

(19) 日本国特許庁(JP)

(12) 公開特許公報(A)

(11) 特許出願公開番号

特開2016-192037
(P2016-192037A)

(43) 公開日 平成28年11月10日 (2016. 11. 10)

(51) Int.Cl.	F I	テーマコード (参考)
G06F 3/06 (2006.01)	G06F 3/06 301N	
	G06F 3/06 540	
	G06F 3/06 304B	
	G06F 3/06 301R	

審査請求 未請求 請求項の数 8 O L (全 20 頁)

(21) 出願番号 特願2015-71244 (P2015-71244)
(22) 出願日 平成27年3月31日 (2015. 3. 31)

(71) 出願人 000005223
富士通株式会社
神奈川県川崎市中原区上小田中4丁目1番1号
(74) 代理人 100092152
弁理士 服部 毅巖
(72) 発明者 渡辺 岳志
神奈川県川崎市中原区上小田中4丁目1番1号 富士通株式会社内
(72) 発明者 池内 和彦
神奈川県川崎市中原区上小田中4丁目1番1号 富士通株式会社内
(72) 発明者 前田 親志
神奈川県川崎市中原区上小田中4丁目1番1号 富士通株式会社内

最終頁に続く

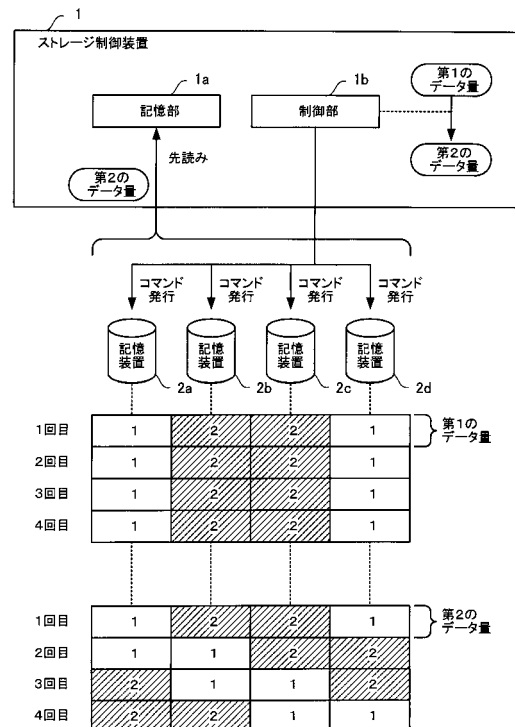
(54) 【発明の名称】 ストレージ制御装置、ストレージ制御プログラム、およびストレージ制御方法

(57) 【要約】

【課題】シーケンシャルリードのスループットを向上できる。

【解決手段】制御部 1 b は、記憶装置 2 a , 2 b , 2 c , 2 d から記憶部 1 a に第 1 のデータ量でデータを先読みすると記憶装置 2 a , 2 b , 2 c , 2 d へのコマンド発行量に偏りが生じるか否かを判定する。制御部 1 b は、コマンド発行量に偏りが生じると判定した場合に、記憶装置 2 a , 2 b , 2 c , 2 d ごとのコマンド発行量の偏りを低減する第 2 のデータ量でデータを先読みする。これにより、ストレージ制御装置 1 は、プリフェッチステージングごとのリードコマンドの発行回数の不均衡を循環させて、リードコマンドの発行回数の不均衡が特定の記憶装置に固定化されることを抑止する。

【選択図】 図 1



【特許請求の範囲】**【請求項 1】**

データ領域と前記データ領域に対応するパリティ領域とがそれぞれ分散して配置される複数の記憶装置を制御するストレージ制御装置であって、

記憶部と、

前記複数の記憶装置から前記記憶部に第 1 のデータ量でデータを先読みすると前記記憶装置ごとのコマンド発行量に偏りが生じる場合に、前記記憶装置ごとのコマンド発行量の偏りを低減する第 2 のデータ量でデータを先読みする制御部と、

を備えるストレージ制御装置。

【請求項 2】

前記制御部は、前記記憶装置ごとのコマンド発行量に偏りが生じる場合を、前記複数の記憶装置で構成される RAID セットの RAID レベルにもとづいて検出する、

請求項 1 記載のストレージ制御装置。

【請求項 3】

前記制御部は、前記記憶装置ごとのコマンド発行量の偏りを、前記複数の記憶装置で構成される RAID セットのうち、前記データ領域に対応する前記記憶装置の数または前記パリティ領域に対応する前記記憶装置の数にもとづいて検出する、

請求項 1 記載のストレージ制御装置。

【請求項 4】

前記第 2 のデータ量は、データを先読みするごとに、コマンド発行量に偏りが生じる前記記憶装置が順次交代するデータ量である、

請求項 1 記載のストレージ制御装置。

【請求項 5】

前記制御部は、前記複数の記憶装置で構成される RAID セットのうち前記データ領域に対応する前記記憶装置の数とキャッシュページサイズとの積に、前記複数の記憶装置の数に所定のシフト量を加えた数を乗じて前記第 2 のデータ量を算出する、

請求項 4 記載のストレージ制御装置。

【請求項 6】

前記制御部は、1 つの前記記憶装置から前記データ領域と前記パリティ領域を含む連続した領域からデータを先読みする場合に、前記パリティ領域で分割される前記データ領域ごとにデータを先読みするコマンドを発行する、

請求項 1 乃至請求項 5 のいずれか 1 つに記載のストレージ制御装置。

【請求項 7】

コンピュータが実行するストレージ制御方法であって、

データ領域と前記データ領域に対応するパリティ領域とがそれぞれ分散して配置される複数の記憶装置からキャッシュメモリに第 1 のデータ量でデータを先読みすると前記記憶装置ごとのコマンド発行量に偏りが生じる場合に、前記記憶装置ごとのコマンド発行量の偏りを低減する第 2 のデータ量でデータを先読みする、

処理を実行するストレージ制御方法。

【請求項 8】

コンピュータに、

データ領域と前記データ領域に対応するパリティ領域とがそれぞれ分散して配置される複数の記憶装置からキャッシュメモリに第 1 のデータ量でデータを先読みすると前記記憶装置ごとのコマンド発行量に偏りが生じる場合に、前記記憶装置ごとのコマンド発行量の偏りを低減する第 2 のデータ量でデータを先読みする、

処理を実行させるストレージ制御プログラム。

【発明の詳細な説明】**【技術分野】****【0001】**

本発明は、ストレージ制御装置、ストレージ制御プログラム、およびストレージ制御方

10

20

30

40

50

法に関する。

【背景技術】

【0002】

ホストからの読出しリクエストにかかる時間を短縮する技術の1つとしてプリフェッチステージング (Prefetch Staging) が知られている。プリフェッチステージングは、ボリュームの単位で読出しリクエストのシーケンシャル性を検出した場合に、読出しリクエストを待たずにあらかじめキャッシュメモリにデータをステージングする。プリフェッチステージングに関しては、キャッシュヒット率の向上を図る制御技術の提案がある。

【0003】

また、ディスク故障からデータを保護する技術の1つとしてRAID (Redundant Arrays of Independent Disks) が知られている。RAIDは、データを複数のディスクに分散させて、データの保護とデータアクセスのパフォーマンスを向上する。また、RAID 5やRAID 6は、分散パリティを使用するストライプセットを有するRAIDレベルとして知られている。

【先行技術文献】

【特許文献】

【0004】

【特許文献1】特開2003-345519号公報

【特許文献2】特開2007-249457号公報

【特許文献3】特開2008-310741号公報

20

【発明の概要】

【発明が解決しようとする課題】

【0005】

しかしながら、固定長のプリフェッチステージングにおいて、ストレージ制御装置は、ディスクごとにリードコマンドを発行する。また、RAID 5やRAID 6においては、ディスクごとのデータがパリティによって分割される。そのため、ストレージ制御装置は、パリティにより分割されたデータ単位でリードコマンドを発行する場合がある。

【0006】

このようなプリフェッチステージングは、たとえば、RAIDグループを構成するあるディスクに対して「N」回のリードコマンドを発行し、別のディスクに対して「N+1」回のリードコマンドを発行する。リードコマンドの発行回数の不均衡は、プリフェッチステージングごとに蓄積され、特定のディスクのレスポンス低下を惹起する。

