(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4892185号
(P4892185)
(45)発行日　平成24年3月7日(2012.3.7)　(24)登録日　平成23年12月22日(2011.12.22)

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

請求項の数 11　外国語出願　(全 25 頁)

(21)出願番号　特願2004-338341(P2004-338341)
(22)出願日　平成16年11月24日(2004.11.24)
(65)公開番号　特開2005-276162(P2005-276162A)
(43)公開日　平成17年10月6日(2005.10.6)
審査請求日　平成19年9月27日(2007.9.27)
(31)優先権主張番号　10/758971
(32)優先日　平成16年1月15日(2004.1.15)
(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　カルニベル　１０８

審査官　菅原　浩二

最終頁に続く

(54)【発明の名称】分散リモートコピーシステム

(57)【特許請求の範囲】
【請求項１】
第一のプライマリボリュームを含む第一のプライマリストレージサブシステム、及び第二のプライマリボリュームを含む第二のプライマリストレージサブシステムであって、前記第一及び第二のプライマリボリュームは、複数の書込みデータを当該複数の書き込みデータの書込み順に保存する、第一及び第二のプライマリストレージサブシステムと、
前記第一及び第二のプライマリストレージサブシステムに接続されるとともに、前記第一及び第二のプライマリストレージサブシステムから前記書込みデータを同期的に受信するように構成された、中間ストレージサブシステムであって、前記第一及び第二のプライマリストレージサブシステムから受信した前記書込みデータの保存の順番を反映する書込み順情報を生成するカウンタを含み、
データ完全性が保証される整合性グループとして定義された第一及び第二の中間ボリュームを含み、前記第一の中間ボリュームは、前記第一のプライマリボリュームから前記書込みデータを受信するように構成され、前記第二の中間ボリュームは前記第二のプライマリボリュームから前記書込みデータを受信するように構成される、中間ストレージサブシステムと、　前記中間ストレージサブシステムの有効カウンタであって、前記有効カウンタは、前記中間ストレージサブシステムに接続されるとともに、前記中間ストレージサブシステムから前記書込みデータを非同期的に受信するように構成された、第一及び第二のセカンダリストレージサブシステムでのコピーのために有効化される準備ができている前記書込みデータの最大順番番号を保持するように構成される有効カウンタと、

前記第一及び第二のセカンダリストレージサブシステムであって、前記第一のセカンダリストレージサブシステムは、前記中間ストレージサブシステムとは別のストレージサブシステムである前記第一のプライマリストレージサブシステムの前記第一のプライマリボリュームをミラーするように構成された第一のセカンダリボリュームを含み、前記第二のセカンダリストレージサブシステムは、前記中間ストレージサブシステムとは別のストレージサブシステムである前記第二のプライマリストレージサブシステムの前記第二のプライマリボリュームをミラーするように構成された第二のセカンダリボリュームを含み、前記第一のプライマリボリュームを有する前記第一のプライマリストレージサブシステム及び前記第二のプライマリボリュームを有する前記第二のプライマリストレージサブシステムとは別のストレージサブシステムである前記中間ストレージサブシステムの前記書込み順情報プロバイダによって生成されたとおりの、前記書込みデータと対応付けられた前記書込み順情報に従って、前記書込みデータは、前記第一及び第二のセカンダリストレージサブシステムに保存され、前記中間ストレージサブシステムから前記準備要求受信した場合、前記準備要求受信前に受信した書き込みデータを回復するために必要な情報である回復用情報を保持し、前記中間ストレージサブシステムから前記有効化を受信した場合、前記回復用情報を破棄する第一及び第二のセカンダリストレージサブシステムと、

を備え、

前記中間ストレージサブシステムは、前記書込みデータを準備するようにという準備要求を送信し、かつ当該準備要求に従って準備された前記書込みデータを前記有効カウンタの値に基づいて有効化するようにという有効化要求を送信する、

リモートコピーシステム。

【請求項２】

各々前記第一及び第二のプライマリサブシステムに設けられる第一及び第二のプライマリビットマップと、

前記中間サブシステムに設けられる第一及び第二の中間ビットマップと、

各々前記第一及び第二のセカンダリサブシステムに設けられる第一及び第二のセカンダリビットマップと

をさらに備え、

前記第一及び第二のプライマリビットマップは前記第一の中間ビットマップに対応付けられ、前記第一及び第二のセカンダリビットマップは前記第二の中間ビットマップに対応付けられ、

前記ビットマップは、再同期プロセスで、ペアボリュームのミラー処理が中断した後に変更のあったデータのみを決定し、かつ当該データをコピーするために使用される、請求項１に記載のリモートコピーシステム。

【請求項３】

前記中間サブシステムは、前記第一及び第二のプライマリサブシステムから前記書込みデータを受信するジャーナルボリュームを含む、請求項１に記載のリモートコピーシステム。

【請求項４】

前記第一及び第二のプライマリサブシステムはディスクアレイユニットである、請求項１に記載のリモートコピーシステム。

【請求項５】

前記プライマリサブシステムはプライマリサイトに設けられ、前記セカンダリサブシステムはセカンダリサイトに設けられ、前記プライマリサイトはプライマリホストを含み、前記セカンダリサイトはセカンダリホストを含み、

前記プライマリサブシステム又は前記プライマリホストのいずれかが障害を被るか又はオフラインにされた場合に、前記セカンダリサブシステムは、第一の保存領域として、前記プライマリサブシステムに取って代わるように構成されている、請求項１に記載のリモートコピーシステム。

【請求項６】

第一のプライマリストレージサブシステムとは別のストレージサブシステムである中間ストレージサブシステムで、第一のプライマリストレージサブシステムから同期的に送信される第一の書込みデータを受信することと、
データ完全性が保証される整合性グループとして定義される複数のボリュームからなる中間ボリュームを含む前記中間ストレージサブシステムで、第一の書込み順情報を前記第一の書込みデータに対応付けることと、
前記第二のプライマリストレージサブシステムとは別のストレージサブシステムである前記中間サブシステムで、第二のプライマリストレージサブシステムから同期的に送信される第二の書込みデータを受信することと、
前記中間ストレージサブシステムで、第二の書込み順情報を前記第二の書込みデータに対応付けることと、
前記中間ストレージサブシステムとは別のストレージサブシステムである第一のセカンダリストレージサブシステムに、前記第一の書込みデータ及び前記第一の書込み順情報を非同期的に送信することと、
前記中間ストレージサブシステムとは別のストレージサブシステムである第二のセカンダリストレージサブシステムに、前記第二の書込みデータ及び前記第二の書込み順情報を非同期的に送信することと、
前記第一及び第二のセカンダリサブシステムに保存するために前記第一、第二及び第三の書込みデータを準備するようにという要求であって、参照用の順番番号を含む準備要求を、前記中間サブシステムから前記第一及び第二のセカンダリサブシステムへ送信することと、
前記準備要求に従って準備された前記書込みデータを有効化するようにという要求を送信することであって、当該有効化要求は、前記中間サブシステムから前記第一及び第二のセカンダリサブシステムへ送信され、準備すべき前記書込みデータを識別し、前記参照用の順番番号よりも小さい又は等しい前記順番番号で前記書込みデータを有効化する、前記有効化要求を送信することと、
を含む、リモートコピーシステムを操作する方法であって、
前記第一及び第二の書込みデータは、各々前記第一及び第二の書込み順情報に従って前記第一及び第二のセカンダリサブシステムに保存され、
前記中間ストレージサブシステムは、ホストユニットに直接的には接続されていない、
リモートコピーシステムを操作する方法。
【請求項７】
前記中間ストレージサブシステムで、前記第一のプライマリサブシステムから第三の書込みデータを、前記第一の書込みデータの後に、同期的に受信することと、
前記第三の書込みデータを第三の書込み順情報に対応付けることと、
前記第三の書込みデータ及び前記第三の書込み順情報を前記第一のセカンダリサブシスムに非同期的に送信することと、
をさらに含み、　前記第一及び第三の書込みデータは、前記第一のセカンダリサブシステムでの保存領域を識別する保存先アドレスと同一の保存先アドレスを有し、
前記第一及び第三の書込み順情報は、前記第三の書込みデータを前記識別された保存領域に保存する前に、前記第一の書込みデータを前記識別された保存領域に保存するために使用される、請求項６に記載の方法。
【請求項８】
リモートコピーシステムに設けられるとともに、複数のプライマリストレージサブシステムに接続され、かつ複数のセカンダリサブシステムに接続される、中間ストレージサブシステムであって、
前記中間ストレージサブシステムとは別のストレージサブシステムである少なくとも一つのプライマリサブシステムから書込みデータを受信するように構成された第一の中間保存領域であって、前記書込みデータは、前記少なくとも一つのプライマリサブシステムから同期的に受信され、前記第一の中間保存領域は、データ完全性が保証される整合性グル

ープとして定義される複数のボリュームの第一の中間保存領域と、
前記少なくとも一つのプライマリサブシステムから受信された前記書込みデータのために書込み順情報を生成するように構成された、カウンタであって、前記書込み順情報は前記書込みデータに対応付けられる、カウンタと、
前記中間サブシステムに設けられた有効カウンタであって、前記セカンダリサブシステムでのコピーのために有効化される準備ができている前記書込みデータの最大順番番号を保持するように構成された、有効カウンタと、
第二の中間保存領域と、
を備える中間ストレージサブシステムであり、
前記中間ストレージサブシステムとは別のストレージサブシステムである前記セカンダリサブシステムのうちの少なくとも一つに前記書込みデータを保存するために、前記書込み順情報を使用することにより、前記少なくとも一つのセカンダリサブシステムは、前記少なくとも一つのプライマリサブシステムをミラーし、
前記中間ストレージサブシステムは、前記書込みデータを準備するようにという要求と、当該準備要求に従って準備された前記書込みデータを、前記有効カウンタの値に基づいて有効化するようにという要求とを送信し、
前記複数のプライマリサブシステムは、第一のプライマリサブシステムに設けられた第一のプライマリボリュームと、第二のプライマリサブシステムに設けられた第二のプライマリボリュームとを含み、
前記第一及び第二の中間保存領域は第一及び第二の中間ボリュームであって、前記第一の中間ボリュームは、前記第一のプライマリボリュームから書込みデータを受信するように構成され、前記第二の中間ボリュームは、前記第二のプライマリボリュームから書込みデータを受信するように構成され、
前記第一の中間ボリュームは、前記第一のプライマリボリュームから受信した前記書込みデータを、第一のセカンダリサブシステムに設けられた第一のセカンダリボリュームに送信するように構成され、前記第二の中間ボリュームは、前記第二のプライマリボリュームから受信した前記書込みデータを、第二のセカンダリサブシステムに設けられた第二のセカンダリボリュームに送信するように構成され、
前記中間ストレージサブシステムは、ホストユニットに直接的には接続されてはいない、
中間ストレージサブシステム。
【請求項９】
前記第一の保存領域は、前記少なくとも一つのプライマリサブシステムから所定の順番で第一及び第二の書込みデータを受信するように構成され、前記第一及び第二の書込みデータには各々、
前記中間サブシステムによって第一及び第二の書込み順情報が提供され、前記第一及び第二の書込みデータは、前記第一及び第二の書込み順情報を用いて、前記所定の順番に従って前記少なくとも一つのセカンダリサブシステムに保存される、請求項８に記載のストレージサブシステム。
【請求項１０】
前記書込み順情報は、順番番号を生成するように構成されたカウンタによって生成され、
当該生成された順番番号は、前記書込みデータが前記少なくとも一つのプライマリサブシステムから受信された順番に従って、前記書込みデータに対応付けられる、請求項８に記載のストレージサブシステム。
【請求項１１】
前記第一の保存領域は、前記複数のプライマリサブシステムから書込みデータを受信するように構成されたジャーナルボリュームである、請求項８に記載のストレージサブシステム。
【発明の詳細な説明】

