

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

PCT

(10) 国際公開番号
WO 2009/008045 A1(51) 国際特許分類:
G06F 3/06 (2006.01)

(21) 国際出願番号: PCT/JP2007/063605

(22) 国際出願日: 2007年7月6日 (06.07.2007)

(25) 国際出願の言語: 日本語

(26) 国際公開の言語: 日本語

(71) 出願人 (米国を除く全ての指定国について): 富士通株式会社 (FUJITSU LIMITED) [JP/JP]; 〒2118588 神奈川県川崎市中原区上小田中4丁目1番1号 Kanagawa (JP).

(72) 発明者; および

(75) 発明者/出願人 (米国についてのみ): 野田 祐司 (NODA, Yuji) [JP/JP]; 〒9291192 石川県かほく市宇野気ヌ

98番地の2 株式会社P F U内 Ishikawa (JP). 中神剛 (NAKAGAMI, Tsuyoshi) [JP/JP]; 〒9291192 石川県かほく市宇野気ヌ98番地の2 株式会社P F U内 Ishikawa (JP).

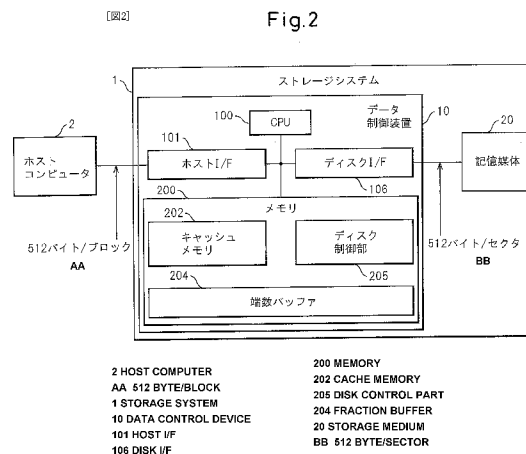
(74) 代理人: 青木 篤, 外(AOKI, Atsushi et al.); 〒1058423 東京都港区虎ノ門三丁目5番1号 虎ノ門37森ビル青和特許法律事務所 Tokyo (JP).

(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, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK,

[続葉有]

(54) Title: STORAGE SYSTEM DATA CONTROL DEVICE AND METHOD, AND PROGRAM FOR THE STORAGE SYSTEM DATA CONTROL

(54) 発明の名称: ストレージシステムのデータ制御装置、方法およびそのためのプログラム



(57) Abstract: A data control device of a storage system which is connected with a host computer communicably and reads and writes data from and on a storage medium by the block. The data control device comprises a host I/F for communicating with the host, a memory for temporarily storing the data from the host, a fraction buffer which regards extra data as fraction data when the extra data exists in front of a leading block data in the range of a write command or a read command from the host and/or the extra data exists in rear of final block data in the range and which stores the fraction data, a disk control part which controls the storage medium for writing and reading second-size data obtained by adding a check code to first-size data received from the host in and from the storage medium having a physical constitution of the first size, and a disk I/F which enables writing and reading data in and from the storage medium.

(57) 要約: ホストコンピュータと通信可能に接続され、記憶媒体にブロック単位でデータを読み書きするストレージシステムのデータ制御装置は、ホストと通信するホストI/Fと、ホストからのデータを一時記憶するメモリと、ホストからのライトコマンドまたはリードコマンドの範囲の先頭ブロックデータの前方に余分なデータがあり、および/または前記範囲の最終ブロックデータの後方に余分なデータが有るとき、その余分なデータを端数データと呼び、端数データを記憶する端数バッファと、ホストから受けた第1サイズのデータにチェックコードを付加した第2サイズの

[続葉有]



SL, SM, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US,
UZ, VC, VN, ZA, ZM, ZW.

CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE,
IS, IT, LT, LU, LV, MC, MT, NL, PL, PT, RO, SE, SI, SK,
TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW,
ML, MR, NE, SN, TD, TG).

(84) 指定国 (表示のない限り、全ての種類の広域保護が可
能): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD,
SL, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY,
KG, KZ, MD, RU, TJ, TM), ヨーロッパ (AT, BE, BG,

添付公開書類:
— 国際調査報告書

明 細 書

ストレージシステムのデータ制御装置、方法およびそのためのプログラム 技術分野

[0001] 本発明は、コンピュータが扱うデータを保存するストレージシステムに関し、特にブロックアクセスのストレージシステムのデータ制御装置、方法およびそのためのプログラムに関する。

背景技術

[0002] 従来のストレージシステムにおいて、ストレージシステム内部で扱うデータの1ブロックは520バイト(データ長512バイト+チェックコード長8バイト)で、そのデータを記憶する記憶媒体が1ブロックにつき512バイトの物理構成のディスクであった場合、ストレージシステム内部で扱うデータの1ブロックのデータ長と記憶媒体の1ブロックのデータ長とが異なるためストレージシステム内部の1ブロック分のデータを直接記憶媒体の1ブロックに格納することができなかった。

[0003] 特許文献1には、データを扱う単位の第1ブロックとデータとチェックコードからなる第2ブロックとの間で第1ブロックのサイズと第2ブロックのサイズとの公倍数分のサイズを有する第3ブロックの単位でディスクにデータを書込むことで隙間をなくし通信速度を向上させる発明が開示されている。

[0004] しかしながら、この発明は公倍数分のサイズに満たないデータをディスクに書込むとき、つまり第3ブロックのサイズ(第1ブロックサイズ $\times$ N)より小さいブロックのサイズ(第1ブロックサイズ $\times$ M)のデータをディスクに書込むときに不足している第1ブロックサイズ $\times$ (N-M)のデータをキャッシュメモリから読出して第1ブロックサイズ $\times$ N(=M+(N-M))のデータをディスクに書込む。したがって、不足分だけ余分なデータの書込み処理を行う必要が生じその分だけ余分な処理時間を要する問題がある。この問題を解決するには、ストレージシステム内部の1ブロック分のデータを直接記憶媒体の1ブロックに格納することが望ましい。このことを実現するには、以下のような方式が考えられる。

[0005] 図1Aは1ブロック分のデータを直接記憶媒体の1ブロックに格納する第1の方式に

よる記憶媒体のデータマップを示す図である。

第1の方式:データ(DATA)とチェックコード(CC)でそれぞれ1ブロックずつ格納する方式

この方式は、2ブロックごとに隙間(GAP)が発生し、読み込み時と書き込み時に隙間(GAP)部分のデータ転送が必要となるため通信速度が低下しかつ記憶容量が半減するという欠点がある。

[0006] 図1Bは1ブロック分のデータを直接記憶媒体の1ブロックに格納する第2の方式による記憶媒体のデータマップを示す図である。

第2の方式:データ(DATA)とチェックコード(CC)を隙間無く連続的に512バイトのブロックを跨らせる方式

この方式は、記憶媒体にデータを書込む際に、ブロック跨りの端数のデータ、例えば3ブロック分のデータ(512バイト)とCC(8バイト)を書込むとき第4ブロック目に端数(FRACTION)のデータの領域が生じるのでその端数のデータの読み込み処理を予めしておかないと書き込み処理ができず、このための読み込み処理に要する時間だけ書き込み処理に時間がかかるという欠点がある。

[0007] 特許文献1:特開2006-195851号公報(明細書の[請求項1]、[請求項8]、[請求項12]、[請求項14]、段落番号[0005]～[0008]、[0012]～[0040]、図面の[図1]～[図6]参照)。

発明の開示

発明が解決しようとする課題

[0008] 本発明は、上記第2の方式における書き込み処理に時間を要する問題を解決するストレージシステムのデータ制御装置、方法およびそのためのプログラムを提供することを目的とする。

本発明はまた、ディスクアレイ技術を用いたブロックアクセスのストレージシステムのデータ制御装置、方法およびそのためのプログラムを提供することを他の目的とする。

課題を解決するための手段

[0009] 上記目的を達成する本発明のストレージシステムのデータ制御装置は、ホストコン

ピュータと通信可能に接続され、記憶媒体にブロック単位でデータを読み書きするストレージシステムのデータ制御装置であって、前記データ制御装置は、前記ホストコンピュータと通信するホストI/Fと、前記ホストコンピュータからのデータを一時記憶するメモリと、前記ホストコンピュータからのライトコマンドまたはリードコマンドの範囲の先頭ブロックデータの前方に余分なデータが有り、および／または前記範囲の最終ブロックデータの後方に余分なデータが有るとき、前記余分なデータを端数データと呼び、該端数データを記憶する端数バッファと、前記ホストコンピュータから受けた第1サイズのデータにチェックコードを付加した第2サイズのデータを、物理構成が前記第1サイズの前記記憶媒体との間で書込みおよび読込みするよう制御するディスク制御部と、前記記憶媒体との間でデータを書込みおよび読込み可能なディスクI/Fと、を備えることを特徴とする。

- [0010] 上記ストレージシステムのデータ制御装置において、前記ディスク制御部は、前記ライトコマンドの範囲のデータを記憶媒体に書込む手段と、前記リードコマンドの範囲のデータを前記記憶媒体から読込み、前記メモリに記憶する手段と、前記メモリに記憶した前記リードコマンドの範囲のデータからチェックコードを消去して前記ホストコンピュータに転送する手段と、を備える。
- [0011] 上記ストレージシステムのデータ制御装置において、前記第1サイズのデータは512バイトであり、前記チェックコードは8バイトであり、前記第2サイズのデータは520バイトである。
- [0012] 上記ストレージシステムのデータ制御装置において、前記ホストコンピュータからのライトコマンドの範囲のデータサイズが前記第1サイズと前記第2サイズの公倍数のときで端数データが無しのとき、前記範囲のデータを前記記憶媒体のバウンダリからバウンダリまでに隙間なく書込む。
- [0013] 上記ストレージシステムのデータ制御装置において、RAID1またはRAID10方式のRAID制御手段を備え、リード処理時に、リードコマンドの範囲の先頭／最終ブロックが端数となった場合、それぞれ1ブロック分のデータを保持し、ライト処理時に、ライトコマンドの範囲の先頭／最終ブロックが端数となった場合、先のRD処理時に端数のデータを保持し、その後のWT処理時に1ブロック分のデータから端数データを

