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

特許第4108074号
(P4108074)

(45)発行日　平成20年6月25日(2008.6.25)　(24)登録日　平成20年4月11日(2008.4.11)

(51)Int.Cl.　F I
G06F 3/06 (2006.01)　G06F 3/06 304F
G06F 13/00 (2006.01)　G06F 13/00 520C

請求項の数 1　外国語出願　(全 13 頁)

(21)出願番号　特願2004-307582 (P2004-307582)
(22)出願日　平成16年10月22日 (2004.10.22)
(65)公開番号　特開2005-301976 (P2005-301976A)
(43)公開日　平成17年10月27日 (2005.10.27)
審査請求日　平成18年10月24日 (2006.10.24)
(31)優先権主張番号　10/792550
(32)優先日　平成16年3月2日 (2004.3.2)
(33)優先権主張国　米国 (US)

早期審査対象出願

(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　100079108
弁理士　稲葉　良幸
(74)代理人　100093861
弁理士　大賀　眞司
(72)発明者　大崎　伸之
アメリカ合衆国カリフォルニア州　キャンベル　エラムアベニュー　１２８１

審査官　梅景　篤

最終頁に続く

(54)【発明の名称】多重リモートストレージでのデータ同期

(57)【特許請求の範囲】
【請求項１】
プライマリストレージサブシステムと、少なくとも一のセカンダリストレージサブシステムを有するストレージシステムにおいて、
前記各ストレージサブシステムはそれぞれ、受信した一つ以上のデータブロックと各データブロックに付与された順序番号を格納するキューを備え、前記各セカンダリストレージサブシステムは、前記プライマリストレージサブシステムに直接、又は、他のセカンダリストレージサブシステムを介して間接的に接続されており、
前記プライマリストレージサブシステムは、ホストシステムで生成された一つ以上のデータブロックを受信すると、各データブロックに対して順序番号を付与し、該プライマリストレージサブシステムから直接データブロックを受信するストレージサブシステムである後続ストレージサブシステムに対して、前記各データブロックと各データブロックに付与された順序番号を送信し、
前記各セカンダリストレージサブシステムは、前記プライマリストレージサブシステムから送信された各データブロックと各データブロックに付与された順序番号を、前記プライマリストレージサブシステム、又は、他のセカンダリストレージサブシステムのいずれか一つから直接受信し、該セカンダリストレージサブシステムから直接データブロックを受信するストレージサブシステムである後続ストレージサブシステムに対して、前記各データブロックと各データブロックに付与された順序番号を送信し、
該セカンダリストレージサブシステムが受信した最新のデータブロックに付与された順

序番号と該セカンダリストレージサブシステムの後続ストレージサブシステムから受信した順序番号の中で最も古いデータブロックに付与された順序番号を、該セカンダリストレージサブシステムに直接データブロックを送信するストレージサブシステムである先行ストレージサブシステムに対して送信し、

前記プライマリストレージサブシステムは、該プライマリストレージサブシステムの後続ストレージサブシステムから受信した順序番号の中から最も古いデータブロックに付与された順序番号を検索して、該順序番号を該後続ストレージサブシステムに報告し、

前記各セカンダリストレージサブシステムは、該セカンダリストレージサブシステムの先行ストレージサブシステムから、前記プライマリストレージサブシステムにより検索された順序番号を受信し、当該順序番号を該セカンダリストレージサブシステムの後続ストレージサブシステムに報告することを特徴とするストレージシステム。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、一般的に、プライマリ(又はローカル)ストレージ施設と本プライマリストレージ施設に保存されているデータの少なくとも一部をミラーする複数のセカンダリ(又はリモート)ストレージ施設より成る、データ処理ストレージシステムに関連する。より具体的には、本発明は、一つのストレージロケーションから他のストレージロケーションにデータをコピー中に中断が発生した場合に、生存ストレージ施設間でデータを同期化する方法及び本方法を備えた装置に関連する。

【背景技術】

【０００２】

企業、政府及び他の組織体でのデータ処理業務の拡大により、夥しい量のデータが保存されるに至っており、その多くは組織体の日々の運営にとって極めて重要なものである。例えば、金融上の大量の取引は今や全面的に電子的に処理されている。航空会社でのビジネスでは、予約情報が万一喪失した場合には、大混乱になるリスクを背負っている。この結果として、データの信頼性を確保する為に、ローカルデータはデータが喪失した場合に備えて、一式以上のデータのコピーを通常しばしばリモートロケーションにバックアップしている。データが重要である程、バックアップ方法も精巧になる。例えば、影響の大きい重要なデータを守る一方法は、データをローカルストレージ施設とは地理的に離れたサイトにバックアップコピーを保存することである。各リモートストレージ施設は、ローカルストレージ施設のミラーイメージを維持して、ローカルストレージ施設のデータイメージが変更される度に、自施設のデータを変更してミラーを維持する。ローカルストレージシステムのデータをミラーする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 Data Storage System"のタイトルの米国特許５，９３３，６５３号に記されている。

