

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7348056号
(P7348056)

(45)発行日 令和5年9月20日(2023.9.20)

(24)登録日 令和5年9月11日(2023.9.11)

(51)国際特許分類		F I			
G 0 6 F	3/06 (2006.01)	G 0 6 F	3/06	3 0 1 A	
G 0 6 F	13/10 (2006.01)	G 0 6 F	3/06	3 0 4 P	
G 0 6 F	13/14 (2006.01)	G 0 6 F	13/10	3 4 0 A	
		G 0 6 F	13/14	3 1 0 H	
		G 0 6 F	3/06	3 0 2 Z	
					請求項の数 10 (全26頁)
(21)出願番号	特願2019-229906(P2019-229906)	(73)特許権者	000005108		
(22)出願日	令和1年12月20日(2019.12.20)		株式会社日立製作所		
(65)公開番号	特開2021-99560(P2021-99560A)		東京都千代田区丸の内一丁目6番6号		
(43)公開日	令和3年7月1日(2021.7.1)	(74)代理人	110001678		
審査請求日	令和4年6月21日(2022.6.21)		藤央弁理士法人		
		(72)発明者	達見 良介		
			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	倉田 成己		
			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		
		審査官	田名網 忠雄		
					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】

ホストからのI/O要求を処理するストレージシステムであって、
複数のコントローラと、
複数の記憶ドライブと、を含み、
前記ホストと前記複数のコントローラとは、ホストパスで接続され、
前記複数のコントローラと前記複数の記憶ドライブとは、ドライブパスで接続され、
論理ボリュームを担当するコントローラが、前記論理ボリュームを指定するI/O要求を処理し、
前記論理ボリュームを担当するコントローラと、前記論理ボリュームに対する複数のホストパスの優先度と、前記論理ボリュームにかかる物理記憶領域を含む記憶ドライブと、
に対して、前記ホストから前記記憶ドライブまでのホストパス及びドライブパスが決定され、
前記複数のコントローラは、
前記論理ボリュームの前記複数のホストパス及び複数のドライブパスの各パスの転送量を監視し、
前記複数のホストパスの優先度の変更後の前記ホストパス及び前記ドライブパスの変更を推定し、推定された前記ホストパス及び前記ドライブパスの変更と前記監視した各パスの転送量とに基づいて、前記優先度の変更後の前記複数のホストパス及び前記複数のドライブパスの各パスの転送量を推定し、

各パスの推定された前記転送量に基づいて、各パスの転送量が所定の条件を満たすように、前記複数のホストパスの優先度を変更する、ストレージシステム。

【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、

前記複数のコントローラは、

前記複数のホストパス及び前記複数のドライブパスの転送量を監視し、

前記複数のホストパス及び前記複数のドライブパスのいずれかのパスにおける転送量が所定の閾値を超える場合に、前記複数のホストパスの優先度を変更する、ストレージシステム。

【請求項 3】

請求項 2 に記載のストレージシステムであって、

前記複数のコントローラは、論理ボリュームとホストとの組み合わせそれぞれについて、前記複数のホストパス及び前記複数のドライブパスの転送量を監視する、ストレージシステム。

【請求項 4】

請求項 2 に記載のストレージシステムであって、

前記複数のコントローラは、前記複数のホストパスの優先度の変更後の、前記複数のホストパス及び前記複数のドライブパスの各パスの推定転送量が前記閾値以下となるように、前記複数のホストパスの優先度を変更する、ストレージシステム。

【請求項 5】

請求項 2 に記載のストレージシステムであって、

前記複数のコントローラは、前記複数のホストパス及び前記複数のドライブパスの各パスと対応する閾値との差異のばらつきが小さくなるように、前記複数のホストパスの優先度を変更する、ストレージシステム。

【請求項 6】

請求項 2 に記載のストレージシステムであって、

前記複数のコントローラは、複数世代の前記複数のホストパス及び前記複数のドライブパスの転送量の監視結果を保持し、

前記複数世代の加重平均と前記閾値とを比較する、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、

前記複数の記憶ドライブの少なくとも一部は、前記複数のコントローラの一部にのみ共有され、

前記複数のコントローラは、コントローラ間のデータ転送が発生しない条件において、前記複数のホストパスの優先度の変更後の、前記複数のホストパス及び前記複数のドライブパスの各パスの転送量を推定する、ストレージシステム。

【請求項 8】

請求項 2 に記載のストレージシステムであって、

前記複数のコントローラは、前記複数のホストパスの優先度の変更後の前記複数のホストパス及び前記複数のドライブパスの各パスの推定転送量が前記閾値以下となる条件において、前記論理ボリュームの前記複数のホストパスにおいて、前記論理ボリュームを担当するコントローラのホストパスの優先度を高くする、ストレージシステム。

【請求項 9】

請求項 1 に記載のストレージシステムであって、

前記複数の記憶ドライブそれぞれのドライブパスは、前記複数のコントローラにおける 1 又は複数の特定のコントローラと設定されている、ストレージシステム。

【請求項 10】

ホストからの I/O 要求を処理する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、複数のコントローラと、複数の記憶ドライブと、を含み、前記ホストと前記複数のコントローラとは、ホストパスで接続され、

10

20

30

40

50

前記複数のコントローラと前記複数の記憶ドライブとは、ドライブパスで接続され、前記制御方法は、

前記複数のコントローラにおいて論理ボリュームを担当するコントローラが、前記論理ボリュームを指定するI/O要求を処理し、前記論理ボリュームを担当するコントローラと、前記論理ボリュームに対するホストパスの優先度と、前記論理ボリュームにかかる物理記憶領域を含む記憶ドライブと、に対して、前記ホストから前記記憶ドライブまでのホストパス及びドライブパスが決定され、

前記複数のコントローラが、前記論理ボリュームの複数のホストパス及び複数のドライブパスの各パスの転送量を監視し、

前記複数のコントローラが、前記複数のホストパスの優先度の変更後の前記ホストパス及び前記ドライブパスの変更を推定し、推定された前記ホストパス及び前記ドライブパスの変更と前記監視した各パスの転送量とに基づいて、前記優先度の変更後の前記複数のホストパス及び前記複数のドライブパスの各パスの転送量を推定し、

前記複数のコントローラが、各パスの推定された前記転送量に基づいて、各パスの転送量が所定の条件を満たすように、前記複数のホストパスの優先度を変更する、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に関する。

【背景技術】

【0002】

一般的なストレージシステムは、複数のコントローラを含み、ホスト計算機から受け取った入出力（I/O）要求に従って、コントローラ内のプロセッサが、ドライブへのデータの入出力を制御する。コントローラとホスト計算機、コントローラ同士、コントローラとドライブとは、ネットワークで接続されている。

【0003】

I/O処理を効率よく実行するため、論理ボリューム毎にI/O処理を担当するプロセッサを割り当てるストレージシステムが存在する。このようなストレージシステムのいくつかは、複数のコントローラを含み、各コントローラにおいて、プロセッサが、ホストインタフェース（I/F）、ドライブとの接続のためのバックエンドI/F及び主記憶メモリと直結される。さらに、複数のコントローラは内部バスで接続される。

【0004】

このようなストレージシステムでは、物理的にどのプロセッサに接続されたバックエンドI/Fを使用するかによって、プロセッサのI/O処理手順やデータパスが変わり、データアクセスのレイテンシやスループットに影響を与える。例えば、リード要求を受けたホストI/Fを含むコントローラと別のコントローラのバックエンドI/Fを使用して、リード対象のデータをドライブから対象のコントローラの主記憶メモリに転送するとする。この場合、対象データを、リード要求を受けたホストI/Fを含むコントローラの主記憶メモリへ転送し、そこからホスト計算機へ転送する必要がある。便宜上、このようなストレージシステムのことをマルチコントローラストレージシステムと呼ぶ。

【0005】

特許文献1では、コントローラ間データ転送によるオーバーヘッドをなくすため、リード要求を受けたホストI/Fと同じコントローラのバックエンドI/Fを使用して、リード対象のデータを対象コントローラの主記憶メモリに転送する方法が開示されている。

【0006】

また、特許文献2では、マルチコントローラストレージシステムにおいて、コントローラ間通信性能の差異に応じて、ホストとストレージシステムとの間のパスの優先度を決定する方法が開示されている。

【先行技術文献】

【特許文献】

10

20

30

40

50

【0007】

【文献】国際公開第2015/056301号

国際公開第2016/013116号

【発明の概要】

【発明が解決しようとする課題】

【0008】

マルチコントローラストレージシステムでは、どのホストI/Fからアクセスがあったかによって、使用するバックエンドI/Fが異なる。一方で、コントローラ毎の負荷分散は、I/O処理を担当するプロセッサの稼働率（処理負荷）に基づき、各論理ボリュームの担当プロセッサを変更する。

10

【0009】

しかし、ストレージシステムにかかるワークロードはさまざまであり、プロセッサの処理負荷が大きくデータ量が小さいものや、逆に、プロセッサの処理負荷が小さくデータ量が大きいものが存在する。その結果、プロセッサの処理負荷によって、各論理ボリュームの担当プロセッサを変更しても、ホストI/FやバックエンドI/Fの処理負荷が均等になるとは限らない。そのため、一部のパスがボトルネックとなり、ストレージシステムの性能が低下し得る。

