

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号

特許第7086093号

(P7086093)

(45)発行日 令和4年6月17日(2022.6.17)

(24)登録日 令和4年6月9日(2022.6.9)

(51)国際特許分類

F I

G 0 6 F 16/11 (2019.01)

G 0 6 F 16/11

G 0 6 F 13/10 (2006.01)

G 0 6 F 13/10 3 4 0 A

G 0 6 F 3/06 (2006.01)

G 0 6 F 3/06 3 0 4 F

請求項の数 8 (全213頁)

(21)出願番号	特願2019-548937(P2019-548937)	(73)特許権者	511175211
(86)(22)出願日	平成30年1月31日(2018.1.31)		ピュア ストレージ, インコーポレイテッド
(65)公表番号	特表2020-514902(P2020-514902 A)		アメリカ合衆国 カリフォルニア 9 4 0 4 1-2 0 5 5, マウンテン ビュー, カストロ ストリート 6 5 0, スイート 4 0 0
(43)公表日	令和2年5月21日(2020.5.21)	(74)代理人	100079108
(86)国際出願番号	PCT/US2018/016125		弁理士 稲葉 良幸
(87)国際公開番号	WO2018/164782	(74)代理人	100109346
(87)国際公開日	平成30年9月13日(2018.9.13)		弁理士 大貫 敏史
審査請求日	令和3年2月1日(2021.2.1)	(74)代理人	100117189
(31)優先権主張番号	62/470,172		弁理士 江口 昭彦
(32)優先日	平成29年3月10日(2017.3.10)	(74)代理人	100134120
(33)優先権主張国・地域又は機関	米国(US)		弁理士 内藤 和彦
(31)優先権主張番号	62/502,060		
(32)優先日	平成29年5月5日(2017.5.5)		
最終頁に続く		最終頁に続く	

(54)【発明の名称】 データセット及び他の管理オブジェクトをクラウドベースのストレージシステムに同期複製すること

(57)【特許請求の範囲】

【請求項1】

データセットがその全体で同期複製される複数のストレージシステムであって、各ストレージシステムがコンピュータメモリ及びコンピュータプロセッサを含み、前記ストレージシステムのそれぞれの中の前記コンピュータメモリが、特定のストレージシステムの前記コンピュータプロセッサによる実行時に、前記特定のストレージシステムに、ポッドにアタッチするステップであって、前記ポッドが前記データセットと、管理オブジェクト及び管理動作の集合と、前記データセットを修正する又は読み取るためのアクセス動作の集合と、複数のメンバーストレージシステムとを含む、アタッチするステップを実施させるコンピュータプログラム命令を含み、管理動作が前記メンバーストレージシステムのいずれかを通して同等に管理オブジェクトを修正する又は問い合わせ、前記データセットを読み取る又は修正するためのアクセス動作が、前記メンバーストレージシステムのいずれかを通して同等に動作し、各メンバーストレージシステムが、前記ストレージシステムによる使用のために記憶され、告知された複数のデータセットの適切な部分集合として、前記データセットの別のコピーを記憶し、任意の1つのストレージシステムを通して実行され、完了された管理オブジェクト又は前記データセットを修正する動作が、前記ポッドを問い合わせるための後続の管理オブジェクト、又は前記データセットを読み取るための後続のアクセス動作で反映される、

複数のストレージシステム。

【請求項 2】

前記ストレージシステムの 1 つ以上が、特定のストレージシステムの前記コンピュータプロセッサによる実行時に、前記特定のストレージシステムに、
前記データセットの一部分を読み取る要求を受け取るステップと、
前記データセットの前記一部分を読み取る前記要求をローカルに処理するステップと
を実施させるコンピュータプログラム命令を含む、
請求項 1 に記載の複数のストレージシステム。

【請求項 3】

前記ストレージシステムの少なくとも 1 つが、クラウドサービスプロバイダによって提供されるクラウドストレージとして実施される、請求項 1 に記載の複数のストレージシステム。

10

【請求項 4】

複数のストレージシステム全体でデータセットを同期複製する方法であって、
前記複数のストレージシステムによって、ポッドにアタッチすることであって、前記ポッドが、前記データセットと、管理オブジェクト及び管理動作の集合と、前記データセットを修正する又は読み取るためのアクセス動作の集合と、複数のメンバーストレージシステムとを含む、アタッチすることを含み、
管理動作が前記メンバーストレージシステムのいずれかを通して同等に管理オブジェクトを修正する又は問い合わせ、
前記データセットを読み取る又は修正するためのアクセス動作が、前記メンバーストレージシステムのいずれかを通して同等に動作し、
各メンバーストレージシステムが、前記ストレージシステムによる使用のために記憶され、告知された複数のデータセットの適切な部分集合として、前記データセットの別のコピーを記憶し、
任意の 1 つのストレージシステムを通して実行され、完了された管理オブジェクト又は前記データセットを修正する動作が、前記ポッドを問い合わせるための後続の管理オブジェクト、又は前記データセットを読み取るための後続のアクセス動作で反映される、
方法。

