(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第3597550号
(P3597550)
(45)発行日　平成16年12月8日(2004.12.8)　(24)登録日　平成16年9月17日(2004.9.17)

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

請求項の数 7　(全 9 頁)

(21)出願番号　特願平5-310612
(22)出願日　平成5年12月10日(1993.12.10)
(65)公開番号　特開平7-160436
(43)公開日　平成7年6月23日(1995.6.23)
審査請求日　平成12年9月4日(2000.9.4)

(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　100084032
弁理士　三品　岩男
(72)発明者　田中　幸一
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　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

審査官　馬場　慎

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】
複数のディスク装置を備え、当該複数のディスク装置のうち少なくとも１つは、他のディスク装置に書き込まれた格納データを読み出すことができないときに、読み出すことができない当該データを再現するための冗長データを書き込む冗長データ領域を有するディスクアレイ装置において、
上記ディスク装置は、格納データを最終的に書き込むための格納データ領域を有し、
上記ディスク装置のうちの少なくとも１つは、格納データを一時的に書き込むためのテンポラリ領域を有し、
上記ディスクアレイ装置は、
外部から受付けた、上記ディスク装置に格納すべき格納データを、冗長データを作成することなく、一旦、上記テンポラリ領域に書き込み、当該テンポラリ領域への書き込みが完了した時点で、外部へ書き込み完了を報告するテンポラリ領域書き込み手段と、
上記書き込み完了の報告後、テンポラリ領域上の上記格納データに基づいて、冗長データを生成する生成手段と、
当該冗長データを上記冗長データ領域に格納し、当該格納データを、上記テンポラリ領域書き込み手段により書き込んだテンポラリ領域を有するディスク装置とは異なるディスク装置にある格納データ領域に書き込むデータ領域書き込み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】

請求項１記載のディスクアレイ装置において、
外部から格納すべき格納データを受付けたときに、テンポラリ領域を有するディスク装置であって、かつ読み出しもしくは書き込み動作を行っていないディスク装置がある場合、当該ディスク装置が有するテンポラリ領域に、当該格納データを書き込む手段を有することを特徴とするディスクアレイ装置。
【請求項３】
請求項１記載のディスクアレイ装置において、
テンポラリ領域を有するディスクを複数有し、
外部から格納すべき格納データを受付けたときに、複数の上記ディスク装置を選択し、当該格納データを書き込むテンポラリ領域を、当該ディスク装置にあるテンポラリ領域とし、複数のテンポラリ領域の各々に同一の格納データを多重に書き込む手段と、
上記テンポラリ領域から上記格納データを読み出す際に、一方のディスク装置から上記格納データを読み出すことができないときは、他方のディスク装置から読み出す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４】
請求項３記載のディスクアレイ装置において、
一方のディスク装置から上記格納データを読み出すことができないときとは、当該ディスク装置が故障しているときであることを特徴とするディスクアレイ装置。
【請求項５】
請求項３記載のディスクアレイ装置において、
一方のディスク装置から上記格納データを読み出すことができないときとは、当該ディスク装置が、読み出しもしくは書き込み動作を行っているときであることを特徴とするディスクアレイ装置。
【請求項６】
請求項１記載のディスクアレイ装置において、
あらかじめ定められたデータ長よりも、上記格納データのデータ長が大きい場合には、上記テンポラリ領域に格納しないで、当該格納データに基づいて、冗長データを生成する生成手段と、
当該冗長データを上記冗長データ領域に格納し、当該格納データを上記格納データ領域に書き込む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７】
請求項１記載のディスクアレイ装置において、
上記テンポラリ領域から上記格納データを読みだし、上記格納データ領域へ格納データを書き込む際に、上記格納データ領域を有するディスク装置で書き込みエラーが発生したときは、当該ディスク装置を切り離し、当該ディスク装置への書き込みは以後行わない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】
【産業上の利用分野】
本発明は、いわゆるＲＡＩＤ（Ｒｅｄｕｎｄａｎｔ　Ａｒｒａｙ　ｏｆ　Ｉｎｅｘｐｅｎｓｉｖｅ　Ｄｉｓｋｓ）と呼ばれるディスクアレイ装置に係わり、ディスク装置のアクセス処理性能、特に、データ書き込み処理の高速化に関する。
【０００２】
【従来の技術】
コンピュータを構成する要素のうち、駆動機構を有するハードディスク装置（ディスク装置またはドライブとも呼ばれる）は、他の電子部品に比べ故障率が高く、また、データを書き込む手段として用いられるため、その故障が与える影響は、大きい。そのため、信頼性の向上が要望されていた。そのような要望に応えるものとして、特開平２－２３６７１４号公報に記載されているものがある。これは、複数のディスク装置を有し、書き込むべき格納データの書き込み時には、格納データの書き込みに加えて、冗長データを生成し、