【技術分野】
【０００１】
0001 本発明はストレージシステムに関わり、より具体的には、リモートコピー機能を果たす分散ストレージシステムに関連する。
【背景技術】
【０００２】
0002 データは全ての計算プロセスの基礎となる資源である。最近のインタネットやｅ－ビジネスの爆発的成長により、データストレージシステムに対する需要は凄まじく増大している。一般的にネットワーク結合型ストレージはＮＡＳ(Network Attached Storage)又はＳＡＮ(Storage Area Network)の二つの形態をとっている。ＮＡＳでは、ファイルフォーマットのデータをストレージサーバとクライアント間で転送するのにイーサネット（イーサネットは登録商標です。）上のＩＰを使用する。ＮＡＳに於いては、ディスクアレイやテープデバイスの如き集合ストレージシステムは、ＴＣＰ／ＩＰのようなメッセージ交信プロトコルを用いて、イーサネット等のＬＡＮ(Local Area Network)を通してメッセージ交換ネットワークに直結される。本ストレージシステムはクライアントーサーバシステムでのサーバとしての働きをする。
【０００３】
0003 一般的にＳＡＮは種々のサーバとストレージ資源の間でデータ移動を行う専用の高速ネットワークである。ＮＡＳとは異なって、クライアントとサーバ間での慣用的メッセージ交換での交信衝突を避ける為に、分離した専用ネットワークが用いられる。ＳＡＮでは、ストレージ資源とプロセッサ又はサーバとの間の直接結合が可能になる。ＳＡＮは単一サーバ専用でも又サーバ間で共用することもできる。単一のローカルサイトに集中することも又地理的に離れたサイトに拡張することもできる。ＳＡＮインタフェースでは、ＦＣ(Fibre Channel)、ＥＳＣＯＮ(Enterprise System Connection)、ＳＣＳＩ(Small Computer System Interface)、ＳＳＡ(Serial Storage Architecture)、ＨＩＰＰＩ(High Performance Parallel Interface)その他将来登場するものも含めて多様で異なったプロトコルが使用可能である。例えば、ＩＥＴＦ(Internet Engineering Task Force)はＴＣＰ／ＩＰ上でブロックレベル転送を可能にするｉＳＣＳＩの新プロトコルを開発中である。又ある会社は、ｉＳＣＳＩをＳＡＮでの主要スタンダードにすべく、ｉＳＣＳＩ－ＴＣＰ／ＩＰプロトコルスタックをホストプロセッサから分離しようとしている。
【０００４】
0004 ストレージシステムのタイプに関わらず、データストレージシステムのユーザは、貴重なデータがストレージユニット(又はストレージシステム)の障害で損失することを防ぐ為に、バックアップデータを維持する事に強い関心がある。従って、データストレージシステムは、プライマリユニットが障害になった時の緊急回復データを保存するバックアップユニットを持つことが普通である。然しながら、障害は、ユニット自身の問題によるばかりではなく、ストレージユニットの設置場所での地震や嵐等の自然災害で発生する事もある。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場合には、両ユニットとも自然災害で損傷を受ける事があり得る。従って、多くのストレージシステムのユーザは、プライマリユニットとバックアップユニットを例えば１００マイル以上も離れて設置している。実際には、プライマリとバックアップユニットを異なる大陸に設置しているユーザもある。
【０００５】
0005 今日では、同期モードと非同期モードの二つの動作モードが、ストレージシステムがデータをバックアップ又はセカンダリサイトにコピーするのに用いられる。同期モードでは、ホストからのプライマリストレージシステムへの書込み要求は、書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、そこからの応答が返って初めて完了する。従って、本モードでは、セカンダリストレージシステムからの応答が返るまでは、プライマリシステムのキャッシュにホストからのデータが保存されている為に、セカンダリストレージシステムでのデータ損失は発生しないことが保証されている。加えて、プライマリ

ストレージシステムの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のセカンダリボリューム(ＳＶＯＬ）は同一に保たれている為に、ＰＶＯＬが損失した場合にはＳＶＯＬが直ちに代わりを務める事ができる。然しながら、このモードでは、プライマリとセカンダリは、例えば、１００マイルも離れて設置する事はできない。距離を離すと、ストレージシステムは、ホストからの書込み要求を効率的に実行することはできない。

【０００６】

0006　非同期モードでは、ホストからのプライマリストレージシステムへの書込み要求は、データを指定プライマリシステムに書き込むのみで終了する。書込みデータはその後に、セカンダリストレージシステムにコピーされる。即ち、プライマリストレージシステムへのデータ書込みは、セカンダリストレージシステムへのデータコピーとは独立したプロセスである。この結果として、プライマリとセカンダリは、例えば、１００マイル以上も離れて設置することができる。然しながら、ＰＶＯＬとＳＶＯＬは常に同じデータを維持している訳ではない為、プライマリシステムがダウンしたらデータが失われる可能性がある。従って、同期モードと非同期モードの長所、即ち、データを失う事無くプライマリとセカンダリシステムが離れて設置できるデータストレージシステムが望まれるところである。非同期モードの好適なリモートコピー方法については、参照記載のＹａｍａｍｏｔｏその他による、米国特許ＮＯ．６,４０８,３７０に公開されている。

【０００７】

【特許文献１】米国特許第６４０８３７０号

【発明の開示】

【発明が解決しようとする課題】

【０００８】

同期モードでは、ホストからのプライマリストレージシステムへの書込み要求は、書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、そこからの応答が返って初めて完了する。従って、本モードでは、セカンダリストレージシステムからの応答が返るまでは、プライマリシステムのキャッシュにホストからのデータが保存されている為に、セカンダリストレージシステムでのデータ損失は発生しないことが保証されている。加えて、プライマリストレージシステムの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のセカンダリボリューム(ＳＶＯＬ）は同一に保たれている為に、ＰＶＯＬが損失した場合にはＳＶＯＬが直ちに代わりを務める事ができる。然しながら、このモードでは、プライマリとセカンダリは、例えば、１００マイルも離れて設置する事はできない。距離を離すと、ストレージシステムは、ホストからの書込み要求を効率的に実行することはできない。

【０００９】

非同期モードでは、ホストからのプライマリストレージシステムへの書込み要求は、データを指定プライマリシステムに書き込むのみで終了する。書込みデータはその後に、セカンダリストレージシステムにコピーされる。即ち、プライマリストレージシステムへのデータ書込みは、セカンダリストレージシステムへのデータコピーとは独立したプロセスである。この結果として、プライマリとセカンダリは、例えば、１００マイル以上も離れて設置することができる。然しながら、ＰＶＯＬとＳＶＯＬは常に同じデータを維持している訳ではない為、プライマリシステムがダウンしたらデータが失われる可能性がある。従って、同期モードと非同期モードの長所、即ち、データを失う事無くプライマリとセカンダリシステムが離れて設置できるデータストレージシステムが望まれるところである。

【課題を解決するための手段】

【００１０】

0007　リモートコピーシステム(リモートコピー機能を備えるストレージシステム)は一つ又は複数のプライマリストレージサブシステムを持つことができる。分散リモートコピーシステムは、複数のプライマリストレージサブシステムを持つシステムを言う。リモートコピーの目的の為に、分散リモートコピーシステムでの複数のプライマリストレージサ

ブシステムのデータは、複数のセカンダリストレージサブシステムにコピーされる。例えば、第一のプライマリサブシステムは第一のセカンダリサブシステムとペアを組みミラーされ、又第二のプライマリサブシステムは第二のセカンダリサブシステムとペアを組みミラーされる。

【００１１】

0008 このようなリモートコピーシステムで、データインテグリテイ(データの完全性)を保つ事は、現在使用可能なストレージデバイスと技術レベルのもとでは、困難な課題である。プライマリストレージサブシステムを使用するデータベースでは、書込み順又はデータインテグリテイを保証する為に、一般的に、書込み順を記録する書込みログを書き込む。第一のログが第一のプライマリサブシステムの為に、第二のログが第二のプライマリサブシステムの為に書き込まれる。第一と第二のログが共に、プライマリサブシステムでのデータベースのデータインテグリテイの為に使用される。同様に、第一と第二のログは、それぞれ、セカンダリサブシステムでの書込み順を保証する為に、セカンダリサブシステムにコピーする必要がある。プライマリサブシステムでの実際の書込み順をミラーする為に、二つのログ間の関係情報を維持しながら、第一と第二のログをセカンダリサブシステムにコピーする必要がある。然しながら、現行技術によるリモートコピーでは、第一と第二のプライマリサブシステムの間には、殆ど協調関係が存在しない為、このことは容易には実現できないことである。

【００１２】

0009 ここで用いる“ストレージシステム”は、データ保存の為に構成されたコンピュータシステムであり、一つ以上のストレージユニット又はディスクアレイの如きストレージサブシステムを備える。従って、ストレージシステムは、一つ以上のホストと一つ以上のストレージサブシステム、又は一つだけのストレージサブシステム又はユニット、又は通信回線で複数のホストに結合された複数のストレージサブシステム又はユニットを備えるコンピュータシステムである。

【００１３】

0010 ここで用いる“ストレージサブシステム”は、データ保存の為に構成され、ストレージ領域及び一つ以上のホストからの要求を処理するためのストレージコントローラを備えるコンピュータシステムである。ストレージサブシステムは、ストレージデバイス、ストレージユニット、ストレージ装置などと呼ばれる事がある。ストレージサブシステムの一例はディスクアレイユニットである。

【００１４】

0011 ここで用いる“ホスト”は、一つ以上のストレージシステム又はストレージサブシステムに結合し、本ストレージシステム又はストレージサブシステムに要求を送信するように構成されたコンピュータシステムである。ホストはサーバ又はクライアントとしての役割を果たす。

【００１５】

0012 ここで用いる“リモートコピーシステム”は、リモートコピー機能を果たすように構成されたコンピュータシステムである。リモートコピーシステムは、ネットワーク又は通信回線で結合した、単一のストレージシステム、ストレージサブシステム又はユニット、あるいは複数のストレージユニット、ストレージシステム、又はストレージサブシステムを指すこともある。従って、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、中間ストレージシステム、又はこれらの組合せを示すことがある。更にリモートコピーシステムは、一つ以上のホストを含むこともできる。

【００１６】

0013 一実施例では、プライマリストレージサブシステム(以降プライマリサブシステムで表す)は中間ストレージサブシステム(以降中間サブシステムで表す)にデータを同期的に送信する。書込みデータの受信を契機に、中間サブシステムは書込み順を指定する情報を含む制御データを生成する。一実施例では、制御データの一部例えば順序番号が生成

され、書込みデータの制御データに付加される。中間サブシステムは、この書込みデータを制御データと共に、セカンダリストレージサブシステム(以降セカンダリサブシステムで表す)に非同期的に送信する。セカンダリサブシステムは、制御データに基づいてこの書込みデータをセカンダリボリュームに保存する。書込み順は、中間サブシステムにより生成された順序番号により維持される。

【００１７】

0014 一実施例では、リモートコピーシステムは、複数の書込みデータを与えられた順で保存する各々第一と第二のプライマリボリュームを備える第一と第二のプライマリサブシステムと；第一と第二のプライマリサブシステムからの書込みデータを受信し、本書込みデータに対して、本第一と第二のプライマリサブシステムで与えられた書込み順を反映し、各書込みデータに付与される書込み順情報、を生成する書込み順情報生成部を備える、中間サブシステムと；各々第一と第二のプライマリボリュームをミラーする第一と第二のセカンダリボリュームを備え、本中間サブシステムからの書込みデータを受信する第一と第二のセカンダリサブシステムと；から成り、各データに付与された書込み順情報に基づいて、本書込みデータを第一と第二のセカンダリサブシステムに保存する。

