

(19) 世界知的所有権機関
国際事務局(43) 国際公開日
2009 年 1 月 15 日 (15.01.2009)

PCT

(10) 国際公開番号
WO 2009/008084 A1

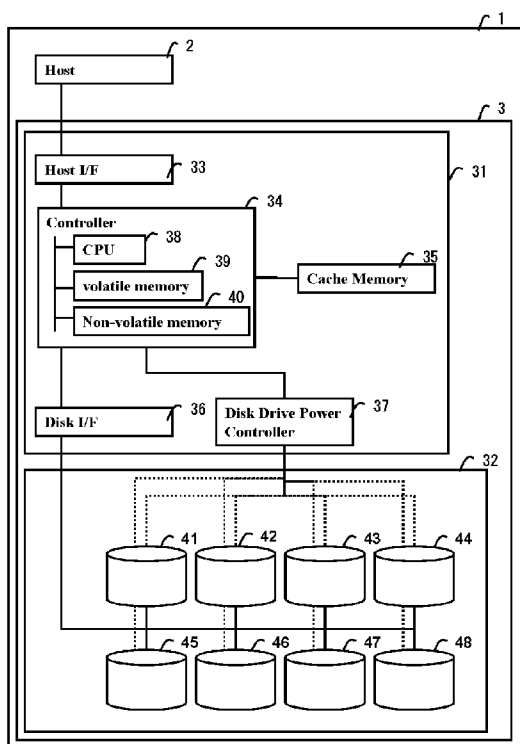
- (51) 国際特許分類:
G06F 3/06 (2006.01)
- (21) 国際出願番号: PCT/JP2007/063901
- (22) 国際出願日: 2007 年 7 月 12 日 (12.07.2007)
- (25) 国際出願の言語: 日本語
- (26) 国際公開の言語: 日本語
- (71) 出願人 (米国を除く全ての指定国について): 富士通株式会社 (FUJITSU LIMITED) [JP/JP]; 〒2118588 神奈川県川崎市中原区上小田中 4 丁目 1 番 1 号 Kanagawa (JP).
- (72) 発明者; および
- (75) 発明者/出願人 (米国についてのみ): 久保田 聡 (KUBOTA, Satoshi) [JP/JP]; 〒2118588 神奈川県川崎市中原区上小田中 4 丁目 1 番 1 号 富士通株式会社内 Kanagawa (JP). 天野 和彦 (AMANO, Kazuhiko) [JP/JP]; 〒2118588 神奈川県川崎市中原区上小田中 4 丁目 1 番 1 号 富士通株式会社内 Kanagawa (JP).
- (74) 代理人: 横山 淳一 (YOKOYAMA, Junichi); 〒2118588 神奈川県川崎市中原区上小田中 4 丁目 1 番 1 号 富士通株式会社内 Kanagawa (JP).
- (81) 指定国 (表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.
- (84) 指定国 (表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), ヨーロッパ (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE,

[続葉有]

(54) Title: DISK ARRAY DEVICE, CONTROL METHOD AND CONTROL PROGRAM

(54) 発明の名称: ディスクアレイ装置、制御方法、および制御プログラム

[図1]



(57) Abstract: [PROBLEMS] To provide a disk array device capable of reducing power consumption in a system having many data read instructions from a host device. [MEANS FOR SOLVING PROBLEMS] The disk array device reads and writes data by an instruction from the host device. The disk array device has a disk device capable of switching over a first mode in which the data can be written and read and a second mode in which the power consumption is less than in the first mode, and a control unit for changing the disk device from the second mode to the first mode when processing of writing data in the disk device is to be performed.

(57) 要約: (課題) ホスト装置からのデータの読出し命令が多いシステムでの、電力の消費を低減することを可能とするディスクアレイ装置を提供する。(解決手段) ディスクアレイ装置は、ホスト装置からの命令により、データの読出しおよび書込みを実行する。ディスクアレイ装置は、ホスト装置からの命令により、データの読出しおよび書込みを実行する。ディスクアレイ装置は、データの書込みおよび読出し処理を実行し得る第一の状態と該第一の状態よりも電力の消費が少ない第二の状態となることを可能とするディスク装置と、データを該ディスク装置に書込む処理を実行するときに、該ディスク装置を該第二の状態から該第一の状態に変更する制御部とを有する。



IS, IT, LT, LU, LV, MC, MT, NL, PL, PT, RO, SE, SI, SK,
TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW,
ML, MR, NE, SN, TD, TG).

添付公開書類:

— 国際調査報告書

明 細 書

ディスクアレイ装置、制御方法、および制御プログラム

技術分野

[0001] 本発明は、ディスクアレイ装置の制御方式に関する。

背景技術

[0002] ディスクアレイ装置は記憶されたデータの保存の可能性を高めるために複数の記憶ユニットを冗長な構成とする。ディスクアレイ装置は複数のディスク装置を有する。複数のディスク装置を管理する手法としてRAIDがある。RAID (Redundant Array of Independent Disks) は、独立した複数のディスク装置を冗長構成とすることにより、データの保存の信頼性を高めるデータ記録方式である。RAIDを構成するディスク装置をメンバディスクという。

[0003] ディスクアレイ装置が動作するためには電力が必要である。近年、ディスクアレイ装置が消費する電力を低減することが重要になっている。ディスクアレイ装置の省電力化のため、予め定めた時間以上ホスト装置からのアクセスがないときに、RAIDグループの冗長なメンバディスク装置を省電力する技術がある。

[0004] 例えば、データ配信サービスを行うシステムでは、ホスト装置からの命令は、データ読出し命令が多い。しかし、従来技術は、一定時間以内にホスト装置からアクセスがあると省電力モードとはならず、ディスクアレイ装置の消費電力を十分に削減できない。

[0005] 従来技術として下記の文献がある。

特許文献1:特開2005－228288号公報

特許文献2:特開2000－293314号公報

発明の開示

[0006] (発明が解決しようとする課題)

本発明の課題は、ホスト装置からのデータの読出し命令が多いシステムでの、電力の消費を低減することを可能とするディスクアレイ装置を提供することである。

(課題を解決するための手段)

本発明の第一の解決手段のディスクアレイ装置は以下となる。ディスクアレイ装置は、ホスト装置からの命令により、データの読出しおよび書込みを実行する。ディスクアレイ装置は、データの書込みおよび読出し処理を実行し得る第一の状態と第一の状態よりも電力の消費が少ない第二の状態となることを可能とするディスク装置と、データをディスク装置に書込む処理を実行するときに、ディスク装置を第二の状態から第一の状態に変更する制御部とを有する。

[0007] 本発明の第二の解決手段のディスクアレイ装置は、第一の解決手段に加え、ホスト装置から受信した書込みデータを格納するメモリを有する。

[0008] 本発明の第三の解決手段のディスクアレイ装置は、ディスク装置を複数有する。複数のディスク装置はデータが冗長に格納され、複数のディスク装置の少なくとも一つのディスク装置がデータの書込みおよび読出し処理を実行し得る状態よりも電力の消費が少ない第二の状態となる。

[0009] 本発明の第四の解決手段のディスクアレイ装置は、第三の解決手段に加え、制御部は、予め定めた時間内でメモリに格納された書込みデータが所定の値よりも少ないときに、複数のディスク装置で冗長となるディスク装置をデータの読出あるいは書込みが可能な状態よりも消費電力が少ない状態に切り替える。

[0010] 本発明の第五の解決手段のディスクアレイ装置は、第四の解決手段に加え、メモリの書込みデータが所定の値を超えると、ディスク装置をデータの書込みが可能な状態に切り替える。

[0011] 本発明の第五の解決手段のディスクアレイ装置は、第一の解決手段のディスクアレイ装置に以下の機能を加える。第二の状態は更に、ディスク装置のモータを停止する第一のモードとディスク装置への全ての電源を断つ第二のモードとを有する。予め定めた第一の時間内でのメモリに格納された書込みデータが所定の値よりも少ないときに、複数のディスク装置で冗長となるディスク装置を第一のモードに切り替え、予め定めた第二の時間内でのメモリに格納された書込みデータが所定の値よりも少ないときに、複数のディスク装置で冗長となるディスク装置を第二のモードに切り替える。

