



(51) 国際特許分類:

G06F 3/06 (2006.01)

(21) 国際出願番号:

PCT/JP2010/068561

(22) 国際出願日:

2010 年 10 月 21 日(21.10.2010)

(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): 櫻井 博 (SAKURAI, Hiroshi) [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: 酒井 昭徳 (SAKAI, Akinori); 〒1006020 東京都千代田区霞が関 3 丁目 2 番 5 号 霞が

関ビルディング 20 階 酒井総合特許事務所 Tokyo (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, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, 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, PE, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

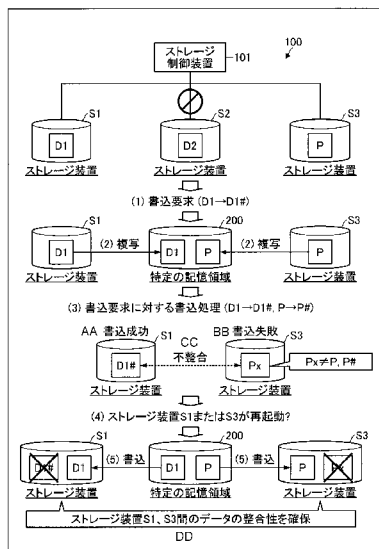
(84) 指定国 (表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), ヨーロッパ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI

[続葉有]

(54) Title: STORAGE CONTROL DEVICE AND STORAGE CONTROL METHOD

(54) 発明の名称: ストレージ制御装置およびストレージ制御方法

[図2]



101 Storage Control Device
 S1, S2, S3 Storage Device
 200 Specific Recording Region
 (1) Write Request (D1→D1#)
 (2) Duplication
 (3) Write Processing Responding to Write Request (D1→D1#, P→P#)
 (4) Storage Device (S1) or (S3) Restarted?
 (5) Write
 AA Write Success
 BB Write Failure
 CC Inconsistency
 DD Data Consistency Secured Between Storage Device (S1) and (S3)

(57) Abstract: This storage control device (101) receives a request to write to a storage device (S1). When a write request has been received, the storage control device (101) duplicates data (D1) recorded at a recording region of the storage device (S1) and parity data (P) that is generated based on the data (D1) to a specific recording region (200). The storage control device (101) determines whether or not storage device (S1) or (S3) to which the write processing responding to the write request has been executed has been restarted. When storage device (S1) or (S3) has been restarted, the storage control device (101) writes the data (D1) duplicated to the specific recording region (200) to the duplication-source recording region of storage device (S1). Furthermore, the storage control device (101) writes the parity data (P) duplicated to the specific recording region (200) to the duplication-source recording region of storage device (S3).

(57) 要約: ストレージ制御装置 (101) は、ストレージ装置 (S1) に対する書込要求を受け付ける。ストレージ制御装置 (101) は、書込要求を受け付けた場合、ストレージ装置 (S1) の記憶領域に記憶されているデータ (D1) と、データ (D1) を生成元とするパリティデータ (P) を、特定の記憶領域 (200) に複写する。ストレージ制御装置 (101) は、書込要求に対する書込処理が実行されたストレージ装置 (S1) または (S3) が再起動されたか否かを判断する。ストレージ制御装置 (101) は、ストレージ装置 (S1) または (S3) が再起動された場合、特定の記憶領域 (200) に複写されたデータ (D1) をストレージ装置 (S1) の複写元の記憶領域に書き込む。さらに、ストレージ制御装置 (101) は、特定の記憶領域 (200) に複写されたパリティデータ (P) をストレージ装置 (S3) の複写元の記憶領域に書き込む。

WO 2012/053085 A1



(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 (条約第 21 条(3))

明 細 書

発明の名称： ストレージ制御装置およびストレージ制御方法

技術分野

[0001] 本発明は、ストレージ装置を制御するストレージ制御装置およびストレージ制御方法に関する。

背景技術

[0002] RAID (Redundant Array of Inexpensive Disks) 3, 5等の冗長化されたストレージシステムは、ディスク装置に障害が発生すると、障害が発生したディスク装置を切り離す。ストレージシステムは、障害が発生したディスク装置が切り離されると、切り離されたディスク装置を除く残余のディスク装置によってデータの記録、更新を行う。

[0003] また、ストレージシステムは、障害が発生したディスク装置が切り離された後、さらに別のディスク装置に障害が発生すると、新たに障害が発生したディスク装置を切り離す。以下の説明では、ストレージシステム内のディスク装置に障害が発生して冗長化を失った状態から、さらに別のディスク装置を切り離した状態を「マルチデッド状態」という。

[0004] ここで、ディスク装置の切り離し要因となる障害としては、例えば、サーマル・オフトラック、コンタミネーション、ノイズ、接触不良等がある。ディスク装置のサーマル・オフトラック、コンタミネーション、ノイズ、接触不良等の障害は、ハードリセットや電源切断後の再投入（電源のOFF/ON）によって回復することが多い。

[0005] すなわち、障害が発生して切り離されたディスク装置に対してハードリセットや電源のOFF/ONを行うことで、該ディスク装置を正常に動作するディスク装置として復帰させることができる。このため、マルチデッド状態となったストレージシステムを復旧する場合、例えば、ストレージシステムのハードリセットを行って、マルチデッド状態となる前の状態に戻すことが

行われる。

- [0006] なお、ディスク装置に障害が発生したストレージシステムを復旧するための技術を開示する文献として、例えば、下記特許文献 1～3 がある。

先行技術文献

特許文献

- [0007] 特許文献1：特開平 1 1－9 5 9 3 3 号公報
特許文献2：特開 2 0 0 5－7 8 4 3 0 号公報
特許文献3：特開 2 0 1 0－2 6 8 1 2 号公報

発明の概要

発明が解決しようとする課題

- [0008] しかしながら、上述した従来技術では、ストレージシステムの復旧時に、ハードリセット等を行ってストレージシステムをマルチデッド状態となる前の状態に戻しても、ディスク装置間のデータの不整合によりデータ化けが発生する場合があるという問題があった。
- [0009] 本発明は、上述した従来技術による問題点を解消するため、ストレージ装置間のデータの整合性を保障することができるストレージ制御装置およびストレージ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

- [0010] 上述した課題を解決し、目的を達成するため、開示のストレージ制御装置およびストレージ制御方法は、データ群ごとに、前記データ群の各データと前記データ群から生成される訂正符号とを、記憶先のストレージ装置が異なるように記憶するシステム内のストレージ装置に対する書込要求を受け付け、前記書込要求を受け付けた場合、前記ストレージ装置の要求先の記憶領域に記憶されているデータと、当該データを生成元とする、前記ストレージ装置に記憶されている訂正符号とを、前記ストレージ装置から特定の記憶領域に複写し、前記データと前記訂正符号が前記特定の記憶領域に複写された後、前記書込要求に対する書込処理が実行されたストレージ装置が再起動され

たか否かを判断し、ストレージ装置が再起動されたと判断された場合、前記特定の記憶領域に複写されたデータを前記ストレージ装置の要求先の記憶領域に書き込むとともに、前記特定の記憶領域に複写された訂正符号を当該訂正符号の複写元のストレージ装置の記憶領域に書き込む。

[0011] 上述した課題を解決し、目的を達成するため、開示のストレージ制御装置およびストレージ制御方法は、データ群ごとに、前記データ群の各データと前記データ群から生成される訂正符号とを、記憶先のストレージ装置が異なるように記憶するシステム内のストレージ装置に対する書込要求を受け付け、前記書込要求を受け付けた場合、前記システム内のストレージ装置にアクセスする制御部に対して、前記ストレージ装置の要求先の記憶領域に記憶されているデータと、当該データを生成元とする、前記ストレージ装置に記憶されている訂正符号とを、特定の記憶領域に複写するよう指示し、指示された前記データと前記訂正符号が前記特定の記憶領域に複写された後、前記書込要求に対する書込処理が実行されたストレージ装置が再起動されたか否かを判断し、ストレージ装置が再起動されたと判断された場合、前記制御部に対して、前記特定の記憶領域に複写されたデータを前記ストレージ装置の要求先の記憶領域に書き込むとともに、前記特定の記憶領域に複写された訂正符号を当該訂正符号の複写元のストレージ装置の記憶領域に書き込むよう指示する。

発明の効果

[0012] 本ストレージ制御装置およびストレージ制御方法によれば、ストレージ装置間のデータの整合性を保障することができるという効果を奏する。

図面の簡単な説明

[0013] [図1]実施の形態1にかかるストレージシステム100の一実施例を示す説明図である。

[図2]実施の形態1にかかるストレージ制御手法の一実施例を示す説明図である。

[図3]実施の形態1にかかるストレージシステム300のハードウェア構成を

示すブロック図である。

[図4]管理データ410, 420, 430の具体例を示す説明図である。

[図5]実施の形態1にかかるストレージ制御装置101の機能的構成を示すブロック図である。

[図6]書込要求600の具体例を示す説明図である。

[図7]書込状態テーブル700の具体例を示す説明図である。

[図8]書込バッファBの具体例を示す説明図である。

[図9]管理データ410, 420, 430の更新例を示す説明図(その1)である。

[図10]検出結果テーブル1000の具体例を示す説明図である。

[図11]HDD1~HDD3のLBA1の記憶内容の変遷例を示す説明図(その1)である。

[図12]HDD1~HDD3のLBA1の記憶内容の変遷例を示す説明図(その2)である。

[図13]HDD1~HDD3のLBA1の記憶内容の変遷例を示す説明図(その3)である。

[図14]HDD1~HDD3のLBA1の記憶内容の変遷例を示す説明図(その4)である。

[図15]HDD1~HDD3のLBA1の記憶内容の変遷例を示す説明図(その5)である。

[図16]実施の形態1にかかるストレージ制御装置のデータ複写処理手順の一例を示すフローチャート(その1)である。

[図17]実施の形態1にかかるストレージ制御装置のデータ複写処理手順の一例を示すフローチャート(その2)である。

[図18]実施の形態1にかかるストレージ制御装置101のシステム復旧処理手順の一例を示すフローチャートである。

[図19]ステップS1803のデータ書込処理の具体的な処理手順の一例を示すフローチャートである。

[図20]実施の形態2にかかるストレージシステム2000のハードウェア構成を示すブロック図である。

[図21]書込状態テーブル2100の具体例を示す説明図である。

[図22]冗長構成時のストレージ制御装置101のストレージ制御処理手順の一例を示すフローチャート（その1）である。

[図23]冗長構成時のストレージ制御装置101のストレージ制御処理手順の一例を示すフローチャート（その2）である。

[図24]冗長構成時のストレージ制御装置101のストレージ制御処理手順の一例を示すフローチャート（その3）である。

[図25]管理データ410, 420, 430の更新例を示す説明図（その2）である。

[図26]非冗長構成時のストレージ制御装置101のストレージ制御処理手順の一例を示すフローチャート（その1）である。

[図27]非冗長構成時のストレージ制御装置101のストレージ制御処理手順の一例を示すフローチャート（その2）である。

[図28]管理データ410, 420, 430の更新例を示す説明図（その3）である。

[図29]ストレージ制御装置101のシステム復旧処理手順の一例を示すフローチャートである。

発明を実施するための形態

[0014] 以下に添付図面を参照して、本発明にかかるストレージ制御装置およびストレージ制御方法の好適な実施の形態を詳細に説明する。

[0015] （実施の形態1）

図1は、実施の形態1にかかるストレージシステム100の一実施例を示す説明図である。図1において、ストレージシステム100は、ストレージ制御装置101とストレージ装置S1～S3と、を含む構成である。ストレージ制御装置101は、ストレージ装置S1～S3を制御するコンピュータである。ストレージ装置S1～S3は、データを記憶する記憶装置である。

ストレージ装置 S 1 ～ S 3 は、例えば、ハードディスク、光ディスク、フラッシュメモリ、磁気テープ等の記憶媒体を含む。

[0016] ここで、ストレージシステム 100 は、RAID 3, 5 等の冗長化されたシステムである。ストレージシステム 100 は、データ群ごとに、データ群の各データと、該データ群の各データから生成される訂正符号とを、記憶先のストレージ装置が異なるように分散して記憶している。データ群は、互いに関連するデータの集合である。各データは、例えば、ビット単位、バイト単位、ブロック（セクタ）単位のデータである。

[0017] 訂正符号は、データ群を構成するデータを生成元とする冗長コードである。訂正符号としては、生成元のデータ群のいずれかのデータを、データ群の残余のデータと訂正符号とから復元可能なものを用いることができる。具体的には、例えば、訂正符号は、パリティ、ハミング符号、ECC（Error Correcting Code）、CRC（Cyclic Redundancy Check）コード等を用いて実現される。以下の説明では、訂正符号の一例として、生成元のデータ群を構成する各データの排他的論理和をとったパリティデータを例に挙げて説明する。

[0018] 図 1 の例では、データ D 1 と、データ D 2 と、データ D 1, D 2 から生成されるパリティデータ P とが、記憶先のストレージ装置が異なるように分散して記憶されている。パリティデータ P は、データ D 1 とデータ D 2 の排他的論理和をとったデータである。以下の説明では、例えば、データ D 1 とデータ D 2 の排他的論理和をとったパリティデータ P を下記式（1）のように表現する。ただし、xor は、排他的論理和を表す。

[0019]
$$P = D1 \text{ xor } D2 \quad \cdots (1)$$

[0020] ストレージシステム 100 では、障害が発生したストレージ装置を切り離す。ストレージ装置を切り離すとは、ストレージ装置に対するアクセスを停止することである。この結果、ストレージシステム 100 は、冗長構成から非冗長構成となる。冗長構成とは、データの冗長性が確保されている状態である。非冗長構成とは、冗長構成から、障害が発生したストレージ装置が切

り離されて、データの冗長性を失った状態である。

[0021] ストレージシステム 100 は、冗長構成から非冗長構成となっても、障害が発生して切り離されたストレージ装置に記憶されているデータ（欠損データ）を、パリティデータと他のストレージ装置に記憶されているデータとを利用して復元できる。このため、ストレージシステム 100 は、冗長構成から非冗長構成となっても、システムの運用を継続して行うことができる。

[0022] 例えば、ストレージ装置 S2 に障害が発生すると、ストレージシステム 100 は冗長構成から非冗長構成となり、ストレージ装置 S2 からデータ D2 を読み出せなくなる。この場合、ストレージシステム 100 は、下記式（2）のように、ストレージ装置 S1 に記憶されているデータ D1 と、ストレージ装置 S3 に記憶されているパリティデータ P の排他的論理和をとってデータ D2 を復元する。

$$\begin{aligned} [0023] \quad D2 &= D1 \quad \text{xor} \quad P = D1 \quad \text{xor} \quad D1 \quad \text{xor} \quad D2 \quad \cdots \\ &\cdot (2) \end{aligned}$$

[0024] また、ストレージシステム 100 は、非冗長構成から、さらにストレージ装置が切り離された状態（マルチデッド状態）となった場合、欠損データを復元できなくなるためシステムダウンする。システムダウンとは、データの読出処理および書込処理を正常に行うことができない状態である。

[0025] ここで、データの書込処理の一例について説明する。ここでは、ストレージ装置 S2 が切り離された非冗長構成のストレージシステム 100 において、ストレージ装置 S1 に記憶されているデータ D1 をデータ D1 # に更新する場合について説明する。

[0026] 具体的には、例えば、まず、ストレージ制御装置 101 が、ストレージ装置 S1 からデータ D1 を読み出す。つぎに、ストレージ制御装置 101 が、ストレージ装置 S3 からパリティデータ P を読み出す。そして、ストレージ制御装置 101 が、各ストレージ装置から読み出したデータ D1 とパリティデータ P の排他的論理和をとってデータ D2 を算出する。これにより、切り離されたストレージ装置 S2 に記憶されているデータ D2 が復元される。

- [0027] つぎに、ストレージ制御装置 101 が、書込対象のデータ D1 # と、算出されたデータ D2 の排他的論理和をとって、新たなパリティデータ P # を算出する ($P \# = D1 \# \text{ xor } D2$)。そして、ストレージ制御装置 101 が、データ D1 # をストレージ装置 S1 に書き込む。また、ストレージ制御装置 101 が、パリティデータ P # をストレージ装置 S3 に書き込む。
- [0028] このように、ストレージシステム 100 では、データ D1 をデータ D1 # に更新する場合、ストレージ装置間のデータの整合性を確保して欠損データ（上述の例ではデータ D2）を復元し、パリティデータの更新も合わせて行う。
- [0029] ここで、ストレージ装置に対するデータの書き込みは、サーマル・オフトラック、コンタミネーション、ノイズ、接触不良等の障害により失敗する場合がある。上述したデータ D1 をデータ D1 # に更新する場合の書き込み失敗例としては、例えば、下記 (a) ~ (c) がある。
- [0030] (a) ストレージ装置 S1 に対するデータ D1 # の書き込みに失敗して、データ D1 # の一部をストレージ装置 S1 に書き込めなかった場合
- [0031] この場合、ストレージ装置 S1 に記憶されているデータは、データ D1 およびデータ D1 # のいずれとも異なるデータ D_x となる ($D_x \neq D1, D1 \#$)。また、ストレージ装置 S1 に対するデータ D1 # の書き込みに失敗すると、ストレージシステム 100 からストレージ装置 S1 が切り離される。この結果、ストレージシステム 100 は、非冗長構成からマルチデッド状態となりシステムダウンする。このため、ストレージ装置 S3 に記憶されているデータは、パリティデータ P のままである ($P = D1 \text{ xor } D2$)。
- [0032] (b) ストレージ装置 S1 に対するデータ D1 # の書き込みに成功した後、ストレージ装置 S3 に対するパリティデータ P # の書き込みに失敗して、パリティデータ P # の一部をストレージ装置 S3 に書き込めなかった場合
- [0033] この場合、ストレージ装置 S1 に記憶されているデータは、データ D1 # となる。一方、ストレージ装置 S3 に記憶されているデータは、パリティデータ P およびパリティデータ P # のいずれとも異なるデータ P_x となる (P

x ≠ P, P #)。また、ストレージ装置 S 3 に対するパリティデータ P #の書き込みに失敗すると、ストレージ装置 S 3 がストレージシステム 100 から切り離される。この結果、ストレージシステム 100 は、非冗長構成からマルチデッド状態となりシステムダウンする。

[0034] (c) ストレージ装置 S 1 に対するデータ D 1 #の書き込みに成功した後、ストレージ装置 S 3 に対するパリティデータ P #の書き込みに失敗して、パリティデータ P #をストレージ装置 S 3 にまったく書き込めなかった場合

[0035] この場合、ストレージ装置 S 1 に記憶されているデータは、データ D 1 #となる。一方、ストレージ装置 S 3 に記憶されているデータは、パリティデータ P のままである。また、ストレージ装置 S 3 に対するパリティデータ P #の書き込みに失敗すると、ストレージ装置 S 3 がストレージシステム 100 から切り離される。この結果、ストレージシステム 100 は、非冗長構成からマルチデッド状態となりシステムダウンする。

[0036] 上述したように、上記 (a) ~ (c) のいずれの場合も、ストレージシステム 100 は、書込失敗により非冗長構成からマルチデッド状態となりシステムダウンする。ここで、サーマル・オフトラック、コンタミネーション、ノイズ、接触不良等のストレージ装置の障害は、該ストレージ装置に対するハードリセットや電源の OFF / ON を行うことによって回復することが多い。

[0037] このため、ストレージシステム 100 では、書き込みに失敗して切り離されたストレージ装置 (ストレージ装置 S 1 または S 3) に対するハードリセットや電源の OFF / ON を行って、システムダウン直前の非冗長構成に戻すことができる。しかし、上記 (a) ~ (c) のいずれの場合も、ストレージシステム 100 をシステムダウン直前の非冗長構成に戻したとしても、ストレージ装置 S 1, S 3 間のデータの整合性がとれていない状態となっている。

[0038] 上記 (a) の場合、ストレージ装置 S 1 に記憶されているデータは「D x ≠ D 1」であり、ストレージ装置 S 3 に記憶されているパリティデータは、

ストレージ装置 S 1 に記憶されていたデータ D 1 に基づく「 $P = D 1 \text{ xor } D 2$ 」である。このため、下記式 (3) のとおり、データ D x とパリティデータ P の排他的論理和をとっても、データ D 2 を復元することができない。

$$[0039] \quad D 2 \neq D x \text{ xor } P = D x \text{ xor } D 1 \text{ xor } D 2 \cdots (3)$$

[0040] 上記 (b) の場合、ストレージ装置 S 1 に記憶されているデータは「D 1 #」であり、ストレージ装置 S 3 に記憶されているパリティデータは「 $P x \neq P \# = D 1 \# \text{ xor } D 2$ 」である。このため、下記式 (4) のとおり、データ D 1 # とパリティデータ P x の排他的論理和をとっても、データ D 2 を復元することができない。

$$[0041] \quad D 2 \neq D 1 \# \text{ xor } P x \cdots (4)$$

[0042] 上記 (c) の場合、ストレージ装置 S 1 に記憶されているデータは「D 1 #」であり、ストレージ装置 S 3 に記憶されているパリティデータは、ストレージ装置 S 1 に記憶されていた D 1 に基づく「P」である。このため、下記式 (5) のとおり、データ D 1 # とパリティデータ P の排他的論理和をとっても、データ D 2 を復元することができない。

$$[0043] \quad D 2 \neq D 1 \# \text{ xor } P = D 1 \# \text{ xor } D 1 \text{ xor } D 2 \cdots (5)$$

[0044] これでは、ストレージシステム 100 をシステムダウン直前の非冗長構成、つまりストレージ装置 S 2 が切り離された状態に戻したとしても、以降において、データ D 2 の読み出しが発生した場合に、データ D 2 を正しく復元することができない。

[0045] そこで、実施の形態 1 では、パリティデータの更新を伴うデータの更新時に、更新前のデータとパリティデータとを特定の記憶領域に退避しておく。そして、実施の形態 1 では、書込失敗の後にストレージ装置が再起動されると、退避したデータとパリティデータを退避元にそれぞれ戻して、書込失敗によるストレージ装置間のデータの不整合を防ぐ。以下、図 2 を用いて、実

施の形態 1 にかかるストレージ制御手法の一実施例について説明する。

[0046] (ストレージ制御手法の一実施例)

図 2 は、実施の形態 1 にかかるストレージ制御手法の一実施例を示す説明図である。ここでは、図 1 に示したストレージ装置 S 2 が切り離された非冗長構成のストレージシステム 100 において、ストレージ装置 S 1 に記憶されているデータ D 1 をデータ D 1 # に更新する場合を例に挙げて説明する。

[0047] (1) ストレージ制御装置 101 は、書込要求を受け付ける。ここで、書込要求には、ストレージ装置 S 1 への書込対象となるデータ D 1 # と、要求先のストレージ装置 S 1 の記憶領域のアドレスが含まれている。ストレージ制御装置 101 は、例えば、外部のコンピュータから書込要求を受け付けてもよく、また、ユーザの操作入力により書込要求を受け付けることにしてもよい。

[0048] (2) ストレージ制御装置 101 は、書込要求を受け付けた場合、書込要求に対応するストレージ装置 S 1 の記憶領域に記憶されているデータ D 1 と、データ D 1 を生成元とするパリティデータ P とを、特定の記憶領域 200 に複写する。具体的には、例えば、ストレージ制御装置 101 が、ストレージ装置 S 1 からデータ D 1 を読み出して特定の記憶領域 200 に書き込む。また、ストレージ制御装置 101 が、ストレージ装置 S 3 からパリティデータ P を読み出して特定の記憶領域 200 に書き込む。

[0049] ここで、特定の記憶領域 200 は、ストレージシステム 100 に記憶されているデータ群の各データおよびデータ群から生成されるパリティデータの記憶先の記憶領域とは異なる不揮発性の記憶領域である。特定の記憶領域 200 は、例えば、ストレージ制御装置 101 に設けられていてもよく、また、各ストレージ装置 S 1 ~ S 3 にそれぞれ設けられていてもよい。

[0050] (3) ストレージ制御装置 101 は、受け付けた書込要求に対する書込処理を実行する。具体的には、例えば、ストレージ制御装置 101 が、書込対象のデータ D 1 # を、ストレージ装置 S 1 の記憶領域に書き込む。また、ストレージ制御装置 101 が、新たなパリティデータ P # ($P \# = D 1 \# \times$

or D2) を、ストレージ装置 S3 のパリティデータ P の記憶先の記憶領域に書き込む。より具体的には、例えば、まず、ストレージ制御装置 101 が、ストレージ装置 S1 のデータ D1 とストレージ装置 S3 のパリティデータ P から、切り離されているストレージ装置 S2 のデータ D2 を復元する。そして、ストレージ制御装置 101 が、書込対象のデータ D1 # と復元したデータ D2 から新たなパリティデータ P # を算出して、ストレージ装置 S3 のパリティデータ P の記憶先の記憶領域に書き込む。

[0051] ここで、ストレージ装置 S1 または S3 に対する書き込みに失敗すると、書き込みを失敗したストレージ装置 S1 または S3 がストレージシステム 100 から切り離されてマルチデッド状態となり、ストレージシステム 100 がシステムダウンする。ここでは、ストレージ装置 S3 に対するパリティデータ P # の書き込みに失敗して、パリティデータ P # の一部をストレージ装置 S3 に書き込めなかった場合を想定する。

[0052] この場合、ストレージシステム 100 をシステムダウン直前の非冗長構成に戻すために、切り離されたストレージ装置 S3 が再起動される。ここでは、ストレージ装置 S3 が再起動された結果、ストレージ装置 S3 が正常に動作するストレージ装置として復帰した場合を想定する。

[0053] (4) ストレージ制御装置 101 は、(3) で切り離したストレージ装置が再起動されたか否かを判断する。具体的には、例えば、ストレージ制御装置 101 が、ストレージ装置 S3 から再起動の完了通知を受け付けることにより、ストレージ装置 S3 が再起動されたと判断する。

[0054] (5) ストレージ制御装置 101 は、ストレージ装置 S1 または S3 が再起動された場合、特定の記憶領域 200 に複写されたデータ D1 をストレージ装置 S1 の複写元の記憶領域に書き込む。さらに、ストレージ制御装置 101 は、特定の記憶領域 200 に複写されたパリティデータ P をストレージ装置 S3 の複写元の記憶領域に書き込む。

[0055] このように、実施の形態 1 にかかるストレージ制御手法によれば、パリティデータの更新を伴うデータの更新時に、更新前のデータとパリティデータ

を、特定の記憶領域 200 に退避する。また、書込失敗でストレージ装置が再起動されると、特定の記憶領域 200 に退避したデータとパリティデータをそれぞれ退避元の記憶領域に戻すことで、書込失敗によるストレージ装置間のデータの不整合を防ぐことができる。

[0056] すなわち、実施の形態 1 では、サーマル・オフトラック等の障害は、ストレージ装置を再起動することで回復することを想定して、書込要求単位で、更新前のデータとパリティデータを特定の記憶領域 200 に一時的に退避しておく。これにより、サーマル・オフトラック等の障害によってストレージ装置に対する書き込みに失敗しても、ストレージ装置の再起動後に、退避したデータとパリティデータを退避元の記憶領域に戻すことで、ストレージ装置間のデータの整合性をとることができる。

[0057] 例えば、上記（3）において、ストレージ装置 S3 に対するパリティデータ P# の書き込みに失敗すると、ストレージ装置 S1、S3 間のデータの不整合が発生して、ストレージ装置 S2 のデータ D2 を復元することができない。そこで、上記（5）において、特定の記憶領域 200 に退避したデータ D1 とパリティデータ P をそれぞれ退避元の記憶領域に戻すことで、書込失敗によるストレージ装置 S1、S3 間のデータの不整合を防ぐとともに、戻したデータ D1 とパリティデータ P とによるデータ D2 の復元が可能となる。

[0058] （ストレージシステム 300 のハードウェア構成）

つぎに、実施の形態 1 にかかるストレージ制御装置 101 を RAID3 のストレージシステム 300 に適用した場合について説明する。RAID3 は、データ群を複数の HDD（Hard Disk Drive）に分割して記録し、データ群から生成されるパリティデータを専用の HDD（いわゆる、パリティドライブ）に記録する方式である。

[0059] 以下の図 3 の例では、ストレージシステム 300 内の HDD の台数が、HDD1～HDD3 の 3 台の場合を例に挙げて説明するが、ストレージシステム 300 に含まれる HDD の台数はこれに限らない。RAID3 のストレージ

ジシステム３００の場合、ＨＤＤの台数として、１台のパリティドライブを含む３台以上であれば任意の台数を採用することができる。

[0060] 図３は、実施の形態１にかかるストレージシステム３００のハードウェア構成を示すブロック図である。図３において、ストレージシステム３００は、ＣＰＵ（Ｃｅｎｔｒａｌ　Ｐｒｏｃｅｓｓｉｎｇ　Ｕｎｉｔ）３０１と、メインメモリ３０２と、ディスプレイコントローラ３０３と、ディスプレイ３０４と、通信コントローラ３０５と、ストレージ制御装置１０１と、ＨＤＤ１～ＨＤＤ３と、を備えている。

[0061] ＣＰＵ３０１は、ストレージシステム３００を統括的に制御する。メインメモリ３０２は、ＲＡＭ（Ｒａｎｄｏｍ　Ａｃｃｅｓｓ　Ｍｅｍｏｒｙ）を含む。ＣＰＵ３０１は、例えば、ストレージシステム３００のＯＳ（Ｏｐｅｒａｔｉｎｇ　Ｓｙｓｔｅｍ）をメインメモリ３０２上で展開して起動する。

[0062] ディスプレイコントローラ３０３は、ＣＰＵ３０１の制御にしたがってディスプレイ３０４の表示制御を行う。ディスプレイ３０４は、カーソル、アイコンあるいはツールボックスをはじめ、文書、画像、機能情報等のデータを表示する。ディスプレイ３０４は、例えば、ＣＲＴ、ＴＦＴ液晶ディスプレイ、プラズマディスプレイ等を採用することができる。