【００１８】

0015 他の実施例では、リモートコピーシステムで、複数のプライマリサブシステムと複数のセカンダリサブシステムに結合される中間サブシステムは、少なくとも一式のプライマリサブシステムからの書込みデータを同期的に受け取る第一のストレージ領域と；少なくとも一式のプライマリサブシステムから受けとった各書込みデータに付与される書込みデータの書込み順情報、を生成する書込み順情報生成部とを持ち、本書込み順情報は、セカンダリサブシステムの少なくとも一式でデータ書込みの為に使用され、少なくとも一式のセカンダリサブシステムが少なくとも一式のプライマリサブシステムのミラーをする。

【００１９】

0016 更に他の実施例では、リモートコピーシステムを操作する方法が提供され、中間サブシステムで第一と第二のプライマリサブシステムから各々第一と第二の書込みデータを同期的に受信し；本第一と第二の書込みデータに各々第一と第二の書込み順情報を付与し；本第一と第二の書込みデータと第一と第二の書込み順情報を各々第一と第二のセカンダリサブシステムに非同期的に送信する；ステップから成り、第一と第二の書込みデータは、各々第一と第二の書込み順情報に従って、第一と第二のセカンダリサブシステムに保存される。

【００２０】

0017 更に他の実施例は、リモートコピーシステムを操作するコンピュータプログラムを格納するコンピュータ読み取り可能な媒体を提供し、本媒体は、中間サブシステムで、第一と第二のプライマリサブシステムでの各々第一と第二のプライマリボリュームから第一と第二の書込みデータを同期的に受信するコードと；本第一と第二の書込みデータに各々第一と第二の書込み順情報を付与するコードと；本第一と第二の書込みデータと本第一と第二の書込み順情報を、各々第一と第二のセカンダリボリュームより成る第一と第二のセカンダリサブシステムに、非同期的に転送するコードとを、格納し、第一と第二の書込みデータは各々第一と第二の書込み順情報に従って、第一と第二のセカンダリサブシステムに保存され、第一と第二のセカンダリボリュームは各々第一と第二のプライマリボリュームのミラーをする。

【００２１】

0018 更にもう一つの実施例では、分散リモートコピーシステムでの中間サブシステムは、第一と第二のプライマリサブシステムで定義される各々第一と第二のプライマリボリュームから書込みデータを同期的に受信する手段と；プライマリサブシステムから受信した書込みデータに書込み順情報を与える書込み順情報生成手段と、から成り、本書込み順情報は、各々第一と第二のセカンダリサブシステムで定義される第一と第二のセカンダリボリュームへの書込みデータの書込みに使用され、本第一と第二のセカンダリボリューム

が各々第一と第二のプライマリボリュームをミラーする。

【発明の効果】

【００２２】

本リモートコピーシステムは好適な分散リモートコピーシステムで、セカンダリストレージサブシステム(以下セカンダリサブシステムと呼ぶ)に亘って、データのインテグリティ(データの完全性)を維持すべく構成されている。

【発明を実施するための最良の形態】

【００２３】

0033 図１Ａ（図１(a)）は、本発明の一実施例による、複数のプライマリストレージサブシステム(以下プライマリサブシステムと呼ぶ)１００ａを持つリモートコピーシステム５０を示す。本リモートコピーシステムは好適な分散リモートコピーシステムで、セカンダリストレージサブシステム(以下セカンダリサブシステムと呼ぶ)に亘って、データのインテグリティ(データの完全性)を維持すべく構成されている。

【００２４】

0034 システム５０は、複数のプライマリサブシステム１００ａ、複数のセカンダリサブシステム１００ｂ、及び一つの中間ストレージサブシステム(以下中間サブシステムと呼ぶ)１００ｃを含む。プライマリサブシステムは第一と第二のプライマリサブシステム１００ａ－１と１００ａ－２を含む。セカンダリサブシステムは第一と第二のセカンダリサブシステム１００ｂ－１と１００ｂ－２を含む。一実施例では、単一の中間サブシステムが用いられる。

【００２５】

0035 本実施例では、プライマリサブシステムはデータを中間サブシステムに同期的にコピーし、中間サブシステムは本データをセカンダリサブシステムに非同期的にコピーする。中間サブシステム１００ｃは、プライマリサブシステム１００ａの近くに設置され、セカンダリサブシステム１００ｂとは十分離れて設置される。例えば、プライマリと中間サブシステムは１０マイル以内に設置され、中間とセカンダリサブシステムは互いに１００マイル以上離れて設置される。一実施例では、プライマリと中間サブシステムは、同じビル又は複合建築体内に設置される。

【００２６】

0036 図１Ｂ（図１(b)）は、読み書き要求を処理するストレージコントローラ６２と書込み要求に対応してデータ保存の為の記録媒体を備えるストレージユニット６３を含む好適なストレージサブシステム６０を示す。ストレージコントローラ６２は、ホストコンピュータに結合する為のホストチャネルアダプタ６４と他のサブシステムに結合する為のサブシステムチャネルアダプタ６６とストレージサブシステム６０内のストレージユニット６３に結合する為のディスクアダプタ６８を有する。本実施例では、これらの各アダプタはデータを送受信する為のポート(図示していない)を備え、本ポートを通してデータ転送をコントロールする為のマイクロプロセッサ(図示していない)を備える。

【００２７】

0037 ストレージコントローラ６２は、更に、ストレージユニット６３との間で読み書きするデータを一時的に保存する為のキャッシュメモリ７０を備える。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(図示していない)である。

【００２８】

0038 本サブシステムはホストコンピュータに対する記憶領域として、複数の論理ボリュームを提供する。ホストコンピュータはストレージサブシステムとの間でデータを読み書きする為に、これら論理ボリュームの識別子を使用する。これら論理ボリュームの識別子はＬＵＮ(Logical Unit Number)と呼ばれる。これら論理ボリュームは単一の物理ストレージデバイスに含まれてもよく又複数の物理ストレージデバイスに跨って定義されても良い。同様に複数の論理ボリュームが単一の物理ストレージデバイスに収容されても良い。ストレージサブシステムに関するより詳細な記述は、２００２／６／５に出願受理された日本出願Ｎｏ．２００２－１６３７０５に対する優先出願、標題"Data Storage Subsys

tem", ２００３／３／２１に出願受理され本代理人の担当する米国特許ＮＯ．１０／３９４,６３１に見られ、ここに参考として含まれる。

【００２９】

0039 図１Ａに戻って、リモートコピーシステム５０は更にプライマリホスト１１０ａとセカンダリホスト１１０ｂを含む。ホストは、各ストレージサブシステムに対して、データの読み書き要求を発行する。プライマリホストは、プライマリサブシステムのボリュームにアクセスするアプリケーションプログラム１０２ａを持つ。セカンダリホストは、セカンダリサブシステムのボリュームにアクセスするアプリケーションプログラム１０２ｂを持つ。プライマリホスト１１０ａ又はプライマリサブシステム１００ａの何れか又は双方が使用不能になれば、アプリケーションプログラム１０２ａはセカンダリホスト１１０ｂにフェイルオーバされ、例えば、エンタプライスビジネスにデータベースを継続使用できる様にする。

【００３０】

0040 各ストレージサブシステムはデータ保存の為のボリュームを有する。プライマリサブシステムはプライマリボリューム即ちＰＶＯＬ１０１ａを有する。プライマリボリュームはプライマリホストがアクセスする生産用データを保存する。プライマリボリュームは第一と第二のＰＶＯＬ１０１ａ－１と１０１ａ－２で成る。セカンダリサブシステムはセカンダリボリューム即ちＳＶＯＬ１０１ｂを有する。セカンダリボリュームはＰＶＯＬの生産用データのバックアップデータを保存する。セカンダリボリュームは第一と第二のＳＶＯＬ１０１ｂ－１と１０１ｂ－２で成る。中間サブシステムは中間ボリュームＩＶＯＬ１０１ｃを有する。中間ボリュームはセカンダリボリュームに保存されるべき生産用データのコピーを一時的に保存する。従って、中間ボリュームは、プライマリボリューム１０１ａに対して“セカンダリボリューム”で、セカンダリボリューム１０１ｂに対して“プライマリボリューム”である。中間ボリュームは、第一と第二のＩＶＯＬ１０１ｃ－１と１０１ｃ－２で成る。

【００３１】

0041 整合性グループ１２０が中間サブシステム１００ｃにて定義される。整合性グループは、データインテグリテイが保証されるボリュームの集合である。ユーザは、典型的に、データインテグリテイが望まれる所定のアプリケーション、例えばデータベース等、に関連するデータを格納するボリュームに対して整合性グループを定義する。

【００３２】

0042 タイマ１３０が中間サブシステム１００ｃに備えられ、データ要求に対して時刻情報を提供する。例えば、タイマはプライマリサブシステム１００ａより受信した書込み要求の制御データにタイムスタンプを与える。書込み要求は、通常、所定のストレージロケーションのアドレス指定と共にここに書き込まれるデータを伴う。一般的に書込み要求は、書き込まれるデータ(書込みデータ)とこれに関連するヘッダ又は管理情報を含む。制御データは、所与の要求（読み出し、または書込み）に関連する管理情報を提供し、当該要求のヘッダの一部とも考えられる。

【００３３】

0043 カウンタ１２１は、中間サブシステム１００ｃに配備され、プライマリサブシステム１００ａから書込み要求を受信した時に制御データに順序番号を付与する。順序番号は書込み順を保持する為に使用される。一般的に、各整合性グループ１２０は自分のカウンタ１２１を持ち、整合性グループ又はデータベースに対する書込み順を維持する。通常、中間サブシステムは書込みデータと共にヘッダ又は制御データを受け取り、順序番号はこのヘッダ又は制御データに付加される。

【００３４】

0044 タイムスタンプ又は順序番号又は双方は、セカンダリサブシステムが書込み順をプライマリサブシステム１００ａと同じにする為に使用することができる。本発明では、本順序番号が書込み順を保証する為に使用される。

【００３５】

0045　有効カウンタ１２２が、セカンダリサブシステム１００ｂで有効化可能な最新順序番号を保存する為に中間サブシステム１００ｃに用意される。

【００３６】

0046　図５を参照するに、各ストレージサブシステムは、ボリューム間のペア関係の情報を保存する為に、整合性グループテーブルを備える。プライマリサブシステム１０１ａはプライマリ整合性グループテーブル５００ａ、即ち各々第一と第の二プライマリサブシステムの為の第一と第二のテーブル５００ａ－１と５００ａ－２を備える。中間サブシステム１０１ｃは第一と第二の整合性グループテーブル５００ｃ－１と５００ｃ－２を備える。セカンダリサブシステム１０１ｂは、セカンダリ整合性グループテーブル５００ｂ、即ち各々第一と第二のセカンダリサブシステムの為の第一と第二のテーブル５００ｂ－１と５００ｂ－２を備える。

【００３７】

0047　各整合性グループテーブルは、グループＩＤ欄ＧＩＤ５１０、グループ属性欄ＧＲＡＴＴＲ５２０、グループ状態欄ＧＲＳＴＳ５３０、中間ポインタ欄ＩＰＴＲ５４０、及びボリューム情報欄ＶＯＬＩＮＦＯ＃５５０を有する。

【００３８】

0048　ＧＩＤ５１０は整合性グループの識別情報を格納する。識別情報は一般に当該ストレージサブシステムでの固有番号である。ＧＩＤ５１０はペアを構成する整合性グループを識別する為に使用される。例えば、二つのＧＩＤ欄が同じ値を持っていれば、二つの整合性グループはペアになっているとされる。

