

(19) 日本国特許庁(JP)

(12) 公開特許公報(A)

(11) 特許出願公開番号

特開2017-199043

(P2017-199043A)

(43) 公開日 平成29年11月2日(2017.11.2)

(51) Int.Cl.	F I	テーマコード (参考)
G 0 6 F 3/06 (2006.01)	G 0 6 F 3/06 3 0 1 M	
	G 0 6 F 3/06 3 0 4 N	
	G 0 6 F 3/06 5 4 O	
	G 0 6 F 3/06 3 0 4 B	
	G 0 6 F 3/06 3 0 1 Z	
審査請求 未請求 請求項の数 20 O L (全 24 頁)		

(21) 出願番号	特願2014-151668 (P2014-151668)	(71) 出願人	000004237
(22) 出願日	平成26年7月25日 (2014. 7. 25)		日本電気株式会社
			東京都港区芝五丁目7番1号
		(74) 代理人	100080816
			弁理士 加藤 朝道
		(72) 発明者	小川 周吾
			東京都港区芝五丁目7番1号 日本電気株式会社社内

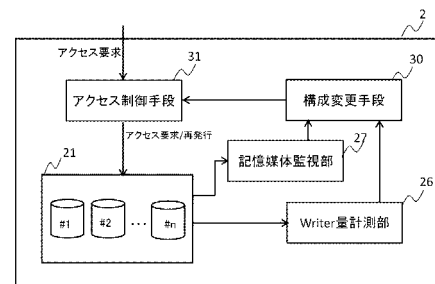
(54) 【発明の名称】 ストレージ装置とシステム及び方法とプログラム

(57) 【要約】

【課題】 応答の遅延、停止が発生した記憶媒体に対するアクセスを回避可能とし、冗長アレイ構成のストレージのアクセス性能の低下を回避可能とする。

【解決手段】 複数の記憶媒体と、各記憶媒体に対する書き込みデータ量を計測する書き込み量計測部と、前記各記憶媒体の負荷状態を監視するとともに、前記各記憶媒体に対するアクセスの応答の遅延、又は前記各記憶媒体の停止状態を監視する記憶媒体監視部と、ストライピングの単位であるチャンクについて、書き込みアクセス対象の前記チャンクの割り当て元の記憶媒体の書き込みデータ量と、前記記憶媒体の負荷状態、応答の遅延又は停止状態の少なくとも1つに基づき、別の記憶媒体を選択し、書き込みアクセス対象の前記チャンクを、前記別の記憶媒体のチャンクに変更する構成変更手段と、新たにアクセス対象となったチャンクに対してアクセス要求を発行するアクセス制御手段を備える。

【選択図】 図 1 2



【特許請求の範囲】**【請求項 1】**

複数の記憶媒体と、

前記各記憶媒体に対する書き込みデータ量を計測する書き込み量計測部と、

前記各記憶媒体の負荷状態を監視するとともに、前記各記憶媒体に対するアクセスの応答の遅延又は停止状態を監視する記憶媒体監視部と、

データを複数の前記記憶媒体に分割して転送するストライピングの単位となる記憶領域であるチャンクについて、書き込みアクセス対象のチャンクの割り当て元の記憶媒体の書き込みデータ量と、前記記憶媒体の負荷状態、応答の遅延又は停止状態の少なくとも1つに基づき、前記書き込みアクセス対象のチャンクの割り当て元の前記記憶媒体の変更が必要であると判断した場合、別の記憶媒体を選択し、前記書き込みアクセス対象のチャンクを、前記別の記憶媒体のチャンクに変更する構成変更手段と、

新たにアクセス対象となったチャンクに対して書き込みアクセス要求を発行するアクセス制御手段と、

を備えた、ことを特徴とするストレージ装置。

【請求項 2】

前記構成変更手段は、書き込みアクセス要求を前記記憶媒体に対して発行する前に、書き込みアクセス対象のチャンクの割り当て元の前記記憶媒体の書き込みデータ量と、他の記憶媒体との書き込みデータ量との差が、予め定められた所定の条件を満たしているか、又は、

前記書き込みアクセス対象のチャンクの割り当て元の前記記憶媒体について、高負荷、応答遅延、応答停止の少なくともいずれか一つに該当していると判定した場合、

前記複数の記憶媒体から、前記書き込みデータ量の条件、高負荷、応答遅延、停止に関与していずれにも該当しない、別の記憶媒体を選択し、

前記書き込みアクセス対象のチャンクの割り当て元の前記記憶媒体から、前記書き込みアクセス対象の前記チャンクを解放し、

前記別の記憶媒体の未使用のチャンクをアクセス対象として割り当てることで、書き込みアクセス先のストライプの構成を変更する、ことを特徴とする請求項 1 記載のストレージ装置。

【請求項 3】

前記複数の記憶媒体が R A I D を構成し、

前記アクセス制御手段が、

ホストからのアクセス要求を各記憶媒体に対するアクセスに変換するアクセス実行部と

、
R A I D ストライプの構成変更によって新たにアクセス対象となったチャンクに対してアクセス要求を発行するアクセス再発行部と、

を備え、

前記構成変更手段は、

前記各記憶媒体に対してアクセス要求を行う記憶媒体アクセス部と、

前記各記憶媒体のチャンクの割り当て、解放を行うチャンク割当部と、

R A I D のストライプ構成を管理する R A I D 管理部と、

書き込みアクセス要求を発行する前に、書き込み先の記憶媒体の書き込みデータ量、又は、前記書き込み先の記憶媒体の状態に応じて、R A I D ストライプの構成を変更する R A I D 構成変更部と、

を備えた、ことを特徴とする請求項 1 又は 2 記載のストレージ装置。

【請求項 4】

前記 R A I D 管理部は、前記 R A I D のストライプの前記ホスト側からみたアドレス、R A I D のレベル、R A I D のストライプのチャンク数、前記チャンクが割り当てられた記憶媒体の識別情報とアドレス情報の対を含む使用チャンク情報を、対応付けて記憶部で管理する、ことを特徴とする請求項 3 記載のストレージ装置。

10

20

30

40

50

【請求項 5】

前記アクセス実行部は、前記ホストからのアクセス要求を前記各記憶媒体に対応するアクセスに変換してアクセスの実行を管理して、前記アクセス要求に関する処理の実行完了の応答を前記ホストに対して行う、ことを特徴とする請求項 3 又は 4 記載のストレージ装置。

【請求項 6】

前記記憶媒体アクセス部は、前記アクセス実行部によって変換又は生成されたアクセス要求を前記記憶媒体に対して発行し、前記記憶媒体からの結果を、前記アクセス実行部に応答として返す、ことを特徴とする請求項 3 乃至 5 のいずれか 1 項に記載のストレージ装置。

10

【請求項 7】

前記チャンク割当部は、前記各記憶媒体の記憶領域をチャンク単位に分割し、各チャンクの使用状況を管理し、RAID ストライプの構成変更に伴うチャンクの解放、及び、チャンクの割り当てを行う、ことを特徴とする請求項 3 乃至 6 のいずれか 1 項に記載のストレージ装置。

【請求項 8】

前記 RAID 管理部は、RAID ストライプを構成するチャンクの組み合わせ、及び記憶領域全体のアドレス空間に対するチャンクのマッピングを管理する、ことを特徴とする請求項 3 乃至 7 のいずれか 1 項に記載のストレージ装置。

【請求項 9】

前記書き込み量計測部は、各記憶媒体に発行された書き込みアクセス要求の情報を参照して、記憶媒体毎の書き込みデータ量を計測する、ことを特徴とする請求項 3 乃至 8 のいずれか 1 項に記載のストレージ装置。

20

【請求項 10】

前記記憶媒体監視部は、前記記憶媒体アクセス部を通して、各記憶媒体に対する発行待ち、又は、発行中のアクセス要求から、負荷を算出し、高負荷状態の記憶媒体を検出し、各記憶媒体のアクセスに対する応答状況を監視して応答遅延又は停止状態の記憶媒体を検出する、ことを特徴とする請求項 3 乃至 9 のいずれか 1 項に記載のストレージ装置。

【請求項 11】

前記 RAID 構成変更部は、前記記憶媒体アクセス部によって各記憶媒体に書き込みアクセス要求を発行する直前に、書き込み先の記憶媒体の書き込みデータ量と、アクセス負荷、応答遅延、又は、停止の有無を確認し、

30

書き込み先の記憶媒体について、他の記憶媒体との書き込みデータ量の差が、予め定められた所定の条件を満たしているか、

書き込み先の記憶媒体について、高負荷、応答遅延又は停止の少なくとも 1 つの条件を満たしている場合には、書き込みアクセス先の RAID ストライプについて、書き込み先の記憶媒体が、高負荷、応答遅延、応答停止に該当しない、他の記憶媒体となるように、チャンクの構成を変更する、ことを特徴とする請求項 3 乃至 10 のいずれか 1 項に記載のストレージ装置。

【請求項 12】

40

前記アクセス再発行部は、書き込み先の記憶媒体において、他の記憶媒体との書き込みデータ量の差が、予め定められた所定の条件を満たすか、書き込み先の記憶媒体について高負荷、応答遅延、応答停止のいずれかの条件を満たしている場合には、前記記憶媒体アクセス部に対して、書き込み先の記憶媒体への書き込みアクセス要求のキャンセルを指示し、前記アクセス実行部に対して、割り当てを変更したチャンクに対するアクセス要求の再発行を指示する請求項 3 乃至 11 のいずれか 1 項に記載のストレージ装置。

【請求項 13】

前記記憶媒体は、ブロック単位で一括消去される書き換え可能な不揮性半導体記憶デバイスからなる、請求項 1 乃至 12 のいずれか 1 項に記載のストレージ装置。

【請求項 14】

50

請求項 1 乃至 1 3 のいずれか 1 項に記載のストレージ装置と、
前記ストレージ装置と通信接続され、前記ストレージ装置に対してアクセス要求を行う
ホストと、
を備えたコンピュータシステム。