[0063] 通信コントローラ３０５は、通信回線を通じてＬＡＮ（Ｌｏｃａｌ　Ａｒｅａ　Ｎｅｔｗｏｒｋ）、ＷＡＮ（Ｗｉｄｅ　Ａｒｅａ　Ｎｅｔｗｏｒｋ）、インターネット等のネットワーク３２０に接続され、ネットワーク３２０を介して外部のコンピュータに接続される。通信コントローラ３０５は、ネットワーク３２０と内部のインターフェースを司り、外部のコンピュータからのデータの入出力を制御する。また、ストレージシステム３００は、キーボードやマウス等の入力装置、スキャナやプリンタ等の出力装置を備えることにしてもよい。

[0064] ストレージ制御装置１０１は、プロセッサ３０６と、メモリ３０７と、不揮発性メモリ３０８と、ＨＤＤコントローラ３０９と、バスコントローラ３

１０と、を備えている。プロセッサ３０６は、メモリ３０７に格納されている本ストレージ制御プログラム等の各種プログラムを実行する。

[0065] プロセッサ３０６は、ＨＤＤコントローラ３０９を介して、ＨＤＤ１～ＨＤＤ３に対するアクセスを制御する。また、プロセッサ３０６は、ＨＤＤコントローラ３０９を介して、ＨＤＤ１～ＨＤＤ３との接続を制御する。以下の説明では、ストレージ制御装置１０１と各ＨＤＤ１～ＨＤＤ３とを接続することを「オンライン」、ストレージ制御装置１０１とＨＤＤ１～ＨＤＤ３とを切り離すことを「オフライン」という。

[0066] 不揮発性メモリ３０８は、電源からの電力の供給が遮断されても記憶内容が消去されない特性を有する記憶媒体である。不揮発性メモリ３０８は、例えばＮＶＲＡＭ（Non Volatile RAM）やＦＬＡＳＨメモリ等である。ＨＤＤコントローラ３０９は、プロセッサ３０６の制御にしたがって、ＨＤＤ１～ＨＤＤ３に対するアクセスを行う。バスコントローラ３１０は、ストレージ制御装置１０１とＣＰＵ３０１とのデータのやりとりを制御する。

[0067] ＨＤＤ１～ＨＤＤ３は、磁性体を塗布した円盤（ハードディスク）を高速回転させることで、磁気ヘッドがデータの読み出し／書き込みを行う記憶装置である。ＨＤＤ１～ＨＤＤ３は、例えば、上述したストレージ装置Ｓ１～Ｓ３に相当する。各ＨＤＤ１～ＨＤＤ３の記憶領域は、データ領域３１１，３１２，３１３とシステム領域３１４，３１５，３１６とに分割されている。

[0068] 各データ領域３１１，３１２，３１３および各システム領域３１４，３１５，３１６の容量は、各ＨＤＤ１～ＨＤＤ３のデータ容量の範囲内で可変である。プロセッサ３０６は、オフライン、オンラインに限らず、ＨＤＤコントローラ３０９を介して、各ＨＤＤ１～ＨＤＤ３のシステム領域３１４～３１６にアクセス可能である。

[0069] ストレージシステム３００において、ＨＤＤ１，ＨＤＤ２は、データ群の各データを分散して記憶するＨＤＤであり、ＨＤＤ３は、データ群の各デー

タの排他的論理和であるパリティデータを記憶するHDD（パリティドライブ）である。HDD 1，HDD 2のデータ領域311，312には、データD1～Dnが分散して記憶されている。また、HDD 3のデータ領域313には、パリティデータP1～Pmが記憶されている。例えば、データD1，D2がHDD 1，HDD 2のデータ領域311，312に分散して記憶され、データD1，D2の排他的論理和であるパリティデータP1がHDD 3のデータ領域313に記憶されている。

[0070] ここで、データは、例えば、ストレージシステム300におけるOS、アプリケーションプログラム、ユーザ情報等を分割したデータである。また、各データのデータサイズは、例えば、ビット単位やバイト単位である。データ群の各データと該データ群から生成されるパリティデータは、例えば、同一のLBA（Logical Block Addressing）番号を割り当てることで管理される。

[0071] LBAとは、ハードディスク内のすべてのセクタに通し番号を振り、その通し番号によってセクタを指定する方式である。すなわち、各HDD 1～HDD 3に対するアクセス（読み出し、書き込み）は、LBA単位で行われる。例えば、データD1，D2およびパリティデータP1にLBA 1が割り当てられているとする。

[0072] この場合、HDD 1のLBA 1を指定することで、データD1が記憶されているHDD 1の記憶領域を特定してアクセスすることができる。また、HDD 2のLBA 1を指定することで、データD2が記憶されているHDD 2の記憶領域を特定してアクセスすることができる。また、HDD 3のLBA 1を指定することで、パリティデータP1が記憶されているHDD 3の記憶領域を特定してアクセスすることができる。

[0073] 各HDD 1～HDD 3のシステム領域314～316には、例えば、ストレージシステム300の管理データが記憶されている。管理データは、例えば、各HDD 1～HDD 3に障害が発生しているか否かを示す情報を含む。各管理データは、プロセッサ306によりHDDコントローラ309を介し

て更新される。ここで、管理データの詳細例について説明する。

[0074] (管理データ 410, 420, 430 の詳細例)

図 4 は、管理データ 410, 420, 430 の詳細例を示す説明図である。図 4 において、管理データ 410 は、HDD 1 のシステム領域 314 に記憶されている管理データである。管理データ 420 は、HDD 2 のシステム領域 315 に記憶されている管理データである。管理データ 430 は、HDD 3 のシステム領域 316 に記憶されている管理データである。

[0075] 各管理データ 410, 420, 430 は、HDD を識別する HDD 番号と、HDD の状態とを関連付けて記憶している。HDD の状態は、初期状態では「正常」であり、HDD に障害が発生すると「正常」から「異常」に更新される。また、HDD の状態は、HDD が障害から復帰すると「異常」から「正常」に更新される。

[0076] また、管理データ 410 は、HDD 1 の起動時刻およびアクセスログを記憶している。管理データ 410 のアクセスログは、HDD コントローラ 309 が HDD 1 に最後にアクセスしたアクセス時刻を含む情報である。管理データ 420 は、HDD 2 の起動時刻およびアクセスログを記憶している。管理データ 420 のアクセスログは、HDD コントローラ 309 が HDD 2 に最後にアクセスしたアクセス時刻を含む情報である。

[0077] 管理データ 430 は、HDD 3 の起動時刻およびアクセスログを記憶している。管理データ 430 のアクセスログは、HDD コントローラ 309 が HDD 3 に最後にアクセスしたアクセス時刻を含む情報である。管理データ 410, 420, 430 によれば、障害が発生している HDD を特定することができる。また、管理データ 410, 420, 430 によれば、各 HDD 1 ~ HDD 3 の起動時刻およびアクセス時刻を特定することができる。

[0078] (ストレージ制御装置 101 の機能的構成)

図 5 は、実施の形態 1 にかかるストレージ制御装置 101 の機能的構成を示すブロック図である。図 5 において、ストレージ制御装置 101 は、受付部 501 と、更新部 502 と、制御部 503 と、複写指示部 504 と、実行

部505と、成否判定部506と、検出部507と、判断部508と、書込指示部509と、対象判定部510と、を含む構成である。各機能部（受付部501～対象判定部510）は、具体的には、例えば、図3に示したメモリ307に記憶されたプログラムをプロセッサ306に実行させることにより、または、HDDコントローラ309、バスコントローラ310により、その機能を実現する。また、各機能部（受付部501～対象判定部510）の処理結果は、特に指定する場合を除いて、メモリ307に記憶される。

[0079] 受付部501は、HDD k のLBA j に対する書込要求を受け付ける機能を有する（例えば、 $k=1, 2, 3$ 、 $j=1, 2, \dots, m$ ）。書込要求は、例えば、要求先のHDD k を特定するHDD番号「 k 」と、要求先のLBA j を特定するLBA番号「 j 」と、書込対象となるデータ（以下、「書込データ」という）と、を含む。具体的には、例えば、受付部501が、バスコントローラ310を介して、CPU301から書込要求を受け付ける。書込要求の具体例については、後述の図6を用いて説明する。

[0080] 更新部502は、書込要求を受け付けた場合、書込要求に対する書込処理が正常に完了したか否かを特定するための書込状態情報を更新する機能を有する。ここで、書込処理が正常に完了したとは、書込要求に伴って発生するデータの書き込みが正常に完了したことを表す。書込状態情報を記憶する書込状態テーブルの具体例については、後述の図7を用いて説明する。

[0081] 図6は、書込要求600の具体例を示す説明図である。図6において、書込要求600は、要求先のHDD k を特定するHDD番号「1」と、要求先のLBA j を特定するLBA番号「1」と、書込データ「D1#」と、を含む。図6の書込要求600は、HDD1のLBA1のデータD1を書込データD1#に更新するためのものである。

[0082] 図7は、書込状態テーブル700の具体例を示す説明図である。図7において、書込状態テーブル700は、書込状態フラグ、HDD番号、LBA番号および書込失敗HDD番号のフィールドを有する。各フィールドに情報を設定することで、書込状態情報がレコードとして書込状態テーブル700に

記憶されている。

- [0083] 書込状態フラグは、書込要求に対する書込処理が正常に完了したか否かを示すフラグである。書込状態フラグは、初期状態では「0」であり、書込要求を受け付けると「0」から「1」に変更される。また、書込状態フラグは、書込要求に対する書込処理が正常に完了すると「1」から「0」に変更される。書込状態フラグが「1」である場合、書込要求に対する書込処理中であることを示す。
- [0084] HDD番号は、要求先のHDD kの識別子である。LBA番号は、要求先のLBA jの識別子である。書込失敗HDD番号は、書込要求に対する書込処理が実行された結果、書込要求に伴って発生するデータの書き込みに失敗したHDDの識別子である。
- [0085] 図7中（7-1）は、図6に示した書込要求600を受け付けた場合の書込状態情報の更新例である。書込要求600を受け付けた場合、更新部502が、書込状態テーブル700の書込状態フラグを「0」から「1」に変更する。また、更新部502が、書込要求600に含まれるHDD番号「1」を、書込状態テーブル700のHDD番号のフィールドに書き込む。また、更新部502が、書込要求600に含まれるLBA番号「1」を、書込状態テーブル700のLBA番号のフィールドに書き込む。
- [0086] 書込状態テーブル700によれば、書込要求を受け付けた要求先のHDD番号とLBA番号を特定することができる。また、書込状態テーブル700によれば、書込要求に対する書込処理が成功したか否かを特定することができる。書込状態テーブル700は、例えば、図3に示した不揮発性メモリ308により実現される。
- [0087] 図5の説明に戻り、制御部503は、HDD1～HDD3に対するアクセスを制御する機能を有する。具体的には、例えば、制御部503は、後述する複写指示部504、実行部505および書込指示部509の指示にしたがって、HDD1～HDD3に対するアクセスを制御する。制御部503は、例えば、図3に示したHDDコントローラ309により実現される。以下の

説明では、制御部503の一例として、HDDコントローラ309を例に挙げて説明する。

[0088] 複写指示部504は、書込要求を受け付けた場合、HDDコントローラ309に対して、要求先のHDD k のLBA j に記憶されているデータ D_i を、書込バッファBに複写するよう指示する機能を有する（ $i = 1, 2, \dots, n$ ）。また、複写指示部504は、書込要求を受け付けた場合、HDDコントローラ309に対して、データ D_i を生成元とするパリティデータ P_j を、書込バッファBに複写するよう指示する機能を有する。

[0089] ここで、書込バッファBは、図2に示した「特定の記憶領域200」に相当し、例えば、各HDD1～HDD3のデータ領域311, 312, 313とは異なる不揮発性の記憶領域である。書込バッファBは、例えば、各HDD1～HDD3のシステム領域314, 315, 316やストレージ制御装置101の不揮発性メモリ308に設けられる。

[0090] 以下の説明では、各HDD1～HDD3のシステム領域314, 315, 316に書込バッファBがそれぞれ設けられている場合を例に挙げて説明する。ここで、書込バッファBの具体例について説明する。

[0091] 図8は、書込バッファBの具体例を示す説明図である。図8において、書込バッファBは、書込バッファB1と、書込バッファB2と、書込バッファB3と、を含む構成である。書込バッファB1は、書込要求を受け付けた要求先のHDD k のLBA j に記憶されているデータ D_i の複写先となる記憶領域である。

[0092] 書込バッファB2は、書込要求を受け付けた要求先のHDD k のLBA j に記憶されているデータ D_i を生成元とするパリティデータ P_j の複写先となる記憶領域である。書込バッファB3は、書込要求に含まれる書込データ（以下、「書込データ D_i #」と表記する）の複写先となる記憶領域である。なお、書込バッファB3についての詳細な説明は後述する。

[0093] 例えば、書込要求600を受け付けた場合、複写指示部504が、HDDコントローラ309に対して、HDD1のLBA1に記憶されているデータ

D 1 を書込バッファ B 1 に複写するよう指示する。この結果、HDD コントローラ 3 0 9 が、HDD 1 の L B A 1 に記憶されているデータ D 1 を読み出して、各 HDD 1 ~ HDD 3 の書込バッファ B 1 に書き込む。これにより、HDD 1 の L B A 1 に記憶されているデータ D 1 が各 HDD 1 ~ HDD 3 の書込バッファ B 1 に複写される。

[0094] また、書込要求 6 0 0 を受け付けた場合、複写指示部 5 0 4 が、HDD コントローラ 3 0 9 に対して、HDD 1 の L B A 1 に記憶されているデータ D 1 を生成元とするパリティデータ P 1 を書込バッファ B 2 に複写するよう指示する。この結果、HDD コントローラ 3 0 9 が、HDD 3 の L B A 1 に記憶されているパリティデータ P 1 を読み出して、各 HDD 1 ~ HDD 3 の書込バッファ B 2 に書き込む。これにより、パリティデータ P 1 が各 HDD 1 ~ HDD 3 の書込バッファ B 2 に複写される。

[0095] また、複写指示部 5 0 4 は、データ D i を生成元とするパリティデータが複数存在する場合は、HDD コントローラ 3 0 9 に対して、データ D i を生成元とする各パリティデータを、書込バッファ B に複写するよう指示する。例えば、R A I D 6 のストレージシステムの場合、パリティデータが 2 種類存在する。このため、R A I D 6 のストレージシステムの場合、複写指示部 5 0 4 が、HDD コントローラ 3 0 9 に対して、データ D i を生成元とする 2 種類のパリティデータを、書込バッファ B に複写するよう指示する。

[0096] 以下、説明の簡単化のため、特に指定する場合を除いて、「HDD コントローラ 3 0 9 に対してデータの読み出し／書き込みを指示して、データを読み出す／書き込む」ことを「HDD コントローラ 3 0 9 を介して、データを読み出す／書き込む」と表現する。

[0097] 図 5 の説明に戻り、実行部 5 0 5 は、データ D i とパリティデータ P j とが書込バッファ B に複写された後、書込要求に対する書込処理を実行する機能を有する。ここで、実行部 5 0 5 によって実行される書込要求に対する書込処理の具体例 (A) ~ (C) について説明する。ただし、書込要求に対する書込処理に成功した場合を例に挙げて説明する。

[0098] (A) 冗長構成時の書込処理

まず、冗長構成時に実行される書込要求600に対する書込処理（書込データD1#のストレージ装置S1への書き込みを要する処理）について説明する。具体的には、例えば、まず、実行部505が、HDDコントローラ309を介して、HDD2のLBA1に記憶されているデータD2（書き込みを要さないデータ）を読み出す。つぎに、実行部505が、書込要求600に含まれる書込データD1#と、読み出されたデータD2との排他的論理和をとって、新たなパリティデータP1#を算出する（ $P1\# = D1\# \oplus D2$ ）。

[0099] そして、実行部505が、HDDコントローラ309を介して、書込データD1#をHDD1のLBA1に書き込む。また、実行部505が、HDDコントローラ309を介して、算出されたパリティデータP1#をHDD3のLBA1に書き込む。この結果、HDD1のLBA1のデータD1が書込データD1#に更新され、HDD3のLBA1のパリティデータP1が新たなパリティデータP1#に更新される。

[0100] (B) 非冗長構成時の書込処理（その1）

つぎに、HDD2がオフラインとなっている非冗長構成時に実行される書込要求600に対する書込処理（データD1を書込データD1#に更新する処理）について説明する。具体的には、例えば、まず、実行部505が、書込要求を受けると、HDDコントローラ309を介して、HDD1のLBA1に記憶されているデータD1を読み出す。また、実行部505が、HDDコントローラ309を介して、HDD3のLBA1に記憶されているパリティデータP1を読み出す。

[0101] そして、実行部505が、読み出されたデータD1とパリティデータP1の排他的論理和をとって、HDD2のLBA1に記憶されているデータD2を算出する。これにより、オフラインとなっているHDD2のLBA1に記憶されているデータD2を復元することができる。このあと、実行部505が、書込要求600に含まれる書込データD1#と、算出されたデータD2

との排他的論理和をとって、新たなパリティデータP 1 #を算出する。

[0102] そして、実行部505が、HDDコントローラ309を介して、書込データD 1 #をHDD 1のLBA 1に書き込む。また、実行部505が、HDDコントローラ309を介して、算出されたパリティデータP 1 #をHDD 3のLBA 1に書き込む。この結果、HDD 1のLBA 1のデータD 1が書込データD 1 #に更新され、HDD 3のLBA 1のパリティデータP 1が新たなパリティデータP 1 #に更新される。

[0103] (C) 非冗長構成時の書込処理 (その2)

つぎに、HDD 2がオフラインとなっている非冗長構成時に実行される、HDD 2のLBA 1のデータD 2を書込データD 2 #に更新する書込要求に対する書込処理について説明する。この場合、HDD 2がオフラインとなっているため、以下の手順によりHDD 3のLBA 1に記憶されているパリティデータP 1の更新を行う。

[0104] 具体的には、例えば、まず、実行部505が、HDDコントローラ309を介して、HDD 1のLBA 1に記憶されているデータD 1を読み出す。このあと、実行部505が、読み出されたデータD 1と、書込要求に含まれる書込データD 2 #との排他的論理和をとって、新たなパリティデータP 1 #を算出する。

[0105] そして、実行部505が、HDDコントローラ309を介して、算出されたパリティデータP 1 #をHDD 3のLBA 1に書き込む。この結果、HDD 3のLBA 1のパリティデータP 1が、データD 2の更新内容(D 2→D 2 #)が反映された新たなパリティデータP 1 #に更新される。

[0106] また、受付部501は、HDD kのLBA jに対する読出要求を受け付ける機能を有する。ここで、読出要求は、例えば、要求先のHDD kを特定するHDD番号「k」と、要求先のLBA jを特定するLBA番号「j」と、を含む。具体的には、例えば、受付部501が、バスコントローラ310を介して、CPU 301から読出要求を受け付ける。

[0107] また、実行部505は、受け付けた読出要求に対する読出処理を実行する

機能を有する。具体的には、例えば、実行部505が、HDDコントローラ309を介して、要求先のHDDkのLBAjに記憶されているデータDiを読み出す。また、要求先のHDDkがオフラインの場合、実行部505が、要求先のHDDkのLBAjに記憶されているデータDiを生成元とするパリティデータPjと他のHDDが記憶するデータとを用いて、データDiを復元する。ここで、HDD2がオフラインとなっている場合を例に挙げて、HDD2のLBA1のデータD2を読み出す読出処理について説明する。

[0108] 具体的には、例えば、まず、読出要求を受けた実行部505が、HDDコントローラ309を介して、HDD1のLBA1に記憶されているデータD1を読み出す。また、実行部505が、HDDコントローラ309を介して、HDD3のLBA1に記憶されているパリティデータP1を読み出す。そして、実行部505が、読み出されたデータD1とパリティデータP1の排他的論理和をとって、HDD2のLBA1に記憶されているデータD2を算出する。これにより、オフラインとなっているHDD2のLBA1に記憶されているデータD2を復元して、読出データとして用いることができる。

[0109] 成否判定部506は、書込要求に対する書込処理の処理結果に基づいて、書込要求に対する書込処理が成功したか否かを判定する機能を有する。具体的には、例えば、成否判定部506が、書込要求に伴って発生するデータの書き込みがすべて正常に完了した場合、書込要求に対する書込処理が成功したと判定する。

[0110] 上記(A)および(B)の例では、書込データD1#およびパリティデータP1#の書き込みが正常に完了した場合、成否判定部506が、書込処理が成功したと判定する。一方、書込データD1#またはパリティデータP1#の少なくともいずれか一方の書き込みが正常に完了しなかった場合、成否判定部506が、書込処理が失敗したと判定する。

[0111] また、上記(C)の例では、パリティデータP1#の書き込みが正常に完了した場合、成否判定部506が、書込処理が成功したと判定する。一方、パリティデータP1#の書き込みが正常に完了しなかった場合、成否判定部

506が、書込処理が失敗したと判定する。

[0112] データの書き込みが正常に完了したか否かの判定は、例えば、HDDkに対して書き込まれる書込データDi#のCRCコードと、実際に書き込まれたデータのCRCコードの一致判定により行われる。ここで、CRCコードが一致する場合、成否判定部506が、書込データDi#の書き込みが正常に完了したと判定する。一方、CRCコードが不一致の場合、成否判定部506が、書込データDi#の書き込みが正常に完了しなかったと判定する。

[0113] また、更新部502は、書込要求に対する書込処理が実行された結果、書込要求に対する書込処理が正常に完了したか否かを特定するための書込状態情報を更新する。具体的には、例えば、更新部502が、上記成否判定部506によって判定された判定結果に基づいて、図7に示した書込状態テーブル700を更新する。

[0114] より具体的には、例えば、書込要求に対する書込処理が成功したと判定された場合、更新部502が、書込状態テーブル700の書込状態フラグを「1」から「0」に変更する。一方、書込要求に対する書込処理が失敗したと判定された場合、更新部502が、書込状態テーブル700の書込失敗HDD番号のフィールドに、データの書き込みが正常に完了しなかったHDDkのHDD番号を書き込む。

[0115] 図7中（7-2）は、書込要求600に対する書込処理が成功した場合の書込状態情報の更新例である。具体的には、例えば、書込要求600に対する書込処理が正常に完了した場合、更新部502が、書込状態テーブル700の書込状態フラグを「1」から「0」に変更する。

[0116] 図7中（7-3）は、書込要求600に対する書込処理が失敗した場合の書込状態情報の更新例である。ここでは、HDD3のLBA1に対するパリティデータP1#の書き込みが正常に完了しなかった場合を例に挙げて説明する。この場合、更新部502が、書込状態テーブル700の書込失敗HDD番号のフィールドに、HDD3のHDD番号「3」を書き込む。

[0117] また、書込要求に対する書込処理が成功した場合、更新部502が、書込

バッファBに記憶されているデータD_iやパリティデータP_jを削除することにしてもよい。また、新たな書込要求を受け付けた場合に、新たなデータD_iやパリティデータP_jを、書込バッファBに上書きするような構成であってもよい。この場合、書込要求に対する書込処理の成功時に、その都度、書込バッファBに記憶されているデータD_iやパリティデータP_jを削除する必要がなくなる。

[0118] 図5の説明に戻り、検出部507は、HDDコントローラ309を介して、障害が発生したHDDを検出する機能を有する。具体的には、例えば、検出部507が、HDD1～HDD3の中から、データの書き込みが正常に行われなかったHDDを、障害が発生したHDDとして検出する。

[0119] 例えば、上記(A)の書込処理が実行された結果、HDD1のLBA1に対する書込データD1#の書き込みが正常に完了しなかった場合、検出部507が、HDD1を障害が発生したHDDとして検出する。また、上記(A)の書込処理が実行された結果、HDD3のLBA1に対するパリティデータP1#の書き込みが正常に完了しなかった場合、検出部507が、HDD3を障害が発生したHDDとして検出する。

[0120] より具体的には、例えば、検出部507が、書込状態テーブル700を参照して、書込失敗HDD番号のフィールドに設定されているHDD番号のHDDを、障害が発生したHDDとして検出する。これにより、データの書き込みが正常に行われなかったHDDを、障害が発生したHDDとして検出することができる。

[0121] また、検出部507が、HDDコントローラ309を介して、データの読み出しが正常に行われなかったHDDを、障害が発生したHDDとして検出する。より具体的には、例えば、検出部507が、HDDから読み出されたデータにリードエラーを示す情報が含まれている場合、該HDDを障害が発生したHDDとして検出する。

[0122] また、検出部507が、HDDコントローラ309を介して、各HDD1～HDD3に診断コマンドを送信することにしてもよい。そして、検出部5

07が、HDDコントローラ309を介して、診断コマンドを送信してから所定時間内に診断コマンドに対する応答信号を各HDD1～HDD3から受信したか否かを判断する。

[0123] 例えば、まず、検出部507が、HDDコントローラ309を介して、HDD1に診断コマンドを送信する。このあと、検出部507が、所定時間内にHDD1から診断コマンドに対する応答信号を受信した場合、HDD1を障害が発生していないHDDとして検出する。一方、所定時間内に診断コマンドに対する応答信号を受信しなかった場合、検出部507が、HDD1を障害が発生したHDDとして検出する。

[0124] また、検出部507は、障害が発生したHDDを検出した場合、障害が発生したHDDをオフライン（切り離す）にする。また、検出部507は、障害が発生したHDDを検出した場合、HDDコントローラ309を介して、図4に示した各HDD1～HDD3の管理データ410、420、430を更新する。ここで、各HDD1～HDD3の管理データ410、420、430の更新例について説明する。

[0125] 図9は、管理データ410、420、430の更新例を示す説明図（その1）である。図9中（9-1）は、HDD2に障害が発生した場合の各管理データ410、420、430の更新例である。検出部507は、障害が発生したHDD2を検出した場合、HDDコントローラ309を介して、管理データ410のHDD2の状態を「正常」から「異常」に更新する。

[0126] また、検出部507が、HDDコントローラ309を介して、管理データ420のHDD2の状態を「正常」から「異常」に更新する。また、検出部507が、HDDコントローラ309を介して、管理データ430のHDD2の状態を「正常」から「異常」に更新する。（9-1）の管理データ410、420、430によれば、HDD1～HDD3のうち、HDD2の状態が「異常」となっていることを特定することができる。

[0127] ここで、管理データ420は、障害が発生したHDD2のシステム領域315に記憶されている。このため、HDD2の障害により、HDDコントロ

ーラ 309 が HDD 2 のシステム領域 315 にアクセスできず、管理データ 420 を更新できない場合がある。(9-1) に示す管理データ 420 の HDD 2 の状態「(異常)」は、HDD コントローラ 309 が管理データ 420 を更新できた場合に、管理データ 420 内の HDD 2 の状態が「異常」となることを表している。

[0128] 図 9 中 (9-2) は、HDD 2 に障害が発生した後、さらに、HDD 3 に障害が発生した場合の各管理データ 410, 430 の更新例である。検出部 507 は、障害が発生した HDD 3 を検出した場合、HDD コントローラ 309 を介して、管理データ 410 の HDD 3 の状態を「正常」から「異常」に更新する。

[0129] また、検出部 507 が、HDD コントローラ 309 を介して、管理データ 430 の HDD 3 の状態を「正常」から「異常」に更新する。ここでは、HDD 2 は切り離されているため、既に障害が発生している HDD 2 の管理データ 420 は更新しない。(9-2) の管理データ 410, 430 によれば、HDD 1 ~ HDD 3 のうち、HDD 2 および HDD 3 の状態が「異常」となっていることを特定することができる。

[0130] ここで、管理データ 430 は、障害が発生した HDD 3 のシステム領域 316 に記憶されている。このため、HDD 3 の障害により、HDD コントローラ 309 が HDD 3 のシステム領域 316 にアクセスできず、管理データ 430 を更新できない場合がある。(9-2) に示す管理データ 430 の HDD 3 の状態「(異常)」は、HDD コントローラ 309 が管理データ 430 を更新できた場合に、管理データ 430 内の HDD 3 の状態が「異常」となることを表している。

[0131] また、検出された検出結果は、例えば、図 10 に示す検出結果テーブル 1000 に記憶することにしてもよい。検出結果テーブル 1000 は、例えば、メモリ 307 によって実現される。ここで、検出結果テーブル 1000 の具体例について説明する。

[0132] 図 10 は、検出結果テーブル 1000 の具体例を示す説明図である。図 1

0において、検出結果テーブル1000は、HDD番号および検出時刻のフィールドを有し、各フィールドに情報を設定することで、検出結果をレコードとして記憶する。ここで、HDD番号は、障害が発生したHDDの識別子である。検出時刻は、障害が発生したHDDが検出された時刻である。

[0133] 図10中(10-1)は、障害が発生したHDDとして、HDD2が検出された場合の検出結果テーブル1000の更新例である。この場合、検出部507が、検出結果テーブル1000のHDD番号のフィールドに、HDD2のHDD番号「2」を書き込む。また、検出部507が、検出結果テーブル1000の検出時刻のフィールドに、HDD2を検出した検出時刻「2010/08/23 02:11:55」を書き込む。

[0134] 図10中(10-2)は、障害が発生したHDDとして、さらに、HDD3が検出された場合の検出結果テーブル1000の更新例である。この場合、検出部507が、検出結果テーブル1000のHDD番号のフィールドに、HDD3のHDD番号「3」を書き込む。また、検出部507が、検出結果テーブル1000の検出時刻のフィールドに、HDD3を検出した検出時刻「2010/08/24 12:09:22」を書き込む。

[0135] ここで、ストレージシステム300は、障害が発生したHDDの台数が2台以上(マルチデッド状態)になると、障害が発生しているHDDのデータを復元できなくなるためシステムダウンする。そこで、判断部508により、ストレージシステム300のシステムダウンの可否を判断する。

[0136] 判断部508は、検出された検出結果に基づいて、障害が発生したHDDの台数 x が、所定数 X 以上か否かを判断する機能を有する。ここで、所定数 X は、障害が発生したHDDの台数が X 台以上となると、ストレージシステム300がマルチデッド状態となる値に設定される。具体的には、例えば、所定数 X は、パリティドライブの台数 Y に1を加算した値「 $X=Y+1$ 」に設定される。