【課題を解決するための手段】

【0010】

本発明の一態様は、ホストからのI/O要求を処理するストレージシステムであって、複数のコントローラと、複数の記憶ドライブと、を含み、前記ホストと前記複数のコントローラとは、ホストパスで接続され、前記複数のコントローラと複数の記憶ドライブとは、ドライブパスで接続され、論理ボリュームを担当するコントローラが、前記論理ボリュームを指定するI/O要求を処理し、前記論理ボリュームを担当するコントローラと、前記論理ボリュームに対するホストパスの優先度と、前記論理ボリュームにかかる物理記憶領域を含む記憶ドライブと、に対して、前記ホストから前記記憶ドライブまでのホストパス及びドライブパスが決定され、前記複数のコントローラは、前記論理ボリュームにおける前記複数のホストパス及び複数のドライブパスの各パスの転送量を監視し、前記複数のホストパスの優先度の変更後の前記ホストパス及び前記ドライブパスの変更を推定し、推定された前記ホストパス及び前記ドライブパスの変更と前記監視した各パスの転送量とに基づいて、前記優先度の変更後の前記複数のホストパス及び複数のドライブパスの各パスの転送量を推定し、各パスの推定された前記転送量に基づいて、各パスの転送量が所定の条件を満たすように、前記複数のホストパスの優先度を変更する。

20

30

【発明の効果】

【0011】

本発明の一態様によれば、ストレージシステムの性能を向上することができる。

【0012】

上記した以外の課題、構成及び効果は、以下の実施形態の説明により明らかにされる。

【図面の簡単な説明】

【0013】

40

【図1】図1は、実施例1における計算機システム全体の構成例を示す図である。

【図2A】図2Aは、実施例1におけるメモリ内の領域構成を示す。

【図2B】図2Bは、実施例1におけるプログラム格納領域が格納している、プロセッサが実行するプログラムを示す。

【図2C】図2Cは、実施例1における制御情報格納領域132000に格納される制御情報の構成例を示す。

【図2D】図2Dは、実施例1におけるキャッシュメモリ領域の構成を示す。

【図2E】図2Eは、実施例1におけるデータ転送用バッファ領域の構成を示す。

【図3A】図3Aは、実施例1におけるドライブパスの転送量の平準化の効果の一例を示す図である。

50

【図 3 B】図 3 B は、実施例 1 におけるドライブパスの転送量の平準化の効果の一例を示す図である。

【図 4】図 4 は、実施例 1 における論理ボリュームアドレス管理テーブルを示す図である。

【図 5】図 5 は、実施例 1 におけるホストパス優先度管理テーブルを示す図である。

【図 6】図 6 は、実施例 1 におけるオーナー権管理テーブルを示す図である。

【図 7】図 7 は、実施例 1 におけるパス管理テーブルを示す図である。

【図 8】図 8 は、実施例 1 におけるパス転送量管理テーブルを示す図である。

【図 9】図 9 は、実施例 1 におけるパス単位時間転送量管理テーブルを示す図である。

【図 10】図 10 は、実施例 1 におけるパス単位時間総転送量管理テーブルを示す図である。

10

【図 11】図 11 は、実施例 1 における I/O コマンド管理テーブルを示す図である。

【図 12】図 12 は、実施例 1 における I/O コマンド処理要求メッセージを示す図である。

【図 13】図 13 は、実施例 1 におけるキャッシュセグメント管理テーブルを示す。

【図 14】図 14 は、実施例 1 におけるコマンド振り分け処理のフローチャートである。

【図 15 A】図 15 A は、実施例 1 における I/O 処理のフローチャートである。

【図 15 B】図 15 B は、実施例 1 における I/O 処理のフローチャートである。

【図 15 C】図 15 C は、実施例 1 における I/O 処理のフローチャートである。

【図 16】図 16 は、実施例 1 におけるパス転送量監視処理のフローチャートである。

【図 17】図 17 は、実施例 1 における優先ホストパス選択処理のフローチャートである。

20

【図 18】図 18 は、実施例 2 における計算機システム全体の構成例を示す図である。

【図 19 A】図 19 A は、実施例 2 におけるホストパス優先度切り替えたときの各パスのデータ転送量計算方式の一例を示す図である。

【図 19 B】図 19 B は、実施例 2 におけるホストパス優先度切り替えたときの各パスのデータ転送量計算方式の一例を示す図である。

【発明を実施するための形態】

【0014】

以下、本開示の実施例を、添付図面を参照しながら説明する。以下の説明では、「プログラム」を主語として処理を説明する場合があるが、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、適宜に記憶部及び／またはインタフェース部等を用いながら行うため、処理の主語が、プロセッサ（或いは、そのプロセッサを有するコントローラのようなデバイス）とされてもよい。

30

【0015】

プログラムは、プログラムソースから計算機のような装置にインストールされてもよい。プログラムソースは、例えば、プログラム配布サーバまたは計算機が読み取り可能な（例えば非一時的な）記録媒体であってもよい。また、以下の説明において、2 以上のプログラムが 1 つのプログラムとして実現されてもよいし、1 つのプログラムが 2 以上のプログラムとして実現されてもよい。

【0016】

以下の説明では、「x x x テーブル」といった表現にて、入力に対して出力が得られる情報を説明することがあるが、当該情報は、どのような構造のデータでもよい。また、以下の説明において、各テーブルの構成は一例であり、1 つのテーブルは、2 以上のテーブルに分割されてもよいし、2 以上のテーブルの全部または一部が 1 つのテーブルであってもよい。

40

【0017】

以下において、論理ボリュームを制御する複数のコントローラを含むストレージシステム（マルチコントローラストレージシステム）における、ストレージシステムとホストとの間のパス選択制御方法を説明する。ストレージシステムは、論理ボリューム毎に、I/O データがどの経路を通して転送されたかをモニタリングする。

【0018】

50

ストレージシステムは、各論理ボリュームに対するホストパス（フロントエンドパス）の優先度を変更した際に、I/Oデータが通るパスの変化により、各ホストパス及び各ドライブパス（バックエンドパス）の転送量がどのように変化するかを推定する。ストレージシステムは、転送量が推定された組み合わせの中から、ホストパス及びドライブパスの転送量が所定の条件を満たす組み合わせを求め、その組み合わせになるように、ストレージシステムとホストとの間において、ホストパスの優先度を変更する。

【0019】

上述のように、各論理ボリュームのI/Oアクセスパターンに応じて、ホスト及びドライブパスのデータ転送量の偏りを抑えて、I/Oスループットのボトルネックを解消する。

【実施例1】

【0020】

実施例1について、図面を参照して説明する。なお、以下に説明する実施例は特許請求の範囲にかかる発明を限定するものではなく、また実施例の中で説明されている特長の組み合わせの全てが発明の解決手段に必須であるとは限らない。

【0021】

図1は、実施例1における計算機システムの構成例を示す図である。計算機システムは、一つのノードからなるストレージシステム100000、ホスト計算機200000及び管理端末300000、並びに、それらを接続する外部ネットワーク400000及び管理ネットワーク500000を含む。図1は、一つのストレージシステム100000、一つのホスト計算機200000及び一つの管理端末300000を示すが、それらの数は任意である。

【0022】

ストレージシステム100000は、複数のストレージコントローラ（以下、コントローラとも呼ぶ）110000を含む。各コントローラ110000は、一つ以上のマイクロプロセッサ（以下、プロセッサとも呼ぶ）120000と、一つ以上のメモリ130000と、一つ以上のホストI/F（フロントエンドI/F）140000と、一つ以上のドライブI/F（バックエンドI/F）150000と、一つ以上の管理I/F160000とを含む。

【0023】

複数のコントローラ110000は、コントローラ間パス180000で接続されており、互いのメモリ130000に、プロセッサ120000または図示されていないDMA（Direct Memory Access）回路などによって、アクセス可能である。以下においては、説明の簡単のため、コントローラの数とは二つであるとして説明するが、コントローラの数が三つ以上でもよい。プロセッサがマルチコアプロセッサの場合、プロセッサ内のいくつかのコアをグループとして、論理的に複数のプロセッサとして管理してもよい。

【0024】

ホストI/F140000は、例えば、SAN（Storage Area Network）などの外部ネットワーク400000を通してホスト計算機200000に接続する。管理I/F160000は、例えば、LAN（Local Area Network）などの管理ネットワーク500000を通して管理端末300000に接続する。外部ネットワーク400000及び管理ネットワーク500000はデータ通信ができるプロトコルであれば任意である。

【0025】

ストレージシステム100000は、複数の記憶ドライブ（以下、ドライブまたは記憶デバイスとも呼ぶ）を格納するドライブボックス170000と、ドライブとコントローラ110000とを接続するドライブI/F150000とを含む。