【００３９】

0049　ＧＲＡＴＴＲ５２０は、当該グループが、プライマリボリュームか又はセカンダリボリュームのどちらとして機能しているかを表示する。例えば、プライマリ又はセカンダリサブシステムの整合性グループは各々ＰＲＩＭＡＲＹ又はＳＥＣＯＮＤＡＲＹを示す。ＰＲＩＭＡＲＹは、当該整合性グループはデータの送信者で、ＳＥＣＯＮＤＡＲＹは、当該整合性グループはデータの受信者であることを示す。従って、中間サブシステムの整合性グループテーブル５００ｃは、プライマリグループに関連する第一グループテーブル５００ｃ－１はＳＥＣＯＮＤＡＲＹを、セカンダリグループに関連する第二グループテーブル５００ｃ－２はＰＲＩＭＡＲＹを示す。

【００４０】

0050　ＧＲＳＴＳ５３０は整合性グループの状態を示す。一般的に、整合性グループはＣＯＰＹ(コピー中)、ＰＡＩＲ(ペア中)、又はＳＵＳＰ(サスペンド中)の何れかの状態にある。ＣＯＰＹ状態は、現在、データが、一つのボリュームからもう一つのボリュームにコピー中、例えばＰＶＯＬからＳＶＯＬにコピー中、であることを示す。ＰＡＩＲ状態は、例えば、ＰＶＯＬとＳＶＯＬはペア間のミラーが完成し、同一の情報を持つことを示す。ＳＵＳＰ状態は、例えば、ＰＶＯＬからＳＶＯＬへのコピーがサスペンド(中断)している事を示す。

【００４１】

0051　ＩＰＴＲ５４０は、中間サブシステムでペアと成っている整合性グループテーブルへのポインタを含む。中間サブシステムでの本テーブルは、第一テーブル５００ｃ－１と第二テーブル５００ｃ－２を関連つける相応しい値が格納される。即ち、第一テーブル５００ｃ－１のＩＰＴＲ５４０は第二テーブル５００ｃ－２をポイントし、第二テーブル５００ｃ－２のＩＰＴＲ５４０は第一テーブル５００ｃ－１をポイントする。プライマリ又はセカンダリサブシステムは、通常唯一つの整合性グループテーブルしか持たない為、本サブシステムでは本テーブルの本欄の値はＮＵＬＬである。

【００４２】

0052　ＶＯＬＩＮＦＯ＃５５０はボリューム情報を格納する。一ボリュームに一エリアが割り当てられる。ボリューム識別欄ＶＯＬＩＤ５５１は所定のストレージサブシステム１００でのボリューム識別子を表示する。各ボリュームはそのストレージサブシステム１００での固有の番号を持つ。ボリューム状態欄ＶＯＬＳＴＳ５５２は、ボリュームの状態

即ち、ＣＯＰＹ，ＰＡＩＲ，又はＳＵＳＰを示す。シリアル番号欄ＰＡＩＲＤＫＣ５５３は、ペアを組むストレージサブシステム１００の識別子又はシリアル番号を格納する。各ストレージサブシステムは各々固有の番号を持つ。ペアボリューム欄ＰＡＩＲＶＯＬＩＤ５５４は、ＰＡＩＲＤＫＣ５５３で指定されるペアを組むストレージサブシステム１００のボリューム識別子を格納する。

【００４３】

0053　図５は、一実施例による、リモートコピーシステム５０での下記のミラー構成を説明する。各プライマリサブシステムは単一のＰＶＯＬを持つとする。各プライマリサブシステム１００ａは、整合性グループテーブルを持ち、ここでは、

・0054　ＧＩＤ＝０、

・0055　ＧＲＡＴＴＲ＝ＰＲＩＭＡＲＹは、当ボリュームがＰＶＯＬである事を示し、

・0056　ＣＴＧはＶＯＬＩＮＦ＃０で示される如く唯一つのＰＶＯＬを持ち、

・0057　ＩＰＴＲ＝ＮＵＬＬは、本ＣＴＧ項目はＩＶＯＬ１０１ｃにて定義されてはいない事を示す。

【００４４】

0058　各セカンダリサブシステム１００ｂは単一のＳＶＯＬを持つとする。各セカンダリサブシステム１００ｂは整合性グループテーブルを持ち、ここでは、

・0059　ＧＩＤ＝０，は中間サブシステム１００ｃの整合性グループテーブルのＧＩＤと同じであり、

・0060　ＧＲＡＴＴＲ＝ＳＥＣＯＮＤＡＲＹは、当該ボリュームがＳＶＯＬであることを示し、

・0061　本ＣＴＧはＶＯＬＩＮＦＯ＃０に示される如く唯一つのＳＶＯＬを含み、

・0062　ＩＰＴＲ＝ＮＵＬＬは、本ＣＴＧ項目はＩＶＯＬ１０１ｃにて定義されてはいない事を示す。

【００４５】

0063　中間サブシステム１００ｃは、プライマリサブシステム１００ａとセカンダリサブシステム１００ｂに結合する。本中間サブシステム１００ｃは二つの整合性グループテーブルを持つ。ここでは、

・0064　ＧＩＤ＝０は、プライマリサブシステム１００ａの整合性グループテーブルに含まれるのと同じＧＩＤであることを示し、

・0065　ＧＲＡＴＴＲ＝ＳＥＣＯＮＤＡＲＹは、第一テーブル５００ｃ－１に対して表示され、ＧＲＡＴＴＲ＝ＰＲＩＭＡＲＹは第二テーブル５００ｃ－２に対して表示され、

・0066　各テーブルは、ＶＯＬＩＮＦＯ＃０とＶＯＬＩＮＦＯ＃１の二つのボリューム情報欄を含み、

・0067　ＩＰＴＲはＩＶＯＬ１０１ｃでの対応する整合性グループテーブルをポイントする。

【００４６】

0068　図２を参照するに、本発明の一実施例による同期型リモートコピー方法に関するプロセス２００は、プライマリサブシステムと中間サブシステムの間で実行される。本プロセス２００は、書込み要求がプライマリホスト１１０ａからプライマリサブシステム１００ａに到着する事により起動される。

【００４７】

0069　プライマリサブシステムが書込み要求を受信すると、本書込みデータを不揮発(又は安定な)記憶領域に格納する(ステップ２０５)。本不揮発記憶領域は、書込みデータを一時的に記憶する、キャッシュメモリ又は磁気ディスク上に割り当てられた記憶領域で良い。本プライマリサブシステムは次いで、書込みデータを中間サブシステム１００ｃに送信し、中間サブシステム１００ｃからの応答を待つ(ステップ２１０)。中間サブシステム１００ｃがプライマリサブシステム１００ａから書込みデータを受信すると、本サブシ

ステムは本書込みデータの制御データの少なくとも一部を生成する(ステップ２１５)。本実施例では、本生成ステップで、書込み要求のヘッダに対して順序番号を生成しこれを付与する。タイムスタンプが更にヘッダに対して付与されてもよい。

【００４８】

0070 一実施例では、制御データには、シリアル番号２９１(ＰＶＯＬ－ＩＤ２９１)、書込みアドレス２９２(ＡＤＤＲ２９２)、長さ２９３(ＬＥＮ２９３)、時刻２９４、及び順序番号(ＳＥＱ２９５)が含まれる。ＰＶＯＬ－ＩＤ２９１は、プライマリサブシステムと本サブシステム内のＰＶＯＬを特定する。プライマリサブシステムと本ＰＶＯＬはシリアル番号を用いて表示される。

【００４９】

0071 ＡＤＤＲ２９２は、ＰＶＯＬでの書込みの開始アドレスを示し、データが本位置より書き込めるようにする。ＬＥＮ２９３は書込みデータの長さを示す。従って、ＡＤＤＲ２９２とＬＥＮ２９３で書込みデータの正確な領域が指定される。

【００５０】

0072 時刻２９４は、プライマリサブシステム１００ａから中間サブシステム１００ｃに書込み要求が到着した時に制御データに付与されるタイムスタンプである。本タイムスタンプは中間サブシステム１００ｃのタイマ１３０により与えられる。

【００５１】

0073 ＳＥＱ２９５は書込み要求に付与された順序番号である。この値は、当該中間サブシステム１００ｃの整合性グループに関連するカウンタ１２１により提供される。本ステップでは、カウンタ１２１からの値を＋１(つまりカウンタを加算)してカウンタ１２１に戻し、得られた値を書込み要求の制御データに付加する。

【００５２】

0074 プロセス２００に戻って、中間サブシステム１００ｃは、プライマリサブシステム１００ａからの書込みデータとこの制御データを不揮発又は安定な記憶領域に保存する（ステップ２２０）。第一の応答がプライマリサブシステム１００ａに送信される(ステップ２２５)。本第一の応答は、書込みデータが中間サブシステムにコピーされたことを示す。不揮発記憶領域に保存されたデータは、続いてＩＶＯＬ１０１ｃに、より長期的に保存される。プライマリサブシステム１００ａは、第二の応答をプライマリホスト１１０ａに送信する(ステップ２３０)。図示されている様に、プロセス２００では、第二の応答は、第一の応答がプライマリサブシステムで受信されて始めて返送される為、同期型コピー動作である。

【００５３】

0075 図３は、本発明の一実施例による、中間サブシステム１００ｃとセカンダリサブシステム１００ｂとの間の非同期型コピー動作を行うプロセス３００を示す。プロセス３００の特徴は、プライマリホスト１１０ａからプライマリサブシステム１００ａに発行された書込み順序に従って、セカンダリサブシステム１００ｂが受信した書込みデータの書込み順序を維持する事である。このデータインテグリテイの保証は、セカンダリサブシステムが受信した書込みデータを先ずＩＮＶＡＬＩＤとして表示し、次いで有効化することを含む。本有効化プロセスは図４に関連して説明する。

【００５４】

0076 プロセス３００は一定間隔、例えば１０秒毎に起動される。中間サブシステム１００ｃの不揮発記憶領域に保存された書込みデータとその制御データは、セカンダリサブシステム１００ｂに送信される(ステップ３０１)。本発明では、送信する書込みデータは対応する制御データのＳＥＱ２９５の昇順に従って選択される。

【００５５】

0077 セカンダリサブシステム１００ｂは、書込みデータとその制御データを保存し、当該データをＩＮＶＡＬＩＤとしてマークする(ステップ３０５)。セカンダリサブシステムは、中間サブシステム１００ｃに応答を返す(ステップ３１０)。

【００５６】

0078　本応答を受け取ると、中間サブシステム１００ｃは、有効カウンタ１２２の現在値が送信済みの制御データのＳＥＱ２９５より小さいかを判定する(ステップ３１５)。小さければこの値は更新される(ステップ３２０)。プロセス３００は次いで、再度起動される迄、一定時間休眠する(ステップ３２５)。有効カウンタ１２２は、かくして、セカンダリサブシステム１００ｂで有効化可能な書込みデータの最新の順序番号を保持する事になる。別の言い方をすれば、有効カウンタ１２２は、セカンダリサブシステムに保存されたＩＮＶＡＬＩＤデータ(有効化待ちのデータ)の最大順序番号数を示す、と言える。

【００５７】

0079　図４は、本発明の一実施例による、セカンダリサブシステム１００ｂでのデータ有効化の為のプロセス４００を示す。本プロセス４００はＰＲＥＰＡＲＥ及びＶＡＬＩＤＡＴＥ状態即ちフェーズを使用する。ＰＲＥＰＡＲＥフェーズでは、全セカンダリサブシステムは、ロールバック動作に必要な回復用情報を保持したまま、中間サブシステム１００ｃより受信したデータを有効化する為の全ての必要な動作を実行する。セカンダリサブシステム１００ｂがプロセス４００の処理を実行できない条件が発生すれば、本回復情報を用いてロールバックする。ＶＡＬＩＤＡＴＥフェーズでは、セカンダリサブシステム１００ｂは、データを有効化し回復情報を破棄する。これら二つのフェーズと回復機構は、所定のデータに対する有効化プロセス４００がセカンダリサブシステム１００ｂの何れかで障害になるかもしれない為、有益なものである。

