

(19) 日本国特許庁 (JP)

(12) 特 許 公 報 (B2)

(11) 特許番号

特許第5444464号
(P5444464)

(45) 発行日 平成26年3月19日 (2014. 3. 19)

(24) 登録日 平成25年12月27日 (2013. 12. 27)

(51) Int. Cl.	F I
G 0 6 F 3 / 0 6 (2006. 01)	G 0 6 F 3 / 0 6 3 0 1 Z
	G 0 6 F 3 / 0 6 5 4 0

請求項の数 10 (全 36 頁)

(21) 出願番号	特願2012-517600 (P2012-517600)	(73) 特許権者	000005108
(86) (22) 出願日	平成22年1月14日 (2010. 1. 14)		株式会社日立製作所
(65) 公表番号	特表2013-506886 (P2013-506886A)		東京都千代田区丸の内一丁目6番6号
(43) 公表日	平成25年2月28日 (2013. 2. 28)	(74) 代理人	110000279
(86) 国際出願番号	PCT/JP2010/000166		特許業務法人ウィルフォート国際特許事務所
(87) 国際公開番号	W02011/086598	(72) 発明者	石崎 太郎
(87) 国際公開日	平成23年7月21日 (2011. 7. 21)		神奈川県小田原市中里322番2号 株式会社日立製作所 R A I Dシステム事業部内
審査請求日	平成24年4月5日 (2012. 4. 5)	(72) 発明者	鈴木 勝喜
			神奈川県小田原市中里322番2号 株式会社日立製作所 R A I Dシステム事業部内

最終頁に続く

(54) 【発明の名称】 ストレージシステム

(57) 【特許請求の範囲】

【請求項1】

ホスト装置と接続されたストレージシステムであって、
前記ホスト装置に仮想ボリュームを提供し、前記ホスト装置からの前記仮想ボリュームに対するライトコマンドとリードコマンドとを処理するコントローラと、
複数の物理記憶装置と、
を有し、
前記コントローラは、
前記仮想ボリュームの有する複数の仮想記憶領域のうちの第1仮想記憶領域へのデータのライトコマンドを受領すると、前記第1仮想記憶領域に、前記複数の物理記憶装置の有する複数の実記憶領域のいずれも割当てられていない場合に、
前記仮想ボリュームについて設定された R A I D レベルとコンビネーションとに基づいた数の物理記憶装置である 2 以上の物理記憶装置を選択し、
前記 2 以上の物理記憶装置から、それぞれ、いずれの仮想記憶領域にも割当てられていない 2 以上の実記憶領域を選択し、
前記選択された 2 以上の実記憶領域を前記第1仮想記憶領域に割り当て、
当該割当てられた 2 以上の実記憶領域に、前記 R A I D レベルまたはコンビネーションにしたがって、前記データに基づく 2 以上のデータを書き込み、
前記コントローラは、前記物理記憶装置毎に、前記物理記憶装置の有する実記憶領域毎に対するアクセス頻度を表すアクセス頻度情報を有しており、

10

20

前記コントローラは、

アクセス頻度情報に基づいて得られる物理記憶装置のアクセス頻度が第1の閾値を超過している物理記憶装置を選択し、その物理記憶装置に対応したアクセス頻度情報を基に、その物理記憶装置の有する実記憶領域の中でアクセス頻度が最も高い第1実記憶領域を選択し、

アクセス頻度情報に基づいて得られる物理記憶装置のアクセス頻度が最も低い物理記憶装置を選択し、その物理記憶装置の有する複数の実記憶領域のうち、いずれの前記仮想記憶領域にも割り当てられていない第2実記憶領域を選択し、前記第1実記憶領域に格納されているデータを前記第2実記憶領域に移行し、前記第1実記憶領域の割当て先の仮想記憶領域に、前記第1実記憶領域に代えて前記第2実記憶領域を割り当てる

ことを特徴とするストレージシステム。

10

【請求項2】

請求項1に記載のストレージシステムであって、

前記コントローラは、前記複数の仮想記憶領域それぞれと、前記複数の仮想記録領域それぞれに割当てられた前記複数の実記憶領域のいずれかとの対応付けの情報である第1情報を有し、

前記コントローラは、前記第1仮想記憶領域に、前記複数の実記憶領域のいずれかが割当てられた場合、前記第1仮想記憶領域と前記割当てられた実記憶領域との対応付けの情報を新たに第1情報に記憶する

ことを特徴とするストレージシステム。

20

【請求項3】

請求項1に記載のストレージシステムであって、

複数の前記物理記憶装置は、前記物理記憶装置の管理単位であるプールを形成し、

前記コントローラは、前記ホストに対して、前記仮想ボリュームに設定された前記RAIDレベルまたは前記コンビネーションとは異なるRAIDレベルまたはコンビネーションが設定された他の仮想ボリュームを提供し、

前記コントローラは、前記第1仮想記憶領域に前記実記憶領域を割当てる場合、前記プールに含まれる前記物理記憶装置を選択し、

前記コントローラは、前記他の仮想ボリュームが有する第2仮想記憶領域に前記実記憶領域を割当てる場合、前記プールに含まれる前記物理記憶装置を選択する

ことを特徴とするストレージシステム。

30

【請求項4】

請求項3に記載のストレージシステムであって、

前記第1仮想記憶領域と、前記第2仮想記憶領域とは同一の物理記憶装置が有する異なる前記実記憶領域が割当てられる

ことを特徴とするストレージシステム。

【請求項5】

請求項1に記載のストレージシステムであって、

前記ストレージシステムに新たに物理記憶装置が追加された場合、

前記コントローラは、アクセス頻度情報に基づいて得られる物理記憶装置のアクセス頻度が前記第1の閾値を超過している物理記憶装置を選択し、その物理記憶装置に対応したアクセス頻度情報を基に、その物理記憶装置の有する実記憶領域の中でアクセス頻度が最も高い第3実記憶領域を選択し、前記第3実記憶領域に格納されているデータを、前記新たに追加された物理記憶装置が有するいずれかの実記憶領域に移行する

ことを特徴とするストレージシステム。

40

【請求項6】

請求項5に記載のストレージシステムであって、

前記コントローラは、前記ストレージシステムが有する前記複数の物理記憶装置の情報である第2情報を有し、

前記コントローラは、前記新たに追加された物理記憶装置の情報を前記第2情報に記憶

50

する
ことを特徴とするストレージシステム。

【請求項 7】

請求項 1 に記載のストレージシステムであって、
前記物理記憶装置に障害が発生すると、
前記コントローラは、

障害が発生した物理記憶装置の実記憶領域が割り当てられている仮想記憶領域に割り当てられている残りの 1 以上の実記憶領域にそれぞれ記憶されている 1 以上のデータから、前記障害が発生した物理記憶装置の実記憶領域に記憶されているデータを修復し、

前記複数の物理記憶装置の有する前記複数の実記憶領域のうちから、いずれの前記仮想記憶領域にも割り当てられていない第 4 実記憶領域を選択し、

修復後のデータを前記第 4 実記憶領域に移行することを特徴とするストレージシステム。

【請求項 8】

ホスト装置と接続されたストレージシステムの制御方法であって、
前記ストレージシステムは、

前記ホスト装置に仮想ボリュームを提供し、前記ホスト装置からの前記仮想ボリュームに対するライトコマンドとリードコマンドとを処理するコントローラと、

複数の物理記憶装置と、
を備え、

前記コントローラは、

前記仮想ボリュームの有する複数の仮想記憶領域のうちの第 1 仮想記憶領域へのデータのライトコマンドを受領し、

前記第 1 仮想記憶領域に、前記複数の物理記憶装置の有する複数の実記憶領域のいずれも割当ていない場合に、

前記仮想ボリュームについて設定された R A I D レベルとコンビネーションとに基づいた数の物理記憶装置である 2 以上の物理記憶装置を選択し、

前記 2 以上の物理記憶装置から、それぞれ、いずれの仮想記憶領域にも割当てられていない 2 以上の実記憶領域を選択し、

前記第 1 仮想記憶領域に割り当て、

当該割当てられた 2 以上の実記憶領域に、前記 R A I D レベルまたはコンビネーションにしたがって、処理されたデータに基づく 2 以上のデータを書き込み、

前記コントローラは、前記物理記憶装置毎に、前記物理記憶装置の有する実記憶領域毎に対するアクセス頻度を表すアクセス頻度情報を有しており、

前記コントローラは、

アクセス頻度情報に基づいて得られる物理記憶装置のアクセス頻度が第 1 の閾値を超過している物理記憶装置を選択し、その物理記憶装置に対応したアクセス頻度情報を基に、その物理記憶装置の有する実記憶領域の中でアクセス頻度が最も高い第 1 実記憶領域を選択し、

アクセス頻度情報に基づいて得られる物理記憶装置のアクセス頻度が最も低い物理記憶装置を選択し、その物理記憶装置の有する複数の実記憶領域のうち、いずれの前記仮想記憶領域にも割り当てられていない第 2 実記憶領域を選択し、前記第 1 実記憶領域に格納されているデータを前記第 2 実記憶領域に移行し、前記第 1 実記憶領域の割当て先の仮想記憶領域に、前記第 1 実記憶領域に代えて前記第 2 実記憶領域を割り当てる

ことを特徴とするストレージシステムの制御方法。

【請求項 9】

請求項 8 に記載のストレージシステムの制御方法であって、

前記物理記憶装置の管理単位であるプールを形成し、

前記ホストに対して、前記仮想ボリュームに設定された前記 R A I D レベルまたは前記コンビネーションとは異なる R A I D レベルまたはコンビネーションが設定された他の仮

10

20

30

40

50

想ボリュームを提供し、

前記第1仮想記憶領域に前記実記憶領域を割当ててする場合、

前記プールに含まれる前記物理記憶装置を選択し、

前記他の仮想ボリュームが有する第2仮想記憶領域に前記実記憶領域を割当てて場合、

前記プールに含まれる前記物理記憶装置を選択する

ことを特徴とするストレージシステムの制御方法。

【請求項10】

請求項9に記載のストレージシステムの制御方法であって、

前記第2仮想記憶領域に前記実記憶領域を割当てて場合、

少なくとも1つの前記第1仮想記憶領域に割当てた前記実記憶領域を有する前記物理記憶装置と同一の前記物理記憶装置が有する、いずれの前記仮想記憶領域にも割当てられていない前記実記憶領域を選択する

ことを特徴とするストレージシステムの制御方法。

10

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、Dynamic Provisioning (Thin Provisioningとも呼ばれる) の技術が適用される記憶制御に関する。

【背景技術】

20

【0002】

現在、企業等が扱うデータを安全に格納するために、RAID (Redundant Array of Independent (or Inexpensive) Disks) を構成するストレージシステムが広く用いられている。さらに、ストレージシステムの容量の利用効率向上のために、シン・プロビジョニングの導入がすすんでいる。

【0003】

RAIDでは、複数の記憶装置からRAIDグループを構成し、RAIDグループ上に論理ボリュームを構成している。シン・プロビジョニングによらないストレージシステムでは、ホストに対して、論理ボリュームが提供され、論理ボリューム作成時から、当該論理ボリュームに論理ボリュームの全記憶容量に等しい容量の記憶装置の記憶領域が割当てられる。これに対し、ホストに対して仮想ボリュームが提供されるシン・プロビジョニングでは、当該仮想ボリューム内の仮想記憶領域には、当初仮想記憶領域に対応する記憶装置の記憶領域は割当てられておらず、仮想記憶領域に書き込み要求が来た際に、仮想記憶領域に対してだけ論理ボリュームに割当てられている記憶装置の記憶領域を割当てている。このため、システム導入時には当面必要な容量の記憶装置のみを備え、必要に応じて記憶装置を増設していくことが容易にできる。

30

【0004】

シン・プロビジョニングにおける技術として、特許文献1では、シン・プロビジョニングを適用したストレージシステムにおいて、複数の記憶装置上のデータの再配置を論理ボリューム単位で行う方法が提案されている。

40

【0005】

また、特許文献2では、記憶装置を増設後、データの冗長構成を変更することにより、記憶装置を記憶装置単位で増設する方法が提案されている。

【先行技術文献】

【特許文献】

【0006】

【特許文献1】特開2008-234158号公報

【特許文献2】特開2009-230352号公報

【発明の概要】

【発明が解決しようとする課題】

50

【0007】

上記特許文献1では、仮想ボリュームに割当てて前から複数の記憶装置でRAIDグループを構成し、RAIDグループ内の論理ボリューム単位でデータの管理が行われている。たとえば、論理ボリューム内のデータを当該RAIDグループの複数の記憶装置に跨って格納することにより、データの冗長性を保っている。そのため、記憶装置の増設は、RAIDグループを構成するために必要な数の記憶装置をひとまとまりとして、1回の増設で複数の記憶装置の増設を行わなければならない。

【0008】

一方、上記特許文献2の方法でも、上記特許文献1と同様に、仮想ページへの割当て前に、複数の記憶装置でRAIDグループを構成し、その中の論理ボリュームを仮想ボリュームに提供することが前提である。そのことから、単一の記憶装置を増設後、RAIDグループの再構成が必要となり、データの冗長構成を再編する必要があるが生じる。

10

【0009】

本発明の1つの目的は、シン・プロビジョニングを適用したストレージシステムにおいて、記憶装置増設の自由度を向上させ、記憶装置増設時のコストを低減する方法を提供することである。

【0010】

また、本発明の他の目的は、シン・プロビジョニングを適用したストレージシステムにおいて、記憶装置の増設が業務性能へ与える影響を押さえることである。

【0011】

20

さらに、本発明の他の目的は、記憶装置の障害発生に備えた特定の記憶装置であるスペアディスクを必要としないストレージシステムを提供することである。

【課題を解決するための手段】

【0012】

シン・プロビジョニング機能を有するストレージシステムにおいて、複数の記憶装置でRAIDグループを構成せず、RAIDグループ内の論理ボリュームを仮想ボリュームに提供する代わりに、記憶装置の記憶領域を仮想ボリュームに直接提供する。すなわち、ストレージシステムは、仮想記憶領域に書き込み要求を受領したとき、まず、当該仮想記憶領域が含まれる仮想ボリュームに設定されたRAIDレベルとデータの冗長構成（データの分割数と生成するパリティの数）を特定する。特定されたRAIDレベルと冗長構成に応じた数の記憶装置を、当該仮想記憶領域に対して選択する。選択された記憶装置のそれぞれから、いずれの仮想記憶領域にも割当てられていない記憶領域を選択し、当該仮想記憶領域に割当てて。この割当てられた記憶領域に対して、データを分割し、パリティと共に書き込む。