【請求項 1 5】

複数の記憶媒体の各々に対する書き込みデータ量を計測し、
前記各記憶媒体の負荷状態を監視するとともに、前記各記憶媒体に対するアクセスの応
答の遅延又は停止状態を監視し、

データを複数の前記記憶媒体に分割して転送するストライピングの単位となる記憶領域
であるチャンクについて、書き込みアクセス対象のチャンクの割り当て元の記憶媒体の書
き込みデータ量と、前記記憶媒体の負荷状態、応答の遅延又は停止状態の少なくとも 1 つ
に基づき、前記書き込みアクセス対象のチャンクの割り当て元の前記記憶媒体の変更が必
要であると判断した場合、別の記憶媒体を選択し、書き込みアクセス対象の前記チャンク
を、前記別の記憶媒体のチャンクに変更し、

新たにアクセス対象となったチャンクに対して書き込みアクセス要求を発行する、こと
を特徴とするストレージのアクセス制御方法。

【請求項 1 6】

書き込みアクセス要求を前記記憶媒体に対して発行する前に、前記書き込みアクセス対
象のチャンクの割り当て元の前記記憶媒体の書き込みデータ量と、他の記憶媒体との書き
込みデータ量との差が、予め定められた所定の条件を満たしているか、又は、

前記書き込みアクセス対象のチャンクの割り当て元の前記記憶媒体について、高負荷、
応答遅延、応答停止の少なくともいずれか一つに該当していると判定した場合、

前記複数の記憶媒体から、前記書き込みデータ量の条件、高負荷、応答遅延、停止に関
していずれにも該当しない別の記憶媒体を選択し、

前記書き込みアクセス対象のチャンクの割り当て元の前記記憶媒体から、前記書き込み
アクセス対象の前記チャンクを解放し、

前記別の記憶媒体の未使用のチャンクをアクセス対象として割り当てることで、書き込
みアクセス先のストライプの構成を変更する、ことを特徴とする請求項 1 5 記載のストレ
ージのアクセス制御方法。

【請求項 1 7】

前記複数の記憶媒体が R A I D を構成し、

書き込みアクセス要求を発行する前に、書き込み先の記憶媒体の書き込みデータ量、又
は、前記記憶媒体の高負荷、応答遅延、停止状態に応じて、R A I D ストライプの構成を
変更する、ことを特徴とする請求項 1 5 又は 1 6 記載のストレージのアクセス制御方法。

【請求項 1 8】

複数の記憶媒体の各々に対する書き込みデータ量を計測する処理と、

前記各記憶媒体の負荷状態を監視するとともに、前記各記憶媒体に対するアクセスの応
答の遅延又は停止状態を監視する処理と、

データを複数の前記記憶媒体に分割して転送するストライピングの単位となる記憶領域
であるチャンクについて、書き込みアクセス対象のチャンクの割り当て元の記憶媒体の書
き込みデータ量と、前記記憶媒体の負荷状態、応答の遅延又は停止状態の少なくとも 1 つ
に基づき、前記書き込みアクセス対象のチャンクの割り当て元の前記記憶媒体の変更が必
要であると判断した場合、別の記憶媒体を選択し、書き込みアクセス対象の前記チャンク
を、前記別の記憶媒体のチャンクに変更する処理と、

新たにアクセス対象となったチャンクに対して書き込みアクセス要求を発行する処理と

、
をコンピュータに実行させるプログラム。

【請求項 1 9】

書き込みアクセス要求を前記記憶媒体に対して発行する前に、前記書き込みアクセス対
象のチャンクの割り当て元の前記記憶媒体の書き込みデータ量と、他の記憶媒体との書き

10

20

30

40

50

込みデータ量との差が、予め定められた所定の条件を満たしているか、又は、

前記書き込みアクセス対象のチャンクの割り当て元の前記記憶媒体について、高負荷、応答遅延、応答停止の少なくともいずれか一つに該当していると判定した場合、

前記複数の記憶媒体から、前記書き込みデータ量の条件、高負荷、応答遅延、停止に関し、

前記書き込みアクセス対象のチャンクの割り当て元の前記記憶媒体から、前記書き込みアクセス対象の前記チャンクを解放し、

前記別の記憶媒体の未使用のチャンクをアクセス対象として割り当てることで、書き込みアクセス先のストライプの構成を変更する処理を、前記コンピュータに実行させる請求項 18 記載のプログラム。

10

【請求項 20】

前記複数の記憶媒体が R A I D を構成しており、

書き込みアクセス要求を発行する前に、書き込み先の記憶媒体の書き込みデータ量、又は、前記記憶媒体の高負荷、応答遅延、停止状態に応じて、R A I D ストライプの構成を変更する処理を前記コンピュータに実行させる請求項 18 又は 19 記載のプログラム。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、ストレージ装置とシステム及び方法とプログラムに関する。

【背景技術】

20

【0002】

< R A I D >

複数の記憶媒体を用いて仮想的に単一の記憶領域を提供し冗長性を持たせ信頼性を向上する技術として、R A I D (Redundant Array of Inexpensive Disks) が広く用いられている。記憶媒体間の冗長構成の実現に関して、複数の R A I D レベルが存在する。例えば R A I D 10、R A I D 5、R A I D 6 等では、通常、各記憶媒体の記憶領域を、チャンク (chunk) と呼ばれる、予め定められた一定サイズ (チャンクサイズ (chunk size) 又はストライプサイズ (stripe size)) の領域で区切る。チャンクはストランピング (データを複数の前記記憶媒体に分割して転送する) の単位となる。仮想的なアドレス空間に異なる記憶媒体のチャンクを順番に割り当てるストライピングを行うことで、複数記憶媒体の並列動作による性能向上と、複数記憶媒体の容量を合わせた記憶領域を実現する。4 台の記憶媒体で構成される R A I D 5 では、データは一定サイズのチャンク単位に分割され、例えばチャンク 1 ~ 3 は記憶媒体 1 ~ 3 に格納され、チャンク 1 ~ 3 のパリティは記憶媒体 4 に格納される。チャンク 1 ~ 3 (パリティのチャンクを含ませる場合、チャンク 1 ~ 4) を「R A I D ストライプ」と呼ぶ。チャンクとパリティの順序はストライプごとに変更され、例えば、次のストライプでは、チャンク 4 ~ 6 は記憶媒体 2 ~ 4 に格納され、チャンク 4 ~ 6 のパリティは記憶媒体 1 に格納される。また、R A I D 0 を除き、記憶媒体間のデータにミラー、パリティ等の冗長性を持たせることで、記憶媒体の障害等によるデータ消失を防止して、信頼性を向上させている。

30

【0003】

40

R A I D は、記憶媒体である H D D (Hard Disk Drive) の性能、容量、信頼性を伸ばすために用いられているが、近年では、R A I D の記憶媒体として、N A N D フラッシュメモリを用いた S S D (Solid State Drive) も用いられつつある。S S D についてその概略を以下に説明する。

【0004】

< S S D >

H D D と同様のインタフェースを備えた S S D に対して、アクセス元となるホストは、H D D と同様に、読み出し (Read) 及び書き込み (Write) アクセスが可能である。しかしながら、よく知られているように、N A N D フラッシュメモリは、H D D とは異なる特性を有する。例えば N A N D フラッシュメモリは、各領域に対する格納データの

50

更新回数増加に従ってビットエラーの発生率が上昇する。そのため、NANDフラッシュメモリが正しくデータを保持可能な条件下における、格納データの更新回数に制限を持つ。

【0005】

SSDでは、NANDフラッシュメモリにおける格納データの更新回数（書き換え回数）の増加に伴うエラー発生を防ぐため（SSDの寿命を延ばす）に、各記憶領域に対する更新回数（データ書換回数）を各記憶領域（記憶デバイス）間で均一化する「ウェアレベリング（wear levelling / wear leveling）」と呼ばれる技術が用いられる。SSDで用いられるNANDフラッシュメモリの各記憶領域における更新回数は、SSDに対するWriteの総量で表される。つまり、SSDに対するWriteの総量がSSDの寿命を決定する。

10

【0006】

また、SSDは大容量化が進む一方で、その寿命は短くなりつつある。短寿命化は、記憶デバイスであるNANDフラッシュメモリの更新回数制限の減少による。NANDフラッシュメモリは、近年、製造プロセスの微細化、及び、メモリセルに格納される情報量の多値化によって、大容量化が実現されているが、更新可能な回数は減少している。例えば、メモリセルに格納されるビットを多値化したMLC（Multi Level Cell）又はTLC（Triple Level Cell）等のNANDフラッシュメモリは、大容量化、低コスト化を実現しているが、更新可能な回数の制限が特に少ない。したがって、SSDの寿命に直結するWrite可能な総データ量も減少する。つまり、SSDに対するデータの更新頻度、更新量の削減は、寿命の観点から、特に重要性を増している。

20

【0007】

< SSDのアクセス特性 >

更に、SSDでは、HDDと同様のアクセス手段を提供するために、NANDフラッシュメモリのアクセス特性を隠蔽するための様々な処理が行われる。例えば、NANDフラッシュメモリでは、格納されたデータを更新する場合、一度格納されたデータを消去する消去（Erase）操作が必要である。NANDフラッシュメモリにおいて、読み出し及び書き込みは、該読み出し及び書き込みアクセスの単位であるページ単位に行われ、消去（Erase）は、ページを複数まとめたブロック単位となる。NANDフラッシュメモリにおいて、あるページに格納されたデータを更新する場合、一括消去の対象となるブロックに含まれる有効なページに対するデータの退避、及び、再格納が、SSDによって行われる。NANDフラッシュメモリにおいて、ブロックの消去に伴うデータ退避、及び、再格納のページ数は、SSDに対する、書き込み先のアドレス、及び、アクセスサイズによって変化する。具体的には、書き込みアクセスのサイズが小さくなるほど、NANDフラッシュメモリのブロックにおける更新対象外のページが増加して、当該ブロックの消去に伴い退避が必要とされるページ数が増加する。