【0026】

ドライブは、例えばHDD（Hard Disk Drive）、SSD（Solid State Drive）またはテープ型の記憶デバイス等であってもよい。複数のドラ

10

20

30

40

50

イブは、一つ以上のドライブの物理記憶領域を基に論理ボリュームを構成し、コントローラ110000に提供することができる。

【0027】

複数のドライブは、冗長化のためにRAID (Redundant Array of Independent Disks) グループを構成してもよく、論理ボリュームはRAIDグループから提供されてもよい。論理ボリュームは、ホスト計算機200000に対して、LU (Logical Unit) として提供可能であり、ホスト計算機200000が指定するアドレスに対して書き込み要求及び読み出し要求を受け付けることが可能である。なお、各論理ボリュームには、I/O処理を担当する一つのプロセッサ120000が設定される。

10

【0028】

ホストI/F140000を通してホストに接続されるパスをホストパス、ドライブI/F150000を通してドライブに接続されるパスをドライブパスと呼ぶ。ホスト計算機200000は、LUに対して複数のホストパスを介してアクセス可能であり、Active/Passiveや、ALUA (Asymmetric Logical Unit Access) など、パス毎に属性やアクセスの優先度が設定可能である。

【0029】

図2Aから図2Eは、メモリ130000の論理構成を示す図である。図2Aは、メモリ130000内の領域構成を示す。メモリ130000は、プログラム格納領域131000、制御情報格納領域132000、キャッシュメモリ領域133000、データ転送用バッファ領域134000及びキャッシュメモリ領域135000を含む。プロセッサ毎に使用できるキャッシュメモリ領域が確保されている。

20

【0030】

図2Bは、プログラム格納領域131000が格納している、プロセッサ120000が実行するプログラムを示す。コマンド振り分け処理プログラム131100は、ホスト計算機200000から送られてきたリード/ライト要求を、リード/ライト対象の論理ボリュームのI/O処理を担当するプロセッサに振り分ける。I/O処理を担当するプロセッサがI/O処理プログラム131200に従って、対象領域のリード/ライト処理を実行する。

【0031】

パス転送量監視プログラム131300は、ホストパス及びドライブパスの転送量をモニタし、集計する。優先ホストパス選択プログラム131400は、集計された結果から、ホストパスの優先度を変更したときの各ホストパス及び各ドライブパスの負荷予測(転送量予測)を行い、負荷の組み合わせが所定の条件を満たすように、各ホストパスの優先度を決定する。

30

【0032】

優先ホストパス切り替えプログラム131500は、優先ホストパス選択プログラム131400によって決定されたホストパスの優先度をホスト計算機200000に通知する。ホスト計算機200000は、通知された優先度に従って、自身のホストパスの優先度を再設定する。

40

【0033】

図2Cは、制御情報格納領域132000に格納される制御情報の構成例を示す。制御情報格納領域132000は、論理ボリュームアドレス管理テーブル132100、ホストパス優先度管理テーブル132200、オーナー権管理テーブル132300、パス管理テーブル132400、パス転送量管理テーブル132500、パス単位時間転送量管理テーブル132600、単位時間総転送量管理テーブル132700、I/Oコマンド管理テーブル132800及びキャッシュセグメント管理テーブル132900を格納している。これらのテーブルの詳細については後述する。

【0034】

図2Dは、キャッシュメモリ領域133000の構成を示しており、キャッシュセグメ

50

ント133100を含む。図2Eは、データ転送用バッファ領域134000の構成を示しており、データバッファ領域134100を含む。

【0035】

図3A及び図3Bは、本実施例に係るホストパスの優先度の変更によるホストパス及びドライブパスの転送量の平準化の例を示すものである。図3Aに示す状態において、ある論理ボリュームに対するコントローラ110000Bのホストパス優先度がActiveに設定されており、コントローラ110000Aのホストパス優先度がPassiveに設定されている。優先度がPassiveのホストパスは、優先度がActiveのホストパスが使用できない場合に使用される。ホスト計算機と論理ボリュームとの間において、使用されるホストパスに応じて使用されるドライブパスが決まる。

10

【0036】

コントローラ110000Aとコントローラ110000Bのドライブパス転送量の間に偏りが存在し、コントローラ110000Bのドライブパス転送量が、コントローラ110000Aのドライブパス転送量より多い。この時、コントローラ110000Bのドライブパスがボトルネックとなり、ストレージシステム全体の性能が低下する。

【0037】

図3Bは、いくつかの論理ボリュームのActiveパスが、コントローラ110000Bからコントローラ110000Aに変更され、I/O要求が図3Aと同様のパターンでストレージシステムに発行された状態を示す。コントローラ110000Aとコントローラ110000Bのドライブパス転送量が平準化され、ボトルネックが解消されている。これにより、ストレージシステム全体の性能を向上することができる。

20

【0038】

図4は、論理ボリュームアドレス管理テーブル132100の構成例を示す図である。論理ボリュームアドレス管理テーブル132100は、論理ボリューム番号132110、論理ボリューム内アドレス132120、ドライブ番号132130、ドライブ内アドレス132140を関連付ける。

【0039】

論理ボリューム番号132110は、論理ボリュームを識別するための番号である。論理ボリューム内アドレス132120は、論理ボリューム内のアドレスを識別するための番号である。ドライブ番号132130は、ドライブ番号を示す。ドライブ内アドレス132140は、ドライブ内のアドレスを示す。論理ボリュームアドレス管理テーブル132100を使用して、論理ボリューム及び論理ボリューム内アドレスから、アクセスするドライブ及びドライブ内のアドレスを算出することができる。

30

【0040】

図5は、ホストパス優先度管理テーブル132200の構成例を示す図である。ホストパス優先度管理テーブル132200は、論理ボリューム番号132210及び各ホストパスの優先度情報132220を関連付ける。論理ボリューム番号132210は、論理ボリュームを識別するための番号である。

【0041】

ホストパスの優先度情報132220は、ストレージシステムに実装されるホストパスそれぞれの優先度を示す。本実施例は、ホストパス優先度をActive及びPassiveで表現するが、SCSIのALUA規格では、ホストパス優先度は、Active/Optimized、Active/Non-Optimized、Standby、Unavailableの4種類で表現される。

40

【0042】

論理ボリューム、ホストパス及びそのホストパスの優先度の組み合わせに対して、ホストから論理ボリュームの物理記憶領域を含むドライブまでのホストパス及びドライブパスが決定される。また、論理ボリュームと、一つのホスト計算機の複数のホストパスの優先度の組み合わせから、それらホストパスそれぞれの当該論理ボリュームのためのデータ転送量の割合が推定可能である。このように、複数のホストパスの優先度の変更後のホスト

50

パス及びドライブパスの変更を推定し、監視しているホストパス及びドライブパスの転送量に基づき、それらホストパスそれぞれの当該論理ボリュームのためのデータ転送量の割合が推定される。

【0043】

なお、任意の名称の優先度及び任意の数の種類の優先度を使用することができる。また、各論理ボリュームに対して、ホストと論理ボリュームとの間において、同じ優先度の複数のホストパスが存在してもよく、ある優先度のパスが存在しなくてもよい。ホストパス優先度管理テーブル132200を使用して、論理ボリューム毎に各ホストパスの優先度を決定することができる。

【0044】

図6は、オーナ権管理テーブル132300の構成例を示す図である。オーナ権管理テーブル132300は、論理ボリューム番号132310及びオーナプロセッサ番号132320を関連付ける。論理ボリューム番号132310は、論理ボリュームを識別するための番号を示す。オーナプロセッサ番号132320は、オーナプロセッサ番号を示す。オーナ権管理テーブル132300を使って、論理ボリュームから、論理ボリュームのオーナプロセッサ番号を特定できる。

【0045】

図7は、パス管理テーブル132400の構成例を示す図である。パス管理テーブル132400は、ポート番号132410、コントローラ132420及び状態132430を関連付ける。ポート番号132410は、ホストパスまたはドライブパスを識別するための番号を示す。コントローラ132420は、上記ホストパスまたはドライブパスがどのコントローラに所属するかを示し、状態132430は、ホストパスまたはドライブパスが実装済み、未実装、または閉塞中のいずれであるかを示す。パス管理テーブル132400を使って、各ホストパス及び各ドライブパスがどのコントローラに所属しているか、利用可能な状態にあるかを調べることができる。

【0046】

図8は、パス転送量管理テーブル132500の構成例を示す図である。パス転送量管理テーブル132500は、ホスト計算機200000、論理ボリューム番号132510、各ホストパスの流入量132520及び流出量132530、並びに各ドライブパスの流入量132540及び流出量132550を関連付ける。

【0047】

この他、ホスト計算機200000からのI/Oに同期しないデータ転送によるデータ量をカウントするためテーブル（付図示）も存在する。ホスト計算機200000からのI/Oに同期しないデータ転送は、例えば、非同期のデステージ処理や、レプリケーションのためのデータコピー、ドライブ間の容量平準化のためのリバランス等のためのデータ転送を含む。

【0048】

論理ボリューム番号132510は、論理ボリュームを識別するための番号を示す。ホストパスの流入量132520は、ある時点から現在までにホストからコントローラに対して流れたデータの転送量を示す。ホストパスの流出量132530は、ある時点から現在までにコントローラからホスト計算機に対して流れたデータの転送量を示す。