【００５８】

0080　本発明の一実施例では、プロセス４００は中間サブシステムにて一定時間間隔で起動される(ステップ４０１)。有効カウンタ１２２は、その値が前回のプロセス以降変化があったか否かを決定するために、チェックされる（ステップ４０２）。変化がなかったら、プロセスはステップ４０１に戻り、所定の時間休眠する。有効カウンタ１２２が変化しなかったと言うことは、前回のセッション以後、セカンダリサブシステムには書込みデータが送られていないことを示す。値が変化していれば、プロセスは先に進む。

【００５９】

0081　中間サブシステム１００ｃは、有効カウンタ１２２の現在値と共に、全てのセカンダリサブシステム１００ｂに対してＰＲＥＰＡＲＥ要求を送信する(ステップ４０３)。この要求により、セカンダリサブシステム１００ｂは、送られてきた有効カウンタ１２２以下の順序番号を持つデータについて有効化作業を開始する。中間サブシステム１００ｃは、次いで、セカンダリサブシステム１００ｂからの応答を待つ(ステップ４０６)。

【００６０】

0082　セカンダリサブシステムでは、ＰＲＥＰＡＲＥ要求の受信を契機に、ＩＮＶＡＬＩＤとしてマークされ、有効カウンタの現在値以下の順序番号を割り当てられたデータの有効化作業を開始する（ステップ４０５）。現在値は基準値として参照される。データはＰＲＥＰＡＲＥＤとしてマークされる(ステップ４１０)。上記のステップには、データを不揮発ストレージの一時的領域から長期的領域にコピーし、本一時的領域を開放し、制御データを開放し、ＩＮＶＡＬＩＤ及びＰＲＥＰＡＲＥＤ状態に関する情報を更新することを含む。

【００６１】

0083　データを回復可能に維持する為に、セカンダリサブシステムは旧データに上書きしないで新データが有効化される迄、旧データを保持する。一つの方法は、新データを永久ストレージ(例えばディスク)の最終ターゲットに書き込まないことである。セカンダリサブシステム１００ｂは、上記が成功裏に実行できたらデータをＰＲＥＰＡＲＥＤとしてマークする。

【００６２】

0084　セカンダリサブシステムは、全ての指定された書込みデータがＰＲＥＰＡＲＥＤとしてマークされたか否かを判定する(ステップ４１５)。判定が成立すれば、セカンダリサブシステム１００ｂは、中間サブシステム１００ｃに全ての指定されたデータはＰＲＥＰＡＲＥＤとしてマークされたことを示す応答を返す(ステップ４２０)。セカンダリサブ

システム１００ｂは、タイマをセットして、ＶＡＬＩＤＡＴＥ要求が中間サブシステム１００ｃから受信されるのを待つ(ステップ４２１)。

【００６３】

0085 中間サブシステム１００ｃが、全てのセカンダリサブシステム１００ｂから、全ての指定された書込みデータがＰＲＥＰＡＲＥＤとしてマークされたことを示す応答を受信したら、中間サブシステムはセカンダリサブシステム１００ｂにＶＡＬＩＤＡＴＥ要求を送信する(ステップ４２５)。中間サブシステムは、次いで、セカンダリサブシステムからＶＡＬＩＤＡＴＥ要求に対する応答を待つ（ステップ４２６）。

【００６４】

0086 セカンダリサブシステムに於いては、各サブシステムは、ＰＲＥＰＡＲＥＤとしてマークされたデータをＶＡＬＩＤに更新して、当該データの回復に関する情報を破棄する(ステップ４３０)。旧データは通常この時点で破棄される。各セカンダリサブシステムは中間サブシステム１００ｃに応答を返す(ステップ４３５)。

【００６５】

0087 ステップ４１５に戻って、幾つかの書込みデータをＰＲＥＰＡＲＥＤとしてマークすることに失敗した場合には、ＦＡＩＬＵＲＥ通知が中間サブシステムに送信される(ステップ４１５)。中間サブシステムは、以下により詳細に記述される、エラー処理を実行する(ステップ４２４)。

【００６６】

0088 プロセス４００は、参照順序番号、即ち有効カウンタ１２２の値、以下の順序番号を持つ書込みデータを有効化するのに用いられる。即ち、セカンダリサブシステム１００ｂのＳＶＯＬ１０１ｂの集合は、ＳＶＯＬ１０１ｂを通して整合しており、参照順序番号以下の順序番号で送信された全てのデータが受信され有効化されたことをセカンダリサブシステムが表示した以降には、書込みデータの損失は存在しない。

【００６７】

0089 時によっては、プロセス４００は失敗し回復処理が必要になる。障害の例は、ステップ４２４で、書込みデータのＰＲＥＰＡＲＥ操作のときに発生する。本実施例では、回復処理は中間サブシステムで実施される。

【００６８】

0090 一般的に、中間サブシステムは、ステップ４２４でのエラーか、セカンダリサブシステムに要求を出してからのタイムアウト(ステップ４０６か４２６)で障害発生を検出する。

【００６９】

0091 障害がＰＲＥＰＡＲＥフェーズで発生した場合には、中間サブシステム１００ｃは、セカンダリサブシステム１００ｂにＡＢＯＲＴ要求を送信し、セカンダリサブシステムはロールバックを行い、ＰＲＥＰＡＲＥＤとしてマークされたデータの回復は、本データとこれの制御データをＩＮＶＡＬＩＤに戻すことにより開始される。

【００７０】

0092 ＶＡＬＩＤＡＴＥフェーズの場合には、障害はセカンダリサブシステムからの応答が既定時間内に受信できない場合に発生する。このような事態は起きそうもないが、もし実際に起きたなら、中間サブシステムはＶＡＬＩＤＡＴＥ要求を再発行する。

【００７１】

0093 中間サブシステム１００ｃが、ＰＲＥＰＡＲＥかＶＡＬＩＤＡＴＥフェーズ中にダウンした場合には、セカンダリサブシステム１００ｂは、リブートされた中間サブシステム１００ｃかホスト１００の一つからのコマンドを待つ。一般に、リブート又は回復した中間サブシステム１００ｃは、プロセス４００を最初からではなく、最新の成功ステップから再開始する。

【００７２】

0094 図８は、本発明の一実施例に従い、プロセス４００が障害なく成功裏に実行された場合のプロセスフローを示す。第一の状態８００で示すように、中間サブシステム１０

０ｃは、セカンダリサブシステム１００ｂ－１，１００ｂ－２、及び１００ｂ－３にＰＲＥＰＡＲＥ要求を発行する(ステップ８００)。本ＲＥＰＡＲＥ要求は、１０４２までの順序番号の書込みデータをＰＲＥＰＡＲＥすることをセカンダリサブシステムに要求する。

【００７３】

0095　第一の状態８００でＰＲＥＰＡＲＥ要求を受信する前には、セカンダリサブシステム１００ｂの第一、第二、及び第三の整合性グループ１２０－１，１２０－２、及び１２０－３は、１０２４までの順序番号の全ての書込みデータはＰＲＥＰＡＲＥＤになっていることを示す。１０１０までの順序番号の全ての書込みデータはＶＡＬＩＤになっている。１０２４より高い順序番号の書込みデータはＩＮＶＡＬＩＤが表示されている。例えば、第一の整合性グループ１２０－１は、サブシステム１００ｂ－１はＩＮＶＡＬＩＤの順序番号１０２９の書込みデータを持っている、ことを示す。

【００７４】

0096　第二の状態８１０は、ＰＲＥＰＡＲＥ要求が完了した後の整合性グループを示す。第一の整合性グループ１２０－１は順序番号１０４２までの書込みデータはＰＲＥＰＡＲＥされていることを示す。従って、順序番号１０２９までの書込みデータは最早ＩＮＶＡＬＩＤではなく、整合性グループ１２０－１は、サブシステム１００ｂ－１はＩＮＶＡＬＩＤデータを持っていないことを示す。整合性グループは、順序番号１０１０迄の書込みデータはＶＡＬＩＤのままであることを示す。

【００７５】

0097　中間サブシステム１００ｃはＶＡＬＩＤＡＴＥ要求をセカンダリサブシステムに送信し、ＰＲＥＰＡＲＥされた書込みデータは本要求によりＶＡＬＩＤになる。第三の状態８２０は、ＶＡＬＩＤＡＴＥ要求がセカンダリサブシステム１００ｂにて実行された後の整合性グループを示す。本整合性グループは、全てのＰＲＥＰＡＲＥされた書込みデータはＶＡＬＩＤになった事を示す。即ち、１０４２迄の順序番号の全ての書込みデータはＶＡＬＩＤになっている。

【００７６】

0098　図９は、プロセス４００で障害が発生した場合を示す。第一の状態９００は、図８の第一の状態８００のミラーイメージである。中間サブシステム１００ｃはセカンダリサブシステム１００ｂにＰＲＥＰＡＲＥ要求を発行する。

【００７７】

0099　第二の状態９１０は、ＰＲＥＰＡＲＥ処理がセカンダリサブシステムの１つ、即ちセカンダリサブシステム１００ｂ－２で成功裏には実行できなかったことを示す。セカンダリサブシステム１００ｂ－２は、要求されたすべての、即ち順序番号１０４２迄の、書込みデータを成功裏にはＰＲＥＰＡＲＥできなかった。セカンダリサブシステム１００ｂ－２の整合性グループ１２０－２は、順序番号１０３８迄しか書込みデータはＰＲＥＰＡＲＥされていないことを示す。本ＰＲＥＰＡＲＥ処理はセカンダリサブシステム１００ｂ－１と１００ｂ－３では成功裏に終了している。これらの整合性グループは、１０４２迄はＰＲＥＰＡＲＥされている。

【００７８】

0100　セカンダリサブシステム１００ｂ－２の障害の結果、中間サブシステムはセカンダリサブシステムに参照順序番号１０３８（１０４２でなく）でＶＡＬＩＤＡＴＥ要求を発行する。本ＶＡＬＩＤＡＴＥ要求は、セカンダリサブシステム１００ｂ－２での書込みデータのＰＲＥＰＡＲＥ処理での問題を反映して、修正されている。

【００７９】

0101　第三の状態９２０は、修正ＶＡＬＩＤＡＴＥ要求がセカンダリサブシステム１００ｂで成功裏に実行できたことを示す。整合性グループ１２０－１と１２０－３は、依然として、ＰＲＥＰＡＲＥＤだがＶＡＬＩＤでない書込みデータを持っていることを示す。然しながら、１２０－２の整合性グループは、全てのＰＲＥＰＡＲＥＤデータはＶＡＬＩＤになっていることを示す。ＶＡＬＩＤになった書込みデータは、順序番号１０３８迄のものであることを示す。

