

(19) 日本国特許庁 (JP)

(12) 特 許 公 報 (B2)

(11) 特許番号

特許第6510635号
(P6510635)

(45) 発行日 令和1年5月8日 (2019.5.8)

(24) 登録日 平成31年4月12日 (2019.4.12)

(51) Int. Cl. F I
G 0 6 F 3 / 0 6 (2006.01)
 G 0 6 F 3 / 0 6 3 0 6 Z
 G 0 6 F 3 / 0 6 3 0 1 Z

請求項の数 12 (全 43 頁)

(21) 出願番号	特願2017-511430 (P2017-511430)	(73) 特許権者	000005108
(86) (22) 出願日	平成27年4月9日 (2015.4.9)		株式会社日立製作所
(86) 国際出願番号	PCT/JP2015/061165		東京都千代田区丸の内一丁目6番6号
(87) 国際公開番号	W02016/163016	(74) 代理人	110002365
(87) 国際公開日	平成28年10月13日 (2016.10.13)		特許業務法人サンネクト国際特許事務所
審査請求日	平成29年8月25日 (2017.8.25)	(72) 発明者	山本 貴大
			東京都千代田区丸の内一丁目6番6号 株
			式会社日立製作所内
		(72) 発明者	藤本 和久
			東京都千代田区丸の内一丁目6番6号 株
			式会社日立製作所内
		(72) 発明者	坪 弘明
			東京都千代田区丸の内一丁目6番6号 株
			式会社日立製作所内

最終頁に続く

(54) 【発明の名称】 ストレージシステム及びデータ制御方法

(57) 【特許請求の範囲】

【請求項 1】

第1の記憶装置と、

前記第1の記憶装置よりも書き込み上限回数が少なく、かつ、単位面積当たりの記憶容量が多い第2の記憶装置と、

ホストに提供する仮想ボリュームに対して前記第1の記憶装置及び前記第2の記憶装置から記憶領域を割り当てるプロセッサとを備え、

前記プロセッサは、

前記仮想ボリュームに対して前記第2の記憶装置から割り当てられている記憶領域のうち、前記ホストからのライト頻度が予め定められたライト閾値よりも多い記憶領域に格納されているデータを前記第2の記憶装置の記憶領域から前記第1の記憶装置の記憶領域に再配置し、

新たに割り当てる記憶領域について、当該記憶領域を提供するホストへの割り当て済み記憶領域のライト頻度に基づいて、新たに割り当てる記憶領域へのライト頻度を予測し、予測したライト頻度に基づいて前記第1の記憶装置と前記第2の記憶装置とのいずれに記憶領域を割り当てるかを決定して割り当てる

ことを特徴とするストレージシステム。

【請求項 2】

前記プロセッサは、

前記仮想ボリュームに対して前記第1の記憶装置から割り当てられている記憶領域のう

10

20

ち、前記ホストからのライト頻度が前記ライト閾値以下である記憶領域に格納されているデータを前記第 1 の記憶装置の記憶領域から前記第 2 の記憶装置の記憶領域に再配置することを特徴とする請求項 1 に記載のストレージシステム。

【請求項 3】

前記プロセッサは、

前記仮想ボリュームに対して前記第 1 の記憶装置又は前記第 2 の記憶装置のうち、何れかの記憶装置から新規に記憶領域を割り当てる場合、予め定められたワークロードタイプに基づいて、前記第 1 の記憶装置又は前記第 2 の記憶装置のうち、何れかの記憶装置から新規に記憶領域を割り当てる

ことを特徴とする請求項 1 に記載のストレージシステム。

10

【請求項 4】

前記プロセッサは、

前記第 1 の記憶装置の書き込み上限回数と、前記第 2 の記憶装置の書き込み上限回数とに基づいて、前記ライト閾値を算出する

ことを特徴とする請求項 1 に記載のストレージシステム。

【請求項 5】

第 1 の記憶装置と、

前記第 1 の記憶装置よりも書き込み上限回数が少なく、かつ、単位面積当たりの記憶容量が多い第 2 の記憶装置と、

ホストに提供する仮想ボリュームに対して前記第 1 の記憶装置及び前記第 2 の記憶装置から記憶領域を割り当てるプロセッサとを備え、

20

前記プロセッサは、

前記仮想ボリュームに対して前記第 2 の記憶装置から割り当てられている記憶領域のうち、前記ホストからのライト頻度が予め定められたライト閾値よりも多い記憶領域に格納されているデータを前記第 2 の記憶装置の記憶領域から前記第 1 の記憶装置の記憶領域に再配置し、

前記ホストからのライト頻度と、前記ライト閾値とに基づいて、寿命を維持するために必要な前記第 1 の記憶装置の推奨容量及び前記第 2 の記憶装置の推奨容量を通知する

ことを特徴とするストレージシステム。

30

【請求項 6】

書き込み上限回数に制限のない第 3 の記憶装置を備え、

前記プロセッサは、

前記第 1 の記憶装置及び前記第 2 の記憶装置から提供される記憶領域を第 1 の階層に設定し、前記第 3 の記憶装置から提供される記憶領域を第 2 の階層に設定し、

前記仮想ボリュームに対して前記第 2 の階層から割り当てられている記憶領域のうち、前記ホストからのライト頻度及びリード頻度を合計した I/O 頻度が予め定められたティア間 I/O 閾値以上である記憶領域に格納されているデータを前記第 2 の階層の記憶領域から前記第 1 の階層の記憶領域に再配置し、

前記仮想ボリュームに対して前記第 1 の階層から割り当てられている記憶領域であって、かつ、前記第 2 の記憶装置から割り当てられている記憶領域のうち、前記ホストからのライト頻度が前記ライト閾値よりも多い記憶領域に格納されているデータを前記第 2 の記憶装置の記憶領域から前記第 1 の記憶装置の記憶領域に再配置する

40

ことを特徴とする請求項 1 に記載のストレージシステム。

【請求項 7】

前記プロセッサは、

前記仮想ボリュームに対して前記第 1 の階層から割り当てられている記憶領域であって、かつ、前記第 1 の記憶装置から割り当てられている記憶領域のうち、前記ホストからのライト頻度が前記ライト閾値以下である記憶領域に格納されているデータを前記第 1 の記憶装置の記憶領域から前記第 2 の記憶装置の記憶領域に再配置する

ことを特徴とする請求項 6 に記載のストレージシステム。

50

【請求項 8】

第 1 の記憶装置と、

前記第 1 の記憶装置よりも書き込み上限回数が少なく、かつ、単位面積当たりの記憶容量が多い第 2 の記憶装置と、

ホストに提供する仮想ボリュームに対して前記第 1 の記憶装置及び前記第 2 の記憶装置から記憶領域を割り当てるプロセッサと、

書き込み上限回数に制限のない第 3 の記憶装置とを備え、

前記プロセッサは、

前記仮想ボリュームに対して前記第 2 の記憶装置から割り当てられている記憶領域のうち、前記ホストからのライト頻度が予め定められたライト閾値よりも多い記憶領域に格納されているデータを前記第 2 の記憶装置の記憶領域から前記第 1 の記憶装置の記憶領域に再配置し、

前記第 1 の記憶装置及び前記第 2 の記憶装置から提供される記憶領域を第 1 の階層に設定し、前記第 3 の記憶装置から提供される記憶領域を第 2 の階層に設定し、

前記仮想ボリュームに対して前記第 2 の階層から割り当てられている記憶領域のうち、前記ホストからのライト頻度及びリード頻度を合計した I/O 頻度が予め定められたティア間 I/O 閾値以上である記憶領域に格納されているデータを前記第 2 の階層の記憶領域から前記第 1 の階層の記憶領域に再配置し、

前記仮想ボリュームに対して前記第 1 の階層から割り当てられている記憶領域であって、かつ、前記第 2 の記憶装置から割り当てられている記憶領域のうち、前記ホストからのライト頻度が前記ライト閾値よりも多い記憶領域に格納されているデータを前記第 2 の記憶装置の記憶領域から前記第 1 の記憶装置の記憶領域に再配置し、

前記仮想ボリュームに対して前記第 1 の階層から割り当てられている記憶領域のうち、前記ホストからのライト頻度が寿命維持のために予め定められたライトデモーション閾値以上である記憶領域に格納されたデータを前記第 2 の階層の記憶領域に再配置することを特徴とするストレージシステム。

【請求項 9】

前記新たに割り当てる記憶領域にかかるワークロードタイプに基づいて、前記記憶領域の配置先を前記第 1 の記憶装置と前記第 2 の記憶装置とのいずれかに決定し、

前記ワークロードタイプにより決定できない場合に、前記ホストのライト頻度を予測する

ことを特徴とする請求項 1 に記載のストレージシステム。

【請求項 10】

第 1 の記憶装置と、

前記第 1 の記憶装置よりも書き込み上限回数が少なく、かつ、単位面積当たりの記憶容量が多い第 2 の記憶装置と、

ホストに提供する仮想ボリュームに対して前記第 1 の記憶装置及び前記第 2 の記憶装置から記憶領域を割り当てるプロセッサとを備えたストレージシステムのデータ制御方法であって、

前記プロセッサが、

前記仮想ボリュームに対して前記第 2 の記憶装置から割り当てられている記憶領域のうち、前記ホストからのライト頻度が予め定められたライト閾値よりも多い記憶領域に格納されているデータを前記第 2 の記憶装置の記憶領域から前記第 1 の記憶装置の記憶領域に再配置する第 1 のステップと、

前記仮想ボリュームに対して前記第 1 の記憶装置から割り当てられている記憶領域のうち、前記ホストからのライト頻度が前記ライト閾値以下である記憶領域に格納されているデータを前記第 1 の記憶装置の記憶領域から前記第 2 の記憶装置の記憶領域に再配置する第 2 のステップとを備え、

新たに割り当てる記憶領域について、当該記憶領域を提供するホストへの割り当て済み記憶領域のライト頻度に基づいて、新たに割り当てる記憶領域へのライト頻度を予測し、

10

20

30

40

50

予測したライト頻度に基づいて前記第 1 の記憶装置と前記第 2 の記憶装置とのいずれに記憶領域を割り当てるかを決定して割り当てる

ことを特徴とするデータ制御方法。

【請求項 1 1】

第 1 の記憶装置と、

前記第 1 の記憶装置よりも書き込み上限回数が少なく、かつ、単位面積当たりの記憶容量が多い第 2 の記憶装置と、

ホストに提供する仮想ボリュームに対して前記第 1 の記憶装置及び前記第 2 の記憶装置から記憶領域を割り当てるプロセッサとを備えたストレージシステムのデータ制御方法であって、

前記プロセッサが、

前記仮想ボリュームに対して前記第 2 の記憶装置から割り当てられている記憶領域のうち、前記ホストからのライト頻度が予め定められたライト閾値よりも多い記憶領域に格納されているデータを前記第 2 の記憶装置の記憶領域から前記第 1 の記憶装置の記憶領域に再配置し、

前記ホストからのライト頻度と、前記ライト閾値とに基づいて、寿命を維持するために必要な前記第 1 の記憶装置の推奨容量及び前記第 2 の記憶装置の推奨容量を通知する

ことを特徴とするデータ制御方法。

【請求項 1 2】

第 1 の記憶装置と、

前記第 1 の記憶装置よりも書き込み上限回数が少なく、かつ、単位面積当たりの記憶容量が多い第 2 の記憶装置と、

ホストに提供する仮想ボリュームに対して前記第 1 の記憶装置及び前記第 2 の記憶装置から記憶領域を割り当てるプロセッサと、

書き込み上限回数に制限のない第 3 の記憶装置とを備えたストレージシステムのデータ制御方法であって、

前記プロセッサが、

前記仮想ボリュームに対して前記第 2 の記憶装置から割り当てられている記憶領域のうち、前記ホストからのライト頻度が予め定められたライト閾値よりも多い記憶領域に格納されているデータを前記第 2 の記憶装置の記憶領域から前記第 1 の記憶装置の記憶領域に再配置し、

前記第 1 の記憶装置及び前記第 2 の記憶装置から提供される記憶領域を第 1 の階層に設定し、前記第 3 の記憶装置から提供される記憶領域を第 2 の階層に設定し、

前記仮想ボリュームに対して前記第 2 の階層から割り当てられている記憶領域のうち、前記ホストからのライト頻度及びリード頻度を合計した I/O 頻度が予め定められたティア間 I/O 閾値以上である記憶領域に格納されているデータを前記第 2 の階層の記憶領域から前記第 1 の階層の記憶領域に再配置し、

前記仮想ボリュームに対して前記第 1 の階層から割り当てられている記憶領域であって、かつ、前記第 2 の記憶装置から割り当てられている記憶領域のうち、前記ホストからのライト頻度が前記ライト閾値よりも多い記憶領域に格納されているデータを前記第 2 の記憶装置の記憶領域から前記第 1 の記憶装置の記憶領域に再配置し、

前記仮想ボリュームに対して前記第 1 の階層から割り当てられている記憶領域のうち、前記ホストからのライト頻度が寿命維持のために予め定められたライトデモーション閾値以上である記憶領域に格納されたデータを前記第 2 の階層の記憶領域に再配置する

ことを特徴とするデータ制御方法。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、ストレージシステム及びデータ制御方法に関し、特に同一種別であって特性の異なる複数の記憶装置を利用して、記憶装置間でデータを自動的に再配置するストレ

10

20

30

40

50

ジシステム及びデータ制御方法に適用して好適なものである。

【背景技術】

【0002】

近年、HDD (Hard Disk Drive) やSSD (Solid State Drive) の普及により、ストレージシステムに搭載される記憶装置の種類が多様化している。特にSSDについては、データ格納方式としてSLC (Single Level Cell)、2bit MLC (Multi Level Cell) 又はTLC (Triple Level Cell) 等があり、同一種別の記憶装置であっても、寿命 (書き込み上限回数) 及びビットコストが異なるという特性がある。

【0003】

寿命は、一般にSLCが最も長く、MLC、TLCの順に短くなる。またビットコストは、TLCが最も低く、MLC、SLCの順に高くなる。よってSSDを搭載するストレージシステムにおいては、その特性を考慮してデータを適切に配置することが重要となる。

10

【0004】

特許文献1及び2には、複数の記憶装置が搭載されたストレージシステムを管理又は制御する技術が開示されている。具体的に特許文献1には、複数の記憶装置をプールとして統合的に管理し、仮想的なボリュームをホスト装置に提供する技術が開示されている。

【0005】

また特許文献2には、HDDとSSDとを異なる階層の記憶装置として管理し、ホスト装置からのアクセス頻度に応じてデータを自動的に再配置し、高階層でのデータヒット率を向上させることにより、ホスト装置に対する応答性能を向上させる技術が開示されている。

20

【先行技術文献】

【特許文献】

【0006】

【特許文献1】 米国特許第7447832号明細書

【特許文献2】 米国特許第8041682号明細書

【発明の概要】

【発明が解決しようとする課題】

【0007】

しかし特許文献1及び2に記載の技術では、同一種別であって特性 (寿命及びビットコスト) の異なる複数の記憶装置 (例えばSSD) が混在したストレージシステムに対して書き込みを行った場合、特性が考慮されることなく、各記憶装置にデータが書き込まれる。

30

【0008】

SSDの特性を考慮せずに書き込みを行うと、書き込み上限回数の少ないSSDの寿命が書き込み上限回数の多いSSDよりも先に尽きることになり、寿命の尽きたSSDに対する交換回数が増え、ストレージシステムの運用管理コストが高くなるという課題がある。

【0009】

本発明は以上の点を考慮してなされたもので、同一種別であって特性の異なる記憶装置の寿命を延ばし、運用管理コストを削減し得るストレージシステム及びデータ制御方法を提案する。

40

【課題を解決するための手段】

【0010】

かかる課題を解決するために、本発明においては、第1の記憶装置と、第1の記憶装置よりも書き込み上限回数が少なく、かつ、単位面積当たりの記憶容量が多い第2の記憶装置と、ホストに提供する仮想ボリュームに対して第1の記憶装置及び第2の記憶装置から記憶領域を割り当てるプロセッサとを備え、プロセッサは、仮想ボリュームに対して第2の記憶装置から割り当てられている記憶領域のうち、ホストからのライト頻度が予め定め