[0137] ストレージシステム300はRAID3のため、パリティドライブの台数 Y は「 $Y=1$ 」である。このため、所定数 X は「 $X=2$ 」となる。また、R

RAID5のストレージシステム（例えば、後述するストレージシステム2000）の場合、パリティドライブの台数 Y は「 $Y=1$ 」である。このため、所定数 X は「 $X=2$ 」となる。また、RAID6のストレージシステムの場合、パリティドライブの台数 Y は「 $Y=2$ 」である。このため、所定数 X は「 $X=3$ 」となる。

[0138] 具体的には、例えば、まず、判断部508が、HDDコントローラ309を介して、各HDD1～HDD3の管理データ410、420、430を参照して、HDD1～HDD3のうち状態が「異常」となっているHDDの台数 x_1 、 x_2 、 x_3 をそれぞれ特定する。

[0139] 図9に示した（9-1）の例では、判断部508が、HDD1の管理データ410を参照して、HDD1～HDD3のうち状態が「異常」となっているHDDの台数「 $x_1=1$ 」を特定する。また、判断部508が、HDD2の管理データ420を参照して、HDD1～HDD3のうち状態が「異常」となっているHDDの台数「 $x_2=1$ 」を特定する。また、判断部508が、HDD3の管理データ430を参照して、HDD1～HDD3のうち状態が「異常」となっているHDDの台数「 $x_3=1$ 」を特定する。

[0140] そして、判断部508が、台数 x_1 、 x_2 、 x_3 のうち最大の値を、障害が発生したHDDの台数 x として特定する。ここでは、台数 x_1 、 x_2 、 x_3 のすべての値が「1」のため、障害が発生したHDDの台数 x は「 $x=1$ 」となる。このため、判断部508が、障害が発生したHDDの台数 x （ $x=1$ ）が所定数 X （ $X=2$ ）未満と判断する。

[0141] また、図9に示した（9-2）の例では、判断部508が、HDD1の管理データ410を参照して、HDD1～HDD3のうち状態が「異常」となっているHDDの台数「 $x_1=2$ 」を特定する。また、判断部508が、HDD2の管理データ420を参照して、HDD1～HDD3のうち状態が「異常」となっているHDDの台数「 $x_2=1$ 」を特定する。また、判断部508が、HDD3の管理データ430を参照して、HDD1～HDD3のうち状態が「異常」となっているHDDの台数「 $x_3=2$ 」を特定する。

- [0142] そして、判断部508が、台数 x 1, x 2, x 3のうち最大の値を、障害が発生したHDDの台数 x として特定する。ここでは、管理データ410あるいは管理データ430に記録されている台数 x 1, x 3の値が「2」で最大のため、障害が発生したHDDの台数 x は「 $x=2$ 」となる。このため、判断部508が、障害が発生したHDDの台数 x ($x=2$) が所定数 X ($X=2$) 以上と判断する。
- [0143] また、障害が発生したHDDの台数 x が所定数 X 以上となった場合、判断部508が、障害が発生したHDDの台数 x が所定数 X 以上となったことをCPU301に通知する。すなわち、ストレージシステム300がマルチデッド状態となったため、判断部508が、ストレージシステム300をシステムダウンさせるようCPU301に要求する。
- [0144] CPU301は、障害が発生したHDDの台数 x が所定数 X 以上となったことが通知された場合、ストレージシステム300をシステムダウンさせる。すなわち、CPU301は、ストレージシステム300がマルチデッド状態となったため、ストレージシステム300をシステムダウンさせる。
- [0145] また、システムダウンしたストレージシステム300を復旧する場合、例えば、ストレージシステム300をシステムダウン直前の非冗長構成に戻す。具体的には、例えば、まず、CPU301またはユーザの操作入力により、ストレージシステム300のハードリセット（または、電源のOFF/ON）を行う。
- [0146] つぎに、プロセッサ306により、障害が発生した複数のHDDについて、障害が発生した順序を特定する。障害が発生したHDDの順序を特定する具体的な処理内容については後述する。このあと、プロセッサ306により、障害が発生した複数のHDDのうち、最後に障害が発生したHDD以外のHDDをオフラインにする。すなわち、プロセッサ306により、障害が発生した複数のHDDのうち、最後に障害が発生したHDDのみをオンラインにする。
- [0147] この際、プロセッサ306により、HDDコントローラ309を介して、

各HDD1～HDD3の管理データ410、420、430内の最後に障害が発生したHDDの状態を更新する。ただし、更新対象となる管理データは、オンラインとなっているHDDの管理データである。図9の(9-2)に示した例では、プロセッサ306が、オンラインのHDD1、HDD3の管理データ410、430のHDD3の状態を「異常」から「正常」に変更する。

[0148] そして、CPU301またはユーザの操作入力により、ストレージシステム300を再起動する。この際、CPU301が、ストレージシステム300の起動通知をプロセッサ306に出力することにしてもよい。これにより、ストレージシステム300をシステムダウン直前の非冗長構成に戻すことができる。なお、ここではCPU301またはユーザの操作入力によってストレージシステム300を再起動することにしたが、プロセッサ306がストレージシステム300の再起動を行うことにしてもよい。

[0149] また、判断部508は、書込要求に対する書込処理が実行されたHDDが再起動されたか否かを判断する機能を有する。ここで、書込要求に対する書込処理が実行されたHDDは、例えば、上記(A)および(B)の書込処理の場合、HDD1、HDD3である。また、上記(C)の書込処理の場合、書込要求に対する書込処理が実行されたHDDはHDD3である。

[0150] 具体的には、例えば、判断部508が、ストレージシステム300の起動通知をCPU301から受け付けた場合、書込要求に対する書込処理が実行されたHDDが再起動されたと判断する。また、判断部508が、書込要求に対する書込処理が実行されたHDDから起動完了通知を受け付けた場合、書込要求に対する書込処理が実行されたHDDが再起動されたと判断することにしてもよい。

[0151] 書込指示部509は、書込要求に対する書込処理が実行されたHDDが再起動されたと判断された場合、HDDコントローラ309に対して、書込バッファBに複写されたデータDiを複写元のLBAに書き込むよう指示する機能を有する。ここで、データDiの複写元のLBAとは、書込要求の要求

先のHDD kのLBA jである。

[0152] また、書込指示部509は、書込要求に対する書込処理が実行されたHDDが再起動されたと判断された場合、HDDコントローラ309に対して、書込バッファBに複写されたパリティデータP_jを複写元のLBAに書き込むよう指示する機能を有する。ここで、パリティデータP_jの複写元のLBAは、要求先のHDD kのLBA jに記憶されているデータD_iを生成元とするパリティデータP_jの記憶先のLBA jである。

[0153] 図8に示した例では、書込指示部509が、HDDコントローラ309に対して、書込バッファB₁に複写されたデータD₁をHDD 1のLBA 1に書き込むよう指示する。この結果、HDDコントローラ309が、書込バッファB₁に記憶されているデータD₁を読み出して、HDD 1のLBA 1に書き込む。この際、HDDコントローラ309は、例えば、各HDD 1～HDD 3の書込バッファB₁のうち、アクセス可能な任意の書込バッファB₁からデータD₁を読み出す。これにより、データD₁がHDD 1の複写元のLBA 1に書き戻される。

[0154] また、書込指示部509が、HDDコントローラ309に対して、書込バッファB₂に複写されたパリティデータP₁をHDD 3のLBA 1に書き込むよう指示する。この結果、HDDコントローラ309が、書込バッファB₂に記憶されているパリティデータP₁を読み出して、HDD 3のLBA 1に書き込む。この際、HDDコントローラ309は、例えば、各HDD 1～HDD 3の書込バッファB₂のうち、アクセス可能な任意の書込バッファB₂からパリティデータP₁を読み出す。これにより、パリティデータP₁がHDD 3の複写元のLBA 1に書き戻される。この結果、HDD 1とHDD 3との間のデータの整合性をとることができる。

[0155] なお、書込指示部509は、書込バッファBに複数のパリティデータが複写されている場合、HDDコントローラ309に対して、書込バッファBに複写された複数のパリティデータを複写元のLBAにそれぞれ書き込むよう指示する。例えば、RAID 6のストレージシステムの場合、複写指示部5

04が、HDDコントローラ309に対して、書込バッファBに複写された2種類のパリティデータを複写元のLBAにそれぞれ書き込むよう指示する。

[0156] ここで、ストレージシステム300をマルチデッド状態から非冗長構成に戻す場合のHDD1～HDD3のLBA1の記憶内容の変遷例について説明する。ここでは、HDD2がオフラインとなっている非冗長構成のストレージシステム300において、書込要求600に対する書込処理（上記（B））が実行された場合を例に挙げて説明する。

[0157] 図11は、HDD1～HDD3のLBA1の記憶内容の変遷例を示す説明図（その1）である。図11の（11-1）において、書込要求600に対する書込処理が実行された結果、HDD1のLBA1に対する書込データD1#の書き込みに成功し、HDD3のLBA1に対する新たなパリティデータP1#の書き込みに失敗している。

[0158] 具体的には、HDD3のLBA1に対してパリティデータP1#の一部を書き込めなかったため、HDD3のLBA1の記憶内容がパリティデータP1#とは異なるパリティデータPxとなっている。この場合、HDD3が、障害が発生したHDDとして検出されてオフラインとなる。このため、ストレージシステム300がマルチデッド状態となりシステムダウンする。

[0159] 図11の（11-2）において、ストレージシステム300のハードリセットが行われた結果、HDD3が再起動されて、ストレージシステム300がマルチデッド状態から非冗長構成に復旧されている。ここで、HDD3のLBA1に記憶されているパリティデータPxは、本来のパリティデータP1#とは異なるデータである。したがって、この時点では、HDD1、HDD3間でデータの不整合が発生している。

[0160] 図11の（11-3）において、書込バッファB1に複写されたデータD1がHDD1のLBA1に書き込まれ、書込バッファB2に複写されたパリティデータP1がHDD3のLBA1に書き込まれる。この結果、HDD1、HDD3は書込み処理開始前の状態に復帰するので、HDD1、HDD3

間でデータの整合性が確保され、オフラインとなっているHDD2のLBA1に記憶されているデータD2を復元することが可能となる。

[0161] <書込失敗時に、複写したデータD_iとパリティデータP_jを書き戻す>

上述した説明では、書込要求に対する書込処理が実行されたHDDが再起動されると、書込バッファBに複写されたデータD_iとパリティデータP_jを複写元のLBAに書き込むことにしたが、これに限らない。

[0162] ストレージシステム300の電源のOFF/ONは、瞬時停電などが原因で発生する場合がある。このため、書込要求に対する書込処理が正常に完了した後、瞬時停電によりストレージシステム300の電源がOFF/ONされ、ストレージシステム300が再起動される場合がある。この場合、書込バッファBに複写されたデータD_iとパリティデータP_jとが再起動後に複写元のLBAに書き込まれると、複写元のLBAの記憶内容が、書込要求に対する書込処理が実行される前の状態に戻ってしまう。

[0163] そこで、書込指示部509は、成否判定部506によって書込処理が失敗したと判定され、かつ、判断部508によってHDDが再起動されたと判断された場合に、書込バッファBに複写されたデータD_iを複写元のLBAに書き込むことにしてもよい。同様に、書込指示部509は、成否判定部506によって書込処理が失敗したと判定され、かつ、判断部508によってHDDが再起動されたと判断された場合に、書込バッファBに複写されたパリティデータP_jを複写元のLBAに書き込むことにしてもよい。

[0164] 具体的には、例えば、まず、書込指示部509が、図7に示した書込状態テーブル700の書込状態フラグが「0」となっているか否かを判断する。ここで、書込状態フラグが「0」の場合には、書込処理が行なわれていない状態にあるため、書込指示部509は、判断部508によってHDDが再起動されたと判断されても、書込バッファBに複写されたデータD_iとパリティデータP_jの複写元のLBAに対する書き込みを行わない。

[0165] 一方、書込状態テーブル700の書込状態フラグが「1」で、かつ、判断部508によってHDDが再起動されたと判断された場合、書込処理中にH

HDDが再起動され、書込要求に対する書込処理が正常に完了していないことを意味する。そのため書込指示部509が、書込バッファBに複写されたデータD_iとパリティデータP_jを複写元のLBAに書き込む。これにより、書込要求に対する書込処理が成功した後、瞬時停電等の理由でHDDが再起動された場合であっても、書込バッファBに複写したデータD_iとパリティデータP_jを複写元のLBAに書き戻すことを防ぐことができる。

[0166] <マルチデッド状態から冗長構成への復旧>

また、上述した説明では、システムダウンしたストレージシステム300を復旧する場合、ストレージシステム300をシステムダウン直前の非冗長構成に戻す場合について説明したが、これに限らない。

[0167] 例えば、上記(A)の冗長構成時の書込処理において、HDD1に対する書込データD1#の書き込みに失敗し、かつ、HDD3に対するパリティデータP1#の書き込みに失敗して、ストレージシステム300がマルチデッド状態となる場合がある。この場合、システムダウンしたストレージシステム300を復旧する際に、ストレージシステム300をマルチデッド状態から冗長構成に戻すことにしてもよい。

[0168] 具体的には、例えば、まず、CPU301またはユーザの操作入力により、システムダウンしたストレージシステム300のハードリセットを行う。つぎに、CPU301またはユーザの操作入力により、ストレージシステム300を再起動する。この際、障害が発生したHDD1、HDD3の両方をオンラインにする。これにより、ストレージシステム300を冗長構成に戻すことができる。

[0169] このあと、書込指示部509は、書込要求に対する書込処理が実行されたHDDが再起動されたと判断された場合、HDDコントローラ309を介して、書込処理の開始時に書込バッファBに複写されたデータD_iを複写元のLBAに書き込む。また、書込指示部509は、書込要求に対する書込処理が実行されたHDDが再起動されたと判断された場合、HDDコントローラ309を介して、書込バッファBに複写されたパリティデータP_jを複写元

のLBAに書き込む。

[0170] ここで、ストレージシステム300をマルチデッド状態から冗長構成に戻す場合のHDD1～HDD3のLBA1の記憶内容の変遷例について説明する。ここでは、冗長構成のストレージシステム300において、書込要求600に対する書込処理（上記（A））が実行された場合を例に挙げて説明する。

[0171] 図12は、HDD1～HDD3のLBA1の記憶内容の変遷例を示す説明図（その2）である。図12の（12-1）において、書込要求600に対する書込処理が実行された結果、HDD1のLBA1に対する書込データD1#の書き込みに失敗し、同時にHDD3のLBA1に対するパリティデータP1#の書き込みに失敗している。

[0172] 具体的には、HDD1のLBA1に対して書込データD1#の一部を書き込めなかったため、HDD1のLBA1の記憶内容が書込データD1#とは異なるデータDxとなっている。また、HDD3のLBA1に対してパリティデータP1#の一部を書き込めなかったため、HDD3のLBA1の記憶内容がパリティデータP1#とは異なるパリティデータPxとなっている。この場合、HDD1、HDD3が、障害が発生したHDDとして検出されてオフラインとなる。このため、ストレージシステム300がマルチデッド状態となりシステムダウンする。

[0173] 図12の（12-2）において、ストレージシステム300のハードリセットが行われた結果、HDD1、3が再起動されて、ストレージシステム300がマルチデッド状態から冗長構成に復旧されている。ここで、HDD1のLBA1に記憶されているデータDxは、当初HDD1に記憶されていたデータD1および書込データD1#のいずれとも異なるデータである。また、HDD3のLBA1に記憶されているパリティデータPxは、当初HDD3に記憶されていたパリティデータP1およびパリティデータP1#のいずれとも異なるデータである。したがって、この時点では、HDD1、HDD3間でデータの不整合が発生している。

[0174] 図12の(12-3)において、書込処理の開始時に書込バッファB1に複写されたデータD1がHDD1のLBA1に書き込まれ、書込バッファB2に複写されたパリティデータP1がHDD3のLBA1に書き込まれている。この結果、HDD1、HDD3間でデータの整合性が確保される。また、以降HDD2がオフラインとなってもHDD2のLBA1に記憶されているデータD2を復元することが可能となる。

[0175] また、冗長構成時の書込要求に対する書込処理において、要求先のHDDkに対するデータDiまたはパリティドライブに対するパリティデータPjの書き込みに失敗して非冗長構成となっても、システムの運用が可能である。このため、書込要求に対する書込処理が実行された結果、ストレージシステム300が冗長構成から非冗長構成となっても、更新部502が、書込状態テーブル700の書込状態フラグを「1」から「0」に変更する。

[0176] <書込要求の要求先のHDDkがオフライン>

また、上記(C)のように、オフラインとなっているHDD2のLBA1のデータD2を書込データD2#に更新する場合、HDD3のLBA1に記憶されているパリティデータP1のみの更新が行われる。このため、オフラインのHDD2に対する書込要求があった場合は、HDD3のLBA1に記憶されているパリティデータP1のみの書込バッファへの複写を行えばよい。

[0177] 対象判定部510は、検出された検出結果に基づいて、書込要求の要求先のHDDkに障害が発生しているか否かを判定する機能を有する。具体的には、例えば、対象判定部510が、HDDコントローラ309を介して、各HDD1～HDD3の管理データ410、420、430を参照して、要求先のHDDkの状態が「異常」となっているか否かを判定する。

[0178] ここで、管理データ410、420、430うち少なくともいずれかの管理データにおいて、要求先のHDDkの状態が「異常」となっている場合、対象判定部510が、HDDkに障害が発生していると判定する。一方、すべての管理データ410、420、430で、要求先のHDDkの状態が「

正常」となっている場合、対象判定部 510 が、HDD k に障害が発生していないと判定する。

[0179] また、対象判定部 510 は、図 10 に示した検出結果テーブル 1000 を参照して、要求先の HDD k に障害が発生しているか否かを判定することにしてもよい。具体的には、例えば、検出結果テーブル 1000 の中に、書込要求に含まれる HDD 番号に対応するレコードがあるか否かを判定する。

[0180] ここで、書込要求に含まれる HDD 番号に対応するレコードがある場合、対象判定部 510 が、要求先の HDD k に障害が発生していると判定する。一方、検出結果テーブル 1000 の中に、書込要求に含まれる HDD 番号に対応するレコードがない場合、対象判定部 510 が、要求先の HDD k に障害が発生していないと判定する。

[0181] 複写指示部 504 は、要求先の HDD k に障害が発生していると判定された場合、HDD コントローラ 309 を介して、要求先の HDD k の LBA j に記憶されているデータ D i を生成元とするパリティデータ P j を書込バッファ B に複写する。また、複写指示部 504 は、要求先の HDD k に障害が発生していると判定された場合、要求先の HDD k の LBA j に記憶されているデータ D i の書込バッファ B への複写は行わない。

[0182] この場合、判断部 508 は、パリティデータ P j が書込バッファ B に複写された後、書込要求に対する書込処理が実行された HDD が再起動されたか否かを判断する。そして、書込指示部 509 は、例えば、書込要求に対する書込処理が実行された HDD が再起動されたと判断された場合、HDD コントローラ 309 を介して、書込バッファ B に複写されたパリティデータ P j を複写元の LBA j に書き込む。

[0183] <パリティドライブがオフライン>

また、パリティドライブがオフラインとなっている場合、書込要求の要求先の HDD k の LBA j のデータ D i のみの更新が行われる。このため、パリティドライブがオフラインとなっている場合は、書込要求の要求先の HDD k の LBA j に記憶されているデータ D i のみの複写を行えばよい。

- [0184] 対象判定部510は、検出された検出結果に基づいて、書込要求の要求先のHDDkのLBAjに記憶されているデータDiを生成元とするパリティデータPjの記憶先のHDDに障害が発生しているか否かを判定する機能を有する。具体的には、例えば、対象判定部510が、HDDコントローラ309を介して、各HDD1～HDD3の管理データ410、420、430を参照して、パリティドライブ（HDD3）の状態が「異常」となっているか否かを判定する。
- [0185] ここで、管理データ410、420、430のうち少なくともいずれかの管理データにおいて、パリティドライブの状態が「異常」となっている場合、対象判定部510が、パリティドライブに障害が発生していると判定する。一方、すべての管理データ410、420、430において、パリティドライブの状態が「正常」となっている場合、対象判定部510が、パリティドライブに障害が発生していないと判定する。
- [0186] また、対象判定部510は、図10に示した検出結果テーブル1000を参照して、パリティドライブに障害が発生しているか否かを判定することにしてもよい。具体的には、例えば、検出結果テーブル1000の中に、パリティドライブであるHDD3のHDD番号「3」に対応するレコードがあるか否かを判定する。
- [0187] ここで、HDD3のHDD番号「3」に対応するレコードがある場合、対象判定部510が、パリティドライブに障害が発生していると判定する。一方、検出結果テーブル1000の中に、HDD3のHDD番号「3」に対応するレコードがない場合、対象判定部510が、パリティドライブに障害が発生していないと判定する。
- [0188] 複写指示部504は、パリティドライブに障害が発生していると判定された場合、HDDコントローラ309を介して、要求先のHDDkのLBAjに記憶されているデータDiを書込バッファBに複写する。また、複写指示部504は、パリティドライブに障害が発生していると判定された場合、要求先のHDDkのLBAjに記憶されているデータDiを生成元とするパリティ

ティデータ P_j の書込バッファ B への複写は行わない。

[0189] この場合、判断部 508 は、データ D_i が書込バッファ B に複写された後、書込要求に対する書込処理が実行された HDD が再起動されたか否かを判断する。そして、書込指示部 509 は、例えば、書込要求に対する書込処理が実行された HDD が再起動されたと判断された場合、HDD コントローラ 309 を介して、書込バッファ B に複写されたデータ D_i を要求先の HDD k の LBA j に書き込む。

[0190] <書込データ D_i # の書込バッファ B への複写>

また、複写指示部 504 は、さらに、HDD コントローラ 309 を介して、書込要求に含まれる書込データ D_i # を書込バッファ B に複写することにしてもよい。例えば、書込要求 600 を受け付けた場合、複写指示部 504 が、HDD コントローラ 309 を介して、HDD 1 の LBA 1 に記憶されているデータ D_1 を各 HDD 1 ~ HDD 3 の書込バッファ B_1 に複写する。また、複写指示部 504 が、HDD コントローラ 309 を介して、HDD 1 の LBA 1 に記憶されているデータ D_1 を生成元とするパリティデータ P_1 を各 HDD 1 ~ HDD 3 の書込バッファ B_2 に複写する。さらに、複写指示部 504 が、HDD コントローラ 309 を介して、書込要求 600 に含まれる書込データ D_1 # を、各 HDD 1 ~ HDD 3 の書込バッファ B_3 (図 8 参照) に複写する。

[0191] この場合、判断部 508 は、データ D_i 、パリティデータ P_j および書込データ D_i # が書込バッファ B に複写された後、書込要求に対する書込処理が実行された HDD が再起動されたか否かを判断する。例えば、判断部 508 が、データ D_1 、パリティデータ P_1 および書込データ D_1 # が書込バッファ B_1 ~ B_3 にそれぞれ複写された後、書込要求 600 に対する書込処理が実行された HDD が再起動されたか否かを判断する。

[0192] そして、書込指示部 509 は、HDD が再起動されたと判断された場合、HDD コントローラ 309 を介して、書込バッファ B に複写されたデータ D_i を要求先の LBA j に書き込む。また、書込指示部 509 は、HDD コン

トローラ 309 を介して、書込バッファ B に複写されたパリティデータ P_j を複写元の LBA_j に書き込む。

[0193] このあと、書込指示部 509 は、HDD コントローラ 309 を介して、書込バッファ B に複写された書込データ D_i # を要求先の LBA_j に書き込む。また、書込指示部 509 は、HDD コントローラ 309 を介して、書込データ D_i # を生成元とするパリティデータ P_j # を、パリティデータ P_j の複写元の LBA_j に書き込む。

[0194] 例えば、書込要求 600 に対する書込処理が実行された HDD が再起動されたと判断された場合、書込指示部 509 が、HDD コントローラ 309 を介して、書込バッファ B₁ に複写されたデータ D₁ を要求先の HDD₁ の LBA₁ に書き込む。また、書込指示部 509 が、HDD コントローラ 309 を介して、書込バッファ B₂ に複写されたパリティデータ P₁ を複写元の HDD₃ の LBA₁ に書き込む。

[0195] このあと、書込指示部 509 が、HDD コントローラ 309 を介して、書込バッファ B₃ に複写された書込データ D₁ # を生成元とするパリティデータ P₁ # を算出する。そして、書込指示部 509 が、HDD コントローラ 309 を介して、書込バッファ B₃ に複写された書込データ D₁ # を要求先の HDD₁ の LBA₁ に書き込む。また、書込指示部 509 が、HDD コントローラ 309 を介して、算出されたパリティデータ P₁ # を、パリティデータ P₁ の複写元の HDD₃ の LBA₁ に書き込む。

[0196] これにより、書込バッファ B に複写したデータ D_i とパリティデータ P_j を書き戻して再起動後の HDD 間のデータの整合性を確保した後、書込バッファ B に複写した書込データ D_i # を用いて、書込要求に対する書込処理を再実行することができる。

[0197] また、更新部 502 は、上記書込指示部 509 によって指示された各種データの書込処理（以下、「データ書込処理」という）が成功した場合、書込状態テーブル 700 の書込状態フラグを「1」から「0」に変更する。すなわち、例えば、書込バッファ B に複写されたデータ D_i とパリティデータ P

」を複写元のLBA」に書き戻したことで、HDD間のデータの整合性が確保されたため、書込状態テーブル700の書込状態フラグを「1」から「0」に変更する。

[0198] また、更新部502は、書込状態テーブル700の書込状態フラグが「1」から「0」に変更された場合、書込バッファBを初期化することにしてもよい。すなわち、書込要求に対する書込処理またはデータ書込処理に成功した場合、更新部502が、書込バッファB1～B3に複写されたデータD_i、パリティデータP_jおよび書込データD_i#を破棄する。

[0199] また、上記書込指示部509によって指示されたデータ書込処理に失敗した場合、以降において、データD_iの読出時に、書込バッファBに複写されたデータD_iを読出データとして用いることにしてもよい。また、データ書込処理に失敗した場合、以降において、パリティデータP_jの読出時に、書込バッファBに複写されたパリティデータP_jを読出データとして用いることにしてもよい。また、データ書込処理に失敗した場合、以降において、書込データD_i#の読出時に、書込バッファBに複写された書込データD_i#を読出データとして用いることにしてもよい。これにより、新たな書込要求を受け付けるまでの間、ストレージシステム300を正常に運用することができる。

[0200] <データD_iとパリティデータP_jの複写が不要なパターン>

また、書込要求に含まれる書込データD_i#を書込バッファBに複写する場合、要求先のHDDkのLBA」のデータD_iと、データD_iを生成元とするパリティデータP_jの書込バッファBへの複写が不要となる場合がある。すなわち、データD_iとパリティデータP_jを複写元のLBA」に書き戻して書込要求を受け付ける直前の状態に戻すことなく、書込要求に対する書込処理を再実行することができる場合がある。

[0201] 具体的には、書込要求に対する書込処理の実行時に、オフラインとなっているHDDのデータを復元する必要がない場合は、データD_iとパリティデータP_jの書込バッファBへの複写が不要となる。ここで、データD_iとパ

リティデータ P_j の書込バッファ B への複写が不要となるパターン 1 ~ パターン 3 について説明する。

[0202] (パターン 1) 冗長構成時に書込要求を受け付けた場合

ストレージシステム 300 が冗長構成の場合、書込バッファ B に複写されたデータ D_i とパリティデータ P_j を複写元の LBA_j に書き込むことなく、書込バッファ B に複写された書込データ $D_i \#$ を用いて、書込要求に対する書込処理を再実行することができる。そこで、上記判断部 508 により、ストレージシステム 300 が冗長構成か否かを判断する。

[0203] ここで、RAID 3, 5, 6 等のストレージシステムは、障害が発生した HDD の台数 x が、パリティドライブの台数 Y 未満の場合に冗長構成となる。このため、判断部 508 が、検出された検出結果に基づいて、障害が発生した HDD の台数 x がパリティドライブの台数 Y 未満か否かを判断する。ストレージシステム 300 のパリティドライブの台数 Y は「 $Y = 1$ 」のため、判断部 508 が、障害が発生した HDD の台数 x が「 $x = 0$ 」か否かを判断する。

[0204] そして、複写指示部 504 は、障害が発生した HDD の台数 x がパリティドライブの台数 Y 未満と判断された場合、HDD コントローラ 309 を介して、書込要求に含まれる書込データ $D_i \#$ を書込バッファ B に複写する。また、複写指示部 504 は、障害が発生した HDD の台数 x がパリティドライブの台数 Y 未満と判断された場合、データ D_i とパリティデータ P_j の書込バッファ B への複写は行わない。

[0205] この場合、判断部 508 は、書込要求に含まれる書込データ $D_i \#$ が書込バッファ B に複写された後、書込要求に対する書込処理が実行された HDD が再起動されたか否かを判断する。そして、書込指示部 509 は、HDD が再起動されたと判断された場合、HDD コントローラ 309 を介して、書込バッファ B に複写された書込データ $D_i \#$ を要求先の LBA_j に書き込む。また、書込指示部 509 は、HDD コントローラ 309 を介して、書込データ $D_i \#$ を生成元とするパリティデータ $P_j \#$ を、パリティデータ P_j の複