30

【0008】

また、SSD内部におけるNANDフラッシュメモリの記憶領域に関する使用状況、SSDに対するアクセスの状況等の条件で、消去（Erase）に伴うページの退避、再格納の処理負荷は、増減する。そのため、SSDに対するアクセス要求の状況から、SSDの内部における負荷の状況を正確に把握することは、困難である。SSD内部の負荷の把握が困難であることは、SSDが意図しない状況で高負荷となり、結果として、一時的に応答の遅延、若しくは停止が発生することを意味する。

40

【0009】

SSDにおけるアクセスの応答遅延は、情報処理システムの性能低下に加えて、システム全体の動作におけるタイムアウトの形で障害として影響を与える可能性が存在する。

【0010】

RAIDを構成するSSDにおいて、応答の遅延、若しくは停止が発生した場合、応答遅延又は停止状態のSSDに対する読み出し（Read）、又は書き込み（Write）が行われると、アクセス要求元に対して応答の遅延、停止が顕在化する。

50

【 0 0 1 1 】

つまり、R A I Dを構成する各S S Dの応答遅延、停止は、R A I Dグループ全体の応答遅延、停止に結び付くことになる。

【 0 0 1 2 】

R A I Dグループ全体の応答遅延、停止を回避するためには、応答の遅延、停止が発生したS S Dに対するアクセスを避けることが必要である。

【 0 0 1 3 】

< 特許文献 1 >

記憶媒体にS S Dを用いたR A I D構成として、例えば特許文献 1 には、複数のストレージデバイスのうちの、読み込み要求が対象とするデータを含む少なくとも第 1 のストレージデバイスを識別し、第 1 のストレージデバイスが相対的に長い応答レイテンシ又は相対的に低いスループットの少なくとも一方を含む可変パフォーマンスを示すであろうことを検出又は予測したことに応答して、前記複数のストレージデバイスのうちの前記第 1 のストレージデバイス以外の 1 又はそれ以上のデバイスから、前記データを取得するように構成された再構成読み込み要求を生成するように構成されたコンピュータシステムが開示されている。特許文献 1 では、応答遅延が発生したS S Dのデータを、代わりにR A I Dグループ内の他のS S Dに格納されている、冗長関係にあるデータを読み出すことで、応答の遅延を回避している。

10

【 0 0 1 4 】

< 特許文献 2 >

また特許文献 2 には、S S DでR A I Dを構成する場合、S S Dにおけるウェアレベリングによる書き込み / 消去回数の平滑化が行われるとともに、R A I D機能による負荷分散によって、各S S Dへの書き込み回数が均一化される可能性が高くなり、複数のS S Dが同時期に故障する可能性が高くなってしまい、交換保守作業が難しくなるため、データ損失の危険性が生じるという問題に対して、複数の不揮発性の半導体記憶装置に対するデータの書き込み頻度に差をつける必要があると判断した場合、各半導体記憶装置に対するデータの書き込み頻度に差をつけるべく、各半導体記憶装置に対してデータと共に誤り訂正符号データをストライプブロック単位で分散させて記録するデータ破損の復元機能にかかるデータおよび誤り訂正符号データの配置情報を記録したマッピング表で管理されているストライプブロックのデータを入れ替えて配置情報を変更する。これにより、各半導体記憶装置に対するデータの書き込み頻度に差がつき、複数の半導体記憶装置が同時期に故障してしまうことを回避させることができるので、データ損失の危険性を低減させるようにした構成が開示されている。

20

30

【 0 0 1 5 】

< 特許文献 3 >

なお、H D DのR A I Dに関して、特許文献 3 には、R A I D 5 レベルのディスクアレイのごとく、データの書き込み時にパリティの読み出しを要求するタイプのディスクアレイにおいて、書き込み要求あるいは読み出し要求の実行を高速にするような、ディスクアレイ内の記憶領域の割り当て方法として、論理的に連続する複数ブロックのデータに対して論理ブロックおよびその論理ブロックのパリティを格納するドライブが全て異なるように制御する構成が開示されている。

40

【 0 0 1 6 】

< 特許文献 4 >

特許文献 4 には、階層レベルを含む階層構成ルールが変更された場合に、コントローラは、ストレージボリュームの階層の階層構成を変更し、プールが、有効であった階層構成ルールを満たさない場合に生じる、物理ストレージデバイスに対する階層レベルの変化に基づいて、コントローラは、プールを複数の仮想ボリュームに割り当てる構成が開示されている。

【 0 0 1 7 】

< 特許文献 5 >

50

なお、R A I D構成ではないが、メモリの負荷に応じて格納先のメモリを変更する関連技術として、例えば特許文献5には、複数のノードの各々は、高負荷状態と判定された第1メモリの一部のデータを、他のノードの第2メモリに移動し、第1メモリと第2メモリとでインタリーブを行うメモリ制御部を備えた構成が開示されている。

【先行技術文献】

【特許文献】

【0018】

【特許文献1】特表2013-541765号公報

【特許文献2】特開2010-15516号公報

【特許文献3】特開平8-185275号公報

【特許文献4】特開2011-65624号公報

【特許文献5】特開2009-237709号公報

【発明の開示】

【発明が解決しようとする課題】

【0019】

以下に関連技術の分析を与える。

【0020】

上記したように、R A I DのH D DをS S Dで置き換える場合、S S DのH D Dとは異なる特性を考慮する必要がある。例えばS S Dの寿命は、接続元ホストからのW r i t e総量に依存する。このため、W r i t e量が限界に近づいたS S Dは、データ損失を防ぐために、交換が必要である。S S Dの交換は新たなコストの発生につながる。よって、可能な限り、S S Dの交換の発生頻度を減らすことが望ましい。つまり、S S DのW r i t eデータ量を削減することが望まれる。

【0021】

また、複数のS S Dを使用する環境において、各S S DのW r i t eデータ量を削減するには、S S Dに対するW r i t eデータ量を、直接減らすだけでなく、各S S Dに対するW r i t eデータ量を均一にすることで、特定のS S Dが早く寿命を迎えないようにすることが望まれる。

【0022】

しかしながら、R A I Dを構成するS S Dでは、各S S Dに対するW r i t eデータ量の均一化は困難である。その理由として、一般的なR A I Dでは、提供するアドレス空間に対する、各S S Dの記憶領域のマッピングが固定であることが挙げられる。つまり、R A I Dの提供するアドレス空間に対する各W r i t eデータの格納先は、W r i t e先のアドレス(W r i t eアクセス対象のアドレス)によって一意に決定される。このため、R A I Dを構成するS S D間において、W r i t eデータ量の偏りを解消するための、W r i t eデータ量の調整は困難である。

【0023】

また、記憶媒体間で冗長性を持つR A I D 5、R A I D 6等のR A I Dレベルを用いる場合には、各記憶媒体のチャンクを組み合わせることでパリティを形成するため、データの格納先となる記憶媒体、アドレスを自由に選択することは困難である。

【0024】

さらに、S S Dは、H D Dとは異なり、外部から負荷を予測することが困難である。このため、外部からのアクセス要求の負荷が低い状況であっても、一時的に応答の遅延、停止が発生し得る。R A I Dを構成するS S Dにおいて、応答の遅延、若しくは停止が発生した場合、応答遅延又は停止状態のS S Dに対するR e a d又はW r i t eが行われると、アクセス要求元に対して、応答の遅延、停止が顕在化する。

【0025】

R A I Dグループ全体の応答遅延、停止を回避するためには、応答の遅延、停止が発生したS S Dに対するアクセスを避けることが必要である。

【0026】

10

20

30

40

50

特許文献 1 によれば、R A I D グループに対する R e a d アクセスは、応答遅延が発生した S S D のデータを、代わりに、R A I D グループ内の他の S S D に格納された冗長関係にあるデータを読み出すことで、応答遅延等を回避することは可能であるが、特定の S S D に対する W r i t e アクセスを回避することは困難である。これは、一般的な R A I D では、

- ・各 S S D に対する記憶領域のマッピングが固定であること、
- ・固定されたチャンクの集合でパリティを形成すること、による。

【 0 0 2 7 】

また、特許文献 2 では、R A I D ストライプと呼ばれる、複数記憶媒体間でパリティを構成するチャンクの集合において、複数のフラッシュメモリで構成された記憶媒体が同時に寿命を迎えないように、チャンクのマッピングを記憶媒体間に入れ替えることで、各記憶媒体の W r i t e データ量を調整している。しかし、データ格納先の変更は、R A I D ストライプ内に限定されており、記憶媒体間で、チャンクのマッピングを交換するものである。このため、応答が停止した S S D と他の S S D との間でデータを入れ替えた場合に、停止した S S D に対するアクセスを避けることは不可能である。

【 0 0 2 8 】

また、特許文献 3 には、R A I D ストライプを動的に形成する方法が開示されている。R A I D ストライプを構成するチャンクを動的に選択する際に、先に選んだチャンクが格納された記憶媒体のいずれとも衝突しないように選択することで、冗長性を持った R A I D ストライプの動的な形成を実現する。しかしながら、特許文献 3 には、構成済みの R A I D ストライプに対して、使用する記憶媒体側の状態に基づき、該 R A I D ストライプを再構成する、ことについて何も記載されていない。

【 0 0 2 9 】

上記特許文献等では、R A I D グループを構成する S S D において発生する、W r i t e データ量に依存する寿命、及び、一時的な応答遅延又は停止の問題を同時に解決することはできない。

【 0 0 3 0 】

本発明は、上記課題に鑑みて創案されたものであって、その目的は、応答の遅延、停止が発生した記憶媒体に対するアクセスを回避可能とし、アクセス性能の低下を抑制可能とする装置、システム、方法、プログラムを提供することにある。

【課題を解決するための手段】

【 0 0 3 1 】

本発明の 1 つの側面によれば、複数の記憶媒体と、前記各記憶媒体に対する書き込みデータ量を計測する書き込み量計測部と、前記各記憶媒体の負荷状態を監視するとともに、前記各記憶媒体に対するアクセスの応答の遅延又は停止状態を監視する記憶媒体監視部と、データを複数の前記記憶媒体に分割して転送するストライピングの単位となる記憶領域であるチャンクについて、書き込みアクセス対象の前記チャンクの割り当て元の記憶媒体の書き込みデータ量と、前記記憶媒体の負荷状態、応答の遅延又は停止状態の少なくとも 1 つに基づき、書き込みアクセス対象のチャンクの割り当て元の前記記憶媒体の変更が必要であると判断した場合、別の記憶媒体を選択し、書き込みアクセス対象の前記チャンクを、前記別の記憶媒体のチャンクに変更する構成変更手段と、新たにアクセス対象となったチャンクに対して書き込みアクセス要求を発行するアクセス制御手段と、を備えたストレージ装置が提供される。

