(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第3613722号
(P3613722)
(45)発行日　平成17年1月26日(2005.1.26)　(24)登録日　平成16年11月12日(2004.11.12)

(51)Int.Cl.[7]　F I
G06F 3/06　G06F 3/06 305F
G06F 3/06 540

請求項の数 1　(全 17 頁)

(21)出願番号　特願平9-50520
(22)出願日　平成9年3月5日(1997.3.5)
(65)公開番号　特開平10-247133
(43)公開日　平成10年9月14日(1998.9.14)
審査請求日　平成13年7月17日(2001.7.17)

(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　100095511
弁理士　有近　紳志郎
(72)発明者　兼田　泰典
神奈川県川崎市麻生区王禅寺１０９９番地
株式会社日立製作所　システム開発研究所内
(72)発明者　山本　彰
神奈川県川崎市麻生区王禅寺１０９９番地
株式会社日立製作所　システム開発研究所内

最終頁に続く

(54)【発明の名称】ディスクアレイ装置

(57)【特許請求の範囲】
【請求項１】
ディスク制御装置と、該ディスク制御装置によりアクセスが制御される複数のディスク装置とを有し、前記複数のディスク装置は、コンピュータによりアクセスされるデータを格納するデータ領域を有する少なくとも２つの第１のディスク装置と、前記２つのディスク装置の互いに対応するデータ領域に格納されたデータに対する冗長データを格納する前記データ領域に対応する冗長データ格納領域を有する第２のディスク装置とを含むディスクアレイ装置であって、
前記第１のディスク装置は、
前記ディスク制御装置が受けるコンピュータからのデータの読み出し要求に対し、該データを格納したデータ領域から前記データを読み出しできないセクタ障害を検出した場合、前記読み出しを中断し、前記ディスク制御装置にエラー報告を行う手段
を有し、
前記ディスク制御装置は、
前記読み出し要求に対し、該データを格納したデータ領域から前記データを、前記データを格納したデータ領域に対応する他のデータ領域から他のデータを、前記データを格納したデータ領域に対応する前記冗長データ領域から冗長データを読み出す手段と、
前記エラー報告を受けた場合、前記読み出した他のデータと前記冗長データとを用いて前記データを復元する手段と、
前記エラー報告を受けることなく前記データの読み出しに成功した場合は前記読み出し

たデータを前記コンピュータに転送し、前記エラー報告を受けた場合は前記復元したデータを前記コンピュータに転送する手段と
を有することを特徴とするディスクアレイ装置。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】
本発明は、ディスクアレイ装置に関し、特に、セクタ障害が発生しても、最大応答時間を低減または保障することが出来るディスクアレイ装置に関する。
【０００２】
【従来の技術】
●ＲＡＩＤとは
論文：”Ａ Ｃａｓｅ ｆｏｒ Ｒｅｄｕｎｄａｎｔ Ａｒｒａｙｓ ｏｆ Ｉｎｅｘｐｅｎｃｉｖｅ Ｄｉｓｋｓ （ＲＡＩＤ）”, Ｄａｖｉｄ Ａ．Ｐａｔｔｅｒｓｏｎ, Ｇａｒｔｈ Ｇｉｂｓｏｎ, ａｎｄ Ｒａｎｄｙ Ｈ．Ｋａｔｚ, Ｃｏｍｐｕｔｅｒ Ｓｃｉｅｎｃｅ Ｄｉｖｉｓｉｏｎ Ｄｅｐａｒｔｍｅｎｔ ｏｆ Ｅｌｅｃｔｒｉｃａｌ Ｅｎｇｉｎｅｅｒｉｎｇ ａｎｄ Ｃｏｍｐｕｔｅｒ Ｓｃｉｅｎｃｅｓ, Ｕｎｉｖｅｒｓｉｔｙ ｏｆ Ｃａｌｆｏｒｎｉａ Ｂｅｒｋｅｌｅｙ）において、複数の磁気ディスク装置を並列に動かして読み出し／書き込みの高速化を図り、冗長構成によって信頼性を上げた外部記憶装置であるディスクアレイ装置（ＲＡＩＤ：Ｒｅｄｕｎｄａｎｔ Ａｒｒａｙｓ ｏｆ Ｉｎｅｘｐｅｎｃｉｖｅ Ｄｉｓｋｓ）が開示されている。
前記論文の中では、ディスクアレイ装置の構成が、ＲＡＩＤレベル１からＲＡＩＤレベル５までの番号で分類されている。この中で、ＲＡＩＤレベル３は、大量のデータを転送するシーケンシャルアクセスに対し性能の向上が見込める。また、ＲＡＩＤレベル５は、小さなサイズの読み書きが大量に発生するランダムアクセスに対し性能の向上が見込める（日経エレクトロニクス１９９３／４／２６ Ｎｏ．５７９ Ｐ７７～９１ 「特集ディスクアレイ装置」）。
【０００３】
●ＲＡＩＤ制御とは
ＲＡＩＤ制御とは、複数のディスク装置にどのようにデータを分散させるかを制御すること及びどのように冗長データを作成しどこに格納するかを制御することである。
図１４は、３台のディスク装置ＤＩＳＫ＃０，ＤＩＳＫ＃１，ＤＩＳＫ＃２で構成するディスクアレイ装置の概念図である。
このディスクアレイ装置は、実際には、３台のディスク装置ＤＩＳＫ＃０，ＤＩＳＫ＃１，ＤＩＳＫ＃２を使用してデータを保持しているが、コンピュータシステムからは、Ｄ０からＤ７までの「ストライプ」と呼ばれる所定量（例えば３２ＫＢ）のデータ格納容量をもつデータ格納領域を連続して有する１台のディスク装置として論理的に認識される。
ディスク装置ＤＩＳＫ＃０にはストライプＤ０，Ｄ２，Ｄ４，Ｄ６が割り当てられ、ディスク装置ＤＩＳＫ＃１にはストライプＤ１，Ｄ３，Ｄ５，Ｄ７が割り当てられている。
ディスク装置ＤＩＳＫ＃２には冗長データ格納領域Ｐ０１，Ｐ２３，Ｐ４５，Ｐ６７が割り当てられている。冗長データ格納領域Ｐ０１にはストライプＤ０とストライプＤ１に格納しているデータの排他的論理和を格納し、冗長データ格納領域Ｐ２３にはストライプＤ２とストライプＤ３に格納しているデータの排他的論理和を格納し、冗長データ格納領域Ｐ４５にはストライプＤ４とストライプＤ５に格納しているデータの排他的論理和を格納し、冗長データ格納領域Ｐ６７にはストライプＤ６とストライプＤ７に格納しているデータの排他的論理和を格納する。
ストライプＤｍとストライプＤｎと冗長データ格納領域Ｐｍｎとをまとめて「ストライプグループ」と呼ぶ。
【０００４】
前記排他的論理和は、冗長データであり、１台のディスク装置が壊れたときにデータを復元するのに使用する。この「１台のディスク装置が壊れた」状態を「縮退」と呼ぶ。縮退