書き込む。
冗長データは、１台のディスク装置が故障して、そのディスク装置上の格納データが読めなくなったときに、他の故障していないディスク装置上の格納データと冗長データとから読めなくなったデータが再現できるように生成される。生成された冗長データは、当該冗長データを生成するのに使われた格納データを格納しているディスク装置とは異なる、別のディスク装置へ書き込まれる。
このように構成された、複数のディスク装置からなるものは、ディスクアレイ装置、もしくは、ＲＡＩＤと呼ばれる。これによれば、アレイを構成する任意のディスク装置が故障しても、他のディスク装置の格納データと前記冗長データから該障害ディスク装置上の格納データを修復でき、ハードディスク装置のデータ保全性の向上が図れる。
【０００３】
【発明が解決しようとする課題】
ＲＡＩＤは、データ保全性が向上する半面、格納データの書き込みには、常に上記冗長データの生成と書き込みとを伴い、そのために書き込み処理時間が増大し、書き込み完了の報告を外部にするのが遅れて、性能が低下してしまうという問題点が指摘されていた。
そのため、ディスク装置への書き込みが少ない（読み込み処理が多くの割合を占める）システムにおいてしか、事実上向かないものであった。
この対策として、ディスク制御装置内にキャッシュメモリを用意し、キャッシュメモリへの書き込みが完了すると、ただちにホストコンピュータに書き込み完了を報告し、その後、冗長データを生成し、格納データと冗長データをディスク装置に書き込むシステムもある。
しかし、このシステムは、高価なキャッシュメモリを用いるために、システムが高価になるという問題がある。
本発明の目的は、コストを上昇させることなく、上記書き込み処理時の性能低下（ライトペナルティ）を防いだディスクアレイ装置を提供することにある。
本発明の他の目的は、上記の目的に加えて、ＲＡＩＤの特徴であるデータの保全性も維持できるディスクアレイ装置を提供することにある。
【０００４】
【課題を解決するための手段】
上記目的を達成するために、複数のディスク装置を備え、当該複数のディスク装置のうち少なくとも１つは、他のディスク装置に書き込まれた格納データを読み出すことができないときに、読み出すことができない当該データを再現するための冗長データを書き込む冗長データ領域を有するディスクアレイ装置において、上記ディスク装置のうちの少なくとも１つは、上記格納すべき格納データを書き込む領域が、当該格納データを一時的に書き込むためのテンポラリ領域と、当該テンポラリ領域に書き込まれた当該格納データを最終的に書き込むための格納データ領域とに分割されており、外部から受付けた、上記ディスク装置に書き込むべき格納データは、冗長データを作成することなく、一旦、上記テンポラリ領域に書き込み、当該テンポラリ領域への書き込みが完了した時点で、外部へ書き込み完了を報告するテンポラリ領域書き込み手段と、上記書き込み完了の報告後、テンポラリ領域上の上記格納データに基づいて、冗長データを生成する生成手段と、当該冗長データを上記冗長データ領域に格納し、当該格納データを上記格納データ領域に書き込むデータ領域書き込み手段とを有することとしたものである。
また、外部から格納すべき格納データを受付けたときに、複数の上記ディスク装置を選択し、当該格納データを書き込むテンポラリ領域を、当該ディスク装置にあるテンポラリ領域とし、複数のテンポラリ領域の各々に同一の格納データを多重に書き込む手段と、上記テンポラリ領域から上記格納データを読み出す際に、一方のディスク装置から上記格納データを読み出すことができないときは、他方のディスク装置から読み出す手段とを有することとしたものである。
また、外部から格納すべき格納データを受付けたときに、当該格納データを書き込むテンポラリ領域を、最終的に当該格納データを書き込むべき格納データ領域を有するディスク