【００８０】
0102 図１０Ａ（図１０(a)）と１０Ｂ（図１０(b)）は、本発明の一実施例に従う、プライマリサイトからセカンダリサイトへのフェイルオーバ処理を示す。フェイルオーバはプライマリサイトがユーザに対して使用不能になったときに実行される。これは、プライマリサブシステム１００ａかプライマリホスト１１０ａの何れか又は何れも使用不能になったときに発生する。プライマリサイトが障害になる、又は保守の為に一時的にオフラインにされた場合には、セカンダリサイトがアプリケーションを実行可能になる。
【００８１】
0103 一般的に、セカンダリサブシステムのボリューム、即ちＳＶＯＬ１０１ｂは、プライマリサブシステム１００ａにペア又はミラーしている場合には、書込み不能である。従って、本発明の一実施例に拠れば、フェイルオーバではプライマリサブシステムへのミラー動作はサスペンドする必要がある。
【００８２】
0104 図１０(a)は、プライマリサイトのみが使用不能で、中間サブシステム１００ｃが使用可能な場合でのフェイルオーバ処理を示す。この状況では、ＩＶＯＬ１０１ｃがプライマリサブシステムから受信した全てのデータを保持している為に、データは損失していない。これらのデータは、中間サブシステム１００ｃがＳＶＯＬ１０１ｂに転送する。この後に、セカンダリサブシステムは使用可能になる。
【００８３】
0105 最初にユーザが、セカンダリホスト１００ｂを使用してフェイルオーバを起動する(ステップ１０００)。セカンダリサブシステム１００ｂはＦＡＩＬＯＶＥＲ要求を中間サブシステム１００ｃに送信する。中間サブシステム１００ｃは全ての未転送データをセカンダリサブシステム１００ｂに送信する(ステップ１００５)。本発明の一実施例では、中間サブシステムからセカンダリサブシステムにデータを送信するのに、プロセス３００が使用される。中間サブシステムは、全ての未転送データがセカンダリサブシステムに転送されたら、ペア状態をＰＳＵＳに変更する(ステップ１００６)。ＰＳＵＳはペア動作がサスペンドしていることを示す。ひとたびセカンダリサブシステムがＦＡＩＬＯＶＥＲ要求の完了を知らされると、セカンダリサブシステム１００ｂは、ペア状態をＰＳＵＳに変更する(ステップ１０１０)。
【００８４】
0106 図１０(b)は中間サブシステム１００ｃが使用不能の場合のフェイルオーバ動作を示す。プライマリサイトは使用不能でも可能でもよい。この状況では、中間サブシステムが、セカンダリサブシステム１００ｂへの未転送データを保持している可能性がある為、データは損失する可能性がある。ユーザは、セカンダリホスト１１０ｂを使用して、フェイルオーバ動作を起動して、セカンダリサブシステムにＦＡＩＬＯＶＥＲ要求を送信する(ステップ１０２０)。中間サブシステム１００ｃが使用不能の為、ＦＡＩＬＯＶＥＲ要求は失敗して、セカンダリサブシステム１００ｂはセカンダリホストにエラーメッセージを返す(ステップ１０２５)。
【００８５】
0107 セカンダリホスト１１０ｂは、全てのセカンダリサブシステムの整合性グループからある種の情報を取得する(ステップ１０３０)。これらの情報には、ＩＮＶＡＬＩＤデータ、ＰＲＥＲＡＲＥＤデータ、及びＶＡＬＩＤデータの各順序番号が含まれる。全てのＰＲＥＲＡＲＥＤデータは、プロセス４００によりＶＡＬＩＤデータに更新する(ステップ１０３５)。
【００８６】
0108 セカンダリホスト１１０ｂは、有効化プロセスが成功したことを確認後、セカンダリサブシステム１００ｂのペア状態をＰＳＵＳに変更開始する(ステップ１０４０)。セカンダリサブシステムの状態はＰＳＵＳに変更される(ステップ１０４５)。
【００８７】
0109 以上で説明したように、リモートコピーシステム５０は三つのデータセンタを持

っている。プライマリと中間サブシステムの双方が同時に障害にならない限り、データは損失しない。プライマリとセカンダリのデータセンタが各々複数のサブシステムを持っていてもデータの整合性は保証される。中間サブシステムは、プライマリサブシステムからの全ての書込みデータを受信しセカンダリサブシステムに転送する。中間サブシステムは、複数のプライマリサブシステムからの書込みデータの書込み順を保持する為の管理情報（例えば、順序番号）を生成する。

【００８８】

0110 図６は、本発明のもう一つの実施例に従う、リモートコピーシステム５０’を示す。システム５０’は、図１のシステム５０に比べて、中間サブシステム１００ｃ’が減少したボリューム数で構成されている。アプリケーションが中間サブシステムにフェイルオーバしないなら、中間サブシステムはプライマリサイトのプライマリボリュームより少数のボリュームで構成することができる。ボリューム数を減らす事でコストを低減できる。中間サブシステム１００ｃ’は、プライマリサブシステム１００ａから受信したデータを保存するために、一つのボリュームを備えている。中間サブシステム１００ｃ’がボリュームを持たないで、一時的にデータを保存する為にキャッシュメモリを使用する、実施例も存在する。

【００８９】

0111 中間サブシステム１００ｃ’はジャーナルボリューム(ＪＮＬＶＯＬ)７００を持つ。ＪＮＬＶＯＬ７００はＩＶＯＬ１０１ｃに相当する。システム５０のＩＶＯＬ１００ｃとは異なり、ＪＮＬＶＯＬ７００はＰＶＯＬ１０１ａのミラーではない。ＪＮＬＶＯＬ７００は、"Remote Copy System"と標題を有し、参照としてここに含まれる米国特許ＮＯ．１０／６０２,２２３（２００３／６／２３出願受理）にてより詳しく説明されているように、プライマリサブシステム１００ａから受信した書込みデータとその制御データの為のバッファ又は一時的記憶領域として働く。

【００９０】

0112 図７は、本発明の一実施例に従う、リモートコピーシステム５０’に対する整合性グループテーブル７０２を示す。本整合性グループテーブル７０２は、ボリューム情報(ＶＯＬＩＮＦＯ#)７５０、ジャーナルボリューム情報(ＪＮＬ＿ＶＯＬＩＤ#)７６０、ジャーナル入力情報(ＪＮＬＩＮ)７７０、及びジャーナル出力情報(ＪＮＬＯＵＴ)７８０を含む。ＶＯＬＩＮＦＯ#７５０は、中間サブシステム１００ｃで定義されるＩＶＯＬ１０１ｃは存在しない事を示す為に、ＮＵＬＬを保存するボリュームＩＤ(ＶＯＬＩＤ)７５１を含む。同様に、プライマリサブシステム１００ａとセカンダリサブシステム１００ｂで定義される、ＰＡＩＲＶＯＬＩＤ(図示していない)は、関連する中間サブシステムで定義されるＩＶＯＬ１０１ｃが存在しない為、ＮＵＬＬを表示する。

【００９１】

0113 ＪＮＬ＿ＶＯＬＩＤ７６０は、プライマリサブシステム１００ａより送信された書込みデータと制御データを保存する為に使用されるボリュームを識別する情報を含む。ＪＮＬＩＮ７７０は、ＪＮＬ＿ＶＯＬＩＤ７６０エントリの識別子とジャーナルボリューム７００でのアドレスを含む。ＪＮＬＩＮ７７０は、次の書込みデータと制御データのセットが保存されるアドレスを表示する。ＪＮＬＯＵＴ７８０は、ＪＮＬ＿ＶＯＬＩＤ７６０エントリの識別子とジャーナルボリューム７００でのアドレスを含む。ＪＮＬＯＵＴ７８０は、次の書込みデータと制御データのセットがセカンダリサブシステム１００ｂに送信されるソースアドレスを表示する。有効なＪＮＬ＿ＶＯＬＩＤ７６０エントリが存在しない（例えば、全てのＪＮＬ＿ＶＯＬＩＤ７６０がＮＵＬＬ）場合には、整合性グループにはジャーナルボリューム７００は割り当てられていない。

【００９２】

0114 リモートコピーシステム５０’では、図２のプロセス２００と実質的に同じプロセスを使用して同期コピー動作を実行する。一つの相違点はステップ２２０関するもので整合性グループテーブルを参照して不揮発記憶媒体を選択することである。本操作は、(１) ＩＶＯＬ１０１ｃ(即ち、ＶＯＬＩＤ７５１！＝ＮＵＬＬか)とＪＮＬＶＯＬ７００（

即ち、ＪＮＬ＿ＶＯＬＩＤ７６０！＝ＮＵＬＬか）が存在するかをチェックする；(２)(１)の結果に従い、下記の手順を実行する。

【００９３】

・0115　プライマリサブシステム１００ａから受信した書込みデータと制御データを中間サブシステム１００ｃのキャッシュメモリに格納する、

・0116　ＪＮＬＶＯＬ７００が存在すれば、書込みデータと制御データをＪＮＬＩＮ７７０で指定されるＪＮＬＶＯＬ７００のアドレスに格納する、

・0117　ＩＶＯＬ１０１ｃが存在すれば、書込みデータをＩＶＯＬ１０１ｃに格納する。

【００９４】

0118　一般的に、中間サブシステム１００ｃは、同期コピー動作を促進する為に、書込みデータをキャッシュメモリに格納し次第、プライマリサブシステム１００ａに応答を返す。

【００９５】

0119　リモートコピーシステム５０’では、図３のプロセス３００と実質的に同じプロセスを使用して非同期コピー動作を実行する。相違点の一つは、図３のステップ３０１に関し、データをセカンダリサブシステム１００ｂに送信すべく選択する場合に、

0120　(１)ＩＶＯＬ１０１ｃ(即ち、ＶＯＬＩＤ７５１！＝ＮＵＬＬか)とＪＮＬＶＯＬ７００(即ち、ＪＮＬ＿ＶＯＬＩＤ７６０！＝ＮＵＬＬか)が存在するかをチェックし、

0121　(２)(１)の結果に従い、下記の手順を実行する。

【００９６】

・0122　キャッシュメモリから最小の順序番号を持つデータを選択する。

【００９７】

・0123　キャッシュメモリにデータが存在せず、更にＪＮＬＶＯＬ７００が存在すれば、ＪＮＬＶＯＬ７００からデータを抽出する。更に、ＪＮＬＶＯＬ７００が存在しなければ、セカンダリサブシステムはミラー処理を維持できず、状態をＰＳＵＳにする。

【００９８】

0124　リモートコピーシステム５０’は、(１)プライマリサブシステム１００ａからの書込みデータとその制御データを不揮発のランダムメモリ(ＮＶＲＡＭ)のみに格納し、(２)書込みデータとその制御データをＪＮＬＶＯＬ７００に格納する、ことによって、中間サブシステムでのボリューム数を低減できる。ＮＶＲＡＭのみを使用するなら中間サブシステムには、ボリュームは不要である。このような構成では、ＮＶＲＡＭの容量が限られ(例えば数ＧＢ)、又ＮＶＲＡＭの容量を増加する事は高価につく為、プライマリボリュームのミラーを維持する事は困難である。ＮＶＲＡＭと共にＪＮＬＶＯＬ７００を使用することは、書込みデータとその制御データをＪＮＬＶＯＬ７００に保存できる為に、より受け入れやすい構成である。この後者の構成でも、必要なボリュームはＩＶＯＬ１０１ｃに比べて低減できる。

【００９９】

0125　図１１は、本発明の一実施例に従う、ミラー動作のサスペンド終了後の再同期時間を最小にするリモートコピーシステム５０”を示す。システム５０”は、プライマリサブシステムに配備される複数のビットマップ１１００ａ、セカンダリサブシステムに配備される複数のビットマップ１１１０ｂ、及び中間サブシステムに配備されるビットマップ１１００ｃと１１１０ｃを備える。

【０１００】

0126　一般的に、ミラー動作は、プライマリと中間サブシステムの間、又はセカンダリと中間サブシステムの間の結合が障害になるか、人的に中断されると、サスペンドされる。障害が回復するかユーザがサスペンドしたミラーボリュームの再同期の指示をしたときに、ミラー動作は回復するか再同期される。再同期に於いては、再同期時間を最小にする為に、サスペンド後に変更があったデータのみが、プライマリからセカンダリサブシステムにコピーされる。本実施例では、この目的の為にビットマップを使用する。同様に、サ

