

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7627183号
(P7627183)

(45)発行日 令和7年2月5日(2025.2.5)

(24)登録日 令和7年1月28日(2025.1.28)

(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 6 Z
G 0 6 F	13/10 (2006.01)	G 0 6 F	3/06 3 0 4 Z
G 0 6 F	16/11 (2019.01)	G 0 6 F	3/06 3 0 5 F
		G 0 6 F	3/06 3 0 6 B
		G 0 6 F	3/06 3 0 5 H
請求項の数 20 (全28頁) 最終頁に続く			
(21)出願番号 特願2021-115921(P2021-115921)		(73)特許権者 390019839	
(22)出願日 令和3年7月13日(2021.7.13)		三星電子株式会社	
(65)公開番号 特開2022-17215(P2022-17215A)		Samsung Electronics	
(43)公開日 令和4年1月25日(2022.1.25)		Co., Ltd.	
審査請求日 令和6年7月9日(2024.7.9)		大韓民国京畿道水原市靈通区三星路 1 2	
(31)優先権主張番号 63/051,158		9	
(32)優先日 令和2年7月13日(2020.7.13)		1 2 9 , Samsung - ro , Ye o	
(33)優先権主張国・地域又は機関		ng tong - gu , Suwon - si	
米国(US)		, Gyeonggi - do , Repub	
(31)優先権主張番号 63/052,854		lic of Korea	
(32)優先日 令和2年7月16日(2020.7.16)		(74)代理人 110000051	
(33)優先権主張国・地域又は機関		弁理士法人共生国際特許事務所	
米国(US)		(72)発明者 奇 亮 ソク	
(31)優先権主張番号 63/057,744		アメリカ合衆国 , カリフォルニア州 9	
(32)優先日 令和2年7月28日(2020.7.28)		4 3 0 3 , パロアルト , アルテア ウ	
最終頁に続く		最終頁に続く	

(54)【発明の名称】 ストレージ装置の動作方法及びストレージ装置を含むシステム

(57)【特許請求の範囲】

【請求項 1】

ストレージ媒体を含むストレージ装置を動作させる方法であって、
前記ストレージ装置によって、前記ストレージ装置を第 1 読み取り専用モードで動作させることによって、部分復元可能な障害の状態にあることを判断する段階と、
前記第 1 読み取り専用モードで前記ストレージ装置を動作させる段階と、を備え、
前記第 1 読み取り専用モードで前記ストレージ装置を動作させる段階は、
前記障害の状態に関連付けられた前記ストレージ装置の領域内の第 1 データ及び第 2 データを識別する段階と、
前記第 1 データの経過時間 (a g e) がしきい値経過時間 (t h r e s h o l d a g e) を超えたことを判断し、前記ストレージ装置内のレスキュースペースのサイズが、前記障害の状態に関連付けられた前記ストレージ装置の領域のサイズよりも小さいことを判断する段階と、
前記判断に基づいて、前記第 1 データを前記ストレージ装置内の前記レスキュースペース内にコピーし、前記第 2 データを前記ストレージ装置の前記領域に保持する段階と、を含むことを特徴とするストレージ装置の動作方法。

【請求項 2】

前記しきい値経過時間は、前記ストレージ装置の前記領域の保存期間に基づくことを特徴とする請求項 1 に記載のストレージ装置の動作方法。

【請求項 3】

前記第 1 読み取り専用モードで前記ストレージ装置を動作させる段階は、

前記ストレージ装置に接続されたホストから第 3 データの項目の読み取りコマンドを受信する段階と、

前記第 3 データの項目をデコーディングする段階と、

前記デコーディングに基づいて、インディケーション (i n d i c a t i o n) をリターンする段階と、

前記ストレージ装置の前記領域のサイズがしきい値サイズ未満であることを判定する段階と、をさらに備えることを特徴とする請求項 1 に記載のストレージ装置の動作方法。

【請求項 4】

前記第 1 読み取り専用モードで前記ストレージ装置を動作させる段階は、

10

使用可能なレスキュースペースのサイズがしきい値サイズよりも小さいことを判断する段階と、

前記ストレージ装置の操作を前記第 1 読み取り専用モードから第 2 読み取り専用モードに切り替える段階と、

前記ストレージ装置に接続されたホストに通知を送る段階と、をさらに備え、

前記通知は、前記ストレージ装置が前記第 2 読み取り専用モードで動作しており、前記ストレージ装置に格納されたデータの項目が期限切れになることを許容することを通知することを特徴とする請求項 3 に記載のストレージ装置の動作方法。

【請求項 5】

前記ホストにより、前記通知を受信する段階と、

20

前記ホストにより、前記ストレージ装置から前記第 3 データの項目を読み取る段階と、

前記ホストにより、前記第 3 データの項目を、他のストレージ装置へ書き込む段階と、をさらに備えることを特徴とする請求項 4 に記載のストレージ装置の動作方法。

【請求項 6】

前記ホストにより、前記ストレージ装置に格納されたデータの量と前記データの期限切れになるまでの時間とに基づいてデータレートを決定する段階をさらに備えることを特徴とする請求項 5 に記載のストレージ装置の動作方法。

【請求項 7】

前記第 1 読み取り専用モードで前記ストレージ装置を動作させる段階は、

前記ストレージ装置に接続されたホストから、第 3 データの項目に対する読み取りコマンドを受信する段階と、

30

前記第 3 データの項目の経過時間が第 2 しきい値経過時間よりも大きいことを判断する段階と、

前記第 3 データの項目の経過時間が前記第 2 しきい値経過時間よりも大きいことに基づいて、インディケーションを前記ホストにリターンする段階と、をさらに備えることを特徴とする請求項 4 に記載のストレージ装置の動作方法。

【請求項 8】

前記第 1 読み取り専用モードで前記ストレージ装置を動作させる段階は、

前記ストレージ装置に接続されたホストから、第 3 データの項目に対する読み取りコマンドを受信する段階と、

40

前記第 3 データの項目をデコーディングする段階と、

前記デコーディングに基づいて、インディケーション (i n d i c a t i o n) をリターンする段階と、をさらに備えることを特徴とする請求項 4 に記載のストレージ装置の動作方法。

【請求項 9】

前記第 1 読み取り専用モードでの動作は、

前記ストレージ装置のページに格納された第 3 データの項目の経過時間が、前記第 1 データの項目の経過時間よりも少ない量だけしきい値経過時間を超えたことを判断する段階と、

前記ストレージ装置のレスキュースペースに前記第 3 データの項目をコピーする段階と

50

、を含むことを特徴とする請求項 1 に記載のストレージ装置の動作方法。

【請求項 1 0】

前記第 1 読み取り専用モードで前記ストレージ装置を動作させる段階は、

前記第 1 データの項目を圧縮する段階をさらに備えることを特徴とする請求項 1 に記載のストレージ装置の動作方法。

【請求項 1 1】

前記第 1 読み取り専用モードで前記ストレージ装置を動作させる段階は、

前記レスキュースペースのための名前空間を作成する段階と、

前記名前空間のサイズを増加させる段階と、をさらに備えることを特徴とする請求項 1 に記載のストレージ装置の動作方法。

【請求項 1 2】

前記ストレージ装置に接続されたホストにより、前記第 1 データの項目に第 1 重要度のインディケータ (i n d i c a t o r) をタグ付けする段階をさらに備えることを特徴とする請求項 1 に記載のストレージ装置の動作方法。

【請求項 1 3】

前記ストレージ装置に接続されたホストにより、前記第 1 データの項目が第 1 ファイルの一部であるとしてタグ付けする段階をさらに備え、

前記第 1 ファイルは、第 1 ファイルタイプのファイルであることを特徴とする請求項 1 に記載のストレージ装置の動作方法。

【請求項 1 4】

前記ストレージ装置に接続されたホストにより、前記第 1 データが第 1 ファイルの一部であるとしてタグ付けする段階と、

前記ストレージ装置に接続されたホストにより、第 3 データが第 1 ファイルの一部であるとしてタグ付けする段階と、をさらに備えることを特徴とする請求項 1 に記載のストレージ装置の動作方法。

【請求項 1 5】

ストレージ装置を含むシステムであって、

前記ストレージ装置は、処理回路を有し、

前記処理回路は、

前記ストレージ装置によって、前記ストレージ装置を第 1 読み取り専用モードで動作させることによって、部分復元可能な障害の状態にあることを判断し、

前記第 1 読み取り専用モードで前記ストレージ装置を動作させるように構成され、

前記第 1 読み取り専用モードでの動作は、

前記障害の状態に関連付けられた前記ストレージ装置の領域内の第 1 データ及び第 2 データを識別し、

前記第 1 データの経過時間がしきい値経過時間を超えたことを判断し、前記ストレージ装置内のレスキュースペースのサイズが、前記障害の状態に関連付けられた前記ストレージ装置の領域のサイズよりも小さいことを判断し、

前記判断に基づいて、前記第 1 データを前記ストレージ装置内の前記レスキュースペース内にコピーし、前記第 2 データを前記ストレージ装置の前記領域に保持することを含むことを特徴とするストレージ装置を含むシステム。

【請求項 1 6】

前記しきい値経過時間は、前記ストレージ装置の前記領域の保存期間に基づくことを特徴とする請求項 1 5 に記載のストレージ装置を含むシステム。

【請求項 1 7】

前記処理回路は、

使用可能なレスキュースペースのサイズがしきい値サイズよりも小さいことを判断し、

第 2 読み取り専用モードで前記ストレージ装置を動作させ、

前記ストレージ装置に接続されたホストに通知を送るようにさらに構成され、

前記通知は、前記ストレージ装置が前記第 2 読み取り専用モードで動作しており、前記

10

20

30

40

50

ストレージ装置に格納されたデータの項目が期限切れになることを許容することを通知することを特徴とする請求項 15 に記載のストレージ装置を含むシステム。

【請求項 18】

ストレージ装置を含むシステムであって、
前記ストレージ装置は、処理手段を有し、
前記処理手段は、
前記ストレージ装置によって、前記ストレージ装置を第 1 読み取り専用モードで動作させることによって、部分復元可能な障害の状態にあることを判断し、
前記第 1 読み取り専用モードで前記ストレージ装置を動作させるように構成され、
前記第 1 読み取り専用モードでの動作は、
前記障害の状態に関連付けられた前記ストレージ装置の領域内の第 1 データ及び第 2 データを識別し、

10

前記第 1 データの経過時間がしきい値経過時間を超えたことを判断し、前記ストレージ装置内のレスキュースペースのサイズが、前記障害の状態に関連付けられた前記ストレージ装置の領域のサイズよりも小さいことを判断し、

前記判断に基づいて、前記第 1 データを前記ストレージ装置内の前記レスキュースペース内にコピーし、前記第 2 データを前記ストレージ装置の前記領域に保持することを含むことを特徴とするストレージ装置を含むシステム。

【請求項 19】

前記しきい値経過時間は、前記ストレージ装置の前記領域の保存期間に基づくことを特徴とする請求項 18 に記載のストレージ装置を含むシステム。

20

【請求項 20】

前記処理手段は、
使用可能なレスキュースペースのサイズがしきい値サイズよりも小さいことを判断し、
第 2 読み取り専用モードで前記ストレージ装置を動作させ、
前記ストレージ装置に接続されたホストに通知を送るように構成され、
前記通知は、前記ストレージ装置が前記第 2 読み取り専用モードで動作しており、前記ストレージ装置に格納されたデータの項目が期限切れになることを許容することを通知することを特徴とする請求項 18 に記載のストレージ装置を含むシステム。

【発明の詳細な説明】

30

【技術分野】

【0001】

本発明は、データストレージシステムに関し、特に、内部障害状態の存在下で回復（復元）力を示すように構成されるストレージ装置の動作方法及びストレージ装置を含むシステムに関する。

【背景技術】

【0002】

ストレージ装置とシステムは、多様なデータ処理システム内で広範な応用性を有している。

ストレージ装置は、温度センサーの障害又は放射線による（*radiation-induced*）ビットフリップ（*bit flip*）を含む相当な範囲の内部障害状態に直面するほど十分に複雑である。