写元のLBA_jに書き込む。

- [0206] これにより、データD_iおよびパリティデータP_jの書込バッファBに対する複写処理、および書込バッファBに複写されたデータD_iおよびパリティデータP_jの複写元のLBA_jに対するデータ書込処理を削減できる。
- [0207] ここで、冗長構成時に書込要求を受け付けた場合のHDD1～HDD3のLBA1の記憶内容の変遷例について説明する。ここでは、冗長構成のストレージシステム300において、書込要求600に対する書込処理（上記（A））が実行された場合を例に挙げて説明する。
- [0208] 図13は、HDD1～HDD3のLBA1の記憶内容の変遷例を示す説明図（その3）である。図13の（13-1）において、書込要求600に対する書込処理が実行された結果、HDD1のLBA1に対する書込データD1#の書き込みに失敗し、HDD3のLBA1に対するパリティデータP1#の書き込みに失敗している。この場合、HDD1、HDD3が、障害が発生したHDDとして検出されてオフラインとなる。このため、ストレージシステム300がマルチデッド状態となりシステムダウンする。
- [0209] 図13の（13-2）において、ストレージシステム300のハードリセットが行われた結果、HDD1、3が再起動されて、ストレージシステム300がマルチデッド状態から冗長構成に復旧されている。ここで、HDD1のLBA1に記憶されているデータD_xは、データD1および書込データD1#のいずれとも異なるデータである。また、HDD3のLBA1に記憶されているパリティデータP_xは、パリティデータP1およびパリティデータP1#のいずれとも異なるデータである。したがって、この時点では、HDD1、HDD3間でデータの不整合が発生している。
- [0210] 図13の（13-3）において、書込バッファB3に複写された書込データD1#がHDD1のLBA1に書き込まれている。また、書込バッファB3に複写された書込データD1#を生成元とするパリティデータP1#（ $P1\# = D1\# \text{ xor } D2$ ）がHDD3のLBA1に書き込まれている。この結果、HDD1、HDD3間でデータの整合性が確保され、HDD2が

オフラインとなってもHDD 2のLBA 1に記憶されているデータD 2を復元することが可能となる。

[0211] (パターン2) 要求先のHDD kがオフライン

要求先のHDD kがオフラインの場合、書込バッファBに複写されたパリティデータP jを複写元のLBA jに書き込むことなく、書込バッファBに複写された書込データD i #を用いて、書込要求に対する書込処理を再実行することができる。

[0212] そこで、複写指示部504は、要求先のHDD kに障害が発生していると判定された場合、HDDコントローラ309を介して、書込要求に含まれる書込データD i #を書込バッファBに複写する。また、複写指示部504は、要求先のHDD kに障害が発生していると判定された場合、データD iとパリティデータP jの書込バッファBへの複写は行わない。

[0213] この場合、判断部508は、書込要求に含まれる書込データD i #が書込バッファBに複写された後、書込要求に対する書込処理が実行されたHDDが再起動されたか否かを判断する。そして、書込指示部509は、HDDが再起動されたと判断された場合、HDDコントローラ309を介して、書込データD i #を生成元とするパリティデータP j #を、パリティデータP jの複写元のLBA jに書き込む。

[0214] これにより、パリティデータP jの書込バッファBに対する複写処理、および書込バッファBに複写されたパリティデータP jの複写元のLBA jに対するデータ書込処理を削減できる。

[0215] ここで、要求先のHDD kがオフラインの書込要求を受け付けた場合のHDD 1～HDD 3のLBA 1の記憶内容の変遷例について説明する。ここでは、HDD 2がオフラインとなっている非冗長構成時に、HDD 2のLBA 1のデータD 2を書込データD 2 #に更新する書込要求に対する書込処理（上記（C））が実行された場合を例に挙げて説明する。

[0216] 図14は、HDD 1～HDD 3のLBA 1の記憶内容の変遷例を示す説明図（その4）である。図14の（14-1）において、HDD 2のLBA 1

のデータD2をデータD2#に更新する書込要求に対する書込処理が実行された結果、HDD3のLBA1に対するパリティデータP1#の書き込みに失敗している。この場合、HDD3が、障害が発生したHDDとして検出されてオフラインとなる。このため、ストレージシステム300がマルチデッド状態となりシステムダウンする。

[0217] 図14の(14-2)において、ストレージシステム300のハードリセットが行われた結果、HDD3が再起動されて、ストレージシステム300がマルチデッド状態から非冗長構成に復旧されている。ここで、HDD3のLBA1に記憶されているパリティデータPxは、パリティデータP1と異なるデータである。したがって、この時点では、HDD1、HDD3間でデータの不整合が発生している。

[0218] 図14の(14-3)において、書込バッファB3に複写された書込データD2#を生成元とするパリティデータP1# ($P1\# = D1 \text{ xor } D2\#$) がHDD3のLBA1に書き込まれている。この結果、HDD1、HDD3間でデータの整合性が確保され、以降において、オフラインとなっているHDD2のLBA1の書込データD2#を復元することが可能となる。

[0219] (パターン3) パリティドライブがオフライン

パリティドライブがオフラインの場合、書込バッファBに複写されたデータDiを複写元のLBAjに書き込むことなく、書込バッファBに複写された書込データDi#を用いて、書込要求に対する書込処理を再実行することができる。

[0220] そこで、複写指示部504は、パリティドライブに障害が発生していると判定された場合、HDDコントローラ309を介して、書込要求に含まれる書込データDi#を書込バッファBに複写する。また、複写指示部504は、パリティドライブに障害が発生していると判定された場合、データDiとパリティデータPjの書込バッファBへの複写は行わない。

[0221] この場合、判断部508は、書込要求に含まれる書込データDi#が書込バッファBに複写された後、書込要求に対する書込処理が実行されたHDD

が再起動されたか否かを判断する。そして、書込指示部509は、HDDが再起動されたと判断された場合、HDDコントローラ309を介して、書込バッファBに複写された書込データD_i#を要求先のLBA_jに書き込む。これにより、データD_iの書込バッファBに対する複写処理、および書込バッファBに複写されたデータD_iの複写元のLBA_jに対するデータ書込処理を削減できる。

[0222] ここで、パリティドライブがオフラインのときに書込要求を受け付けた場合のHDD1～HDD3のLBA1の記憶内容の変遷例について説明する。ここでは、HDD3がオフラインとなっている非冗長構成時に、書込要求600に対する書込処理が実行された場合を例に挙げて説明する。

[0223] 図15は、HDD1～HDD3のLBA1の記憶内容の変遷例を示す説明図（その5）である。図15の（15-1）において、書込要求600に対する書込処理が実行された結果、HDD1のLBA1に対する書込データD₁#の書き込みに失敗している。この場合、HDD1が、障害が発生したHDDとして検出されてオフラインとなる。このため、ストレージシステム300がマルチデッド状態となりシステムダウンする。

[0224] 図15の（15-2）において、ストレージシステム300のハードリセットが行われた結果、HDD1が再起動されて、ストレージシステム300がマルチデッド状態から非冗長構成に復旧されている。ここで、HDD1のLBA1に記憶されているデータD_xは、データD₁および書込データD₁#のいずれとも異なるデータである。

[0225] 図15の（15-3）において、書込バッファB3に複写された書込データD₁#がHDD1のLBA1に書き込まれている。

[0226] また、上述した説明では、書込バッファBを、各HDD1～HDD3のシステム領域314、315、316にそれぞれ設ける場合を例に挙げて説明したが、不揮発性メモリ308に設けることにしてもよい。不揮発性メモリ308に書込バッファBを設けた場合、複写指示部504や書込指示部509が、HDDコントローラ309を介することなく、書込バッファBに直接ア

クセスできる。また、複写指示部504が、データD_iやパリティデータP_j等を書込バッファBに複写するために、複数のHDD1～HDD3にアクセスする必要がない。

[0227] このため、不揮発性メモリ308に書込バッファBを設ける場合、各HDD1～HDD3に書込バッファBを設ける場合に比べて、複写指示部504によるデータ複写処理および書込指示部509によるデータ書込処理の高速化を図ることができる。一方、各HDD1～HDD3のシステム領域314、315、316に書込バッファBを設ける場合は、不揮発性メモリ308に記憶するデータ量を抑えることができる。

[0228] <障害が発生したHDDの順序特定>

ここで、障害が発生した複数のHDDについて、障害が発生した順序を特定する具体的な処理内容の一例について説明する。まず、図9中(9-2)に示した各HDD1～HDD3の管理データ410、420、430を用いて、障害が発生した順序を特定する場合について説明する。

[0229] 具体的には、例えば、プロセッサ306が、HDDコントローラ309を介して、各HDD1～HDD3の管理データ410、420、430を参照する。そして、プロセッサ306が、管理データ410、420、430の少なくともいずれかの管理データにおいて、状態が「異常」となっているHDD2、HDD3を特定する。

[0230] つぎに、プロセッサ306が、HDDコントローラ309を介して、特定されたHDD2の管理データ420のアクセスログを参照して、HDD2に最後にアクセスしたアクセス時刻(以下、「アクセス時刻t1」という)を特定する。また、プロセッサ306が、HDDコントローラ309を介して、特定されたHDD3の管理データ430のアクセスログを参照して、HDD3に最後にアクセスしたアクセス時刻(以下、「アクセス時刻t2」という)を特定する。

[0231] そして、プロセッサ306が、特定されたアクセス時刻t1とアクセス時刻t2とを比較して、HDD2、HDD3について、障害が発生した順序を

特定する。具体的には、例えば、プロセッサ 306 が、HDD 2, HDD 3 について、アクセス時刻が古い順に障害が発生したと特定する。

[0232] ここでは、アクセス時刻 t_1 のほうがアクセス時刻 t_2 に比べて古い場合を想定する。このため、プロセッサ 306 が、HDD 2 に障害が発生した後、HDD 3 に障害が発生したことを特定する。また、管理データ 410, 420, 430 のアクセスログから障害が発生した順序を特定する場合を例に挙げて説明したが、各 HDD 1 ~ HDD 3 の起動時刻から障害が発生した順序を特定することにしてもよい。具体的には、例えば、プロセッサ 306 が、HDD 2, HDD 3 について、起動時刻が古い順に障害が発生したと特定することにしてもよい。

[0233] つぎに、図 10 に示した検出結果テーブル 1000 を用いて、障害が発生した順序を特定する場合について説明する。各 HDD 1 ~ HDD 3 の管理データ 410, 420, 430 は、HDD の障害により適切に更新されていない場合がある。このため、プロセッサ 306 は、検出結果テーブル 1000 を参照して、障害が発生した複数の HDD について、障害が発生した順序を特定することにしてもよい。

[0234] 図 10 に示した (10-2) の例では、まず、プロセッサ 306 が、検出結果テーブル 1000 の HDD 番号を参照して、障害が発生した HDD 2, HDD 3 を特定する。そして、プロセッサ 306 が、検出結果テーブル 1000 の検出時刻を参照して、障害が発生した複数の HDD 2, HDD 3 について、障害が発生した順序を特定する。

[0235] 具体的には、例えば、プロセッサ 306 が、HDD 2, HDD 3 について、検出時刻が古い順に障害が発生したと特定する。ここでは、HDD 2 の検出時刻「2010/08/23 02:11:55」のほうが、HDD 3 の検出時刻「2010/08/24 12:09:22」に比べて古い。このため、プロセッサ 306 が、HDD 2 に障害が発生した後、HDD 3 に障害が発生したことを特定する。

[0236] (ストレージ制御装置 101 のストレージ制御処理手順)

つぎに、実施の形態 1 にかかるストレージ制御装置 101 のストレージ制御処理手順について説明する。

[0237] 図 16 および図 17 は、実施の形態 1 にかかるストレージ制御装置のデータ複写処理手順の一例を示すフローチャートである。図 16 のフローチャートにおいて、まず、受付部 501 により、HDD k の LBA j に対する書込要求を受け付けたか否かを判断する（ステップ S 1601）。

[0238] ここで、書込要求を受け付けるのを待つ（ステップ S 1601 : No）、受け付けた場合（ステップ S 1601 : Yes）、判断部 508 により、障害が発生した HDD の台数 x が「 $x = 0$ 」か否かを判断する（ステップ S 1602）。

[0239] ここで、「 $x = 0$ 」の場合（ステップ S 1602 : Yes）、更新部 502 により、書込状態テーブル 700 の書込状態フラグを「0」から「1」に変更する（ステップ S 1603）。さらに、更新部 502 により、受け付けた書込要求に含まれる HDD 番号「 k 」と LBA 番号「 j 」を、書込状態テーブル 700 に書き込む（ステップ S 1604）。

[0240] そして、複写指示部 504 により、HDD コントローラ 309 を介して、書込要求に含まれる書込データ $D_i \#$ を、各 HDD 1 ~ HDD 3 の書込バッファ B3 に複写する（ステップ S 1605）。このあと、実行部 505 により、受け付けた書込要求に対する書込処理を実行して（ステップ S 1606）、図 17 に示すステップ S 1701 に移行する。

[0241] また、ステップ S 1602 において、障害が発生した HDD の台数 x が「 $x \neq 0$ 」の場合（ステップ S 1602 : No）、更新部 502 により、書込状態テーブル 700 の書込状態フラグが「1」か否かを判断する（ステップ S 1607）。

[0242] ここで、書込状態フラグが「0」の場合（ステップ S 1607 : No）、更新部 502 により、書込状態テーブル 700 の書込状態フラグを「0」から「1」に変更する（ステップ S 1608）。さらに、更新部 502 により、受け付けた書込要求に含まれる HDD 番号「 k 」と LBA 番号「 j 」を、

書込状態テーブル700に書き込む（ステップS1609）。

[0243] つぎに、対象判定部510により、パリティドライブに障害が発生しているか否かを判定する（ステップS1610）。ここで、パリティドライブに障害が発生している場合（ステップS1610：Yes）、ステップS1605に移行する。

[0244] 一方、パリティドライブに障害が発生していない場合（ステップS1610：No）、対象判定部510により、書込要求の要求先のHDDkに障害が発生しているか否かを判定する（ステップS1611）。ここで、書込要求の要求先のHDDkに障害が発生している場合（ステップS1611：Yes）、ステップS1605に移行する。

[0245] 一方、書込要求の要求先のHDDkに障害が発生していない場合（ステップS1611：No）、ステップS1612に移行する。そして、複写指示部504により、HDDコントローラ309を介して、書込要求の要求先のHDDkのLBAjに記憶されているデータDiを、各HDD1～HDD3の書込バッファB1に複写する（ステップS1612）。

[0246] また、複写指示部504により、HDDコントローラ309を介して、書込バッファB1に複写されたデータDiを生成元とするパリティデータPjを、各HDD1～HDD3の書込バッファB2に複写する（ステップS1613）。さらに、複写指示部504により、書込要求に含まれる書込データDi#を、各HDD1～HDD3の書込バッファB3に複写して（ステップS1614）、ステップS1606に移行する。

[0247] また、ステップS1607において、書込状態フラグが「1」の場合（ステップS1607：Yes）、ステップS1615に移行する。そして、検出部507により、HDDコントローラ309を介して、オンラインのHDD1～HDD3の管理データ410、420、430の、書き込みに失敗したHDDの状態を「正常」から「異常」に変更する（ステップS1615）。書き込みに失敗したHDDは、例えば、書込状態テーブル700の書込失敗HDD番号のフィールドに設定されているHDD番号から特定される。

- [0248] このあと、検出部507により、書き込みに失敗したHDDをオフラインにする（ステップS1616）。そして、判断部508により、障害が発生したHDDの台数が2台以上となったことをCPU301に通知することで、ストレージシステム300がシステムダウンして（ステップS1617）、本フローチャートによる一連の処理を終了する。
- [0249] なお、ステップS1607において、書込状態フラグが「1」となるのは（ステップS1607：Yes）、後述する図18に示すステップS1804において、データ書込処理に失敗した場合（ステップS1804：No）である。
- [0250] 図17のフローチャートにおいて、まず、成否判定部506により、図16に示したステップS1606において実行された書込要求に対する書込処理が成功したか否かを判定する（ステップS1701）。ここで、書込要求に対する書込処理が成功した場合（ステップS1701：Yes）、更新部502により、書込状態テーブル700の書込状態フラグを「1」から「0」に変更して（ステップS1702）、ステップS1704に移行する。
- [0251] 一方、書込要求に対する書込処理が失敗した場合（ステップS1701：No）、更新部502により、書込状態テーブル700の書込失敗HDD番号のフィールドに、データの書き込みに失敗したHDDのHDD番号を書き込む（ステップS1703）。このあと、検出部507により、障害が発生したHDDが検出されたか否かを判断する（ステップS1704）。
- [0252] ここで、障害が発生したHDDが検出されなかった場合（ステップS1704：No）、本フローチャートによる一連の処理を終了する。一方、障害が発生したHDDが検出された場合（ステップS1704：Yes）、ステップS1705に移行する。そして、検出部507により、HDDコントローラ309を介して、オンラインのHDD1～HDD3の管理データ410、420、430の、障害が発生したHDDの状態を「正常」から「異常」に変更する（ステップS1705）。
- [0253] このあと、検出部507により、HDDコントローラ309を介して、障

害が発生したHDDをオフラインにする（ステップS 1706）。そして、判断部508により、障害が発生したHDDの台数 x が「 $x \geq 2$ 」か否かを判断する（ステップS 1707）。ここで、「 $x < 2$ 」の場合（ステップS 1707：No）、更新部502により、書込状態テーブル700の書込状態フラグを「1」から「0」に変更して（ステップS 1708）、本フローチャートによる一連の処理を終了する。

[0254] 一方、「 $x \geq 2$ 」の場合（ステップS 1707：Yes）、ステップS 1709に移行する。そして、判断部508により、障害が発生したHDDの台数が2台以上となったことをCPU301に通知することで、ストレージシステム300がシステムダウンして（ステップS 1709）、本フローチャートによる一連の処理を終了する。

[0255] これにより、パリティデータ P_j の更新を伴うデータ D_i の更新時に、更新前のデータ D_i 、パリティデータ P_j および書込データ $D_i \#$ を書込バッファBに退避しておくことができる。また、冗長構成時に書込要求を受け付けた場合は、書込データ $D_i \#$ のみを書込バッファBに退避しておくことができる。また、要求先のHDD k またはパリティドライブがオフラインの場合は、書込データ $D_i \#$ のみを書込バッファBに退避しておくことができる。

[0256] つぎに、実施の形態1にかかるストレージ制御装置101のシステム復旧処理手順について説明する。ここでは、システムダウンしたストレージシステム300をシステムダウン直前の非冗長構成に復旧する場合を例に挙げて説明する。ただし、ストレージシステム300のハードリセットが行われた後、障害が発生した複数のHDDのうち最後にオフラインとなったHDDをオンラインにして、ストレージシステム300が再起動された場合を想定する。

[0257] 図18は、実施の形態1にかかるストレージ制御装置101のシステム復旧処理手順の一例を示すフローチャートである。図18のフローチャートにおいて、まず、判断部508により、書込要求に対する書込処理が実行され

たHDDが再起動されたか否かを判断する（ステップS1801）。

[0258] ここで、HDDが再起動されるのを待つ（ステップS1801：No）、再起動された場合（ステップS1801：Yes）、書込指示部509により、書込状態テーブル700の書込状態フラグが「1」か否かを判断する（ステップS1802）。ここで、書込状態フラグが「0」の場合（ステップS1802：No）、本フローチャートによる一連の処理を終了する。

[0259] 一方、書込状態フラグが「1」の場合（ステップS1802：Yes）、書込指示部509により、データ書込処理を実行する（ステップS1803）。データ書込処理の具体的な処理手順については後述する。このあと、成否判定部506により、データ書込処理が成功したか否かを判断する（ステップS1804）。

[0260] ここで、データ書込処理に成功した場合（ステップS1804：Yes）、更新部502により、書込状態テーブル700の書込状態フラグを「1」から「0」に変更して（ステップS1805）、本フローチャートによる一連の処理を終了する。一方、データ書込処理に失敗した場合（ステップS1804：No）、本フローチャートによる一連の処理を終了する。

[0261] つぎに、ステップS1803のデータ書込処理の具体的な処理手順の一例について説明する。ステップS1803のデータ書込処理は、図16に示したステップS1606の書込処理を、どのステップを経由して実行したかによって異なる。ここでは一例として、図16に示した（ステップS1611：No）を経由した場合のデータ書込処理の具体的な処理手順について説明する。

[0262] 図19は、ステップS1803のデータ書込処理の具体的な処理手順の一例を示すフローチャートである。図19のフローチャートにおいて、まず、書込指示部509により、HDDコントローラ309を介して、書込バッファB1に複写されたデータDiを要求先のLBAjに書き込む（ステップS1901）。

[0263] つぎに、書込指示部509により、HDDコントローラ309を介して、

書込バッファB 2に複写されたパリティデータP_jを複写元のLBA_jに書き込む（ステップS 1 9 0 2）。このあと、書込指示部5 0 9により、HDDコントローラ3 0 9を介して、書込バッファB 3に複写された書込データD_i #を生成元とするパリティデータP_j #を算出する（ステップS 1 9 0 3）。

[0264] つぎに、書込指示部5 0 9により、HDDコントローラ3 0 9を介して、書込バッファB 3に複写された書込データD_i #を要求先のLBA_jに書き込む（ステップS 1 9 0 4）。そして、書込指示部5 0 9により、HDDコントローラ3 0 9を介して、算出されたパリティデータP_j #を、パリティデータP_jの複写元のLBA_jに書き込んで（ステップS 1 9 0 5）、図1 8に示したステップS 1 8 0 4に移行する。

[0265] これにより、書込バッファBに退避したデータD_iとパリティデータP_jを退避元に書き戻して再起動後のHDD間のデータの整合性を確保した後、書込バッファBに退避した書込データD_i #を用いて、書込要求に対する書込処理を再実行することができる。

[0266] また、ステップS 1 8 0 4のデータ書込処理に失敗した場合は、新たな書込要求を受け付けるまでの間、書込バッファBに退避されたデータD_i、パリティデータP_jおよび書込データD_i #を読出データとして用いることができる。また、ステップS 1 8 0 4のデータ書込処理に失敗した場合、HDDを再起動しても書込要求に対する書込処理に失敗したことを示すメッセージを、図3に示したディスプレイ3 0 4に表示して、ユーザに報知することにしてもよい。なお、上述したシステム復旧処理の実行後、障害が発生しているHDDを交換してリビルドを行うことにより、ストレージシステム3 0 0を冗長構成に戻すことにしてもよい。

[0267] 以上説明した実施の形態1にかかるストレージ制御装置1 0 1によれば、書込要求に対する書込処理の実行前に、要求先のHDD_kのLBA_jに記憶されているデータD_iを書込バッファBに複写することができる。また、ストレージ制御装置1 0 1によれば、書込要求に対する書込処理の実行前に、

データ D_i を生成元とするパリティデータ P_j を書込バッファ B に複写することができる。これにより、パリティデータ P_j の更新を伴うデータ D_i の更新時に、更新前のデータ D_i とパリティデータ P_j を書込バッファ B に退避しておくことができる。

[0268] また、ストレージ制御装置 101 によれば、書込要求に対する書込処理が実行された HDD が再起動された場合、書込バッファ B に複写されたデータ D_i を要求先の HDD k の LBA j に書き込むことができる。また、ストレージ制御装置 101 によれば、書込要求に対する書込処理が実行された HDD が再起動された場合、書込バッファ B に複写されたパリティデータ P_j を複写元の LBA j に書き込むことができる。これにより、サーマル・オフトラック等の障害により HDD に対する書き込みに失敗しても、HDD の再起動後に、書込バッファ B に退避したデータ D_i とパリティデータ P_j を退避元の LBA j に書き戻して、HDD 間のデータの整合性をとることができる。

[0269] また、ストレージ制御装置 101 によれば、書込要求に対する書込処理に失敗し、かつ、書込処理が実行された HDD が再起動された場合に、書込バッファ B に複写したデータ D_i とパリティデータ P_j を複写元の LBA j に書き込むことができる。これにより、書込要求に対する書込処理が成功した後、瞬時停電等の理由で HDD が再起動されたとしても、書込バッファ B に退避したデータ D_i とパリティデータ P_j を退避元の LBA j に書き戻すことを防ぐことができる。

[0270] また、ストレージ制御装置 101 によれば、書込要求に対する書込処理の実行前に、書込要求に含まれる書込データ $D_i \#$ を、書込バッファ B に複写することができる。また、ストレージ制御装置 101 によれば、書込バッファ B に複写したデータ D_i とパリティデータ P_j を複写元の LBA j に書き戻した後、書込バッファ B に複写した書込データ $D_i \#$ を用いて、書込要求に対する書込処理を再実行することができる。これにより、データ D_i とパリティデータ P_j を退避元に書き戻して再起動後の HDD 間のデータの整合

性を確保した後、書込要求に対する書込処理を再実行することができる。

[0271] また、ストレージ制御装置 101 によれば、要求先の HDD k がオフラインの場合、書込データ Di # を書込バッファ B に複写し、書込失敗による HDD の再起動後に、複写した書込データ Di # を用いて、書込要求に対する書込処理を再実行することができる。

[0272] また、ストレージ制御装置 101 によれば、パリティドライブがオフラインの場合、書込データ Di # を書込バッファ B に複写し、書込失敗による HDD の再起動後に、複写した書込データ Di # を用いて、書込要求に対する書込処理を再実行することができる。

[0273] また、ストレージ制御装置 101 によれば、冗長構成時に書込要求を受け付けた場合、書込データ Di # を書込バッファ B に複写し、書込失敗による HDD の再起動後に、複写した書込データ Di # を用いて、書込要求に対する書込処理を再実行することができる。

[0274] これらのことから、実施の形態 1 にかかるストレージ制御装置 101 によれば、ストレージシステムの復旧時における HDD 間のデータの整合性を保障することができる。これにより、書込失敗による HDD 間のデータの不整合が原因で発生する読出データのデータ化けを防いで、システムが起動しなくなったり、システムが誤動作することを回避することができる。

[0275] なお、上述した説明では、ストレージ制御装置 101 が制御部 503 を備えることとしたが、各 HDD 1 ~ HDD 3 が制御部 503 を備えることにしてもよい。この場合、各 HDD 1 ~ HDD 3 の制御部 503 は、例えば、ストレージ制御装置 101 の複写指示部 504、実行部 505 および書込指示部 509 の指示にしたがって、各 HDD 1 ~ HDD 3 に対するアクセスを制御する。

[0276] (実施の形態 2)

つぎに、実施の形態 2 にかかるストレージ制御装置 101 について説明する。実施の形態 2 では、ストレージ制御装置 101 を RAID 5 のストレージシステム 2000 に適用する。RAID 5 は、データ群の各データを複数

のHDDに分散して記録するとともに、データ群から生成されるパリティデータを複数のHDDに分散して記録する方式である。

[0277] すなわち、RAID5のストレージシステム2000では、RAID3のストレージシステムのようにパリティデータを専用のHDD（パリティドライブ）に記録するのではなく、複数のHDDにパリティデータを分散して記録する。なお、実施の形態1で説明した箇所と同一箇所については、同一符号を付して図示および説明を省略する。

[0278] （ストレージシステム2000のハードウェア構成）

図20は、実施の形態2にかかるストレージシステム2000のハードウェア構成を示すブロック図である。図20において、ストレージシステム2000は、CPU301と、メインメモリ302と、ディスプレイコントローラ303と、ディスプレイ304と、通信コントローラ305と、ストレージ制御装置101と、HDD1～HDD3と、を備えている。ストレージ制御装置101は、プロセッサ306と、メモリ307と、不揮発性メモリ308と、HDDコントローラ309と、バスコントローラ310と、を備えている。

[0279] ストレージシステム2000において、各HDD1～HDD3に対するアクセスはストリップ単位で行われる。ここで、ストリップとは、ハードディスクの複数のブロックを束ねた記憶領域である。ブロックとは、ハードディスクを同心円状に区切ったトラックから分割された複数のセクタを束ねた記憶領域である。

[0280] 図20において、HDD1のデータ領域311は、ストリップA、ストリップCおよびストリップP3を含む構成である。また、HDD2のデータ領域312は、ストリップB、ストリップP2およびストリップEを含む構成である。また、HDD3のデータ領域313は、ストリップP1、ストリップDおよびストリップFを含む構成である。

[0281] ここで、HDD1のストリップAにはデータa0が記憶されている。また、HDD2のストリップBにはデータb0が記憶されている。そして、HD

D 3 のストリップ P 1 には、ストリップ A のデータ a 0 とストリップ B のデータ b 0 の排他的論理和をとったパリティデータ p 1 が記憶されている。

[0282] HDD 1 のストリップ C にはデータ c 0 が記憶されている。また、HDD 3 のストリップ D にはデータ d 0 が記憶されている。そして、HDD 2 のストリップ P 2 には、ストリップ C のデータ c 0 とストリップ D のデータ d 0 の排他的論理和をとったパリティデータ p 2 が記憶されている。

[0283] HDD 2 のストリップ E にはデータ e 0 が記憶されている。また、HDD 3 のストリップ F にはデータ f 0 が記憶されている。そして、HDD 1 のストリップ P 3 には、ストリップ E のデータ e 0 とストリップ F のデータ f 0 の排他的論理和をとったパリティデータ p 3 が記憶されている。

[0284] (書込状態テーブル 2 1 0 0 の具体例)

つぎに、実施の形態 2 にかかる書込状態テーブル 2 1 0 0 の具体例について説明する。書込状態テーブル 2 1 0 0 は、書込要求に対する書込処理が正常に完了したか否かを特定するための書込状態情報を記憶する。また、書込状態テーブル 2 1 0 0 は、例えば、図 2 0 に示した不揮発性メモリ 3 0 8 により実現される。

[0285] 図 2 1 は、書込状態テーブル 2 1 0 0 の具体例を示す説明図である。図 2 1 において、書込状態テーブル 2 1 0 0 は、書込状態フラグおよびストリップ番号のフィールドを有する。各フィールドに情報を設定することで、書込状態情報がレコードとして記憶されている。

[0286] 書込状態フラグは、書込要求に対する書込処理が正常に完了したか否かを示すフラグである。書込状態フラグは、初期状態では「0」であり、書込要求を受け付けると「0」から「1」に変更される。また、書込状態フラグは、書込要求に対する書込処理が正常に完了すると「1」から「0」に変更される。ストリップ番号は、書込要求の要求先のストリップの識別子である。

[0287] (ストレージ制御装置 1 0 1 のストレージ制御処理手順)