(発明の効果)

本発明により、ホスト装置とディスクアレイ装置との間でデータの読出し命令が多い

システムでの、電力の消費を低減することを可能とするディスクアレイ装置を提供することができる。

図面の簡単な説明

- [0012] [図1]本実施例のストレージシステム1の構成図である。
[図2]RAIDグループテーブル50の構成例である。
[図3]LUNテーブル60の構成例である。
[図4]READ命令を受信した時の処理のフローチャートである。
[図5]Write命令を受信した時に実行する処理のフローチャートである。
[図6]省電力モード判定処理のフローチャートである。
[図7]省電力モードに移行する移行処理のフローチャートである。
[図8]通常モードに移行する処理のフローチャートである。
[図9]異常検出時の処理のフローチャートである。
[図10]異常なディスク装置がある場合の図6に加える処理のフローチャート
[図11]異常なディスク装置がある場合の図7を変更した処理のフローチャート

符号の説明

- [0013] 1 ストレージシステム
2 ホスト装置
3 ディスクアレイ装置
31 管理部
32 ディスク装置格納部
33 ホストインターフェース
34 制御モジュール
35 キャッシュメモリ
36 ディスクインターフェース
37 ディスク電源制御モジュール

発明を実施するための最良の形態

- [0014] 以下、本発明の実施例について説明する

図1は、本実施例のストレージシステム(Storage System)1の構成図である。スト

レージシステム1は、ホスト装置(Host Computer)2とディスクアレイ装置(Disk Array Device)3とから構成される。

- [0015] ホスト装置2では、種々のアプリケーションソフトウェアが動作する。ホスト装置2は、例えば、アプリケーションソフトウェアで必要となるデータを読み出す命令をディスクアレイ装置3に送信したり、アプリケーションソフトウェアで生成されたデータを書き込む命令をディスクアレイ装置3に送信したりする。
- [0016] ディスクアレイ装置3は、ホスト装置2から送られるデータをディスク装置格納部32のディスク装置格納する処理、ホスト装置2からの要求に応じて格納したデータを読み出す処理を行う。
- [0017] 本実施例のディスクアレイ装置3は、ホスト装置2とネットワークによって接続する。本実施例のディスクアレイ装置3は、複数のディスク装置を管理する管理部31と複数のディスク装置が格納されたディスク装置格納部32とを有する。
- [0018] 管理部31はホストインターフェース(Host I/F)33、制御モジュール(Controller)34、キャッシュメモリ(Cache memory)35、ディスクインターフェース(Disk I/F)36、およびディスク電源制御モジュール(Disk Drive Power Controller)37を有する。
- [0019] ホストインターフェース33は、ホスト装置2と制御モジュール34との間を接続するインターフェースである。
- [0020] 制御モジュール34は、CPU38、揮発メモリ(Volatile Memory)39、および不揮発メモリ(Non-Volatile Memory)40を有する。制御モジュール34は、ディスクアレイ装置3の全体を制御する。例えば、ホスト装置2との間の制御、ホスト装置2から受け取ったデータを一時格納するキャッシュメモリ35の制御、ディスク装置格納部32の複数のディスク装置で構成されるRAIDの制御等を行う。制御モジュール34はホストI/F33、およびディスクアレイ装置3内のディスクIF36に接続する。CPU38は、複数のディスク装置41乃至48を管理するための制御プログラムが実行される。Volatile Memory39では、電源の供給が絶たれると、記憶された内容が失われる。Volatile Memory39は制御プログラムが展開される記憶領域である。また、Volatile Memory39はCPU38が制御プログラムを実行することによって出力される種々の情報を一

時的に格納する。また、Volatile Memory39は、RAIDグループテーブル50、および論理ユニット番号(Logical Unit Number(以下、LUNとする)テーブル60を格納する。Non-Volatile Memory40では、電源が供給されない場合でも記憶された内容が保持される。Non-Volatile Memory40は制御プログラムが格納される。

[0021] キャッシュメモリ35は、ホストから受信した書き込み命令を一時的に記憶するための記憶領域である。ディスクIF36は、制御モジュール34とディスク装置格納部32との間を接続するインターフェースである。ディスク電源制御モジュール37は、ディスク装置格納部32の複数のディスク装置41乃至48の電源の状態を制御する。

[0022] ディスク装置格納部32は複数のディスク装置41、42、43、44、45、46、47、および48を有する。各ディスク装置41乃至48は、制御モジュール34からの指示に応じてデータを格納する、あるいは制御モジュール34からの指示に応じてデータを制御モジュール34に出力する。ディスク装置格納部32の複数のディスク装置41乃至48は、RAIDとして機能する。本実施例のRAIDレベルは「1+0」とする。RAIDレベル「1+0」は、RAIDレベル「1」とRAIDレベル「0」を階層的に組み合わせた仕組みである。RAIDレベル「1」はミラーリングである。RAIDレベル「0」はストライピングである。RAIDレベル「1+0」は、RAIDレベル「1」のミラーリングを複数構成し、複数のミラーリング構成を単位としてRAIDレベル「0」のストライピングを構成する。なお、冗長構成であれば他のRAIDレベル、例えば、RAIDレベル「1」、「5」等であっても実施可能である。

[0023] 次に、Volatile Memory39上に展開されるテーブルについて説明する。ディスクアレイ装置3は、複数のRAIDグループを有する場合がある。ディスクアレイ装置3は、複数のRAIDグループを管理するためのテーブルを有する。図2は、RAIDグループを管理するためのRAIDグループテーブル50である。RAIDグループテーブル50は、RAIDグループを識別するRAIDグループ情報51、RAIDグループのRAIDレベルを特定するRAIDレベル情報52、RAIDグループのメンバディスク数の内、ミラーリングの関係にあるディスク装置を特定するミラー情報53、およびRAIDグループのメンバディスクに対応するディスク装置を特定するDiskID情報54を列項目として

有する。ディスクアレイ装置3は、列項目を一つのまとまりとしたレコードで管理する。

[0024] また、ディスクアレイ装置3は、LUNとRAIDグループとを対応付けるテーブルを有する。本実施例では、LUNは、ディスクアレイ装置3が複数のRAIDグループを持つ場合に、RAIDグループを特定するための識別番号である。ホスト装置2はLUNによって読出しあるいは書込みの対象となるデータを特定する。図3は、LUNとRAIDグループとを対応付けたLUNテーブル60である。LUNテーブル60は、LUNを識別するためのLUN情報61、LUNに対応するRAIDグループを特定するRAIDグループ情報62、ホスト装置2からのREAD命令のREADデータの量を記憶するRead情報63、およびホスト装置2からのWRITE命令のWRITEデータの量を記憶するWRITE情報64を列項目として有する。ディスクアレイ装置3は、列項目を一つのまとまりとしたレコードで管理する。

[0025] 次に、ストレージシステム1でのRead処理について説明する。図4はディスクアレイ装置3がREAD命令を受信した時の処理のフローチャートである。

[0026] ディスクアレイ装置3の制御モジュール34は、ホスト装置2からREAD命令を受信する(S01)。READ命令は、読出しの対象のLUN情報、アドレス情報を有する。制御モジュール34は、読出しの対象のディスク装置を決定する(S02)。READ命令の対象のデータがキャッシュメモリ35に格納されている場合は、制御モジュール34はキャッシュメモリ35から対象のデータを読み出す。

[0027] 例えば、制御モジュール34は特定されたLUNが属するRAIDグループをLUNテーブル60で特定する。そして、制御モジュール34はデータが格納されたアドレスに対応するRAIDグループに属するディスク装置からデータを読み出す(S03)。なお、省電力モードのRAIDグループに対してRead処理を実行する場合、制御モジュール34は、冗長構成となるメンバディスクの内、動作しているディスク装置を特定する。制御モジュール34は、そして制御モジュール34は、特定した動作中のディスク装置からデータを読み出す。制御モジュール34は読み出したデータをホスト装置2に送信する(S04)。また、制御モジュール34は読み出したデータ量をLUNテーブル60のLUNに対応するレコードのRead情報63に格納する。