40

ストレージ装置、又はそのようなストレージ装置を使用しているシステム（例えば、*RAID*（*Redundant Array of Independent Disks*））の設計時に、これらの障害状態を考慮しなければ、ストレージ装置の単一障害が全体のストレージ装置、あるいは、ストレージ装置を含む（*RAID*のようなもの）より大規模なシステムが動作しなくなる可能性がある。

【0003】

したがって、ストレージ装置とストレージ装置を含むシステムの回復（復元）動作のためのシステム及び方法の開発が課題となっている。

50

【先行技術文献】

【特許文献】

【0004】

【文献】米国特許第7451346号明細書

【文献】米国特許第8589723号明細書

【文献】米国特許第8769535号明細書

【文献】米国特許第9710317号明細書

【文献】米国特許第9728277号明細書

【文献】米国特許第9830236号明細書

【文献】米国特許第10095506号明細書

10

【文献】米国特許第10102082号明細書

【文献】米国特許第10263842号明細書

【文献】米国特許第10445200号明細書

【文献】米国特許第10572161号明細書

【文献】米国特許出願公開第20050120267号明細書

【文献】米国特許出願公開第20060143507号明細書

【文献】米国特許出願公開第20070101188号明細書

【文献】米国特許出願公開第20130047028号明細書

【文献】米国特許出願公開第20140195847号明細書

【文献】米国特許出願公開第20150100720号明細書

20

【文献】米国特許出願公開第20180011649号明細書

【文献】米国特許出願公開第20190377637号明細書

【文献】米国特許出願公開第20190384670号明細書

【文献】米国特許出願公開第20200135280号明細書

【発明の概要】

【発明が解決しようとする課題】

【0005】

本発明は上記従来のストレージ装置とストレージ装置を含むシステムにおける課題に鑑みてなされたものであって、本発明の目的は、多様な回復（復元）モードにおいて動作することができ、内部障害にもかかわらず、部分的な機能を維持するか、又はすべての機能にリターンすることができる障害回復力のある読み取り専用モードを備えるストレージ装置の動作方法及びストレージ装置を含むシステムを提供することにある。

30

【課題を解決するための手段】

【0006】

上記目的を達成するためになされた本発明によるストレージ装置の動作方法は、ストレージ媒体を含むストレージ装置を動作させる方法であって、前記ストレージ装置によって、前記ストレージ装置を第1読み取り専用モードで動作させることによって、部分復元可能な障害の状態にあることを判断する段階と、前記第1読み取り専用モードで前記ストレージ装置を動作させる段階と、を備え、前記第1読み取り専用モードで前記ストレージ装置を動作させる段階は、前記障害の状態に関連付けられた前記ストレージ装置の領域内の第1データ及び第2データを識別する段階と、前記第1データの経過時間（age）がしきい値経過時間（threshold age）を超えたことを判断し、前記ストレージ装置内のレスキュースペースのサイズが、前記障害の状態に関連付けられた前記ストレージ装置の領域のサイズよりも小さいことを判断する段階と、前記判断に基づいて、前記第1データを前記ストレージ装置内の前記レスキュースペース内にコピーし、前記第2データを前記ストレージ装置の前記領域に保持する段階と、を含むことを特徴とする。

40

【0007】

前記しきい値経過時間は、前記ストレージ装置の前記ページの保存期間に基づくことが好ましい。

前記第1読み取り専用モードで前記ストレージ装置を動作させる段階は、前記ストレージ

50

装置に接続されたホストから第3データの項目の読み取りコマンドを受信する段階と、前記第3データの項目をデコーディングする段階と、前記デコーディングに基づいて、インディケーション (i n d i c a t i o n) をリターンする段階と、前記ストレージ装置の前記領域のサイズがしきい値サイズ未満であることを判定する段階と、をさらに備えることが好ましい。

前記第1読み取り専用モードで前記ストレージ装置を動作させる段階は、使用可能なレスキュースペースのサイズがしきい値サイズよりも小さいことを判断する段階と、前記ストレージ装置の操作を前記第1読み取り専用モードから第2読み取り専用モードに切り替える段階と、前記ストレージ装置に接続されたホストに通知を送る段階と、をさらに備え、前記通知は、前記ストレージ装置が前記第2読み取り専用モードで動作しており、前記ストレージ装置に格納されたデータの項目が期限切れになることを許容することを通知するが好ましい。

10

前記ホストにより、前記通知を受信する段階と、前記ホストにより、前記ストレージ装置から前記第3データの項目を読み取る段階と、前記ホストにより、前記第3データの項目を、他のストレージ装置へ書き込む段階と、をさらに備えることが好ましい。

前記ホストにより、前記ストレージ装置に格納されたデータの量と前記データの期限切れになるまでの時間とに基づいてデータレートを決定する段階をさらに備えることが好ましい。

【0008】

前記第1読み取り専用モードで前記ストレージ装置を動作させる段階は、前記ストレージ装置に接続されたホストから、第3データの項目に対する読み取りコマンドを受信する段階と、前記第3データの項目の経過時間が第2しきい値経過時間よりも大きいことを判断する段階と、前記第3データの項目の経過時間が前記第2しきい値経過時間よりも大きいに基づいて、インディケーションを前記ホストにリターンする段階と、をさらに備えることが好ましい。

20

前記第1読み取り専用モードで前記ストレージ装置を動作させる段階は、前記ストレージ装置に接続されたホストから、第3データの項目に対する読み取りコマンドを受信する段階と、前記第3データの項目をデコーディングする段階と、前記デコーディングに基づいて、インディケーション (i n d i c a t i o n) をリターンする段階と、をさらに備えることが好ましい。

30

前記第1読み取り専用モードでの動作は、前記ストレージ装置のページに格納された第3データの項目の経過時間が、前記第1データの項目の経過時間よりも少ない量だけしきい値経過時間を超えたことを判断する段階と、前記ストレージ装置のレスキュースペースに前記第3データの項目をコピーする段階と、を含むことが好ましい。

前記第1読み取り専用モードで前記ストレージ装置を動作させる段階は、前記第1データの項目を圧縮する段階をさらに備えることが好ましい。

【0009】

前記第1読み取り専用モードで前記ストレージ装置を動作させる段階は、前記レスキュースペースのための名前空間を作成する段階と、前記名前空間のサイズを増加させる段階と、をさらに備えることが好ましい。

40

前記ストレージ装置に接続されたホストにより、前記第1データの項目に第1重要度のインディケ이터 (i n d i c a t o r) をタグ付けする段階をさらに備えることが好ましい。

前記ストレージ装置に接続されたホストにより、前記第1データの項目が第1ファイルの一部であるとしてタグ付けする段階をさらに備え、前記第1ファイルは、第1ファイルタイプのファイルであることが好ましい。

前記ストレージ装置に接続されたホストにより、前記第1データが第1ファイルの一部であるとしてタグ付けする段階と、前記ストレージ装置に接続されたホストにより、第3データが第1ファイルの一部であるとしてタグ付けする段階と、をさらに備えることが好ましい。

50

【 0 0 1 0 】

上記目的を達成するためになされた本発明によるストレージ装置を含むシステムは、ストレージ装置を含むシステムであって、前記ストレージ装置は、処理回路を有し、前記処理回路は、前記ストレージ装置によって、前記ストレージ装置を第1読み取り専用モードで動作させることによって、部分復元可能な障害の状態にあることを判断し、前記第1読み取り専用モードで前記ストレージ装置を動作させるように構成され、前記第1読み取り専用モードでの動作は、前記障害の状態に関連付けられた前記ストレージ装置の領域内の第1データ及び第2データを識別し、前記第1データの経過時間がしきい値経過時間を超えたことを判断し、前記ストレージ装置内のレスキュースペースのサイズが、前記障害の状態に関連付けられた前記ストレージ装置の領域のサイズよりも小さいことを判断し、前記判断に基づいて、前記第1データを前記ストレージ装置内の前記レスキュースペース内にコピーし、前記第2データを前記ストレージ装置の前記領域に保持することを含むことを特徴とする。

10

【 0 0 1 1 】

前記しきい値経過時間は、前記ストレージ装置の前記領域の保存期間に基づくことが好ましい。

前記処理回路は、使用可能なレスキュースペースのサイズがしきい値サイズよりも小さいことを判断し、第2読み取り専用モードで前記ストレージ装置を動作させ、前記ストレージ装置に接続されたホストに通知を送るようにさらに構成され、前記通知は、前記ストレージ装置が前記第2読み取り専用モードで動作しており、前記ストレージ装置に格納されたデータの項目が期限切れになることを許容することを通知することが好ましい。

20

【 0 0 1 2 】

また、上記目的を達成するためになされた本発明によるストレージ装置を含むシステムは、ストレージ装置を含むシステムであって、前記ストレージ装置は、処理手段を有し、前記処理手段は、前記ストレージ装置によって、前記ストレージ装置を第1読み取り専用モードで動作させることによって、部分復元可能な障害の状態にあることを判断し、前記第1読み取り専用モードで前記ストレージ装置を動作させるように構成され、前記第1読み取り専用モードでの動作は、前記障害の状態に関連付けられた前記ストレージ装置の領域内の第1データ及び第2データを識別し、前記第1データの経過時間がしきい値経過時間を超えたことを判断し、前記ストレージ装置内のレスキュースペースのサイズが、前記障害の状態に関連付けられた前記ストレージ装置の領域のサイズよりも小さいことを判断し、前記判断に基づいて、前記第1データを前記ストレージ装置内の前記レスキュースペース内にコピーし、前記第2データを前記ストレージ装置の前記領域に保持することを含むことを特徴とする。

30

【 0 0 1 3 】

前記しきい値経過時間は、前記ストレージ装置の前記領域の保存期間に基づくことが好ましい。

前記処理手段は、使用可能なレスキュースペースのサイズがしきい値サイズよりも小さいことを判断し、第2読み取り専用モードで前記ストレージ装置を動作させ、前記ストレージ装置に接続されたホストに通知を送るように構成され、前記通知は、前記ストレージ装置が前記第2読み取り専用モードで動作しており、前記ストレージ装置に格納されたデータの項目が期限切れになることを許容することを通知することが好ましい。

40

【発明の効果】

【 0 0 1 4 】

本発明に係るストレージ装置の動作方法及びストレージ装置を含むシステムによれば、多様な回復（復元）モードにおいて動作することができ、内部障害にもかかわらず、部分的な機能を維持するか、又はフル（full）機能にリターンすることができる、障害回復力のある読み取り専用モードを備えるストレージ装置の動作方法及びストレージ装置を含むシステムを提供することができる。

【図面の簡単な説明】

50

【 0 0 1 5 】

【図 1】本発明の一実施形態によるデータストレージシステムの概略構成を示すブロック図である。

【図 2 A】本発明の一実施形態によるストレージ装置の障害状態を説明するためのテーブルである。

【図 2 B】本発明の一実施形態によるストレージ装置の動作モードを説明するためのテーブルである。

【図 2 C】本発明の一実施形態による、ホストによってストレージ装置に送られるためのコマンドを説明するためのテーブルである。

【図 2 D】本発明の一実施形態による回復（復元）モードの階層構造を説明するためのテーブルである。

10

【図 3 A】本発明の一実施形態による障害回復力のあるストレージ装置の動作の第 1 方法を説明するためのフローチャートである。

【図 3 B】本発明の一実施形態による障害回復力のあるストレージ装置の動作の第 2 方法を説明するためのフローチャートである。

【図 4】本発明の一実施形態によるストレージ装置の概略的なデータレイアウトを説明するための図である。

【図 5 A】本発明の一実施形態による障害回復の読み取り専用モードでストレージ装置の動作の第 1 方法を説明するためのフローチャートである。

【図 5 B】本発明の一実施形態による障害回復の読み取り専用モードでストレージ装置の動作の第 2 方法を説明するためのフローチャートである。

20

【図 6 A】本発明の一実施形態による書き込み動作を実行する「RAID - 0」システムの概略的なデータレイアウト図である。

【図 6 B】本発明の一実施形態による読み取り動作を実行する「RAID - 0」システムの概略的なデータレイアウト図である。

【図 6 C】本発明の一実施形態による再マッピング（remapping）と書き込み動作を実行する「RAID - 0」システムの概略的なデータレイアウト図である。

【図 7 A】本発明の一実施形態による「RAID - 0」ストレージシステムの動作の第 1 方法を説明するためのフローチャートである。

【図 7 B】本発明の一実施形態による「RAID - 0」ストレージシステムの動作の第 2 方法を説明するためのフローチャートである。

30

【発明を実施するための形態】

【 0 0 1 6 】

次に、本発明に係るストレージ装置の動作方法及びストレージ装置を含むシステムを実施するための形態の具体例を図面を参照しながら説明する。

【 0 0 1 7 】

添付した図面に関連して以下説明する詳細な説明は、本開示に基づいて提供されるメモリリソース管理システム及び方法の例としての実施形態を説明するためのものであり、本開示が構成・活用される唯一の形態を示すものではない。

