(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2016 年 2 月 4 日(04.02.2016)

WIPO | PCT

(10) 国際公開番号
WO 2016/017002 A1

(51) 国際特許分類:
*G06F 3/06* (2006.01)

(21) 国際出願番号: PCT/JP2014/070224

(22) 国際出願日: 2014 年 7 月 31 日(31.07.2014)

(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: 株式会社日立製作所(HITACHI, LTD.) [JP/JP]; 〒1008280 東京都千代田区丸の内一丁目６番６号 Tokyo (JP).

(72) 発明者: 松井 章(MATSUI, Akira); 〒1008280 東京都千代田区丸の内一丁目６番６号 株式会社日立製作所内 Tokyo (JP).

(74) 代理人: 特許業務法人第一国際特許事務所(PATENT CORPORATE BODY DAI-ICHI KOKUSAI TOKKYO JIMUSHO); 〒1080014 東京都港区芝４丁目１０番５号 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, BN, 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, IR, IS, JP, KE, KG, 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, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, 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.

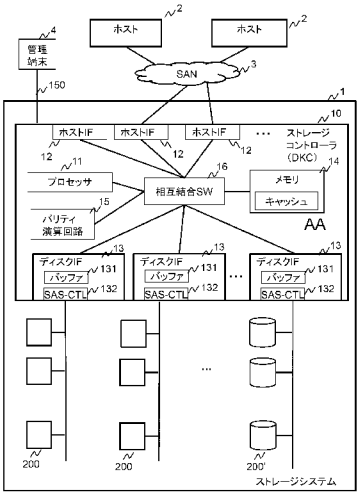
(84) 指定国 (表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッパ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM,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

[続葉有]

(54) **Title**: STORAGE SYSTEM

(54) 発明の名称 : ストレージシステム

図1




1 Storage system
2 Host
4 Management terminal
10 Storage controller (DKC)
11 Processor
12 Host IF
13 Disk IF
14 Memory
15 Parity calculation circuit
16 Mutual connection SW
131 Buffer
AA Cache

(57) **Abstract**: A storage system according to the present invention comprises a controller and a plurality of storage devices, wherein (n + m) storage devices constitute a RAID group, write data from a host computer is stored in the n storage devices, and redundant data generated from n data is stored in the m storage devices. When any fault has occurred in at least one of the storage devices, the controller reads compressed data and redundant data from each of the storage devices in which no fault has occurred among the storage devices constituting the RAID group, and transmits the read compressed data to the storage device for which data restoration is to be performed.

(57) 要約: 本発明のストレージシステムは、コントローラと、複数の記憶デバイスとを備え、（ｎ＋ｍ）台の記憶デバイスによりＲＡＩＤグループを構成し、ホスト計算機からのライトデータをｎ台の記憶デバイスに、ｎ個のデータから生成された冗長データをｍ個の記憶デバイスに格納している。少なくとも１台の記憶デバイスに障害が発生した場合、コントローラはＲＡＩＤグループを構成する記憶デバイスの中で障害の発生していない記憶デバイスの各々から、圧縮された状態のデータと冗長データを読み出して、読み出された圧縮状態のデータを、データ復旧先の記憶デバイスに送信する。

WO 2016/017002 A1

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 ： ストレージシステム

## 技術分野

[0001] 本発明は、ストレージシステムにおけるデータ復旧技術に関する。

## 背景技術

[0002] ストレージシステムの中には、いわゆるＲＡＩＤ（Ｒｅｄｕｎｄａｎｔ Ａｒｒａｙｓ ｏｆ Ｉｎｅｘｐｅｎｓｉｖｅ／Ｉｎｄｅｐｅｎｄｅｎｔ Ｄｉｓｋｓ）技術を用いて、システムを高可用化しているものが多い。ＲＡＩＤ技術とは、ホスト計算機などの上位装置から受け付けたライトデータからパリティなどの冗長データを算出し、ライトデータとパリティとをそれぞれ異なる記憶デバイスに分散格納する技術である。ＲＡＩＤ技術を採用することで、一部の記憶デバイスに障害が発生し、その記憶デバイスからデータを読み出せなくなった場合でも、その他の記憶デバイスに格納されている情報を用いて、データを再生成することができる。

[0003] ＲＡＩＤ技術ではパリティ算出のために、ストレージシステムに搭載されるコントローラ（ストレージコントローラ）の処理負荷、あるいはストレージシステム内の構成要素間（たとえばストレージコントローラと記憶デバイス間）のデータ転送量が増加する。処理負荷やデータ転送量の増加を抑止するために、従来から様々な技術が考えられてきた。たとえば特許文献１には、ストレージコントローラと記憶デバイスの間で発生するデータ転送量を抑制するために、記憶デバイス側にパリティ生成等の機能を有するストレージシステムが開示されている。

## 先行技術文献

## 特許文献

[0004] 特許文献1：米国特許出願公開第２０１３／０２９０６２９号明細書

## 発明の概要

## 発明が解決しようとする課題

[0005] ＲＡＩＤ技術の特徴として、先に述べたとおり、障害が発生した記憶デバイスに格納されていたデータを復旧（再生）できる点が挙げられる。ただしデータの復旧のためには、障害が発生していない記憶デバイスに格納されているデータをすべて読み出し、読み出されたデータに対して所定の演算を施すことでデータを再生成し、再生成されたデータを、新たな記憶デバイス（スペアドライブまたはスペアデバイス）に書き込む、という処理を行う必要があり、これらの処理の過程では、ストレージコントローラと記憶デバイスの間で大量のデータ転送が発生する。

[0006] 特に近年、記憶デバイスの記憶容量が増大しており、記憶デバイス内のデータをすべて読み出すだけでも、長時間を要する。そのため、ＲＡＩＤ技術を用いたデータ復旧処理も長時間化する傾向がある。データ復旧処理中は冗長性がない状態であるため、復旧処理中に別の記憶デバイスに障害が発生した場合、データ復旧が不可能になる。特許文献１には、記憶デバイス側に設けられたパリティ生成機能を用いてデータ復旧を行うことが開示されているが、記憶デバイス内のデータをすべて読み出す必要があることには変わりはなく、データ復旧処理の時間を大幅に削減することは困難である。

## 課題を解決するための手段

[0007] 本発明の一態様に係るストレージシステムは、コントローラと、複数の記憶デバイスとを備え、（ｎ＋ｍ）台の記憶デバイスによりＲＡＩＤグループを構成し、ホスト計算機からのライトデータをｎ台の記憶デバイスに、ｎ個のデータから生成された冗長データをｍ個の記憶デバイスに格納している。少なくとも１台の記憶デバイスに障害が発生した場合、コントローラはＲＡＩＤグループを構成する記憶デバイスの中で障害の発生していない記憶デバイスの各々から、圧縮された状態のデータと冗長データを読み出して、読み出された圧縮状態のデータを、データ復旧先の記憶デバイスに送信する。

## 発明の効果

[0008] 本発明により、データ復旧時の転送データ量を削減することができ、データ復旧時間を短縮することができる。

## 図面の簡単な説明

[0009] [図1]本発明の実施例に係るストレージシステムの構成図である。

[図2]ＦＭＰＫの構成図である。

[図3]ＲＡＩＤグループ内の記憶領域の内容の説明図である。

[図4]ＲＡＩＤグループの記憶空間と論理ユニット（ＬＵ）の関係を説明する図である。

[図5]デバイス管理テーブルの構成を示す図である。

[図6]ＲＧ管理テーブルの構成を示す図である。

[図7]ＬＵ管理テーブルの構成を示す図である。

[図8]ＦＭＰＫ内の記憶領域の管理方法の説明図である。

[図9]マッピングテーブルの構成を示す図である。

[図10]状態管理テーブルの構成を示す図である。

[図11]ホストからのライトデータにＤＩＦが付加される過程を表した概念図である。

[図12]圧縮Ｒｅａｄコマンドのフォーマットを表した図である。

[図13]圧縮Ｒｅａｄコマンド３０００に対する応答情報のフォーマットを表した図である。

[図14]圧縮コピーＷｒｉｔｅコマンドのフォーマットを表した図である。

[図15]圧縮パリティ演算Ｗｒｉｔｅコマンドのフォーマットを表した図である。

[図16]圧縮コピーＷｒｉｔｅコマンドに対する応答情報のフォーマットを表した図である。

[図17]データ復旧処理の全体の流れを表したフローチャートである。

[図18]コピー管理テーブルの構成を示す図である。

[図19]コピー復旧処理の詳細フローチャートである。

[図20]コレクションコピーの詳細フローチャートである。

[図21]圧縮Ｒｅａｄコマンドを受信した時に行われる処理の流れ図である。

[図22]圧縮コピーＷｒｉｔｅコマンドを受信した時に行われる処理の流れ図

である。

[図23]圧縮パリティ演算Ｗｒｉｔｅコマンドを受信した時に行われる処理の流れ図である。

[図24]ステージング情報管理テーブルの構成を示す図である。

[図25]圧縮中間パリティ演算コマンドのフォーマットを表した図である。

[図26]パリティコミットコマンドのフォーマットを表した図である。

[図27]圧縮中間パリティ演算コマンドを受信した時に行われる処理の流れ図である。

[図28]パリティコミットコマンドを受信した時に行われる処理の流れ図である。

[図29]実施例２に係るストレージシステムで行われるコレクションコピーの詳細フローチャートである。

[図30]実施例３に係るストレージシステムでサポートされるストライプラインの例を表した図である。

[図31]実施例３に係るストレージシステムで行われるコレクションコピーの詳細フローチャートである。

[図32]実施例３に係るストレージシステムでサポートされる圧縮パリティ演算Ｗｒｉｔｅコマンドのフォーマットを表した図である。

## 発明を実施するための形態

[0010] 以下、図面を参照して、本発明の一実施形態に係るストレージシステムについて説明する。なお、本発明は、以下に説明する実施形態に限定されるものではない。

## 実施例 1

[0011] 図１は、実施例１に係るストレージシステム１の構成を示す。ストレージシステム１は、ストレージコントローラ（ＤＫＣ）１０と、ストレージコントローラ１０に接続された複数の記憶デバイス（２００、２００’）を有する。

[0012] 記憶デバイス２００、２００’は、ホスト２などの上位装置からのライト

データを格納するための最終記憶媒体である。本実施例のストレージシステム１は、最終記憶装置として、磁気ディスクを記録媒体とするＨＤＤ（Ｈａｒｄ　Ｄｉｓｋ　Ｄｒｉｖｅ）２００’の他、フラッシュメモリなどの不揮発性半導体メモリを記憶媒体として用いるＦＭＰＫ（Ｆｌａｓｈ　Ｍｅｍｏｒｙ　ＰａｃＫａｇｅ）２００を用いることができる。ただしその他の記憶デバイスを用いることも可能である。また、記憶デバイス２００、２００’は一例として、ＳＡＳ（Ｓｅｒｉａｌ　Ａｔｔａｃｈｅｄ　ＳＣＳＩ）規格に従って、ストレージコントローラ１０（以下、「ＤＫＣ１０」と略記する）との通信を行う。

[0013] ＤＫＣ１０は複数のＦＭＰＫを、１又は複数のＲＡＩＤ（Ｒｅｄｕｎｄａｎｔ　Ａｒｒａｙｓ　ｏｆ　Ｉｎｅｘｐｅｎｓｉｖｅ／Ｉｎｄｅｐｅｎｄｅｎｔ　Ｄｉｓｋｓ）グループ１４５として管理する。

[0014] ＤＫＣ１０には、１以上のホスト２と、管理端末４が接続される。ＤＫＣ１００とホスト２とは、一例としてファイバチャネルを用いて形成されるＳＡＮ（Ｓｔｏｒａｇｅ　Ａｒｅａ　Ｎｅｔｗｏｒｋ）１を介して接続される。ＤＫＣ１０と管理端末４とは、一例としてイーサネットを用いて形成されるネットワーク１５０を介して接続される。

[0015] ＤＫＣ１０は少なくとも、プロセッサ１１、ホストインタフェース（図中では「ホストＩＦ」と表記）１２、ディスクインタフェース（図中では「ディスクＩＦ」と表記）１３、メモリ１４、パリティ演算回路１５を有する。そしてプロセッサ１１、ホストＩＦ１２、ディスクＩＦ１３、メモリ１４、パリティ演算回路１５は、相互結合スイッチ（相互結合ＳＷ）１６を介して相互接続されている。これらの構成要素は、高性能化及び高可用性の確保のため、ＤＫＣ１０内に複数搭載されている。ただしこれらの構成要素が１つだけＤＫＣ１０内に設けられている構成でもよい。

[0016] ディスクＩＦ１３は少なくとも、バッファ１３１、記憶デバイス２００、２００’と通信するためのインタフェースコントローラ１３２（図中では「ＳＡＳ－ＣＴＬ」と表記されている）、及び転送回路（非図示）を有する。

インタフェースコントローラ１３２は、記憶デバイス２００、２００’の用いているプロトコル（一例ではＳＡＳ）を、ＤＫＣ１０内部で用いられている通信プロトコル（一例としてＰＣＩ－Ｅｘｐｒｅｓｓ）に変換するためのものである。本実施例では、記憶デバイス２００、２００’がＳＡＳ規格に従った通信を行うため、インタフェースコントローラ１３２にはＳＡＳコントローラ（以下、「ＳＡＳ－ＣＴＬ」と略記する）が用いられる。図１では、１つのディスクＩＦ１３にＳＡＳ－ＣＴＬ１３２が１つのみ記載されているが、１つのディスクＩＦ１３に複数のＳＡＳ－ＣＴＬ１３２が存在する構成を採用してもよい。

[0017] ディスクＩＦ１３は、たとえばＦＭＰＫ２００からデータを読み出し、読み出したデータをバッファ１３１に一時的に格納する。バッファ１４３に格納されたデータは、転送回路によって、メモリ１４あるいは他のディスクＩＦ１３のバッファへと送信される。バッファ１３１はたとえば揮発性の半導体メモリから構成されるが、不揮発メモリを用いて構成されていても良い。

[0018] ホストＩＦ１２は、ディスクＩＦ１３と同様に、インタフェースコントローラとバッファ、転送回路（非図示）を少なくとも有する。インタフェースコントローラは、ホスト２とＤＫＣ１０間で用いられている通信プロトコル（たとえばファイバチャネル）と、ＤＫＣ１０内部で用いられている通信プロトコルを変換するためのものである。バッファはホスト２からのデータを一時的に格納するためのものである（逆にメモリ１４からホスト２に転送するべきデータを一時的に格納するために用いられることもある）。

[0019] パリティ演算回路１５は、ＲＡＩＤ技術で必要とされる冗長データの生成を行うハードウェアである。パリティ演算回路１５により生成される冗長データの例としては、排他的論理和（ＸＯＲ）、リードソロモン符号等がある。

[0020] プロセッサ１１は、ホストＩＦ１２から到来するＩ／Ｏ要求の処理を行う。メモリ１４は、プロセッサ１１が実行するプログラムや、プロセッサが使用するストレージシステム１の各種管理情報を記憶するために用いられる。

またメモリ１４は、記憶デバイス２００、２００’に対するＩ／Ｏ対象データを一時的に記憶するためにも用いられる。以下、記憶デバイス２００、２００’に対するＩ／Ｏ対象データを一時的に記憶するために用いられる、メモリ１４中の記憶領域を、「キャッシュ」と呼ぶ。メモリ１４はＤＲＡＭ、ＳＲＡＭ等の揮発性記憶媒体で構成されるが、別の実施形態として、不揮発性メモリを用いてメモリ１４を構成してもよい。

[0021] ストレージシステム１は、先にも述べたとおり、ＦＭＰＫ２００、ＨＤＤ２００’等の、複数種類の記憶デバイスを搭載できる。ただし以下では特に断りのない限り、ストレージシステム１にＦＭＰＫ２００のみが搭載されている構成を前提として説明する。

[0022] 図２を用いて、ＦＭＰＫ２００の構成について説明する。ＦＭＰＫ２００は、デバイスコントローラ（ＦＭコントローラ）２０１と複数のＦＭチップ２１０から構成される。ＦＭコントローラ２０１は、メモリ２０２、プロセッサ２０３、データの圧縮伸長を行うための圧縮伸長回路２０４、パリティを計算するためのパリティ演算回路２０５、ＳＡＳ－ＣＴＬ２０６、ＦＭ－ＩＦ２０７を備える。メモリ２０２、プロセッサ２０３、圧縮伸長回路２０４、パリティ演算回路２０５、ＳＡＳ－ＣＴＬ２０６、ＦＭ－ＩＦ２０７は、内部接続スイッチ（内部接続ＳＷ）２０８を介して相互接続されている。

[0023] ＳＡＳ－ＣＴＬ２０６は、ＦＭＰＫ２００とＤＫＣ１０間の通信を行うためのインタフェースコントローラである。ＳＡＳ－ＣＴＬ２０６は、伝送線（ＳＡＳリンク）を介して、ＤＫＣ１０のＳＡＳ－ＣＴＬ１３２に接続される。またＦＭ－ＩＦ２０７は、ＦＭコントローラ２０１とＦＭチップ２１０間の通信を行うためのインタフェースコントローラである。

[0024] プロセッサ２０３は、ＤＫＣ１０から到来する各種コマンドに係る処理を行う。メモリ２０２には、プロセッサ２０３が実行するプログラムや、各種管理情報が記憶される。メモリ２０２には、ＤＲＡＭ等の揮発性メモリが用いられる。ただしメモリ２０２に不揮発性メモリを使用しても良い。

[0025] 圧縮伸長回路２０４は、データの圧縮、または圧縮されたデータの伸長を

行う機能を備えたハードウェアである。またパリティ演算回路２０５は、ＤＫＣ１０が備えるパリティ演算回路１５と同様の機能を持つハードウェア、つまりＲＡＩＤ技術で必要とされる冗長データの生成機能を有する。

[0026] ＦＭチップ２１０は、不揮発性半導体メモリチップで、一例としてＮＡＮＤ型フラッシュメモリである。フラッシュメモリは周知のとおり、ページ単位でデータの読み出し・書き込みが行われ、またデータ消去は、複数ページの集合であるブロック単位で行われる。そして一度書き込みが行われたページは上書きが出来ず、一度書き込みが行われたページに対して再度書き込みを行うためには、当該ページを含むブロック全体を消去する必要がある。そのため、ＦＭＰＫ２００は、ＦＭＰＫ２００が接続されるＤＫＣ１０に対しては、ＦＭチップ２１０の有する記憶領域をそのまま提供することはせず、論理的な記憶空間を提供する。

[0027] なお、ＦＭＰＫ２００は、圧縮伸長回路２０４により、ＤＫＣ１０からのライトデータを圧縮してＦＭチップ２１０に格納することができる。ただし、ＤＫＣ１０に対しては原則として、透過的にデータ圧縮を行う。ＦＭＰＫ２００はＤＫＣ１０に対し、所定のサイズの記憶空間（論理アドレス空間）を提供する。ＤＫＣ１０がＦＭＰＫ２００にデータを書き込む際には、この論理アドレス空間上のアドレスとライト対象データのサイズを指定したライトコマンドを発行する。一例として、ＤＫＣ１０がＦＭＰＫ２００に対し、論理アドレス空間の先頭（アドレス０）に４８ＫＢのデータを書き込むライトコマンド（及び４８ＫＢのライトデータ）を送信したとする。またＦＭＰＫ２００がこの４８ＫＢのデータを圧縮した結果、８ＫＢの圧縮データになり、この８ＫＢの圧縮データがＦＭチップ２１０に格納されたとする。この状態において、ＤＫＣ１０がこのデータを読み出す場合には、論理アドレス空間の先頭（アドレス０）とリードデータサイズ（たとえば４８ＫＢ）を指定したリードコマンドを発行することで、先ほど格納した４８ＫＢのデータを読み出すことができる。読み出しの過程で、ＦＭＰＫ２００は圧縮伸長回路２０４により、８ＫＢの圧縮データを４８ＫＢに伸長（復元）し、伸長さ

れたデータをＤＫＣ１０に返送するよう動作するからである。そのためＤＫＣ１０は、（実際にはデータが圧縮されて格納されている場合でも）論理アドレス空間上にデータが非圧縮状態で格納されているかのように認識する。

[0028] 上では、データの圧縮、伸長が、圧縮伸長回路２０４というハードウェアで行われる例について説明したが、必ずしもデータの圧縮、伸長を、ハードウェアを用いて行わなければならないわけではない。プロセッサ２０３が圧縮、伸長を行うプログラムを実行することによって、データの圧縮、伸長を行うようにしてもよい。またパリティ演算回路２０５についても同様で、パリティ演算回路２０５というハードウェアでパリティの演算を行わなければならないわけではない。プロセッサ２０３がパリティ演算を行うプログラムを実行することによって、パリティ演算を行うようにしてもよい。

[0029] さらに、上ではＦＭＰＫ２００が圧縮／伸長機能、パリティ演算機能を備えている例を説明したが、圧縮／伸長機能、またはパリティ演算機能を有していないＦＭＰＫ２００がストレージシステム１内に存在してもよい。ＦＭＰＫ２００が圧縮／伸長機能を有していない場合には、データは圧縮して格納されない。またＦＭＰＫ２００がパリティ演算機能を有していない場合には、ＤＫＣ１０がパリティの生成を行う。

[0030] 続いて、ストレージシステム１で用いられる記憶領域の概念について説明する。ストレージシステム１は、複数のＦＭＰＫ２００を１つのＲＡＩＤ（Ｒｅｄｕｎｄａｎｔ Ａｒｒａｙｓ ｏｆ Ｉｎｅｘｐｅｎｓｉｖｅ／Ｉｎｄｅｐｅｎｄｅｎｔ Ｄｉｓｋｓ）グループとして管理する。そしてＲＡＩＤグループ内で１つ（あるいは２つ）のＦＭＰＫ２００に障害が発生してデータアクセスできなくなった場合に、残りのＦＭＰＫ２００内のデータを用いて、障害が発生したＦＭＰＫ２００に格納されていたデータを復旧できるようにしている。また、ＲＡＩＤグループ内の一部の記憶領域（あるいは全記憶領域）を、ホスト２などの上位装置に提供する。

[0031] ＲＡＩＤグループ内の記憶領域について、図３を用いて説明する。図３において、ＦＭＰＫ＃ｘ（ｘは０～３の数値）は、ＦＭＰＫ２００がＤＫＣ１

０に提供している記憶空間を表している。ＤＫＣ１０は、複数（図３の例では４つ）のＦＭＰＫ２００から１つのＲＡＩＤグループ２０を構成し、ＲＡＩＤグループ２０に所属する各ＦＭＰＫ（ＦＭＰＫ＃０（２００－０）～ＦＭＰＫ＃３（２００－３））上の記憶空間を、ストライプブロックと呼ぶ複数の固定サイズの記憶領域に分割して管理している。

[0032] また図３では、ＲＡＩＤグループ２０のＲＡＩＤレベル（ＲＡＩＤ技術におけるデータ冗長化方式を表すもので、一般的にはＲＡＩＤ１～ＲＡＩＤ６の種類がある）がＲＡＩＤ５である場合の例を表している。ＤＫＣ１０は、図３においてＲＡＩＤグループ２０内の、「０」、「１」、「Ｐ」などのボックスがストライプブロックを表しており、１つのＲＡＩＤグループ２０内の各ストライプブロックのサイズ（以下では「ストライプサイズ」と呼ばれる）は同じである。ストライプサイズとしては、たとえば６４ＫＢ、２５６ＫＢ、５１２ＫＢなどが用いられる。また、各ストライプブロックに付されている、「１」等の番号のことを、「ストライプブロック番号」と呼ぶ。

[0033] 図３で、ストライプブロックのうち、「Ｐ」と記載されているストライプブロックは、冗長データの格納されるストライプブロックであり、これを「パリティストライプ」と呼ぶ。一方、数字（０、１等）が記載されているストライプブロックは、ホスト２等の上位装置から書き込まれるデータ（冗長データではないデータ）が格納されるストライプブロックである。このストライプブロックのことは、「データストライプ」と呼ばれる。

[0034] 図３に示されたＲＡＩＤグループ２０では、たとえばＦＭＰＫ＃３（２００－３）の先頭に位置するストライプブロックはパリティストライプ３０１－３である。そしてＤＫＣ１０がこのパリティストライプ３０１－３に格納される冗長データを作成する際、各ＦＭＰＫ２００（ＦＭＰＫ＃０（２００－０）～ＦＭＰＫ＃２（２００－２））の先頭に位置するデータストライプ（ストライプブロック３０１－０、３０１－１、３０１－２）に格納されるデータに対して所定の演算（たとえば排他的論理和（ＸＯＲ）等）を施すことによって、冗長データを生成する。

[0035] 以下、パリティストライプと、当該パリティストライプに格納される冗長データを生成するために用いられるデータストライプのセット（たとえば図３中の要素３００）のことを、「ストライプライン」と呼ぶ。実施例１に係るストレージシステム１の場合、図３に示されているストライプライン３００のように、１つのストライプラインに属する各ストライプブロックは、ＦＭＰＫ２００－０～２００－３の記憶空間の同じ位置（アドレス）に存在するという規則に基づいて、ストライプラインが構成される。

[0036] なお、先に説明したストライプブロック番号は、データストライプに付される番号で、ＲＡＩＤグループ内で一意な番号である。図３に示されているように、ＤＫＣ１０は、ＲＡＩＤグループを構成する各ＦＭＰＫ２００の先頭に位置するデータストライプに０、１、２の番号を付し、以下順に、３、４、５．．．の連続番号を付して、データストライプを管理している。以下、ストライプブロック番号がｎ番（ｎは０以上の整数値）であるデータストライプのことを、「データストライプｎ」と表記する。

[0037] またストレージシステム１では、ＲＡＩＤグループに属する各記憶デバイス２００（２００’）に対し、ＲＡＩＤグループ内で一意な番号を付して管理する。この一意な番号は「ＲＡＩＤグループ内位置番号」または「位置番号」と呼ばれる。具体的には、ＲＡＩＤグループ内の先頭ストライプラインに、データストライプ０～データストライプｋ（ｋ＞０）が含まれている場合、データストライプｍ（０≦ｍ≦ｋ）が格納される記憶デバイス２００（２００’）の位置番号が「ｍ」と定められる。

[0038] そしてＲＡＩＤ５のように１ストライプライン内に１つのパリティストライプが存在するＲＡＩＤ構成の場合、パリティストライプが格納される記憶デバイス２００（２００’）の位置番号が「ｋ＋１」と定められる。さらにＲＡＩＤ６のように、１ストライプライン内に２つのパリティストライプが存在するＲＡＩＤ構成の場合、パリティストライプが格納される２つの記憶デバイス２００（２００’）の位置番号はそれぞれ、「ｋ＋１」及び「ｋ＋２」と定められる。

[0039] 図３に示されているＲＡＩＤグループの場合、先頭ストライプライン３００内のデータストライプ０～２が格納される３つの記憶デバイス２００（２００’）の位置番号は０、１、２と定められる。そしてデータストライプ０～２と同一ストライプライン内の冗長データ（パリティ）の格納される記憶デバイス２００（２００’）の位置番号は３と定められる。

[0040] また、ＤＫＣ１０は、ホスト２等の上位装置に対し、論理ユニット（ＬＵ）と呼ばれる記憶空間を、１以上提供する。ＲＡＩＤグループにより形成される記憶空間（以下、「ＲＡＩＤグループの記憶空間」と呼ぶ）と、論理ユニットの関係について、図４を用いて説明する。ＲＡＩＤグループの記憶空間とは、図３に示されているＲＡＩＤグループ内の領域のうち、データストライプ０から順に、データストライプのみを並べることによって形成される記憶空間である。ＤＫＣ１０は、１つのＲＡＩＤグループの記憶空間上の連続領域を論理ユニットと定義することができる。１つのＲＡＩＤグループの記憶空間上に複数の論理ユニットを定義してもよいし、１つのＲＡＩＤグループの記憶空間全体を１つの論理ユニットとして定義することもできる。図４では、ＲＡＩＤグループの記憶空間上に２つの論理ユニット（ＬＵ＃０、ＬＵ＃１）が定義されている例を表している。

[0041] 一例として、ストレージシステム１が、ホスト２からＬＵ＃０の先頭から３ストライプブロック分の領域に対するデータ書き込み要求（ライトコマンド）を受け付けた場合に行われる処理の概要を説明する。なお、ＬＵ＃０は図３に示されているＲＡＩＤグループに定義されているものとする。この場合、ＤＫＣ１０はライトコマンドとともに受信した３ストライプブロック分のデータを、１ストライプブロックごとに分割する。以下、分割されて生成された３つのデータをそれぞれ、「データ０」、「データ１」、「データ２」と呼ぶ。ＤＫＣ１０はデータ０～データ２を用いて冗長データ（パリティ）を生成し、データ０～データ２及び生成されたパリティをそれぞれ、異なる記憶デバイス２００に格納する。ＬＵ＃が図３に示されているＲＡＩＤグループ内に定義されている場合、データ０～データ２及びパリティはそれぞ

れ、ＦＭＰＫ（２００－０）～ＦＭＰＫ（２００－３）に格納される。

[0042] ＬＵ＃０が定義されているＲＡＩＤグループのＲＡＩＤレベルが、図３のようにＲＡＩＤ５である場合、パリティはデータ０～データ２の排他的論理和（ＸＯＲ）を計算することで生成できる。パリティの計算は、ＤＫＣ１０のパリティ演算回路１５を用いて行われる。ただしＲＡＩＤグループを構成する記憶デバイス２００が図２のように、パリティ演算回路２０５等のパリティ生成機能を備えている場合には、記憶デバイス２００の有するパリティ生成機能を用いてパリティを生成することもできる。

[0043] ＲＡＩＤグループ、論理ユニット（ＬＵ）を管理するために、ＤＫＣ１０はメモリ１４に、デバイス管理テーブル、ＲＡＩＤグループ管理テーブル（ＲＧ管理テーブル）、ＬＵ管理テーブルという３種類の管理情報を有する。図５は、デバイス管理テーブルＴ１０００の例を示している。

[0044] デバイス管理テーブルＴ１０００は、ストレージシステム１に搭載された各記憶デバイス２００（または２００’）についての情報を管理するためのテーブルである。デバイス管理テーブルＴ１０００内の各行（以下では、テーブル内の行のことを「レコード」と呼ぶ）に、ストレージシステム１に搭載された各記憶デバイス２００（２００’）の情報が格納される。デバイス管理テーブルＴ１０００のレコードは、デバイス＃（Ｔ１００１）、デバイス種別（Ｔ１００２）、所属ＲＧ＃（Ｔ１００３）、デバイスステータス（Ｔ１００４）、圧縮機能サポート（Ｔ１００５）、パリティ演算機能サポート（Ｔ１００６）、サイズ（Ｔ１００７）の項目を有する。

[0045] ＤＫＣ１０は、ストレージシステム１に搭載された各記憶デバイス２００（または２００’）に、一意な識別番号を付して管理しており、この識別番号は「デバイス番号」（または「デバイス＃」）と呼ばれる。デバイス＃（Ｔ１００１）には、記憶デバイス２００（２００’）のデバイス＃が格納される。デバイス種別（Ｔ１００２）は、記憶デバイス２００（２００’）の種類についての情報を格納するための項目である。本実施例の場合、デバイス種別（Ｔ１００２）には、「ＦＭＰＫ」または「ＨＤＤ」のいずれかの情

報が格納される。あるレコードのデバイス種別（Ｔ１００２）に「ＦＭＰＫ」が格納されている場合、当該レコードで管理される記憶デバイスがＦＭＰＫ２００であることを表しており、また「ＨＤＤ」が格納されている場合、当該レコードで管理される記憶デバイスがＨＤＤ２００’であることを表している。所属ＲＧ＃（Ｔ１００３）については後述する。

[0046] デバイスステータス（Ｔ１００４）には、記憶デバイスの状態が格納される。デバイスステータス（Ｔ１００４）に「正常」が格納されている場合、当該レコードで管理される記憶デバイスは正常に稼働していることを表している。デバイスステータス（Ｔ１００４）に「閉塞」が格納されている場合、当該レコードで管理される記憶デバイスは、障害が発生した等の理由により、稼働していない（閉塞状態にある）ことを表している。