時におけるリード処理については後述する。
【０００５】
●リード処理
コンピュータシステムが例えばストライプＤ４とＤ５に格納されているアプリケーションの読み出しの要求をディスクアレイ装置に発行すると、ディスクアレイ装置はディスク装置ＤＩＳＫ＃１のストライプＤ４とディスク装置ＤＩＳＫ＃２のストライプＤ５の内容を読み出してコンピュータシステムに転送する。
【０００６】
●ライト処理
コンピュータシステムが例えばストライプＤ６とストライプＤ７にデータを格納するライト要求をディスクアレイ装置に発行すると、ディスクアレイ装置はストライプＤ６とストライプＤ７のデータの排他的論理和を演算し、次にストライプＤ６とストライプＤ７にデータを書き込むと共に冗長データ格納領域Ｐ６７に前記排他的論理和を書き込む。
【０００７】
●縮退時におけるリード処理
例えばディスク装置ＤＩＳＫ＃２が故障しているときにコンピュータシステムがストライプＤ４とＤ５に格納されているアプリケーションの読み出しの要求をディスクアレイ装置に発行すると、ディスクアレイ装置はディスク装置ＤＩＳＫ＃１のストライプＤ４とディスク装置ＤＩＳＫ＃２のストライプＤ５の内容を読み出そうとする。しかし、ディスク装置ＤＩＳＫ＃１のストライプＤ４の内容は読み出せるが、故障しているディスク装置ＤＩＳＫ＃２からストライプＤ５の内容は読み出せない。すると、ディスクアレイ装置はディスク装置ＤＩＳＫ＃３の冗長データ格納領域Ｐ４５の内容を読み出し、ストライプＤ４と冗長データ格納領域Ｐ４５の内容の排他的論理和を演算し、ストライプＤ５の内容を復元する。そして、ストライプＤ４の内容と復元したストライプＤ５の内容をコンピュータシステムに転送する。なお、ディスク装置ＤＩＳＫ＃２が故障していることをディスクアレイ装置が予め知っていれば、ディスク装置ＤＩＳＫ＃２に対するリード処理は省略する。
【０００８】
●ディスク装置
ディスク装置は、回転する円板状の記憶媒体（磁気記録媒体や光記録媒体や光磁気記憶媒体など）と、その記憶媒体上を半径方向に移動するヘッドを具備して構成されている。記憶媒体は、いくつものトラックと呼ばれる同心円状の領域に区切られ、各トラックはセクタという単位に区切られている。
ディスク装置は、セクタ単位（多くは５１２バイト）にデータを管理する。あるセクタを読み出すとき、ヘッドを目的のトラックに移動させ、ヘッドの下に目的のセクタが来るまで待ち、記憶媒体の回転にあわせてデータを読み出す。読み出したデータは、シリアルデータなので、パラレルデータに変換し、バッファに保持する。バッファに一定量のデータが蓄えられると、ＳＣＳＩやファイバチャネルやＩＤＥのインタフェースを介してデータを転送する。書き込みは、上記と逆の手順により行われる。
ディスク装置からデータを読み出す、または、ディスク装置にデータを書き込む、ために必要な時間は、
（１）ヘッドが目的のトラックに移動するまでの時間（シーク時間）
（２）ヘッドの下に目的のセクタが来るまでの待ち時間（回転待ち時間）
（３）データを読み書きする時間（データ転送時間）
によって決まる。
データの読み出しの要求を発行してから読み出し終了が報告されるまでの時間を応答時間と呼ぶ。
【０００９】
●セクタ障害とリトライ
セクタ障害とは、記憶媒体上の一部分を読み書きできなくなる障害である。セクタ障害が発生する原因としては、記憶媒体上の傷や磁性体の劣化などが考えられる。ある程度のセ

クタ障害は、ＥＣＣ（Ｅｒｒｏｒ　Ｃｏｒｒｅｃｔｉｎｇ　Ｃｏｄｅ）を付加することで復旧可能である。復旧可能なセクタ障害の程度は、ＥＣＣの種類による。
ＥＣＣによる復旧が不可能な場合、ディスク装置は、読み書きを再試行する。これをリトライという。すなわち、リトライとは、読み書きのできなかったセクタに対して、１回転待ってから当該セクタに対して読み書きを再び試みることである。リトライが行われると、応答時間が伸びてしまう。リトライを数回行っても読み書きできない場合は、エラーとして読み書きを中止する。

【００１０】

●交代セクタ

交代セクタとは、障害の発生したセクタの代りに使用するセクタである。書き込み時にセクタ障害が発生した場合、そのセクタにデータを書き込むことができずエラーとなる。そこで、交代セクタを要求すると、別のセクタが交代セクタとして割り当てられる。そこで、エラーの発生したセクタの代りに、交代セクタを利用する。

【００１１】

【発明が解決しようとする課題】

従来のディスクアレイ装置の制御方法およびディスク装置では、次に述べるように応答時間が伸びてしまう問題点があった。

●データを復元するためのデータを読み出すための応答時間の伸び

セクタ障害の発生を認識すると、読み出しに失敗したデータを復元するためのデータと冗長データの読み出しを開始し、それらデータを読み出してからデータを復元する。このため、データを復元するためのデータと冗長データの読み出しに要する時間が応答時間に直列に入るため、応答時間が長くなってしまう。

●リトライによる応答時間の伸び

セクタ障害が発生しリトライが起こると、ヘッドの下に目的のセクタが来るまで１回転待たなければならない。例えば、７２００ｒｐｍで記憶媒体が回転している場合、１回転の時間は約８．３ｍｓ（＝６０ｓｅｃ÷７２００ｒｐｍ）であるから、Ｎ回のリトライが発生した場合、応答時間が（８．３×Ｎ）ｍｓ伸びてしまう。

●交代セクタによる応答時間の伸び

セクタを連続に読み書きしているシーケンシャルリードライト処理時に、障害セクタに当たると、その障害セクタに対応する交代セクタのあるトラックに移動し、交代セクタがヘッドの下に来るのを待って読み書きし、交代セクタが終了したら、元のトラックに戻って読み書きを続ける。つまり、交代セクタに当たると、余分なシーク時間と回転待ちが発生し、応答時間が伸びてしまう。

●エラー報告による応答時間の伸び

エラーが発生すると、その時点でデータ転送を中止し、エラー報告を行う。エラー報告を受けると、エラーが発生したセクタ以後のデータに対して新たに要求を発行する。しかし、かかる要求の発行に伴う処理時間の増加のため、平均的に見れば、応答時間が伸びてしまう。

【００１２】

そこで、本発明の目的は、セクタ障害が発生しても、最大応答時間を低減または保障することが出来るディスクアレイ装置を提供することにある。

【００１３】

【課題を解決するための手段】

第１の観点では、本発明は、ディスク制御装置と、該ディスク制御装置によりアクセスが制御される複数のディスク装置とを有し、前記複数のディスク装置は、コンピュータによりアクセスされるデータを格納するデータ領域を有する少なくとも２つの第１のディスク装置と、前記２つのディスク装置の互いに対応するデータ領域に格納されたデータに対する冗長データを格納する前記データ領域に対応する冗長データ格納領域を有する第２のディスク装置とを含むディスクアレイ装置であって、前記第１のディスク装置は、前記ディスク制御装置が受けるコンピュータからのデータの読み出し要求に対し、該データを格納