[0028] ここで、モードについて定義する。本実施例のディスク装置は、通常モードと省電力

モードの2種類のモードを持つとする。通常モードは、記録媒体上のデータの読出しあるいは書き込みのための処理をディスク装置が実行可能な状態である。省電力モードは、記録媒体上のデータの読出しあるいは書き込みのための処理を実行するためには、それらの処理を実行可能な状態に復旧する必要がある状態である。省電力モードでは、ディスク装置の機能の一部は停止している。例えば、電源Offの状態である。あるいは、ディスク装置の制御基板は電源Onの状態であるが、記録媒体が回転していない状態である。通常モードのディスク装置は、データの読出しが可能である。省電力モードのディスク装置は、通常モードよりも電力消費量が少ない。

[0029] 次に、ストレージシステム1でのWrite処理について説明する。図5はディスクアレイ装置3がWrite命令を受信した時に実行する処理のフローチャートである。

[0030] ディスクアレイ装置3の制御モジュール34は、ホスト装置2からWrite命令を受信する(S11)。WRITE命令は、格納すべきデータとデータ格納先のLUNおよびアドレス情報とを有する。制御モジュール34は、S11で受信した格納すべきデータとデータ格納先のLUNおよびアドレス情報とをキャッシュメモリ35に格納する(S12)。制御モジュール34は、Write処理の完了をホスト装置2に送信する(S13)。したがって、S13の時点では、ホスト装置2から受信したWrite命令に関するデータはディスク装置には書き込まれていない。制御モジュール34は、キャッシュメモリ35に格納したデータ量をLUNテーブル60のLUNに対応するレコードのWrite情報64に格納する。

[0031] 制御モジュール34は、キャッシュメモリ35に格納されたデータを適宜ディスク装置に格納する処理を実行する。制御モジュール34は、キャッシュメモリ35からディスク装置に書き込むためのデータ、およびデータを書き込む先のアドレス情報を読み出す(S14)。制御モジュール34は、アドレス情報からデータを書き込む対象のディスク装置を特定する(S15)。制御モジュール34は、対象のディスク装置にデータを書き込む(S16)。

[0032] 以上によって、ディスクアレイ装置3はWrite処理を実行する。なお、制御モジュール34が実行するS11乃至S13の処理とS14乃至S16の処理とは連続して実行する必要はない。ディスクアレイ装置3は、キャッシュメモリ35に格納されたデータを適当なタイミングでディスク装置に書き込む処理を実行する。

- [0033] 次に、ストレージシステム1での省電力モードへの切り替え処理について説明する。図6は、ディスクアレイ装置3が省電力モード判定処理のフローチャートである。省電力モード判定処理では、ディスクアレイ装置3はディスク装置を省電力モードに切り替えてよいか否かを判定する。
- [0034] 制御モジュール34は、単位時間が経過したか否かを判定する(S21)。制御モジュール34がディスク装置を省電力モードに切り替える処理の開始は、種々のタイミングを設定することが可能である。例えば、制御モジュール34は管理するディスク装置に対して最後にWriteBack処理を実行してからタイマーを起動する。そして、制御モジュール34は、タイマーが予め定めた時間を経過すると、ディスク装置を省電力モードに切り替える処理を開始する。また、制御モジュール34は、単位時間毎に省電力モードに切り替える処理を開始する。
- [0035] 制御モジュール34は、処理対象のLUNを特定する(S22)。制御モジュール34は、キャッシュメモリ35に一時的に格納されているデータ量を読み出す(S23)。制御モジュール34は、LUNテーブル60のWrite情報64を読み出す。一般的に、キャッシュメモリ35に一時的に格納されているデータは、WriteBack処理を実行する必要がある。キャッシュメモリ35は高速にデータを格納することが可能であるが、記憶領域は有限である。そのため、制御モジュール34は、キャッシュメモリ35に格納されたデータ量が予め定めたデータ量(所定量)を超えて記憶しているか否かを判定する。所定量は、ディスクアレイ装置3のキャッシュメモリ35の記憶可能な領域の大きさ、あるいはホスト装置2やストレージシステム1で扱うアプリケーションの用途に応じて適宜決定される。例えば、頻繁にデータの更新があるシステムであるにも関わらずキャッシュメモリ35に多くのデータを格納すると、システムが不安定になる恐れがある。一方、ホスト装置2はデータを読み出す処理を主に実行し、ディスク装置にデータを書き込む処理がほとんどないシステムの用途の場合、キャッシュメモリ35に多くのデータがあつたとしても、システムが不安定となる可能性は低い。ディスク装置に格納されたデータの読み出すことが主な処理であり、ディスク装置にデータを書き込む処理がほとんどないようなシステムの用途は、例えば、コンテンツ配信サーバがある。コンテンツ配信サーバでは、送信対象のコンテンツ(データ)をディスク装置に格納した後は、コンテンツ

の読み出しが主な処理となる。一方、コンテンツ配信サーバは、格納されたコンテンツを確実に保護する必要がある。コンテンツを確実に保護するために、コンテンツ配信サーバはRAIDによってコンテンツを冗長に格納する。コンテンツ配信サーバは、例えば、映像配信サーバ、音楽配信サーバなどがある。

[0036] 制御モジュール34は、キャッシュメモリ35に一時的に格納されているデータ量が所定値以上か否かを判定する(S24)。例えば、制御モジュール34は、予め定めた閾値(所定値)とLUNテーブル60のWrite情報64のデータ量とを比較することにより判定する。S24にて所定値以上ではないと判定した場合(S24:No)、制御モジュール34はディスク装置を省電力モードに移行させる処理を実行するためのフラグをOnにする(S25)。一方、S24にて所定値以上であると判定した場合(S24:Yes)、制御モジュール34は処理を終了する。なお、制御モジュール34がWriteBack処理の実行を開始するか否かは省電力モードに移行する処理とは別であるため、本実施例では説明を省略する。

[0037] 制御モジュール34は、ディスクアレイ装置3が有する全てのLUNについて判定をするまで(S26:No)、S22乃至S25の処理を実行する。一方、制御モジュール34は、ディスクアレイ装置3が有する全てのLUNについて判定をすると(S26:Yes)、フラグがOnか否かを判定する(S27)。フラグがOnの場合(S27:Yes)、省電力モードに移行する条件を満たすLUNがあるため、制御モジュール34は省電力モード移行処理を実行する(S28)。一方、フラグがOffの場合(S27:No)、省電力モードに移行する条件を満たすLUNは無いため制御モジュール34は処理を終了する。

[0038] 次に、S28でのストレージシステム1での省電力モード切り替え処理について説明する。図7は、省電力モードに移行する移行処理のフローチャートである。

[0039] 制御モジュール34は、LUNに対応するRAIDグループを特定する(S31)。

[0040] 一つのRAIDグループは複数のLUNによってアクセスされる場合がある。そこで、制御モジュール34は、RAIDグループにアクセスする全てのLUNが省電力モードに移行する条件を満たすか否かを判定する(S32)。制御モジュール34は、RAIDグループに属する全てのLUNのWrite情報64が閾値を超えているか否かを判定する。RAIDグループのLUNが省電力モードを満たさないことは、RAIDグループに対し

て単位時間内に所定値以上のWrite命令が実行されていることである。所定値以上のWrite命令が実行されているにも関わらず省電力モードに移行すると、キャッシュメモリ35の記憶領域が圧迫され、ディスクアレイ装置3の動作は不安定になる。