[0047] デバイスステータス（Ｔ１００４）に「障害復旧中（復旧元）」または「障害復旧中（復旧先）」が格納されている場合、当該レコードで管理される記憶デバイスの所属するＲＡＩＤグループについてデータ復旧処理が行われていることを表している。詳細は後述するが、たとえば１台の記憶デバイスに障害が発生した場合、データ復旧処理では障害が発生した記憶デバイスの代わりとなるデバイス（以下、これを「スペアデバイス」と呼ぶ）を１台用意する。そしてＤＫＣ１０は、障害が発生した記憶デバイスのデバイスステータス（Ｔ１００４）には「障害復旧中（復旧元）」を格納する。またＤＫＣ１０は、スペアデバイスとされる記憶デバイスのデバイスステータス（Ｔ１００４）に「障害復旧中（復旧先）」を格納する。そして障害が発生した記憶デバイスに格納されていたデータを復旧し、スペアデバイスに書き込むことで、データ復旧を行う。

[0048] 圧縮機能サポート（Ｔ１００５）、パリティ演算機能サポート（Ｔ１００６）にはそれぞれ、当該レコードで管理される記憶デバイスの、圧縮機能のサポート有無、パリティ演算機能のサポート有無についての情報が格納される。圧縮機能サポート（Ｔ１００５）に「サポート」が格納されている場合には、当該レコードで管理される記憶デバイスが圧縮機能を有していること

を意味し、「未サポート」が格納されている場合には、当該レコードで管理される記憶デバイスが圧縮機能を有していないことを意味する。パリティ演算機能サポート（Ｔ１００６）にも同様に、「サポート」または「未サポート」のいずれかの情報が格納され、「サポート」が格納されている場合には、当該レコードで管理される記憶デバイスがパリティ演算機能を有していることを意味する。

[0049] 圧縮機能サポート（Ｔ１００５）、パリティ演算機能サポート（Ｔ１００６）に格納される情報は、ストレージシステム１の管理者が、管理端末を用いて設定してもよい。あるいは別の実施形態として、ＤＫＣ１０が各記憶デバイスに対して、各記憶デバイスが有する機能を問い合わせるコマンドを発行することによって、各記憶デバイスが圧縮機能及び／またはパリティ演算機能を有しているか問い合わせ、ＤＫＣ１０がこの問い合わせ結果を圧縮機能サポート（Ｔ１００５）、パリティ演算機能サポート（Ｔ１００６）に反映するようにしてもよい。

[0050] サイズ（Ｔ１００７）には、記憶デバイスの容量、具体的には記憶デバイス２００（２００’）がＤＫＣ１０に対して提供している記憶空間のサイズが格納される。この記憶空間のサイズについての情報は、ＤＫＣ１０が記憶デバイス２００（２００’）に対して、サイズの問い合わせを行うコマンドを発行することで、記憶デバイス２００（２００’）から取得可能な情報である。なお、記憶デバイス２００がＦＭＰＫ２００のように圧縮機能を有する場合、記憶空間のサイズは、記憶デバイス２００自身が有する記憶媒体（たとえばＦＭチップ２１０）の合計サイズよりも大きいこともある。

[0051] 図６は、ＲＧ管理テーブルＴ１１００の例を示している。デバイス管理テーブルＴ１０００と同様、ＲＧ管理テーブルＴ１１００内の複数のレコードの各々に、ＲＡＩＤグループについての情報が格納されている。

[0052] ＤＫＣ１０は、ストレージシステム１内に定義された各ＲＡＩＤグループに一意な識別番号を付して管理しており、この識別番号は「ＲＡＩＤグループ番号」または「ＲＧ＃」と呼ばれる。ＲＧ＃（Ｔ１１０１）には、ＲＡＩ

Ｄグループ番号（ＲＧ＃）が格納される。

所属デバイス＃（Ｔ１１０２）は、ＲＡＩＤグループに含まれている記憶デバイス２００（２００’）のデバイス＃が記憶される。たとえばＤＫＣ１０が、デバイス＃が０、１、２、３、４番の記憶デバイスを用いてＲＡＩＤグループを形成した時（仮にそのＲＧ＃が０番であったとする）、ＤＫＣ１０は、ＲＧ＃（Ｔ１１０１）が０のレコードの、所属デバイス＃（Ｔ１１０２）の欄に、「０、１、２、３、４」を格納する。また、図５のデバイス管理テーブルＴ１０００内の各レコードの所属ＲＧ＃（Ｔ１００３）には、各記憶デバイスの所属しているＲＡＩＤグループ番号が格納される。

[0053] 先にも述べたが、ストレージシステム１は、ＲＡＩＤグループに属する各記憶デバイス２００（２００’）に、位置番号を対応付けて管理する。そのため、所属デバイス＃（Ｔ１１０２）に登録された各ＦＭＰＫ２００には、位置番号（Ｔ１１０２’）が対応付けられる。図６において、レコードＴ１１００－１で管理されているＲＡＩＤグループ（ＲＧ＃（Ｔ１１０１）が１）にはデバイス＃が８～１５番のＦＭＰＫ２００が所属しているが、各ＦＭＰＫ２００には、位置番号０～７の番号が対応付けられている。

[0054] ＲＡＩＤ構成（Ｔ１１０３）は、ＲＡＩＤグループの構成についての情報が格納される項目であり、少なくとも、ＲＡＩＤ技術によるデータ保護方式を表すＲＡＩＤレベル、冗長データを生成する時に用いられるデータストライプの数、生成されるパリティストライプの数の情報が格納される。図６の例では、レコードＴ１１００－１のＲＡＩＤ構成（Ｔ１１０３）は「ＲＡＩＤ５（３Ｄ＋１Ｐ）」だが、これはＲＡＩＤレベルが５であること、そして３つのデータストライプから１つのパリティを生成するＲＡＩＤ構成であることを表している。

[0055] ＲＧ容量（Ｔ１１０４）、ストライプサイズ（Ｔ１１０５）にはそれぞれ、ＲＡＩＤグループに格納できるデータの量（サイズ）、ストライプサイズが格納される。本実施例では、ＲＧ容量（Ｔ１１０４）に格納される容量は、ＲＡＩＤグループ内全データストライプの合計サイズであり、パリティの

サイズは含まれない。ただし別の実施形態として、パリティのサイズを含んだ容量が格納されるようにしてもよい。

[0056] ＲＧステータス（Ｔ１１０６）には、ＲＡＩＤグループの状態（「正常」、「障害復旧中」、「障害復旧失敗」のいずれかの状態）が格納される。ＲＡＩＤグループの状態の意味は、デバイスステータス（Ｔ１００４）と同様であり、ＲＧステータス（Ｔ１１０６）に「正常」が格納されている場合には、ＲＡＩＤグループが正常に稼働していることを表している。また「障害復旧中」が格納されている場合には、ＲＡＩＤグループの復旧処理が行われていることを表しており、「障害復旧失敗」が格納されている場合にはＲＡＩＤグループが閉塞状態にあることを表している。

[0057] 圧縮（Ｔ１１０７）、パリティ演算（Ｔ１１０８）にはそれぞれ、「実施」または「未実施」の情報が格納される。圧縮（Ｔ１１０７）に「実施」が格納されている場合、当該レコードで管理されるＲＡＩＤグループでは、記憶デバイス（ＦＭＰＫ２００）が備える圧縮機能を用いたデータ圧縮が行われ、圧縮データがＦＭＰＫ２００に格納されていることを表している（逆に「未実施」が格納されている場合には、ＦＭＰＫ２００がデータ圧縮を行っていないことを意味する）。パリティ演算（Ｔ１１０８）に「実施」が格納されている場合、当該レコードで管理されるＲＡＩＤグループに格納されるパリティは、ＦＭＰＫ２００が備えるパリティ演算機能を用いて算出されることを表している。

[0058] ＤＫＣ１０は各ＲＡＩＤグループについて、ＲＡＩＤグループに所属する全記憶デバイスが圧縮機能をサポートしている場合（所属デバイス＃（Ｔ１１０２）で特定される全記憶デバイスの、圧縮機能サポート（Ｔ１００５）が「サポート」である場合）、圧縮（Ｔ１１０７）に「実施」を設定する。同様に、ＲＡＩＤグループに所属する全記憶デバイスがパリティ演算機能をサポートしている場合（所属デバイス＃（Ｔ１１０２）で特定される全記憶デバイスの、パリティ演算機能サポート（Ｔ１００６）が「サポート」である場合）、パリティ演算（Ｔ１１０８）に「実施」を設定する。また別の実

施形態として、ストレージシステム１の管理者が、管理端末を用いて圧縮（Ｔ１１０７）、パリティ演算（Ｔ１１０８）に「実施」または「未実施」を設定するようにしてもよい。

[0059] 図７は、ＬＵ管理テーブルＴ１２００の例を示している。デバイス管理テーブルＴ１０００と同様、ＬＵ管理テーブルＴ１２００内の各レコードそれぞれに、１つのＬＵについての情報が格納されている。各ＬＵは論理ユニット番号（ＬＵＮ）という一意な識別番号を有しており、ＬＵ＃（Ｔ１２０１）にはＬＵの論理ユニット番号が格納される。

[0060] 先に述べたとおり、ＤＫＣ１０は、ＲＡＩＤグループ内の連続領域をＬＵとして定義する。ＬＵ管理テーブルＴ１２００には、ＬＵが定義されているＲＡＩＤグループのＲＧ＃（Ｔ１２０２）、ＬＵが定義されているＲＡＩＤグループ内のオフセットアドレス（Ｔ１２０３）、ＬＵのサイズ（Ｔ１２０４）が格納される。

[0061] 続いて、ＦＭＰＫ２００内の記憶領域の管理方法について、図８を用いて説明する。ＦＭＰＫ２００は、ＤＫＣ１０に提供する記憶空間（論理アドレス空間）を、論理ページと呼ばれる所定サイズの領域単位で管理する。各論理ページを特定するために、各ページには一意な番号が付されている。この番号のことを「論理ページ番号」と呼ぶ。論理アドレス空間上の先頭に位置する論理ページの論理ページ番号が０番で、それ以降に続く論理ページには、連続した番号が付されている。論理ページのサイズは一例として、１６セクタ（８ＫＢ）である。

[0062] また、本実施例では、ＦＭＰＫ２００内のＦＭチップ２１０に存在する記憶領域は「物理ページ」と呼ばれる。物理ページは、フラッシュメモリにおける、アクセス（リード、ライト）の最小単位である。そのためＦＭコントローラ２０１がＦＭチップ２１０に対してデータの読み書きを行う場合には、物理ページ単位での読み書きを行う。ＦＭＰＫ２００には複数のＦＭチップ２１０が搭載され、各ＦＭチップ２１０に複数の物理ページが存在しているが、ＦＭＰＫ２００では全ＦＭチップ２１０内の各物理ページに対して、

ユニークな番号を付して管理している。この番号のことを「物理ページ番号」と呼ぶ。アクセス対象のデータが格納されている物理ページの物理ページ番号が特定されれば、当該物理ページが存在するＦＭチップ２１０、及びＦＭチップ２１０内位置が一意に特定できる。

[0063] 物理ページのサイズと論理ページのサイズは、同一でも良いし異なっていてもよい。本実施例では、物理ページのサイズは、５２８×１６バイト（＝８ＫＢ＋２５６バイト）とする。論理ページのサイズ（８ＫＢ）よりも２５６バイト大きい理由は、各データには後述するＤＩＦ、及びＥＣＣが付加されるためである。

[0064] 先にも述べたが、ＦＭＰＫ２００は圧縮機能を有する。ＦＭＰＫ２００がデータを圧縮して格納する場合、ＦＭＰＫ２００は論理ページ単位で圧縮を行う。以下、１つの論理ページ上のデータを圧縮することで生成されるデータのことを、「圧縮ページ」と呼ぶ。圧縮ページのサイズは５２０バイトの倍数であり、最小で５２０バイト、最大で（５２０×１６）バイトである。

[0065] 圧縮により、圧縮ページのサイズは物理ページ以下になる。そのため、１つの物理ページに複数の圧縮ページを格納することが可能である。また、１つの圧縮ページが複数の物理ページに跨って格納されることもある。なお、後述するように、圧縮ページがＦＭチップ２１０に格納される際、実際には８バイトのＥＣＣが１または複数付与された状態で格納される。上で説明している圧縮ページのサイズは、ＥＣＣが付与されていない状態でのサイズである。

[0066] 図９に、ＦＭＰＫ２００が管理するマッピングテーブルＴ２１００の一例を示す。各行（レコード）には、各論理ページについての情報が格納される。レコードには、論理ページ番号（Ｔ２１０１）、物理ページ番号（Ｔ２１０２）、サイズ（Ｔ２１０３）、オフセット（Ｔ２１０４）の項目が含まれる。マッピングテーブルＴ２１００は、メモリ２０２上に格納されている。

[0067] 論理ページ番号（Ｔ２１０１）には、当該レコードで管理される論理ページの論理ページ番号が格納される。物理ページ番号（Ｔ２１０２）には、当

該レコードで管理される論理ページがマッピングされた物理ページの物理ページ番号が格納される。

[0068] 物理ページには、圧縮されたデータ（圧縮ページ）が格納されるため、マッピングテーブルＴ２１００では、圧縮ページの格納されている、物理ページ内の領域を特定する情報も管理される。その情報が、サイズ（Ｔ２１０３）とオフセット（Ｔ２１０４）である。オフセット（Ｔ２１０４）は物理ページの先頭アドレスを０としたときの、相対アドレスが格納される。そしてオフセット（Ｔ２１０４）とサイズ（Ｔ２１０３）で特定される領域に圧縮ページが格納されることを表している。

[0069] たとえば図９において、レコードＴ２１００－２（論理ページ番号Ｔ２１０１が１０００のレコード）は、論理ページ番号が１０００の論理ページは、物理ページ番号（Ｔ２１０２）が１０番の物理ページにマッピングされ、かつこの論理ページのデータ（圧縮ページ）は２ＫＢに圧縮されていること、そして物理ページ番号（Ｔ２１０２）が１０番の物理ページの先頭から４ＫＢの位置から始まる２ＫＢの大きさの領域に格納されていることを表している。

[0070] また、１つの圧縮ページが複数の物理ページに跨って格納されることもある。図９において、レコードＴ２１００－１（論理ページ番号Ｔ２１０１が１のレコード）は、論理ページが、２つの物理ページにマッピングされている例を表している。レコードＴ２１００－１で特定される論理ページ（論理ページ番号Ｔ２１０１が１番）は、物理ページ番号（Ｔ２１０２）が５００番の物理ページと４２番の物理ページにマッピングされている。且つ論理ページのデータは、２ＫＢ（＝１ＫＢ＋１ＫＢ）に圧縮されており、物理ページ番号（Ｔ２１０２）が５００番の物理ページの先頭から７ＫＢの位置から始まる１ＫＢの大きさの領域、及び物理ページ番号（Ｔ２１０２）が４２番の物理ページの先頭から０ＫＢの位置から始まる１ＫＢの大きさの領域に跨って格納されていることを表している。

[0071] また、論理ページに対するアクセス（リード、ライト）がなかった場合に

は、物理ページへのマッピングは行われない。図９において、レコードＴ２１００－０は、論理ページ（Ｔ２１０１が２番）に物理ページが割り当てられていない（物理ページ番号Ｔ２１０２がＮＵＬＬ）場合の、レコードの例を表している。ＦＭＰＫ２００は、ＤＫＣ１０から論理ページに対するアクセスを受け付けた時点で、初めて論理ページに対して物理ページをマッピングする。

[0072] 続いて、マッピングテーブルＴ２１００以外にＦＭＰＫ２００が管理している情報について説明する。図１０は状態管理テーブルＴ２０００の内容を示している。状態管理テーブルＴ２０００は、物理容量Ｔ２００１、データ圧縮Ｔ２００２、論理容量Ｔ２００３、接続ＤＫＣ種別Ｔ２００４、所属ＲＡＩＤグループ構成Ｔ２００５、ＲＡＩＤグループ内位置Ｔ２００６、所属ＲＡＩＤグループ番号（所属ＲＧ＃）Ｔ２００７の項目を有している。

[0073] 物理容量Ｔ２００１は、ＦＭＰＫ２００の有するＦＭチップ２１０の合計記憶容量である。データ圧縮Ｔ２００２には「有」または「無」のいずれかの情報が格納され、「有」が格納されている場合には、ＦＭＰＫ２００はＤＫＣ１０からのライトデータを圧縮してＦＭチップ２１０に格納する。データ圧縮Ｔ２００２の設定は、ＤＫＣ１０がＦＭＰＫ２００を用いてＲＡＩＤグループを定義する時等に、ＤＫＣ１０（または管理者）が「有」または「無」を設定する。

[0074] 論理容量Ｔ２００３は、ＦＭＰＫ２００がＤＫＣ１０に提供しているアドレス空間の容量である。ＦＭＰＫ２００にデータが圧縮されずに格納される場合には、原則として物理容量Ｔ２００１の値と論理容量Ｔ２００３の値は等しい。ＦＭＰＫ２００にデータが圧縮されて格納される場合には、論理容量Ｔ２００３の値は物理容量Ｔ２００１より大きくなる。ＤＫＣ１０（または管理者）により、データ圧縮Ｔ２００２には「有」が設定されると、ＦＭＰＫ２００は論理容量Ｔ２００３に仮の値（たとえば物理容量Ｔ２００１の８倍の値等）を格納し、ＤＫＣ１０に対して、仮の値と等しいサイズの記憶空間を提供する。またデータがＦＭチップ２１０に格納されていくにしたが

って、論理容量Ｔ２００３に等しい量のデータを格納できないとＦＭＰＫ２００が判断した場合には、論理容量Ｔ２００３のサイズを小さくする等を行ってよい。逆に、圧縮によりデータサイズが予想以上に小さくなったために、論理容量Ｔ２００３よりも多い量のデータを格納できるとＦＭＰＫ２００が判断した場合、論理容量Ｔ２００３のサイズを大きくする等の処理を行ってよい。

[0075] 接続ＤＫＣ種別Ｔ２００４は、ＦＭＰＫ２００が接続されているストレージシステム１の種別（機種名等）が格納される。ＦＭＰＫ２００がストレージシステム１に接続されたことを契機に、ＤＫＣ１０からストレージシステム１の種別についての情報がＦＭＰＫ２００に渡される。ＦＭＰＫ２００は渡された情報を接続ＤＫＣ種別Ｔ２００４に格納する。

[0076] 所属ＲＡＩＤグループ構成Ｔ２００５、ＲＡＩＤグループ内位置Ｔ２００６、所属ＲＧ＃（Ｔ２００７）は、ＦＭＰＫが所属するＲＡＩＤグループについての情報で、ＲＧ管理情報Ｔ１１００に格納されているＲＡＩＤ構成Ｔ１００３，所属デバイス＃（Ｔ１１０２）、ＲＧ＃（Ｔ１１０１）の情報と同様の情報が格納される。これらの情報は、ＤＫＣ１０がＦＭＰＫ２００を用いてＲＡＩＤグループを定義する際に、ＤＫＣ１０からＦＭＰＫ２００に通知される。

[0077] 続いて、保証コードについて説明する。ストレージコントローラ（ＤＫＣ）１０は、ホスト２から受信したライトデータをＦＭＰＫ２００に格納する過程で、エラー検出用の情報である検証用情報を付加して、データとこの検証用情報とをドライブ１２１に格納する。なお、この検証用情報は、ホスト２が論理ユニットにアクセスする際の最小アクセス単位である、１ディスクブロック（１セクタともいう。また１ディスクブロック（セクタ）のサイズは、一般的には５１２バイトであり、本実施例のストレージシステム１においても、１ディスクブロック（セクタ）のサイズは５１２バイトとする）ごとに付加される。以下ではこの検証用情報のことをＤＩＦと呼ぶ。

[0078] さらにＦＭＰＫ２００内でも、データに更なる検証用情報を付加する処理

が行われる。以下では、ＤＫＣ１０がデータに付加する検証用情報の事を「ＤＫＣ－ＤＩＦ」と呼び、ＦＭＰＫ２００内でＦＭＰＫ２００がデータに付加する検証用情報のことを「ＰＫ－ＤＩＦ」と呼ぶ。また両者を区別せずに呼ぶ場合には、「ＤＩＦ」と表記される。また、ＦＭＰＫ２００がＦＭチップ２１０にデータを格納する際にも、更なる検証用情報を付加するが、この検証用情報のことは「ＥＣＣ」と呼ばれる。

[0079] 図１１を用いて、ＤＫＣ－ＤＩＦ、ＰＫ－ＤＩＦについて説明する。図１１は、ホスト２からのライトデータにＤＩＦが付加される過程を表した概念図である。ホスト２からライトデータ５００がストレージシステム１に到来すると、ＤＫＣ（ストレージコントローラ）１０はＦＭＰＫ２００に当該ライトデータを送信する前に、ＤＫＣ－ＤＩＦ５１１を付加する。ライトデータ５０１は、ＦＭＰＫ２００に送信される直前の、ライトデータの状態（ＤＫＣ－ＤＩＦ５１１が付加されている状態）を表している。ＤＫＣ１０は１セクタ（５１２バイト）毎にＤＫＣ－ＤＩＦ５１１を付加し、ＦＭＰＫ２００に対して、ＤＫＣ－ＤＩＦ５１１の付加されたライトデータ５０１を送信する。

[0080] 続いてＦＭＰＫ２００内でのデータの流れについて説明する。以下では特に、ＦＭＰＫ２００でデータが圧縮される場合について説明する。この場合、ＳＡＳ－ＣＴＬ２０６に到着したデータ（ライトデータ５０１）は、圧縮伸長回路２０４へと渡される。ライトデータ５０２は、ＳＡＳ－ＣＴＬ２０６から圧縮伸長回路２０４に渡されるデータの形式を表している。ＳＡＳ－ＣＴＬ２０６が圧縮伸長回路２０４にデータを渡す際、１セクタ分のライトデータ毎に、ＰＫ－ＤＩＦ５２１を付加する。

[0081] ここで、ＤＫＣ－ＤＩＦ及びＰＫ－ＤＩＦに含まれる情報について説明する。１セクタ（５１２バイト）のデータに対して付されるＤＫＣ－ＤＩＦ５１１のサイズは、８バイトである。ＤＫＣ－ＤＩＦ５１１には、ＣＲＣ（Ｃｙｃｌｉｃ　Ｒｅｄｕｎｄａｎｃｙ　Ｃｈｅｃｋ）、ＲＡＩＤグループ番号、シーケンス番号、アドレス情報が含まれる。

[0082] ＣＲＣは、データ５１０に所定の演算を施して生成される情報である。ＳＡＳ－ＣＴＬ２０６がＤＫＣ１０から、ＤＫＣ－ＤＩＦ５１１の付加されたライトデータ５０１を受信すると、データ５１０に所定の演算を施してＣＲＣを算出する。そして算出されたＣＲＣと、ＤＫＣ－ＤＩＦ５１１内のＣＲＣとが一致するか判定する（以下、この判定のことを「ＣＲＣチェック」と呼ぶ）。両者が一致しない場合、ＤＫＣ１０からＳＡＳ－ＣＴＬ２０６にデータが転送されてくる過程で、障害等の要因でデータ内容が変更されたことを意味する。そのため両者が一致しなかった場合には、ＤＫＣ１０にエラーを返却し、ライトデータ５０１の書き込み処理を中断する。

[0083] アドレス情報は、データ５１０の書き込まれるＦＭＰＫ２００上の論理記憶空間上のアドレスである（またはアドレスの一部がアドレス情報に含まれる。たとえばアドレスが４バイトを超えるものである場合、アドレスの下位４バイトのみがアドレス情報として用いられる）。ＳＡＳ－ＣＴＬ２０６が、ＤＫＣ１０からＤＫＣ－ＤＩＦ５１１の付加されたライトデータ５０１を受信する場合、それとともにライトデータ５０１をＦＭＰＫ２００に格納することを指示する命令（いわゆるＷＲＩＴＥコマンド）も受信する。ＷＲＩＴＥコマンドにも、データ５１０の書き込み先である、ＦＭＰＫ２００上の論理記憶空間上のアドレス情報が含まれているので、ＳＡＳ－ＣＴＬ２０６はＤＩＦに含まれる内のアドレス情報と、ＷＲＩＴＥコマンドに含まれているアドレス情報とが一致するか判定する。両者が一致しない場合には、ＦＭＰＫ２００はＤＫＣ１０にエラーを返却し、ライトデータ５０１の書き込み処理を中断する。

[0084] ＲＡＩＤグループ番号は、データ５１０の書き込まれるＦＭＰＫ２００が所属するＲＡＩＤグループの番号（ＲＧ＃）である。ＦＭＰＫ２００はＤＫＣ１０からあらかじめ、自身が所属するＲＡＩＤグループの番号の情報を受信している。そのため、ＳＡＳ－ＣＴＬ２０６が、ＤＫＣ１０からＤＫＣ－ＤＩＦ５１１の付加されたライトデータ５０１を受信すると、ＤＫＣ－ＤＩＦ５１１に含まれているＲＡＩＤグループ番号と、あらかじめ受信している

ＲＡＩＤグループ番号とを比較することができる。両者が一致しない場合、ＦＭＰＫ２００はＤＫＣ１０にエラーを返却し、ライトデータ５０１の書き込み処理を中断する。

[0085] シーケンス番号は一種の連続番号である。ＤＫＣ１０がＦＭＰＫ２００に複数セクタ分のデータを書き込む際、隣り合うデータ５１０に付されるＤＫＣ－ＤＩＦ５１１には、連続したシーケンス番号が格納されている。たとえばＤＫＣ１０が１０セクタ分のデータを書き込む場合、先頭のデータ５１０に付与されるＤＫＣ－ＤＩＦ５１１には、シーケンス番号０が格納され、次のデータ５１０に付与されるＤＫＣ－ＤＩＦ５１１には、シーケンス番号１が格納されている。そのため、連続した複数セクタのデータがライト（またはリード）される時、ＳＡＳ－ＣＴＬ２０６は隣り合ったセクタのシーケンス番号が連続番号であるか判定する。連続番号が付されていない場合には、ＦＭＰＫ２００はＤＫＣ１０にエラーを返却し、ライトデータ５０１の書き込み処理を中断する。

[0086] ＰＫ－ＤＩＦ５２１にも同様に、データから算出されたＣＲＣが含まれる。ＰＫ－ＤＩＦ５２１に含まれるＣＲＣは、データ５１０とＤＫＣ－ＤＩＦ５１１で構成される５２０バイトのデータから算出されるＣＲＣである。

[0087] 図１１の説明に戻る。圧縮伸長回路２０４に対して、ＤＫＣ－ＤＩＦ５１１、ＰＫ－ＤＩＦ５２１の付加されたデータが渡されると、圧縮伸長回路２０４はデータ圧縮を行う。なお、先に述べたとおり、圧縮は１論理ページ分のデータ毎に行われる。また圧縮の際、データ５１０に付加されているＤＫＣ－ＤＩＦ５１１とＰＫ－ＤＩＦ５２１も併せて圧縮される。つまり圧縮伸長回路２０４は、「データ５１０とＤＫＣ－ＤＩＦ５１１とＰＫ－ＤＩＦ５２１」のセットを１６個まとめて圧縮する。なお圧縮伸長回路２０４は、サイズが５２０バイトの倍数になるように圧縮データを生成する。

[0088] 圧縮データ５３０－０が生成されると、圧縮伸長回路２０４は圧縮データ５３０－０にＰＫ－ＤＩＦ５３１を付加する。ＰＫ－ＤＩＦ５３１は、５２０バイトのデータ（圧縮されたデータ）ごとに付される。ＰＫ－ＤＩＦ５３

１はＰＫ－ＤＩＦ５２１と同様に、データ（圧縮データ５３０－０）から算出されたＣＲＣを含んでいる。また圧縮伸長回路２０４は、圧縮を行う前に、データ５１０とＤＫＣ－ＤＩＦ５１１からＣＲＣを算出する。そして算出されたＣＲＣとＰＫ－ＤＩＦ５２１に含まれるＣＲＣとが一致するか判定する。両者が一致しない場合にはＤＫＣ１０にエラーを返却し、ライトデータ５０１の書き込み処理を中断する。

[0089] 圧縮伸長回路２０４により生成された、圧縮データ５３０－０及びそのＰＫ－ＤＩＦ５３１は、ＦＭ－ＩＦ２０７を経由してＦＭチップ２１０に書き込まれる。ＦＭ－ＩＦ２０７に圧縮データ５３０－０及びそのＰＫ－ＤＩＦ５３１が到来すると、ＦＭ－ＩＦ２０７はＰＫ－ＤＩＦ５３１に含まれているＣＲＣのチェックを行う。チェック方法は圧縮伸長回路２０４で行われているものと同様で、圧縮データ５３０－０からＣＲＣを算出し、算出されたＣＲＣとＰＫ－ＤＩＦ５３１に含まれているＣＲＣが一致するか判定する。両者が一致しない場合、ＤＫＣ１０にエラーを返却し、ライトデータ５０１の書き込み処理を中断する。

[0090] ＣＲＣのチェックでエラーが発生しなかった場合には、ＦＭ－ＩＦ２０７は圧縮データ５３０－０に付与されているＰＫ－ＤＩＦ５３１を除去する。そして圧縮データ５３０－０から別のエラーチェックコードを生成する。このエラーチェックコードは「ＥＣＣ」と呼ばれる。ＥＣＣ５４１はＰＫ－ＤＩＦ５３１と同様、５２０バイトの圧縮データ５３０－０毎に付与される。そしてＦＭ－ＩＦ２０７は、ＥＣＣ５４１の付与された圧縮データ５３０－０をＦＭチップ２１０に書き込む。

[0091] データがＦＭチップ２１０から読み出される場合には、上で説明した処理と逆の処理が行われる。ＦＭ－ＩＦ２０７がＦＭチップ２１０から、ＥＣＣ５４１の付加された圧縮データ５３０－０を読み出し、ＥＣＣ５４１のチェックを行う（圧縮データ５３０－０から算出されたＥＣＣと、ＥＣＣ５４１の比較を行う）。その後ＥＣＣ５４１を圧縮データ５３０－０から除去して、ＰＫ－ＤＩＦ５３１を付与し、圧縮伸長回路２０４へＰＫ－ＤＩＦ５３１

の付与された圧縮データ５３０－０が渡される。圧縮伸長回路２０４では、ＰＫ－ＤＩＦ５３１に含まれるＣＲＣのチェックを行い、その後圧縮データ５３０－０を伸長し、「データ５１０とＤＫＣ－ＤＩＦ５１１とＰＫ－ＤＩＦ５２１」のセットを（１または複数）生成する。

[0092] 生成された「データ５１０とＤＫＣ－ＤＩＦ５１１とＰＫ－ＤＩＦ５２１」のセットが、ＳＡＳ－ＣＴＬ２０６を経由してＤＫＣ１０に転送される際、ＳＡＳ－ＣＴＬ２０６は、ＰＫ－ＤＩＦ５２１に含まれるＣＲＣのチェックを行い、その後「データ５１０とＤＫＣ－ＤＩＦ５１１とＰＫ－ＤＩＦ５２１」のセットからＰＫ－ＤＩＦ５２１を除去し、データ５１０とＤＫＣ－ＤＩＦ５１１をＤＫＣ１０へと転送する。

[0093] 上で説明したデータの流れは、データが圧縮伸長回路２０４によって圧縮される場合の例である。ただしＦＭＰＫ２００は、データを圧縮せずにＦＭチップ２１０に格納することもできる。その場合には、ＳＡＳ－ＣＴＬ２０６でＰＫ－ＤＩＦ５２１の付加されたデータ５１０は圧縮伸長回路２０４を経由せずにＦＭ－ＩＦ２０７に渡される。ＦＭ－ＩＦ２０７では、データ５１０及びそのＤＫＣ－ＤＩＦ５１１とＰＫ－ＤＩＦ５２１が到来すると、ＰＫ－ＤＩＦ５２１に含まれているＣＲＣのチェックを行う。チェック方法は上で説明したものと同様である。

