(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2006-106975
(P2006-106975A)
(43)公開日　平成18年4月20日(2006.4.20)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 3/06	(2006.01)	G06F	3/06	305F	5B065
		G06F	3/06	540	

審査請求 未請求　請求項の数 6　ＯＬ　（全 9 頁）

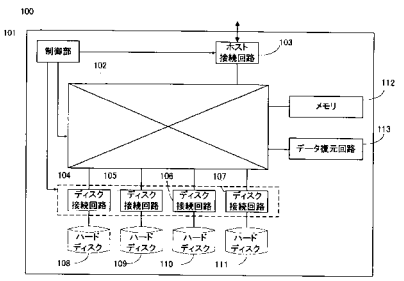
(21)出願番号　特願2004-290501 (P2004-290501)
(22)出願日　平成16年10月1日 (2004.10.1)

(71)出願人　000001007
キヤノン株式会社
東京都大田区下丸子３丁目３０番２号
(74)代理人　100090273
弁理士　國分　孝悦
(72)発明者　外山　猛
東京都大田区下丸子３丁目３０番２号　キヤノン株式会社内
(72)発明者　伊藤　博康
東京都大田区下丸子３丁目３０番２号　キヤノン株式会社内

Ｆターム(参考)　5B065 BA01　CA11　CA30　CH05　EA01
EA15

(54)【発明の名称】ディスクアレイ装置及びその制御方法

(57)【要約】
【課題】　ディスクアレイ装置の読み出し応答時間の短縮を図る。
【解決手段】　ハードディスク装置の所定領域ごとの障害情報を記憶し、ホストからの要求データが記録されている領域が読み出し障害を発生する領域であるか否かを該障害情報に基づいて判断し、その判断結果に基づいて、当該領域からデータを読み出す第１の読み出し方式、又は、他のハードディスク装置から復元用データを読み出して該要求データを復元する第２の読み出し方式のいずれかを選択して実行するようにした。
【選択図】　図１



【特許請求の範囲】
【請求項１】
ハードディスク装置の所定領域ごとの障害情報を記憶した記憶手段と、
上位装置からの第１のデータが記録されている対応領域が読み出し障害を発生する領域であるか否かを前記障害情報に基づいて判断する判断手段と、
前記判断手段の判断結果に基づいて、前記対応領域から前記第１のデータを読み出す第１の読み出し方式、又は、前記要求データを復元するための第２のデータを読み出した後、該第２のデータから前記第１のデータを復元する第２の読み出し方式のいずれかを選択して実行する制御手段とを有することを特徴とするディスクアレイ装置。
【請求項２】
前記記憶手段に記憶された障害情報は、読み出しの実行結果に応じて更新されることを特徴とする請求項１に記載のディスクアレイ装置。
【請求項３】
前記ハードディスク装置は、自ら読み出しの再試行を実行しないように設定されていることを特徴とする請求項１又は２に記載のディスクアレイ装置。
【請求項４】
ハードディスク装置の所定領域ごとの障害情報を記憶し、
上位装置からの第１のデータが記録されている対応領域が読み出し障害を発生する領域であるか否かを前記障害情報に基づいて判断し、
該判断結果に基づいて、前記対応領域から前記第１のデータを読み出す第１の読み出し方式、又は、前記要求データを復元するための第２のデータを読み出した後、該第２のデータから前記第１のデータを復元する第２の読み出し方式のいずれかを選択して実行すること特徴とするディスクアレイ装置の制御方法。
【請求項５】
ハードディスク装置の所定領域ごとの障害情報を記憶し、
上位装置からの第１のデータが記録されている対応領域が読み出し障害を発生する領域であるか否かを前記障害情報に基づいて判断し、
該判断結果に基づいて、前記対応領域から前記第１のデータを読み出す第１の読み出し方式、又は、前記要求データを復元するための第２のデータを読み出した後、該第２のデータから前記第１のデータを復元する第２の読み出し方式のいずれかを選択して実行する処理をコンピュータにて実行させることを特徴とするコンピュータプログラム。
【請求項６】
請求項５に記載のコンピュータプログラムを記録したことを特徴とするコンピュータ読み取り可能な記録媒体。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、ディスクアレイ装置、その制御方法、コンピュータプログラム、及びコンピュータ読み取り可能な記録媒体に関する。
【背景技術】
【０００２】
近年、複数のハードディスク装置を並列に動作させて読み書き速度を高速化し、冗長構成により信頼性を高めたディスクアレイ装置が普及している。
【０００３】
ディスクアレイ装置には、ＲＡＩＤ（Redundant Arrays of Inexpensive Disks）と呼ばれる技術が採用されている。ＲＡＩＤには、用途に応じて様々な構成が存在する。図５にＲＡＩＤ５の構成例を示す。５０１～５０４は、それぞれ別個のハードディスク装置である。Ｄ０～Ｄ８は、それぞれストライプと呼ばれる所定量のデータ格納領域であり、ディスクアレイ装置におけるデータ管理制御の単位となっている。
【０００４】