30

【発明の効果】

【0013】

本発明の一実施形態によれば、記憶装置の増設コストを最小限に抑えながら、記憶容量を拡張することができる。

【図面の簡単な説明】

【0014】

40

【図1】本発明の一実施形態に係るストレージシステムの構成を示す。

【図2】本実施形態に係る動的容量拡張の概要を示す。

【図3】DPプールの第1の構成例を示す。

【図4】DPプールの第2の構成例を示す。

【図5】メモリ123に記憶されるコンピュータプログラム及び情報を示す。

【図6】DP共通仕様テーブル503の構成を示す。

【図7】プール管理テーブル505の構成を示す。

【図8】仮想LU管理テーブル507の構成を示す。

【図9】PDEV管理テーブル509の構成を示す。

【図10】実ページ管理テーブル511の構成を示す。

50

【図 1 1】 マッピング管理テーブル 5 1 3 の構成を示す。
 【図 1 2】 実アクセス状況テーブル 5 1 5 の構成を示す。
 【図 1 3】 仮想アクセス状況テーブル 5 1 7 の構成を示す。
 【図 1 4】 D P プールの生成の流れを示すフローチャートである。
 【図 1 5】 仮想 L U 2 0 1 の生成の流れを示すフローチャートである。
 【図 1 6】 仮想ページに対するライトの流れの概要を示す。
 【図 1 7】 仮想ページに対するライトの流れを示すフローチャートである。
 【図 1 8】 仮想ページに対するリードの流れを示すフローチャートである。
 【図 1 9】 複数の R A I D タイプに従って格納されたデータが一つの D P プールに混在している様子を示す。

10

【図 2 0】 一つの仮想ページに割り当てられる P D E V が P D E V 番号に依存せずに選択された場合のデータユニット（データ要素及びパリティの総称）の配置の一例を示す。

【図 2 1】 P D E V の増設の際に行われる処理のフローチャートである。

【図 2 2】 再配置処理のフローチャートを示す。

【図 2 3】 P D E V が増設された場合に行われる再配置の一例を示す。

【図 2 4】 P D E V の減設の際に行われる処理のフローチャートである。

【図 2 5】 アクセス頻度に基づく再配置の一例を示す。

【図 2 6】 P D E V に障害が発生した場合のリストア処理の概要を示す。

【図 2 7】 リストア優先タイプのリストア処理のフローチャートである。

【図 2 8】 交換優先タイプのリストア処理のフローチャートである。

20

【発明を実施するための形態】

【0 0 1 5】

図 1 は、本発明の一実施形態に係るストレージシステムの構成を示す。

【0 0 1 6】

一以上（例えば二つ）のホスト装置（以下、ホスト） 1 0 1 がストレージシステム 1 0 3 に接続されている。ホスト 1 0 1 とストレージシステム 1 0 3 との間の通信は、例えば、ファイバチャネル（F C）又は i S C S I（S C S I は Small Computer System Interface の略）に従うプロトコルで行われる。

【0 0 1 7】

ストレージシステム 1 0 3 は、仮想 L U（L U は Logical Unit の略）を管理する。仮想 L U とは、Dynamic Provisioning（Thin Provisioning と呼ばれる）の技術に従う仮想的な L U である。仮想 L U は、複数の仮想ページ（仮想的な記憶領域）を有する。各仮想ページには、一以上のアドレス（例えば L B A（Logical Block Address））が属する。

30

【0 0 1 8】

ホスト 1 0 1 は、仮想 L U に対する I / O（Input/Output）を行うための I / O コマンドを送信する。I / O コマンドは、I / O 先情報を有する。I / O 先情報は、仮想 L U の L U N（Logical Unit Number）と、仮想 L U の仮想ページのアドレスとを含む。ホスト 1 0 1 は、一般的には計算機であるが、計算機に代えて、別のストレージシステムであってもよい。

【0 0 1 9】

ストレージシステム 1 0 3 は、複数の物理記憶デバイス（P D E V） 1 1 7 と、複数の P D E V 1 1 7 に接続された記憶制御装置（以下、「D K C」と表記する） 1 1 1 とを有する。D K C 1 1 1 は、多重化（例えば二重化）されている。複数の P D E V 1 1 7 は、同種の P D E V で構成されていても良いし、複数種類の P D E V が混在していても良い。P D E V の種類としては、例えば、H D D（Hard Disk Drive）と S S D（Solid State Drive）がある。

40

【0 0 2 0】

D K C 1 1 1 は、ホスト 1 0 1 から I / O コマンド（ライトコマンド又はリードコマンド）を受信し処理する。D K C は、例えば、ライトコマンドを受信した場合、受信したライトコマンドから特定される一以上のライト先の仮想ページの各々に、複数の実ページ（

50

実体的な記憶領域)を割り当てる。以下、仮想ページに実ページを割り当てることを、便宜上、「動的容量拡張」と言う。

【0021】

D K C 1 1 1 に、一以上のバックエンド接続デバイス 1 1 3 を介して、複数の P D E V 1 1 7 が接続されている。バックエンド接続デバイス 1 1 3 は、多重化(例えば二重化)されている。バックエンド接続デバイス 1 1 3 としては、例えば、F C - A L (Fibre Channel-Arbitrated Loop) 或いは S A S (Serial Attached SCSI) エクスパンダなどを採用することができる。一つのバックエンド接続デバイス 1 1 3 に、複数の P D E V 1 1 7 のうちの二以上の P D E V 1 1 7 が接続される。

【0022】

少なくとも一つの D K C 1 1 1 に、管理装置 1 0 5 が接続されている。管理装置 1 0 5 は、例えば L A N (Local Area Network) を通じて、少なくとも一つの D K C 1 1 1 に接続されている。管理装置 1 0 5 は、例えば計算機であり、L U に関する各種設定及び参照を行うことができる。

【0023】

各 D K C 1 1 1 は、F E - I F (フロントエンドインタフェース) 1 2 1 と、B E - I F (バックエンドインタフェース) 1 2 7 と、コントローラとを有する。コントローラは、例えば、転送回路 1 2 9 と、メモリ 1 2 3 と、M P (マイクロプロセッサ) 1 2 5 とを有する。コントローラは、それらの要素に加えて、例えば、M P 1 2 5 に代えてパリティを生成するためのハードウェア回路を有しても良い。

【0024】

転送回路 1 2 9 には、F E - I F 1 2 1、メモリ 1 2 3、M P 1 2 5 及び B E - I F 1 2 7 が接続されている。転送回路 1 2 9 は、それらの要素 1 2 1、1 2 3、1 2 5 及び 1 2 7 間でやり取りされるデータ或いはコマンドを中継する。

【0025】

F E - I F 1 2 1 は、ホスト 1 0 1 に接続されるインタフェース装置である。F E - I F 1 2 1 は、複数の通信ポートを有しており、各通信ポートに、ホスト 1 0 1 が接続される。F E - I F は、ホスト 1 0 1 から、I / O コマンドを受信し、受信した I / O コマンドを、M P 1 2 5 に転送する。

【0026】

B E - I F 1 2 7 は、P D E V 1 1 7 に接続されるインタフェース装置である。B E - I F 1 2 7 は、複数の通信ポートを有しており、各通信ポートに、バックエンド接続デバイス 1 1 3 を通じて P D E V 1 1 7 が接続される。

【0027】

メモリ 1 2 3 は、一以上のメモリで構成されている。メモリ 1 2 3 は、情報(例えばテーブル)及びコンピュータプログラムを記憶する。メモリ 1 2 3 は、キャッシュ領域を有する。キャッシュ領域には、ホスト 1 0 1 と P D E V 1 1 7 との間でやり取りされるデータが一時格納される。

【0028】

M P 1 2 5 は、ホスト 1 0 1 からの I / O コマンドを処理する。また、M P 1 2 5 は、管理装置 1 0 5 からの要求に応答して、情報(例えば L U に関する情報)をメモリ 1 2 3 に設定したり、メモリ 1 2 3 に格納されている情報(例えば L U に関する情報)を管理装置 1 0 5 に提供したりする。

【0029】

図 2 は、本実施形態に係る動的容量拡張の概要を示す。

【0030】

仮想 L U 2 0 1 と D P (Dynamic Provisioning) プール 2 0 3 が D K C 1 1 1 によって管理されている。一つの D P プール 2 0 3 には、仮想 L U 2 0 1 が関連付けられ、D P プール 2 0 3 からその仮想 L U 2 0 1 に実ページが割り当てられる。

【0031】

仮想LU201は、複数の仮想ページで構成されており、ホスト101から認識される。仮想LU201について、RAIDレベルが定義される。仮想LU201のRAIDレベル、コンビネーション及び容量などを、例えば管理装置105（図1参照）から設定可能である。ここで、RAIDレベルとは、RAIDの技術をその機能によって、分類したものであり、RAID5、RAID6のように表される。たとえば、RAID5では、書き込み要求の来たデータを n 個（ n は自然数）に分割し、当該データから1つのパリティデータを作成して、 $(n+1)$ 台の記憶装置に分散して書き込む。また、RAID6では、分割した n 個のデータに対し、2種類のパリティデータを作成し、 $(n+2)$ 台の記憶装置に分散して書き込まれる。また、コンビネーションとは、RAIDを構成するために任意に定められたデータ要素への分割数 n と、RAIDレベルによって定められるパリティの数 m （ m は0以上の整数）との組み合わせであり、 $nD+mP$ (or D) と表される。つまり、仮想ページに書き込まれたデータは、 n 個のデータ要素に分割され、 m 個のパリティが生成され、 $(n+m)$ 台の記憶装置に分散して書き込まれる。

10

【0032】

DPプール203は、ストレージシステム103が有する複数のPDEV117のうちの一部又は全部のPDEV117で構成されている。DPプール203は、多数の実ページで構成されている。

【0033】

ここで、本実施形態の一つの特徴として、DPプール203にはRAIDの定義が無く、実ページが、複数のPDEVで構成されるRAIDグループからではなく、個々のPDEVから提供されることがある。

20

【0034】

一般に、Dynamic Provisioning (Thin Provisioning) の技術によれば、仮想LUに割り当てられる実ページは、予め定義されたRAIDグループの記憶空間を基に形成されている。具体的には、例えば、プールは、一又は複数のRAIDグループで構成されている。各RAIDグループは、RAIDレベル及びコンビネーションに基づき二以上のPDEVで構成されている。実ページは、一つのRAIDグループに基づいて構成されており、一つの実ページが、RAIDグループを構成する二以上のPDEVに予め跨っている。このため、仮想ページに割り当てられた実ページに対するライトでは、ライト対象のデータが、その実ページを提供するRAIDグループのRAIDレベル及びコンビネーションに基づいて書き込まれる。

30

【0035】

それに対し、本実施形態によれば、RAIDグループの定義は無く、DPプール203を構成する個々のPDEV117は、RAIDグループを構成するメンバではない。実ページが、個々のPDEVから提供される。また、RAIDレベル及びコンビネーションは、仮想LU201について設定されている。仮想LU201の仮想ページに、その仮想LU201のRAIDレベル及びコンビネーションに基づいて決定された数の実ページがDPプール203から割り当てられ、その仮想ページに対するライトの対象のデータが、割り当てられた複数の実ページに書き込まれる。

【0036】

40

RAID技術によれば、ライト対象のデータを構成する複数のデータ要素のうちの二以上のデータ要素（或いは、ライト対象のデータ及びそのデータの複製）が、一つのPDEV117に格納されることはない。このため、本実施形態では、一つのPDEV117から二以上の実ページが同一の仮想ページに割り当てられることはない。つまり、一つの仮想ページには、それぞれ異なるPDEV117から実ページが割り当てられる。従って、DPプールに属する複数のPDEVの使用率（PDEVの容量に対する割当て済み実ページの総容量の比率）が均一になるように、仮想ページに複数の実ページが割り当てられることが望ましい。なぜなら、例えば一つのPDEVにたくさんの空きの実ページがあっても、その一つのPDEVから複数の実ページを一つの仮想ページに割り当てることができず、故に、結果として、空きの実ページが余ってしまう可能性があるからである。また、

50

このとき、複数のPDEVの使用率で無く複数のPDEVの空き容量が均一になるように実ページが割当てても良い。

【0037】

仮想LU201及びDPプール203は、一つであっても複数であっても良い。図2の例によれば、仮想LU201として、二つの仮想LU#1及び#2があり、DPプール203として、一つのDPプール#1があり、DPプール#1に仮想LU#1及び#2が関連付けられているとする。また、仮想LU#1のRAIDレベル及びコンビネーションがRAID5(2D+1P)であり、仮想LU#2のRAIDレベル及びコンビネーションがRAID6(4D+2P)であるとする。また、DPプール#1が、六つのPDEV#01～#06で構成されているとする。

10

【0038】

仮想LU#1のRAIDレベル及びコンビネーション：RAID5(2D+1P)によれば、仮想LU#1の仮想ページV11に対するライトの対象となるデータは、DKC111によって、二つのデータ要素に分割され、且つ、一つのパリティが生成される。そのため、仮想ページV11には、三つのPDEV#01～#03から三つの実ページR11、R22及びR31が割り当てられる。DKC111によって、仮想ページV11と実ページR11、R22及びR31との対応関係が管理される。そして、三つの実ページR11、R22及びR31のうちの二つに、二つのデータ要素がそれぞれ書き込まれ、残りの一つに、パリティが書き込まれる。

20

【0039】

また、仮想LU#2のRAIDレベル及びコンビネーション：RAID6(4D+2P)によれば、仮想LU#2の仮想ページV21に対するライトの対象となるデータは、四つのデータ要素に分割され、且つ、二つのパリティが生成される。そのため、仮想ページV21には、六つのPDEV#01～#06から六つの実ページR18、R23、R33、R41、R51及びR61が割り当てられる。DKC111によって、仮想ページV21と実ページR18、R23、R33、R41、R51及びR61との対応関係が管理される。そして、六つの実ページR18、R23、R33、R41、R51及びR61のうちの四つに、四つのデータ要素がそれぞれ書き込まれ、残りの二つに、二つのパリティがそれぞれ書き込まれる。