【0049】

ドライブパスの流入量132540は、ある時点から現在までにドライブからコントローラに対して流れたデータの転送量を示す。ドライブパスの流出量132550は、ある時点から現在までにコントローラからドライブに対して流れたデータの転送量を示す。パス転送量管理テーブル132500を使用して、各ホスト計算機における論理ボリューム毎のある時点から現在までのホストパスデータ転送量及びドライブパスデータ転送量を適切に算出することができる。

【0050】

図9は、パス単位時間転送量管理テーブル132600の構成例を示す図である。パス

10

20

30

40

50

単位時間転送量管理テーブル132600は、論理ボリューム番号132610、各ホストパスの流入量132620及び流出量132630、並びに各ドライブパスの流入量132640及び流出量132650を関連付ける。

【0051】

論理ボリューム番号132610は、論理ボリュームを識別するための番号を示す。ホストパスの流入量132620は、ホストからコントローラに対して流れるデータの単位時間当たりの転送量を示す。ホストパスの流出量132630は、コントローラからホスト計算機に対して流れるデータの単位時間当たりの転送量を示す。ドライブパスの流入量132640は、ドライブからコントローラに対して流れるデータの単位時間当たりの転送量を示す。ドライブパスの流出量132650は、コントローラからドライブに対して流れるデータの単位時間当たりの転送量を示す。

10

【0052】

パス単位時間転送量管理テーブル132600を使用して、各ホスト計算機における論理ボリューム毎の単位時間当たりのホストパスデータ転送量及びドライブパスデータ転送量を算出することができる。パス単位時間転送量管理テーブル132600は、パス転送量管理テーブル132500のデータから生成される。パス単位時間転送量管理テーブル132600は、1世代のみの時間区間の単位時間転送量を格納してもよく、複数世代の時間区間の単位時間転送量を格納していてもよい。

【0053】

図10は、単位時間総転送量管理テーブル132700の構成を示す図である。単位時間総転送量管理テーブル132700は、ポートの種類132710、スループット132720及び過負荷閾値132730を関連付ける。ポートの種類132710は、ホストパス／ドライブパス、ポート番号、転送方向の組み合わせの種類を示す。過負荷閾値132740は、ポート種類ごとに、単位時間当たりの転送量が多く過負荷と判定される閾値を示す。この閾値は、管理端末300000を通して、管理者が設定してもよいし、ストレージシステム100000全体の構成情報や稼働情報に基づいて、自動的に設定されてもよい。

20

【0054】

図11は、I/Oコマンド管理テーブル132800の構成例を示す図である。I/Oコマンド管理テーブル132800は、I/Oコマンド管理番号132810、I/O処理中フラグ132820、I/Oコマンド受領ホストパス番号132830、I/Oコマンドパラメータ132840を関連付ける。I/Oコマンド管理テーブル132800は、各プロセッサのI/O処理実行の状態を示し、具体的には、各プロセッサ内で、どれだけのI/Oが処理されているか、どのI/O処理がどのFEポートで受信されたI/Oであるか、を管理している。

30

【0055】

図12は、I/Oコマンド処理要求メッセージ131110の構成例を示す図である。I/Oコマンド処理要求メッセージ131110は、I/Oコマンド管理番号131111を示し、コマンド振り分け処理プログラム131100を実行するプロセッサ120000とI/O処理プログラム131200を実行するプロセッサとの間で、通信するために使用される。これらの具体的な処理は、フロー図を使って後述する。

40

【0056】

図13は、キャッシュセグメント管理テーブル132900の構成例を示す図である。キャッシュセグメント管理テーブル132900は、キャッシュセグメント番号132910、論理ボリューム番号132920、論理ボリューム内セグメント番号132930及びキャッシュセグメント属性132940を関連付ける。

【0057】

キャッシュセグメント番号132910は、キャッシュメモリ領域133000上のキャッシュセグメント133100を識別するための値を示す。論理ボリューム番号132920及び論理ボリューム内セグメント番号132930は、キャッシュセグメント番号

50

132910にどの論理ボリュームのどのアドレスのデータが格納されているかを示す。論理ボリューム内セグメント番号132930は、論理ボリューム内のアドレスをキャッシュセグメント単位で区切った時の識別番号を示す。

【0058】

キャッシュセグメント属性132940は、キャッシュセグメント番号132910のキャッシュの状態を示す。クリーンは、キャッシュセグメント上のデータがドライブに格納済みの状態である。ダーティは、キャッシュセグメント上のデータがドライブに格納されていない状態である。フリーはキャッシュセグメントがどの論理ボリューム及び論理ボリューム内セグメントにも割り当てられていない状態である。

【0059】

キャッシュセグメント管理テーブル132900を使用して、どのキャッシュセグメントにどの論理ボリュームのどのアドレスのデータが格納されているか、そのデータがドライブに格納済みかどうか判定できる。ストレージシステム100000は、キャッシュセグメント管理テーブル132900の他に、論理ボリュームと論理ボリューム内セグメント番号のデータがキャッシュセグメント上に存在するかどうかを高速に検索するためのキャッシュディレクトリ等を持っていてもよい。

【0060】

図14は、コマンド振り分け処理プログラム131100により実行される処理のフローチャートである。まず、ホスト計算機200000からいずれかのコントローラに接続されているホストI/F140000を経由して、同じコントローラ上のメモリ130000にI/Oコマンドが登録される。ステップ131101で、同じコントローラ上のプロセッサ120000に実行されるコマンド振り分け処理プログラム131100は、I/Oコマンドが登録されたことを検出し、ステップ131102へ進む。

【0061】

ステップ131102で、コマンド振り分け処理プログラム131100は、オーナー権管理テーブル132300のI/Oオーナープロセッサ番号132320から、アクセス対象論理ボリュームのI/O処理担当プロセッサを同定し、ステップ131103へ進む。

【0062】

ステップ131103で、コマンド振り分け処理プログラム131100は、上記I/O処理担当プロセッサと同一のコントローラのメモリ上のI/Oコマンド管理テーブル132800において、未使用のI/Oコマンド管理番号132810を検索する。各コントローラの構成は、不図示の管理情報に格納されている。コマンド振り分け処理プログラム131100は、検索したI/Oコマンド管理番号に対応するI/O処理中フラグ132820をONに書き換え、I/Oコマンド受領ホストパス番号132830とI/Oコマンドパラメータ132840を更新し、ステップ131104へ進む。

【0063】

ステップ131104で、コマンド振り分け処理プログラム131100は、上記I/O処理担当プロセッサと同じコントローラのメモリに上記I/Oコマンド管理番号を記したI/Oコマンド処理要求メッセージ131110を送信し、I/Oコマンド振り分け処理を終了する。

【0064】

図15A、図15B、図15Cは、I/O処理プログラム131200により実行される処理のフローチャートであり、図15Aは、ホストI/O要求処理の前半部分、図15Bはリード処理、図15Cはライト処理を示す。

【0065】

上記のI/Oコマンド処理要求メッセージ131110を受信したI/O処理担当プロセッサにより実行されるI/O処理プログラム131200は、ステップ131201で上記のコマンド振り分け処理プログラム131100により振り分けられたI/Oコマンドを検出する。

【0066】

10

20

30

40

50

ステップ131202では、I/O処理プログラム131200は、上記I/Oコマンドから、要求処理がリード処理なのかライト処理なのかを判定する。要求処理がリード処理の場合(131202: YES)、I/O処理プログラム131200はステップ131210へ進み、要求処理がライト処理の場合(131202: NO)、I/O処理プログラム131200はステップ131220へ進む。

【0067】

リード処理131210から説明する。ステップ131211で、I/O処理プログラム131200は、アクセス対象のデータがキャッシュセグメント管理テーブル132900または上記のキャッシュディレクトリによって、キャッシュセグメント上に存在するかどうか(キャッシュヒットするかどうか)を判定する。対象のデータがキャッシュセグメント上に存在しない場合(131211: NO)、I/O処理プログラム131200は、ステップ131212へ進む。対象のデータがキャッシュセグメント上に存在する場合(131211: YES)、I/O処理プログラム131200は、ステップ131216へ進む。

10

【0068】

ステップ131212で、I/O処理プログラム131200は、論理ボリュームアドレス管理テーブル132100を参照し、アクセス対象の論理ボリューム番号132110及び論理ボリューム内アドレス132120に対応する、ドライブ番号132130及びドライブ内アドレス132140を特定し、ステップ131213へ進む。

【0069】

20

ステップ131213で、I/O処理プログラム131200は、ドライブ内のデータをコントローラのメモリ130000へ転送するためのドライブパス番号を取得する。その際、コントローラ間のデータ転送が発生しないように、I/O処理プログラム131200は、I/Oコマンド管理テーブル133800からI/Oを受領したホストパス番号132830を取得し、パス管理テーブル132400からI/O受領したコントローラと同じコントローラのドライブパス番号(ポート番号)132410を選択する。複数のドライブパスがある場合、I/O処理プログラム131200は、ラウンドロビンで選択する等、ドライブパスの負荷が偏らないように選択する。