装置とは異なるディスク装置にあるテンポラリ領域から選択する手段を有することとしたものである。

【０００５】

【作用】

外部からの格納すべき格納データは、冗長データを作成することなく一旦テンポラリ領域に書き込み、この書き込みが完了した時点で外部（例えば、ホストコンピュータ）へ書き込み完了を報告し、テンポラリ領域上の格納データを、その後読み出して、冗長データを生成し、格納データと、冗長データとをそれぞれのデータ領域に書き込む。これにより、キャッシュメモリを用いずに、外部へのレスポンスが高速化される。

また、複数のディスク装置上のテンポラリ領域に２重書きをし、その後、同一格納データへの読み込み要求を受けた場合、テンポラリ領域から該当格納データを読み込むこととすると、格納データの保全性が高まる。

このときに、テンポラリ領域からの読み出しにおいて、２重書きしているディスク装置のうちで、現在読み出し／書き込み動作を行っていないディスク装置（このディスク装置はドライブビジーになっていないと呼ばれる）を選択することにより、ドライブビジーによる遅延を回避し、高速に書き込み動作を行うことができる。

また、テンポラリ領域から格納データ領域に格納データを書き込む際に、古い格納データとテンポラリ領域にある新しい格納データと冗長データとを読みだして、新たに冗長データを生成する必要がある。当該格納データを書き込むテンポラリ領域を、最終的に当該格納データを書き込むべき格納データ領域を有するディスク装置とは異なるディスク装置にあるテンポラリ領域から選択すると、各ドライブを多重動作できることにより、テンポラリ領域から格納データ領域への書き込みが高速化する。

このときに、選択するテンポラリ領域として、現在読み出し／書き込み動作を行っていないディスク装置にあるものから選択することにより、ドライブビジーによる遅延を回避し、高速に書き込み動作を行える。

【０００６】

【実施例】

以下、本発明に係わるディスクアレイ装置の一実施例について説明する。

図１は、本実施例に係わるディスクアレイ装置の全体構成を示した図である。図中ディスクアレイ装置１０１は、制御回路１０２、上位インタフェース回路１０３、小型ディスクインタフェース回路１０４、磁気ディスク装置１１１より構成される。また制御回路１０２は、マイクロプロセッサ１０５、プログラム用メモリ１０６、データバッファ１０７、冗長情報生成回路１０８、不揮発性メモリ１０９から構成されている。また、不揮発性メモリ１０９は、テンポラリ領域管理テーブル１１０を有する。

図２は、ディスクアレイ装置内の磁気ディスク装置の構成を示した図である。ディスクアレイ装置１０１内の全ての磁気ディスク装置１１１は、ユーザデータを最終的に格納するデータ領域２０１と、ユーザデータを一時的に書き込むテンポラリ領域２０２より構成される。データ領域２０１とテンポラリ領域２０２のサイズは任意である。

図３は、本実施例に係わるソフトウエア構成図である。これは、図１に示す、マイクロプロセッサ１０５と、マイクロプロセッサ１０５が実行するマイクロプログラムが格納されているプログラム用メモリ１０６とにより実行される。ただし、バックグラウンドデータ書き込み処理３０４のうち冗長データの生成は、冗長情報生成回路１０８が行う。スケジュラー３０１によりディスクアレイ装置１０１全体が制御される。スケジュラー３０１によりホストＩ／Ｆ制御部３０２、ホストコマンド実行部３０３、バックグランドデータ書き込み処理部３０４は制御される。また、ホストコマンド実行部３０３は、コマンド判定部３０５を通して、リードコマンド処理３０６、ライトコマンド処理３０７、制御系コマンド処理３０８を利用することによりホストコマンドを実行する。

