す如く、ディスクコントローラ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。

【 0 0 5 3 】

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。

【 0 0 5 4 】

I/Oリードライト要求は典型的に、コマンドエン트리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。

【 0 0 5 5 】

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。

【 0 0 5 6 】

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られる。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。

【 0 0 5 7 】

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。

【 0 0 5 8 】

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。

【 0 0 5 9 】

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。

【 0 0 6 0 】

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。

【 0 0 6 1 】

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 1 1 0 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 1 1 0 からロールバックキュー 1 1 1 に移される。

【 0 0 6 2 】

ロールバックキュー 1 1 1 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 1 0 4 とリモートストレージシステム 1 0 5 , 1 0 6 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。

【 0 0 6 3 】

リモートストレージシステム 1 0 5 , 1 0 6 も基本的には同じ目的の為に、ロールバックキュー 1 1 1 を含む同一のキュー構造を有する。

【 0 0 6 4 】

ロールバックキュー 1 1 1 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。

【 0 0 6 5 】

コマンドエントリは、ロールバックキュー 1 1 1 から関連データがストレージボリューム 1 4 2 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 1 1 2 に移動される。

【 0 0 6 6 】

ローカルストレージボリューム 1 4 2 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 1 0 4 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 1 0 5 , 1 0 6 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。

【 0 0 6 7 】

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 1 1 3 に格納される。

【 0 0 6 8 】

上述したように、図 2 は、ホストプロセッサ 1 0 1 より受信しシーケンス番号 1 5 が割り当てられ、従ってワーキングキュー 1 1 0 に格納される、I / O 要求からのコマンドエントリを示す。

【 0 0 6 9 】

シーケンス番号 1 3 と 1 4 のコマンドエントリはこの時点でロールバックキュー 1 1 1 に保持され、関連データのストレージボリューム 1 4 2 への書き込み待ちの状態になっている。

【 0 0 7 0 】

中間キュー 1 1 2 は、シーケンス番号 1 0 , 1 1、及び 1 2 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 1 0 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 1 4 2 に書き込み済みで、従ってライトヒストリキュー 1 1 3 に格納されている。

【 0 0 7 1 】

リモートストレージシステム 1 0 5 , 1 0 6 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 1 0 5 はシーケンス番号 1 0 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、全データが受信の間は、ワーキングキューに格納されている。

【 0 0 7 2 】

シーケンス番号 1 0 はローカルストレージシステム 1 0 4 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 1 2 1 に移動される。

【 0 0 7 3 】

中間キュー 1 2 2 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 1 0 5 のストレージボリューム 1 4 0 に書き込み中である。

【 0 0 7 4 】

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 1 2 3 に移動される。ライトヒストリキュー 1 2 3 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。

【 0 0 7 5 】

他のリモートストレージシステム 1 0 6 も同様なデータキュー (1 3 0 、 1 3 1 、 1 3 2 及び 1 3 3) を有する。図 2 では、リモートストレージ 1 0 6 は現在シーケンス番号 1 2 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 1 3 0 に格納されている。

【 0 0 7 6 】

ロールバックキュー 1 3 1 は現在シーケンス番号 8 ~ 1 1 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 1 3 4 のみならず、ストレージボリューム 1 4 2 にも保持される事が好ましい。

【 0 0 7 7 】

ローカルストレージシステム 1 0 4 は更に、メモリ 1 3 4 中にリモートコピーステータステーブル 1 1 4 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 1 0 5 、 1 0 6 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。

【 0 0 7 8 】

例えば、リモートストレージシステム 1 0 5 (テーブル 1 1 4 では “ S 1 ” 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 1 0 6 (“ S 2 ”) 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 1 1 である事が判る。リモートコピーステータステーブル 1 1 4 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 1 2 1 , 1 3 1 とライトヒストリキュー 1 2 3 , 1 3 3 に関する情報も保持している。

【 0 0 7 9 】

かくして、リモートコピーステータステーブル 1 1 4 では、リモートストレージシステム 1 0 5 , 1 0 6 のロールバックキュー 1 2 1 , 1 3 1 は各々 4 データメッセージエントリ分の “ 長さ ” を有し、1 0 M B のデータを保持できる事がわかる。