ディスクアレイ装置を介してハードディスク装置をアクセスする装置（以下、「ホスト」と称する）は、各ストライプデータがいずれのハードディスク装置に記録されているかの情報は知る必要はなく、ホストからの論理アドレスを各ハードディスク装置上のアドレスに変換する作業はディスクアレイ装置が担う。

【０００５】

Ｐ０２、Ｐ３５、Ｐ６８は、それぞれ冗長データの格納領域である。Ｐ０２にはストライプＤ０～Ｄ２内のデータの排他的論理和が格納されている。Ｐ３５には、ストライプＤ３～Ｄ５内のデータの排他的論理和が格納されている。Ｐ６８には、ストライプＤ６～Ｄ８内のデータの排他的論理和が格納されている。

【０００６】

このようにＲＡＩＤ５では、いずれかのハードディスク装置が壊れた場合でも、残りのハードディスク装置から壊れたハードディスク装置のデータを復元することが可能な構成となっている。

【０００７】

一方で、ハードディスク装置に対しては、信頼性とともに、読み出し速度向上の要求が年々高まっており、この要求に答えるべく各種の改良案が提案されている。例えば、特許文献１には、ストライプデータを読み出す際、当該ストライプデータの復元に必要なデータ（以下、復元用データと称する）も並列に読み出しておき、前記ストライプデータの読み出しエラーが発生した場合にはリトライをすることなく前記他のデータにより前記ストライプデータを復元する技術が提案されている。

【０００８】

【特許文献１】特開平１０－２４７１３３号公報

【発明の開示】

【発明が解決しようとする課題】

【０００９】

上記従来例は、ストライプデータの読み出しと当該ストライプデータの復元用データの読み出しとを並行して行うことによりディスクアレイ装置の読み出し応答時間を短縮しようとしている。

【００１０】

しかしながら、並列に読み出されたデータは、共にディスクアレイ装置内の共有のデータ経路（例えば、上記公報における図１の１０３）を流れることになるので、読み出しエラーが発生しない場合の読み出し応答時間はむしろ長くなってしまうという問題がある。

【００１１】

そのため、読み出しエラーの発生率が低い場合には、読み出しのスループットはむしろ悪くなってしまう。

【００１２】

本発明は上記のような点に鑑みてなされたものであり、読み出し障害の発生により読み出し応答時間が長くなってしまうのを防止することを目的とする。

【課題を解決するための手段】

【００１３】

本発明のディスクアレイ装置は、ハードディスク装置の所定領域ごとの障害情報を記憶した記憶手段と、上位装置からの第１のデータが記録されている対応領域が読み出し障害を発生する領域であるか否かを前記障害情報に基づいて判断する判断手段と、前記判断手段の判断結果に基づいて、前記対応領域から前記第１のデータを読み出す第１の読み出し方式、又は、前記要求データを復元するための第２のデータを読み出した後、該第２のデータから前記第１のデータを復元する第２の読み出し方式のいずれかを選択して実行する制御手段とを有する点に特徴を有する。

本発明のディスクアレイ装置の制御方法は、ハードディスク装置の所定領域ごとの障害情報を記憶し、上位装置からの第１のデータが記録されている対応領域が読み出し障害を発生する領域であるか否かを前記障害情報に基づいて判断し、該判断結果に基づいて、前