【0070】

次に、I/O処理プログラム131200は、ステップ131214へ進む。ステップ131214で、I/O処理プログラム131200は、ドライブからデータをステージングするために、選択されたドライブパス番号(ポート番号)132410が所属するコントローラ内のデータ転送用バッファ領域134000においてデータバッファ領域134100を確保する。I/O処理プログラム131200は、上記のドライブ番号132130及びドライブ内アドレス132140のデータを、確保したデータバッファ領域134100に読み出し、ステップ131215へ進む。

30

【0071】

ステップ131215で、I/O処理プログラム131200は、パス転送量管理テーブル132500のアクセスした論理ボリューム番号132110に対応するドライブパスINPUT転送量132540をステージングのために転送したデータ量分だけ加算する。I/O処理プログラム131200は、ステップ131216へ進む。

40

【0072】

ステップ131216で、I/O処理プログラム131200は、上記のデータバッファ領域134100からホストI/F140000を使用して、対象データをホスト計算機200000へ転送し、ステップ131217へ進む。

【0073】

ステップ131217で、I/O処理プログラム131200は、ホスト計算機200000へ転送したデータ分の転送量を、パス転送量管理テーブル132500に加算する。具体的には、I/O処理プログラム131200は、当該論理ボリューム番号を論理ボリューム番号132510から検索し、対応するホストパス番号のOUTPUT転送量13

50

2530を加算するし、リード処理を終了する。

【0074】

ステップ131218では、I/O処理プログラム131200は、キャッシュヒットしたキャッシュセグメント番号132910が、ホスト計算機200000からI/Oコマンドを受領したコントローラと同じメモリ上に存在するかを判定する。同じメモリ上に存在するときには、I/O処理プログラム131200は、ステップ131216へ進む。ステップ131216で、I/O処理プログラム131200は、データバッファ領域134100の代わりにキャッシュセグメント133100を使用して、データをホストへ転送する。

【0075】

キャッシュヒットしたキャッシュセグメント番号132910が、ホスト計算機200000からI/Oコマンドを受領したコントローラと同じメモリ上に存在しないとき、I/O処理プログラム131200は、ステップ131219へ進む。ステップ131219で、I/O処理プログラム131200は、ホスト計算機200000からI/Oコマンドを受領したコントローラ内のデータ転送用バッファ領域134000においてデータバッファ領域134100を確保し、キャッシュセグメントから、確保したデータバッファへデータを転送する。その後、フローは、ステップ131216へ進む。

【0076】

次にライト処理について説明する。ステップ131221で、I/O処理プログラム131200は、I/Oコマンドを受領したホストパスと同じコントローラのメモリ130000上に、ホスト計算機200000からのライトデータ格納用のデータバッファ領域134100を確保し、ステップ131222へ進む。

【0077】

ステップ131222で、I/O処理プログラム131200は、ホスト計算機200000にデータ転送を要求し、上記のデータバッファ領域134100にライトデータを転送し、ステップ131223へ進む。

【0078】

ステップ131223で、I/O処理プログラム131200は、ホスト計算機200000から転送したデータ分の転送量をパス転送量管理テーブル132500に加算する。具体的には、I/O処理プログラム131200は、当該論ボリューム番号を論理ボリューム番号132510から検索し、対応するホストパス番号のINPUT転送量132520を加算する。次にステップ131224へ進む。

【0079】

ステップ131224で、I/O処理プログラム131200は、ライトデータをドライブに格納するまでの間に、コントローラ間のキャッシュメモリ領域133000でライトデータを2重化してデータが保護された状態にするため、キャッシュセグメント133100を2つのコントローラそれぞれに確保する。I/O処理プログラム131200は、キャッシュセグメント管理テーブル132900の中からキャッシュセグメント属性132940がフリーなレコードを選択し、ライト対象の論理ボリューム番号132920及び論理ボリューム内セグメント番号132930を登録する。次に、I/O処理プログラム131200は、ステップ131225へ進む。

【0080】

ステップ131225で、I/O処理プログラム131200は、上記ライトデータが格納されたデータバッファ領域134100から、上記確保した二つのキャッシュセグメント133100へデータを転送する。ドライブヘデステージする前のデータをダーティデータと呼ぶ。次に、I/O処理プログラム131200は、ステップ131226へ進む。

【0081】

ステップ131226で、I/O処理プログラム131200は、キャッシュセグメント管理テーブル132900の当該セグメント番号に対応するスロットステータスをダー

10

20

30

40

50

ティに更新し、ステップ131227へ進む。ステップ131227で、I/O処理プログラム131200は、上記データバッファ領域134100を解放する。

【0082】

その後、I/O処理プログラム131200は、ステップ131228へ進み、ホスト計算機200000へGoodステータスを送信し、ライト処理を完了する。キャッシュセグメント上のダーティデータは、I/O処理とは非同期に、ドライブヘデステージされる。データは、RAIDグループを構成する複数のドライブに、冗長性を持つように格納されてもよい。

【0083】

図16は、パス転送量監視プログラム131300により実行される処理のフローチャートである。パス転送量監視プログラム131300は、所定周期で起動し、各ホストパス及び各ドライブパスの単位時間当たりの転送量を計算する。

【0084】

ステップ131301で、パス転送量監視プログラム131300は、論理ボリュームに対する繰り返し判定を行う。パス転送量監視プログラム131300は、ストレージシステム100000に実装されている全ての論理ボリュームに対してステップ131302からステップ131309が実行されたか判定する。実行済みであれば、パス転送量監視プログラム131300は、繰り返し処理を抜けて、ステップ131310に進む。未実行の論理ボリュームが1つ以上存在すれば、パス転送量監視プログラム131300は、未実行の論理ボリュームを1つ選択し、ステップ131302に進む。

【0085】

ステップ131302で、パス転送量監視プログラム131300は、ホストパスに対する繰り返し判定を行う。パス転送量監視プログラム131300は、ストレージシステム100000に実装されている全てのホストパスに対して、ステップ131303からステップ131305が実行されたか判定する。実行済みであれば、パス転送量監視プログラム131300は、繰り返し処理を抜けて、ステップ131306に進む。未実行のホストパスが1つ以上存在すれば、パス転送量監視プログラム131300は、未実行のホストパスを1つ選択し、ステップ131303に進む。

【0086】

ステップ131303では、パス転送量監視プログラム131300は、パス転送量管理テーブル132500から選択した論理ボリュームと選択したホストパスに対応する、INPUT及びOUTPUTデータ転送量を、ホスト計算機(ホスト計算機なしも含む)毎に取得する。次に、ステップ131304に進む。

【0087】

ステップ131304で、パス転送量監視プログラム131300は、取得したINPUT及びOUTPUTデータ転送量を、パス転送量監視プログラム131300の実行周期で割り算して、単位時間当たりのINPUT及びOUTPUT転送量をホスト毎に算出する。パス転送量監視プログラム131300は、計算結果をパス単位時間転送量管理テーブル132600に書き込む。次に、パス転送量監視プログラム131300は、ステップ131305に進む。

【0088】

ステップ131305で、パス転送量監視プログラム131300は、パス転送量管理テーブル132500の対象の論理ボリューム及びホストパスに対応するINPUT及びOUTPUT転送量をサーバ(サーバなしも含む)毎にクリアし、ステップ131302に戻る。

【0089】

ステップ131306からステップ131309において、パス転送量監視プログラム131300は、上記のステップ131302からステップ131305で、ホストパスに対して実行した処理と同様の処理をドライブパスに対して実行する。全てのドライブパスに対する処理が終了したら、パス転送量監視プログラム131300は、ステップ13

10

20

30

40

50

1301に戻る。

【0090】

ステップ131310で、バス転送量監視プログラム131300は、各ホストパス、各ドライブパスに対して、全ての論理ボリューム及びホスト計算機(ホスト計算機なしも含む)の単位時間当たりのINPUT及びOUTPUT転送量をそれぞれ合計し、単位時間総転送量管理テーブル132700のスループット132720に値を書き込む。

【0091】

バス転送量監視プログラム131300によって、各ホスト計算機がどの論理ボリュームにどのホストパス及びドライブパスを使用して、どのくらいアクセスしているかと、各ホストパス及び各ドライブパスの単位時間当たりの総転送量がわかる。

【0092】

図17は、優先ホストパス選択プログラム131400により実行される処理のフローチャートである。優先ホストパス選択プログラム131400は、所定周期で起動し、各論理ボリュームに対するホストパス及びドライブパスの負荷状況を判定し、過負荷なパスが存在しなくなるように、ホスト毎にホストパスの優先度を決定する。

【0093】

ステップ131401で、優先ホストパス選択プログラム131400は、単位時間総転送量管理テーブル132700において、各ホストパス及び各ドライブパスのスループットを参照し、ステップ131402へ進む。

【0094】

ステップ131402で、優先ホストパス選択プログラム131400は、ホスト及びドライブパスそれぞれの転送量(スループット)が、対応する過負荷閾値132730を超えていないか判定する。過負荷閾値を超えているパスがあれば(131402: YES)、優先ホストパス選択プログラム131400は、ステップ131403に進む。超えているパスがなければ(131402: NO)、優先ホストパス選択プログラム131400は、処理を終了する。転送量が過負荷閾値を超えるパスが存在する場合にホストパス優先度の変更を行うことで、ボトルネックの発生によるストレージシステムの性能低下を適切に避けることができる。