除いた残りの端数データを新たにバッファにコピーし、最終的に1ブロック分のデータを保持する。

[0014] 上記ストレージシステムのデータ制御装置において、RAID5またはRAID6方式のRAID制御手段を備え、リード処理時に、リードコマンドの範囲の先頭／最終ブロックが端数となった場合、それぞれ1ブロック分のデータを保持し、旧データ／旧Parity読み込み時に、端数のデータを読み込み、旧データと新データをEXCLUSIVEORすることにより新Parityを取得し、端数のデータを保持し、新データ／新Parity書き込み時に、1ブロック分のデータから端数データを除いた残りの端数データを新たにバッファへコピーし、最終的に1ブロック分のデータを保持する。

[0015] 上記目的を達成する本発明のストレージシステムのデータ制御方法は、ホストコンピュータから受領するデータを記憶媒体に保存するストレージシステムのデータ制御方法であって、前記ホストコンピュータから受けた第1サイズのデータにチェックコードを付加して第2サイズのデータを生成するステップと、物理構成が前記第1サイズである前記記憶媒体との間で前記第2サイズのデータで書き込みおよび読み込みをするステップと、を備えることを特徴とする。

[0016] 上記ストレージシステムのデータ制御方法において、前記ホストコンピュータからのライトコマンドまたはリードコマンドの範囲の先頭ブロックデータの前方に余分なデータが有り、および／または前記範囲の最終ブロックデータの後方に余分なデータが有るとき、前記余分なデータを端数データと呼び、該端数データを端数バッファに記憶するステップと、前記ライトコマンドの範囲のデータを記憶媒体に書き込むステップと、前記リードコマンドの範囲のデータを記憶媒体から読み込み、メモリに記憶するステップと、前記メモリに記憶した前記リードコマンドの範囲のデータからチェックコードを消去して前記ホストコンピュータに転送するステップと、を備える。

[0017] 上記目的を達成する本発明のストレージシステムのデータ制御を行うプログラムは、上記ストレージシステムのデータ制御方法をコンピュータに実行させることを特徴とする。

図面の簡単な説明

[0018] [図1A]1ブロック分のデータを直接記憶媒体の1ブロックに格納する第1の方式による

記憶媒体のデータマップを示す図である。

[図1B]1ブロック分のデータを直接記憶媒体の1ブロックに格納する第2の方式による記憶媒体のデータマップを示す図である。

[図2]本発明によるストレージシステムのデータ制御装置の一実施形態のブロック構成図である。

[図3]記憶媒体に格納されるデータのマップを示す図である。

[図4]図2に示すデータ制御装置のライト処理の前半フローチャートである。

[図5]図2に示すデータ制御装置のライト処理の後半フローチャートである。

[図6]図2に示すデータ制御装置のリード処理のフローチャートである。

[図7]本発明を適用したRAID制御方式のデータ制御装置の一実施形態のブロック構成図である。

[図8]図7に示すデータ制御装置の520/512変換制御部における動作のフローチャートである。

[図9]ハッシュテーブルを示す図である。

[図10]バッファコントロールブロック(BCB)を示す図である。

[図11A]制御方式RAID0を示す図である。

[図11B]制御方式RAID10を示す図である。

[図11C]制御方式RAID5を示す図である。

[図11D]制御方式RAID6を示す図である。

符号の説明

- [0019] 1、71 ストレージシステム
 2、72 ホストコンピュータ
 10、70 データ制御装置
 20、80 記憶媒体
 100 CPU
 200 メモリ
 101、701 ホストI/F
 202、702 キャッシュメモリ

204、707 端数バッファ

205、705 ディスク制御部

106、706 ディスクI/F

703 RAID制御部

704 520/512変換制御部

発明を実施するための最良の形態

[0020] 以下、添付図面を参照しつつ本発明の実施形態について詳細に説明する。

図2は本発明によるストレージシステムのデータ制御装置の一実施形態のブロック構成図である。図2に示すストレージシステム1は、データ制御装置10と記憶媒体20を有し、ホストコンピュータ(以下、ホストと記す)2と通信可能に接続されている。記憶媒体20は1セクタ512バイトの物理構成を有する。データ制御装置10はCPU100とホスト2と通信するホストI/F101、記憶媒体20との間でデータの書込みおよび読み込みを行なうディスクI/F106およびメモリ200からなり、メモリ200は、ホスト2からのデータを一時記憶するキャッシュメモリ202、端数データを記憶する端数バッファ204、およびホスト2から受けた512バイトのデータを520バイトのデータで扱い、512バイトのデータで記憶媒体20との間で読み込みおよび書込みを行なうよう制御するディスク制御部205からなる。ここで、ホストコンピュータ2からのライトコマンドの範囲の先頭ブロックデータの前方にバウンダリまで余分なデータが有り、および/またはその範囲の最終ブロックデータの後方にバウンダリまで余分なデータが有るとき、これらの余分なデータを端数データと呼ぶ。ここで、バウンダリとは、記憶媒体の物理的境界を言う。

[0021] 図3は記憶媒体に格納されるデータのマップを示す図である。図3の最上段のLBA(Logical Block Address) #0から最下段のLBA #67に向けて、下記のようにデータが格納される。図3において、左端および右端がバウンダリである。

本発明では、先に説明した第2の方式を採用し、アクセス性能と記憶容量を稼ぐために、520バイトのデータを連続的(隙間無く)に512バイトのブロックを跨らせて配置する。

LBA #0: 第1ブロックデータ(以下、ブロックと記す)

LBA # 1: 第1ブロックのチェックコード(以下、CCと記す) + 第2ブロック

LBA # 2: 第2ブロック + 第2ブロックのCC + 第3ブロック

:

:

LBA # 64: 第63ブロック + 第63ブロックのCC + 第64ブロック

LBA # 65: 第64ブロック + 第64ブロックのCC

LBA # 66: 第65ブロック

[0022] 図3から判るように、64個のブロックデータをディスクに書込むために65セクタを使用している。また、図3から判るように、第2ブロックが先頭ブロックの場合、第1ブロックのCCが端数データであり、第3ブロックが先頭ブロックの場合、第2ブロックデータの後端と第2ブロックのCCが端数データであり、第63ブロックが最終ブロックの場合、第63ブロックのCCと第64ブロックの先端が端数データであり、第64ブロックが最終ブロックの場合、第64ブロックのCCが端数データである。ここで、CCは8バイトのデータ長であり、その内2バイトはCRC (Cyclic Redundancy Check: 巡回冗長検査) であり、4バイトは論理アドレス (Logical Address) であり、残りの2バイトはRAID No. またはVoL. No. であり、これらNo. はホスト2から見たディスクを特定するものである。

[0023] 図4は図2に示すデータ制御装置のライト処理の前半フローチャートである。

ステップS1で、ホスト2からライトコマンドを受領する。

ステップS2で、メモリ200内にキャッシュメモリ(以下、キャッシュと記す)202の利用エリアを獲得する。

ステップS3で、ホスト2から受領したブロック単位のデータに対しブロック毎にCCを付加してキャッシュ202に記憶する。

ステップS4で、ホスト2にデータ受領の完了を通知する。

[0024] ステップS5で、受領したデータのブロックアドレスが連続か否かを判定し、判定結果がYESのときはステップS6に進み、判定結果がNOのときはステップS11に進む。

ステップS6で、受領したデータにステップS8で蓄積した前回までに受領したデータを加える。ただし、第1回目の処理では蓄積された受領データは無い。

ステップS7で、受領したデータのサイズが1ブロックのデータサイズ512バイトと、1

ブロックのデータにチェックコードCCのデータを加算したデータサイズ520バイトとの公倍数か否かを判定し、判定結果がYESのときはステップS11に進み、判定結果がNOのときはステップS8に進む。

ステップS8で、前回までに受領した蓄積されたデータに今回受領したデータを蓄積する。

ステップS9で、所定時間経過したか否かを判定し、判定結果がYESのときはステップS11に進み、判定結果がNoのときはステップS1に戻り、ステップS1～S9を繰り返し実行する。ここで、所定時間とは、ホスト2からのライトコマンドが発行されなくなるまでの時間である。

[0025] 図5は図2に示すデータ制御装置のライト処理の後半フローチャートである。

ステップS11で、受領したデータの書込先ディスクNo. とロジカルブロックアドレス(LBA) No. を算出する。ステップS5で、受領したブロックのアドレスが不連続の場合、ブロック毎にステップS11～S18を実行する。

ステップS12で、受領したデータ、すなわちライトコマンド範囲のデータの先頭ブロックおよび／または最終ブロックに端数データが有るか無いかを判定し、判定結果がYESのときはステップS13に進み、判定結果がNOのときはステップS17に進む。例えば、受領したデータのブロックアドレスが#0～#5であり、ブロックアドレス#0のデータがCCを含まないデータのみであり、ライトコマンド範囲のデータのブロックアドレスが#1～#3であった場合、先頭ブロックアドレス#1および最終ブロックアドレス#3は共にCCを含む、すなわちブロックの先頭に端数データを含むので判定結果がYESとなる。

[0026] ステップS13で、端数データ対応の登録済端数バッファが端数バッファ204内に存在するか否かを判定し、判定結果がYESのときはすでに登録済みと判定しステップS17に進み、判定結果がNOのときは未登録と判定しステップS14に進む。個々のバッファは後述するバッファコントロールブロックBCBで管理される。

ステップS14で、端数バッファは未登録と判定されたので端数データを格納するための端数バッファを獲得する。

[0027] ステップS15で、書込み先ディスクNo. のディスクのLBANo. のセクタから端数デ

ータを読み込む。ステップS12の具体例で説明すると、図3に示すLBA # 1の第1ブロックのCCとLBA # 4の第4ブロックのCCその後方の第5ブロックのデータ前方が端数データとなる。

ステップS16で、ステップS14で算出した端数バッファにステップS15で読込んだ端数データを書込む。

- [0028] ステップS17で、先頭ブロックおよび／または最終ブロックのデータと端数バッファに書込まれた端数データとから書込み先ディスクの1セクタに書込む1ブロック分のデータを作成する。