以下の説明は、図に示した実施形態と関連して、本発明の特徴を提示する。

40

しかし、同一又は同等の機能と構造がまた、本発明の範囲内に含まれるように意図される他の実施形態によって達成されることが理解されるべきである。

本明細書で類似の図面符号は、類似の構成要素又は特徴を示す。

【 0 0 1 8 】

本明細書で使用する、「～の一部」は、物体の「少なくとも一部」を意味し、したがって物体の全部又は全部より少ないことを意味することができる。

このように、物体の「一部」は、物体全体を特別な場合として含んでいるが、すなわち、物体全体が物体の一部に対する一例である。

ここで使用される用語「又は」は、「及び／又は」として解釈されるべきであり、例えば、「A 又は B」は、「A」、「B」又は「A 及び B」のうち、いずれか 1 つを意味する。

50

【 0 0 1 9 】

本明細書の背景のセクションで提供されている背景は、脈絡（コンテキスト）を設定するためにだけ含まれており、このセクションのコンテンツは、従来技術であると認められていない。

説明される任意のコンポーネント又はコンポーネントの任意の組み合わせ（例えば、ここに含まれている任意のシステムダイアグラムで）は、ここに含まれている任意のフローチャートの動作のうち、いずれか1つを実行するために使用される。

なお、（i）前記動作は、例としてのものであり、明示的にカバーされていない多様な追加の段階を含むことができ、（ii）前記動作の時間的順序は変更され得る。

【 0 0 2 0 】

10

本明細書で使用する、方法（例えば、調整）又は第1数量（例えば、第1変数）が、第2数量（例えば、第2変数）に「基づく」と言及されるとき、これは、第2数量が方法に対する入力又は第1数量に影響を与えるが、例えば、第2数量は、第1数量を計算する関数への入力（例えば、唯一の入力又は1つ以上の入力のうちの1つ）であるか、第1数量は、第2数量と同じ値を有するか、又は第2数量と同じであり得る（例えば、メモリ内の同じ位置又は位置に格納される）。

【 0 0 2 1 】

ただし、用語「第1」、「第2」、「第3」などが、本明細書では、多様なエレメント、構成要素、領域、レイヤー及び/又はセクションを説明するために使用され、これらのエレメント、構成要素、領域、レイヤー及び/又はセクションは、これらの用語により限定されないと理解されるだろう。

20

これらの用語は、1つのエレメント、構成要素、領域、レイヤー又はセクションを他のエレメント、構成要素、領域、レイヤー、又はセクションと区別するためにのみ使用される。したがって、本明細書に記載された第1エレメント、構成要素、領域、レイヤー又はセクションは、本発明の概念の技術的思想と範囲を逸脱することなく、第2エレメント、構成要素、領域、レイヤー又はセクションと称される。

【 0 0 2 2 】

「すぐ下（beneath）」、「下（below）」、「下位（lower）」、「下部（under）」、「上部（above）」、「上位（upper）」などのような空間的に相対的な用語は、説明の便宜のために、図面に示したような1つのエレメント、特徴の異なるエレメント又は特徴に対して存在する関係を説明するために使用される。

30

そのような空間的に相対的な用語は、図面に示した方向に加えて、使用中又は動作中の装置の他の方向を含むように意図されたものであることを理解するはずである。たとえば、図面の装置が裏返された場合は、他のエレメント若しくは特徴の「下」、「下」又は「すぐ下」として説明されたエレメントは、前記他の装置又は特徴の「上部」に向かうことになる。

したがって、例としての用語「下」と「下部」は、上部及び下部の方向の両方を含み得る。

装置は、別の方法で配置されることがあり（例えば、90度回転されるか、又は他の方向に）、ここで使用された空間的に相対的な記述語（descriptor）は、それに

40

さらに、1つのレイヤーが2つのレイヤーの間にあるものとして述べられるときに、これは2つのレイヤーの間の唯一のレイヤーであり得るか、又は1つ以上の介在するレイヤー（intervening layers）が存在する可能性がある」と理解するだろう。

【 0 0 2 3 】

本明細書で使用する用語は、特定の実施形態を説明するためのものであり、本発明を限定するものではない。

本明細書で使用する用語「実質的に」、「約」及びこれと類似した用語は、程度（degree）の用語ではなく、近似値の用語として使用され、当業者によって認知されることが可能な測定又は計算された値の固有な偏差を考慮するように意図されたものである。

50

【 0 0 2 4 】

本明細書で使用する、単数形「a」と「an」は、文脈上明らかに別の意味を示していると判定されない限り、複数形も含むように意図する。

本明細書で使用する、「含む (comprises)」及び/又は「含んでいる (comprising)」という用語は、言及された特徴、整数、段階、動作、エレメント、及び/又は構成要素 (コンポーネント) の存在を特定するが、1つ以上の他の特徴、整数、段階、動作、エレメント、構成要素、及び/又はそのグループの存在若しくは追加を排除しないということも理解されるだろう。

本明細書で使用する用語「及び/又は」は、1つ以上の関連されて、リストされた項目の1つ以上の任意かつすべての組み合わせを含む。「の少なくとも1つ」のような表現は、エレメントのリストの前に来るときに、エレメントの全リストを変更し、リストの個々のエレメントを変更しない。なお、本発明の実施形態を説明するとき、「できる (may)」という用語は、「本発明の1つ以上の実施形態」を表す。

また、「例としての」という用語は、例えば、例又は例示を示すものであると意図される。

本明細書で使用する「使用する (use)」、「使用している (using)」、「使用された (used)」という用語は、それぞれ「活用する (utilize)」、「活用している (utilizing)」、「活用された (utilized)」という用語と同義語として見なされる。

【 0 0 2 5 】

コンポーネント又はレイヤーが、他の構成要素又はレイヤー「に位置する」、「に接続される」、「に結合される」又は「に隣接する」と述べられるときに、前記構成要素又はレイヤーが、他の構成要素又はレイヤーに直接位置するか、接続・結合されるか、又は隣接したものであり、1つ以上の介在する構成要素又はレイヤーが存在することができる。

対照的に、構成要素又はレイヤーが、他の構成要素又はレイヤー「のすぐ上に」、「に直接接続される」、「に直接結合される」又は「に隣接する」と述べられるとき、介在する構成要素又はレイヤーが存在しない。

【 0 0 2 6 】

本明細書で引用する任意の数値範囲は、引用された範囲内に含まれている同じ数値精度のすべてのサブ範囲を含むように意図される。

たとえば、「1.0 ~ 10.0」の範囲や「1.0と10.0の間」の範囲は、記載された最小値1.0と記載された最大値10.0の間の(これらを含む)、すなわち1.0以上の最小値と10.0以下の最大値を有するすべてのサブ範囲、例えば、2.4 ~ 7.6の範囲を含む。

本明細書に記載した任意の最大値の限定は、その中に含まれているすべてのより低い数値限定を含むように意図され、本明細書に記載された任意の最小値の限定は、その中に含まれているすべてのより高い数値限定を含むように意図される。

【 0 0 2 7 】

本発明の実施形態に係る障害回復力のあるストレージ装置は、多様な回復 (復元) モードにおいて動作することができ、内部障害にもかかわらず、部分的な機能を維持するか、又はフル (full) 機能にリターンすることができる。

例えば、温度センサーの障害に対応してストレージ装置は、減少した速度で動作するか、書き込み動作のための減少した速度で動作することができるか、又は書き込み動作をほとんど実行しないことがあり得る。

ストレージ媒体にエラーを引き起こす可能性がある一時的な障害に応答して、ストレージ装置は、ストレージ媒体を再びフォーマットすることができる。

ストレージ装置を、エラーを起こさない一時的な障害に応答して、ストレージ装置は、電源をパワーサイクリングすることができる (power cycling)。

【 0 0 2 8 】

ストレージ装置が部分的機能を維持することができる方法の1つは、いくつかの状況に

において、読み取り専用モードで動作するものであり、読み取り専用モードでは、ストレージ装置が、ホストからの読み取り要請に応答することができるが、ホストから書き込み要請を受信すると、エラーメッセージをリターンすることができる。

また、期限切れ予定のデータをレスキュースペースに移動することができるため、期限切れにより、データが失われない。

「RAID - 0」のストレージシステムでは、ストレージ装置が読み取り専用モードで動作することに切り替えることができる可能性は、ストレージ装置が読み取り専用モードで動作することに切り替えた後、実行される任意の書き込み操作（動作）に対するストライプを構成するストリップを再分配することにより、達成される。

【0029】

図1は、本発明の実施形態による1つ以上のストレージ装置110（又は「ドライブ」）に接続されるホスト105を含むコンピューティングシステムの概略構成を示すブロック図である。

ストレージ装置110は、図に示すように、ボリュームマネージャー115（以下で、詳細に説明する）を介してホスト105に接続されるか、又はホスト105に直接接続される。

【0030】

一実施形態では、ストレージ装置110は、内部障害状態を経験する可能性があり、ストレージ装置は、障害状態のシステムレベルの影響を緩和するために、以下で詳細に記述するように、多様な復元動作（操作）を示すことができる。

それぞれのストレージ装置110は、ソリッドステートドライブ（SSD）であり、コントローラ120（又は「制御回路」）及びストレージ媒体125（例えば、ナンド（NAND）型フラッシュメモリ）を含み得る。

ストレージ装置110で消去可能な最小単位は、「ブロック」と呼ばれ、ストレージ装置110で最小の書き込み可能な単位は「ページ」と称される。

【0031】

各ストレージ装置110は、2.5インチ、1.8インチ、MO-297、MO-300、M.2、及びEDSFF（Enterprise and Data Center SSD Form Factor）を含むが、これらに限定されない永続性ストレージ装置に適した複数のフォームファクタの内のいずれか1つのフォームファクタを有することができ、PCI（Peripheral Component Interconnect）、PCIe（PCI express）、SCSI（Ethernet、Small Computer System Interface）、SATA、（Serial ATA Attachment）、及びSAS（Serial Attached SCSI）を含む永続性ストレージ装置に適した複数のインターフェースの内のいずれか1つである電氣的インターフェースを有することができ、インターフェースを介してホスト105に接続される。

【0032】

ストレージ媒体125は、保存期間（ストレージ媒体125の使用履歴に左右されることがあり、それに応じてストレージ媒体125内で変化し得る）を有し、保存期間よりも長く格納されたデータ（すなわち、保存期間を超える経過時間を有するデータ）は、信頼できなくなって有効期間が切れたと言うことができる。

データは、ブロックコードの可能性のあるエラー訂正コードを用いて、ストレージ媒体125に格納される。

ストレージ媒体125からデータを読み取るとき、ストレージ媒体125からコードブロック（code block）と称される多量のローデータを読み取ってデコーディングを試みる。

試みが失敗した場合、追加の試み（読み取りの再試み）が行われる。

【0033】

使用に応じて、ストレージ媒体125の一部である、例えば、ブロックは、保存期間が

10

20

30

40

50

許容することができないほど短くなる地点までに低下する可能性があり、そのブロックは、「不良ブロック」として分類され得る。

このような状況が、全体のストレージ媒体 1 2 5 を利用不可能にすることを防止するために、「不良ブロック管理の予備スペース」と呼ばれる予備空間 (space、以下、空間と記す) を存在させ (例えば、各フラッシュメモリダイ又は各フラッシュメモリプレーンに含まれる)、フラッシュメモリダイ又はフラッシュメモリプレーン内部のコントローラ 1 2 0 又は他のコントローラは、予備ブロックを使用し始め、不良ブロックの使用を中断することができる。