【0095】

ステップ131403で、優先ホストパス選択プログラム131400は、バス単位時間転送量管理テーブル132600を参照し、どのホスト計算機のどの論理ボリュームのホストパスの属性を変更すると、閾値を超えなくなるかを算出する。このとき、ホストパス優先度管理テーブル132200で、未定義状態や使用不可状態のパスは使用されない。

【0096】

例えば、ホスト計算機番号1と論理ボリューム番号0の組み合わせのパス優先度設定において、ホストパス番号0の優先度がActiveであり、ホストパス番号1の優先度がPassiveとする。ホストパス番号0とドライブパス番号0の組み合わせ及びホストパス番号1とドライブパス番号1の組み合わせが、それぞれ、ホスト計算機番号1と論理ボリューム番号0との間の一つのパスを構成する。

【0097】

バス単位時間転送量管理テーブル132600が、ホストパス番号0の流入量100MB/s、ホストパス番号0の流出量300MB/s、ドライブパス番号0の流入量が0MB/s、ドライブパス番号0の流出量300MB/sを示しているものとする。ホストパス番号1及びドライブパス番号1の論理ボリューム番号0に対する転送量は0である。

【0098】

ActiveパスとPassiveパスを、ホストパス番号0とホストパス番号1の間で入れ替えたとき、各パスの転送量の変化は、単位時間総転送量管理テーブル132700の転送量に対して、以下のように計算できる。ホストパス番号0の流入量から100MB/sをマイナスし、ホストパス番号0の流出量から300MB/sをマイナスし、ドライブパス番号0の流出量から300MB/sをマイナスする。さらに、ホストパス番号1

10

20

30

40

50

の流入量に100MB/sをプラスし、ホストパス番号1の流出量に300MB/sをプラスし、ドライブパス番号1の流出量に300MB/sをプラスする。

【0099】

優先ホストパス選択プログラム131400は、例えば、過負荷閾値を超えるホストパス及びドライブパスを順次選択し、ホストパス優先度の変更により他のパスの転送量が過負荷閾値を超えるように変化しない範囲で、選択した各パスの転送量が過負荷閾値以下又は最も小さくなるように、対応するホストパス優先度を変更する。

【0100】

パス間の平準化をより進めるために、優先ホストパス選択プログラム131400は、例えば、各パスの過負荷閾値に対する単位時間当たりの転送量の割合のばらつきが、最も小さくなるものを求めてもよい。算出方法は、全ての組み合わせパターンを検索してもよいし、組み合わせ最適化問題として数理的な解法を用いてもよい。また、過負荷閾値を超えないような組み合わせが存在しない場合、優先ホストパス選択プログラム131400は、各パスの過負荷閾値に対する単位時間当たりの転送量の割合が、最も小さくなるものを選択してもよい。

【0101】

優先ホストパス選択プログラム131400は、論理ボリュームの複数のホストパスにおいて、オーナー権を有するコントローラのホストパスの優先的に選択してもよい。これにより、ストレージシステムの負荷を低減できる。例えば、いずれのパスの転送量も過負荷閾値以下でとなる条件において、ホスト計算機と論理ボリュームの組み合わせに対する複数のホストパスを選択できる場合、オーナー権を有するコントローラのホストパスを選択する。

【0102】

優先ホストパス選択プログラム131400は、選択したホストパス優先度の組み合わせをホストパス優先度管理テーブル132200に格納する。各単位時間転送量の監視結果を複数世代分（複数周期分）格納している場合には、予め設定された重みを使用して世代間で加重平均を取った値と過負荷閾値を比較することで、ホストパスの優先度が切り替わる頻度が上がり、切り替えによる処理コスト増加で性能が不安定になることを避けることができる。ホストパスの優先度の決め方は、ここに記した最適化方式に限定されない。

【0103】

次に、優先ホストパス選択プログラム131400は、ステップ131404へ進む。ステップ131404で、優先ホストパス選択プログラム131400は、ステップ131403で更新したホストパス優先度を各ホスト計算機200000に通知し、処理を終了する。

【0104】

ホストパス優先度の変更頻度（転送量が過負荷閾値を超える頻度）が閾値を超える場合、優先ホストパス選択プログラム131400の起動周期を所定時間だけ長くして、ホストパス優先度の変更頻度を低くしてもよい。これにより、優先度変更による処理負荷の増加を低減できる。なお、ホストパス優先度の変更は、何れかのパスの転送量が閾値を超えることと異なる条件に応じて（例えば所定周期で）実行されてもよい。

【0105】

以上のように、本実施例は、ホストパス及びドライブパスの負荷に応じて、ホストパスの優先度の変更によるホストパスとドライブパスの転送量を推定することで、ホストパスの優先度を適切に変更することができる。

【実施例2】

【0106】

実施例2について説明する。以下においては、実施例1との差分について主に説明し、実施例1と共通の部分の不要な説明を省略する。図18は、実施例2における計算機システムの構成例を示す図である。実施例2のストレージシステムは、複数のノードを含み、コントローラ間を接続するコントローラ間ネットワーク600000及びコントローラエ

10

20

30

40

50

／F190000を含み、そのコントローラの接続数が拡張されている。ドライブボックスは、一部のコントローラでのみ共有されている。

【0107】

図18は、例として、四つのコントローラ110001～110004を示す。コントローラ110001及び110002は一つのノードに含まれ、コントローラ110003及び110004は他のノードに含まれる。ドライブボックス170001はコントローラ110001、110002に共有され、ドライブボックス170002はコントローラ110003、110004に共有されている。

【0108】

この構成では、ドライブパスが直接接続されていないコントローラのメモリヘドライブからデータを転送する場合、ドライブが接続されたコントローラのメモリヘデータを転送したのち、コントローラI／F190000及びコントローラ間ネットワーク60000を通してデータを転送する必要がある。

【0109】

このシステムは、使用するホストパスを切り替えた先でコントローラ間データ転送が必要になるかならないかに応じて、選択されるドライブパスが変化させる。そのため、ホストパス優先度を切り替えたときに、各パスの流入量及び流出量がどのように変化するかの計算法が、実施例1と異なる。これにより、ホストパスの優先度の変更によるホストパスとドライブパスの転送量を適切に推定できる。

【0110】

図19A及び19Bは、実施例2において、ホストパス優先度を変更したときの、データ転送量変化予測のための計算法の一例を示した図である。1台のホスト計算機200000に対して、6台のコントローラ110001～110006が接続されている。ドライブボックス170001～170003は、それぞれ、2台のコントローラに共有されている。

【0111】

図19Aに、ある論理ボリュームのホストパス優先度を切り替える前の、現在の各パスの転送量（スループット）を示す。コントローラ1（110001）のホストパスがActiveとなっており、コントローラ4（110004）のホストパスがPassiveとなっている。

【0112】

コントローラ1（110001）に接続されているドライブボックス（ドライブ）からは、コントローラ1（110001）を通して直接ホスト計算機200000にデータ転送できる。そのため、コントローラ1（110001）とドライブボックス170001を共有するコントローラ2（110002）のドライブパスの転送量は0MB／sとなっている。

【0113】

その他のコントローラに接続されているドライブボックスからホスト計算機200000へのデータ転送は、接続されているどちらのコントローラを使用しても、必ずコントローラ間でデータ転送を必要とする。そのため、図19A及び19Bの例において、一つのドライブボックスの二つのドライブパスの転送量は等しい。なお、コントローラ間でデータ転送を必要とする同一ドライブボックスのドライブパスの転送量は同一とは限らず、コントローラ間のパス選択アルゴリズムやコントローラの状態に依存する。

【0114】

図19Bでは、優先ホストパス選択プログラム131400が、ステップ131403においてホストパスの優先度を選択する際に、ActiveパスとPassiveパスを切り替えた場合の各パスの転送量の変化の計算法を示す。図19Aの状態から、コントローラ4（110004）のパスをActiveに、コントローラ1（110001）のパスをPassiveに変更したとする。

【0115】

10

20

30

40

50

A c t i v eに変更されたコントローラ4 (110004) のホストパスの転送量は、変更前のコントローラ1 (110001) のホストパスと同様に、1000MB/sと計算される。コントローラ4 (110004) に接続されているドライブボックス170002からのデータはコントローラ4 (110004) を通って直接ホスト計算機200000に転送される。そのため、コントローラ4 (110004) のドライブパスの転送量は200MB/s、コントローラ3 (110003) のドライブパスの転送量は0MB/sと計算される。

【0116】

コントローラ1 (110001) 及びコントローラ2 (110002) については、どちらのドライブパスを使用しても、コントローラ間のデータ転送が必要となるため、均等に使用されると仮定して、どちらも250MB/sと計算される。コントローラ5 (110005) 及びコントローラ6 (110006) については、パスの優先度変更前と状況が変わっていないため、そのまま変化がないとして計算される。

