

(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2015-191407

(P2015-191407A)

(43) 公開日 平成27年11月2日(2015.11.2)

(51) Int. Cl. F I テーマコード (参考)
G06F 3/06 (2006.01)
 G06F 3/06 305C
 G06F 3/06 301Z

審査請求 未請求 請求項の数 7 O L (全 18 頁)

(21) 出願番号 特願2014-67626 (P2014-67626)
 (22) 出願日 平成26年3月28日 (2014. 3. 28)

(71) 出願人 000005223
 富士通株式会社
 神奈川県川崎市中原区上小田中4丁目1番1号
 (74) 代理人 100092152
 弁理士 服部 毅巖
 (72) 発明者 石川 圭也
 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
 (72) 発明者 塚本 新菜
 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

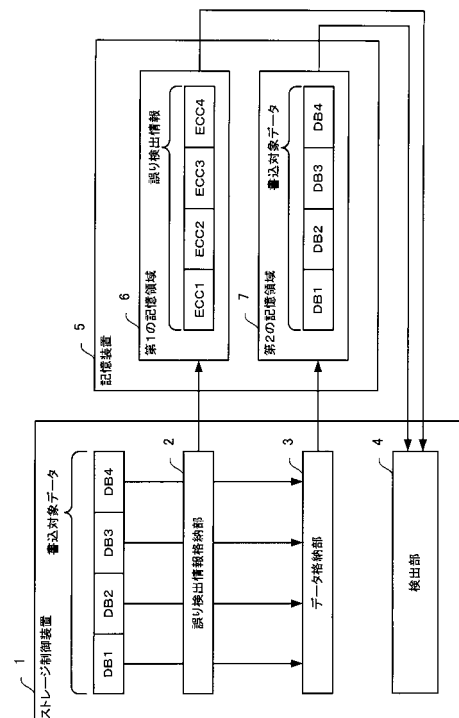
(54) 【発明の名称】 ストレージ制御装置、制御プログラム、および制御方法

(57) 【要約】

【課題】データ書込失敗の検出精度を高める。

【解決手段】ストレージ制御装置 1 は、記憶装置 5 を制御対象とする。記憶装置 5 は、第 1 の記憶領域 6 と、第 1 の記憶領域 6 と異なる第 2 の記憶領域 7 を含む。誤り検出情報格納部 2 は、書込対象データについて誤り検出情報として、データブロックごとに ECC を生成する。誤り検出情報格納部 2 は、生成した ECC 1 から ECC 4 を第 1 の記憶領域 6 に格納する。データ格納部 3 は、DB 1 から DB 4 を第 2 の記憶領域 7 に格納する。検出部 4 は、第 1 の記憶領域 6 から読み出した誤り検出情報と、第 2 の記憶領域 7 から読み出した書込対象データとからデータブロックごとの誤り検出をおこなう。

【選択図】 図 1



【特許請求の範囲】**【請求項 1】**

書込対象データについてアクセス単位となるデータブロックごとに誤り検出情報を生成し、記憶装置の第 1 の記憶領域に前記誤り検出情報を格納する誤り検出情報格納部と、前記記憶装置の前記第 1 の記憶領域と異なる第 2 の記憶領域に前記書込対象データを格納するデータ格納部と、前記第 1 の記憶領域から読み出した前記誤り検出情報と、前記第 2 の記憶領域から読み出した前記書込対象データとから前記データブロックごとの誤り検出をおこなう検出部と、

を備えることを特徴とするストレージ制御装置。

10

【請求項 2】

前記記憶装置は、ディスクドライブ装置であって、前記第 1 の記憶領域は、前記ディスクドライブ装置のメタ情報を格納するメタ情報記憶領域であって、前記第 2 の記憶領域は、前記ディスクドライブ装置のユーザデータを格納するユーザデータ記憶領域である、

ことを特徴とする請求項 1 記載のストレージ制御装置。

【請求項 3】

前記書込対象データは、キャッシュメモリに記憶されていることを特徴とする請求項 2 記載のストレージ制御装置。

20

【請求項 4】

前記キャッシュメモリは、前記誤り検出情報を格納するキャッシュメモリ誤り検出情報格納部と、前記キャッシュメモリの前記誤り検出情報を格納する記憶領域と異なる書込対象データ記憶領域に前記書込対象データを格納するキャッシュメモリデータ格納部と、

を備えることを特徴とする請求項 3 記載のストレージ制御装置。

【請求項 5】

前記記憶装置は、ディスクドライブ装置であって、前記誤り検出情報の格納先である誤り検出情報格納先情報と、前記書込対象データの格納先である書込対象データ格納先情報とを、前記第 1 の記憶領域および前記第 2 の記憶領域と異なる第 3 の記憶領域に格納する格納先情報格納部を備え、

前記誤り検出情報格納部は、前記誤り検出情報格納先情報にもとづいて前記第 1 の記憶領域に前記誤り検出情報を格納し、

前記データ格納部は、前記書込対象データ格納先情報にもとづいて前記第 2 の記憶領域に前記書込対象データを格納する、

ことを特徴とする請求項 1 記載のストレージ制御装置。

30

【請求項 6】

記憶装置の制御プログラムであって、コンピュータに、

書込対象データについてアクセス単位となるデータブロックごとに誤り検出情報を生成し、前記記憶装置の第 1 の記憶領域に前記誤り検出情報を格納し、

前記記憶装置の前記第 1 の記憶領域と異なる第 2 の記憶領域に前記書込対象データを格納し、

前記第 1 の記憶領域から読み出した前記誤り検出情報と、前記第 2 の記憶領域から読み出した前記書込対象データとから前記データブロックごとの誤り検出をおこなう、

処理を実行させることを特徴とする制御プログラム。

40

【請求項 7】

記憶装置の制御方法であって、コンピュータが、

書込対象データについてアクセス単位となるデータブロックごとに誤り検出情報を生成し、前記記憶装置の第 1 の記憶領域に前記誤り検出情報を格納し、

50

前記記憶装置の前記第 1 の記憶領域と異なる第 2 の記憶領域に前記書込対象データを格納し、

前記第 1 の記憶領域から読み出した前記誤り検出情報と、前記第 2 の記憶領域から読み出した前記書込対象データとから前記データブロックごとの誤り検出をおこなう、

処理を実行することを特徴とする制御方法。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、ストレージ制御装置、制御プログラム、および制御方法に関する。

【背景技術】

【0002】

情報処理システムに使用されるストレージ装置として、データアクセスの高速性やデータの障害耐性を高める構成を有するディスクアレイ装置 (RAID (Redundant Arrays of Inexpensive Disks) 装置ともいう) が多く用いられている。このディスクアレイ装置は、種類の異なる汎用的なオペレーティングシステムによって稼動するオープン系システムで多く利用されるため、固定長レコードのデータフォーマットを有している。

【0003】

データをストレージ装置に記憶する際には、記憶したデータにエラーが生じていないかを判定するためのエラーチェックコードがデータに付される。たとえば、固定長データフォーマットでは、論理ブロックごとにエラーチェックコードが付される。