ステップS18で、ステップS17で作成したデータを書込み先ディスクのライトコマンドで指定された書込先ディスクNo. のLBANo. のセクタに書込む。

- [0029] 図6は図2に示すデータ制御装置のリード処理のフローチャートである。

ステップS21で、ホスト2からリードコマンドを受領する。

ステップS22で、リードコマンドから読込元ディスクNo. とロジカルブロックアドレス(LBA) No. を算出する。

ステップS23で、キャッシュメモリ202の利用エリアを獲得する。

- [0030] ステップS24で、リードコマンドの指定する読込み範囲のデータ内に端数データが有るか無いかを判定し、判定結果がYESのときはステップS25に進み、判定結果がNOのときはステップS26に進む。

ステップS25で、読込み範囲のデータ内の端数データを格納する端数バッファを獲得する。

ステップS26で、ステップS22で算出した読込元ディスクNo. とLBANo. のセクタからデータを読み込み、端数データが有る(ステップS24で判定結果がYESの)ときはその端数データを端数バッファに書込み、端数以外のデータはキャッシュメモリに記憶する。

ステップS27で、ステップS26で記憶したキャッシュメモリのデータからチェックコードを消去したデータをホスト2に転送する。

- [0031] 図7は本発明を適用したRAID制御方式のデータ制御装置の一実施形態のブロック構成図である。図7に示すストレージシステム71は、データ制御装置70と記憶媒体

80を有し、ホストコンピュータ(以下、ホストと記す)72と通信可能に接続されている。記憶媒体80は1セクタ512バイトの物理構成を有する。データ制御装置70は、ホスト2と通信するホストI/F701、ホスト2からのデータを不図示のキャッシュメモリに一時記憶するよう制御するキャッシュ制御部702、後述するRAID制御部703、520/512変換制御部704、ホスト2から受けた512バイトのデータを520バイトのデータで扱い、512バイトのデータで記憶媒体80との間で読込みおよび書込みを行なうよう制御するディスク制御部705、記憶媒体80との間でデータの書込みおよび読込みを行なうディスクI/F706および端数データを記憶する端数バッファ707からなる。

[0032] 図7に示すRAID制御部703の処理について以下に説明する。

RAID制御部703はRAIDの形態RAID0/10/5/6に応じて以下に記すリード処理、ライト処理およびパトロール処理を実行する。

1. 端数データの保持条件について

リード処理:

RAID0/10/5/6: リード(RD)コマンドの範囲の先頭/最終ブロックが端数となった場合、それぞれ1ブロック分のデータを端数バッファに保持するよう管理する。

[0033] ライト処理:

RAID0/10: ライト(WT)コマンドの範囲の先頭/最終ブロックが端数となった場合、先のRD処理で端数のデータを端数バッファに保持するよう管理する。ここで、端数のデータとは、ホスト2からのライトコマンドの先頭ブロックの余分なデータと最終ブロックの余分なデータを意味する。

その後のWT処理時に1ブロック分のデータから端数データを除いた残りの端数データを新たにバッファにコピーし、最終的に1ブロック分のデータを保持管理する。

[0034] RAID5/6: RAID5の旧データ/旧Parity読込み時に、端数のデータを読込み、旧データと新データをEXCLUSIVEORすることにより新Parityを得る。端数のデータを保持し、新データ/新Parity書込み時に、1ブロック分のデータから端数データを除いた残りの端数データを新たにバッファへコピーし最終的に1ブロック分のデータを保持管理する。

[0035] Patrol処理:

RAID0/10/5/6: Patrolで実施するDiskのRDチェック時に、先頭ブロックと最後ブロックの片方の端数データのみを保持しているバッファがあった場合は、もう一方の端数データを保管し、その端数データを含む1ブロックのデータとして保持管理する。

[0036] 2. 端数データの消去条件について

以下の条件のときに消去する。

1) キャッシュ (Cache) 制御部からのHost-Write情報(ホストからライトデータを受信した情報)を元に、端数データ保持と端数データの領域に対するライトデータをホストから受信した場合、例えば、バッファコントロールブロック(BCB)で管理しているディスク対象の端数データを消去する。

2) 新たに端数データを読込む際にバッファが枯渇していた場合に、使用頻度の一番低い端数データを消去する。

[0037] 図8は図7に示すデータ制御装置の520/512変換制御部704における動作のフローチャートである。

ステップS81で、RAID制御部703から受領したデータの書込先ディスクNo. とロジカルブロックアドレス(LBA)No. を算出する。

Logical Block Address (LBA) 変換処理を実行する。すなわち、上位から520byte/sectorで要求されるため、Diskの物理アドレスへ(512/sector)へ変換する。

[0038] ステップS82で、受領したデータの先頭ブロックと最終ブロックに端数が有るか無いかを判定し、判定結果がYESのときはステップS83に進み、判定結果がNOのときはディスク制御部705の処理に移る。

ステップS83で、コマンドはライトWTかリードRDか判定し、RDであればステップS800に進み、WTであればステップS84に進む。

[0039] ステップS84で、RD-WTシーケンス処理を実行する。すなわち、上位からのライトコマンドで、先頭/最終ブロックの端数が発生した場合に以下のシーケンス処理を行なう。

1) 端数データのリード (Read)

2) 上位からのライトコマンドを512/secに拡張し、1) でReadした端数データと合わせて

ライト(Write)する。

[0040] ステップS800で、以下のように端数データを管理する。

- Readした端数データをハッシュテーブルで保持管理する。
- 上位からのRead/Writeコマンド時には、ハッシュテーブルに保持している端数データがHitしていないか否かをチェックする。
- 端数データのインバリデート処理を行なう。

[0041] ステップS801で、定期的にプリフェッチデモンを起動する。

プリフェッチデモンは、Cache制御部702が保持しているディスク(Disk)未書込みデータ(将来Diskへ書込まれるデータ)をチェックし、端数データが必要なものがある場合に、Diskアクセスの負荷を監視し、負荷が低いケースは端数データを予め読込む処理を実行する。

[0042] 図9はハッシュテーブルを示す図である。

ハッシュ(Disk No.とそのDiskのLBAを元に端数バッファのEntryを求める)

ハッシュテーブルには、BCB(Buffer Control Block)をリンクするポインタのみが格納されている。

ハッシュエントリが衝突した場合はBCBにBCBを待ち行列形式でリンクする。

[0043] ハッシュエントリの数を256とした場合、インデックス(Index)は、ディスクNo. (8bit)の下位4ビット(bit)+LBAの下位4ビット(bit)で表すことができる。

bit 7 4 3 0

Index | Disk No. の下位4ビット | LBAの下位4ビット

ハッシュエントリの衝突は、Indexが同じ、例えば0でも、Disk No. の上位ビットおよび/またはLBAの上位ビットは通常0であるがこれが1となると衝突が起きる。このようなIndexが同じ場合、上位ビット0対応のBCBと上位ビット1対応のBCBとをリンクする。

[0044] 図10はバッファコントロールブロック(BCB)を示す図である。バッファコントロールブロックには、キューイングポインタ(Queuing Pointer)、ディスクナンバ(Disk No.)、ディスクロジカルブロックアドレス(Disk LBA)、データバリッドフラグData Valid Flag、バッファアドレス(Buffer Address)およびバッファオフセット(Buffer Offset)を格納し、

バッファを管理する。

– Queuing Pointer:

ハッシュテーブル、または、BCBへのリンクポインタを格納

– Disk No.

Diskの論理番号を格納

[0045] – Disk LBA

Buffer内に保持しているデータのアドレス(DiskのLBA) を格納

– Data Valid Flag

Buffer内に保持しているデータの有効性(Fフラグ:Offsetより前半が有効、Bフラグ:Offsetより後半が有効)を示すBitのデータを格納

512バイト全てが有効な場合は、FフラグとBフラグの両方をセットして格納

[0046] – Buffer Address

バッファのアドレスを格納

– Buffer Offset

端数データのバウンダリをバッファの先頭からのバイトオフセットでセットしたデータを格納

[0047] 図11A、11B、11C、11Dは、それぞれ制御方式RAID0、RAID10、RAID5、RAID6を順に示す図である。

図11Aに示すように、制御方式RAID0は、複数のディスクにデータを分散し高速処理を図るものである。

図11Bに示すように、制御方式RAID10は、個々にミラーリングしたディスクグループ(上下の対)に対し上段のディスクグループと下段のディスクグループをそれぞれストライピングし、データ保護と高速処理を図るものである。

[0048] 図11Cに示すように、制御方式RAID5は、RAID0にチェック用のディスクを付加し、データ保護と高速処理を図るものである。

図11Dに示すように、制御方式RAID6は、RAID0にチェック用のディスクを2つ付加し、さらなるデータ保護と高速処理を図るものである。

[0049] 以上説明したように、本発明によれば、従来の問題、すなわち書込み処理に時間を

要する問題を解決するため、データを記憶媒体へ書込む際に発生するブロック跨りの端数データの読込み処理を、予め読込み処理時においてメモリ上に保持・管理することで、極力減らす制御を行い性能の向上を図ることができる。

[0050] また、大容量データの書込みやシーケンシャルな書込みに際しては、520バイトと512バイトの跨りが発生しないサイズ、すなわち520バイト×64ブロック(512バイト×65ブロック相当)までデータ(33280バイト)が溜まるのを待って記憶媒体のディスクへ書込むことで端数のデータが発生しないように制御する。

[0051] 本発明によれば、1ブロックが512バイト固定のディスクに対し、例えばRAID (Redundant Arrays of Inexpensive Disks) コントローラが扱う1ブロック520(512+8)バイト、うち8バイトはブロック毎のチェックコード(BCC)のデータを効率良くディスクへ書込むストレージシステムのデータ制御装置、方法およびそのためのプログラムを提供することができる。

[0052] 本発明のストレージシステムのデータ制御装置は、データ制御装置内で扱う全てのデータの各ブロック単位(512バイト)に、チェックコード(8バイト)を付加し管理・記憶することで、データの正当性を論理的に保証する機能をもっている。

[0053] また、本発明によれば、記憶媒体として1ブロック長が520バイトのチェックコード付専用のディスクではなく1ブロックのデータ長が512バイト固定である一般的なATA/SATA(Serial Advanced Technology Attachment)ディスクを採用でき、安価で高信頼という特徴をもったストレージシステムが提供できる。