図４は、本実施例に係わるテンポラリ領域管理テーブル１１０の構成を示した図である。図中、本管理テーブルは、上位装置指定情報４０１、制御部管理情報４０２、テンポラリ領域管理情報４０３、およびテンポラリ領域のうち使われているものと空いているものを

確認するためのネクストポインタ４０４から構成されている。ネクストポインタ４０４には、２種類のチェインがあり、１つは、テンポラリ領域のうち使われているものを結んだチェインであり、もう１つは、空いているものを結んだチェインである。
上位装置指定情報４０１は、論理ＬＢＡ（ＬＢＡ：Ｌｏｇｉｃａｌ Ｂｌｏｃｋ Ａｄｄｒｅｓｓ）４０５、論理ＬＥＮ（ＬＥＮ：ＬＥＮｇｔｈ）４０６から構成される。制御部管理情報４０２は、物理ドライブＮｏ．４０７、物理ＬＢＡ４０８、物理ＬＥＮ４０９から構成される。テンポラリ領域管理情報４０３は、テンポラリ領域として使用される物理ドライブＮｏ．４１０、その磁気ディスク装置内の物理ＬＢＡ４１１、および物理ＬＥＮ４１２から構成される。ネクストポインタ４０４を確認することによりテンポラリ領域の空きを確認することが可能である。
次に上位装置からライトコマンドを受取り、テンポラリ領域に対して２重書きを行う各処理について図５，図７を用いて説明する。
上位装置５０１からライトデータ５０２を受け取ると、制御部５０３は、まずライトデータ５０２のブロック長を確認し（図７、７０１）、しきい値よりブロック長が大きかった場合は、テンポラリ領域に格納せずに、通常のライト処理（７０２）を実施する。つまり上位装置５０１からのライトデータ５０２は、通常のライト領域５０８へ格納し、また、作成されたパリティデータは、通常のパリティライト領域５０９へ書き込む。ライトデータ５０２のサイズが大きいときは、ライトペナルティは小さいから直接、最終的な格納場所に格納することとしたものである。
しきい値よりブロック長が小さい場合は、テンポラリ領域をネクストポインタ４０４（図４）を使用して確保する（７０３）。確保する際には、実際にデータを書く物理的な磁気ディスク装置と重ならないように管理している。異なるディスク装置とすることにより、バックグラウンドで処理するときに、テンポラリ領域にある新データと実際にデータを書く位置にある旧データとが異なるディスク装置であるために、並列に読みだすことができ、冗長データを生成する処理が高速化する。さらに、テンポラリ領域を確保する際は、格納データを書き込むテンポラリ領域を、現に読み出しもしくは書き込み動作を行っていないディスク装置にあるテンポラリ領域から選択するようにする。こうすることにより、ドライブビジーによる遅延を回避し、高速に書き込み動作を行うことができる。
もし、テンポラリ領域に空きが無く確保できなかった場合（７０４）には、通常のライト処理（７０２）を実施する。空きテンポラリ領域が確保できた場合、テンポラリ領域５１１のうちの領域５０６／５０７に対して２重書きを実施する。
次にテンポラリ領域に対して２重書きしているユーザデータを読み込み、パリティデータを生成し、データ領域に書き込む処理について図６，図８を用いて説明する。
制御部６０３は、バックグランド処理で、まずテンポラリ領域管理テーブル６１１よりすでに２重書きされているユーザデータが格納されている領域６０５／６０６のアドレスを確保し（図８、８０１）、今現在リードおよびライト処理を行っていない方のテンポラリ領域から、データをリードする（８０２）。つまり、２重書きされているユーザデータが格納されている領域６０５／６０６のうち、例えばデータが格納されている領域６０６が格納されている磁気ディスク装置がリード処理で使用されている場合、別のデータが格納されている領域６０５をリードすることになる。こうすることにより、ドライブビジーによる遅延を回避し、高速に読み出し動作を行うことができる。
磁気ディスク装置の障害により１つのテンポラリ領域からデータがリードできなかった場合（８０３）には、２重書きしている別のテンポラリ領域からデータをリードする（８０４）。このように、２重書きしていることにより、ディスク装置が故障してもデータを保持することができ、データの保全性が高まる。
テンポラリ領域から読み出されたデータと、変更する必要のあるパリティデータ６０８により、冗長情報生成回路６０４で新しいパリティデータを生成し（８０５）、データは、６０７へ、パリティデータは、６０８へ書き込まれる（８０６）。データ領域へ格納データを書き込む場合に、上記格納データ領域を有するディスク装置で書き込みエラーが発生したときは、当該ディスク装置を自動的に切り離し、このディスク装置への書き込みは以