【０００３】

【特許文献１】米国特許５，９３３，６５３号

【特許文献２】米国特許６，０９２，０６６号

【発明の開示】

【発明が解決しようとする課題】

【０００４】

リモートストレージ施設に転送される更新データは、しばしばキューされ、リモートコピー処理のオーバヘッドを減らす為にグループ化して、インターネット等のネットワーク転送媒体を通して、送信される。この為、リモートサイトで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常に同一であるわけではない。ローカルデータをミラーするのに、二つ以上のリモートストレージが使用された場合には、少なくとも全てが更新される迄は、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しまっている。このような異なったデータイメージの存在は、ローカル施設が障害になった場合には、問題になる。ローカルストレージ施設が障害になると、あるリモートストレージ施設は、完全に正確

では無くともローカルストレージ施設の障害前のデータをより最新まで反映し、他の施設はより古い中間データイメージのままで、最新の更新処理は全く追従できない状態であることがあり得る。かくして、ローカルストレージ施設で障害が発生した場合には、システムを再起動する前に、全てのリモートストレージ施設が同一の最新データイメージを持つように、再同期することが必要になる。

【０００５】

もう一つの検討しなければならない問題は、リモートコピー動作中に“停止(Suspension)”が発生した場合での回復である。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、リモートコピー動作中のネットワーク停止又はその他の介入障害等の予期せぬ事態での停止が発生した場合も、再同期が必要になる。リモートコピーでの再同期の一方法は、"Method and Apparatus for Independent Operation of a Remote Data Facility"のタイトルの米国特許６,０９２,０６６号に記されている。しかしながら、この特許に記されている技術では、限定された状況での再同期が可能になるのみである。例えば、リンク障害、キャッシュオーバフロー、更にドライブ障害等の二つの障害が重なった、より以上に複雑なシステムサスペンションが発生した場合には、システム全体の再初期化を回避する為にすぐ使用できる再同期方法は存在しない。このような事態の場合、これらの技術は少なくとも二セットのコピーが残るように保障していない為、再同期の為には通常、ボリューム全体のコピーが必要になってしまう。

【０００６】

プライマリサイトが災害等の問題で障害になり、プライマリストレージサブシステムのデータが使用不能になったら、コンピュータシステムは、セカンダリストレージサブシステムのデータを使用してジョブを開始する。何れかのセカンダリストレージサブシステムのデータを使用してジョブを開始するのに先立って、他のセカンダリストレージサブシステムのデータ全てが同じデータを持つように同期を行う必要がある。ストレージシステムが同期されていないと、本ストレージシステム内のデータは不整合状態になってしまう。幾つかのストレージシステムが使用されている場合には、各ストレージシステムは他のストレージシステムについてコピーの進行状態（即ち、どのデータのコピーが終了しているか）について知る手段がない。ストレージシステム間のデータの相違については、人手によってつかむことは実質上不可能である。この結果として、生産用に使用する一つのストレージシステムの全データを他の全てのストレージシステムにコピーすることが必要になり得る。結果として、不必要に大量のデータが転送され、完了するのに長時間が必要になってしまう。

【課題を解決するための手段】

【０００７】

本発明の一様態では、プライマリストレージサブシステムと、少なくとも一のセカンダリストレージサブシステムを有するストレージシステムは、前記各ストレージサブシステムはそれぞれ、受信した一つ以上のデータブロックと各データブロックに付与された順序番号を格納するキューを備え、前記各セカンダリストレージサブシステムは、前記プライマリストレージサブシステムに直接、又は、他のセカンダリストレージサブシステムを介して間接的に接続されており、前記プライマリストレージサブシステムは、ホストシステムで生成された一つ以上のデータブロックを受信すると、各データブロックに対して順序番号を付与し、該プライマリストレージサブシステムから直接データブロックを受信するストレージサブシステムである後続ストレージサブシステムに対して、前記各データブロックと各データブロックに付与された順序番号を送信し、前記各セカンダリストレージサブシステムは、前記プライマリストレージサブシステムから送信された各データブロックと各データブロックに付与された順序番号を、前記プライマリストレージサブシステム、又は、他のセカンダリストレージサブシステムのいずれか一つから直接受信し、該セカンダリストレージサブシステムから直接データブロックを受信するストレージサブシステムである後続ストレージサブシステムに対して、前記各データブロックと各データブロックに付与された順序番号を送信し、該セカンダリストレージサブシステムが受信した最新の