【0117】

このようにして、切り替え後の各パスの転送量を計算し、切り替え前の値との差分を単位時間総転送量管理テーブルのスループットと足し合わせることで、各パスの総転送量の変化を得ることができる。優先ホストパス選択プログラム131400は、これを各ホスト計算機と論理ボリュームの組み合わせに対して実行し、過負荷閾値を超えないようなホストパス優先度の組み合わせを検索する。

【0118】

なお、本発明は上記した実施例に限定されるものではなく、様々な変形例が含まれる。例えば、上記した実施例は本発明を分かりやすく説明するために詳細に説明したものであり、必ずしも説明したすべての構成を備えるものに限定されるものではない。また、ある実施例の構成の一部を他の実施例の構成に置き換えることが可能であり、また、ある実施例の構成に他の実施例の構成を加えることも可能である。また、各実施例の構成の一部について、他の構成の追加・削除・置換をすることが可能である。

【0119】

また、上記の各構成・機能・処理部等は、それらの一部又は全部を、例えば集積回路で設計する等によりハードウェアで実現してもよい。また、上記の各構成、機能等は、プロセッサがそれぞれの機能を実現するプログラムを解釈し、実行することによりソフトウェアで実現してもよい。各機能を実現するプログラム、テーブル、ファイル等の情報は、メモリや、ハードディスク、SSD (Solid State Drive) 等の記録装置、または、ICカード、SDカード等の記録媒体に置くことができる。

【0120】

また、制御線や情報線は説明上必要と考えられるものを示しており、製品上必ずしもすべての制御線や情報線を示しているとは限らない。実際には殆どすべての構成が相互に接続されていると考えてもよい。

【符号の説明】

【0121】

100000 ストレージシステム
110000-110006 コントローラ
120000 プロセッサ
130000 メモリ
131000 プログラム格納領域
131100 コマンド振り分け処理プログラム
131110 コマンド処理要求メッセージ
131200 I/O処理プログラム
131300 パス転送量監視プログラム
131400 優先ホストパス選択プログラム
131500 優先ホストパス切り替えプログラム

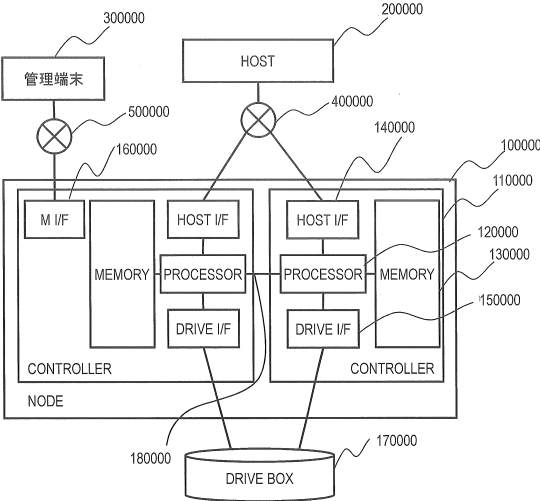
- 1 3 2 0 0 0 制御情報格納領域
- 1 3 2 1 0 0 論理ボリュームアドレス管理テーブル
- 1 3 2 2 0 0 ホストバス優先度管理テーブル
- 1 3 2 3 0 0 オーナ権管理テーブル
- 1 3 2 4 0 0 パス管理テーブル
- 1 3 2 5 0 0 パス転送量管理テーブル
- 1 3 2 6 0 0 パス単位時間転送量管理テーブル
- 1 3 2 7 0 0 単位時間総転送量管理テーブル
- 1 3 2 8 0 0 コマンド管理テーブル
- 1 3 2 9 0 0 キャッシュセグメント管理テーブル
- 1 3 3 0 0 0 キャッシュメモリ領域
- 1 3 4 0 0 0 データ転送用バッファ領域
- 1 3 4 1 0 0 データバッファ領域
- 1 3 5 0 0 0 キャッシュメモリ領域
- 1 4 0 0 0 0 ホスト I / F
- 1 5 0 0 0 0 ドライブ I / F
- 1 6 0 0 0 0 管理 I / F
- 1 7 0 0 0 0 - 1 7 0 0 0 3 ドライブボックス
- 1 8 0 0 0 0 コントローラ間バス
- 1 9 0 0 0 0 コントローラ I / F
- 2 0 0 0 0 0 ホスト計算機
- 3 0 0 0 0 0 管理端末
- 4 0 0 0 0 0 外部ネットワーク
- 5 0 0 0 0 0 管理ネットワーク
- 6 0 0 0 0 0 コントローラ間ネットワーク

10

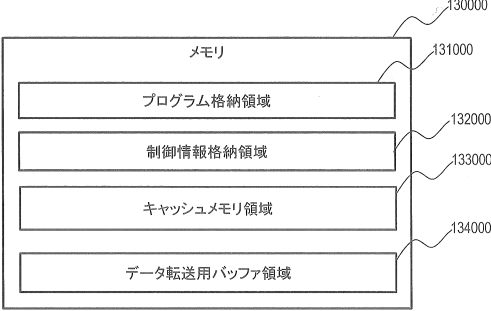
20

【図面】

【図 1】



【図 2 A】

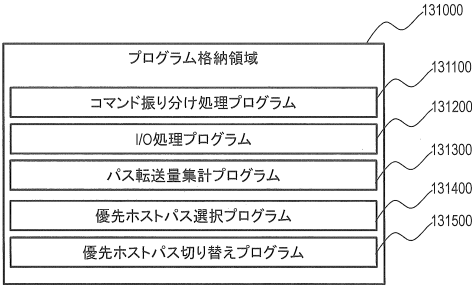


30

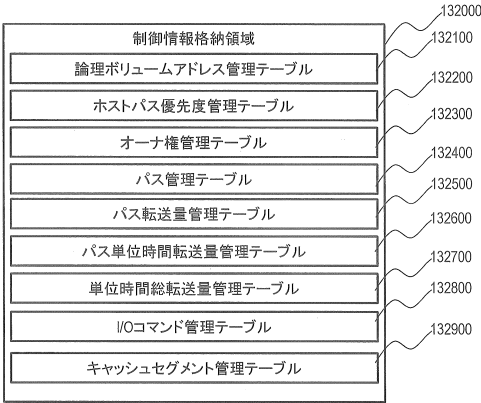
40

50

【図 2 B】

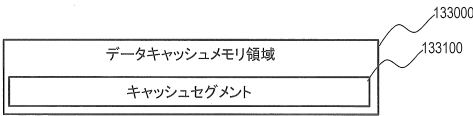


【図 2 C】

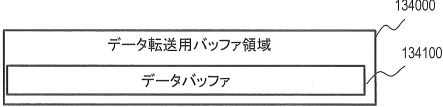


10

【図 2 D】

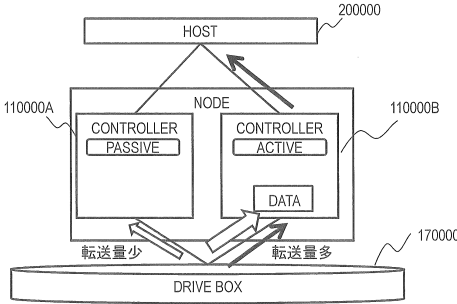


【図 2 E】

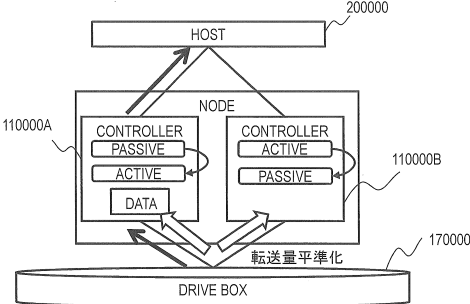


20

【図 3 A】



【図 3 B】



30

40

50

【図 4】

132110 132120 132130 132140 132100

論理ボリューム番号	論理ボリューム内アドレス	ドライブ番号	ドライブ内アドレス
00	0x0000	00	0x0400
00	0x0100	01	0x0200

論理ボリュームアドレス管理テーブル

【図 5】

132200 132210 132220

論理ボリューム番号	ホストバス #0	ホストバス #1	ホストバス #2	ホストバス #3	...
00	A	P	N	N	
01	N	A	P	P	

A: ACTIVE, P: PASSIVE, N: 未定義

ホストバス優先度管理テーブル

【図 6】

132300 132310 132320

論理ボリューム番号	オーナープロセッサ番号
00	0
01	1

オーナー権管理テーブル

【図 7】

132400 132410 132420 132430

ポート番号	コントローラ	状態
ホストバス#0	0	実装
ホストバス#1	1	実装
ホストバス#2	0	未実装
⋮	⋮	⋮
ドライブバス#0	0	実装
ドライブバス#1	1	実装
ドライブバス#2	0	閉塞
⋮	⋮	⋮

バス管理テーブル

【図 8】

132500 132510 132520 132530 132540 132550

論理ボリューム番号	ホストバス#0 INPUT	ホストバス#0 OUTPUT	...	ドライブバス#0 INPUT	ドライブバス#0 OUTPUT	...
00	1.0GB	0.3GB	...	0.3GB	0.0GB	...
01	0.0GB	1.2GB	...	1.2GB	0.0GB	...