記対応領域から前記第１のデータを読み出す第１の読み出し方式、又は、前記要求データを復元するための第２のデータを読み出した後、該第２のデータから前記第１のデータを復元する第２の読み出し方式のいずれかを選択して実行する点に特徴を有する。

本発明のコンピュータプログラムは、ハードディスク装置の所定領域ごとの障害情報を記憶し、上位装置からの第１のデータが記録されている対応領域が読み出し障害を発生する領域であるか否かを前記障害情報に基づいて判断し、該判断結果に基づいて、前記対応領域から前記第１のデータを読み出す第１の読み出し方式、又は、前記要求データを復元するための第２のデータを読み出した後、該第２のデータから前記第１のデータを復元する第２の読み出し方式のいずれかを選択して実行する処理をコンピュータにて実行させる点に特徴を有する。

本発明のコンピュータ読み取り可能な記録媒体は、上記本発明のコンピュータプログラムを記録した点に特徴を有する。

【発明の効果】

【００１４】

本発明によれば、読み出し障害の発生により読み出し応答時間が長くなってしまうのを防止することができる。それと同時に、従来例に対して、ディスクアレイ装置内のデータ経路の帯域に必要以上の負荷がかからないので、その空いている帯域を他の並列動作に割り当てることが可能となる。

【発明を実施するための最良の形態】

【００１５】

以下、添付図面を参照して、本発明の好適な実施形態について説明する。図１は、本実施形態におけるディスクアレイ装置の構成を示した図である。１０１は、装置全体の制御を司る制御部であり、ＣＰＵ、制御プログラムを格納するＲＯＭ、制御に必要なワークメモリを格納するＲＡＭ等から構成される。

【００１６】

１０２は、スイッチ回路であり、制御部１０１による設定により、ホスト接続回路１０３、ディスク接続回路１０４～１０７、メモリ１１２、データ復元回路の各ユニット間のデータ経路を形成する。

【００１７】

１０３は、不図示のホストと接続するためのホスト接続回路であり、例えば、PCI Expressのような高速シリアルインターフェースによりホストとの間のインターフェース制御を行う。

【００１８】

１０４～１０７はディスク接続回路であり、例えば、Serial ATAのような高速シリアルインターフェースによりハードディスク装置１０８～１１１との間のインターフェース制御を行う。

【００１９】

１０８～１１１は、ハードディスク装置である。制御部１０１はハードディスク装置１０８～１１１によりＲＡＩＤ５のディスクアレイを構成する。

【００２０】

１１２は、メモリであり、不図示のメモリコントローラによる制御によりメモリへのデータの書き込み、及び、メモリからのデータの読み出しが行われる。

【００２１】

１１３は、データ復元回路であり、復元用データを用いてストライプデータの復元を行う。データ復元回路１１３は内部に所定サイズのメモリを有しており、該メモリ内のデータと入力されたデータとの排他的論理和を計算した結果を該メモリに出力する。

【００２２】

次に、図２を参照して、本実施形態におけるハードディスク装置からのデータの読み出し動作を説明する。図２において、メモリ１１２、ハードディスク装置１０８～１１１、及びデータ復元回路１１３は、図１と同じ符号を用いている。