30

【0040】

一つのDPプール203は、一以上の属性が同じ物理記憶デバイス(PDEV)で構成されていても良いし、属性が異なるPDEVで構成されていても良い。属性としては、例えば、PDEVのI/O性能や信頼性(例えばMTBF(Mean Time Between Failures))がある。PDEVのI/O性能としては、例えば、PDEVそれ自体の種類(例えば、HDD、SSDなど)、インタフェースの種類(例えば、FC(Fibre Channel)、SATA(Serial ATA)、SAS(Serial Attached SCSI))、ディスクの回転速度などがある。

【0041】

図3に示すように、一つのDPプール203は、異なる種類の属性を有するPDEV(図3の例ではHDDとSSD)で構成されていても良い。仮想ページは、異なる属性を有する複数の実ページから構成された場合、各PDEVからデータを書き込みや読み出しを行う際、I/O性能の低いPDEVに合わせられて性能が低下するため各仮想ページは同一の属性を有するPDEVから構成されることが望ましい。しかし、各仮想LUは、1つのプールに所属している。そのため、同一プール内に属性が異なるPDEVを混在させることで、アクセス頻度等に応じて、同一仮想LUの仮想ページごとに属性が異なるPDEV内の実ページを割当てることができる。また、同一仮想ページに対しては同一の属性を有するPDEVから実ページを割当てた場合も、ある属性のPDEVの容量が不足したとき、属性が異なるPDEVから実ページを割当てることができる。そして、当該容量不足の属性のPDEVを増設後に、異なる属性のPDEVに割当てたデータ要素を増設したPDEVに割当てることができる。この種の構成のDPプール203によれば、ディスクの

40

50

容量を効率的に消費することができる。

【0042】

また、図4に示すように、一つのDPプール203は、複数の属性のうちの所定の一種類以上の属性が同じPDEVで構成されても良い。図4の例によれば、PDEVの性能、PDEVのインタフェース、及び、PDEVそれ自体の種類の全てが同じであるPDEVで一つのDPプール203が構成されている。この種の構成のDPプール203によれば、仮想LU201に対するアクセスについて所望のI/O性能を見込むことが出来る。例えば、DPプール#Eに関連付けられていた仮想LUのI/O性能を上げることが望ましい場合には、その仮想LUを、DPプール#Eを構成するPDEVよりもI/O性能の高いPDEVで構成されているDPプール（例えばDPプール#D）に関連付けるといったことが行われても良い。なお、仮想LUが関連付けられるDPプールが別のDPプールに変更される場合、変更元のDPプールから変更先のDPプールにデータ要素及びパリティが移動される。

10

【0043】

DPプール203に関する情報や、仮想ページと複数の実ページとの対応関係を表す情報などが、DKC111によって管理される。それらの情報は、例えば、DKC111のメモリ123（例えばキャッシュ領域）に格納される。

【0044】

図5は、メモリ123に記憶されるコンピュータプログラム及び情報を示す。

【0045】

メモリ123は、制御プログラム501を記憶する。制御プログラム501は、MP125で実行される。以下の説明においてプログラム501によって行われる処理は、実際にはMP125によって行われる。

20

【0046】

メモリ123は、DP共通仕様テーブル503、プール管理テーブル505、仮想LU管理テーブル507、PDEV管理テーブル509、実ページ管理テーブル511、マッピング管理テーブル513、実アクセス状況テーブル515及び仮想アクセス状況テーブル517を記憶する。実ページ管理テーブル511は、DPプール203毎に存在する。これらのテーブルのうちの少なくとも一つのテーブルにおける情報は、テーブル以外の構造で管理されても良い。

30

【0047】

図6は、DP共通仕様テーブル503の構成を示す。

【0048】

DP共通仕様テーブル503は、DPプールの仕様に関する共通の情報を有する。具体的には、例えば、テーブル503は、DPプールの仕様に関する項目毎に、項目200と値201とを有する。項目200は、項目の名称であり、値201は、その項目についての値である。図6の例によれば、作成可能最大プール数（ストレージシステム103に作成可能なDPプール203の最大数）の50であり、実ページ容量が32KB（キロバイト）である。

【0049】

なお、本実施形態では、全てのDPプールにおいて、実ページの容量は同じである。そのため、実ページの容量、RAIDレベル及びコンビネーションを基に、仮想LU201を構成する仮想ページの容量が決定される。コンビネーションが $nD+mP$ のとき、仮想ページには、実ページ容量の n 倍の容量のデータを記憶でき、 $(n+m)$ 倍の容量が割当てられる。例えば、RAID6（ $4D+2P$ ）の仮想LU201の仮想ページの容量は、RAID5（ $2D+1P$ ）の仮想LU201の仮想ページの容量の2倍である。なぜなら、前述の通り、実ページの容量は同じであり、RAID6（ $4D+2P$ ）の仮想LU201の仮想ページに割り当てられる実ページの数、RAID5（ $2D+1P$ ）の仮想LU201の仮想ページに割り当てられる実ページの数、RAID6（ $4D+2P$ ）の仮想LU201の仮想ページには、六つの実ページが割り当て

40

50

られ、RAID5 (2D+1P) の仮想LU201の仮想ページには、三つの実ページが割り当てられる。

【0050】

図7は、プール管理テーブル505の構成を示す。

【0051】

プール管理テーブル505は、DPプール203に関する情報を有する。具体的には、例えば、テーブル505は、DPプール203毎に、プール番号210、PDEV番号リスト211、実割当容量212、使用容量213及び先頭PDEV番号214を有する。

【0052】

プール番号210は、DPプール203の番号である。

10

【0053】

PDEV番号リスト211は、DPプール203を構成するPDEVの番号のリストである。

【0054】

実割当容量212は、DPプール203を構成するPDEVの総容量である。従って、DPプール203に新たにPDEVが追加される場合、実割当容量212は、追加されたPDEVの容量分増えることになり、一方、DPプール203からPDEVが減設された場合、実割当容量212は、減設されたPDEVの容量分減ることになる。

【0055】

使用容量213は、DPプール203から割り当てられている実ページの総容量である。

20

【0056】

先頭PDEV番号214は、実ページの次の割当元となるPDEVの番号である。例えば、本実施形態では、仮想ページに割り当てられる二以上の実ページを提供する二以上のPDEVは、PDEV番号の順である。具体的には、例えば、DPプールがPDEV#01～#04の四つのPDEVで構成されていて、仮想LUのRAIDレベル及びコンビネーションがRAID5 (2D+1P) であり、且つ、その仮想LUにおける未割当ての仮想ページ（実ページが割り当てられていない仮想ページ）に初めてライトが発生する場合、先頭PDEV番号214は「1」であり、それ故、ライト先の仮想ページには、PDEV#01、#02及び#03からそれぞれ実ページが割り当てられる。その後、先頭PDEV番号214が「4」に更新される。そして、その仮想LUにおける未割当ての仮想ページに二回目にライトが発生する場合、先頭PDEV番号214が「4」であるため、ライト先の仮想ページには、PDEV#04、#01及び#02から、それぞれ実ページが割り当てられる。その後、先頭PDEV番号214が「3」に更新される。このように、本実施形態では、仮想ページに割り当てられる二以上の実ページの提供元の二以上のPDEVは、PDEV番号の順序を基に決定される。なお、DPプールを構成する複数のPDEVの番号は、連続でも非連続でもよい。また、仮想ページに割り当てられる二以上の実ページの提供元の二以上のPDEVは、番号の小さい順に代えて、番号の大きい順でも良い。PDEVの番号順に、提供元の二以上のPDEVが決定されるので、DPプール内のPDEVが均等に使用される。

30

40

【0057】

図8は、仮想LU管理テーブル507の構成を示す。

【0058】

仮想LU管理テーブル507は、仮想LU201に関する情報を有する。仮想LUの設定及び確認を行うことに使用できる。具体的には、例えば、テーブル507は、仮想LU201毎に、LUN220、所属プール番号221、RAIDレベル222、コンビネーション223及び容量224を有する。

【0059】

LUN220は、仮想LU201のLUNである。

【0060】

50

所属プール番号 2 2 1 は、仮想 L U 2 0 1 が所属する D P プール（仮想 L U 2 0 1 が関連付けられている D P プール） 2 0 3 の番号である。

【 0 0 6 1 】

R A I D レベル 2 2 2 は、仮想 L U 2 0 1 の R A I D レベルである。

【 0 0 6 2 】

コンビネーション 2 2 3 は、仮想 L U 2 0 1 のコンビネーションである。

【 0 0 6 3 】

容量 2 2 4 は、仮想 L U 2 0 1 の容量である。

【 0 0 6 4 】

図 9 は、D P プール作成等に使用する P D E V 管理テーブル 5 0 9 の構成を示す。

10

【 0 0 6 5 】

P D E V 管理テーブル 5 0 9 は、P D E V 1 1 7 に関する情報を有する。具体的には、例えば、テーブル 5 0 9 は、P D E V 1 1 7 毎に、P D E V 番号 2 3 0、所属プール番号 2 3 1、I F 種別 2 3 2、P D E V 性能 2 3 3、容量 2 3 4、空き容量 2 3 5 及びデバイス種別 2 3 6 を有する。

【 0 0 6 6 】

P D E V 番号 2 3 0 は、P D E V 1 1 7 の番号である。

【 0 0 6 7 】

所属プール番号 2 3 1 は、P D E V 1 1 7 が所属する D P プール（P D E V 1 1 7 が構成要素の一つとなっている D P プール） 2 0 3 の番号である。「n u l l」は、いずれのプールにも割当てられていないことを示す。

20

【 0 0 6 8 】

I F 種別 2 3 2 は、P D E V 1 1 7 のインタフェースの種別（例えば、S A S、S A T A など）である。

【 0 0 6 9 】

P D E V 性能 2 3 3 は、P D E V 1 1 7 のデバイス種別 2 3 6 に応じた表記となる。例えば、P D E V 1 1 7 のデバイス種別 2 3 6 が H D D であれば、P D E V 性能 2 3 3 は、ディスクの回転速度（例えば 7 2 0 0 回転／秒）等で表される。

【 0 0 7 0 】

容量 2 3 4 は、P D E V 1 1 7 の容量、言い換えれば、P D E V 1 1 7 が提供する実ページの総容量である。

30

【 0 0 7 1 】

空き容量 2 3 5 は、P D E V 1 1 7 が提供する実ページのうち仮想ページに割り当てられておらず、仮想ページに割当可能な実ページの総容量である。

【 0 0 7 2 】

デバイス種別 2 3 6 は、P D E V 1 1 7 それ自体の種別である。デバイス種別 2 3 6 としては、例えば、H D D、S S D がある。

【 0 0 7 3 】

図 1 0 は、実ページに対する仮想ページの関係を示す実ページ管理テーブル 5 1 1 の構成を示す。実ページ管理テーブル 5 1 1 は仮想ページに対して、実ページを割当てる際等

40

に使用する。

【 0 0 7 4 】

実ページ管理テーブル 5 1 1 は、D P プール毎に存在する。実ページ管理テーブル 5 1 1 は、そのテーブル 5 1 1 に対応した D P プール内の実ページに関する情報を有する。具体的には、例えば、テーブル 5 1 1 は、実ページ毎に、P D E V 番号 2 4 0、実ページ番号 2 4 1、L U N 2 4 2、仮想ページ番号 2 4 3 及びアクセスカウンタ 2 4 4 を有する。

【 0 0 7 5 】

P D E V 番号 2 4 0 は、実ページを提供する P D E V 1 1 7 の番号である。

【 0 0 7 6 】

実ページ番号 2 4 1 は、実ページの番号である。

50

【0077】

LUN 242は、実ページの割当先の仮想ページを有する仮想LUN 201のLUNである。

【0078】

仮想ページ番号243は、実ページの割当先の仮想ページの番号である。

【0079】

アクセスカウンタ244は、実ページに対して行われたI/O（アクセス）の回数である。なお、本実施形態のように代替情報として例えば実アクセス状況テーブル515があれば、アクセスカウンタ244は無くても良い。

【0080】

図11は、ライトコマンドを受領した仮想ページに実ページが割当てられているか判定する際等に使用するマッピング管理テーブル513の構成を示す。

【0081】

マッピング管理テーブル513は、仮想ページと実ページとの対応関係を表す。具体的には、例えば、テーブル513は、仮想ページ毎に、LUN 250、仮想ページ番号251、実ページ識別情報252及びアクセスカウンタ253を有する。

【0082】

LUN 250は、仮想ページを有する仮想LUN 201のLUNである。

【0083】

仮想ページ番号251は、仮想ページの番号である。

【0084】

実ページ識別情報252は、仮想ページに割り当てられている実ページの識別情報、例えば、実ページの番号と、その実ページを有するPDEV 117の番号との組合せである。

【0085】

アクセスカウンタ253は、仮想ページに対して行われたI/O（アクセス）の回数である。なお、本実施形態のように代替情報として例えば仮想アクセス状況テーブル517があれば、アクセスカウンタ253は無くても良い。

【0086】

図12は、再配置等をする記憶装置を決定する際等に参照する実アクセス状況テーブル515の構成を示す。

【0087】

実アクセス状況テーブル515は、一以上のDPプール203に共通であり、実ページ毎のアクセス状況を表す情報を有する。具体的には、例えば、テーブル515は、実ページ毎にセルを有し、セルが、実ページに対するアクセスの頻度を表す情報（以下、実アクセス頻度情報）を有する。実アクセス頻度情報は、例えば、その情報に対応する実ページの使用の有無を表す情報を含み、ライトの頻度とリードの頻度とを別々に表す。ライト頻度は、単位期間当たりのライトの回数、具体的には、例えば、1分、1時間、1日及び1週間毎のライトの回数である。リード頻度は、単位期間当たりのリードの回数、具体的には、例えば、1分、1時間、1日及び1週間毎のリードの回数である。

【0088】

いずれかのDPプール203にPDEV 117が一つ追加された場合、その追加されたPDEV 117に対応するレコードがテーブル515に追加される。また、いずれかのDPプール203からPDEV 117が一つ減らされた場合、その減らされたPDEV 117に対応するレコードがテーブル515から除去される。