[0041] 全てのLUNが省電力モードに移行する条件を満たさない場合(S32:No)、制御モジュール34は、RAIDグループを省電力モードに移行することができない。したがって、制御モジュール34は処理を終了する。一方、全てのLUNが省電力モードに移行する条件を満たす場合(S32:Yes)、制御モジュール34は、RAIDグループを構成するメンバディスクの内、冗長な関係となるディスク装置を特定する(S33)。制御モジュール34は、現在の対象となるLUNからLUN情報61を特定する。制御モジュール34は、LUN情報61からRAIDグループ情報62を特定する。制御モジュール34はRAIDグループ情報62からRAIDグループ51を特定し、ディスク情報54を特定する。

[0042] なお、S32において、複数種類の省電力モードを設定することが可能である。省電力モードは、例えば、ディスク装置のモータの回転を停止させる第一の段階、およびディスク装置のモータの回転を停止させ、かつディスク装置の制御基板の電気的な処理も停止させる第二の段階を設定することが可能である。管理者は、第一の段階を満たす条件、及び第二の段階を満たす条件を予め設定する。条件は、例えば、単位時間当たりのWriteBack処理を要するデータがキャッシュメモリに格納される量によって決定する。この場合、第一の省電力モードの段階に移行するデータ量の閾値は、第二の省電力モードの段階に移行するデータ量の閾値よりも大きくなる。なお、条件は、Write Backの閾値を固定値(単位データ量)とし、時間の長短によって決定することとしてもよい。この場合、第一の省電力モードの段階に移行する時間の閾値は、第二の省電力モードの段階に移行する時間の閾値よりも短くなる。第二の省電力モードは、第一の省電力モードよりも通常モードに復旧する可能性が低いためである。

[0043] 制御モジュール34は、RAIDグループの全てのLUNが第一の段階の条件を満たす場合に第一の段階の省電力モードに移行する処理を実行し、RAIDグループの全てのLUNが第二の段階の条件を満たす場合に第二の段階の省電力モードに移

行する処理を実行する。ディスク装置の制御基板の電氣的な処理を停止させかつモータの回転を停止させた第二の省電力モードのほうが、モータの回転を停止させるのみの第一の省電力モードよりも省電力効果は高。しかし第二の省電力モードは、ディスク装置が通常モードに復旧するための時間を多く必要とする。複数の段階に省電力モードを分けることにより、省電力効果があがる。

[0044] 例えば、図1のディスク装置格納部32に格納された各ディスク装置がRAID1+0の構成であり、以下の関係であるとする。ディスク装置41とディスク装置45とがミラーリング1の関係、ディスク装置42とディスク装置46とがミラーリング2の関係、ディスク装置43とディスク装置47とがミラーリング3の関係、およびディスク装置44とディスク装置48とがミラーリング4の関係である。ミラーリング1、ミラーリング2、ミラーリング3、およびミラーリング4がストライピングの関係である。この場合、ディスク装置41とディスク装置45、ディスク装置42とディスク装置46、ディスク装置43とディスク装置47、およびディスク装置44とディスク装置48がミラーリングを構成するため冗長な関係である。なお、ミラーリングの関係は、2つのディスク装置で構成するだけでなく、3つ以上のディスク装置で構成される場合もある。

[0045] 制御モジュール34は、冗長な関係となるディスク装置の片方を省電力モードに移行する(S34)。省電力モードに移行するディスク装置はRAIDの冗長度に応じて定める。RAIDの冗長度はRAIDレベルに応じて固定である。従って制御モジュール34はRAIDグループテーブル50のRAIDレベル情報52によって、冗長度を取得する。なお、冗長構成は、RAIDグループテーブル50のミラー情報53から取得する。また、制御モジュール34は、例えば、冗長関係の内、省電力モード移行するディスク装置の情報を予め有する。あるいは、制御モジュール34は、通常モードから省電力モードに移行する度に、省電力モードに移行するディスク装置を切り替えるようにすることも可能である。

[0046] 制御モジュール34は、ディスク電源制御モジュール37に省電力モードに移行するディスク装置の情報、および、省電力モードの種類情報を送信する。ディスク電源制御モジュール37は対象のディスク装置を省電力モードに切り替える。

[0047] 次に、ストレージシステム1での省電力モード時でのWrite処理について説明する。

省電力モードのディスク装置はデータの書き込みは実行できない。したがって、制御モジュール34はWriteBack処理を実行することは出来ない。そこで、制御モジュール34は、図5のS11乃至S13の処理のみを実行する。S14乃至S16については、書き込み対象のRAIDグループのディスク装置が、省電力モードから通常モードに移行した後で、制御モジュール34がWriteBack処理を実行する。制御モジュール34はキャッシュメモリ35に格納したデータ量をLUNテーブル60のLUNに対応するレコードのWrite情報64に格納する。

[0048] 次に、ストレージシステム1での省電力モードからの復旧処理について説明する。

図8は通常モードに移行する処理のフローチャートである。

[0049] 制御モジュール34は、予め定めた時間毎に図8の処理を開始する構成としてもよい。あるいは、制御モジュール34は、Write命令の量を記憶し、予め定めたWrite命令の量を超えた場合に図8の処理を開始する構成としても良い。時間あるいはWrite命令の量は、ストレージシステム1の扱うアプリケーションの種類、キャッシュメモリ35の領域の大きさによって管理者が定める。

[0050] 制御モジュール34は、S42以下の処理の対象となるLUNを特定する(S41)。制御モジュール34は、LUNテーブル60のWrite情報64から、LUNに対応するキャッシュメモリ35に格納されたWrite Back処理を必要とするデータの量を検出する(S42)。制御モジュール34は、LUNの対象となるWrite Back処理を必要とするデータの量が所定値以上か否かを判定する(S43)。データ量が所定値未満の場合(S43: No)、省電力モードを維持するため、制御モジュール34はS46の処理に移行する。一方、データ量が所定値以上の場合(S43: No)、制御モジュール34は、LUNに対応するディスク装置を通常モードに移行するための処理を実行する。

[0051] ディスク装置を通常モードに移行するための処理はS44、S45である。まず、制御モジュール34は、省電力モードとなっているディスク装置を特定する(S44)。制御モジュール34は、例えば、ミラー情報53、および予め設定された省電力モードに移行するディスク装置の情報によって特定する。制御モジュール34は、S45で特定したディスク装置を通常モードに移行する(S45)。制御モジュール34は、ディスク電源制御モジュール37に対象のディスク装置を通常モードに移行すべき命令を出力する。

ディスク電源制御モジュール37は対象のディスク装置を通常モードに移行させる。

[0052] 次に、制御モジュール34は、全LUNに対する処理が完了したか否かを判定する(S47)。全LUNに対する処理が完了するまで(S47:No)、制御モジュール34はS41からの処理を実行する。全LUNに対する処理が完了すると(S47:Yes)、制御モジュール34は処理を終了する。

[0053] 次に、冗長構成となるディスク装置のいずれかが省電力モードである場合に、制御モジュール34が動作中のディスク装置の異常を検知した場合の処理について説明する。動作中のディスク装置が故障すると、制御モジュール34は冗長構成となる他のディスク装置を省電力モードから通常モードに変更する必要がある。

[0054] 図9は、制御モジュール34がディスク装置の異常を検知した時の処理のフローチャートである。制御モジュール34は、通常モードで動作中のディスク装置の異常を検知する(S51)。制御モジュール34は、異常なディスク装置を記憶する(S52)。制御モジュール34は、異常を検知したディスク装置と冗長の関係となるディスク装置を特定する(S53)。具体的には、制御モジュール34は、異常を検知したディスク装置とRAIDグループテーブル50のミラー情報53とから冗長の関係となるディスク装置を特定する。制御モジュール34は、冗長の関係となるディスク装置を起動する(S54)。制御モジュール34は、冗長の関係となるディスク装置からデータを読み出す(S55)。以上の処理によって、冗長構成の動作中のディスク装置が異常となっても、ディスクアレイ装置3はデータをホスト2に送信することが可能となる。異常が検知されたディスク装置は、別途交換する。