データブロックに付与された順序番号と該セカンダリストレージサブシステムの後続ストレージサブシステムから受信した順序番号の中で最も古いデータブロックに付与された順序番号を、該セカンダリストレージサブシステムに直接データブロックを送信するストレージサブシステムである先行ストレージサブシステムに対して送信し、前記プライマリストレージサブシステムは、該プライマリストレージサブシステムの後続ストレージサブシステムから受信した順序番号の中から最も古いデータブロックに付与された順序番号を検索して、該順序番号を該後続ストレージサブシステムに報告し、前記各セカンダリストレージサブシステムは、該セカンダリストレージサブシステムの先行ストレージサブシステムから、前記プライマリストレージサブシステムにより検索された順序番号を受信し、当該順序番号を該セカンダリストレージサブシステムの後続ストレージサブシステムに報告するものである。

【０００８】

他の実施例では、複数のストレージサブシステムでは、第二のストレージサブシステムに接続され、第二のサブシステムから一つ以上のデータブロックを受信し保存する少なくとも一式の第三のストレージサブシステムを含む。第二のストレージサブシステムは第三のストレージサブシステムの先行ストレージサブシステムである。第三のストレージサブシステムは第二のストレージサブシステムの後続ストレージサブシステムである。

【０００９】

本発明の具体的な実施例では、ストレージサブシステムは、テーブル内の最小最終順序番号を検索するプロセッサを備える。プロセッサは、テーブル内の最小最終順序番号と各後続ストレージサブシステムで特定されている最小最終順序番号を比較して、本ストレージサブシステムの最小最終順序番号を決定する。プロセッサは、自サブシステムの最小最終順序番号を、先行ストレージサブシステムが存在すれば先行サブシステムに報告し、存在しなければ各後続ストレージサブシステムに報告する。ストレージサブシステムに先行ストレージサブシステムが存在すれば、ストレージサブシステムのプロセッサは、先行ストレージサブシステムからの最小最終順序番号の受信を契機に、受信した最小最終順序番号以下の順序番号を持つデータブロックをキューから削除する。ストレージサブシステムに先行ストレージサブシステムが存在しなければ、ストレージサブシステムのプロセッサは、ストレージシステムの最小最終順序番号を決定次第、決定された最小最終順序番号以下の順序番号を持つデータブロックをキューから削除する。いずれの場合でも、受信又は決定した最小最終順序番号を後続(存在すれば)するストレージサブシステムに伝える。

【００１０】

本発明のある実施例では、各ストレージサブシステムはプロセッサを持ち、一つ以上のストレージサブシステムの障害に於いては、自システムの最終順序番号レコードに保存されている最終順序番号を他の生存ストレージサブシステム内の最終順序番号レコードに保存されている最終順序番号と比較し、何れかの生存ストレージサブシステムに保存されている最終順序番号が、自システムの最終順序番号より大きい場合には、当該システムの自システムより大きい最終順序番号を持つデータブロックを、自ストレージサブシステムのキューにコピーする。

【００１１】

本発明の他の様態は、複数のストレージサブシステムを持つストレージシステムに保存されたデータを管理する方法に向けられる。本方法は、一つのストレージサブシステムから一つ以上のデータブロックをサブシステムに接続されている後続ストレージサブシステムにコピーし、各コピーされたデータブロックとこれに付与された順序番号を後続するストレージサブシステムのメモリ内のキューに保存し；後続するストレージサブシステムのメモリ内の最終順序番号レコードに最終順序番号を保存し；各ストレージサブシステムに各後続するストレージサブシステムで最小最終順序番号として特定された各最終順序番号をメモリ内のテーブルの各対応する欄に保存する。

【００１２】

本発明の他の様態は、複数のストレージサブシステムを持つストレージシステムに保存

されたデータを管理する方法に向けられる。本方法は、一式のストレージサブシステムからこれに接続された後続ストレージサブシステムに一つ以上のデータブロックをコピーし、データブロックとこれに付与された順序番号を後続ストレージサブシステムのメモリ内のキューに保存し；ストレージサブシステムのデータブロックの最終順序番号から最小最終順序番号を決定し；決定された最小最終順序番号以下の順序番号を持つデータブロックを各ストレージサブシステム内のキューから削除する。