【0089】

図13は、再配置等をする仮想ページを決定する際等に参照する仮想アクセス状況テーブル517の構成を示す。

【0090】

仮想アクセス状況テーブル517は、仮想ページ毎のアクセス状況を表す情報を有する

10

20

30

40

50

。具体的には、例えば、テーブル517は、仮想ページ毎にセルを有し、セルが、仮想ページに対するアクセスの頻度を表す情報（以下、仮想アクセス頻度情報）を有する。仮想アクセス頻度情報は、例えば、その情報に対応する仮想ページの使用の有無を表す情報を含み、ライトの頻度とリードの頻度との両方を表す。ライト頻度は、単位期間当たりのライトの回数、具体的には、例えば、1分、1時間、1日及び1週間毎のライトの回数である。リード頻度は、単位期間当たりのリードの回数、具体的には、例えば、1分、1時間、1日及び1週間毎のリードの回数である。

【0091】

いずれかのDPプール203に仮想LU201が一つ関連付けられた場合、その関連付けられた仮想LU201に対応するレコードがテーブル517に追加される。また、い

10

【0092】

なお、仮想アクセス頻度情報及び前述の実アクセス頻度情報のうちの少なくとも一方は、ライト頻度とリード頻度とを別々に表すことに代えて、ライトかリードに関わらずアクセスの頻度を表しても良いし、ライト頻度のみを表しても良いし、リード頻度のみを表しても良い。

【0093】

また、仮想アクセス頻度情報及び実アクセス頻度情報のうちの少なくとも一方が、仮想領域或いは実ページにアクセスが発生する都度に更新されてもよいし、所定のスケジュールで（例えば定期的に）更新されてもよい。例えば、制御プログラム501が、処理されたコマンド履歴をメモリ123で管理してよい。コマンド履歴は、処理されたI/Oコマンド毎に、例えば、種別、日時（例えば、受信日時、処理完了日時、或いは、I/Oコマンドが有するタイムスタンプが表す日時）、アクセス先の仮想ページの番号、アクセス先の実ページの識別情報（例えばPDEV番号と実ページ番号との組合せ）を有してよい。制御プログラム501は、I/Oコマンドを処理した場合、そのI/Oコマンドに関する種別、日時、アクセス先の仮想ページの番号、及びアクセス先の実ページの識別情報を、コマンド履歴に追加してよい。制御プログラム501は、そのコマンド履歴を基に、仮想アクセス頻度情報及び実アクセス頻度情報を更新してよい。

20

【0094】

図14は、DPプールの生成の流れを示すフローチャートである。

30

【0095】

（S10）

制御プログラム501が、管理装置105から、プール生成指示を受ける。プール生成指示は、パラメータとして、例えば、以下の（p1）～（p3）のうちの、少なくとも（p1）がある。（p1）は、ユーザによる入力が必要のパラメータであり、（p2）及び（p3）は、ユーザが任意に入力するパラメータである。

（p1）ユーザ（例えば管理者）が要求する容量（ユーザ要求プール容量）。ユーザ要求プール容量は、実割当容量212とは異なる。なぜなら、実割当容量212は、DPプールに所属するPDEVの総容量であるからである。

40

（p2）PDEV117の属性を表す情報（PDEV情報）。例えば、デバイス種別、IF種別及びPDEV性能。

（p3）プールに付与する番号（プール番号）。

【0096】

（S11）

制御プログラム501が、プール管理テーブル505から特定されるDPプール203の数が、作成可能最大プール数の値201より小さいか否かを判断する。この判断の結果が肯定的の場合、S12が行われる。この判断の結果が否定的の場合、制御プログラム501は、エラーを表す情報を管理装置105に送信し、管理装置105が、エラーを表示する（S16）。

50

【0097】

(S12)

制御プログラム501が、ユーザ要求プール容量以上の総容量となるようにPDEV117を選択する。具体的には、例えば、制御プログラム501は、下記の処理を行う。

(S12-1)： 制御プログラム501は、PDEV管理テーブル509を参照し、所属プール番号213が「null」になっているPDEV（いずれのDPプール203にも所属していないPDEV）を特定する。特定されたPDEVが、作成されるDPプールの構成要素の候補である。なお、プール生成指示のパラメータとして、PDEV情報があれば、所属プール番号213が「null」であり、且つ、PDEV情報に適合するPDEVが、DPプールの構成要素の候補として特定される。以下、この(S12-1)で特定されたPDEVを、「候補PDEV」と言う。

10

(S12-2)： 制御プログラム501は、複数の候補PDEVから、ユーザ要求プール容量以上の総容量となるようにPDEVを選択する。この結果、選択されたPDEVの総容量（DPプールの実割当容量）は、ユーザ要求プール容量と同じか又はそれより大きい。なお、PDEVの選択のためにRAIDレベル及びコンビネーションが考慮されて良い。例えば、ユーザ要求プール容量が1TB（テラバイト）の場合、制御プログラム501は、選択されたPDEVの総容量がユーザ要求プール容量の1.3倍（1.3TB）となるように、複数の候補PDEVからPDEVを選択して良い。作成されるDPプールに、異なるRAIDレベル及びコンビネーションの仮想LUが関連付けられる場合には、パリティ比率が最も高いRAIDレベル及びコンビネーションのそのパリティ比率を基に、PDEVが選択されて良い。また、DPプールに関連付けられる仮想LUのRAIDレベル及びコンビネーションが決められている場合には、そのRAIDレベル及びコンビネーションでのパリティ比率を基に、PDEVが選択されて良い。なお、「パリティ比率」とは、仮想ページに対するI/O対象のデータのサイズに対する、生成されるパリティの総サイズの割合である。

20

【0098】

(S13)

制御プログラム501は、作成するDPプールに関する情報を、プール管理テーブル505に追加する。作成するDPプールに関する情報として、下記が追加される。

＊プール番号210： 所定の規則に従って決定された番号（例えば、プール管理テーブル505に登録されているプール番号210のうち最も大きい番号の次の番号）。或いは、S10での上記パラメータ（p3）が表すプール番号。

30

＊PDEV番号リスト211： S12で選択されたPDEVの番号。

＊実割当容量212： S12で選択されたPDEVの総容量。

＊使用容量213：ゼロ。なぜなら、当該DPプールから実ページが割り当てられていないため。

＊先頭PDEV番号214： S12で選択されたPDEVの番号のうちの、所定の規則に従って決定された番号。例えば、S12で選択されたPDEVの番号のうちの最も小さい番号。

【0099】

(S14)

制御プログラム501が、PDEV管理テーブル509における、S12で選択されたPDEVの所属プール番号231を、「null」から、作成されるDPプールの番号（S13での番号）に更新する。

40

【0100】

(S15)

制御プログラム501が、作成されるDPプールを構成する各PDEV（S12で選択された各PDEV）の記憶空間を、実ページ容量毎に分割し、それにより得られる各実ページを管理する。具体的には、制御プログラム501は、作成されるDPプールに対応した実ページ管理テーブル511を作成しメモリ123に格納する。作成された実ページ管

50

理テーブル 5 1 1 が有する各レコードは、下記情報を有する。

＊ P D E V 番号 2 4 0 : S 1 2 で選択されたいずれかの P D E V の番号。

＊ 実ページ番号 2 4 1 : S 1 2 で選択された P D E V が提供する実ページに付与されたいずれかの通し番号。

＊ L U N 2 4 2 : null。いずれの実ページがいずれの仮想ページにも割り当てられていないため。

＊ 仮想ページ番号 2 4 3 : null。いずれの実ページがいずれの仮想ページにも割り当てられていないため。

＊ アクセスカウンタ 2 4 4 : ゼロ。いずれの実ページに対して I / O が発生していないため。

10

【 0 1 0 1 】

図 1 5 は、仮想 L U 2 0 1 の生成の流れを示すフローチャートである。

【 0 1 0 2 】

(S 2 0)

制御プログラム 5 0 1 が、管理装置 1 0 5 から、仮想 L U 生成指示を受ける。仮想 L U 生成指示は、パラメータとして、例えば、以下の (p 1 1) ~ (p 1 5) のうちの、少なくとも (p 1 5) がある。(p 1 5) は、ユーザによる入力が必要のパラメータであり、(p 1 1) ~ (p 1 4) は、ユーザが任意に入力するパラメータである。(p 1 1) ~ (p 1 4) のうち、ユーザから入力されていないパラメータは、所定の規則によって決定される値 (デフォルト値) である。

20

(p 1 1) 作成する仮想 L U に付与する L U N。デフォルト値は、例えば、仮想 L U の L U N として付与可能な番号群のうちの最も小さい番号である。

(p 1 2) 作成する仮想 L U の数。デフォルト値は、「 1 」である。

(p 1 3) 作成する仮想 L U の所属先の D P プールの番号 (プール番号)。デフォルト値は、作成されている D P プールの番号のうちの最も小さい番号である。

(p 1 4) R A I D レベル (コンビネーションの設定も同時に行う)。デフォルト値は、例えば、R A I D 5 (2 D + 1 P) である。

(p 1 5) 仮想 L U の容量。

【 0 1 0 3 】

(S 2 1)

制御プログラム 5 0 1 が、上記 (p 1 2) が表す数の仮想 L U 毎に、作成する仮想 L U に関する情報を、仮想 L U 管理テーブル 5 0 7 に追加する。作成する仮想 L U に関する情報として、下記が追加される。

＊ L U N 2 2 0 : 上記パラメータ (p 1 1) が表す L U N。

＊ 所属プール番号 2 2 1 : 上記パラメータ (p 1 3) が表すプール番号。上記パラメータ (p 1 3) が表すプール番号の D P プールに、上記 (p 1 2) が表す数の仮想 L U が関連付けられる。

＊ R A I D レベル 2 2 2 : 上記パラメータ (p 1 4) が表す R A I D レベル。

＊ コンビネーション 2 2 3 : 上記パラメータ (p 1 4) が表すコンビネーション。

＊ 容量 2 2 4 : 上記パラメータ (p 1 5) が表す容量。

40

【 0 1 0 4 】

(S 2 2)

制御プログラム 5 0 1 が、上記 (p 1 2) が表す数の仮想 L U 毎に、作成する仮想 L U の各仮想ページに関する情報を、マッピング管理テーブル 5 1 3 に追加する。各仮想ページについて追加される情報は、例えば下記の通りである。

＊ L U N 2 5 0 : 仮想ページを有する仮想 L U の L U N。

＊ 仮想ページ番号 2 5 1 : 仮想ページの番号。

＊ 実ページ識別情報 2 5 2 : null。いずれの仮想ページに実ページが割り当てられていないため。

＊ アクセスカウンタ 2 5 3 : ゼロ。いずれの仮想ページに対して I / O が発生していな

50

いため。

【0105】

図16は、仮想ページに対するライトの流れの概要を示す。以下、実ページが割り当てられていない仮想ページに対するライトの流れと、実ページが割り当てられている仮想ページに対するライトの流れとを説明する。

【0106】

<ケース1：実ページが割り当てられていない仮想ページに対するライト>

以下の処理が行われる。

(w11) 制御プログラム501が、仮想LU#11の仮想ページV11を指定したライトコマンドをホスト#1から受け、マッピング管理テーブル513を参照し、ライト先の仮想ページV11に実ページが割り当てられていないことを特定する。

10

(w12) 制御プログラム501が、仮想LU管理テーブル507を参照して、仮想ページV11を有する仮想LU#11のRAIDレベルがRAID5(2D+1P)であることを特定する。

(w13) 制御プログラム501が、仮想LU#11が所属するDPプール#11に対応した実ページ管理テーブル511を参照し、LUN242及び仮想ページ番号243がnullである実ページから、PDEV番号240が異なる三つの実ページ((w12)で特定されたRAIDレベル(RAID5(2D+1P))に基づく数の実ページ)を選択する。ここでは、PDEV#11内の実ページR11、PDEV#12内の実ページR21、及び、PDEV#13内の実ページR31が選択される。

20

(w14) 制御プログラム501が、(w13)で選択した実ページR11、R21及びR31を、仮想ページV11に割り当て、マッピング管理テーブル513を更新する。

(w15) 制御プログラム501が、三つの実ページR11、R21及びR31のうちの二つに、仮想ページV11に対するライト対象のデータを分割することにより得られた二つのデータ要素を書き込み、それら三つの実ページのうちの残りの一つに、それら二つのデータ要素に基づくパリティを書き込む。

【0107】

<ケース2：実ページが割り当てられている仮想ページに対するライト>

以下の処理が行われる。

(w21) 制御プログラム501が、仮想LU#12の仮想ページV22を指定したライトコマンドをホスト#2から受け、マッピング管理テーブル513を参照し、ライト先の仮想ページV22に割り当てられている実ページを特定する。仮想ページV22を有する仮想LU#12のRAIDレベルがRAID6(3D+2P)なので、仮想ページV22には五つの実ページが割り当てられており、(w21)では、それら五つの実ページが特定される。具体的には、PDEV#13内の実ページR32、PDEV#14内の実ページR41、PDEV#15内の実ページR51、PDEV#16内の実ページR61、及び、PDEV#17内の実ページR71が選択される。

30

(w22) 制御プログラム501が、(w21)で特定した五つの実ページR32、R41、R51、R61及びR71のうちの三つに、仮想ページV22に対するライト対象のデータを分割することにより得られた三つのデータ要素を書き込み、それら五つの実ページのうちの残りの二つに、それら三つのデータ要素に基づく二つのパリティを書き込む。

40

【0108】

図17は、仮想ページに対するライトの流れを示すフローチャートである。

【0109】

(S30)

制御プログラム501が、ライト先情報を有するライトコマンドとライト対象のデータとを受ける。ライト先情報は、LUNとLBAとを有する。ライト対象のデータは、メモリ123(キャッシュ領域)に書き込まれる。

【0110】

(S31)

50

制御プログラム501が、マッピング管理テーブル513を参照して、ライト先情報から特定される仮想ページ（ライト先仮想ページ）に実ページが割り当てられているか否かを判断する。

【0111】

＜（S31）の判断の結果が肯定的の場合（ライト先仮想ページについてのPDEV識別情報252がnullではない場合）＞。

【0112】

これは、上記ケース2に相当する。具体的には、S32～S36が行われることなく、S37が行われる。すなわち、制御プログラム501が、ライト先仮想ページ（例えば、図16での仮想ページV22）に割り当てられている実ページ（例えば、図16での実ページR32、R41、R51、R61及びR71）に、ライト先仮想ページを有する仮想LU（例えば、図16での仮想LU#12）のRAIDレベルに基づいて、ライト対象のデータ（例えば、そのデータを分割して得られた二以上のデータ要素と、一以上のパリティ）を書き込む。そして、制御プログラム501は、以下のライト状況更新処理：