したデータ領域から前記データを読み出しできないセクタ障害を検出した場合、前記読み出しを中断し、前記ディスク制御装置にエラー報告を行う手段を有し、前記ディスク制御装置は、前記読み出し要求に対し、該データを格納したデータ領域から前記データを、前記データを格納したデータ領域に対応する他のデータ領域から他のデータを、前記データを格納したデータ領域に対応する前記冗長データ領域から冗長データを読み出す手段と、前記エラー報告を受けた場合、前記読み出した他のデータと前記冗長データとを用いて前記データを復元する手段と、前記エラー報告を受けることなく前記データの読み出しに成功した場合は前記読み出したデータを前記コンピュータに転送し、前記エラー報告を受けた場合は前記復元したデータを前記コンピュータに転送する手段とを有することを特徴とするディスクアレイ装置を提供する。

上記第１の観点のディスクアレイ装置では、障害が発生するか否かに関わらず、データの復元に必要なデータと冗長データとを読み出しておく。これにより、データを復元するためのデータと冗長データの読み出しに要する時間が応答時間に並列に入り、直列に入らないため、最大応答時間を低減することが出来る。

【００２３】

【発明の実施の形態】

以下、本発明の実施形態を図を用いて説明する。なお、これにより本発明が限定されるものではない。

【００２４】

《第１の実施形態》

図１は、本発明の第１の実施形態のディスクアレイ装置を含むコンピュータシステムの構成図である。

●コンピュータシステムの構成

このコンピュータシステムは、ホストコンピュータ１０と、そのホストコンピュータ１０に接続されたディスクアレイ装置１００と、そのディスクアレイ装置１００に接続された複数のディスク装置２５１～２５５とを具備して構成されている。

前記ホストコンピュータ１０とディスクアレイ装置１００とは、ＳＣＳＩバス２０を介して接続されているが、ＰＣＩ（Ｐｅｒｉｐｈｅｒａｌ Ｃｏｍｐｏｒｎｅｎｔ Ｉｎｔｅｒｃｏｎｎｅｃｔ）やファイバーチャネル，ＩＥＥＥ１３９４などの接続手段を用いて接続してもよい。

【００２５】

前記ディスクアレイ装置１００は、前記ホストコンピュータ１０に接続するためのホスト接続手段１１０と、前記ディスク装置２５１～２５５に接続するためのディスク接続手段２００と、前記ホスト接続手段１１０や前記ディスク接続手段２００を介して転送するデータを保持する記憶手段１５０と、前記ホスト接続手段１１０と前記ディスク接続手段２００と前記記憶手段１５０を接続するデータバス１２０と、前記記憶手段１５０に接続されたデータ復元手段１６０と、前記記憶手段１５０に接続された冗長データ生成手段１７０と、前記各手段を接続する制御バス１３０と、その制御バス１３０を介して前記各手段を制御する制御手段１９０とを具備している。

前記ディスク接続手段２００とディスク装置２５１～２５５とは、ＳＣＳＩバス２０１を介して接続されているが、ＳＣＳＩバス以外の接続手段を用いても接続してもよい。

【００２６】

●ディスク装置の構成

図２は、ディスク装置（２５１～２５５）の構成図である。

ディスク装置（２５１～２５５）は、円板状の記憶媒体３００と、その記憶媒体３００を回転させるモータ３２０と、そのモータ３００の回転数を一定に保つモータ制御手段３２５と、前記記憶媒体３００上を半径方向に移動可能なヘッド３１０と、そのヘッド３１０を前記記憶媒体３００上に位置付けるヘッド制御手段３１５と、前記ヘッド３１０から読み出した信号を整形する波形整形手段３５０と、読み出したデータの正当性をチェックするＥＣＣ手段３５２と、読み出したデータを一時的に保持しておくバッファ３５４と、そ

のバッファ３５４のデータを外部へ転送するための転送手段３５６と、前記各手段を制御するコントローラ３９０とを具備している。前記転送手段３５６は、前記ＳＣＳＩバス２０１に接続される。

【００２７】

図３は、記憶媒体３００のデータ格納構造の説明図である。

データの格納は、記録媒体３００上に作られる同心円状のトラック３０１と、そのトラック３０１を区切って作られるセクタ３０２という単位で行われる。各セクタ３０２は、位置情報領域３０３と、データ領域３０４と、ＥＣＣ領域３０５とで構成されている。前記位置情報領域３０３には、トラック番号やセクタ番号が記録されている。この位置情報領域３０３内の情報をチェックすることにより、目的のトラックにヘッド３１０が正しく移動できていることを確認し、目的のセクタが来るのを待つ。前記データ領域３０４には、データを格納する。前記ＥＣＣ領域３０５には、前記データ領域３０４にデータを書き込んだ時に、そのデータから演算したＥＣＣを保持する。データ領域３０４からデータを読み出した時、その読み出したデータが正しいか否かを、ＥＣＣ領域３０５からＥＣＣを読み出して、チェックする。データに対するＥＣＣが一致しない場合には、可能ならＥＣＣを用いてデータの誤りを訂正する。

【００２８】

●データと冗長データの配置

図４は、データと冗長データの配置を示す説明図である。

ディスク装置２５１にはストライプＤ０，Ｄ４，Ｄ８，…が割り当てられ、ディスク装置２５２にはストライプＤ１，Ｄ５，Ｄ９，…が割り当てられ、ディスク装置２５３にはストライプＤ２，Ｄ６，Ｄ１０，…が割り当てられ、ディスク装置２５４にはストライプＤ３，Ｄ７，Ｄ１１，…が割り当てられ、データは、ストライプＤ０，Ｄ１，Ｄ２，Ｄ３，…の順にディスク装置２５１～２５４を巡回するように保持する。

ディスク装置２５５には冗長データ格納領域Ｐ１－３，Ｐ４－７，Ｐ８－１１，…が割り当てられて、冗長データは、それら冗長データ格納領域Ｐ１－３，Ｐ４－７，Ｐ８－１１，…に保持する。冗長データ格納領域Ｐ１－３に格納する冗長データは、ストライプＤ０～Ｄ３に保持したデータの排他的論理和である。他の冗長データ格納領域Ｐ４－７，Ｐ８－１１　に格納する冗長データも同様である。

【００２９】

●リード処理

ホストコンピュータ１０は、例えばストライプＤ５の読み出しの要求を発行する。

制御手段１９０は、ホスト接続手段１１０を介して、前記要求を受け取る。そして、ストライプＤ５を含むストライプグループすなわちストライプＤ４，Ｄ５，Ｄ６，Ｄ７および冗長データ格納領域Ｐ４－７　の内容を読み出すように、制御バス１３０を介して、５つのディスク接続手段２００に指令する。

各ディスク接続手段２００は、要求された内容を各ディスク装置２５１～２５５から読み出し、データバス１２０を介して、記憶手段１５０に転送する。

制御手段１９０は、記憶手段１５０に転送されたストライプＤ５のデータを、ホスト接続手段１１０を介して、ホストコンピュータ１０に転送する。

【００３０】

●セクタ障害時のリード処理

ホストコンピュータ１０は、例えばストライプＤ５の読み出しの要求を発行する。

制御手段１９０は、ホスト接続手段１１０を介して、前記要求を受け取る。そして、ストライプＤ５を含むストライプグループすなわちストライプＤ４，Ｄ５，Ｄ６，Ｄ７および冗長データ格納領域Ｐ４－７　の内容を読み出すように、制御バス１３０を介して、５つのディスク接続手段２００に指令する。