50

られたライト閾値よりも多い記憶領域に格納されているデータを第1の記憶装置の記憶領域に再配置する。

【0011】

またかかる課題を解決するために、本発明においては、第1の記憶装置と、第1の記憶装置よりも書き込み上限回数が少なく、かつ、単位面積当たりの記憶容量が多い第2の記憶装置と、ホストに提供する仮想ボリュームに対して第1の記憶装置及び第2の記憶装置から記憶領域を割り当てるプロセッサとを備えたストレージシステムのデータ制御方法であって、プロセッサが、仮想ボリュームに対して第2の記憶装置から割り当てられている記憶領域のうち、ホストからのライト頻度が予め定められたライト閾値よりも多い記憶領域に格納されているデータを第1の記憶装置の記憶領域に再配置する第1のステップと、仮想ボリュームに対して第1の記憶装置から割り当てられている記憶領域のうち、ホストからのライト頻度がライト閾値以下である記憶領域に格納されているデータを第2の記憶装置の記憶領域に再配置する第2のステップとを備える。

10

【発明の効果】

【0012】

本発明によれば、同一種別であって特性の異なる記憶装置の寿命を延ばし、運用管理コストを削減することができる。

【図面の簡単な説明】

【0013】

【図1】第1の実施の形態における計算機システムの全体構成図である。

20

【図2】ストレージシステムの論理構成図である。

【図3】TLC-MLC間で行われるページ再配置処理の概念構成図である。

【図4】TLC-TLC間で行われるページ再配置処理の概念構成図である。

【図5】共有メモリの内部構成図である。

【図6】ローカルメモリの内部構成図である。

【図7】ページ毎モニタテーブルの論理構成図である。

【図8】パリティグループ毎モニタテーブルの論理構成図である。

【図9】パリティグループ毎再配置管理テーブルの論理構成図である。

【図10】プール毎再配置管理テーブルの論理構成図である。

【図11】ホストI/O処理のフローチャートである。

30

【図12】デスページ処理のフローチャートである。

【図13】寿命情報採取処理のフローチャートである。

【図14】閾値決定処理のフローチャートである。

【図15】再配置処理のフローチャートである。

【図16】TLC-MLC間再配置処理のフローチャートである。

【図17】ライトリバランス処理のフローチャートである。

【図18】性能リバランス処理のフローチャートである。

【図19】新規割り当て決定処理のフローチャートである。

【図20】推奨容量算出方法の概念図である。

【図21】パラメータ設定画面の画面構成の一例である。

40

【図22】警告画面の画面構成の一例である。

【図23】第2の実施の形態における計算機システムの全体構成図である。

【図24】ストレージシステムの論理構成図である。

【図25】ライトデモーション処理によるページ再配置処理の概念構成図である。

【図26】パリティグループ毎再配置管理テーブルの論理構成図である。

【図27】プール毎再配置管理テーブルの論理構成図である。

【図28】再配置処理のフローチャートである。

【図29】ティア間再配置処理のフローチャートである。

【図30】ライトデモーション処理のフローチャートである。

【図31】新規割り当て決定処理のフローチャートである。

50

【図 3 2】ライト追加可能量及びライト削減要求量の算出方法の概念図である。

【発明を実施するための形態】

【0014】

以下、図面を参照しながら本発明の一実施の形態を説明する。なお以下に説明する実施の形態は、本発明を実現するための一例であって、本発明の技術的範囲を限定するものではない。また各図において共通の構成については、同一の参照番号を付してその説明を省略する。また各図において複数の同一部材を図示する場合は、「201A」、「201B」のようにアルファベットを付して区別する一方で、総称して呼ぶ場合には「201」のようにアルファベットを省略する。さらに本発明の実施の形態は、汎用コンピュータ上で稼動するソフトウェアに実装してもよいし、専用ハードウェアに実装してもよい。またソフトウェアとハードウェアとの組み合わせに実装してもよい。以下の説明では、管理用の情報をテーブル形式で説明するが、管理用の情報は必ずしもテーブルによるデータ構造で表現されなくてもよく、「リスト」、「DB」、「キュー」などについて単に「情報」と呼ぶことがある。また「プログラム」を主語（動作主体）として本発明の実施の形態における各処理について説明する場合がある。プログラムはプロセッサによって実行されることで定められた処理をメモリ及び通信ポート（通信制御装置）を用いながら行うため、プロセッサを主語とした説明としてもよい。プログラムの一部又は全部は専用ハードウェアで実現してもよく、またモジュール化されていてもよい。各種プログラムはプログラム配布サーバや記憶メディアによって各計算機にインストールされてもよい。

【0015】

（1）第1の実施の形態

第1の実施の形態は、記憶装置として半導体メモリ（SSD）のみを搭載するストレージシステムにおいて、特性の異なる半導体メモリ間（例えばTLC-MLC間）でデータを再配置することにより、半導体メモリの長寿命化を実現するとともに、ビットコストを低減しようとするものである。さらには同一種別の半導体メモリ間（例えばTLC-TLC間）でデータを再配置することにより、負荷分散を実現して長寿命化を実現しようとするものである。

【0016】

（1-1）計算機システムの全体構成

図1は、第1の実施の形態における計算機システム1の全体構成を示す。計算機システム1は、ホスト101、管理サーバ102、ネットワーク103及びストレージシステム104から構成される。ホスト101は、例えば一般的なサーバであり、ネットワーク103を介してストレージシステム104のポート105に接続される。ホスト101は、ストレージシステム104に対してリード要求又はライト要求を発行して、データの読み書きを実行する。

【0017】

ネットワーク103は、例えばSAN（Storage Area Network）やイーサネット（登録商標）等の通信回線である。管理サーバ102は、ネットワーク103を介して、ストレージシステム104の保守I/F106又はポート105に接続される。ストレージ管理者は、管理サーバ102を操作して、ストレージシステム104を運用する上で必要な各種設定や管理を行う。

【0018】

次にストレージシステム104の内部構成について説明する。ストレージシステム104は、ポート105、保守I/F106、プロセッサパッケージ107、キャッシュメモリ108、共有メモリ109、ドライブ111及びドライブ112等を備える。これらは内部ネットワーク110を介して互いに通信可能に接続される。

【0019】

キャッシュメモリ108は、ストレージシステム104のI/O処理のスループットやレスポンスを向上させるために、データを一時的なキャッシュとして格納するための高速アクセス可能なメモリである。プロセッサパッケージ107は、ローカルメモリ113及

びプロセッサ 114 を備えて構成される。なおここではプロセッサパッケージ 107 は、107A 及び 107B の 2 つを図示しているが、数はこれに限定されない。

【0020】

プロセッサ 114 は、ホスト 101 からのリード要求及びライト要求を処理するために、ドライブ 111 及び 112 と、キャッシュメモリ 108 との間のデータの転送処理を実行する。共有メモリ 109 は、プロセッサ 114 がリード要求又はライト要求を処理し、またストレージシステム 104 の機能（ボリュームのコピー機能など）を実行する上で必要な制御用の情報を格納するメモリである。共有メモリ 109 は、複数のプロセッサパッケージ 107（ここでは 107A 及び 107B）のプロセッサ 114 により共有される情報を格納する。

10

【0021】

ローカルメモリ 113 は、プロセッサ 114 がリード要求又はライト要求を処理し、またストレージシステム 104 の機能を実行する上で必要な制御用の情報を格納するメモリである。ローカルメモリ 113 は、各プロセッサ 114 により占有される情報を格納する。ローカルメモリ 113 には、例えばプロセッサ 114 により実行するプログラムが格納される。

【0022】

ドライブ 111 及び 112 は、複数の半導体メモリ（SSD：Solid State Drive）から構成される記憶装置である。ここでのドライブ 111 は、1 つのセルに 2 ビットの情報を格納する MLC SSD（Multi Level Cell SSD）であり、ドライブ 112 は、1 つのセルに 3 ビットの情報を格納する TLC（Triple Level Cell SSD）である。

20

【0023】

MLC SSD と、TLC SSD との違いは、上記の通りセル当たりの情報量が異なることに加えて、寿命（書き込み上限回数）及びビットコストが異なる。寿命は、一般にセル当たりの情報量が多くなるほど短くなる。よって MLC SSD の方が TLC SSD よりも長い。またビットコストは、単位面積当たりの情報量が多いほど低く抑えられるため、MLC SSD の方が TLC SSD よりも高い。

【0024】

すなわち MLC SSD は、寿命が長くてビットコストが高く、TLC SSD は、寿命が短くてビットコストが低いと言える。なおここでは MLC SSD 及び TLC SSD を利用してドライブ 111 及び 112 を構成するとしているが、これに限らず、SLC SSD（Single Level Cell SSD）を利用するとしてもよい。

30

【0025】

ドライブ 111 及び 112 について、以降は同一種別であって特性の異なる記憶装置であることを強調する場合にはドライブ 111 を MLC SSD と呼び、ドライブ 112 を TLC SSD と呼ぶ場合がある。また複数の MLC SSD から構成されるパリティグループを MLC PG と呼び、複数の TLC SSD から構成されるパリティグループを TLC PG と呼ぶ場合がある。

【0026】

（1-2）ストレージシステムの論理構成

40

図 2 は、ストレージシステム 104 の論理構成を示す。プロセッサ 114 がホスト 101 に提供する仮想ボリューム 202A 及び 202B は、ホスト 101 から認識される論理的な記憶領域であり、ホスト 101 からのリード要求又はライト要求の発行対象となる記憶領域である。

【0027】

プール 204 は、1 個以上のプールボリューム 206A～206E により構成される。プールボリューム 206A～206E は、それぞれ MLC SSD 又は TLC SSD の何れかの記憶領域から構成される。一又は複数のプールボリューム 206A～206E により、パリティグループ（PG：Parity Group）205A 又は 205B が形成される。

【0028】

50

なおここでのパリティグループ205Aは、MLC SSDのみで構成され、パリティグループ205BはTLC SSDのみで構成されているが、必ずしもこれに限らず、MLC SSD及びTLC SSDが混在して1つのパリティグループを構成するとしてもよい。プールボリューム206は、パリティグループ205の一部又は全部の領域が切り出されて使用される。

【0029】

プロセッサ114は、例えばホスト101が仮想ボリューム202Aに対してライト要求を発行した場合、この仮想ボリューム202Aにおいてライト要求の対象となった仮想的な記憶領域に対し、未使用の実記憶領域をプール204から所定単位（ページ）で割り当てる。

10

【0030】

なおページとは、データを書き込む際の最小単位の記憶領域である。ここでは仮想ボリューム202A又は202Bに対して割り当てられた仮想ページを201A～201Eとして図示しており、これらのページ201A～201Eに対して割り当てるプールボリューム206A～206Eにおける実ページを207A～207Fとして図示している。

【0031】

次回ホスト101から同じページ201Aに対してリード要求又はライト要求が発行された場合、プロセッサ114は、既に割り当てられているプールボリューム206Aの記憶領域207Aに対してI/O処理を実行することにより、あたかもホスト101が仮想ボリューム202Aに対してI/O処理を実行しているように処理することができる。

20

【0032】

すなわち仮想ボリューム202A又は202Bを用いて、使用する部分のみプールボリューム206A～206Eの記憶領域（ページ）207A～207Fを割り当てることにより、限られた記憶容量を効率的に使用することができる。

【0033】

ここで、各仮想ボリューム202A又は202Bを構成するページ201毎に、ホスト101からの単位時間当たりの書き込み回数（これをライト頻度と呼ぶ）は異なる。よって例えばライト頻度が高いページ201Aは、特性として書き込み上限回数の多いMLC SSDで構成されたパリティグループ205Aに配置することで、TLC SSDのような書き込み上限回数の少ない記憶装置の寿命を延ばすことができる。

30

【0034】

上記の再配置は、具体的にはページ207Cに格納されたデータを未使用のページ207Bにコピーし、仮想ボリューム202Aのページ201Cとプールボリューム206Bのページ207Cとの対応付けを仮想ボリューム202Aのページ201Cとプールボリューム206Aのページ207Bとの対応付けに変更することにより実行する。

【0035】

（1-3）ページ配置処理の概念構成

図3は、TLC-MLC間で行われるページ再配置処理の概念構成を示す。ここではMLC SSDで構成されたパリティグループ（以下、MLC PGと呼ぶ）205Aと、TLC SSDで構成されたパリティグループ（以下、TLC PGと呼ぶ）205Bとの間で、ページ301を再配置する。ページ301の再配置先は、MLC PG205A及びTLC 205B上のページ301毎に採取しているモニタ情報に基づいて決定される。

40

【0036】

具体的には、ページ301毎にライト回数を一定期間採取し、周期の満了後、モニタ情報に基づいて算出されたMLC PG205A及びTLC PG205B毎のライト頻度と、TLC-MLC間ライト閾値302とに基づいてページ301の再配置先を決定する。

【0037】

図3では、TLC-MLC間ライト閾値302が2000以上のライト負荷のページ3

50

01をMLC PG205Aに再配置し（矢印303A）、2000未満のライト負荷のページ301をTLC PGに再配置する（矢印303B）。

【0038】

このように寿命の異なるSSDで構成されたMLC PG205AとTLC PG205Bとの間で、ライト頻度の高いページを書き込み上限回数の多いSSDで構成されたMLC PG205Aに再配置し、ライト頻度の低いページを書き込み上限回数の少ないSSDで構成されたTLC PG205Bに再配置する。

【0039】

この結果、書き込み上限回数の少ないSSDで構成されたTLC PG205Bにおけるライト回数を削減することができ、書き込み上限回数の少ないSSD（ここではTLC SSD）の寿命を延ばすことができる。

10

【0040】

図4は、同一特性のドライブ間（TLC-TLC間又はMLC-MLC間）で行われるページ再配置処理の概念構成を示す。同一特性のドライブ間で行われるページ再配置処理は、書き込み上限回数が同程度であるSSDで構成されたパリティグループ間でのライト頻度を調整することを目的として行われる。

【0041】

実際にはパリティグループ間のライト頻度を調整するライトリバランスと、ライトリバランスにより崩れたパリティグループ間のI/O頻度を調整する性能リバランスとの2種類のリバランス処理を行う。

20

【0042】

ここではTLC PG間で行うライトリバランス及び性能リバランスについて説明する。まずライトリバランスでは、寿命を保証するために削減が必要なライト頻度（これをライト削減要求量と呼ぶ）又は寿命を保証することができる範囲内で追加可能なライト頻度（これをライト追加可能量と呼ぶ）をパリティグループ毎に算出する。

【0043】

このときライト削減要求量が正の値のパリティグループをライト削減PG205Bとし、ライト追加可能量が正の値のパリティグループをライト追加PG205Cとする。そしてライト削減PG205Bのライト高負荷なページ401Aとライト追加PG205Cのライト低負荷なページ401Bとを互いに移動して（矢印403）、パリティグループ間のライト頻度を調整する。

30

【0044】

ライトリバランスの際の移動対象となるページは、モニタ情報を元にした閾値に基づいて決定される。閾値には、ライト高負荷なページを決定する閾値（これをライト削減閾値と呼ぶ）405と、ライト低負荷なページを決定する閾値（これをライト追加閾値と呼ぶ）406とがある。ライトリバランスの際、ライト削減PG205Bでは、ライト削減閾値405以上のライト頻度のページ401Aを移動対象とし、ライト追加PG205Cでは、ライト追加閾値406以下のライト頻度のページ401Bを移動対象とする。そしてそれぞれのページを移動することでライト頻度を調整する。

【0045】

移動するページ数は、1回のページ移動で移動するライト頻度を算出し、目標とするライト削減要求量を満たすために必要な移動ページ数（これを移動計画ページ数と呼ぶ）を予め算出することにより決定される。そしてこの算出された移動計画ページ数分だけページの移動（再配置）が行われる。

40

【0046】

次に性能リバランスでは、ライトリバランスにより移動したライト頻度の分だけリード頻度をライトリバランスとは逆の方向に移動することで、各パリティグループのリード頻度及びライト頻度を合算したI/O頻度をパリティグループ間で調整する。具体的には、ライト削減PG205Bのリード低負荷なページ402Bとライト追加PG205Cのリード高負荷なページ402Aとを移動して（矢印404）、パリティグループ間のリード

50

頻度を調整する。

【0047】

性能リバランスの際の移動対象となるページは、モニタ情報を元にした閾値に基づいて決定される。閾値には、リード高負荷なページを決定する閾値（これをリード削減閾値と呼ぶ）407とリード低負荷なページを決定する閾値（これをリード追加閾値と呼ぶ）408がある。

【0048】

性能リバランスの際、ライト削減PG205Bでは、リード追加閾値408以下のリード頻度のページ402Bを移動対象とし、ライト追加PG205Cでは、リード削減閾値407以上のリード頻度のページ402Aを移動対象とする。そしてそれぞれのページを移動することでリード頻度を調整する。移動するページ数は、ライトリバランスと同様、移動計画ページ数を予め算出することにより決定される。そしてこの移動計画ページ数分だけページの移動（再配置）が行われる。

