(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2004-13367
(P2004-13367A)
(43)公開日　平成16年1月15日(2004.1.15)

(51)Int.Cl.[7]	F I		テーマコード（参考）
G06F 3/06	G06F 3/06	304F	5B065
G06F 12/00	G06F 12/00	520E	5B082
	G06F 12/00	545M	

審査請求 未請求　請求項の数 8　OL　（全 16 頁）

(21)出願番号　特願2002-163705(P2002-163705)
(22)出願日　平成14年6月5日(2002.6.5)

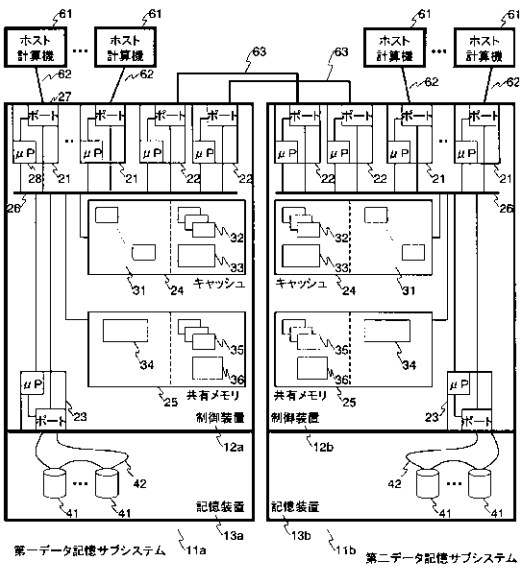
(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地
(74)代理人　100075096
弁理士　作田　康夫
(72)発明者　大野　洋
神奈川県小田原市中里３２２番地２号　株式会社日立製作所ＲＡＩＤシステム事業部内
(72)発明者　占部　喜一郎
神奈川県小田原市中里３２２番地２号　株式会社日立製作所ＲＡＩＤシステム事業部内

最終頁に続く

(54)【発明の名称】データ記憶サブシステム

(57)【要約】
【課題】データ記憶サブシステム間のリモートコピーにおいて、コピー先で論理ボリュームに対する属性情報を使用するためにはホスト計算機間で別途属性情報の交換の通信を行わなければならなかった。また、属性情報の更新と通常のデータの更新の間の順序関係を維持することは困難であった。
【解決手段】データ記憶サブシステム間で、データのリモートコピー機能と同一のデータ転送パスを使用して、属性情報をコピーする。また、更新データおよび更新属性情報を合わせた通しの更新順序番号を付与してコピーを行い、受信側のデータ記憶サブシステムにおいてこの更新順序番号に従ってデータならびに属性情報を更新する。
ホスト計算機側では全く意識することなしに、属性情報がバックアップ側のデータ記憶サブシステムにコピーされ、またデータおよび属性情報の更新の順序を維持したままコピー先のデータ記憶サブシステム上でこれらの更新が反映される。
【選択図】　図１

【図1】　全体構成図



【特許請求の範囲】
【請求項１】
２以上のデータ記憶サブシステムが相互に接続された場合において、
前記データ記憶サブシステムが有する、データのリモートコピー機能を用いて、
各データ記憶サブシステムが、自己の属性情報を、他のデータ記憶サブシステムへ、コピーすることを特徴とするデータ記憶サブシステム。
【請求項２】
磁気ディスク装置その他の記憶装置と、前記記憶装置に対して、データの入出力を行う制御装置とからなるデータ記憶サブシステムであって、
ホスト計算機に接続された一のデータ記憶サブシステムを含む、複数のデータ記憶サブシステムが前記制御装置を介して接続された構成において、
前記ホスト計算機から前記一のデータ記憶サブシステムに対し書込まれたデータを、別のデータ記憶サブシステムに対してコピーする際に、
前記ホスト計算機から、前記一のデータ記憶サブシステムに対して発行されたデータ書込みコマンド以外の制御指示コマンドの内容に従い、当該データ記憶サブシステム内の制御装置が処理した結果発生する属性情報を、前記別のデータ記憶サブシステムに対してコピーすることを特徴とするデータ記憶サブシステム。
【請求項３】
請求項２記載のデータ記憶サブシステムにおいて、
前記属性情報のコピーは、前記データのコピーと同じ転送パスを用いて行なわれるデータ記憶サブシステム。
【請求項４】
請求項２記載のデータ記憶サブシステムにおいて、
前記記憶装置は、一の論理ボリューム又は２以上の区切られた論理ボリュームを有し、
前記ホスト計算機は、前記論理ボリュームを指定して、データの読み書きを行うことが可能であり、
前記属性情報は、次のうち少なくともいづれか１の情報を含むデータ記憶サブシステム：
１）論理ボリュームを識別するために付与するタグ情報、
２）論理ボリュームへアクセス中のホスト計算機のホストＩＤ情報、
３）論理ボリュームに対して特定のアクセス権を持つホスト計算機の情報、
４）論理ボリュームに対して特定のアクセス権を持つホスト計算機のポートの情報、
５）論理ボリュームのリザーブ制御情報、又は、
６）論理ボリュームへの更新を許可するキー情報。
【請求項５】
請求項２記載のデータ記憶サブシステムにおいて、
前記記憶装置は、一の論理ボリューム又は２以上の区切られた論理ボリュームを有し、
前記ホスト計算機は、前記論理ボリュームを指定して、データの読み書きを行うことが可能であり、
前記一のデータ記憶サブシステムは、自己に接続された前記ホスト計算機から、当該データ記憶サブシステムの論理ボリュームに対して発行された、データ書込みコマンド及び前記制御指示コマンドを送出し、
前記別のデータ記憶サブシステムは、前記データ書込みコマンドに対応するデータ及び前記制御指示コマンドに対応する属性情報を、これらの順序性を保って、当該別のデータ記憶サブシステムの論理ボリュームに対して格納するデータ記憶サブシステム。
【請求項６】
請求項２記載のデータ記憶サブシステムにおいて、
前記記憶装置は、一の論理ボリューム又は２以上の区切られた論理ボリュームを有し、
前記ホスト計算機は、前記論理ボリュームを指定して、データの読み書きを行うことが可能であり、
前記一のデータ記憶サブシステムは、自己が接続されている前記ホスト計算機上で動作す

るプログラムが、前記論理ボリュームに対して、データ以外の任意の情報を付与する機能と、前記プログラムからの指示で付与された前記データ以外の任意の情報を格納する機能とを有するデータ記憶サブシステム。