【0007】

特定のディスクのレスポンス低下は、プリフェッチステージングの性能低下につながり、シーケンシャルリードのスループットを低下させる。

1つの側面では、本発明は、シーケンシャルリードのスループットを向上できるストレージ制御装置、ストレージ制御プログラム、およびストレージ制御方法を提供することを目的とする。

【課題を解決するための手段】

【0008】

上記目的を達成するために、以下に示すような、ストレージ制御装置が提供される。ストレージ制御装置は、データ領域とデータ領域に対応するパリティ領域とがそれぞれ分散して配置される複数の記憶装置を制御する。ストレージ制御装置は、記憶部と、制御部とを備える。制御部は、複数の記憶装置から記憶部に第1のデータ量でデータを先読みすると記憶装置ごとのコマンド発行量に偏りが生じる場合に、記憶装置ごとのコマンド発行量の偏りを低減する第2のデータ量でデータを先読みする。

【発明の効果】

【0009】

1態様によれば、ストレージ制御装置、ストレージ制御プログラム、およびストレージ制御方法において、シーケンシャルリードのスループットを向上できる。

50

【図面の簡単な説明】

【0010】

【図1】第1の実施形態のストレージ制御装置の構成の一例を示す図である。

【図2】第2の実施形態のストレージシステムの構成の一例を示す図である。

【図3】第2の実施形態のストレージ装置の構成の一例を示す図である。

【図4】第2の実施形態のRAIDグループと、RAIDグループから割り当てられるボリュームの一例を示す図である。

【図5】第2の実施形態のCMの処理機能の構成の一例を示すブロック図である。

【図6】第2の実施形態のプリフェッチステージング処理のフローチャートを示す図である。

10

【図7】第2の実施形態のシーケンシャル性検出用テーブルの一例を示す図である。

【図8】第2の実施形態のデータ量決定処理のフローチャートを示す図である。

【図9】規定値をプリフェッチステージングデータ量としてRAID5におけるディスクセットからデータを読み出す参考例を示す図である。

【図10】規定値をプリフェッチステージングデータ量としてRAID6におけるディスクセットからデータを読み出す参考例を示す図である。

【図11】リードコマンド発行回数の差が蓄積するプリフェッチステージングの参考例を示す図である。

【図12】第2の実施形態のリードコマンド発行回数の一例を示す図である。

【図13】第2の実施形態のリードコマンド発行回数の均一化の一例を示す図である。

20

【発明を実施するための形態】

【0011】

以下、図面を参照して実施の形態を詳細に説明する。

〔第1の実施形態〕

まず、第1の実施形態のストレージ制御装置について図1を用いて説明する。図1は、第1の実施形態のストレージ制御装置の構成の一例を示す図である。

【0012】

ストレージ制御装置1は、複数の記憶装置2a, 2b, 2c, 2dを制御する。記憶装置2a, 2b, 2c, 2dは、たとえばSSD (Solid State Drive) やHDD (Hard Disk Drive) などであり、複数のSSDやHDDを収容したDE (Drive Enclosure) など

30

であってよい。

【0013】

記憶装置2a, 2b, 2c, 2dのそれぞれは、データ領域と、データ領域に対応するパリティ領域とがそれぞれ分散して配置される。たとえば、記憶装置2a, 2b, 2c, 2dは、RAID5やRAID6を構成するディスクである。

【0014】

ストレージ制御装置1は、記憶部1aと、制御部1bとを備える。記憶部1aは、記憶装置2a, 2b, 2c, 2dから先読み(プリフェッチステージング)したデータを格納する。記憶部1aは、たとえば、キャッシュメモリである。

【0015】

40

制御部1bは、記憶装置2a, 2b, 2c, 2dから記憶部1aに第1のデータ量でデータを先読みすると記憶装置2a, 2b, 2c, 2dへのコマンド発行量(数)に偏りが生じるか否かを判定する。第1のデータ量は、所定のプリフェッチステージング範囲に相当するデータ量である。なお、制御部1bは、必ずしも第1のデータ量を具体的に想定することを要しない。制御部1bは、所定のデータ量(すなわち、第1のデータ量)でデータを先読みする場合に、記憶装置2a, 2b, 2c, 2dのコマンド発行量に偏りが生じる可能性をRAIDレベルから判定することができる。たとえば、制御部1bは、RAIDレベルからコマンド発行量に偏りが生じるか否かを判定ことができ、制御部1bは、RAIDレベルがRAID5やRAID6のときにコマンド発行量に偏りが生じると判定できる。

50

【 0 0 1 6 】

制御部 1 b は、コマンド発行量に偏りが生じると判定した場合に、記憶装置 2 a , 2 b , 2 c , 2 d ごとのコマンド発行量の偏りを低減する第 2 のデータ量でデータを先読みする。

【 0 0 1 7 】

ここで、データの先読みごとのリードコマンドの発行回数の不均衡について説明する。たとえば、第 1 のデータ量でデータを先読みしたとき、記憶装置 2 a , 2 d に対する 1 回目のコマンド (リードコマンド) の発行量が「 1 」回であり、記憶装置 2 b , 2 c に対する 1 回目のコマンドの発行量が「 2 」回である。第 1 のデータ量でデータを先読みしたとき、2 回目以降のコマンド発行量は、1 回目のコマンド発行量と同じである。

10

【 0 0 1 8 】

このように、第 1 のデータ量によるデータの先読みは、記憶装置 2 a , 2 b , 2 c , 2 d に対するコマンドの発行量が固定化するとき、データの先読みごとに記憶装置ごとのコマンド発行量の不均衡が拡大する。たとえば、4 回の先読みによるコマンドの総発行量は、記憶装置 2 a , 2 d が「 4 」回であり、記憶装置 2 b , 2 c が「 8 」回である。したがって、記憶装置 2 b , 2 c は、レスポンス低下によりプリフェッチステージングの性能を低下させる原因となる場合がある。

【 0 0 1 9 】

次に、データの先読みごとのリードコマンドの発行回数の不均衡解消について説明する。たとえば、第 2 のデータ量でデータを先読みしたとき、記憶装置 2 a , 2 d に対する 1 回目のコマンドの発行量が「 1 」回であり、記憶装置 2 b , 2 c に対する 1 回目のコマンドの発行量が「 2 」回である。続いて、第 2 のデータ量でデータを先読みしたとき、記憶装置 2 a , 2 b に対する 2 回目のコマンドの発行量が「 1 」回であり、記憶装置 2 c , 2 d に対する 2 回目のコマンドの発行量が「 2 」回である。

20

【 0 0 2 0 】

続いて、第 2 のデータ量でデータを先読みしたとき、記憶装置 2 b , 2 c に対する 3 回目のコマンドの発行量が「 1 」回であり、記憶装置 2 a , 2 d に対する 3 回目コマンドの発行量が「 2 」回である。続いて、第 2 のデータ量でデータを先読みしたとき、記憶装置 2 c , 2 d に対する 4 回目のコマンドの発行量が「 1 」回であり、記憶装置 2 a , 2 b に対する 4 回目コマンドの発行量が「 2 」回である。

30

【 0 0 2 1 】

したがって、第 2 のデータ量でデータを先読みするとき、記憶装置 2 a , 2 b , 2 c , 2 d に対するコマンドの総発行量は、記憶装置 2 a , 2 b , 2 c , 2 d のすべてが「 6 」回である。したがって、レスポンス低下によりプリフェッチステージングの性能を低下させる原因となる記憶装置が存在しない。すなわち、第 2 のデータ量は、データを先読みするごとに、コマンド発行回数に偏りが生じる記憶装置が順次交代するデータ量である。

【 0 0 2 2 】

これにより、ストレージ制御装置 1 は、プリフェッチステージングごとのリードコマンドの発行回数の不均衡を循環させて、リードコマンドの発行回数の不均衡が特定の記憶装置に固定化されることを抑止する。したがって、ストレージ制御装置 1 は、特定の記憶装置におけるレスポンス低下によるプリフェッチステージングの性能低下を抑止することができる。すなわち、ストレージ制御装置 1 は、シーケンシャルリードのスループットを向上できる。

40

【 0 0 2 3 】

[第 2 の実施形態]