【 0 0 3 4 】

図 2 A は、それぞれのケース識別子 (「ケース ID」) とラベル付けされた障害状態 (又は「障害条件」) を説明するためのテーブルである。

10

ケース 1 は、ストレージ装置 1 1 0 は、もはや読み取り又は書き込みの動作 (操作) を実行することができず、電源をパワーサイクリングするか、又はストレージ媒体を再フォーマットすることで、解決することができない任意の障害状態 (条件) である。

ストレージ装置 1 1 0 は、このような方法で動作する状態は、例えば、異なる障害のメカニズム (failure mechanism) にそれぞれ対応する多様なサブ状態を有し得る。

このような状態又は障害の状態 (ストレージ装置 1 1 0 は、もはや読み取り又は書き込みの操作を実行することができず、電源をパワーサイクリングするか、又はストレージ媒体を再フォーマットすることで、解決することができない) は、例えば、損傷されたコントローラのファームウェアの一部 (この場合、コントローラが、損傷されたコマンドが実行されない安全モードで再起動することができる)、又はストレージ装置 1 1 0 の処理回路の障害 (例えば、ストレージ媒体との相互作用を管理するが、ホスト 1 0 5 との通信を担当していない処理回路の障害) によって引き起こされる。

20

このようなタイプの障害状態が発生すれば、ストレージ装置 1 1 0 は、ホスト 1 0 5 からの任意の読み取り又は書き込みのコマンドに対してエラーメッセージとともに応答することができる。

【 0 0 3 5 】

ケース 2 は、

(i) ストレージ装置 1 1 0 がもはや読み取り又は書き込みの動作を実行することができず、

30

(i i) ストレージ装置 1 1 0 の電源をパワーサイクリングするか、又はストレージ媒体を再フォーマットすることで復元可能な任意の障害状態である。

このような障害の状態は、例えば、ストレージ装置 1 1 0 のコントローラ 1 2 0 のプログラムの実行エラー (例えば、コントローラ 1 2 0 のランダムアクセスメモリ (RAM) でビットフリップの結果として範囲を超えたポインタ、又はビットフリップの結果として間違っ

たデータを書き込んでいない場合 (例えば、コントローラによってストレージ媒体に、最新の書き込みの後、プログラムの実行エラーが発生した場合)、ストレージ装置の電源をパワーサイクリングすることがストレージ装置 1 1 0 を、通常の動作に復元するのに十分であり、プログラムの実行エラーによってコントローラ 1 2 0 がエラーデータをストレージ媒体 1 2 5 に書き込んだ場合には、ストレージ媒体 1 2 5 を再フォーマットするのは、ストレージ装置 1 1 0 を通常の動作に復元するのに十分であり得る。

40

【 0 0 3 6 】

ケース 3 は、ストレージ装置 1 1 0 を読み取り専用モードで動作させることによって緩和されることがあり、ストレージ媒体 1 2 5 を再フォーマットするのが、全機能を復元しない任意の障害状態である。

このような障害の例は、

(i) 温度センサーの障害、

50

(i i) 読み取り専用モードに転換されたストレージ媒体 1 2 5 の一部を含む。

温度センサーの障害の場合には、温度センサーの測定値が範囲を超えている（例えば、臨界温度を超えた）と判定して障害を感知することができ、この場合、ストレージ装置 1 1 0 の過熱のリスクが、読み取り動作よりも多くの電力を消費することができる書き込み動作を回避することによって、減少されることが可能である。

ストレージ媒体 1 2 5 の一部の読み取り専用モードへの移行は、例えば、フラッシュメモリプレーン又はダイが、ランタイムの不良ブロックの管理に使用されている不良ブロック管理の予備空間を使いつくす場合には、フラッシュメモリのストレージ媒体 1 2 5 に対して発生する可能性がある。

【 0 0 3 7 】

10

例えば、ストレージ装置 1 1 0 は、読み取り動作の実行を試みつつ、データ項目をデコーディングする試みに失敗し、データを格納するブロックが不良ブロックであると判断し、データを不良ブロックから不良ブロック管理の予備空間へ移動するとき、残りの不良ブロック管理の予備空間がしきい値サイズよりも小さく、それに応じてプレーン又はダイの信頼性を確保するには不十分であると判断、決定し得る。

その次に、ストレージ装置 1 1 0 は、不良ブロック管理がもはや実行されていないと決定して、読み取り専用モードに切り替えられ得る。

読み取り専用モードの動作は、以下で詳細に説明される。

本明細書で使用するように、「データ項目 (data item) 」は、1 つの動作で処理される任意の量のデータであり、例えば、コードブロックをデコーディングした結果のデータは、データ項目であり得る。

20

本明細書で使用するように、「データ項目をデコーディングして、デコーディングに基づくインディケーション (indication) をリターンすること」は、データ項目をデコーディングしようとする試みが失敗して、デコーディングの試みが失敗したというインディケーションをリターンするのを意味する

【 0 0 3 8 】

ケース 4 は、ライトスルー (write - through) モードでストレージ装置 1 1 0 を動作させることによって緩和される任意の障害状態である。

例えば、ストレージ装置 1 1 0 の電源サブライバックアップキャパシタ (power supply backup capacitor) が故障した場合には、装置は、ホストから受信された任意の書き込みコマンドに応答して、コマンドの完了をホスト 1 0 5 に送る前にストレージ媒体 1 2 5 への書き込みを完了する可能性があるため、ストレージ媒体 1 2 5 への書き込みが完了される前に電源が故障した場合、ホストは、記録が正常に完了されたことを不正確に通知されない。

30

ライトスルーモードで動作するのは、性能の低減をもたらす可能性がある（スループットとレイテンシの側面から）。

【 0 0 3 9 】

ケース 5 は、電力損失を減少させる方法でストレージ装置 1 1 0 を作動させることによって緩和される任意の障害状態である。

例えば、温度センサーの障害の場合には、ストレージ装置 1 1 0 は、前述したように、読み取り専用モードで動作するか、又は動作（例えば、読み取り動作よりももっと多くの電力を消費することができる書き込み操作）が実行される速度を減少させて、ストレージ装置 1 1 0 での電力損失を低減させる。

40

例えば、ストレージ装置 1 1 0 は、ストレージ媒体に対する 1 番目の書き込みを実行した後、減少された性能に対応する間隔（待機のために書き込み操作が実行される速度が減少する）の間に待機し、その後、ストレージ媒体に他の書き込み（例えば、1 秒）を実行する。

【 0 0 4 0 】

ケース 6 は、ストレージ装置 1 1 0 を読み取り専用モードで動作させることによって緩和され、ストレージ媒体 1 2 5 を再フォーマットすることがフル機能を復元する任意の障

50

害状態である。

図 2 A のテーブルの 3 番目のカラムは、それぞれの場合に対して有効なユーザーデータが利用可能であるか否か、及びストレージ装置 110 は、例えば、ストレージ媒体 125 を再フォーマットすることで、フル機能にリターンされうるか否かを示す。

【0041】

図 2 A にリストされたケースによって例示されるように、本発明の実施形態では、3つのレベルの障害復元性（障害回復力）、すなわち完全回復（復元）モード、部分回復（復元）モード及び脆弱モードは、ストレージ装置 110 によって達成される。

完全（フル）回復（復元）モードで、ストレージ装置 110 は、「自己回復」機能で動作することができ、ストレージ装置 110 は、パワーサイクリング（power cycling）又はストレージ媒体 125 のフォーマットのような動作を再設定することで、フル機能（装置からユーザーデータが失われることがあるとしても）を復元する。

【0042】

部分的な回復（復元）モードで、ストレージ装置 110 は、障害状態が存在する場合、より低い性能、より小さい容量又は減少された機能で動作する。

例えば、前述したように、電源バックアップキャパシタが故障すれば、データがストレージ媒体 125 に記録された後にのみ（つまり、同期式書き込みが行われ得る）、すべての書き込みが完了されることがあり（つまり、コマンド完了がホスト 105 に送られ得る）、それに応じてストレージ装置 110 の動作を遅らせ、その性能を低下させる。

このような状況では、ユーザーデータは保存される。

ストレージ装置 110 は、RTBB（bad block management runtime bad block）のための予備空間が消尽されると、より小さな容量で動作する。

このような状況で、ストレージ装置 110 の影響を受けるダイは、ディスク空間から除外されることがあり、全体のディスク容量が減少する。

失われた空間のユーザーデータが失われ得る。

たとえば、IO 決定論（IO determinism）のセット又はゾーニングされた名前空間（namespace）のゾーンが、もはや新しいデータの書き込みを許容することができない場合、セット又はゾーンがディスク空間から除外され得るが、残りのディスク空間は、読み取りと書き込みの動作に対して使用可能な状態にする。

ゾーン又はセットのユーザーデータが失われうる。

【0043】

ストレージ装置 110 が書き込み動作を許容せず、読み取り専用モードに切り替える場合、ストレージ装置 110 は、減少された機能で動作する。

ストレージ装置 110 は、2つのタイプの読み取り専用モード、すなわち、持続可能な読み取り専用モード（「1 番目の読み取り専用モード」と呼ばれることがある）と持続不可能な読み取り専用モード（「2 番目の読み取り専用モード」と呼ばれることがある）で動作する。

持続可能な読み取り専用モードで、ストレージ装置 110 は、ストレージ媒体 125 の保存期間を超えて読み取り要請をサービスし続ける。

持続可能な読み取り専用モードで動作することが実現可能でないとき、例えば、レスキュースペースを設定するために使用されていないストレージ空間が十分でないとき、持続不可能な読み取り専用モードを使用する（以下、詳細に説明する）。

【0044】

持続不可能な読み取り専用モードに切り替えるときに、ストレージ装置 110 は、ストレージ装置 110 が第 2（持続不可能な）読み取り専用モードで動作しており、ストレージ装置 110 に格納されたデータ項目が期限切れ（例えば、そのそれぞれの保存期間が終了するとき）になることが許容されるという通知をホスト 105 に送る。

持続不可能な読み取り専用モードで、ストレージ装置 110 は、ストレージ媒体 125 の保存期間中に継続して読み取り要請に対してサービスでき、ストレージ装置 110 が、

データのインテグリティ問題（例えば、読み取り動作中のデータをデコーディングする1つ以上失敗した試行によって検出される）に遭遇すれば、ストレージ装置110は、有効ではないデータ領域を報告する。

脆弱モードで動作するストレージ装置110は、前述したように、読み取り又は書き込みの操作（動作）を実行することができず、正常な終了（graceful exit）を実行して、ホストからのコマンドを受信し続けてエラーをリターンする。

【0045】

一部の実施形態で、

（i）実行空間（P）、

（ii）性能低下空間（UP）、

（iii）読み取り専用空間（RO）、

（iv）揮発性読み取り専用空間（VRO）と、

（v）アクセス不可能空間（IA）

と呼ばれる5つの論理ブロックアドレス（LBA）の空間タイプが定義される。

実行（P）空間は、性能の犠牲なしに通常の方法で読み書きできる有効なデータを含むLBA空間である。

性能低下（UP）空間は、通常の方法で読み書きすることができるが、性能が低下された（例えば、低下された書き込みの性能）、有効なデータを含むLBA空間である。

読み取り専用（RO）空間は、読み取り専用の有効なデータを含むLBA空間である。

ストレージ装置110は、このようなタイプのLBA空間に向かうホストからの書き込みコマンドに対してエラーメッセージで応答する。

読み取り専用空間のデータは、保存期間を超える期間の間、有効である。

揮発性読み取り専用（VRO）空間は、読み取り専用であり、ストレージ装置110は、このようなタイプのLBA空間に向かうホストからの書き込みコマンドに対してエラーメッセージで応答する。

このようなタイプのLBA空間上にあるデータは、一時的に有効であり、有効期限が切れるとき、つまり、ストレージ媒体125内のデータの経過時間がストレージ媒体125の保存期間に達すると、無効になる。

アクセス不可能（IA）空間は、ホストからアクセスができない、有効ではないデータを含むLBA空間である。

【0046】

図2Bのテーブルは、図2Aのテーブルの1番目のコラムで識別され、障害状態の各々でストレージ装置110に存在することができるLBA空間のタイプの組み合わせを示す。

アスタリスク（*）で識別されるモードは、LBA空間に対する詳細な情報を検索するために、ホストが特徴獲得コマンド（get feature command）を使用することを支援する。