【 0 0 3 2 】

本発明によれば、上記ストレージ装置と、ホストを備えたコンピュータシステムが提供される。

【 0 0 3 3 】

本発明の他の側面によれば、複数の記憶媒体の各々に対する書き込みデータ量を計測し、前記各記憶媒体の負荷状態を監視するとともに、前記各記憶媒体に対するアクセスの応答の遅延又は停止状態を監視し、

10

20

30

40

50

データを複数の前記記憶媒体に分割して転送するストライピングの単位となる記憶領域であるチャンクについて、書き込みアクセス対象の前記チャンクの割り当て元の記憶媒体の書き込みデータ量と、前記記憶媒体の負荷状態、応答の遅延又は停止状態の少なくとも１つに基づき、書き込みアクセス対象のチャンクの割り当て元の前記記憶媒体の変更が必要であると判断した場合、別の記憶媒体を選択し、書き込みアクセス対象の前記チャンクを、前記別の記憶媒体のチャンクに変更し、

新たにアクセス対象となったチャンクに対して書き込みアクセス要求を発行する、ストレージ制御方法が提供される。

【 0 0 3 4 】

本発明の他の側面によれば、複数の記憶媒体の各々に対する書き込みデータ量を計測する処理と、

前記各記憶媒体の負荷状態を監視するとともに、前記各記憶媒体に対するアクセスの応答の遅延又は停止状態を監視する処理と、

データを複数の前記記憶媒体に分割して転送するストライピングの単位となる記憶領域であるチャンクについて、書き込みアクセス対象の前記チャンクの割り当て元の記憶媒体の書き込みデータ量と、前記記憶媒体の負荷状態、応答の遅延又は停止状態の少なくとも１つに基づき、書き込みアクセス対象のチャンクの割り当て元の前記記憶媒体の変更が必要であると判断した場合、別の記憶媒体を選択し、書き込みアクセス対象の前記チャンクを、前記別の記憶媒体のチャンクに変更する処理と、

新たにアクセス対象となったチャンクに対して書き込みアクセス要求を発行する処理と、をコンピュータに実行させるプログラムが提供される。

【 0 0 3 5 】

本発明によれば、該プログラムを記録したコンピュータ読み出し可能な記録媒体（半導体メモリ、磁／光ディスク等）が提供される。

【発明の効果】

【 0 0 3 6 】

本発明によれば、応答の遅延、停止が発生した記憶媒体に対するアクセスを回避可能とし、冗長アレイ構成のストレージのアクセス性能の低下を回避可能としている。

【図面の簡単な説明】

【 0 0 3 7 】

【図 1】本発明の一実施形態のシステム構成の一例を例示する図である。

【図 2】発明の一実施形態のストレージの構成を例示する図である。

【図 3】本発明の一実施形態におけるアクセス要求テーブルを例示する図である。

【図 4】本発明の一実施形態におけるアクセステーブルを例示する図である。

【図 5】本発明の一実施形態における R A I D 管理テーブルを例示する図である。

【図 6】本発明の一実施形態の動作を説明するフローチャートである。

【図 7】本発明の一実施形態の動作を説明するフローチャートである。

【図 8】本発明の一実施形態の動作を説明するフローチャートである。

【図 9】本発明の一実施形態におけるストレージの記憶媒体の構成を例示する図である。

【図 10】本発明の一実施形態における R A I D 管理テーブルを例示する図である。

【図 11】本発明の一実施形態における各記憶媒体の W r i t e データ量を例示する図である。

【図 12】本発明の基本概念を説明する図である。

【発明を実施するための形態】

【 0 0 3 8 】

本発明の基本概念を説明し、つづいて実施形態について説明する。図 12 を参照すると、本発明に係るストレージ装置は、その一つの形態において、複数の記憶媒体（21）と、前記各記憶媒体（21）に対する書き込みデータ量を計測する W r i t e 量（書き込み量）計測部（26）と、前記各記憶媒体の負荷状態を監視するとともに、前記各記憶媒体に対するアクセスの応答の遅延又は停止状態を監視する記憶媒体監視部（27）と、スト

10

20

30

40

50

ライピング（データを複数の記憶媒体に分割して並列転送する）の単位である記憶領域（チャンク）について、書き込みアクセス対象の前記チャンクの割り当て元の記憶媒体の書き込みデータ量と、前記記憶媒体の負荷状態、応答の遅延又は停止状態の少なくとも１つに基づき、書き込みアクセス対象のチャンクの割り当て元の前記記憶媒体がデータ書き込み先として不適であり、変更が必要であると判断した場合、別の記憶媒体（書き込みデータ量の条件、負荷状態、応答の遅延又は停止のいずれにも該当せず、データ書き込み先としてより適切な別の記憶媒体）を選択し、書き込みアクセス対象の前記チャンクを、前記別の記憶媒体のチャンクに変更（したがって書き込みアクセス先のストライプのチャンク構成を変更）する構成変更手段（３０）と、新たにアクセス対象となったチャンクに対して書き込みアクセス要求を発行するアクセス制御手段（３１）と、を備えている。

10

【００３９】

なお、本発明の一形態において、Write量計測部（２６）、記憶媒体監視部（２７）、構成変更手段（３０）、アクセス制御手段（３１）は、例えばストレージ装置（２）を構成するコンピュータ上で実行されるプログラムによりその機能、処理を実現するようにしてもよい。Write量計測部（２６）、記憶媒体監視部（２７）、構成変更手段（３０）、アクセス制御手段（３１）は、RAIDコントローラあるいはRAIDサーバに実装してもよい。

【００４０】

本発明の一形態において、構成変更手段（３０）は、後述する図２の記憶媒体アクセス部（２３）、RAID構成変更部（２８）、チャンク割当部（２４）、RAID管理部（２５）で構成してもよい。

20

【００４１】

また、本発明の一形態において、アクセス制御手段（３１）は、図２のアクセス実行部（２２）とアクセス再発行部（２９）で構成してもよい。

【００４２】

本発明によれば、記憶媒体の応答遅延や応答停止等による影響を回避し、複数の記憶媒体に対する書き込みデータ量を調整することで、統合された仮想的な記憶領域を形成する。例えばRAID等による記憶媒体間のデータ冗長化を実現しつつ、各記憶媒体のWriteデータ量の調整、及び記憶媒体の一時的な応答遅延、又は応答停止の影響を受けずに、複数記憶媒体からなる記憶領域を構成可能としている。

30

【００４３】

前述したように、RAIDを構成する記憶媒体（例えばSSD）において、応答の遅延、又は停止が発生した場合、応答遅延又は停止状態の記憶媒体に対するRead、又はWriteが行われると、アクセス要求元に対して応答の遅延、停止が顕在化する。各記憶媒体の応答遅延、停止がRAIDグループ全体の応答遅延、停止に結び付く。

【００４４】

RAIDグループ全体の応答遅延、停止を回避するためには、応答の遅延、停止が発生した記憶媒体に対するアクセスを避けることが必要である。

【００４５】

そこで、本発明によれば、データの格納先となる記憶媒体（例えばSSD）及びアドレスを、データ格納時点の各記憶媒体（例えばSSD）に対するWriteデータ量、及び、記憶媒体の応答遅延、停止状態に応じて、RAIDストライプの構成を、動的に変更する手段を提供する。

40

【００４６】

本発明によれば、複数記憶媒体間のデータ冗長化を実現しつつ、各記憶媒体のWriteデータ量の調整、及び記憶媒体の一時的な応答遅延、又は応答停止の影響を受けずに記憶領域を構成する。

【００４７】

一実施形態によれば、ホストからのアクセス要求を記憶媒体（図２の２１）に対するアクセスに変換してアクセス処理の実行を管理するアクセス実行部（図２の２２）と、

50

各記憶媒体に対してアクセス要求を行う記憶媒体アクセス部（図２の２３）と、
各記憶媒体から記憶領域の割り当てを行うチャンク割当部（図２の２４）と、
ＲＡＩＤストライプの構成を管理するＲＡＩＤ管理部（図２の２５）と、
各記憶媒体に対するＷｒｉｔｅデータ量を計測するＷｒｉｔｅ量計測部（図２の２６）
と、

各記憶媒体の負荷状態、及び応答遅延又は停止状態を監視する記憶媒体監視部（図２の
２７）と、

Ｗｒｉｔｅアクセス要求を発行する前に、Ｗｒｉｔｅ先の記憶媒体のＷｒｉｔｅデータ
量、又は記憶媒体の状態に応じてＲＡＩＤストライプの構成を変更するＲＡＩＤ構成変更
部（図２の２８）と、

ＲＡＩＤストライプの構成変更によって新たにアクセス対象となったチャンクに対して
アクセス要求を発行するアクセス再発行部（図２の２９）を備える。

【００４８】

アクセス実行部（図２の２２）は、ホストからのアクセス要求を各記憶媒体（図２の２
１）に対応するＲｅａｄ又はＷｒｉｔｅアクセスに変換する。また、Ｗｒｉｔｅアクセス
の場合は、同時に、パリティ等の冗長データを更新するために、記憶媒体に対して必要な
アクセスを生成する。更に、変換又は生成されたアクセスの実行状況を管理して、アクセ
ス要求に関する処理の実行完了の応答を、ホストに対して行う。

【００４９】

記憶媒体アクセス部（図２の２３）は、アクセス実行部（図２の２２）によって変換又
は生成されたアクセス要求を記憶媒体に対して発行し、結果をアクセス実行部（図２の２
２）に応答として返す。