20

【請求項 5】

前記複数のストレージシステムの 1 つである特定のストレージシステムによって、前記データセットの一部分を読み取る要求を受け取ることと、
前記特定のストレージシステムによって、前記データセットの前記一部分を読み取る前記要求をローカルに処理することと
をさらに含む、請求項 4 に記載の方法。

30

【請求項 6】

前記ストレージシステムの少なくとも 1 つが、クラウドサービスプロバイダによって提供されるクラウドストレージとして実施される、請求項 4 に記載の方法。

【請求項 7】

複数のストレージシステム全体でデータセットを同期複製するための装置であって、前記装置は、コンピュータプロセッサと、前記コンピュータプロセッサに動作可能なように結合されたコンピュータメモリであって、その中に、
前記コンピュータプロセッサによる実行時に、前記装置に、
ポッドにアタッチするステップであって、前記ポッドが、前記データセットと、管理オブジェクト及び管理動作の集合と、前記データセットを修正する又は読み取るためのアクセス動作の集合と、複数のメンバーストレージシステムを含む、アタッチするステップ
を実行させるコンピュータプログラム命令を配置させた前記コンピュータメモリとを備え、
管理動作が前記メンバーストレージシステムのいずれかを通して同等に管理オブジェクトを修正する又は問い合わせ、
前記データセットを読み取る又は修正するためのアクセス動作が、前記メンバーストレ

40

50

ジシステムのいずれかを通して同等に動作し、
各メンバーストレージシステムが、前記ストレージシステムによる使用のために記憶され、告知された複数のデータセットの適切な部分集合として、前記データセットの別のコピーを記憶し、
任意の1つのストレージシステムを通して実行され、完了された管理オブジェクト又は前記データセットを修正する動作が、前記ポッドを問い合わせるための後続の管理オブジェクト、又は前記データセットを読み取るための後続のアクセス動作で反映される、
装置。

【請求項 8】

前記コンピュータプロセッサによる実行時に、前記装置に、
前記データセットの一部分を読み取る要求を受け取るステップと、
前記データセットの前記一部分を読み取る前記要求をローカルに処理するステップと
を実施させるコンピュータプログラム命令をさらに備える、請求項 7 に記載の装置。

【発明の詳細な説明】

【図面の簡単な説明】

【0001】

【図 1 A】 【0001】 一部の実施態様に係るデータストレージのための第 1 の例のシステムを示す図である。

【図 1 B】 【0002】 一部の実施態様に係るデータストレージのための第 2 の例のシステムを示す図である。

【図 1 C】 【0003】 一部の実施態様に係るデータストレージのための第 3 の例のシステムを示す図である。

【図 1 D】 【0004】 一部の実施態様に係るデータストレージのための第 4 の例のシステムを示す図である。

【図 2 A】 【0005】 一部の実施形態に従って、複数のストレージノード、及びネットワーク接続ストレージを提供するために各ストレージノードに結合された内部ストレージを有するストレージクラスタの斜視図である。

【図 2 B】 【0006】 いくつかの実施形態に従って複数のストレージノードを結合する相互接続スイッチを示すブロック図である。

【図 2 C】 【0007】 いくつかの実施形態に係るストレージノードのコンテンツ及び不揮発性ソリッドステートストレージユニットの 1 つのコンテンツを示すマルチレベルブロック図である。

【図 2 D】 【0008】 いくつかの実施形態に係るいくつかの上記図のストレージノード及びストレージユニットの実施形態を使用するストレージサーバ環境を示す図である。

【図 2 E】 【0009】 いくつかの実施形態に従って、制御プレーン、計算プレーン、及びストレージプレーン、並びに基本的な物理リソースと対話するオーソリティを示す、ブレードハードウェアブロック図である。

【図 2 F】 【0010】 いくつかの実施形態に係るストレージクラスタのブレードでのエラスティシティ (elasticity) ソフトウェア層を示す図である。

【図 2 G】 【0011】 いくつかの実施形態に係る、オーソリティ、及びストレージクラスタのブレードでのストレージリソースを示す図である。

【図 3 A】 【0012】 本開示のいくつかの実施形態に従って、クラウドサービスプロバイダとのデータ通信のために結合されるストレージシステムの図を説明する図である。