【0049】

（1-4）メモリの内部構成

図5は、共有メモリ109の内部構成を示す。共有メモリ109には、ページ毎モニタテーブル501、パリティグループ毎モニタテーブル502、パリティグループ毎再配置管理テーブル503、プール毎再配置管理テーブル504、キャッシュ管理テーブル505、ダイナミックマッピングテーブル506及び論理物理アドレス変換テーブル507が格納される。

【0050】

ページ毎モニタテーブル501は、I/O回数を含む各ページのモニタ情報を管理するテーブルであり、パリティグループ毎モニタテーブル502は、I/O回数を含む各パリティグループのモニタ情報を管理するテーブルである。またパリティグループ毎再配置管理テーブル503は、パリティグループ毎のページ再配置に関する制御情報を管理するテーブルである。

【0051】

プール毎再配置管理テーブル504は、プール毎のページ再配置に関する制御情報を管理するテーブルであり、キャッシュ管理テーブル505は、キャッシュメモリ108にデータを格納する際にキャッシュメモリ108にあるデータのダーティ／クリーン状態管理を保持するテーブルである。

【0052】

またプール毎再配置管理テーブル504は、仮想ボリューム202に対して書き込むデータをキャッシュする場合には、キャッシュメモリ108のアドレスと、対応する仮想ボリューム202のページ201を特定するアドレスとを対応づけて管理するテーブルである。

【0053】

ダイナミックマッピングテーブル506は、仮想ボリューム202の各ページ201と、各ページ201に割り当てられているプールボリューム206のページ207と、各ページ201のモニタ情報との対応関係を管理するテーブルである。

【0054】

論理物理アドレス変換テーブル507は、パリティグループと、プールボリュームと、プールボリュームのデータを格納する物理ディスクに対応するパリティグループのアドレスとの対応関係を管理するテーブルである。

【0055】

図6は、ローカルメモリ113の内部構成を示す。ローカルメモリ113には、ホストI/O処理プログラム601、デステージ処理プログラム602、寿命情報採取処理プログラム603、閾値決定処理プログラム604、再配置処理プログラム604A及び新規割り当て決定処理プログラム605が格納される。これらの各種プログラムは、プロセッサ114により実行される。

【0056】

ホスト I/O 処理プログラム 601 は、ホスト 101 からの I/O 要求を受領した場合に仮想ボリューム 202 に対するリード要求又はライト要求を処理するプログラムである。デステージ処理プログラム 602 は、キャッシュメモリ 108 上の物理ディスクに未反映のデータを物理ディスクに格納するプログラムである。この処理は、ホスト 101 からの I/O 要求に対する処理とは非同期に実行される。

【0057】

寿命情報採取処理プログラム 603 は、ドライブ 111 及び 112 に対して所定周期でコマンドを発行して寿命情報を採取し、採取した情報を共有メモリ 109 に反映するプログラムである。閾値決定処理プログラム 604 は、所定周期で採取したモニタ情報とドライブ 111 及び 112 の寿命情報とに基づいて、ページ再配置のための閾値を算出するプログラムである。

【0058】

再配置処理プログラム 604A は、閾値決定処理プログラム 604 により呼び出されるプログラムであり、閾値決定処理プログラム 604 により決定された各種閾値に基づいて、ページを再配置するプログラムである。新規割り当て決定処理プログラム 605 は、ホスト I/O 処理プログラム 601 に同期して実行され、仮想ボリューム 202 における新規の仮想ページに対して実ページの割り当て先であるパリティグループ 205 を閾値に基づいて決定するプログラムである。

【0059】**(1-5) テーブル構成**

図 7 は、ページ毎モニタテーブル 501 の論理構成を示す。ページ毎モニタテーブル 501 は、ページ番号欄 701、ライト I/O カウンタ欄 702、リード I/O カウンタ欄 703、合計 I/O カウンタ欄 704 及び新規ライトフラグ欄 705 から構成される。

【0060】

ページ番号欄 701 には、ページ 201 を特定するページ番号が格納され、ライト I/O カウンタ欄 702 には、一定周期のライト回数が格納される。またリード I/O カウンタ欄 703 には、一定周期のリード回数が格納され、合計 I/O カウンタ欄 704 には、一定周期のリード回数及びライト回数の合計 I/O 回数が格納される。

【0061】

周期は、上記した閾値決定処理プログラム 604 がモニタ情報を採取する周期と同じであり、閾値決定処理プログラム 604 は、この一定期間のモニタ情報を処理対象とする。また新規ライトフラグ欄 705 には、ページが新規割り当てページか否かを示すフラグが格納される。

【0062】

図 8 は、パリティグループ毎モニタテーブル 502 の論理構成を示す。パリティグループ毎モニタテーブル 502 は、パリティグループ番号欄 801、最大ライト頻度欄 802、最小ライト頻度欄 803、最大リード頻度欄 804、最小リード頻度欄 805、リード/ライト比率欄 806、ライト追加可能量欄 807、ライト削減要求量欄 808、新規ライト I/O カウンタ欄 809、新規ライト比率欄 810、平均 I/O 頻度欄 811 及び割り当てページ数欄 812 から構成される。

【0063】

パリティグループ番号欄 801 には、パリティグループ 205 を特定するパリティグループ番号が格納され、最大ライト頻度欄 802 には、パリティグループ内のページの最大ライト頻度が格納される。また最小ライト頻度欄 803 には、パリティグループ内のページの最小ライト頻度が格納される。

【0064】

最大リード頻度欄 804 には、パリティグループ内のページの最大リード頻度が格納され、最小リード頻度欄 805 には、パリティグループ内のページの最小リード頻度が格納される。またリード/ライト比率欄 806 には、パリティグループに対するリード回数と

10

20

30

40

50

ライト回数との比率が格納される。

【0065】

ライト追加可能量欄807には、寿命を保証することができる範囲内でパリティグループに追加可能なライト頻度が格納され、ライト削減要求量欄808には、寿命を保証するためにパリティグループから削減が必要なライト頻度が格納される。ライト追加可能量欄807及びライト削減要求量欄808には、パリティグループのライト頻度及びSSDの寿命情報に基づいて算出される値が格納され、何れかが正の値となる。

【0066】

ライト追加可能量欄807及びライト削減要求量欄808に格納される値は、SSDから採取できる寿命情報がライト追加可能率及びライト削減要求率である場合、下記式1及び2を計算して算出することができる。

10

【0067】

なおライト追加可能率とは、現状のライト頻度を100%として、追加可能なライト頻度の割合であり、ライト削減要求率とは、現状のライト頻度を100%として寿命を維持するために削減すべきライト頻度の割合である。

【0068】

【数1】

ライト追加可能量 [IOPH (Input Output Per Hour)]
 =Min (パリティグループを構成する全SSDのライト追加可能率 [%])
 ×パリティグループのライト頻度 [IOPH]
 (1)

20

【0069】

【数2】

ライト削減要求量 [IOPH]
 =Max (パリティグループを構成する全SSDのライト削減要求率 [%])
 ×パリティグループのライト頻度 [IOPH]
 (2)

【0070】

新規ライトI/Oカウンタ欄809には、パリティグループに対する新規ライト回数が格納され、新規ライト比率欄810には、パリティグループにおける書き込み処理のうち、更新ライトと新規ライトとの比率が格納される。また平均I/O頻度欄811には、パリティグループにおける各ページの平均I/O頻度が格納され、割り当てページ数欄812には、パリティグループに割り当てられているページ数が格納される。

30

【0071】

図9は、パリティグループ毎再配置管理テーブル503の論理構成を示す。パリティグループ毎再配置管理テーブル503は、パリティグループ番号欄901、メディアタイプ欄902、移動元PG種別欄903、移動先PG欄904、ライト削減閾値欄905、ライト追加閾値欄906、リード削減閾値欄907、リード追加閾値欄908、移動計画ページ数欄909A及び909B、移動実績ページ数欄910A及び910B並びに新規ライト可能量欄911から構成される。

40

【0072】

パリティグループ番号欄901には、パリティグループを特定するパリティグループ番号が格納され、メディアタイプ欄902には、パリティグループを構成するSSDの特性の情報が格納される。また移動元PG種別欄903には、ライト削減PG又はライト追加PGの何れかを示す情報が格納される。移動元PG種別は、パリティグループ毎モニタテーブル502のライト追加可能量欄807又はライト削減要求量欄808に格納される情報に基づいて決定される。

【0073】

50

具体的には、ライト追加可能量欄 807 に正の値が格納されている場合、このパリティグループの移動元 P G 種別欄 903 には、ライト追加 P G を示す情報が格納される。またライト削減要求量欄 808 に正の値が格納されている場合、このパリティグループの移動元 P G 種別 903 には、ライト削減 P G を示す情報が格納される。

【0074】

移動先 P G 欄 904 には、ページ再配置実行時の移動先のパリティグループ番号が格納される。移動先のパリティグループ番号は、再配置進捗度が最低のパリティグループを移動先として決定してもよい。再配置進捗度は、移動計画ページ数欄 909A 及び 909B と、移動実績ページ数欄 910A 及び 910B とを用いて、下記式 3 を計算して算出することができる。

【0075】

【数 3】

再配置進捗度 = 移動実績ページ数 / 移動計画ページ数 × 100 (3)

【0076】

ライト削減閾値欄 905 には、ライト削減 P G でのライトリバランス時の移動対象のページを決定するための閾値が格納され、ライト追加閾値欄 906 には、ライト追加 P G でのライトリバランス時の移動対象のページを決定するための閾値が格納される。

【0077】

リード削減閾値欄 907 には、ライト追加 P G での性能リバランス時の移動対象ページを決定するための閾値が格納され、リード追加閾値欄 908 には、ライト削減 P G での性能リバランス時の移動対象ページを決定するための閾値が格納される。ライト削減閾値、ライト追加閾値、リード削減閾値及びリード追加閾値は、下記式 4 を計算して算出することができる。

【0078】

【数 4】

ライト削減閾値 [IOPH] = 最大ライト頻度 [IOPH] × 0.7

ライト追加閾値 [IOPH] = 最小ライト頻度 [IOPH] × 1.3

リード削減閾値 [IOPH] = 最大リード頻度 [IOPH] × 0.7

リード追加閾値 [IOPH] = 最小リード頻度 [IOPH] × 1.3

..... (4)

【0079】

移動計画ページ数 (ライトリバランス) 欄 909A には、同一特性の SSD により構成されるパリティグループ間でライト頻度を調整するために必要なページの移動数が格納され、移動計画ページ数 (性能リバランス) 欄 909B には、同一特性の SSD により構成されるパリティグループ間でリード頻度を含む I/O 頻度を調整するために必要なページの移動数が格納される。ライトリバランス用の移動計画ページ数は、下記式 5 を計算して算出することができる。

【0080】

【数 5】

ライトリバランス用の移動計画ページ数 [ページ]

= 目標移動量 [IOPH] ÷ 1 回のページ移動で変化する I/O 頻度 [IOPH/ページ]
目標移動量 [IOPH]

= Min (ライト追加可能量 [IOPH]、ライト削減要求量 [IOPH])

1 回のページ移動で変化する I/O 頻度 [IOPH/ページ]

= (ライト追加 P G の移動対象ページの平均 I/O 頻度 [IOPH]

− ライト削減 P G の移動対象ページの平均 I/O 頻度 [IOPH]) ÷ 2

..... (5)

10

20

30

40

50

【0081】

また性能リバランス用の移動計画ページ数は、下記式6を計算して算出することができる。

【0082】

【数6】

性能リバランス用の移動計画ページ数 [ページ]

= 目標移動量 [IOPH] ÷ 1回のページ移動で変化するI/O頻度 [IOPH/ページ]

目標移動量 [IOPH]

= ((ライトリバランスで移動したライト頻度 [IOPH] × 2) +

(ライト追加PGの平均I/O頻度 [IOPH]

− ライト削減PGの平均I/O頻度 [IOPH])) ÷ 2

1回のページ移動で変化するI/O頻度 [IOPH]

= (ライト追加PGの移動対象ページの平均I/O頻度 [IOPH]

− ライト削減PGの移動対象ページの平均I/O頻度 [IOPH]) ÷ 2

..... (6)

10

【0083】

移動実績ページ数 (ライトリバランス) 欄 910Aには、ライトリバランスのために他のパリティグループに移動したページ数が格納され、移動実績ページ数 (性能リバランス) 欄 910Bには、性能リバランスのために他のパリティグループに移動したページ数が格納される。新規ライト可能量欄 912には、パリティグループで処理することのできる新規ライト回数が格納される。新規ライト可能量は、下記式7を計算して算出することができる。

20

【0084】

【数7】

新規ライト可能量 [I/O]

= パリティグループの移動後のライト頻度 [IOPH]

× 新規割り当て比率 [%] × 再配置周期 [H]

..... (7)

30

【0085】

図10は、プール毎再配置管理テーブル504の論理構成を示す。プール毎再配置管理テーブル504は、プール番号欄1001、寿命制御再配置周期欄1002、TLC-MLC間ライト閾値欄1003、新規ライト閾値欄1004、ワークロードタイプ欄1005、同種ドライブ間新規割り当てポリシ欄1006、新規割り当てバッファ欄1007及び再配置バッファ欄1008から構成される。

【0086】

プール番号欄1001には、プールを特定するプール番号が格納され、寿命制御再配置周期欄1002には、プールにおいて寿命制御によるページの再配置を実行する周期が格納される。またTLC-MLC間ライト閾値欄1003には、TLC PGとMLC PGとの間でページを再配置する際にどちらの種類のパリティグループにページを配置するのかを決定するための閾値が格納される。TLC-MLC間ライト閾値は、下記式8を計算して算出することができる。

40

【0087】

【数 8】

$TLC-MLC$ 間ライト閾値 $[IOPH]$
 $= Ave$ (全 TLC PG の限界ライト閾値 $[IOPH]$)
 TLC PG の限界ライト閾値 $[IOPH]$
 $= IF$ (当該パリティグループの移動元 PG 種別がライト追加 PG ?)
 当該パリティグループのライト頻度 $[IOPH]$ + ライト追加可能量 $[IOPH]$
 $ELSE$ IF (当該パリティグループの移動元 PG 種別がライト削減 PG ?)
 当該 PG のライト頻度 $[IOPH]$ - ライト削減要求量 $[IOPH]$
 (8)

10

【0088】

新規ライト閾値欄 1004 には、プールにおいて新規ページに対する割り当て処理を行う際、 TLC PG と MLC PG のどちらから新規ページに対する実ページの割り当てを行うかを決定するための閾値が格納される。新規ライト閾値は、ページ移動時のペナルティ (= 1 ページあたりのデータサイズ) としてストレージシステム 104 が指定してもよい。

【0089】

ワークロードタイプ欄 1005 には、プールに関連付けられている仮想ボリュームに対してホスト 101 から発行される I/O 要求の特性の情報が格納される。ワークロードタイプには、例えば `Write intensive`、`Read intensive`、`Unknown` 等がある。

20

【0090】

`Write intensive` は、ホスト 101 からの I/O 要求のうち、ライト比率が高いことを意味し、`Read intensive` は、ホスト 101 からの I/O 要求のうち、リード比率が高いことを意味し、`Unknown` は、ホスト 101 からの I/O 要求のうち、リード/ライト比率が不明であることを意味する。ワークロードタイプは、ユーザが指定してもよいし、ストレージシステム 104 が自動で決定してもよい。

【0091】

同種ドライブ間新規割り当てポリシー欄 1006 には、新規割り当て処理で TLC PG 内又は MLC PG 内でどのパリティグループから新規ページに対する実ページの割り当てを行うかを決定するための情報が格納される。新規割り当てポリシーには、例えばパリティグループ間で順番に割り当てるラウンドロビン、容量の多いパリティグループから割り当てる容量優先又はライト追加可能量の多いパリティグループから割り当てる寿命優先がある。新規割り当てポリシーは、ユーザが指定してもよいし、ストレージシステム 104 が自動で決定してもよい。

30

【0092】

新規割り当てバッファ欄 1007 には、プールで新規割り当て処理のために使用するバッファのプール容量に対する比率が格納される。再配置バッファ欄 1008 には、プールでページの再配置処理のために使用するバッファのプール容量に対する比率が格納される。なおこのプール毎再配置管理テーブル 504 内には、再配置の制御情報のデフォルト値を格納するエントリが 1 つ存在し、ユーザからの指定がない制御情報にはデフォルト値が格納される。

40

【0093】

(1-6) フローチャート

図 11 は、ホスト I/O 処理のフローチャートを示す。このホスト I/O 処理は、ストレージシステム 104 がホスト 101 からの I/O 要求を受領したことを契機として、プロセッサ 114 とホスト I/O 処理プログラム 601 との協働により実行される。説明の便宜上、処理主体をホスト I/O 処理プログラム 601 として説明する。