【００５０】

チャンク割当部（図２の２４）は、各記憶媒体の記憶領域を一定のチャンク単位に分割
して、各チャンクの使用状況を管理する。また、チャンク割当部（図２の２４）は、ＲＡ
ＩＤストライプの構成変更に伴うチャンクの割り当て要求に対して、指定された記憶媒体
の未使用のチャンクを割り当てる。

【００５１】

ＲＡＩＤ管理部（図２の２５）は、各ＲＡＩＤストライプを構成するチャンクの組み合
わせ、及び記憶領域全体のアドレス空間に対するチャンクのマッピングを管理する。

【００５２】

Ｗｒｉｔｅ量計測部（図２の２６）は、各記憶媒体に対して実際に発行されたＷｒｉｔ
ｅアクセス要求の情報を参照して、記憶媒体毎のＷｒｉｔｅデータ量を計測する。

【００５３】

記憶媒体監視部（図２の２７）は、記憶媒体アクセス部（図２の２３）を通して、各記
憶媒体に対する発行待ち、あるいは、発行中のアクセス要求から、負荷を算出し、高負荷
状態（算出された負荷が所定の閾値以上）の記憶媒体を検出する。また、記憶媒体監視部
（図２の２７）は、各記憶媒体のアクセスに対する応答状況を監視して応答遅延、あるい
は停止状態の記憶媒体を検出する。

【００５４】

ＲＡＩＤ構成変更部（図２の２８）は、記憶媒体アクセス部（図２の２３）によって各
記憶媒体にＷｒｉｔｅアクセス要求を発行するにあたり、例えば、発行の直前に、Ｗｒｉ
ｔｅ先（Ｗｒｉｔｅアクセス対象のアドレス）の記憶媒体のＷｒｉｔｅデータ量と、アクセ
ス負荷、応答遅延、あるいは、停止の有無を確認する。Ｗｒｉｔｅ先の記憶媒体につい
て、他の記憶媒体とのＷｒｉｔｅデータ量の差が事前に決定された条件を満たすか、又は
、Ｗｒｉｔｅ先の記憶媒体について、高負荷、応答遅延、応答停止のいずれかの条件を満
たした場合は、Ｗｒｉｔｅアクセス要求先となるＲＡＩＤストライプについて、Ｗｒｉ
ｔｅ先の記憶媒体が、前記条件を満たさない他の記憶媒体となるように、チャンクの構成を
変更する。

【００５５】

10

20

30

40

50

アクセス再発行部（図２の２９）は、Write先の記憶媒体において、他の記憶媒体とのWriteデータ量の差が事前に決定された条件を満たすか、又は、Write先の記憶媒体について高負荷、応答遅延、応答停止のいずれかの条件を満たしている場合には、

- ・記憶媒体アクセス部（図２の２３）に対して、Write先の記憶媒体へのWriteアクセス要求のキャンセルを指示し、
- ・アクセス実行部（図２の２２）に対して、割り当てを変更したチャンクに対するアクセス要求の再発行を指示する。

【００５６】

本発明によれば、RAIDによる記憶媒体間のデータ冗長化を実現しつつ、各記憶媒体のWriteデータ量の調整、及び記憶媒体の一時的な応答遅延、又は応答停止の影響を受けずに、複数の記憶媒体からなる記憶領域の構成を実現することができる。以下、実施形態の詳細について図面を参照して説明する。

【００５７】

<システム構成例>

図１は、本発明の一実施形態の構成を例示する図である。図１を参照すると、本発明の一実施形態においてシステム全体の構成はデータにアクセスする１台以上のホスト１、１つ又は複数の記憶媒体から構成されるストレージ２、ホスト１とストレージ２を接続するネットワーク３を含む。ネットワーク３は、ストレージ２に対して複数のホスト１が接続可能な形態であってもよい。あるいは、ホスト１とストレージ２がネットワーク３により直接接続される形態であってもよい。またストレージ２は、ネットワーク３によって接続された複数の装置又はシステム（分散システム等）によって構成してもよい。

【００５８】

<ストレージの構成例>

図１のストレージ２の構成の一例について説明する。図２を参照すると、ストレージ２は、記憶媒体２１と、アクセス実行部２２と、記憶媒体アクセス部２３と、チャンク割当部２４と、RAID管理部２５と、Write量計測部２６と、記憶媒体監視部２７と、RAID構成変更部２８と、アクセス再発行部２９を備える。

【００５９】

記憶媒体２１は、ストレージ２の記憶領域全体のデータを格納し、NANDフラッシュメモリで構成されたSSDの複数の記憶媒体（＃１、＃２、・・・＃ｎ、ｎは２以上の整数）から構成される。複数の記憶媒体の各々は、NANDフラッシュメモリで構成されたSSDからなり（SSD＃１～SSD＃ｎ）、RAIDを構成する。

【００６０】

アクセス実行部２２は、ホスト１からのアクセス要求を受け取り、RAID管理部２５からアクセス先アドレスが含まれるRAIDストライプの構成情報を取得して、対応する記憶媒体２１へのアクセス要求に変換する。また、アクセス実行部２２は、ホスト１からのアクセス要求がWriteである場合は、RAIDストライプ内の冗長データの更新を行うアクセス要求を生成する。アクセス実行部２２は、ホスト１から受け取ったアクセス要求の情報、及び変換、生成された記憶媒体２１に対するアクセス要求の対応関係を、アクセス実行部２２内のアクセス要求テーブル２２１に格納する。更に、アクセス実行部２２は、ホスト１からの各アクセス要求に対応する処理の実行を管理し、記憶媒体２１に対するアクセス要求の発行を記憶媒体アクセス部２３に指示し、記憶媒体２１に対するアクセス結果を受け取ってホスト１に対して結果を応答する。

【００６１】

図３は、アクセス実行部２２に含まれるアクセス要求テーブル２２１の構成を例示する図である。アクセス要求テーブル２２１は、アクセスの種類、アクセス先アドレス（先頭の０ｘはヘキサデシマル（１６進）表示であることを示す）、アクセスサイズ、対応するアクセスIDの各欄を含む。対応するアクセスIDは、変換、又は生成された記憶媒体２１に対するアクセスを表すID（識別子）である。アクセス実行部２２は、ホスト１から

の各アクセス要求に対して、アクセス要求に関する情報をアクセス要求テーブル 221 に格納する。また、図 3 においては、図示していないが、Write アクセス要求に付随する Write データは別途管理する。なお、アクセス要求テーブル 221 のデータ形式は、図 3 に示した形式に限定されるものでないことは勿論である。例えば記憶媒体 21 に対するアクセスを表す識別子は数字の ID に限らず、任意の形式であってよい。また、アクセス要求テーブル 221 は、アクセス実行部 22 等からアクセスできれば、アクセス実行部 22 の外部に設ける構成としてもよいことは勿論である。

【0062】

再び図 2 を参照すると、記憶媒体アクセス部 23 は、アクセス実行部 22 によって変換、生成された記憶媒体 21 に対するアクセス要求を、実際に記憶媒体 21 に対して発行する。また、記憶媒体アクセス部 23 は、記憶媒体 21 に対する各アクセス要求の情報を、記憶媒体アクセステーブル 231 に格納し、発行待ちのアクセス要求を管理する。

【0063】

図 4 は、記憶媒体アクセス部 23 に含まれる記憶媒体アクセステーブル 231 の構成を例示する図である。記憶媒体アクセステーブル 231 は、アクセス ID、記憶媒体 ID、アクセス種類、アドレス、サイズの各欄を備えている。アクセス ID は、アクセス実行部 22 によって変換、生成された記憶媒体 21 に対するアクセス要求を一意に表す ID 等の識別子である。記憶媒体 ID は、アクセス要求の発行先となる記憶媒体を表す ID 等の識別子である。アクセス種類は、アクセスの種類 (Read, Write) である。アドレス、サイズは、アクセス先アドレスとアクセスサイズである。また、図 4 においては、図示していないが、Write アクセス要求に付随する Write データは別途管理する。なお、記憶媒体アクセステーブル 231 のデータ形式は、図 4 に示した形式に限定されるものでないことは勿論である。例えばアドレス、及びアクセスサイズの表記は、図 4 に示されるチャンク単位以外に、バイト単位、一般的な記憶媒体のアクセス単位であるセクタ (512B) 単位などの任意の形式であって良い。記憶媒体アクセステーブル 231 は、記憶媒体アクセス部 23 等からアクセスできれば、記憶媒体アクセス部 23 の外部に設ける構成としてもよいことは勿論である。

【0064】

再び図 2 を参照すると、チャンク割当部 24 は、各記憶媒体 21 の記憶領域を一定サイズのチャンクに区切って管理し、各記憶媒体 21 において、チャンク毎の RAID ストラップに対する割り当て有無を管理する。また、チャンク割当部 24 は、RAID 構成変更部 28 が RAID ストラップに対して割り当てるチャンクを変更する場合、RAID 構成変更部 28 によって指定された記憶媒体 21 の記憶領域から、未使用のチャンクを割り当てる。

【0065】

RAID 管理部 25 は、RAID ストラップに対する各記憶媒体のチャンクの割り当て情報を RAID 管理テーブル 251 として管理する。

【0066】

図 5 は、RAID 管理部 25 に含まれる RAID 管理テーブル 251 の構成を例示する図である。RAID 管理テーブル 251 は、アドレス、RAID レベル、チャンク数、使用チャンク (記憶媒体 ID - アドレス) の各欄を含む。アドレスは、ストレージ 2 がホスト 1 に対して提供する、記憶領域の空間における、RAID ストラップのマッピング先アドレスである。チャンク数は、RAID ストラップを構成するチャンク数である。使用チャンク (記憶媒体 ID - アドレス) は、各 RAID ストラップを構成するチャンクの記憶媒体の ID と記憶媒体内のアドレスの組み合わせである。RAID レベルが一定ではない場合は、RAID ストラップ毎に、前記各情報を追加で記録する。なお、RAID 管理テーブル 251 のデータ形式は、図 5 に示した形式に限定されるものでないことは勿論である。例えば各 RAID ストラップのサイズが一定であり、且つ、ストレージ 2 の記憶領域の先頭アドレスから順番にマッピングされた RAID ストラップの情報が格納されている場合、RAID ストラップのマッピング先アドレスの情報 (第 1 欄) を省略することが可