【請求項７】

請求項２記載のデータ記憶サブシステムにおいて、

前記制御指示コマンドに代わり、前記データ書込みコマンドが一部に有する制御指示情報が用いられるデータ記憶サブシステム。

【請求項８】

請求項２記載のデータ記憶サブシステムにおいて、

前記記憶装置は、一の論理ボリューム又は２以上の区切られた論理ボリュームを有し、

前記一のデータ記憶サブシステムは、前記制御指示コマンドを、特定の論理ボリュームに対する書込みデータとして扱う機能を有し、

前記ホスト計算機からの制御指示コマンドに代わり、前記特定の論理ボリュームに書き込まれたデータを用いるデータ記憶サブシステム。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、ホスト計算機のデータを格納する外部記憶装置群であって、相互に附近地又は遠隔地に設置された複数の外部記憶装置の間で、情報の複製を行なう技術に係り、特に、各外部記憶装置の属性情報の取扱いに関する。

【０００２】

【従来の技術】

近年の計算機システムでは、使用されるデータの容量が指数関数的に増大し、データの更新頻度も高くなっている。このデータをどのようにバックアップし、また障害発生時にどれだけ迅速に正常稼働状態に復旧できるかが、データストレージ技術における重要課題となっている。この課題に対する一つの解として、磁気ディスクアレイを搭載したデータ記憶サブシステム（外部記憶装置）を遠隔地に複数台設置して、それらの間を通信パスで接続し、一方のデータ記憶サブシステム上で更新されたデータを、ホスト計算機を経由せずに、自動的に他のデータ記憶サブシステムに常時コピーする、リモートコピー技術がある。

【０００３】

このリモートコピー技術では、コピー先データ記憶サブシステムにもホスト計算機を接続することにより、第一のデータ記憶サブシステムが使用不能となる故障や災害が発生した場合に、コピー先のデータ記憶サブシステム上にコピーされているデータを使用して業務を引き継ぐディザスタリカバリが実現できる。この引き継ぎの際、データの整合性が重要になるが、コピーの中断がどの時点で発生するかは予測不能である。このため、当該データを使用するアプリケーションが整合性を回復できるようにするため、コピー先のデータ記憶サブシステム上のデータの更新順序を、コピー元のデータ記憶サブシステムに対して行われた更新順序と同一に維持することが必要になる。

【０００４】

このようなデータ更新順序維持を実現する従来技術として、特開平６－２９０１２５号公報記載の技術があげられる。また、このようなデータ更新順序維持を、ホスト計算機の介在無く、かつ各データのコピー完了確認を待たない非同期方式で実現する従来技術として、特開平１１－８５４０８号公報記載の技術が挙げられる。

【０００５】

一方、データ記憶サブシステムに接続されるホスト計算機上で動作するプログラムは、データの取扱い方法が高機能化しており、データ記憶サブシステムに対しても、単なるデータの格納以外に付加的な制御を要求するようになってきている。このような制御は、ホスト計算機がデータ記憶サブシステムにアクセスする際のデータ格納領域の単位となる、論理ボリュームを対象として行われるのが一般的である。

【０００６】
付加的な制御の例として、複数のホスト計算機からアクセスされる可能性のある論理ボリュームにおけるアクセス制御があげられる。これは当該論理ボリュームに対する参照や更新といった操作の許可・不許可を、ホスト計算機単位あるいはプログラムのインスタンス単位に行い、なおかつ動的に設定変更が可能であることが望ましい。アクセス制御をホスト計算機単位で行う場合、データ記憶サブシステム側では、ホスト計算機のＩＤやホスト計算機に接続されたポートのＩＤによりホスト計算機を識別し、アクセス制御を行う。また、最も簡単なホスト計算機単位のアクセス制御としては、単一のホスト計算機にのみアクセスを許可するリザーブ機能がある。一方、プログラムのインスタンス単位で行う場合、アクセスを許可するためのキー値を登録する機能の提供と、その後のデータ読み書き等のアクセスコマンドに当該キー値が付加されている要求のみアクセスを許可する制御方法で実現できる。
【０００７】
別な制御の例として、論理ボリュームの状態等を報告する機能がある。例えば、ある時点で当該論理ボリュームを使用中のホスト計算機のＩＤ情報が登録されていて、別なホスト計算機からの要求に応じてその情報を返すことで、後からアクセスしたホスト計算機が当該論理ボリュームの使用状況を確認でき、その後の動作の決定に利用できる。また、論理ボリュームを認識するのに、データ記憶サブシステム固有のＬＵＮ（Ｌｏｇｉｃａｌ　Ｕｎｉｔ　Ｎｕｍｂｅｒ）の代わりに、接続されたホスト計算機側で都合の良い適当なタグ情報を付与して利用することがある。
【０００８】
一方で、データ記憶サブシステムが提供する制御機能ではなく、プログラム側での独自の処理を行う場合がある。たとえば、複数のホスト計算機上に分散した、プログラムの複数インスタンスが連携して動作する場合に、共有するデータ記憶サブシステム経由で、連携に必要な情報を交換することがある。この場合、データ記憶サブシステムの論理ボリュームへデータを書込む代わりに、データ記憶サブシステム上の作業メモリ領域上に必要な情報を読み書きする方法があり、これにより物理的なディスクへの書込み処理が無い高速なデータ交換を実現することができる。この場合、当該データ記憶サブシステムは、任意の作業用情報の読み書きを行うという制御コマンドを提供することになる。
【０００９】
これらのデータ記憶サブシステムの制御処理で使用するあるいは処理の結果発生する情報は、データ記憶サブシステム内の処理プロセッサ間で共有されるメモリ領域に保持される。以下、本明細書中では、これらの情報のことを「属性情報」と呼ぶ。属性情報は、前記の各例の場合、アクセス設定のリスト、リザーブ制御情報、アクセス許可のキー値、使用中ホストのホストＩＤ、論理ボリュームのタグ情報、プログラムが読み書きする任意の作業用情報そのもの、などとなる。
【００１０】
これらの付加的な制御は各データ記憶サブシステムの中で閉じたものであり、この結果発生する属性情報はリモートコピーの対象外であった。しかし、リモートコピーのコピー先でディザスタリカバリ等の引継ぎを実行するときに、より正確にデータを復元するためには、引継ぎ直前時点でのデータ記憶サブシステム上の属性情報も利用できた方が好ましい。
【００１１】
従来このような属性情報の利用を行う際には、各データ記憶サブシステムに接続されたホスト計算機間で通信を行うことにより、これらの情報を交換していた。例えば、特開平１１－３０５９４７では、ホスト計算機から発するアテンション報告指示情報を、磁気ディスク制御装置が受け、これをリモートの磁気ディスク制御装置へ転送し、これを受け取ったリモートの磁気ディスク制御装置が、リモートのホスト計算機へ通知する技術が開示されている。
【００１２】