【発明の効果】

【００１３】

本発明を実施する事により、より進歩したデータ処理ストレージシステムが可能になり、本ストレージシステムでは、データは一つのストレージサブシステムから他の後続するストレージサブシステムに下り方向にコピーされ、コピー状態に関する情報は一つのストレージサブシステムから他の先行するストレージサブシステムに上り方向に伝えられる。本ストレージサブシステムは、典型的には、一つのプライマリストレージサブシステムと複数のセカンダリストレージサブシステムで構成される。データがプライマリストレージサブシステムから一つ以上のセカンダリストレージサブシステムに下り方向にコピーされる時には、各セカンダリストレージサブシステムは、コピー状態の更新情報を時刻順に最終順序番号を用いてキューの中に保存し、更に受信データの最終順序番号に関する情報を先行するストレージサブシステムに上り方向に報告する。ストレージサブシステム又はサイトが障害になると、本最終データを受信したストレージサブシステムのキューが、最終順序番号として決定され特定される。他のストレージサブシステムの各々に於いては、受信されていないデータの順序番号が検出され、これらのデータが当該ストレージサブシステムに送信される。各ストレージサブシステム内の最終順序番号は共有される為、各ストレージシステムは、他のストレージサブシステムとの同期に不必要な消去可能データをキュー内で、認識することが出来る。手短に言うと、各ストレージサブシステムのキューに保存された受信データの最終順序番号は、他のストレージサブシステムと交換又は共有され、障害時のデータ同期の為にストレージサブシステム間でデータをコピーするのに使用される。障害時のデータ同期の為に各ストレージサブシステムのキュー内に保存されるデータ量は最少に維持され、障害時のデータ同期の為にストレージサブシステム間でコピーしなければならないデータ量も減少する。

【発明を実施するための最良の形態】

【００１４】

図１は、ホストシステム１０１と複数のストレージサブシステムを持つストレージシステム１００を示す。ストレージシステムには一式のプライマリストレージサブシステム１０２と複数のセカンダリストレージサブシステム１１１、１１２が存在する。ホストシステム１０１は、入出力が転送される入出力インタフェース１０３を通してプライマリストレージサブシステム１０２に結合している。プライマリストレージサブシステム１０２に入力されたデータ又は情報は、メモリ１０５を使用してＣＰＵ１０６により処理され、ディスクやＲＡＩＤ又はＪＢＯＤ(Just a Bunch of Disks)と呼ばれるストレージ媒体に保存される。メモリ

１０５は、以下に詳細に述べるように、データブロックとこれに付与された順序番号を保存する為のキュー１０８、最終順序番号を保存する為のレコード１０９、及び各後続ストレージシステムの最終順序番号を保存する為のテーブル１１３を収容する。内部バス１１０は内部データの転送に用いられる。ネットワークインタフェース１０４は、ネットワークに接続され、セカンダリストレージサブシステム等の他のシステムがデータをコピーし、情報を送受信する為の通信を可能にする。データは、プライマリストレージサブシステム１０２からセカンダリストレージサブシステム１１１、１１２への下り方向にコピーされる。ディスク１０７からのデータは、プライマリストレージサブシステム１０２から、ネットワークに結合するネットワークインタフェース１０４を通して、一つ以上の対象セカンダリストレージサブシステムに流れる。他のストレージシステム（セカンダリストレージサブシステム）は、メモリ１０５、プロセッサ１０６、及び記憶媒体１０７を含めて

同様な構成を持つことが望ましい。

【００１５】

図２は、ホストシステム２０８、プライマリストレージサブシステム２０１、及び複数のセカンダリストレージサブシステム２０２～２０５を含むストレージシステム２００の論理システム構成を示す概略図である。図１で示すように、プライマリストレージサブシステム２０１は、ホストシステム２０８からの入出力が実行されるシステムである。このストレージシステム２００では、プライマリストレージサブシステム２０１のデータは、セカンダリストレージサブシステム２０２～２０５にコピーされる事になっている。ストレージシステム２００は、枝分かれ形式のストレージシステム構造になっており、セカンダリストレージサブシステム２０２、２０３は第一レベルのセカンダリストレージサブシステムで、セカンダリストレージサブシステム２０４、２０５は第二レベルのセカンダリストレージサブシステムである。データは、プライマリストレージサブシステム２０１から、プライマリストレージサブシステム２０１に直接後続するシステムとして、第一レベルのセカンダリストレージサブシステム２０２と２０３にコピーされる。データは次いで、第一レベルのセカンダリストレージサブシステム２０２に直接後続するシステムとして、第二レベルのセカンダリストレージサブシステム２０４と２０５に、第一レベルのセカンダリストレージサブシステム２０２からコピーされる。別の言い方をすると、プライマリストレージサブシステム２０１は、第二ストレージシステム２０２と２０３に先行する第一ストレージシステムで、一方、第三ストレージシステム２０４と２０５は、第二ストレージシステム２０２に後続するストレージシステムである。各ストレージシステムは、結合を自動検出するか人為入力によって、結合しているストレージサブシステムを認識している。

【００１６】

コピープロセスに於いては、データブロックは、一つのストレージシステムから他のストレージシステムに、ホストシステム２０８がプライマリストレージサブシステム２０１に書き込んだ順序通りに、転送される。各データブロックには、プライマリストレージサブシステム２０１にて、典型的には＃１、＃２・・・の昇順に、順序番号が付与される。