[0094] その後ＦＭ－ＩＦ２０７は、ＰＫ－ＤＩＦ５２１の付加されたデータ５１０からＰＫ－ＤＩＦ５２１を除去し、ＥＣＣを生成して付加する。ここでのＥＣＣは、データ５１０とＤＫＣ－ＤＩＦ５１１で構成される５２０バイトのデータから生成されるものである。そしてＥＣＣの付加されたデータ５１０及びＤＫＣ－ＤＩＦ５１１をＦＭチップ２１０に格納する。

[0095] なお、上で説明したＤＫＣ－ＤＩＦ５１１、ＰＫ－ＤＩＦ５２１、ＰＫ－ＤＩＦ５３１に含まれる情報は一例であって、上で説明した以外の検証用情報が含まれるようにしてもよい。またＤＫＣ－ＤＩＦ５１１は、ＦＭＰＫ２００が接続されるＤＫＣ１０によって付加される情報であるので、ＤＫＣ１０の種類（機種）によって、ＤＫＣ－ＤＩＦ５１１のフォーマットが異なる

場合がある。たとえばＣＲＣやアドレス情報の長さがＤＫＣ１０の種類によって異なっている場合があり得る。またＤＫＣ－ＤＩＦ５１１内における、アドレス情報、シーケンス番号、ＣＲＣの並び順がＤＫＣ１０の種類によって異なっていることもあり得る。本実施例に係るＦＭＰＫ２００は、各機種（ＤＫＣ）のＤＫＣ－ＤＩＦのフォーマットについての情報（ＣＲＣやアドレス情報の格納されている位置など）を把握している。また接続されるＤＫＣから、ＤＫＣの種類（機種）情報を受信することで、ＣＲＣ，アドレス情報、シーケンス番号、ＲＡＩＤグループ番号の格納位置を認識可能に構成されている。

[0096] 続いて、ＦＭＰＫ２００がサポートするコマンドの種類と、コマンドのフォーマットについて説明する。ＦＭＰＫ２００は、ＤＫＣ１０等の上位装置からコマンドを受信し、受信したコマンドに含まれている指示情報（パラメータ）の内容に従った処理（データのリード、ライト等）を行う。ＦＭＰＫ２００は、ＳＳＤやＨＤＤ等の周知の記憶デバイスと同様、データの読み出しを指示するリードコマンド、データの書き込みを指示するライトコマンドもサポートしているが、それ以外のコマンドもサポートしている。ここではＦＭＰＫ２００でサポートされているコマンドのうち、本実施例で行われるデータ復旧処理で用いられるコマンドの内容について説明する。以下、ＦＭＰＫ２００にコマンドを発行する発行元の装置は、ＤＫＣ１０であるという前提で説明する。

[0097] ［圧縮Ｒｅａｄコマンド］

このコマンドは、リードデータを圧縮された状態で、ＤＫＣ１０等のコマンド発行元に対して返却することを指示するためのコマンドである。圧縮Ｒｅａｄコマンドに含まれているパラメータについて図１２を用いて説明する。

[0098] 圧縮Ｒｅａｄコマンド３０００には、オペコード（Ｏｐｃｏｄｅ）３００１、Ｒｅａｄ開始オフセット３００２、Ｒｅａｄサイズ３００３、バッファアドレス３００４、転送サイズ３００５、のパラメータが含まれる。オペコ

ード（Ｏｐｃｏｄｅ）３００１は、ＦＭＰＫ２００がサポートする全コマンドに共通に含まれている情報で（ただしオペコード３００１に含まれている情報の内容は、コマンドによって異なる）、ＦＭＰＫ２００は受信したコマンドのオペコード３００１を参照することで、受信したコマンドの種類を識別する。当然ながら、圧縮Ｒｅａｄコマンド３０００に含まれているオペコード３００１には、圧縮Ｒｅａｄコマンドであることが識別できる情報が格納されている。

[0099] Ｒｅａｄ開始オフセット３００２、Ｒｅａｄサイズ３００３は、ＤＫＣ１０がリードしたいデータ（リード対象データ）が格納されている、ＦＭＰＫ２００上の論理アドレス空間上の領域を特定するための情報である。Ｒｅａｄ開始オフセット３００２には、リード対象データの格納されている、ＦＭＰＫ２００上の論理アドレス空間上領域の先頭アドレスが、Ｒｅａｄサイズ３００３にはリード対象データのサイズが指定される。本実施例では、アドレスを特定する情報として、論理ブロックアドレス（ＬＢＡ）が用いられる。ただし別の実施形態として、アドレスを特定する情報として、論理ページ番号を用いてもよい。またＲｅａｄサイズ３００３には、セクタ数が指定される。ただし別の実施形態として、他の単位（たとえば論理ページ数、バイト数等）が指定されるようにしてもよい。またＲｅａｄサイズ３００３には、ＦＭＰＫ２００内でデータが圧縮されて格納されている、いないにかかわらず、非圧縮時のデータサイズが指定される。

[0100] バッファアドレス３００４、転送サイズ３００５は、リード対象データの転送先の領域（領域の先頭アドレスと領域のサイズ）を特定する情報であり、ＤＫＣ１０がＦＭＰＫ２００に圧縮Ｒｅａｄコマンド３０００を発行する場合には、バッファアドレス３００４にはバッファ１３１のアドレスを指定する。

[0101] ＦＭＰＫ２００は、ＤＫＣ１０から圧縮Ｒｅａｄコマンド３０００を受信すると、圧縮状態のリードデータをＤＫＣ１０に転送（ＤＫＣ１０の、バッファアドレス３００４、転送サイズ３００５で特定されるバッファ１３１上

領域に転送）するとともに、応答情報をＤＫＣ１０に返却する。応答情報には、受信したコマンドに係る処理が正常に行われたか否かを表す情報や、サイズについての情報が含まれる。圧縮Ｒｅａｄコマンド３０００に対する応答情報のフォーマットについて、図１３を用いて説明する。

[0102] 圧縮Ｒｅａｄコマンド３０００に対する応答情報には、転送結果３０１１、Ｒｅａｄサイズ３０１２、バッファ使用サイズ３０１３が含まれる。転送結果３０１１には、「成功」または「エラー」の情報が含まれており、転送結果３０１１が「成功」の場合には、圧縮Ｒｅａｄコマンド３０００に係る処理が正常に行われたことを意味する。

[0103] Ｒｅａｄサイズ３０１２には、リード対象データのサイズ（非圧縮時）が格納される。原則として、圧縮Ｒｅａｄコマンド３０００のＲｅａｄサイズ３００３と同じ値が格納されている。またバッファ使用サイズ３０１３には、ＤＫＣ１０に転送された、圧縮状態のリードデータのサイズが格納される。

[0104] ［圧縮コピーＷｒｉｔｅコマンド］

このコマンドは、ＤＫＣ１０が先に説明した圧縮Ｒｅａｄコマンド３０００を用いて、ＦＭＰＫ２００から読み出したデータ（圧縮状態）を、ＦＭＰＫ２００に格納する際に用いられる。圧縮コピーＷｒｉｔｅコマンドに含まれているパラメータについて、図１４を用いて説明する。

[0105] 圧縮コピーＷｒｉｔｅコマンド３１００には、オペコード（Ｏｐｃｏｄｅ）３１０１、Ｗｒｉｔｅ開始オフセット３１０２、Ｗｒｉｔｅサイズ３１０３、転送元アドレス３１０４、転送サイズ３１０５、のパラメータが含まれる。オペコード（Ｏｐｃｏｄｅ）３１０１は、圧縮Ｒｅａｄコマンドの説明の際に述べたとおり、受信したコマンドの種類をＦＭＰＫ２００が識別するための情報が含まれている。

[0106] Ｗｒｉｔｅ開始オフセット３１０２、Ｗｒｉｔｅサイズ３１０３は、ライト対象データの書き込み先の領域を特定するための情報で、Ｗｒｉｔｅ開始オフセット３１０２には、ライト対象データの書き込み先となる、（ＦＭＰ

Ｋ２００が提供する）論理アドレス空間の先頭アドレスが、Ｗｒｉｔｅサイズ３１０３にはライト対象データのサイズが指定される。なお、圧縮コピーＷｒｉｔｅコマンド３１００が発行される時、ＤＫＣ１０から圧縮されたデータがＦＭＰＫ２００に送信されるが、ここで指定されるＷｒｉｔｅ開始オフセット３１０２、Ｗｒｉｔｅサイズ３１０３は、非圧縮時のライトデータの格納される領域（論理アドレス空間上の領域）が指定される。

[0107] 転送元アドレス３１０４、転送サイズ３１０５は、ＦＭＰＫ２００に転送すべき、圧縮状態のライト対象データが格納されている領域を特定するための情報である。通常、ＤＫＣ１０が圧縮コピーＷｒｉｔｅコマンド３１００をＦＭＰＫ２００に発行する際、ライト対象データはバッファ１３１に格納されている。そのため転送元アドレス３１０４、転送サイズ３１０５にはそれぞれ、バッファ１３１上の圧縮状態のライト対象データが格納されている領域の先頭アドレス、圧縮状態のライト対象データのサイズが指定される。

[0108] 圧縮コピーＷｒｉｔｅコマンド３１００に対する応答情報は、図１６に示されているように、転送結果３０１１のみが含まれている。転送結果３０１１は、圧縮Ｒｅａｄコマンド３０００に対する応答情報に含まれている転送結果３０１１と同じである。つまり「成功」または「エラー」の情報が含まれている。転送結果３０１１が「成功」の場合には、圧縮コピーＷｒｉｔｅコマンド３１００に係る処理が正常に行われたことを意味する。

[0109] ［圧縮パリティ演算Ｗｒｉｔｅコマンド］

このコマンドは、ＤＫＣ１０が先に説明した圧縮Ｒｅａｄコマンド３０００を用いて、ＦＭＰＫ２００から読み出したデータ（圧縮状態）を、ＦＭＰＫ２００に送信するとともに、ＦＭＰＫ２００に対し、送信したデータを用いて冗長データ（パリティ）を算出することを指示するためのコマンドである。なお、以下では、圧縮パリティ演算Ｗｒｉｔｅコマンドとともに ＦＭＰＫ２００に送信されるデータ（ＦＭＰＫ２００から読み出した圧縮状態のデータ）のことを「ライト対象データ」と呼ぶ。圧縮パリティ演算Ｗｒｉｔｅコマンドに含まれているパラメータについて、図１５を用いて説明する。

[0110] 圧縮パリティ演算Ｗｒｉｔｅコマンド３２００には、オペコード（Ｏｐｃｏｄｅ）３２０１、Ｗｒｉｔｅ開始オフセット３２０２、Ｗｒｉｔｅサイズ３２０３、バッファアドレス３２０４、転送サイズ３２０５、ＲＡＩＤグループ内位置３２０６、のパラメータが含まれる。

[0111] オペコード（Ｏｐｃｏｄｅ）３２０１は、受信したコマンドの種類をＦＭＰＫ２００が識別するための情報が含まれている。

[0112] Ｗｒｉｔｅ開始オフセット３２０２、Ｗｒｉｔｅサイズ３２０３は、パリティ演算を行って生成されたデータ（以下、これをパリティ演算結果と呼ぶ）の格納先を特定するための情報で、Ｗｒｉｔｅ開始オフセット３２０２には、パリティ演算結果の書き込み先となる、（ＦＭＰＫ２００が提供する）論理アドレス空間の先頭アドレスが、Ｗｒｉｔｅサイズ３２０３にはパリティ演算結果のサイズが指定される。なお、圧縮コピーＷｒｉｔｅコマンド３１００と同様、ここで指定されるＷｒｉｔｅ開始オフセット３２０２、Ｗｒｉｔｅサイズ３２０３には、非圧縮時のパリティ演算結果の格納される領域（論理アドレス空間上の領域）が指定される。

[0113] バッファアドレス３２０４、転送サイズ３２０５は、圧縮コピーＷｒｉｔｅコマンドの転送元アドレス３１０４、転送サイズ３１０５と同様である。つまりライト対象データが格納されている領域を特定するための情報である。通常、ＤＫＣ１０が圧縮パリティ演算Ｗｒｉｔｅコマンド３２００をＦＭＰＫ２００に発行する際、ライト対象データはバッファ１３１に格納されている。そのためバッファアドレス３２０４、転送サイズ３２０５にはそれぞれ、ライト対象データが格納されているバッファ１３１上領域の先頭アドレス、ライト対象データデータのサイズ（圧縮状態のサイズ）が指定される。

[0114] ＲＡＩＤグループ内位置３２０６には、（圧縮状態の）ライト対象データが、本来格納されていたＦＭＰＫ２００の位置番号が格納される。たとえば図３に示されているＲＡＩＤグループ２０を例にとって説明する。ＦＭＰＫ＃０（位置番号０番）から読み出した圧縮状態のデータを、圧縮パリティ演算Ｗｒｉｔｅコマンドを発行することによってＦＭＰＫ＃３に送信する場合

、ＤＫＣ１０はＲＡＩＤグループ内位置３２０６に、圧縮状態のデータが格納されていたＦＭＰＫ＃０の位置番号（つまり０番）が格納された、圧縮パリティ演算Ｗｒｉｔｅコマンドを作成して、ＦＭＰＫ＃３に作成したコマンドを発行する。

[0115] 圧縮パリティ演算Ｗｒｉｔｅコマンド３２００に対する応答情報は、圧縮コピーＷｒｉｔｅコマンド３１００に対する応答情報と同じである。つまり図１６に示されているように、転送結果３０１１のみが含まれている。

[0116] なお、ＦＭＰＫ２００は、上で説明した圧縮Ｒｅａｄコマンド、圧縮コピーＷｒｉｔｅコマンド、圧縮パリティ演算Ｗｒｉｔｅの他に、状態管理テーブルＴ２０００への情報設定を行うためのコマンド（以下では、「情報設定コマンド」と呼ぶ）、ＦＭＰＫ２００の故障部位診断コマンドも用意されている。状態管理テーブルＴ２０００への情報設定を行うためのコマンドは、所属ＲＡＩＤグループ構成（Ｔ２００５）等の、状態管理テーブルＴ２０００に設定すべき情報を送信するだけであるため、詳細は省略する。また、ＦＭＰＫ２００の故障部位の診断用コマンドは、ＤＫＣ１０が、障害の発生したＦＭＰＫ２００に対して発行するコマンドである。このコマンドを受信したＦＭＰＫ２００は、ＦＭＰＫ２００がＤＫＣ１０に対して提供している記憶アドレス空間の中で、障害が発生しているためＤＫＣ１０がアクセス不可能（リード、ライト不可能）な論理ページの一覧を、コマンド発行元（ＤＫＣ１０）に返却する。

[0117] 続いて、ＦＭＰＫ２００に障害が発生した時に、ストレージシステム１で行われるデータ復旧処理について説明する。以下では一例として、ＲＡＩＤグループ番号が０のＲＡＩＤグループに属する、デバイス＃が１番のＦＭＰＫ２００に障害が発生した場合を例にとって説明する。

[0118] 図１７は、ＤＫＣ１０が実行するデータ復旧処理の全体の流れを表したフローチャートである。ＤＫＣ１０は、ＦＭＰＫ２００にアクセス（リードまたはライト）した際に、アクセスが失敗した（エラーが返却された）場合に、図１７のデータ復旧処理を開始する。ストレージシステム１のメモリ１４

には、データ復旧処理を実行するためのプログラム（データ復旧プログラム）が格納されており、プロセッサ１１がこのデータ復旧プログラムを実行することにより、データ復旧処理が実行される。

[0119] データ復旧処理では最初にスペアデバイスの選択が行われる。プロセッサ１１は、デバイス管理テーブルＴ１０００を参照し、所属ＲＧ＃（Ｔ１００３）が「未割当（スペア）」であるＦＭＰＫ２００を１つ選択する（Ｓ２０）。以下この選択されたＦＭＰＫ２００を、「復旧先デバイス」または「スペアデバイス」と呼ぶ。なお、復旧先デバイスの選択の際、障害が発生したＦＭＰＫ２００（以下、このＦＭＰＫ２００を「復旧元デバイス」と呼ぶ）と同等のＦＭＰＫ２００を選択する。具体的には、圧縮機能サポートＴ１００５、パリティ演算機能サポートＴ１００６、サイズＴ１００７の内容が同じＦＭＰＫ２００を選択する。

[0120] ただし、復旧元デバイスで、圧縮機能サポートＴ１００５またはパリティ演算機能サポートＴ１００６が「サポート」であったが、圧縮機能サポートＴ１００５またはパリティ演算機能サポートＴ１００６が「サポート」されているスペアデバイスが存在しない場合には、圧縮機能サポートＴ１００５またはパリティ演算機能サポートＴ１００６が「未サポート」であるＦＭＰＫ２００をスペアデバイスとして選択する。ただしこの場合、ＦＭＰＫ２００が有する圧縮機能、パリティ演算機能を用いたデータ復旧処理を行うことができず、周知のストレージ装置で行われている、ストレージコントローラによるデータ復旧処理を行う。

[0121] また、デバイス管理テーブルＴ１０００内の、復旧先デバイスのレコードについて、所属ＲＧ＃（Ｔ１００３）に、障害の発生したＦＭＰＫ２００の属するＲＡＩＤグループ番号と同じ番号を格納し、デバイスステータスＴ１００４に「障害復旧中（復旧先）」を格納する。さらに障害の発生したＦＭＰＫ２００についてのレコードのデバイスステータスＴ１００４に「障害復旧中（復旧元）」を格納する。復旧先デバイスとして、デバイス＃が４番のＦＭＰＫ２００（以下、デバイス＃がｘ（ｘは整数値）のＦＭＰＫ２００の

ことを、「ＦＭＰＫ＃ｘ」と表記する）が選択された場合、デバイス管理テーブルＴ１０００は、図５に示された状態になる。

[0122] 本実施例におけるデータ復旧処理では、復旧元デバイスのデータを復旧するために、大きくは以下の２通りの方法が用いられる。１つ目の方法は、復旧元デバイスからデータを読み出し、それを復旧先デバイスに書き込む（コピーする）方法である。復旧元デバイスに障害が発生した場合、全記憶領域がアクセス不可能な状態になることは少ない。そのため、復旧元デバイスの記憶空間の中でも、ＤＫＣ１０からアクセス可能（リード可能）な領域が存在することがある。その場合、復旧元デバイスの記憶空間のうち、ＤＫＣ１０からリード可能な領域については、このリード可能な領域からデータを読み出して復旧先デバイスにコピーすることによってデータ復旧を行う。この方法のことを、以下では「コピー復旧」と呼ぶ。ただしこの方法は、ＤＫＣ１０からアクセス不可能（リード不可能）な領域については利用できない。

[0123] ２つ目の方法は、復旧元デバイスが所属しているＲＡＩＤグループ内の各デバイスからデータを読み出し、読み出された各データを用いて所定の演算を行うことで、復旧元デバイスに格納されていたデータを再生成する方法である。この方法のことは、以下では「コレクション（ｃｏｒｒｅｃｔｉｏｎ）」あるいは「コレクションコピー」と呼ばれる。データの再生成方法は、たとえば特許文献１に記載の演算を行えばよい。コレクションを行う場合、本実施例に係るストレージシステム１では３つの方式を採り得るが、この３つの方式については後述する。

[0124] Ｓ３０では、プロセッサ１１は復旧先デバイスに対して、情報設定コマンドを送信することにより、復旧先デバイスの状態管理テーブルＴ２０００に、所属ＲＡＩＤグループ構成（Ｔ２００５），ＲＡＩＤグループ内位置（Ｔ２００６）、データ圧縮（Ｔ２００２）の設定を行う。所属ＲＡＩＤグループ構成（Ｔ２００５）には、復旧先デバイスが所属するＲＡＩＤグループのＲＡＩＤ構成と同じもの（ＲＧ管理テーブルＴ１１００内のＴ１１０３に格納されている情報）が設定される。ＲＡＩＤグループ内位置（Ｔ２００６）

には、障害が発生したＦＭＰＫ２００のＲＡＩＤグループ内位置（ＲＧ管理テーブルＴ１１００内のＴ１１０２に格納されている情報）が設定される。またデータ圧縮（Ｔ２００２）についても、障害が発生したＦＭＰＫ２００と同じ情報が設定される。つまり障害が発生したＦＭＰＫ２００（の所属しているＲＡＩＤグループ）がデータ圧縮を行っていた場合には、データ圧縮（Ｔ２００２）に「有」が設定される。障害が発生したＦＭＰＫ２００でデータ圧縮が行われていなかった場合には、データ圧縮（Ｔ２００２）に「無」が設定される。

[0125] Ｓ４０では、プロセッサ１１は復旧元デバイスに対して、故障部位診断コマンドを発行する。このコマンドを受信した復旧元デバイスは、プロセッサ１１に対して診断結果を返送する。診断結果には先に述べたとおり、アクセス不可能な論理ページの一覧が含まれている。

[0126] Ｓ５０では、プロセッサ１１はコピー管理テーブルを作成する。コピー管理テーブルは、プロセッサ１１がデータ復旧処理を行う際に把握しておくべき情報を集約しているテーブルである。図１８を用いてコピー管理テーブルＴ１５００で管理される情報の内容を説明する。コピー管理テーブルＴ１５００には、障害ＲＧ＃（Ｔ１５０１）、復旧元デバイス（Ｔ１５０２）、復旧先デバイス（Ｔ１５０３）、コレクション方式（Ｔ１５０４），復旧デバイス容量（Ｔ１５０５）、コピー方法ビットマップ（Ｔ１５０６）、復旧完了済みオフセット（Ｔ１５０７）、の項目が含まれる。

[0127] 障害ＲＧ＃（Ｔ１５０１）には、データ復旧処理による復旧対象であるＲＡＩＤグループのＲＧ＃が格納される。復旧元デバイス（Ｔ１５０２）、復旧先デバイス（Ｔ１５０３）にはそれぞれ、復旧元デバイスのデバイス＃、復旧先デバイスのデバイス＃が格納される。コレクション方式（Ｔ１５０４）には、データ復旧処理で実施されるコレクション方式についての情報が格納される。本実施例に係るストレージシステム１では、以下で説明する３通りのコレクション方式が選択可能である。

[0128] １つ目の方式（以下では「方式１」と呼ぶ）は、周知のストレージ装置で

実施されるコレクション方法と同じものである。具体的にはＤＫＣ１０が、ＲＡＩＤグループ内の、復旧元ＦＭＰＫ２００以外の正常なＦＭＰＫ２００からデータを読み出し、読み出されたデータからＤＫＣ１０内のパリティ演算回路１５を用いて、復旧元デバイスに格納されていたデータを再生成する。そしてＤＫＣ１０は再生成されたデータを復旧先に書き出す。

[0129] ２つ目の方式（以下では「方式２」と呼ぶ）は、復旧先ＦＭＰＫ２００がパリティ演算機能を有している場合に適用可能な方法で、特許文献１にも開示されている方法である。具体的にはＤＫＣ１０が、ＲＡＩＤグループ内の復旧元ＦＭＰＫ２００以外の正常なＦＭＰＫ２００からデータを読み出し、読み出されたデータを復旧先ＦＭＰＫ２００に送信する。復旧先ＦＭＰＫ２００では、自身の有するパリティ演算機能（パリティ演算回路２０５）を用いて、ＤＫＣ２００から送信されたデータからパリティを算出することでデータを再生成する。なお、方式１、方式２は公知の方法であるので、本実施例では説明を省略する。

[0130] ３つ目の方式（以下では「方式３」と呼ぶ）は、復旧先ＦＭＰＫ２００がパリティ演算機能、及び圧縮機能を有している場合に適用可能な方法である。以下では、方式３によるデータ復旧が行われるという前提で説明を行う。そのため方式３の具体的な内容についても、図１７以降の処理の流れを説明する過程で説明する。

[0131] コピー管理テーブルＴ１５００の説明に戻る。コレクション方式（Ｔ１５０４）には、上で説明した「方式１」、「方式２」、「方式３」のいずれかが格納される。「方式１」、「方式２」、「方式３」のいずれが格納されるかは、データ復旧対象のＲＡＩＤグループに所属するＦＭＰＫ２００が、パリティ演算機能、データ圧縮機能を備えているか否かで決定される。ＦＭＰＫ２００が、パリティ演算機能、データ圧縮機能の両方を備えている場合には、プロセッサ１１は、コレクション方式（Ｔ１５０４）に「方式３」を設定する。ＦＭＰＫ２００が、パリティ演算機能を有し、データ圧縮機能を有していない場合には、プロセッサ１１はコレクション方式（Ｔ１５０４）に

「方式２」を設定する。ＦＭＰＫ２００が、パリティ演算機能、データ圧縮機能のいずれも有していない場合には、プロセッサ１１はコレクション方式（Ｔ１５０４）に「方式１」を設定する。

[0132] コピー方法ビットマップ（Ｔ１５０６）には、データ復旧処理を実施するデータについての情報が格納される。先にも述べたが、ＦＭＰＫ２００に障害が発生した時、ＦＭＰＫ２００内の全論理ページがアクセス不可能（具体的にはリード不可能）になっていないこともある。そのため、本実施例に係るストレージシステム１では、Ｓ４０における診断の結果に基づいて、復旧元ＦＭＰＫ２００の中で、ＤＫＣ１０がアクセス不可能な論理ページについてのみ、コレクションを行い、アクセス可能な論理ページについては、コピー復旧によるデータ復旧を行う。

[0133] 復旧元ＦＭＰＫ２００の論理アドレス空間の大きさが、ｎ論理ページ分である場合、ＤＫＣ１０はコピー方法ビットマップ（Ｔ１５０６）として、ｎビットのサイズのビットマップを用意する。そしてコピー方法ビットマップ（Ｔ１５０６）のｋビット目（１≦ｋ≦ｎ）のビットは、復旧元ＦＭＰＫ２００の論理アドレス空間の、ｋ番目の論理ページについて、コレクションを行うか否かを表している。Ｓ４０における診断の結果、復旧元ＦＭＰＫ２００の論理アドレス空間の、ｋ番目の論理ページがアクセス不可能である場合、コピー方法ビットマップ（Ｔ１５０６）のｋビット目のビットに１が格納される（つまりｋ番目の論理ページについては、コレクションによるデータ復旧が行われる）。ｋ番目の論理ページがアクセス可能である場合は、コピー方法ビットマップ（Ｔ１５０６）のｋビット目のビットに０が格納される（つまりｋ番目の論理ページについては、コピー復旧が行われる）。プロセッサ１１は、Ｓ７０において、このビットマップの内容に基づいて、次に行うべき処理を決定する。

[0134] 復旧完了済みオフセット（Ｔ１５０７）には、データ復旧が完了した、論理アドレス空間のアドレスが格納される。本実施例では復旧完了済みオフセット（Ｔ１５０７）に格納するアドレスとして、論理ページ番号を用いる。

ただしその他のアドレス情報（たとえばＬＢＡ等）を用いてもよい。

[0135] また本実施例では、データ復旧は、復旧元デバイスの論理空間の先頭アドレス（論理ページ番号が０の論理ページ）から順に、データ復旧を行う。そのためＳ５０では、プロセッサ１１は復旧完了済みオフセット（Ｔ１５０７）に、初期値として０を格納する。そして１論理ページ分のデータ復旧が完了したら、プロセッサ１１は復旧完了済みオフセット（Ｔ１５０７）に、データ復旧の完了した論理ページのページ数（１）を加算する。

[0136] Ｓ５０の説明に戻る。Ｓ５０でプロセッサ１１は、コピー管理テーブルＴ１５００の障害ＲＧ＃（Ｔ１５０１）、復旧元デバイス（Ｔ１５０２）、復旧先デバイス（Ｔ１５０３）、コレクション方式（Ｔ１５０４），復旧デバイス容量（Ｔ１５０５）、コピー方法ビットマップ（Ｔ１５０６）、復旧完了済みオフセット（Ｔ１５０７）に、情報を格納する。ＲＡＩＤグループ番号が０のＲＡＩＤグループに属する、デバイス＃が１番のＦＭＰＫ２００に障害が発生した場合、プロセッサ１１は障害ＲＧ＃（Ｔ１５０１）に０を、復旧元デバイス（Ｔ１５０２）に１（ＦＭＰＫ＃１）を格納する。また復旧先デバイスとしてＦＭＰＫ＃４が選択された場合には、復旧先デバイス（Ｔ１５０３）には４（ＦＭＰＫ＃４）が格納される。

[0137] 障害が発生したＲＡＩＤグループ（ＲＧ＃０）の状態が図６に示されたものであった場合、そして復旧先デバイスであるＦＭＰＫ＃４の属性（特に圧縮機能サポートＴ１００５，パリティ演算機能サポートＴ１００６）が図５に示されたものであった場合を想定する。この場合、ＲＧ＃０は圧縮を実施しており、またＦＭＰＫ２００の有するパリティ演算機能を用いたパリティ生成を行っている。またＦＭＰＫ＃４は、圧縮機能、パリティ演算機能のいずれもサポートしている。そのため、プロセッサ１１はコレクション方式Ｔ１５０４として「方式３」を格納する。また、復旧デバイス容量（Ｔ１５０５）には、復旧先デバイスのサイズが格納される。

[0138] コピー方法ビットマップ（Ｔ１５０６）には、上で説明したとおり、Ｓ４０における診断の結果に基づいて設定が行われる。復旧元ＦＭＰＫ２００の

論理アドレス空間の、ｋ番目の論理ページがアクセス不可能である場合、コピー方法ビットマップ（Ｔ１５０６）のｋビット目のビットに１が格納される。ｋ番目の論理ページがアクセス可能である場合は、コピー方法ビットマップ（Ｔ１５０６）のｋビット目のビットに０が格納される。

[0139] またＳ５０では、プロセッサ１１は復旧完了済みオフセット（Ｔ１５０７）を初期化（０を格納）する。

[0140] Ｓ６０ではプロセッサ１１は、コピー方法ビットマップ（Ｔ１５０６）の、（復旧完了済みオフセット（Ｔ１５０７）＋１）番目のビットを選択し、Ｓ７０ではデータ復旧を実施する論理ページの復旧方法を決定する。選択されたビットが０であれば（Ｓ７０：コピー復旧）、コピー復旧を行う（Ｓ７１）。選択されたビットが１であれば（Ｓ７０：コレクション）、コレクションコピーを行う（Ｓ７２）。Ｓ７１、Ｓ７２の処理内容については後述する。

[0141] Ｓ７１またはＳ７２の処理が完了すると、プロセッサ１１はＳ８０の処理を行う。Ｓ８０ではプロセッサ１１は、復旧完了済みオフセット（Ｔ１５０７）に、復旧されたデータのサイズを加算する。論理ページ単位での復旧の場合、１が加算される。Ｓ９０でプロセッサは、復旧元デバイスの全領域について、データ復旧が完了したか判断する。データ復旧完了の判断は、復旧デバイス容量（Ｔ１５０５）から復旧元デバイスの終端論理ページ番号を求め（復旧デバイス容量（Ｔ１５０５）を論理ページサイズで除算する）、復旧完了済みオフセット（Ｔ１５０７）で示される論理ページ番号が、復旧元デバイスの終端論理ページ番号に達したか判断すればよい。

[0142] データ復旧が完了していない場合（Ｓ９０：ＮＯ）、プロセッサ１１は再びＳ６０以降の処理を実施する。データ復旧が完了した場合（Ｓ９０：ＹＥＳ）、プロセッサ１１は管理情報の更新を行う。具体的には、デバイス管理テーブルＴ１０００内の復旧元デバイスのレコードについて、デバイスステータス（Ｔ１００４）を「閉塞」に変更する。またデバイス管理テーブルＴ１０００内の、復旧先デバイスのレコードについて、デバイスステータス（