【図 3 B】 【0013】 本開示のいくつかの実施形態に係るストレージシステムの図を説明する図である。

【図 4】 【0014】 本開示のいくつかの実施形態に係るポッドをサポートする複数のストレージシステムを示すブロック図を説明する図である。

【図 5】 【0015】 本開示のいくつかの実施形態に係るポッドをサポートする複数のストレージシステムを示すブロック図を説明する図である。

【図 6】 【0016】 本開示のいくつかの実施形態に係るポッドをサポートする複数のスト

10

20

30

40

50

レージシステムを示すブロック図を説明する図である。

【図 7】 [0017] 本開示のいくつかの実施形態に従って、2つ以上のストレージシステム間で同期複製関係を確立する例の方法を示すフローチャートを説明する図である。

【図 8】 [0018] 本開示のいくつかの実施形態に従って、2つ以上のストレージシステム間で同期複製関係を確立する追加の例の方法を示すフローチャートを説明する図である。

【図 9】 [0019] 本開示のいくつかの実施形態に従って、2つ以上のストレージシステム間で同期複製関係を確立する追加の例の方法を示すフローチャートを説明する図である。

【図 10】 [0020] 本開示のいくつかの実施形態に従って、2つ以上のストレージシステム間で同期複製関係を確立する追加の例の方法を示すフローチャートを説明する図である。

10

【図 11】 [0021] 本開示のいくつかの実施形態に従って、複数のストレージシステム全体で同期されるデータセットに向けられた I/O 動作にサービスを提供するための例の方法を示すフローチャートを説明する図である。

【図 12】 [0022] 本開示のいくつかの実施形態に従って、複数のストレージシステム全体で同期されるデータセットに向けられた I/O 動作にサービスを提供するための追加の例の方法を示すフローチャートを説明する図である。

【図 13】 [0023] 本開示のいくつかの実施形態に従って、複数のストレージシステム全体で同期されるデータセットに向けられた I/O 動作にサービスを提供するための追加の例の方法を示すフローチャートを説明する図である。

【図 14】 [0024] 本開示のいくつかの実施形態に従って、複数のストレージシステム全体で同期されるデータセットに向けられた I/O 動作にサービスを提供するための追加の例の方法を示すフローチャートを説明する図である。

20

【図 15】 [0025] 本開示のいくつかの実施形態に従って、データセットを同期複製するストレージシステム間で仲介するための例の方法を示すフローチャートを説明する図である。

【図 16】 [0026] 本開示のいくつかの実施形態に従って、データセットを同期複製するストレージシステム間で仲介するための例の方法を示すフローチャートを説明する図である。

【図 17】 [0027] 本開示のいくつかの実施形態に従って、データセットを同期複製するストレージシステム間で仲介するための例の方法を示すフローチャートを説明する図である。

30

【図 18】 [0028] 本開示のいくつかの実施形態に従って、データセットを同期複製するストレージシステムのための回復のための例の方法を示すフローチャートを説明する図である。

【図 19】 [0029] 本開示のいくつかの実施形態に従って、データセットを同期複製するストレージシステムのための回復のための例の方法を示すフローチャートを説明する図である。

【図 20】 [0030] 本開示のいくつかの実施形態に従って、データセットを同期複製するストレージシステムのための回復のための例の方法を示すフローチャートを説明する図である。

40

【図 21】 [0031] 本開示のいくつかの実施形態に従って、データセットを同期複製するストレージシステムのための再同期のための例の方法を示すフローチャートを説明する図である。

【図 22】 [0032] 本開示のいくつかの実施形態に従って、データセットを同期複製するストレージシステムのための再同期のための追加の例の方法を示すフローチャートを説明する図である。

【図 23】 [0033] 本開示のいくつかの実施形態に従って、データセットを同期複製するストレージシステムのための再同期のための追加の例の方法を示すフローチャートを説明する図である。

【図 24】 [0034] 本開示のいくつかの実施形態に従って、データセットを同期複製す

50

るストレージシステムのための再同期のための追加の例の方法を示すフローチャートを説明する図である。

【図 2 5】 [0035] 本開示のいくつかの実施形態に従って、データセットを同期複製するストレージシステムのための再同期のための追加の例の方法を示すフローチャートを説明する図である。

【図 2 6】 [0036] 本開示のいくつかの実施形態に従って、データセットを同期複製するストレージシステムのための再同期のための追加の例の方法を示すフローチャートを説明する図である。

【図 2 7】 [0037] 本開示のいくつかの実施形態に従って、データセットを同期複製するストレージシステムのための再同期のための追加の例の方法を示すフローチャートを説明する図である。