次に、第 2 の実施形態のストレージシステムについて図 2 を用いて説明する。図 2 は、第 2 の実施形態のストレージシステムの構成の一例を示す図である。

【 0 0 2 4 】

ストレージシステム 5 は、ホスト 6 と、ネットワーク (伝送路) 7 と、ストレージ装置 1 0 とを含む。ネットワーク 7 は、たとえば 1 または複数のストレージ装置 1 0 と、1 ま

50

たは複数のホスト6を接続する。ストレージ装置10は、ネットワーク7を介してホスト6と通信可能に接続する。ホスト6は、ライトリクエストを発行してストレージ装置10にデータを書き込み、またリードリクエストを発行してストレージ装置10からデータを読み出すことができる。

【0025】

次に、第2の実施形態のストレージ装置について図3を用いて説明する。図3は、第2の実施形態のストレージ装置の構成の一例を示す図である。

ストレージ装置10は、チャンネルアダプタ11, 12と、リモートアダプタ13, 14と、コントローラモジュール(以下、CM(Controllor Module))20, 30と、ドライブエンクロージャ(以下、DE)50を含む。

10

【0026】

ストレージ装置10は、チャンネルアダプタ11, 12を介してホスト6と接続する。チャンネルアダプタ11は、CM20に対応して備えられ、チャンネルアダプタ12は、CM30に対応して備えられる。ストレージ装置10は、リモートアダプタ13, 14を介して他のストレージ装置と接続できる。リモートアダプタ13は、CM20に対応して備えられ、リモートアダプタ14は、CM30に対応して備えられる。

【0027】

DE50は、複数のHDDを含む。なお、DE50は、HDDに代えて、あるいはHDDとともにSSDなどその他の記憶装置を含むものであってもよい。

CM20とCM30は、相互に接続して負荷を分担することができる。なお、ストレージ装置10は、2つのCM20, 30を含むが、これに限らず、CM20とCM30のいずれか一方を含むものであってもよいし、3つ以上のCMを含むものであってもよく、たとえば、4つあるいは8つのCMを含むものであってもよい。

20

【0028】

CM20は、プロセッサ21と、メモリ22と、ディスクアダプタ23, 24を含む。CM30は、プロセッサ31と、メモリ32と、ディスクアダプタ33, 34を含む。なお、CM30はCM20と同様の構成を有するため、以下、CM20の説明をもってCM30の説明に代える。

【0029】

プロセッサ21、メモリ22、およびディスクアダプタ23, 24は、図示しないバスを介して接続されている。プロセッサ21は、CM20全体を制御し、階層制御を含むストレージ制御をおこなう。なお、プロセッサ21は、マルチプロセッサであってもよい。プロセッサ21は、たとえばCPU(Central Processing Unit)、MPU(Micro Processing Unit)、DSP(Digital Signal Processor)、ASIC(Application Specific Integrated Circuit)、またはPLD(Programmable Logic Device)である。またプロセッサ21は、CPU、MPU、DSP、ASIC、PLDのうちの2以上の要素の組み合わせであってもよい。

30

【0030】

メモリ22は、HDDからデータを読み出したときにデータを保持するほか、HDDにデータを書き込むときのバッファとなる。また、メモリ22は、ユーザデータや制御情報を格納する。

40

【0031】

たとえば、メモリ22は、RAM(Random Access Memory)や不揮発性メモリを含む。RAMは、CM20の主記憶装置として使用される。RAMには、プロセッサ21に実行させるオペレーティングシステム(Operating System)のプログラムやファームウェア、アプリケーションプログラムの少なくとも一部が一時的に格納される。また、RAMには、プロセッサ21による処理に必要な各種データが格納される。また、RAMは、各種データの格納に用いるメモリと別体にキャッシュメモリを含む。なお、キャッシュメモリは、HDDから先読みしたデータを格納する。

【0032】

50

不揮発性メモリは、ストレージ装置 10 の電源遮断時においても記憶内容を保持する。不揮発性メモリは、たとえば、EEPROM (Electrically Erasable and Programmable Read Only Memory) やフラッシュメモリなどの半導体記憶装置や、HDD などである。また、不揮発性メモリは、CM20 の補助記憶装置として使用される。不揮発性メモリには、オペレーティングシステムのプログラムやファームウェア、アプリケーションプログラム、および各種データが格納される。

【0033】

バスに接続される周辺機器としては、入出力インタフェース、および通信インタフェースがある。入出力インタフェースは、入出力装置と接続して入出力をおこなう。入出力インタフェースは、HDD などの記憶装置から送られてくる信号やデータをプロセッサ 21 やメモリ 22 に送信する。また、入出力インタフェースは、たとえば、プロセッサ 21 から受信した信号を、他の制御部や、CM20 と接続する出力装置に出力する。通信インタフェースは、ストレージ装置 10 内の他の CM (CM30) との間でデータの送受信をおこなう。

10

【0034】

ディスクアダプタ 23, 24 は、HDD とのインタフェース制御 (アクセス制御) をおこなう。

以上のようなハードウェア構成によって、ストレージ装置 10、あるいは CM20, 30 の処理機能を実現することができる。なお、第 1 の実施形態に示したストレージ制御装置 1 も、ストレージ装置 10、あるいは CM20, 30 と同様のハードウェアにより実現することができる。

20

【0035】

次に、DE50 が収容する HDD により構成される RAID グループと、RAID グループから割り当てられるボリュームについて図 4 を用いて説明する。図 4 は、第 2 の実施形態の RAID グループと、RAID グループから割り当てられるボリュームの一例を示す図である。

【0036】

DE50 は、HDD 501 から HDD 503 と、HDD 511 から HDD 515 を含む。HDD 501 から HDD 503 は、RAID 5 の RAID グループ #0 を構成する。RAID グループ #0 は、所定の記憶容量を割り当てたボリューム #0 とボリューム #1 を含む。HDD 511 から HDD 515 は、RAID 6 の RAID グループ #1 を構成する。RAID グループ #1 は、所定の記憶容量を割り当てたボリューム #2 とボリューム #3 を含む。ボリューム #0 からボリューム #3 は、ホスト 6 が認識可能な論理ボリュームである。

30

【0037】

なお、RAID グループ #0 は、RAID 5 であることから、分散パリティを使用するストライプセットを有する RAID レベルである。また、RAID グループ #1 は、RAID 6 であることから、二重分散パリティを使用するストライプセットを有する RAID レベルである。

40

【0038】

次に、CM の処理機能の構成について図 5 を用いて説明する。図 5 は、第 2 の実施形態の CM の処理機能の構成の一例を示すブロック図である。

CM20 は、記憶部 201 と、制御部 202 としての機能を備える。CM20 は、ストレージ制御装置の一形態である。記憶部 201 は、メモリ 22 によって実現される。制御部 202 は、たとえば、CM20 のプロセッサ 21 が所定のプログラムを実行することで実現される。

【0039】

記憶部 201 は、プリフェッチデータ記憶部 2011 と、シーケンシャル性検出用テーブル記憶部 2012 と、RAID 構成情報記憶部 2013 としての機能を含む。プリフェッチデータ記憶部 2011 は、記憶装置から先読みしたデータを記憶する。プリフェッチ

50

データ記憶部 2011 は、メモリ 22 のうちキャッシュメモリによって実現される。シーケンシャル性検出用テーブル記憶部 2012 は、シーケンシャル性検出用テーブルを記憶する。シーケンシャル性検出用テーブルについては、図 7 を用いて後で説明する。RAID 構成情報記憶部 2013 は、RAID 構成情報を記憶する。RAID 構成情報は、RAID グループごとの RAID レベルと、RAID グループを構成する記憶装置群と、RAID グループから記憶領域の割り当てのあるボリュームとを特定可能な情報とを含む。

【0040】

制御部 202 は、データ量決定部 2021 と、プリフェッチステージング部 2022 とを含む。データ量決定部 2021 は、先読みするデータ量（プリフェッチステージングデータ量）を決定する。データ量決定部 2021 は、後述するデータ量決定処理をプロセッサ 21 が実行することで実現される。プリフェッチステージング部 2022 は、データ量決定部 2021 が決定したプリフェッチステージングデータ量で先読みをおこなう。プリフェッチステージング部 2022 は、後述するプリフェッチステージング処理をプロセッサ 21 が実行することで実現される。

10

【0041】

次に、第 2 の実施形態のプリフェッチステージング処理について図 6 を用いて説明する。図 6 は、第 2 の実施形態のプリフェッチステージング処理のフローチャートを示す図である。