【先行技術文献】

【特許文献】

【0004】

【特許文献 1】特開 2006 - 107311 号公報

【特許文献 2】特開 2007 - 115390 号公報

【発明の概要】

【発明が解決しようとする課題】

【0005】

ストレージ制御装置は、ホストからライト要求を受け付けたデータをメモリ上に一旦保持してからストレージ装置に記憶させる。このとき、異常が生じてストレージ装置に正常なデータを書き込むことができなくなることがある。

【0006】

たとえば、データ異常の態様によっては、ストレージ装置に論理ブロック単位でデータが書き込まれず、書き込み前のデータが更新されずに保持される場合がある。このとき、ストレージ装置は、更新前のデータと、更新前のデータに対応するエラーチェックコードを保持している状態である。ストレージ制御装置は、この状態でステージングをおこなうと、キャッシュメモリに読み出したエラーチェックコードと、キャッシュメモリに読み出したデータから算出したエラーチェックコードとが一致するため正常なデータであると判断する。ストレージ制御装置は、ストレージ装置のデータ書込の失敗を検出することができないことから、更新前のデータをホストに回答して、ホストでデータ化けを生じ得る。

【0007】

1 つの側面では、本発明は、データ書込失敗の検出精度を高めたストレージ制御装置、制御プログラム、および制御方法を提供することを目的とする。

【課題を解決するための手段】

【0008】

上記目的を達成するために、以下に示すような、ストレージ制御装置が提供される。ストレージ制御装置は、誤り検出情報格納部と、データ格納部と、検出部とを備える。誤り検出情報格納部は、書込対象データについてアクセス単位となるデータブロックごとに誤り検出情報を生成し、記憶装置の第 1 の記憶領域に誤り検出情報を格納する。データ格納部は、記憶装置の第 1 の記憶領域と異なる第 2 の記憶領域に書込対象データを格納する。

10

20

30

40

50

検出部は、第 1 の記憶領域から読み出した誤り検出情報と、第 2 の記憶領域から読み出した書込対象データとからデータブロックごとの誤り検出をおこなう。

【発明の効果】

【0009】

1 態様によれば、ストレージ制御装置、制御プログラム、および制御方法において、データ書込失敗の検出精度を高める。

【図面の簡単な説明】

【0010】

【図 1】第 1 の実施形態のストレージ制御装置の構成の一例を示す図である。

【図 2】第 2 の実施形態のストレージサーバの構成の一例を示す図である。

10

【図 3】第 2 の実施形態のキャッシュメモリの構成の一例を示す図である。

【図 4】第 2 の実施形態のディスクモジュールの構成の一例を示す図である。

【図 5】第 2 の実施形態のデータブロックの一例を示す図である。

【図 6】第 2 の実施形態の C A のハードウェア構成の一例である。

【図 7】第 2 の実施形態のライト要求受付処理のフローチャートを示す図である。

【図 8】第 2 の実施形態のライトバック処理のフローチャートを示す図である。

【図 9】第 2 の実施形態のブロックライト失敗の一例を示す図である。

【図 10】ブロックライト失敗の比較例を示す図である（その 1）。

【図 11】第 2 の実施形態のリード要求受付処理のフローチャートを示す図である。

【図 12】ブロックライト失敗の比較例を示す図である（その 2）。

20

【図 13】第 3 の実施形態のストレージ制御装置の構成の一例を示す図である。

【図 14】第 3 の実施形態のライト要求受付処理のフローチャートを示す図である。

【発明を実施するための形態】

【0011】

以下、図面を参照して実施の形態を詳細に説明する。

[第 1 の実施形態]

まず、第 1 の実施形態のストレージ制御装置について図 1 を用いて説明する。図 1 は、第 1 の実施形態のストレージ制御装置の構成の一例を示す図である。

【0012】

ストレージ制御装置 1 は、記憶装置 5 へのデータの書込と、記憶装置 5 からのデータの読出しを制御する。ストレージ制御装置 1 は、図示しない外部装置からデータの書込指示を受け付けて記憶装置 5 にデータを書き込み、また、外部装置からデータの読出指示を受け付けて記憶装置 5 からデータを読み出して外部装置に応答する。

30

【0013】

ストレージ制御装置 1 は、誤り検出情報格納部 2 と、データ格納部 3 と、検出部 4 とを含む。記憶装置 5 は、ストレージ制御装置 1 が制御対象とするストレージ装置であり、第 1 の記憶領域 6 と第 2 の記憶領域 7 を含む。第 1 の記憶領域 6 と第 2 の記憶領域 7 は、異なる記憶領域である。なお、記憶装置 5 は、ストレージ制御装置 1 にとって外部記憶装置であってもよいし、内部記憶装置であってもよい。また、記憶装置 5 は、H D D (Hard Disk Drive) や S S D (Solid State Drive) などのディスクドライブであってもよいし、データを一時的に格納するキャッシュメモリなどであってもよい。

40

【0014】

誤り検出情報格納部 2 は、書込対象データについて誤り検出情報として、データブロックごとに E C C (Error Check Code) を生成する。データブロックは、記憶装置 5 のアクセス単位である。たとえば、記憶装置 5 が H D D であるときに、アクセス単位は、L B A (Logical Block Address: 論理ブロックアドレス) である。誤り検出情報格納部 2 は、書込対象データをデータブロックに分割する。なお、図 1 は、D B 1, D B 2, D B 3, D B 4 の 4 つのデータブロックを示すが、データブロックの数は書込対象データの大きさにしたがうものであり、図示した 4 つに限らない。

【0015】

50

誤り検出情報格納部 2 は、D B 1 から E C C 1 を生成し、データブロック D B 2 から E C C 2 を生成し、データブロック D B 3 から E C C 3 を生成し、データブロック D B 4 から E C C 4 を生成する。誤り検出情報格納部 2 は、生成した E C C 1 から E C C 4 を第 1 の記憶領域 6 に格納する。

【 0 0 1 6 】

データ格納部 3 は、D B 1 から D B 4 を第 2 の記憶領域 7 に格納する。検出部 4 は、第 1 の記憶領域 6 から読み出した誤り検出情報と、第 2 の記憶領域 7 から読み出した書込対象データとからデータブロックごとの誤り検出をおこなう。たとえば、検出部 4 は、第 1 の記憶領域 6 から E C C 1 を読み出し、第 2 の記憶領域 7 から D B 1 を読み出す。検出部 4 は、第 2 の記憶領域 7 から読み出した D B 1 から E C C 1 を生成し、第 1 の記憶領域 6 から読み出した E C C 1 と比較することで、D B 1 の誤り検出をおこなうことができる。

10

【 0 0 1 7 】

これにより、ストレージ制御装置 1 は、第 1 の記憶領域 6 が記憶する誤り検出情報と、第 2 の記憶領域 7 が記憶する書込対象データのうち少なくともいずれか一方のデータが、対応する記憶領域に正しく書き込まれていないこと（データ書込失敗）を検出できる。したがって、ストレージ制御装置 1 は、データ書込失敗の検出精度を向上することができる。