【００１７】

図３は、セカンダリストレージサブシステムの一つなどのストレージサブシステムでのコピープロセスを示す。サブシステムは、データブロックとそれの順序番号をサブシステムに直接接続する先行サブシステムから受信する（ステップ３０１）。ここで使用されているように、ストレージサブシステム間の関係を記述するのに、便宜上、“先行する”なる用語は“直接に先行する”意味で、“後続する”なる用語は“直接に後続する”意味とする。ステップ３０２に於いて、データブロックとその順序番号は、書き込み順序を保持する為に、最初に、サブシステムのメモリ中のキュー（例えば、図１のメモリ１０５内のキュー１０８を参照）に保存し、データは保存媒体（例えば、図１のディスク１０７を参照）に保存する。キュー１０８に収めたデータは、この直後か又は一定時間経過後にディスク１０７に収める。本データは又、キュー１０８に収める前にディスク１０７に格納しても良い。入力データの最終ブロックに伴う最終順序番号は、最終順序番号レコード（例えば、図１のレコード１０９を参照）に収める（ステップ３０３）。ストレージサブシステムが、一式以上の後続ストレージサブシステムを持っておれば、入力データとこれの順序番号は、この各後続ストレージサブシステムに転送する（ステップ３０５）。例えば、図２のストレージサブシステム２０２は、後続ストレージサブシステム２０４と２０５を持ち、ストレージサブシステム２０１はストレージサブシステム２０２の先行ストレージサブシステムである。ストレージサブシステムが後続ストレージサブシステムを持っていなければ、受信した最終順序番号は、上りの流れとして先行ストレージサブシステムに戻される（ステップ３０６）。

【００１８】

図４は図３のコピープロセスに関連するコピーステータス更新プロセスを説明するフロー図である。コピーステータス更新プロセスは、ストレージサブシステムが後続ストレー

ジサブシステムから最終順序番号を受信したときに開始する（ステップ４０１）。ストレージサブシステムは、テーブル内の、各後続ストレージサブシステムに対応する最終順序番号欄（図１のテーブル１１３を参照）を受信した順序番号で更新する（ステップ４０２）。このテーブル１１３は、各後続ストレージサブシステムの最終順序番号を保有する。図６はテーブルの一例を示す。ストレージサブシステムは、このテーブル内の最小最終順序番号を検索する（ステップ４０３）。最小最終順序番号は、ストレージシステムの配下に於ける最古のデータブロックに伴う順序番号である。ステップ４０３でのテーブル検索は、ストレージサブシステムが後続ストレージサブシステムの一つから最終順序番号を受信し次第、又は、例えば、プライマリストレージサブシステム２０１又は図２のホストシステム２０８で事前設定されたある一定時間たってから開始しても良い。ストレージシステムが構成されてコピーが開始した直後の期間に於いては、テーブル１１３の各最終順序番号欄は、対応する後続ストレージサブシステムの最終順序番号で満たされてはいない、事態が発生し得る。この場合には、本検索は、各欄が対応する後続ストレージサブシステムの最終順序番号で満たされてから、開始される。もとより、ストレージサブシステムがただ一つの後続ストレージサブシステムしか持っていない場合には、本検索及び比較処理は省略することができる。次いで、ストレージサブシステムがプライマリストレージサブシステムの場合には、決定された最小最終順序番号を全ての後続ストレージサブシステムに報告し（ステップ４０４）、先行ストレージサブシステムが存在すれば、先行ストレージサブシステムに本決定された最小最終順序番号を連絡する（ステップ４０５）。最小最終順序番号には、これを識別する為にヘッダーを加えても良い。図４に示すステップ４０４と４０５はこの順番に実行する必要はなく、この順序は逆転可能なことに注意すること。

【００１９】

図５は、コピーステータス更新プロセスでのコピーステータス更新情報に基づいて、データを削除するプロセスを示すフロー図である。データ削除プロセスは、ストレージサブシステムがこれの先行ストレージシサブステムから最小最終順序番号を受信することを契機に開始する（ステップ５０１）。ストレージサブシステムは、キュー（例えば、図１のキュー１０８を参照）内のデータブロックで、受信した最小最終順序番号以下の順序番号を持つ全てのデータブロックを削除しても良い（ステップ５０２）。このデータ削除は、ストレージサブシステムが自らの先行ストレージサブシステムから最小最終順序番号を受信した直後に実行しても、一定時間経過してから実行しても良い。先行ストレージサブシステムを持たないプライマリストレージサブシステム２０１に於いては、最小最終順序番号は、テーブルを検索して、各後続ストレージサブシステムから受信した最小最終順序番号の最小値を見つけることにより、図４のステップ４０３で決定される。プライマリストレージサブシステムは、上記で決まった最小最終順序番号以下の順序番号を持つキュー１０８内のデータブロックを削除できる。何れの場合でも、受信した又は決定した最小最終順序番号は各後続（存在すれば）するセカンダリストレージサブシステムに伝えられ（ステップ５０３）、本最小最終順序番号がステップ５０１で受信される。