【発明が解決しようとする課題】
従来技術では、属性情報を使用するためには、ホスト計算機間で、別途、属性情報の交換の通信を行わなければならず、またその実装にはリモートコピーの状態を考慮することが必要となり、処理が複雑になる。本発明の実施例では、ホスト計算機側では全く意識することなしに、属性情報をバックアップ側のデータ記憶サブシステムにもコピーし、バックアップ側で自動的に利用可能としている。
【００１３】
また従来技術では、属性情報の更新はホスト計算機間で、データの更新はデータ記憶サブシステム間で各々個別に行われるため、属性情報の更新とデータの更新の間の順序関係を維持することは、基本的に不可能であった。ホスト計算機上のプログラムでデータ記憶サブシステムと密に連携して管理することで実現は可能だが、このような連携は実装が難しいだけでなく、オーバヘッドが大きいため性能も出ない。本発明の実施例では、ホスト計算機側では全く意識することなしに、データおよび属性情報の更新の順序を維持したまま、しかも高速に、コピー先のデータ記憶サブシステム上にこれらの更新が反映される機能を実現している。
【００１４】
上述した、属性情報とリモートコピーの状態とを完全に連携させるためには、リモートコピー中断後の再開時やコピー方向反転時での属性情報の再同期の方法も解決されなければならない。本発明の実施例では、ホスト計算機側では全く意識することなしに、リモートコピー中断後の再開時やコピー方向反転時での属性情報の差分コピーを実行できるようにし、短時間で属性情報の再同期を終了している。
【００１５】
さらに上述した機能を、複数のデータ記憶サブシステムにまたがるリモートコピーに適用して、耐障害性をより高くしたいというニーズがある。本発明の実施例では、複数のデータ記憶サブシステムにまたがる属性情報のコピーを実現している。
【００１６】
【課題を解決するための手段】
データ記憶サブシステム間で、データのリモートコピー機能と同一のデータ転送パスを使用して、属性情報をコピーする機能を設ける。
【００１７】
更新データおよび更新属性情報を合わせた通しの更新順序番号を付与してコピーを行い、受信側のデータ記憶サブシステムにおいてこの更新順序番号に従って、データならびに属性情報を更新する機能を設ける。
【００１８】
リモートコピーの中断中に更新された属性情報を、コピー元・コピー先両方のデータ記憶サブシステムにおいて記憶し、リモートコピーの再開後に更新部分のみをコピーして、再同期する機能を設ける。
【００１９】
コピー先となるデータ記憶サブシステムの数だけコピー元で属性情報更新情報を複製し、各コピー先に別々に送信する機能を設ける。また１次コピー先で受け取った属性更新情報を、２次コピー先にリモートコピー機能により送信する機能を設ける。
【００２０】
【発明の実施の形態】
＜はじめに＞
図９により、本発明の前提となる同期転送方式のリモートコピー、および非同期転送方式のリモートコピーについて説明する。
【００２１】
同期転送方式によるリモートコピーでは、第一データ記憶サブシステム１１ａは、ホスト計算機６１から更新データブロックを受領し、そのデータブロックを第二データ記憶サブシステム１１ｂに転送する。１１ｂは、当該データブロックの受領を１１ａに通知する。

１１ａは、当該通知受領後ホスト計算機に対し更新データブロックの書込み完了通知を行う。以上の処理のいずれかを失敗した場合には、１１ａは、ホスト計算機に書込み障害として報告する。

【００２２】

非同期転送方式によるリモートコピーでは、第一データ記憶サブシステム１１ａは、ホスト計算機６１から更新データブロックを受領し、その時点で当該ホスト計算機に更新データブロックの書込み完了通知を行う。１１ａは、自身のスケジュールで、ホスト計算機側の処理と非同期に、第二データ記憶サブシステム１１ｂへ当該データブロックを転送する。遠隔地へのデータ経路の複雑化、中途経路のボトルネック化により、データ転送中の当該データの順序性は保証されない。

【００２３】

図９では、例えば、データ転送経路上の順序は＃１、＃４、＃３、＃２である。更新順序の保存は、１１ｂにおいて、１１ｂのソート機能により、第一データ記憶サブシステムがホストからデータを受領した順序を保って、更新することにより達成される。つまり、１１ｂにおいては、＃１、＃２、＃３、＃４の順序であり、この更新処理の途中に、第一データ記憶サブシステムやデータ経路に不慮の災害が発生しても、第二データ記憶サブシステムにおけるデータ更新順序は守られている。このため、１１ｂに接続されるホスト計算機６１では、データベースやジャーナルファイルシステムなどは、矛盾なく回復処理を行なうことができる。非同期転送方式はホスト計算機の高性能処理、データ記憶サブシステム間の距離拡大、という特長を持ち、かつ順序性を保証したリモートサイトへの書込み制御により任意時点のデータベースやジャーナルファイルシステムの整合性を確保できる特長を有す。コピー元からの送信順序が任意であっても、リモート側で一旦、受信し、格納した上で、対応する情報が揃ったところからリモート側の更新処理を行なうためである。

【００２４】

＜実施例１＞

図１は第１の実施例の全体構成を示している。二つのデータ記憶サブシステム（以下、単に適宜「サブシステム」と略記する）１１ａ、１１ｂが、データ転送パス６３を介して接続され、互いに附近地又は遠隔地に設置されている。

【００２５】

各サブシステムは、ホストアクセスパス６２を介して、複数のホスト計算機６１と接続されている。サブシステムは、データの読み書き要求の処理や、更新データの格納を制御する制御装置１２（以下、適宜、１２ａ、１２ｂの総称として記す。以下同様）と、実際のデータを最終的に格納する記録媒体を含む記憶装置１３の二つの部分から構成される。

【００２６】

制御装置１２は、ホスト計算機と接続されるチャネルアダプタ２１、他のサブシステムと接続されるチャネルアダプタ２２、記憶装置と接続されるディスクアダプタ２３を持つ。各アダプタは、実際のデータ送受を行なうポート２７と、データ送受を制御するマイクロプロセッサ２８から構成される。

【００２７】

図１では、一つのアダプタに一つのポートと一つのマイクロプロセッサが搭載されているが、いずれも個数は任意であり、個数が増えるほど処理能力が高まることになる。

【００２８】

また制御装置１２には、記憶装置１３から取り出した、あるいは記憶装置１３にこれから格納すべきデータを保持するキャッシュメモリ２４（以下、適宜「キャッシュ」と略記する）と、各アダプタのマイクロプロセッサが処理を実行するためのワーキングエリア等として使用される共有メモリ２５がある。各アダプタは、キャッシュおよび共有メモリとバス２６で接続されていて、これらのメモリ領域とデータをやり取りすることができる。