【００２３】
まず、ハードディスク装置１０８内の領域２０１のデータを読み出す動作（以下、「直接読み出し」と称する）を説明する。制御部１０１はスイッチ回路１０２を制御して、ディスク接続回路１０４とメモリ１１２との間のデータ経路を形成し、領域２０１のデータをメモリ１１２内の領域２０７に転送する。
【００２４】
その後、制御部１０１はスイッチ回路１０２を制御して、メモリ１１２とホスト接続回路１０３との間のデータ経路を形成し、領域２０２のデータをホストへ転送する。
【００２５】
次に、領域２０１のデータの復元用データ、すなわち、領域２０３、領域２０４、及び領域２０５のデータを読み出して領域２０１のデータを復元する動作（以下、「復元読み出し」と称する）を説明する。
【００２６】
制御部１０１はスイッチ回路１０２を制御してハードディスク装置１０９とデータ復元回路１１３との間のデータ経路を形成し、領域２０３のデータをデータ復元回路１１３内の領域２０６に転送する。
【００２７】
この転送処理の際、領域２０３からのデータは領域２０６内のデータとの排他的論理和をとられてから領域２０６に記録される。
【００２８】
領域２０６内は予めゼロに初期化されているので、この転送処理により領域２０６には領域２０３のデータと同じデータが記録される。
【００２９】
次に、制御部１０１は、スイッチ回路１０２を制御してハードディスク装置１１０とデータ復元回路１１３内の領域２０６に転送する。
【００３０】
この転送処理の際、領域２０４からのデータは領域２０６内のデータと排他的論理和をとられた後、領域２０６に記録される。
【００３１】
次に、制御部１０１は、スイッチ回路１０２を制御してハードディスク装置１１１とデータ復元回路１１３内の領域２０６に転送する。
【００３２】
この転送処理の際、領域２０５からのデータは領域２０６内のデータと排他的論理和を取られた後、領域２０６に記録される。
【００３３】
領域２０６には領域２０１のデータが復元されているので、制御部１０１は領域２０６のデータを領域２０７に転送する。
【００３４】
その後、制御部１０１はスイッチ回路１０２を制御して、メモリ１１２とホスト接続回路１０３との間のデータ経路を形成し、領域２０７のデータをホストへ転送する。
【００３５】
このように、ディスクアレイ装置１００は、ホストからの読み出し要求に対して、要求データが記録されているハードディスク装置から読み出す直接読み出し方式と、該要求データの復元用データを他のハードディスク装置から読み出した後、要求データを復元する復元読み出し方式とを有する。
【００３６】
いずれの方式を選択するかは、ホストからの要求データの記録されている領域が読み出し障害を発生する領域であるか否かの判断結果に基づく。すなわち、ホストからの要求データが記録されている領域が読み出し障害を発生する領域でないと判断した場合は、直接読み出し方式を選択する。ホストからの要求データが記録されている領域が読み出し障害

を発生する領域であると判断した場合は、復元読み出し方式を選択する。ホストからの要求データの記録されている領域が読み出し障害を発生する領域であるか否かの判断は、当該領域において過去読み出し障害が発生したか否かの情報を記録した障害領域情報を参照して行われる。

【００３７】

次に、図３を参照して、本実施形態における障害領域情報のフォーマットを説明する。３０１～３０４は、ディスクアレイ装置１００内の不図示の不揮発性メモリ内にハードディスク装置１０８～１１１ごとに設けられた障害領域情報である。

【００３８】

ディスクアレイ装置１００では、各ハードディスク装置のデータ領域を所定サイズのデータブロックごとに管理している。例えば、３０５に示すように各データブロックが障害領域であるか否かをビット情報で表したビットマップとして管理する。そして、ある領域で、直接読み出し方式で読み出し障害が発生したら、該領域を含むデータブロックのビットを１にセットする。以後、該データブロックは読み出し障害の発生する領域として判断される。

【００３９】

次に、図４のフローチャートを参照して、本実施形態におけるディスクアレイ装置の読み出し動作を説明する。ハードディスク装置１０８～１１１は、セクタの読み出しがエラーした場合に自らリトライを実行する機能を備えている。ディスクアレイ装置１００は、ハードディスク装置１０８～１１１の各々について、この自らリトライを実行する機能が動作しないモードに予め設定しておく。これにより、ディスクアレイ装置１００がハードディスク装置からデータの読み出しを開始してから読み取り障害の有無を判断するまでの時間を短縮することができる。

【００４０】

まず、ステップＳ４０１では、ホストからの読み出し要求を、ホスト接続回路１０３を介して受信する。

【００４１】

ステップＳ４０２では、要求データがいずれのハードディスク装置に記録されているかを調べた後、該要求データが記録された領域が読み出し障害の発生する領域であるか否かを障害領域情報３０１～３０４を参照して判断する。

【００４２】

例えば、要求データがハードディスク装置１０８に記録されている場合は、障害領域情報３０１を参照して判断がなされる。

【００４３】

以下の説明では、要求データがハードディスク装置１０８に記録されている場合の例を適宜挙げて補足する。

【００４４】

ステップＳ４０２の判断が肯定判断の場合はステップＳ４０３に進み、直接読み出し方式により要求データがハードディスク装置１０８から読み出される。読み出されたデータは、メモリ１１２上に展開される。