【0094】

まずホスト I/O 処理プログラム 601 は、ホスト 101 からの I/O 処理要求を受領

50

すると、受領した I / O 処理要求が仮想ボリューム 202 に対してデータを書き込むライト要求であるか否かを判断する (S1101)。

【0095】

ライト要求である場合 (S1101: Y)、ホスト I / O 処理プログラム 601 は、ダイナミックマッピングテーブル 506 を参照して、ライト対象の仮想ページに対して実ページが割り当て済みであるか否かを判断する。そして未割り当ての場合には、未使用の実ページを割り当てる新規割り当て決定処理を実行する (S1109)。新規割り当て決定処理の詳細については後述する (図19)。

【0096】

新規割り当て決定処理を実行した後、次いでホスト I / O 処理プログラム 601 は、仮想ボリューム 202 上のライト対象のアドレスに対応した領域がキャッシュメモリ 108 上に確保されているか否かを確認し、確保されていない場合にはキャッシュメモリ 108 上の領域を確保する (S1110)。

【0097】

次いでホスト I / O 処理プログラム 601 は、ホスト 101 に対してライト用のデータを送信するように通知する。ホスト I / O 処理プログラム 601 は、ホスト 101 からデータが送信されると、このデータをキャッシュメモリ 108 上の確保した領域に書き込む (S1111)。そしてホスト I / O 処理プログラム 601 は、まだドライブ 111 又は 112 に書き込みが完了していない領域であることを示すダーティフラグをキャッシュ管理テーブル 505 において ON に設定する。

【0098】

ここで、ダーティフラグは、キャッシュメモリ 108 にのみデータが格納されており、ドライブ 111 又は 112 にはまだ格納されていない状態を示す情報である。ダーティフラグは、キャッシュメモリ 108 の領域を管理するキャッシュ管理テーブル 505 において ON 又は OF に設定される。キャッシュ管理テーブル 505 は、後述するデステージ処理 (図12) において参照される。

【0099】

そしてダーティフラグが ON である場合にはキャッシュメモリ 108 上の領域に格納されているデータは、ドライブ 111 又は 112 に書き込まれる。ドライブ 111 又は 112 に書き込まれた後は、ダーティフラグは OFF に設定され、リード処理に対応してドライブ 111 又は 112 から読み込んだデータをキャッシュメモリ 108 に格納した場合を含めて、キャッシュ管理テーブル 505 にはクリーンフラグが ON に設定される。

【0100】

このようにキャッシュ管理テーブル 505 は、キャッシュメモリ 108 上のアドレスと対応する仮想ボリューム 202 のアドレスと、キャッシュメモリ 108 上のデータの状態とを少なくとも管理する。なおキャッシュメモリ 108 のアドレスに対応する仮想ボリューム 202 のアドレスは、仮想ボリューム 202 のデータを置くためにキャッシュメモリ 108 上の領域を確保した場合にのみ、有効な値である仮想ボリューム 202 のアドレスが格納される。

【0101】

以上の処理を行った後、ホスト I / O 処理プログラム 601 は、ホスト 101 に I / O 処理 (ライト処理) が完了したことを通知して (S1112)、本処理を終了する。

【0102】

ステップ S1101 に戻り、ホスト I / O 処理プログラム 601 は、ホスト 101 から受領した I / O 要求が仮想ボリューム 202 からデータを読み込むリード要求である場合 (S1101: N)、キャッシュ管理テーブル 505 を参照して、リード要求に対応する仮想ボリューム 202 上のアドレスに対応したデータがキャッシュメモリ 108 上に格納されているか否かを判断する (S1102)。

【0103】

リード要求に対応する仮想ボリューム 202 上のアドレスに対応したデータがキャッシ

10

20

30

40

50

メモリ 108 上に格納されている場合をキャッシュヒットと呼ぶ。キャッシュヒットである場合 (S1102:Y)、ホスト I/O 処理プログラム 601 は、キャッシュメモリ 108 上のデータをホスト 101 に転送するとともに (S1108)、I/O 処理 (リード処理) が完了したことをホスト 101 に通知して、本処理を終了する。

【0104】

これに対し、キャッシュヒットしなかった場合 (S1102:N)、ホスト I/O 処理プログラム 601 は、リード対象の仮想ボリューム 202 のアドレスに対応したデータを格納するための領域をキャッシュメモリ 108 上に確保する (S1103)。次いでホスト I/O 処理プログラム 601 は、ダイナミックマッピングテーブル 506 を参照して、リード対象の仮想ボリューム 202 のアドレスにプール 204 から実ページが割り当てら

10

【0105】

仮想ボリューム 202 に実ページが割り当てられていない場合、ホスト I/O 処理プログラム 601 は、ダイナミックマッピングテーブル 506 を参照して、デフォルト値を格納するページを用いてデフォルト値の格納ページのドライブ 111 又は 112 のアドレスを算出し、ドライブ 111 又は 112 からデフォルト値をキャッシュメモリ 108 の領域に転送する (S1105)。

【0106】

ここで、デフォルト値の場合は、ダイナミックマッピングテーブル 506 の仮想ボリューム 202 及び論理アドレスに対応する、プールボリューム番号及び論理アドレスには、デフォルト値格納ページのあるプールボリューム番号及び論理アドレスが設定されている。デフォルト値格納ページは、プール 204 に 1 つ以上あればよい。容量効率を考えればプール 204 にデフォルト値ページは 1 又は 2 つである。

20

【0107】

デフォルト値格納ページのアドレスと対応付けられている仮想ボリューム 202 の論理アドレスは、ホスト 101 から新規にデータの書き込みがあった際に、ホスト 101 のデータ書き込み用のページで未だどの仮想ボリューム 202 のアドレスにも対応付けられていない未使用のページと対応付け直される。

【0108】

これに対し、仮想ボリューム 202 に実ページが割り当てられている場合、ホスト I/O 処理プログラム 601 は、ダイナミックマッピングテーブル 506 を参照して、プールボリューム番号及び論理アドレスを取得し、更に論理物理アドレス変換テーブル 507 を参照して、物理ドライブ番号及び物理開始アドレスを算出することで、リード対象の仮想ボリューム 202 のアドレスに対応したデータが格納されているドライブ 111 又は 112 のアドレスを算出する (S1104)。

30

【0109】

次にホスト I/O 処理プログラム 601 は、算出したアドレスからデータをキャッシュメモリ 108 上の領域に転送する (S1105)。そしてリード時にページ毎モニタテーブル 501 を参照して、リード I/O カウンタの数値をカウントアップする (S1106)。

40

【0110】

そしてホスト I/O 処理プログラム 601 は、ドライブ 111 又は 112 からキャッシュメモリ 108 上に格納したデータをホスト 101 に転送するとともに (S1107)、I/O 処理 (リード処理) が完了したことをホスト 101 に通知して、本処理を終了する。

【0111】

以上の処理により、ホスト I/O 処理を行うとともに、必要なモニタ情報を採取することができる。

【0112】

図 12 は、デスページ処理のフローチャートを示す。このデスページ処理は、ホスト I

50

／O処理（図11）とは非同期で、プロセッサ114とデステージ処理プログラム602との協働により適宜実行される。説明の便宜上、処理主体をデステージ処理プログラム602として説明する。

【0113】

まずデステージ処理プログラム602は、キャッシュ管理テーブル505を参照して、ダーティフラグのON又はOFFを確認し、ドライブ111又は112に書き込みがなされていない未反映データがキャッシュメモリ108上にあるか否かを判断する（S1201）。

【0114】

デステージ処理プログラム602は、未反映データがキャッシュメモリ108上にある場合、キャッシュ管理テーブル505から仮想ボリューム番号及び論理アドレスを取得する。そしてこの仮想ボリューム番号及び論理アドレスを元にダイナミックマッピングテーブル506を参照して、プールボリューム番号及び論理アドレスを取得する。

10

【0115】

このときプールボリューム番号及び論理アドレスがデフォルト値格納ページのアドレスであった場合、デステージ処理プログラム602は、新規データを書き込むためにダイナミックマッピングテーブル506から新規空きページを割り当てる。そしてこの割り当てページのプールボリューム番号及び論理アドレスをダイナミックマッピングテーブル506の対応する仮想ボリューム番号及び論理アドレスに対応付けて格納する。

【0116】

既にページが割り当てられている場合は、デフォルト値のプールボリューム番号及び論理アドレスとは異なるプールボリューム番号及び論理アドレスの値が仮想ボリュームの論理アドレスに対応付けて格納されている。デステージ処理プログラム602は、プールボリューム番号及び論理アドレスを取得した後、論理物理アドレス変換テーブル507を参照して、ドライブ111又は112のアドレスを算出する（S1202）。

20

【0117】

次いでデステージ処理プログラム602は、算出したドライブ111又は112のアドレスに対して、キャッシュメモリ108上の未反映データを書き込む（S1203）。そしてダイナミックマッピングテーブル506のページ番号に対応するページ毎モニタテーブル501を参照して、ライトI／Oカウンタの数値をカウントアップする（S1204）。

30

【0118】

次いでデステージ処理プログラム602は、ページ毎モニタテーブル501の新規ライトフラグ欄705を参照して、デステージ処理対象のページが新規割り当てページであるか否かを判断する（S1205）。新規割り当てページである場合（S1205：Y）、デステージ処理プログラム602は、このページの格納先のパリティグループに対応するパリティグループ毎モニタテーブル502を参照して、新規ライトI／Oカウンタの数値をカウントアップする（S1206）。

【0119】

これに対し、デステージ処理プログラム602は、ページ処理対象のページが新規割り当てページでない場合（S1205：N）、ステップS1201に移行する。ステップS1201においてデステージ処理プログラム602は、キャッシュメモリ108上に未反映データがさらにあるか否かを判断する。そしてさらなる未反映データがない場合には（S1201：N）、本処理を終了する。

40

【0120】

以上の処理により、キャッシュメモリ108上のデータを非同期にドライブ111又は112に格納するとともに、必要なモニタ情報を採取することができる。

【0121】

図13は、寿命情報採取処理のフローチャートである。この寿命情報採取処理は、プロセッサ114と寿命情報採取処理プログラム603との協働により一定周期で実行される

50

。説明の便宜上、処理主体を寿命情報採取処理プログラム603として説明する。

【0122】

まず寿命情報採取処理プログラム603は、ドライブ111又は112に対して寿命採取のコマンドを発行する(S1301)。次いで寿命情報採取処理プログラム603は、寿命情報としてライト追加可能率又はライト削減要求率を受信する(S1302)。そして受信した寿命情報をローカルメモリ113に格納して(S1303)、本処理を終了する。

【0123】

以上の処理により、ドライブ111又は112から寿命情報を採取することができる。

【0124】

図14は、閾値決定処理のフローチャートである。この閾値決定処理は、プロセッサ114と閾値決定処理プログラム604との協働により一定周期で実行される。一定周期の情報は、プール毎再配置管理テーブル504の寿命制御再配置周期欄1002に格納される。説明の便宜上、処理主体を閾値決定処理プログラム604として説明する。

【0125】

まず閾値決定処理プログラム604は、全てのパリティグループについて、ページ毎のモニタ情報の集計が完了したか否かを判断する(S1401)。すなわちパリティグループ毎モニタテーブル502の各欄に情報が格納されているか否かを判断する。

【0126】

集計が完了していない場合(S1401:N)、閾値決定処理プログラム604は、寿命情報採取処理プログラム603を呼び出してドライブ111又は112から寿命情報を採取し(S1402)、ページ毎のモニタ情報をパリティグループ毎に集計する(S1403)。集計が完了している場合(S1401:Y)、閾値決定処理プログラム604は、ページ再配置のための各種閾値を算出する(S1404)。

【0127】

ここで算出される各種閾値は、パリティグループ毎モニタテーブル502、パリティグループ毎再配置管理テーブル503及びプール毎再配置管理テーブル504の各欄に格納される。そして閾値決定処理プログラム604は、再配置処理プログラム604Aを呼び出して再配置処理を実行した後(S1405)、本処理を終了する。

【0128】

以上の処理により、パリティグループ毎にモニタ情報を集計し、集計したモニタ情報に基づいて、各種閾値を算出することができる。そして算出した閾値を用いて再配置処理を実行することができる。

【0129】

図15は、再配置処理のフローチャートである。この再配置処理は、プロセッサ114と閾値決定処理プログラム604により呼び出される再配置処理プログラム604Aとの協働により実行される。説明の便宜上、処理主体を再配置処理プログラム604Aとして説明する。

【0130】

まず再配置処理プログラム604Aは、閾値決定処理プログラム604により算出されたTLC-MLC間ライト閾値に基づいて、TLC PGとMLC PGとの間でページを再配置するTLC-MLC間再配置処理を実行する(S1501)。

【0131】

このTLC-MLC間再配置処理により、TLC SSD(ドライブ112)の寿命を延ばすことができる。またビットコストを削減することができる。

【0132】

次に再配置処理プログラム604Aは、パリティグループ毎再配置管理テーブル503を参照し、ライトリバランス用の移動計画ページ数に基づいて、同一特性のSSDで構成されたパリティグループ間(TLC PG間又はMLC PG間)でページを再配置するライトリバランス処理を実行する(S1502)。

10

20

30

40

50

【0133】

このライトリバランス処理により、同一特性のSSDで構成されたパリティグループ間でライト負荷を分散して、寿命を平準化することができる。

【0134】

次に再配置処理プログラム604Aは、パリティグループ毎再配置管理テーブル503を参照して、ライト削減PGのライトリバランス用の移動計画ページ数が0以下であるか否かを判断する(S1503)。ライト削減PGのライトリバランス用の移動計画ページ数が0以下である場合(S1503:Y)、再配置処理プログラム604Aは、ライト頻度の調整が完了していると判断して、次に性能リバランス処理を実行する(S1504)。

10

【0135】

これに対し、ライト削減PGのライトリバランス用の移動計画ページ数が0よりも大きい場合(S1503:N)、再配置処理プログラム604Aは、現在のパリティグループの構成では、ライト頻度を調整できず、寿命を維持することができないため、警告画面を表示して、寿命を保証するために追加すべきTLC SSD又はMLC SSDの容量をユーザに通知する(S1505)。

【0136】

以上の処理により、パリティグループ間でのライト頻度及びリード頻度を調整することができる。またライト頻度及びリード頻度を調整することができない場合には寿命を維持するために必要なTLC SSD又はMLC SSDの容量をユーザに通知することができる。

20

【0137】

図16は、TLC-MLC間再配置処理のフローチャートである。このTLC-MLC間再配置処理は、プロセッサ114と再配置処理プログラム604Aとの協働により実行される。説明の便宜上、処理主体を再配置処理プログラム604Aとして説明する。

【0138】

まず再配置処理プログラム604Aは、閾値決定処理プログラム604により算出されたTLC-MLC間ライト閾値に基づいて、全てのパリティグループ内の各ページをTLC PG又はMLC PGに再配置する。このとき再配置について未判定のページがあるか否かを判断する(S1601)。

30

【0139】

未判定のページがある場合(S1601:Y)、再配置処理プログラム604Aは、パリティグループ毎再配置管理テーブル503を参照して、この未判定の対象ページが所属するパリティグループのメディアタイプがTLCであるか否かを判断する(S1602)。メディアタイプがTLCである場合(S1602:Y)、再配置処理プログラム604Aは、対象ページのライト頻度がTLC-MLC間ライト閾値以上であるか否かを判断する(S1603)。

【0140】

対象ページのライト頻度がTLC-MLC間ライト閾値以上である場合(S1603:Y)、再配置処理プログラム604Aは、このページをTLC PGからMLC PGに移動する(S1604)。これに対し、対象ページのライト頻度がTLC-MLC間ライト閾値未満である場合(S1603:N)、再配置処理プログラム604Aは、何もせずにステップS1601に移行する。

40

【0141】

ステップS1602に戻り、メディアタイプがMLCである場合(S1602:N)、対象ページのライト頻度がTLC-MLC間ライト閾値未満であるか否かを判断する(S1605)。対象ページのライト頻度がTLC-MLC間ライト閾値未満である場合(S1605:Y)、再配置処理プログラム604Aは、このページをMLC PGからTLC PGに移動する(S1606)。

【0142】

50

これに対し、対象ページのライト頻度がTLC-MLC間ライト閾値以上である場合（S1605：N）、再配置処理プログラム604Aは、何もせずにステップS1601に移行する。再配置処理プログラム604Aは、全てのパリティグループ内の各ページについて判定を終えると、本処理を終了する。以上の処理により、TLC-MLC間でページを再配置して、TLC SSD（ドライブ112）の寿命を延ばすとともに、ビットコストを削減することができる。