【００２９】

記憶装置１３は、複数の磁気ディスクドライブ４１を持つ。これらの磁気ディスクドライブは、ディスク入出力パス４２により制御装置１２側のディスクアダプタ２３に接続され

る。
【００３０】
サブシステムは、ホスト計算機に対して論理ボリュームと呼ばれるデータ記憶領域が存在するように動作し、ホスト計算機側では対象となる論理ボリュームを指定して、データの読み書きを行なう。この論理ボリュームの識別子のことをＬＵＮ（Ｌｏｇｉｃａｌ　Ｕｎｉｔ　Ｎｕｍｂｅｒ）と呼ぶ。制御装置１２では、記憶装置４１内の磁気ディスクドライブの記憶領域を分割して、各論理ボリュームに対応づけ、最終的にこれらの対応領域にデータを格納する。
【００３１】
＜一般的な処理＞
制御装置１２の処理について簡単に説明する。ここでは、第一サブシステム側から第二サブシステム側に対して、リモートコピーを行なっている場合を説明する。ただし、リモートコピーの向きは論理ボリューム毎に設定・変更でき、ある時点で両方向のリモートコピーが共存することもある。
【００３２】
ホスト計算機６１からデータ書込み要求を受け取ったチャネルアダプタ２１は、キャッシュ２４上の対応領域にキャッシュエントリ３１としてデータを格納する。これらのキャッシュエントリは、ディスクアダプタ２３上のマイクロプロセッサが、他のマイクロプロセッサの処理とは独立したスケジューリングにより、対応する磁気ディスクドライブの領域にデータを書込み、データを最終的に保存する。
【００３３】
同期転送方式によるリモートコピーを行なっている場合は、書込み処理を行なったチャネルアダプタ２１から、第二サブシステムに接続されるチャネルアダプタ２２に転送要求が出され、前記で説明した同期転送が実行される。この際、第二サブシステム側のチャネルアダプタ２２は、転送されたデータ内容により、キャッシュ上の該当キャッシュエントリ３１を更新していく。
【００３４】
非同期転送方式によりリモートコピーを行なっている場合は、第一サブシステム側のチャネルアダプタ２１上のマイクロプロセッサは、更新データ内容をキャッシュエントリ３１以外にデータ更新情報３２としてキャッシュ２４上の別領域に保存し、この時点で書込み要求をしたホスト計算機６１には処理完了を報告する。第二サブシステムに接続されるチャネルアダプタ２２上のマイクロプロセッサは、他のマイクロプロセッサと独立したスケジューリングにより、データ更新情報３２を第二サブシステムに、ホスト計算機からの書込み要求順序の発行順序番号を付加して、転送する。
【００３５】
第二サブシステム側のチャネルアダプタ２２は、転送されたデータをキャッシュ３１上の別領域に更新データ３２として一旦保存し、さらに別なスケジューリングの処理により、発行順序番号が揃った部分について、当該発行順序番号の順で、キャッシュエントリ３１の更新を行なっていく。
【００３６】
いずれの方式のリモートコピーでも、第二サブシステム側でのキャッシュエントリ３１は、ディスクアダプタ２３上のマイクロプロセッサが、他のマイクロプロセッサとは独立したスケジューリングにより、対応する磁気ディスクドライブの領域にデータを書込み、データを最終的に保存する。
【００３７】
以上の説明で明らかなように、いずれのサブシステムでも、磁気ディスクドライブへの書込みは、対ホスト計算機の処理やリモートコピー処理とは独立して動作する。このため、以下、本明細書中では、ディスクアダプタ２３ならびに記憶装置１３の構成要素に関する議論は行なわない。
【００３８】

一方、ホスト計算機６１は、サブシステムに対して、磁気ディスク装置に記憶するためのデータの読み書き以外の処理要求を行う場合がある。これらの要求はチャネルアダプタ上のマイクロプロセッサが、必要に応じて他のマイクロプロセッサと協調しながら、処理を行なう。処理の結果は、一般に論理ボリューム毎に管理される属性情報として、共有メモリ上の属性情報テーブル３４上に格納される。このようなデータ以外の処理要求としては、従来の技術の項で述べたように、論理ボリュームのリザーブ設定などがあげられる。

【００３９】

＜本発明における処理＞

本発明に係る制御装置固有の処理について説明する。本発明による制御装置では、前記のホスト計算機からの書込みデータの他に、前記属性情報をもリモートコピー機能の一部として転送し、異なるサブシステム間で属性情報の常時一致化を行なうものである。ホスト計算機は、属性情報を発生させるような要求について、通常のデータの書込み要求を含めて、途中で打ち切らずに終わりまで、要求発行順にサブシステム側で処理されることを期待する。このため、リモートコピーにおいても、属性情報を発生させるような要求と、通常のデータの書込み要求とを通した、要求発行順に対応データ又は属性情報が更新されるように構成する。

【００４０】

以後の説明では、ホスト計算機から行われる通常のデータの書込み要求のことを「データ更新」、属性情報を発生させるような、データ書込み以外の要求のことを「制御指示」と呼ぶこととする。また、リモートコピーのコピー元（以下の説明の大部分では第一サブシステム側）を「正側」、コピー先（以下の説明の大部分では第二サブシステム側）を「副側」と呼ぶ場合がある。

【００４１】

図２に属性情報テーブル３４の構成を示す。テーブルは、論理ボリューム毎の行に、当該論理ボリュームの属性情報が格納される構成となっている。一番左の列に論理ボリュームのＩＤであるＬＵＮが格納され、続いて、リザーブ状態を表すフラグ、アクセスを許可するキー、使用中のホストＩＤ、その他の属性情報を格納する列がある。

【００４２】

表の一部の領域は、ユーザ用領域の列として予約され、サブシステムに接続されたホスト計算機上のプログラムが、専用のアプリケーションインタフェース（以下、「ＡＰＩ」と略記する。）を使って、当該論理ボリュームに関して設定・記憶させたい任意の属性情報を格納することができる。これにより、サブシステムの制御装置ではサポートしていない、プログラム独自の競合管理等を実現することが容易になる。

【００４３】

図３に、同期転送によるリモートコピー構成での、ホスト計算機６１からの要求に対する処理フローを示す。各要求は、ホスト計算機６１から第一サブシステム１１ａのチャネルアダプタ２１に対して、ＳＣＳＩプロトコルのコマンドとして発行される。チャネルアダプタのマイクロプロセッサ２８（図１）は、コマンドの種別を判断する。データ更新コマンドであれば正側キャッシュ２４上のキャッシュエントリ３１として書込み（Ｓ１１）、制御指示コマンドであれば、制御指示の内容を実行し（Ｓ１６）、その結果発生する論理ボリュームに対する属性情報は、共有メモリ２５上の属性情報テーブル３４に格納される（Ｓ１７）。続いて、チャネルアダプタ２１からの要求により、チャネルアダプタ２２上のマイクロプロセッサ２８は、データ転送パス６３を経由して、副側のチャネルアダプタ２２に更新情報を転送する（Ｓ１２）。副側のチャネルアダプタ２２上のマイクロプロセッサ２８は、更新情報にあるフラグから当該更新内容がデータか属性情報か判断できるので、データならば対応する副側キャッシュ２４上のキャッシュエントリ３１として書込み（Ｓ１３）、属性情報ならば副側共有メモリ２５上の属性情報テーブル３４に格納される（Ｓ１８）。これらの処理が完了すると、副側のチャネルアダプタから正側のチャネルアダプタに転送完了通知を行い（Ｓ１４）、正側のチャネルアダプタは、要求を発行したホスト計算機に完了通知を行なう。ここで説明した一連の処理は、ある論理ボリュームに対