【００２０】

本実施例では、コピーステータス更新プロセスとデータ削除プロセスの採用によって、サブシステム障害に於いて、不必要なデータ転送を行うことなく複数ストレージサブシステム間でデータ同期を行う方法が提供される。好都合なことに、プライマリストレージサブシステムのデータバックアップに、何台のストレージサブシステムが配置されているか、又これらのお互いの接続構成に関する全体のストレージシステム構造について、各ストレージサブシステムは認識する必要はない。各ストレージサブシステムは、サブシステム障害時に必要な情報を交換し同期に必要な冗長情報（即ち、最終順序番号を保存しサブシステム障害時のデータ同期に必要なデータをコピーする）を保持するのに、(コピーデータを受け取る)直接先行するストレージサブシステムと、(コピーデータを送信する)直接後続するストレージサブシステムの情報のみを知ればよい。

【００２１】

図７はコピープロセス中の各ストレージサブシステムのある瞬間の構成を抽出した構成図である。ホストシステム７０１はプライマリストレージサブシステム７０２に接続され、本システム７０２にデータを書き込む。プライマリストレージサブシステム７０２のデータは、これに接続されている第一レベルの後続セカンダリストレージサブシステム７０３と７０４にコピーされる。第一レベルのセカンダリストレージサブシステム７０４のデータは、これに接続されている第二レベルの後続セカンダリストレージサブシステム７０５と７０６にコピーされる。

【００２２】

図７はコピー動作が進行中の各ストレージサブシステムのある瞬間の構成を示す。データブロック７０１０１～７０１０５は発生順に並んでおり、ホストシステム７０１で生成されプライマリストレージサブシステム７０２のキュー１０８に書き込まれる。図７に示す通り、データブロック７０２０１～７０２０５は、キュー１０８内の書き込み済ブロックである。この時点では、データブロック７０２０１～７０２０４は、セカンダリストレージサブシステム７０３にコピー済で、データブロック７０３０１～７０３０４に対応するキューに格納済である。データブロック７０２０１～７０２０３は、セカンダリストレージサブシステム７０４にコピー済で、データブロック７０４０１～７０４０３に対応するキューに格納済である。セカンダリストレージサブシステム７０４内のデータブロック７０４０１～７０４０２は、セカンダリストレージサブシステム７０６にコピー済で、データブロック７０６０１～７０６０２に対応するキューに格納済である。データブロック７０４０１は、セカンダリストレージサブシステム７０５にコピー済で、データブロック７０５０１に対応するキューに格納済である。

【００２３】

最終順序番号レコード１０９と各後続装置の最終順序番号のテーブル１１３は、それぞれ７０４０４と７０４０６として、ストレージサブシステム７０４のみについて示されている。ストレージサブシステム７０４が後続ストレージサブシステム７０５、７０６から最終順序番号（各々、レコード７０４０７内の＃１、レコード７０４０８内の＃２）を受信すると、本ストレージサブシステム７０４は、テーブル７０４０６内の最小最終順序番号（レコード７０４０７内の＃１）を先行するストレージサブシステム７０２に報告する。

【００２４】

ストレージサブシステム７０２に於いては、最小最終順序番号＃１はストレージサブシステム７０４から受け取ったもので、最終順序番号＃４（これはストレージサブシステム７０３での最終順序番号でもある）はストレージサブシステム７０３から受け取ったものである。ストレージサブシステム７０２での最小最終順序番号はかくして＃１で、本番号＃１は後続ストレージサブシステム７０３、７０４に報告される。従って、ストレージサブシステム７０２、７０３、７０４は報告された最小最終順序番号（＃１）以下の順序番号を持つデータブロックをいつでも削除可能である。次いで、ストレージサブシステム７０４はこの最小最終順序番号を第二レベルの後続ストレージサブシステム７０５と７０６に報告し、ストレージサブシステム７０５と７０６は報告された最小最終順序番号（＃１）以下の順序番号を持つデータブロックをいつでも削除可能である。

【００２５】

図８は図７で示すデータ状態にデータ削除プロセスを適用した後の各ストレージサブシステムのデータ状態を抽出した構成図である。より具体的には、図８は、各データブロック#(８０２０１、８０３０１、８０４０１、８０５０１、８０６０１)が各対応するストレージサブシステム（８０２、８０３、８０４、８０５、８０６)の各キューより削除された後の状態を示す。ストレージサブシステム障害の場合での、ストレージサブシステム間でのデータ同期のメカニズムが示されている。