後行わないこととする。
尚、以上の実施例においては、記憶媒体として磁気ディスク装置を用いたディスクアレイ装置の例を示したが、他の記録媒体、たとえば、光ディスクを用いた場合でも同様に実現できる。
【０００７】
【発明の効果】
本発明によれば、いわゆるＲＡＩＤ４，５におけるライトペナルティによる書き込み処理性能の低下を防ぐことができるようになり、ディスク装置への書き込みが少ない（読み込み処理が多くの割合を占める）システムであるかどうかによらず、ＲＡＩＤ利用によるシステムの性能向上を図ることが可能となる。
また、従来のＯＳ（Ｕｎｉｘ等）と同じインタフェースをとることができ、アプリケーションに対して、テンポラリ領域を意識させないので、既存のアプリケーションを活用することが可能であると言う効果がある。
【図面の簡単な説明】
【図１】本発明の一実施例に係わるディスクアレイ装置のブロック図。
【図２】本発明の一実施例に係わる磁気ディスク装置内の格納領域の説明図。
【図３】本発明の一実施例に係わるソフトウエア構成図。
【図４】本発明の一実施例に係わるテンポラリ領域管理テーブルの説明図。
【図５】本発明の一実施例に係わる書き込み動作の説明図。
【図６】本発明の一実施例に係わるバックグランド時の書き込み動作の説明図。
【図７】本発明の一実施例に係わるテンポラリ領域へのデータ書き込み処理のフロー図。
【図８】本発明の一実施例に係わるテンポラリ領域からデータ領域へのデータ書き込み処理のフロー図。
【符号の説明】
１０１　ディスクアレイ装置
１０２　制御回路
１０３　上位インタフェース回路
１０４　小型ディスクインタフェース回路
１０５　マイクロプロセッサ
１０６　プログラム用メモリ
１０７　データバッファ
１０８　冗長情報生成回路
１０９　不揮発性メモリ
１１０　テンポラリ領域管理テーブル
１１１　磁気ディスク装置
２０１　データ領域
２０２　テンポラリ領域
３０１　スケジュラー
３０２　ホストＩ／Ｆ制御部
３０３　上位装置コマンド実行部
３０４　バックグランドデータ書き込み処理
３０５　コマンド判定部
３０６　リードコマンド処理
３０７　ライトコマンド処理
３０８　制御系コマンド処理