＊ライト先仮想ページに対応したアクセスカウンタ253の値を1インクリメントする；
＊ライト先仮想ページに対応した仮想アクセス頻度情報（仮想アクセス状況テーブル517内の情報）を更新する（具体的には、ライト頻度（1分、1時間、1日、1週間）を増加する更新を行う）；

＊ライト先仮想ページに割り当てられている各実ページに対応したアクセスカウンタ244の値を1インクリメントする；

＊ライト先仮想ページに割り当てられている各実ページに対応した実アクセス頻度情報（実アクセス状況テーブル515内の情報）を更新する（具体的には、ライト頻度（1分、1時間、1日、1週間）を増加する更新を行う）、

を行う。なお、ライト先仮想ページに割り当てられている実ページは、マッピング管理テーブル513から特定可能である。ライト先仮想ページを有する仮想LU（以下、ライト先仮想LU）は、ライト先情報が有するLUNから識別可能である。ライト先仮想LUのRAIDレベルは、仮想LU管理テーブル507から特定可能である。

【0113】

＜（S31）の判断の結果が否定的の場合（ライト先仮想ページについてのPDEV識別情報252がnullである場合）＞。

【0114】

これは、上記ケース1に相当する。そのため、S32～S36が行われ、その後で、S37が行われる。

【0115】

（S32）

制御プログラム501が、ライト先仮想LU（例えば、図16での仮想LU#11）に対応したRAIDレベル222及びコンビネーション223（例えば、RAID5（2D+1P））を特定する。制御プログラム501が、特定されたRAIDレベル222及びコンビネーション223を基に、選択するPDEVの数Kを決定する。例えば、RAIDレベルが表すコンビネーションが、「nD+mP(or D)」の場合（nは自然数、mは0以上の整数）、選択するPDEVの数Kは、（n+m）である。具体的には、例えば、コンビネーションが「2D+1P」の場合、 $K = 2 + 1 = 3$ である。

【0116】

（S33）

制御プログラム501が、ライト先仮想LUに対応した所属プール番号221と同じ値のプール番号210に対応する先頭PDEV番号214を特定する。

【0117】

（S34）

制御プログラム501が、ライト先仮想LUに対応した所属プール番号221と同じ値のプール番号210に対応するPDEV番号リスト211から、S32で選択した数Kと

10

20

30

40

50

同数の PDEV 番号を選択する。その際、制御プログラム 501 は、まず、S33 で特定した先頭 PDEV 番号 214 と同じ PDEV 番号（例えば #11）を選択し、その PDEV 番号を先頭に、順次、PDEV 番号の小さい順に、選択された PDEV 番号の数が K に達するまで、PDEV 番号を選択していく。もし、選択された PDEV 番号の数が K に達する前に、選択した PDEV 番号が、ライト先仮想 LU が所属する DP プール（例えば、図 16 での DP プール #11）に対応した PDEV 番号リスト 211 のうちの最も大きい番号になった場合には、その PDEV 番号リスト 211 のうちの最も小さい番号が選択される。

【0118】

なお、この（S34）では、PDEV 番号の順序に従って PDEV 番号（PDEV）を選択することに代えて、以下のようにして PDEV 番号（PDEV）が選択されても良い。

＊空き容量 235 の多い PDEV を優先的に選択すること。これにより、DP プールにおける PDEV を均等に使用することが期待できる。

＊IF 種別 232 及び PDEV 性能 233 の両方又は片方が同一の PDEV を選択すること。これにより、一つの仮想ページに割り当てられる実ページに対する I/O 性能が同じであることが期待される。

【0119】

（S35）

制御プログラム 501 が、S34 で選択した各 PDEV 番号から特定される各 PDEV から、一つの空き実ページを選択する。具体的には、例えば、制御プログラム 501 は、ライト先仮想 LU が所属する DP プールに対応した実ページ管理テーブル 511 を基に、選択した PDEV 番号毎に、null である LUN 242 及び仮想ページ番号 243 に対応したいずれかの実ページ番号 241 を選択する。図 16 の例によれば、PDEV 番号 #11 について実ページ番号 R11 が選択され、PDEV 番号 #12 について実ページ番号 R21 が選択され、PDEV 番号 #13 について実ページ番号 R31 が選択される。

【0120】

（S36）

制御プログラム 501 が、ライト先仮想ページに対応した実ページ識別情報 252 を、null から、S35 で選択した実ページ番号 241 と、その番号 241 に対応した PDEV 番号 240 との組合せ（例えば、#11-R11、#12-R21、#13-R31）に更新する。つまり、制御プログラム 501 は、ライト先仮想ページ（例えば、図 16 での仮想ページ V11）に、S35 で選択された二以上の空き実ページを割り当てる。

【0121】

（S37）

制御プログラム 501 が、S36 で割り当てた二以上の実ページに、ライト先仮想 LU の RAID レベル及びコンビネーションに基づいて、ライト対象のデータを書き込む。図 16 の例によれば、二つのデータ要素と一つのパリティが、三つの実ページ R11、R21 及び R31 に書き込まれる。そして、制御プログラム 501 は、前述したようなライト状況更新処理を行う。これにより、ライト先仮想ページに対応したアクセスカウンタ 253 の値、ライト先仮想ページに対応した仮想アクセス頻度情報、ライト先仮想ページに割り当てられている各実ページに対応したアクセスカウンタ 244 の値、及び、ライト先仮想ページに割り当てられている各実ページに対応した実アクセス頻度情報が、更新される。

【0122】

図 18 は、仮想ページに対するリードの流れを示すフローチャートである。

【0123】

（S40）

制御プログラム 501 が、リード元情報を有するリードコマンドを受ける。リード元情報は、例えば、LUN 及び LBA を有する。

【0124】

(S41)

制御プログラム501が、マッピング管理テーブル513を参照して、リード元情報から特定される仮想ページ（リード元仮想ページ）に実ページが割り当てられているか否かを判断する。S41の判断の結果が否定的の場合、S42が行われる。S41の判断の結果が肯定的の場合、S43が行われる。

【0125】

(S42)

制御プログラム501が、エラーを、リードコマンドの送信元のホストに送信する。

【0126】

(S43)

制御プログラム501が、リード元仮想ページに割り当てられている二以上の実ページから、二以上のデータ要素（及び一以上のパリティ）を読み出し、読み出した二以上のデータ要素で構成されるデータ（つまり、リード対象のデータ）を、リードコマンドの送信元のホストに送信する（リード対象のデータは、パリティで復元されたデータ要素を含んだデータであっても良い。）。そして、制御プログラム501は、以下のリード状況更新処理：

- ＊リード元仮想ページに対応したアクセスカウンタ253の値を1インクリメントする；
- ＊リード元仮想ページに対応した仮想アクセス頻度情報（仮想アクセス状況テーブル517内の情報）を更新する（具体的には、リード頻度（1分、1時間、1日、1週間）を増加する更新を行う）；
- ＊リード元仮想ページに割り当てられている各実ページに対応したアクセスカウンタ244の値を1インクリメントする；
- ＊リード元仮想ページに割り当てられている各実ページに対応した実アクセス頻度情報（実アクセス状況テーブル515内の情報）を更新する（具体的には、リード頻度（1分、1時間、1日、1週間）を増加する更新を行う）、

【0127】

以上、これまで説明したように、本実施形態では、DPプール203を構成する実ページは、RAIDグループから提供される実体的な記憶領域ではなく、一つのDEVから提供される記憶領域である。そして、一つのDPプール203には、異なるRAIDレベル及びコンビネーションが定義された複数の仮想LUを関連付けることが可能である。このため、図19に例示するように、一つのDPプール#40に、異なるRAIDレベル及びコンビネーションでそれぞれ格納されたデータ（例えば、RAID5（3D+1P）で格納されたデータ（データ要素a、b及びc）とパリティPのセットと、RAID6（5D+2P）で格納されたデータ（データ要素e、f、g、h及びi）とパリティP1及びP2のセット）が混在する。この場合、もし、或る仮想LUについて定義されているRAIDレベル又はコンビネーションが別のRAIDレベル又はコンビネーションに変更された場合には、その仮想LUに割り当てられている実ページについてのみ、データの再構築が行われれば良い。例えば、仮想LUのRAIDレベル（又はコンビネーション）が第1のRAIDレベル（又はコンビネーション）から第2のRAIDレベル（又はコンビネーション）に変更された場合、制御プログラム501が、実ページが割り当てられている仮想ページ毎に、以下の処理（k1）～（k3）：

- （k1）仮想ページに割り当てられているv個の実ページからv個のデータユニットを読み出し、v個のデータユニットを基にデータを構築する（本実施形態の説明で言う「データユニット」とは、実ページに格納されているデータ要素又はパリティの総称である）；
- （k2）第2のRAIDレベル及びコンビネーションに基づくw個の実ページを、仮想LUが関連付けられているDPプールのw個のDEVからそれぞれ選択し、仮想ページに、（k1）でのv個の実ページに代えて、選択したw個の実ページを割り当てる；
- （k3）割り当てたw個の実ページに、（k1）で構築したデータに基づくw個のデータ

10

20

30

40

50

ユニットを書き込む、
を行う。なお、(k2)では、RAIDレベル(又はコンビネーション)が変更される仮想LUが関連付けられているDPプールでの各PDEVの使用率が均一になるように、w個の実ページが選択されて良い。これにより、仮想LUについて定義されているRAIDレベル(又はコンビネーション)の変更を契機に、その仮想LUが関連付けられているDPプールに属する各PDEVの使用率の均一化が期待できる。このとき、PDEVの使用率ではなく空き容量が均一になるようにw個の実ページが選択されても良い。

【0128】

また、上述の説明では、仮想ページに割り当てられる実ページの提供元のPDEVは、PDEV番号の順番で選択されるが、例えば、図20に示すように、ランダムにPDEV
10
が選択されて良い。図20の例によれば、RAID5(3D+1P)に従うデータ要素a、b及びcとパリティPも、RAID6(5D+2P)に従うデータ要素e、f、g、h及びiとパリティP1及びP2も、PDEV番号に依存せずに選択された各PDEVにおける実ページに格納されている。なお、PDEV番号に依存しないが、PDEVの使用率(もしくは空き容量)は考慮されて良い。DPプールにおける複数のPDEVの使用率(もしくは空き容量)を均等にするためである。

【0129】

また、制御プログラム501は、同一のバックエンド接続デバイス113に接続されて
20
いる複数のPDEVから、一つの仮想ページに複数の実ページを割り当てて良い。これにより、それら複数の実ページへのアクセスの性能の均一化が期待できる。なお、もし、同一のバックエンド接続デバイス113に接続されている複数のPDEVの使用率(もしくは空き容量)が均等ではなければ、複数のバックエンド接続デバイス113に接続されている複数のPDEVから複数の実ページが仮想ページに割り当てられて良い。

【0130】

ところで、本実施形態では、或るHDD内の実ページから別のPDEV内の実ページにデータユニットが移動される再配置処理が行われる。

【0131】

例えば、PDEVが増設又は減設された時に行われる。具体的には、例えば、DPプ
30
ール203に空きの実ページの総容量が不足した場合(例えば、DPプール203の実割当容量212に対する使用容量213の比率が所定の閾値以上になった場合)、空きの実ページを増やすために、PDEVの増設が行われる。DPプール203に対するPDEVの増設は、PDEV単位で行うことができる。なお、空きの実ページの総容量が不足したことは、例えば、制御プログラム501が、プール管理テーブル505を参照することにより、検出することができる。制御プログラム501は、空きの実ページの総容量が不足したことを検出した場合、空きの実ページの総容量が不足したDPプールの番号と警告とを、管理装置105に送信する。管理装置105は、警告とDPプール番号とを表示する。ユーザは、その警告及びDPプール番号を見て、PDEVを増設し、増設したPDEVを、警告と共に表示されたDPプール番号から識別されるDPプールのメンバにさせる。なお、PDEVの増設は、空きの実ページの総容量が不足したときに限らず、他のタイミングで(例えば所定の計画に従って)行うこともできる。
40

【0132】

図21は、PDEVの増設の際に行われる処理のフローチャートである。

【0133】

(S1801)

PDEVの増設が行われた場合(例えば、制御プログラム501が、PDEVが増設され
50
たことを検出した場合)、制御プログラム501は、増設されたPDEV(以下、増設PDEV)に関する情報(レコード)を、PDEV管理テーブル509に追加する。PDEVに関する情報のうち、IF種別232、PDEV性能233、容量234及び空き容量235は、ユーザから管理装置105を通じて入力されても良いし、増設PDEVから取得されても良い。PDEV番号230は、ユーザから管理装置105を通じて入力され

ても良いし、制御プログラム501によって決められても良い。制御プログラム501は、PDEV管理テーブル509における、増設PDEVに関する情報に、所属プール番号231、すなわち、増設PDEVがメンバとなるDPプールの番号を追加する。この所属プール番号231は、ユーザから管理装置105を通じて入力された番号でも良いし、制御プログラム501が決定した番号（例えば、空きの実ページの総容量が不足していると検出されたDPプールの番号）でも良い。

【0134】

(S1802)

制御プログラム501が、S1801で追加した所属プール番号231と同じプール番号210（図7参照）に対応したPDEV番号リスト211（以下、図21の説明において「対象リスト211」と言う）に、増設PDEVの番号を追加し、且つ、そのプール番号210に対応した実割当容量212に、増設PDEVの容量を追加する。また、制御プログラム501は、S1801で追加した所属プール番号231に対応する実ページ管理テーブル511に、増設PDEVが有する各実ページに対応する情報（レコード）を追加する。

10

【0135】

(S1803)

制御プログラム501が、PDEV管理テーブル509で管理されている容量234及び空き容量235のうち、対象リスト211に属するPDEVに対応した容量234及び空き容量235を参照し、対象リスト211に属する各PDEVの使用率（容量234に対する空き容量235の比率）を算出する。そして、制御プログラム501は、使用率が所定の閾値を超過しているPDEVを特定する。

20

【0136】

(S1804)

制御プログラム501が、S1803で特定されたPDEVの使用率が所定の閾値以下となるまで、再配置処理を実行する。「再配置処理」とは、一つのDPプールにおいて或るPDEV内の実ページから別のPDEV内の実ページにデータユニットを移動する処理である。

【0137】

(S1805)

制御プログラム501が、増設PDEVを有するDPプールにおける全てのPDEVの使用率が均等になるまで再配置処理を実行する。このPDEVの増設の際に行われる処理では、PDEVの使用率の代わりに、空き容量を基に、再配置処理の対象のPDEVが決定されても良い。