【００２６】

一例として、プライマリストレージサブシステム８０２が障害又は使用不能になったとする。ストレージシステム構成で存在している最終データブロックは、ストレージサブシ

ステム８０３のデータブロック８０３０４で、順序番号＃４（即ち、最大の最終順序番号）を持っている。本順序番号は他の全てのストレージサブシステムに配布される。この配布は、データコピートランザクションに使用されるネットワークを通して、各ストレージサブシステムに配布論理を導入することによって行うか、あるいは各ストレージサブシステムと結合するメカニズムを有するサーバを追加して行うことができる。他のストレージサブシステム（８０４、８０５、８０６）は順序番号＃４迄のデータブロックが必要である。ストレージサブシステム８０４は自らの最終順序番号（＃３）をレコード１０９内に持っており、データブロック８０３０４のみをコピーすればよいことを知っている。このことをネットワーク又はサーバを通して連絡して、データブロック８０３０４のコピー要求が発行され、本データブロックがストレージシステム８０４にコピーされる。ストレージサブシステム８０５はデータブロック８０３０４、８０３０３、８０３０２を、ストレージサブシステム８０６はデータブロック８０３０４、８０３０３を、各々コピーする必要があることを知っている。コピーは、ストレージサブシステム８０４の場合と同じメカニズムを使用して実行される。

【００２７】

他の例として、ストレージサブシステム８０４が障害か使用不能だとする。更に、ストレージサブシステム８０４、８０６が使用不能だと仮定する。ストレージシステム構成に残っている最終データブロックは、ストレージシステム８０２の順序番号＃５のデータブロック８０２０５である。ストレージサブシステム８０３はデータブロック８０２０５の、ストレージサブシステム８０５はデータブロック８０２０２～８０２０５のコピーをそれぞれ取得する必要がある。

【００２８】

各ストレージサブシステムは、最新データを持っているストレージサブシステムによって同期化される。データブロックはキュー内に冗長に保存され、どのストレージサブシステムが障害になっても、実質的にデータが失われることはない。キューの大きさを減少させる為に、不要になったデータブロックは削除される。

【００２９】

以上に述べた装置及び方法は、本発明の原理の応用を説明したもので、本発明の精神と請求項で定義された範囲を逸脱する事無く、多数の異なった実装や改変が可能である。従って、本発明の範囲は、これまでの記述を参照して決めるのではなく、添付する請求項の全範囲に従って判断する必要がある。

【図面の簡単な説明】

【００３０】

【図１】図１は、本発明の一実施例による複数のストレージサブシステムを持つストレージシステムを説明する概略図である。

【図２】図２は、本発明の一実施例によるストレージシステムの論理的システム構成を説明する概略図である。

【図３】図３は、本発明の一実施例によるコピープロセスを示すフロー図である。

【図４】図４は、本発明の一実施例によるコピープロセスに関連するコピーステータス更新プロセスを示すフロー図である。

【図５】図５は、本発明の一実施例によるコピーステータス更新プロセスのコピーステータス更新情報に基づくデータ削除プロセスを示すフロー図である。

【図６】図６は、本発明の一実施例によるコピーステータス更新情報を含むテーブルを示す。

【図７】図７は、本発明の一実施例によるコピープロセス中でのストレージシステム内の各ストレージサブシステムのある瞬間の構成を抽出した図である。

【図８】図８は、本発明の一実施例による、コピープロセスと関連コピーステータス更新プロセス及びデータ削除プロセスにおけるストレージシステム内での各ストレージサブシステムのデータ状態を抽出した図である。

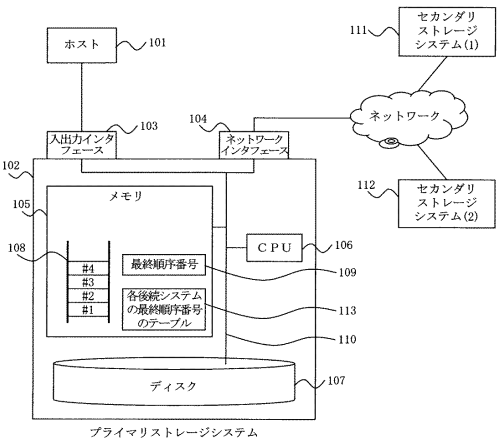
【符号の説明】

【００３１】

１０１　ホスト
１０２　プライマリストレージサブシステム
１０３　入出力インタフェース
１０４　ネットワークインタフェース
１０５　メモリ
１０６　ＣＰＵ
１０７　ディスク
１０８　キュー
１０９　最終順序番号
１１０　内部バス
１１１　セカンダリストレージサブシステム(１)
１１２　セカンダリストレージサブシステム(２)
１１３　各後続システムの最終順序番号のテーブル