【0143】

図17は、ライトリバランス処理のフローチャートである。このライトリバランス処理は、プロセッサ114と再配置処理プログラム604Aとの協働により実行される。説明の便宜上、処理主体を再配置処理プログラム604Aとして説明する。

10

【0144】

まず再配置処理プログラム604Aは、閾値決定処理プログラム604により算出されたライトリバランス用移動計画ページ数、ライト削減閾値及びライト追加閾値に基づいて、同一特性のSSDで構成されたパリティグループ間（TLC PG間又はMLC PG間）でページを再配置する。このとき再配置について未判定のページがあるか否かを判断する（S1701）。

【0145】

未判定のページがない場合には（S1701：N）、再配置処理プログラム604Aは、本処理を終了する。これに対し、未判定のページがある場合（S1701：Y）、再配置処理プログラム604Aは、パリティグループ毎再配置管理テーブル503を参照して、この未判定の対象ページが所属するパリティグループの移動元PG種別がライト削減PGであるか否かを判断する（S1702）。

20

【0146】

対象ページが所属するパリティグループの移動元PG種別がライト削減PGである場合（S1702：Y）、再配置処理プログラム604Aは、ページ毎モニタテーブル501のライトI/Oカウンタを参照して、この対象ページのライト頻度を取得する。そして取得したライト頻度がライト削減閾値以上であるか否かを判断する（S1703）。

【0147】

対象ページのライト頻度がライト削減閾値未満である場合（S1703：N）、再配置処理プログラム604Aは、ステップS1701に移行する。これに対し、対象ページのライト頻度がライト削減閾値以上である場合（S1703：Y）、再配置処理プログラム604Aは、この対象ページの移動先のパリティグループを決定する。

30

【0148】

移動先のパリティグループを決定する際、再配置処理プログラム604Aはパリティグループ毎再配置管理テーブル503を参照して、ライト追加PGのうち、ライトリバランス用の移動実績ページ数がライトリバランス用の移動計画ページ数未満であるパリティグループが存在するか否かを判断する（S1704）。

【0149】

ライト追加PGのうち、ライトリバランス用の移動実績ページ数がライトリバランス用の移動計画ページ数未満であるパリティグループが存在する場合（S1704：Y）、再配置処理プログラム604Aは、このパリティグループにライト負荷の高い対象ページを移動しても寿命を維持できると判断して、このパリティグループを移動先PGとしてパリティグループ毎再配置管理テーブル503に登録し、対象ページをこのパリティグループに移動する（S1705）。

40

【0150】

これに対し、ライト追加PGのうち、ライトリバランス用の移動実績ページ数がライトリバランス用の移動計画ページ数未満であるパリティグループが存在しない場合（S1704：N）、再配置処理プログラム604Aは、この対象ページについては判定を終え、ステップS1701に移行する。

【0151】

50

ステップS1702に戻り、対象ページが所属するパリティグループの移動元PG種別がライト削減PGでない場合（S1702：N）、すなわち対象ページが所属するパリティグループの移動元PG種別がライト追加PGである場合、再配置処理プログラム604Aは、ページ毎モニタテーブル501のライトI/Oカウンタ欄702を参照して、この対象ページのライト頻度を取得する。そして取得したライト頻度がライト追加閾値未満であるか否かを判断する（S1706）。

【0152】

対象ページのライト頻度がライト追加閾値以上である場合には（S1706：N）、再配置処理プログラム604Aは、ステップS1701に移行する。これに対し、対象ページのライト頻度がライト追加閾値未満である場合（S1706：Y）、再配置処理プログラム604Aは、この対象ページの移動先のパリティグループを決定する。

10

【0153】

移動先のパリティグループを決定する際の処理は、上記のステップS1704及びS1705と同様であるため説明を省略する。以上の処理により、同一特性のSSDで構成された異なるパリティグループ間でページを再配置してライト負荷を分散し、寿命を平準化することができる。

【0154】

図18は、性能リバランス処理のフローチャートである。この性能リバランス処理は、プロセッサ114と再配置処理プログラム604Aとの協働により実行される。説明の便宜上、処理主体を再配置処理プログラム604Aとして説明する。

20

【0155】

なお性能リバランス処理は、ライトリバランス処理（図17）で調整したライト頻度と同程度のリード頻度のページをライトリバランス処理で移動させた方向とは逆方向に移動する点で、ライトリバランス処理と異なり、他の基本的な処理内容は同様である。

【0156】

まず再配置処理プログラム604Aは、閾値決定処理プログラム604により算出された性能リバランス用の移動計画ページ数、ライト削減閾値及びライト追加閾値に基づいて、同一特性のSSDで構成されたパリティグループ間（TLC PG間又はMLC PG間）でページを再配置する。このとき再配置について未判定のページがあるか否かを判断する（S1801）。

30

【0157】

未判定のページがない場合には（S1801：N）、再配置処理プログラム604Aは、本処理を終了する。これに対し、未判定のページがある場合（S1801：Y）、再配置処理プログラム604Aは、パリティグループ毎再配置管理テーブル503を参照して、この未判定の対象ページが所属するパリティグループの移動元PG種別がライト削減PGであるか否かを判断する（S1802）。

【0158】

対象ページが所属するパリティグループの移動元PG種別がライト削減PGである場合（S1802：Y）、再配置処理プログラム604Aは、ページ毎モニタテーブル501のリードI/Oカウンタを参照して、この対象ページのリード頻度を取得する。そして取得したリード頻度がリード追加閾値未満であるか否かを判断する（S1803）。

40

【0159】

対象ページのリード頻度がリード追加閾値以上である場合には（S1803：N）、再配置処理プログラム604Aは、ステップS1801に移行する。これに対し、対象ページのリード頻度がリード追加閾値未満である場合（S1803：Y）、再配置処理プログラム604Aは、この対象ページの移動先のパリティグループを決定する。

【0160】

移動先のパリティグループを決定する際、再配置処理プログラム604Aはパリティグループ毎再配置管理テーブル503を参照して、ライト追加PGのうち、性能リバランス用の移動実績ページ数が性能リバランス用の移動計画ページ数未満であるパリティグルー

50

プが存在するか否かを判断する（S1804）。

【0161】

ライト追加PGのうち、性能リバランス用の移動実績ページ数が性能リバランス用の移動計画ページ数未満であるパリティグループが存在する場合（S1804：Y）、再配置処理プログラム604Aは、このパリティグループにリード負荷の低い対象ページを移動させても高負荷にならないと判断して、このパリティグループを移動先PGとしてパリティグループ毎再配置管理テーブル503に登録し、対象ページをこのパリティグループに移動する（S1805）。

【0162】

これに対し、ライト追加PGのうち、性能リバランス用の移動実績ページ数が性能リバランス用の移動計画ページ数未満であるパリティグループが存在しない場合（S1804：N）、再配置処理プログラム604Aは、この対象ページについては判定を終え、ステップS1801に移行する。

10

【0163】

ステップS1802に戻り、対象ページが所属するパリティグループの移動元PG種別がライト削減PGでない場合（S1802：N）、すなわち対象ページが所属するパリティグループの移動元PG種別がライト追加PGである場合、再配置処理プログラム604Aは、ページ毎モニタテーブル501のリードI/Oカウンタ欄703を参照して、この対象ページのリード頻度を取得する。そして取得したリード頻度がリード削減閾値以上であるか否かを判断する（S1806）。

20

【0164】

対象ページのリード頻度がリード削減閾値未満である場合には（S1806：N）、再配置処理プログラム604Aは、ステップS1801に移行する。これに対し、対象ページのリード頻度がリード削減閾値以上である場合（S1806：Y）、再配置処理プログラム604Aは、この対象ページの移動先のパリティグループを決定する。

【0165】

移動先のパリティグループを決定する際の処理は、上記のステップS1804及びS1805と同様であるため説明を省略する。以上の処理により、同一特性のSSDで構成された異なるパリティグループ間でページを再配置してリード負荷を分散し、I/O頻度を平準化することができる。

30

【0166】

図19は、新規割り当て決定処理のフローチャートを示す。この新規割り当て決定処理は、ホスト101から新規の仮想ページに対するライト要求を受領したことを契機として、プロセッサ114と新規割り当て決定処理プログラム605との協働により実行される。説明の便宜上、処理主体を新規割り当て決定処理プログラム605として説明する。

【0167】

まず新規割り当て決定処理プログラム605は、プール毎再配置管理テーブル504を参照して、新規ページの割り当て対象の仮想ボリューム202に記憶領域を提供するプール204のワークロードタイプが「Unknown」であるか否かを判断する（S1901）。

40

【0168】

プール204のワークロードタイプが「Unknown」ではない場合（S1901：N）、新規割り当て決定処理プログラム605は、更に「Write intensive」であるか否かを判断する（S1906）。ワークロードタイプが「Write intensive」である場合（S1906：Y）、新規割り当て決定処理プログラム605は、新規ページに対するライト頻度は多いと予測して、書き込み上限回数が比較的多いMLC PGを新規ページに対する実ページの割り当て先のパリティグループに設定する（S1907）。

【0169】

これに対し、ワークロードタイプが「Read intensive」である場合（S

50

1906:N)、新規割り当て決定処理プログラム605は、新規ページに対するライト頻度は少ないものと予測して、書き込み上限回数が少ないTLC PGを新規ページに対する実ページの割り当て先のパリティグループに設定する(S1904)。

【0170】

ステップS1901に戻り、ワークロードタイプが「Unknown」である場合(S1901:Y)、新規割り当て決定処理プログラム605は、ホスト101からのI/O特性が分からないため、新規ページに対する将来のライト頻度を予測して、割り当て先を決定する。まずは新規ページの予測ライト頻度を算出する(S1902)。

【0171】

例えば新規割り当て決定処理プログラム605は、ホスト101からの1ページ当たりの平均ライト頻度をモニタ情報として採取して見積もることにより、予測ライト頻度を算出する。

【0172】

次いで新規割り当て決定処理プログラム605は、プール毎再配置管理テーブル504を参照して、予測ライト頻度が新規ライト閾値未満であるか否かを判断する(S1903)。予測ライト頻度が新規ライト閾値未満の場合(S1903:Y)、新規割り当て決定処理プログラム605は、書き込み上限回数の少ないTLC PGを割り当て先のパリティグループに設定する(S1904)。

【0173】

これに対し、予測ライト頻度が新規ライト閾値以上の場合(S1903:N)、新規割り当て決定処理プログラム605は、書き込み上限回数の多いMLC PGを割り当て先のパリティグループに設定する(S1907)。

【0174】

次いで新規割り当て決定処理プログラム605は、割り当て先に設定した特性のパリティグループについて、新規ページの割り当てが可能か否かを判断するため、パリティグループ毎モニタテーブル502及びパリティグループ毎再配置管理テーブル503を参照して、新規ライト可能量が新規ライトI/Oカウンタよりも大きいかなんかを割り当て先に設定した特性のパリティグループごとに判断する(S1905)。

【0175】

新規ライト可能量が新規ライトI/Oカウンタよりも大きいパリティグループが存在する場合(S1905:Y)、新規割り当て決定処理プログラム605は、このパリティグループから新規ページを割り当てるために、ページ毎モニタテーブル501を参照して、このパリティグループにおける何れかのページについて新規ライトフラグを設定し(S1909)、本処理を終了する。

【0176】

これに対し、新規ライト可能量が新規ライトI/Oカウンタよりも大きいパリティグループが存在しない場合(S1905:N)、新規割り当て決定処理プログラム605は、寿命を維持するにあたり新規に許容可能なライト頻度を超えているため、ユーザに推奨容量を通知する(S1908)。

【0177】

そして新規割り当て決定処理プログラム605は、判断対象のパリティグループから新規ページを割り当てるために、何れかのパリティグループにおける何れかのページについて新規ライトフラグを設定し(S1909)、本処理を終了する。

【0178】

以上の処理により、ホスト101から新規ページにデータを書き込むライト要求を受領した場合、新規ページに対するライト頻度を考慮して、寿命特性に応じたSSDにより構成されたパリティグループから新規ページに対して実ページを割り当てることができる。

【0179】

(1-7) 推奨容量算出方法の概念構成

図20は、ホスト101からのライト頻度に対して寿命を保証するにあたり推奨される

10

20

30

40

50

書き込み上限回数の異なる記憶装置の容量比率の算出方法の考え方を示す。図20は、プール204内の各ページ207（又は仮想ボリューム202内の各ページ201）のライト頻度の分布を表す。グラフ2007は、左からライト頻度が多い順番に全ページ207を並べたときの各ページ207のライト頻度を示す。縦軸はライト頻度であり、横軸はページ数である。

【0180】

TLC-MLC間ライト閾値2005は、TLC PGとMLC PGのどちらにページを配置するかを決める閾値であり、TLC-MLC間ライト閾値2005とグラフ2007との交点がTLCとMLCの推奨される容量比率（2006）となる。TLC-MLC間ライト閾値2005は、ユーザが指定してもよいし、ストレージシステム104が算出してもよい。

10

【0181】

ストレージシステム104が算出する場合、推奨容量比率は、顧客要件のライト頻度（=Whost）、プール204（又は仮想ボリューム202）容量（=C）及び各SSDの寿命を保証するにあたり許容できるライト頻度の相関グラフ2001、2002を利用し、下記式9を計算して算出することができる。

【0182】

ここでグラフ2001は、寿命を保証するにあたり単位容量あたりに許容できるライト頻度（=Wtlc）を傾きとしたTLC SSDの容量と許容できるライト頻度との相関を表す。またグラフ2002は、寿命を保証するにあたり単位容量あたりに許容できるライト頻度（=Wmlc）を傾きとしたMLC SSDの容量と許容できるライト頻度との相関を表す。

20

【0183】

【数9】

$$\begin{aligned} & \text{TLC SSD推奨容量} \\ & = ((Wmlc \times C) - Whost) \div (Wmlc - Wtlc) \div C \\ & \text{MLC SSD推奨容量} \\ & = (C - \text{TLC推奨容量}) \div C \\ & \dots\dots\dots (9) \end{aligned}$$

30

【0184】

以上により、顧客要件ライト頻度とプール容量とを満たすTLC SSDとMLC SSDとの容量比率を算出することができる。

【0185】

（1-8）画面構成

図21は、プール毎にパラメータを設定する際の画面構成の一例を示す。プール単位のGUI画面2101は、設定対象のプール204を特定できるプール番号を表示する領域2102と、寿命制御再配置のON/OFFを設定する領域2103と、寿命制御再配置をONにした場合の詳細設定のON/OFFを設定する領域2104と、詳細設定の内容を設定する領域2105とから構成される。本画面で設定した情報は、プール毎再配置管理テーブル504の各欄に格納される。

40

【0186】

寿命制御再配置領域2103の設定がOFFの場合、閾値決定処理プログラム604が閾値決定処理を実行することはないが、よって再配置処理プログラム604Aがページの再配置を実行することはないが、寿命制御の精度の低下を防ぐため、モニタ情報は寿命制御再配置のON/OFFにかかわらず採取される。

【0187】

寿命制御再配置領域2103の設定がONの場合、上記説明してきた通りページの再配置が行われる。この場合、詳細設定領域2104の項目設定領域2105が入力可能に表示される。詳細設定領域2104の設定がOFFの場合、項目設定領域2105の各種パ

50

ラメータには、デフォルト値又はストレージシステム 104 内で自動的に算出された値が設定される。

【0188】

再配置周期領域 2106 には、寿命制御のための再配置を実行する周期が設定される。この周期はユーザが指定することができる。例えばユーザが「7 days」と指定すると、7 日周期でページの再配置が実行される。

【0189】

ワークロードタイプ領域 2107 には、ホスト 101 からの I/O 特性が設定される。この I/O 特性はユーザが指定することができる。ホスト 101 からの I/O 特性を予め把握している場合は、ユーザが I/O 特性を指定することにより、新規割り当て先の SSD の特性を明示的に指定することができる。

10

【0190】

具体的には「Write intensive」が指定された場合、ホスト 101 の I/O 特性は、ライト高負荷であるため、書き込み上限回数の比較的多い MLC PG から新規ページに対して実ページの割り当てが行われる。

【0191】

また「Read intensive」が指定された場合、ホスト 101 の I/O 特性は、ライト低負荷であるため、書き込み上限回数の少ない TLC PG から新規ページに対して実ページの割り当てが行われる。

【0192】

20

またホスト 101 の I/O 特性をユーザが把握していない場合、ユーザは「Unknown」を指定する。この場合、ストレージシステム 104 が新規ページに対して実ページの割り当て先のパリティグループを自動的に決定することになる。