10

【図 2 8】 [0038] 本開示のいくつかの実施形態に従って、データセットを同期複製するストレージシステムのための再同期のための追加の例の方法を示すフローチャートを説明する図である。

【図 2 9】 [0039] 本開示のいくつかの実施形態に従って、データセットを同期複製するストレージシステムのための再同期のための追加の例の方法を示すフローチャートを説明する図である。

【図 3 0】 [0040] 本開示のいくつかの実施形態に従って同期複製されたストレージシステムへの接続性を管理するための例の方法を示すフローチャートを説明する図である。

【図 3 1】 [0041] 本開示のいくつかの実施形態に従って同期複製されたストレージシステムへの接続性を管理するための追加の例の方法を示すフローチャートを説明する図である。

20

【図 3 2】 [0042] 本開示のいくつかの実施形態に従って同期複製されたストレージシステムへの接続性を管理するための追加の例の方法を示すフローチャートを説明する図である。

【図 3 3】 [0043] 本開示のいくつかの実施形態に従って同期複製されたストレージシステムへの接続性を管理するための追加の例の方法を示すフローチャートを説明する図である。

【図 3 4】 [0044] 本開示のいくつかの実施形態に従って同期複製されたストレージシステムへの接続性を管理するための追加の例の方法を示すフローチャートを説明する図である。

30

【図 3 5】 [0045] 本開示のいくつかの実施形態に従って同期複製されたストレージシステムへの接続性を管理するための追加の例の方法を示すフローチャートを説明する図である。

【図 3 6】 [0046] 本開示のいくつかの実施形態に従って同期複製されたストレージシステムへの接続性を管理するための追加の例の方法を示すフローチャートを説明する図である。

【図 3 7】 [0047] 本開示のいくつかの実施形態に係る仲介サービスのための自動ストレージシステム構成のための例の方法を示すフローチャートを説明する図である。

【図 3 8】 [0048] 本開示のいくつかの実施形態に係る仲介サービスのための自動ストレージシステム構成のための例の方法を示すフローチャートを説明する図である。

40

【図 3 9】 [0049] 本開示のいくつかの実施形態に係る仲介サービスのための自動ストレージシステム構成のための例の方法を示すフローチャートを説明する図である。

【図 4 0】 [0050] 本開示のいくつかの実施形態に係る仲介サービスのための自動ストレージシステム構成のための例の方法を示すフローチャートを説明する図である。

【図 4 1】 [0051] 本開示のいくつかの実施形態に従って、ともにストレージデータの論理ボリューム又は論理ボリュームの一部を表す場合があるメタデータオブジェクトの構造化された集合体として実装されてよいメタデータ表現の図を説明する図である。

【図 4 2 A】 [0052] 本開示のいくつかの実施形態に従ってデータセットを同期複製するストレージシステムの間でメタデータを同期させるための例の方法を示すフローチャー

50

トを説明する図である。

【図 4 2 B】 [0053] 本開示のいくつかの実施形態に従ってデータセットを同期複製するストレージシステムの間でメタデータを同期させるための例の方法を示すフローチャートを説明する図である。

【図 4 3】 [0054] 本開示のいくつかの実施形態に従ってデータセットを同期複製するストレージシステムの間でアクティブメンバーシップを決定するための例の方法を示すフローチャートを説明する図である。

【図 4 4】 [0055] 本開示のいくつかの実施形態に従ってデータセットを同期複製するストレージシステムの間でアクティブメンバーシップを決定するための例の方法を示すフローチャートを説明する図である。

10

【図 4 5】 [0056] 本開示のいくつかの実施形態に従ってデータセットを同期複製するストレージシステムの間でアクティブメンバーシップを決定するための例の方法を示すフローチャートを説明する図である。

【図 4 6】 [0057] 本開示のいくつかの実施形態に従ってデータセットを同期複製するストレージシステムの間でアクティブメンバーシップを決定するための例の方法を示すフローチャートを説明する図である。

【図 4 7】 [0058] 本開示のいくつかの実施形態に従ってデータセットを同期複製するストレージシステムの間でアクティブメンバーシップを決定するための例の方法を示すフローチャートを説明する図である。

【図 4 8】 [0059] 本開示のいくつかの実施形態に従ってデータセットを同期複製するストレージシステムの間でメタデータを同期させるための例の方法を示すフローチャートを説明する図である。