【 0 0 1 8 】

[第 2 の実施形態]

次に、第 2 の実施形態のストレージサーバについて図 2 を用いて説明する。図 2 は、第 2 の実施形態のストレージサーバの構成の一例を示す図である。

20

【 0 0 1 9 】

ストレージサーバ 1 0 は、複数の H D D 2 5 , 2 6 を備え、ホスト（ホストコンピュータ）1 1 , 1 2 に高速、大容量、高信頼性のディスクシステムを提供する。ストレージサーバ 1 0 は、ホスト 1 1 , 1 2 からライト要求を受け付けて H D D 2 5 , 2 6 にデータを書き込み、また、ホスト 1 1 , 1 2 からリード要求を受け付けて H D D 2 5 , 2 6 からデータを読み出して応答する。H D D 2 5 , 2 6 は、ストレージ装置の一形態であり、ストレージサーバ 1 0 は、ストレージ装置を制御するストレージ制御装置の一形態である。

【 0 0 2 0 】

ストレージサーバ 1 0 は、C A (Channel Adapter) 1 3 , 1 4、スイッチ 1 5 , 1 6 , 2 3 , 2 4、C M (Controller Module) 1 7 , 1 8、キャッシュメモリ 1 9 , 2 0、D A (Device Adapter) 2 1 , 2 2、H D D 2 5 , 2 6 を含む。

30

【 0 0 2 1 】

C A 1 3 , 1 4 は、ストレージサーバ 1 0 の I / F (Interface) 制御モジュールである。C A 1 3 , 1 4 は、F C (Fibre Channel)、i S C S I (Internet Small Computer System Interface)、F C L I N K、O C L I N Kなどの各種インタフェースを制御し、データを内部バスへ転送する、プロトコルチップとして機能する。C A 1 3 は、2 以上あり、ホスト 1 1 とスイッチ 1 5 を冗長構成で接続する。C A 1 4 は、2 以上あり、ホスト 1 2 とスイッチ 1 6 を冗長構成で接続する。

【 0 0 2 2 】

40

C A 1 3 , 1 4 は、ホスト 1 1 , 1 2 からのコマンドと、キャッシュメモリ 1 9 , 2 0 へのアクセス処理を担う。C A 1 3 は、スイッチ 1 5 を介して C M 1 7 を接続する。C A 1 4 は、スイッチ 1 6 を介して C M 1 8 を接続する。C A 1 3 , 1 4 は、ホスト 1 1 , 1 2 から転送されるユーザデータにディスクフォーマットに応じたデータ保護のため B C C (Block Check Code) を付加する。B C C は、データブロック単位の E C C である。また、C A 1 3 , 1 4 は、H D D 2 5 , 2 6 から読み出されたデータをチェックし、B C C を外してホスト 1 1 , 1 2 に転送する。

【 0 0 2 3 】

C M 1 7 , 1 8 は、データの排他制御、キャッシュメモリの管理、H D D のリード/ライト制御など、ストレージサーバ 1 0 を統括的に制御する制御モジュールである。C M 1

50

7は、キャッシュメモリ19を内蔵し、キャッシュメモリ19を管理する。CM18は、キャッシュメモリ20を内蔵し、キャッシュメモリ20を管理する。ストレージサーバ10は、CM17とCM18とを備える冗長構成を有し、一方に障害が発生した場合に他方が処理を引き継ぐことができる。また、ユーザデータは、CM17とCM18の2系統でミラーリングされ、ミラーリングされたユーザデータは、CM17とCM18で共有される。

【0024】

DA21, 22は、CM17, 18と、複数のHDD25, 26をスイッチ23, 24を介して接続し、HDD25, 26を制御する。HDD25, 26は、ドライブモジュールの一形態であって、たとえば、SSDなどに置き換えることができる。HDD25, 26は、たとえば、DA21, 22、およびスイッチ23, 24とともにディスクアレイモジュールを構成する。

10

【0025】

次に、第2の実施形態のキャッシュメモリについて図3を用いて説明する。図3は、第2の実施形態のキャッシュメモリの構成の一例を示す図である。キャッシュメモリ19とキャッシュメモリ20は、同様の構成であるため、キャッシュメモリ19について説明する。

【0026】

キャッシュメモリ19は、BCC領域191とユーザデータ領域192を含む。ユーザデータ領域192は、ユーザデータを記憶する記憶領域である。BCC領域191は、メタデータを記憶する記憶領域の1つであり、ユーザデータ領域192と異なる記憶領域である。

20

【0027】

ユーザデータ領域192は、データブロックごとにユーザデータを記憶する。ここでいうデータブロックは、HDD25, 26のアクセス単位であり、たとえば、LBA(Logical Block Address)である。ユーザデータ領域192は、たとえば、データブロックごとにデータ1、データ2、...、データ8を記憶する。

【0028】

BCC領域191は、各データブロックのBCCを記憶する。BCCは、誤り検出情報の1つであり、たとえば、6バイトのBID(Block ID)と、2バイトのCRC(Cyclic Redundancy Check)とから構成される。BIDは、データブロックを一意に識別する識別子である。BCC領域191は、たとえば、データ1に対応するBCC1、データ2に対応するBCC2、...、データ8に対応するBCC8を記憶する。

30

【0029】

このように、キャッシュメモリ19において、ユーザデータと、ユーザデータのBCCとは、それぞれアクセス単位の異なる記憶領域に記憶される。そのため、ユーザデータと、ユーザデータのBCCとが同一の記憶領域に記憶されることによる、ユーザデータと、ユーザデータのBCCとが同時にデータ書込失敗となる危険を低減する。

【0030】

次に、第2の実施形態のディスクモジュールについて図4を用いて説明する。図4は、第2の実施形態のディスクモジュールの構成の一例を示す図である。HDD25とHDD26は、ディスクモジュールの一形態であり、より詳しくは、ディスクドライブ装置の一形態である。HDD25とHDD26は、同様の構成であるため、HDD25について説明する。

40

【0031】

HDD25は、メタ領域251とユーザデータ領域252を含む。ユーザデータ領域252は、ユーザデータを記憶する記憶領域である。メタ領域251は、メタデータを記憶する記憶領域の1つであり、ユーザデータ領域252と異なる記憶領域である。メタデータは、各データブロックのBCCを含む。また、メタデータは、領域管理に関する領域管理情報を含み、領域管理情報は、たとえば、iノード、Vデータ、ファイルデータなどの

50

割付情報がある。

【0032】

ユーザデータ領域252は、データブロックごとにユーザデータを記憶する。ここでいうデータブロックは、HDD25, 26のアクセス単位であり、たとえば、LBAに対応する物理アドレスである。ユーザデータ領域252は、たとえば、データブロックごとにデータ1、データ2、...、データ8を記憶する。