【0042】

プリフェッチステージング処理は、プリフェッチステージングの実行条件を判定して、実行条件が成立する場合にプリフェッチステージングを実行する処理である。プリフェッチステージング処理は、ホスト 6 からリードリクエストを受け付けたことを契機に、制御部 202 が実行する処理である。

20

【0043】

[ステップ S11] プリフェッチステージング部 2022（制御部 202）は、シーケンシャル性検出用テーブルを参照する。なお、ここでいうシーケンシャル性とは、リードリクエストが要求する読出しデータの連続性を指す。

【0044】

ここで、シーケンシャル性検出用テーブルについて図 7 を用いて説明する。図 7 は、第 2 の実施形態のシーケンシャル性検出用テーブルの一例を示す図である。

30

シーケンシャル性検出用テーブル 300 は、シーケンシャル性検出対象のボリューム（以下、対象ボリューム）に対応して所定数のエントリを有する。シーケンシャル性検出用テーブル 300 は、項目「ボリューム番号」と、項目「先頭 LBA（Logical Block Address）」と、項目「最終 LBA」と、項目「シーケンシャル性検出回数」と、項目「タイムスタンプ」とを含む。項目「ボリューム番号」は、対象ボリュームを特定可能な識別情報を示し、対象ボリュームに対応するエントリを特定可能にする。対象ボリュームに対応するエントリは、項目「先頭 LBA」と、項目「最終 LBA」と、項目「シーケンシャル性検出回数」と、項目「タイムスタンプ」をパラメータ（エントリパラメータ）に含む。

【0045】

項目「先頭 LBA」は、シーケンシャル性検出開始時の対象ボリュームの先頭 LBA を示し、シーケンシャル性検出開始の契機となったリードリクエストから特定される。項目「最終 LBA」は、シーケンシャル性検出中の対象ボリュームの最終 LBA を示す。最終 LBA は、シーケンシャル性検出中に受け付けた最後のリードリクエスト、あるいはプリフェッチステージングの実行結果から特定される。項目「シーケンシャル性検出回数」は、シーケンシャル性検出中に対象ボリュームのシーケンシャル性を検出した回数を示す。項目「タイムスタンプ」は、対象ボリュームのエントリを最後に更新した時刻を示す。

40

【0046】

たとえば、項目番号「#V1」で特定される対象ボリュームは、先頭 LBA「#H1」、最終 LBA「#T1」、シーケンシャル性検出回数「#S1」、タイムスタンプ「#TS1」であることを示す。

50

【 0 0 4 7 】

再び、図 6 を用いたプリフェッチステージング処理の説明に戻る。

[ステップ S 1 2] プリフェッチステージング部 2 0 2 2 は、シーケンシャル性検出用テーブルに、受け付けたリードリクエストが読出し対象とするボリュームを対象ボリュームとするエントリ（対象エントリ）があるか否かを判定する。プリフェッチステージング部 2 0 2 2 は、対象エントリがある場合にステップ S 1 6 にすすみ、対象エントリがない場合にステップ S 1 3 にすすむ。

【 0 0 4 8 】

[ステップ S 1 3] プリフェッチステージング部 2 0 2 2 は、シーケンシャル性検出用テーブルのエントリ数が上限に達している場合に、項目「タイムスタンプ」を参照して最も古いエントリを削除する。なお、プリフェッチステージング部 2 0 2 2 は、シーケンシャル性検出用テーブルのエントリ数が上限に達していない場合に、最も古いエントリを削除することを要しない。

10

【 0 0 4 9 】

[ステップ S 1 4] プリフェッチステージング部 2 0 2 2 は、受け付けたリードリクエストが読出し対象とするボリュームを対象ボリュームとする新規エントリを追加する。

[ステップ S 1 5] プリフェッチステージング部 2 0 2 2 は、新規エントリに対象ボリュームのエントリパラメータを設定する。プリフェッチステージング部 2 0 2 2 は、エントリパラメータを設定した後、プリフェッチステージング処理を終了する。

【 0 0 5 0 】

[ステップ S 1 6] プリフェッチステージング部 2 0 2 2 は、対象エントリの項目「最終 L B A」について、受け付けたリードリクエストの最終 L B A で更新する。

[ステップ S 1 7] プリフェッチステージング部 2 0 2 2 は、対象エントリの項目「シーケンシャル性検出回数」について、シーケンシャル性検出回数を「1」インクリメントして更新する。

20

【 0 0 5 1 】

[ステップ S 1 8] プリフェッチステージング部 2 0 2 2 は、シーケンシャル性検出回数があらかじめ設定する閾値を超えたか否かを判定する。プリフェッチステージング部 2 0 2 2 は、シーケンシャル性検出回数が閾値を超えた場合にシーケンシャル性を検出したとしてステップ S 1 9 にすすむ。一方、プリフェッチステージング部 2 0 2 2 は、シーケンシャル性検出回数が閾値を超えない場合にプリフェッチステージング処理を終了する。

30

【 0 0 5 2 】

[ステップ S 1 9] データ量決定部 2 0 2 1（制御部 2 0 2）は、データ量決定処理を実行する。データ量決定処理は、プリフェッチステージングデータ量を決定する処理である。データ量決定処理の詳細については、図 8 を用いて後で説明する。

【 0 0 5 3 】

[ステップ S 2 0] プリフェッチステージング部 2 0 2 2 は、ステップ S 1 9 において決定したプリフェッチステージングデータ量にしたがいプリフェッチステージングを実行する。

【 0 0 5 4 】

[ステップ S 2 1] プリフェッチステージング部 2 0 2 2 は、対象エントリの項目「最終 L B A」について、プリフェッチステージングの実行結果にしたがう最終 L B A で更新する。

40

【 0 0 5 5 】

[ステップ S 2 2] プリフェッチステージング部 2 0 2 2 は、対象エントリの項目「シーケンシャル性検出回数」について、シーケンシャル性検出回数を初期値にリセット（たとえば、ゼロクリア）する。プリフェッチステージング部 2 0 2 2 は、シーケンシャル性検出回数をリセットした後、プリフェッチステージング処理を終了する。

【 0 0 5 6 】

次に、第 2 の実施形態のデータ量決定処理について図 8 を用いて説明する。図 8 は、第

50

2の実施形態のデータ量決定処理のフローチャートを示す図である。

データ量決定処理は、プリフェッチステージングデータ量を決定する処理である。データ量決定処理は、プリフェッチステージング処理のステップS19においてデータ量決定部2021が実行する処理である。

【0057】

【ステップS31】データ量決定部2021は、対象ボリュームのRAIDレベルを検出する。データ量決定部2021は、RAID構成情報を参照することで、対象ボリュームに記憶領域を割り当てたRAIDグループを特定するとともに、当該RAIDグループのRAIDレベルを特定することができる。

【0058】

【ステップS32】データ量決定部2021は、対象ボリュームのRAIDレベルがRAID5またはRAID6であるか否かを判定する。データ量決定部2021は、対象ボリュームのRAIDレベルがRAID5またはRAID6でない場合にステップS33にすすみ、RAID5またはRAID6である場合にステップS34にすすむ。

【0059】

【ステップS33】データ量決定部2021は、プリフェッチステージングデータ量として規定値（第1の実施形態の第1のデータ量に相当）を取得する。規定値は、あらかじめ設定される値であり、たとえば、記憶装置など各種性能を考慮して設定される値である。

【0060】

【ステップS34】データ量決定部2021は、対象ボリュームのRAID構成を取得する。これにより、データ量決定部2021は、対象ボリュームのRAIDグループを構成する記憶装置数（ディスク数）を取得することができる。

【0061】

【ステップS35】データ量決定部2021は、プリフェッチステージングデータ量として変更値（第1の実施形態の第2のデータ量に相当）を算出する。変更値は、記憶装置に発行するリードコマンドの発行量の偏りを低減可能な値である。データ量決定部2021は、たとえば、(1)式により変更値Vを算出することができる。

【0062】

$$V = (d \times n + 1) \times (d - p) \times c \quad \dots \quad (1)$$

ただし、 d = (RAIDグループの構成ディスク数)であり、 n = (任意の自然数)であり、 p = (パリティディスク数)であり、 c = (キャッシュページサイズ)である。なお、パリティディスク数 p は、RAID5において「1」であり、RAID6において「2」である。また、任意の自然数 n は、変更値Vが規定値を超える最小の値とする。なお、(1)式中の定数「1」は、コマンド発行量に偏りが生じる記憶装置を順次交代させるためのシフト量である。