つぎに、実施の形態 2 にかかるストレージ制御装置 1 0 1 のストレージ制御処理手順について説明する。ここでは、ストレージシステム 2 0 0 0 を銀

行等のオンラインシステムに適用して、ユーザがATM (Automated Teller Machine) を用いて現金の入出金を行う場合を例に挙げて説明する。

[0288] 前提として、ストレージシステム2000の電源が投入された結果、ストリップAのデータa0とストリップBのデータb0が読み込まれてOSが正常に起動した場合を想定する。また、初期状態において、各HDD1～HDD3の状態は「正常」である。以下、冗長構成時のストレージ制御装置101のストレージ制御処理手順について説明する。

[0289] <冗長構成時のストレージ制御処理手順>

図22～図24は、冗長構成時のストレージ制御装置101のストレージ制御処理手順の一例を示すフローチャートである。図22のフローチャートにおいて、まず、受付部501により、ストリップBのデータb0をデータb1に更新するための書込要求を受け付ける（ステップS2201）。書込要求は、要求先のストリップBのストリップ番号「B」と、書込対象となるデータb1を含む。この例では、データb1は、OSが起動したことを示すイベントログである。

[0290] つぎに、更新部502により、図21に示した書込状態テーブル2100の書込状態フラグを「0」から「1」に変更する（ステップS2202）。さらに、更新部502により、受け付けた書込要求に含まれるストリップ番号「B」を、書込状態テーブル2100のストリップ番号のフィールドに書き込む（ステップS2203）。

[0291] そして、複写指示部504により、HDDコントローラ309を介して、書込要求に含まれるデータb1を、各HDD1～HDD3の書込バッファB3に複写する（ステップS2204）。このあと、実行部505により、受け付けた書込要求に対する書込処理を実行する（ステップS2205）。

[0292] 具体的には、例えば、実行部505が、HDDコントローラ309を介して、書込要求に含まれるデータb1をストリップBに書き込む。また、実行部505が、HDDコントローラ309を介して、書込要求に含まれるデー

タ b 1 とストリップ A に記録されているデータ a 0 との排他的論理和をとったパリティデータ p 1 (# 1) をストリップ P 1 に書き込む。ここでは、書込要求に対する書込処理に成功した場合を想定する。

[0293] つぎに、更新部 5 0 2 により、書込状態テーブル 2 1 0 0 の書込状態フラグを「 1 」から「 0 」に変更する（ステップ S 2 2 0 6）。このあと、プロセッサ 3 0 6 により、ストリップ A のデータ a 0 とストリップ B のデータ b 1 を読み込んで、運用プログラムを起動して（ステップ S 2 2 0 7）、図 2 3 に示すステップ S 2 3 0 1 に移行する。

[0294] 図 2 3 のフローチャートにおいて、まず、受付部 5 0 1 により、ストリップ C のデータ c 0 をデータ c 1 に更新するための書込要求を受け付ける（ステップ S 2 3 0 1）。書込要求は、要求先のストリップ C のストリップ番号「 C 」と、書込対象となるデータ c 1 を含む。この例では、データ c 1 は、 1 0 0, 0 0 0 円の入金データである。

[0295] つぎに、更新部 5 0 2 により、書込状態テーブル 2 1 0 0 の書込状態フラグを「 0 」から「 1 」に変更する（ステップ S 2 3 0 2）。さらに、更新部 5 0 2 により、受け付けた書込要求に含まれるストリップ番号「 C 」を、書込状態テーブル 2 1 0 0 のストリップ番号のフィールドに書き込む（ステップ S 2 3 0 3）。

[0296] そして、複写指示部 5 0 4 により、HDD コントローラ 3 0 9 を介して、書込要求に含まれるデータ c 1 を、各 HDD 1 ~ HDD 3 の書込バッファ B 3 に複写する（ステップ S 2 3 0 4）。このあと、実行部 5 0 5 により、受け付けた書込要求に対する書込処理を実行する（ステップ S 2 3 0 5）。

[0297] 具体的には、例えば、実行部 5 0 5 が、HDD コントローラ 3 0 9 を介して、書込要求に含まれるデータ c 1 をストリップ C に書き込む。また、実行部 5 0 5 が、HDD コントローラ 3 0 9 を介して、書込要求に含まれるデータ c 1 とストリップ D に記録されているデータ d 0 との排他的論理和をとったパリティデータ p 2 (# 1) を計算して、ストリップ P 2 に書き込む。ここでは、書込要求に対する書込処理に成功した場合を想定する。

- [0298] つぎに、更新部502により、書込状態テーブル2100の書込状態フラグを「1」から「0」に変更して（ステップS2306）、図24に示すステップS2401に移行する。この結果、ストリップCに100,000円の入金データが記録される。
- [0299] 図24のフローチャートにおいて、まず、受付部501により、ストリップDのデータd0をデータd1に更新するための書込要求を受け付ける（ステップS2401）。書込要求は、要求先のストリップDのストリップ番号「D」と、書込対象となるデータd1を含む。データd1は、5,000円の出金データである。
- [0300] つぎに、更新部502により、書込状態テーブル2100の書込状態フラグを「0」から「1」に変更する（ステップS2402）。さらに、更新部502により、受け付けた書込要求に含まれるストリップ番号「D」を、書込状態テーブル2100のストリップ番号のフィールドに書き込む（ステップS2403）。
- [0301] そして、複写指示部504により、HDDコントローラ309を介して、書込要求に含まれるデータd1を、各HDD1～HDD3の書込バッファB3に複写する（ステップS2404）。このあと、実行部505により、受け付けた書込要求に対する書込処理を実行する（ステップS2405）。
- [0302] 具体的には、例えば、実行部505が、HDDコントローラ309を介して、書込要求に含まれるデータd1をストリップDに書き込む。また、実行部505が、HDDコントローラ309を介して、書込要求に含まれるデータd1とストリップCに記録されているデータc1との排他的論理和をとったパリティデータp2（#2）を計算して、ストリップP2に書き込む。
- [0303] ここでは、書込要求に対する書込処理に成功する場合と、書込要求に対する書込処理に失敗する場合の両方を想定する。失敗例として、ストリップDに対する書込要求に含まれるデータd1の書込中に、瞬時停電などによりストレージシステム2000の電源が切断されて、データd1の一部をストリップDに書き込めなかった場合を想定する。

- [0304] この場合、ストリップDに記憶されているデータは、データd 0およびデータd 1のいずれとも異なるデータd xである。一方、ストリップP 2へのパリティデータ書込みが成功した場合、ストリップP 2に記憶されているパリティデータp 2（# 2）は、データd 1を用いて計算されたものである。したがって、この時点では、ストリップDとストリップP 2との間でデータの不整合が発生している。また、瞬時停電によりストレージシステム2 0 0 0の電源が切断された場合、CPU 3 0 1によってストレージシステム2 0 0 0の電源が再投入されて、HDD 1～HDD 3が再起動される。
- [0305] 図2 4のフローチャートの説明に戻り、このあと、判断部5 0 8により、HDDが再起動されたか否かを判断する（ステップS 2 4 0 6）。ここで、HDDが再起動された場合（ステップS 2 4 0 6：Y e s）、書込指示部5 0 9により、HDDコントローラ3 0 9を介して、書込バッファB 3に複写されたデータd 1をストリップDに書き込む（ステップS 2 4 0 7）。
- [0306] さらに、書込指示部5 0 9により、HDDコントローラ3 0 9を介して、パリティデータp 2（# 2）を計算してストリップP 2に書き込む（ステップS 2 4 0 8）。パリティデータp 2（# 2）は、書込バッファB 3に複写されたデータd 1と、ストリップCに記憶されているデータc 1とから計算される。そして、更新部5 0 2により、書込状態テーブル2 1 0 0の書込状態フラグを「1」から「0」に変更して（ステップS 2 4 0 9）、本フローチャートによる一連の処理を終了する。
- [0307] 一方、ステップS 2 4 0 6において、HDDが再起動されなかった場合（ステップS 2 4 0 6：N o）、更新部5 0 2により、書込状態テーブル2 1 0 0の書込状態フラグを「1」から「0」に変更して（ステップS 2 4 0 9）、本フローチャートによる一連の処理を終了する。
- [0308] このように、書込要求に対する書込処理の実行中に、瞬時停電等により電源が切断された場合であっても、電源投入後に、書込バッファB 3に複写されたデータd 1を用いて、書込要求に対する書込処理を再実行することができる。書込処理の再実行の結果、ストリップDに5, 0 0 0円の出金データ

が記録される。このため、以降において、残高照会があった場合、ストリップCのデータc 1とストリップDのデータd 1から95, 000円の残高データを参照することができる。

[0309] <非冗長構成時のストレージ制御処理手順>

つぎに、非冗長構成時のストレージ制御装置101のストレージ制御処理手順について説明する。ここでは、HDD1が故障してアクセス不能となった場合を例に挙げて説明する。ここで、各HDD1～HDD3のシステム領域314～316に記憶されている管理データ410, 420, 430の更新例について説明する。

[0310] 図25は、管理データ410, 420, 430の更新例を示す説明図（その2）である。図25において、HDD1に障害が発生した場合の各HDD1～HDD3の管理データ410, 420, 430の更新例が示されている。図25中、上段はHDD1～HDD3が正常の状態であるときの管理データ410, 420, 430であり、下段はHDD1に障害が発生した後の管理データ410, 420, 430である。

[0311] 具体的には、管理データ410のHDD1の状態が「正常」から「（異常）」に更新されている。なお、「（異常）」は、HDDコントローラ309が管理データ410を更新できた場合に、管理データ410内のHDD1の状態が「異常」となることを表している。また、管理データ420のHDD1の状態が「正常」から「異常」に更新されている。また、管理データ430のHDD1の状態が「正常」から「異常」に更新されている。

[0312] 図26および図27は、非冗長構成時（HDD1異常時）のストレージ制御装置101のストレージ制御処理手順の一例を示すフローチャートである。図26のフローチャートにおいて、まず、受付部501により、ストリップCのデータc 1をデータc 2に更新するための書込要求を受け付ける（ステップS2601）。書込要求は、要求先のストリップCのストリップ番号「C」と、書込対象となるデータc 2を含む。データc 2は、500, 000円の入金データである。

- [0313] つぎに、更新部502により、書込状態テーブル2100の書込状態フラグを「0」から「1」に変更する（ステップS2602）。さらに、更新部502により、受け付けた書込要求に含まれるストリップ番号「C」を、「1」に変更された書込み状態フラグに対応する書込状態テーブル2100のストリップ番号のフィールドに書き込む（ステップS2603）。
- [0314] そして、複写指示部504により、HDDコントローラ309を介して、書込要求に含まれるデータc2を、各HDD1～HDD3の書込バッファB3に複写する（ステップS2604）。ただし、HDD1の書込バッファB3にアクセス不能な場合は、HDD1の書込バッファB3に対するデータc2の複写は行われない。このあと、実行部505により、受け付けた書込要求に対する書込処理を実行する（ステップS2605）。
- [0315] 具体的には、例えば、実行部505が、HDDコントローラ309を介して、ストリップDからデータd1を読み出す。そして、実行部505が、データd1と書込要求に含まれるデータc2との排他的論理和をとったパリティデータp2（#3）を計算して、ストリップP2に書き込む。ここでは、書込要求に対する書込処理に成功した場合を想定する。
- [0316] つぎに、更新部502により、書込状態テーブル2100の書込状態フラグを「1」から「0」に変更して（ステップS2606）、図27に示すステップS2701に移行する。この結果、ストリップP2に500,000円の入金データが反映されたパリティデータp2（#3）が書き込まれる。
- [0317] 図27のフローチャートにおいて、まず、受付部501により、ストリップDのデータd1をデータd2に更新するための書込要求を受け付ける（ステップS2701）。書込要求は、要求先のストリップDのストリップ番号「D」と、書込対象となるデータd2を含む。データd2は、100,000円の出金データである。
- [0318] つぎに、更新部502により、書込状態テーブル2100の書込状態フラグを「0」から「1」に変更する（ステップS2702）。さらに、更新部502により、受け付けた書込要求に含まれるストリップ番号「D」を、書

込状態テーブル2100のストリップ番号のフィールドに書き込む（ステップS2703）。

[0319] そして、複写指示部504により、HDDコントローラ309を介して、書込要求の要求先のストリップDに記憶されているデータd1を、各HDD1～HDD3の書込バッファB1に複写する（ステップS2704）。ただし、HDD1の書込バッファB1にアクセス不能な場合は、HDD1の書込バッファB1に対するデータd1の複写は行われない。

[0320] また、複写指示部504により、HDDコントローラ309を介して、ストリップDのデータd1を生成元とするストリップP2に記録されているパリティデータp2（#3）を、各HDD1～HDD3の書込バッファB2に複写する（ステップS2705）。ただし、HDD1の書込バッファB2にアクセス不能な場合は、HDD1の書込バッファB2に対するパリティデータp2（#3）の複写は行われない。

[0321] さらに、複写指示部504により、書込要求に含まれるデータd2を、各HDD1～HDD3の書込バッファB3に複写する（ステップS2706）。上記同様に、HDD1の書込バッファB3にアクセス不能な場合は、HDD1の書込バッファB3に対するデータd2の複写は行われない。このあと、実行部505により、受け付けた書込要求に対する書込処理を実行する（ステップS2707）。

[0322] 具体的には、例えば、実行部505が、ストリップP2に記録されているパリティデータp2（#3）とストリップDに記憶されているデータd1との排他的論理和をとって、ストリップCのデータc2を復元する。つぎに、実行部505が、HDDコントローラ309を介して、書込要求に含まれるデータd2をストリップDに書き込む。さらに、実行部505が、HDDコントローラ309を介して、復元されたデータc2と書込要求に含まれるデータd2との排他的論理和をとったパリティデータp2（#4）を計算して、ストリップP2に書き込む。

[0323] 図27のフローチャートに戻り、つぎに、成否判定部506により、書込

要求に対する書込処理が成功したか否かを判定する（ステップS 2 7 0 8）。ここで、書込要求に対する書込処理が成功した場合（ステップS 2 7 0 8 : Y e s）、更新部5 0 2により、書込状態テーブル2 1 0 0の書込状態フラグを「1」から「0」に変更して（ステップS 2 7 0 9）、本フローチャートによる一連の処理を終了する。

[0324] この結果、ストリップDに1 0, 0 0 0円の出金データが記録される。このため、以降において、残高照会があった場合は、ストリップDのデータd 2とストリップP 2のパリティデータp 2（# 3）から4 9 5, 0 0 0円の残高データを復元、参照することができる。

[0325] 一方、書込要求に対する書込処理が失敗した場合（ステップS 2 7 0 8 : N o）、検出部5 0 7により、書き込みに失敗したHDDを障害が発生したHDDとして検出する（ステップS 2 7 1 0）。そして、検出部5 0 7により、HDDコントローラ3 0 9を介して、オンラインのHDDの管理データの障害が発生したHDDの状態を「正常」から「異常」に変更する（ステップS 2 7 1 1）。

[0326] つぎに、検出部5 0 7により、HDDコントローラ3 0 9を介して、障害が発生したHDDをオフラインにする（ステップS 2 7 1 2）。そして、判断部5 0 8により、障害が発生したHDDの台数xが「 $x \geq 2$ 」となったか否かを判断する。図2 7の例では、HDDの台数xが2以上になったと判断されている（ステップS 2 7 1 3）。最後に、判断部5 0 8により、障害が発生したHDDの台数が2台以上となったことをCPU 3 0 1に通知することで、ストレージシステム2 0 0 0がシステムダウンする（ステップS 2 7 1 4）。

[0327] ここで、各HDD 1～HDD 3のシステム領域3 1 4～3 1 6に記憶されている管理データ4 1 0, 4 2 0, 4 3 0の更新例について説明する。ここでは、図2 7に示したステップS 2 7 0 7において、HDD 2に障害が発生して、ストリップP 2に対するパリティデータp 2（# 4）の書き込みに失敗した場合を例に挙げて説明する。具体的には、HDD 2のストリップP 2

にパリティデータ p 2 (# 4) の一部を書き込めなかったため、ストリップ P 2 の記憶内容がパリティデータ p 2 (# 4) とは異なるパリティデータ p 2 x となっている。この場合、HDD 2 が、障害が発生した HDD として検出されてオフラインとなる。

[0328] 図 28 は、管理データ 410, 420, 430 の更新例を示す説明図 (その 3) である。図 28 において、HDD 2 に障害が発生した場合の各 HDD 1 ~ HDD 3 の管理データ 410, 420, 430 の更新例が示されている。図 28 中、上段は HDD 2 に障害が発生する前の管理データ 410, 420, 430 であり、下段は HDD 2 に障害が発生した後の管理データ 410, 420, 430 である。

[0329] 具体的には、管理データ 420 の HDD 2 の状態が「正常」から「(異常)」に更新されている。なお、「(異常)」は、HDD コントローラ 309 が管理データ 420 を更新できた場合に、管理データ 420 内の HDD 2 の状態が「異常」となることを表している。また、管理データ 430 の HDD 2 の状態が「正常」から「異常」に更新されている。一方、既に切り離されている HDD 1 では、管理データ 410 の HDD 2 の状態は「正常」のままとなっている。

[0330] <システム復旧処理手順>

つぎに、ストレージ制御装置 101 のシステム復旧処理手順について説明する。ここでは、システムダウンしたストレージシステム 2000 を非冗長構成に復旧する場合を例に挙げて説明する。前提として、ストレージシステム 2000 のハードリセットを行った後、最後にオフラインとなった HDD 2 をオンラインにして再起動する場合を想定する。この場合、各 HDD 2, HDD 3 の管理データ 420, 430 の HDD 2 の状態が「異常」から「正常」に変更される。なお、HDD 1 の管理データ 410 の HDD 2 の状態は「正常」のままである。

[0331] 図 29 は、ストレージ制御装置 101 のシステム復旧処理手順の一例を示すフローチャートである。図 29 のフローチャートにおいて、まず、判断部

508により、書込要求に対する書込処理が実行されたHDD、この例ではHDD2が再起動されたか否かを判断する（ステップS2901）。

[0332] ここで、HDD2が再起動されるのを待つ（ステップS2901：No）、HDD2が再起動された場合（ステップS2901：Yes）、書込指示部509により、書込状態テーブル2100の書込状態フラグが「1」か否かを判断する（ステップS2902）。ここで、書込状態フラグが「0」の場合（ステップS2902：No）、本フローチャートによる一連の処理を終了する。

[0333] 一方、書込状態フラグが「1」の場合（ステップS2902：Yes）、書込指示部509により、HDDコントローラ309を介して、書込バッファB1に複写されたデータd1をHDD3のストリップDに書き込む（ステップS2903）。また、書込指示部509は、HDDコントローラ309を介して、書込バッファB2に複写されたパリティデータp2（#3）を、再起動したHDD2のストリップP2に書き込む（ステップS2904）。

[0334] このあと、書込指示部509は、HDDコントローラ309を介して、書込バッファB3に複写されたデータd2をストリップDに書き込む（ステップS2905）。また、書込指示部509は、HDDコントローラ309を介して、データd2を生成元とするパリティデータp2（#4）を計算してストリップP2に書き込む（ステップS2906）。

[0335] そして、成否判定部506により、ステップS2903～ステップS2906のデータ書込処理が成功したか否かを判断する（ステップS2907）。ここで、データ書込処理が成功した場合（ステップS2907：Yes）、更新部502により、書込状態テーブル2100の書込状態フラグを「1」から「0」に変更して（ステップS2908）、本フローチャートによる一連の処理を終了する。

[0336] この結果、ストリップDに10,000円の出金データが記録される。このため、以降において、残高照会があった場合は、ストリップDのデータd2とストリップP2のパリティデータp2（#4）から495,000円の

残高データを参照することができる。

- [0337] 一方、ステップS 2 9 0 7において、データ書込処理が失敗した場合（ステップS 2 9 0 7：N o）、本フローチャートによる一連の処理を終了する。この場合、書込バッファB 1～B 3に複写されたデータd 1、パリティデータp 2（# 3）およびデータd 2を読出データとして用いることができる。例えば、ストリップCの読出要求を受け付けた場合、書込バッファB 1に複写されたデータd 1と書込バッファB 2に複写されたパリティデータp 2（# 3）からストリップCのデータc 2を復元することができる。
- [0338] 以上説明した実施の形態2にかかるストレージ制御装置1 0 1によれば、パリティデータの更新を伴うデータの更新時に、更新前のデータとパリティデータを書込バッファB 1， B 2に退避しておくことができる。また、ストレージ制御装置1 0 1によれば、書込失敗によるH D Dの再起動後に、書込バッファB 1， B 2に退避したデータとパリティデータを退避元のストリップに書き戻して、H D D間のデータの整合性をとることができる。また、ストレージ制御装置1 0 1によれば、書込バッファB 1， B 2に退避したデータとパリティデータを退避元のストリップに書き戻した後、書込バッファB 3に退避したデータを用いて、書込要求に対する書込処理を再実行することができる。
- [0339] なお、本実施の形態で説明したストレージ制御方法は、予め用意されたプログラムをパーソナル・コンピュータやワークステーション等のコンピュータで実行することにより実現することができる。本ストレージ制御プログラムは、ハードディスク、フレキシブルディスク、C D－R O M、M O、D V D等のコンピュータで読み取り可能な記録媒体に記録され、コンピュータによって記録媒体から読み出されることによって実行される。また、本ストレージ制御プログラムは、インターネット等のネットワークを介して配布してもよい。
- [0340] また、本実施の形態で説明したストレージ制御装置1 0 1は、スタンダードセルやストラクチャードA S I C（A p p l i c a t i o n S p e c i

f i c I n t e g r a t e d C i r c u i t) 等の特定用途向け I C (以下、単に「A S I C」と称す。)や F P G A 等の P L D (P r o g r a m m a b l e L o g i c D e v i c e) によっても実現することができる。具体的には、例えば、上述したストレージ制御装置 1 0 1 の機能を H D L 記述によって機能定義し、その H D L 記述を論理合成して A S I C や P L D に与えることにより、ストレージ制御装置 1 0 1 を製造することができる。

符号の説明

- [0341] 1 0 0, 3 0 0, 2 0 0 0 ストレージシステム
 1 0 1 ストレージ制御装置
 4 1 0, 4 2 0, 4 3 0 管理データ
 5 0 1 受付部
 5 0 2 更新部
 5 0 3 制御部
 5 0 4 複写指示部
 5 0 5 実行部
 5 0 6 成否判定部
 5 0 7 検出部
 5 0 8 判断部
 5 0 9 書込指示部
 5 1 0 対象判定部
 7 0 0, 2 1 0 0 書込状態テーブル
 B, B 1, B 2, B 3 書込バッファ
 S 1 ~ S 3 ストレージ装置

請求の範囲

[請求項1]

データ群ごとに、前記データ群の各データと前記データ群から生成される訂正符号とを、記憶先のストレージ装置が異なるように記憶するシステム内のストレージ装置に対する書込要求を受け付ける受付部と、

前記受付部が前記書込要求を受け付けた場合、前記ストレージ装置の要求先の記憶領域に記憶されているデータと、当該データを生成元とする、前記ストレージ装置に記憶されている訂正符号とを、前記ストレージ装置から特定の記憶領域に複写する複写部と、

前記複写部によって前記データと前記訂正符号が前記特定の記憶領域に複写された後、前記書込要求に対する書込処理が実行されたストレージ装置が再起動されたか否かを判断する判断部と、

前記判断部によってストレージ装置が再起動されたと判断された場合、前記特定の記憶領域に複写されたデータを前記ストレージ装置の要求先の記憶領域に書き込むとともに、前記特定の記憶領域に複写された訂正符号を当該訂正符号の複写元のストレージ装置の記憶領域に書き込む書込部と、

を備えることを特徴とするストレージ制御装置。

[請求項2]

前記書込要求に対する書込処理の処理結果に基づいて、前記書込処理が成功したか否かを判定する成否判定部を備え、

前記書込部は、

前記成否判定部によって前記書込処理が失敗したと判定され、かつ、前記判断部によってストレージ装置が再起動されたと判断された場合、前記特定の記憶領域に複写されたデータを前記ストレージ装置の要求先の記憶領域に書き込むとともに、前記特定の記憶領域に複写された訂正符号を前記複写元のストレージ装置の記憶領域に書き込むことを特徴とする請求項1に記載のストレージ制御装置。

[請求項3]

前記複写部は、

さらに、前記書込要求に含まれる書込対象データを前記特定の記憶領域に複写し、

前記書込部は、

前記特定の記憶領域に複写されたデータを前記ストレージ装置の要求先の記憶領域に書き込み、かつ、前記特定の記憶領域に複写された訂正符号を前記複写元のストレージ装置の記憶領域に書き込んだ後、前記特定の記憶領域に複写された書込対象データを前記ストレージ装置の要求先の記憶領域に書き込むとともに、当該書込対象データを生成元とする訂正符号を前記複写元のストレージ装置の記憶領域に書き込むことを特徴とする請求項 2 に記載のストレージ制御装置。

[請求項 4]

前記システムの中から、障害が発生したストレージ装置を検出する検出部を備え、

前記複写部は、

前記書込要求を受け付けたストレージ装置が、前記検出部によって検出された障害が発生したストレージ装置の場合、前記書込要求に含まれる書込対象データを前記特定の記憶領域に複写し、

前記判断部は、

前記複写部によって前記書込対象データが前記特定の記憶領域に複写された後、前記書込要求に対する書込処理が実行されたストレージ装置が再起動されたか否かを判断し、

前記書込部は、

前記判断部によってストレージ装置が再起動されたと判断された場合、前記特定の記憶領域に複写された書込対象データを生成元とする訂正符号を算出して、前記ストレージ装置の要求先の記憶領域に記憶されているデータを生成元とする訂正符号の記憶先のストレージ装置の記憶領域に書き込むことを特徴とする請求項 3 に記載のストレージ制御装置。

[請求項 5]

前記複写部は、

前記要求先の記憶領域に記憶されているデータを生成元とする訂正符号の記憶先のストレージ装置が、前記検出部によって検出された障害が発生したストレージ装置の場合、前記書込要求に含まれる書込対象データを前記特定の記憶領域に複写し、

前記判断部は、

前記複写部によって前記書込対象データが前記特定の記憶領域に複写された後、前記書込要求に対する書込処理が実行されたストレージ装置が再起動されたか否かを判断し、

前記書込部は、

前記判断部によってストレージ装置が再起動されたと判断された場合、前記特定の記憶領域に複写された書込対象データを前記ストレージ装置の要求先の記憶領域に書き込むことを特徴とする請求項4に記載のストレージ制御装置。

[請求項6]

前記複写部は、

前記検出部によって検出されたストレージ装置の台数が、前記訂正符号を記憶するストレージ装置の台数未満の場合、前記書込要求に含まれる書込対象データを前記特定の記憶領域に複写し、

前記判断部は、

前記複写部によって前記書込対象データが前記特定の記憶領域に複写された後、前記書込要求に対する書込処理が実行されたストレージ装置が再起動されたか否かを判断し、

前記書込部は、

前記判断部によってストレージ装置が再起動されたと判断された場合、前記特定の記憶領域に複写された書込対象データを前記要求先の記憶領域に書き込むとともに、当該書込対象データを生成元とする訂正符号を算出して、前記要求先の記憶領域に記憶されているデータを生成元とする訂正符号を記憶する記憶領域に書き込むことを特徴とする請求項5に記載のストレージ制御装置。

[請求項7] 前記特定の記憶領域は、前記データ群の各データおよび前記データ群から生成される訂正符号の記憶先の記憶領域とは異なる不揮発性の記憶領域であることを特徴とする請求項1～6のいずれか一つに記載のストレージ制御装置。

[請求項8] 前記特定の記憶領域は、前記システム内の各ストレージ装置に設けられることを特徴とする請求項7に記載のストレージ制御装置。

[請求項9] データ群ごとに、前記データ群の各データと前記データ群から生成される訂正符号とを、記憶先のストレージ装置が異なるように記憶するシステム内のストレージ装置に対する書込要求を受け付ける受付部と、

前記受付部が前記書込要求を受け付けた場合、前記システム内のストレージ装置にアクセスする制御部に対して、前記ストレージ装置の要求先の記憶領域に記憶されているデータと、当該データを生成元とする、前記ストレージ装置に記憶されている訂正符号とを、特定の記憶領域に複写するよう指示する複写指示部と、

前記複写指示部によって指示された前記データと前記訂正符号が前記特定の記憶領域に複写された後、前記書込要求に対する書込処理が実行されたストレージ装置が再起動されたか否かを判断する判断部と、

前記判断部によってストレージ装置が再起動されたと判断された場合、前記制御部に対して、前記特定の記憶領域に複写されたデータを前記ストレージ装置の要求先の記憶領域に書き込むとともに、前記特定の記憶領域に複写された訂正符号を当該訂正符号の複写元のストレージ装置の記憶領域に書き込むよう指示する書込指示部と、

を備えることを特徴とするストレージ制御装置。

[請求項10] データ群ごとに、前記データ群の各データと前記データ群から生成される訂正符号とを、記憶先のストレージ装置が異なるように記憶するシステム内のストレージ装置に対する書込要求を受け付ける受付工

程と、

前記受付工程が前記書込要求を受け付けた場合、前記ストレージ装置の要求先の記憶領域に記憶されているデータと、当該データを生成元とする、前記ストレージ装置に記憶されている訂正符号とを、前記ストレージ装置から特定の記憶領域に複写する複写工程と、

前記複写工程によって前記データと前記訂正符号が前記特定の記憶領域に複写された後、前記書込要求に対する書込処理が実行されたストレージ装置が再起動されたか否かを判断する判断工程と、

前記判断工程によってストレージ装置が再起動されたと判断された場合、前記特定の記憶領域に複写されたデータを前記ストレージ装置の要求先の記憶領域に書き込むとともに、前記特定の記憶領域に複写された訂正符号を当該訂正符号の複写元のストレージ装置の記憶領域に書き込む書込工程と、