図2Cは、ストレージ装置110にその状態及び関連するパラメータの値についてストレージ装置110にクエリ（query）するために、ストレージ装置110に伝達される「特徴獲得」コマンドの引数（argument）を示す（「IOPS」は、毎秒入力及び出力の動作を意味する）。

【0047】

図2Dのテーブルは、ホスト105がストレージ装置110に、その状態についてクエリするために使用できるNVMe（non volatile memory express）コマンドと、ストレージ装置110が障害回復特徴のカテゴリ内で応答するために使用できるリストされた定数（enumerated constants）の階層構造とを示している。

【0048】

図3Aは、本発明の一実施形態による障害回復力のあるストレージ装置の動作の第1方法を説明するためのフローチャートである。

ステップS305で、ホスト105は、ストレージ装置からデータを送信又は受信し、

10

20

30

40

50

ステップS310で、ホスト105は、ストレージ装置でエラーが発生したかを判断し、ステップS315で、ストレージ装置110は、内部診断を実行してその障害復元状態（例えば、完全（full）復元、部分復元、又は脆弱モード）を判断し、ステップS320で、ストレージ装置110は、診断に基づいて、その性能、容量、機能（capability）（例えば、読み取り専用モードへの移行）を変更し、ステップS325で、ストレージ装置110は、アプリケーション・プログラミング・インターフェース（API）に基づいて、ホスト105からの要請に応じて状態を掲示し、ステップS330で、ホスト105は、上記状態に基づいて与えられた帯域幅で所定のタイプのデータをストレージ装置110又は他のストレージ装置110にルーティングする。

【0049】

図3Bは、本発明の一実施形態による障害回復力のあるストレージ装置の動作の第2方法を説明するためのフローチャートである。

上記方法は、ステップS335で、ストレージ装置により、ストレージ装置が電源をパワーサイクリング（power cycling）するか、又はストレージ媒体をフォーマットすることにより、回復が可能な第1障害状態にあることを判断する段階と、ステップS340で、ストレージ装置により、ストレージ装置が減少した性能で、減少した容量で、又は読み取り専用モードで、ストレージ装置を動作させることによって、部分復元可能な第2障害状態にあることを判断する段階と、ステップS345で、減少した性能で、減少した容量で、又は読み取り専用モードで、ストレージ装置を動作させる段階と、を備える。

【0050】

ここに開示した実施形態に対する多数の実施形態と変形が、また、構成され得る。

FPGA（Field Programmable Gate Array）やエンベデッドプロセッサは、内部ブロックのチェックを実行して、ストレージ装置110の状態に対する非同期アップデートをホスト105に送ることができる。

イベントが発生し、ホスト105に送信され得る（例えば、温度や装置内部の他のパラメータ）。

ホスト105は、通知を提供するための装置ドライバの特徴が存在しない場合、所定のスケジュールに基づいてストレージ装置110をポーリングすることができる。

FPGAやエンベデッドプロセッサは、ストレージ装置110の履歴性能をモニタリング（監視）し、マシンラーニングを使用して予測分析（例えば、与えられた障害回復状態にある可能性）を提供することができる。

コマンドがNVMe仕様で導入され得るが、例えば、NVMe仕様の遠隔測定情報が（ストレージ装置110の状態を報告するために）拡張され得る。

一部の実施形態は、例えば、イーサネット（登録商標）ストレージ装置、又は、キーバリュ（KV）ストレージ装置で具現化される。

【0051】

一部の実施形態で、ホストは、

（i）ストレージ装置110の状態（例えば、イメージ、ビデオ、テキスト、又は高い優先順位若しくは低い優先順位データのようなファイルタイプ）に基づいて異なるデータタイプを送信して（例えば、部分的に脆弱なモードで考えられている装置には、高い優先順位データやリアルタイムデータが記録されないことがあり得る）、

（ii）ストレージ装置110が部分的に脆弱な状態で、より低い性能状態にある場合、送信速度を減少させ、

（iii）ストレージ装置110が部分的に脆弱で、低い容量状態にある場合、減少されたデータの総量を送信し、

（iv）ストレージ装置110が部分的に脆弱で持続不可能な読み取り専用モードにある場合、最大速度でデータを読み取り、データを別の場所に格納して保存期間を超えないようにし（このような状況では、ホストは、コピーするデータの量と保存期間に基づいて、必要なデータレートを計算することができる）、

10

20

30

40

50

(v) データがエラーを有し得るので、脆弱なストレージ装置 110 からデータの「読み取り」を無視し、ホスト 105 によって受信されたデータを簡単に削除し、

(vi) ホストとストレージ装置 110 との間でこのようなイベントのタイミングを制御するメッセージに基づいて、パワーサイクリング又はフォーマットされた完全復元ストレージ装置 110 のキャッシュに読み取り/書き込みの入力及び出力を一時的に再ルーティングする。

【0052】

容量が減少した部分的に脆弱な SSD 上の FPGA は、入ってくるデータの書き込みをフィルタリングしてそのデータの一部だけをストレージ装置 110 に書き込む。

いくつかの場合では、フィルタリングは、圧縮を含み得る。

10

このような FPGA は、ホスト 105 から多様な種類のデータ（例えば、イメージ、ビデオ、テキスト、又は高い優先順位若しくは低い優先順位データのようなファイルタイプ）を受信してストレージ装置 110 の状態に応じてフィルタリングする。

例えば、FPGA は、部分的に脆弱なモードにあるストレージ装置 110 に高い優先順位のデータが記録されてはいけなと判断（決定）する。

FPGA は、拒否メッセージをホスト 105 に送信し、拒否の理由を提供する。

代替的には、FPGA は、部分復元の低容量の状態ストレージ装置 110 に記録するための特定のタイプのデータ（例えば、イメージデータ）をフィルタリングする。

例えば、ストレージ装置 110 が、性能を失う場合（たとえば、減少された書き込み速度で動作）、レイテンシに敏感な（latency-sensitive）読み取りと書き込みが拒否され得る。

20

【0053】

一部の実施形態で、前述したように、ストレージ装置 110 は、「持続可能な」（又は「第 1」）読み取り専用モードと「持続不可能な」（又は「第 2」）読み取り専用モードと呼ばれる 2 つの読み取り専用モードの内のいずれか 1 つで動作する。

持続不可能な読み取り専用モードで、ストレージ装置 110 は、ホスト 105 から受信することができる任意の書き込みコマンドを単に拒否し、

(i) 有効期限が切れていないデータに対する読み取りコマンドを実行し、

(ii) すべての読み取りコマンドを実行しようと試み、それに応じて、デコーディングが成功しなかった場合、成功的なデコーディングの試みによって発生したデータとエラーコード（データ項目が読み取られないことを示すインディケーション（indication）、例えば、「0」）をリターンする。

30

【0054】

持続可能な読み取り専用モードで、ストレージ装置 110 は、ストレージ媒体 125 の保存期間を超えて、外部の読み取り要請に対してサービスを提供する。

ストレージ装置 110 は、ホスト 105 からの任意の書き込みコマンドにตอบสนองして、エラーメッセージをリターンすることを除いては、いかなる措置も取らないが、以下で説明するように、内部目的のためにストレージ媒体 125 に書き込みを行う。

保存期間後のデータ保存を可能にするためには、ストレージ装置 110 は、期限切れ期限が間近に迫ったデータ（例えば、保存媒体 125 の保存期間に基づいてしきい値経過時間（例えば、ストレージ媒体 125 の保存期間の 80%）を超える経過時間を有するデータ）を移動する空間に「レスキュースペース」と呼ばれる空間を割り当て、それに応じてデータの期限切れは、レスキュースペースの保存期間だけ延期される。

40

レスキュースペースのデータは、有効期限が切れる予定の場合、データの損失を防止するために必要に応じて、頻繁にレスキュースペース内で再び移動する。

いくつかの実施形態で、ストレージ装置 110 は、例えば、このような移動を実行するために低負荷（low load）期間を利用して、期限切れ前に、データをよく移動させる。

前述したように、持続可能な読み取り専用モードへの移行は、不良ブロック管理の予備空間の枯渇によりトリガーされることがある。

50

このような状況では、影響を受けるブレーン又はダイは、読み取り専用になり、ここに格納されたすべてのデータがレスキュースペースに再配置されると、ストレージ装置 110 の物理アドレスの空間から除外され得る。

【0055】

ストレージ装置 110 は、全体の LBA 空間にわたって、ホスト 105 からの読み出しコマンドに応答する。

ストレージ装置 110 の物理的空間が減少することができるが（例えば、ストレージ媒体 125 の一部が読み取り専用になって、そのデータがレスキュースペースに移動されているため（又は移動されたため））、装置の論理空間は変更されないため、ホスト 105 が PBA (Physical Block Address) 空間にマッピングすることができない LBA 空間内で読み取りコマンドを送信すると、ストレージ装置 110 は、0 (zero) をリターンする。

10

【0056】

レスキュースペースは、オーバープロビジョニング空間、又はストレージ装置 110 の未使用空間から割り当てられる。

IO 決定論のセット又はゾーニングされた名前空間のゾーンが書き込み可能でない場合、すべてのデータが再配置されると、そのセット又はゾーンが読み取り専用になって物理アドレスの空間から除外され得る。

レスキュースペースは、名前空間を使用して、動的に生成されうる。

1 つ以上の名前空間がユーザーデータに使用され得、追加の名前空間がレスキュースペースのために生成され得る。

20

ストレージ装置 110 が、持続可能な読み取り専用モードに切り替えるとき、ユーザーデータののための 1 つ以上の名前空間は、サイズが減少させ、レスキュースペースのための名前空間のサイズが増加する。

【0057】

障害に影響を受けるストレージ媒体 125 の領域（例えば、不良ブロック管理の予備空間が使いつくされた領域）のサイズが、利用可能なレスキュースペースのサイズを超えると、ストレージ装置 110 は、最も早く期限切れになるデータ（例えば、最も古いデータ）だけを移動するか、重要な、又は所定のファイルタイプ（例えば、イメージ、ビデオ、ドキュメント）に関連するものとしてタグ付けされたデータだけを移動するか、又は、隣接するタイプ (contiguous type) のデータ（例えば、ファイルに関連するすべてのデータ）を移動し、他のデータは残しておく。

30

重要であると考えられているファイルタイプのセットは、ユーザーが構成する。

KV ストレージ装置 110 を使用している一部の実施形態で、データの重要度は、キーバリューのペア (key value pairs) で圧縮又はエンコーディングされる。

【0058】

一部の実施形態で、ストレージ装置 110 は、利用可能なレスキュースペースの量がしきい値サイズよりも小さく、それによって、持続可能な読み取り専用モードでの動作をサポートするのに不十分であると判断（決定）されると、その結果、ストレージ装置 110 は、持続不可能な読み取り専用モードに切り替える。

40

このような判断（決定）は、障害が最初に感知されたとき、行われ得、又は、障害が最初に感知されたとき、ストレージ装置 110 が最初に持続可能な読み取り専用モードに切り替え、一定時間後にレスキュースペースが、利用可能なレスキュースペースのしきい値サイズ未満に減少するほど消費され（例えば、ストレージ媒体 125 の障害部分でのデータの継続的な期限切れによって）、ストレージ装置 110 は、持続可能な読み取り専用モードから持続不可能な読み取り専用モードに切り替える。

持続可能な読み取り専用モードから持続不可能な読み取り専用モードに切り替えるとき、ストレージ装置 110 は、ホスト 105 に、このような転換を通知し、ホスト 105 は、データが期限切れになる前に、ストレージ装置から他のストレージ装置に可能な限り多くのデータをコピーしようと試みる。

50

【 0 0 5 9 】

場合によっては、データは、一日や1時間の期間にわたって失われる可能性がある。

従って、ストレージ装置 1 1 0 の F P G A は、ストレージ装置 1 1 0 が持続可能な読み取り専用モードに切り替えるときに、移転 (t r a n s f e r) を容易にするために、ドライブ上のデータの属性 (例えば、古くなった、タグ付けされた、連続の、又はファイルタイプの) をモニタリング (監視) する。