し同時には一つしか実行されないように管理され、後着の処理は保留ないしエラーとなる。これにより、副側でも、データならびに属性情報の更新が、両者を通じた要求発行順に更新される。

【００４４】

＜更新情報の転送＞

更新情報の転送について、さらに説明する。本実施例では、データ転送パス６３上でも、ＳＣＳＩプロトコルを使用する。正側チャネルアダプタ２１上のポート２７がイニシエータとなり、副側のチャネルアダプタ２１上のポート２７に対して、更新情報を送信する。

【００４５】

図５には、データの更新情報を送信する場合のＳＣＳＩコマンドの構造を示す。オペレーションコードにはリモートコピーライトがセットされ、その他に、宛先の論理ボリュームを示すターゲットＬＵＮ番号、更新データの論理ボリューム内データ格納先頭アドレス、転送長（データサイズ）が格納される。同期転送の場合、順序番号は使用しない。本実施例では、更新情報としてデータ以外に属性情報も送られることになるので、これを区別するためのフラグが追加されており、ここに「データ」を示すフラグを設定する。このコマンドが副側で受け付けられると、続くデータ転送フェーズでデータ本体が送られることになる。

【００４６】

図６には、属性情報の更新情報を送信する場合のＳＣＳＩコマンドの構造を示す。基本的にはデータの更新情報と同様の構造を持ち、オペレーションコードにはリモートコピーライトがセットされ、その他に、宛先の論理ボリュームを示すターゲットＬＵＮ番号、更新情報の属性情報テーブル内アドレス、転送長（属性情報サイズ）が格納される。同期転送の場合、順序番号は使用しない。前記のフラグに「属性情報」を示すフラグを設定する。このコマンドが副側で受け付けられると、続くデータ転送フェーズで属性情報本体が送られることになる。なお、属性情報は一般に容量が小さいので、ＳＣＳＩコマンドの特定の領域（ベンダ固有領域やアドレスフィールドの一部など）に更新情報本体も格納することも可能である。また、テーブル内アドレスと転送長の組み合せの代わりに属性情報の種別を示すＩＤを使うことも可能である。

【００４７】

図４に、非同期転送によるリモートコピー構成での、ホスト計算機６１からの要求に対する処理フローを示す。同期転送の場合と同様に、第一サブシステム１１ａのチャネルアダプタ２１（図１）のマイクロプロセッサ２８は、コマンドの種別を判断し、データ更新コマンドであれば正側キャッシュ２４上のキャッシュエントリ３１として書込み（Ｓ２１）、制御指示コマンドであれば制御指示の内容を実行して（Ｓ２４）発生する属性情報を正側共有メモリ２５上に格納する（Ｓ２５）。さらにデータの更新の場合は正側キャッシュ２４上の別領域にデータ更新情報３２を書込み（Ｓ２２）、制御指示の場合には発生した属性情報の更新内容に関する情報を正側共有メモリ２５上に属性情報更新情報３５として書込む（Ｓ２６）。Ｓ２２、Ｓ２６で付けられる順序番号は、データの更新情報、属性情報の更新情報の双方を合わせた通し番号とする。論理ボリューム毎、または設定された論理ボリュームグループ毎に、通し番号を設定する。

【００４８】

ここで各更新情報の格納内容は、図５、図６で示した更新情報の転送コマンドに含まれる内容と更新内容本体を合せたものである。各更新情報を格納する際、順序番号として、当該論理ボリュームに対するデータ更新および属性更新の両者を併せた通しでの処理順番を示す連続番号を書込む。また、論理ボリューム毎に順序番号を付与する代わりに、あらかじめ正側・副側で同一に設定された論理ボリュームを複数含む「論理ボリュームグループ」の中で、当該グループに対する更新処理の番号を付与することもできる。これにより複数論理ボリュームを併用するプログラムに対しても、リモートコピーの更新の順序性を保証する。

【００４９】

順序性を保証するとは、送信されたデータや属性情報の全てが、順序正しく受信側でソート処理等により復元されることを意味する。言い換えれば、ホストから要求された順序を維持しつつ、コピー先のデータ及び属性情報が更新されていくことをいう。つまり、何らかの原因でデータや属性情報のコピーが中断した場合であっても、コピー元で、ある特定時点までに更新されたデータや属性情報の全てが欠落なくコピー先に反映されていて、かつ、コピー元で当該時点以降に更新されたデータ又は属性情報はコピー先では一切反映されていないことを意味している。伝送の途中で、データ又は属性情報が喪失した場合には、その伝送（コピー動作）をエラーと見なして処理を中断するか、何らかの手段で再送して復旧させる。

【００５０】

続いて、チャネルアダプタ２２上のマイクロプロセッサ２８は、独自のスケジューリングにより動作し、キャッシュ上のデータ更新情報３２および共有メモリ上の属性情報更新情報３５を、任意の順序で副側のチャネルアダプタ２２に転送する（Ｓ２７）。副側のチャネルアダプタ２２上のマイクロプロセッサ２８は、受信した更新情報を、データならば副側キャッシュ２４上のデータ更新情報３２として、属性情報ならば副側共有メモリ２５上の属性情報更新情報３５として、一旦、書込む（Ｓ２８）。

【００５１】

副側チャネルアダプタ上のマイクロプロセッサは、さらに別な独自のスケジューリングで、前記の各更新情報を論理ボリュームないし論理ボリュームグループ内での順序番号によりソートし（Ｓ２９）、順序番号の不連続が無くなった部分から、同期処理での処理（Ｓ１３ないしＳ１８）と同様の更新処理を行う（Ｓ３０）。この処理により、副側でも、論理ボリュームないしは論理ボリュームグループ内で、データならびに属性情報の更新が、両者を通した要求発行順で更新されることとなる。

【００５２】

なお、一つの制御指示の結果、複数の属性情報を生成する場合もあり得る。この場合、同期転送であれば全ての属性情報の転送処理（図３：Ｓ１７、Ｓ１２、Ｓ１８の繰り返し）を完了させ、非同期転送であれば全ての属性情報の格納処理（図４：Ｓ２５、Ｓ２６の繰り返し）を完了させてから、ホスト計算機に完了通知を行う。