【0193】

同種ドライブ間新規割り当てポリシー領域 2108 には、新規ページに対する実ページの割り当て時に TLC PG 又は MLC PG のうち、何れの特性のパリティグループからページを割り当ててかを決定するポリシーが設定される。

【0194】

例えばユーザにより「ラウンドロビン」が指定された場合、各パリティグループから均等にページが割り当てられる。また「容量優先」が指定された場合、容量が少なパリティグループから優先してページが割り当てられる。また「寿命優先」が指定された場合、寿命が長いパリティグループから優先してページが割り当てられる。

30

【0195】

バッファサイズ領域 2109 及び 2110 には、特性が MLC PG であるパリティグループの容量に対するバッファの割合が設定される。新規割り当て用領域 2109 には、特性が MLC PG であるパリティグループから新規ページを割り当ての際に使用するバッファが設定される。新規割り当て時に MLC PG から割り当てられるはずのライト高負荷なページが MLC PG の残容量が足りないことによって TLC PG から割り当てられることを防ぐ効果がある。

【0196】

40

新規割り当てバッファは、再配置周期毎に確保しなおされる。このため、新規割り当てバッファは、周期内に予想されるホスト 101 から新規ページに対して書き込まれるデータ量に基づいて最適なサイズを見積もることができる。

【0197】

再配置バッファ用領域 2110 には、ページの再配置時に使用されるバッファが設定される。再配置バッファにより、再配置時に単位時間当たりに移動できるデータサイズを調整する。このため、再配置バッファを多くとることにより、再配置時のスループットを増やす効果がある。

【0198】

図 22 は、ストレージシステム 104 がユーザに通知する警告画面の画面構成の一例を

50

示す。警告画面は、ホスト101からのライト頻度が多く、現在の構成では目標期間よりもSSDの寿命が短くなること又は寿命を保証するにあたり過剰にSSDを搭載していることをユーザに通知することができる。

【0199】

画面を表示する契機は、周期的に実行されるページ再配置処理の完了後にストレージシステム104が自動で表示するとしてもよいし、ユーザがストレージシステム104に対する操作により、任意のタイミングで表示させてもよい。後者の場合、任意の操作が行われたタイミングでのモニタ情報に基づいて推奨容量が算出される。

【0200】

プール単位のGUI画面2201は、設定対象のプール204を特定できるプール番号を表示する領域2202と、警告の内容を通知する領域2203から構成される。警告の内容を通知する領域2203は、現在のドライブ構成から追加が必要又は削減が可能なTLC又はMLCの容量を通知する領域2204及び現在のホスト101からのI/O要求の情報から推奨されるMLCとTLCの容量を通知する領域2205から構成される。なお通知する容量の情報は、TLCとMLCの容量の比率で表してもよい。

【0201】

(1-9) 第1の実施の形態による効果

以上のように第1の実施の形態におけるストレージシステム104によれば、TLC-MLC間でページを再配置することにより、寿命の異なるSSDから構成されたパリティグループ間でライト頻度を調整することができる。またライトリバランス処理を行うことにより、同一特性のSSDで構成されたパリティグループ間でライト頻度を調整することができる。よって寿命劣化の激しいSSDに対するライト頻度のページを寿命劣化の緩やかなSSDに移動して、SSDの保守交換回数を削減することができる。またストレージシステム104のコストを削減することができる。

【0202】

(2) 第2の実施の形態

第2の実施の形態は、記憶装置として半導体メモリ(SSD)だけでなく、ハードディスクドライブ(HDD)を搭載するストレージシステムにおいて、記憶装置(SSD及びHDD)を性能に応じた階層に分類し、ホストからのアクセス頻度に応じて適切な階層の記憶装置にデータを配置する階層制御を行う点で、第1の実施の形態と異なる。

【0203】

(2-1) 計算機システムの全体構成

図23は、第2の実施の形態における計算機システム1Aの全体構成を示す。計算機システム1Aは、SAS(Serial Attached SCSI)規格のHDD(ドライブ2301)を搭載している点及びSSD(ドライブ111及び112)をティア1に設定し、SAS規格のHDD(ドライブ2301)をティア2に設定するとともに、各ページのI/O頻度に応じてティア間でデータを再配置する点で、第1の実施の形態と異なる。

【0204】

例えばティア1からはI/O頻度が100[IOPS]のページを割り当てるように設定されており、ティア2からはI/O頻度が10[IOPS]のページを割り当てるように設定されているとする。一方でI/O頻度が50[IOPS]のページがティア2から割り当てられており、I/O頻度が20[IOPS]のページがティア1から割り当てられているとする。

【0205】

この場合、ストレージシステム104Aは全体として10(ティア2の上限IOPH)+20=30[IOPS]の性能しか発揮することができない。そこでI/O頻度が50[IOPH]のページをティア2からティア1に移動(再配置)すると、ストレージシステム104Aは全体として50+20=70[IOPS]の性能を発揮することができるようになる。

【0206】

(2-2) ストレージシステムの論理構成

図24は、ストレージシステム104Aの論理構成を示す。ストレージシステム104Aは、プールボリューム206A～206CがMLC SSD又はTLC SSDから構成され、プールボリューム206D及び206EがSAS HDDから構成されており、各プールボリューム206A～206Eがティア1又は2に分類されている点で、第1の実施の形態と異なる。

【0207】

そして第2の実施の形態のストレージシステム104Aは、第1の実施の形態において説明してきたページの再配置に加えて、SSDから構成されるティア1においてホスト101からのライト負荷が高くなり、SSDの寿命を保証することができなくなった場合、

10

ライト負荷の高い例えばページ207Dをティア2に移動するライトデモーション処理を実行する。これによりSSDの寿命劣化を防ぐことができる。

【0208】

(2-3) ページ配置処理の概念構成

図25は、ライトデモーション処理によるページ再配置処理の概念構成を示す。ここではMLC PG205A又はTLC PG205Bから構成されるティア1と、SAS HDDで構成されたパリティグループ（以下、SAS PGと呼ぶ）205Cから構成されるティア2との間で、ページ301を再配置する。ライトデモーション処理は、閾値決定処理時にティア1内でのページ再配置だけではSSDの寿命を維持することができないと判断された場合に実行される。

20

【0209】

具体的には、閾値決定処理においてライトリバランスによる移動計画ページ数を算出した後、ライト削減PGのライト削減要求量を満たすだけのページを移動することができなかった場合、移動することができなかった分のライト頻度に基づいて、ライトデモーション処理による移動計画ページ数を算出する。そしてライトリバランス実行時に、ライトデモーション処理による移動計画ページ数分だけMLC PG205A又はTLC PG205BからSAS PG205Cにライトデモーション閾値2501以上のライト頻度のページを移動する。

【0210】

(2-4) テーブル構成

30

図26は、第2の実施の形態におけるパリティグループ毎再配置管理テーブル503Aの論理構成を示す。パリティグループ毎再配置管理テーブル503Aは、パリティグループ毎再配置管理テーブル503（図9）の各欄に加えて、ライトデモーション閾値欄2601、ライトデモーション用移動計画ページ数欄2602、ライトデモーション用移動実績ページ数欄2603及びティアレベル欄2604から構成される。

【0211】

ライトデモーション閾値欄2601には、ライトデモーション処理対象のページを決定するための閾値が格納され、ライトデモーション用移動計画ページ数欄2602には、ライトデモーション処理により移動するページ数が格納される。またライトデモーション用移動実績ページ数欄2603には、ライトデモーション処理により移動したページ数が格納される。ティアレベル欄2604には、パリティグループが所属する階層順序（例えばティア1、ティア2、ティア3）が格納される。なおここでは階層順序の値が小さいティアほど高性能のドライブで構成されている。

40

【0212】

図27は、第2の実施の形態におけるプール毎再配置管理テーブル504Aの論理構成を示す。プール毎再配置管理テーブル504Aは、プール毎再配置管理テーブル504（図10）の各欄に加えて、ティア間I/O閾値欄2701から構成される。ティア間I/O閾値欄2701には、ページをどの階層に配置するのかを決定するための閾値が格納される。

【0213】

50

(2-5) フローチャート

図28は、第2の実施の形態における再配置処理のフローチャートを示す。この再配置処理は、プロセッサ114と閾値決定処理プログラム604により呼び出される再配置処理プログラム604Aとの協働により実行される。説明の便宜上、処理主体を再配置処理プログラム604Aとして説明する。

【0214】

なお前提として閾値決定処理プログラム604は、閾値決定処理(図14)においてティア間I/O閾値及びライトデモーション用移動計画ページ数を算出し、各種閾値をパリティグループ毎再配置管理テーブル503A及びプール毎再配置管理テーブル504Aにそれぞれ格納しているものとする。

10

【0215】

まず再配置処理プログラム604Aは、ティア間再配置処理を実行する(S2801)。ここでは再配置処理プログラム604Aは、プール毎再配置管理テーブル504Aのティア間I/O閾値欄2701を参照して、対象ページを配置するティアを決定する。

【0216】

その後再配置処理プログラム604Aは、第1の実施の形態における再配置処理(図15)と同様にTLC-MLC間再配置処理を実行し(S1501)、次いでライトリバランス処理を実行する(S1502)。ライトリバランス処理を実行した後、再配置処理プログラム604Aは、パリティグループ毎再配置管理テーブル503Aを参照して、ライトデモーション用移動計画ページ数が0であるか否かを判断する(S2802)。

20

【0217】

ライトデモーション用移動計画ページ数が0でない場合(S2802:N)、再配置処理プログラム604Aは、ライトリバランス処理を実行しただけではライト頻度を調整しきれなかったため、ライトデモーション処理を実行する(S2803)。その後再配置処理プログラム604Aは、性能リバランスを実行して(S1504)、本処理を終了する。

【0218】

これに対し、ライトデモーション用移動計画ページ数が0である場合(S2802:Y)、再配置処理プログラム604Aは、ライトデモーション処理を実行する必要はないため、性能リバランスを実行して(S1504)、本処理を終了する。以上の処理により、ホスト101からのライト頻度が過多であるためティア1を構成するSSDの寿命を維持することができない場合、ティア2にライト負荷が高いページを移動して、ティア1内のSSDの寿命を維持することができる。

30

【0219】

図29は、ティア間再配置処理のフローチャートを示す。このティア間再配置処理は、プロセッサ114と再配置処理プログラム604Aとの協働により実行される。説明の便宜上、処理主体を再配置処理プログラム604Aとして説明する。

【0220】

まず再配置処理プログラム604Aは、閾値決定処理プログラム604により算出されたティア間I/O閾値に基づいて、全てのパリティグループ内の各ページをティア1又は2に再配置する。このとき再配置について未判定のページがあるか否かを判断する(S2901)。

40

【0221】

未判定のページがある場合(S2901:Y)、再配置処理プログラム604Aは、パリティグループ毎再配置管理テーブル503Aを参照して、この未判定の対象ページが所属するティアがティア2であるか否かを判断する(S2902)。

【0222】

対象ページが所属するティアがティア2である場合(S2902:Y)、再配置処理プログラム604Aは、プール毎再配置管理テーブル504Aを参照して、対象ページのI/O頻度がティア間I/O閾値以上であるか否かを判断する(S2903)。

50

【0223】

対象ページのI/O頻度がティア間I/O閾値以上である場合（S2903：Y）、再配置処理プログラム604Aは、対象ページをティア2からティア1に移動する（S2904）。これに対し、対象ページのI/O頻度がティア間I/O頻度未満である場合（S2903：N）、再配置処理プログラム604Aは、何もせずにステップS2901に移行する。

【0224】

ステップS2902に戻り、対象ページが所属するティアがティア1の場合（S2902：N）、再配置処理プログラム604Aは、プール毎再配置管理テーブル504Aを参照して、対象ページのI/O頻度がティア間I/O閾値未満であるか否かを判断する（S2905）。 10

【0225】

対象ページのI/O頻度がティア間I/O閾値未満である場合（S2905：Y）、再配置処理プログラム604Aは、対象ページをティア1からティア2に移動する（S2906）。これに対し、対象ページのI/O頻度がティア間I/O頻度以上である場合（S2905：N）、再配置処理プログラム604Aは、何もせずにステップS2901に移行する。

【0226】

再配置処理プログラム604Aは、全てのパリティグループ内の各ページについて判定を終えると、本処理を終了する。以上の処理により、ホスト101からのI/O頻度及び各ティアの性能に応じて、各ティアにページを再配置することができる。 20

【0227】

図30は、ライトデモーション処理のフローチャートを示す。このライトデモーション処理は、プロセッサ114と再配置処理プログラム604Aとの協働により実行される。説明の便宜上、処理主体を再配置処理プログラム604Aとして説明する。

【0228】

まず再配置処理プログラム604Aは、閾値決定処理プログラム604により算出されたライトデモーション用移動計画ページ数、ライト削減閾値及びライトデモーション閾値に基づいて、異なるティア間でページを再配置する。このとき再配置について未判定のページがあるか否かを判断する（S3001）。 30

【0229】

未判定のページがない場合（S3001：N）、再配置処理プログラム604Aは本処理を終了する。これに対し未判定のページがある場合（S3001：Y）、再配置処理プログラム604Aは、パリティグループ毎再配置管理テーブル503Aを参照して、この未判定の対象ページが所属するパリティグループの移動元PG種別がライト削減PGであるか否かを判断する（S3002）。

【0230】

対象ページが所属するパリティグループの移動元PG種別がライト削減PGでない場合（S3002：N）、すなわちライト追加PGである場合、再配置処理プログラム604Aは対象ページに対してライトデモーション処理を実行する必要はないと判断して、ステップS3001に移行する。 40

【0231】

これに対し、対象ページが所属するパリティグループの移動元PG種別がライト削減PGである場合（S3002：Y）、再配置処理プログラム604Aは、ページ毎モニターテーブル501のライトI/Oカウンタ欄702を参照して、この対象ページのライト頻度を取得する。

【0232】

そして再配置処理プログラム604Aは、パリティグループ毎再配置管理テーブル503Aのライト削減閾値欄905及びライトデモーション閾値欄2601を参照して、ライト削減閾値及びライトデモーション閾値を取得する。 50

【0233】

そして取得した対象ページのライト頻度がライトデモーション閾値以上であり、かつ、ライト削減閾値未満であるか否かを判断する（S3003）。対象ページのライト頻度がライトデモーション閾値以上であり、かつ、ライト削減閾値未満でない場合（S3003：N）、再配置処理プログラム604AはステップS3001に移行する。

【0234】

これに対し、対象ページのライト頻度がライトデモーション閾値以上であり、かつ、ライト削減閾値未満である場合（S3003：Y）、再配置処理プログラム604Aは、対象ページが所属するパリティグループのライトデモーション用移動実績ページ数がライトデモーション用移動計画ページ数未満であるか否かを判断する（S3004）。 10

【0235】

対象ページが所属するパリティグループのライトデモーション用移動実績ページ数がライトデモーション用移動計画ページ数未満でない場合（S3004：N）、再配置処理プログラム604Aは、ステップS3001に移行する。

【0236】

これに対し、対象ページが所属するパリティグループのライトデモーション用移動実績ページ数がライトデモーション用移動計画ページ数未満である場合（S3004：Y）、再配置処理プログラム604Aは、対象ページをティア1からティア2に移動する（S3005）。再配置処理プログラム604Aは、未判定の全てのページについて判定を終えると、本処理を終了する。 20

【0237】

図31は、新規割り当て決定処理のフローチャートを示す。この新規割り当て決定処理は、ホスト101から新規ページに対するライト要求を受領したことを契機として、プロセッサ114と新規割り当て決定処理プログラム605との協働により実行される。説明の便宜上、処理主体を新規割り当て決定処理プログラム605として説明する。

【0238】

ステップS1901～S1905までは、第1の実施の形態における新規割り当て決定処理（図19）と同様であるため、ここでの説明は省略する。ステップS1905において、新規割り当て決定処理プログラム605は、割り当て先に設定した特性のティア1のパリティグループについて、新規ライト可能量が新規ライトI/Oカウンタよりも大きい 30
か否かを判断する（S1905）。

【0239】

新規ライト可能量が新規ライトI/Oカウンタよりも大きいパリティグループが存在しない場合（S1905：N）、再配置処理プログラム604Aは、ティア1からこれ以上ページを割り当てると、SSDの寿命を維持することができなくなると判断して、新規ページに対する実ページの割り当て先をティア2に設定する（S3101）。