F P G A は、優先順位に基づいてデータをキュー (q u e u e) に入れるか、又はレスキュースペースに移転するデータを整理する (o r d e r) ことができ、データがレスキュースペースに移動するときに、データを圧縮することができる。一部の実施形態は、例えば、イーサネット (登録商標) ストレージ装置、又はキーバリュ (K V) ストレージ装置で具現化される。

10

【 0 0 6 0 】

図 4 は、本発明の一実施形態による、ストレージ装置 1 1 0 の概略的なデータレイアウトを説明するための図である。

ブロックの寿命テーブル 4 0 5 は、ストレージ装置 1 1 0 の各ブロックの保存期間をリストし、マッピングテーブル 4 1 0 は、論理ブロックアドレスから物理ブロックアドレスへのマッピングをリストする。

ストレージ装置 1 1 0 が、持続可能な読み取り専用モードに切り替わるときに、ストレージ装置 1 1 0 は、障害のある使用空間 4 1 5 (障害状態によって影響を受ける物理空間) のデータをレスキュースペース 4 2 0 にコピーする。

20

通常の使用空間 4 2 5 のデータは、ストレージ装置 1 1 0 が (持続可能な読み取り専用モードにある代わりに) 通常の動作モードにあったかのように処理される。

障害状態は、物理空間の損失をもたらすので、使用された L B A 空間 4 3 0 のサイズでの対応する減少があり、以前に利用可能な L B A 空間の一部は、未使用の L B A 空間 4 3 5 となる。

コントローラ 1 2 0 によって実行されるファームウェア 4 4 0 は、前述したように、データが移動され、テーブルがアップデートされる。

【 0 0 6 1 】

図 5 A は、本発明の一実施形態で、持続可能な障害回復の読み取り専用モードでストレージ装置の動作の第 1 方法を説明するためのフローチャートである。

30

ステップ S 5 0 5 で、ホスト 1 0 5 は、ストレージ装置からデータを送受信し、ステップ S 5 1 0 で、ホスト 1 0 5 は、ストレージ装置でエラーが発生したかを決定し、ステップ S 5 1 5 で、ストレージ装置 1 1 0 は内部診断実行して持続可能な読み取り専用モードに転換し、ステップ S 5 2 0 で、F P G A はストレージ装置 1 1 0 上の有効なデータ属性 (n a t u r e) を (たとえば、データタイプ又は経過時間に基づいて) 選択的にモニタリングし、ステップ S 5 2 5 で、ストレージ装置 1 1 0 は、データの経過時間が保存期間に近づくと有効なデータをレスキュースペースに再配置し、ステップ S 5 3 0 で、ストレージ装置 1 1 0 は、レスキュースペースの状態と内容に対する情報を、データをフラッシュ (f l u s h) するホスト 1 0 5 に選択的に送る。

【 0 0 6 2 】

40

図 5 B は、本発明の一実施形態による持続可能な障害回復の読み取り専用モードでストレージ装置の動作の第 2 方法を説明するためのフローチャートである。

上記方法は、ステップ S 5 3 5 において、ストレージ装置によって、第 1 読み取り専用モードでストレージ装置を動作させることにより、ストレージ装置が部分復元可能な障害の状態にあることを判断し、ステップ S 5 4 0 で、ストレージ装置のページに格納された第 1 データ項目の経過時間がしきい値経過時間を超えたことを判断し、第 1 データ項目をストレージ装置のレスキュースペースにコピーすることにより、ストレージ装置を第 1 読み取り専用モードで動作させる。

【 0 0 6 3 】

一部の実施形態で、前述したように、ストレージ装置 1 1 0 のアレイとボリュームマネ

50

ージャ－１１５を含む「ＲＡＩＤ－０」システムは、「ＲＡＩＤ－０」システムのストレージ装置１１０のいずれか１つを読み取り専用モード（例えば、持続可能な読み取り専用モード）への移行に対応するように構成される。

通常の動作で、ボリュームマネージャ－１１５は、ストレージ装置１１０のアレイにかけてデータをストライピングする（*stripping*）責任があるが、例えば、ストレージ装置１１０のアレイのそれぞれのストレージ装置１１０に各ストライプ（各ストライプは、このようなストリップで構成される）のいずれか１つのストリップを書き込む責任がある。

【００６４】

このようなシステムでは、ストレージ装置１１０のアレイの内の任意のストレージ装置は、読み取り専用モード（又は「読み取り専用状態」）に移行するとき、「ＲＡＩＤ－０」システムは、動作モード（「エマージェンシーモード」と呼ばれることがある）に転換し、ストレージ装置１１０のアレイのためのボリュームマネージャ－１１５は、

（*i*）障害のあるストレージ装置１１０からメタデータ及び回復されたユーザーデータに対して残っている、影響を受けないストレージ装置１１０（読み取り／書き込みの状態にある）のそれぞれにレスキュースペースを割り当て、

（*ii*）マッピングテーブル（「エマージェンシーマッピングテーブル」と呼ばれることがある）を維持する。

各ストレージ装置１１０上のレスキュースペースは、 n 個のストリップを格納することができ、ここで、

$n = R / (\text{ストリップサイズ})$ であり、

$R = C / M$ であり、

C は、ストレージ装置１１０のアレイのストレージ装置のそれぞれの容量であり、 M は、ストレージ装置のアレイでストレージ装置の個数である。

【００６５】

ボリュームマネージャ－１１５は、ホストで（例えば、ハードウェア、ソフトウェア、ファームウェア、又はハードウェア、ソフトウェア、及びファームウェアの組み合わせで）、又は「ＲＡＩＤ－０」システムのレイドコントローラ（例えば、ホストとは別途のエンクロージャー（筐体）に収容されることがある）で具現化される。

一実施形態では、ボリュームマネージャ－１１５は、ＦＰＧＡで具現化される。

「ＲＡＩＤ－０」システムは、自己完結型であり得、ホストの観点から、「ＲＡＩＤ－０」システムが単一のストレージ装置１１０のように動作するようにストレージ装置１１０のアレイを仮想化する。

一般的に、ボリュームマネージャ－は、ボリュームマネージャ－によって実行されるものとして、ここで説明する動作を実行するように（例えば、適切なソフトウェア又はファームウェアによって）構成される処理回路（以下で、詳細に説明する）であり得る。

【００６６】

「ＲＡＩＤ－０」システムがエマージェンシーモードで動作しており、ホスト１０５から書き込みコマンドを受信し、ストレージ装置１１０のアレイにストライプを書き込むことを要求する場合、ボリュームマネージャ－１１５は、緊急（エマージェンシー）時のマッピングテーブルを確認してストライプが「登録されたか」を判断し、すなわち、エントリがストライプに対して既に作成されたか否かを判断する。

まだエントリが生成されていない場合（つまり、ストライプが「オープンマッピングされた（*open mapped*）」、つまり登録されていない場合）、ボリュームマネージャ－１１５は、障害のあるストレージ装置１１０（読み取り専用モードに転換されたストレージ装置１１０）に一般的に書かれたストリップが、書かれた位置をインディケート（指示）するためにエマージェンシーマッピングテーブルにエントリを生成する。

エマージェンシーマッピングテーブルが、既にストライプに対するエントリを含む場合、エントリは、一般的に障害のあるストレージ装置１１０に書かれたストリップが書き込む位置を決定するために使用される。

10

20

30

40

50

どのような場合でも、ボリュームマネージャー 115 は、図 6 A に示すように、ストレージ装置 110 のアレイにそれぞれのストリップを記録することで、障害のあるストレージ装置 110 に一般的に書かれた、ストリップ 605 を他のストレージ装置 110 のレスキュースペースに記録する。

【0067】

読み取りコマンドが受信されてストレージ装置 110 のアレイからストライプを読み取ることを要求する場合、ボリュームマネージャー 115 は、エマージェンシー（緊急）時のマッピングテーブルを確認して、エントリがストライプに対して作成されたのか否かを判断する。

エントリが全く作成されていない場合、図 6 B に示すように、ボリュームマネージャー 115 は、正常な動作で障害のあるストレージ装置 110 を含む、それぞれのストレージ装置 110 からストリップを読み取るようにストライプを読み取る。

エマージェンシーマッピングテーブルがストライプに対するエントリが含む場合、エントリは、一般的に障害のあるストレージ装置 110 から読み込まれるストリップが読み取る位置を決定するのに使用される。

【0068】

一般的に、障害のあるストレージ装置 110 に書き込まれるストリップの再マッピングは、例えば、次のように達成される。

ストレージ装置 110 のアレイの各ストレージ装置 110 は、「0」と「M-1」との間の数字であり得るドライブの識別番号（又は「ドライブ ID」）を有し、ここで、M はストレージ装置 110 のアレイでストレージ装置 110 の個数である。

ボリュームマネージャー 115 は、ドライブの識別番号を再び割り当てるが、例えば、登録されたストライプ（登録されていないストライプに対する読み取り動作は、元のドライブの識別番号を使用し続ける）への読み取り又は書き込みの動作を実行するのに使用される代替のドライブの識別番号をストレージ装置 110 のアレイの各ストレージ装置 110 に割り当てる。

【0069】

以下の式を使用して代替のドライブの識別番号を生成することができる。

（ドライブ ID）<（障害のあるドライブ ID）である場合

新しいドライブ ID = ドライブ ID

そうでなければ、

新しいドライブ ID = $((\text{ドライブ ID} - 1) + (M - 1)) \bmod (M - 1)$

【0070】

この式の効果に応じて、

（i）障害のあるストレージ装置の元のドライブの識別番号よりも小さい識別番号を有する各ストレージ装置にそれぞれの元のドライブの識別番号を割り当て、

（ii）第 1 ストレージ装置の元ドライブの識別番号よりも大きい識別番号を有する各ストレージ装置に「1」よりも小さいそれぞれの元のドライブの識別番号を割り当てる。

【0071】

代替のドライブ番号を使用して、一般的に障害のあるストレージ装置 110 に書かれたストリップが書かれることができるターゲットドライブは、

式 “ターゲットドライブ ID = $s i d \% (M - 1)$ ”

を使用して識別され、ここでターゲットドライブ ID は、ターゲットドライブの代替のドライブの識別番号であり、「s i d」は、一般的に障害のあるストレージ装置 110 に書かれたストリップのストリップ識別子であり、「%」は、モジュロ演算子（modulo operator）である。

【0072】

図 6 C は、本発明の一実施形態による再マッピング（remapping）と書き込み動作を実行する「RAID-0」システムの概略的なデータレイアウト図であり、ドライブ 1 が読み取り専用モードに転換された 4 つのストレージ装置 110（すなわち、M = 4

10

20

30

40

50

)を有する「RAID - 0」システムの概略図である。

ストライプ1が書き込まれるターゲットドライブIDは、以下の式によって暗示的に決定される。

$$\text{ターゲットドライブID} = \text{ストリップID} \% (M - 1) = 1$$

【0073】

つまり、ターゲットドライブは、代替のドライブの識別番号1（つまり、ドライブ2）を有するストレージ装置110である。

ドライブ内において、レスキュースペースは、ストリップサイズと同じサイズのストリップ（「レスキューストリップ」又は「R - ストリップ」と呼ばれる）に分割される。

エマージェンシーマッピングテーブルは、エントリ（1, 0）（1番目のエレメント（要素）がストライプID（ストライプ1）であり、2番目の要素がターゲットドライブのR - ストリップIDである）を含む。

したがって、エマージェンシーマッピングテーブルのエントリ（1, 0）は、ストリップ（1, 1）がR - ストリップ（1, 0）にマッピングされたことを意味する。

【0074】

「RAID - 0」システムは、ストレージ装置110のアレイの内の1つ以上によって、持続可能な読み取り専用モード又は持続不可能な読み取り専用モードへの移行に対応するように構成される。

持続不可能な読み取り専用モードに切り替える場合、障害のあるストレージ装置110の登録されたストリップは、他の装置のレスキュースペースに永続的に書かれる。

持続不可能な読み取り専用モードに切り替える場合は、例えば、障害のあるストレージ装置からすべてのデータを移動するのに十分な速度で、ボリュームマネージャー115は、障害のあるストレージ装置110から影響を受けないストレージ装置110に、データの有効期限が切れる前にデータを移行する。