【0063】

たとえば、 d = 「5」、 p = 「2」、 c = 「64 kB (byte)」、規定値が「1024 (= 16c) kB」のとき、変更値V ($n = 1$)は「1152 kB」である。また、変更値V ($n = 2$)は「2112 kB」であり、変更値V ($n = 1$)と変更値V ($n = 2$)のいずれもが規定値「1024 kB」を超える。したがって、この場合、データ量決定部2021は、変更値Vとして「1152 kB」を算出する。

【0064】

【ステップS36】データ量決定部2021は、ステップS33で取得した規定値、またはステップS35で算出した変更値を、プリフェッチステージングデータ量として設定し、データ量決定処理を終了する。

【0065】

なお、データ量決定部2021がステップS32で実行したRAIDレベルの判定は、規定値をプリフェッチステージングデータ量としてプリフェッチステージングを実行した場合に生じる記憶装置ごとのコマンド発行量の偏りが生じる状態（検出対象状態）の検出

10

20

30

40

50

に相当する。このように、データ量決定部 2021 は、必ずしも規定値を用いて検出対象状態を検出することを要しない。データ量決定部 2021 は、あらかじめ設定した検出条件によって、検出対象状態を検出することができる。

【0066】

たとえば、データ量決定部 2021 は、RAID レベルが RAID 5 または RAID 6 であるかに代えて、RAID グループを構成する記憶装置群におけるデータディスクの数（データ領域に対応する記憶装置の数）を検出対象状態の検出条件とすることができる。また、データ量決定部 2021 は、RAID レベルが RAID 5 または RAID 6 であるかに代えて、RAID グループを構成する記憶装置群におけるパリティディスクの数（パリティ領域に対応する記憶装置の数）を検出対象状態の検出条件とすることができる。具体的には、データ量決定部 2021 は、RAID グループを構成する記憶装置群のパリティディスクの数が「2」であることを、検出対象状態の検出条件とすることができる。パリティディスクの数「2」は、RAID レベルが RAID 6 に相当する。

10

【0067】

また、データ量決定部 2021 は、プリフェッチステージング処理の実行ごとに、プリフェッチステージングデータ量を決定したが、これに代えてボリューム作成時に 1 度おこなうようにしてもよい。これにより、プリフェッチステージング部 2022 は、ボリューム作成時に決定したプリフェッチステージングデータ量を取得して、プリフェッチステージングを実行するようにしてもよい。

20

【0068】

次に、変更値が記憶装置に発行するリードコマンドの発行量の偏りを低減可能な値となることについて、図 9 から図 13 を用いて説明する。

まず、規定値をプリフェッチステージングデータ量として RAID 5 におけるディスクセットからデータを読み出す場合について、図 9 を用いて説明する。図 9 は、規定値をプリフェッチステージングデータ量として RAID 5 におけるディスクセットからデータを読み出す参考例を示す図である。

【0069】

図示する RAID グループは、RAID レベルが RAID 5 であり、Disk # 0 から Disk # 2 の 3 つのディスクから構成される。Disk # 0 から Disk # 2 は、1 つのストライプに 2 つのデータディスクと 1 つのパリティディスクとを含む。データブロック（D + 番号で表記（たとえば、「D00」））をハッチングなしの四角枠で示し、パリティブロック（P + 番号で表記（たとえば、「P0」））をハッチングありの四角枠で示す。データブロックとパリティブロックは、それぞれキャッシュページサイズ（たとえば、64 kB）の大きさを有する。図示するプリフェッチステージング範囲は、プリフェッチステージングデータ量が 18 キャッシュページサイズであることを示す。

30

【0070】

プリフェッチステージング範囲のデータをキャッシュメモリにステージングするとき、プリフェッチステージング部 2022 は、各ディスクにリードコマンドを発行する。プリフェッチステージング部 2022 は、データブロックを読み出し対象とし、パリティブロックの読み出し対象としない。一般に、パリティブロックを読み飛ばした方がディスクからデータを読み出す際のスループットが高いことによる。

40

【0071】

すなわち、プリフェッチステージング部 2022 は、1 つのディスクについて連続するデータブロックを 1 つのリードコマンドで読み出す。したがって、プリフェッチステージング部 2022 は、Disk # 0 について 3 回、Disk # 1 について 4 回、Disk # 2 について 3 回のリードコマンドを発行してプリフェッチステージングを実行する。Disk # 0 について 3 回のリードコマンドの発行は、データブロック「D00, D03」とデータブロック「D06, D09」とデータブロック「D12, D15」を読み出し対象とするリードコマンドである。また、Disk # 1 について 4 回のリードコマンドの発行は、データブロック「D01」とデータブロック「D04, D07」とデータブロック「D

50

「D 1 3」とデータブロック「D 1 6」を読み出し対象とするリードコマンドである。

【0072】

このようなプリフェッチステージング範囲は、次のプリフェッチステージングにおいても各ディスクに対するリードコマンド発行回数が同じである。すなわち、Disk # 0, Disk # 2のリードコマンド発行回数とDisk # 1のリードコマンド発行回数の差が蓄積し、Disk # 0, Disk # 2と比較してDisk # 1のデータ読み出しにかかる負荷が大きくなる。

【0073】

次に、規定値をプリフェッチステージングデータ量としてRAID 6におけるディスクセットからデータを読み出す場合について、図10を用いて説明する。図10は、規定値をプリフェッチステージングデータ量としてRAID 6におけるディスクセットからデータを読み出す参考例を示す図である。

10

【0074】

図示するRAIDグループは、RAIDレベルがRAID 6であり、Disk # 0からDisk # 4の5つのディスクから構成される。Disk # 0からDisk # 4は、1つのストライプに3つのデータディスクと2つのパリティディスクを含む。データブロック(D + 番号で表記(たとえば、「D 0 0」))をハッチングなしの四角枠で示し、パリティブロック(P + 番号で表記(たとえば、「P 0」)またはQ + 番号で表記(たとえば、「Q 0」))をハッチングありの四角枠で示す。データブロックとパリティブロックは、それぞれキャッシュページサイズ(たとえば、64kB)の大きさを有する。図示するプリフェッチステージング範囲は、プリフェッチステージングデータ量が15キャッシュページサイズであることを示す。

20

【0075】

プリフェッチステージング範囲のデータをキャッシュメモリにステージングするとき、プリフェッチステージング部2022は、各ディスクにリードコマンドを発行する。プリフェッチステージング部2022は、Disk # 0について1回、Disk # 1について2回、Disk # 2について2回、Disk # 3について1回、Disk # 4について1回のリードコマンドを発行してプリフェッチステージングを実行する。Disk # 0について1回のリードコマンドの発行は、データブロック「D 0 0, D 0 4, D 0 8」を読み出し対象とするリードコマンドである。また、Disk # 1について2回のリードコマンドの発行は、データブロック「D 0 1, D 0 5」とデータブロック「D 1 2」を読み出し対象とするリードコマンドである。

30

【0076】

このようなプリフェッチステージング範囲は、次のプリフェッチステージングにおいても各ディスクに対するリードコマンド発行回数が同じである。すなわち、Disk # 0, Disk # 3, Disk # 4のリードコマンド発行回数とDisk # 1, Disk # 2のリードコマンド発行回数の差が蓄積し、Disk # 0, Disk # 3, Disk # 4と比較してDisk # 1, Disk # 2のデータ読み出しにかかる負荷が大きくなる。

【0077】

プリフェッチステージングを繰り返すことによるリードコマンド発行回数の差の蓄積について図11を用いて説明する。図11は、リードコマンド発行回数の差が蓄積するプリフェッチステージングの参考例を示す図である。

40

【0078】

図示するRAIDグループは、図10で図示したRAIDレベルがRAID 6のRAIDグループである。1回目のプリフェッチステージングでDisk # 0からDisk # 4へのリードコマンド発行回数がそれぞれ、1回、2回、2回、1回、1回であったところ、2回目のプリフェッチステージングでもDisk # 0からDisk # 4に同じ回数のリードコマンドが発行されていることを示す。したがって、2回目のプリフェッチステージングでDisk # 0, Disk # 3, Disk # 4へのリードコマンド発行回数が2回、Disk # 1, Disk # 2へのリードコマンド発行回数が4回となっている。このような