【図１】

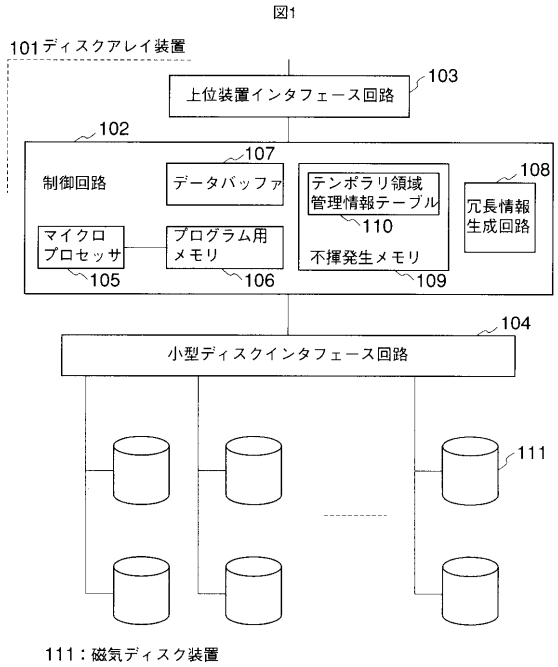


図1
101ディスクアレイ装置
上位装置インタフェース回路
103
102
制御回路
107
データバッファ
テンポラリ領域
管理情報テーブル
110
108
冗長情報
生成回路
マイクロ
プロセッサ
105
プログラム用
メモリ
106
不揮発メモリ
109
104
小型ディスクインタフェース回路
111
111：磁気ディスク装置

【図２】

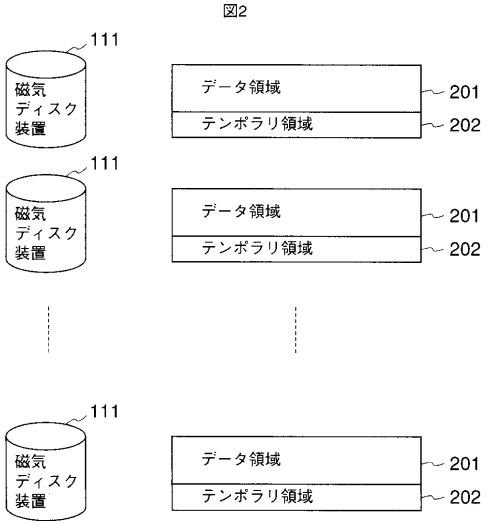


図2
111
磁気
ディスク
装置
データ領域
201
テンポラリ領域
202

【図３】

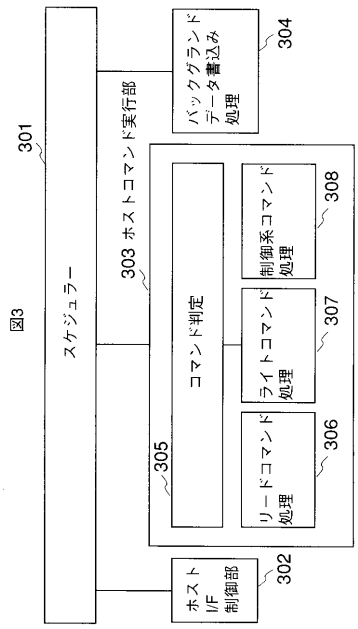


図3
301
スケジューラー
303 ホストコマンド実行部
バックグランド
データ書込み
処理
304
コマンド判定
制御系コマンド
処理
308
ライトコマンド
処理
307
リードコマンド
処理
306
305
ホスト
I/F
制御部
302

【図４】

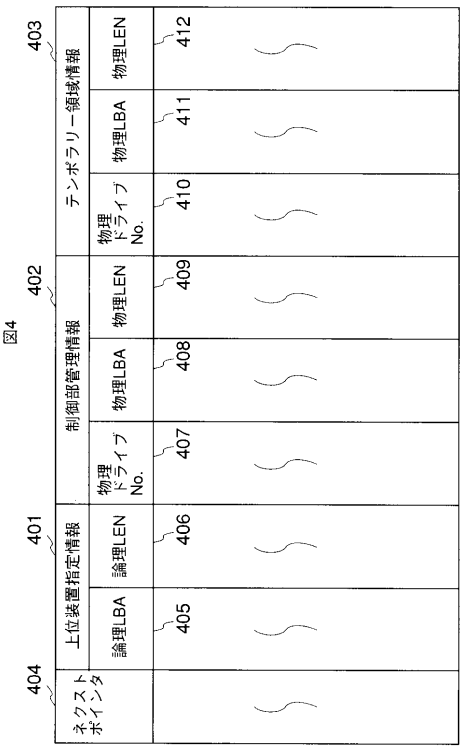
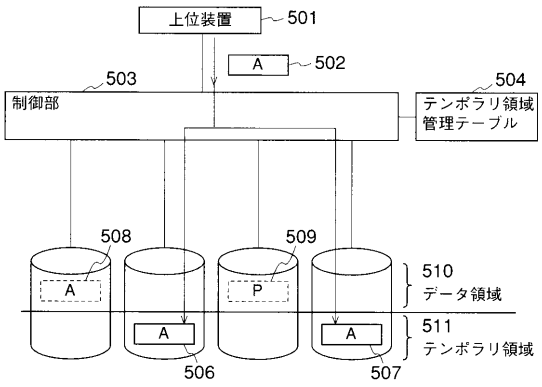