をコンピュータが実行することを特徴とするストレージ制御方法。

[請求項11]

データ群ごとに、前記データ群の各データと前記データ群から生成される訂正符号とを、記憶先のストレージ装置が異なるように記憶するシステム内のストレージ装置に対する書込要求を受け付ける受付工程と、

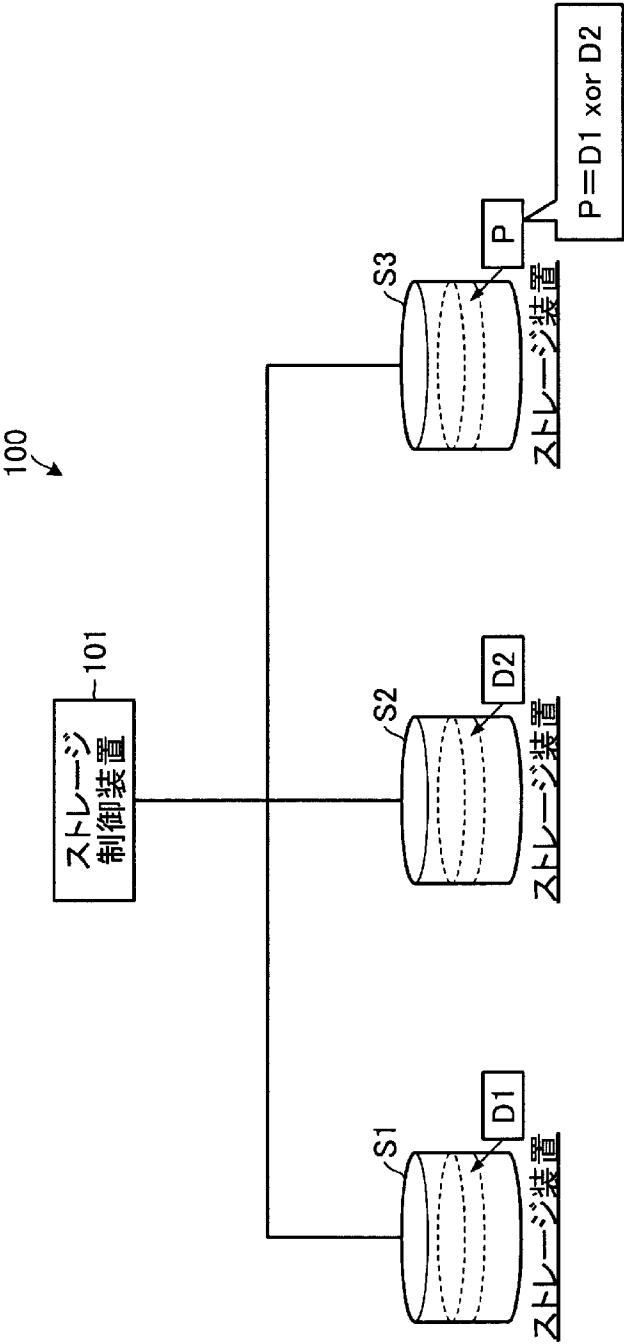
前記受付工程によって前記書込要求を受け付けた場合、前記システム内のストレージ装置にアクセスする制御部に対して、前記ストレージ装置の要求先の記憶領域に記憶されているデータと、当該データを生成元とする、前記ストレージ装置に記憶されている訂正符号とを、特定の記憶領域に複写するよう指示する複写指示工程と、

前記複写指示工程によって指示された前記データと前記訂正符号が前記特定の記憶領域に複写された後、前記書込要求に対する書込処理が実行されたストレージ装置が再起動されたか否かを判断する判断工程と、

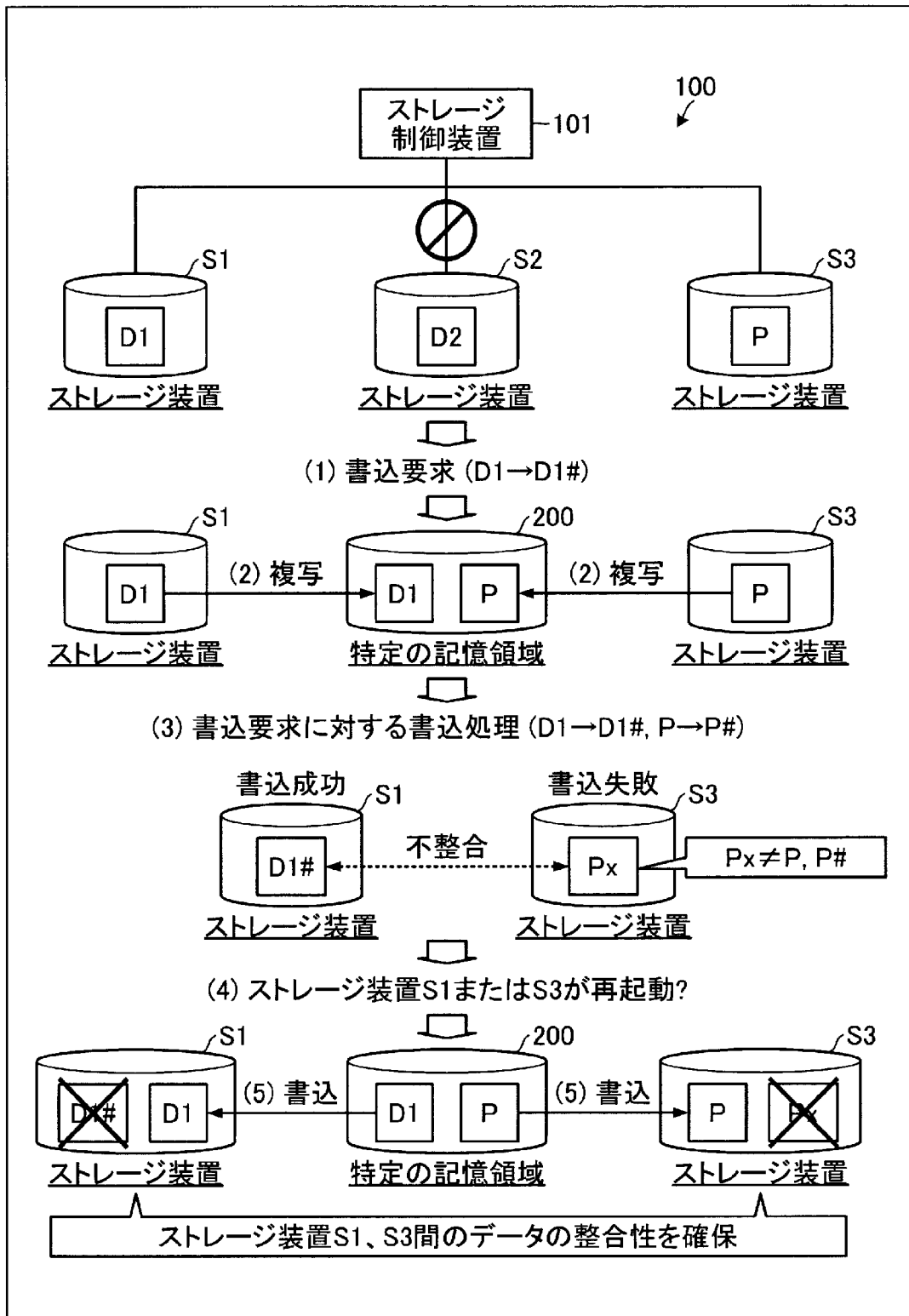
前記判断工程によってストレージ装置が再起動されたと判断された

場合、前記制御部に対して、前記特定の記憶領域に複写されたデータを前記ストレージ装置の要求先の記憶領域に書き込むとともに、前記特定の記憶領域に複写された訂正符号を当該訂正符号の複写元のストレージ装置の記憶領域に書き込むよう指示する書込指示工程と、
をコンピュータが実行することを特徴とするストレージ制御方法。

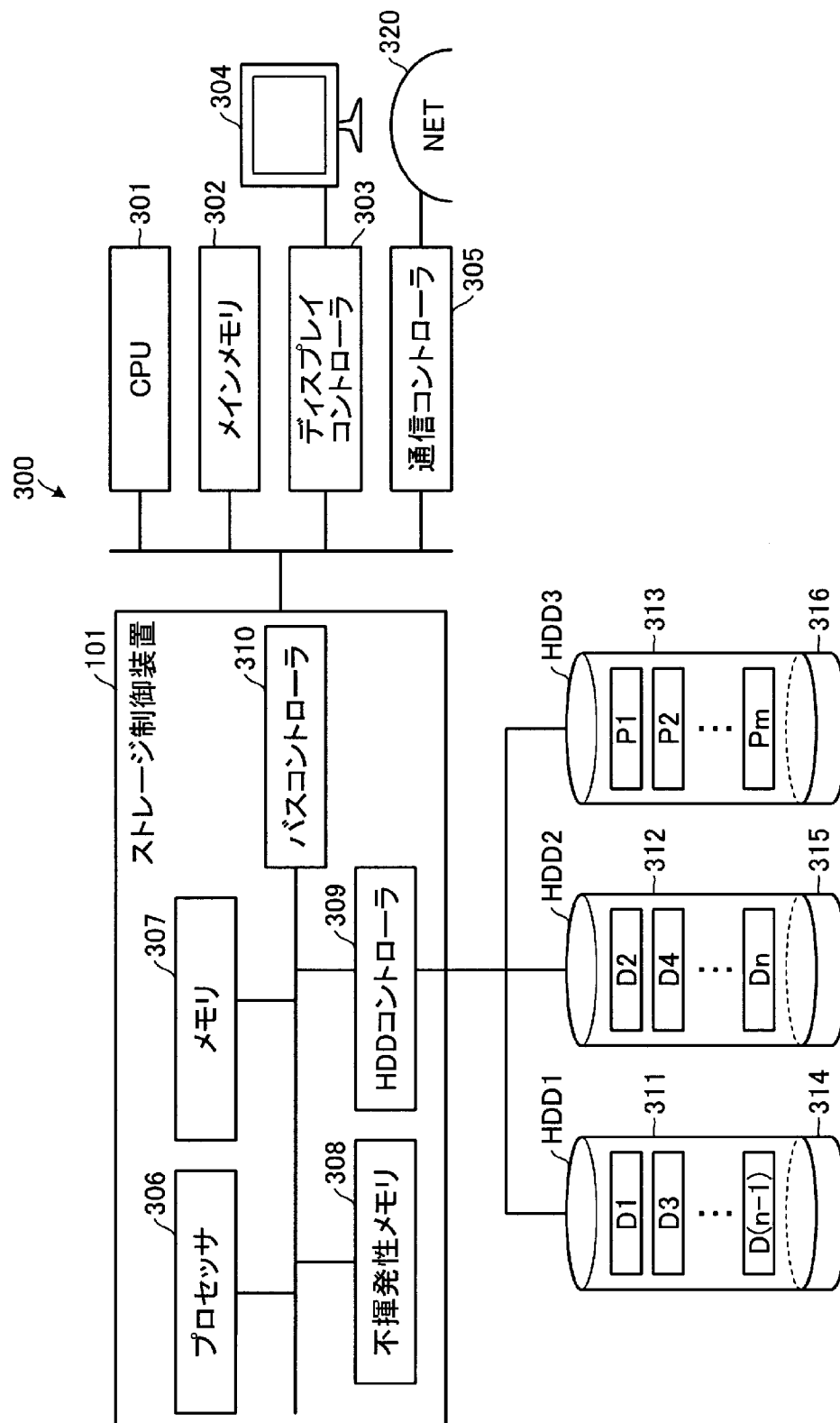
[図1]



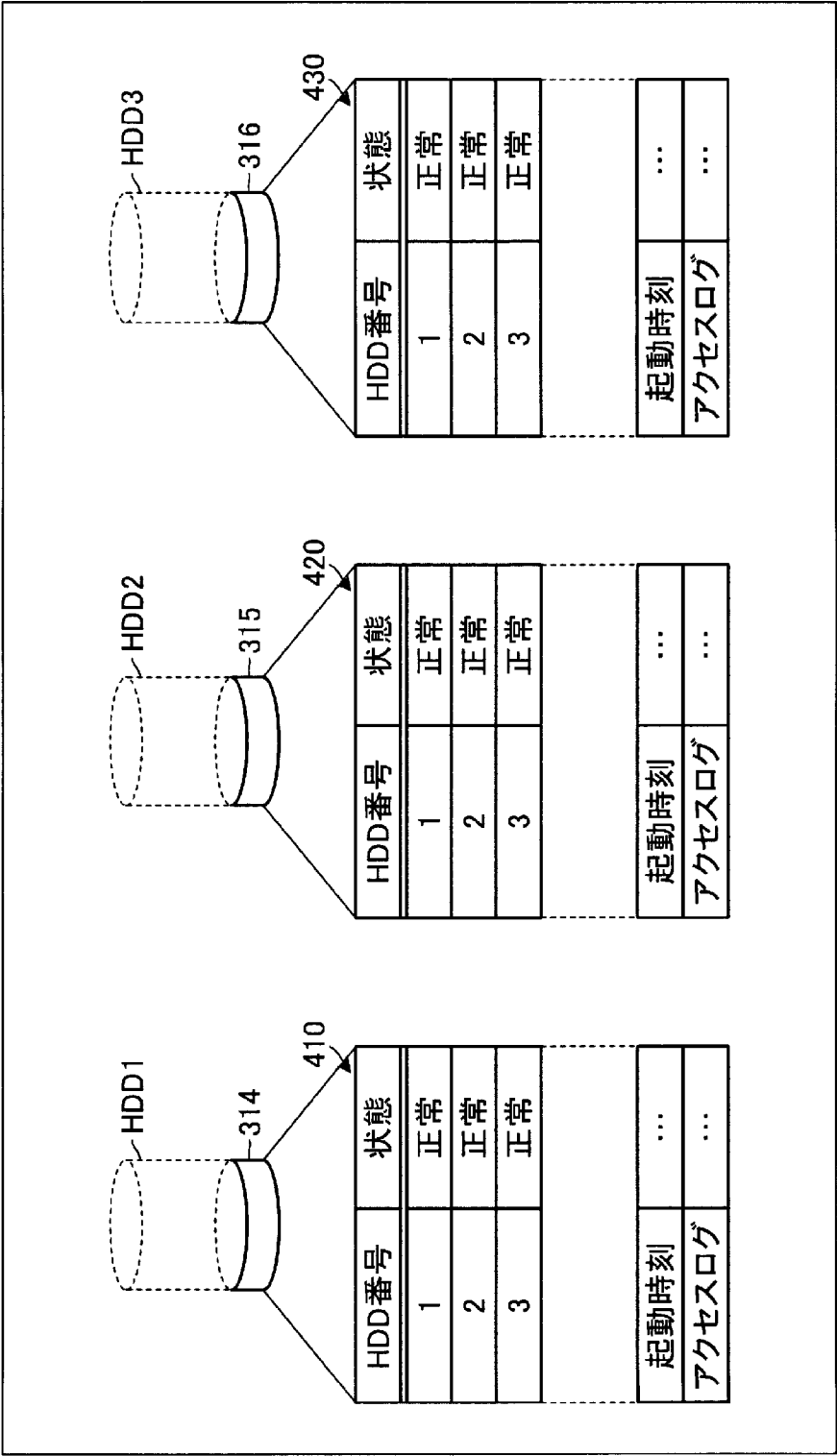
[図2]



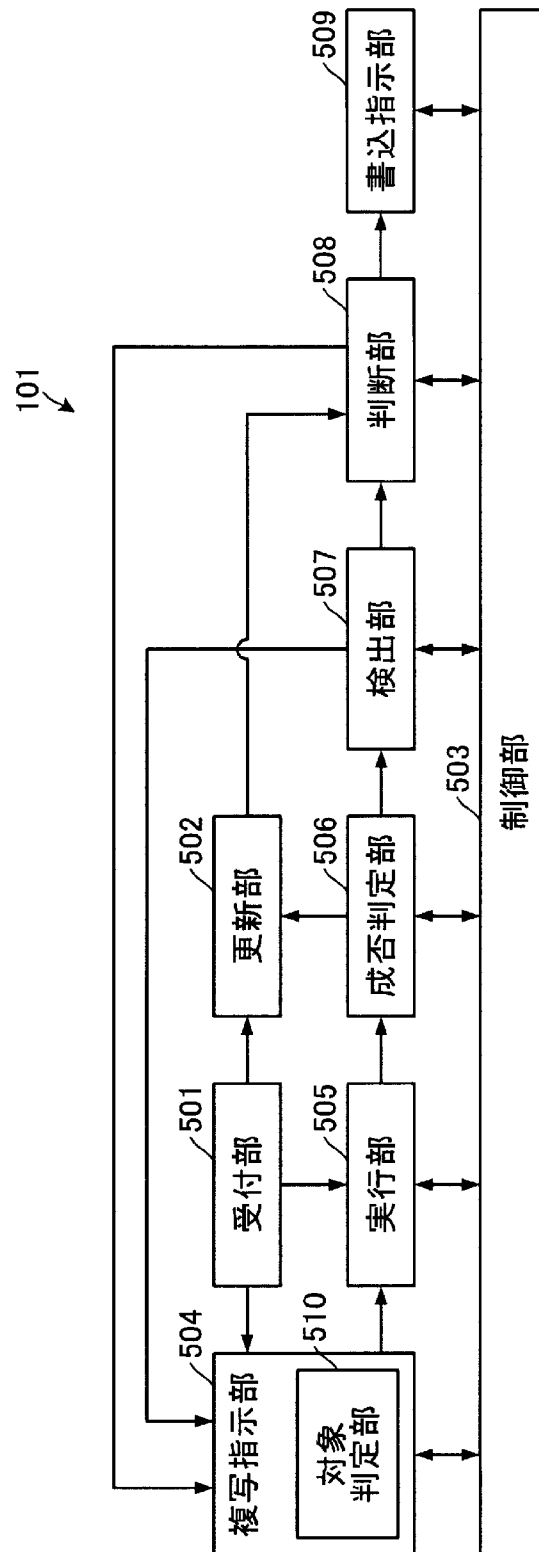
[図3]



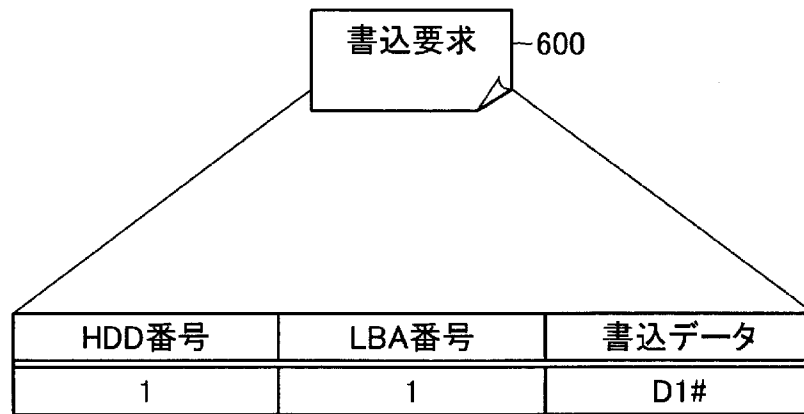
[図4]



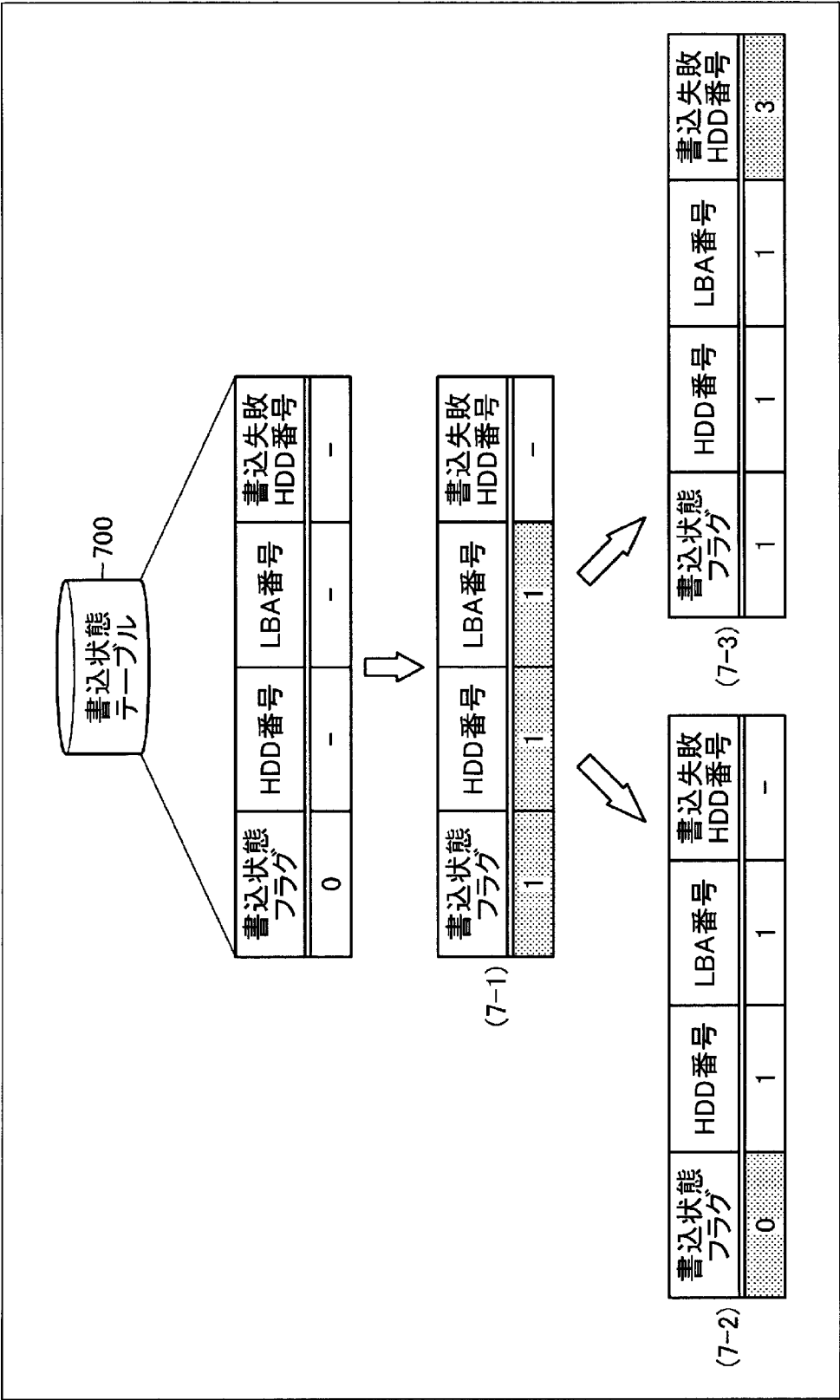
[図5]



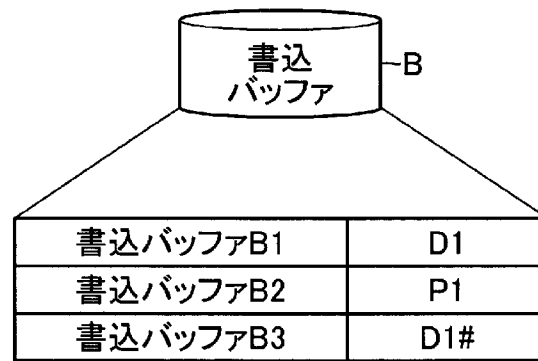
[図6]



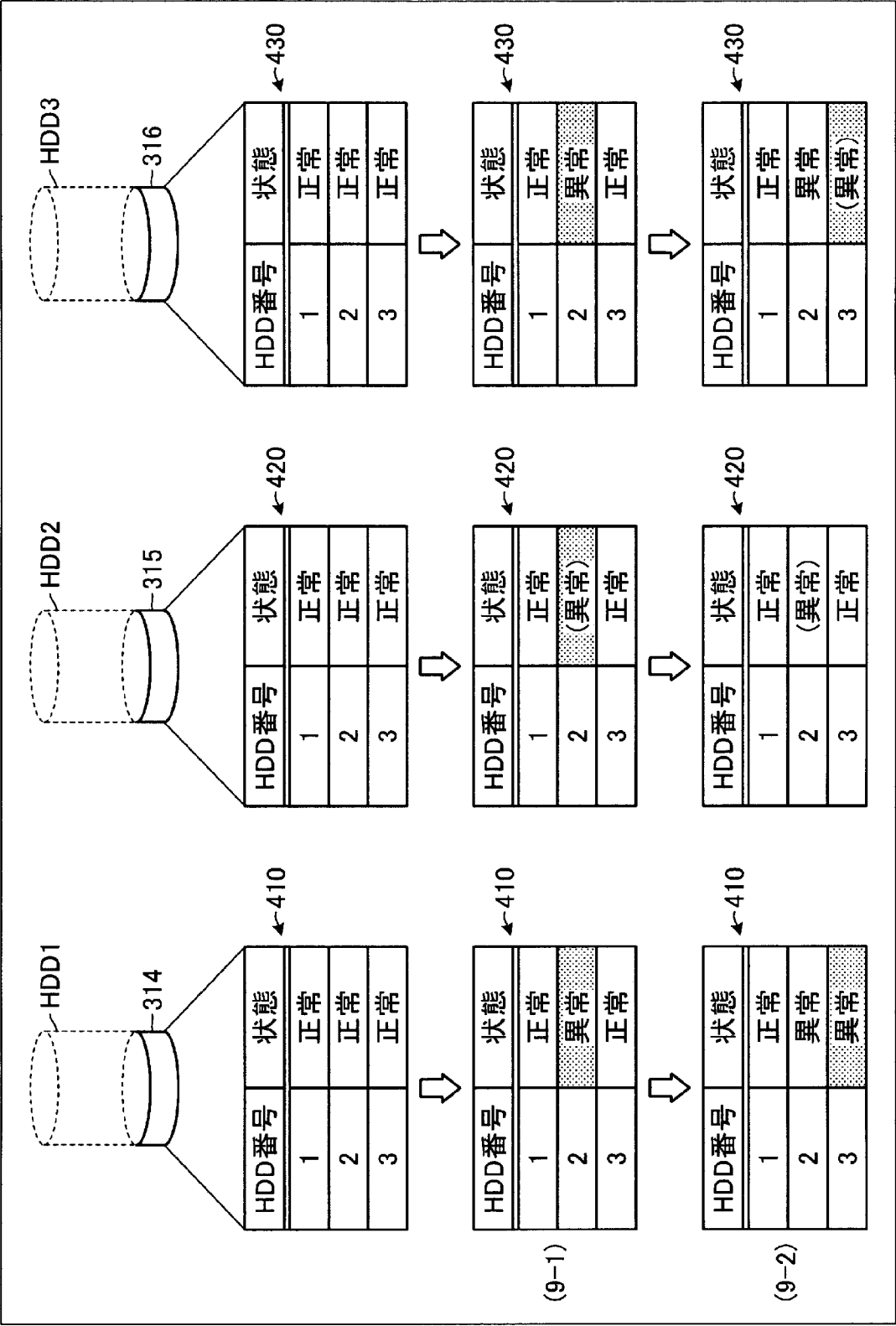
[図7]



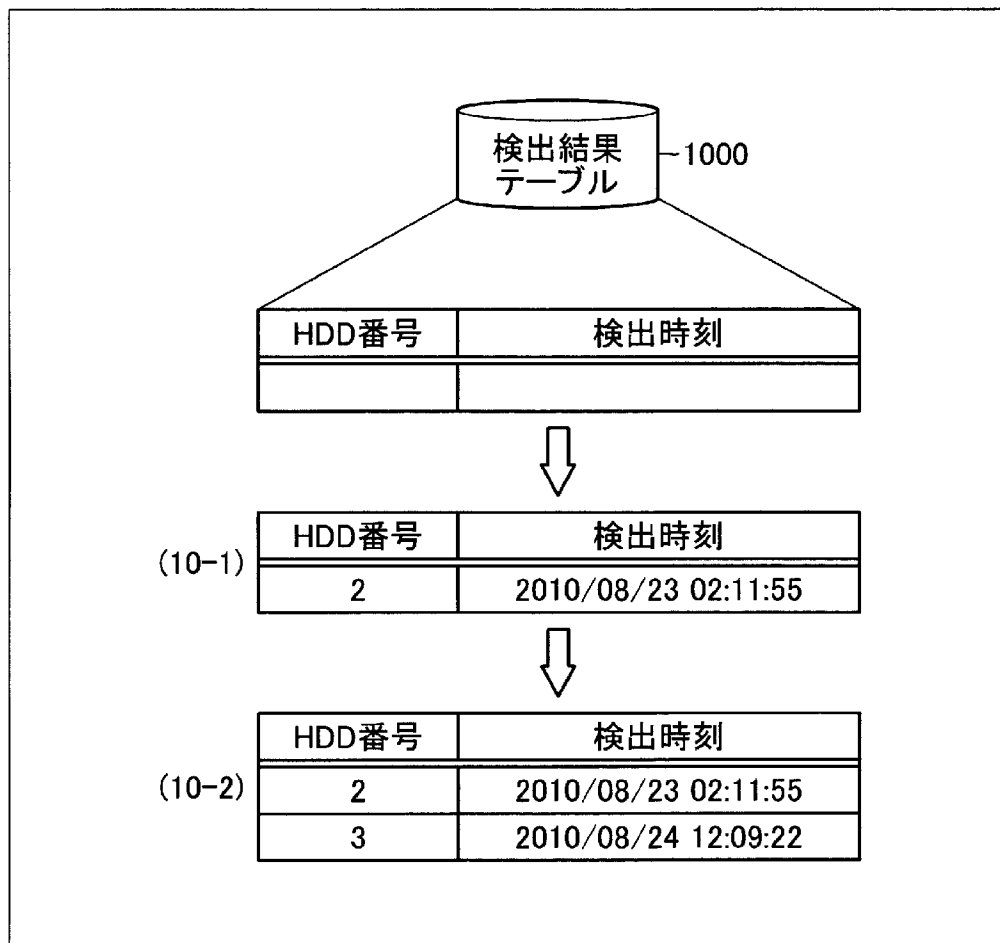
[図8]