【００５３】

ここまでの説明では、ホスト計算機からのデータ更新ならびに制御指示は、各々、単一のコマンドに単一の要求が含まれることを前提とした。その他に、単一のコマンドに複数の制御指示が含まれていたり、データ更新コマンドに制御指示が含まれていたりという場合もあり得る。この場合、同期転送であれば、全ての制御指示に関する処理（Ｓ１６、Ｓ１７、Ｓ１２、Ｓ１８、Ｓ１４の繰り返し）を行なった後、データ更新の処理（Ｓ１１、Ｓ１２、Ｓ１３、Ｓ１４）を行い、ホスト計算機に完了通知を行う（Ｓ１５）。非同期転送であれば、全ての制御指示に関する処理（Ｓ２４、Ｓ２５、Ｓ２６の繰り返し）を行なった後、データ更新の処理（Ｓ２１、Ｓ２２）を行い、ホスト計算機に完了通知を行う（Ｓ２３）。

【００５４】

＜中断と逆転＞

図７を使用して、リモートコピーの中断と逆転に関する処理について説明する。リモートコピーのコピー状態は、論理ボリューム毎に管理されている。「正常」状態とは、正副で同期が取れている、すなわち非更新データは全て一致していて更新データのみコピーが実行されている状態を指す（図７（ａ））。この状態では、データの一貫性を保つため、副側のホスト計算機は、当該論理ボリュームの内容を更新できない。

【００５５】

正常状態にあるとき、プログラム等からの指示により、リモートコピーを中断させることができる（図７（ｂ））。この「中断」状態では、副側のホスト計算機からも当該論理ボリュームの内容を更新可能であり、正側でプログラムは動作したまま、副側でバックアップを取得したり、開発中のプログラムを実行テストしたり、といったことが可能である。

中断状態からは、プログラム等からの指示により、同じコピーの向きでリモートコピーを再開させることと（図７（ｃ））、逆向きでリモートコピーを再開させるテイクオーバ（図７（ｄ））が可能である。

【００５６】

いずれの場合も、両サブシステム上で「中断」状態の間に更新されたデータないし属性情報の位置をそれぞれ記録しており（図７（ｂ）、３３、３６）、いずれかのサブシステムにおいて更新されたデータないし属性情報についてのみ正側の現在値を副側にコピーすることにより、当該論理ボリュームのコピー状態を「正常」状態に復旧させる。この動作を再同期と呼び、コピー内容を更新情報に限定することで、「正常」状態復旧までの時間が短縮される。再同期が完了するまでは、副側の論理ボリュームは整合性はないため、更新順序に意味は無く、順序は無関係にコピーを行う。

【００５７】

「中断」状態の間の更新位置の記録について、さらに説明する。キャッシュ２４上にデータの更新位置を記録するビットマップ３３が存在し（図１）、対応する論理ボリュームの管理単位（トラック・シリンダ・複数のシリンダを含むグループ）ごとに更新フラグビットを持つ。「中断」状態の間にデータが更新された場合、当該データを含む管理単位に対応する更新フラグビットをセットする。再同期時には、いずれかのサブシステムで更新フラグビットのセットされている管理単位について、当該管理単位全体のデータを正側から副側へコピーする。

【００５８】

一方、本実施例ではこれに属性情報についての更新管理・再同期が追加される。属性情報の更新管理については、共有メモリ２５上に属性情報の更新位置を記録する属性情報更新テーブル３６が存在する。これは図２で示した属性情報テーブル３４と同様の行と列の構成を持っていて、各列の内容が属性情報そのものの代わりに更新されたことを示す更新フラグビットになっている。

【００５９】

各サブシステムでは、リモートコピーが「中断」状態に遷移するタイミングで、当該論理ボリュームに対応する行の全ての更新フラグをクリアする。「中断」状態の間は、属性情報の更新を行うチャネルアダプタ上のプロセッサが、属性情報本体を更新するのと同時に、対応する属性情報更新テーブル上の更新フラグをセットする。再同期時には、いずれかのサブシステムで属性情報更新テーブル上の更新フラグがセットされている全てについて、当該フラグが指す属性情報の正側の現在値を副側へコピーして、内容の一致化を行なう。

【００６０】

図７（ｃ）及び（ｄ）いずれの場合も、いずれかのデータ記憶サブシステムで更新された、データ及び属性情報を、全て正側から副側へコピーして再同期を行う。再同期完了後、正常状態に遷移する。

【００６１】

＜実施例２＞

次に本発明の第二の実施例を図８に示す。

図８では、第一から第四までのデータ記憶サブシステム１１ａ～ｄが配置されている。第一の実施例と全く同一の構成要素、すなわち制御装置１２・記憶装置１３・ディスクアダプタ２３・各アダプタ上のポート２７およびマイクロプロセッサ２８・磁気ディスクドライブ４１・ディスク入出力パス４２については、説明を省略する。また、以後の説明では、第一サブシステムをコピー元とするリモートコピーの動作のみを説明するため、第一サブシステム以外へのホスト計算機の接続も図示しない。

【００６２】

本実施例では、第一サブシステム１１ａと第二サブシステム１１ｂ・第三サブシステム１１ｃとの間、および第三サブシステム１１ｃと第四サブシステム１１ｄとの間が、データ転送パス６３にて接続されている。このような構成で、第一データ記憶サブシステム１１

ａに接続されたホスト計算機が使用する論理ボリュームを、第二から第四までの三台のサブシステムにリモートコピーすることが可能である。具体的には、第一サブシステムから第二・第三サブシステムに並行してリモートコピーを実施し、また第三サブシステムは、リモートコピー機能で第一サブシステムから受信した更新情報を、さらに第四サブシステムに転送する。

【００６３】

第一サブシステムの動作を説明する。

同期転送の場合、図３に示した処理フローのうち、データ転送パスにて更新情報を送信する処理で、リモートコピーのコピー先となる複数の副側サブシステムへ全て送信し（Ｓ１２）、全てのサブシステムから転送完了通知が来た時点でホスト計算機に完了通知を行なう（Ｓ１５）。

【００６４】

一方、非同期転送の場合、図４に示した処理フローのうち、正側キャッシュ２４へのデータ更新情報３２あるいは正側共有メモリ２５への属性情報更新情報３５への書込み処理で、それぞれリモートコピーのコピー先となるサブシステムの台数分だけ領域を確保して同一内容を書込み（Ｓ２２、Ｓ２６）、以後の更新情報を送信する処理はコピー先毎に独立して行なう。

【００６５】

いずれの場合でも、リモートコピーの中断については、コピー先の各サブシステム毎に独立に行われるため、中断中の更新情報を管理する更新ビットマップ３３ならびに更新属性情報テーブル３６は、コピー先の副側サブシステムの台数分だけ用意される。