各ディスク接続手段２００は、要求された内容を各ディスク装置２５１～２５５から読み出そうとする。ここで、ストライプＤ５の読み出し中にディスク装置２５２でセクタ障害が発生したとする。

ディスク装置２５２は、リトライを実施しない。
制御手段１９０は、制御バス１３０を介して、ストライプＤ５でセクタ障害が発生したことを認識すると、ストライプＤ４，Ｄ６，Ｄ７および冗長データＰ４－７　の内容が記憶手段１５０に転送されるのを待って、データ復元手段１６０を起動する。
データ復元手段１６０は、ストライプＤ４，Ｄ６，Ｄ７および冗長データＰ４－７　の内容を用いてストライプＤ５のデータを記録手段１５０に復元する。
制御手段１９０は、復元されたストライプ５のデータを、記憶手段１５０から、ホスト接続手段１１０を介して、ホストコンピュータ１０に転送する。
【００３１】
●応答時間
図５は、応答時間の説明図である。
応答時間は、処理時間５１０，５２０，５３３（５３１，５３２＋５３９，５３３，５３４および５３５は、シークや回転待ちなどに依存するため若干の差が生じるが、それらのうちで最長のもの），５９０，５９５および５９９の和になる。
処理時間５１０では、ホスト接続手段１１０が、ホストコンピュータ１０からの読み出しの要求を受け取り、その要求を解釈する。
処理時間５２０では、ディスク装置２５１～２５５に対して読み出しを指令する。
処理時間５３１～５３５では、各ディスク装置２５１～２５５で読み出しが行われる。そして、処理時間５３１，５３３，５３４および５３５では、読み出しが成功し、読み出された内容が記憶手段１５０に転送される。一方、処理時間５３２では読み出しが失敗し、処理時間５３９でセクタ障害を認識し、リトライを実施せず、読み出しを中断する。
処理時間５９０では、読み出しに失敗したデータをデータ復元手段１６０が復元する。
処理時間５９５では、復元したデータをホストコンピュータに転送する。
処理時間５９９では、ホストコンピュータに終了を報告する。
【００３２】
比較のために、図６に、従来方法による応答時間を示す。
応答時間は、処理時間５１０，５２０，５３２，５３９，５４０＋５３９，５４０＋５３９，５４０＋５３９，５２０’，５３３（５３１，５３３，５３４および５３５のうちで最長のもの），５９０，５９５及び５９９の和になる。
処理時間５１０では、ホスト接続手段１１０が、ホストコンピュータ１０からの読み出しの要求を受け取り、その要求を解釈する。
処理時間５２０では、ディスク装置２５２に対して読み出しを指令する。
処理時間５３２では、ディスク装置２５２で読み出しが行われるが、読み出しに失敗し、処理時間５３９でセクタ障害を認識し、リトライを行う。
処理時間５４０では１回目のリトライを実施するが、読み出しに失敗し、処理時間５３９でセクタ障害を認識し、さらにリトライを行う。
次の処理時間５４０では２回目のリトライを実施するが、やはり読み出しに失敗し、処理時間５３９でセクタ障害を認識し、さらにリトライを行う。
次の処理時間５４０では３回目のリトライを実施するが、やはり読み出しに失敗し、処理時間５３９でセクタ障害を認識する。リトライの回数制限を３回とすると、この時点で読み出しを中断する。
処理時間５２０’では、ディスク装置２５１，２５３～２５５に対して読み出しを指令する。
処理時間５３１，５３３～５３５では、各ディスク装置２５１，２５３～２５５で読み出しが行われ、読み出された内容が記憶手段１５０に転送される。
処理時間５９０では、読み出しに失敗したデータをデータ復元手段１６０が復元する。
処理時間５９５では、復元したデータをホストコンピュータに転送する。
処理時間５９９では、ホストコンピュータに終了を報告する。
【００３３】
図５（本発明）と図６（従来）とを比較すれば明らかなように、図５（本発明）では、全

てのディスク装置に対するデータと冗長データの読み出しが並行して行われているため、セクタ障害が発生しても、応答時間は長くならない。これに対して、図６（従来）では、３回のリトライが行われた上に、セクタ障害の発生を認識した後で他のデータと冗長データの読み出しを開始するため、応答時間が長くなってしまう。

【００３４】

図７および図８は、全てのディスク装置に対するデータと冗長データの読み出しを並行して行なうと共にリトライを３回まで許す場合の応答時間を示している。そして、図７は３回のリトライでも読み出しに失敗した場合であり、図８は３回のリトライで読み出しに成功した場合である。

図７の場合は、図５と図６の中間的な応答時間となる。

図８の場合は、図５の場合と同等の応答時間となる。

【００３５】

●リトライしないディスク装置を用いた場合の応答時間

図９は、セクタ障害を検出してもリトライを行わないディスク装置を使用し、且つ、セクタ障害の発生を認識した後で他のデータと冗長データの読み出しを開始する場合の応答時間を示している。

応答時間は、処理時間５１０，５２０，５３２，５３９，５２０’，５３３（５３１，５３３，５３４および５３５のうちで最長のもの），５９０，５９５及び５９９の和になる。

処理時間５１０では、ホスト接続手段１１０が、ホストコンピュータ１０からの読み出しの要求を受け取り、その要求を解釈する。

処理時間５２０では、ディスク装置２５２に対して読み出しを指令する。

処理時間５３２では、ディスク装置２５２で読み出しが行われるが、読み出しに失敗し、処理時間５３９でセクタ障害を認識し、読み出しを中断し、セクタ障害を報告する。

処理時間５２０’では、ディスク装置２５１，２５３～２５５に対して読み出しを指令する。

処理時間５３１，５３３～５３５では、各ディスク装置２５１，２５３～２５５で読み出しが行われ、読み出された内容が記憶手段１５０に転送される。

処理時間５９０では、読み出しに失敗したデータをデータ復元手段１６０が復元する。

処理時間５９５では、復元したデータをホストコンピュータに転送する。

処理時間５９９では、ホストコンピュータに終了を報告する。

図９の場合は、図５と図６の中間的な応答時間となる。

【００３６】

●リトライアボートと応答時間

図１０および図１１は、全てのディスク装置に対するデータと冗長データの読み出しを並行して行なうと共にディスク処理の最大予定終了時間Ｔｍａｘを設けた場合の応答時間を示している。そして、図１０はリトライ中にディスク処理の最大予定終了時間Ｔｍａｘに達してリトライをアボートした場合であり、図１１はディスク処理の最大予定終了時間Ｔｍａｘ内のリトライで読み出しに成功した場合である。なお、ディスク処理の最大予定終了時間Ｔｍａｘは、使用するディスク装置の最大シーク時間，回転数，最低データ転送速度にあわせて予め制御手段１９０に設定しておく。

【００３７】

図１０の場合の応答時間は、処理時間５１０，５２０，５３２，５３９，５４０，５３９，５４０，５９０，５９５および５９９の和になる。

処理時間５１０では、ホスト接続手段１１０が、ホストコンピュータ１０からの読み出しの要求を受け取り、その要求を解釈する。

処理時間５２０では、ディスク装置２５１～２５５に対して読み出しを指令する。

処理時間５３１～５３５では、各ディスク装置２５１～２５５で読み出しが行われる。そして、処理時間５３１，５３３，５３４および５３５では、読み出しが成功し、読み出された内容が記憶手段１５０に転送される。一方、処理時間５３２では読み出しが失敗し、