スペンド後に変更があったデータのみが中間からセカンダリサブシステムにコピーされる。

【０１０１】

0127 ミラーボリュームの各ペアに二つのビットマップが用意される。ビットマップ１１００はプライマリボリュームにアサインされ、ビットマップ１１１０はセカンダリボリュームにアサインされる。従って、本発明の一実施例では、ビットマップ１１００ａはプライマリサブシステムの各プライマリボリュームにアサインされる。対応するビットマップ１１１０ｃは中間サブシステムでのセカンダリボリューム(又はセカンダリボリュームとして稼動している中間ボリューム１０１ｃ)にアサインされる。

【０１０２】

0128 中間サブシステムでは、中間ボリュームは、セカンダリサブシステムに対してプライマリボリュームである為、ビットマップ１１００ｃを有する。セカンダリサブシステムは、中間サブシステムのビットマップ１１００ｃに対応するビットマップ１１１０ｂを有する。

【０１０３】

0129 ビットマップ１１００と１１１０は、ミラーがサスペンドしてからのボリュームになされた変更を追跡する。これらビットマップは、対応するミラーがサスペンドして初めて活性化する。例えば、ミラー(ＰＶＯＬ１０１ａ、ＩＶＯＬ１０１ｃ)のみがサスペンドしている場合には、ＰＶＯＬ１０１ａのビットマップ１１００ａとＩＶＯＬ１０１ｃのビットマップ１１１０ｃのみが活性化され、ＩＶＯＬ１０１ｃのビットマップ１１００ｃとＳＶＯＬ１０１ｂのビットマップ１１１０ｂは非活性のままである。

【０１０４】

0130 再同期化に於いては、全ての活性化しているビットマップを、再同期化するソースストレージサブシステムに転送し、次いでマージ（ＯＲ）される。ソースストレージサブシステムは、そのときのアプリケーションに従って、プライマリサブシステムか中間サブシステムである。再同期でのソースストレージサブシステムは、マージされたビットマップに基づいてデータをコピーする。例えば、図１２に於いては、ＰＶＯＬ１０１ａとＩＶＯＬ１０１ｃ間のミラーは継続され、ＩＶＯＬ１０１ｃとＳＶＯＬ１０１ｂがサスペンドしている。再同期化では、中間サブシステム１０１ｃがセカンダリサブシステムのビットマップ１１１０ｂを取得して、これをＩＶＯＬ１０１ｃのビットマップ１１００ｃとマージ（ＯＲ）する。マージされたビットマップに基づく差分データがＩＶＯＬ１０１ｃからＳＶＯＬ１０１ｂにコピーされる。リモートコピーシステム５０”では、２セットの異なったビットマップが用意されているので、第一のミラー(ＰＶＯＬ１０１ａとＩＶＯＬ１０１ｃ)は、第二のミラー(ＩＶＯＬ１０１ｃとＳＶＯＬ１０１ｂ)とは独立に再同期化される。

【０１０５】

0131 これまでの詳細な記述は、本発明の具体的実施例を説明する為のもので、制約を加えることを意図しているものではない。本発明の範囲内で多数の改造と変形が可能である。従って、本発明は、添付の請求範囲によって定義されるものである。

【図面の簡単な説明】

【０１０６】

【図１（ａ）】0019 図１(a)は、本発明の一実施例による、複数のプライマリサブシステムを持つリモートコピーシステムを示す。

【図１（ｂ）】0020 図１(b)は、データの読み書き要求を処理するストレージコントローラと、本書込み要求に従って、データを保存する記録媒体を持つストレージユニットを含むストレージサブシステムの好適な一実施例を示す。

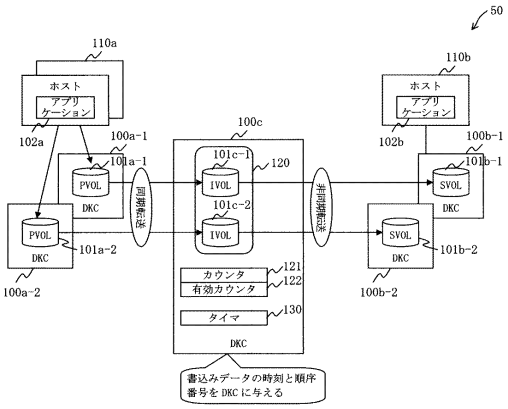
【図２】0021 図２によれば、本発明の一実施例による、同期的リモートコピー方法に関するプロセスは、プライマリサブシステムと中間サブシステムとの間で実行される。

【図３】0022 図３は、本発明の一実施例による、中間サブシステムとセカンダリサブシステムとの間の非同期リモートコピー動作の為のプロセスを説明する。

【図４】0023　図４は、本発明の一実施例による、セカンダリサブシステムでのデータ有効化の為のプロセスを説明する。
【図５】0024　図５は、本発明の一実施例による、リモートコピーシステムの為のミラー構成を説明する。
【図６】0025　図６は、本発明のもう一つの実施例による、リモートコピーシステムを説明する。
【図７】0026　図７は、本発明の一実施例による、リモートコピーシステムの為の整合性グループテーブルを説明する。
【図８】0027　図８は、本発明の一実施例による、障害なしに成功裏に実行されたプロセスのプロセスフローを示す。
【図９】0028　図９は、プロセス中に障害が発生した場合のプロセスフローを示す。
【図１０（ａ）】0029　図１０(a)は、プライマリサイトのみが使用不能で、中間サブシステムが使用可能な場合のフェイルオーバ動作を示す。
【図１０（ｂ）】0030　図１０(b)は、中間サブシステムが使用不能になった場合のフェイルオーバ動作を示す。
【図１１】0031　図１１は、本発明の一実施例による、ミラー動作がサスペンドした後の再同期時間を最小限にするリモートコピーシステムを示す。
【図１２】0032　図１２は、ＰＶＯＬ１０１ａとＩＶＯＬ１０１ｃがミラー継続し、ＩＶＯＬ１０１ｃとＳＶＯＬ１０１ｂ間のミラーがサスペンドした状態を説明する。
【符号の説明】
【０１０７】
１００a-1、１００a-2、１００b-1、１００b-2、１００c…ＤＫＣ
１０１a-1、１０１a-2…ＰＶＯＬ
１０１b-1、１０１b-2…ＳＶＯＬ
１０１c-1、１０１c-2…ＩＶＯＬ
１０２a、１０２b…アプリケーション
１１０a、１１０b…ホスト
１２１…カウンタ
１２２…有効カウンタ
１３０…タイマ
５０…リモートコピーシステム
６０…ストレージサブシステム
６２…ストレージコントローラ
６３…ストレージユニット
６４…ホストチャネルアダプタ
６６…ストレージチャネルアダプタ
６８…ディスクアダプタ
７０…キャッシュメモリ