[0055] また、異常なディスク装置を検知している場合での、制御モジュール34が実行する省電力モードに変更する際の処理は以下とする。

[0056] 制御モジュール34は、省電力モードに移行可能か否かを判定するとき、省電力モードに移行可能なRAIDグループ内の冗長構成となるディスク装置に異常があるか否かについても判定する。例えば、制御モジュール34は、図6のフローチャートのS22とS23との間に制御モジュール34は冗長構成となるディスク装置に異常があるか否かを判定する。図10は異常なディスク装置がある場合の図6に加える処理のフローチャートである。

- [0057] 制御モジュール34は、図9のS52で記憶した異常なディスク装置の識別番号を読み出す(S61)。制御モジュール34は、図6のS22で特定した処理対象のLUNに対応するRAIDグループを構成するディスク装置の識別番号は、S61の異常なディスク装置に存在するか否かを判定する(S62)。
- [0058] RAIDグループのディスク装置に異常がない場合に(S62:No)、制御モジュール34は、図6のS23以下の他の省電力モードへの移行条件を満たすか否かを判定する。一方、RAIDグループのディスク装置に異常がある場合に(S62:Yes)、制御モジュール34は省電力モードには移行しないことを決定する(S63)。制御モジュール34は、異常なディスク装置とミラーリングの関係となるディスク装置群をミラー情報53で特定し、記憶する。
- [0059] また、RAIDグループ内の冗長構成となるディスク装置に異常があるときの制御モジュール34が実行する省電力モードに移行する処理は以下となる。制御モジュール34は、図7のフローチャートのS33とS34の代わりに以下の処理を実行する。図11は異常なディスク装置がある場合の図7を変更した処理のフローチャートである。
- [0060] 図7のS32でRAIDグループの全てのLUNが省電力モードに移行する条件を満たすとき(S32:Yes)、制御モジュール34は、処理の対象のRAIDグループ内に対応するミラー情報を検索する(S71)。制御モジュール34は、対象のRAIDグループ情報51のミラー情報53と図10のS63で記憶したミラー情報とを比較する(S72)。
- [0061] 対象のRAIDグループが異常なディスク装置を含むミラー情報を含む場合(S72:Yes)、制御モジュール34は検出したミラー情報に含まれるディスク装置以外の冗長構成の片方のディスク装置を省電力モードに移行する(S73)。一方、対象のRAIDグループが異常なディスク装置を含むミラー情報を含まない場合(S72:No)、制御モジュール34は冗長構成の片方のディスク装置を省電力モードに移行する(S74)。
- [0062] 以上によって、ディスクアレイ装置3は、異常なディスク装置がある場合に、異常なディスク装置を除いたディスク装置のみを対象として省電力モードに移行することが可能となる。なお、異常が検知されたディスク装置について、例えばディスク装置の交換によって、異常が解消されると、制御モジュール34はS52で記憶した異常なディスク装置のID情報、およびS63の異常なディスク装置のミラー情報を更新する。

[0063] 以上により、ディスクアレイ装置3は、ディスクアレイ装置3内のアクセスがある各ディスク装置についても適切に省電力モードに移行することが可能となる。

[0064] 本実施例では、それぞれのLUNが省電力モードに移行するための条件を満たすか否かについて、制御モジュール34が毎回判定する。しかし、例えば、テーブルに各LUNの省電力モードの状態情報、省電力モードに移行することが可能か否かを示す情報を格納することも可能である。なお、制御モジュール34が毎回判定することにより、各LUNの省電力モードの状態情報、省電力モードに移行することが可能か否かを示す情報を格納するための記憶領域を設定する必要がない。

[0065] 以上により、ディスク装置に格納されたデータの読出す処理が多く、ディスク装置にデータを書き込む処理が少ない用途のサーバで省電力の効果が大きい。ディスク装置に格納されたデータの読出す処理が多く、ディスク装置にデータを書き込む処理が少ない用途のサーバの例は、映像配信サーバ、音楽配信サーバなどのコンテンツを配信するサーバである。

産業上の利用可能性

[0066] 本発明は、ホスト装置からの命令により、データの読出しおよび書込みを実行するディスクアレイ装置に利用可能である。

請求の範囲

- [1] ホスト装置からの命令により、データの読出しおよび書込みを実行するディスクアレイ装置であって、
- データの書込みおよび読出し処理を実行し得る第一の状態と該第一の状態よりも電力の消費が少ない第二の状態となることを可能とするディスク装置と、
- データを該ディスク装置に書込む処理を実行するときに、該ディスク装置を該第二の状態から該第一の状態に変更する制御部と
- を有することを特徴とするディスクアレイ装置。
- [2] 該ホスト装置から受信した書込みデータを格納するメモリを更に有す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。
- [3] 該ディスク装置を複数有し、
- 該複数のディスク装置はデータが冗長に格納され、
- 該複数のディスク装置の少なくとも一つのディスク装置がデータの書込みおよび読出し処理を実行し得る状態よりも電力の消費が少ない第二の状態となることを特徴とする請求項2に記載のディスクアレイ装置。
- [4] 該ホスト装置から受信した書込みデータを格納するメモリを更に有し、
- 該制御部は、予め定めた時間内で該メモリに格納された書込みデータが所定の値よりも少ないときに、該複数のディスク装置で冗長となるディスク装置をデータの読出あるいは書込が可能な状態よりも消費電力が少ない状態に切り替える
- ことを特徴とする請求項3に記載のディスクアレイ装置。
- [5] 該メモリの書込みデータが該所定の値を超えると、ディスク装置を該データの書込みが可能な状態に切り替えることを特徴とする請求項4に記載のディスクアレイ装置。
- [6] 該第二の状態は更に、該ディスク装置のモータを停止する第一のモードと該ディスク装置への全ての電源を断つ第二のモードとがあり、
- 予め定めた第一の時間内での該メモリに格納された書込みデータが所定の値よりも少ないときに、該複数のディスク装置で冗長となるディスク装置を該第一のモードに切り替え、
- 予め定めた第二の時間内での該メモリに格納された書込みデータが所定の値より

も少ないときに、該複数のディスク装置で冗長となるディスク装置を該第二のモードに切り替えることを特徴とする請求項4に記載のディスクアレイ装置。

- [7] データの書込みおよび読出し処理を実行し得る第一の状態、あるいは該第一の状態よりも電力の消費が少ない第二の状態となるディスク装置を有し、ホスト装置からの命令によりデータの読出しおよび書込みを実行するディスクアレイ装置の制御方法であって、