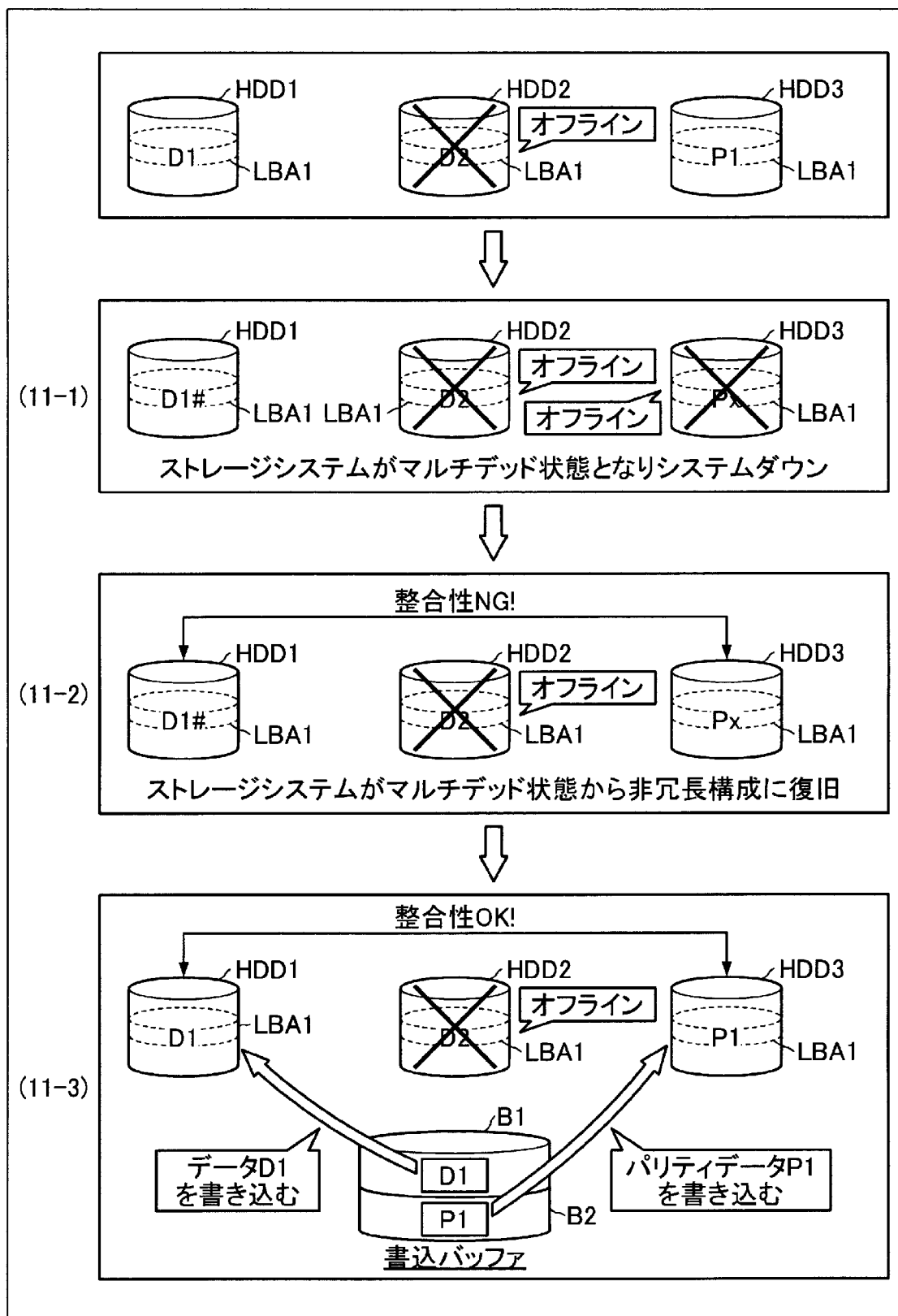
[図9]



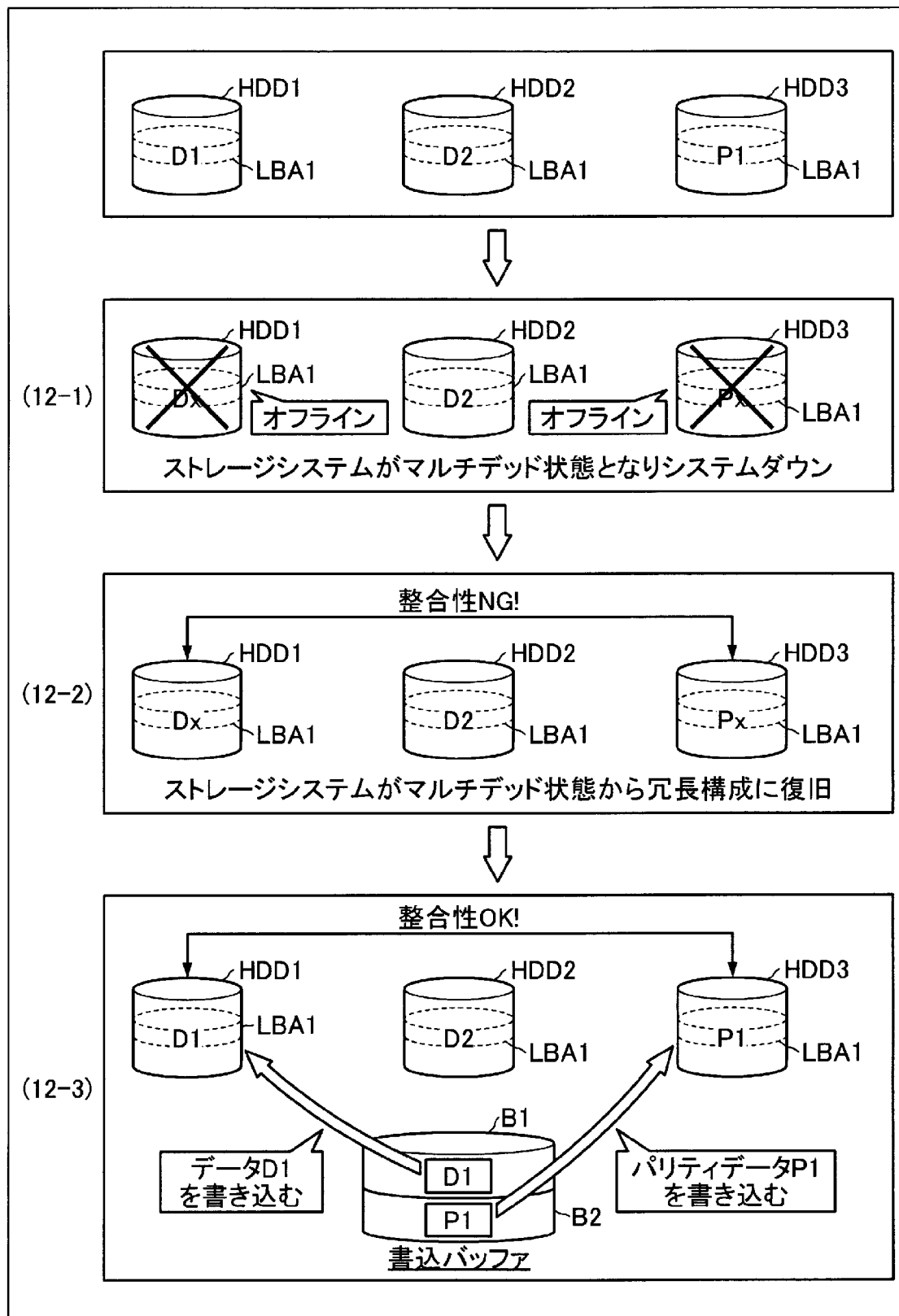
[図10]



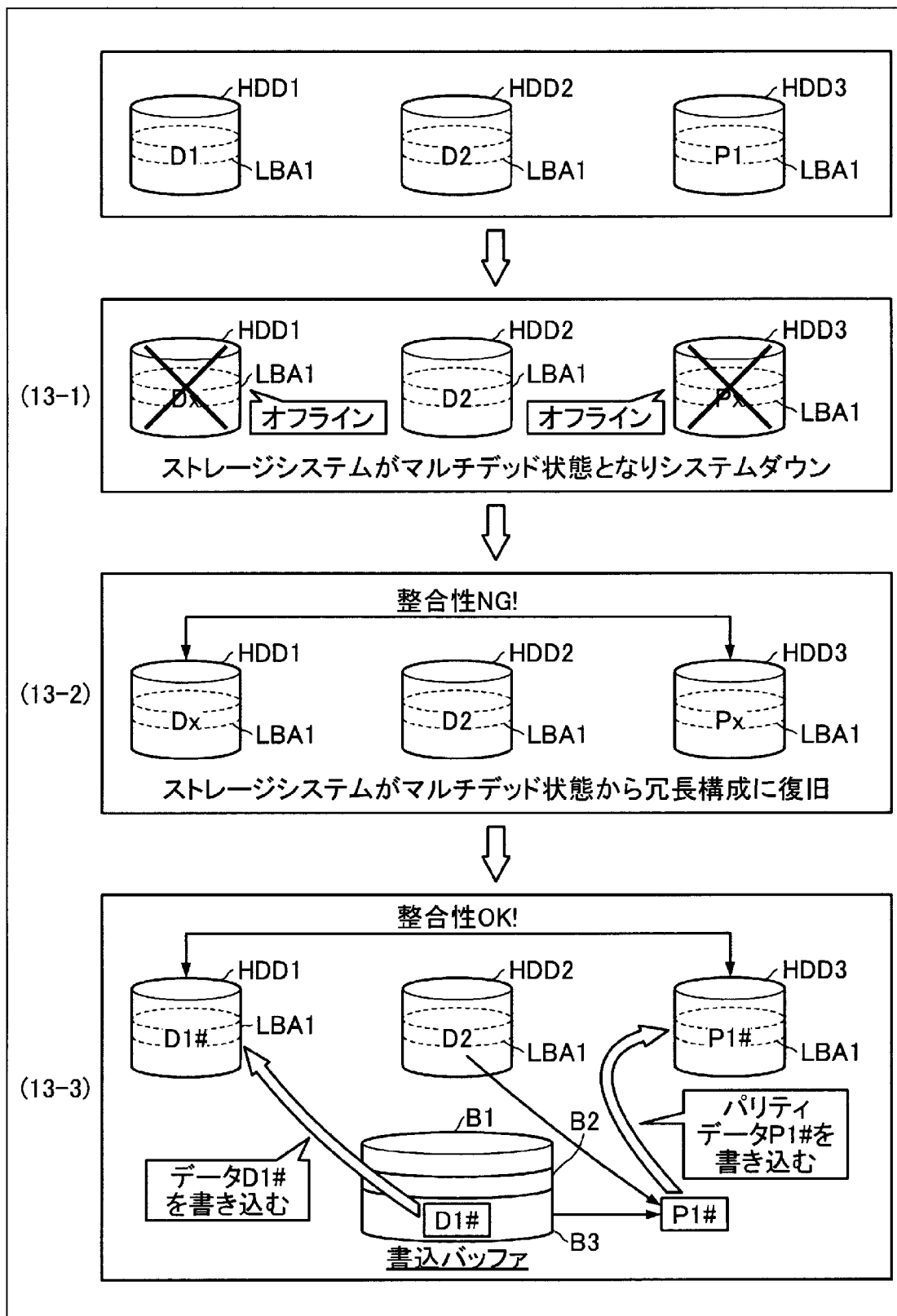
[図11]



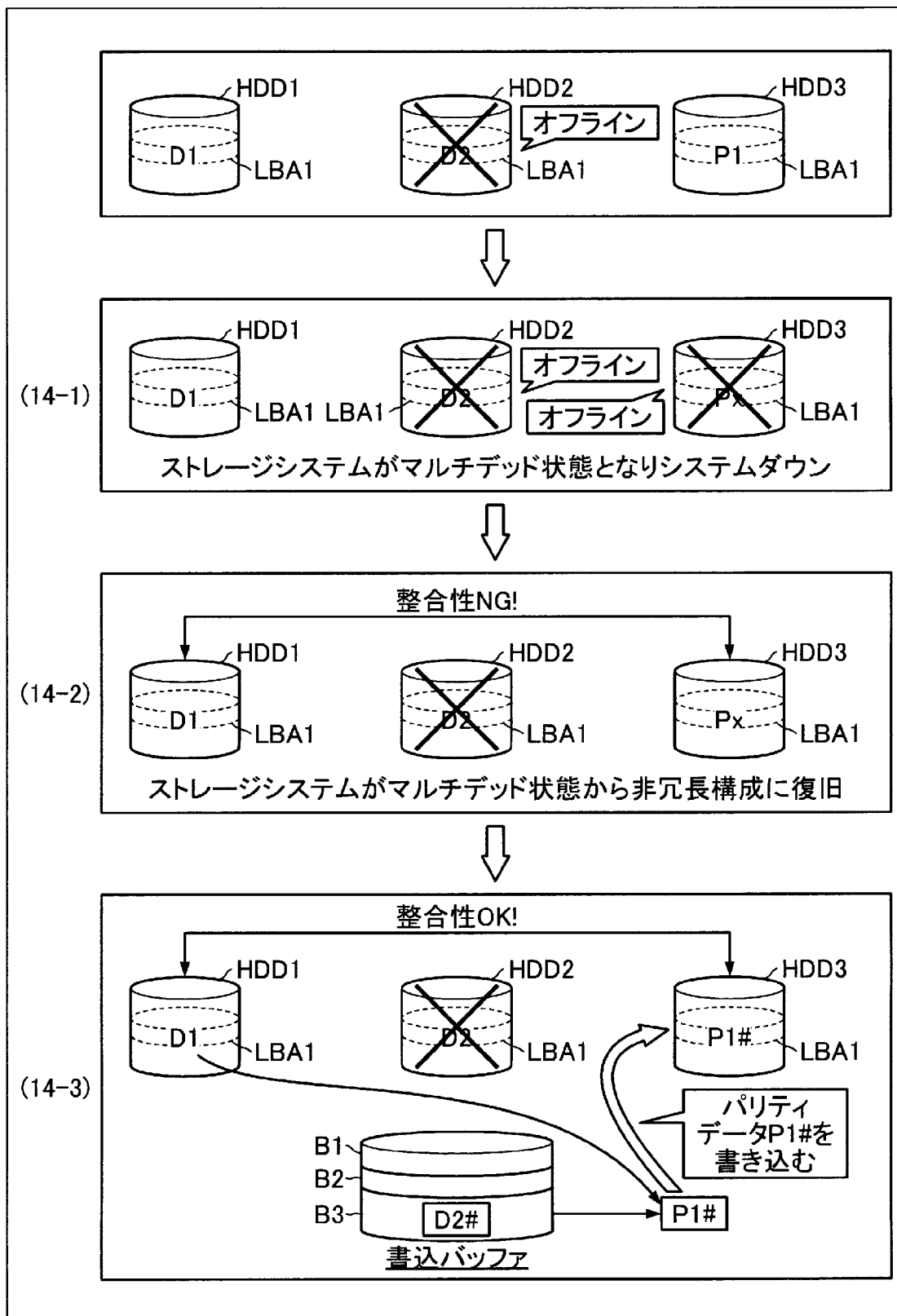
[図12]



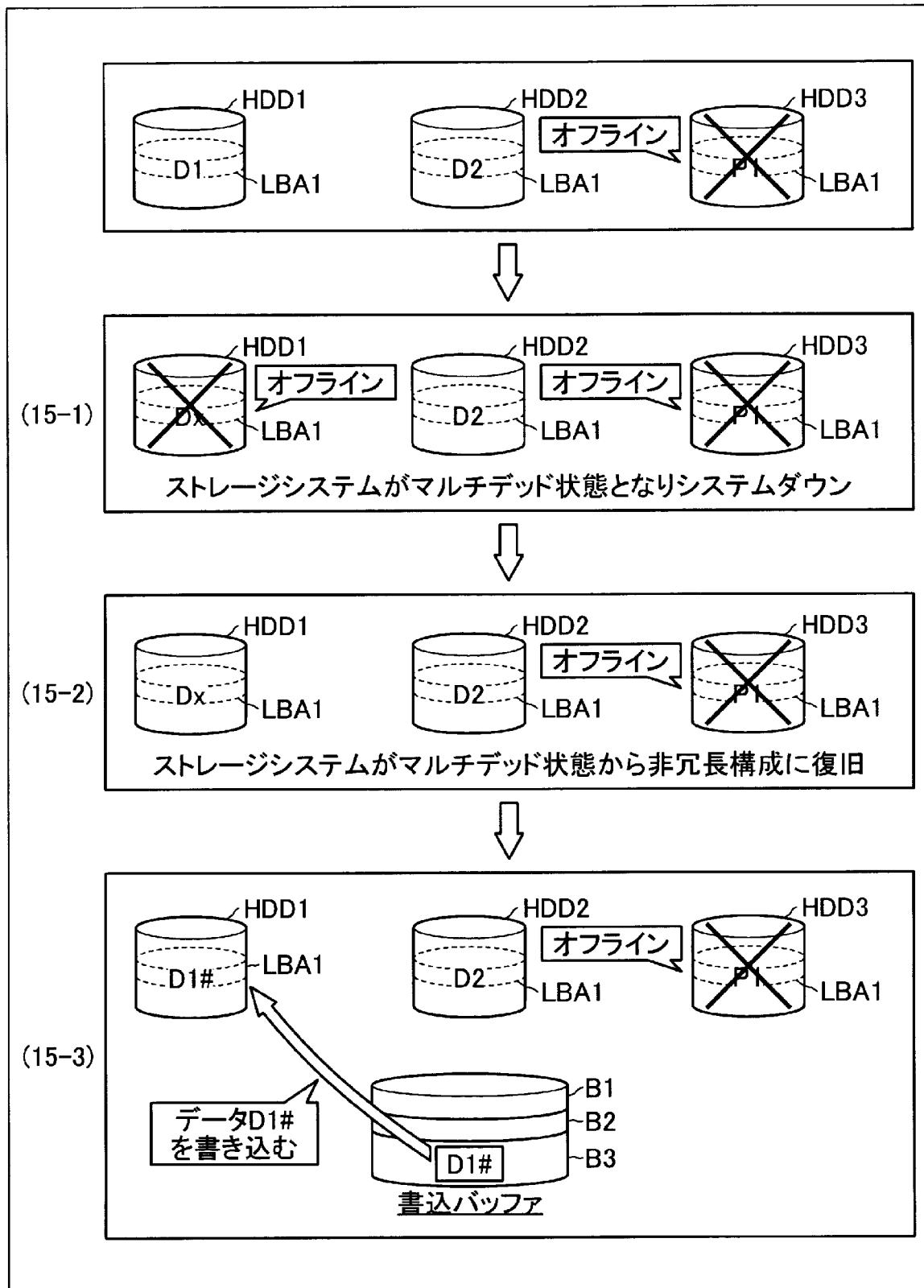
[図13]



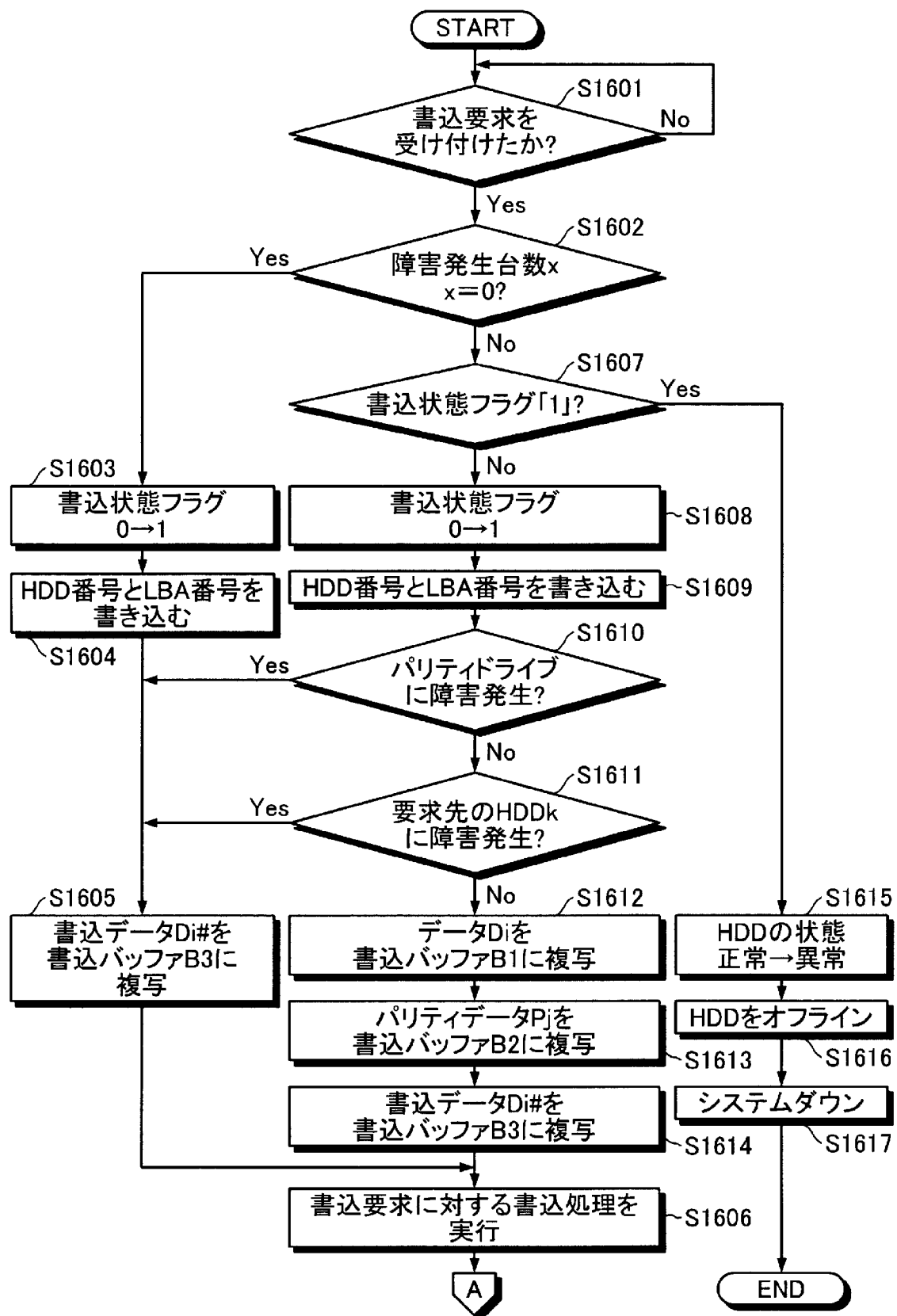
[図14]



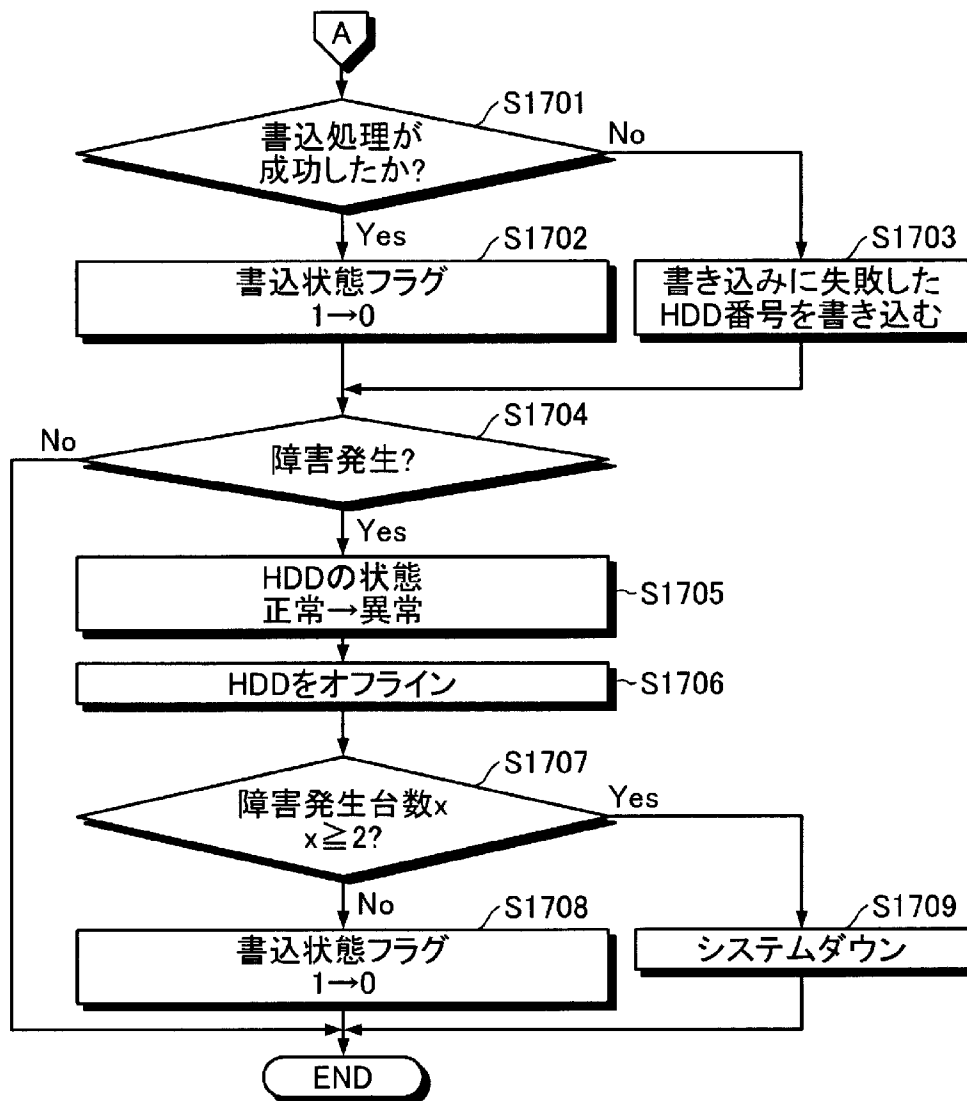
[図15]



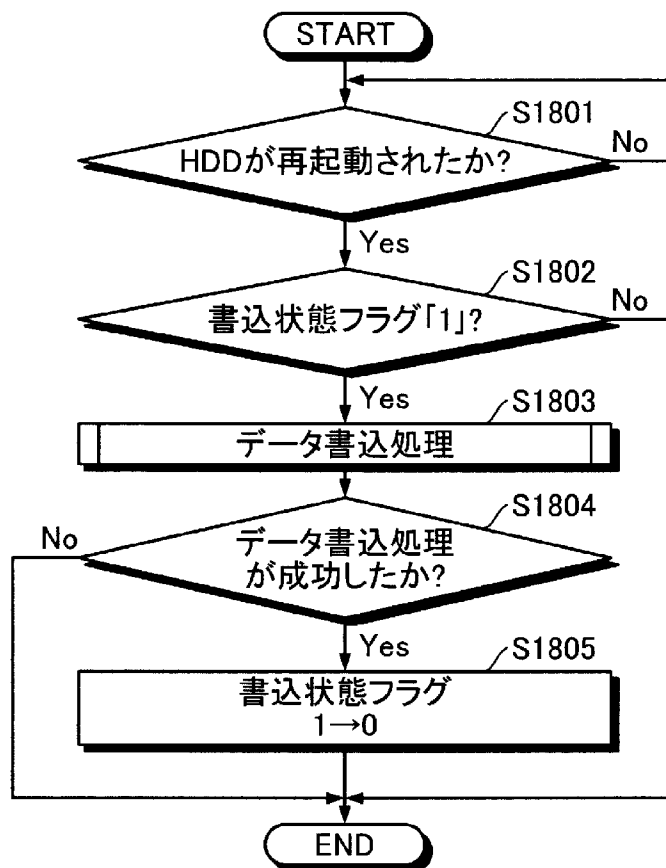
[図16]



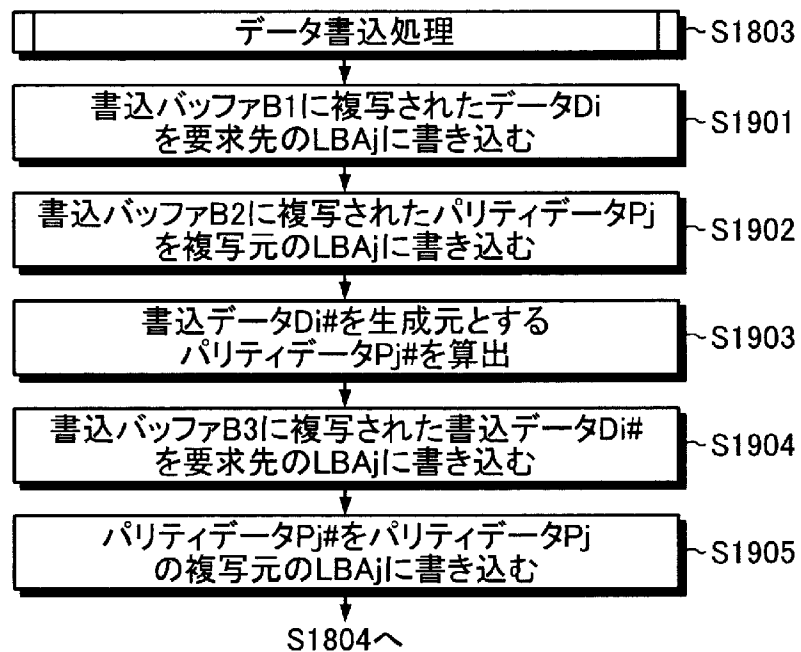
[図17]