【0033】

メタ領域251は、各データブロックのBCCを記憶する。メタ領域251は、たとえば、データ1に対応するBCC1、データ2に対応するBCC2、...、データ8に対応するBCC8を記憶する。

10

【0034】

このように、HDD25において、ユーザデータと、ユーザデータのBCCとは、それぞれアクセス単位の異なる記憶領域に記憶される。そのため、ユーザデータと、ユーザデータのBCCとが同一の記憶領域に記憶されることによる、ユーザデータと、ユーザデータのBCCとが同時にデータ書込失敗となる危険を低減する。

【0035】

次に、第2の実施形態のデータブロックについて図5を用いて説明する。図5は、第2の実施形態のデータブロックの一例を示す図である。

データブロックは、512byteのサイズであり、この512byteごとに8byteのBCCが生成される。データブロックは、一意に特定可能なLBAが付される。たとえば、ユーザデータ領域252は、LBA#1、LBA#2、LBA#3、LBA#4、...を含み、ユーザデータを記憶する。また、メタ領域251は、LBA#mを含み、BCCを記憶する。なお、HDD25, 26は、256byte、64byteなど、アクセス単位(512byte)より小さな単位で、データブロック内のデータを区分可能にする。

20

【0036】

次に、第2の実施形態のCA13のハードウェア構成について図6を用いて説明する。図6は、第2の実施形態のCAのハードウェア構成の一例である。CA13とCA14は、同様の構成であるため、CA13について説明する。

【0037】

CA13は、プロセッサ101によって装置全体が制御されている。プロセッサ101には、バス109を介してRAM(Random Access Memory)102と複数の周辺機器が接続されている。プロセッサ101は、マルチプロセッサであってもよい。プロセッサ101は、たとえばCPU(Central Processing Unit)、MPU(Micro Processing Unit)、DSP(Digital Signal Processor)、ASIC(Application Specific Integrated Circuit)、またはPLD(Programmable Logic Device)である。またプロセッサ101は、CPU、MPU、DSP、ASIC、PLDのうちの2以上の要素の組み合わせであってもよい。

30

【0038】

RAM102は、CA13の主記憶装置として使用される。RAM102には、プロセッサ101に実行させるOSのプログラムやファームウェア、アプリケーションプログラム(制御プログラムなど)の少なくとも一部が一時的に格納される。また、RAM102には、プロセッサ101による処理に必要な各種データが格納される。また、RAM102は、各種データの格納に用いるメモリと別体にキャッシュメモリを含むものであってもよい。

40

【0039】

バス109に接続されている周辺機器としては、不揮発性メモリ103、入出力インタフェース104、機器接続インタフェース105、通信インタフェース106、およびDMA(Direct Memory Access)107がある。

【0040】

50

不揮発性メモリ 103 は、CA 13 の電源遮断時においても記憶内容を保持する。不揮発性メモリ 103 は、たとえば、EEPROM (Electrically Erasable Programmable Read-Only Memory) や SSD などである。また、不揮発性メモリ 103 は、CA 13 の補助記憶装置として使用される。不揮発性メモリ 103 には、オペレーティングシステムのプログラムやファームウェア、アプリケーションプログラム、および各種データが格納される。

【0041】

入出力インタフェース 104 は、入出力装置 110 と接続して入出力をおこなう。機器接続インタフェース 105 は、光学ドライブ装置 111 やメモリ装置 112 と接続する。通信インタフェース 106 は、ホスト 11 と接続する。DMA 107 は、キャッシュメモリ 19 との間でデータ転送をおこなう。

10

【0042】

CA 13 は、たとえばコンピュータ読み取り可能な記録媒体に記録されたプログラムを実行することにより、第 2 の実施形態の処理機能を実現する。CA 13 に実行させる処理内容を記述したプログラムは、様々な記録媒体に記録しておくことができる。たとえば、CA 13 に実行させるプログラムを不揮発性メモリ 103 に格納しておくことができる。プロセッサ 101 は、不揮発性メモリ 103 内のプログラムの少なくとも一部を RAM 102 にロードし、プログラムを実行する。また CA 13 に実行させるプログラムを、図示しない光ディスク、光磁気記録媒体、メモリカードなどの可搬型記録媒体に記録しておくこともできる。

20

【0043】

光学ドライブ装置 111 は、レーザ光などを利用して、光ディスクに記録されたデータの読み取りをおこなう。光ディスクは、光の反射によって読み取り可能なようにデータが記録された可搬型の記録媒体である。光ディスクには、DVD (Digital Versatile Disc)、DVD-RAM、CD-ROM (Compact Disc Read Only Memory)、CD-R (Recordable) / RW (ReWritable) などがある。光磁気記録媒体には、MO (Magneto-Optical disk) などがある。

【0044】

メモリ装置 112 は、機器接続インタフェース 105 との通信機能を搭載した記録媒体である。また、メモリ装置 112 は、メモリカードへのデータの書き込み、またはメモリカードからのデータの読出しをおこなうメモリリーダーライタに代えてもよい。メモリカードは、カード型の記録媒体である。

30

【0045】

可搬型記録媒体に格納されたプログラムは、たとえばプロセッサ 101 からの制御により、不揮発性メモリ 103 にインストールされた後、実行可能となる。またプロセッサ 101 が、可搬型記録媒体から直接プログラムを読み出して実行することもできる。

【0046】

以上のようなハードウェア構成によって、第 2 の実施形態の CA 13 の処理機能を実現することができる。なお、CA 14、CM 17、18、第 1 の実施形態に示したストレージ制御装置 1 も、図示した CA 13 と同様のハードウェアにより実現することができる。

40

【0047】

次に、第 2 の実施形態のライト要求受付処理について図 7 を用いて説明する。図 7 は、第 2 の実施形態のライト要求受付処理のフローチャートを示す図である。ライト要求受付処理は、CA 13、14 がホスト 11 からライト要求を受け付けて実行する処理である。以下、CA 13 がライト要求受付処理を実行する場合について説明するが、CA 14 についても同様である。

【0048】

[ステップ S 11] CA 13 は、ホスト 11 から書込対象データを受信する。

[ステップ S 12] CA 13 は、書込対象データについてデータブロックごとの BCC を生成する。

50

【 0 0 4 9 】

[ステップ S 1 3] C A 1 3 は、キャッシュメモリ 1 9 のユーザデータ領域 1 9 2 に書込対象データを書き込む。

[ステップ S 1 4] C A 1 3 は、キャッシュメモリ 1 9 の B C C 領域 1 9 1 に、生成した B C C を書き込む。

【 0 0 5 0 】

[ステップ S 1 5] C A 1 3 は、ホスト 1 1 に確認応答を送信し、ライト要求受付処理を終了する。

次に、第 2 の実施形態のライトバック処理について図 8 を用いて説明する。図 8 は、第 2 の実施形態のライトバック処理のフローチャートを示す図である。ライトバック処理は、C A 1 3 , 1 4 がライト要求受付処理を終了した後に、C M 1 7 , 1 8 が実行する処理である。以下、C M 1 7 がライトバック処理を実行する場合について説明するが、C M 1 8 についても同様である。