データを該ディスク装置に書込む処理を実行するときに、該ディスク装置を該第二の状態から該第一の状態に変更し、

該ディスク装置に該データの書込み処理を実行することを特徴とするディスクアレイ装置の制御方法。

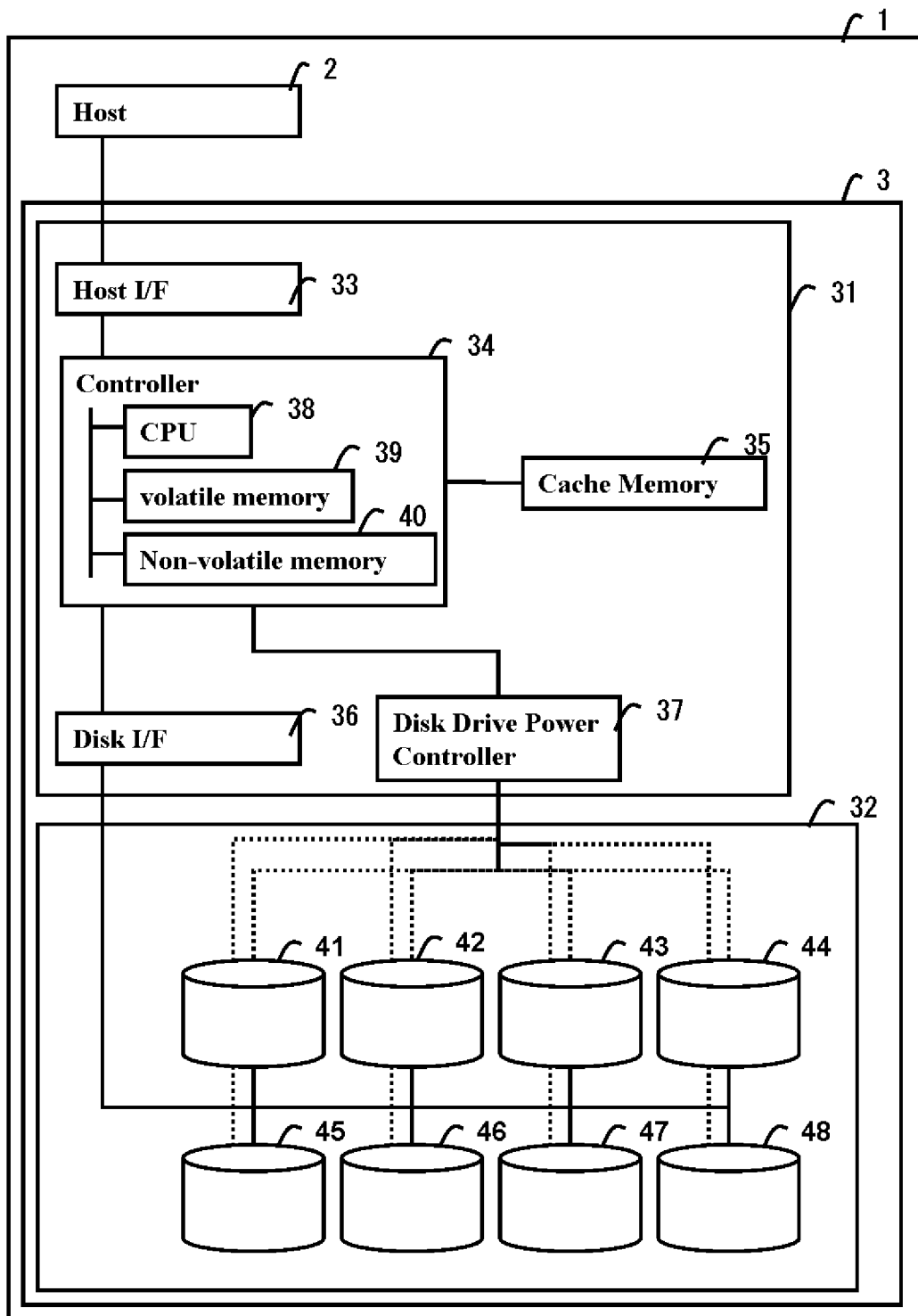
- [8] データの書込みおよび読出し処理を実行し得る第一の状態、あるいは該第一の状態よりも電力の消費が少ない第二の状態となるディスク装置と、該ディスク装置の制御を行う制御部を有し、ホスト装置からの命令により、データの読出しおよび書込み処理をディスクアレイ装置に実行させる制御プログラムであって、

該ディスクアレイ装置の該制御部に、

データを該複数のディスク装置に書込む処理を実行するときに、該複数のディスク装置を該第二の状態から該第一の状態に変更するステップ、

該ディスク装置に該データの書込み実行するステップ、
を実行させることを特徴とする制御プログラム。

[図1]



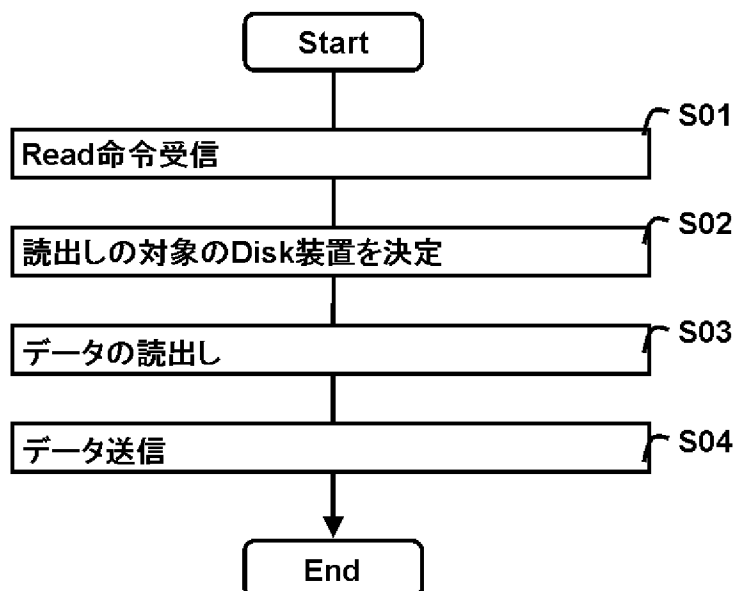
[図2]

RAID Group	RAID Level	mirror	Disk ID
1	1	(1,2) (3,4)	1,2,3,4
2	5	-	5,6,7,8
3	1+0	(9,10) (11,12)	9,10,11,12
4	1+0	(13,14) (15,16)	13,14,15,16

[図3]

LUN	RAID Group	Read (Byte)	Write (Byte)
1	1	10,000	5,000
2	1	100,000	0
3	2	100,000	0
4	3	100,000	0
5	2	100,000	10,000

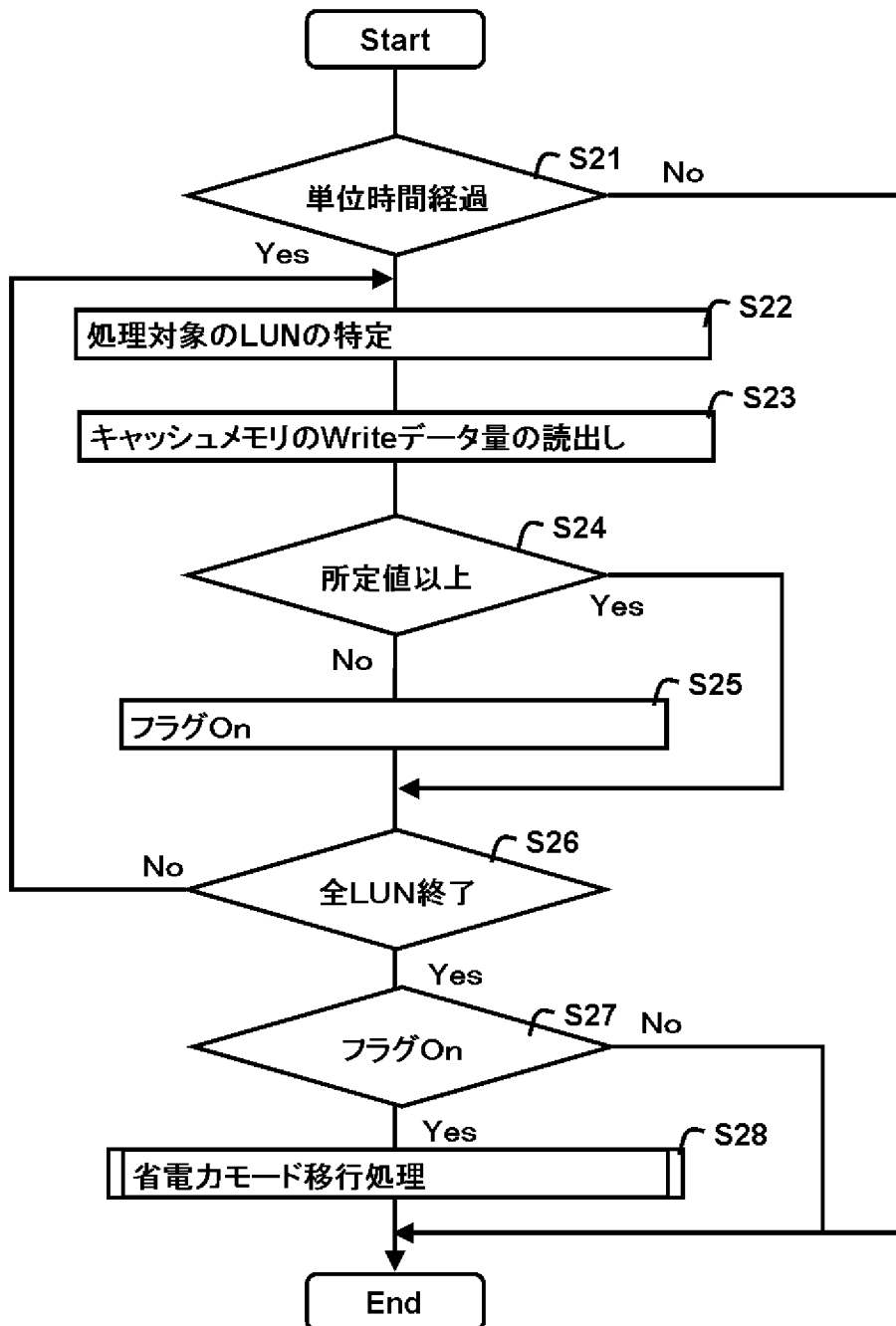
[図4]