処理時間５３９でセクタ障害を認識し、リトライを行う。
処理時間５４０では１回目のリトライを実施するが、読み出しに失敗し、処理時間５３９でセクタ障害を認識し、さらにリトライを行う。
次の処理時間５４０では２回目のリトライを実施するが、ディスク装置２５１～２５５に要求を発行してからディスク処理の最大予定終了時間Ｔｍａｘが経過してもディスク処理が終了しない。
そこで、処理時間５４９では、制御手段１９０は、ディスク装置５３２のディスク処理をアボート（中断）する。
処理時間５９０では、読み出しに失敗したデータをデータ復元手段１６０が復元する。
処理時間５９５では、復元したデータをホストコンピュータに転送する。
処理時間５９９では、ホストコンピュータに終了を報告する。
【００３８】
図１１の場合の応答時間は、処理時間５１０，５２０，５３２，５３９，５４０，５９５および５９９の和になる。
処理時間５１０では、ホスト接続手段１１０が、ホストコンピュータ１０からの読み出しの要求を受け取り、その要求を解釈する。
処理時間５２０では、ディスク装置２５１～２５５に対して読み出しを指令する。
処理時間５３１～５３５では、各ディスク装置２５１～２５５で読み出しが行われる。そして、処理時間５３１，５３３，５３４および５３５では、読み出しが成功し、読み出された内容が記憶手段１５０に転送される。一方、処理時間５３２では読み出しが失敗し、
処理時間５３９でセクタ障害を認識し、リトライを行う。
処理時間５４０では１回目のリトライを実施し、ディスク処理の最大予定終了時間Ｔｍａｘが経過する前に読み出しに成功する。
処理時間５９５では、データをホストコンピュータに転送する。
処理時間５９９では、ホストコンピュータに終了を報告する。
【００３９】
図１０，図１１のようにディスク処理の最大予定終了時間Ｔｍａｘを設定すると、セクタ障害時でも最大応答時間を保障することが出来る。すなわち、図１０の場合が最大応答時間であり、これ以上に応答時間が伸びることはない。
【００４０】
《第２の実施形態》
第２の実施形態のシステム構成は、図１のシステム構成と同様である。但し、第２の実施形態では、ディスク装置２５１～２５５として、図１２に示すディスク装置を使用する。
【００４１】
●ディスク装置
図１２は、第２の実施形態で使用するディスク装置の構成図である。
このディスク装置は、図２のディスク装置の構成に加えて、ダミーデータを生成するダミーデータ生成手段４００と、ダミーデータを生成した論理ブロック番号（またはヘッド番号とトラック番号とセクタ番号）を保持するエラー保持手段４１０と、交代セクタを使用するか否かを切り替える交代セクタ処理切り替え手段４２０を設けた構成になっている。
【００４２】
●ダミーデータ生成手段４００とエラー保持手段４１０
ディスク装置は、転送手段３５６を介して、ディスクアレイ装置１００からの読み出し要求を受け取る。ＳＣＳＩの場合、読み出し要求には、読み出し対象の論理ブロック番号が与えられている。コントローラ３９０は、与えられた論理ブロック番号をヘッド番号とトラック番号とセクタ番号とに変換し、ヘッド制御手段３１５を使用してヘッド３１０を目的のトラックに移動させ、ヘッド３１０の下に目的のセクタが来たらデータの読み出しを開始する。ヘッド３１０で読み出したデータとそれに続くＥＣＣは、波形整形手段３５０を介して、ＥＣＣ手段３５２に入力される。
ＥＣＣ手段３５２は、読み出したデータに誤りがないことを確認するとデータをバッファ

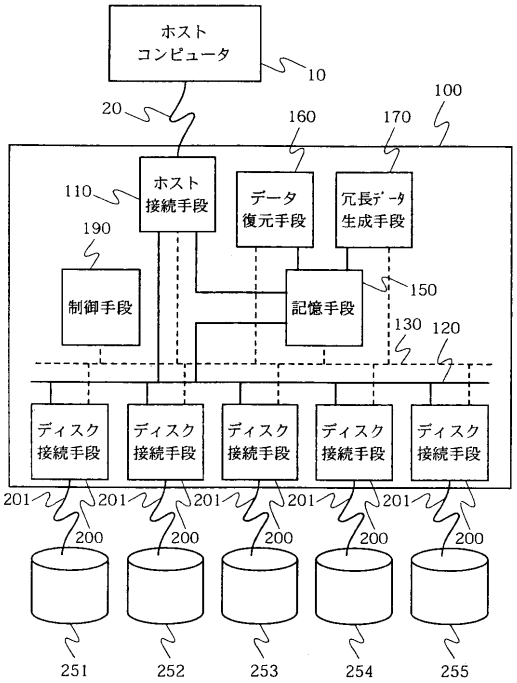
３５４に転送するが、読み出したデータに誤りがあることを検出すると訂正を試みる。訂正できた場合には、正しくなったデータをバッファ３５４に転送する。もし、訂正できなかった場合には、ダミーデータ生成手段４００を起動する。
ダミーデータ生成手段４００は、１セクタ分の“０”のデータをバッファ３５４に転送するとともに、セクタ障害の発生したセクタの論理ブロック番号をエラー保持手段４１０に保持させる。
転送手段３５６は、バッファ３５４内のデータを転送する。従って、ダミーデータ生成手段４００が生成したデータ“０”もディスクアレイ装置１００に転送される。
【００４３】
以上のように、セクタ障害が発生したときにダミーデータを転送するようにすると、データ転送を中断する必要がない。
【００４４】
ディスクアレイ装置１００へのエラーの報告は、ディスクアレイ装置１００に、読み出し終了を報告する前に、エラー保持手段４１０に保持した論理ブロック番号をまとめて報告する。なお、セクタのデータを転送する毎に、そのセクタのデータがダミーデータであるか否かの情報を付加する方法によりエラー報告を行ってもよい。
【００４５】
●交代セクタ処理切り替え手段４２０
交代セクタ処理切り替え手段４２０により、交代セクタを使用するように切り替えると、交代セクタへのヘッド移動や回転待ちが発生する。これに対して、交代セクタをエラーとして処理するように切り替えると、読み出し要求のあったセクタが交代セクタに当たる場合にはダミーデータ生成手段４００を起動し、１セクタ分の“０”のデータをバッファ３５４に転送するとともに、エラー保持手段４１０にダミーデータを転送したセクタの論理ブロック番号を保持する。従って、交代セクタへのヘッド移動や回転待ちが発生しない。
【００４６】
●リード処理
セクタ障害が発生しない場合のリード処理については、第１の実施形態と同様である。
【００４７】
●セクタ障害時のリード処理
ホストコンピュータ１０は、例えばストライプＤ５の読み出しの要求を発行する。
制御手段１９０は、ホスト接続手段１１０を介して、前記要求を受け取る。そして、ストライプＤ５を含むストライプグループすなわちストライプＤ４，Ｄ５，Ｄ６，Ｄ７および冗長データ格納領域Ｐ４－７　の内容を読み出すように、制御バス１３０を介して、５つのディスク接続手段２００に指令する。
各ディスク接続手段２００は、要求された内容を各ディスク装置２５１～２５５から読み出そうとする。ここで、ストライプＤ５の読み出し中にディスク装置２５２でセクタ障害が発生したとする。
ディスク装置２５２は、リトライを実施せず、セクタ障害が発生したセクタのデータとしてダミーデータ“０”を転送する。このため、ディスクアレイ装置１００は、セクタ障害の発生を認識せずに処理を続行する。
【００４８】
ディスク装置２５２は、ストライプＤ５の読み出しを完了すると、エラー保持手段４１０に格納した、ダミーデータを転送した論理ブロック番号をまとめてディスクアレイ装置１００に転送する。この論理ブロック番号を受け取って初めてディスクアレイ装置１００はセクタ障害が発生していたことを認識する。
そこで、ディスクアレイ装置１００の制御手段１９０は、ダミーデータ“０”が転送されたセクタについてのみ、データ復元手段１６０を起動し、ストライプＤ５内の障害セクタのデータを復元する。そして、この復元が終了した時点で、制御手段１９０は、ホスト接続手段１１０を介して、ストライプＤ５のデータを、記憶手段１５０からホストコンピュータ１０に転送する。