10

【 0 0 5 1 】

[ステップ S 2 1] C M 1 7 は、キャッシュメモリ 1 9 のユーザデータ領域 1 9 2 から書込対象データを読み出す。

[ステップ S 2 2] C M 1 7 は、キャッシュメモリ 1 9 の B C C 領域 1 9 1 から B C C を読み出す。

【 0 0 5 2 】

[ステップ S 2 3] C M 1 7 は、H D D 2 5 のユーザデータ領域 2 5 2 に書込対象データを書き込む。

20

[ステップ S 2 4] C M 1 7 は、H D D 2 5 のメタ領域 2 5 1 に B C C を書き込む。

【 0 0 5 3 】

[ステップ S 2 5] C M 1 7 は、H D D 2 5 のユーザデータ領域 2 5 2 からデータブロックごとに書込対象データを読み出して B C C を生成し、H D D 2 5 のメタ領域 2 5 1 から読み出した B C C と比較する。C M 1 7 は、生成した B C C と、読み出した B C C とが一致するか否かを判定する。C M 1 7 は、生成した B C C と、読み出した B C C とが一致しない場合にステップ S 2 1 にすすみ、リトライ動作をおこなう。一方、C M 1 7 は、生成した B C C と、読み出した B C C とが一致する場合にライトバック処理を終了する。

30

【 0 0 5 4 】

このようなストレージサーバ 1 0 は、ステップ S 2 3 における H D D 2 5 へのデータの書き込みに失敗した場合があっても、書込対象データと B C C とは異なる領域に記憶されていることから、書込対象データの書込の失敗（ブロックライト失敗）を看過することがない。

【 0 0 5 5 】

ここで、第 2 の実施形態のブロックライト失敗について図 9 を用いて説明する。図 9 は、第 2 の実施形態のブロックライト失敗の一例を示す図である。

ストレージサーバ 1 0 (C A 1 3) は、ホスト 1 1 から、D A T A 1、D A T A 2、D A T A 3、および D A T A 4 を書込対象データとするライト要求を受け付ける。

【 0 0 5 6 】

40

ストレージサーバ 1 0 (C A 1 3) は、D A T A 1、D A T A 2、D A T A 3、および D A T A 4 をデータブロックサイズに整える。たとえば、ストレージサーバ 1 0 (C A 1 3) は、データブロックサイズに満たない D A T A 4 についてパディングをおこなう。ストレージサーバ 1 0 (C A 1 3) は、D A T A 1 から B C C 1 を生成し、D A T A 2 から B C C 2 を生成し、D A T A 3 から B C C 3 を生成し、パディングされた D A T A 4 から B C C 4 を生成する。ストレージサーバ 1 0 (C A 1 3) は、D A T A 1、D A T A 2、D A T A 3、および D A T A 4 をキャッシュメモリ 1 9 のユーザデータ領域 1 9 2 に書き込む。ストレージサーバ 1 0 (C A 1 3) は、B C C 1、B C C 2、B C C 3、および B C C 4 をキャッシュメモリ 1 9 の B C C 領域 1 9 1 に書き込む。

【 0 0 5 7 】

50

ストレージサーバ10 (CM17) は、キャッシュメモリ19のユーザデータ領域192からDATA1、DATA2、DATA3、およびDATA4を読み出す。ストレージサーバ10 (CM17) は、キャッシュメモリ19のBCC領域191からBCC1、BCC2、BCC3、およびBCC4を読み出す。

【0058】

このとき、HDD25のユーザデータ領域252は、LBA#1にDATA01を記憶し、LBA#2にDATA02を記憶し、LBA#3にDATA03を記憶し、LBA#4にDATA04を記憶する。また、HDD25のメタ領域251は、DATA01に対応するBCC01、DATA02に対応するBCC02、DATA03に対応するBCC03、およびDATA04に対応するBCC04を記憶する。

10

【0059】

ストレージサーバ10 (CM17) は、元のデータを上書きしてHDD25のユーザデータ領域252を更新する。ここで、ストレージサーバ10 (CM17) は、LBA#2でDATA2の書き込みに失敗(ブロックライト失敗)したとする。これにより、LBA#1は、DATA1を記憶し、LBA#2は、DATA02を記憶し、LBA#3は、DATA3を記憶し、LBA#4は、DATA4を記憶する。すなわち、LBA#2は、ブロックライト失敗によりDATA02を記憶したままである。

【0060】

また、ストレージサーバ10 (CM17) は、元のデータを上書きしてHDDのメタ領域251を更新する。これにより、HDD25のメタ領域251は、BCC1、BCC2、BCC3、およびBCC4を記憶する。これにより、LBA#2のBCCは、HDD25のメタ領域251にBCC2が記憶され、LBA#2が記憶するDATA02から生成するBCC02と一致しない。

20

【0061】

したがって、ストレージサーバ10 (CM17) は、HDD25のユーザデータ領域252におけるブロックライト失敗を検出できる。したがって、ストレージサーバ10は、データ書込失敗の検出精度を向上することができる。

【0062】

ここで、ブロックライト失敗の比較例について図10を用いて説明する。図10は、ブロックライト失敗の比較例を示す図である(その1)。

30

ストレージサーバは、ホストから、DATA1、DATA2、DATA3、およびDATA4を書込対象データとするライト要求を受け付ける。

【0063】

ストレージサーバは、DATA1、DATA2、DATA3、およびDATA4をデータブロックサイズに整える。ストレージサーバは、DATA1からBCC1を生成し、DATA2からBCC2を生成し、DATA3からBCC3を生成し、パディングされたDATA4からBCC4を生成する。ストレージサーバは、DATA1とBCC1、DATA2とBCC2、DATA3とBCC3、およびパディングされたDATA4とBCC4をキャッシュメモリに書き込む。

【0064】

40

ストレージサーバは、キャッシュメモリからDATA1とBCC1、DATA2とBCC2、DATA3とBCC3、およびパディングされたDATA4とBCC4を読み出す。

【0065】

このとき、HDDは、LBA#1にDATA01とBCC01を記憶し、LBA#2にDATA02とBCC02を記憶し、LBA#3にDATA03とBCC03を記憶し、LBA#4にパディングされたDATA04とBCC04を記憶する。

【0066】

ストレージサーバは、元のデータを上書きしてHDDを更新する。ここで、ストレージサーバは、LBA#2でDATA2とBCC2の書き込みに失敗(ブロックライト失敗)

50

したとする。これにより、L B A # 1 は、D A T A 1 と B C C 1 を記憶し、L B A # 2 は、D A T A 0 2 と B C C 0 2 を記憶し、L B A # 3 は、D A T A 3 と B C C 3 を記憶し、L B A # 4 は、パディングされた D A T A 4 と B C C 4 を記憶する。すなわち、L B A # 2 は、ブロックライト失敗により D A T A 0 2 と B C C 0 2 を記憶したままである。