30

【0138】

図22は、再配置処理のフローチャートを示す。なお、以下の説明では、複数のデータユニットを基に生成されるデータ（仮想領域に対してI/Oが行われるデータ）を、「I/Oデータ」と言う。

【0139】

(S1901)

制御プログラム501が、データユニットの移動元の実ページと、データユニットの移動先の実ページとを選択する。この再配置処理が、図21のS1804での再配置処理であれば、移動元の実ページは、図21のS1803で特定されたPDEV（使用率が所定の閾値を超過しているPDEV）から選択され、移動先の実ページは、少なくとも一つの空きの実ページにデータユニットが格納されても使用率が所定の閾値を超過しないPDEV（例えば、使用率が最も低いPDEV）から選択される。一方、この再配置処理が、図21のS1805での再配置処理であれば、移動元の実ページは、使用率が最も高いPDEVから選択され、移動先の実ページは、使用率が最も低いPDEVから選択される。

40

【0140】

(S1902)

50

制御プログラム501が、RAIDが維持されるか否かを判断する。「RAIDが維持されるか否か」とは、RAID技術に基づいてI/Oデータの生成が可能か否かである。具体的には、移動先の実ページを有するPDEV内の別の実ページに、移動元の実ページに格納されているデータユニットに基づいて生成されるI/Oデータの基になる別のデータユニットが存在するか否かである。より具体的には、例えば、制御プログラム501は、移動元の実ページの割当先の仮想ページに対応する実ページ識別情報252に、移動先の実ページを有するPDEVの番号が含まれておらず、RAIDが維持されるか否かを判断する。この判断の結果が肯定的の場合、S1903が行われ、この判断の結果が否定的の場合、S1904が行われる。

【0141】

10

(S1903)

制御プログラム501は、移動元の実ページから移動先の実ページにデータユニットを移動する。そして、制御プログラム501は、移動元の実ページの割当先であった仮想ページに対応する実ページ識別情報252に含まれている、移動元実ページの識別情報(PDEV番号と実ページ番号との組合せ)を、移動先実ページの識別情報に更新する。つまり、制御プログラム501は、移動先の実ページを、移動元の実ページに代えて仮想ページに割り当てる。

【0142】

(S1904)

制御プログラム501は、移動先の実ページを別のPDEVから選択する。例えば、別のPDEVは、この再配置処理が図21のS1804及びS1805のどちらであるかに関わらず、S1901で選択した移動先実ページを有するPDEVの次に使用率が低いPDEVでよい。また、この再配置処理では、PDEVの使用率の代わりに空き容量を基にPDEVが選択されても良い。

20

【0143】

以上が、PDEVが増設された場合に行われる処理の説明である。図21のS1805によれば、例えば、図23に示すようにしてデータユニットが移動される。例えば、一つのDPプールに属するPDEV#01～#09において、PDEV#01及び#05が、四つのデータユニットを記憶していて、残りのPDEV#02～#04及び#06～#09が、三つのデータユニットを記憶しているとする。ここで、増設PDEV#10がDP

30

プールに所属させられた場合、PDEV#01及び#05からそれぞれ一つのデータユニットが増設PDEV#10に移動される。これにより、PDEV#01～#10の使用率

がなるべく均一になる。

【0144】

PDEVの増設に関する処理は、上述の説明に限らず、下記のようなバリエーションが考えられる。下記の処理においても、PDEVの使用率の代わりに空き容量を基にPDEVが選択されても良い。

【0145】

(第1のバリエーション)

S1803～S1805が行われたい。このバリエーションでは、制御プログラム501は、各ライトコマンドの処理において、ライト先の仮想ページに割り当てる実ページを、必ず、増設PDEVから選択するようにする。例えば、DPプール#20にPDEV#51～#53があり、PDEV#54が増設された場合、制御プログラム501は、ライトコマンドの処理では、必ず、ライト先の仮想ページに割り当てる二以上の実ページにPDEV#54内の実ページが含まれるようにする。なお、ライトコマンドの処理において増設PDEVから実ページを必ず選択することは、増設PDEVの使用率が、増設PDEVを有するDPプール内の他のPDEVの使用率と同程度(例えば、DPプールでのPDEVの使用率の平均と同じ、又は、その平均との差が所定値未満)になるまで行われる。その後は、DPプールに属する複数のPDEVが均等に選択される。

40

【0146】

50

(第2のバリエーション)

移動先の実ページが、移動元の実ページのアクセス頻度（実アクセス状況テーブル515から特定されるアクセス頻度）に基づいて決定される。例えば、移動元実ページのアクセス頻度の割りに、移動元実ページを有するPDEVのI/O性能が高ければ、移動先の実ページは、移動元の実ページを有するPDEV（例えばSSD）よりもI/O性能の低いPDEV（例えばHDD）から選択されて良い。逆に、例えば、移動元実ページのアクセス頻度の割りに、移動元実ページを有するPDEVのI/O性能が低ければ、移動先の実ページは、移動元の実ページを有するPDEV（例えばHDD）よりもI/O性能の高いPDEV（例えばSSD）から選択されて良い。

【0147】

10

図24は、PDEVの減設の際に行われる処理のフローチャートである。

【0148】

(S2701)

制御プログラム501が、ユーザから管理装置105を通じて、減設対象のPDEVの番号の入力を受ける。

【0149】

(S2702)

制御プログラム501が、実ページ管理テーブル511を参照して、減設対象のPDEVが有する複数の実ページのうち仮想ページに割り当てられている一以上の実ページをLUN242及び仮想ページ番号243がnullでない実ページから特定する。

20

【0150】

(S2703)

制御プログラム501が、RAIDを維持し、且つ、使用率が均等になるように、データユニットを移動する。具体的には、例えば、制御プログラム501が、S2702で特定された実ページ（移動元の実ページ）毎に、減設対象のPDEV以外のPDEVのうち使用率が最も低いPDEVから一つの実ページ（移動先の実ページ）を選択し、選択した移動先実ページに、移動元の実ページからデータユニットを移動する。移動元の実ページを有するPDEVのI/O性能と、移動先の実ページを有するPDEVのI/O性能が同じであることが好ましい。

【0151】

30

(S2704)

制御プログラム501が、S2702で特定された実ページ（移動元の実ページ）毎に、移動元実ページが割り当てられた仮想ページに、移動元実ページに代えて移動先実ページを割り当てる。具体的には、例えば、制御プログラム501は、S2702で特定された実ページ（移動元の実ページ）毎に、移動元実ページが割り当てられている仮想ページに対応した実ページ識別情報252における、移動元実ページの識別情報を、移動先実ページの識別情報に更新する。

【0152】

(S2705)

制御プログラム501が、減設対象のPDEVに関する情報を、メモリ123に格納されているテーブルから削除する。具体的には、例えば、制御プログラム501は、以下の処理：

40

＊PDEV管理テーブル509から、減設対象のPDEVに対応するレコードを削除する；

＊プール管理テーブル505における、減設対象PDEVが属するDPプールに対応したPDEV番号リスト211から、減設対象のPDEVの番号を削除し、且つ、実割当容量212から、減設対象PDEVの容量を引く。また、減設対象のPDEVの番号が先頭PDEV番号214であった場合、先頭PDEV番号214の値を減設対象PDEVの番号の次に大きいPDEV番号の値に更新する。；

＊減設対象PDEVが属するDPプールに対応する実ページ管理テーブル511から、減

50

設対象 P D E V が有する各実ページに対応する情報（レコード）を削除する、を行う。また、この P D E V の減設の際に行われる処理では、P D E V の使用率の代わりに、空き容量を基に、P D E V が選択されても良い。

【0153】

ところで、ストレージシステム 103 の稼動中において、ストレージシステム 103 が単位時間当たりに受け付ける I / O コマンドの数、種別及び I / O 先情報に応じて、特定の実ページのアクセス頻度が高くなってその実ページを有する P D E V の稼働率（負荷）が高くなることもある。

【0154】

そこで、本実施形態では、I / O コマンドの処理に応じて更新される実アクセス状況テーブル 515 及び／又は仮想アクセス状況テーブル 517 に基づいて、再配置処理が行われる。再配置処理については、例えば以下のような三つの例（EX1）～（EX3）が考えられる。

【0155】

（EX1）制御プログラム 501 は、実アクセス状況テーブル 515 及び／又は仮想アクセス状況テーブル 517 を参照して、アクセス頻度或いは稼働率が第1の閾値を超過している P D E V（以下、高負荷 P D E V）が存在するか否かを判断する。P D E V のアクセス頻度は、例えば、P D E V が有する実ページのアクセス頻度の平均である。高負荷 P D E V が特定された場合、制御プログラム 501 は、高負荷 P D E V 内のデータユニットを、稼働率が最も低い P D E V 内の空きの実ページに移動する。移動対象のデータユニットを記憶している実ページは、アクセス頻度が最も高い仮想ページに割り当てられている実ページ、又は、アクセス頻度が最も高い実ページである。図25の例によれば、高負荷 P D E V #05 内のデータユニット「J」が、稼働率が最も低い P D E V #10 に移動される。その結果、高負荷 P D E V #05 の稼働率が40%から20%に下がり、一方、P D E V #10 の稼働率が10%から20%に上がる。データユニットの移動は、複数の P D E V の稼働率が均一になるように行われることが好ましい。

【0156】

（EX2）制御プログラム 501 は、実アクセス状況テーブル 515 及び／又は仮想アクセス状況テーブル 517 を参照して、アクセス頻度或いは稼働率が第2の閾値（第2の閾値<第1の閾値）を下回っている P D E V（以下、低負荷 P D E V）が存在するか否かを判断する。低負荷 P D E V が特定された場合、制御プログラム 501 は、稼働率が最も高い P D E V 内のデータユニットを、低負荷 P D E V 内の空きの実ページに移動する。移動対象のデータユニットを記憶している実ページは、アクセス頻度が最も高い仮想ページに割り当てられている実ページ、又は、アクセス頻度が最も高い実ページである。

【0157】

（EX3）制御プログラム 501 は、実アクセス状況テーブル 515 及び／又は仮想アクセス状況テーブル 517 を参照して、実ページ毎又は仮想ページ毎のリード比率及びライト比率を算出する。リード比率及びライト比率とは、アクセス全体に対するリード及びライトの比率である。具体的には、例えば、リード比率及びライト比率は、リード頻度及びライト頻度と、リード頻度とライト頻度との合計を基に算出することができる。制御プログラム 501 は、以下の（X1）又は（X2）の条件の合致が得られた場合、算出されたリード率又はライト率を有する実ページ内のデータユニットを、算出されたリード率又はライト率に相応しい属性（例えば I / O 性能）を有する実ページに移動する。

（X1）算出されたリード比率又はライト比率が、そのリード比率又はライト比率を有する実ページを提供する P D E V の属性（例えば I / O 性能）に適していない。

（X2）算出されたリード比率又はライト比率が、そのリード比率又はライト比率を有する仮想ページに割り当てられている実ページを提供する P D E V の属性（例えば I / O 性能）に適していない。

【0158】

仮想 L U の所属先が、第1の D P プールから第2の D P プールに切り替えられることが

10

20

30

40

50

ある。この場合、第1のDPプール内の実ページが割り当てられている仮想ページ毎に、以下の(f1)及び(f2)の処理：

(f1) 仮想ページに割り当てられている、第1のDPプール内の複数の実ページから、第2のDPプール内の複数の実ページに、それぞれデータユニットを移動する；

(f2) 複数の移動先の実ページを、複数の移動元の実ページに代えて仮想ページに割り当てる、

を行う。なお、(f1)での移動先の実ページは、仮想ページのアクセス頻度（仮想アクセス状況テーブル517から特定されるアクセス頻度）に適した属性（例えばI/O性能）を有するPDEVから選択される。例えば、仮想ページのアクセス頻度が所定の閾値以上であれば、I/O性能が高いPDEVから移動先の実ページが選択され、一方、仮想ページのアクセス頻度が所定の閾値より低ければ、I/O性能が低いPDEVから移動先の実ページが選択されてよい。

10

【0159】

本実施形態では、前述した通り、DPプールの実ページは、RAIDグループからではなくPDEVから提供される。このため、PDEVに障害が発生した場合には、障害が発生したPDEV（以下、障害PDEV）が有する複数の実ページのうちの、仮想ページに割り当てられている実ページについてのみ、データユニットのリストアがされればよい。ここで、リストアとは、同一仮想ページに割当てられた実ページのうち、障害PDEV内の実ページ以外の実ページに書き込まれたデータ要素およびパリティから、障害PDEV内の当該実ページに格納されていたデータを修復することをいう。具体的には、例えば、図26に示すように、障害PDEVにデータ要素「b」のみが格納されていた場合、制御プログラム501は、他のデータ要素「a」及び「c」とパリティ「P」とを基に、データ要素「b」をリストアし、リストアしたデータ要素「b」を、障害の発生していないPDEVに書き込む。書き込み先となるPDEVは、DPプールでの使用率（もしくは空き容量）が均等になるように選択される。

20

【0160】

障害PDEV内のデータユニットのリストアに関する処理として、リストア優先タイプのリストア処理と、交換優先タイプのリストア処理が考えられる。

【0161】

リストア優先タイプのリストア処理は、例えば次の通りである。すなわち、制御プログラム501が、障害PDEVのうちの、一以上の仮想ページに割り当てられている一以上の実ページ内のデータユニットをリストアし、リストアされた一以上のデータユニットを、複数の他のPDEVに分散しておく。障害PDEVが別のPDEVに交換された後、制御プログラム501が、複数のPDEVの使用率が均一になるように再配置処理を行うことで、複数の他のPDEVから交換後のPDEVにデータユニットを移動する。

30

【0162】

交換優先タイプのリストア処理は、例えば次の通りである。すなわち、障害PDEVが別のPDEVに交換される。制御プログラム501が、障害PDEVのうちの、一以上の仮想ページに割り当てられていた一以上の実ページ内のデータユニットをリストアし、リストアされた一以上のデータユニットを、交換後のPDEVに書き込む。

40

【0163】

以下、それぞれのリストア処理を詳細に説明する。

【0164】

図27は、リストア優先タイプのリストア処理のフローチャートである。

【0165】

(S2501)

制御プログラム501は、実ページ管理テーブル511を参照し、障害PDEVのうちの割当て済み実ページを特定する。制御プログラム501は、特定された割当て済み実ページ毎に、以下の(S2502)～(S2504)を行う。