10

20

30

40

50

能である。また、R A I D 管理テーブル 2 5 1 は、R A I D 管理部 2 5 等からアクセスできれば、R A I D 管理部 2 5 の外部に設ける構成としてもよいことは勿論である。

【 0 0 6 7 】

再び図 2 を参照すると、W r i t e 量計測部 2 6 は、記憶媒体アクセス部 2 3 から実際に各記憶媒体 2 1 に対して発行された W r i t e アクセスを監視し、各記憶媒体 2 1 に対する W r i t e データ量を集計する。

【 0 0 6 8 】

記憶媒体監視部 2 7 は、記憶媒体アクセス部 2 3 が各記憶媒体 2 1 に対して行ったアクセスに対する各記憶媒体 2 1 からの応答時間を監視し、応答遅延、あるいは応答停止を検出する。

10

【 0 0 6 9 】

また記憶媒体監視部 2 7 は、応答遅延、応答停止の検出と併せて、アクセス要求の発行待ち状況、発行中のアクセス要求のいずれか、又は両方から各記憶媒体 2 1 のアクセス負荷を算出するようにしてもよい。

【 0 0 7 0 】

R A I D 構成変更部 2 8 は、記憶媒体アクセス部 2 3 にて記憶媒体 2 1 に対して発行直前の W r i t e アクセス要求に着目し、W r i t e 先の記憶媒体 2 1 に対して、W r i t e 量計測部 2 6、及び記憶媒体監視部 2 7 の情報から、データの格納先として、当該 W r i t e 先の記憶媒体 2 1 が適切であるか否かを確認する。

【 0 0 7 1 】

R A I D 構成変更部 2 8 は、確認の結果、W r i t e 先の記憶媒体 2 1 が、データの格納先として適していないと判断した場合には、R A I D ストライプにおける、W r i t e 先のチャンクの割り当て元である記憶媒体 2 1 を変更する。

20

【 0 0 7 2 】

具体的には、例えば、W r i t e 先の記憶媒体 2 1 が、応答遅延、応答停止、又は高負荷のいずれかの状態である場合、新たにデータを格納することは困難であると判断して、W r i t e 先のチャンクを、上記応答遅延、応答停止、又は高負荷のいずれにも該当しない別の記憶媒体 2 1 のものに変更する。

【 0 0 7 3 】

あるいは R A I D 構成変更部 2 8 は、W r i t e 先の記憶媒体 2 1 に対する W r i t e データ量について、W r i t e データ量が最も少ないか、又は、最も多い他の記憶媒体 2 1 と比較して、差分が事前に定めた上限値、又は、下限値を超えた場合に、W r i t e 先のチャンクを、別の記憶媒体 2 1 のものに変更する構成としてもよい。

30

【 0 0 7 4 】

アクセス再発行部 2 9 は、R A I D 構成変更部 2 8 において、R A I D ストライプの構成を変更しなかった場合には、記憶媒体アクセス部 2 3 に対して、発行直前の W r i t e アクセス要求に対する実際の発行を許可する。

【 0 0 7 5 】

一方、R A I D 構成変更部 2 8 において、R A I D ストライプの構成が変更された場合には、アクセス再発行部 2 9 は、

40

- ・記憶媒体アクセス部 2 3 に対して、発行直前の W r i t e アクセス要求のキャンセルを指示し、
- ・アクセス実行部 2 2 に対して、割り当てを変更した R A I D ストライプ内のチャンクに対する W r i t e 要求の再発行を指示する。

【 0 0 7 6 】

< 一実施形態の動作例 >

図 6 は、本発明の一実施形態においてホスト 1 からストレージ 2 がアクセス要求を受け取り、結果を応答する動作を説明する流れ図である。以下、図 6、及び図 1 乃至図 5 を参照して、実施形態の動作を説明する。

【 0 0 7 7 】

50

<ステップ S 1 1 >

アクセス実行部 2 2 は、ホスト 1 からストレージ 2 に対するアクセス要求を受け取り、アクセス要求テーブル 2 2 1 に登録する。

【0078】

<ステップ S 1 2 >

アクセス実行部 2 2 は、R A I D 管理部 2 5 の R A I D 管理テーブル 2 5 1 を参照して、アクセス先となる R A I D ストライプを決定し、R A I D ストライプを構成するチャンクに対する R e a d、W r i t e アクセスに変換する。

【0079】

チャンクに対する R e a d、W r i t e アクセスへの変換手順は、一般に知られる R A I D と同様であり、本発明には含まれない。アクセス実行部 2 2 は、変換した R e a d、W r i t e アクセスの実行を記憶媒体アクセス部 2 3 に要求する。

【0080】

また、アクセス実行部 2 2 は、ホスト 1 からのアクセス要求を変換後の、チャンクに対するアクセスの I D をアクセス要求テーブル 2 2 1 に登録する。

【0081】

<ステップ S 1 3 >

記憶媒体アクセス部 2 3 は、アクセス実行部 2 2 から受け取った、チャンクに対するアクセス要求を記憶媒体 2 1 に対して発行する。

【0082】

<ステップ S 1 4 >

記憶媒体アクセス部 2 3 は、記憶媒体 2 1 からのアクセス要求に対する結果（応答）をアクセス実行部 2 2 に応答する。

【0083】

<ステップ S 1 5 >

アクセス実行部 2 2 は、記憶媒体アクセス部 2 3 からのアクセス要求の結果を受け取ると、アクセス要求テーブル 2 2 1 を参照して、ホスト 1 からのアクセス要求に対応した全てのアクセスが完了しているか確認する。

【0084】

<ステップ S 1 6 >

ホスト 1 からのアクセス要求に対応したすべてのアクセスが完了していない場合（ステップ S 1 6 の N o 判定）、アクセス実行部 2 2 は他のアクセス完了まで待機して、ステップ S 1 4 からの処理を実行する。

【0085】

<ステップ S 1 7 >

ホスト 1 からのアクセス要求に対応したすべてのアクセスが完了している場合（ステップ S 1 6 の Y e s 判定）、アクセス実行部 2 2 は、ホスト 1 にアクセス要求の結果を応答して返し、アクセス要求の情報を、アクセス要求テーブル 2 2 1 から削除し、手順を終了する。

【0086】

以上のステップ S 1 1 から S 1 7 までの処理を行い、ホスト 1 からストレージ 2 がアクセス要求を受け取り、結果を応答する動作が完了する。

【0087】

<アクセス要求の発行と結果の応答の手順>

図 7 及び図 8 は、図 6 のステップ S 1 3 において、記憶媒体アクセス部 2 3 がアクセス実行部 2 2 から各記憶媒体 2 1 に対するアクセス要求を受け取り、各記憶媒体 2 1 に対するアクセス要求を発行する手順を説明する流れ図である。なお、図 7 及び図 8 は、図面作成の都合で分図されている。図 7 及び図 8、図 1 乃至図 6 を参照して、アクセス要求の発行と結果の応答の手順を説明する。

【0088】

10

20

30

40

50

<ステップ S 2 0 1 >

記憶媒体アクセス部 2 3 は、アクセス実行部 2 2 から受け取った各記憶媒体 2 1 に対するアクセス要求を記憶媒体アクセステーブル 2 3 1 に登録する。

【 0 0 8 9 】

<ステップ S 2 0 2 >

記憶媒体アクセス部 2 3 は、記憶媒体アクセステーブル 2 3 1 に登録されたアクセス要求から、各記憶媒体 2 1 に対して発行するアクセス要求を選択する。

【 0 0 9 0 】

<ステップ S 2 0 3 >

記憶媒体アクセス部 2 3 は、ステップ S 2 0 2 で選択したアクセス要求が R e a d であるか否かを判定する。

【 0 0 9 1 】

<ステップ S 2 0 4 >

ステップ S 2 0 2 で選択したアクセス要求が R e a d である場合（ステップ S 2 0 3 の Y e s 判定）、記憶媒体アクセス部 2 3 は、選択したアクセス要求を記憶媒体 2 1 に発行して手順を終了する。

【 0 0 9 2 】

<ステップ S 2 0 5 >

ステップ S 2 0 2 で選択したアクセス要求が W r i t e である場合（ステップ S 2 0 3 の N o 判定）、記憶媒体アクセス部 2 3 は、R A I D 構成変更部 2 8 に対して W r i t e 先の記憶媒体 2 1、及び W r i t e アドレス、サイズの情報を通知する。

【 0 0 9 3 】

<ステップ S 2 0 6 >

R A I D 構成変更部 2 8 は、記憶媒体アクセス部 2 3 から通知された W r i t e アクセス要求の発行先記憶媒体 2 1 について、記憶媒体監視部 2 7 に応答遅延、又は応答停止が発生していないか問い合わせる。

【 0 0 9 4 】

また、R A I D 構成変更部 2 8 は、記憶媒体アクセス部 2 3 の記憶媒体アクセステーブル 2 3 1 を参照して、W r i t e アクセス要求先である記憶媒体 2 1 が高負荷状態であるか否かを確認する。

【 0 0 9 5 】

<ステップ S 2 0 7 >

ステップ S 2 0 6 の判断の結果、W r i t e 先の記憶媒体 2 1 が応答遅延、応答停止、高負荷のいずれかの状態であると判定された場合（ステップ S 2 0 7 の Y e s 判定）、ステップ S 2 0 8（図 8）に進む。ステップ S 2 0 6 の判断の結果、W r i t e 先の記憶媒体 2 1 が応答遅延、応答停止、高負荷の状態ではない場合（ステップ S 2 0 7 の N o 判定）、ステップ S 2 0 9（図 8）に進む。

【 0 0 9 6 】

<ステップ S 2 0 8（図 8）>

R A I D 構成変更部 2 8 は、記憶媒体アクセス部 2 3 から通知された W r i t e アクセス要求の新たな発行先を現在の発行先以外の記憶媒体 2 1 から選択する。