20

【発明を実施するための形態】

【0002】

【0060】 本開示の実施形態に従って、データセット及び他の管理オブジェクトをクラウドベースのストレージシステムに同期複製するための例の方法、装置、及び製品が、図 1 A で始まる添付の図面を参照して説明される。図 1 A は、一部の実施態様に係るデータストレージのための例のシステムを示す。(本明細書では「ストレージシステム」とも呼ばれる) システム 100 は、制限よりむしろ説明の目的で多数の要素を含む。システム 100 が、他の実施態様において同じ又は異なる方法で構成された同じ要素、より多くの要素、又はより少ない要素を含んでよいことに留意されたい。

30

【0003】

【0061】 システム 100 は、いくつかのコンピューティングデバイス 164 A ~ 164 B を含む。例えばデータセンタ内のサーバ、ワークステーション、パーソナルコンピュータ、ノートブック等の(本明細書で「クライアントデバイス」とも呼ばれる)コンピューティングデバイスが実施されてよい。コンピューティングデバイス 164 A ~ 164 B は、ストレージエリアネットワーク(『SAN』) 158 又はローカルエリアネットワーク(『LAN』) 160 を通して、1つ以上のストレージレイ 102 A ~ 102 B へのデータ通信のために結合されてよい。

【0004】

40

【0062】 SAN 158 は、さまざまなデータ通信ファブリック、デバイス、及びプロトコルで実装されてよい。例えば、SAN 158 のためのファブリックは、ファイバチャネル、イーサネット、インフィニバンド、シリアルアタッチドスモールコンピュータシステムインタフェース(『SAS』)等を含んでよい。SAN 158 との使用のためのデータ通信プロトコルは、アドバンスドテクノロジーアタッチメント(『ATA』)、ファイバチャネルプロトコル、スモールコンピュータシステムインタフェース(『SCSI』)、インターネットスモールコンピュータシステムインタフェース(『iSCSI』)、HyperSCSI、不揮発性メモリエキスプレス(『NVMe』)オーバーファブリック(over fabrics)等を含んでよい。SAN 158 が、制限のためではなく説明のために提供されていることに留意されたい。他のデータ通信結合は、コンピューテ

50

ィングデバイス１６４Ａ～１６４Ｂとストレージアレイ１０２Ａ～１０２Ｂとの間で実装されてよい。

【０００５】

【００６３】 また、ＬＡＮ１６０は、さまざまなファブリック、デバイス、及びプロトコルで実装されてもよい。例えば、ＬＡＮ１６０のためのファブリックは、イーサネット（８０２．３）及び無線（８０２．１１）等を含んでよい。ＬＡＮ１６０での使用のためのデータ通信プロトコルは、トランスミッションコントロールプロトコル（『ＴＣＰ』）、ユーザデータグラムプロトコル（『ＵＤＰ』）、インターネットプロトコル（『ＩＰ』）、ハイパーテキスト転送プロトコル（『ＨＴＴＰ』）、ワイヤレスアクセスプロトコル（『ＷＡＰ』）、ハンドヘルド機器転送プロトコル（『ＨＤＴＰ』）、セッションイニシエーションプロトコル（『ＳＩＰ』）、リアルタイムプロトコル（『ＲＴＰ』）等を含んでよい。

10

【０００６】

【００６４】 ストレージアレイ１０２Ａ～１０２Ｂは、コンピューティングデバイス１６４Ａ～１６４Ｂに永続データストレージを提供する場合がある。実施態様では、ストレージアレイ１０２Ａは、シャシ（不図示）に含まれてよく、ストレージアレイ１０２Ｂは、別のシャシ（不図示）に含まれてよい。ストレージアレイ１０２Ａ及び１０２Ｂは、（本明細書では「コントローラ」とも呼ばれる）１つ以上のストレージアレイコントローラ１１０を含んでよい。ストレージアレイコントローラ１１０は、コンピュータハードウェア、コンピュータソフトウェア、又はコンピュータハードウェアとソフトウェアの組合せを含んだ自動化された計算機のカモジュールとして実施されてよい。一部の実施態様では、ストレージアレイコントローラ１１０は、多様なストレージタスクを実施するように構成されてよい。ストレージタスクは、コンピューティングデバイス１６４Ａ～１６４Ｂから受け取られたデータをストレージアレイ１０２Ａ～１０２Ｂに書き込むこと、ストレージアレイ１０２Ａ～１０２Ｂからデータを消去すること、ストレージアレイ１０２Ａ～１０２Ｂからデータを取り出し、データをコンピューティングデバイス１６４Ａ～１６４Ｂに提供すること、ディスクの活用及び性能をモニタし、報告すること、例えばレイド（『ＲＡＩＤ』）又はレイドのようなデータ冗長性動作等の冗長性動作を実行すること、データを圧縮すること、データを暗号化すること等を含んでよい。