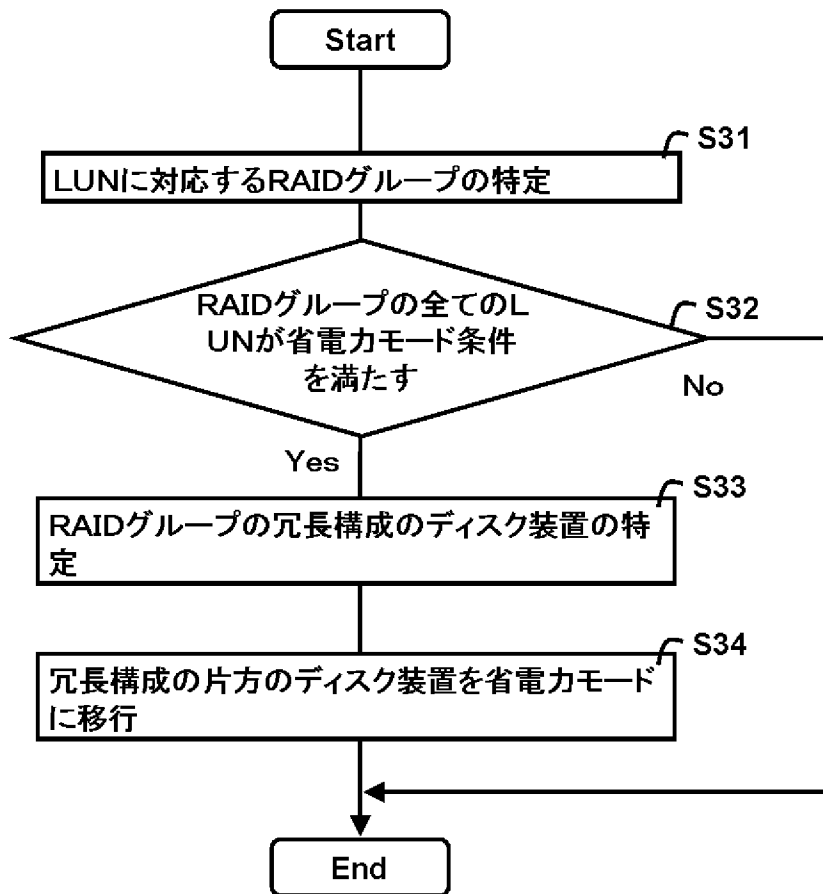
[図5]



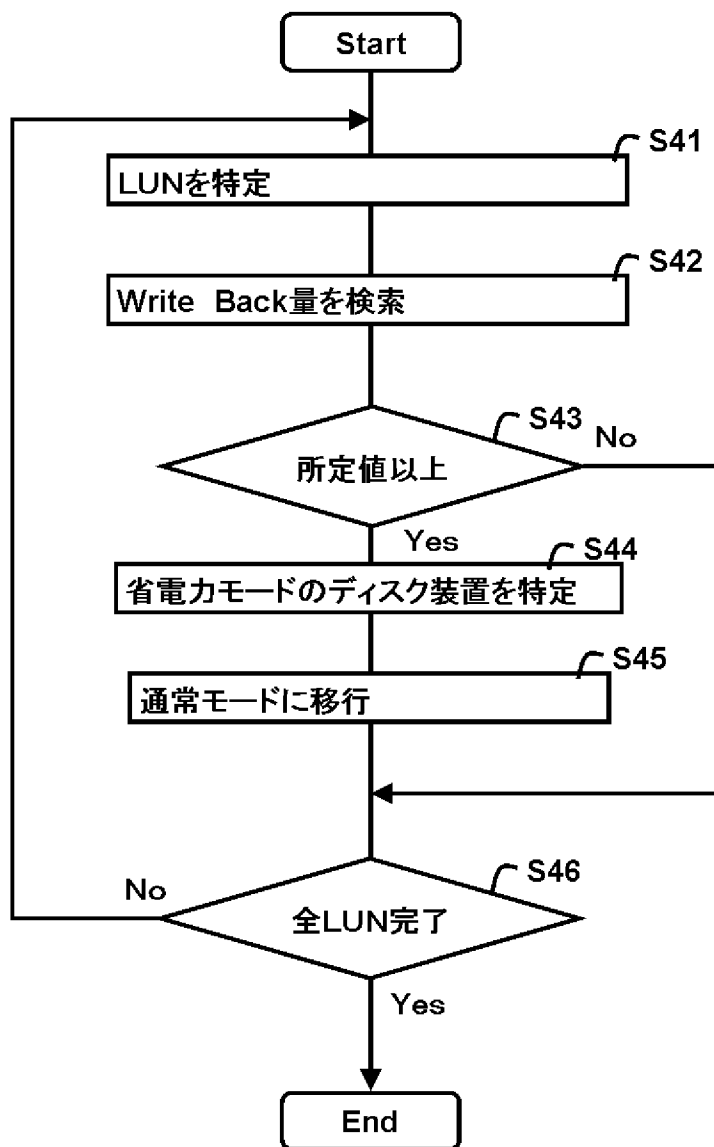
[図6]



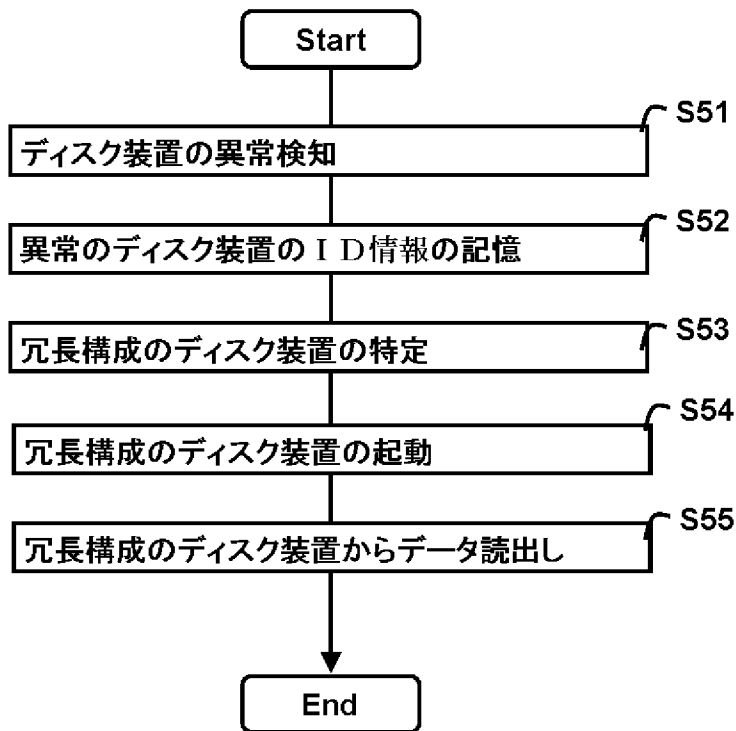
[図7]