【 0 0 6 7 】

このような、L B A # 2 の B C C は、L B A # 2 が記憶する D A T A 0 2 から生成する B C C 0 2 と、L B A # 2 が記憶する B C C 0 2 とが一致する。

すなわち、ストレージサーバは、H D D のブロックライト失敗を検出できない。このような、ストレージサーバは、データ書込失敗の検出精度が不十分である。一方、第 2 の実施形態のストレージサーバ 1 0 は、図 9 を用いて説明したように、H D D 2 5 のユーザデータ領域 2 5 2 におけるブロックライト失敗を検出できることから、データ書込失敗の検出精度が比較例と比較して高い。

10

【 0 0 6 8 】

次に、第 2 の実施形態のリード要求受付処理について図 1 1 を用いて説明する。図 1 1 は、第 2 の実施形態のリード要求受付処理のフローチャートを示す図である。リード要求受付処理は、C A 1 3 , 1 4 がホスト 1 1 からリード要求を受け付けて実行する処理である。以下、C A 1 3 がリード要求受付処理を実行する場合について説明するが、C A 1 4 についても同様である。

【 0 0 6 9 】

[ステップ S 3 1] C M 1 7 は、H D D 2 5 のユーザデータ領域 2 5 2 から読出対象データを読み出して、キャッシュメモリ 1 9 のユーザデータ領域 1 9 2 に書き込む（ステージング）。また、C M 1 7 は、H D D 2 5 のメタ領域 2 5 1 から読出対象データの B C C を読み出して、キャッシュメモリ 1 9 の B C C 領域 1 9 1 に書き込む。

20

【 0 0 7 0 】

[ステップ S 3 2] C A 1 3 は、キャッシュメモリ 1 9 のユーザデータ領域 1 9 2 から読出対象データを読み出して、データブロックごとの B C C を生成する。

[ステップ S 3 3] C A 1 3 は、キャッシュメモリ 1 9 の B C C 領域 1 9 1 から B C C を読み出す。

【 0 0 7 1 】

[ステップ S 3 4] C A 1 3 は、ステップ S 3 2 で生成した B C C と、ステップ S 3 3 で読み出した B C C とが一致するか否かを判定する。C A 1 3 は、生成した B C C と読み出した B C C とが一致する場合にステップ S 3 5 にすすみ、一致しない場合にステップ S 3 6 にすすむ。

30

【 0 0 7 2 】

[ステップ S 3 5] C A 1 3 は、ホストにリード応答を送信して、リード要求受付処理を終了する。

[ステップ S 3 6] C A 1 3 は、ホストにエラー応答を送信して、リード要求受付処理を終了する。

【 0 0 7 3 】

このように、ストレージサーバ 1 0 は、H D D 2 5 のユーザデータ領域 2 5 2 におけるブロックライト失敗を検出できる。したがって、ストレージサーバ 1 0 は、データ書込失敗の検出精度を向上することができる。

40

【 0 0 7 4 】

再び、ブロックライト失敗の比較例について図 1 2 を用いて説明する。図 1 2 は、ブロックライト失敗の比較例を示す図である（その 2 ）。

ストレージサーバ 1 0 は、ホストから、D A T A 1 、D A T A 2 、D A T A 3 、および D A T A 4 を読出対象データとするリード要求を受け付ける。このとき、H D D は、L B A # 2 のブロックライト失敗により、L B A # 1 に D A T A 1 と B C C 1 を記憶し、L B A # 2 に D A T A 0 2 と B C C 0 2 を記憶し、L B A # 3 に D A T A 3 と B C C 3 を記憶し、L B A # 4 に D A T A 4 と B C C 4 を記憶する。

50

【 0 0 7 5 】

ストレージサーバは、HDDからDATA 1とBCC 1、DATA 0 2とBCC 0 2、DATA 3とBCC 3、およびパディングされたDATA 4とBCC 4をキャッシュメモリに読み出す。

【 0 0 7 6 】

このような、ストレージサーバは、キャッシュメモリが記憶するDATA 0 2から生成するBCC 0 2と、キャッシュメモリが記憶するBCC 0 2とが一致する。すなわち、ストレージサーバは、HDDのブロックライト失敗を検出できない。

【 0 0 7 7 】

そのため、ストレージサーバは、キャッシュメモリから読み出したDATA 1、DATA 0 2、DATA 3、およびDATA 4をホストに回答する。ホストは、DATA 2がDATA 0 2に置き換わっていることにより、データ化けを検出することとなる。

【 0 0 7 8 】

一方、第2の実施形態のストレージサーバ10は、図11を用いて説明したように、HDD 25のユーザデータ領域252におけるブロックライト失敗を、リード要求応答時においても検出できることから、データ書込失敗の検出精度が比較例と比較して高い。また、ストレージサーバ10は、データ化けに対応するホストの処理負荷を軽減することができる。

【 0 0 7 9 】

[第3の実施形態]

次に、第3の実施形態のストレージ制御装置について図13を用いて説明する。図13は、第3の実施形態のストレージ制御装置の構成の一例を示す図である。

【 0 0 8 0 】

ノード# 1、ノード# 2、・・・、ノード# 8は、それぞれがストレージ制御装置の1形態であり、複数のノードで全体として性能向上を図るスケールアウトストレージを実現する。なお、各ノードは、第2の実施形態のストレージサーバ10と同様のハードウェア構成である。

【 0 0 8 1 】

各ノードは、それぞれキャッシュメモリと1または2以上のHDDを備える。ノード# 1は、キャッシュメモリ50とHDD 54, 55, 56を備える。ノード# 2は、図示しないキャッシュメモリとHDD 60, 61, 62を備える。ノード# 8は、図示しないキャッシュメモリとHDD 70, 71, 72を備える。なお、各ノードは、図示しないCAがホストインタフェース制御部として機能し、CAを介してホストと接続する。また、各ノードは、図示しないCMが接続制御部として機能し、CMを介してノード同士を接続する。

【 0 0 8 2 】

ノード# 1は、キャッシュメモリ50に、メタ情報領域51と、BCC領域52と、ユーザデータ領域53とを備える点で、第2の実施形態のストレージサーバ10と異なる。

メタ情報領域51は、HDDがメタ領域に記憶するメタ情報を記憶する領域である。ユーザデータ領域53は、ユーザデータをデータブロックごとに記憶する領域である。BCC領域52は、ユーザデータ領域53が記憶するユーザデータのデータブロックごとのBCCを記憶する領域である。

【 0 0 8 3 】

また、ノード# 1は、HDD 54のメタ領域にメタ情報を記憶し、ノード# 1からノード# 8は、データとBCCを分散してHDDに記憶する。メタ情報は、データ位置情報(書込対象データ格納先情報)と、BCC位置情報(誤り検出情報格納先情報)を含む。データ位置情報は、データの記憶位置を示し、たとえば、データ1がノード# 1のHDD 55の所定アドレスに記憶されていることを示す。BCC位置情報は、BCCの記憶位置を示し、たとえば、データ1からデータ8に対応するそれぞれのBCCがノード# 2のHDD 60の所定アドレスに記憶されていることを示す。