50

リードコマンド発行回数の差は、負荷の大きなディスクが固定されて、プリフェッチステージングを繰返すごとに拡大する。

【0079】

データ量決定部2021は、このようなコマンド発行量の偏りが生じる場合に、プリフェッチステージングデータ量を規定値から変更値に差し替える。

プリフェッチステージングデータ量を規定値から変更値に差し替えたときの、リードコマンド発行回数について図12を用いて説明する。図12は、第2の実施形態のリードコマンド発行回数の一例を示す図である。

【0080】

図示するRAIDグループは、図10で図示したRAIDレベルがRAID6のRAIDグループである。図10と異なるのは、プリフェッチステージング範囲の大きさである。図10においては、プリフェッチステージング範囲の大きさが15キャッシュページサイズであったが、図12におけるプリフェッチステージング範囲の大きさは、18キャッシュページサイズであることを示す。すなわち、変更値に差し替えられたプリフェッチステージングデータ量が18キャッシュページサイズである。

10

【0081】

これにより、プリフェッチステージング部2022は、1回目のプリフェッチステージングにおいて、Disk#0からDisk#4へのリードコマンド発行回数は、それぞれ2回、2回、2回、1回、1回である。また、プリフェッチステージング部2022は、2回目のプリフェッチステージングにおいて、Disk#0からDisk#4へのリードコマンド発行回数は、それぞれ2回、2回、1回、1回、2回である。

20

【0082】

したがって、プリフェッチステージング部2022は、プリフェッチステージングの複数回の実行において、Disk#0からDisk#4へのリードコマンド発行回数が固定化されることを抑止する。

【0083】

次に、プリフェッチステージングを5回実行したときの、Disk#0からDisk#4へのリードコマンド発行回数について図13を用いて説明する。図13は、第2の実施形態のリードコマンド発行回数の均一化の一例を示す図である。

【0084】

1回目および2回目のプリフェッチステージングにおけるDisk#0からDisk#4へのリードコマンド発行回数は、図12を用いて説明したとおりである。3回目のプリフェッチステージングにおいて、Disk#0からDisk#4へのリードコマンド発行回数は、それぞれ2回、1回、1回、2回、2回である。4回目のプリフェッチステージングにおいて、Disk#0からDisk#4へのリードコマンド発行回数は、それぞれ1回、1回、2回、2回、2回である。5回目のプリフェッチステージングにおいて、Disk#0からDisk#4へのリードコマンド発行回数は、それぞれ1回、2回、2回、2回、1回である。

30

【0085】

したがって、プリフェッチステージング部2022は、プリフェッチステージングの5回の実行において、Disk#0からDisk#4へのリードコマンド発行回数の合計をすべて8回として均一化することができる。このように、ストレージ装置10は、Disk#0からDisk#4への1回のプリフェッチステージングの実行によるリードコマンドの発行回数の不均衡を、1周期(不均衡是正周期)5回のプリフェッチステージングの実行により是正することができる。

40

【0086】

このような不均衡是正周期は、1回のプリフェッチステージングの実行によるリードコマンドの発行回数の不均衡が固定化される規定値を変更値に差し替えることにより実現できる。

【0087】

50

これにより、ストレージ装置 10 は、D i s k # 0 から D i s k # 4 のうち特定のディスクにおけるレスポンス低下によるプリフェッチステージングの性能低下を抑止することができる。すなわち、ストレージ装置 10 は、シーケンシャルリードのスループットを向上できる。

【0088】

[変形例]

第 2 の実施形態では、データ量決定処理のステップ S 3 5 において、(1) 式を用いて変更値 V を算出したが、変更値 V の算出式はこれに限らない。

【0089】

データ量決定部 2021 は、たとえば、(2) 式により変更値 V を算出することができる。

$$V = (d \times n + k) \times (d - p) \times c \quad \dots \quad (2)$$

ただし、d = (R A I D グループの構成ディスク数) であり、n = (任意の自然数) であり、k = (d と互いに素な関係にある自然数)、p = (パリティディスク数) であり、c = (キャッシュページサイズ) である。なお、パリティディスク数 p は、R A I D 5 において「 1 」であり、R A I D 6 において「 2 」である。また、任意の自然数 n は、変更値 V が規定値を超える最小の値とする。なお、(2) 式中の定数「 k 」は、コマンド発行量に偏りが生じる記憶装置を順次交代させるためのシフト量である。

【0090】

たとえば、d = 「 5 」のとき、k は、「 2 」や「 3 」などの値をとり得る。したがって、たとえば、d = 「 5 」、k = 「 2 」、p = 「 2 」、c = 「 6 4 k B (b y t e) 」、規定値が「 1 0 2 4 (= 1 6 c) k B 」のとき、変更値 V (n = 1) は「 1 3 4 4 k B 」である。また、変更値 V (n = 2) は「 2 3 0 4 k B 」であり、変更値 V (n = 1) と変更値 V (n = 2) のいずれもが規定値「 1 0 2 4 k B 」を超える。したがって、この場合、データ量決定部 2021 は、変更値 V として「 1 3 4 4 k B 」を算出する。

【0091】

このようにして、データ量決定部 2021 は、1 回のプリフェッチステージングの実行によるリードコマンドの発行回数の不均衡を是正可能な変更値 V を容易に算出することができる。

【0092】

なお、上記の処理機能は、コンピュータによって実現することができる。その場合、ストレージ制御装置 1、ストレージ装置 10、C M 2 0、3 0 が有すべき機能の処理内容を記述したプログラムが提供される。そのプログラムをコンピュータで実行することにより、上記処理機能がコンピュータ上で実現される。処理内容を記述したプログラムは、コンピュータで読み取り可能な記録媒体に記録しておくことができる。コンピュータで読み取り可能な記録媒体としては、磁気記憶装置、光ディスク、光磁気記録媒体、半導体メモリなどがある。磁気記憶装置には、ハードディスク装置 (H D D)、フレキシブルディスク (F D)、磁気テープなどがある。光ディスクには、D V D (Digital Versatile Disk)、D V D - R A M、C D (Compact Disc) - R O M、C D - R (Recordable) / R W (Re Writable) などがある。光磁気記録媒体には、M O (Magneto-Optical disk) などがある。

【0093】

プログラムを流通させる場合には、たとえば、そのプログラムが記録された D V D、C D - R O M などの可搬型記録媒体が販売される。また、プログラムをサーバコンピュータの記憶装置に格納しておき、ネットワークを介して、サーバコンピュータから他のコンピュータにそのプログラムを転送することもできる。

【0094】

プログラムを実行するコンピュータは、たとえば、可搬型記録媒体に記録されたプログラムもしくはサーバコンピュータから転送されたプログラムを、自己の記憶装置に格納する。そして、コンピュータは、自己の記憶装置からプログラムを読み取り、プログラムに

10

20

30

40

50

従った処理を実行する。なお、コンピュータは、可搬型記録媒体から直接プログラムを読み取り、そのプログラムに従った処理を実行することもできる。また、コンピュータは、ネットワークを介して接続されたサーバコンピュータからプログラムが転送されるごとに、逐次、受け取ったプログラムに従った処理を実行することもできる。

【0095】

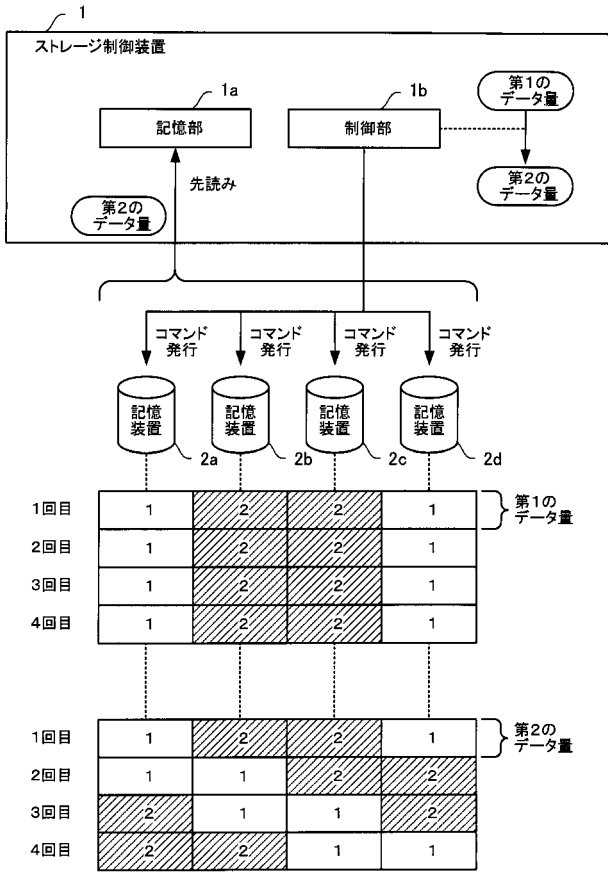
また、上記の処理機能の少なくとも一部を、DSP、ASIC、PLDなどの電子回路で実現することもできる。