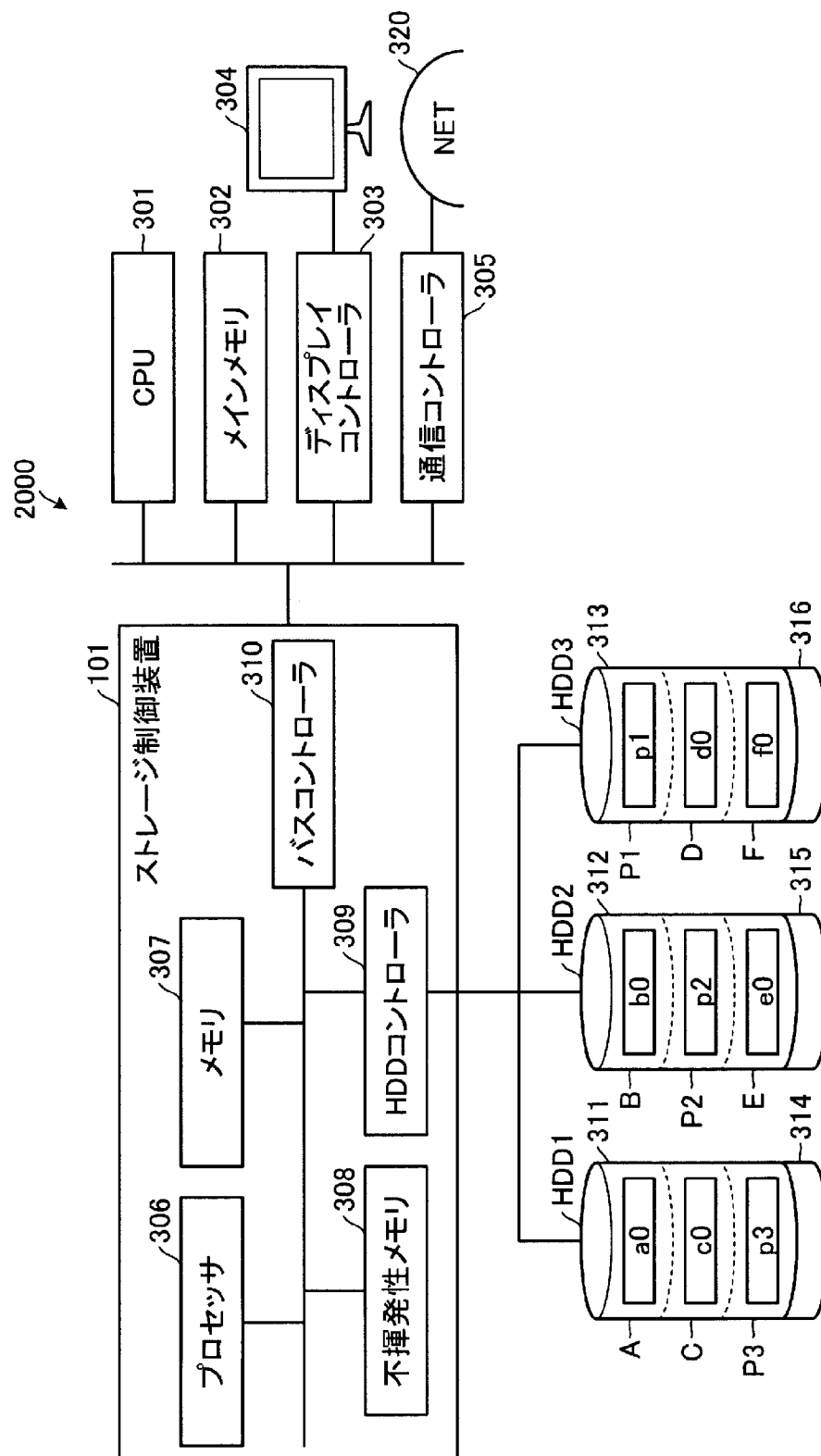
[図18]



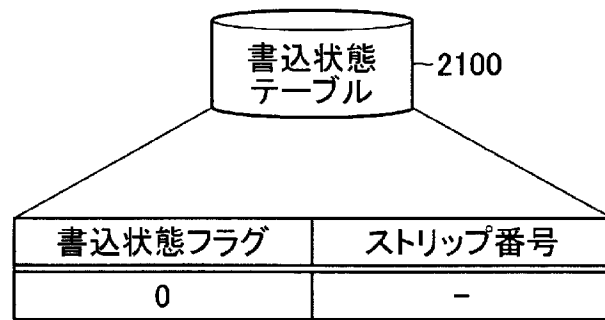
[図19]



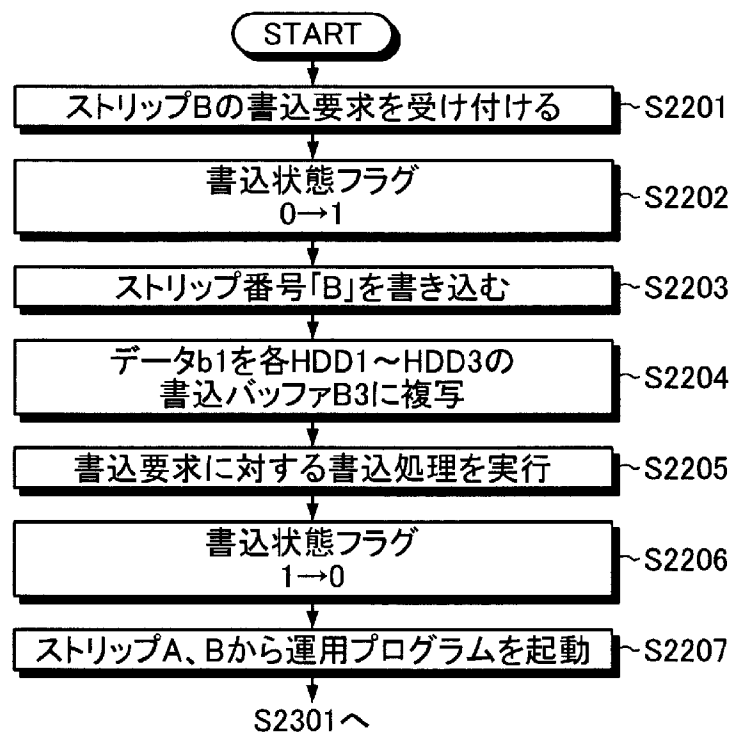
[図20]



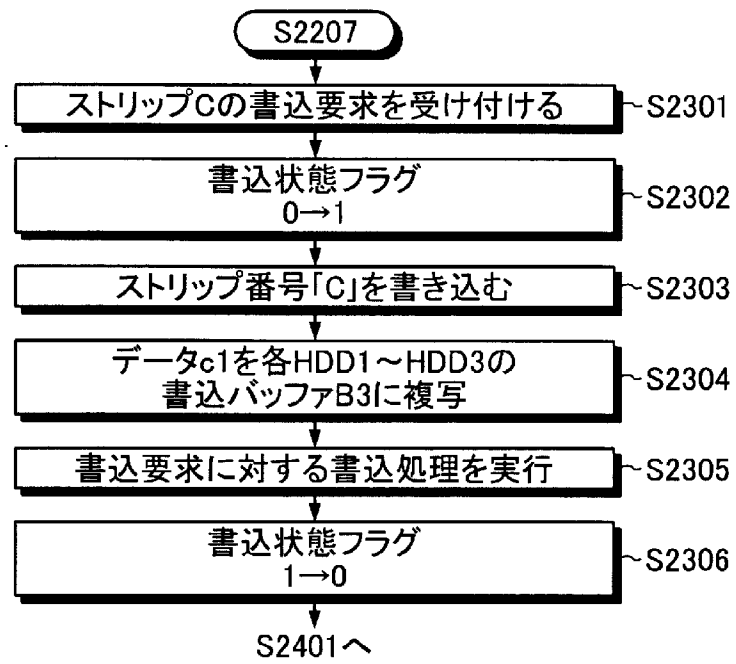
[図21]



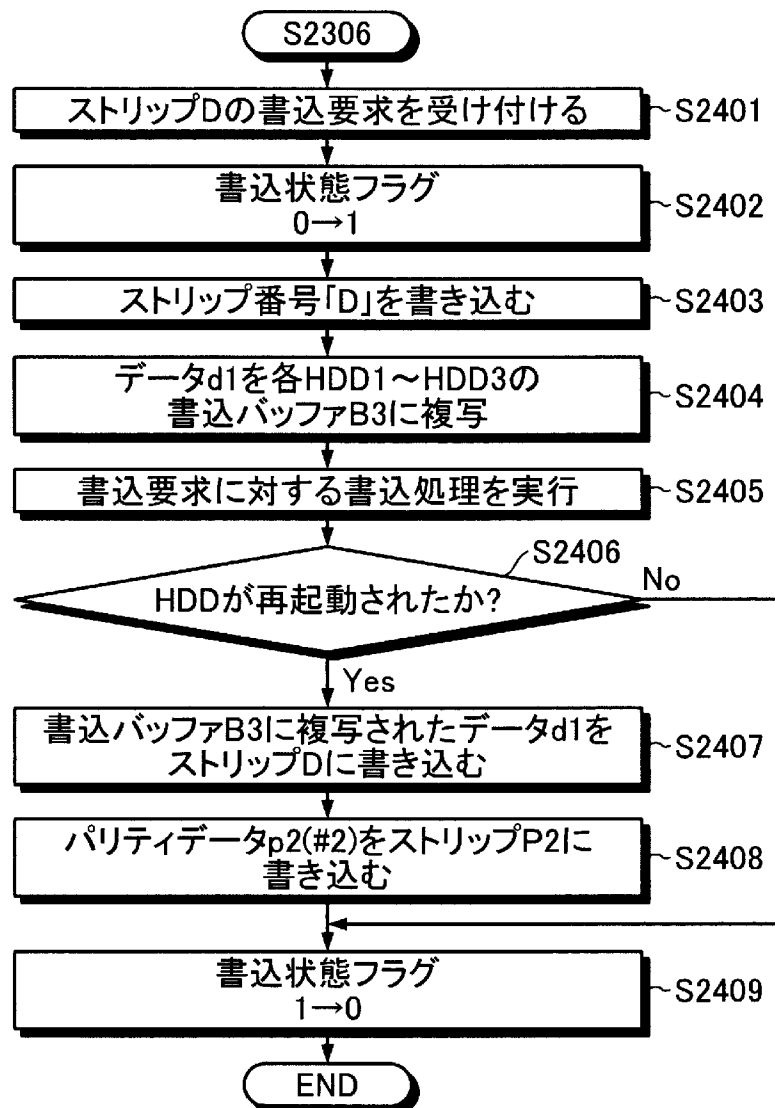
[図22]



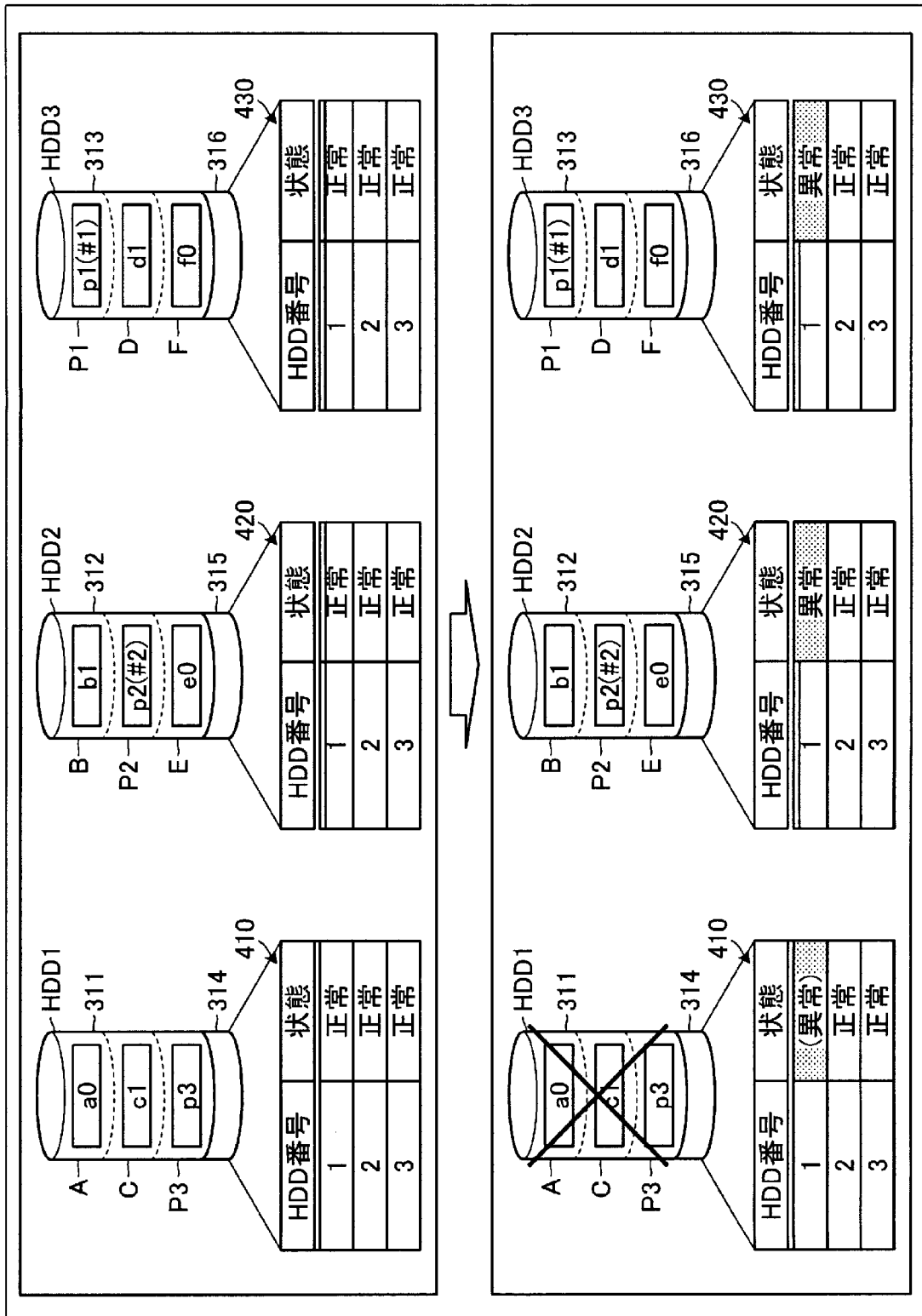
[図23]



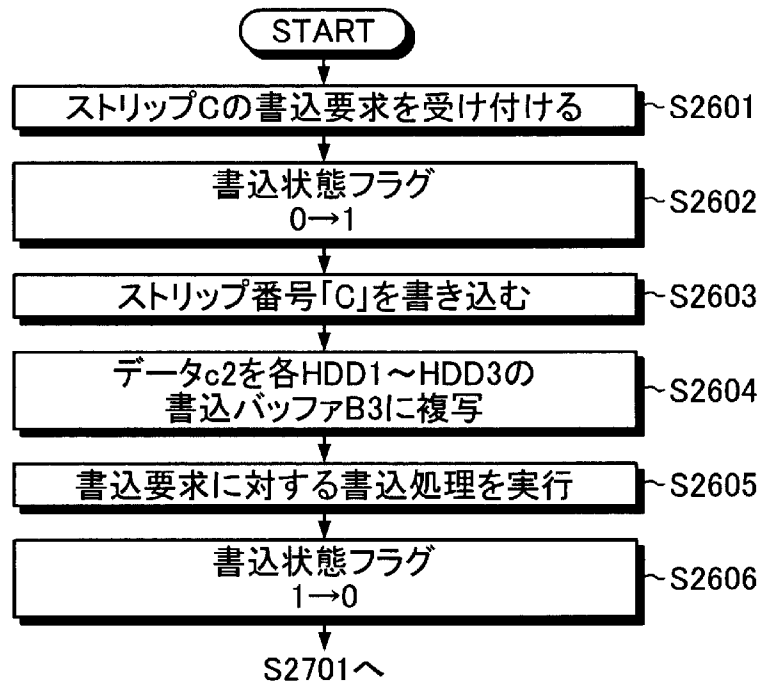
[図24]



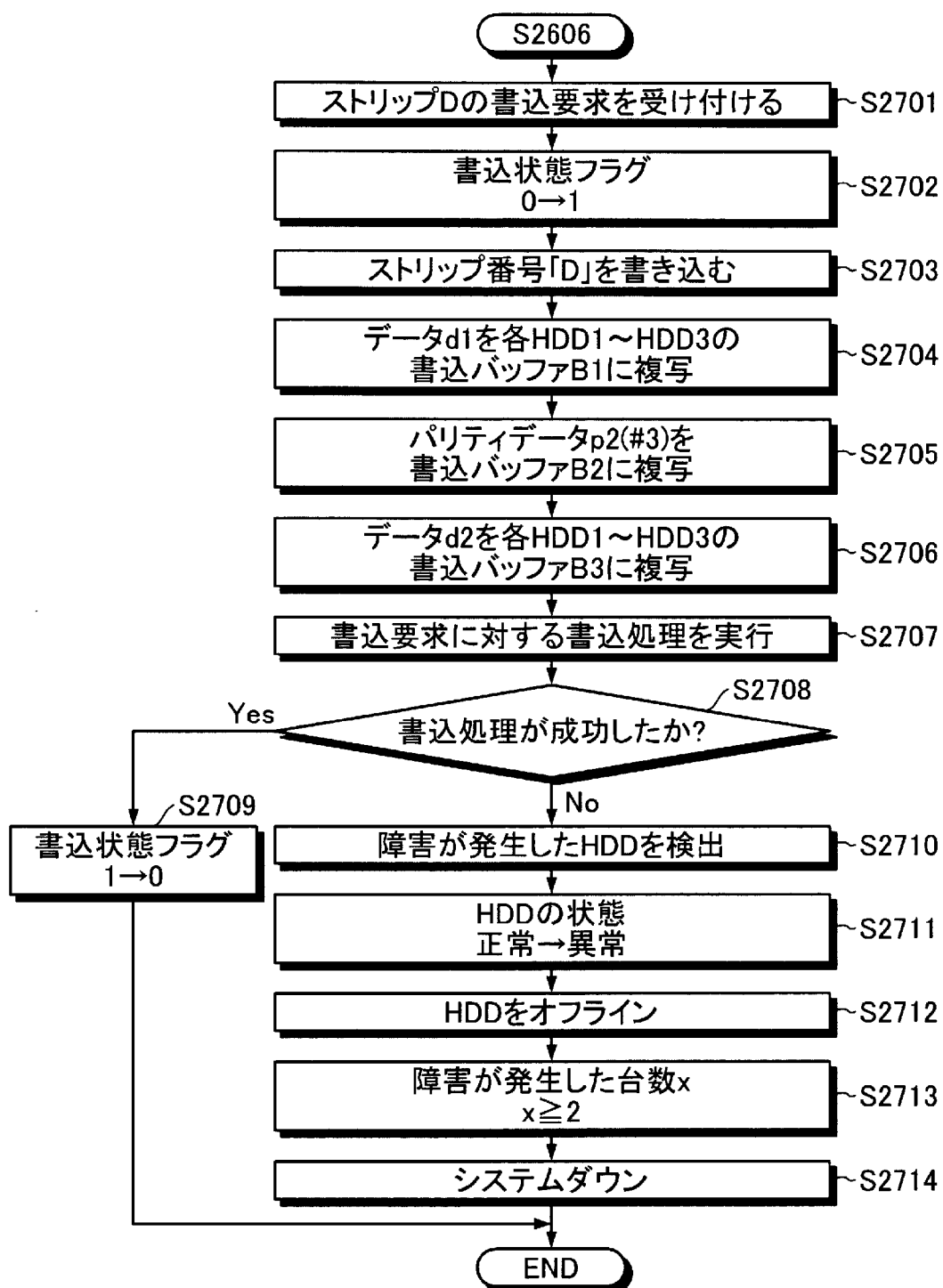
[図25]



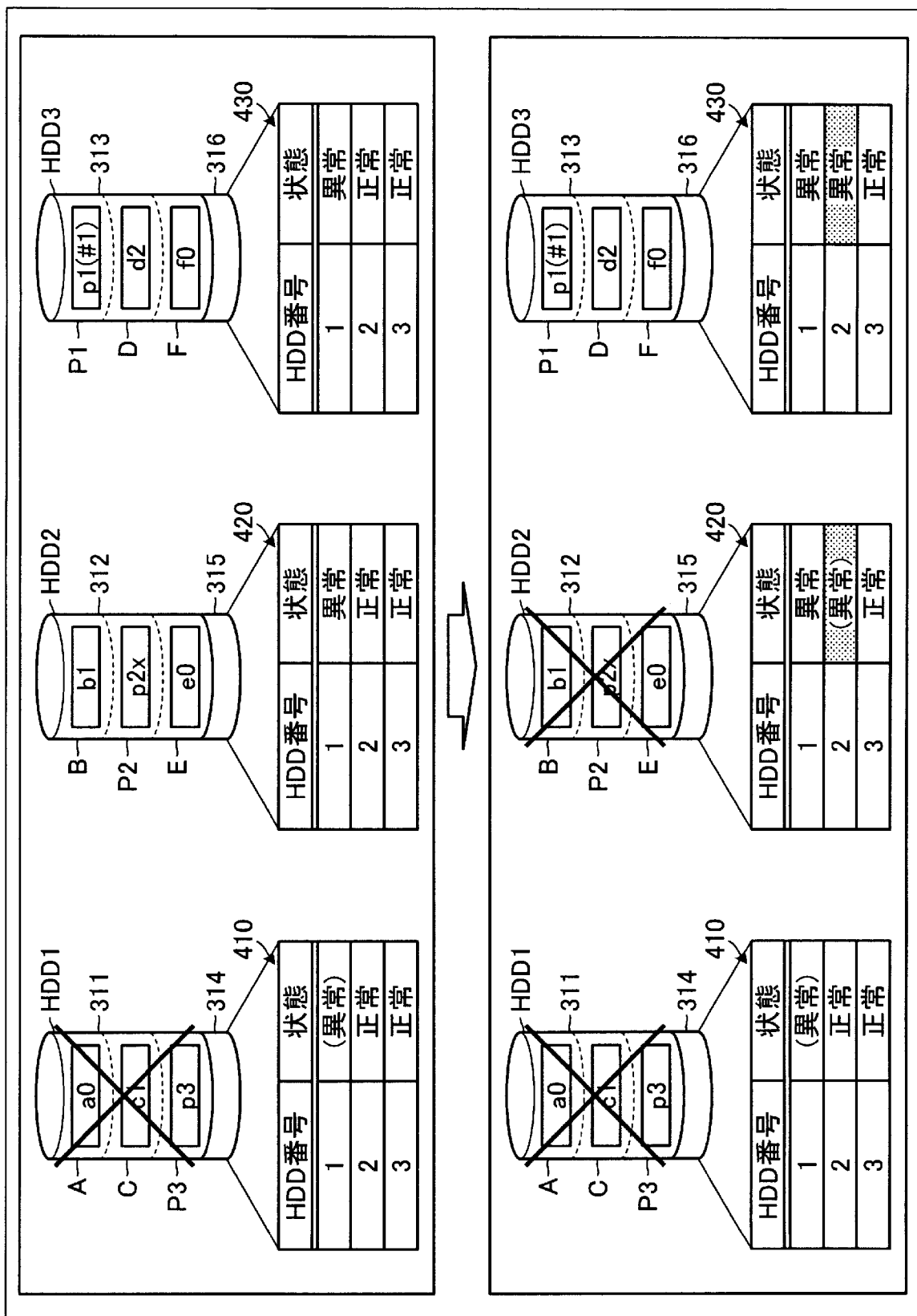
[図26]



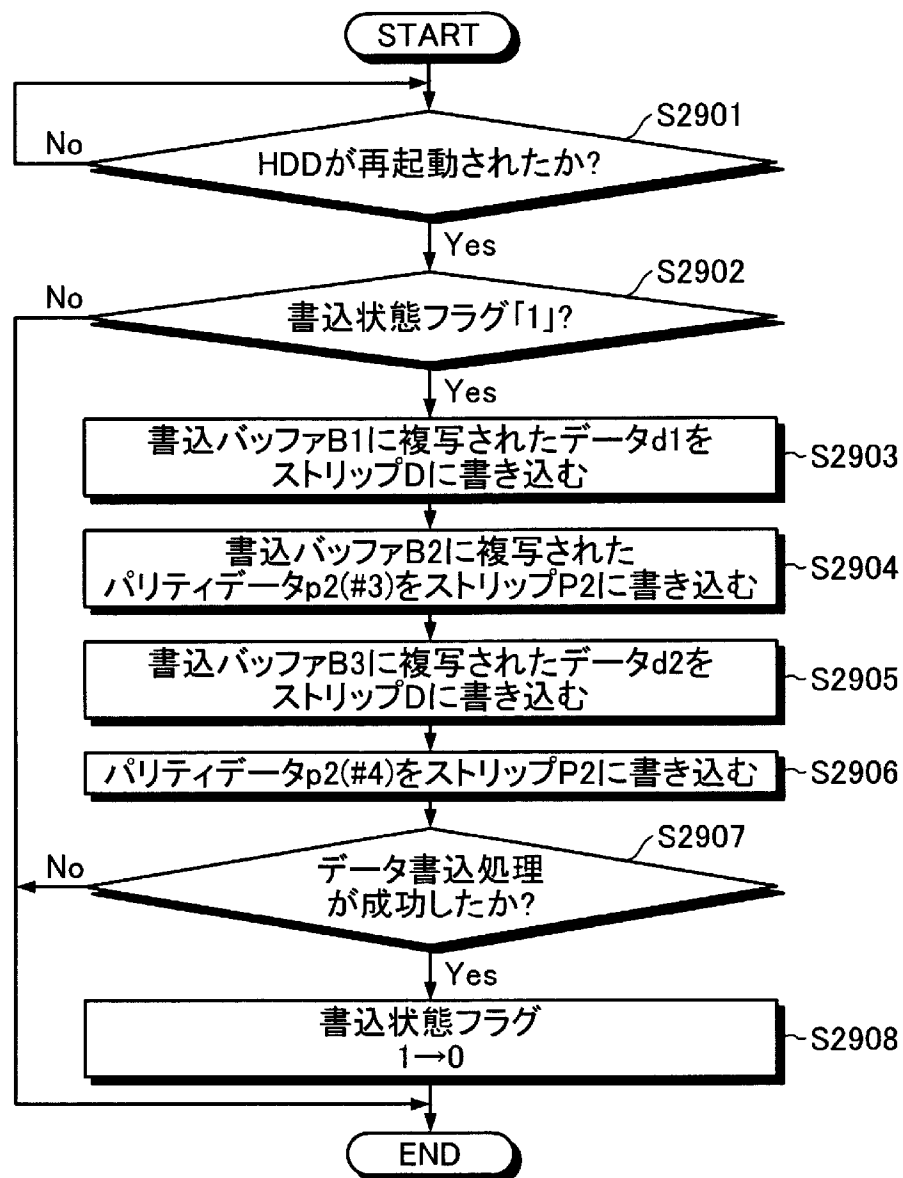
[図27]



[図28]



[図29]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2010/068561

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 1996-2011

Kokai Jitsuyo Shinan Koho 1971-2011 Toroku Jitsuyo Shinan Koho 1994-2011

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.
Y	JP 8-263227 A (Symbios Logic Inc.), 11 October 1996 (11.10.1996), paragraphs [0014], [0016], [0025] to [0037]; fig. 6 & US 5533190 A & EP 718766 A2 & DE 69528443 D & DE 69528443 T	1-11
Y	JP 9-190370 A (Oki Electric Industry Co., Ltd.), 22 July 1997 (22.07.1997), page 1 (Family: none)	1-11

☒ 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
12 January, 2011 (12.01.11)Date of mailing of the international search report
25 January, 2011 (25.01.11)Name and mailing address of the ISA/
Japanese Patent Office

Authorized officer

Facsimile No.

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2010/068561

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	JP 2004-504645 A (Wild File, Inc.), 12 February 2004 (12.02.2004), pages 1 to 2 & US 6016553 A & US 6199178 B1 & US 6240527 B1 & US 6363487 B1 & US 6732293 B1 & US 2002/0049883 A1 & WO 1999/012101 A2 & WO 2001/033357 A1 & DE 19882659 T & AU 9383298 A & AU 1237701 A	1-11
A	JP 6-119126 A (Fujitsu Ltd.), 28 April 1994 (28.04.1994), paragraphs [0039], [0055] to [0067] & US 5787460 A & US 5954822 A & WO 1993/023803 A1 & DE 4345320 A & DE 4392143 C	1-11

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	1 9 2 2 - 1 9 9 6 年
日本国公開実用新案公報	1 9 7 1 - 2 0 1 1 年
日本国実用新案登録公報	1 9 9 6 - 2 0 1 1 年
日本国登録実用新案公報	1 9 9 4 - 2 0 1 1 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 請求項の番号
Y	JP 8-263227 A (シンバイオス・ロジック・インコーポレイテッド) 1996.10.11, 段落 0014, 0016, 0025-0037, 第 6 図 & US 5533190 A & EP 718766 A2 & DE 69528443 D & DE 69528443 T	1-11
Y	JP 9-190370 A (沖電気工業株式会社) 1997.07.22, 第 1 頁 (ファミリーなし)	1-11
Y	JP 2004-504645 A (ワイルド ファイル, インコーポレイテッド) 2004.02.12, 第 1-2 頁 & US 6016553 A & US 6199178 B1	1-11

☒ 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

1 2 . 0 1 . 2 0 1 1

国際調査報告の発送日

2 5 . 0 1 . 2 0 1 1

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 1 0 0 - 8 9 1 5

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)

菅原 浩二

電話番号 0 3 - 3 5 8 1 - 1 1 0 1 内線 3 5 6 5

5 I

9 4 6 0

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 請求項の番号
A	& US 6240527 B1 & US 6363487 B1 & US 6732293 B1 & US 2002/0049883 A1 & WO 1999/012101 A2 & WO 2001/033357 A1 & DE 19882659 T & AU 9383298 A & AU 1237701 A JP 6-119126 A (富士通株式会社) 1994. 04. 28, 段落 0039, 0055-0067 & US 5787460 A & US 5954822 A & WO 1993/023803 A1 & DE 4345320 A & DE 4392143 C	1-11