サーバーは、データ量とデータ期限切れ前の残り時間を基準に速度を計算する。

【0075】

図7Aは、本発明の一実施形態による「RAID - 0」ストレージシステムの動作の第1方法を説明するためのフローチャートである。

ステップS705で、「RAID - 0」システムのストレージ装置110に障害があり読み取り専用モードに切り替え、ステップS710で、影響を受けるストレージ装置110は内部診断を実行し、その障害回復状態が部分的に復元され、読み取り専用であると判断し、ステップS715で、ボリュームマネージャー115は影響を受けるストレージ装置110が読み取り専用モードにあると判断して（「正常に機能する」）、影響を受けないストレージ装置のIDを再び割り当て、ステップS720で、ボリュームマネージャー115は書き込み動作を受信し、影響を受ける装置のストリップがターゲット（影響を受けない）ストレージ装置110にリダイレクトされ（*redirected*）、全体のストリップが影響を受けないストレージ装置の新しいドライブIDをベースにターゲット（影響を受けない）のレスキュースペースに書かれたことを示すように、一時的マッピングテーブルに対するエントリを追加し、ステップS725で、ボリュームマネージャー115は、ホスト105から読み取りコマンドを受信し、影響を受けるストレージ装置のストリップをターゲット（影響を受けない）ストレージ装置のレスキュースペースから読み取りながらRAIDシステムの正常に機能する、影響を受けないストレージ装置110からストライプのすべてのストリップを読み取る。

【0076】

図7Bは、本発明の一実施形態による「RAID - 0」ストレージシステムの動作の第2方法を説明するためのフローチャートである。

上記方法は、ステップS730で、第1ストレージ装置が読み取り専用の状態にあり、第2ストレージ装置が読み取り/書き込みの状態にあると判断し、ステップS735で、第1ストライプの一部を第2ストレージ装置に書き込み、第1ストライプに対するマッピングテーブルにエントリを生成することにより、ストレージシステムに対する第1ストラ

10

20

30

40

50

イブの書き込み動作を実行し、ステップ S 7 4 0 で、第 1 ストレージ装置及び第 2 ストレージ装置から第 2 ストライプの一部を読み取ることにより、ストレージシステムからの第 2 ストライプの第 1 読み取り動作を実行し、ステップ S 7 4 5 で、マッピングテーブルが第 1 ストライプに対するエントリを含むと判断し、第 2 ストレージ装置から第 1 ストライプの一部を読み取ることにより、ストレージシステムから第 1 ストライプの第 2 読み取り動作を実行する。

【 0 0 7 7 】

ホスト 1 0 5 又はコントローラ 1 2 0 のような本明細書で説明した、多様な装置は、処理回路であるか、又は処理回路を含み得る。

本明細書では、用語「処理回路」又は「コントローラ手段」は、データ又はデジタル信号を処理するのに採用されるハードウェア、ファームウェア、及びソフトウェアの任意の組み合わせを意味するために使用される。

処理回路のハードウェアには、例えば、特定用途向け集積回路 (A S I C)、汎用又は特殊目的の中央処理装置 (C P U)、デジタル信号プロセッサ (D S P)、グラフィックス処理装置 (G P U)、及びフィールドプログラマブルゲートアレイ (F P G A) のようなプログラマブルロジック装置を含み得る。

処理回路で、本明細書で使用しているように、各機能は、その機能を実行するように構成され、すなわち、ハードワイヤードの (h a r d - w i r e d) ハードウェアによって、又は非一時的ストレージ媒体に格納されたコマンドを実行するように構成される C P U のような、より汎用のハードウェアによって実行される。

処理回路は、単一の印刷回路基板 (P C B) 上で製作されるか、又は 1 つ以上の相互接続される P C B の上で分布される。

処理回路は、他の処理回路を含み、例えば、処理回路は、P C B 上の相互接続される 2 つの処理回路、すなわち、F P G A、C P U を含み得る。

【 0 0 7 8 】

ストレージ装置の復元操作 (動作) のためのシステム及び方法の例としての実施形態を本明細書で具体的に説明及び図に示した。

しかし、本発明は、上述の実施形態に限られるものではない。本発明の技術的範囲から逸脱しない範囲内で多様に変更実施することが可能である。

【 符号の説明 】

【 0 0 7 9 】

- 1 0 5 ホスト
- 1 1 0 ストレージ装置
- 1 1 5 ボリュームマネージャー
- 1 2 0 コントローラ
- 1 2 5 ストレージ媒体
- 4 0 5 ブロックの寿命テーブル
- 4 1 0 マッピングテーブル
- 4 1 5 障害のある使用空間
- 4 2 0 レスキュースペース
- 4 2 5 通常の使用空間
- 4 3 0 使用された L B A 空間
- 4 3 5 未使用の L B A 空間
- 4 4 0 ファームウェア

10

20

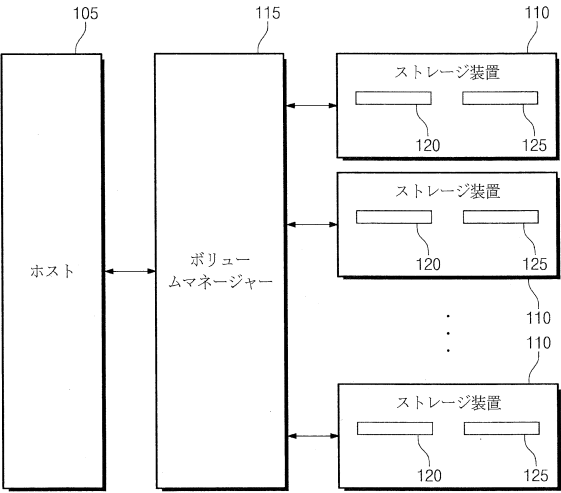
30

40

50

【図面】

【図 1】



【図 2 A】

ケース ID	動作状態	ユーザーデータ	再使用/方法
1	致命的エラー (フェイタルエラー)	無効	不可能
2	致命的エラー (フェイタルエラー)	無効	可能 フォーマットNVM
3	読み取り専用モード	有効	不可能
4	同時書き込みモード	有効	不可能
5	温度安全モード	有効	不可能
6	読み取り専用モード	有効	可能 フォーマットNVM

10

【図 2 B】

モード	P	UP	RO	VRO	IA
① FR_MODE_POWER_CYCLE	100%				
① FR_MODE_REFORMAT	100%				
① FR_MODE_CAPACITY_READ_ONLY*	X%		(100-X)%		
① FR_MODE_CAPACITY*	X%				(100-X)%
① FR_MODE_PERFORMANCE		100%			
① FR_MODE_READ_ONLY			100%		
① FR_MODE_PARTIAL_READ_ONLY*			X%		(100-X)%
① FR_MODE_TEMP_READ_ONLY*				100%	
① FR_MODE_TEMP_PARTIAL_READ_ONLY*				X%	(100-X)%
① FR_MODE_VULNERABLE					100%

【図 2 C】

コマンド	説明
FR_INFO_RESILIENCY_TYPE	故障の場合、障害復元性のタイプをリターンする。
FR_INFO_RETENTION_PERIOD	NAND媒体による再プログラミングなしに平均保存期間をリターンする。 この情報は、一時的読み取り専用モード (FR_MODE_TEMP_READ_ONLY) のために使用される。 又は一時的部分読み取り専用モード (FR_MODE_PARTIAL_READ_ONLY) のために使用される。 これは障害発生時点から装置にあるデータに対する上限の保存時間である。
FR_INFO_IOPS	データラングリティのため残っている最大時間をリターンする。 この情報は、一時的読み取り専用モード (FR_MODE_TEMP_READ_ONLY) のために使用される。 又は一時的部分読み取り専用モード (FR_MODE_PARTIAL_READ_ONLY) のために使用される。 これは障害発生時点から装置にあるデータに対する下限の保存時間である。 注意：時間単位はバトリール間隔に基づいて決定される。
FR_INFO_EARLIEST_EXPIRY	障害の場合、最大の使用可能なIOPSの百分率 (PERCENTAGE) をリターンする。 この情報は低い性能モード (FR_MODE_PERFORMANCE) のために使用される。
FR_INFO_EW	障害の場合、最大の使用可能なIOPSの百分率 (PERCENTAGE) をリターンする。 この情報は低い性能モード (FR_MODE_PERFORMANCE) のために使用される。
FR_INFO_SPACE	障害の場合、使用可能な空間をリターンする。 この情報は読み取り専用モードにおける低容量 (FR_MODE_CAPACITY_READ_ONLY) 又は低容量 (FR_MODE_CAPACITY) のタイプのために使用される。

20

30

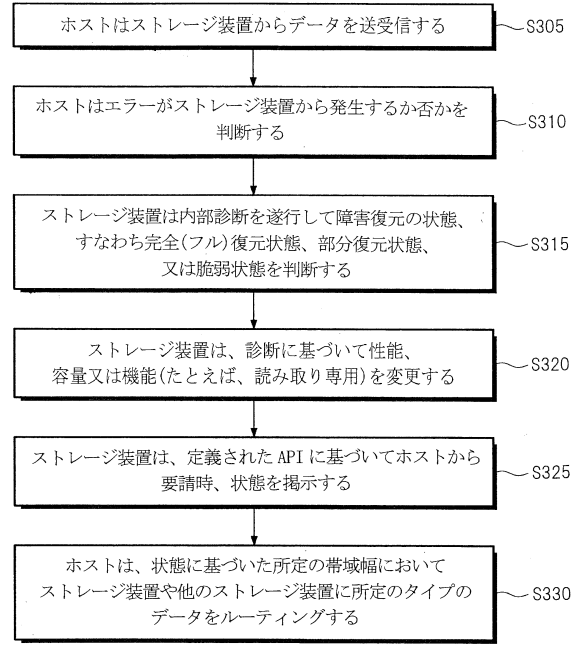
40

50

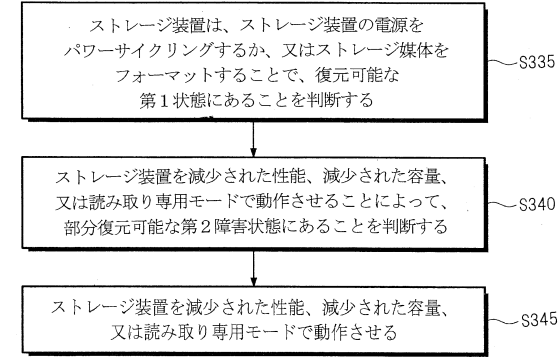
【図 2 D】

・FAULT_RESILIENT_FEATURE	これは、この装置がサポートする各クラスで障害復元クラス及び特徴をリターンする。 - FR_FULLY_RESILIENT - FR_POWER_CYCLE: 装置はパワーサイクリングにより復元される。 - FR_REFORMAT: 装置は再フォーマットすることで復元される。 - FR_PARTIAL_RESILIENT - FR_PERFORMANCE_LOSS: 装置の性能は減少される。 - FR_LOPS: IOPSは減少される。 - FR_BW: IOPSは減少される。 - FR_CAPACITY_LOSS: 装置の性能は減少される。 - FR_SUSTAINABLE_READ_ONLY: 装置はMAND保持時間より長く読み取り専用であり得る。 - FR_UNSTAINABLE_READ_ONLY: 装置はMAND保持時間まで読み取り専用であり得る。 - FR_SPACE_LOSS - FR_SPACE: 装置のストレージ空間は減少される。 - VULNERABLE - GRACEFUL_EXIT
・FAULT_RESILIENT_FEATURE	これは障害復元が行われた後、装置の状態をリターンする。 - FR_LOPS: 使用可能なIOPSは4%である。 - FR_BW: 使用可能なBWは1%である。 - FR_SPACE: 使用可能なストレージ空間は2%である。 - FR_RETENTION: この装置で使用された媒体の平均保持期間
・FAULT_RESILIENT_VOLATILE_BLOCKS (H)	これは、次の時間の間、保持期間に到達するBA範囲のリストをリターンする。 - これは持続不可能な読み取り専用のために再配置されるべきなブロックを決定するのに使用される。
・FAULT_RESILIENT_INVALID_DATA_BLOCKS	これは、障害復元モードに転換した後、無効化されたBA範囲のリストをリターンする。