10

20

30

40

50

【 0 0 8 4 】

次に、第3の実施形態のライト要求受付処理について図14を用いて説明する。図14は、第3の実施形態のライト要求受付処理のフローチャートを示す図である。ライト要求受付処理は、ノードがホストからライト要求を受け付けて実行する処理である。以下、図13に示すノード#1がライト要求受付処理を実行する場合について説明するが、ノード#2からノード#7についても同様である。

【 0 0 8 5 】

[ステップS41] ノード#1は、ホストから書込対象データを受信する。

[ステップS42] ノード#1は、HDD54からメタ情報を読み出して、キャッシュメモリ50のメタ情報領域51に書き込む。これにより、ノード#1は、メタ情報領域51に、データ位置情報とBCC位置情報とを記憶する。

10

【 0 0 8 6 】

[ステップS43] ノード#1は、キャッシュメモリ50のユーザデータ領域53に、ホストから受信した書込対象データを書き込む。これにより、ノード#1は、ユーザデータ領域53に書込対象データ(データ1、データ2、・・・、データ8)を記憶する。

【 0 0 8 7 】

[ステップS44] ノード#1は、書込対象データ(データ1、データ2、・・・、データ8)からデータブロックごとのBCCを生成する。

[ステップS45] ノード#1は、キャッシュメモリ50のBCC領域52に、生成したBCCを書き込む。

20

【 0 0 8 8 】

[ステップS46] ノード#1は、ライト要求に対する確認応答をホストに送信する。

[ステップS47] ノード#1は、メタ情報のうちのデータアドレス(データ位置情報)を取得する。

【 0 0 8 9 】

[ステップS48] ノード#1は、データアドレスから特定される担当ノードにデータを送信する。たとえば、ノード#1は、ノード#2にデータ2を送信し、ノード#8にデータ8を送信する。データを受信した担当ノードは、データアドレスから特定されるHDDの所定アドレスにデータを書き込む。なお、ノード#1は、データ1の担当ノードであるので、ノード#1がHDD55にデータ1を書き込む。

30

【 0 0 9 0 】

[ステップS49] ノード#1は、メタ情報のうちのBCCアドレス(BCC位置情報)を取得する。

[ステップS50] ノード#1は、BCCアドレスから特定される担当ノードにBCCを送信する。たとえば、ノード#1は、ノード#2にBCCを送信する。BCCを受信した担当ノードは、BCCアドレスから特定されるHDDの所定アドレスにデータを書き込む。ノード#1は、担当ノードにBCCを送信した後、ライト要求受付処理を終了する。

【 0 0 9 1 】

このように、ノードは、ホストからライト要求を受け付けても、HDDのメタ領域について書き換えをおこなわない。したがって、ノードは、HDDからキャッシュメモリに読み出したメタ情報をメタ領域に書き戻すことを要しない。これにより、ノードは、メタ情報をメタ領域に書き戻すディスクアクセスを削減する。このようなノードは、データ書込失敗の検出精度を向上することができる他、ディスクアクセス性能の向上を図ることができる。

40

【 0 0 9 2 】

なお、プログラムを流通させる場合には、たとえば、そのプログラムが記録されたDVD、CD-ROMなどの可搬型記録媒体が販売される。また、プログラムをサーバコンピュータの記憶装置に格納しておき、ネットワークを介して、サーバコンピュータから他のコンピュータにそのプログラムを転送することもできる。

【 0 0 9 3 】

50

プログラムを実行するコンピュータは、たとえば、可搬型記録媒体に記録されたプログラムもしくはサーバコンピュータから転送されたプログラムを、自己の記憶装置に格納する。そして、コンピュータは、自己の記憶装置からプログラムを読み取り、プログラムにしたがった処理を実行する。なお、コンピュータは、可搬型記録媒体から直接プログラムを読み取り、そのプログラムにしたがった処理を実行することもできる。また、コンピュータは、ネットワークを介して接続されたサーバコンピュータからプログラムが転送されるごとに、逐次、受け取ったプログラムにしたがった処理を実行することもできる。

【0094】

また、上記の処理機能の少なくとも一部を、DSP、ASIC、PLDなどの電子回路で実現することもできる。

【符号の説明】

【0095】

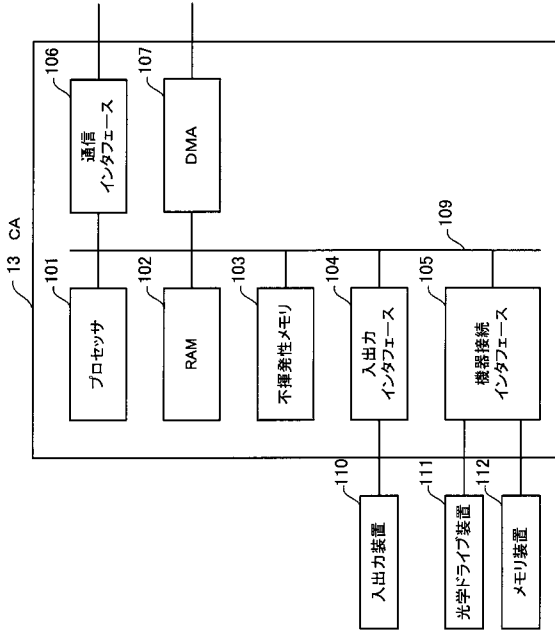
- 1 ストレージ制御装置
- 2 誤り検出情報格納部
- 3 データ格納部
- 4 検出部
- 5 記憶装置
- 6 第1の記憶領域
- 7 第2の記憶領域
- 10 ストレージサーバ
- 11, 12 ホスト
- 13, 14 CA
- 15, 16, 23, 24 スイッチ
- 17, 18 CM
- 19, 20, 50 キャッシュメモリ
- 21, 22 DA
- 25, 26, 54, 55, 56, 60, 61, 62, 70, 71, 72 HDD
- 101 プロセッサ
- 102 RAM
- 103 不揮発性メモリ
- 104 入出力インタフェース
- 105 機器接続インタフェース
- 106 通信インタフェース
- 107 DMA
- 109 バス