Ｔ１００４）を「正常」に変更する。またプロセッサ１１は、ＲＧ管理テーブルＴ１１００内の、所属デバイス（Ｔ１１０２）の情報を変更する。具体的には所属デバイス（Ｔ１１０２）に登録されている情報の中から、復旧元デバイスのデバイス＃を削除し、復旧先デバイスのデバイス＃を追加する（Ｓ１００）。

[0143] 以上がデータ復旧処理の全体の流れである。なお、上では原則として、論理ページごとにデータ復旧を行う例を説明したが、データ復旧の単位は論理ページに限られない。論理ページよりも大きい単位（たとえば論理ページの整数倍単位）でデータ復旧を行うようにしてもよいし、論理ページより小さい単位でデータ復旧を行っても、データ復旧を行うことができる。

[0144] 続いて、Ｓ７１の処理（つまり復旧元デバイスから復旧先デバイスにデータをコピーすることによる復旧処理）の詳細を図１９を用いて説明する。以下では、論理ページ単位でデータをコピーする場合について説明する。最初にプロセッサ１１は、バッファ１３１にリードデータを格納するための領域として、１論理ページ分の領域を確保する。そして、復旧元デバイスに対して圧縮Ｒｅａｄコマンドを発行する（Ｓ２１０）。圧縮Ｒｅａｄコマンドのパラメータとして、復旧完了済みオフセットＴ１５０７（論理ページ番号）をＬＢＡに変換した値をＲｅａｄ開始オフセット３００２に設定する。またＲｅａｄサイズ３００３には１論理ページ分のサイズを用いる。そしてバッファアドレス３００４、転送サイズ３００５には、先に確保したバッファ１３１上の領域の情報を用いる。

[0145] 圧縮Ｒｅａｄコマンドを発行した後、プロセッサ１１は復旧元デバイスから圧縮Ｒｅａｄコマンドに対する応答情報を受信する（Ｓ２２０）。応答情報に含まれる転送結果３０１１が「エラー」であった場合（Ｓ２３０：ＮＯ）、プロセッサ１１はコレクションコピーを実施して（Ｓ２８０）、処理を終了する。コレクションコピーで行われる処理の詳細は後述する。

[0146] 応答情報に含まれる転送結果３０１１が「成功」であった場合（Ｓ２３０：ＹＥＳ）、圧縮コピーＷｒｉｔｅコマンドを復旧先デバイスに発行するこ

とにより、Ｓ２１０、Ｓ２２０で読み出したデータの復旧先デバイスへの書き込みを指示する（Ｓ２５０）。ここで発行される圧縮コピーＷｒｉｔｅコマンドのパラメータである、Ｗｒｉｔｅ開始オフセット３１０２、Ｗｒｉｔｅサイズ３１０３、転送元アドレス３１０４にはそれぞれ、Ｓ２１０で発行された圧縮Ｒｅａｄコマンドのパラメータ、つまり、Ｒｅａｄ開始オフセット３００２、Ｒｅａｄサイズ３００３、バッファアドレス３００４と同じ値が指定される。また圧縮コピーＷｒｉｔｅコマンドの転送サイズ３１０５には、Ｓ２２０で受信した応答情報に含まれているバッファ使用サイズ３０１３の値が指定される。

[0147] Ｓ２６０でプロセッサ１１は、復旧先デバイスから圧縮コピーＷｒｉｔｅコマンドに対する応答情報を受信する。応答情報に含まれる転送結果３０１１が「成功」であった場合（Ｓ２７０：ＹＥＳ）には、処理を終了する。応答情報に含まれる転送結果３０１１が「エラー」であった場合（Ｓ２７０：ＮＯ）、プロセッサ１１はコレクションコピーを実施して（Ｓ２８０）、処理を終了する。

[0148] 続いて、Ｓ７２（またはＳ２８０）の処理（コレクションコピー）の詳細を図２０を用いて説明する。なお、先にも述べたが、本実施例に係るストレージシステム１では、コレクションとして３つの方式が選択可能であるが、以下では方式３におけるコレクション方式のみ説明する。つまり復旧対象ＲＡＩＤグループに属するＦＭＰＫ２００が、パリティ演算機能、データ圧縮機能を備えていることが前提である。また、特に断りのない限り、ＲＡＩＤ構成がＲＡＩＤ４やＲＡＩＤ５のように、１ストライプライン内のパリティストライプ数が１であるＲＡＩＤグループに係るコレクション方法を説明する。また、図１９の処理と同様、以下では、論理ページ単位でコレクションコピーを行う場合について説明する。

[0149] まずプロセッサ１１は、復旧対象ＲＡＩＤグループの中で、まだＳ４１０以降の処理を行っていないＦＭＰＫ２００のうち、正常なＦＭＰＫ２００を１つ選択する（Ｓ４００）。続いて変数ｒ、ｗを用意し、両方の変数に０を

代入することで初期化を行う（Ｓ４１０）。なお変数ｒは、ＦＭＰＫ２００からのデータ読み出しに失敗した場合に、再試行を行った回数を記録するために用いられる。また変数ｗは復旧先デバイスへのデータ書き込みに失敗した場合に、再試行を行った回数を記録するために用いられる。

[0150] Ｓ４２０でプロセッサ１１は、リードデータを格納するための領域として、１論理ページ分の領域をバッファ１３１に確保し、さらにＳ４００で選択したＦＭＰＫ２００に対して圧縮Ｒｅａｄコマンドを発行する。圧縮Ｒｅａｄコマンドのパラメータとして指定する内容は、Ｓ２１０で指定されたものと同様である。

[0151] Ｓ４３０では、プロセッサ１１は圧縮Ｒｅａｄコマンドを発行したＦＭＰＫ２００から応答情報を受信する。応答情報に含まれる転送結果３０１１が「エラー」であった場合（Ｓ４４０：ＮＯ）、プロセッサ１１はＳ４５０の処理を行う。応答情報に含まれる転送結果３０１１が「成功」であった場合（Ｓ４４０：ＹＥＳ）、プロセッサ１１はＳ４８０の処理を実行する。

[0152] 転送結果３０１１が「エラー」であった場合（Ｓ４４０：ＮＯ）、プロセッサ１１は変数ｒが一定値以上であるか判定し（Ｓ４５０）、変数ｒがまだ一定値以上でない場合（Ｓ４５０：ＮＯ）、ｒに１を加算し（Ｓ４６０），再びＳ４２０の処理を実行する。変数ｒが一定値以上である場合（Ｓ４５０：ＹＥＳ）、プロセッサ１１はＳ５４０の処理を行う。Ｓ５４０では、ＲＡＩＤグループの状態（ＲＧステータスＴ１１０６）を「障害復旧失敗」に変更し、データ復旧処理を中断する。また管理端末に、データ復旧が失敗した旨を表示する。あるいはホスト２に、データ復旧が失敗した旨を通知するようにしてもよい。

[0153] Ｓ４４０の判定で、転送結果３０１１が「成功」の場合（Ｓ４４０：ＹＥＳ）、プロセッサ１１は圧縮パリティ演算Ｗｒｉｔｅコマンドを復旧先デバイスに発行することにより、Ｓ４２０，Ｓ４３０で読み出したデータの復旧先デバイスへの書き込みを指示する（Ｓ４８０）。ここで発行される圧縮パリティ演算Ｗｒｉｔｅコマンドのパラメータとして、Ｗｒｉｔｅ開始オフセ

ット３２０２、Ｗｒｉｔｅサイズ３２０３、バッファアドレス３２０４にはそれぞれ、Ｓ４２０で発行された圧縮Ｒｅａｄコマンドのパラメータ、つまり、Ｒｅａｄ開始オフセット３００２、Ｒｅａｄサイズ３００３、バッファアドレス３００４と同じ値が指定される。また転送サイズ３２０５には、Ｓ４３０で受信した応答情報に含まれているバッファ使用サイズ３０１３の値が指定される。そしてＲＡＩＤグループ内位置３２０６には、Ｓ４００で選択したＦＭＰＫ２００（つまりＳ４２０で圧縮Ｒｅａｄコマンドを発行したＦＭＰＫ２００）の位置番号（ＲＧ管理テーブルＴ１１００の位置番号（Ｔ１１０２’）を参照することで特定可能である）が指定される。

[0154] Ｓ４９０でプロセッサ１１は、復旧先デバイスから圧縮パリティ演算Ｗｒｉｔｅコマンドに対する応答情報を受信する。応答情報に含まれる転送結果３０１１が「成功」であった場合（Ｓ５００：ＹＥＳ）には、ＲＡＩＤグループを構成する全ての正常なＦＭＰＫ２００についてＳ４１０～Ｓ５００の処理を行ったか判定し（Ｓ５５０）、全ての正常なＦＭＰＫ２００に対して処理が完了している場合（Ｓ５５０：ＹＥＳ）には、処理を終了する。ＲＡＩＤグループを構成する全ての正常なＦＭＰＫ２００のうち、まだＳ４１０～Ｓ５００の処理が行われていないＦＭＰＫ２００が存在する場合には（Ｓ５５０：ＮＯ）、プロセッサ１１は再びＳ４００以降の処理を実施する。

[0155] Ｓ５００の判定で、転送結果３０１１が「エラー」の場合（Ｓ５００：ＮＯ）、プロセッサ１１は変数ｗが一定値以上であるか判定し（Ｓ５１０）、変数ｗがまだ一定値以上でない場合（Ｓ５１０：ＮＯ）、ｗに１を加算し（Ｓ５２０），再びＳ４２０の処理を実行する。変数ｗが一定値以上である場合（Ｓ５１０：ＹＥＳ）、プロセッサ１１はＳ５４０の処理を行う。Ｓ５４０では、ＲＡＩＤグループの状態（ＲＧステータスＴ１１０６）を「障害復旧失敗」に変更し、データ復旧処理を中断する。

[0156] なお、上で説明した処理は、１ストライプライン内のパリティストライプ数が１であるＲＡＩＤグループについてのデータ復旧処理であるため、Ｓ５５０ではＲＡＩＤグループを構成する全ての正常なＦＭＰＫ２００について

Ｓ４１０～Ｓ５００の処理を行ったかの判定が行われている。一方、１ストライプライン内にｎ個のデータストライプと複数（たとえばＲＡＩＤ６の場合には２個）のパリティストライプが存在するＲＡＩＤグループ（一例としてＲＡＩＤ６）についてのデータ復旧処理が行われる場合、Ｓ５５０では、ｎ個のＦＭＰＫ２００についてＳ４１０～Ｓ５００の処理を行ったか判定されればよい。

[0157] 以上が、データ復旧処理において、ＤＫＣ１０で実施される処理の流れである。次に、ＤＫＣ１０がＦＭＰＫ２００に対して圧縮Ｒｅａｄコマンド等のコマンドを発行した時に、ＦＭＰＫ２００が実行する処理の流れについて説明する。なお、ＦＭＰＫ２００がＤＫＣ１０等の上位装置からコマンドを受信した時、ＦＭＰＫ２０２のメモリ２０２に格納されているコマンド処理用プログラムをプロセッサ２０３が実行することによって、そのコマンドに係る処理が行われる。

[0158] まずＦＭＰＫ２００がＤＫＣ１０から、圧縮Ｒｅａｄコマンドを受信した時に行われる処理の流れを、図２１を用いて説明する。ＦＭＰＫ２００がＤＫＣ１０から圧縮Ｒｅａｄコマンドを受信すると、プロセッサ２０３は変数ｕ、ｃを用意し、これらの変数を初期化（０を代入）する（Ｓ１０２０）。なお、変数ｕは主として、ＦＭＰＫ２００がＤＫＣ１０に返却する応答情報に含まれる、Ｒｅａｄサイズ３０１２の算出のために用いられ、変数ｃはバッファ使用サイズ３０１３の算出のために用いられる。

[0159] 続いてＳ１０３０でプロセッサ２０３は、圧縮Ｒｅａｄコマンドで指定されたリード対象データの格納されている、ＦＭチップ２１０上アドレス（正確には、物理ページの物理ページ番号及び物理ページ内オフセット）の算出を行う。具体的には圧縮Ｒｅａｄコマンドのパラメータに含まれている、Ｒｅａｄ開始オフセット３００２に変数ｕの値を加算して算出されるアドレスから、論理ページ番号を算出する。そしてマッピングテーブルＴ２１００を参照することによって、算出された論理ページ番号がマッピングされている物理ページの物理ページ番号（Ｔ２１０２）、物理ページ内オフセット（Ｔ

２１０４）、そしてサイズ（Ｔ２１０３）を求める。１論理ページが複数の物理ページに跨って格納されている場合には、物理ページ番号（Ｔ２１０２）、物理ページ内オフセット（Ｔ２１０４）、サイズ（Ｔ２１０３）のセットが複数求められる。

[0160] 続いてプロセッサ２０３は、Ｓ１０３０で求められた物理ページ番号（Ｔ２１０２）、物理ページ内オフセット（Ｔ２１０４）、サイズ（Ｔ２１０３）で特定される領域（ＦＭチップ２１０上の領域）からデータを読み出す（Ｓ１０４０）。先にも述べたが、ＦＭチップ２１０に格納されているデータには、５２０バイトのデータ毎にＥＣＣが付されている。データ読み出しの過程において、ＦＭ－ＩＦ２０７は、ＥＣＣを用いてデータのチェックを行う。データのチェックにおいて、エラーが発生しなかった場合（データに付されているＥＣＣと、データから算出されるＥＣＣとが一致している場合）には、ＦＭ－ＩＦ２０７は、５２０バイトのデータ毎に付されているＥＣＣを除去する。代わりにＦＭ－ＩＦ２０７が５２０バイトのデータ毎にＰＫ－ＤＩＦを作成してデータに付加し、ＰＫ－ＤＩＦの付加されたデータを、メモリ２０２上に格納する。その後プロセッサ２０３に読み込みが成功した旨を通知する。

[0161] また、ＦＭチップ２１０の最小アクセス（リード、ライト）単位は、物理ページであるので、Ｓ１０４０では物理ページ単位でのデータ読み出しが行われる。そのため、１論理ページが複数物理ページに跨って格納されている場合、論理ページが圧縮された状態で物理ページに格納され、かつその物理ページに他の論理ページのデータが格納されている場合などには、アクセス対象の論理ページ（あるいは論理ページのデータが圧縮された、圧縮ページ）以外の情報も一緒に読み出される。その場合には、物理ページのデータを読み出してメモリ２０２に格納したあと、アクセス対象の論理ページ以外の情報をメモリ２０２から削除する。

[0162] 一方データのチェックにおいて、エラーが発生した場合（データに付されているＥＣＣと、データから算出されるＥＣＣとが一致しない場合）には、

ＦＭ－ＩＦ２０７はプロセッサ２０３に、読み込みが失敗した旨を通知する。

[0163] プロセッサ２０３は、ＦＭ－ＩＦ２０７から読み込みが失敗した旨を受信した場合（Ｓ１０５０：ＮＯ）、ＤＫＣ１０に返却する応答情報の作成を行う（Ｓ１１６０）。ここで作成される応答情報は、転送結果３０１１に「エラー」が格納された応答情報である。そしてプロセッサ２０３は、作成された応答情報をＤＫＣ１０に返却し（Ｓ１１５０）、処理を終了する。

[0164] 一方ＦＭ－ＩＦ２０７から読み込みが成功した旨をプロセッサ２０３が受信した場合（Ｓ１０５０：ＹＥＳ）、Ｓ１０６０以降の処理が行われる。Ｓ１０６０でプロセッサ２０３は、リードデータが圧縮されているか否かであるか判定する。リードデータが圧縮されているか否かの判定は、Ｓ１０３０で求められたサイズ（Ｔ２１０３）が論理ページのサイズと同じか否かを参照することで判定できる。リードデータが圧縮されていない場合（サイズ（Ｔ２１０３）が論理ページのサイズと同じ場合）、プロセッサ２０３は圧縮伸長回路２０４を用いて、リードデータを圧縮する。圧縮伸長回路２０４ではデータの圧縮後、圧縮されたデータにＰＫ－ＤＩＦを付与し、メモリ２０２に格納する。

[0165] リードデータが圧縮されていた場合（Ｓ１０６０：ＮＯ）には、Ｓ１０７０は実行せず、Ｓ１０８０以降の処理が行われる。なお、以下では、Ｓ１０７０で圧縮されたリードデータのサイズ（Ｓ１０７０が実行されなかった場合には、Ｓ１０４０で読み出されたリードデータのサイズ）をｃ’とする。

[0166] Ｓ１０８０でプロセッサ２０３は、ｃ＋ｃ’が、圧縮Ｒｅａｄコマンドのパラメータで指定されている転送サイズ３００５以下であるかを判定する。ｃ＋ｃ’が転送サイズ３００５を超過している場合（Ｓ１０８０：ＮＯ）、プロセッサ２０３は、転送結果３０１１に「成功」が格納された応答情報を作成し（Ｓ１１４０）、作成された応答情報をＤＫＣ１０に返却し（Ｓ１１５０）、処理を終了する。ｃ＋ｃ’が転送サイズ３００５以下である場合（Ｓ１０８０：ＹＥＳ）、Ｓ１０９０以降の処理が行われる。

[0167] 変数ｃには、これまでに何回かＳ１０３０～Ｓ１０７０が実行されている場合、ＤＫＣ１０のバッファ１３１に転送済みのデータの合計量が格納されている。一方ｃ’は、これからＤＫＣ１０のバッファ１３１に転送すべきデータのサイズである。そのためｃ＋ｃ’が転送サイズ３００５を超過している場合に、Ｓ１０４０で読み出したデータ（またはＳ１０７０で圧縮したデータ）を転送すると、ＤＫＣ１０で確保したバッファ１３１上領域のサイズを超えたデータが転送されることになる。そのためＦＭＰＫ２００では、Ｓ１０８０の判定を行うことにより、ＤＫＣ１０に返送されるデータの量が、圧縮Ｒｅａｄコマンドのパラメータで指定された転送サイズ３００５を超過しないようにしている。また原則として、ＤＫＣ１０が圧縮Ｒｅａｄコマンドを発行する時、バッファ１３１上に大き目のサイズの領域（伸長時のリードデータサイズと同量の領域等）を確保する。

[0168] Ｓ１０９０では、プロセッサ２０３は、Ｓ１０４０で読み出されたデータ（Ｓ１０７０が実行されている場合には、Ｓ１０７０で圧縮されたデータ）をＤＫＣ１０のバッファ１３１に転送する。なお、図示を省略しているが、ＳＡＳ－ＣＴＬ２０６は転送の過程で、データに付されたＰＫ－ＤＩＦのチェックを行う。チェック結果が正常な場合には、ＳＡＳ－ＣＴＬ２０６はデータに付されたＰＫ－ＤＩＦを除去して、ＰＫ－ＤＩＦの除去されたデータをＤＫＣ１０に転送する。チェック結果が正常でない場合には、転送結果３０１１に「エラー」が含まれた応答情報をＤＫＣ１０に返却して、処理を終了する。

[0169] Ｓ１０９０が終了した後、プロセッサ２０３は変数ｃにｃ’を加算する（Ｓ１１００）。またＳ１１１０ではプロセッサ２０３は、変数ｕに対し、Ｓ１０４０で読み出したデータのサイズ（非圧縮時のサイズ）を加算する。なお、ここで説明している処理では、Ｓ１０３０、Ｓ１０４０の処理において、１論理ページ分のデータを読み出しているので、Ｓ１１１０では１論理ページサイズがｕに加算される。ただし別の実施形態として、Ｓ１０３０、Ｓ１０４０の処理で、複数論理ページ分のデータを読み出す、あるいは論理ペ

ージとは関係のない単位でデータを読み出すようにしてもよい。その場合、Ｓ１１１０では読み出したデータのサイズ（ただし非圧縮時サイズ）を変数ｕに加算する。

[0170] Ｓ１１２０ではプロセッサ２０３は、変数ｕがＲｅａｄサイズ３００３未満であるか判定する。ｕがＲｅａｄサイズ３００３未満の場合（Ｓ１１２０：ＹＥＳ）、プロセッサ２０３は再びＳ１０３０からの処理を行う。ｕがＲｅａｄサイズ３００３未満でない場合（Ｓ１１２０：ＮＯ）、プロセッサ２０３は応答情報を作成する（Ｓ１１４０）。ここで作成する応答情報は、転送結果３０１１に「成功」が、Ｒｅａｄサイズ３０１２に変数ｕの値が、バッファ使用サイズ３０１３には変数ｃの値が格納されたものである。その後プロセッサ２０３は応答情報をＤＫＣ１０に返却して、処理を終了する。

[0171] 次に、ＦＭＰＫ２００がＤＫＣ１０から、圧縮コピーＷｒｉｔｅコマンドを受信した時に行われる処理の流れを、図２２を用いて説明する。ＦＭＰＫ２００がＤＫＣ１０から圧縮コピーＷｒｉｔｅコマンドを受信すると、プロセッサ２０３はライトデータを受信する（圧縮コピーＷｒｉｔｅコマンドのパラメータである、転送元アドレス３１０４及び転送サイズ３１０５で特定される、バッファ１３１上領域からライトデータを取得する）。そして受信したライトデータを、メモリ２０２に格納する。さらにプロセッサ２０３は圧縮伸長回路２０４を用いて、ライトデータを伸長し、伸長されたデータもメモリ２０２に格納する（Ｓ１５２０）。先に述べたが、圧縮コピーＷｒｉｔｅコマンドは、ＤＫＣ１０が圧縮Ｒｅａｄコマンド３０００を用いて、ＦＭＰＫ２００から読み出したデータ（圧縮状態）をＦＭＰＫ２００に格納する際に用いられる。そのため、圧縮コピーＷｒｉｔｅコマンドによってライト対象となるデータは、圧縮された状態でＦＭＰＫ２００に到来する。プロセッサ２０３は圧縮された状態のデータをＦＭチップ２１０に格納する前に、ＤＩＦ（ＤＫＣ－ＤＩＦ）を用いたデータの検証（以下、ＤＩＦを用いた検証のことを「ＤＩＦのチェック」と呼ぶ。また、ＤＫＣ－ＤＩＦを用いたデータの検証は「ＤＫＣ－ＤＩＦのチェック」と呼び、ＰＫ－ＤＩＦを用い

たデータの検証は「ＰＫ－ＤＩＦのチェック」と呼ぶ）を行うため、一旦データの伸長を行う。

[0172] 続いてプロセッサ２０３は、伸長データに付加されているＰＫ－ＤＩＦとＤＫＣ－ＤＩＦを用いたデータの検証（チェック）を行う（Ｓ１５４０）。ＰＫ－ＤＩＦのチェックでは、伸長データから生成したＣＲＣと、ＰＫ－ＤＩＦに含まれているＣＲＣとを比較し、両者が一致しているか否かを判定する処理が行われる。

[0173] ＤＫＣ－ＤＩＦのチェックでは、主に以下のチェックを行う。

ａ） 伸長されたデータから生成されるＣＲＣと、伸長されたデータに付加されているＤＫＣ－ＤＩＦ内のＣＲＣが一致するかチェックする。

ｂ） 圧縮コピーＷｒｉｔｅコマンド（または圧縮パリティ演算Ｗｒｉｔｅコマンド）のパラメータに含まれているＷｒｉｔｅ開始オフセット３１０２と、ＤＫＣ－ＤＩＦに含まれているアドレス情報が一致するかチェックする。

ｃ） ＦＭＰＫ２００の状態管理テーブルＴ２０００に格納されている所属ＲＧ＃（Ｔ２００７）と、ＤＫＣ－ＤＩＦに含まれているＲＡＩＤグループ番号が同一かチェックする。

ｄ） 伸長されたデータに複数の５１２バイトデータが含まれている場合、各５１２バイトデータに付加されているＤＫＣ－ＤＩＦ内のシーケンス番号が、連続した番号であるかチェックする。

[0174] なお、ＦＭＰＫ２００が接続されるストレージシステム１の種類によって、ＤＫＣ－ＤＩＦのフォーマットが異なる場合がある。つまりストレージシステム１の種類によって、ＣＲＣやアドレス情報の格納されているＤＫＣ－ＤＩＦ内の位置が異なる場合がある。そのためプロセッサ２０３は、ＦＭＰＫ２００の状態管理テーブルＴ２０００に格納されている、接続ＤＫＣ種別Ｔ２００４の内容に基づいて、ＤＫＣ－ＤＩＦ内のどの位置にＣＲＣ、アドレス情報、シーケンス番号が格納されているかを特定する。

[0175] Ｓ１５４０の処理の結果、ＤＩＦのチェック結果が正常でなかった場合（

Ｓ１５５０：ＮＯ）、プロセッサ２０３は転送結果３０１１に「エラー」が含まれた応答情報を作成し（Ｓ１５９０）、作成した応答情報をＤＫＣ１０に返却して（Ｓ１５８０）、処理を終了する。ＤＩＦのチェック結果が正常であった場合（Ｓ１５５０：ＹＥＳ）、プロセッサ２０３は状態管理テーブルＴ２０００のデータ圧縮（Ｔ２００２）を参照することで、ＦＭＰＫ２００でデータ圧縮を行うことになっているか否か判断する（Ｓ１５６０）。

[0176] データ圧縮（Ｔ２００２）に「有」が格納されている場合（Ｓ１５６０：ＹＥＳ）、ＦＭチップ２１０には、メモリ２０２に格納されたライトデータ（圧縮データ）と伸長データのうち、圧縮データを書き込む処理を行う（Ｓ１５６１）。Ｓ１５６１では、プロセッサ２０３は圧縮データに付加されているＰＫ－ＤＩＦのチェックを行い、チェックが終了したらＰＫ－ＤＩＦを削除する。ただしＰＫ－ＤＩＦのチェック結果が正常でなかった場合には、Ｓ１５５０，Ｓ１５９０と同様に、ＤＫＣ１０に対して転送結果３０１１が「エラー」である応答情報を返却して処理を終了する。その後ＦＭチップ２１０に圧縮データを書き込む。なお、ＦＭチップ２１０にデータ（圧縮データ）を書き込む際、プロセッサ２０３は未使用の物理ページ（マッピングテーブルＴ２１００において、どの論理ページＴ２１０１にもマッピングされていない物理ページ）を選択して、この選択された物理ページにデータを書き込む。この処理は周知のフラッシュメモリで行われている処理と同じである。

[0177] ＦＭチップ２１０に圧縮データを書き込む過程で、ＦＭ－ＩＦ２０７は、圧縮データからＥＣＣを生成し、圧縮データにＥＣＣを付加し、ＥＣＣの付加された圧縮データをＦＭチップ２１０に書き込む。なおＥＣＣの生成と付与は、先に述べたとおり、５２０バイトのデータ毎に行われる。ＦＭチップ２１０へのデータ書き込みが終了した時点で、マッピングテーブルＴ２１００の内容を更新する。

[0178] マッピングテーブルＴ２１００の更新の概略は以下の通りである。Ｗｒｉｔｅ開始オフセット３１０２から論理ページ番号を算出する。マッピングテ

ーブルＴ２１００中のレコードのうち、算出された論理ページ番号と論理ページ番号（Ｔ２１０１）が等しいレコードが、更新されるべきレコードである。このレコードの物理ページ番号（Ｔ２１０２）、オフセット（Ｔ２１０４）、サイズ（Ｔ２１０３）に対して、圧縮データの書き込まれた物理ページ番号、物理ページ内のオフセット、及び圧縮データのサイズを書き込むことで、マッピングテーブルＴ２１００の更新が行われる。

[0179] その後、プロセッサ２０３は転送結果３０１１に「成功」が含まれた応答情報を作成し（Ｓ１５７０）、作成した応答情報をＤＫＣ１０に返却して（Ｓ１５８０）、処理を終了する。また、伸長されたデータ及び圧縮状態のデータの両方が、メモリ２０２に格納されているが、処理終了時に両方のデータを削除する。

[0180] 一方データ圧縮（Ｔ２００２）に「無」が格納されている場合（Ｓ１５６０：ＮＯ）、ＦＭＰＫ２００でデータ圧縮を行わないことを意味する。そのため、伸長データがＦＭチップ２１０に書き込まれる（Ｓ１５６２）。Ｓ１５６２では、プロセッサ２０３は伸長データに付加されているＰＫ－ＤＩＦを削除し、ＦＭチップ２１０に伸長データを書き込む。ＦＭチップ２１０に伸長データを書き込む過程では、Ｓ１５６１と同様、ＦＭ－ＩＦ２０７は、伸長データからＥＣＣを生成し、伸長データにＥＣＣを付加し、ＥＣＣの付加された伸長データをＦＭチップ２１０に書き込む。またマッピングテーブルＴ２１００の内容の更新を行う点は、Ｓ１５６１と同様である。

[0181] その後、プロセッサ２０３は転送結果３０１１に「成功」が含まれた応答情報を作成し（Ｓ１５７０）、作成した応答情報をＤＫＣ１０に返却して（Ｓ１５８０）、処理を終了する。また、メモリ２０２に格納されている、伸長されたデータ及び圧縮状態のデータの両方を、処理終了時に削除する。

[0182] なお、上の処理でＦＭチップ２１０に圧縮データを書き込む処理を行う場合（Ｓ１５６１）、圧縮データ（圧縮ページ）は概して物理ページのサイズより小さい。そのため１物理ページに複数の圧縮データを書き込む方が、記憶領域を効率的に使用できる。しかしＦＭチップ２１０の最小書き込み単位

は物理ページであるため、圧縮コピーＷｒｉｔｅコマンドを受信する毎に、圧縮データを物理ページに書き込む処理を行うと、その物理ページに未使用の領域が残っていても、書き込みができない。

[0183] そのため、別の実施形態として、圧縮コピーＷｒｉｔｅコマンドを受信するたびに、Ｓ１５６１においてＦＭチップ（物理ページ）に圧縮データを書き込むのではなく、メモリ２０２上で圧縮データにＥＣＣを付与した時点で、ＤＫＣ１０に応答情報を返却して（Ｓ１５７０，Ｓ１５８０）、処理を終了してもよい。そして圧縮コピーＷｒｉｔｅコマンドを複数回受信した結果、物理ページサイズに等しい（あるいはそれ以上の）量の圧縮データが蓄積された時点で、圧縮データを物理ページに格納する。このようにすると、ＦＭチップ２１０の記憶領域を効率的に使用できる。この場合、停電などによりＦＭＰＫ２００への電源供給が途絶すると、メモリ２０２に蓄積されたデータが消失する可能性がある。そのため、ＦＭＰＫ２００にバッテリを備える等して、電源供給の途絶時にもメモリ２０２の内容が揮発しないように構成することが望ましい。また、Ｓ１５６２の処理を行う場合（伸長データをＦＭチップ２１０に書き込む場合）でも、複数のデータを（たとえば１物理ブロック分まとめて）ＦＭチップ２１０に書き込むようにしてもよい。

[0184] さらに別の実施形態として、ＤＫＣ１０が圧縮コピーＷｒｉｔｅコマンドをＦＭＰＫ２００に発行する際に、できるだけ物理ページサイズと等しい量のライトデータを送信するようにしてもよい。ＤＫＣ１０は、各圧縮データのサイズを認識しているので（圧縮Ｒｅａｄコマンドの終了時に、応答情報としてＦＭＰＫ２００から圧縮データのサイズを受領するため）、Ｓ２５０において圧縮コピーＷｒｉｔｅコマンドを発行する際に、できるだけ物理ページサイズと等しい量の複数の圧縮データをまとめてＦＭＰＫ２００に送信するとよい。このようにすると、ＦＭＰＫ２００では圧縮コピーＷｒｉｔｅコマンドを受信するたびに、Ｓ１５６１においてＦＭチップ（物理ページ）に圧縮データを書き込んだとしても、物理ページに効率的に圧縮データを格納できる。

[0185] 次に、ＦＭＰＫ２００がＤＫＣ１０から、圧縮パリティ演算Ｗｒｉｔｅコマンドを受信した時に行われる処理の流れを、図２３を用いて説明する。以下では、圧縮パリティ演算Ｗｒｉｔｅコマンドとともに ＦＭＰＫ２００に送信されるデータ（正常なＦＭＰＫ２００から読み出した圧縮状態のデータ）のことを「ライト対象データ」と呼ぶ。また、圧縮パリティ演算Ｗｒｉｔｅコマンドのパラメータである、「Ｗｒｉｔｅ開始オフセット３２０２及びＷｒｉｔｅサイズ３２０３」で特定される領域のことを「ライト対象領域」と呼ぶ。