【００４９】
以上の処理によれば、障害セクタに対してのみデータの復元を行うから、データ復元に必要な時間を短縮することができる。なお、第１の実施形態では、ストライプ全体に対してデータの復元が必要であり、データ復元に必要な時間が比較的長くなる。
【００５０】
●交代セクタ時のリード処理
図１３に示すように、記録媒体３００のストライプＤ５の中の１セクタが障害セクタであり、交代セクタが当てられている場合を考える。
交代セクタを利用する場合、ストライプＤ５の先頭セクタがヘッド３１０の下に来るのを待って読み出しを開始し、障害セクタに当たると、ヘッド３１０を交代セクタのあるトラック３０１に移動させ、交代セクタがヘッド３１０の下に来るのを待って読み出し、交代セクタの読み出しが終了すると、ストライプＤ５のあるトラック３０１にヘッドを移動して、続きのセクタがヘッド３１０の下に来るのを待って続きの読み出しを行う。従って、交代セクタを含まないストライプの読み出しに比べて応答時間が伸びる。
【００５１】
交代セクタを利用しない場合の動作は次のようになる。
ホストコンピュータ１０は、例えばストライプＤ５の読み出しの要求を発行する。
制御手段１９０は、ホスト接続手段１１０を介して、前記要求を受け取る。そして、ストライプＤ５を含むストライプグループすなわちストライプＤ４，Ｄ５，Ｄ６，Ｄ７および冗長データ格納領域Ｐ４－７　の内容を読み出すように、制御バス１３０を介して、５つのディスク接続手段２００に指令する。
各ディスク接続手段２００は、要求された内容を各ディスク装置２５１～２５５から読み出そうとする。ここで、ストライプＤ５の読み出し中に障害セクタがあったとする。
ディスク装置２５２は、交代セクタを利用せず、ダミーデータ転送手段４００を用いて、そのセクタのデータとしてダミーデータ“０”を転送する。このため、ディスクアレイ装置１００は、セクタ障害の発生を認識せずに処理を続行する。
【００５２】
ディスク装置２５２は、ストライプＤ５の読み出しを完了すると、エラー保持手段４１０に格納した、ダミーデータを転送した論理ブロック番号をまとめてディスクアレイ装置１００に転送する。この論理ブロック番号を受け取って初めてディスクアレイ装置１００は障害セクタがあったことを認識する。
そこで、ディスクアレイ装置１００の制御手段１９０は、ダミーデータ“０”が転送されたセクタについてのみ、データ復元手段１６０を起動し、ストライプＤ５内の障害セクタのデータを復元する。そして、この復元が終了した時点で、制御手段１９０は、ホスト接続手段１１０を介して、ストライプＤ５のデータを、記憶手段１５０からホストコンピュータ１０に転送する。
以上のように、交代セクタを使わずにエラーとして扱うようにすれば、応答時間の悪化を防ぐことが出来る。
【００５３】
【発明の効果】
本発明のディスクアレイ装置によれば、次の効果が得られる。
（１）　セクタ障害が発生するか否かに関わらず、データの復元に必要なデータと冗長データとを読み出しておくことにより、セクタ障害発生時の最大応答時間を低減することが出来る。
（２）　ディスク装置のリトライを使用せず、データ復元手段によりデータを復元することにより、最大応答時間を低減することが出来る。
（３）　交代領域を使用せず、データ復元手段によりデータを復元することにより、交代領域へのシーク時間と回転待ちの時間が入らないため、最大応答時間を低減することが出来る。
【図面の簡単な説明】

【図１】本発明の第１の実施形態にかかるコンピュータシステムの構成図である。
【図２】第１の実施形態に用いるディスク装置の構成図である。
【図３】記録媒体上でのトラックとセクタとセクタ内領域とを示す説明図である。
【図４】データと冗長データの配置を示す図である。
【図５】第１の実施形態における応答時間を示すタイムチャートである。
【図６】従来のディスクアレイ装置における応答時間を示すタイムチャートである。
【図７】リトライを３回で打ち切った場合の応答時間を示すタイムチャートである。
【図８】３回目のリトライで読み取りに成功した場合の応答時間を示すタイムチャートである。
【図９】リトライアボートによる応答時間を示すタイムチャートである。
【図１０】ディスク処理の最大予定終了時間でリトライを打ち切った場合の応答時間を示すタイムチャートである。
【図１１】ディスク処理の最大予定終了時間内のリトライで読み取りに成功した場合の応答時間を示すタイムチャートである。
【図１２】第２の実施形態に用いるディスク装置の構成図である。
【図１３】障害セクタと交代セクタの説明図である。
【図１４】ディスクアレイ装置のデータと冗長データの配置を論理イメージと物理イメージで示す説明図である。
【符号の説明】
１０…ホストコンピュータ，
２０…ＳＣＳＩバス，
１００…ディスクアレイ装置，１１０…ホスト接続手段，１２０…データバス，
１３０…制御バス，１５０…記憶手段，１６０…データ復元手段，１７０…冗長データ生成手段，２００…ディスク接続手段，
２０１…ＳＣＳＩバス，
２５１～２５５…ディスク装置，３００…記録媒体，３０１…トラック，３０２…セクタ，３０３…位置情報領域，３０４…データ領域，３０５…ＥＣＣ領域，
３１０…ヘッド，３１５…ヘッド制御手段，３２０…モータ，３２５…モータ制御手段，
３５０…波形整形手段，３５２…ＥＣＣ手段，３５４…バッファ，３５６…転送手段，３９０…コントローラ，４００…ダミーデータ生成手段，４１０…エラー保持手段，４２０…交代セクタ処理切り替え手段。