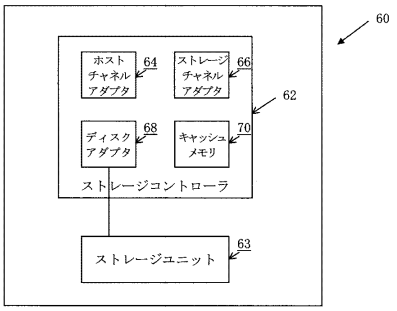
【図１（ａ）】

【図１(a)】



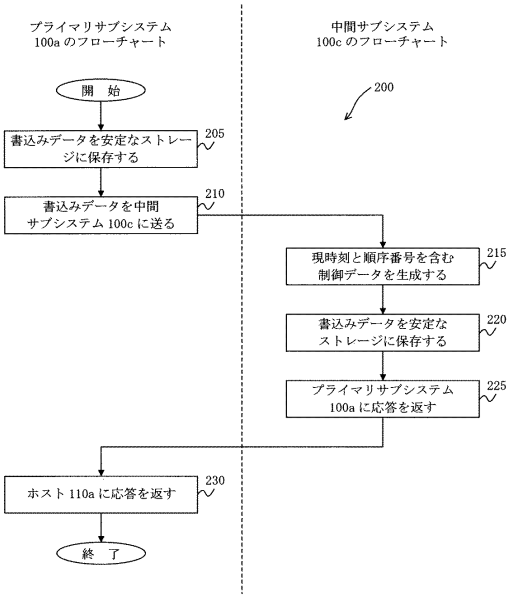
【図１（ｂ）】

【図１(b)】



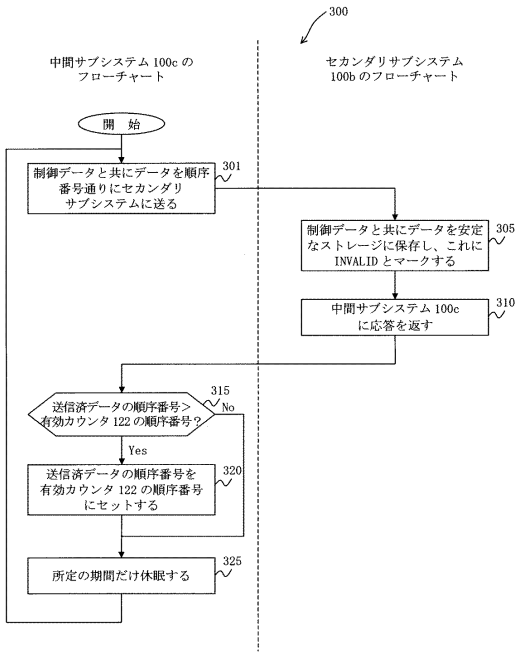
【図２】

【図２】



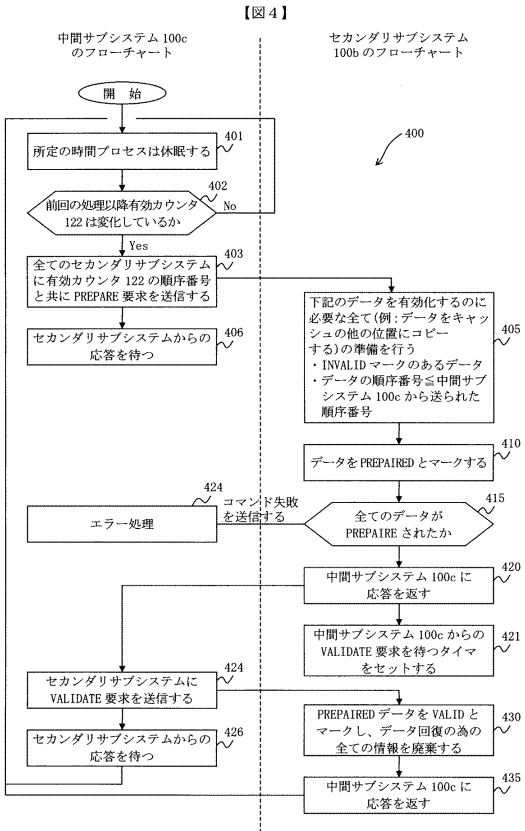
【図３】

【図３】

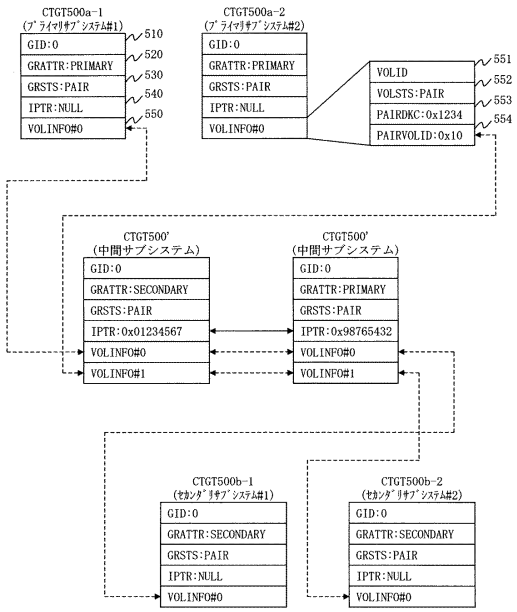


【図４】

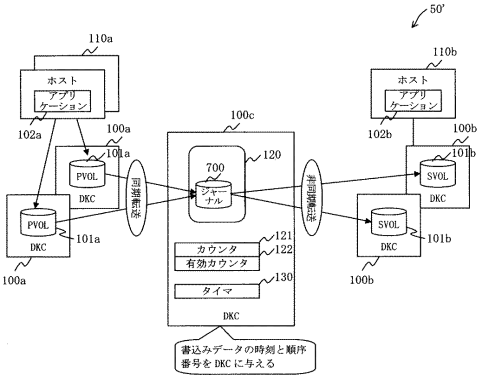
【図４】



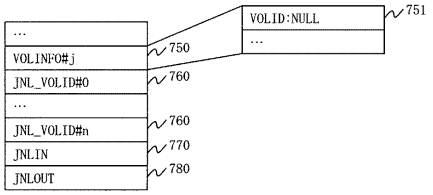
【図５】


【図５】
CTGT500a-1
(プライマリサブシステム#1)
GID:0
GRATTR:PRIMARY
GRSTS:PAIR
IPTR:NULL
VOLINFO#0
510
520
530
540
550
CTGT500a-2
(プライマリサブシステム#2)
GID:0
GRATTR:PRIMARY
GRSTS:PAIR
IPTR:NULL
VOLINFO#0
VOLID
VOLSTS:PAIR
PAIRDKC:0x1234
PAIRVOLID:0x10
551
552
553
554
CTGT500'
(中間サブシステム)
GID:0
GRATTR:SECONDARY
GRSTS:PAIR
IPTR:0x01234567
VOLINFO#0
VOLINFO#1
CTGT500'
(中間サブシステム)
GID:0
GRATTR:PRIMARY
GRSTS:PAIR
IPTR:0x98765432
VOLINFO#0
VOLINFO#1
CTGT500b-1
(セカンダリサブシステム#1)
GID:0
GRATTR:SECONDARY
GRSTS:PAIR
IPTR:NULL
VOLINFO#0
CTGT500b-2
(セカンダリサブシステム#2)
GID:0
GRATTR:SECONDARY
GRSTS:PAIR
IPTR:NULL
VOLINFO#0

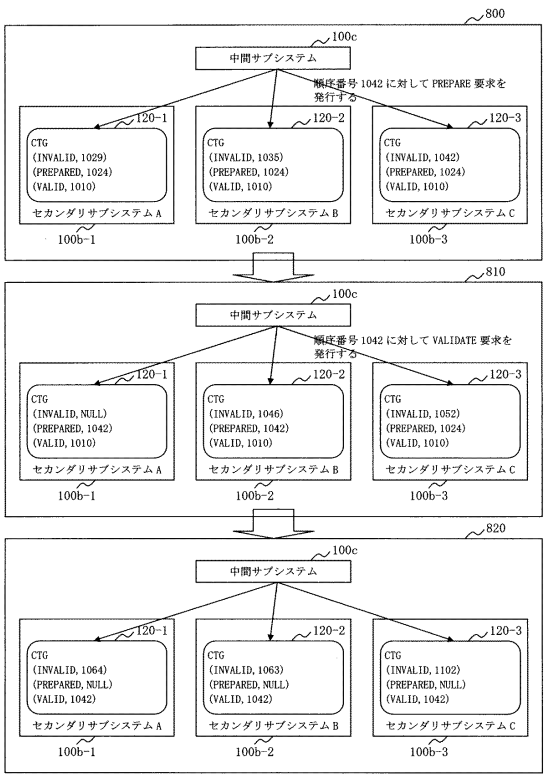
【図６】


【図６】
50'
110a
ホスト
アプリケーション
102a
100a
101a
PVOL
DKC
PVOL
DKC
100a
同期転送
100c
700
ジャーナル
120
カウンタ
有効カウンタ
121
122
タイマ
130
DKC
非同期転送
110b
ホスト
アプリケーション
102b
100b
101b
SVOL
DKC
SVOL
DKC
101b
100b
非込みデータの時刻と順序番号をDKCに与える

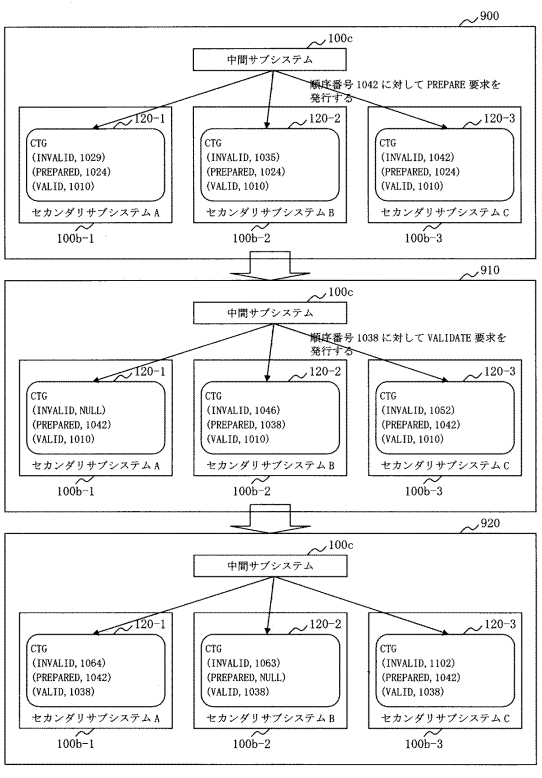
【図７】


【図７】
...
VOLINFO#j
JNL_VOLID#0
...
JNL_VOLID#n
JNLIN
JNLOUT
750
760
760
770
780
VOLID:NULL
...
751

【図８】

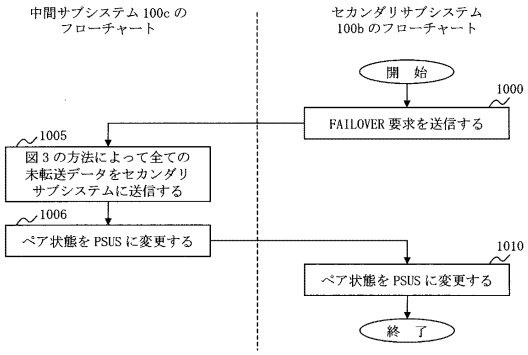

【図８】
800
100c
中間サブシステム
順序番号1042に対してPREPARE要求を発行する
120-1
CTG
(INVALID, 1029)
(PREPARED, 1024)
(VALID, 1010)
120-2
CTG
(INVALID, 1035)
(PREPARED, 1024)
(VALID, 1010)
120-3
CTG
(INVALID, 1042)
(PREPARED, 1024)
(VALID, 1010)
セカンダリサブシステムA
セカンダリサブシステムB
セカンダリサブシステムC
100b-1
100b-2
100b-3
810
100c
中間サブシステム
順序番号1042に対してVALIDATE要求を発行する
120-1
CTG
(INVALID, NULL)
(PREPARED, 1042)
(VALID, 1010)
120-2
CTG
(INVALID, 1046)
(PREPARED, 1042)
(VALID, 1010)
120-3
CTG
(INVALID, 1052)
(PREPARED, 1024)
(VALID, 1010)
セカンダリサブシステムA
セカンダリサブシステムB
セカンダリサブシステムC
100b-1
100b-2
100b-3
820
100c
中間サブシステム
120-1
CTG
(INVALID, 1064)
(PREPARED, NULL)
(VALID, 1042)
120-2
CTG
(INVALID, 1063)
(PREPARED, NULL)
(VALID, 1042)
120-3
CTG
(INVALID, 1102)
(PREPARED, NULL)
(VALID, 1042)
セカンダリサブシステムA
セカンダリサブシステムB
セカンダリサブシステムC
100b-1
100b-2
100b-3

【図９】


【図９】
900
100c
中間サブシステム
順序番号1042に対してPREPARE要求を発行する
120-1
CTG
(INVALID, 1029)
(PREPARED, 1024)
(VALID, 1010)
120-2
CTG
(INVALID, 1035)
(PREPARED, 1024)
(VALID, 1010)
120-3
CTG
(INVALID, 1042)
(PREPARED, 1024)
(VALID, 1010)
セカンダリサブシステムA
セカンダリサブシステムB
セカンダリサブシステムC
100b-1
100b-2
100b-3
910
100c
中間サブシステム
順序番号1038に対してVALIDATE要求を発行する
120-1
CTG
(INVALID, NULL)
(PREPARED, 1042)
(VALID, 1010)
120-2
CTG
(INVALID, 1046)
(PREPARED, 1038)
(VALID, 1010)
120-3
CTG
(INVALID, 1052)
(PREPARED, 1042)
(VALID, 1010)
セカンダリサブシステムA
セカンダリサブシステムB
セカンダリサブシステムC
100b-1
100b-2
100b-3
920
100c
中間サブシステム
120-1
CTG
(INVALID, 1064)
(PREPARED, 1042)
(VALID, 1038)
120-2
CTG
(INVALID, 1063)
(PREPARED, NULL)
(VALID, 1038)
120-3
CTG
(INVALID, 1102)
(PREPARED, 1042)
(VALID, 1038)
セカンダリサブシステムA
セカンダリサブシステムB
セカンダリサブシステムC
100b-1
100b-2
100b-3

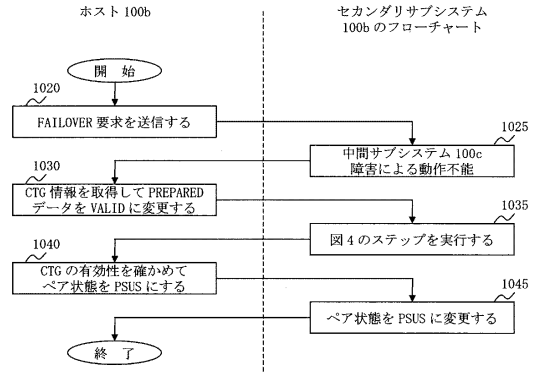
【図１０（ａ）】

【図１０(a)】



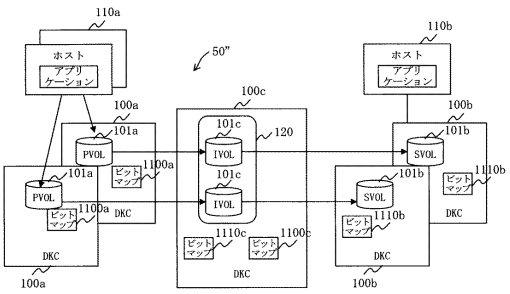
【図１０（ｂ）】

【図１０(b)】



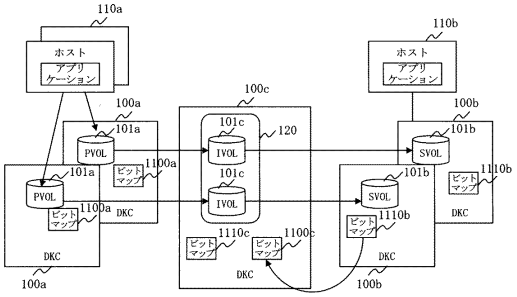
【図１１】

【図１１】



【図１２】

【図１２】



フロントページの続き

(56)参考文献　特開２０００－３０５８５６（ＪＰ，Ａ）
特開２００３－１２２５０９（ＪＰ，Ａ）
特開２００１－２８２６２８（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　　３／０６