【0240】

ステップS1909は、第1の実施の形態における新規割り当て決定処理（図19）と同様であるため説明は省略する。以上の処理により、寿命維持のために必要な分のライト頻度をティア1から削減することができる。なお新規割り当て時にホスト101からのライト要求がシーケンシャルライトである場合、新規割り当て決定処理プログラムを実行せずに、ティア2からページを割り当てるようにしてもよい。 40

【0241】

（2-6）第2の実施の形態による効果

以上のように第2の実施の形態におけるストレージシステム104Aによれば、ホスト101からティア1に対するライト頻度をティア2に移動することにより、一定期間当たりに許容できるライト頻度を越えたSSDの寿命劣化を緩やかにすることができる。よってSSDの保守交換回数を削減でき、ストレージシステム104Aのコストを削減することができる。

【0242】

(3) 第3の実施の形態

第3の実施の形態では、第1の実施の形態において説明したライト追加可能量及びライト削減要求量(図8)を算出する手法について説明する。第1の実施の形態においては、SSDからライト追加可能率及びライト削減要求率を採取し、これらを用いて上記式1及び2を計算することによりライト追加可能量及びライト削減要求量を算出するとしたが、ライト追加可能率及びライト削減要求率をSSDから直接採取することができない場合、以下の手法を用いてライト追加可能量及びライト削減要求量を算出することができる。

【0243】

図32は、SSDの寿命情報として磨耗指標(Wear out Indicator)を取得し、磨耗指標からライト追加可能量及びライト削減要求量を算出するための考え方を示す。グラフの縦軸は、SSDの磨耗指標(寿命率)3201であり、横軸は、SSDを使用しはじめてからの経過年数(経過時間率)3202である。

10

【0244】

寿命率3201は、SSDの消去回数から算出される磨耗指標を意味しており、一般的にS. M. A. R. T. 情報として取得できることが知られている。磨耗指標は、値が100(=L_e)に達したとき、SSDの寿命を意味し、保守交換が必要となる。経過時間率は、目標寿命期間(例えば、3年や5年)を100%(=T_e)とした経過時間の割合を意味する。

【0245】

直線3215は、区間Δt1(=T₁-T₀)の寿命率の変化ΔL1(=L₁-L₀)を示しており、直線の傾きは、区間Δt1のライト頻度(=W₀)を表している。直線3215のライト頻度が継続すると、目標寿命期間を達成する前にSSDの寿命が尽きる。

20

【0246】

このため、直線3216で示すライト頻度(=W₁)までライト頻度を削減し、区間Δt2(=T_e-T₁)の間の寿命率の増加量がΔL(=L_e-L₁)となるように寿命劣化の速度を調整する。また一般的にSSDに対するライト頻度は、WA(Write Amplification)により、ホスト101がSSDに対して発行したライト要求の回数よりもSSD内部のフラッシュチップに対するライト回数の方が大きくなることが知られている。SSDのWA値は、(フラッシュチップに対するライト回数÷ホスト101からのライト要求の回数)により算出される。

30

【0247】

以上の数値を用いてストレージシステム104は、寿命を維持するために必要なライト削減要求量を算出する。ライト削減要求量は、下記式10を計算して算出することができる。

【0248】

【数10】

$$\begin{aligned} & \text{ライト削減要求量 [IOPH]} \\ = & W_0 \times WA \times (1 - ((\Delta t_1 \div \Delta L_1) \times (\Delta L_2 \div \Delta t_2))) \\ & \dots \dots (10) \end{aligned}$$

40

【0249】

またライト頻度が目標寿命期間を達成するにあたり余裕がある場合、ストレージシステム104は、ライト追加可能量を算出する。ライト追加可能量は、下記式11を計算して算出することができる。

【0250】

【数11】

$$\begin{aligned} & \text{ライト追加可能量 [IOPH]} \\ = & W_0 \times WA \times ((\Delta t_1 \div \Delta L_1) \times (\Delta L_2 \div \Delta t_2)) - 1 \\ & \dots \dots (11) \end{aligned}$$

50

【0251】

以上により、一般的に取得可能なSSDの寿命情報からSSD PGに対するライト追加可能量及びライト削減要求量（図8）を算出することができる。そして算出したライト追加可能量及びライト削減要求量に基づいてページを再配置することにより、ストレージシステム104のコストを削減することができる。

【0252】

なお上記説明してきた実施の形態においては、「書き込み」上限回数の少ないSSDのライト回数を削減するようにページを再配置することで、書き込み上限回数の少ないSSDの寿命を延ばす構成について説明してきたが、「書き換え」上限回数の少ないSSDについても上記構成を採用することにより、同様に寿命を延ばすことができる。SSDにおける書き換えとは、複数ページから構成されるブロックを一旦消去し、その後ブロック内の各ページにデータを書き込む一連の処理をいう。よって書き換えを1回行くと、消去と書き込みとを両方行うことになる。すなわち書き換え上限回数の少ないSSDと、書き込み上限回数の少ないSSDとを同様に扱うことで、書き換え上限回数の少ないSSDの寿命を延ばすことができる。

10

【0253】

また上記説明してきた実施の形態においては、書き込み上限回数の少ないSSDの「ライト頻度」（ライト回数）を削減するようにページを再配置することで、書き込み上限回数の少ないSSDの寿命を延ばす構成について説明してきたが、ライトする「データ量」を削減するようにページを再配置することでも、同様に書き込み上限回数の少ないSSDの寿命を延ばすことができる。ライトするデータ量が大きい場合には複数ページにデータをライトする必要があるため、よってライト回数も増加するためである。

20

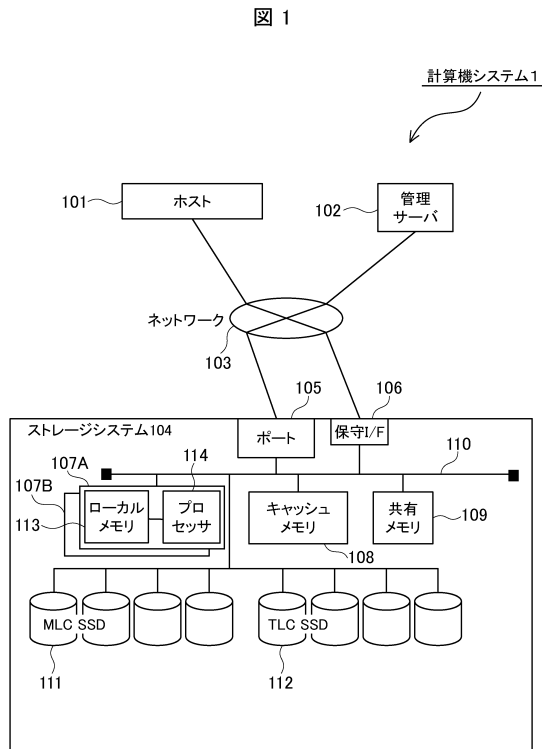
【符号の説明】

【0254】

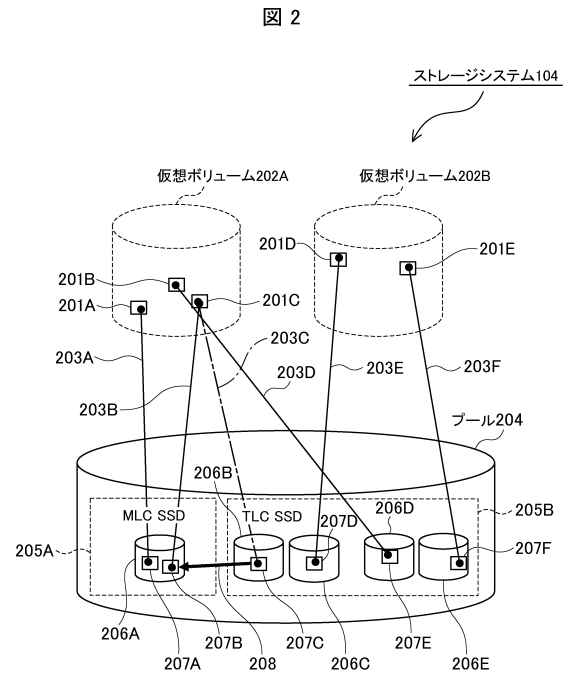
1、1A	計算機システム
104、104A	ストレージシステム
114	プロセッサ
111	MLC SSD
112	TLC SSD
2301	SAS HDD
202	仮想ボリューム
201、301、401、207	ページ
206	プールボリューム
205	パリティグループ
302	TLC-MLC間ライト閾値

30

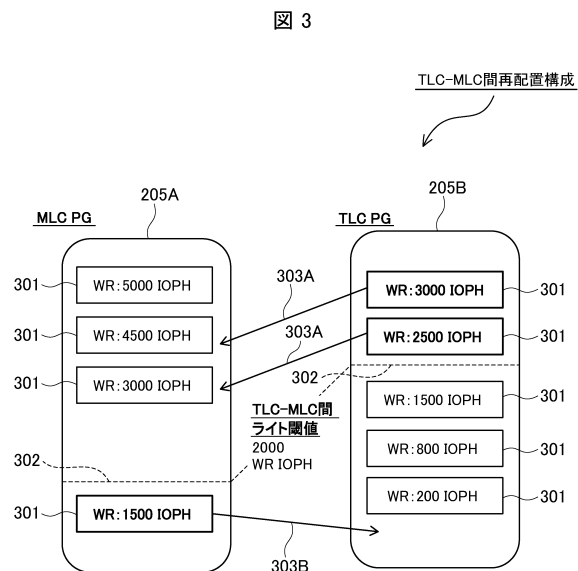
【図 1】



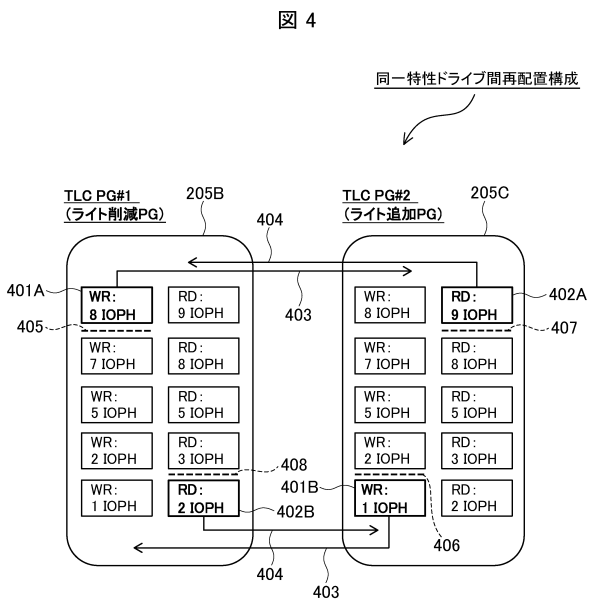
【図 2】



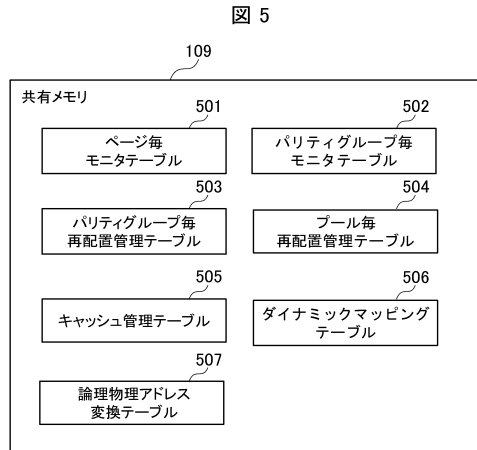
【図 3】



【図 4】



【図 5】



【図 7】

図 7

ページ毎モニターテーブル501

701	ページ番号	0	1	2	3
702	ライト I/O カウンタ	2000	12100	156	56
703	リード I/O カウンタ	5010	6520	4600	300
704	合計 I/O カウンタ	7010	18620	4756	356
705	新規ライトフラグ	0	1	1	0

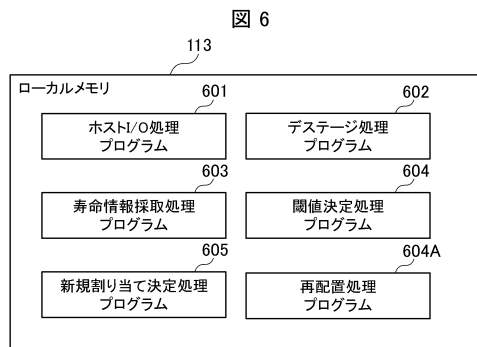
【図 8】

図 8

パリティグループ毎モニターテーブル502

801	パリティグループ番号	0	1	2
802	最大ライト頻度	22187	1962	4703
803	最小ライト頻度	5120	453	1085
804	最大リード頻度	4118	8018	17862
805	最小リード頻度	950	1850	4122
806	リード/ライト比率	16/84	80/20	79/21
807	ライト追加可能量	0	1790	2170
808	ライト削減要求量	2447	0	0
809	新規ライト I/O カウンタ	896	202	69
810	新規ライト比率	16%	40%	6%
811	平均 I/O 頻度	8800	12000	9000
812	割り当てページ数	458	128	386

【図 6】



【図 9】

図 9

パリティグループ毎再配置管理テーブル503

901	パリティグループ番号	0	1	2
902	メディアタイプ	TLC	TLC	MLC
903	移動元PG種別	ライト削減PG	ライト追加PG	ライト追加PG
904	移動先PG	1	0	3
905	ライト削減閾値	15531	Invalid	Invalid
906	ライト追加閾値	Invalid	589	1411
907	リード削減閾値	Invalid	5613	12503
908	リード追加閾値	1236	Invalid	Invalid
909A	移動計画ページ数 (ライトリバランス)	156	70	1486
910A	移動実績ページ数 (ライトリバランス)	154	56	562
909B	移動計画ページ数 (性能リバランス)	84	134	896
910B	移動実績ページ数 (性能リバランス)	0	0	232
911	新規ライト可能量	846	1026	2394

【図 10】

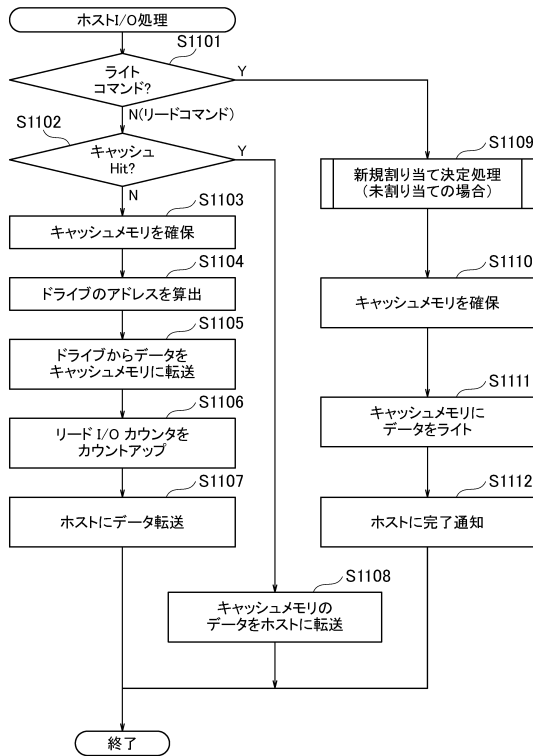
図 10

プール毎再配置管理テーブル504

1001	プール番号	0	1	2
1002	寿命制御再配置周期 [hour]	168	24	72
1003	TLC-MLC間ライト閾値	7040	9541	4530
1004	新規ライト閾値	7550	8205	4260
1005	ワークロードタイプ	Write intensive	Unknown	Read intensive
1006	同種ドライブ間新規割り当てポリシー	寿命優先	容量優先	ラウンドロビン
1007	新規割り当てバッファ	3	3	5
1008	再配置バッファ	2	2	3