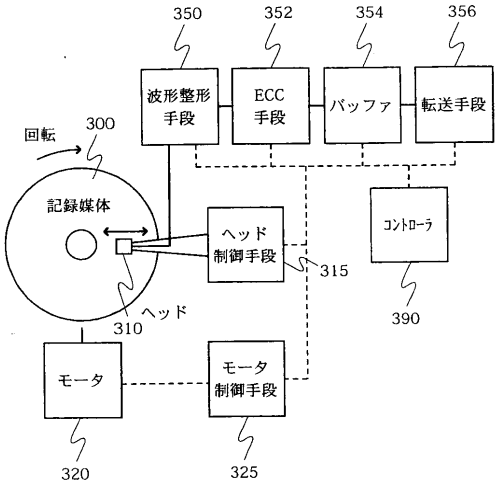
【図１】

図1



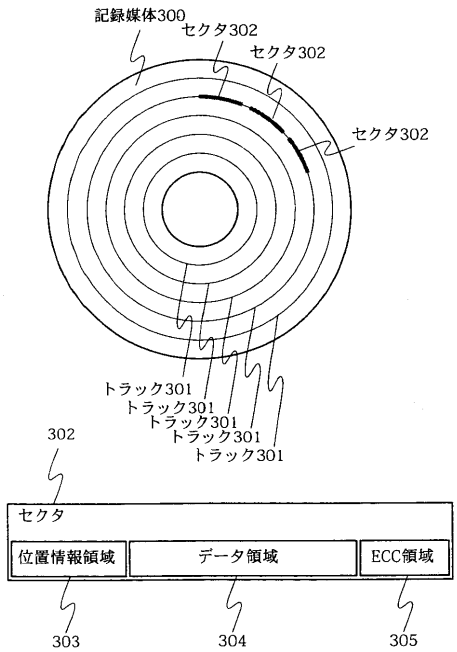
【図２】

図2



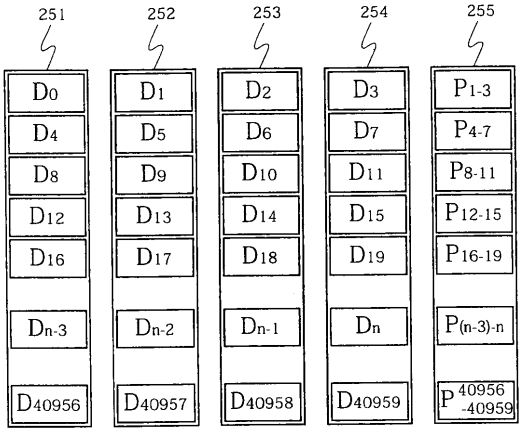
【図３】

図3



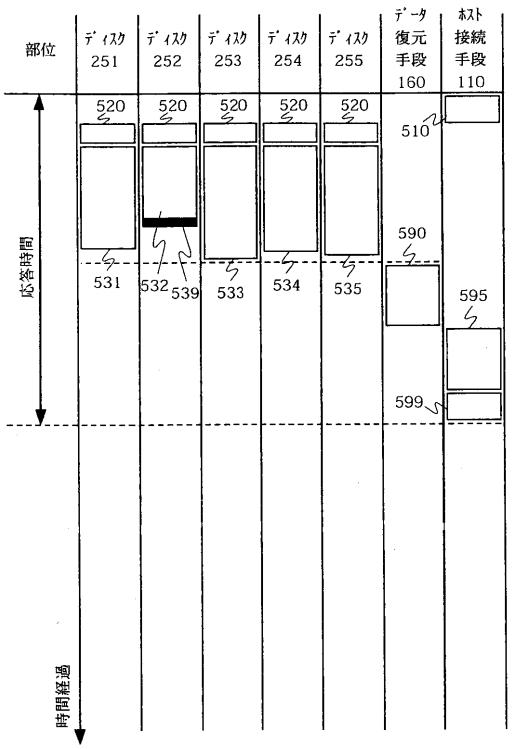
【図４】

図4



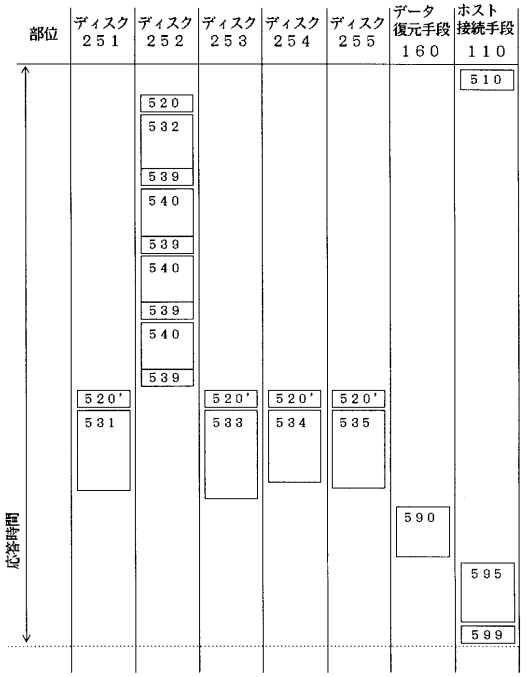
【図５】

図５



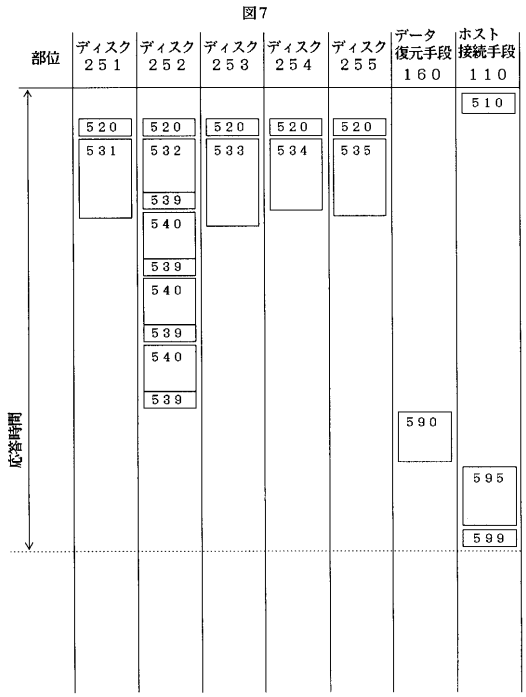
【図６】

図６



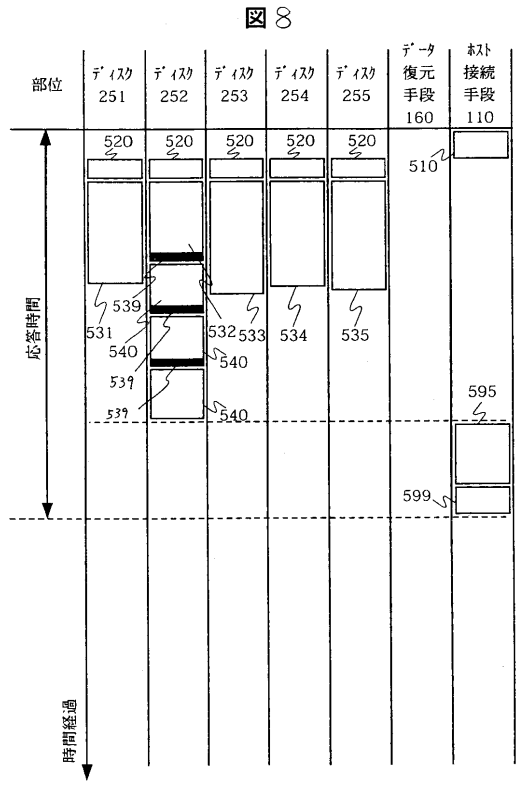
【図７】

図７



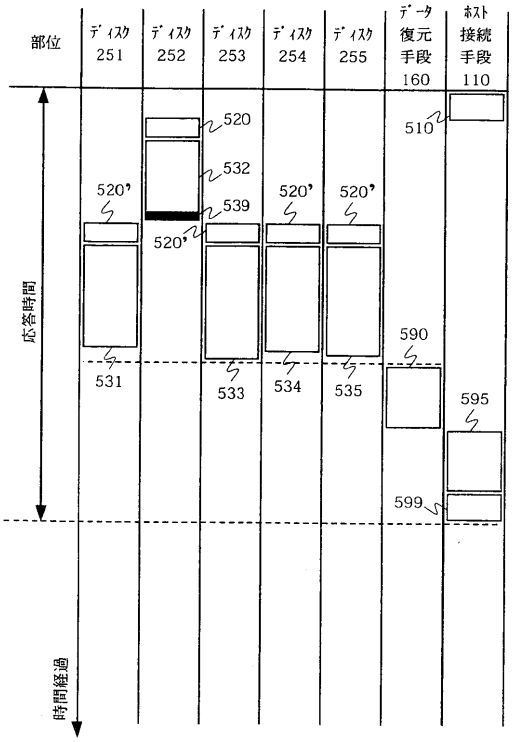
【図８】

図８



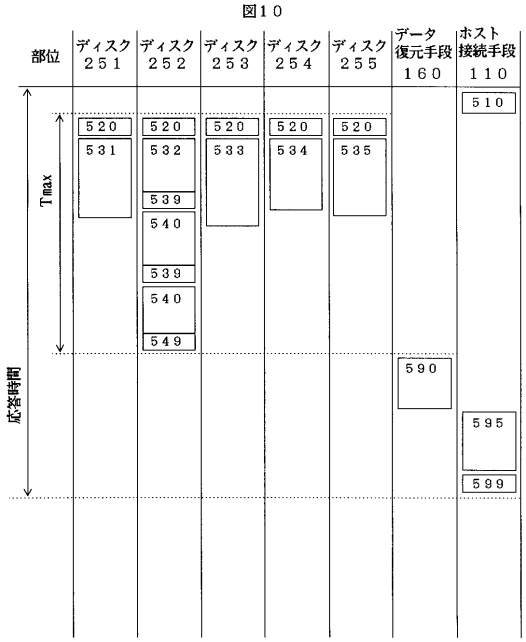
【図９】

図9



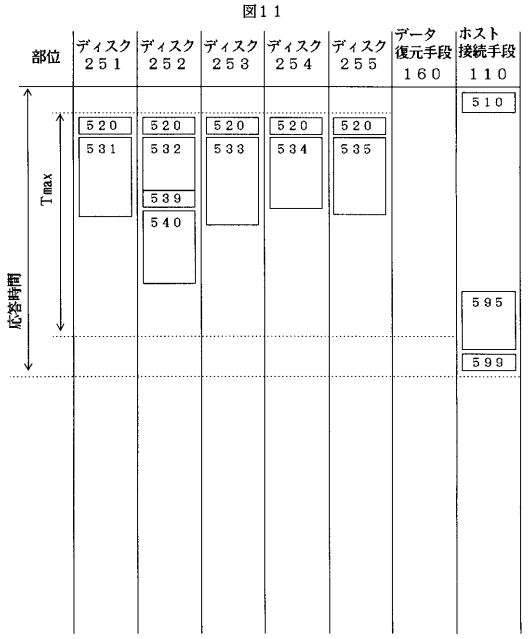
【図１０】

図10



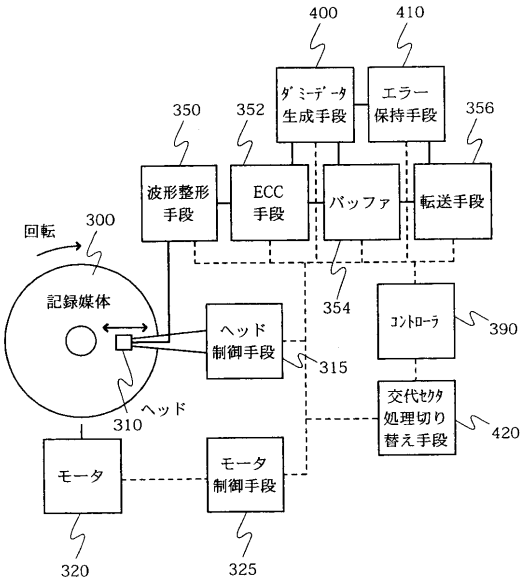