【符号の説明】

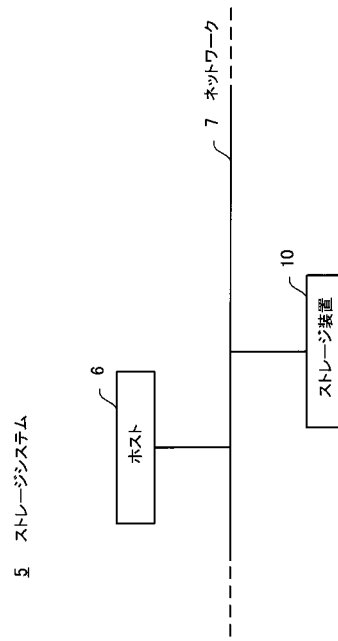
【0096】

- | | | |
|---|---------------------|----|
| 1 | ストレージ制御装置 | 10 |
| 1 a | 記憶部 | |
| 1 b | 制御部 | |
| 2 a , 2 b , 2 c , 2 d | 記憶装置 | |
| 5 | ストレージシステム | |
| 6 | ホスト | |
| 7 | ネットワーク | |
| 1 0 | ストレージ装置 | |
| 1 1 , 1 2 | チャンネルアダプタ | |
| 1 3 , 1 4 | リモートアダプタ | |
| 2 0 , 3 0 | コントローラモジュール (C M) | 20 |
| 2 1 , 3 1 | プロセッサ | |
| 2 2 , 3 2 | メモリ | |
| 2 3 , 2 4 , 3 3 , 3 4 | ディスクアダプタ | |
| 5 0 | ドライブエンクロージャ (D E) | |
| 5 0 1 , 5 0 2 , 5 0 3 , 5 1 1 , 5 1 2 , 5 1 3 , 5 1 4 , 5 1 5 | H D D | |
| 2 0 1 | 記憶部 | |
| 2 0 2 | 制御部 | |
| 2 0 1 1 | プリフェッチデータ記憶部 | |
| 2 0 1 2 | シーケンシャル性検出用テーブル記憶部 | |
| 2 0 1 3 | R A I D 構成情報記憶部 | 30 |
| 2 0 2 1 | データ量決定部 | |
| 2 0 2 2 | プリフェッチステージング部 | |