20

【０００７】

【００６５】 ストレージアレイコントローラ１１０は、フィールドプログラマブルゲートアレイ（『ＦＰＧＡ』）、プログラマブルロジックチップ（『ＰＬＣ』）、特定用途向け集積回路（『ＡＳＩＣ』）、システムオンチップ（『ＳＯＣ』）、又は処理装置、中央演算処理装置、コンピュータメモリ、若しくは多様なアダプタ等の離散構成要素を含む任意のコンピューティングデバイスとして、を含んだ、さまざまな方法で実装されてよい。ストレージアレイコントローラ１１０は、例えば、ＳＡＮ１５８又はＬＡＮ１６０を介して通信をサポートするように構成されたデータ通信アダプタを含んでよい。一部の実施態様では、ストレージアレイコントローラ１１０は、ＬＡＮ１６０に独立して結合されてよい。実施態様では、ストレージアレイコントローラ１１０は、データ通信のためのストレージアレイコントローラ１１０を、中央平面（不図示）を通して（本明細書では「ストレージリソース」とも呼ばれる）永続記憶装置リソース１７０Ａ～１７０Ｂに結合するＩ／Ｏコントローラ等を含んでよい。永続記憶装置リソース１７０Ａ～１７０Ｂは、（本明細書では「ストレージデバイス」とも呼ばれる）任意の数のストレージドライブ１７１Ａ～１７１Ｆ、及び任意の数の不揮発性ランダムアクセスメモリ（『ＮＶＲＡＭ』）デバイス（不図示）を含んでよい。

30

40

【０００８】

【００６６】 一部の実施態様では、永続記憶装置リソース１７０Ａ～１７０ＢのＮＶＲＡＭデバイスは、ストレージアレイコントローラ１１０から、ストレージドライブ１７１Ａ～１７１Ｆに記憶されるデータを受け取るように構成されてよい。一部の例では、データは、コンピューティングデバイス１６４Ａ～１６４Ｂから生じる場合がある。一部の例では

50

、NVRAMデバイスにデータを書き込むことは、ストレージドライブ171A～171Fにデータを直接的に書き込むよりもより迅速に実施されてよい。実施態様では、ストレージアレイコントローラ110は、ストレージドライブ171A～171Fに書き込まれることが分かっているデータのための迅速にアクセス可能なバッファとしてNVRAMデバイスを活用するように構成されてよい。バッファとしてNVRAMデバイスを使用する書込み要求の待ち時間は、ストレージアレイコントローラ110がストレージドライブ171A～171Fに直接的にデータを書き込むシステムに比して改善されてよい。一部の実施態様では、NVRAMデバイスは、高帯域幅、低待ち時間RAMの形をとるコンピュータメモリで実装されてよい。NVRAMデバイスは、NVRAMデバイスに対する主電源損失後に、RAMの状態を維持する固有の電源を受け入れてよい、又は含んでよいため、「不揮発性」と呼ばれる。係る電源は、電池、1つ以上のコンデンサ等であってよい。電力損失に応じて、NVRAMデバイスは、RAMのコンテンツを、例えばストレージドライブ171A～171F等の永続記憶装置に書き込むように構成されてよい。

【0009】

【0067】 実施態様では、ストレージドライブ171A～171Fは、永続的にデータを記録するように構成された任意のデバイスを指す場合があり、「永続的に」又は「永続的な」は、電力損失後に記録されているデータを維持するデバイスの能力を指す。一部の実施態様では、ストレージドライブ171A～171Fは、非ディスクストレージ媒体に相当する場合がある。例えば、ストレージドライブ171A～171Fは、1つ以上のソリッドステートドライブ（『SSD』）、フラッシュメモリベースのストレージ、任意のタイプのソリッドステート不揮発性メモリ、又は任意の他のタイプの非機械的なストレージデバイスであってよい。他の実施態様では、ストレージドライブ171A～171Fは、例えばハードディスクドライブ（『HDD』）等の機械的なつまり回転するハードディスクを含む場合がある。