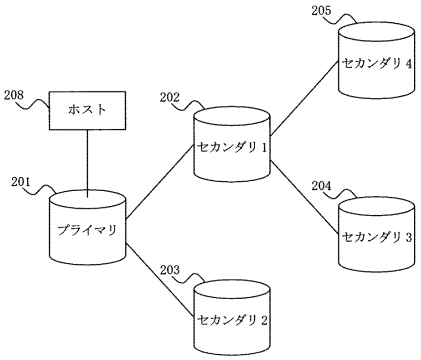
【図１】

図１



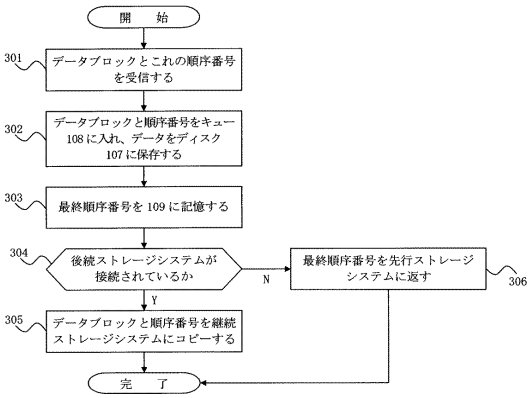
【図２】

図２



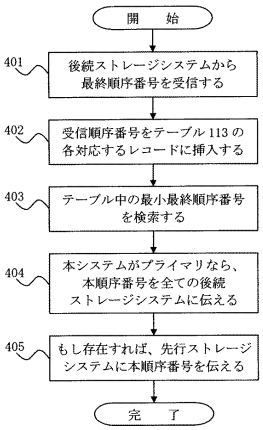
【図３】

図３



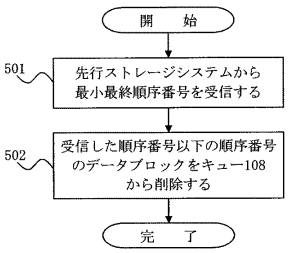
【図４】

図４



【図５】

図５



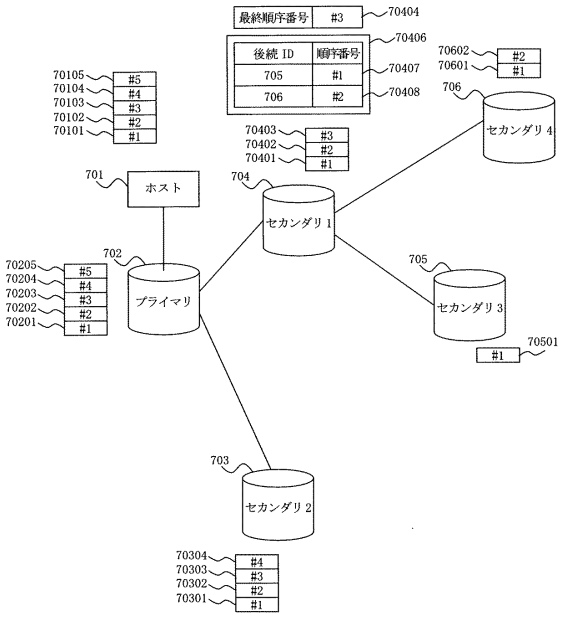
【図６】

図６

後続システム番号	最終順序番号
２０４	＃３
２０５	＃２

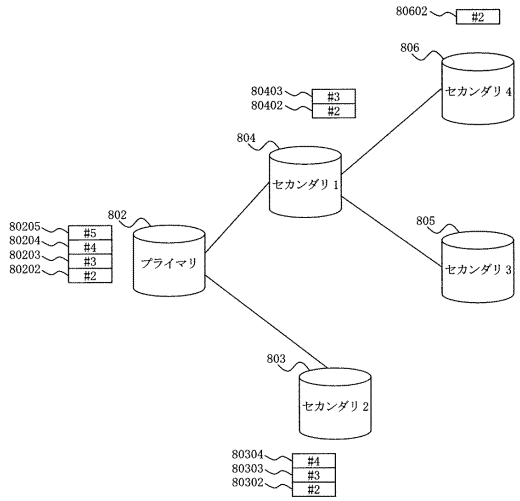
【図７】

図７



【図８】

図８


80602
#2
806
セカンダリ４
80403
80402
#3
#2
804
セカンダリ１
802
80205
80204
80203
80202
#5
#4
#3
#2
プライマリ
805
セカンダリ３
803
セカンダリ２
80304
80303
80302
#4
#3
#2

フロントページの続き

(56)参考文献 特開２００３－２６３２８０（ＪＰ，Ａ）
特開２００３－３４５５２４（ＪＰ，Ａ）
特開２００３－１２２６５９（ＪＰ，Ａ）
特開平０６－０５３９７３（ＪＰ，Ａ）
特開２００１－１６８９０７（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　３／０６－　３／０８
１２／００
１３／００
１３／１０－１３／１４