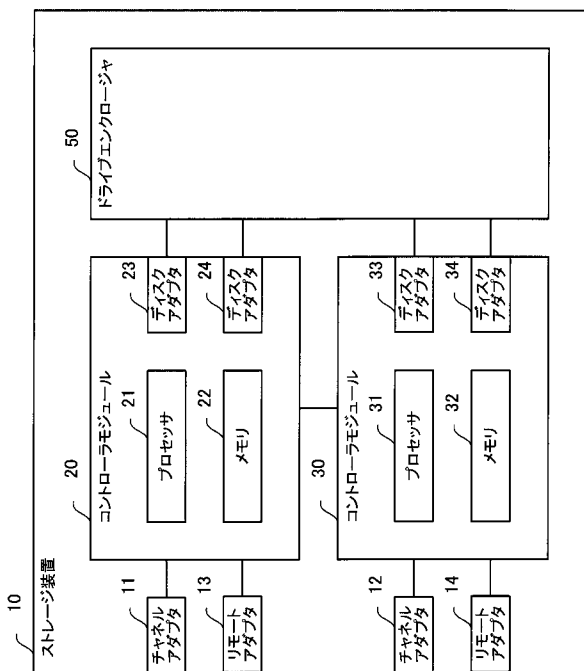
【 図 1 】



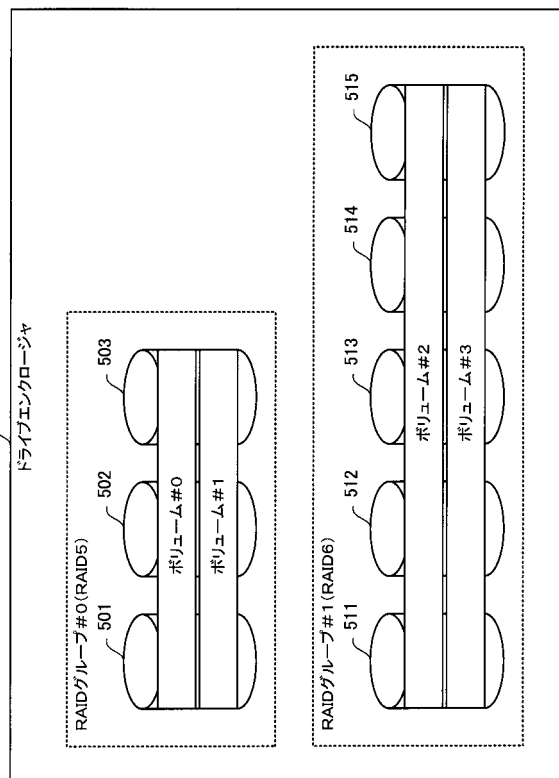
【 図 2 】



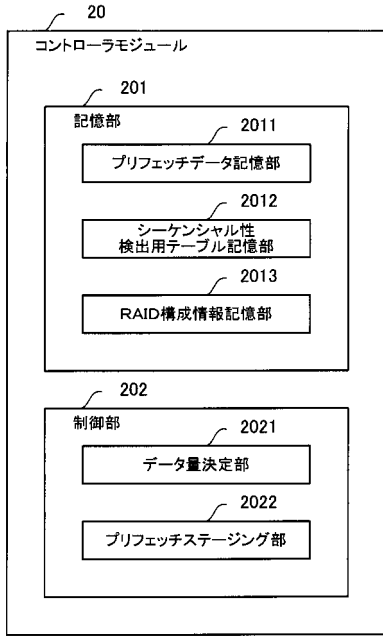
【 図 3 】



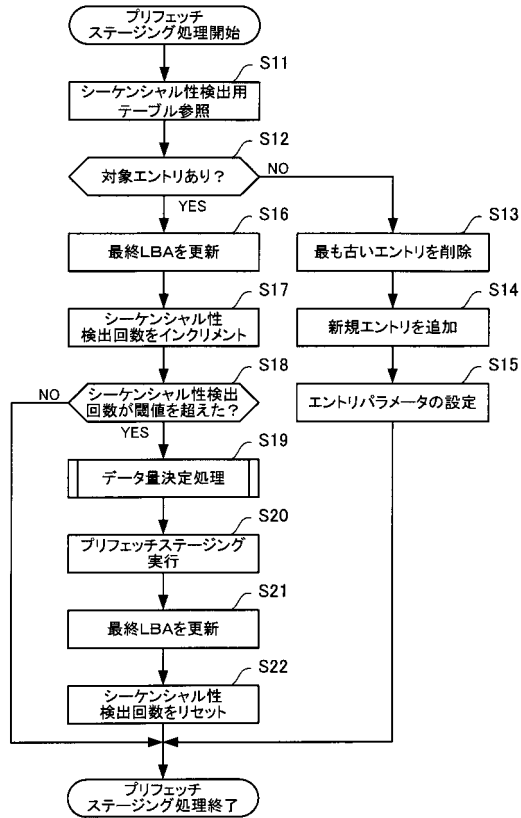
【 図 4 】



【図5】



【図6】

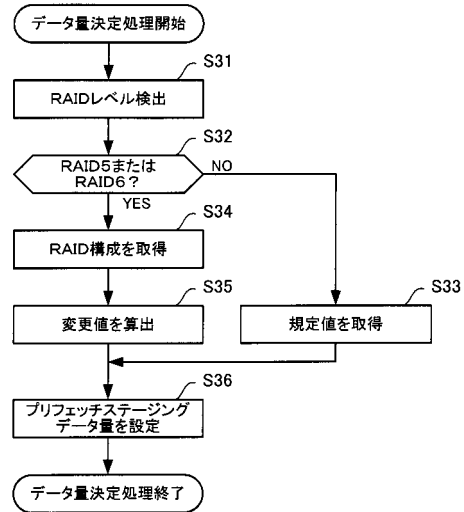


【図7】

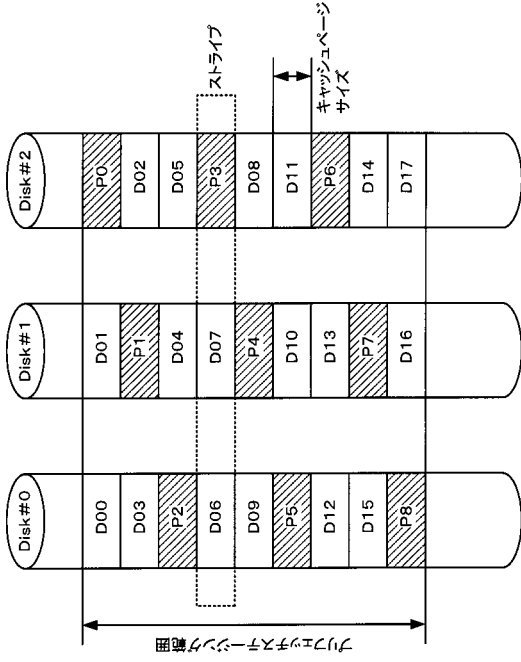
ボリューム番号	先頭LBA	最終LBA	シーケンシャル性検出回数	タイムスタンプ
#V1	#H1	#T1	#S1	#TS1
#V2	#H2	#T2	#S2	#TS2
...

300 シーケンシャル性検出用テーブル

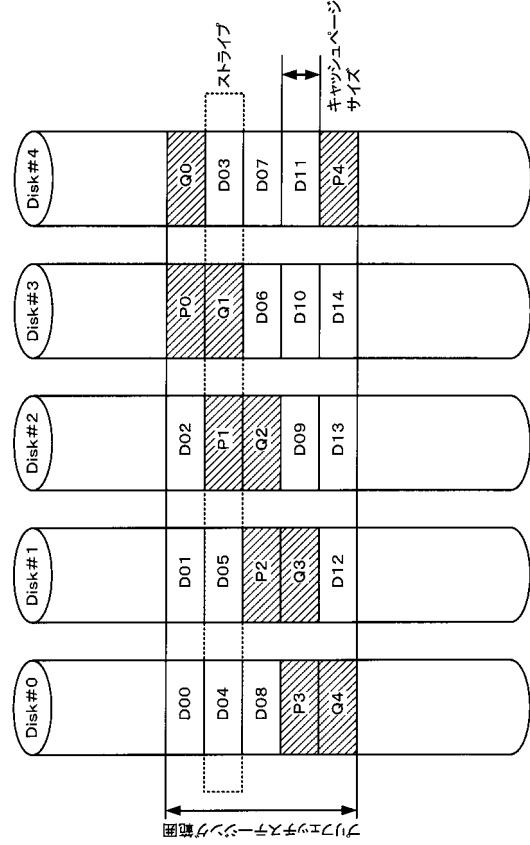
【図8】



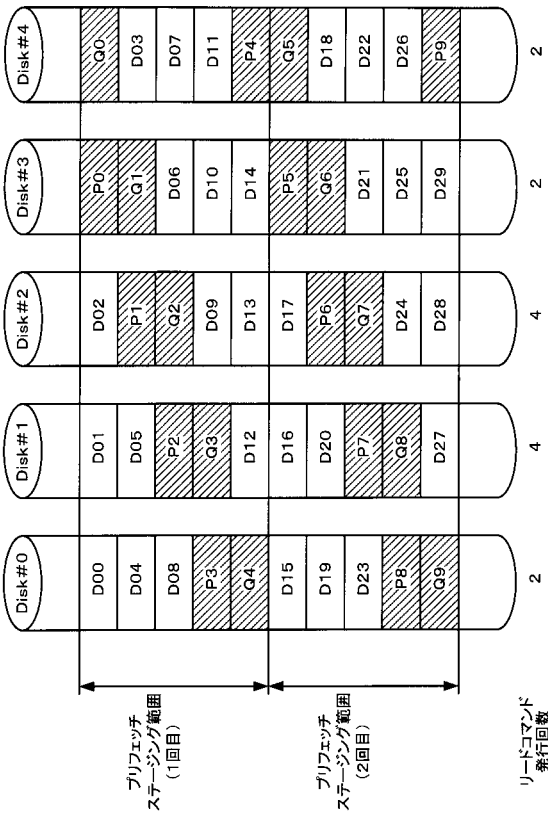
【 図 9 】



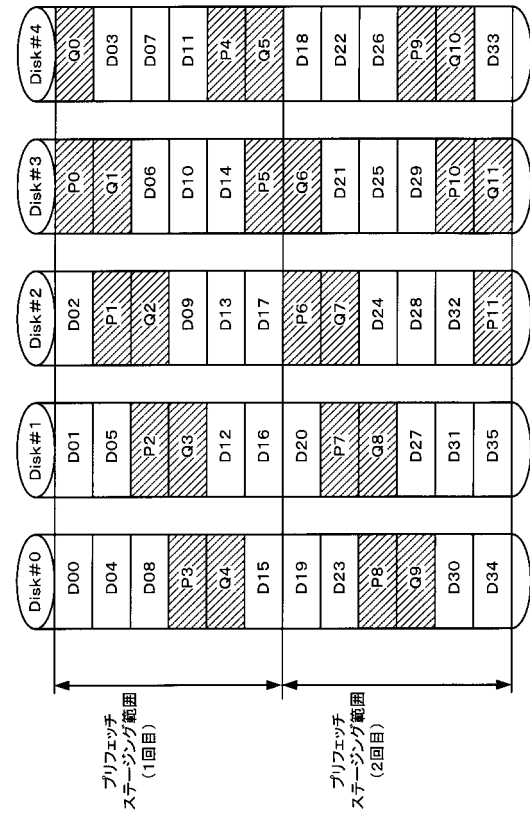
【 図 10 】



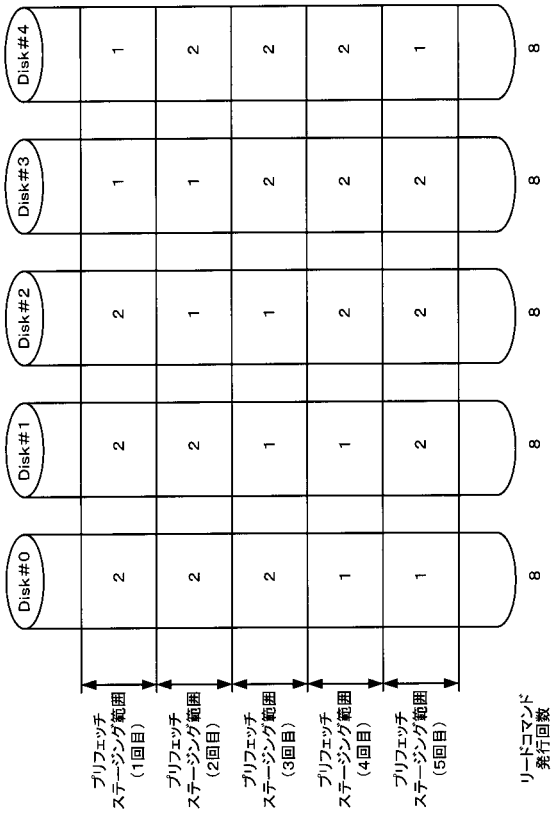
【 図 11 】



【 図 12 】



【 図 1 3 】



フロントページの続き

- (72)発明者 浦田 一宏
神奈川県川崎市中原区上小田中4丁目1番1号 富士通株式会社内
- (72)発明者 土山 由嘉莉
神奈川県川崎市中原区上小田中4丁目1番1号 富士通株式会社内
- (72)発明者 周 广宇
神奈川県川崎市中原区上小田中4丁目1番1号 富士通株式会社内