10

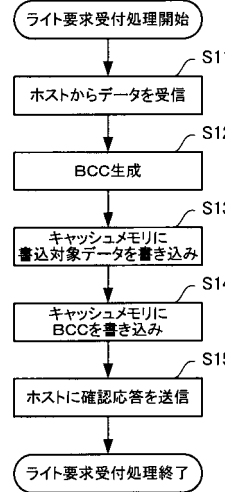
20

30

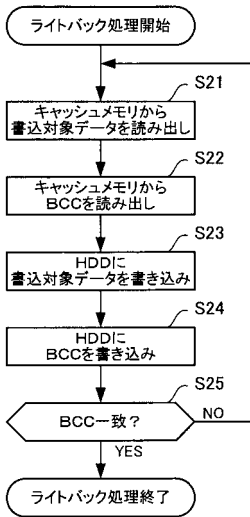
【図 6】



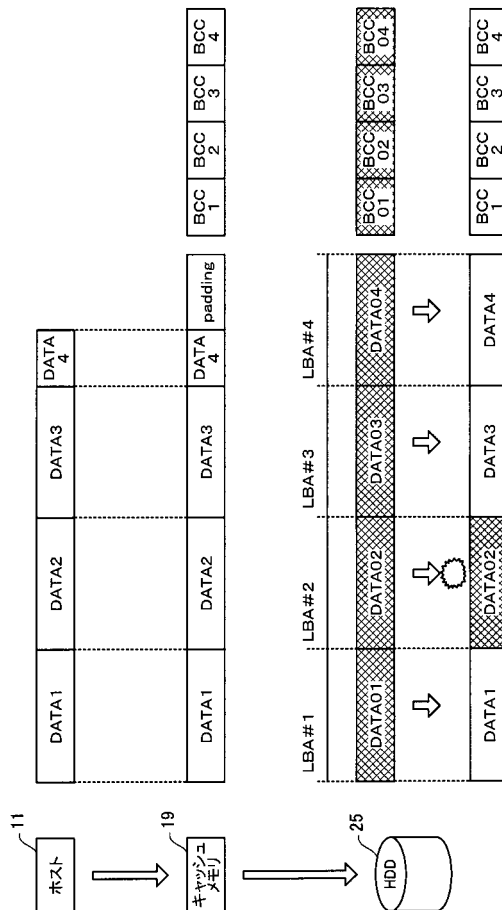
【図 7】



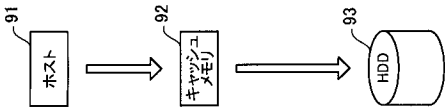
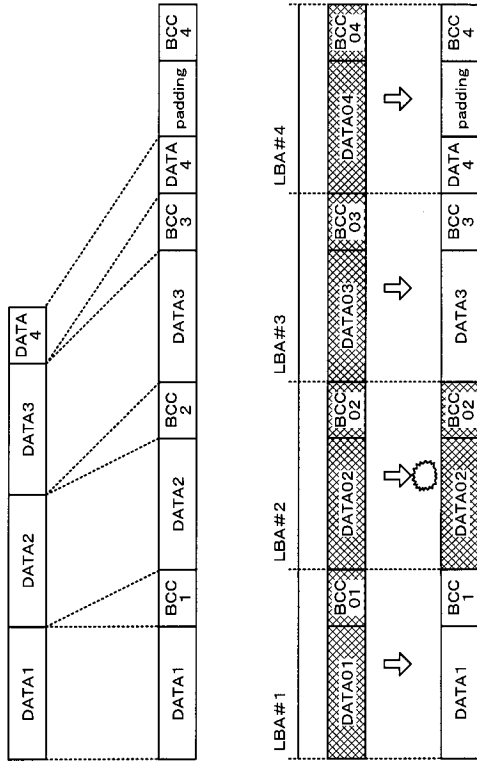
【図 8】



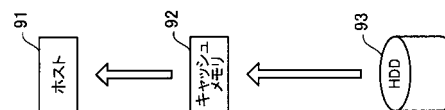
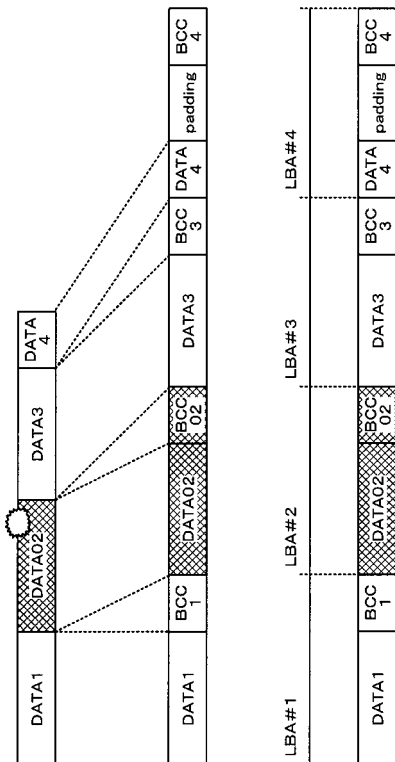
【図 9】



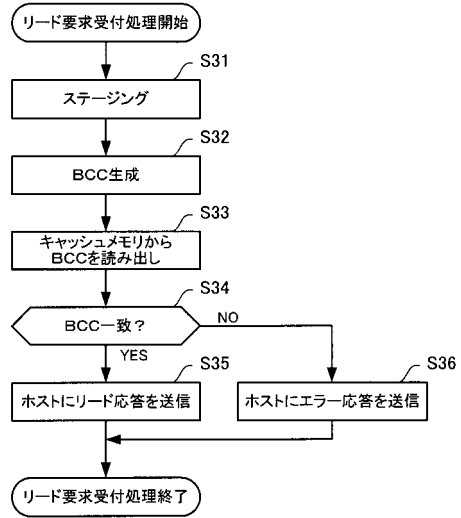
【図 10】



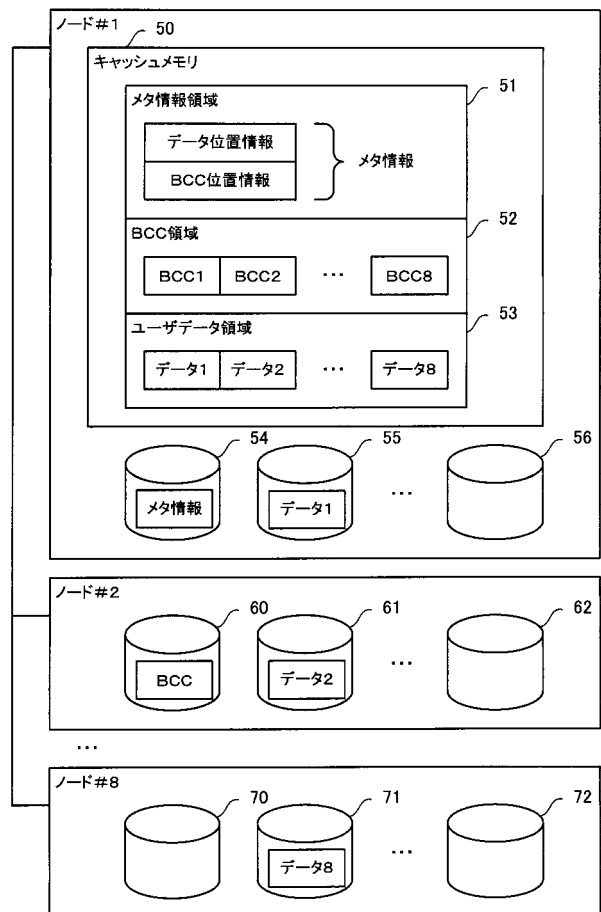
【図 12】



【図 11】



【図 13】



【 図 1 4 】