図4
403
テンポラリー領域情報
物理LEN
412
物理LBA
411
物理
ドライブ
No.
410
402
制御部管理情報
物理LEN
409
物理LBA
408
物理
ドライブ
No.
407
401
上位装置指定情報
論理LEN
406
論理LBA
405
404
ネクスト
ポインタ

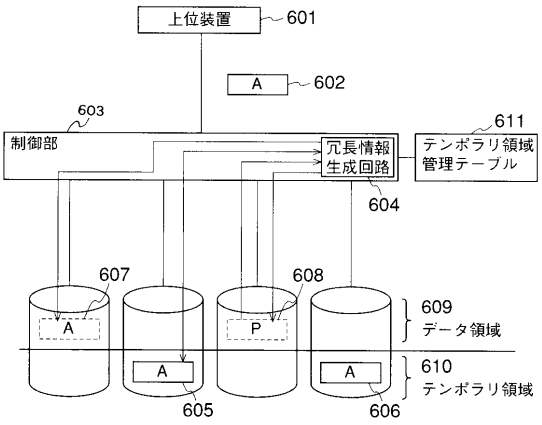
【図５】

図5



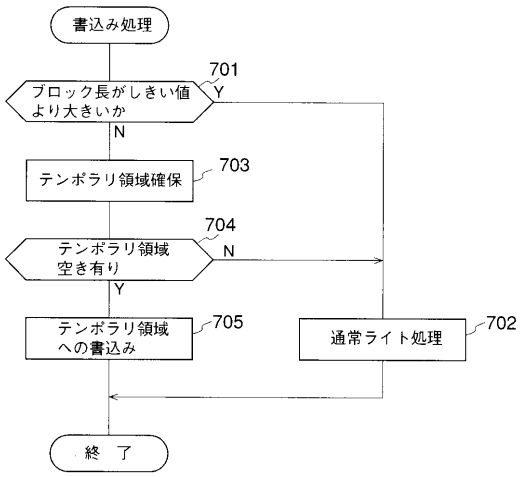
【図６】

図6



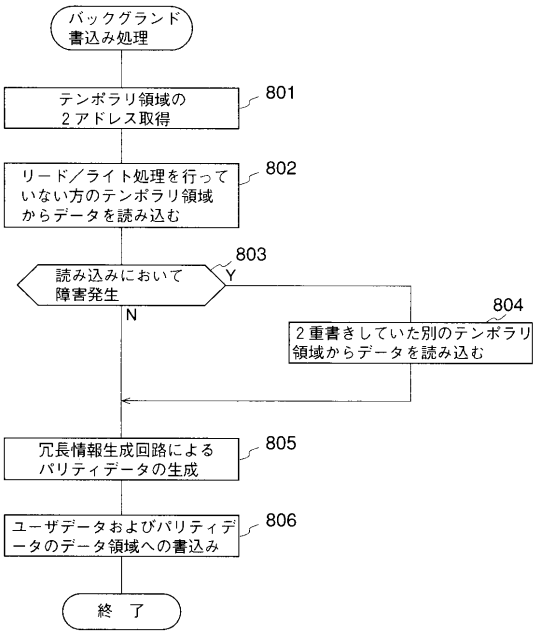
【図７】

図7



【図８】

図8



フロントページの続き

(56)参考文献 特開平０６－３３２６２３（ＪＰ，Ａ）
特開平０６－３３２６３２（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.[7]，ＤＢ名)
G06F 3/06