【 0 0 9 7 】

R A I D 構成変更部 2 8 は、記憶媒体監視部 2 7 を参照して、応答遅延、応答停止、高負荷のいずれでもない記憶媒体から、W r i t e 量計測部 2 6 から取得した各記憶媒体 2 1 の W r i t e データ量を比較して、最も W r i t e データ量の少ない記憶媒体を新たな W r i t e 先として選択してステップ S 2 1 1 以降の処理を実行する。

【 0 0 9 8 】

<ステップ S 2 0 9（図 8）>

R A I D 構成変更部 2 8 は、記憶媒体アクセス部 2 3 から通知された W r i t e アクセス要求の発行先の記憶媒体 2 1 について、W r i t e 量計測部 2 6 から取得した各記憶媒

10

20

30

40

50

体 2 1 の W r i t e データ量を比較して、W r i t e 先として、他に適切な記憶媒体 2 1 が存在するか判断する。

【 0 0 9 9 】

また、R A I D 構成変更部 2 8 は、上記判断において、記憶媒体監視部 2 7、及び記憶媒体アクセス部 2 3 の記憶媒体アクセステーブル 2 3 1 を参照して、応答遅延、応答停止、高負荷の状態ではない記憶媒体 2 1 のみを、適切な W r i t e 先として、判断する。

【 0 1 0 0 】

前記手順の一例として、例えば W r i t e 先の記憶媒体の W r i t e データ量と、各記憶媒体 2 1 の W r i t e データ量の最小値とを比較して、W r i t e 先の記憶媒体の W r i t e データ量が事前に定められた閾値を上回る場合に、W r i t e データ量が最小の記憶媒体 2 1 を、W r i t e 先として適切と判断するようにしてもよい。

10

【 0 1 0 1 】

< ステップ S 2 1 0 (図 8) >

ステップ S 2 0 9 の判断の結果、他に適切な W r i t e 先となる記憶媒体 2 1 が存在する場合 (ステップ S 2 1 0 の Y e s 判定)、R A I D 構成変更部 2 8 は、ステップ S 2 0 9 で判断された適切な W r i t e 先となる記憶媒体 2 1 を、新たな W r i t e 先としてステップ S 2 1 1 以降の処理を継続する。

【 0 1 0 2 】

< ステップ S 2 1 1 (図 8) >

R A I D 構成変更部 2 8 は、チャンク割当部 2 4 に対して、新たな W r i t e 先となる記憶媒体 2 1 について未使用のチャンクの割り当てを要求して、チャンク割当部 2 4 から割り当てられたチャンクのアドレスを受け取る。また、R A I D 構成変更部 2 8 は、チャンク割当部 2 4 に対して、チャンクを新たに割り当てることで使用されなくなった、元の W r i t e 先のチャンクに対する解放を通知する。

20

【 0 1 0 3 】

< ステップ S 2 1 2 (図 8) >

R A I D 構成変更部 2 8 は、R A I D 管理部 2 5 に対して、ステップ S 2 0 2 で選択された W r i t e アクセス要求の W r i t e 先チャンクを、新たな W r i t e 先の記憶媒体 2 1、及びステップ S 2 1 3 で割り当てられたチャンクのアドレスに変更する。

【 0 1 0 4 】

< ステップ S 2 1 3 (図 8) >

R A I D 構成変更部 2 8 は、アクセス再発行部 2 9 に対してステップ S 2 0 2 で選択された W r i t e アクセス要求の再発行を指示する。

30

【 0 1 0 5 】

アクセス再発行部 2 9 は、記憶媒体アクセス部 2 3 に対して、ステップ S 2 0 2 で選択された W r i t e アクセス要求の中止を指示する。また、アクセス再発行部 2 9 は、アクセス実行部 2 2 に対して、ステップ S 2 1 3 で変更した R A I D ストラップの構成に従ってステップ S 2 0 2 で選択された W r i t e アクセス要求の再発行を指示して手順を終了する。

【 0 1 0 6 】

< ステップ S 2 1 4 (図 8) >

ステップ S 2 0 9 の判断の結果、他に適切な W r i t e 先となる記憶媒体 2 1 が存在せず、現在の W r i t e 先が最適と判断された場合 (ステップ S 2 1 0 の N o 判定)、R A I D 構成変更部 2 8 は、記憶媒体アクセス部 2 3 を通して、ステップ S 2 0 2 で選択された W r i t e アクセス要求の記憶媒体 2 1 に対する発行を行う。また、記憶媒体アクセス部 2 3 は、W r i t e 量計測部 2 6 に対して W r i t e 要求先の記憶媒体 2 1 に対して新たに W r i t e データのサイズ分の W r i t e が行われることを通知する。

40

【 0 1 0 7 】

< ステップ S 2 1 5 (図 8) >

記憶媒体 2 1 に対して発行した W r i t e アクセスの応答が返らない状態で事前に定め

50

られた一定時間を経過した場合（ステップS 2 1 5のN o判定）、記憶媒体アクセス部 2 3は、前記発行中のアクセスに対して、ステップS 2 0 8の処理を実行する。

【0 1 0 8】

記憶媒体 2 1に対して発行したW r i t eアクセスについて事前に定められた一定時間内に応答が返ってきた場合（ステップS 2 1 5のY e s判定）、手順を終了する。

【0 1 0 9】

以上のステップS 2 0 1からS 2 1 5までの処理を行うことで、記憶媒体アクセス部 2 3がアクセス実行部 2 2から各記憶媒体 2 1に対するアクセス要求を受け取り、各記憶媒体 2 1に対するアクセス要求を発行する動作が完了する。

【0 1 1 0】

< R A I Dストライプの構成変更の一具体例 >

図 9乃至図 1 1を参照して、W r i t e発行時の記憶媒体間のW r i t eデータ量の差に応じた、R A I Dストライプの構成変更を行う動作例を説明する。図 9は、図 2の記憶媒体 2 1（図 1のストレージ 2に含まれる）を表している。ストレージ 2は、同種の 6 個の記憶媒体 2 1 A ~ 1 1 Fを備えている。ストレージ 2、ホスト 1は、チャンクサイズを単位としたアドレス空間を使用する。記憶媒体 2 1 A ~ 2 1 Fは、この例の初期状態において、いずれも応答遅延、応答停止、高負荷の状態ではないとする。

【0 1 1 1】

また、この例では、W r i t e先の記憶媒体と、記憶媒体 2 1 A ~ 2 1 FのW r i t eデータ量の最小値との差が 1 0チャンク以上である場合に、R A I D構成を変更して記憶媒体間のW r i t eデータ量の差を削減するものとする。

【0 1 1 2】

ホスト 1からのアクセス要求を受けたストレージ 2のアクセス実行部 2 2（図 2）において、異なる記憶媒体に分かれたチャンクに対するアクセスに変換される。チャンクに対するアクセスへの変換は、図 1 0に示された、R A I D管理テーブル 2 5 1を参照して行う。

【0 1 1 3】

なお、この例では、各 R A I Dストライプの R A I Dレベルは全て R A I D 5、ストライプを構成するチャンク数は全て 5 とする。また、使用チャンクのリストの末尾が各 R A I Dストライプのパリティ情報を格納するチャンクとする。

【0 1 1 4】

ホスト 1からのアクセス要求が、アドレス 0 x 0 0 0 0 0 0 0 3 に対するチャンク 1 個分のW r i t eである場合、図 1 0に例示した R A I D管理テーブル 2 5 1の情報から、記憶媒体 2 1 Dのアドレス 0 x 0 0 0 0 のチャンクに対するW r i t eデータ本体のW r i t eアクセスと、記憶媒体 2 1 Eのアドレス 0 x 0 0 0 0 のチャンクに対するパリティを更新するためのW r i t eアクセスの 2 つに変換される。

【0 1 1 5】

記憶媒体アクセス部 2 3によって、記憶媒体 2 1 Dと記憶媒体 2 1 Eに対するW r i t eアクセスが発行される。

【0 1 1 6】

まず、記憶媒体アクセス部 2 3は、記憶媒体 2 1 Dに対するW r i t eアクセス発行前に、記憶媒体 2 1 Dが応答遅延・応答停止・高負荷状態でないことを確認する。

【0 1 1 7】

記憶媒体 2 1 Dは、応答遅延・応答停止・高負荷状態のいずれでもないため、次に各記憶媒体のW r i t eデータ量を、図 1 1に示した、W r i t e量計測部 2 6が取得した各記憶媒体のW r i t eデータ量情報 2 6 1（例えばW r i t e量計測部 2 6の不図示の記憶部に保持される）から取得する。

【0 1 1 8】

図 1 1の記憶媒体のW r i t eデータ量情報 2 6 1は、各記憶媒体に対するW r i t eデータ量をチャンク数で表す。

10

20

30

40

50

【0119】

Write先である記憶媒体21DのWriteデータ量は20チャンクである。一方で、最小のWriteデータ量である記憶媒体21FのWriteデータ量は10チャンクである。このため、RAIDストライプに割り当てられた記憶媒体21Dのアドレス0x0000のチャンクを解放して、新たに記憶媒体21Fの未使用のチャンク(0x0010とする)を割り当てる。

【0120】

また、RAID管理部25は、RAID管理テーブル251のRAIDストライプの構成情報を更新する。記憶媒体21Dのアドレス0x0000のチャンクに対するWriteを中止して、アクセス再発行部29は、新たに記憶媒体21Fのアドレス0x0010のチャンクに対するWriteアクセスを生成する。

10

【0121】

そして、RAIDストライプのパリティに相当するチャンクを更新する、記憶媒体21Eのアドレス0x0000のチャンクに対するWriteアクセス、及び新たに生成された記憶媒体21Fのアドレス0x0010のチャンクに対するWriteアクセスは、各記憶媒体が応答遅延、応答停止、高負荷のいずれの状態でもなく、記憶媒体のWriteデータ量情報261から、Write先の各記憶媒体のWriteデータ量がWriteデータ量の最小値を、10以上、上回ることはない。この結果、記憶媒体21Eと記憶媒体21Fに対して、Writeアクセスが発行される。