【００４５】

ステップＳ４０４では、ステップＳ４０３において読み出し障害が発生したか否かを判断する。この判断が否定判断の場合はステップＳ４０５に進み、読み出したデータをホストへ転送する。また、ステップＳ４０４の判断が否定判断の場合はステップＳ４０６に進み、ホストからの要求データが記憶されている領域が読み出し障害を発生する領域であることを示すべく障害領域情報３０１を変更する。

【００４６】

ステップＳ４０７の判断では、再読み出しを実行すべきか否かを判断する。再読み出しを実行すべきか否かは、予めディスクアレイ装置１００において設定されているものとする。ステップＳ４０７の判断が否定判断であれがステップＳ４０８に進み、ホストへエラ

ーを報告する。ステップＳ４０７の判断が肯定判断の場合、及び、ステップＳ４０２の判断が肯定判断の場合は、ステップＳ４０９に進む。

【００４７】

ステップＳ４０７では、復元読み出し方式によりデータの読み出しが実行される。すなわち、ホストからの要求データの復元用データをハードディスク装置１０９～１１１から読み出し、該要求データを復元する。復元されたデータは、メモリ１１２上に展開され、その後、ステップＳ４０５でホストに転送される。

【００４８】

このように、本実施形態によれば、ハードディスク装置の所定領域ごとに障害の有無を示す障害情報を記憶し、ホストからの要求データが記録されている領域が読み出し障害を発生する領域であるか否かを該障害情報に基づいて判断し、その判断結果に基づいて、当該領域からデータを読み出す直接読み出し方式、又は、他のハードディスク装置から復元用データを読み出して該要求データを復元する復元読み出し方式のいずれかを選択して実行するようにした。

【００４９】

これにより、読み出し障害の発生により、読み出し応答時間が長くなってしまうのを防止することができる。それと同時に、直接読み出しと復元用読み出しのいずれか一方のみを実行するので、すなわち、直接読み出しと復元読み出しを同時並列で実行しないので、ディスクアレイ装置内のデータ経路の帯域に空きが生じ、その空いている帯域を他の並列動作に割り当てることが可能となる。

【００５０】

なお、本発明の目的は、上述した実施形態の機能を実現するソフトウェアのプログラムコードを記録した記憶媒体を、システム或いは装置に供給し、そのシステム或いは装置のコンピュータ（又はＣＰＵやＭＰＵ）が記憶媒体に格納されたプログラムコードを読み出し実行することによっても、達成されることは言うまでもない。

【００５１】

この場合、記憶媒体から読み出されたプログラムコード自体が上述した実施形態の機能を実現することになり、プログラムコード自体及びそのプログラムコードを記憶した記憶媒体は本発明を構成することになる。

【００５２】

プログラムコードを供給するための記憶媒体としては、例えば、フレキシブルディスク、ハードディスク、光ディスク、光磁気ディスク、ＣＤ－ＲＯＭ、ＣＤ－Ｒ、磁気テープ、不揮発性のメモリカード、ＲＯＭ等を用いることができる。

【００５３】

また、コンピュータが読み出したプログラムコードを実行することにより、上述した実施形態の機能が実現されるだけでなく、そのプログラムコードの指示に基づき、コンピュータ上で稼動しているＯＳ(基本システム或いはオペレーティングシステム)などが実際の処理の一部又は全部を行い、その処理によって上述した実施形態の機能が実現される場合も含まれることは言うまでもない。

【００５４】

さらに、記憶媒体から読み出されたプログラムコードが、コンピュータに挿入された機能拡張ボードやコンピュータに接続された機能拡張ユニットに備わるメモリに書き込まれた後、そのプログラムコードの指示に基づき、その機能拡張ボードや機能拡張ユニットに備わるＣＰＵ等が実際の処理の一部又は全部を行い、その処理によって上述した実施形態の機能が実現される場合も含まれることは言うまでもない。