【0010】

【0068】 一部の実施態様では、ストレージアレイコントローラ110は、ストレージアレイ102A～102B内のストレージドライブ171A～171Fからデバイス管理責任を取り除くために構成されてよい。例えば、ストレージアレイコントローラ110は、ストレージドライブ171A～171Fの1つ以上のメモリブロックの状態を記述する場合がある制御情報を管理してよい。制御情報は、例えば、特定のメモリブロックが故障し、もはや書き込まれるべきではないこと、特定のメモリブロックがストレージアレイコントローラ110のためのブートコードを含むこと、特定のメモリブロックで実行されてきたプログラム—消去（『P/E』）サイクルの数、特定のメモリブロックに記憶されるデータの年齢、特定のメモリブロックに記憶されるデータのタイプ等を示してよい。一部の実施態様では、制御情報は、メタデータとして関連付けられたメモリブロックとともに記憶されてよい。他の実施形態では、ストレージドライブ171A～171Fの制御情報は、ストレージアレイコントローラ110によって選択されるストレージドライブ171A～171Fの1つ以上の特定のメモリブロックに記憶されてよい。選択されたメモリブロックは、選択されたメモリブロックが制御情報を含むことを示す識別子でタグ付けされてよい。識別子は、制御情報を含むメモリブロックを迅速に識別するためにストレージドライブ171A～171Fと併せてストレージアレイコントローラ110によって活用されてよい。例えば、ストレージコントローラ110は、制御情報を含むメモリブロックの場所を突き止めるためのコマンドを発行してよい。制御情報は非常に大きいので、制御情報の部分が複数の場所に記憶される場合があること、制御情報が、例えば冗長性のために複数の場所に記憶されてよいこと、又は制御情報が、それ以外の場合、ストレージドライブ171A～171Fの複数のメモリブロック全体で分散されてよいことに留意されたい。

【0011】

【0069】 実施態様では、ストレージアレイコントローラ110は、ストレージドライブ171A～171Fから、ストレージドライブ171A～171Fの1つ以上のメモリブロックの状態を記述する制御情報を取り出すことによってストレージアレイ102A～1

02Bのストレージドライブ171A～171Fからデバイス管理責任を取り除いてよい。ストレージドライブ171A～171Fから制御情報を取り出すことは、例えばストレージアレイコントローラ110が、特定のストレージドライブ171A～171Fに対する制御情報の場所についてストレージドライブ171A～171Fに問い合わせることによって実施されてよい。ストレージドライブ171A～171Fは、ストレージドライブ171A～171Fが制御情報の場所を識別できるようにする命令を実行するように構成されてよい。命令は、ストレージドライブ171A～171Fと関連付けられた、又はそれ以外の場合位置するコントローラ（不図示）によって実行されてよく、ストレージドライブ171A～171Fに各メモリブロックの一部分を走査させて、ストレージドライブ171A～171Fのために制御情報を記憶するメモリブロックを識別してよい。ストレージドライブ171A～171Fは、ストレージドライブ171A～171Fのための制御情報の場所を含む応答メッセージをストレージアレイコントローラ110に送信することによって応答してよい。応答メッセージを受け取ることに応えて、ストレージアレイコントローラ110は、ストレージドライブ171A～171Fのための制御情報の場所と関連付けられたアドレスに記憶されたデータを読み取る要求を発行してよい。

【0012】

【0070】 他の実施態様では、ストレージアレイコントローラ110は、制御情報を受け取ることに応えてストレージドライブ管理動作を実行することによってストレージドライブ171A～171Fからデバイス管理責任をさらに取り除いてよい。ストレージドライブ管理動作は、例えばストレージドライブ171A～171F（例えば、特定のストレージドライブ171A～171Fと関連付けられたコントローラ（不図示））によって通常実行される動作を含んでよい。ストレージドライブ管理動作は、例えば、データがストレージドライブ171A～171Fの中の故障したメモリブロックに書き込まれないことを確実にすること、適切な磨耗率が達成されるように等、データがストレージドライブ171A～171Fの中のメモリブロックに書き込まれることを確実にすることを含んでよい。

【0013】

【0071】 実施態様では、ストレージアレイ102A～102Bは、2つ以上のストレージアレイコントローラ110を実装してよい。例えば、ストレージアレイ102Aは、ストレージアレイコントローラ110A及びストレージアレイコントローラ110Bを含んでよい。所与のインスタンスで、ストレージシステム100の単一のストレージアレイコントローラ110（例えば、ストレージアレイコントローラ110A）は、（本明細書では「一次コントローラ」とも呼ばれる）一次ステータスで指定されてよく、他のストレージアレイコントローラ110（例えば、ストレージアレイコントローラ110A）は、（本明細書では「二次コントローラ」とも呼ばれる）二次ステータスで指定されてよい。一次コントローラは、永続記憶装置リソース170A～170Bのデータを改変する許可等の特定の権利を有してよい（例えば、永続記憶装置リソース170A～170Bにデータを書き込むこと）。一次コントローラの権利の少なくともいくつかは、二次コントローラの権利を置き換える場合がある。例えば、二次コントローラは、一次コントローラが権利を有するとき、永続記憶装置リソース170A～170Bのデータを改変する許可を有さない場合がある。ストレージアレイコントローラ110のステータスは変化する可能性がある。例えば、ストレージアレイコントローラ110Aは二次ステータスで指定されてよく、ストレージアレイコントローラ110Bは一次ステータスで指定されてよい。