[0186] Ｓ２０２０～Ｓ２０５０は、図２２のＳ１５２０～Ｓ１５５０と同じである。またＳ２１５０は、Ｓ１５９０と同じである。ただし図２２のＳ１５２０の実行後には、伸長データと圧縮状態のデータの両方がメモリ２０２に格納されていたが、Ｓ２０２０では、伸長データのみをメモリ２０２に格納し、ＤＫＣ１０から受信した圧縮状態のデータは、メモリ２０２から削除してもよい。

[0187] Ｓ２０５０の判定の後、ＤＩＦのチェック結果が正常であった場合（Ｓ２０５０：ＹＥＳ）、Ｓ２０６０以降の処理が行われる。Ｓ２０６０ではプロセッサ２０３は、今回のライト対象領域に対して、以前にも圧縮パリティ演算Ｗｒｉｔｅコマンドを受信しているか判定する。この判定方法は後述する。今回のライト対象領域に対する圧縮パリティ演算Ｗｒｉｔｅコマンドを初めて受信した場合（Ｓ２０６０：ＮＯ）、ライト対象データを格納する領域をメモリ２０２上に確保し、Ｓ２０２０で伸長されたデータを、確保されたメモリ２０２上領域に格納する（Ｓ２０７０）。その後プロセッサ２０３はＳ２０９０以降の処理を行う。

[0188] プロセッサ２０３は、ライト対象領域と、ライト対象データ（またはライト対象データから生成されたパリティ）が格納されているメモリ２０２上の領域との関係を、図２４に示されているようなステージング情報管理テーブルＴ２５００で管理する。ステージング情報管理テーブルＴ２５００は、論理ページ番号Ｔ２５０１で特定される論理ページに対するライト対象データ

（あるいはライト対象データを用いて生成されるパリティ）が、アドレスＴ２５０２で特定されるメモリ２０２上領域に格納された状態にあることを管理するテーブルである。また、回数Ｔ２５０３には、論理ページ番号Ｔ２５０１で特定される論理ページをライト対象領域として指定した圧縮パリティ演算Ｗｒｉｔｅコマンドを受信した回数が記録される。

[0189] 初期状態では、アドレスＴ２５０２には無効値（ＮＵＬＬ）が、回数Ｔ２５０３には０が格納されている。そしてライト対象データを格納する領域がメモリ２０２上に確保されると、プロセッサ２０３は確保されたメモリ２０２上のアドレスを、アドレスＴ２５０２に格納する。そして回数Ｔ２５０３に１を加算する。

[0190] そのため、過去に論理ページ番号Ｔ２５０１で特定される論理ページをライト対象領域とする圧縮パリティ演算Ｗｒｉｔｅコマンドを受信していない場合、その論理ページに対応するアドレスＴ２５０２には無効値（ＮＵＬＬ）が、回数Ｔ２５０３には０が格納された状態にある。Ｓ２０６０では、プロセッサ２０３は回数Ｔ２５０３の値を参照することにより（アドレスＴ２５０２の値を参照してもよい）、今回のライト対象領域に対して、過去に圧縮パリティ演算Ｗｒｉｔｅコマンドを受信しているか判定する。回数Ｔ２５０３が０以外の値であれば、過去に圧縮パリティ演算Ｗｒｉｔｅコマンドを受信していると判定する。

[0191] 過去に圧縮パリティ演算Ｗｒｉｔｅコマンドを受信している場合（Ｓ２０６０：ＹＥＳ）、メモリ２０２に格納されているデータと、Ｓ２０２０で伸長されたデータとから、パリティが計算される（Ｓ２０８０）。Ｓ２０８０ではプロセッサ２０３は、メモリ２０２上のアドレスＴ２５０２で特定される領域に格納されているデータと、Ｓ２０２０で生成された伸長データとを、パリティ演算回路２０５によって演算する。そしてこの演算結果（以下ではこれを「中間パリティ」と呼ぶ）を、アドレスＴ２５０２で特定されるメモリ２０２上領域に格納し、回数Ｔ２５０３に１を加算する（Ｓ２０８０）。なお、上の説明で分かる通り、パリティ生成は伸長されたデータに対して

行われる。圧縮されたデータからパリティを計算しても、データの再生成はできないからである。

[0192] Ｓ２０８０で、パリティ演算回路２０５によって行われる演算は、たとえばＦＭＰＫ２００が所属するＲＡＩＤグループのＲＡＩＤレベルがＲＡＩＤ５の場合、排他的論理和（ＸＯＲ）である。一方ＦＭＰＫ２００が所属するＲＡＩＤグループのＲＡＩＤレベルがＲＡＩＤ６の場合、排他的論理和が計算されることもあれば、リードソロモン符号（ガロア体の多項式演算）の計算を行うこともある。いずれの計算を行うべきかは、ＲＡＩＤグループ内のどのＦＭＰＫ２００のデータを再生成する必要があるか、に依存する。

[0193] またリードソロモン符号の計算を行う場合、たとえば特許文献１に記載されているように、データに所定の係数を乗じる積演算を行う必要がある。そしてこの乗じるべき所定の係数の値も、ＲＡＩＤグループ内の位置に依存する。そのため、Ｓ２０８０ではプロセッサ２０３は、状態管理テーブルＴ２０００に格納されている、所属ＲＡＩＤグループ構成Ｔ２００５、ＲＡＩＤグループ内位置Ｔ２００６、及び圧縮パリティ演算Ｗｒｉｔｅコマンドのパラメータに含まれているＲＡＩＤグループ内位置３２０６をもとに、データ再生成方法（排他的論理和、またはリードソロモン符号）、及び積演算で用いられる係数を決定し、それを用いたパリティ生成を行う。

[0194] Ｓ２０９０で、プロセッサ２０３はデータの復元が完了したか判定する。たとえば圧縮パリティ演算Ｗｒｉｔｅコマンドを受領したＦＭＰＫ２００の所属するＲＡＩＤグループの構成が、ＲＡＩＤ５（３Ｄ＋１Ｐ）だった場合（ストライプラインが４ストライプブロックで構成される場合）、３つのＦＭＰＫ２００から読み出されたデータのＸＯＲを計算すると、データを再生成できる。つまり圧縮パリティ演算Ｗｒｉｔｅコマンドを３回受信し、ＸＯＲ演算を３回行った場合には、データの復元が完了している（アドレスＴ２５０２で特定されるメモリ２０２上領域には、復元されたデータが格納されている）、と判定可能である。

[0195] そのため本実施例のＦＭＰＫ２００では、プロセッサ２０３はＳ２０９０

において、回数Ｔ２５０３に記録されている、ライト対象アドレスに対する圧縮パリティ演算Ｗｒｉｔｅコマンドを受信している回数を参照することによって、データの復元が完了したか否かを判定する。たとえば圧縮パリティ演算Ｗｒｉｔｅコマンドを受領したＦＭＰＫ２００の所属するＲＡＩＤグループの構成がＲＡＩＤ５（ｎＤ＋１Ｐ）の場合（ｎは１以上の整数値）には、ｎ回圧縮パリティ演算Ｗｒｉｔｅコマンドを受信したか判定する。またＦＭＰＫ２００の所属するＲＡＩＤグループが、ＲＡＩＤ６（ｎＤ＋２Ｐ）のように、複数のパリティストライプを格納するＲＡＩＤ構成であった場合も、ｎ回圧縮パリティ演算Ｗｒｉｔｅコマンドを受信したか否かを判定することで、データの復元が完了しているか否かを判定可能である。

[0196] データの復元が完了していない場合（Ｓ２０９０：ＮＯ）、つまりまだ圧縮パリティ演算Ｗｒｉｔｅコマンドを所定回数分受信していない場合には、プロセッサ２０３は、転送結果３０１１に「成功」が格納された応答情報を作成し（Ｓ２１７０）、作成された応答情報をＤＫＣ１０に返却し（Ｓ２１３０）、処理を終了する。なおこの場合、Ｓ２０７０またはＳ２０８０でメモリ２０２に格納したデータが、停電などの要因で消失することを防ぐために、ＦＭチップ２１０に格納するようにしても良い。

[0197] データの復元が完了した場合（Ｓ２０９０：ＹＥＳ）、つまり圧縮パリティ演算Ｗｒｉｔｅコマンドを所定回数分受信した場合には、プロセッサ２０３は、メモリ２０２に格納されているデータ（復元データと呼ぶ）に、ＤＫＣ－ＤＩＦを付与し、さらにＰＫ－ＤＩＦの付与を行う（Ｓ２１００）。

[0198] Ｓ２１１０では、プロセッサ２０３は状態管理テーブルＴ２０００のデータ圧縮（Ｔ２００２）を参照することで、ＦＭＰＫ２００でデータ圧縮を行うことになっているか否か判断する。これはＳ１５６０と同様の処理である。

[0199] データ圧縮（Ｔ２００２）に「有」が格納されている場合（Ｓ２１１０：ＹＥＳ）、ＦＭＰＫ２００でデータ圧縮を行うことになっているので、Ｓ２１１０の処理によってＤＩＦが付与されたデータを圧縮し、その後ＦＭチッ

プ２１０に圧縮されたデータを書き込む（Ｓ２１１１）。なお、ＦＭチップ２１０へのデータ書き込み後、マッピングテーブルＴ２１００の更新も行う。

[0200] その後プロセッサ２０３は、転送結果３０１１に「成功」が格納された応答情報を作成し（Ｓ２１２０）、作成された応答情報をＤＫＣ１０に返却し（Ｓ２１３０）、処理を終了する。

[0201] データ圧縮（Ｔ２００２）に「無」が格納されている場合（Ｓ２１１０：ＮＯ）、ＦＭＰＫ２００でデータ圧縮を行わない。そのため、Ｓ２１１０においてＤＩＦの付与されたデータをそのままＦＭチップ２１０に書き込む（Ｓ２１１２）。なおＦＭチップ２１０へのデータ書き込み後、マッピングテーブルＴ２１００の更新も行う点は、Ｓ２１１１の処理と同様である。その後プロセッサ２０３は、転送結果３０１１に「成功」が格納された応答情報を作成し（Ｓ２１２０）、作成された応答情報をＤＫＣ１０に返却し（Ｓ２１３０）、処理を終了する。

[0202] 実施例１に係るストレージシステム１では、データ復旧時に、復旧元となる記憶デバイス（障害の発生した記憶デバイス、あるいは障害が発生した記憶デバイスと同一ＲＡＩＤグループに属する記憶デバイス）から圧縮状態でデータを読み出して、復旧先の記憶デバイス（スペアデバイス）にデータを送信するので、復旧用のデータの送信時間を短縮することができる。また、復旧先の記憶デバイスでコレクションを行うため、ストレージコントローラ側でデータコレクションを行う場合に比べて、ストレージコントローラの負荷を低減することができる。さらに、復旧先の記憶デバイスでデータを伸長してＤＩＦのチェックを行うため、データ転送に係るエラーの検出も可能である。

## 実施例 2

[0203] 続いて実施例２に係るストレージシステムについて説明する。実施例２に係るストレージシステム１の構成は、実施例１に係るストレージシステム１と同じである。

[0204] 実施例１と実施例２に係るストレージシステムの違いは、データ復旧処理時、特にコレクション時にＦＭＰＫ２００に対して発行されるコマンドに違いがある。実施例１に係るストレージシステム１では、コレクション時に復旧先デバイスに対して圧縮パリティ演算Ｗｒｉｔｅコマンドという、１種類のコマンドのみを発行していた。一方実施例２に係るストレージシステム１では、コレクション時に復旧先デバイスに対して、圧縮中間パリティ演算コマンド、そしてパリティコミットコマンドという、２種類のコマンドを発行する。以下、実施例２に係るストレージシステム１で行われる、コレクション時の処理の流れについて説明していく。

[0205] まず、圧縮中間パリティ演算コマンド、そしてパリティコミットコマンドという、２種類のコマンドについて説明する。圧縮中間パリティ演算コマンドは、ＤＫＣ１０が実施例１で説明した圧縮Ｒｅａｄコマンド３０００を用いて、ＦＭＰＫ２００から読み出したデータ（圧縮状態）を、ＦＭＰＫ２００に送信するとともに、ＦＭＰＫ２００に対し、送信したデータとＦＭＰＫ２００に格納されているデータとからパリティを算出することを指示するためのコマンドであるという点で、圧縮パリティ演算Ｗｒｉｔｅコマンドと類似している。

[0206] ただし圧縮パリティ演算Ｗｒｉｔｅコマンドを受信したＦＭＰＫ２００は、最終的には生成されたデータをＦＭチップ２１０に格納する（かつマッピングテーブルＴ２１００の更新を行う）が、圧縮中間パリティ演算コマンドがＦＭＰＫ２００に発行された場合、ＦＭＰＫ２００は生成されたデータをＦＭチップ２１０に格納することは行わない。ＦＭＰＫ２００はパリティコミットコマンドを受信した時にはじめて、生成されたデータをＦＭチップ２１０に格納する。

[0207] 圧縮中間パリティ演算コマンド、そしてパリティコミットコマンドのコマンドフォーマットを、図２５、図２６を用いて説明する。圧縮中間パリティ演算コマンドに含まれるパラメータは、圧縮パリティ演算Ｗｒｉｔｅコマンドに含まれるパラメータと同じで、オペコード（Ｏｐｃｏｄｅ）３２０１’

、Ｗｒｉｔｅ開始オフセット３２０２’、Ｗｒｉｔｅサイズ３２０３’、バッファアドレス３２０４’、転送サイズ３２０５’、ＲＡＩＤグループ内位置３２０６’、のパラメータが含まれる。オペコード３２０１’の値を除き（圧縮中間パリティ演算コマンドのオペコード３２０１’の値は当然、圧縮パリティ演算Ｗｒｉｔｅコマンド等の他のコマンドのオペコードとは異なる値である）、各パラメータの内容は圧縮パリティ演算Ｗｒｉｔｅコマンドに含まれるパラメータと同じであるため、説明は省略する。

[0208] 図２６はパリティコミットコマンドのコマンドフォーマットを示している。パリティコミットコマンドにはオペコード３３０１、Ｗｒｉｔｅ開始オフセット３３０２、Ｗｒｉｔｅサイズ３３０３が含まれる。パリティコミットコマンドを受信したＦＭＰＫ２００で行われる処理の流れは後述する。また圧縮中間パリティ演算コマンドとパリティコミットコマンドに対する応答情報のフォーマットは、圧縮パリティ演算Ｗｒｉｔｅコマンド等の応答情報と同じものであるため、ここでの説明は省略する。

[0209] 続いて、ＦＭＰＫ２００がＤＫＣ１０から、圧縮中間パリティ演算コマンドを受信した時に行われる処理の流れを、図２７を用いて説明する。Ｓ２０２０～Ｓ２０８０までの処理は、図２３のＳ２０２０～Ｓ２０８０と同じである。つまり中間パリティを生成して、メモリ２０２に格納する処理が行われる。

[0210] Ｓ２０８０またはＳ２０７０の処理が終了した後、プロセッサ２０３は、転送結果３０１１に「成功」が格納された応答情報を作成し（Ｓ２１７０）、作成された応答情報をＤＫＣ１０に返却し（Ｓ２１３０）、処理を終了する。つまり、圧縮パリティ演算Ｗｒｉｔｅコマンドの処理のうち、図２３のＳ２０９０、Ｓ２１００、Ｓ２１１０、Ｓ２１１１，Ｓ２１１２，Ｓ２１２０が行われない点が、圧縮中間パリティ演算コマンドと圧縮パリティ演算Ｗｒｉｔｅコマンドの違いである。図２３のＳ２０９０、Ｓ２１００、Ｓ２１１０、Ｓ２１１１，Ｓ２１１２，Ｓ２１２０に相当する処理は、ＦＭＰＫ２００がパリティコミットコマンドを受信した時に行われる。

[0211] 続いて、ＦＭＰＫ２００がＤＫＣ１０から、パリティコミットコマンドを受信した時に行われる処理の流れを、図２８を用いて説明する。まずプロセッサ２０３は、コマンドのパラメータに含まれる、Ｗｒｉｔｅ開始オフセット３３０２、Ｗｒｉｔｅサイズ３３０３から、受信したパリティコミットによって処理対象となる領域（以下、この領域をコミット対象領域と呼ぶ）の論理ページ番号を算出する（Ｓ２０１０）。

[0212] 続いてプロセッサ２０３はステージング情報管理テーブルＴ２５００を参照し、コミット対象領域に対応するデータ（中間パリティ）が、メモリ２０２上に格納されているか判定する（Ｓ２０６０’）。これはステージング情報管理テーブルＴ２５００内の論理ページ番号（Ｔ２５０１）が、Ｓ２０１０で特定された論理ページ番号と等しいレコードについて、アドレス（Ｔ２５０２）に有効値（ＮＵＬＬ以外の値）が格納されているかを判定すればよい。

[0213] アドレス（Ｔ２５０２）にＮＵＬＬが格納されている場合（Ｓ２０６０’：ＮＯ）、これまでに、コミット対象領域に対する圧縮中間パリティ演算コマンドを受信していない（中間パリティの生成が行われていない）ことを意味する。そのためプロセッサ２０３は、転送結果３０１１に「エラー」が含まれた応答情報を作成し（Ｓ２１５０）、作成した応答情報をＤＫＣ１０に返却して（Ｓ２１３０）、処理を終了する。

[0214] アドレス（Ｔ２５０２）に有効値が格納されている場合（Ｓ２０６０’：ＹＥＳ）、メモリ２０２に格納されているデータ（コミット対象領域に対応するデータ（中間パリティ））にＰＫ－ＤＩＦとＤＫＣ－ＤＩＦを付与する（Ｓ２１００’）。これは図２３のＳ２１００と同じ処理である。以下、Ｓ２１１０’、Ｓ２１１１’、Ｓ２１１２’、Ｓ２１２０’、Ｓ２１３０’は、図２３のＳ２１１０～Ｓ２１３０と同じ処理である。つまりＰＫ－ＤＩＦとＤＫＣ－ＤＩＦの付与されたデータを、（必要があれば圧縮した後）ＦＭチップ２１０に書き込み、ＤＫＣ１０に処理が成功した旨の応答情報を返却する。

[0215] 次に、実施例２に係るストレージシステム１で行われるコレクションコピーの処理の流れを、図２９を用いて説明する。この処理は、実施例１におけるコレクションコピー（図２０）とほとんど同じであるので、相違点のみを説明する。

まず実施例１におけるコレクションコピー（図２０）では、ＤＫＣ１０のプロセッサ１１は、圧縮パリティ演算Ｗｒｉｔｅコマンドを復旧先デバイスに発行していた（Ｓ４８０）。実施例２におけるコレクションコピー処理では、圧縮中間パリティ演算コマンドを復旧先デバイスに発行する（Ｓ４８０’）。

[0216] また実施例２におけるコレクションコピー処理では、ＲＡＩＤグループを構成する全ての正常なＦＭＰＫ２００についてＳ４１０～Ｓ５００の処理を行った場合に、復旧先デバイスにパリティコミットコマンドを発行し（Ｓ６００）、その後処理を終了する。それ以外の点は、実施例１におけるコレクションコピー処理と同じである。また、実施例１で説明した、図１７、図１９、図２１、図２２の処理は、実施例２に係るストレージシステム１でも同じである。

[0217] 実施例１におけるコレクションコピー処理では、復旧先デバイスは圧縮パリティ演算Ｗｒｉｔｅコマンドを受信した回数を、領域（論理ページ番号等）毎に記憶しておき、圧縮パリティ演算Ｗｒｉｔｅコマンドを受信した回数が所定回数（ＲＡＩＤグループを構成する正常なＦＭＰＫ２００の数）に達したことを、復旧先デバイス自身が判断して、ＦＭＰＫ２００に復元データを格納していた。実施例２におけるコレクションコピー処理では、ＤＫＣ１０が復旧先デバイスに、ＦＭＰＫ２００に復元データを格納する契機を通知するため、ＦＭＰＫ２００では圧縮パリティ演算Ｗｒｉｔｅコマンド（圧縮中間パリティ演算コマンド）を受信した回数を記憶しておく必要がない。そのため、実施例２に係るＦＭＰＫ２００では、ステージング情報管理テーブルＴ２５００で、必ずしも回数（Ｔ２５０３）を管理しておく必要はない。

## 実施例 3

[0218] 続いて実施例３に係るストレージシステムについて説明する。実施例３に係るストレージシステム１の構成は、実施例１に係るストレージシステム１と同じである。

[0219] 実施例１に係るストレージシステム１では、同一ストライプラインに属するストライプブロックはすべて、各記憶デバイス２００（２００’）上の同じ位置（アドレス）に格納されることが前提であった。実施例３に係るストレージシステム１では、同一ストライプラインに属するストライプブロックがそれぞれ、各記憶デバイス２００（２００’）上の異なる位置（アドレス）に格納される構成を許すものである。

[0220] 実施例３に係るストレージシステム１でサポートされる、ストライプラインの構成例について、図３０を用いて説明する。同一ストライプラインに属するストライプブロックが、すべて記憶デバイス２００（２００’）上の同じ位置（アドレス）に格納されるという構成は、記憶デバイス２００（２００’）の障害時におけるデータ復旧を可能にするための必須条件ではない。データ復旧の観点からすれば、同一ストライプラインに属する各ストライプブロックがすべて、異なる記憶デバイス２００（２００’）に格納されるように配置されていればよい。実施例３に係るストレージシステム１では、この規則が守られるように、各ストライプライン内の各ストライプブロックを記憶デバイス２００（２００’）に配置する。それ以外の制約はない。

[0221] 図３０において、ストライプライン３００－１は、ＦＭＰＫ２００－１，２００－２，２００－３，２００－４に跨って定義されている。そしてストライプライン３００－１内の各ストライプブロックのＦＭＰＫ２００上の位置は、いずれも異なっている。またストライプライン３００－２は、ＦＭＰＫ２００－３，２００－４，２００－５，２００－６、２００－７に跨って定義されている。そしてストライプライン３００－２内の各ストライプブロックのＦＭＰＫ２００上の位置は、同じものもあれば（たとえばＦＭＰＫ２００－３上のストライプブロックとＦＭＰＫ２００－７上のストライプブロック）、異なっているものもある。

[0222] 実施例３に係るストレージシステム１では、同一ストライプラインに属する各ストライプブロックがすべて、異なる記憶デバイス２００（２００’）に格納されるように配置されるという規則さえ守られれば、各ストライプラインは任意の記憶デバイス２００（２００’）に存在してもよい。

[0223] この場合、１つのＦＭＰＫ２００，たとえばＦＭＰＫ２００－３に障害が発生してアクセスできなくなったとしても（その他のＦＭＰＫは正常に動作しているとする）、ＦＭＰＫ２００－１，２００－２，２００－４がアクセス可能であるため、ストライプライン３００－１については、ＦＭＰＫ２００－１，２００－２，２００－４に存在するストライプブロックからコレクションが可能である。同様に、ストライプライン３００－２についても、ＦＭＰＫ２００－４，２００－５，２００－６、２００－７に存在するストライプブロックからコレクションが可能である。

[0224] 以下で、実施例３に係るＤＫＣ１０が実行するデータ復旧処理について説明する。この処理は、実施例１（または実施例２）に係るストレージシステム１で行われるものと多くの部分で共通するので、図１７～２４を用いて処理の流れを説明する。また、実施例１におけるデータ復旧処理と同じ処理が行われる箇所については説明を省略し、実施例１と相違する点について中心に説明する。なお、実施例１、２では、論理ページ単位にコレクションまたはコピー復旧を行っていたが、実施例３では、ストライプブロック単位でコレクションまたはコピー復旧を行う例について説明する。ただし実施例３に係るストレージシステム１でも、論理ページ単位にコレクションまたはコピー復旧を行うことは可能である。

[0225] 先にも述べたが、実施例１または２に係るストレージシステム１では、同一ストライプラインに属するストライプブロックが、すべて記憶デバイス２００（２００’）上の同じ位置（アドレス）に格納される。そのためデータ復旧処理もその前提で行われていた。実施例３に係るストレージシステムでは、同一ストライプラインに属するストライプブロックが、記憶デバイス２００（２００’）上の異なる位置に存在し得るため、復旧元デバイスのデー

タをコレクションにより再生成する際、コレクションに必要なデータが、どの記憶デバイスのどのアドレスに格納されているか、特定する必要がある。

[0226] 実施例１または２に係るストレージシステム１では、複数の記憶デバイス２００（２００’）を１つのＲＡＩＤグループという概念でまとめて管理している。そして、異なるＲＡＩＤグループ内に存在する記憶デバイスに跨ってストライプラインが定義されることはなかった。実施例３に係るストレージシステム１では、ストライプラインに属するストライプブロックは、上で述べた規則に従っている限り、任意の記憶デバイスに存在して良いため、ＲＡＩＤグループという概念はない。つまり、実施例１または２における、ＲＧ管理テーブルＴ１１００は、実施例３に係るストレージシステム１には存在しない。

[0227] 代わりにＤＫＣ１０は、ストライプラインごとに、ストライプラインに属するストライプブロックの存在する記憶デバイス２００（２００’）及び記憶デバイス内アドレス、及びストライプラインのＲＡＩＤ構成についての情報を管理するテーブルを有する（以下、これをストライプライン管理テーブルと呼ぶ。またストライプライン管理テーブルの内容はＲＧ管理テーブルＴ１１００と類似するため、詳細説明は省略する）。また実施例１に係るストレージシステム１では、デバイス管理テーブルＴ１０００に、記憶デバイスの所属するＲＡＩＤグループの情報（所属ＲＧ＃（Ｔ１００３））が格納されていたが、実施例３に係るストレージシステム１では所属ＲＧ＃（Ｔ１００３）に代えて、デバイス管理テーブルＴ１０００に、記憶デバイスに格納されているストライプブロックが属するストライプラインについての情報（ストライプラインの識別番号等）のリストを格納して管理する。

[0228] これにより、ストライプラインを構成する１（または２）の記憶デバイス２００（２００’）にアクセスできなくなった場合、アクセスできなくなった記憶デバイス２００（２００’）に属するストライプライン（１または複数）を特定し、特定されたストライプラインを構成するストライプブロックが存在する、複数の記憶デバイス２００（２００’）及び記憶デバイス２０

０（２００’）内位置を特定する。実施例３に係るデータ復旧処理では、この処理が行われる以外は、実施例１に係るデータ復旧処理と大きな違いはない。

[0229] 実施例１で説明した図１７の処理（データ復旧処理の全体の流れ）は、実施例３においても同様の処理が行われるが、上でも述べたとおり、実施例１に係るストレージシステム１と実施例３に係るストレージシステム１では、管理している管理テーブルに違いがあるため、管理情報の更新（Ｓ１００）の処理が異なる。また実施例３では、ストライプブロック単位でコレクションまたはコピー復旧を行うため、コピー管理テーブルＴ１５００の、コピー方法ビットマップＴ１５０６の１ビットは、１ストライプブロックに相当する点、そしてＳ８０の処理で、復旧完了済オフセット（Ｔ１５０７）に復旧されたデータサイズを加算する際、（ストライプサイズ÷論理ページサイズ）を加算する点が、実施例１で説明したものとは異なる。それ以外の点は、実施例１で説明したものと同様の処理が行われる。

[0230] また、Ｓ７２で行われる処理（コレクションコピー）も、実施例１で説明したものとは若干の違いがある。以下では、実施例３に係るストレージシステム１で行われるコレクションコピーの処理の流れについて、図３１を用いて説明する。

[0231] まずプロセッサ１１は、復旧元デバイス内の、復旧完了済みオフセットＴ１５０７（論理ページ番号）で特定される領域が属するストライプライン（以下、復旧対象ストライプラインと呼ぶ）を特定し、及び復旧対象ストライプラインに属する全ストライプブロックが存在する記憶デバイス（ＦＭＰＫ２００）のうち、正常な記憶デバイスのデバイス＃、及び当該正常な記憶デバイス内の、復旧対象ストライプラインに属するストライプブロックが格納されているアドレスをすべて特定する（Ｓ４００’）。これはストライプライン管理テーブル及びデバイス管理テーブルＴ１０００を参照することで特定可能である。

[0232] 続いてプロセッサ１１は、Ｓ４００’で特定された「デバイス＃、アドレ

ス」のセットのうち１つを選択する（Ｓ４０５’）。その後Ｓ４１０（変数ｒ、ｗの初期化）を実行する。

[0233] 続いてプロセッサ１１はＳ４２０’で、バッファ１３１にリードデータを格納するための領域として、１ストライプブロック分の領域を確保するとともに、Ｓ４０５’で選択されたデバイス＃のＦＭＰＫ２００に対し、同じくＳ４０５’で選択されたアドレスをパラメータ（Ｒｅａｄ開始オフセット３００２）として指定した圧縮Ｒｅａｄコマンドを発行する。またこの時、圧縮ＲｅａｄコマンドのＲｅａｄサイズ３００３には、１ストライプブロックのサイズが指定される。続いて行われるＳ４３０～Ｓ４６０の処理は、実施例１で説明したものと同じである。

[0234] Ｓ４８０’でプロセッサは、圧縮パリティ演算Ｗｒｉｔｅコマンドを復旧先デバイスに発行する。実施例３に係るストレージシステム１でサポートされる圧縮パリティ演算Ｗｒｉｔｅコマンドに含まれるパラメータについて、図３２を用いて説明する。パラメータのうち、オペコード３２０１～ＲＡＩＤグループ内位置３２０６までのパラメータは、実施例１で説明したものと同じである。

[0235] Ｓ４８０’で発行される、圧縮パリティ演算Ｗｒｉｔｅコマンドで指定されるパラメータであるが、Ｗｒｉｔｅ開始オフセット３２０２には、復旧完了済みオフセットＴ１５０７（論理ページ番号）をＬＢＡに変換した値が設定される。またＷｒｉｔｅサイズ３２０３には１ストライプブロック分のサイズが指定される。そしてバッファアドレス３２０４には、Ｓ４２０’で確保されたバッファ１３１上の領域の情報が指定される。また、転送サイズ３２０５には、Ｓ４２０’、Ｓ４３０で読み出されたデータのサイズ（これはＳ４３０で受信した、圧縮Ｒｅａｄコマンドの応答情報（バッファ使用サイズ３０１３）に含まれている）が指定される。

[0236] 図３２に示されているとおり、実施例３に係るストレージシステム１でサポートされる圧縮パリティ演算Ｗｒｉｔｅコマンドには、リード元データアドレス３２０７が追加されている。実施例１で説明したとおり、圧縮パリテ

ィ演算Ｗｒｉｔｅコマンドは、ＤＫＣ１０が圧縮Ｒｅａｄコマンド３０００を用いて読み出したデータ（圧縮状態）を、復旧先ＦＭＰＫ２００に送信するとともに、パリティ算出を復旧先ＦＭＰＫ２００に指示するためのコマンドである。リード元データアドレス３２０７には、この読み出したデータ（圧縮状態）の格納されていたアドレス（正常なストライプブロックの格納されていたアドレス）が指定される。

[0237] Ｓ４８０’の後に行われる、Ｓ４９０～Ｓ５２０、Ｓ５４０の処理は、実施例１で説明したものと同じであるので、ここでの説明は省略する。Ｓ５００の判定の後、プロセッサ１１は、Ｓ４００’で特定されたすべてのストライプブロックに対して、Ｓ４０５’～Ｓ５００の処理を実施したか判定する（Ｓ５５０’）。まだＳ４０５’～Ｓ５００の処理が実施されていないストライプブロックが残っている場合（Ｓ５５０’：ＮＯ）、プロセッサ１１は再びＳ４０５’からの処理を繰り返す。すべてのストライプブロックに対して、Ｓ４０５’～Ｓ５００の処理を実施している場合（Ｓ５５０’：ＹＥＳ）は、処理を終了する、