【００６６】

以上の処理により、複数台のサブシステムに対して、データならびに属性情報のリモートコピーを実現する。

【００６７】

続いて第三サブシステムの動作を説明する。

第一サブシステムとのリモートコピーの関係が同期転送ならば、第一サブシステムから受信した更新情報を、ホスト計算機からのデータ更新ないしは制御指示とみなし第四サブシステムに対するリモートコピーを実施する。第四サブシステムに対するリモートコピーの処理は、第一の実施例で説明した同期ないしは非同期転送の場合の処理と同一である。

【００６８】

一方、第一サブシステムとのリモートコピーの関係が非同期転送ならば、既に要求元のホスト計算機との同期関係は崩れているので、第四サブシステムに対しては非同期転送によるリモートコピーが一般的に使用される。この場合は、第一サブシステムから受信した更新情報を、自サブシステムへの更新に使うために一時記憶する他に、二次のリモートコピーを実施するコピー先の副側サブシステム台数分だけ別に記憶する。この別に記憶した更新情報を使用して、第一の実施例の図４の処理フローのうち、更新情報をデータ転送パスで転送する処理（Ｓ２７）以降の処理を実施する。キャッシュ２４上には自サブシステムでの更新用と第四サブシステムへの送信用の二種類の受信更新情報３２が格納され、共有メモリ２５上には自サブシステムでの更新用と第四サブシステムへの送信用の二種類の受信更新情報３５が格納される。

【００６９】

いずれの場合でも、リモートコピーの中断については、受信側（第一サブシステムとの間）・送信側（第四サブシステムとの間）で独立に行われるため、中断中の更新情報を管理する更新ビットマップ３３ならびに更新属性情報テーブル３６は、リモートコピーの相手サブシステム台数分だけ用意される。

【００７０】

以上の処理により、リモートコピーの副側サブシステムにおいて二次的に別のサブシステムに対して、データならびに属性情報のリモートコピーを実現することができる。

【００７１】

本実施例の仕組みを複数組み合わせることにより、任意の数のサブシステム間で、データならびに属性情報のリモートコピーを実現することができる。

【００７２】

【発明の効果】

本発明によれば、ホスト計算機側では全く意識することなしに、属性情報がバックアップ側のデータ記憶サブシステムにコピーされ、バックアップ側で自動的に利用可能となる。これにより、ホスト計算機間での複雑な属性情報の交換が不要となる。

【００７３】

本発明によれば、ホスト計算機側では全く意識することなしに、データおよび属性情報の更新の順序を維持したまま、しかも高速に、コピー先のデータ記憶サブシステム上にこれらの更新が反映される機能が実現できる。これにより、属性情報の更新とデータの更新の間の順序関係を維持することができ、障害時の復旧処理等が容易になる。

【００７４】

本発明によれば、ホスト計算機側では全く意識することなしに、リモートコピー中断後の再開時やコピー方向反転時での属性情報の差分コピーを実現できる。これにより、短時間で属性情報の再同期を終了できる。

【００７５】

本発明によれば、複数のデータ記憶サブシステムにまたがる属性情報のコピーを実現できる。これにより、複数のデータ記憶サブシステムにまたがるリモートコピーの適用が容易となり、耐障害性をより高くしつつ、システムの構築が容易となる。

【図面の簡単な説明】

【図１】本発明の複合データ記憶サブシステムの全体構成図である。

【図２】属性情報テーブルの構成である。

【図３】同期転送リモートコピーの場合の処理フローである。

【図４】非同期転送リモートコピーの場合の処理フローである。

【図５】データ転送パス上のデータ更新情報のライトコマンドである。

【図６】データ転送パス上の属性情報更新情報のライトコマンドである。

【図７】リモートコピーの中断に関する処理を示す図である。

【図８】本発明の別な実施例の全体構成図である。

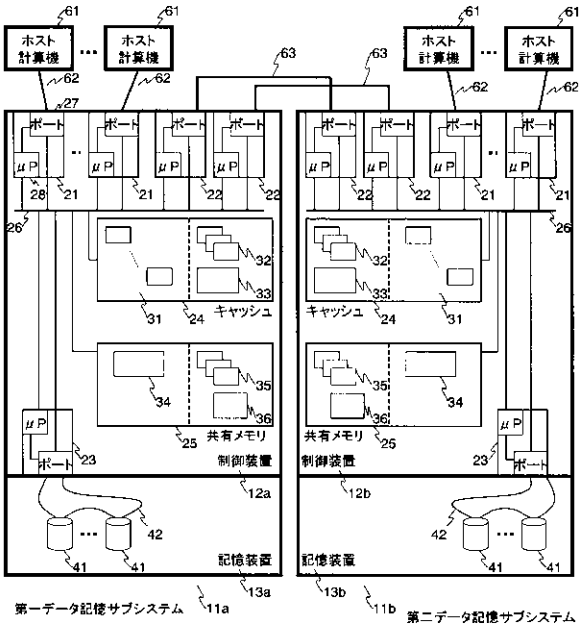
【図９】同期転送方式と非同期転送方式を説明する図である。

【符号の説明】

１１ａ、１１ｂ、１１ｃ、１１ｄ…データ記憶サブシステム、
１２ａ、１２ｂ…制御装置、
１３ａ、１３ｂ…記憶装置、　２１、２２…チャネルアダプタ、
２３…ディスクアダプタ、　２４…キャッシュ、
２５…共有メモリ、　２６…バス、
２７…ポート、　２８…マイクロプロセッサ、
３１…キャッシュエントリ、　３２…データ更新情報、
３３…更新情報、　３４…属性情報テーブル、
３５…属性情報更新テーブル、　３６…更新属性情報テーブル、
４１…磁気ディスクドライブ、　４２…ディスク入出力パス、
６１…ホスト計算機、　６２…ホストアクセスパス、
６３…データ転送パス。

【図１】

【図1】 全体構成図



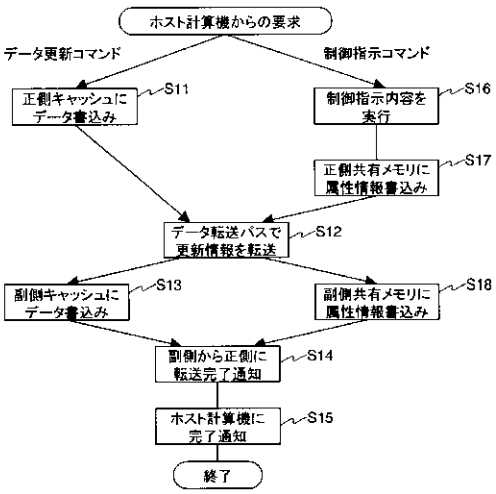
【図２】

【図2】 属性情報テーブルの構成

LUN	リザーブ制御情報	更新アクセス許可キー	使用中ホストID	・・・・・・・・・・	ユーザ用領域
00:00	0	321A	—	・・・・・・・・・・	
00:01	1, TID=7	—	—	・・・・・・・・・・	3FC00129 ・・・・・・
⋮	⋮	⋮	⋮	⋱	⋮
FF:FF	0	—	00-00-3C	・・・・・・・・・・	