【 0 0 8 0 】

リモートストレージシステム 1 0 5 , 1 0 6 各々のライトヒストリキュー 1 2 3 , 1 3 3 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 1 2 3 , 1 3 3 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 1 1 4 に表示されても良い。

【 0 0 8 1 】

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。

【 0 0 8 2 】

図 3 はプライマリストレージボリューム 2 0 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 2 1 , 2 2 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。

【 0 0 8 3 】

2 3 , 2 4 の矢印は、プライマリストレージボリューム 2 0 とセカンダリストレージボ

リユーム 2 1 , 2 2 との間のリモートコピーリンクを示す。

【 0 0 8 4 】

矢印 2 5 が示す図の次の部分では、リモートコピーリンクの一つ 2 3 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。

【 0 0 8 5 】

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 2 0 とセカンダリストレージボリューム 2 1) は自らのビットマップ情報を生成する。

【 0 0 8 6 】

これらのビットマップ情報は、図 3 では B p , B s 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。

【 0 0 8 7 】

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。

【 0 0 8 8 】

第一のシナリオは矢印 2 7 により図の左側に示され、第二のシナリオは矢印 2 8 で示され、第三のシナリオは矢印 2 9 で示されている。矢印 2 7 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 2 1 が障害になった場合である。この場合は、矢印 3 0 で示されるように、プライマリストレージボリューム 2 0 とセカンダリストレージボリューム 2 2 のペアが残っている。

【 0 0 8 9 】

図 3 の矢印 2 8 で示される第二のシナリオでは、セカンダリストレージ 2 2 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 3 1 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 2 0 と稼動可能なセカンダリストレージ 2 1 間のリンクが回復されるのを待つ。

【 0 0 9 0 】

リンクが回復すると、プライマリストレージボリューム 2 0 とセカンダリストレージボリューム 2 1 は 2 つのビットマップ B p と B s の情報を用いて再同期化を行う事が出来る。

【 0 0 9 1 】

図 3 の右側部分に、矢印 2 9 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 2 0 が障害になり、ビットマップ B p 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 (S 1) 2 0 とセカンダリストレージボリューム (S 2) 2 1 を再同期化するための唯一の手段である。

【 0 0 9 2 】

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。

【 0 0 9 3 】

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) 4 0 は、セカンダリストレージボリューム (S 1) 4 2 , (S 2) 4 4 , (S 3) 4 6 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。

【 0 0 9 4 】

矢印 4 1 は、セカンダリストレージボリューム 4 2 とセカンダリストレージボリューム 4 4 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 B p と B s がセカンダリストレージボリューム 4 2 と 4 4 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。

【 0 0 9 5 】

矢印 1 では、プライマリストレージボリューム 4 0 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 S 1 と S 2 及び S 2 と S 3 は同じままである。

【 0 0 9 6 】

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。

【 0 0 9 7 】

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 4 0 はセカンダリストレージボリューム 4 2 及び 4 4 との関連して動作を継続する。

【 0 0 9 8 】

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 B p からの変更情報は失われる。

【 0 0 9 9 】

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 4 0 とセカンダリストレージボリューム 4 4 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 4 0 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。

【 0 1 0 0 】

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 B s 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 - 1 0 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 4 6 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。

【 0 1 0 1 】

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。

【 0 1 0 2 】

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。

【 0 1 0 3 】

本発明の実施例によれば、ペアの停止を検出した場合、停止されたり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 (B p 又は B s) 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。

【 0 1 0 4 】

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 5 0 はセカンダリストレージボリューム 5 2、5 4 とペア状態になっている。矢印 5 6 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 5 0 とセカンダリストレージボリューム 5 2 の間での停止の発生を示している。