請求の範囲

- [1] ホストコンピュータと通信可能に接続され、記憶媒体にブロック単位でデータを読み書きするストレージシステムのデータ制御装置であって、
- 前記データ制御装置は、
- 前記ホストコンピュータと通信するホストI/Fと、
- 前記ホストコンピュータからのデータを一時記憶するメモリと、
- 前記ホストコンピュータからのライトコマンドまたはリードコマンドの範囲の先頭ブロックデータの前方に余分なデータが有り、および／または前記範囲の最終ブロックデータの後方に余分なデータが有るとき、前記余分なデータを端数データと呼び、該端数データを記憶する端数バッファと、
- 前記ホストコンピュータから受けた第1サイズのデータにチェックコードを付加した第2サイズのデータを、物理構成が前記第1サイズの前記記憶媒体との間で書込みおよび読み取りを行うよう制御するディスク制御部と、
- 前記記憶媒体との間でデータを書込みおよび読み取り可能なディスクI/Fと、
- を備えることを特徴とするストレージシステムのデータ制御装置。
- [2] 前記ディスク制御部は、
- 前記ライトコマンドの範囲のデータを記憶媒体に書込む手段と、
- 前記リードコマンドの範囲のデータを前記記憶媒体から読み取り、前記メモリに記憶する手段と、
- 前記メモリに記憶した前記リードコマンドの範囲のデータからチェックコードを消去して前記ホストコンピュータに転送する手段と、
- を備える請求項1に記載のストレージシステムのデータ制御装置。
- [3] 前記第1サイズのデータは512バイトであり、
- 前記チェックコードは8バイトであり、
- 前記第2サイズのデータは520バイトである、
- 請求項1または2に記載のストレージシステムのデータ制御装置。
- [4] 前記ホストコンピュータからのライトコマンドの範囲のデータサイズが前記第1サイズと前記第2サイズの公倍数のときで端数データが無しのとき、前記範囲のデータを前

記記憶媒体のバウンダリからバウンダリまでに隙間なく書込む、
請求項3に記載のストレージシステムのデータ制御装置。

- [5] RAID1またはRAID10方式のRAID制御手段を備え、
リード処理時に、リードコマンドの範囲の先頭／最終ブロックが端数となった場合、
それぞれ1ブロック分のデータを保持し、
ライト処理時に、ライトコマンドの範囲の先頭／最終ブロックが端数となった場合、先のRD処理時に端数のデータを保持し、
その後のWT処理時に1ブロック分のデータから端数データを除いた残りの端数データを新たにバッファにコピーし、最終的に1ブロック分のデータを保持する、
請求項1乃至4の何れか1項に記載のストレージシステムのデータ制御装置。
- [6] RAID5またはRAID6方式のRAID制御手段を備え、
リード処理時に、リードコマンドの範囲の先頭／最終ブロックが端数となった場合、
それぞれ1ブロック分のデータを保持し、
旧データ／旧Parity読み込み時に、端数のデータを読み込み、旧データと新データをEXCLUSIVEORすることにより新Parityを取得し、
端数のデータを保持し、新データ／新Parity書き込み時に、1ブロック分のデータから端数データを除いた残りの端数データを新たにバッファへコピーし、最終的に1ブロック分のデータを保持する、
請求項1乃至4の何れか1項に記載のストレージシステムのデータ制御装置。
- [7] ホストコンピュータから受領するデータを記憶媒体に保存するストレージシステムのデータ制御方法であって、
前記ホストコンピュータから受けた第1サイズのデータにチェックコードを付加して第2サイズのデータを生成するステップと、
物理構成が前記第1サイズである前記記憶媒体との間で前記第2サイズのデータで書き込みおよび読み込みをするステップと、
を備えることを特徴とするストレージシステムのデータ制御方法。
- [8] 前記ホストコンピュータからのライトコマンドまたはリードコマンドの範囲の先頭ブロックデータの前方に余分なデータが有り、および／または前記範囲の最終ブロックデ

ータの後方に余分なデータが有るとき、前記余分なデータを端数データと呼び、該端数データを端数バッファに記憶するステップと、

前記ライトコマンドの範囲のデータを記憶媒体に書込むステップと、

前記リードコマンドの範囲のデータを記憶媒体から読み込み、メモリに記憶するステップと、

前記メモリに記憶した前記リードコマンドの範囲のデータからチェックコードを消去して前記ホストコンピュータに転送するステップと、

を備える請求項7に記載のストレージシステムのデータ制御方法。

- [9] ホストコンピュータから受領するデータを記憶媒体に保存するストレージシステムのデータ制御を行うプログラムであって、

前記ホストコンピュータから受けた第1サイズのデータにチェックコードを付加して第2サイズのデータを生成するステップと、

物理構成が前記第1サイズである前記記憶媒体との間で前記第2サイズのデータで書き込みおよび読み込みをするステップと、

をコンピュータに実行させることを特徴とするプログラム。

- [10] 前記ホストコンピュータからのライトコマンドまたはリードコマンドの範囲の先頭ブロックデータの前方に余分なデータが有り、および／または前記範囲の最終ブロックデータの後方に余分なデータが有るとき、前記余分なデータを端数データと呼び、該端数データを端数バッファに記憶するステップと、

前記ライトコマンドの範囲のデータを記憶媒体に書込むステップと、

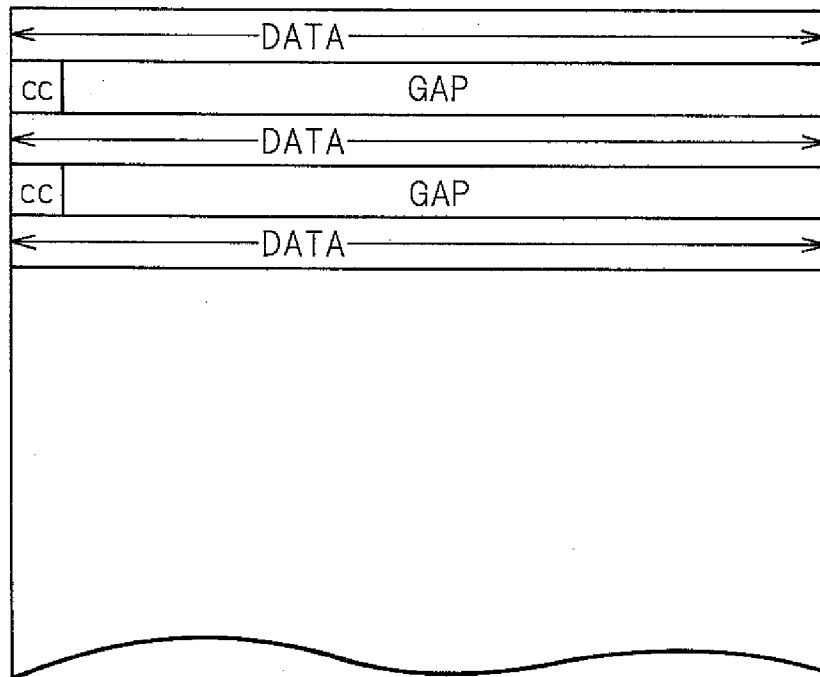
前記リードコマンドの範囲のデータを記憶媒体から読み込み、メモリに記憶するステップと、

前記メモリに記憶した前記リードコマンドの範囲のデータからチェックコードを消去して前記ホストコンピュータに転送するステップと、

をコンピュータに実行させることを特徴とする請求項9に記載のプログラム。

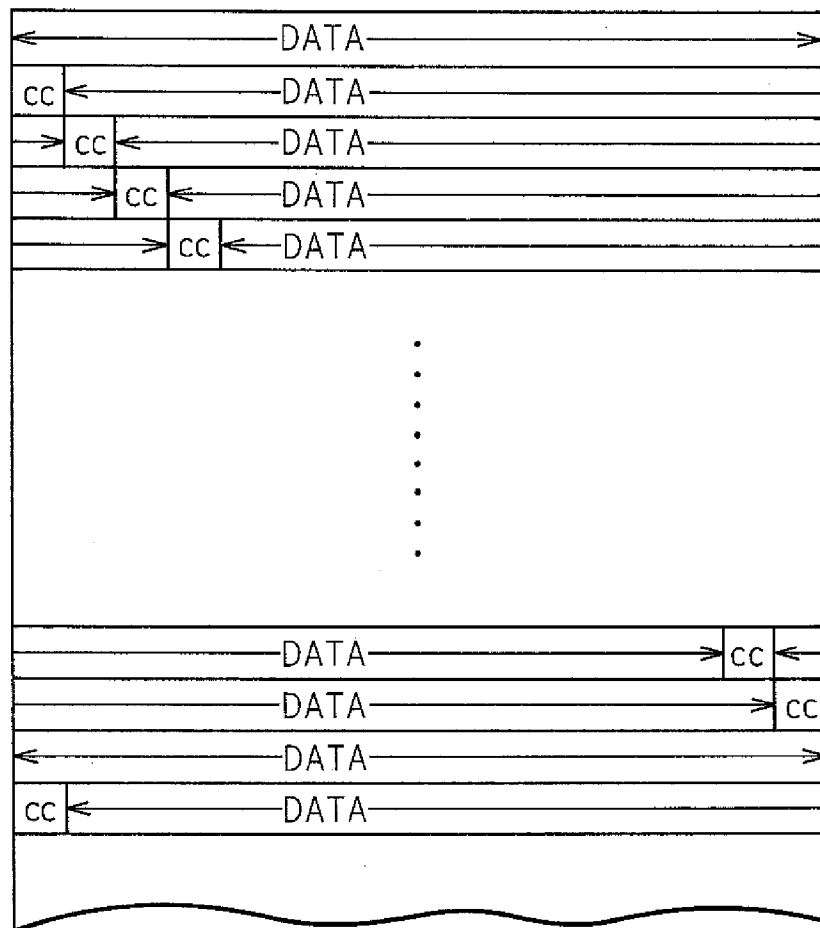
[図1A]