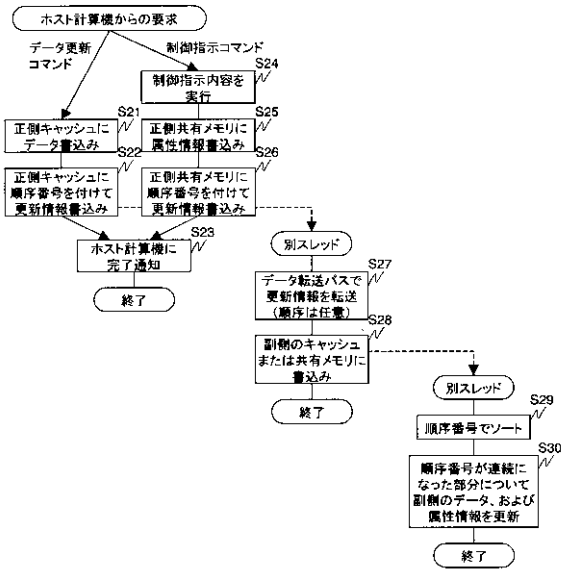
【図３】

【図3】 同期転送の場合の処理フロー



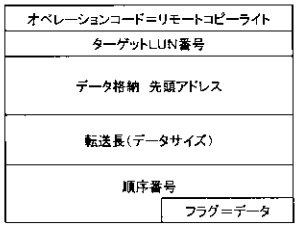
【図４】

【図4】 非同期転送の場合の処理フロー



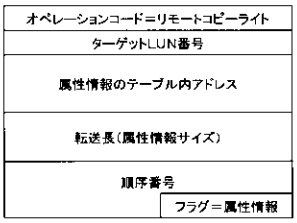
【図５】

【図5】 データ転送パス上のコマンド(データ)



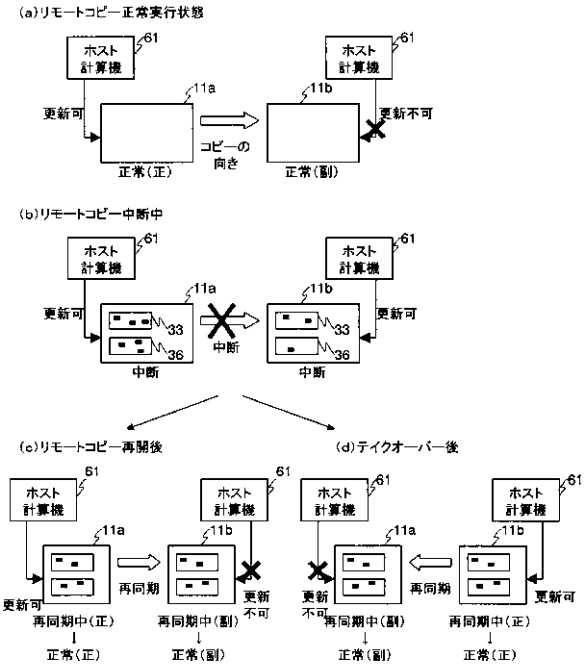
【図６】

【図6】 データ転送パス上のコマンド(属性情報)



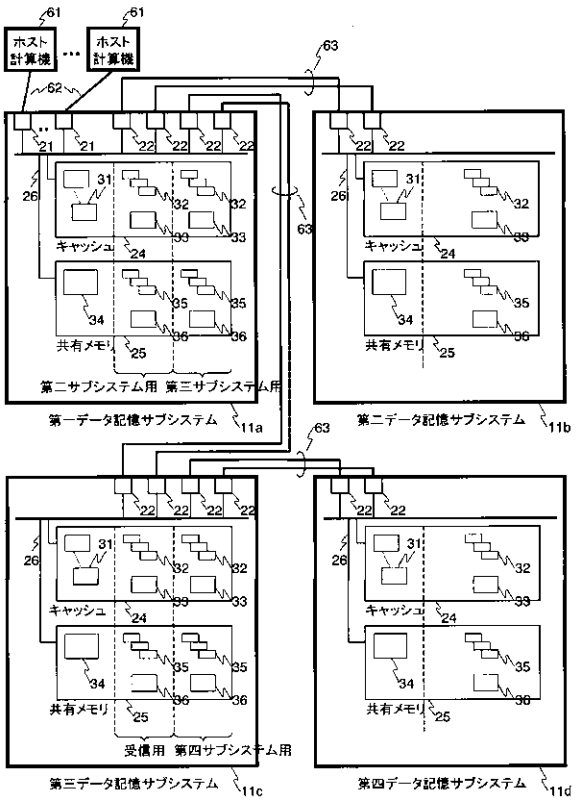
【図７】

【図7】 リモートコピーの中断



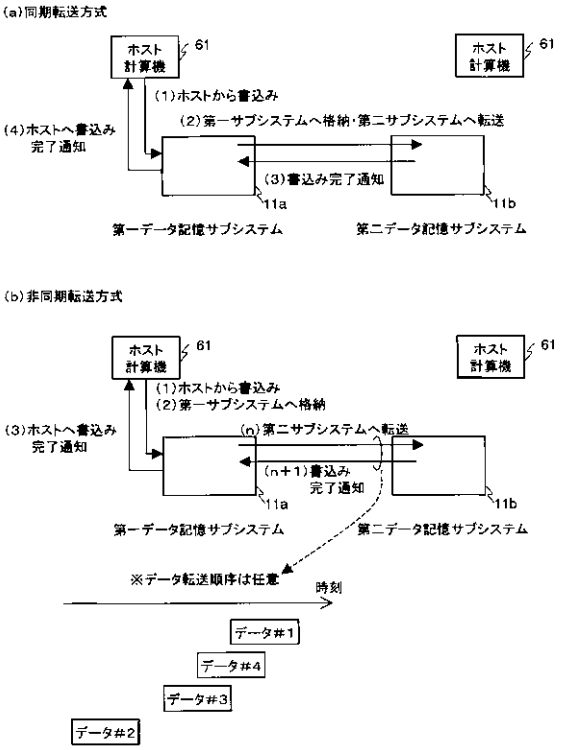
【図８】

【図8】 別な実施例の全体構成図



【図９】

【図9】 同期転送方式と非同期転送方式の説明



フロントページの続き

(72)発明者　中野　俊夫
神奈川県小田原市中里３２２番地２号　株式会社日立製作所ＲＡＩＤシステム事業部内
(72)発明者　田渕　英夫
神奈川県小田原市中里３２２番地２号　株式会社日立製作所ＲＡＩＤシステム事業部内

Ｆターム(参考) 5B065 BA01　CA12　CC08　CE22　EA34　EA35
5B082 EA11　HA03