【0166】

50

(S 2 5 0 2)

制御プログラム 5 0 1 は、割当て済み実ページの割当先の仮想ページに割り当てられている一以上の他の実ページからデータユニットを読み出す。

【0 1 6 7】

(S 2 5 0 3)

制御プログラム 5 0 1 は、S 2 5 0 2 で読み出された一以上のデータユニットを基に、障害 P D E V 内の実ページが記憶しているべきデータユニットをリストアする。

【0 1 6 8】

(S 2 5 0 4)

制御プログラム 5 0 1 は、R A I D が維持され、且つ、障害 P D E V が属する D P プール内の P D E V の使用率が均一になるよう、その D P プール内の他の P D E V から、S 2 5 0 3 でリストアしたデータユニットのライト先の空きの実ページを選択し、選択した実ページに、リストアしたデータユニットを書き込む。制御プログラム 5 0 1 は、障害 P D E V 内の実ページに代えてその書き込み先の実ページを仮想ページに割り当てる。

【0 1 6 9】

(S 2 5 0 5)

障害 P D E V における全ての割当て済み実ページについて S 2 5 0 2 ～ S 2 5 0 4 が行われた後、障害 H D D が別の H D D に交換される。

【0 1 7 0】

(S 2 5 0 6)

制御プログラム 5 0 1 は、D P プール内の P D E V の使用率が均一になるよう、D P プール内の複数の他の P D E V から交換後の P D E V にデータユニットを移動する。制御プログラム 5 0 1 は、データユニットの移動元の実ページに代えて交換後の P D E V 内の移動先実ページを仮想ページに割り当てる。交換優先タイプのリストア処理において、P D E V の使用率の代わりに空き容量が均一になるよう移動先の P D E V を選択しても良い。

【0 1 7 1】

このリストア処理の変形例として、以下が考えられる。例えば、制御プログラム 5 0 1 が、ホスト 1 0 1 からリードコマンドを受信し、且つ、そのリード元の仮想ページに割り当てられている複数の実ページに障害 P D E V 内の実ページが含まれていることを検出した場合に、リードコマンドに応答して読み出された一以上のデータユニットを基に、S 2 5 0 3 及び S 2 5 0 4 が行われて良い。障害 P D E V における全ての割当て済み実ページについて S 2 5 0 3 及び S 2 5 0 4 が行われた後に、S 2 5 0 5 及び S 2 5 0 6 が行われて良い。

【0 1 7 2】

図 2 8 は、交換優先タイプのリストア処理のフローチャートである。

【0 1 7 3】

(S 2 6 0 1)

障害 P D E V が別の P D E V に交換される。

【0 1 7 4】

(S 2 6 0 2)

制御プログラム 5 0 1 が、実ページ管理テーブル 5 1 1 を参照し、障害 P D E V のうちの割当て済み実ページを特定する。制御プログラム 5 0 1 は、特定された割当て済み実ページ毎に、以下の (S 2 6 0 3) ～ (S 2 6 0 5) を行う。

【0 1 7 5】

(S 2 6 0 3)

制御プログラム 5 0 1 は、割当て済み実ページの割当先の仮想ページに割り当てられている一以上の他の実ページからデータユニットを読み出す。

【0 1 7 6】

(S 2 6 0 4)

制御プログラム 5 0 1 は、S 2 6 0 3 で読み出された一以上のデータユニットを基に、

10

20

30

40

50

障害 P D E V 内の実ページが記憶しているべきデータユニットをリストアする。

【0177】

(S2605)

制御プログラム501は、リストアしたデータユニットを、交換後の P D E V に書き込む。

【0178】

リストア優先とするか交換優先とするかは、ストレージシステム103に対するアクセスの状況に応じて決められても良いし、ユーザから予め指定されていても良い。

【0179】

以上、本発明の一つの実施形態を説明したが、本発明は、この実施形態に限定されるものでなく、その要旨を逸脱しない範囲で種々変更可能であることはいうまでもない。例えば、各要素は、番号或いは名称に代えて又は加えて他種の識別情報で識別されて良い。また、例えば、パリティの生成は、制御プログラム501に代えて、ハードウェア回路によって行われて良い。また、例えば、実ページ毎のアクセスカウンタ244及び仮想ページ毎のアクセスカウンタ253のうちの少なくとも一方が、アクセスカウンタに代えて、ライト用のカウンタと、リード用のカウンタとの少なくとも一方であっても良い。また、例えば、実アクセス状況テーブル515は無くても良い。この場合には、仮想ページのアクセス頻度を、仮想ページに対応する実ページのアクセス頻度とみなすことができる。

【符号の説明】

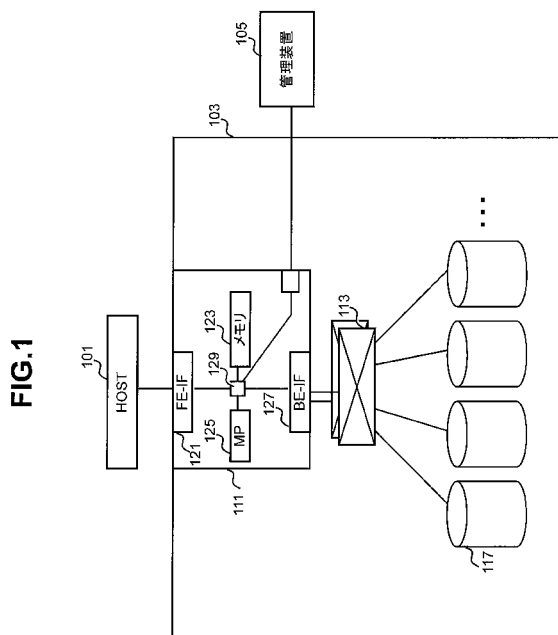
【0180】

103…ストレージシステム

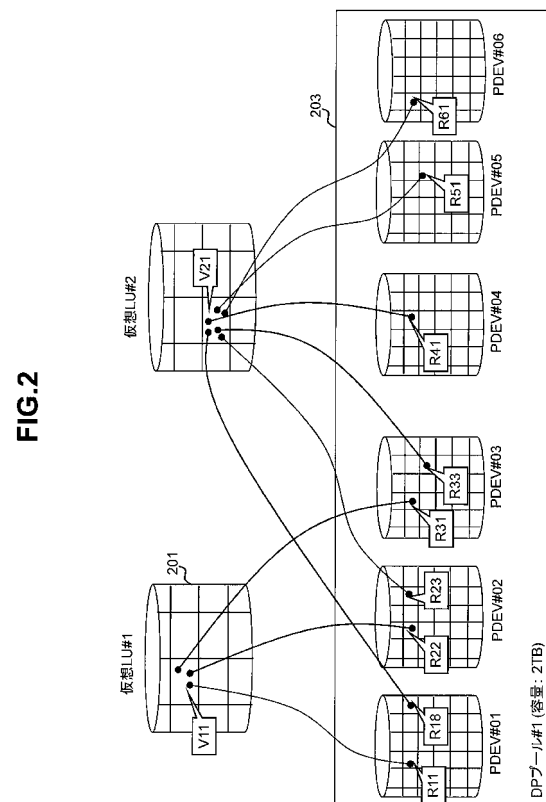
10

20

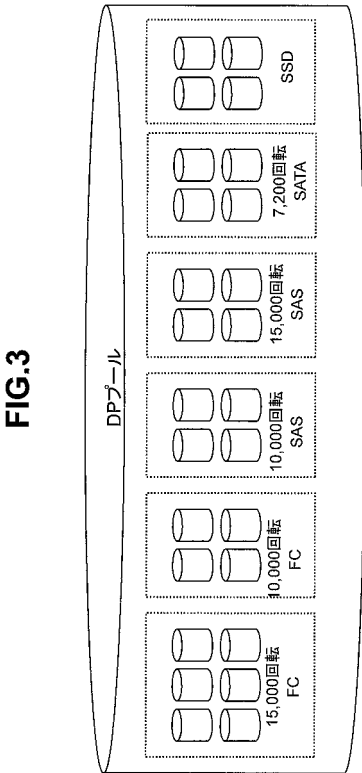
【図1】



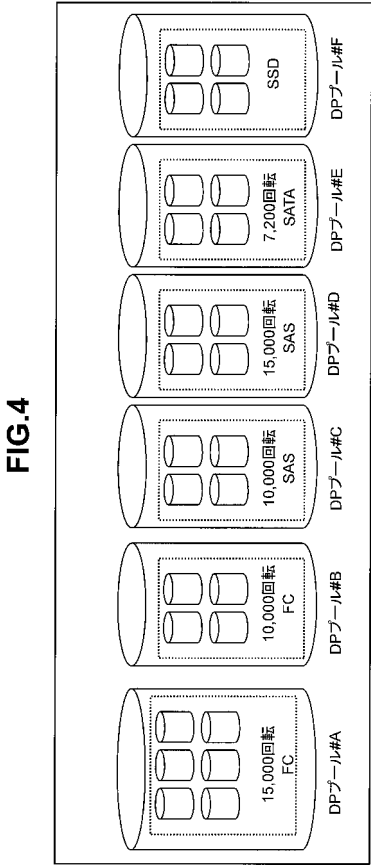
【図2】



【図 3】

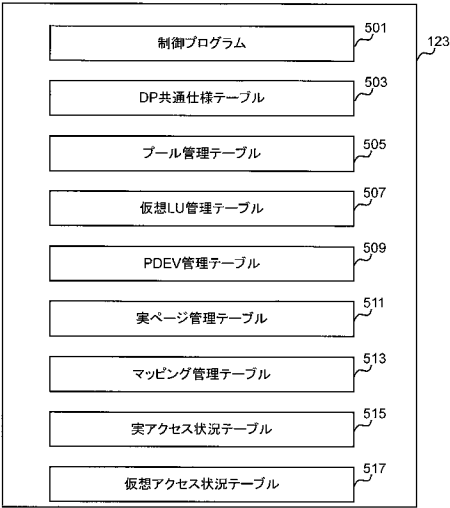


【図 4】



【図 5】

FIG.5



【図 7】

FIG.7

プール番号	PDEV番号リスト	実割当容量	使用容量	先頭PDEV番号	
				214	505
01	01, 02, 03	2TB	0MB	01	
02	04, 05, 06, 07	3TB	100MB	06	
...	...	...	...	...	

【図 6】

FIG.6

200		503	201
項目	値		
作成可能最大プール数	50		
実ページ容量	32KB		

【図 8】

FIG.8

220		221	222		223	224
LUN		所属グループ番号	RAIDレベル	コンピネーション	容量	
01	01		RAID5	2D+1P	1TB	
02	02		RAID6	3D+2P	2TB	
...	...		...	...	...	

【図 10】

FIG.10

240		241	242	243	244	511
PDEV番号		実ページ番号	LUN	仮想ページ番号	アクセスカウンタ	
01		R11	01	V11	5	
01		R12	null	null	0	
...		...	...	...	...	

【図 9】

FIG.9

230		231	232	233	234	235	236
PDEV番号		所属グループ番号	IF種別	PDEV性能	容量	空き容量	デバイス種別
01		01	SAS	7,200rpm	500GB	200GB	HDD
02		01	SAS	7,200rpm	500GB	300GB	HDD
03		null	SATA	7,200rpm	500GB	500GB	HDD
...		...	...	...	...	...	...

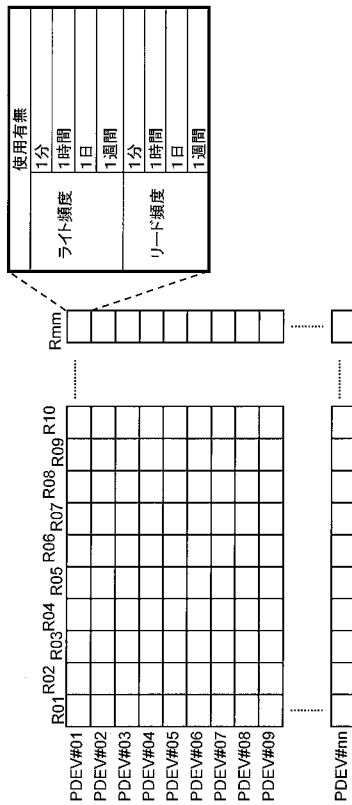
【図 11】

FIG.11

250		251	252	253
LUN		仮想ページ番号	実ページ識別情報	アクセスカウンタ
01		V11	{01, R11}, {02, R22}, ...	5
01		V12	null	0
...		...	...	...