Fig. 1A



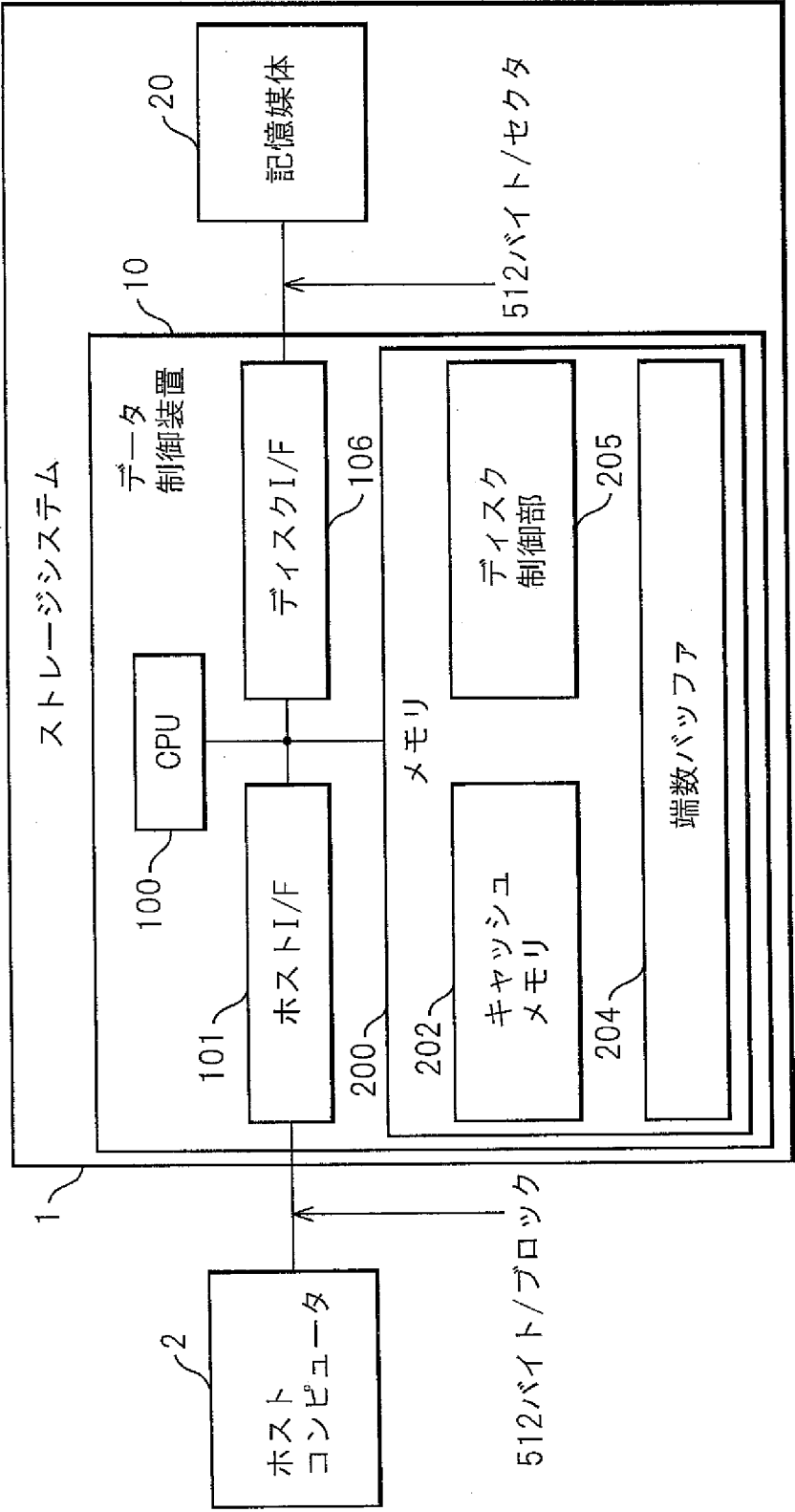
[図1B]

Fig. 1B



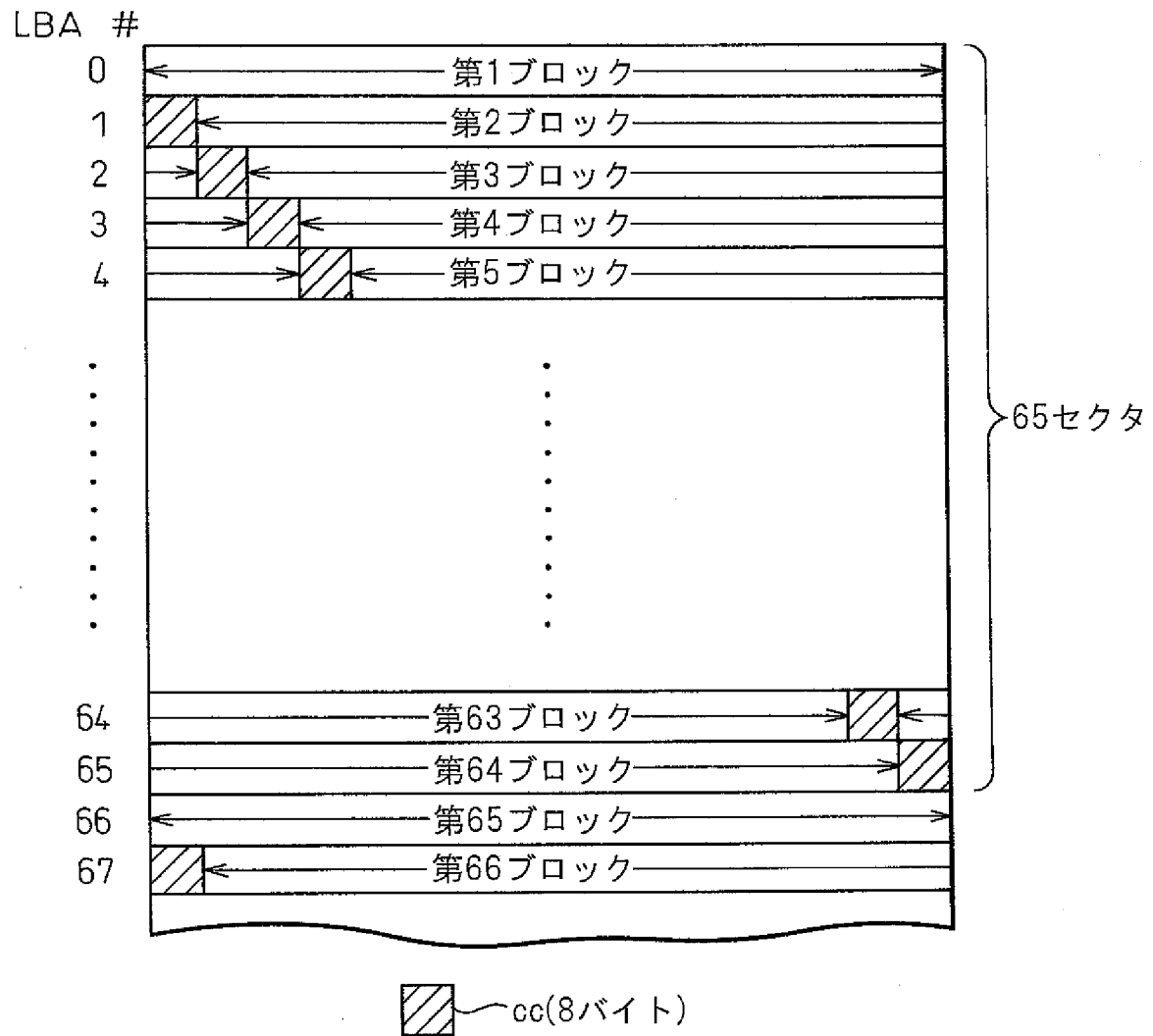
[図2]

Fig.2



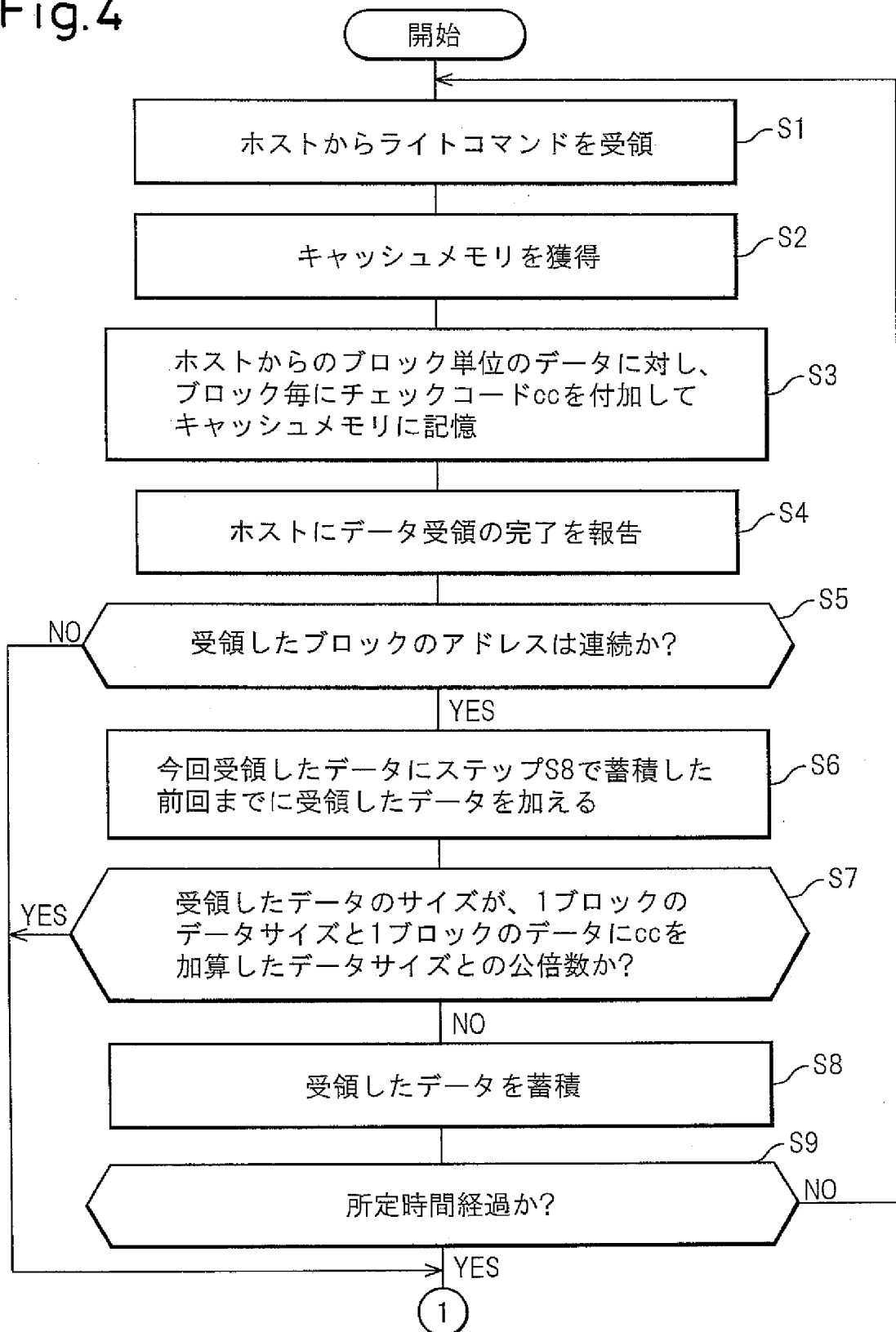
[図3]