バス転送量管理テーブル

【図 9】

132600 132610 132620 132630 132640 132650

論理ボリューム番号	ホストバス#0 INPUT	ホストバス#0 OUTPUT	...	ドライブバス#0 INPUT	ドライブバス#0 OUTPUT	...
00	1.0GB/s	0.3GB/s	...	0.3GB/s	0.0GB/s	...
01	0.0GB/s	1.2GB/s	...	1.2GB/s	0.1GB/s	...

バス単位時間転送量管理テーブル

10

20

30

40

50

【図 1 0】

132700		
132710	132720	132730
ポート種類	スループット	過負荷閾値
ホストバス#0 INPUT	1.5GB/s	1.5GB/s
ホストバス#0 OUTPUT	0.8GB/s	1.5GB/s
ホストバス#1 INPUT	1.2GB/s	3.0GB/s
ホストバス#1 OUTPUT	1.0GB/s	3.0GB/s
⋮	⋮	⋮
ドライブバス#0 INPUT	1.5GB/s	1.5GB/s
ドライブバス#0 OUTPUT	0.8GB/s	1.5GB/s
ドライブバス#1 INPUT	1.2GB/s	3.0GB/s
ドライブバス#1 OUTPUT	1.0GB/s	3.0GB/s
⋮	⋮	⋮

単位時間総転送量管理テーブル

【図 1 1】

132800			
132810	132820	132830	132840
I/Oコマンド 管理番号	I/O処理中フラグ	I/Oコマンド受領ホストバス 番号	I/Oコマンド パラメータ
1	OFF	NULL	NULL
2	ON	1	ZZZ
3	ON	1	YYY
4	ON	2	XXX

I/Oコマンド管理情報

10

【図 1 2】

131110	
131111	I/Oコマンド管理番号
	2

I/Oコマンド処理要求メッセージ

【図 1 3】

132900			
132910	132920	132930	132940
キャッシュセグメント 番号	論理ボリューム 番号	論理ボリューム内 セグメント番号	キャッシュセグメント 属性
1	なし	なし	フリー
2	2	1	クリーン
3	1	1	クリーン
4	3	4	ダーティ
5	4	5	ダーティ

キャッシュセグメント管理テーブル

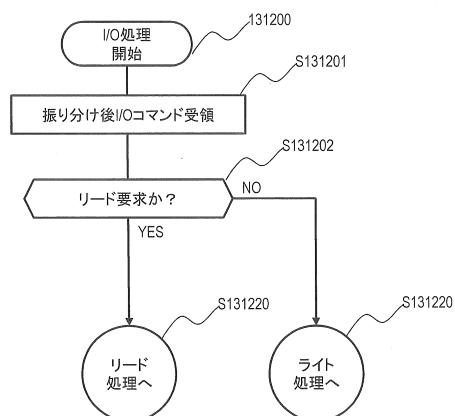
20

30

40

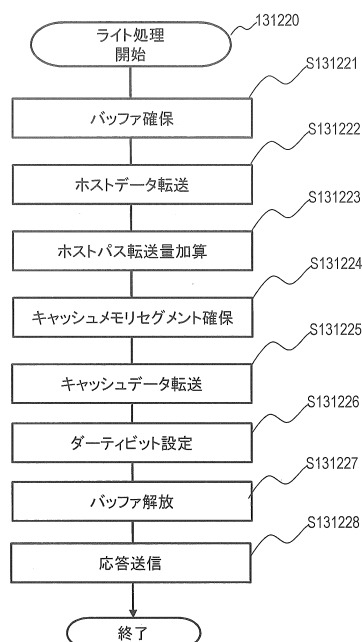
50

【図 15 A】



10

【図 15 C】



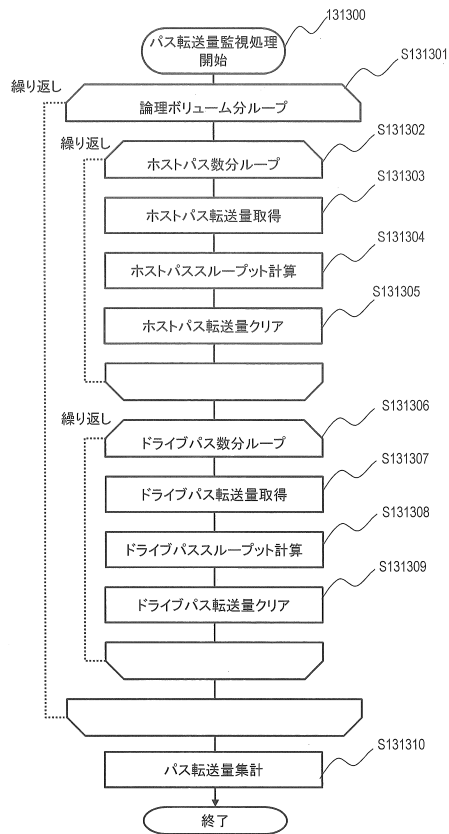
20

30

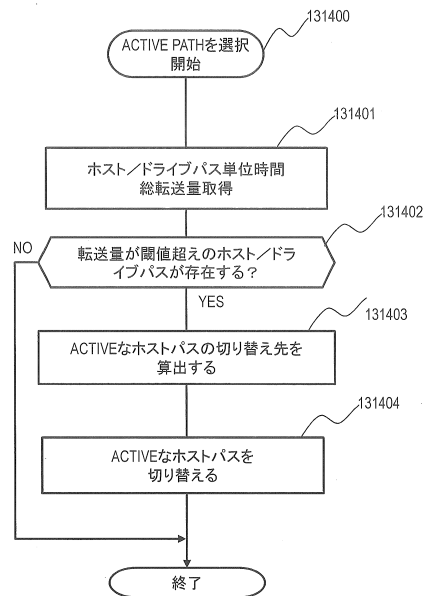
40

50

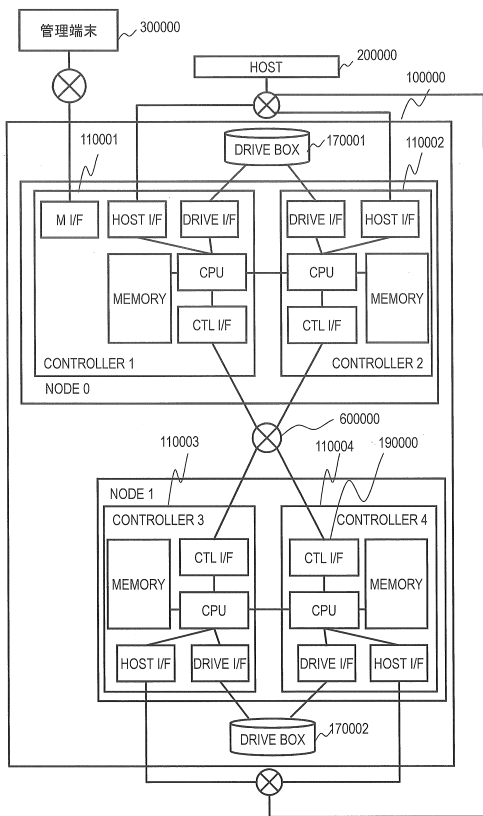
【図 16】



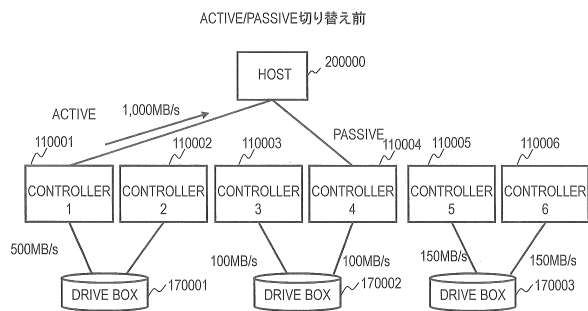
【図 17】



【図 18】



【図 19 A】



10

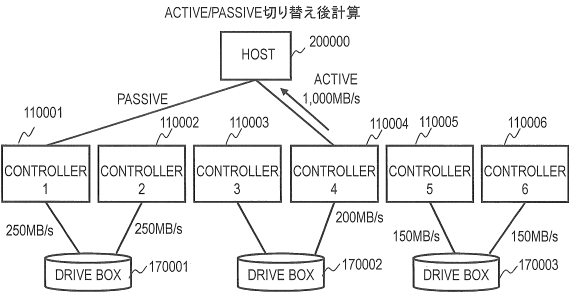
20

30

40

50

【図 19 B】



10

20

30

40

50

フロントページの続き

- (56)参考文献 特開 2016-076184 (JP, A)
 国際公開第 2014/049718 (WO, A1)
 国際公開第 2016/013116 (WO, A1)
 特表 2011-530728 (JP, A)
 特開 2010-079626 (JP, A)
 特開 2006-040026 (JP, A)
- (58)調査した分野 (Int.Cl., DB名)
 G06F 3/06-3/08
 G06F 13/10-13/14
 G06F 16/10-16/188