【 0 1 0 5 】

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 5 0 は異なるセカンダリストレージボリューム 5 4 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 5 4 は、プライマリストレージボリューム 5 0 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 B p ' は、ビットマップ B p 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 5 4 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 B p 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 B p が損失しても、セカンダリ ストレージボリューム 5 2 と 5 4 はプライマリビットマップ B p のコピー B p ' 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 B p 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 6 0 はプライマリビットマップ B p のコピー B p ' を受け取り、セカンダリ ストレージ ボリューム 6 6 はビットマップ B s のコピー B s ' 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 6 2 の障害では、ビットマップ B p 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 6 4 はプライマリ ストレージ ボリューム 6 0 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 6 2 の代わりに セカンダリ ストレージボリューム 6 4 の障害で発生しても、セカンダリ ストレージボリューム 6 6 は セカンダリ ストレージボリューム 6 2 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 1 0 4 (図 1) がホスト プロセッサ 1 0 1 (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 1 0 5 , 1 0 6 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 1 0 1 からのローカルストレージシステム 1 0 4 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 1 0 5 , 1 0 6 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】

CPU 133 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。

【 0 1 3 9 】

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。

【 0 1 4 0 】

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。

【 0 1 4 1 】

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。

【 0 1 4 2 】

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。

【 0 1 4 3 】

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。

【 0 1 4 4 】

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。

【 0 1 4 5 】

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。

【 0 1 4 6 】

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。

【 0 1 4 7 】

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。

【 0 1 4 8 】

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。

【 0 1 4 9 】

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。

【 0 1 5 0 】

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。

【 0 1 5 1 】

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ク接続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 S 3 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 S 2）に対して、“ 応答 ” を送信する。これは、図 1 1 の動作 1 0 0 6 に示されている。

【 0 1 6 4 】

セカンダリストレージボリューム S 2 が “ 応答 ” を受信すると、セカンダリストレージボリューム S 2 と S 3 との間のリモートコピーリンクを再開する。

【 0 1 6 5 】

この結果、セカンダリストレージボリューム S 2 は、例えば図 1 1 に示すような要求 2 0 ~ 2 5 の実行等の他の用途に用いられる事が可能となる。

【 0 1 6 6 】

図 1 1 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。

【 0 1 6 7 】

（ 2 ）本ビットマップのコピーを有するべき第二のストレージボリュームを選択する。

【 0 1 6 8 】

（ 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。

【 0 1 6 9 】

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。

【 0 1 7 0 】

（ 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。

【 0 1 7 1 】

（ 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。

【 0 1 7 2 】

（ 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。

【 0 1 7 3 】

図 1 2 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 7 0 1 で障害が検出される。障害検出は多数の方法でなされる。

【 0 1 7 4 】

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。

【 0 1 7 5 】

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 1 0 5 , 1 0 6 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。

【 0 1 7 6 】

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。

【 0 1 7 7 】

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、ステップ 7 0 3 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。

【 0 1 7 8 】

例えば、図 2 において、ローカルストレージシステム 1 0 4 が障害になり、ステップ 7 0 2 にて、リモートストレージシステム 1 0 5 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 1 0 6 は、ロールバックキュー 1 3 1 内にデータメッセージ 8、9、1 0、及び 1 1 がありライトヒストリキュー 1 3 3 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。

【 0 1 7 9 】

障害検出時には、リモートストレージシステム 1 0 5 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 1 0 6 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。

【 0 1 8 0 】

ステップ 1 1 0 1 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 7 0 4 に移る。

【 0 1 8 1 】

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 1 1 0 2 に示され、処理は図 1 3 で示され、そこで説明される。

【 0 1 8 2 】

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。

【 0 1 8 3 】

典型的には、この決定は、リモートコピー環境の管理者がユーザによって、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。

【 0 1 8 4 】

ステップ 7 0 4 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 7 0 5 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。

【 0 1 8 5 】

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 7 0 6 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。

【 0 1 8 6 】

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 1 0 6 は、ロールバックコマンドを受領すると、シーケンス番号 1 0 と 1 1 のデータメッセージを破棄して、本処理はステップ 7 1 5 で完結する。

【 0 1 8 7 】

もしこれに対して、ステップ 7 0 4 にて、ロールフォワードが決定されたなら、ステップ 7 0 7 にて、ロールフォワードシーケンス番号が決定される。

【 0 1 8 8 】

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。

【 0 1 8 9 】

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 1 0 5 の場合は、リモートストレージシステム 1 0 6 がシーケンス番号 1 0 と 1 1 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、CPU 133 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】

このような場合は、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・・・I/O 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

1 4 2 . . . ストレージボリューム