Fig.3



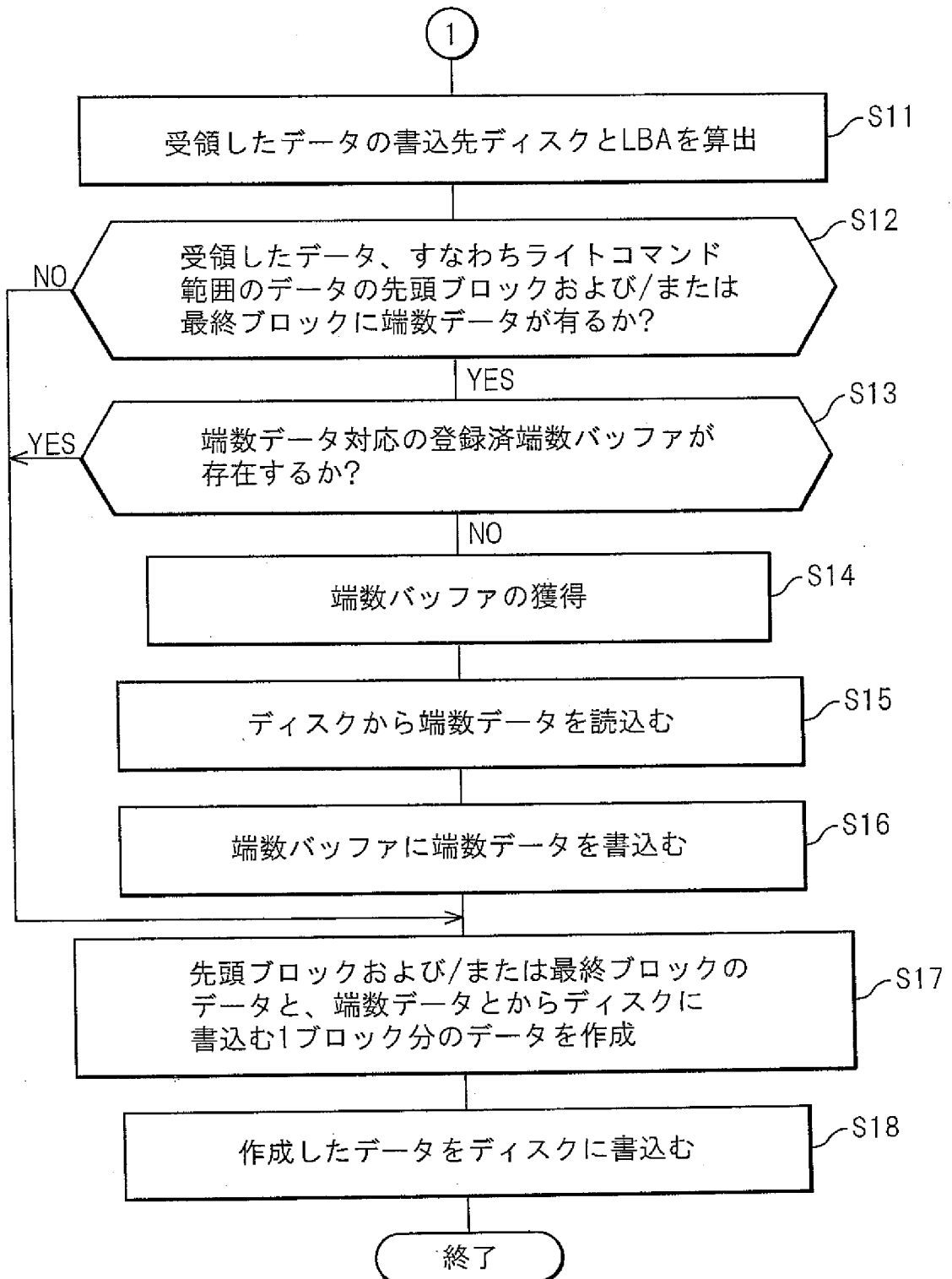
[図4]

Fig.4



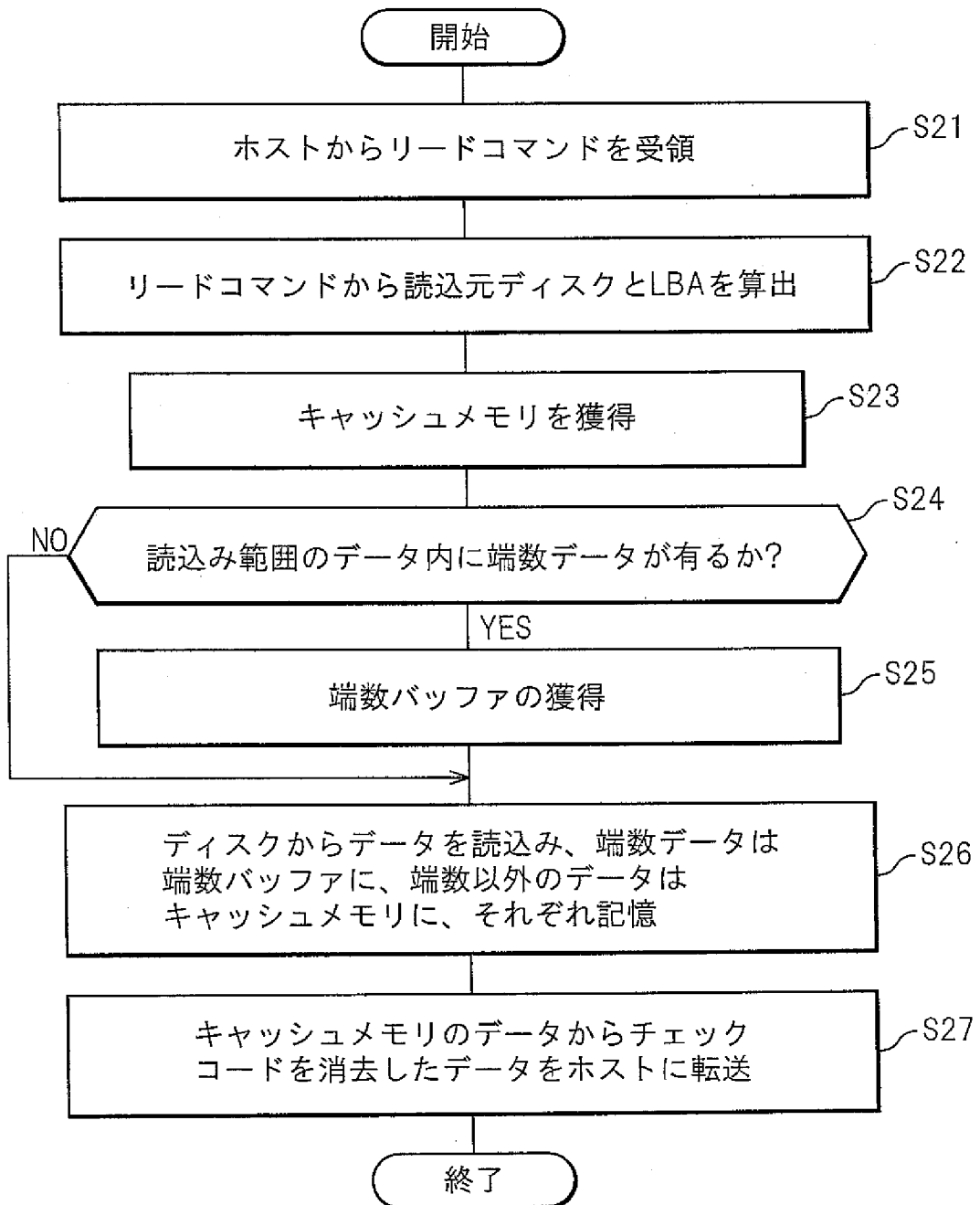
[図5]