【図面の簡単な説明】

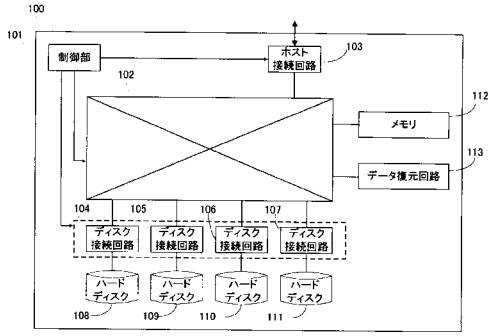
【００５５】

【図１】本実施形態におけるディスクアレイ装置の構成を示した図である。

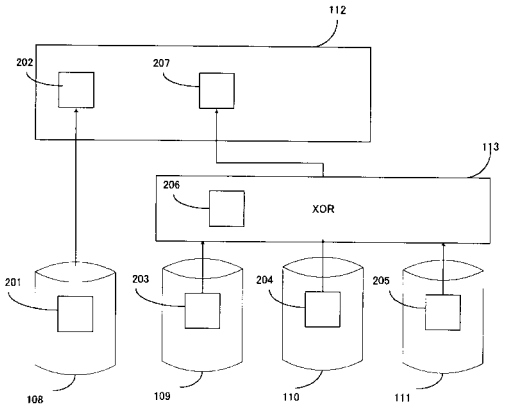
【図２】本実施形態におけるハードディスク装置からのデータの読み出し動作を説明する図である。

【図３】本実施形態における障害領域情報のフォーマットを説明する図である。
【図４】本実施形態におけるディスクアレイ装置の読み出し動作のフローチャートである。
【図５】ＲＡＩＤ５の動作を説明する図である。
【符号の説明】
【００５６】
１００　ディスクアレイ装置
１０１　制御部
１０２　スイッチ回路
１０３　ホスト接続回路
１０４～１０７　ディスク接続回路
１０８～１１１　ハードディスク装置
１１２　メモリ
１１３　データ復元回路

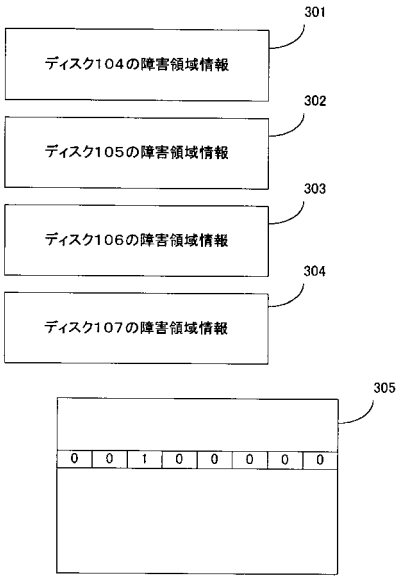
【図１】



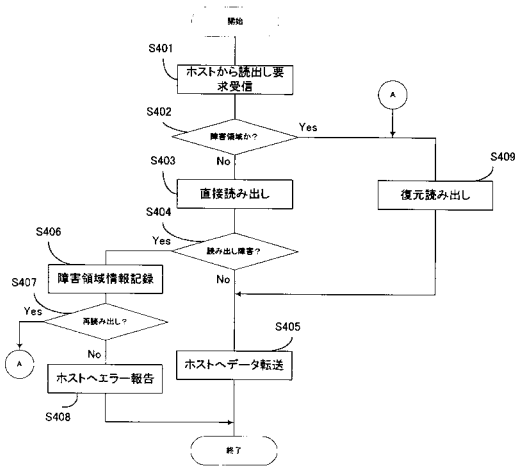
【図２】



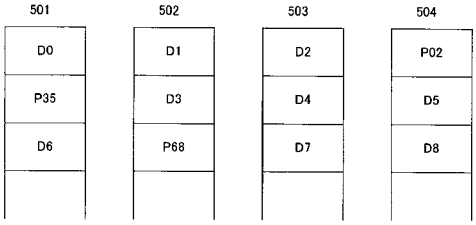
【図３】



【図４】


開始
S401
ホストから読出し要求受信
S402
障害領域か？
Yes
No
A
S403
直接読み出し
S409
復元読み出し
S404
読み出し障害？
Yes
No
S406
障害領域情報記録
S407
再読み出し？
Yes
No
A
S405
ホストへデータ転送
ホストへエラー報告
S408
終了

【図５】


501
502
503
504
D0
D1
D2
P02
P35
D3
D4
D5
D6
P68
D7
D8