[0238] 次に、実施例３に係るＦＭＰＫ２００が、ＤＫＣ１０から圧縮パリティ演算Ｗｒｉｔｅコマンドを受信した時に行われる処理の流れを説明する。この処理は、実施例１において説明したものとほとんど同じであるため、図２３を用いて相違点についてのみ説明する。

[0239] 実施例３に係るＦＭＰＫ２００で行われる、圧縮パリティ演算Ｗｒｉｔｅコマンドに係る処理は、ＤＫＣ－ＤＩＦのチェック（Ｓ２０４０）に関する処理のみが異なり、後は同じである。

[0240] 実施例３に係るＦＭＰＫ２００でのＤＫＣ－ＤＩＦのチェックは、主に以下のチェックを行う。

ａ） 伸長されたデータから生成されるＣＲＣと、伸長されたデータに付加されているＤＫＣ－ＤＩＦ内のＣＲＣが一致するかチェックする。

ｂ’） 圧縮パリティ演算Ｗｒｉｔｅコマンドのパラメータに含まれている、リード元データアドレス３２０７と、ＤＫＣ－ＤＩＦに含まれているア

ドレス情報が一致するかチェックする。

ｄ）　伸長されたデータに複数の５１２バイトデータが含まれている場合、各５１２バイトデータに付加されているＤＫＣ－ＤＩＦ内のシーケンス番号が、連続しているかチェックする。

[0241] 上のａ）、ｄ）は、実施例１において行われるＤＫＣ－ＤＩＦのチェックと同じである。一方ｂ’）に関して、実施例１に係るＦＭＰＫ２００では、圧縮コピーＷｒｉｔｅコマンド（または圧縮パリティ演算Ｗｒｉｔｅコマンド）のパラメータに含まれているＷｒｉｔｅ開始オフセット３１０２と、ＤＫＣ－ＤＩＦに含まれているアドレス情報が一致するかチェックしていた。ただし実施例３に係るストレージシステム１では、同一ストライプラインに属するストライプブロックが、記憶デバイス２００（２００’）上の異なる位置に存在し得るため、Ｗｒｉｔｅ開始オフセット３１０２と、ＤＫＣ－ＤＩＦに含まれているアドレス情報を比較することに意味がない。そのため、実施例３に係るストレージシステム１で用いられる圧縮パリティ演算Ｗｒｉｔｅコマンドのパラメータには、リード元データアドレス３２０７が含められている。そしてＳ２０４０で行われるＤＫＣ－ＤＩＦのチェックの際、ＤＫＣ－ＤＩＦに含まれているアドレス情報とリード元データアドレス３２０７を比較するようにしている。その他の点は、実施例１で説明した処理と同じである。

[0242] 以上が実施例３に係るストレージシステム１で行われるデータ復旧処理の説明である。実施例３に係るストレージシステム１では、同一ストライプラインに属するストライプブロックがそれぞれ、記憶デバイス２００（２００’）上の異なる位置（アドレス）に格納される構成を許しているため、データ配置の自由度が高まる。

[0243] 特に記憶デバイスが圧縮機能を備える場合、格納されるデータの内容によって、記憶デバイスに格納可能なデータ量が異なる。ストライプブロックの格納される記憶デバイス及び記憶デバイス内の位置（アドレス）が固定されていると、圧縮によって記憶デバイスに格納可能な容量が増加したとしても

、増加した記憶領域を有効に活用できないことがある。

[0244] 実施例３に係るストレージシステムの場合、同一ストライプラインに属する各ストライプブロックがすべて、異なる記憶デバイス２００（２００’）に格納されるという規則が守られる限り、任意の記憶デバイスにストライプブロックを格納することができるため、圧縮によって特定の記憶デバイスの格納可能容量が増加した場合、その記憶デバイスに多くのストライプブロックを格納する等の工夫により、圧縮によって増加した記憶領域を有効に活用しやすくなる。

[0245] 以上、本発明の実施例を説明してきたが、これは本発明の説明のための例示であって、本発明を上で説明した実施例に限定する趣旨ではない。本発明は、他の種々の形態でも実施可能である。たとえば実施例に記載のストレージシステム１では、ホスト計算機（ホスト２）からのライトデータを格納する最終記憶媒体が、フラッシュメモリを用いた記憶装置であるＦＭＰＫである構成について説明したが、最終記憶媒体はフラッシュメモリを用いた記憶装置に限定されるものではない。たとえばＰｈａｓｅ　Ｃｈａｎｇｅ　ＲＡＭやＲｅｓｉｓｔａｎｃｅ　ＲＡＭ等の不揮発メモリを採用した記憶装置であってもよい。

[0246] また、上の説明において、実施例１または２に係るストレージシステム１では、１つのＲＡＩＤグループの記憶領域が、１または複数の論理ユニットに対応付けられている構成であるという前提で説明したが、本発明はこの構成に限定されるものではない。たとえば１つの論理ボリュームが、複数のＲＡＩＤグループに対応付けられる構成を採用することもできる。

[0247] また、実施例の説明において、ホスト計算機に提供する論理ユニットの記憶領域とＲＡＩＤグループの記憶領域とは、静的にマッピングされている（定義時点で、論理ユニットの各記憶領域がマッピングされるＲＡＩＤグループ上の記憶領域が一意に決定される）ものとして説明したが、論理ユニットの記憶領域とＲＡＩＤグループ（またはストライプライン）の記憶領域の関係が固定的である構成に限定されるものではない。たとえば周知の技術であ

るＴｈｉｎ－Ｐｒｏｖｉｓｉｏｎｉｎｇ技術を用いて論理ユニットを定義し、ホスト計算機から論理ユニット上の記憶領域に対してライト要求を受け付けた時点ではじめて、当該記憶領域に対してＲＡＩＤグループ（またはストライプライン）の記憶領域を割り当てる等の構成を採用することもできる。

[0248] また実施例においてプログラムとして記載されている構成物は、ハードワイヤードロジックなどを用いたハードウェアによって実現してもよい。また実施例中の各プログラムを、ＣＤ－ＲＯＭ、ＤＶＤ等の記憶媒体に格納して提供する形態をとることも可能である。

## 符号の説明

[0249] １： ストレージシステム

２： ホスト

３： ＳＡＮ

１０： ストレージコントローラ（ＤＫＣ）

１１： プロセッサ

１２： ホストＩＦ

１３： ディスクＩＦ

１４： メモリ

１５： パリティ演算回路

１６： 相互結合スイッチ

２０： ＲＡＩＤグループ

２００： 記憶デバイス（ＦＭＰＫ）

２００’： 記憶デバイス（ＨＤＤ）

２０１： ＦＭコントローラ

２０２： メモリ

２０３： プロセッサ

２０４： 圧縮伸長回路

２０５： パリティ演算回路

２０６： ＳＡＳ－ＣＴＬ

２０７：　ＦＭ－ＩＦ２０７

２０８：　内部接続スイッチ

２１０：　ＦＭチップ

# 請求の範囲

[請求項1] ホスト計算機と接続されるストレージコントローラと、前記ストレージコントローラに接続される複数の記憶デバイスとを有するストレージシステムにおいて、

前記ストレージシステムは、前記複数の記憶デバイスの中の（ｎ＋ｍ）台の記憶デバイスからＲＡＩＤグループを構成しており、

前記ホスト計算機からのライト要求とともに受け取ったｎ個のデータから、該ｎ個のデータを復元するためのｍ個の冗長データを生成し、前記ｎ個のデータと前記ｍ個の冗長データを、前記ＲＡＩＤグループを構成する（ｎ＋ｍ）台の前記記憶デバイスに格納するよう構成されており、

前記ＲＡＩＤグループを構成する記憶デバイスの１つに障害が発生した時、前記ストレージコントローラは、

前記複数の記憶デバイスの中から１台の復旧先記憶デバイスを選択し、

前記ＲＡＩＤグループの中で障害の発生していない記憶デバイスのそれぞれから、前記データ及び冗長データを圧縮状態で読み出し、

前記読み出された圧縮状態の前記データ及び冗長データを前記復旧先記憶デバイスに送信する、

ことを特徴とする、ストレージシステム。

[請求項2] 前記記憶デバイスは、記憶媒体とデバイスコントローラを有し、

前記ＲＡＩＤグループを構成する記憶デバイスの１つに障害が発生した時、前記ストレージコントローラは前記ＲＡＩＤグループの中で障害の発生していない記憶デバイスのそれぞれに圧縮Ｒｅａｄコマンドを発行し、

前記デバイスコントローラは、前記ストレージコントローラから前記圧縮Ｒｅａｄコマンドを受信すると、前記記憶デバイスに格納されているデータを圧縮状態で読み出して、前記ストレージコントローラ

に前記圧縮状態のデータを転送する、

ことを特徴とする、請求項１に記載のストレージシステム。

[請求項3] 前記記憶媒体に前記データが非圧縮状態で格納されている場合、

前記デバイスコントローラは、前記ストレージコントローラから前記圧縮Ｒｅａｄコマンドを受信すると、前記記憶媒体に格納されているデータを圧縮した後、前記ストレージコントローラに前記圧縮状態のデータを転送する、

ことを特徴とする、請求項２に記載のストレージシステム。

[請求項4] 前記ストレージシステムは、前記ｎ個のデータと前記ｍ個の冗長データを圧縮状態で、前記（ｎ＋ｍ）台の記憶デバイスに格納するよう構成されており、

前記デバイスコントローラは、前記ストレージコントローラから前記圧縮Ｒｅａｄコマンドを受信すると、前記記憶媒体に格納されている圧縮状態のデータを前記ストレージコントローラに転送する、

ことを特徴とする、請求項２に記載のストレージシステム。

[請求項5] 前記ストレージコントローラは、前記圧縮状態のデータまたは冗長データを前記復旧先記憶デバイスに送信する際、圧縮パリティ演算Ｗｒｉｔｅコマンドを前記復旧先記憶デバイスに発行し、

前記圧縮パリティ演算Ｗｒｉｔｅコマンドを受信した前記復旧先記憶デバイスは、前記圧縮状態のデータまたは冗長データを伸長し、前記伸長されたデータまたは冗長データと過去に前記圧縮パリティ演算Ｗｒｉｔｅコマンドを受信した際に生成された中間パリティとから、新たなパリティを算出し、前記新たなパリティを中間パリティとして前記記憶デバイスに格納する、

ことを特徴とする、請求項２に記載のストレージシステム。

[請求項6] 前記圧縮パリティ演算Ｗｒｉｔｅコマンドを受信した前記復旧先記憶デバイスは、過去に前記圧縮パリティ演算Ｗｒｉｔｅコマンドを受信していなかった場合、前記圧縮パリティ演算Ｗｒｉｔｅコマンドと

ともに受信した前記圧縮状態のデータ又は冗長データを伸長し、前記伸長されたデータを前記中間パリティとして前記記憶デバイスに格納する、

ことを特徴とする、請求項５に記載のストレージシステム。

[請求項7] 前記圧縮パリティ演算Ｗｒｉｔｅコマンドを受信した前記復旧先記憶デバイスは、前記圧縮パリティ演算Ｗｒｉｔｅコマンドを受信した回数を記憶しており、前記圧縮パリティ演算Ｗｒｉｔｅコマンドを受信した回数が所定回数に達した時点で、前記新たなパリティを復元されたデータとして、前記記憶媒体に格納する、

ことを特徴とする、請求項５に記載のストレージシステム。

[請求項8] 前記復旧先記憶デバイスは、コミットコマンドを受信すると、

前記記憶デバイスに格納された中間パリティを復元されたデータとして、前記記憶媒体に格納する、

ことを特徴とする、請求項５に記載のストレージシステム。

[請求項9] 前記ストレージコントローラは、前記ホスト計算機からのライト要求とともに受け取ったデータに検証用情報を付加して、前記記憶デバイスに格納するよう構成されており、

前記圧縮パリティ演算Ｗｒｉｔｅコマンドを受信した前記復旧先記憶デバイスは、前記圧縮状態のデータまたは冗長データを伸長し、前記伸長されたデータに付加されている前記検証用情報を用いてデータの検証を行い、

前記検証結果が正常でなかった場合、前記ストレージコントローラにエラーを返却する、

ことを特徴とする、請求項５に記載のストレージシステム。

[請求項10] 前記ストレージコントローラは、前記ホスト計算機からのライト要求とともに受け取ったデータに前記検証用情報を付加する際、前記検証用情報に前記記憶デバイス上のアドレス情報を格納するよう構成されており、

前記圧縮パリティ演算Ｗｒｉｔｅコマンドには、前記圧縮状態のデータまたは冗長データの格納されていた、前記記憶デバイス上のアドレス情報が含まれており、

前記圧縮パリティ演算Ｗｒｉｔｅコマンドを受信した前記復旧先記憶デバイスは、前記圧縮状態のデータまたは冗長データを伸長し、前記伸長されたデータに付加されている前記検証用情報に含まれている前記アドレス情報と、前記圧縮パリティ演算Ｗｒｉｔｅコマンドに含まれている前記アドレス情報とが一致していない場合、前記ストレージコントローラにエラーを返却する、

ことを特徴とする、請求項９に記載のストレージシステム。

[請求項11] 前記ストレージコントローラは、前記ＲＡＩＤグループを構成する記憶デバイスの１つに障害が発生した時、

前記障害が発生した記憶デバイスの記憶領域のうち、アクセス可能な記憶領域が存在する場合、前記ＲＡＩＤグループの中で障害の発生していない記憶デバイスのそれぞれから前記データ及び冗長データを読み出すことに代えて、

前記アクセス可能な記憶領域からデータを圧縮状態で読み出し、

前記読み出された圧縮状態の前記データを復旧先記憶デバイスに格納する、

ことを特徴とする、請求項１に記載のストレージシステム。

[請求項12] 記憶媒体とデバイスコントローラを有し、ストレージコントローラからのデータアクセス要求を受け付ける記憶デバイスであって、

前記記憶デバイスは、前記ストレージコントローラから圧縮パリティ演算Ｗｒｉｔｅコマンドと圧縮状態のデータを受信すると、

前記圧縮状態のデータを伸長し、前記伸長されたデータと、過去に前記圧縮パリティ演算Ｗｒｉｔｅコマンドを受信した際に生成された中間パリティとから、新たなパリティを算出し、前記新たなパリティを中間パリティとして前記記憶デバイスに格納する、

ことを特徴とする、記憶デバイス。

[請求項13] 前記デバイスコントローラは、前記ストレージコントローラから前記圧縮Ｒｅａｄコマンドを受信すると、前記記憶デバイスに格納されているデータを圧縮状態で読み出して、前記ストレージコントローラに前記圧縮状態のデータを転送する、

ことを特徴とする、請求項１２に記載の記憶デバイス。

[請求項14] 前記圧縮パリティ演算Ｗｒｉｔｅコマンドを受信した前記記憶デバイスは、前記圧縮パリティ演算Ｗｒｉｔｅコマンドを受信した回数を記憶しており、前記圧縮パリティ演算Ｗｒｉｔｅコマンドを受信した回数が所定回数に達した時点で、前記新たなパリティを復元されたデータとして、前記記憶媒体に格納する、

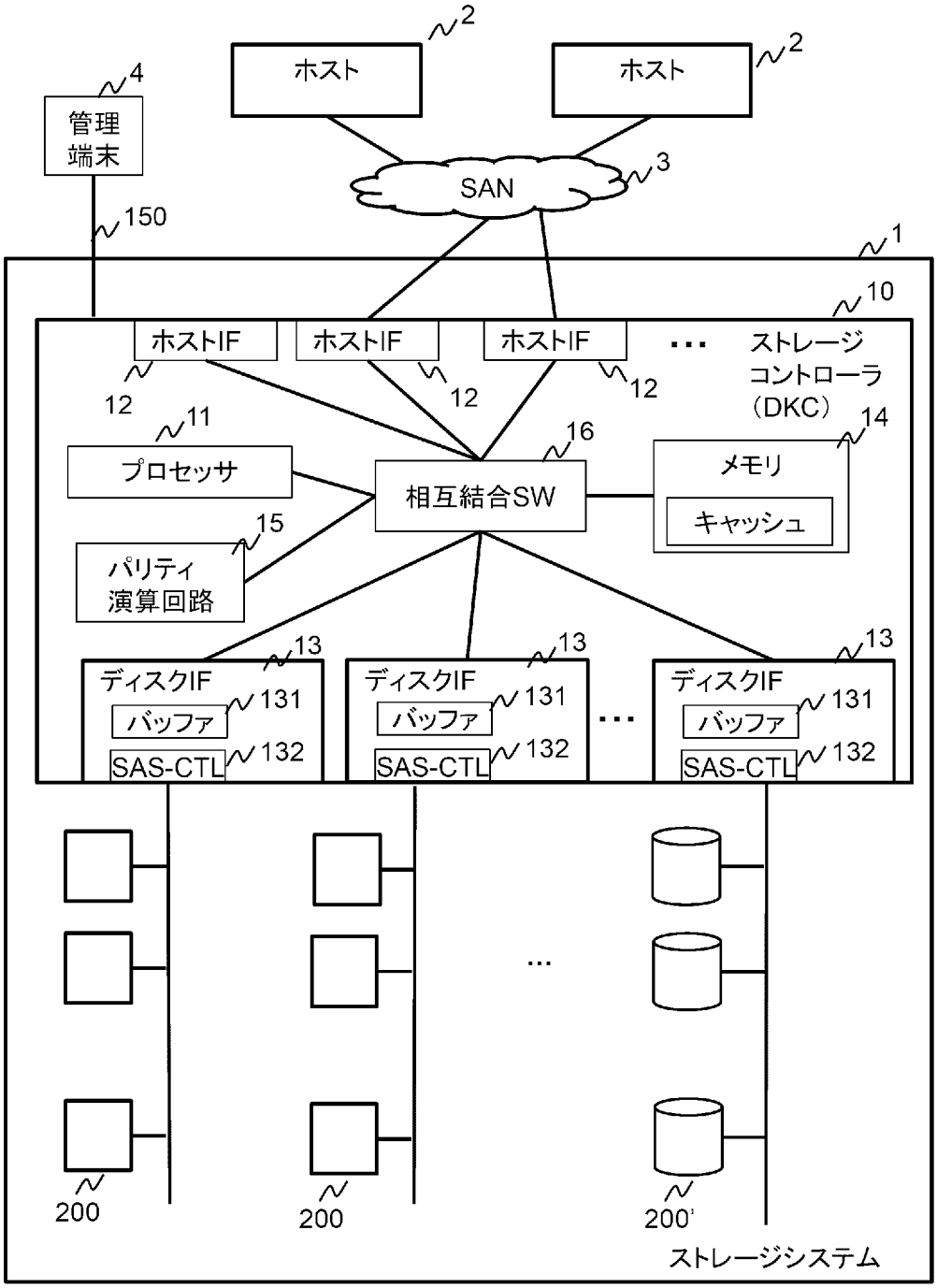
ことを特徴とする、請求項１２に記載の記憶デバイス。

[請求項15] 前記復旧先記憶デバイスは、コミットコマンドを受信すると、前記記憶デバイスに格納された中間パリティを復元されたデータとして、前記記憶媒体に格納する、

ことを特徴とする、請求項１２に記載の記憶デバイス。

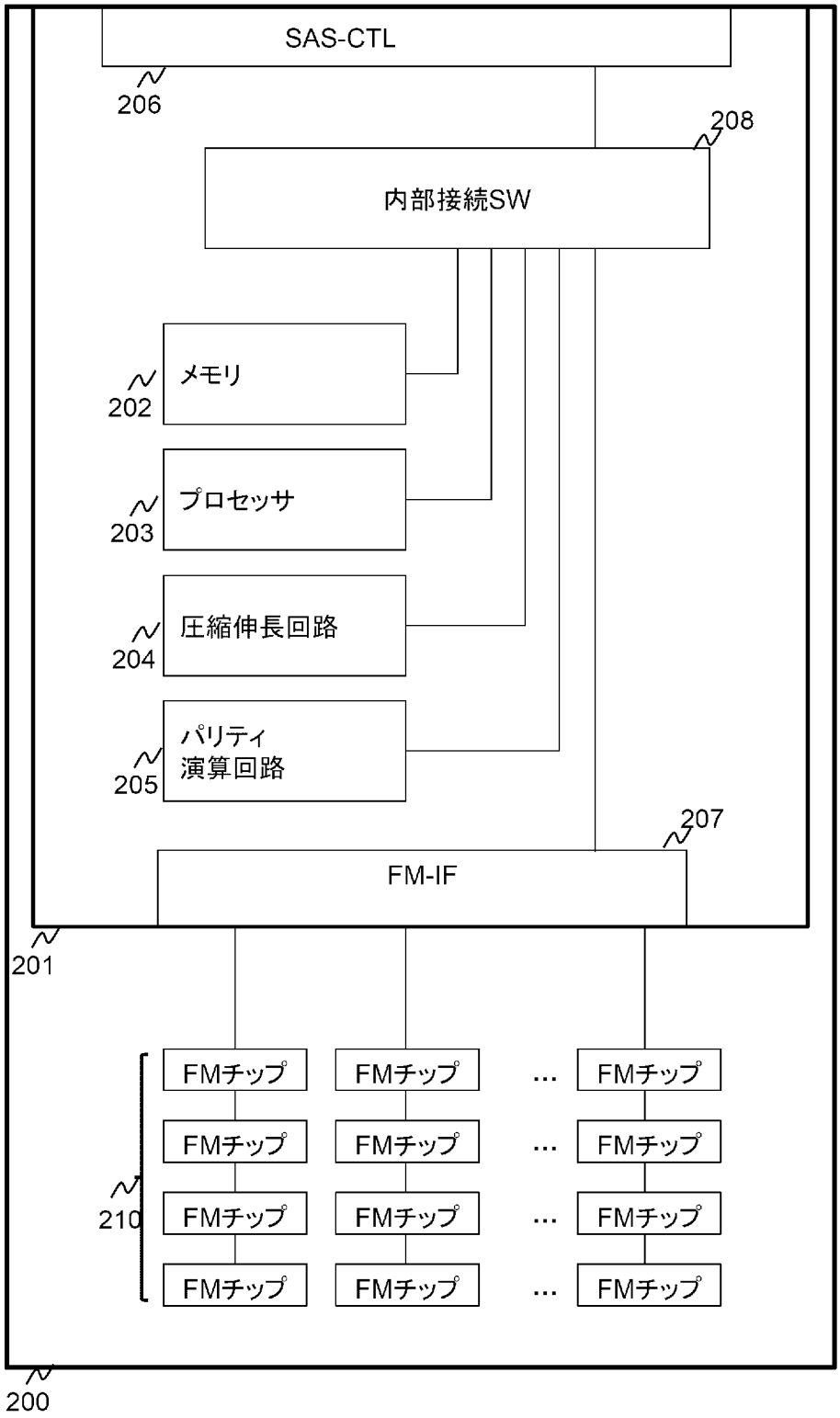
[図1]

図1


ホスト
ホスト
管理
端末
SAN
ホストIF
ホストIF
ホストIF
ストレージ
コントローラ
（DKC）
プロセッサ
相互結合SW
メモリ
キャッシュ
パリティ
演算回路
ディスクIF
バッファ
SAS-CTL
ディスクIF
バッファ
SAS-CTL
ディスクIF
バッファ
SAS-CTL
200
200
200'
ストレージシステム

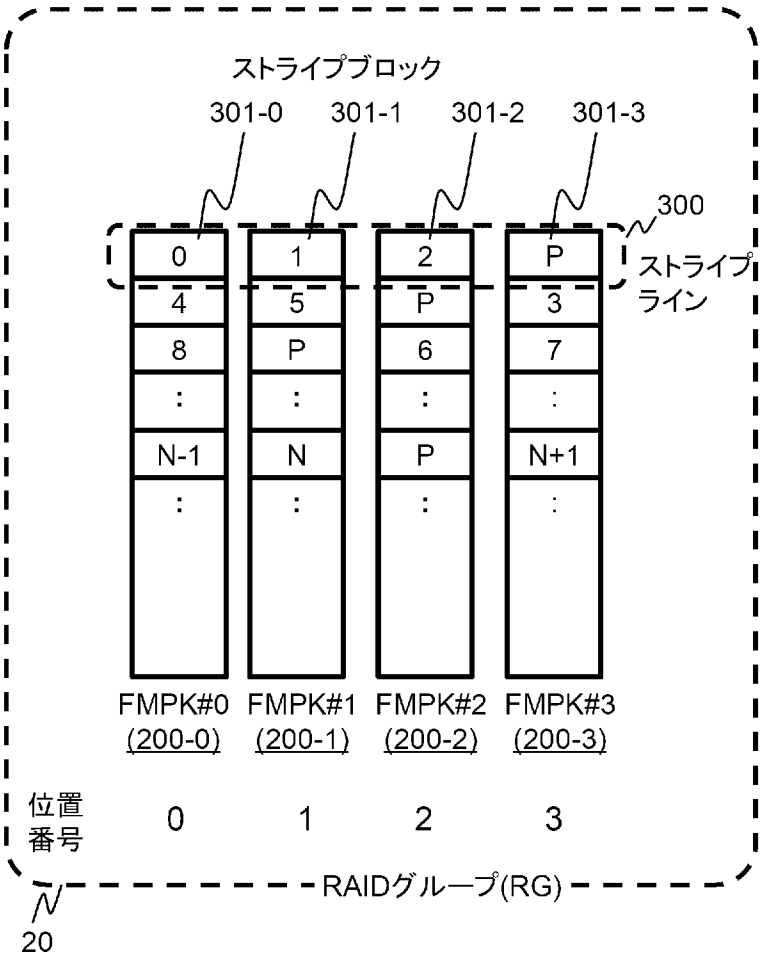
[図2]

図2


SAS-CTL
206
208
内部接続SW
メモリ
202
プロセッサ
203
圧縮伸長回路
204
パリティ
演算回路
205
207
FM-IF
201
FMチップ
210
200

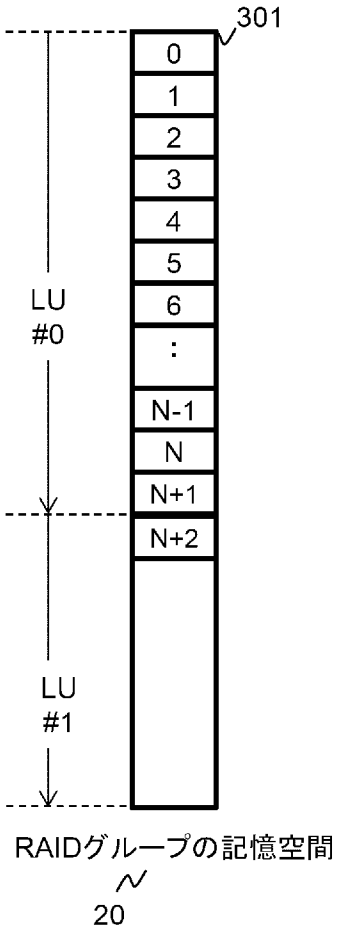
[図3]

図3


ストライプブロック
301-0
301-1
301-2
301-3
300
ストライプライン
0
1
2
P
4
5
P
3
8
P
6
7
:
:
:
:
N-1
N
P
N+1
:
:
:
:
FMPK#0 (200-0)
FMPK#1 (200-1)
FMPK#2 (200-2)
FMPK#3 (200-3)
位置番号
0
1
2
3
RAIDグループ(RG)
20

[図4]

図4



[図5]

図5

T1000

| T1001 デバイス # | T1002 デバイス種別 | T1003 所属RG# | T1004 デバイスステータス | T1005 圧縮機能サポート | T1006 パリティ演算機能サポート | T1007 サイズ |
|---|---|---|---|---|---|---|
| 0 | FMPK | 0 | 正常 | サポート | サポート | 10240GB |
| 1 | FMPK | 0 | 障害復旧中（復旧元） | サポート | サポート | 10240GB |
| 2 | FMPK | 0 | 正常 | サポート | サポート | 10240GB |
| 3 | FMPK | 0 | 正常 | サポート | サポート | 10240GB |
| 4 | FMPK | 0 | 障害復旧中（復旧先） | サポート | サポート | 10240GB |
| 5 | FMPK | 未割当(スペア) | 正常 | サポート | サポート | 10240GB |
| 6 | FMPK | 未割当 | 正常 | サポート | サポート | 5120GB |
| 7 | FMPK | 未割当 | 閉塞 | サポート | サポート | 10240GB |
| 8 | FMPK | 1 | 正常 | サポート | サポート | : |
| : | : | : | : | : | : | : |
| 16 | HDD | 2 | 正常 | 未サポート | 未サポート | : |
| : | : | : | : | : | : | |

[図6]

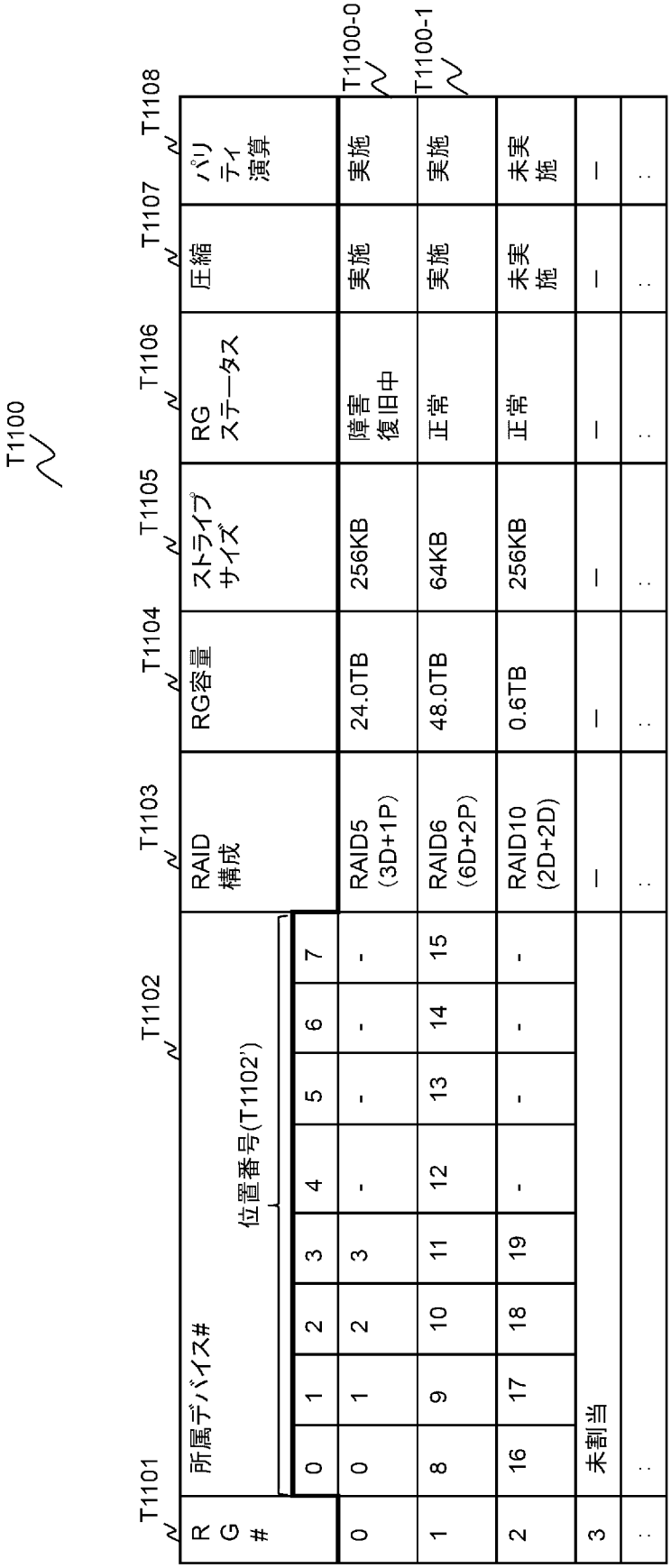

T1100
T1101
T1102
T1103
T1104
T1105
T1106
T1107
T1108
T1100-0
T1100-1
RG #
所属デバイス#
位置番号(T1102')
0
1
2
3
4
5
6
7
RAID 構成
RG容量
ストライプ サイズ
RG ステータス
圧縮
パリティ演算
0
0
1
2
3
-
-
-
RAID5 (3D+1P)
24.0TB
256KB
障害 復旧中
実施
実施
1
8
9
10
11
12
13
14
15
RAID6 (6D+2P)
48.0TB
64KB
正常
実施
実施
2
16
17
18
19
-
-
-
-
RAID10 (2D+2D)
0.6TB
256KB
正常
未実施
未実施
3
未割当
—
—
—
—
—
..