Fig.5



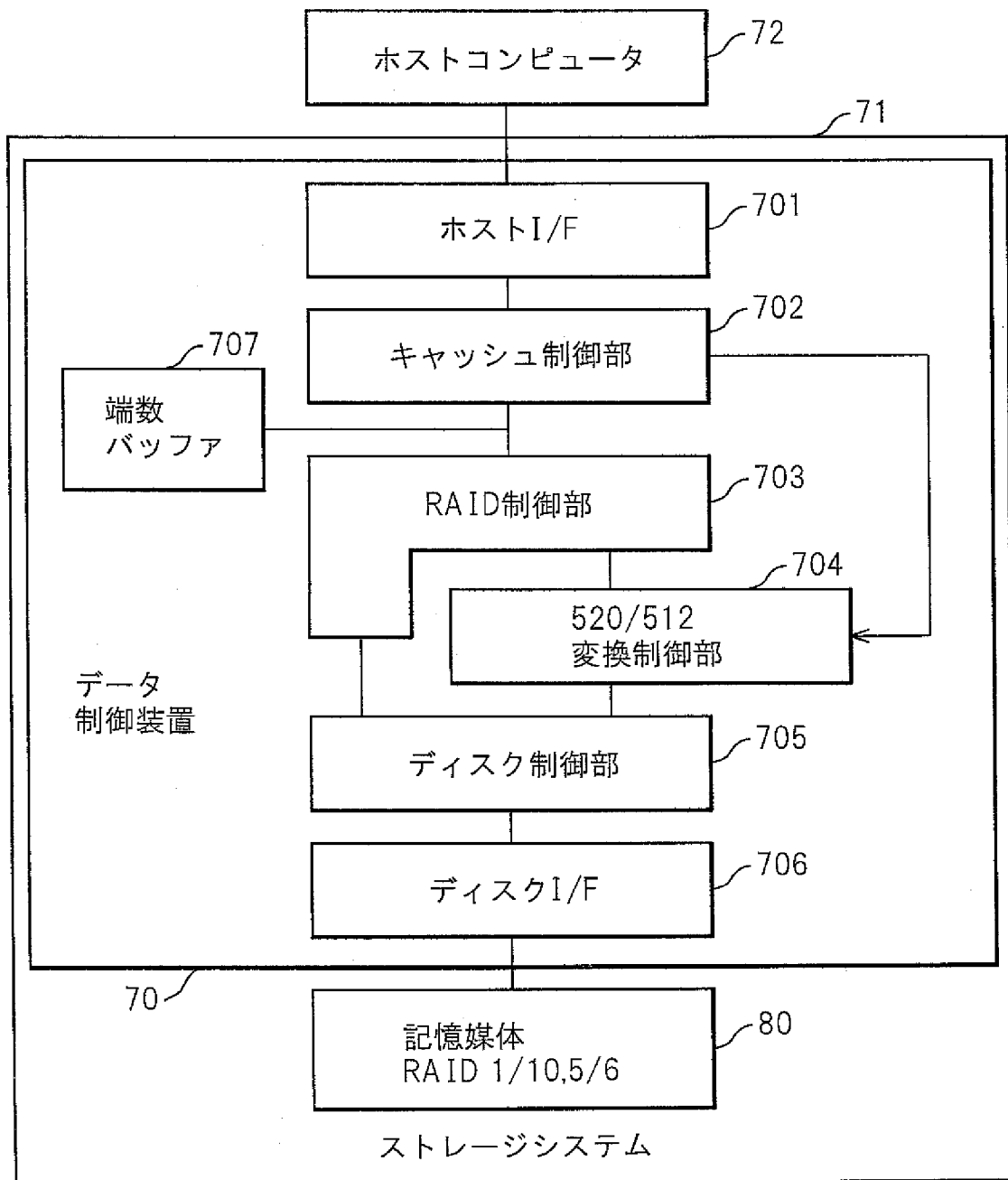
[図6]

Fig.6



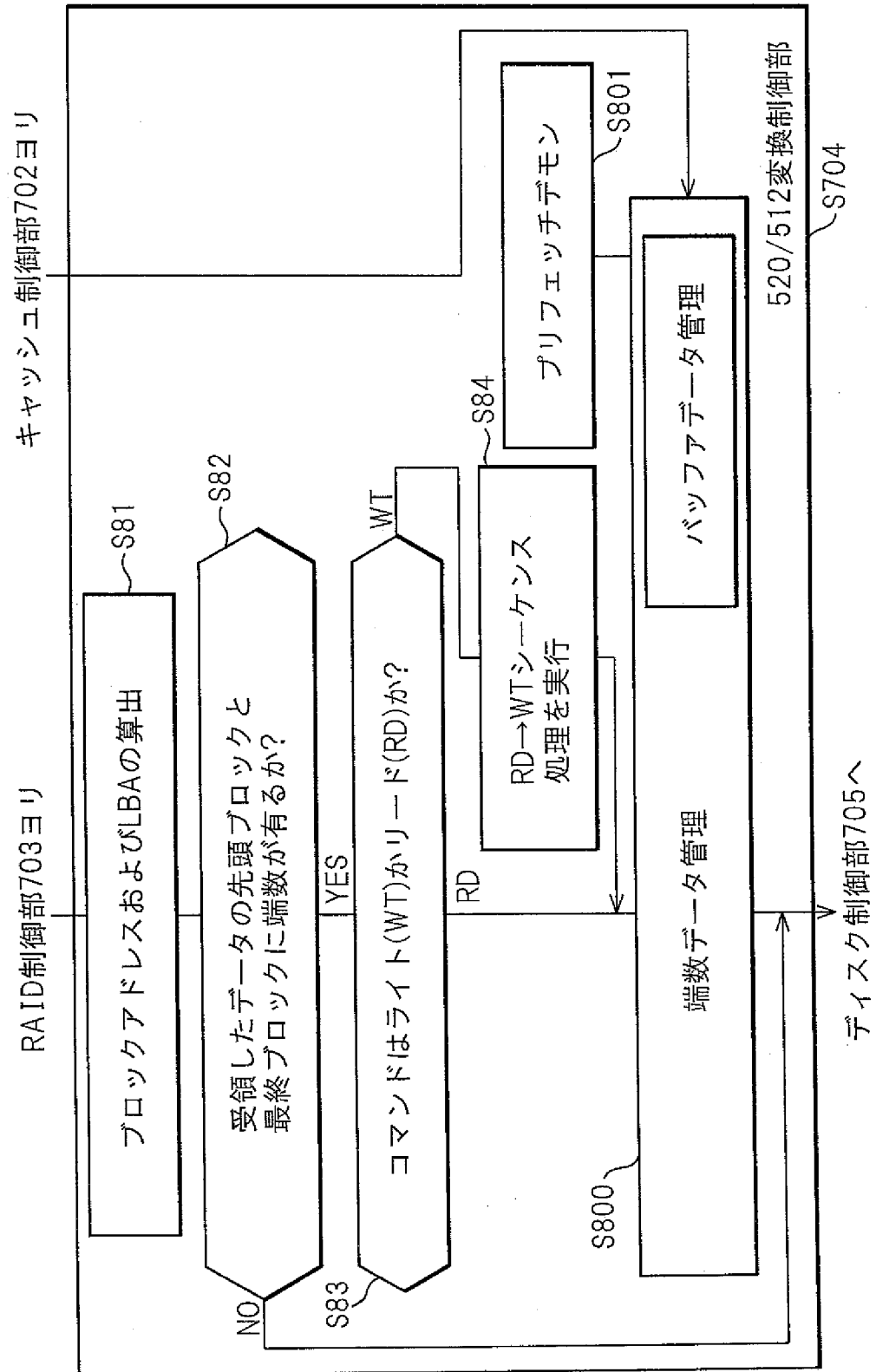
[図7]

Fig.7



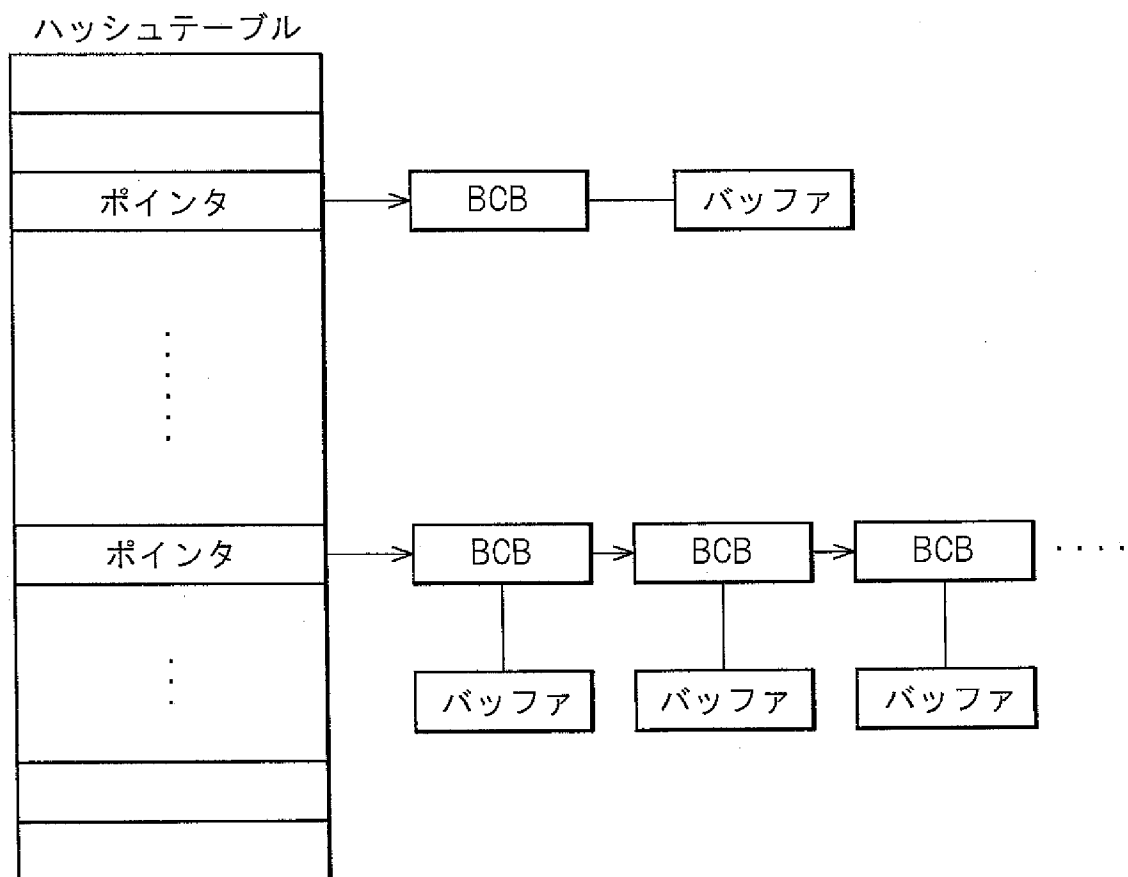
[図8]