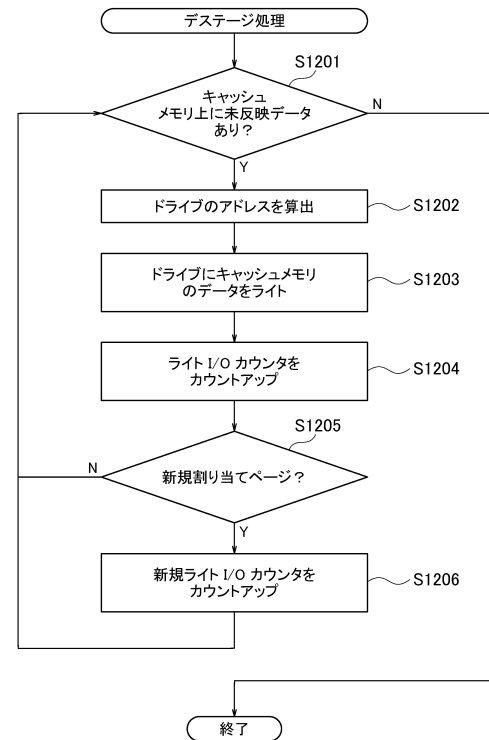
【図 1 1】

図 11



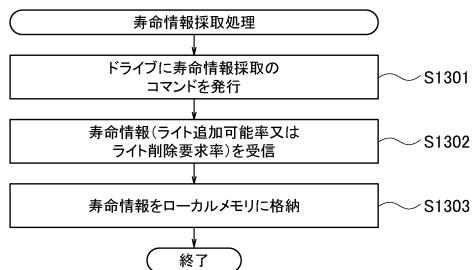
【図 1 2】

図 12



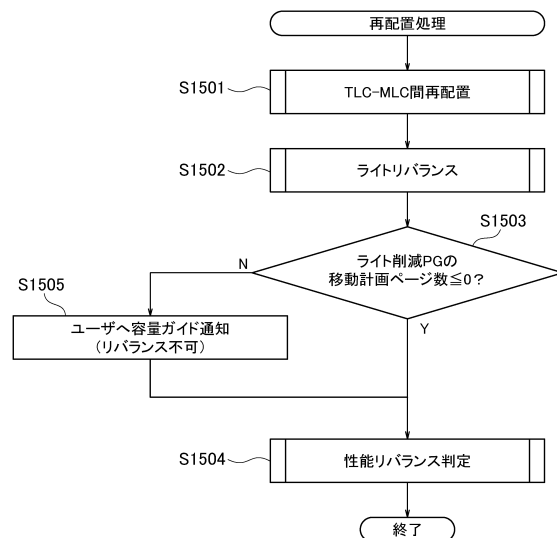
【図 1 3】

図 13



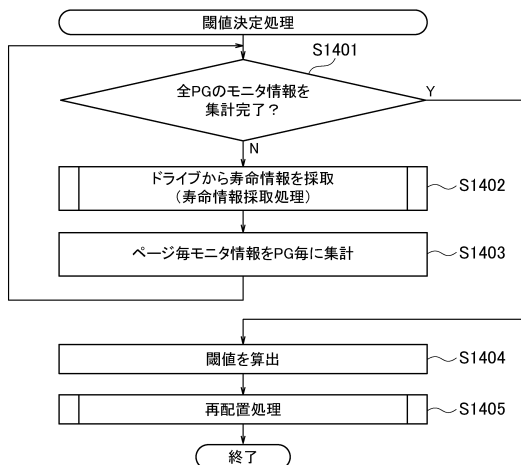
【図 1 5】

図 15

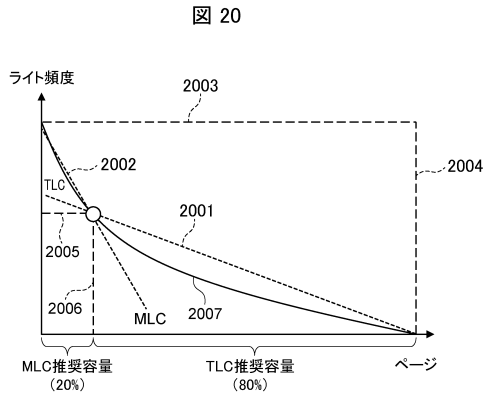


【図 1 4】

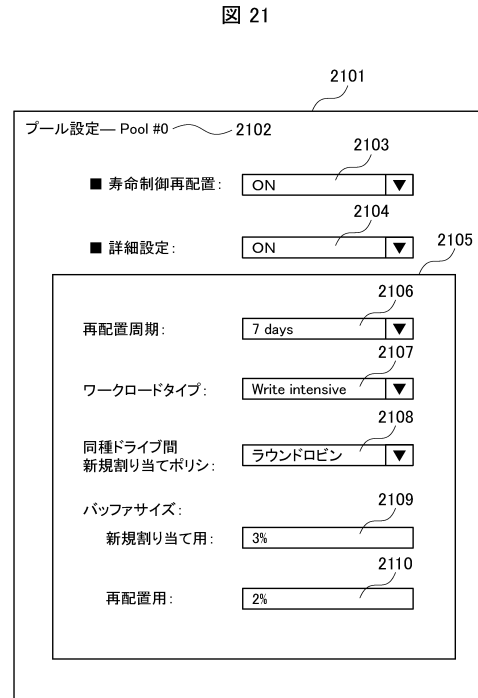
図 14



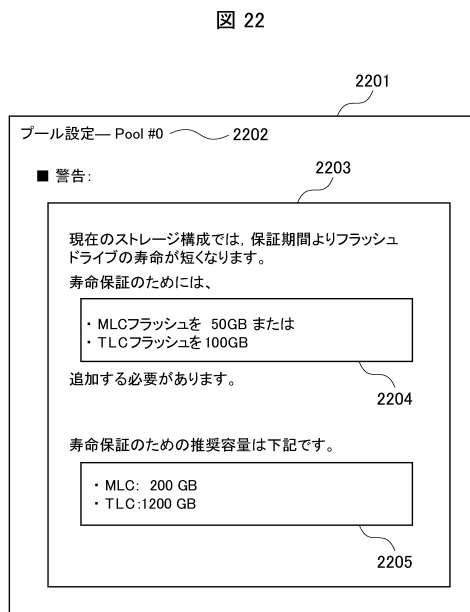
【図 20】



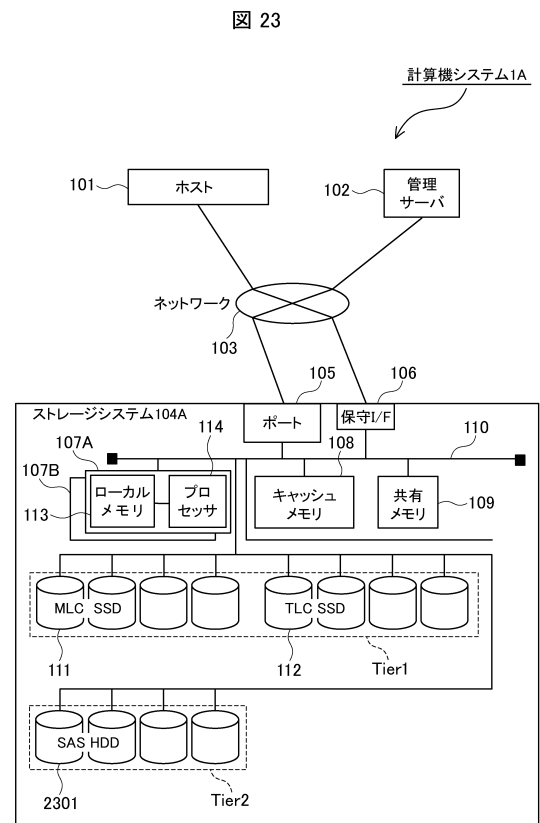
【図 2 1】



【図 2 2】

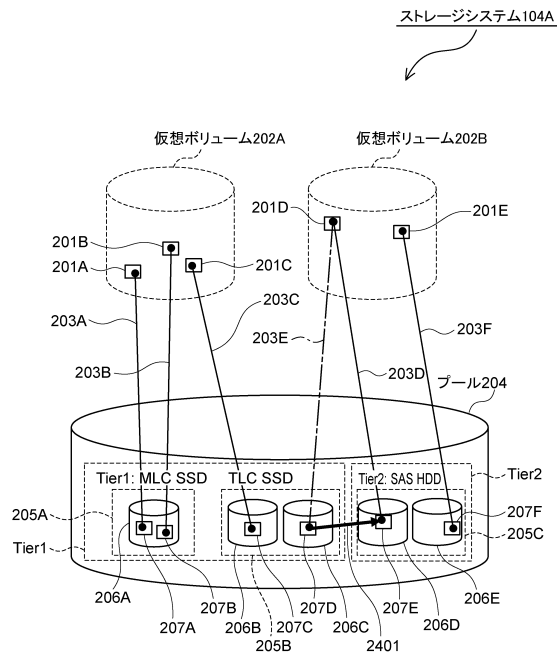


【图 2 3】



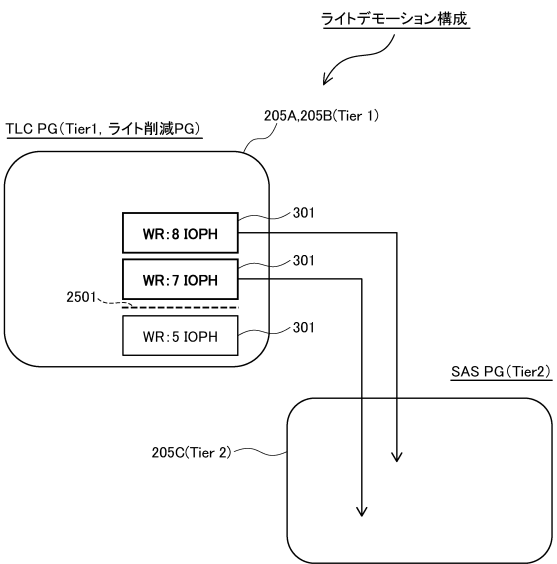
【図 2 4】

図 24



【図 2 5】

図 25



【図 2 6】

図 26

パリティグループ毎再配置管理テーブル503A

901	パリティグループ番号	0	1	2
902	メディアタイプ	TLC	SAS	MLC
903	PGタイプ	ライト削減PG	ライト追加PG	ライト追加PG
904	移動先PG	1	-	3
905	ライト削減閾値	15531	-	-
906	ライト追加閾値	-	-	-
907	リード削減閾値	-	-	-
908	リード追加閾値	1236	-	-
909	移動計画ページ数 (ライトリバランス)			
910	移動実績ページ数 (ライトリバランス)			
911	移動計画ページ数 (性能リバランス)			
912	移動実績ページ数 (性能リバランス)			
913	新規ライト可能量			
2601	ライトデモーション閾値			
2602	移動計画ページ数 (ライトデモーション)			
2603	移動実績ページ数 (ライトデモーション)			
2604	Tier Level	1	2	1

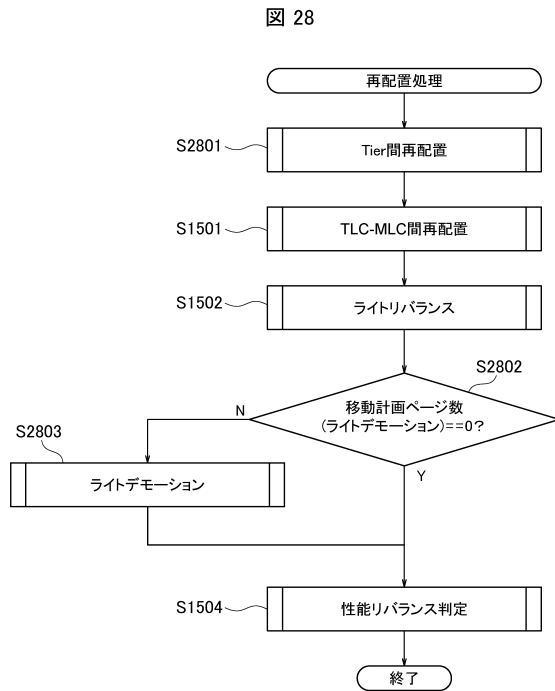
【図 2 7】

図 27

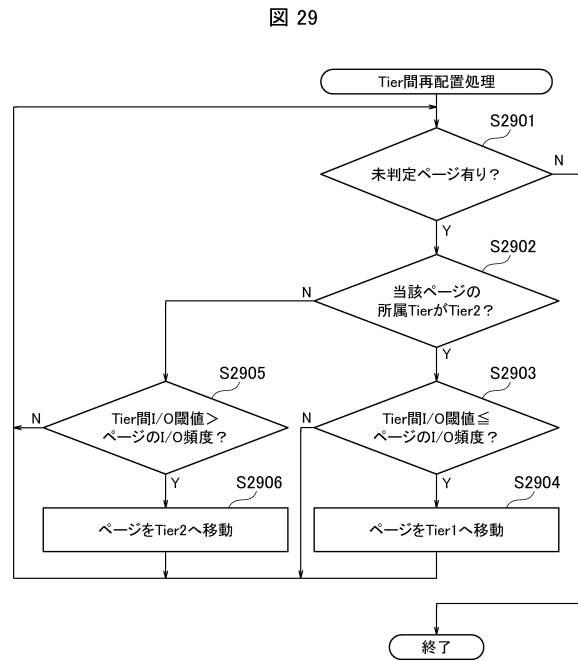
プール毎再配置管理テーブル504A

1001	プール番号	0	1	2
1002	寿命制御再配置周期 [hour]	168	24	72
1003	TLC-MLC間ライト閾値	7040	9541	4530
1004	新規ライト閾値	7550	8205	4260
1005	ワークロードタイプ	Write intensive	Unknown	Read intensive
1006	同種ドライブ間新規割り当てポリシー	寿命優先	容量優先	ラウンドロビン
1007	新規割り当てバッファ	3	3	5
1008	再配置バッファ	2	2	3
2701	Tier間I/O閾値	5680	7520	2602

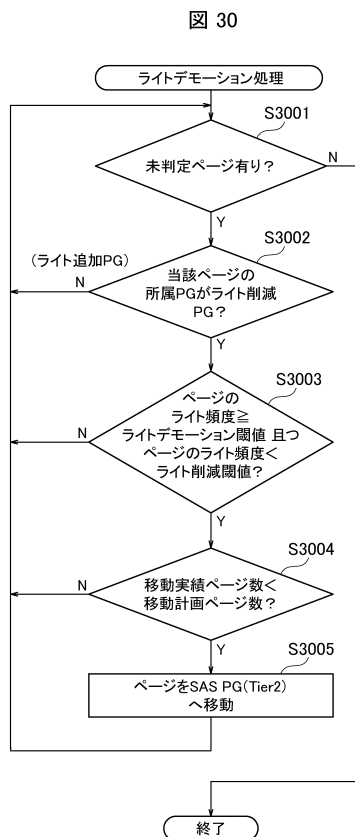
【図 28】



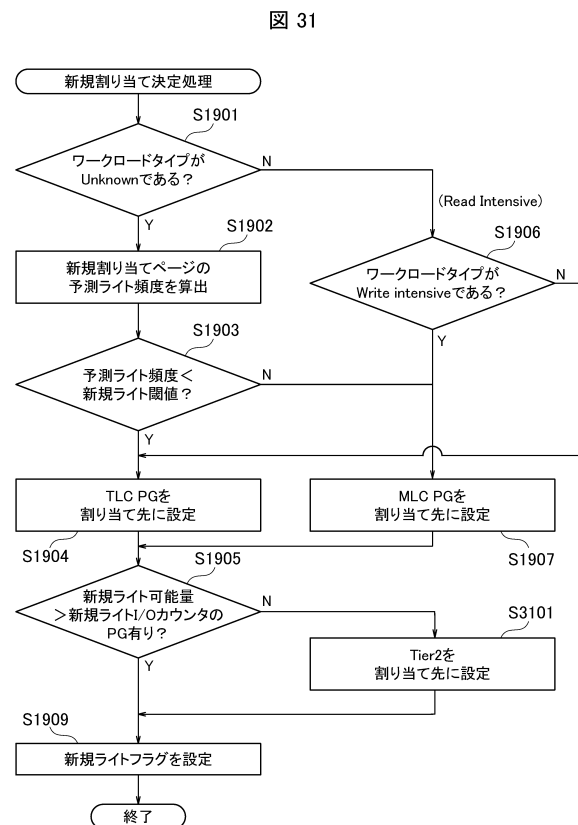
【図 29】



【図 30】

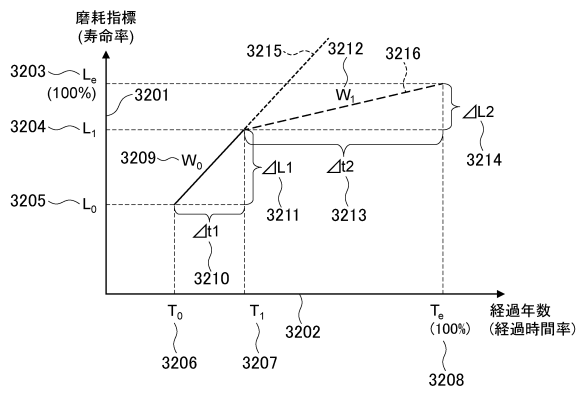


【図 31】



【図 3 2】

図 32



フロントページの続き

審査官 田上 隆一

(56)参考文献 特表2015-505078 (JP, A)
特表2013-536478 (JP, A)

(58)調査した分野(Int.Cl., DB名)
G06F 3/06