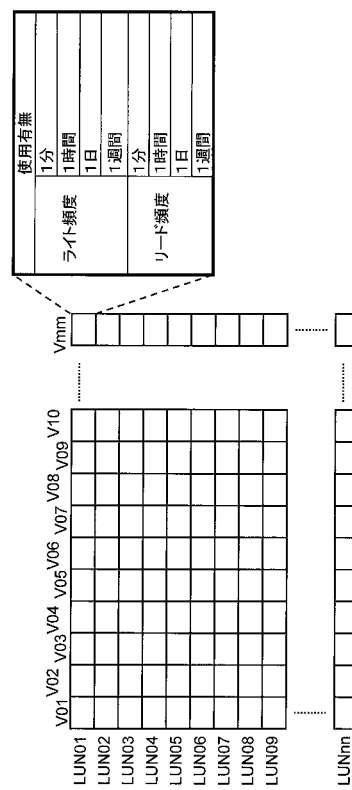
【図 1 2】

FIG.12



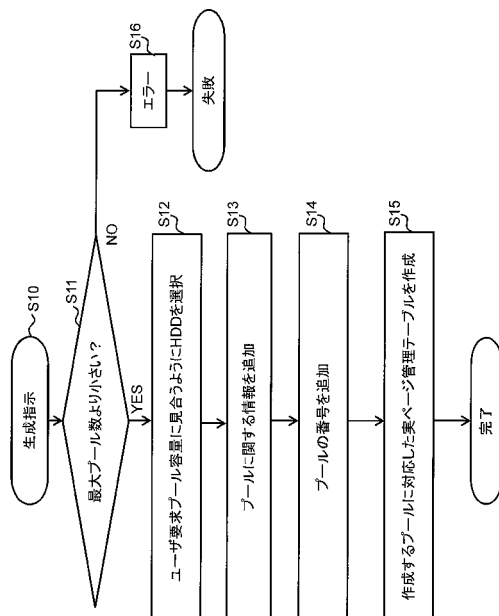
【図 1 3】

FIG.13



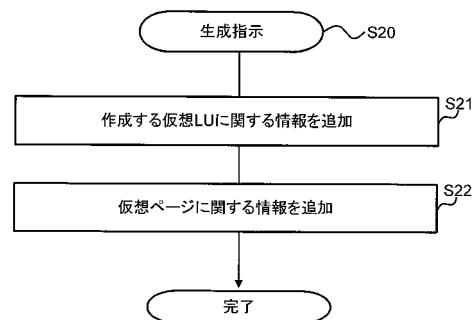
【図 1 4】

FIG.14

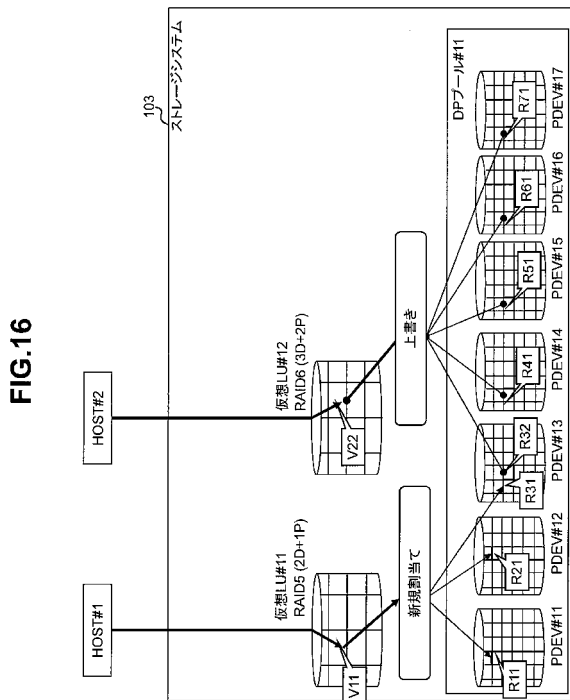


【図 1 5】

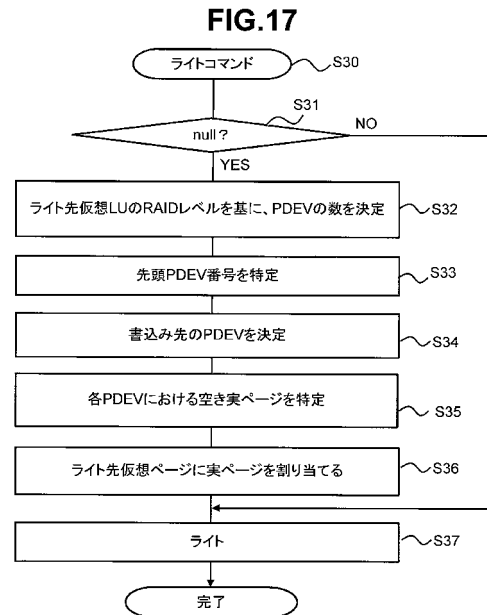
FIG.15



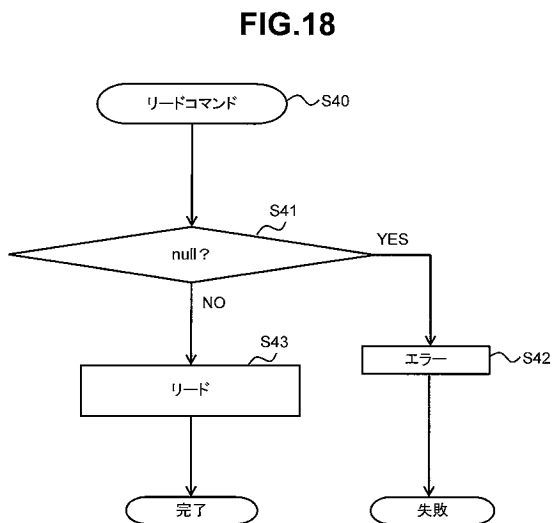
【図 1 6】



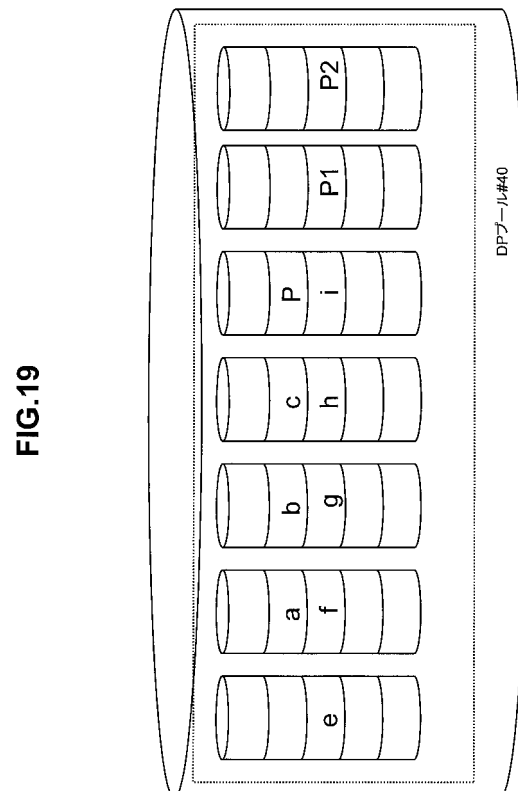
【図 17】



【図 18】

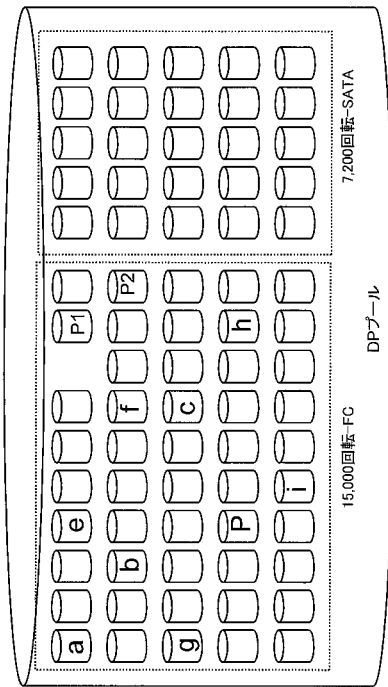


【图 19】



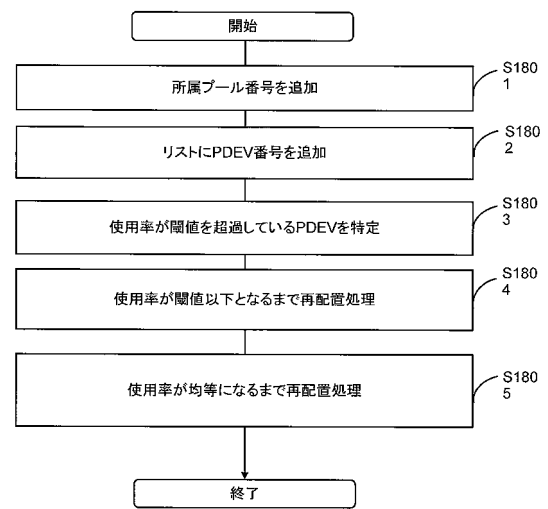
【図 20】

FIG.20



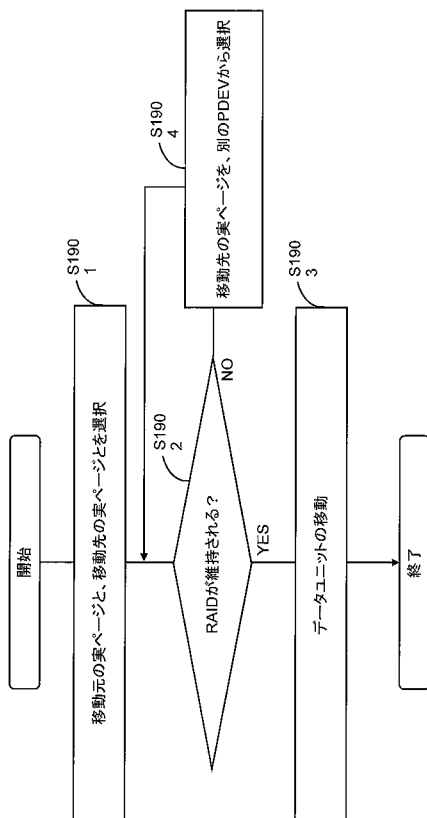
【図 21】

FIG.21



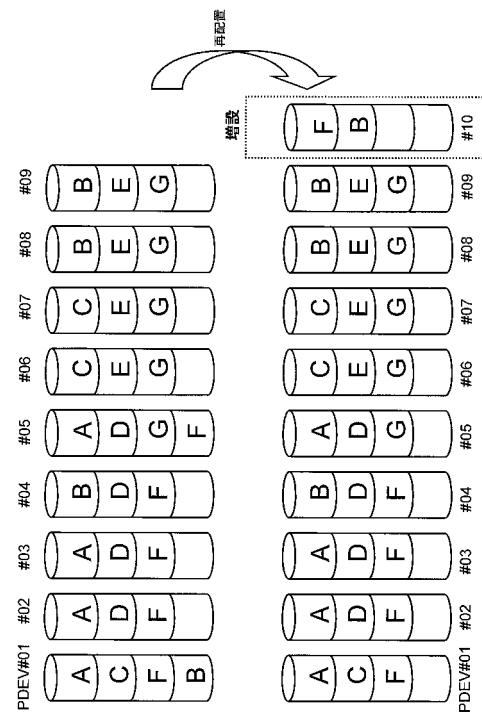
【図 22】

FIG.22



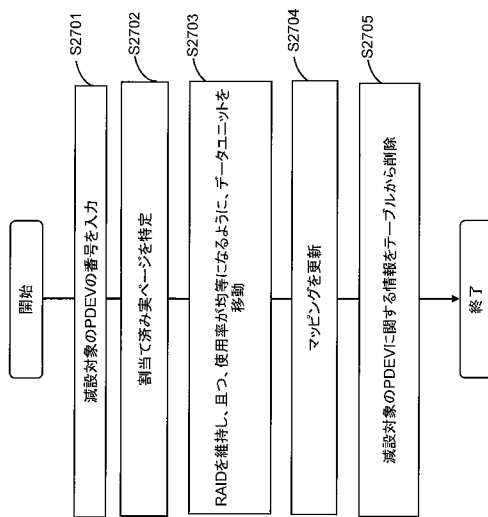
【図 23】

FIG.23



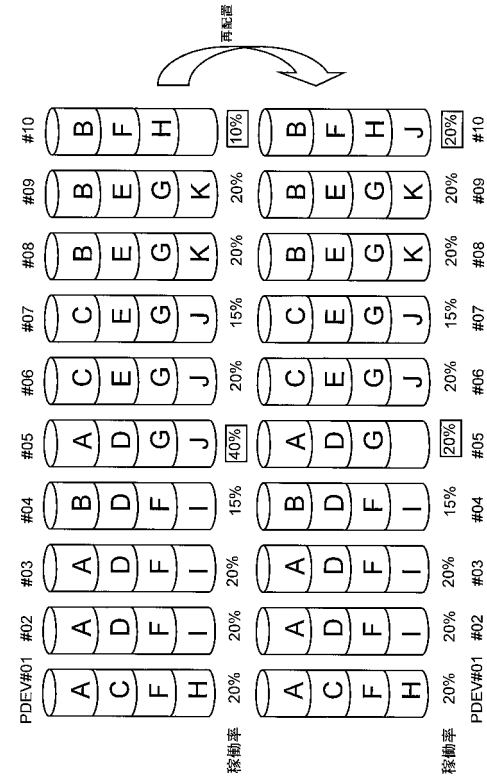
【図 24】

FIG.24



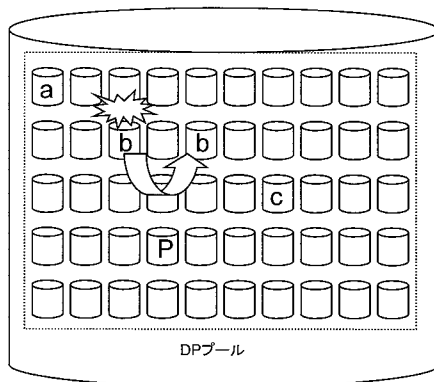
【図 25】

FIG.25



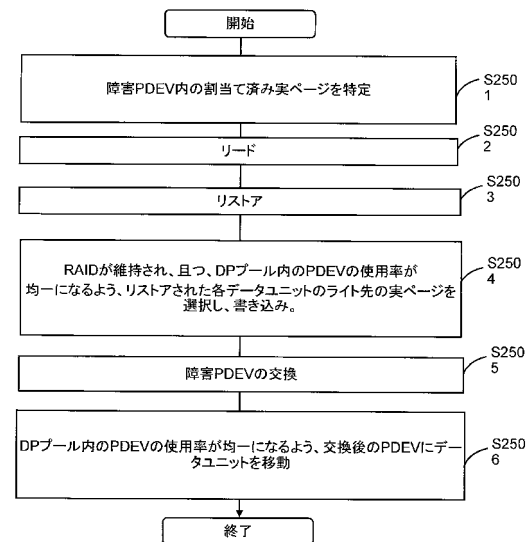
【図 26】

FIG.26



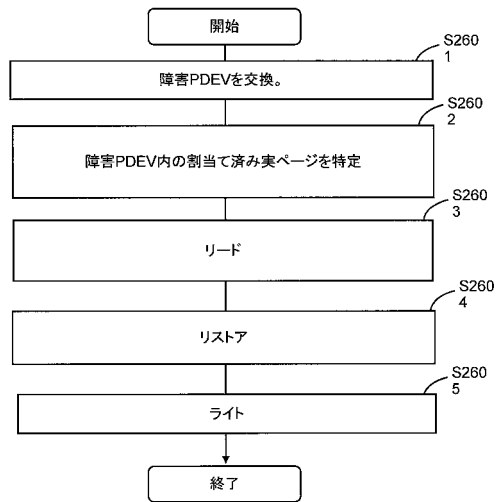
【図 27】

FIG.27



【図 28】

FIG.28



フロントページの続き

審査官 稲葉 崇

- (56)参考文献 特表2006-505831 (JP, A)
米国特許出願公開第2008/0275928 (US, A1)
特開2008-146574 (JP, A)
国際公開第2003/077111 (WO, A1)
- (58)調査した分野(Int.Cl., DB名)
G06F 3/06-3/08
G06F 13/10-13/14
G06F 12/00