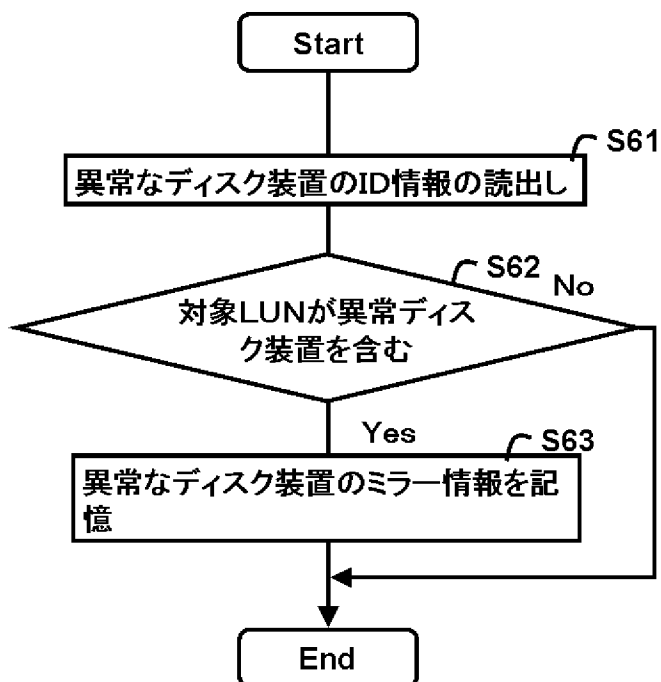
[図8]



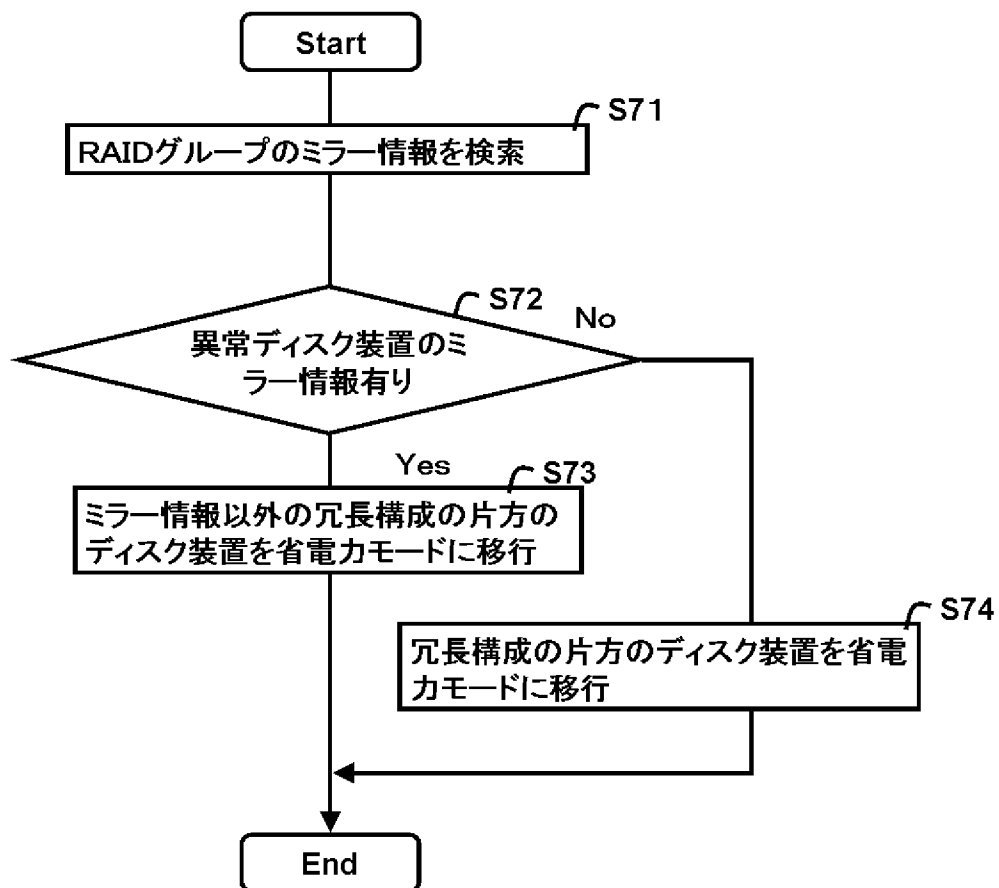
[図9]



[図10]



[図11]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2007/063901

A. CLASSIFICATION OF SUBJECT MATTER

G06F3/06 (2006.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F3/06

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Jitsuyo Shinan Koho	1922-1996	Jitsuyo Shinan Toroku Koho	1996-2007
Kokai Jitsuyo Shinan Koho	1971-2007	Toroku Jitsuyo Shinan Koho	1994-2007

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X Y	JP 2005-228288 A (Hitachi, Ltd.), 25 August, 2005 (25.08.05), Par. Nos. [0011] to [0058] & US 2005-0160221 A1 & GB 2410345 A & GB 2411269 A & DE 102004027671 A1 & FR 2865287 A1	1-3, 7, 8 4-6
Y	JP 2007-156597 A (Hitachi, Ltd.), 21 June, 2007 (21.06.07), Par. Nos. [0093] to [0165] (Family: none)	4-6
A	JP 11-338644 A (Hitachi, Ltd.), 10 December, 1999 (10.12.99), Par. Nos. [0026] to [0031] (Family: none)	1-8

☒ Further documents are listed in the continuation of Box C.

☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search
07 September, 2007 (07.09.07)

Date of mailing of the international search report
18 September, 2007 (18.09.07)

Name and mailing address of the ISA/
Japanese Patent Office

Authorized officer

Facsimile No.

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2007/063901

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 2000-293314 A (Hitachi, Ltd.), 20 October, 2000 (20.10.00), Par. Nos. [0017] to [0050]; Fig. 1 & US 6583947 B1 & US 2003-0193732 A1 & US 2003-0193733 A1 & US 2005-0259345 A1 & US 2006-0193073 A1	1-8

A. 発明の属する分野の分類（国際特許分類（I P C））

Int.Cl. G06F3/06(2006.01)i

B. 調査を行った分野

調査を行った最小限資料（国際特許分類（I P C））

Int.Cl. G06F3/06

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	1922-1996年
日本国公開実用新案公報	1971-2007年
日本国実用新案登録公報	1996-2007年
日本国登録実用新案公報	1994-2007年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

C. 関連すると認められる文献

引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求の範囲の番号
X	JP 2005-228288 A（株式会社日立製作所）2005.08.25, 段落【0011】－【0058】	1-3, 7, 8
Y	& US 2005-0160221 A1 & GB 2410345 A & GB 2411269 A & DE 102004027671 A1 & FR 2865287 A1	4-6
Y	JP 2007-156597 A（株式会社日立製作所）2007.06.21, 段落【0093】－【0165】（ファミリーなし）	4-6

☒ C欄の続きにも文献が列挙されている。☐ パテントファミリーに関する別紙を参照。

* 引用文献のカテゴリー

「A」特に関連のある文献ではなく、一般的技術水準を示すもの
「E」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
「L」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
「O」口頭による開示、使用、展示等に言及する文献
「P」国際出願日前で、かつ優先権の主張の基礎となる出願

の日の後に公表された文献

「T」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
「X」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
「Y」特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
「&」同一パテントファミリー文献

国際調査を完了した日

07.09.2007

国際調査報告の発送日

18.09.2007

国際調査機関の名称及びあて先

日本国特許庁（I S A/J P）
郵便番号100-8915
東京都千代田区霞が関三丁目4番3号

特許庁審査官（権限のある職員）

梅景 篤

5 I

3801

電話番号 03-3581-1101 内線 3565

C (続き) . 関連すると認められる文献		
引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求の範囲の番号
A	JP 11-338644 A (株式会社日立製作所) 1999.12.10, 段落【0026】－【0031】 (ファミリーなし)	1－8
A	JP 2000-293314 A (株式会社日立製作所) 2000.10.20, 段落【0017】－【0050】、図1 & US 6583947 B1 & US 2003-0193732 A1 & US 2003-0193733 A1 & US 2005-0259345 A1 & US 2006-0193073 A1	1－8