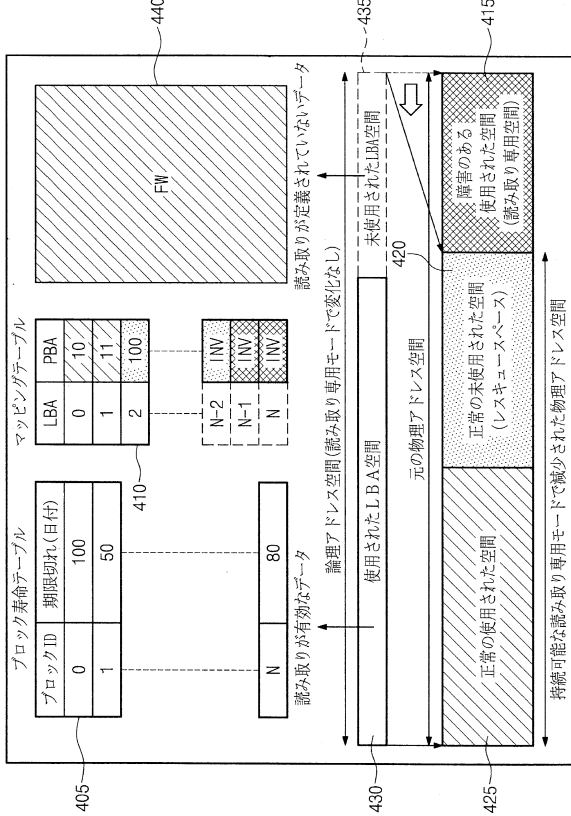
【図 3 A】



【図 3 B】



【図 4】



10

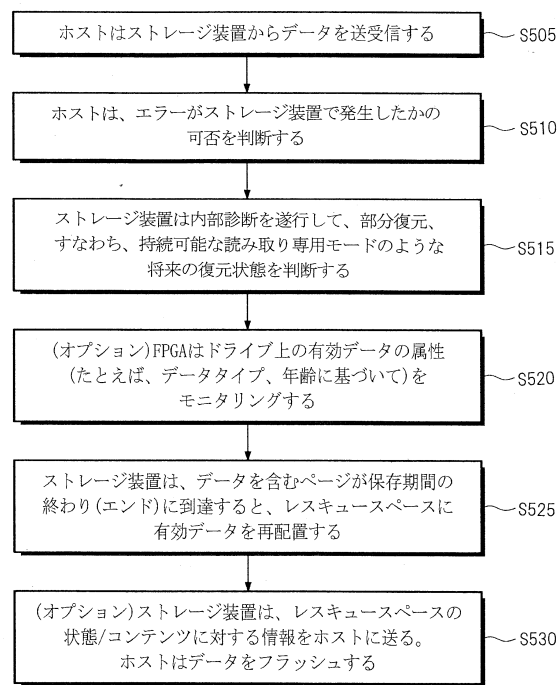
20

30

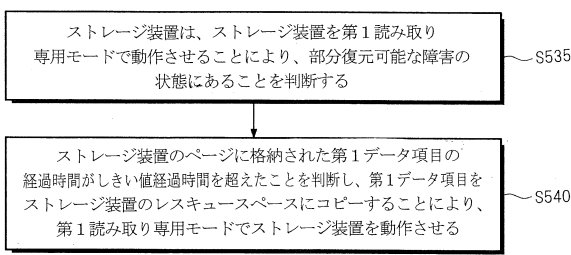
40

50

【図 5 A】



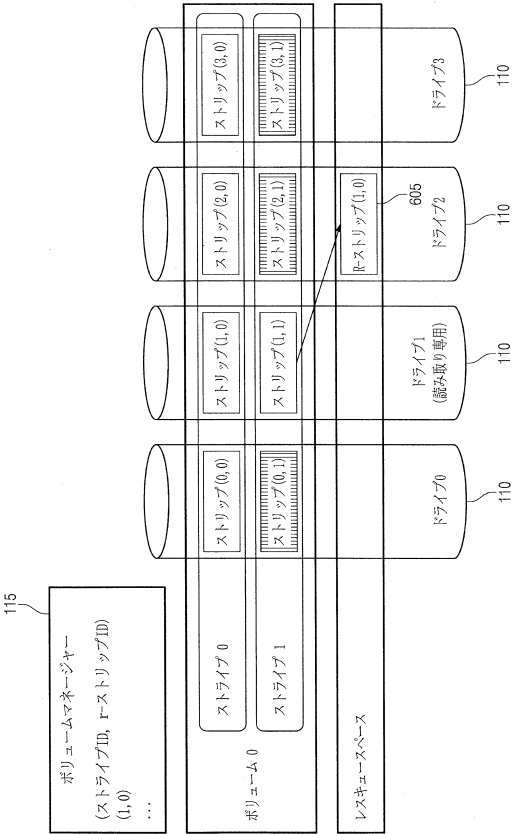
【図 5 B】



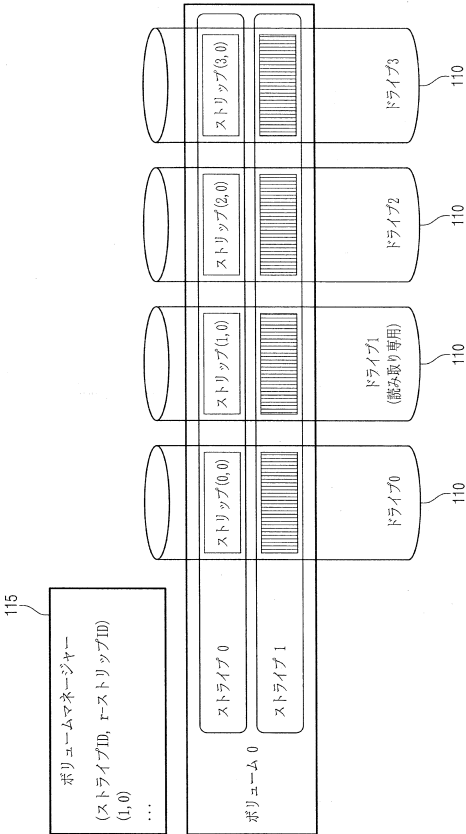
10

20

【図 6 A】



【図 6 B】

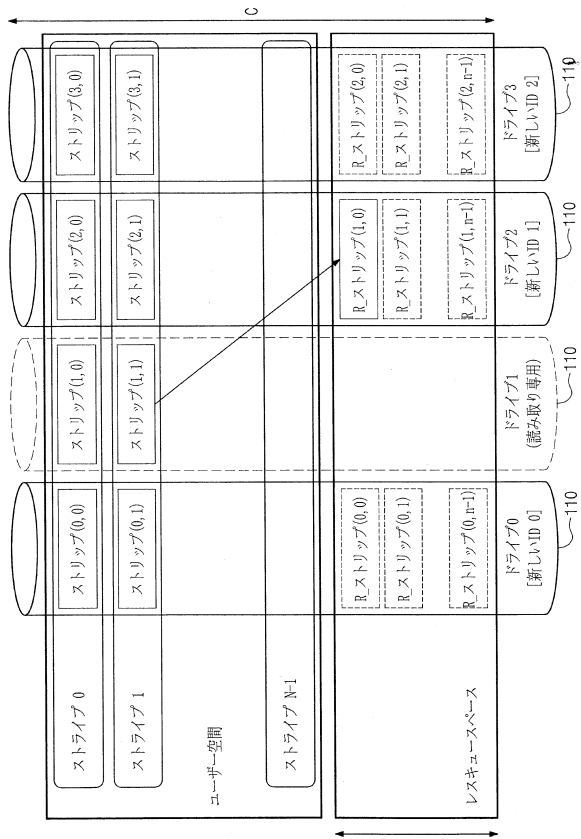


30

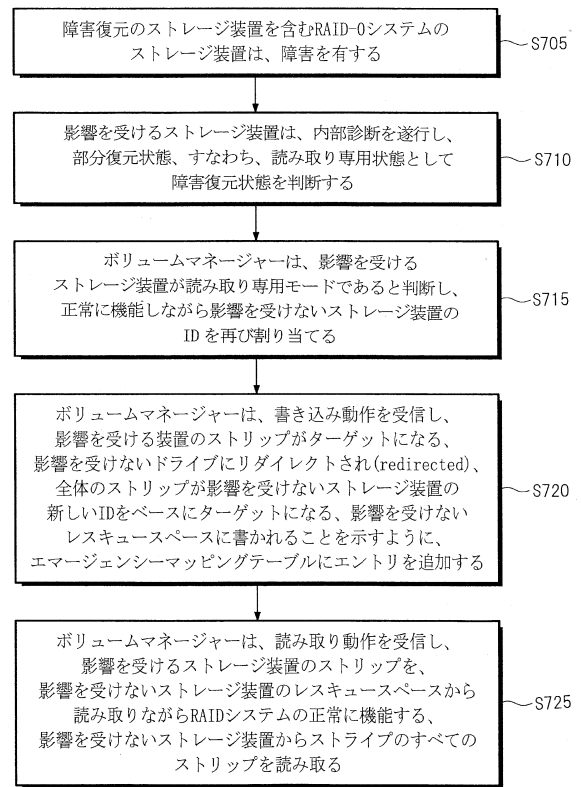
40

50

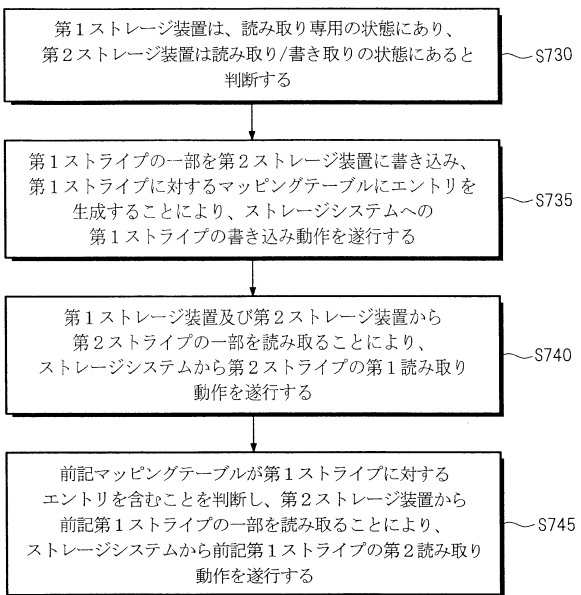
【図 6 C】



【図 7 A】



【図 7 B】



10

20

30

40

50

フロントページの続き

(51)国際特許分類 F I
G 0 6 F 13/10 3 4 0 A
G 0 6 F 16/11

(33)優先権主張国・地域又は機関

米国(US)

(31)優先権主張番号 17/109,041

(32)優先日 令和2年12月1日(2020.12.1)

(33)優先権主張国・地域又は機関

米国(US)

早期審査対象出願

オーク, 8 7 3

(72)発明者 李 東 起

アメリカ合衆国, カリフォルニア州 9 5 0 7 0, サラトガ, マリニー ブレイス, 2 0 0 5 9

(72)発明者 柳 星 旭

アメリカ合衆国, カリフォルニア州 9 4 3 0 3, パロアルト, オルテガ コート, 3 7 2 5

(72)発明者 エーサン ナジャファバディ

アメリカ合衆国, カリフォルニア州 9 5 1 3 4, サンノゼ, エスタンシア ドライブ, 1 8
5, ユニット 3 5 8

審査官 豊田 真弓

(56)参考文献 米国特許第 0 8 9 2 4 7 7 5 (US, B 1)
米国特許第 0 7 5 7 4 6 2 3 (US, B 1)
米国特許出願公開第 2 0 1 8 / 0 0 1 1 3 7 6 (US, A 1)
特開平 0 6 - 2 7 4 4 0 4 (JP, A)
特開 2 0 2 1 - 0 3 3 8 5 6 (JP, A)
特開 2 0 0 6 - 3 5 0 5 9 9 (JP, A)
特開平 1 1 - 1 8 4 6 4 4 (JP, A)
特開平 1 0 - 3 0 1 7 2 1 (JP, A)
特開平 0 6 - 1 7 5 7 9 2 (JP, A)

(58)調査した分野 (Int.Cl., DB 名)

G 0 6 F 3 / 0 6
G 0 6 F 1 3 / 1 0
G 0 6 F 1 6 / 1 1