図6

[図7]

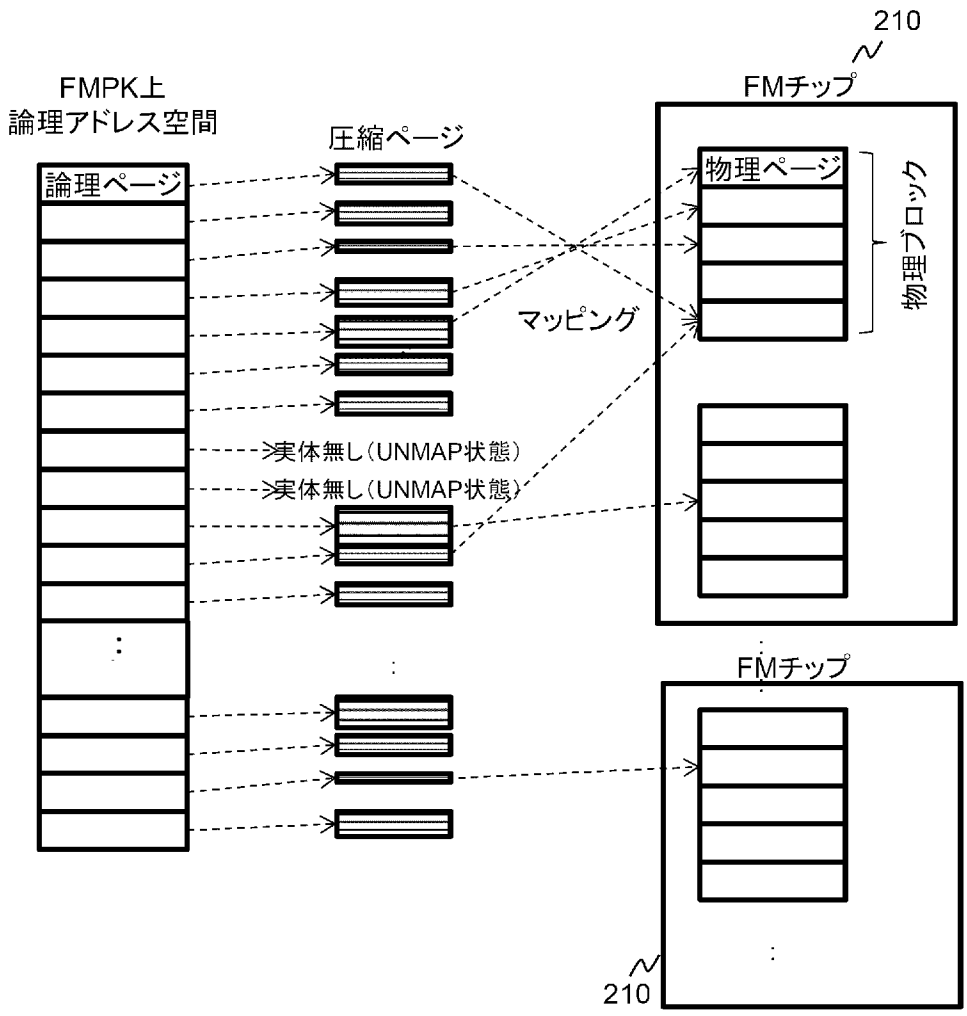
図7

T1200

| T1201 LU# | T1202 RG# | T1203 RG内LU開始オフセット | T1204 LUサイズ |
|---|---|---|---|
| 0 | 0 | 0x0 | 100GB |
| 1 | 0 | 0xC800000 | 500GB |
| 2 | 0 | 0x4B000000 | 50GB |
| : | : | : | : |
| 5 | 2 | 0x20000 | 10GB |
| : | : | : | : |

[図8]

図8



[図9]

図9

T2100

| T2101 論理ページ番号 | T2102 物理ページ番号 | T2103 サイズ（圧縮サイズ） | T2104 オフセット |
|---|---|---|---|
| 0 | 10 | 4KB | 0KB |
| 1 | 500 | 1KB | 7KB |
| | 42 | 1KB | 0KB |
| 2 | NULL | — | — |
| : | : | : | : |
| 1000 | 10 | 2KB | 4KB |
| : | : | : | : |

T2100-1

T2100-0

T2100-2

[図10]

図10

T2000

| 項目 | 内容 |
|---|---|
| 物理容量（非圧縮容量） | 1634GB |
| データ圧縮(実施有無) | 有 |
| 論理容量 | 10240GB |
| 接続DKC種別 | Hitachi VSP |
| 所属RAIDグループ構成 | RAID5(3D+1P) |
| RAIDグループ内位置 | 2 |
| 所属RG# | 0 |
| : | : |

T2001

T2002

T2003

T2004

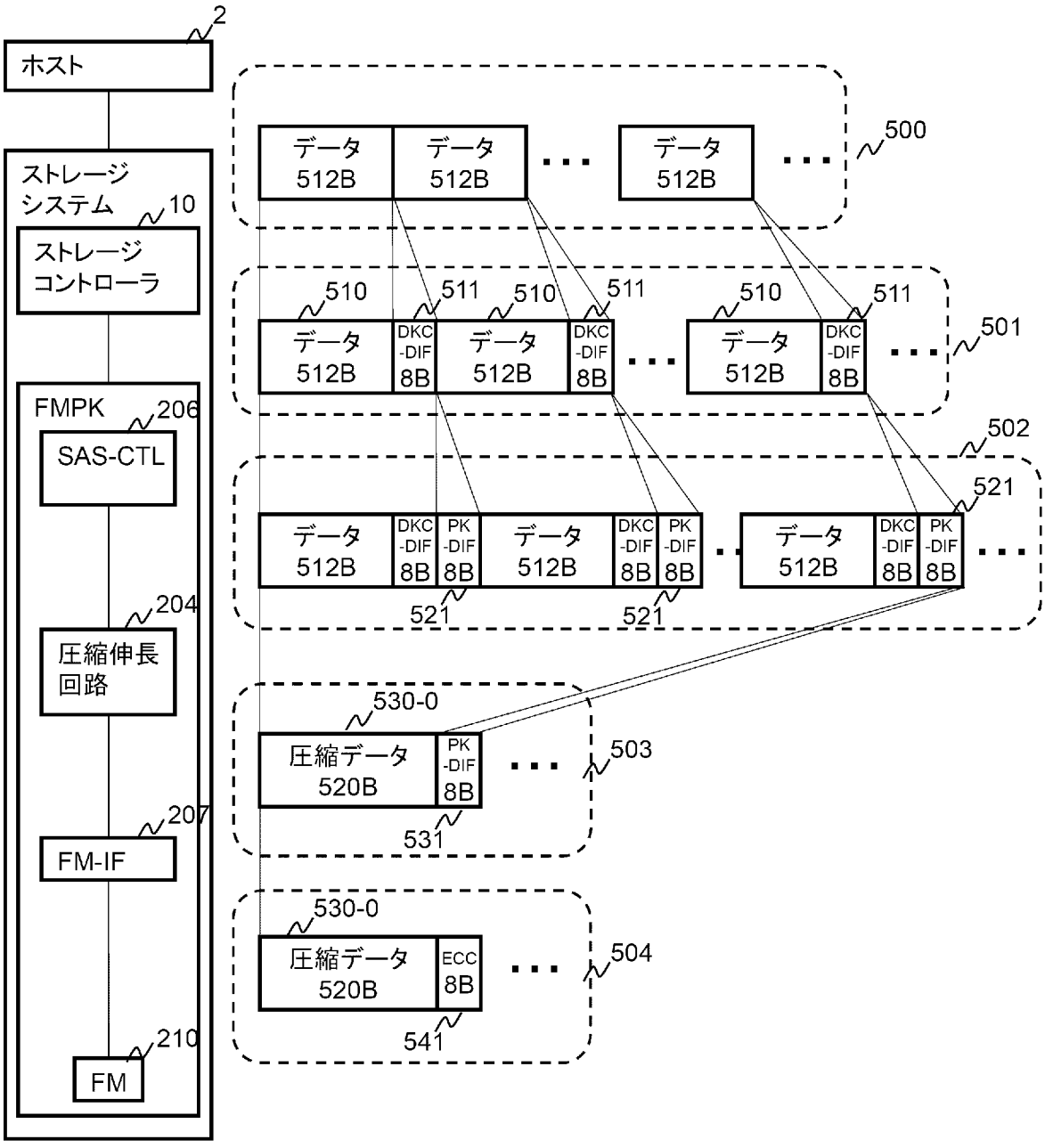
T2005

T2006

T2007

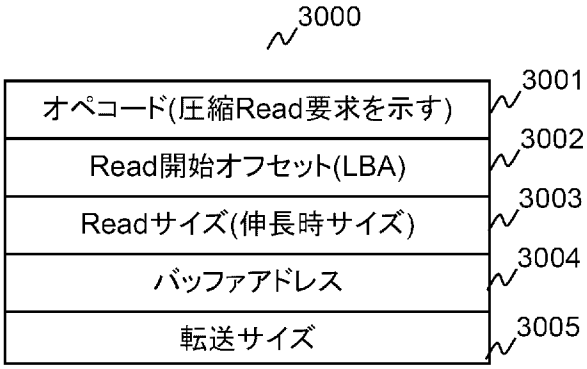
[図11]

図11


2
ホスト
ストレージ
システム
10
ストレージ
コントローラ
FMPK
206
SAS-CTL
204
圧縮伸長
回路
207
FM-IF
210
FM
500
データ
512B
501
510
511
DKC
-DIF
8B
502
521
PK
-DIF
503
530-0
圧縮データ
520B
531
504
ECC
541

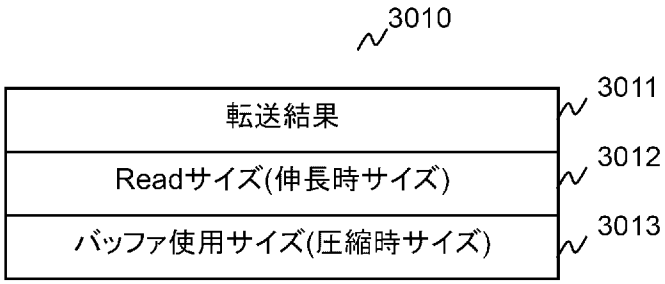
[図12]

図12



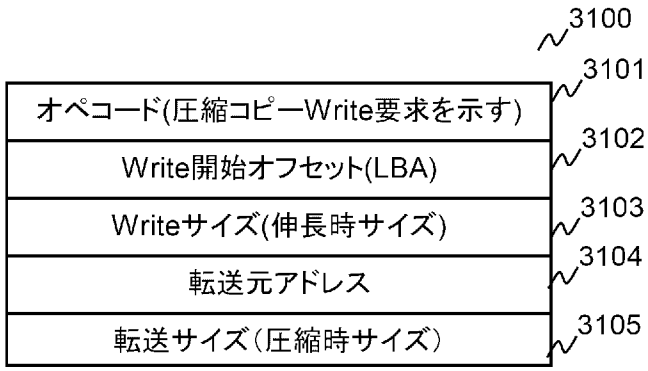

[図13]

図13



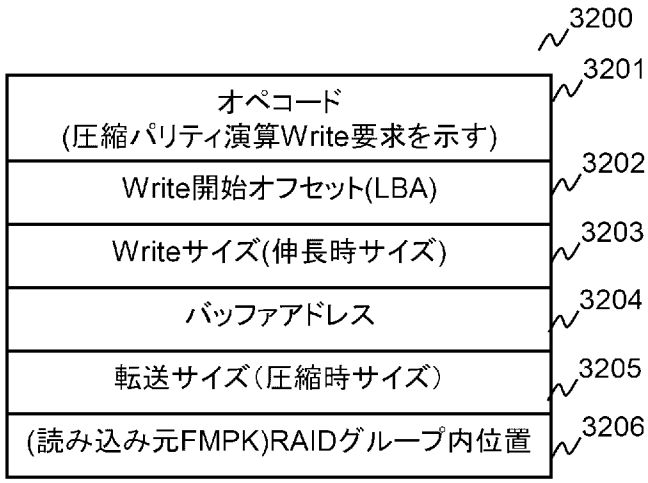

[図14]

図14



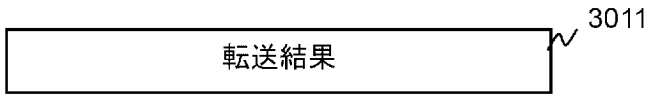

[図15]

図15



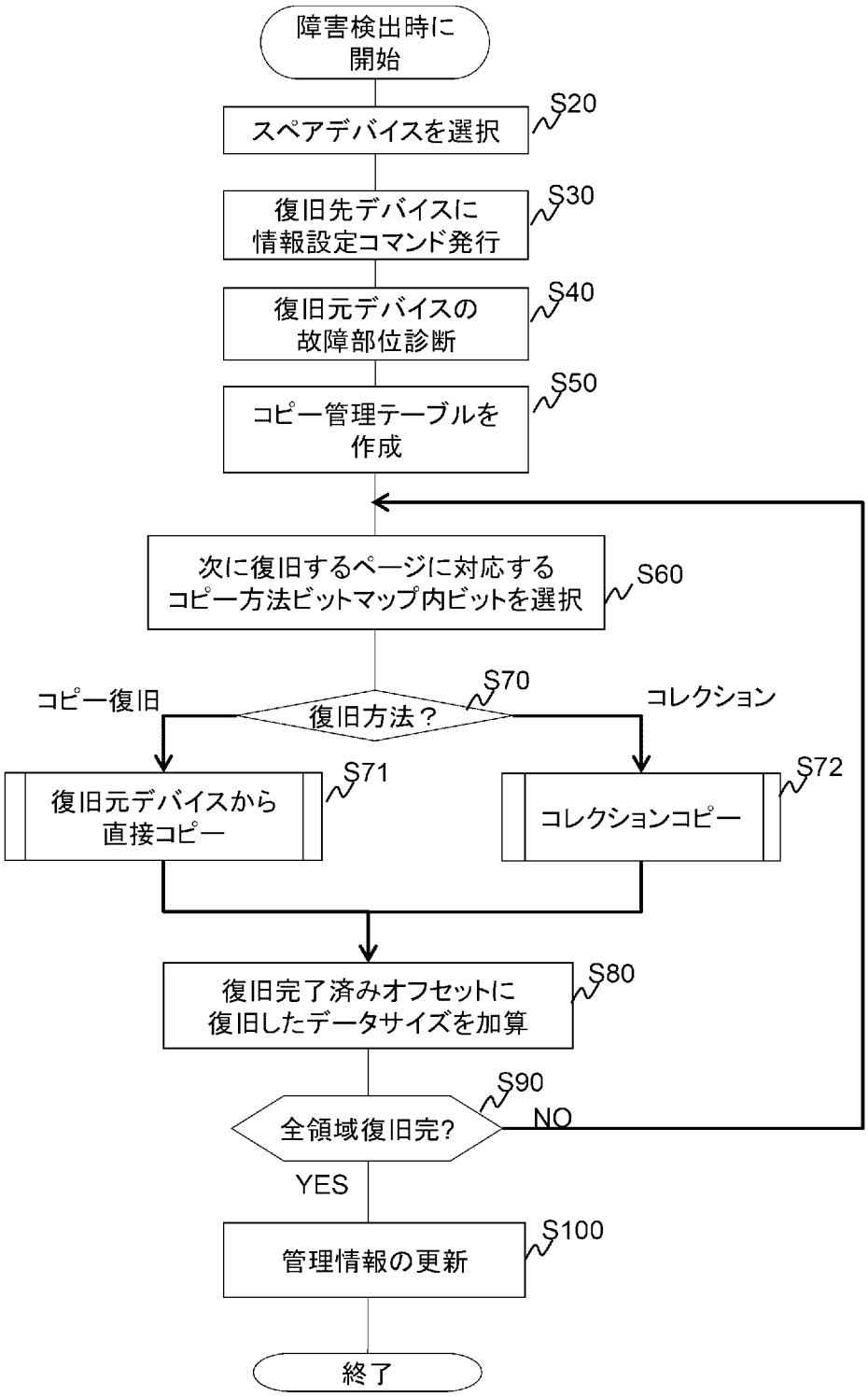
[図16]

図16




[図17]

図17



[図18]

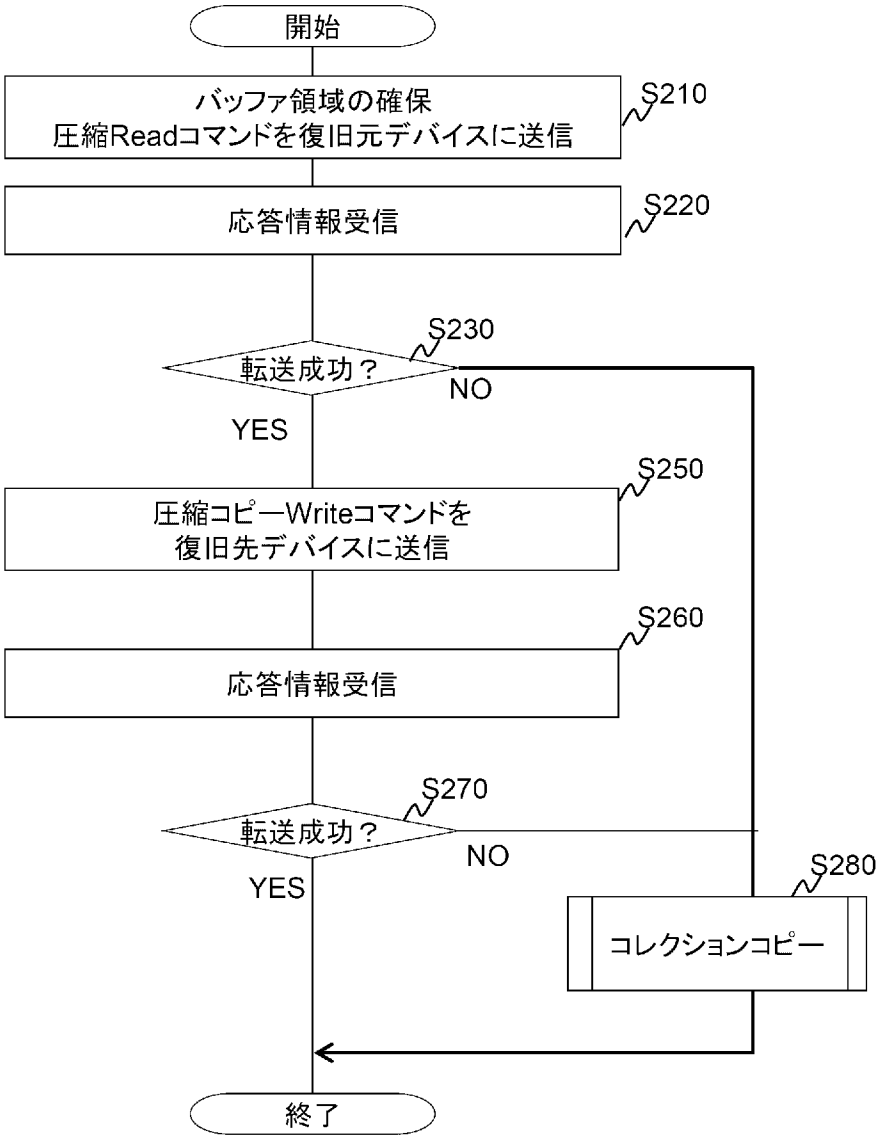
図18

T1500

| 項目 | 内容 | |
|---|---|---|
| 障害RG# | 0 | T1501 |
| 復旧元デバイス | FMPK#1 | T1502 |
| 復旧先デバイス | FMPK#4 | T1503 |
| コレクション方式 | 方式3<br>(FMPK連携圧縮コピー) | T1504 |
| 復旧デバイス容量 | 10240GB | T1505 |
| コピー方法ビットマップ<br>0:復旧元から直接コピー<br>1:コレクションコピー | 0000001000111110000・・・ | T1506 |
| 復旧完了済みオフセット | 0x10500000 | T1507 |

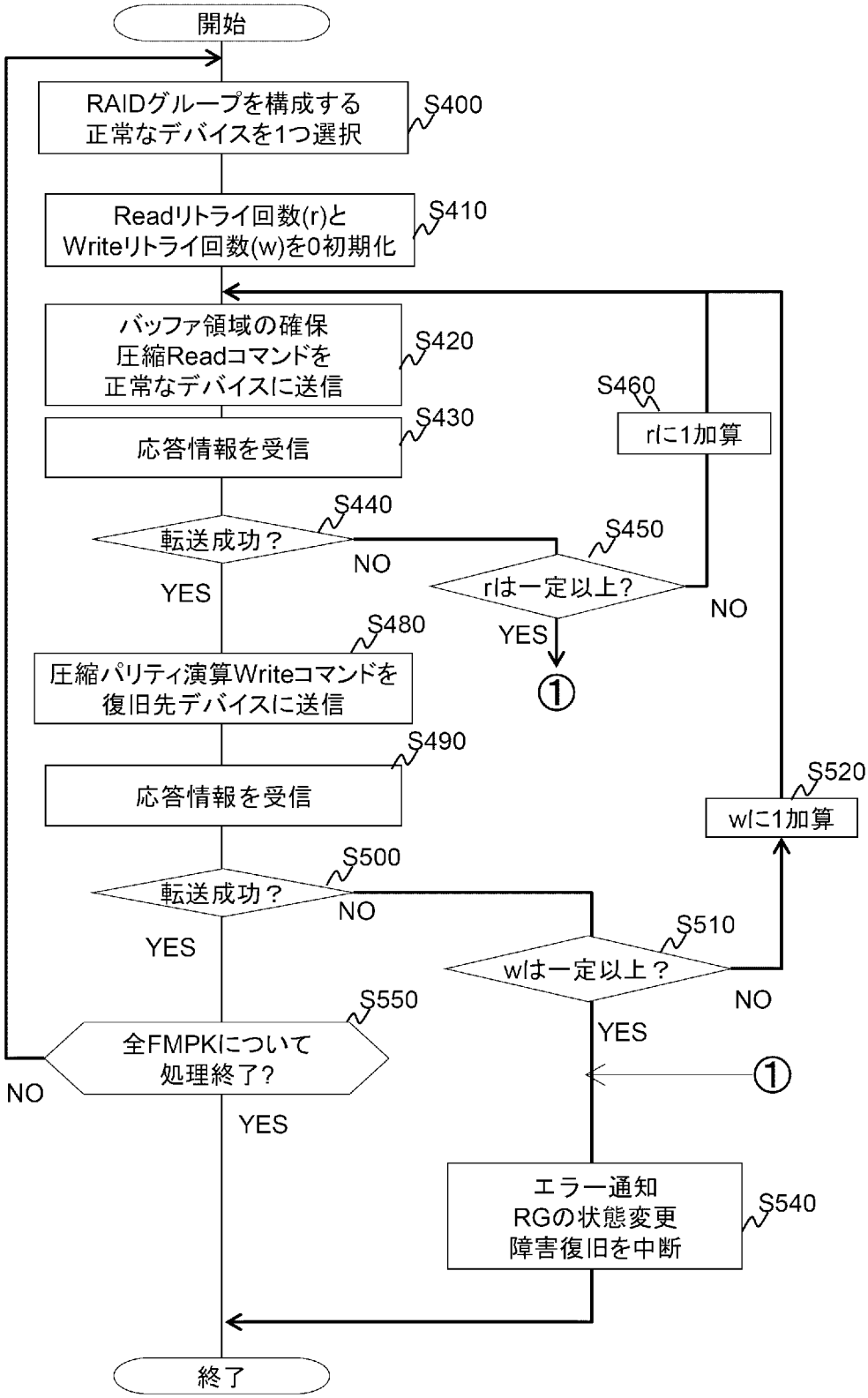
[図19]

図19


開始
バッファ領域の確保
圧縮Readコマンドを復旧元デバイスに送信
S210
応答情報受信
S220
転送成功？
S230
NO
YES
圧縮コピーWriteコマンドを
復旧先デバイスに送信
S250
応答情報受信
S260
転送成功？
S270
NO
YES
コレクションコピー
S280
終了

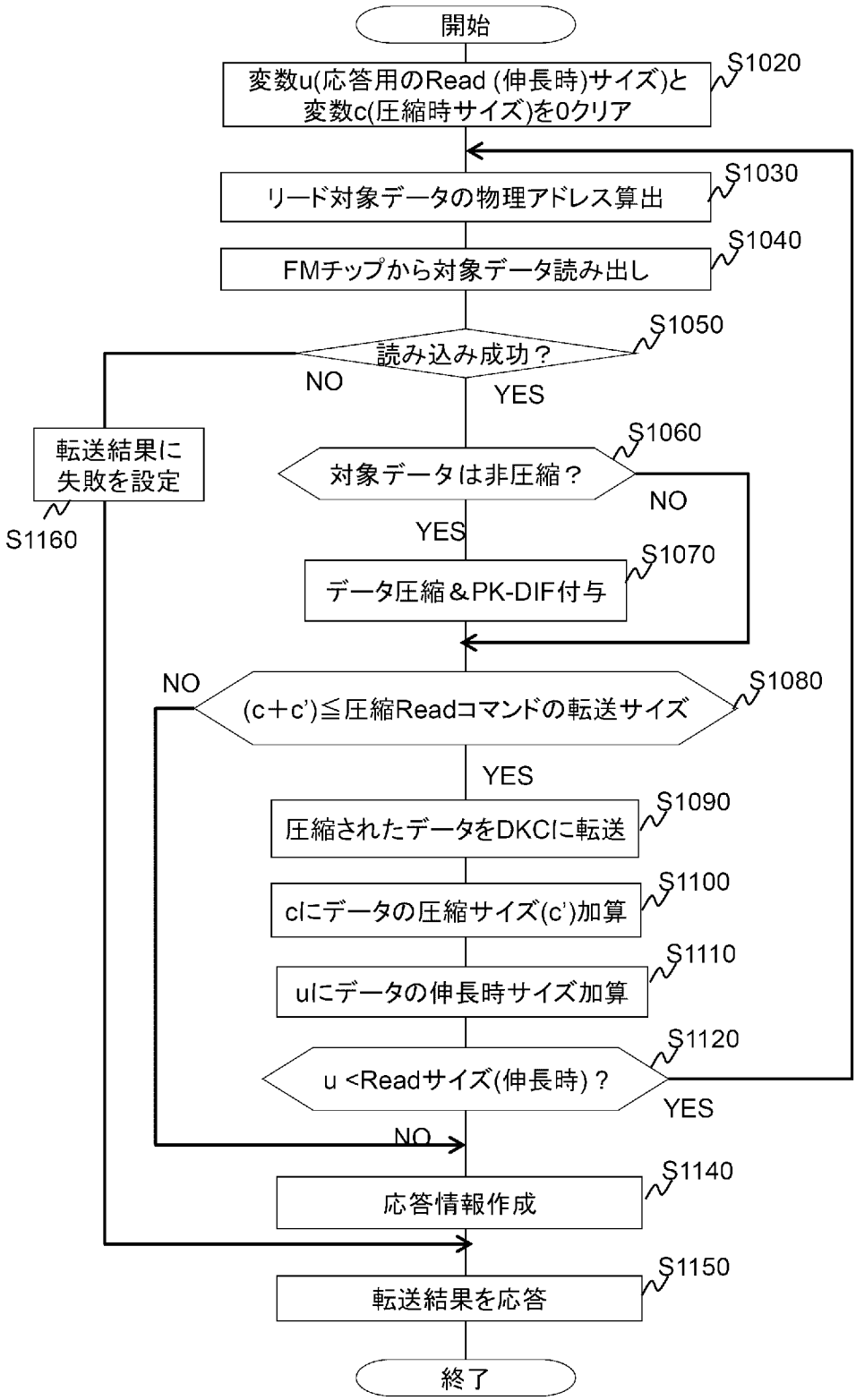
[図20]

図20


開始
RAIDグループを構成する
正常なデバイスを1つ選択
S400
Readリトライ回数(r)と
Writeリトライ回数(w)を0初期化
S410
バッファ領域の確保
圧縮Readコマンドを
正常なデバイスに送信
S420
応答情報を受信
S430
転送成功？
S440
NO
YES
rは一定以上?
S450
NO
YES
S460
rに1加算
①
圧縮パリティ演算Writeコマンドを
復旧先デバイスに送信
S480
応答情報を受信
S490
転送成功？
S500
NO
YES
wは一定以上？
S510
NO
YES
S520
wに1加算
全FMPKについて
処理終了?
S550
NO
YES
①
エラー通知
RGの状態変更
障害復旧を中断
S540
終了

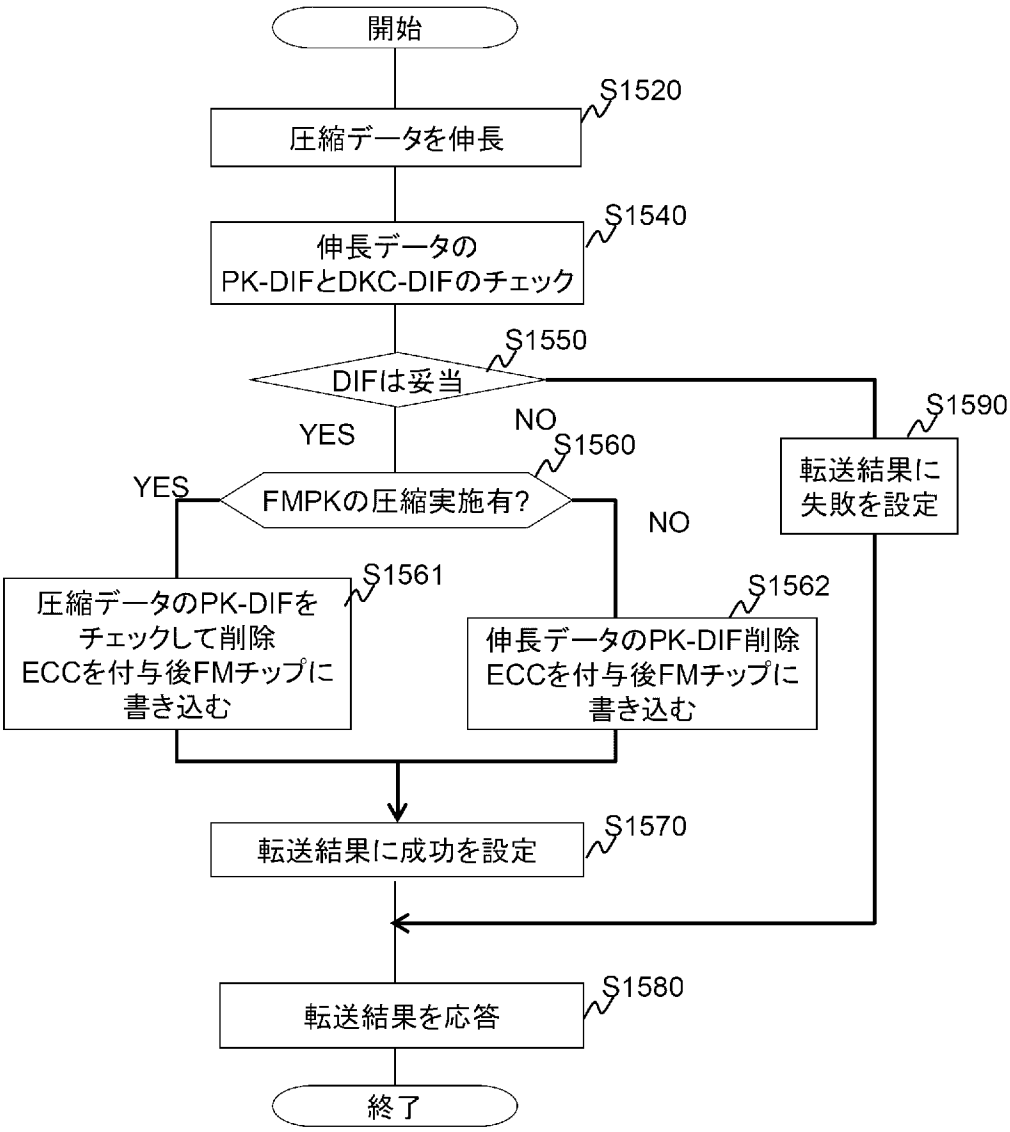
[図21]