【0122】

20

上記2つのWriteアクセスの完了によって、アドレス0x00000003に対するチャンク1個分のWriteアクセス要求に対するすべてのアクセス(データ書き込みパリティ書き込み)が完了する。ストレージ2のアクセス実行部22は、ホスト1に対してWriteアクセス要求に対する処理の完了を応答する。

【0123】

以上の動作によってWrite発行時の記憶媒体間のWriteデータ量の差に応じた、RAIDストライプの構成変更を伴ったWriteアクセス処理が完了する。

【0124】

上記実施形態によれば、書き込み先の記憶媒体のWriteデータ量、及び、該記憶媒体の応答遅延、停止状態に応じて、より適した記憶媒体に変更し、新たにアクセス対象のチャンクを該記憶媒体に割り当てることで、例えばSSDで構成したRAIDシステムのアクセス性能の低下を回避可能としている。

30

【0125】

なお、上記の特許文献の各開示を、本書に引用をもって繰り込むものとする。本発明の全開示(請求の範囲を含む)の枠内において、さらにその基本的技術思想に基づいて、実施形態ないし実施例の変更・調整が可能である。また、本発明の請求の範囲の枠内において種々の開示要素(各請求項の各要素、各実施例の各要素、各図面の各要素等を含む)の多様な組み合わせ乃至選択が可能である。すなわち、本発明は、請求の範囲を含む全開示、技術的思想にしたがって当業者であればなし得るであろう各種変形、修正を含むことは勿論である。

40

【符号の説明】

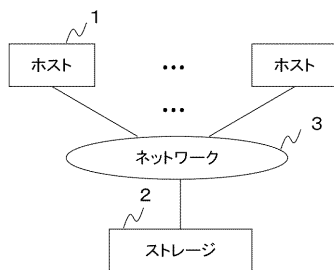
【0126】

- 1 ホスト
- 2 ストレージ
- 21、21A～21F 記憶媒体
- 22 アクセス実行部
- 221 アクセス要求テーブル
- 23 記憶媒体アクセス部
- 231 記憶媒体アクセステーブル
- 24 チャンク割当部

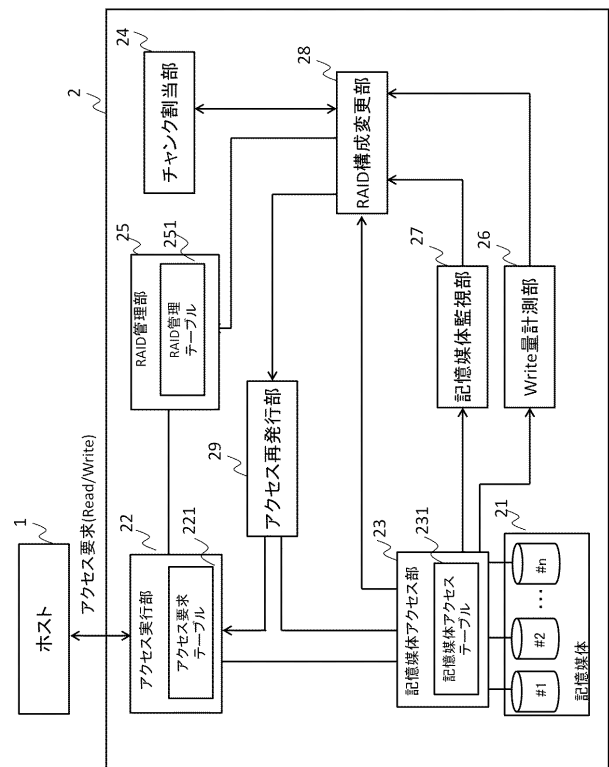
50

- 2 5 R A I D 管理 部
- 2 5 1 R A I D 管理 テーブル
- 2 6 W r i t e 量 計 測 部
- 2 6 1 記憶媒体の W r i t e データ量 情報
- 2 7 記憶媒体 監視 部
- 2 8 R A I D 構成 変更 部
- 2 9 アクセス 再 発行 部
- 3 ネットワーク
- 3 0 構成 変更 手段
- 3 1 アクセス 制御 手段

【 図 1 】



【 図 2 】



【 図 3 】

種類	アドレス	サイズ	対応アクセスID
Read	0x00000000	1	10
Write	0x00000010	1	11, 12
Write	0x00000020	2	13, 14, 15
⋮	⋮	⋮	⋮

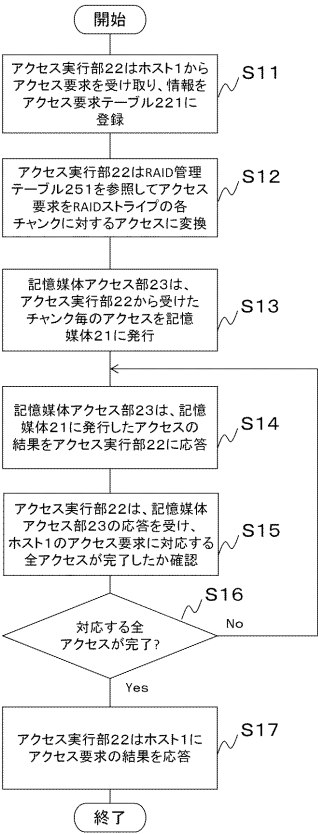
【 図 4 】

アクセスID	記憶媒体ID	アクセス種類	アドレス	サイズ
10	1	Read	0x00000000	1
11	2	Write	0x00000010	1
12	8	Write	0x00000020	1
13	3	Write	0x00000100	1
14	4	Write	0x00000110	1
15	8	Write	0x00000030	1
⋮	⋮	⋮	⋮	⋮

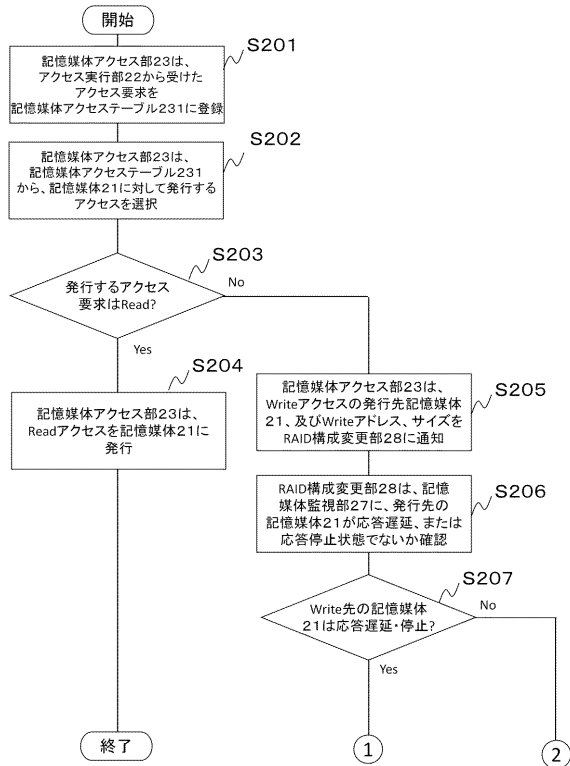
【 図 5 】

アドレス	RAIDレベル	チャンク数	使用チャンク (記憶媒体ID-アドレス)
0x00000000	RAID5	9	(1-0x0000), (2-0x0000), ..., (9-0x0001)
0x00000008	RAID5	9	(1-0x0001), (2-0x0010), ..., (9-0x0101)
0x00000010	RAID6	10	(1-0x0002), (2-0x0011), ..., (10-0x0008)
0x00000018	RAID5	9	(3-0x0020), (2-0x0030), ..., (11-0x0100)
⋮	⋮	⋮	⋮

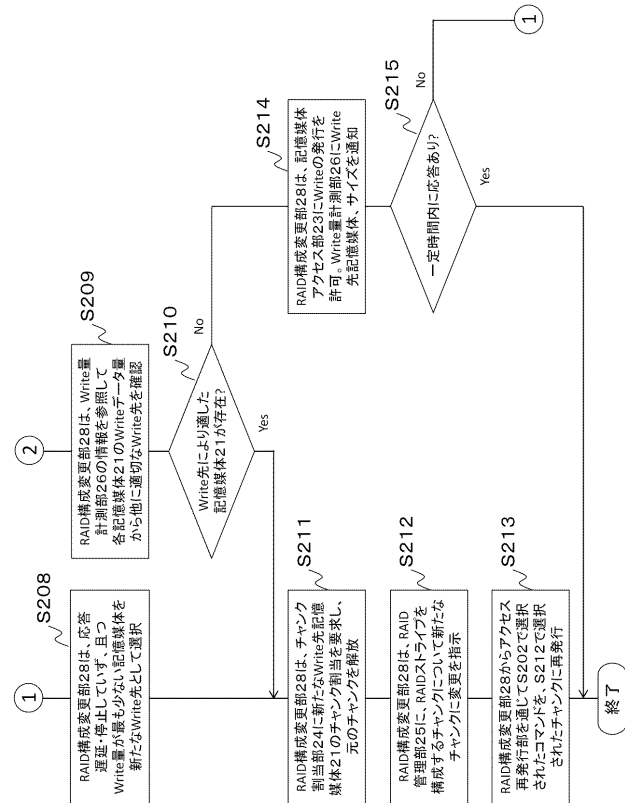
【 図 6 】



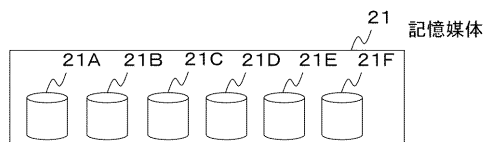
【 図 7 】



【 図 8 】



【 図 9 】



【 図 1 0 】

アドレス	使用チャネル(記憶媒体-アドレス)
0x00000000	(A-0x0000), (B-0x0000), (C-0x0000), (D-0x0000), (E-0x0000)
0x00000004	(A-0x0001), (B-0x0001), (C-0x0001), (D-0x0001), (F-0x0000)
0x00000008	(A-0x0002), (B-0x0002), (D-0x0002), (E-0x0001), (F-0x0001)

【図 1 1】

261

記憶媒体	A	B	C	D	E	F
Writeデータ量	17	11	12	20	18	10

【図 1 2】