Fig.8



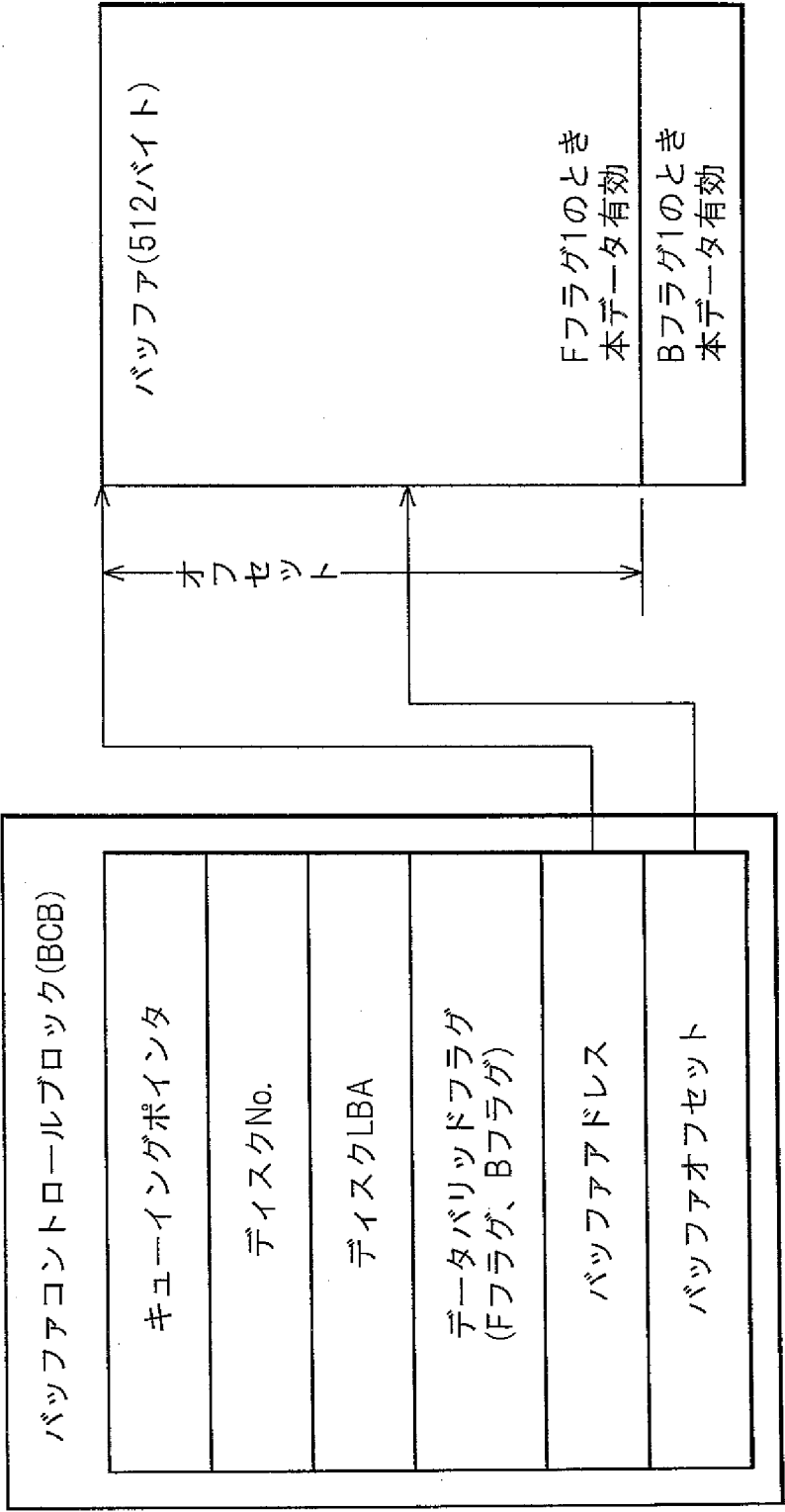
[図9]

Fig.9



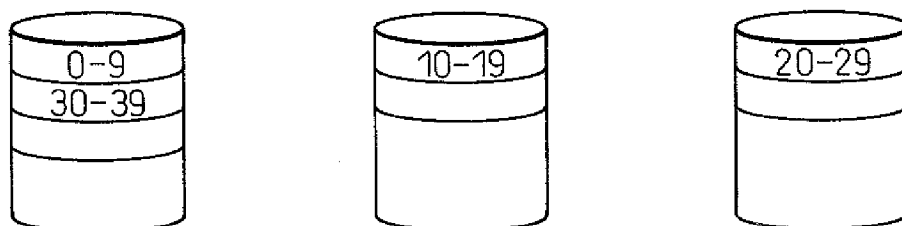
[図10]

Fig. 10



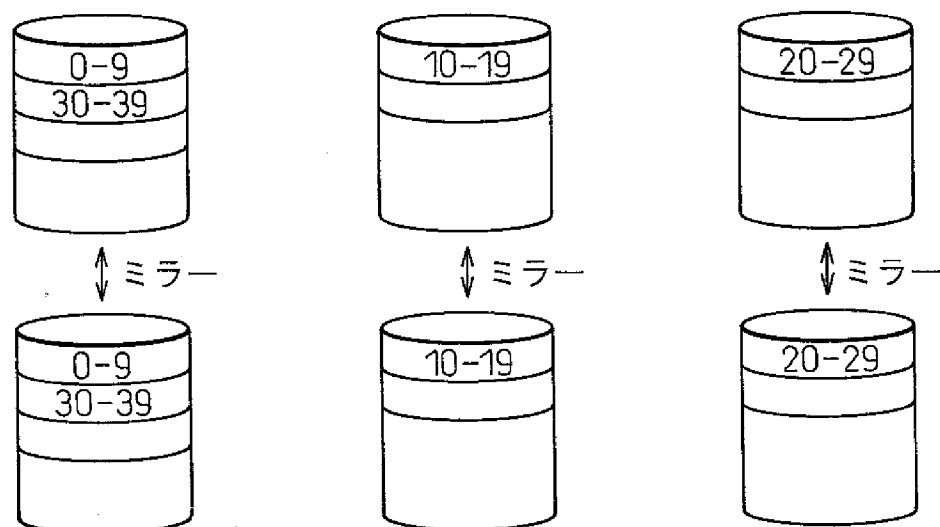
[図11A]

Fig. 11A



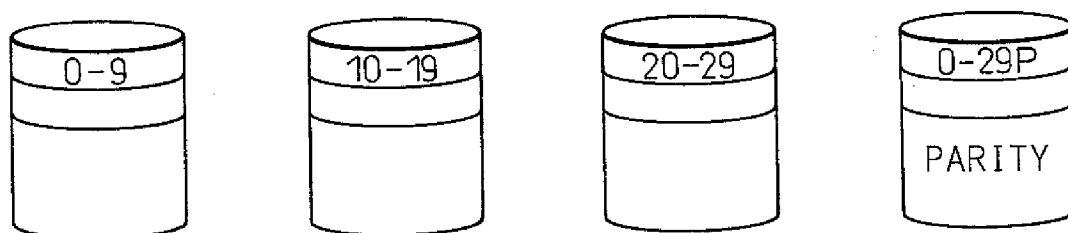
[図11B]

Fig. 11B



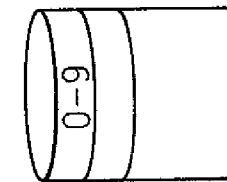
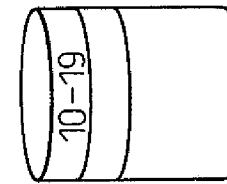
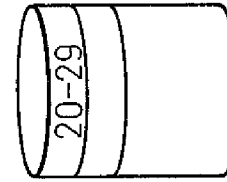
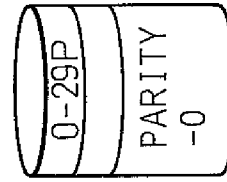
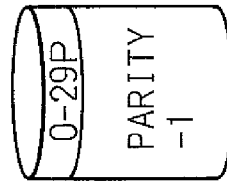
[図11C]

Fig. 11C



[図11D]

Fig. 11D



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2007/063605

A. CLASSIFICATION OF SUBJECT MATTER

G06F3/06 (2006.01) i

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F3/06

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

Jitsuyo Shinan Koho 1922-1996 Jitsuyo Shinan Toroku Koho 1996-2007

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

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X Y	JP 2006-195851 A (Hitachi, Ltd.), 27 July, 2006 (27.07.06), Par. Nos. [0012] to [0071]; all drawings & US 2006-0161756 A1	7, 9 1-6, 8, 10
Y	JP 2000-194509 A (Matsushita Electric Industrial Co., Ltd.), 14 July, 2000 (14.07.00), Par. Nos. [0030] to [0044] (Family: none)	1-6, 8, 10
A	JP 2005-149551 A (Ricoh Co., Ltd.), 09 June, 2005 (09.06.05), Par. Nos. [0059] to [0075] & JP 3637346 B2 & EP 1684288 A1 & WO 2005-045827 A1	1-10

☒ 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
18 September, 2007 (18.09.07)Date of mailing of the international search report
25 September, 2007 (25.09.07)Name and mailing address of the ISA/
Japanese Patent Office

Authorized officer

Facsimile No.

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2007/063605

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 2005-293118 A (Toshiba Solution Kabushiki Kaisha), 20 October, 2005 (20.10.05), Par. Nos. [0015] to [0078] (Family: none)	1-10
A	JP 2006-99916 A (Teac Corp.), 13 April, 2006 (13.04.06), Par. Nos. [0002] to [0006] & US 2006-0069859 A1	1-10

A. 発明の属する分野の分類（国際特許分類（I P C））

Int.Cl. G06F3/06(2006.01)i

B. 調査を行った分野

調査を行った最小限資料（国際特許分類（I P C））

Int.Cl. G06F3/06

最小限資料以外の資料で調査を行った分野に含まれるもの

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

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

C. 関連すると認められる文献

引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求の範囲の番号
X Y	JP 2006-195851 A（株式会社日立製作所）2006.07.27, 段落【0012】－【0071】、全図 & US 2006-0161756 A1	7, 9 1-6, 8, 10
Y	JP 2000-194509 A（松下電器産業株式会社）2000.07.14, 段落【0030】－【0044】（ファミリーなし）	1-6, 8, 10

☒ 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

18.09.2007

国際調査報告の発送日

25.09.2007

国際調査機関の名称及びあて先

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

特許庁審査官（権限のある職員）

梅景 篤

5 I

3801

電話番号 03-3581-1101 内線 3565

C (続き) . 関連すると認められる文献		
引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求の範囲の番号
A	JP 2005-149551 A (株式会社リコー) 2 0 0 5 . 0 6 . 0 9 , 段落【0 0 5 9】－【0 0 7 5】 & JP 3637346 B2 & EP 1684288 A1 & WO 2005-045827 A1	1－1 0
A	JP 2005-293118 A (東芝ソリューション株式会社) 2 0 0 5 . 1 0 . 2 0 , 段落【0 0 1 5】－【0 0 7 8】 (ファミリーなし)	1－1 0
A	JP 2006-99916 A (ティアック株式会社) 2 0 0 6 . 0 4 . 1 3 , 段落【0 0 0 2】－【0 0 0 6】 & US 2006-0069859 A1	1－1 0