図21


開始
変数u(応答用のRead (伸長時)サイズ)と
変数c(圧縮時サイズ)を0クリア
S1020
リード対象データの物理アドレス算出
S1030
FMチップから対象データ読み出し
S1040
読み込み成功？
S1050
NO
YES
転送結果に
失敗を設定
S1160
対象データは非圧縮？
S1060
NO
YES
データ圧縮＆PK-DIF付与
S1070
NO
(c＋c')≦圧縮Readコマンドの転送サイズ
S1080
YES
圧縮されたデータをDKCに転送
S1090
cにデータの圧縮サイズ(c')加算
S1100
uにデータの伸長時サイズ加算
S1110
u <Readサイズ(伸長時)？
S1120
YES
NO
応答情報作成
S1140
転送結果を応答
S1150
終了

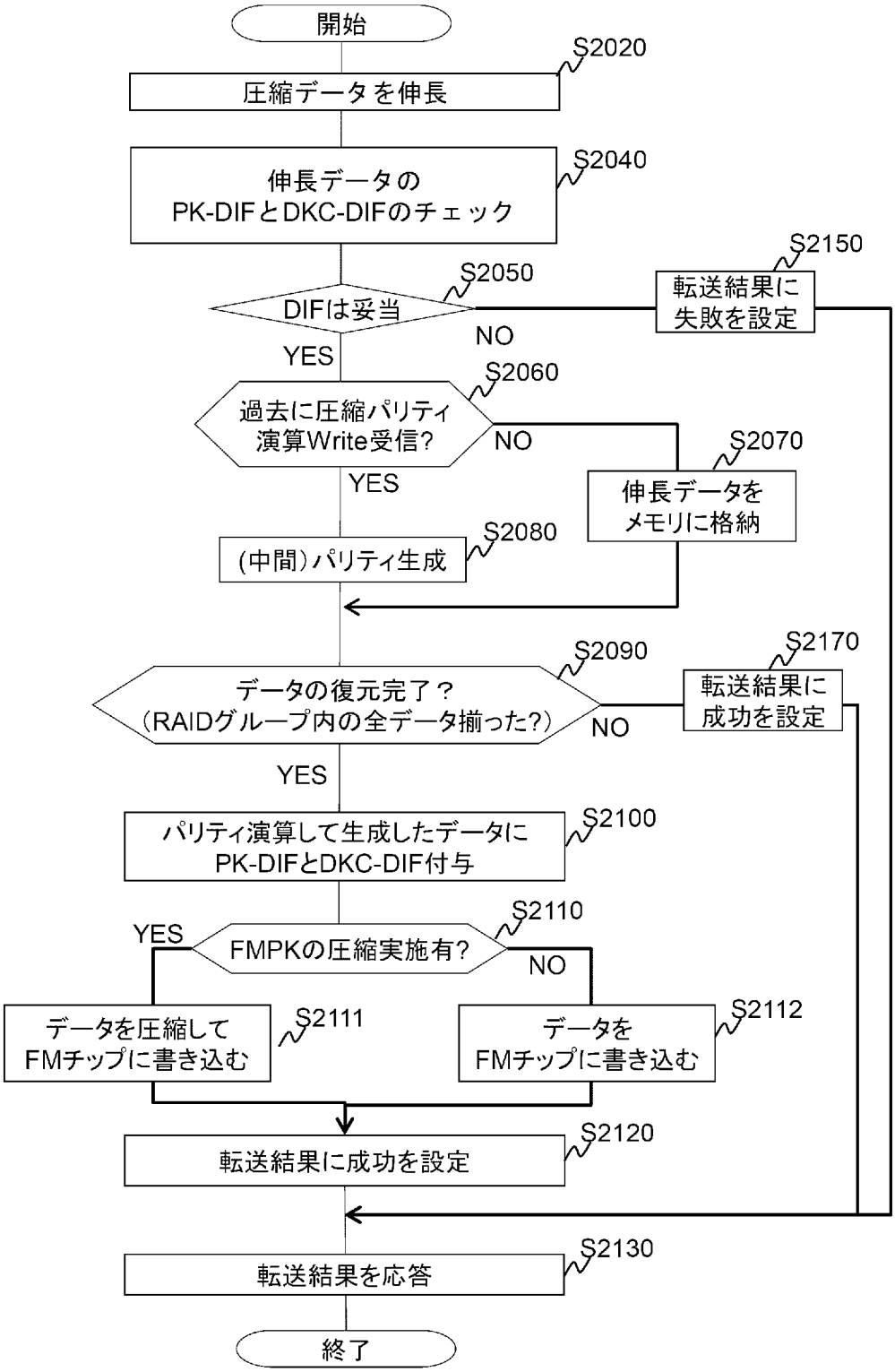
[図22]

図22


開始
S1520
圧縮データを伸長
S1540
伸長データの
PK-DIFとDKC-DIFのチェック
S1550
DIFは妥当
YES
NO
S1590
転送結果に
失敗を設定
S1560
YES
FMPKの圧縮実施有?
NO
S1561
圧縮データのPK-DIFを
チェックして削除
ECCを付与後FMチップに
書き込む
S1562
伸長データのPK-DIF削除
ECCを付与後FMチップに
書き込む
S1570
転送結果に成功を設定
S1580
転送結果を応答
終了

[図23]

図23


開始
S2020
圧縮データを伸長
S2040
伸長データの
PK-DIFとDKC-DIFのチェック
S2150
S2050
DIFは妥当
転送結果に
失敗を設定
NO
YES
S2060
過去に圧縮パリティ
演算Write受信?
NO
S2070
伸長データを
メモリに格納
YES
S2080
(中間)パリティ生成
S2090
S2170
データの復元完了?
(RAIDグループ内の全データ揃った?)
転送結果に
成功を設定
NO
YES
S2100
パリティ演算して生成したデータに
PK-DIFとDKC-DIF付与
S2110
YES
FMPKの圧縮実施有?
NO
S2111
データを圧縮して
FMチップに書き込む
S2112
データを
FMチップに書き込む
S2120
転送結果に成功を設定
S2130
転送結果を応答
終了

[図24]

図24

T2500

| T2501 論理ページ番号 | T2502 アドレス | T2503 回数 |
|---|---|---|
| 0 | 0x80000000 | 1 |
| 1 | 0x80001000 | 2 |
| 2 | NULL | 0 |
| : | : | |

[図25]

図25

3200'

| | |
|---|---|
| オペコード(圧縮中間パリティ演算コマンドを示す) | 3201' |
| Write開始オフセット(LBA) | 3202' |
| Writeサイズ(伸長時サイズ) | 3203' |
| バッファアドレス | 3204' |
| 転送サイズ(圧縮時サイズ) | 3205' |
| (読み込み元FMPK)RAIDグループ内位置 | 3206' |

[図26]

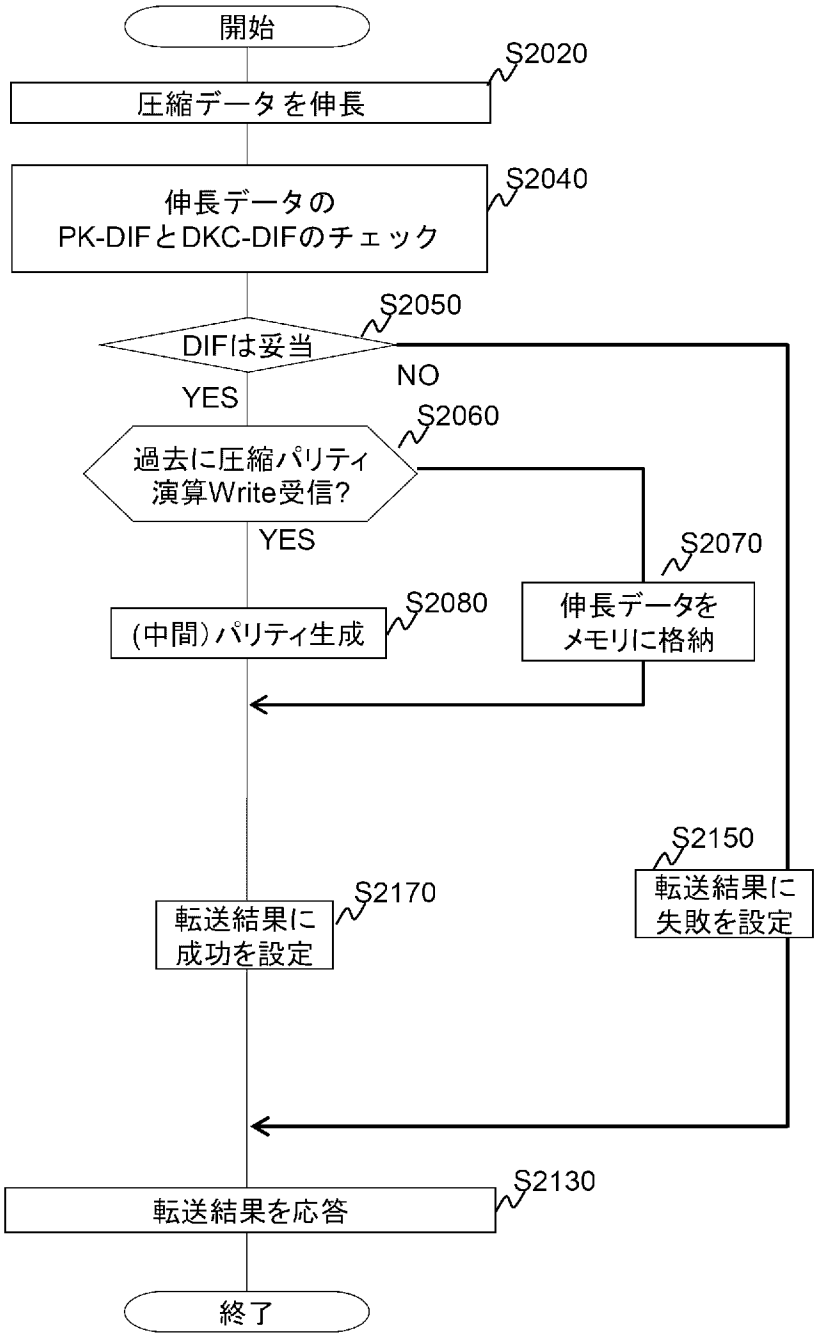
図26

3300

| | |
|---|---|
| オペコード(パリティコミットコマンドを示す) | 3301 |
| Write開始オフセット(LBA) | 3302 |
| Writeサイズ(伸長時サイズ) | 3303 |

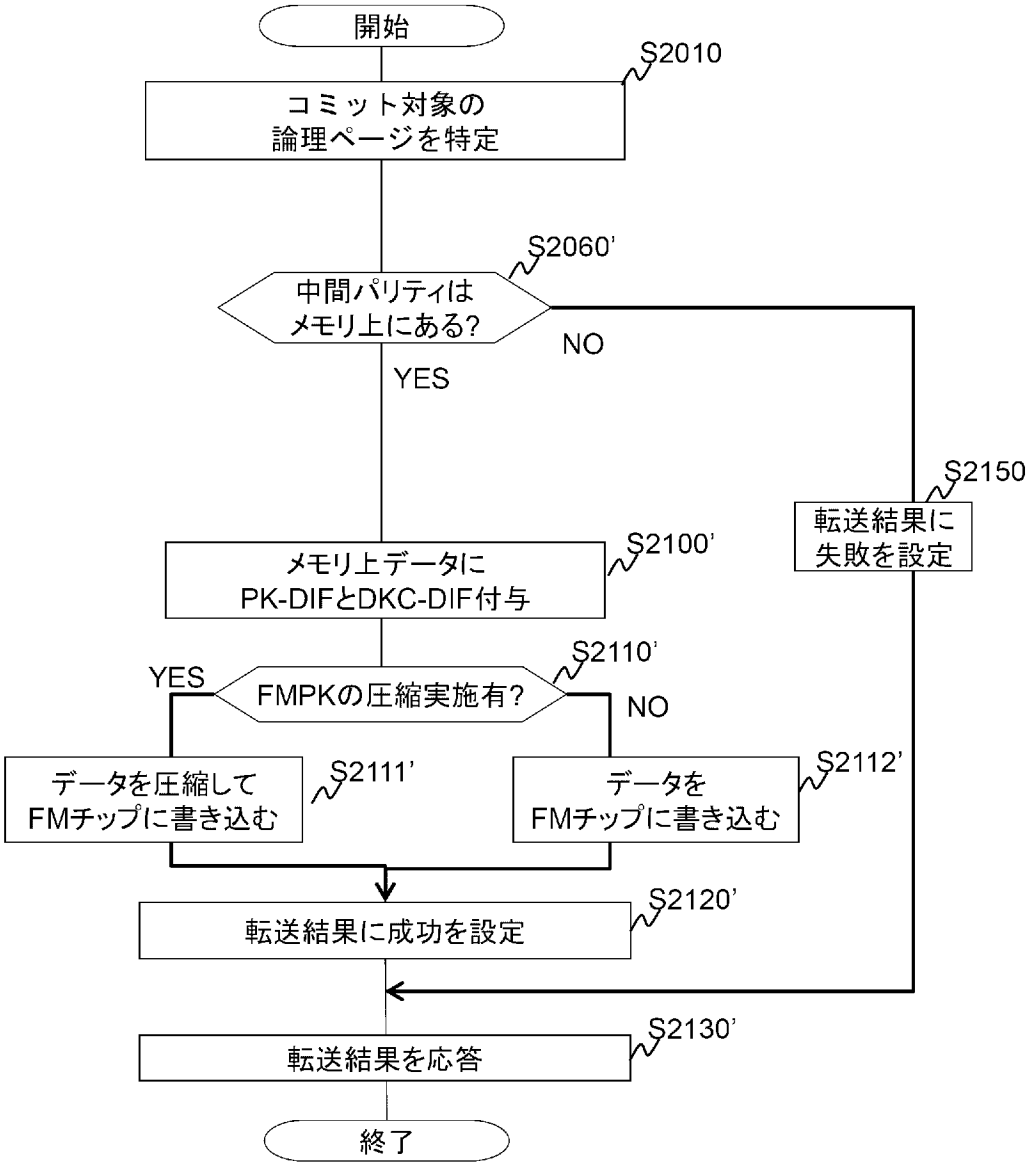
[図27]

図27


開始
S2020
圧縮データを伸長
S2040
伸長データの
PK-DIFとDKC-DIFのチェック
S2050
DIFは妥当
NO
YES
S2060
過去に圧縮パリティ
演算Write受信?
YES
S2070
伸長データを
メモリに格納
S2080
(中間)パリティ生成
S2150
転送結果に
失敗を設定
S2170
転送結果に
成功を設定
S2130
転送結果を応答
終了

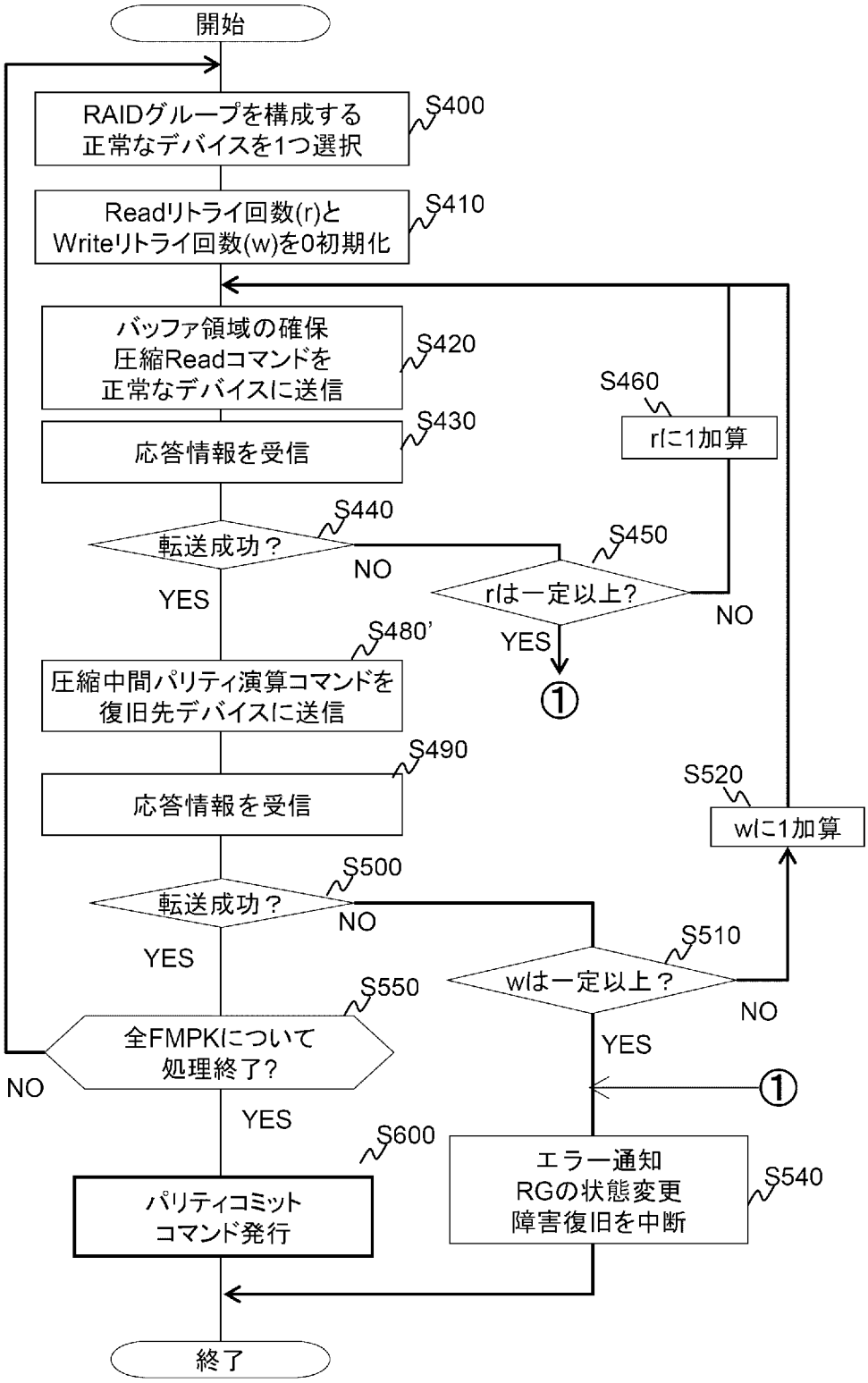
[図28]

図28


開始
S2010
コミット対象の
論理ページを特定
S2060'
中間パリティは
メモリ上にある?
NO
YES
S2150
転送結果に
失敗を設定
S2100'
メモリ上データに
PK-DIFとDKC-DIF付与
S2110'
YES
FMPKの圧縮実施有?
NO
データを圧縮して
FMチップに書き込む
S2111'
データを
FMチップに書き込む
S2112'
S2120'
転送結果に成功を設定
S2130'
転送結果を応答
終了

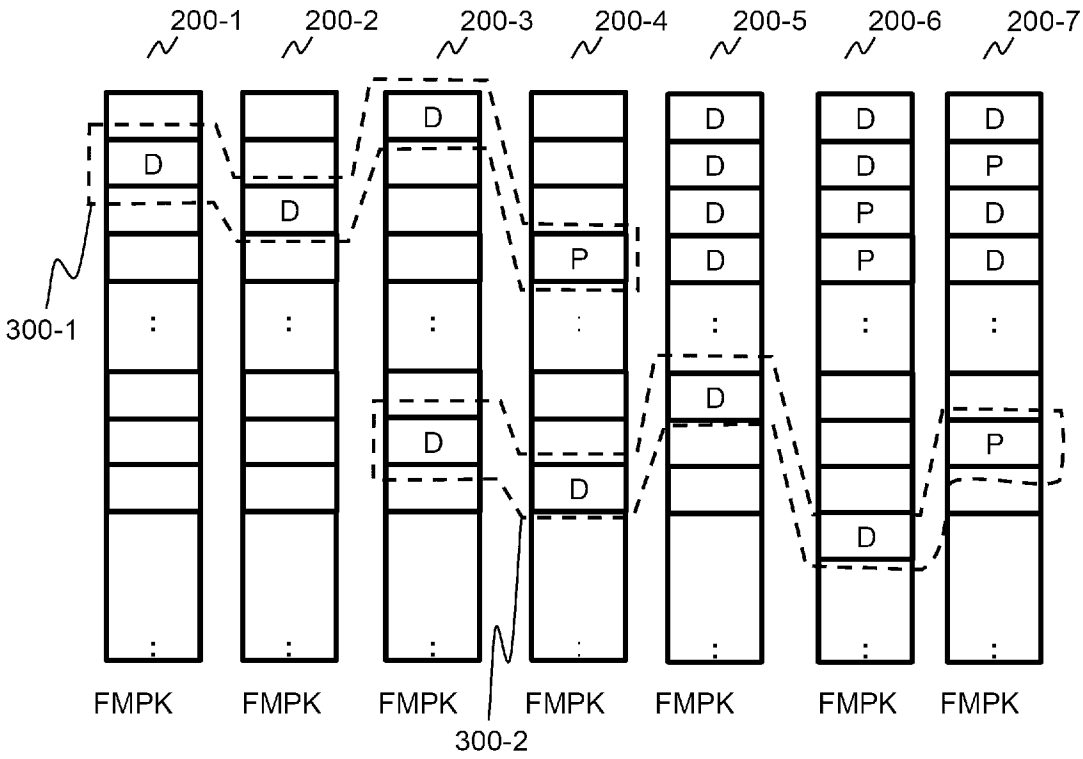
[図29]

図29


開始
RAIDグループを構成する
正常なデバイスを1つ選択
S400
Readリトライ回数(r)と
Writeリトライ回数(w)を0初期化
S410
バッファ領域の確保
圧縮Readコマンドを
正常なデバイスに送信
S420
応答情報を受信
S430
転送成功？
S440
NO
YES
rは一定以上?
S450
NO
YES
rに1加算
S460
①
圧縮中間パリティ演算コマンドを
復旧先デバイスに送信
S480'
応答情報を受信
S490
転送成功？
S500
NO
YES
wは一定以上？
S510
NO
YES
wに1加算
S520
①
全FMPKについて
処理終了?
S550
NO
YES
パリティコミット
コマンド発行
S600
エラー通知
RGの状態変更
障害復旧を中断
S540
終了

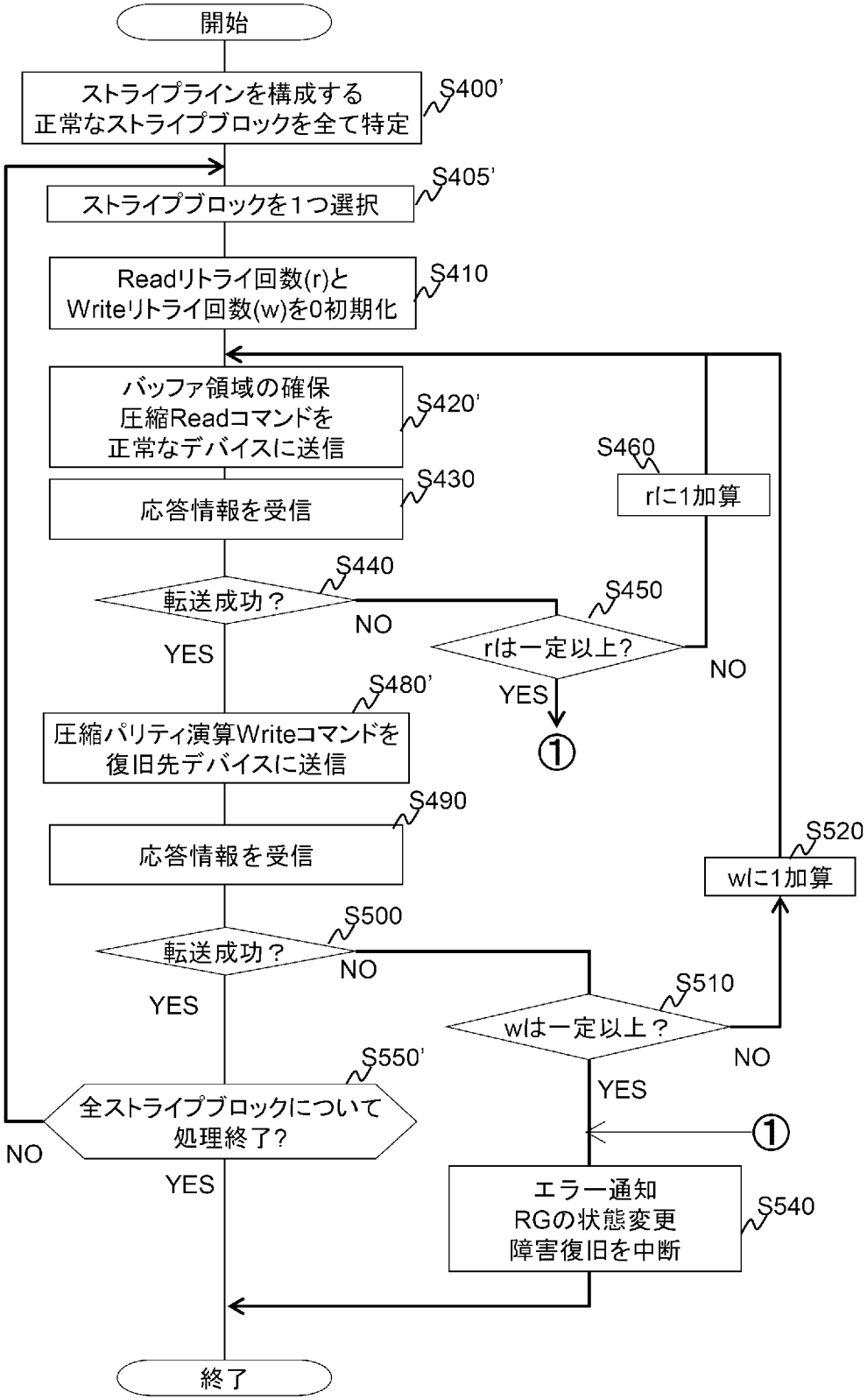
[図30]

図30



[図31]

図31


開始
ストライプラインを構成する
正常なストライプブロックを全て特定
S400'
ストライプブロックを1つ選択
S405'
Readリトライ回数(r)と
Writeリトライ回数(w)を0初期化
S410
バッファ領域の確保
圧縮Readコマンドを
正常なデバイスに送信
S420'
応答情報を受信
S430
転送成功？
S440
NO
YES
rは一定以上?
S450
NO
YES
S460
rに1加算
①
圧縮パリティ演算Writeコマンドを
復旧先デバイスに送信
S480'
応答情報を受信
S490
転送成功？
S500
NO
YES
wは一定以上？
S510
NO
YES
S520
wに1加算
全ストライプブロックについて
処理終了?
S550'
NO
YES
①
エラー通知
RGの状態変更
障害復旧を中断
S540
終了

[図32]

図32


3200
3201
オペコード
(圧縮パリティ演算Write要求を示す)
3202
Write開始オフセット(LBA)
3203
Writeサイズ(伸長時サイズ)
3204
バッファアドレス
3205
転送サイズ（圧縮時サイズ）
3206
(読み込み元FMPK)RAIDグループ内位置
3207
リード元データアドレス

# INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2014/070224

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 | 1996-2014 |
| Kokai Jitsuyo Shinan Koho | 1971-2014 | Toroku Jitsuyo Shinan Koho | 1994-2014 |

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<br>A | JP 2012-519319 A (Hitachi, Ltd.),<br>23 August 2012 (23.08.2012),<br>paragraphs [0001] to [0081]; fig. 1 to 15<br>& US 2011/0238885 A1 & US 2012/0297244 A1<br>& WO 2010/137178 A1 | 1,2,4,11<br>3,5-10,12-15 |
| Y<br>A | JP 5-011934 A (NEC Corp.),<br>22 January 1993 (22.01.1993),<br>paragraphs [0007] to [0022]; fig. 1 to 4<br>(Family: none) | 1,2,4,11<br>3,5-10,12-15 |
| Y | JP 7-311661 A (Fujitsu Ltd.),<br>28 November 1995 (28.11.1995),<br>paragraphs [0004] to [0101]; fig. 1 to 56<br>& US 5859960 A1 & DE 19515661 A<br>& KR 10-0226211 B | 4 |

☒ 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<br>03 October, 2014 (03.10.14) | Date of mailing of the international search report<br>14 October, 2014 (14.10.14) |
| Name and mailing address of the ISA/<br>Japanese Patent Office<br><br>Facsimile No. | Authorized officer<br><br>Telephone No. |

**INTERNATIONAL SEARCH REPORT**

International application No.
PCT/JP2014/070224

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 10-254634 A (Hitachi, Ltd.), 25 September 1998 (25.09.1998), paragraphs [0039] to [0041]; fig. 9 (Family: none) | 11 |

国際調査報告

国際出願番号　ＰＣＴ／ＪＰ２０１４／０７０２２４

Ａ．　発明の属する分野の分類（国際特許分類（ＩＰＣ））

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（ＩＰＣ））

Int.Cl.　G06F3/06

最小限資料以外の資料で調査を行った分野に含まれるもの

| | |
|---|---|
| 日本国実用新案公報 | １９２２－１９９６年 |
| 日本国公開実用新案公報 | １９７１－２０１４年 |
| 日本国実用新案登録公報 | １９９６－２０１４年 |
| 日本国登録実用新案公報 | １９９４－２０１４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<br>Ａ | ＪＰ　２０１２－５１９３１９　Ａ（株式会社日立製作所）<br>２０１２．０８．２３，段落【０００１】－【００８１】，<br>図１－１５　＆　ＵＳ　２０１１／０２３８８５　Ａ１<br>＆　ＵＳ　２０１２／０２９７２４４　Ａ１<br>＆　ＷＯ　２０１０／１３７１７８　Ａ１ | 1，2，4，11<br>3，5-10，<br>12-15 |
| Ｙ<br>Ａ | ＪＰ　５－０１１９３４　Ａ（日本電気株式会社）<br>１９９３．０１．２２，段落【０００７】－【００２２】，図１－４<br>（ファミリーなし） | 1，2，4，11<br>3，5-10，<br>12-15 |

☑　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 | 国際調査報告の発送日 |
|---|---|
| ０３．１０．２０１４ | １４．１０．２０１４ |


国際調査機関の名称及びあて先
日本国特許庁（ＩＳＡ／ＪＰ）
郵便番号１００－８９１５
東京都千代田区霞が関三丁目４番３号

特許庁審査官（権限のある職員）　５Ｔ　３２４２
古河　雅輝
電話番号　０３－３５８１－１１０１　内線　３５６８

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 |
|---|---|---|
| Ｙ | ＪＰ　７－３１１６６１　Ａ（富士通株式会社）<br>１９９５．１１．２８，段落【０００４】－【０１０１】，<br>図１－５６　＆　ＵＳ　５８５９９６０　Ａ１<br>＆　ＤＥ　１９５１５６６１　Ａ<br>＆　ＫＲ　１０－０２２６２１１　Ｂ | 4 |
| Ｙ | ＪＰ　１０－２５４６３４　Ａ（株式会社日立製作所）<br>１９９８．０９．２５，段落【００３９】－【００４１】，図９<br>（ファミリーなし） | 11 |