【図１１】

図11



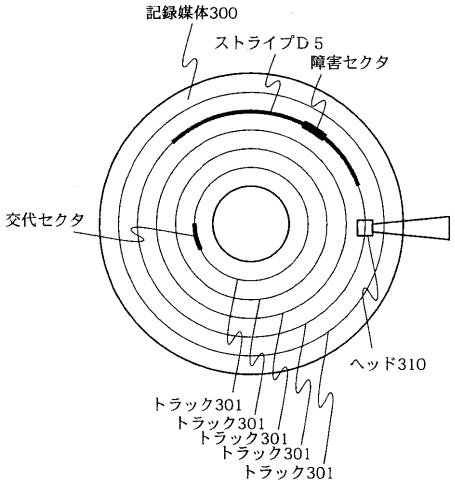
【図１２】

図12



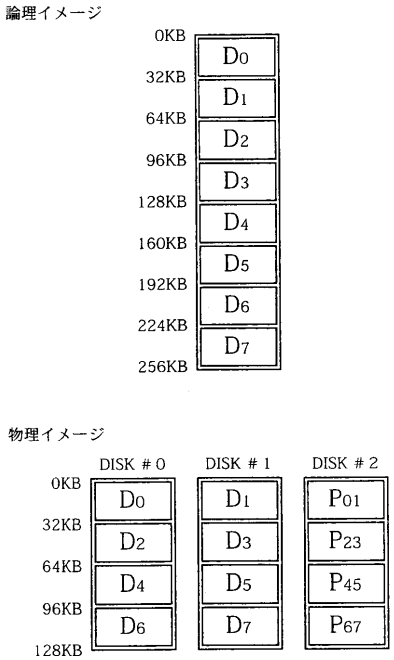
【図１３】

図13



【図１４】

図14



フロントページの続き

(72)発明者　大枝　高
神奈川県川崎市麻生区王禅寺１０９９番地　株式会社日立製作所　システム開発研究所内
(72)発明者　松並　直人
神奈川県川崎市麻生区王禅寺１０９９番地　株式会社日立製作所　システム開発研究所内
(72)発明者　荒川　敬史
神奈川県川崎市麻生区王禅寺１０９９番地　株式会社日立製作所　システム開発研究所内
(72)発明者　村岡　健司
神奈川県小田原市国府津２８８０番地　株式会社日立製作所　ストレージシステム事業部内
(72)発明者　高野　雅弘
神奈川県小田原市国府津２８８０番地　株式会社日立製作所　ストレージシステム事業部内

審査官　藤井　浩

(56)参考文献　特開平０７－２００１９１（ＪＰ，Ａ）
特開平０８－１７１４６３（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.^{7}，ＤＢ名)
G06F　3/06　-　3/08