【0014】

【0072】 一部の実施態様では、ストレージアレイコントローラ110A等の一次コントローラは、1つ以上のストレージアレイ102A～102Bのための一次コントローラとしての機能を果たしてよく、ストレージアレイコントローラ110B等の第2のコントローラは、1つ以上のストレージアレイ102A～102Bのための二次コントローラとしての機能を果たしてよい。例えば、ストレージアレイコントローラ110Aは、ストレージアレイ102A及びストレージアレイ102Bのための一次コントローラであってよく、ストレージアレイコントローラ110Bは、ストレージアレイ102A及び102Bの

ための二次コントローラであってよい。一部の実施態様では、（「ストレージ処理モジュール」とも呼ばれる）ストレージコントローラ110C及び110Dは、一次ステータスも二次ステータスも有さない場合がある。ストレージ処理モジュールとして実装されるストレージアレイコントローラ110C及び110Dは、一次コントローラと二次コントローラ（例えば、それぞれストレージアレイコントローラ110A及び110B）と、ストレージアレイ102Bとの間の通信インタフェースの機能を果たしてよい。例えば、ストレージアレイ102Aのストレージアレイコントローラ110Aは、SAN158を介してストレージアレイ102Bに書込み要求を送信してよい。書込み要求は、ストレージアレイ102Bのストレージアレイコントローラ110C及び110Dの両方によって受け取られてよい。ストレージアレイコントローラ110C及び110Dは、例えば適切なストレージドライブ171A～171Fに書込み要求を送信する等、通信を容易にする。一部の実施態様において、ストレージ処理モジュールは、一次コントローラ及び二次コントローラによって制御されるストレージドライブの数を増加させるために使用されてよいことに留意されたい。

10

【0015】

【0073】 実施態様では、ストレージアレイコントローラ110は、1つ以上のストレージドライブ171A～171Fに、及びストレージアレイ102A～102Bの一部として含まれる1つ以上のNVRAMデバイス（不図示）に中央平面（不図示）を介して通信で結合される。ストレージアレイコントローラ110は、1つ以上のデータ通信リンクを介して中央平面に結合されてよく、中央平面は、1つ以上のデータ通信リンクを介してストレージドライブ171A～171F及びNVRAMデバイスに結合されてよい。本明細書に説明されるデータ通信リンクは、データ通信リンク108A～108Dによって集合的に示され、例えばペリフェラルコンポーネントインターコネクトエクスプレス（『PCIe』）バスを含んでよい。

20

【0016】

【0074】 図1Bは、一部の実施態様に係るデータストレージのための例のシステムを示す。図1Bに示されるストレージアレイコントローラ101は、図1Aに関して説明されるストレージアレイコントローラ110に類似してよい。一例では、ストレージアレイコントローラ101は、ストレージアレイコントローラ110A又はストレージアレイコントローラ110Bに類似してよい。ストレージアレイコントローラ101は、制限よりむしろ説明のために多数の要素を含む。ストレージアレイコントローラ101は、他の実施態様で同じように又は異なるように構成された同じ要素、より多くの要素、又はより少ない要素を含む場合があることに留意されたい。図1Aの要素は、ストレージアレイコントローラ101の特徴を示すのに役立つように以下に含まれてよいことに留意されたい。

30

【0017】

【0075】 ストレージアレイコントローラ101は、1つ以上の処理装置104及びランダムアクセスメモリ（『RAM』）111を含んでよい。処理装置104（又はコントローラ101）は、例えばマイクロプロセッサ、中央演算処理装置等の1つ以上の汎用処理装置を表す。より詳細には、処理装置104（又はコントローラ101）は、複数命令セットコンピューティング（『CISC』）マイクロプロセッサ、縮小命令セットコンピューティング（『RISC』）マイクロプロセッサ、超長命令語（『VLIW』）マイクロプロセッサ、又は他の命令セットを実施するプロセッサ、又は命令セットの組合せを実施するプロセッサであってよい。また、処理装置104（又はコントローラ101）は、例えば特定用途向け集積回路（『ASIC』）、フィールドプログラマブルゲートアレイ（『FPGA』）、デジタルシグナルプロセッサ（『DSP』）、ネットワークプロセッサ等の1つ又は複数の専用処理装置であってよい。

40

【0018】

【0076】 処理装置104は、例えばダブルデータレート4（『DDR4』）バス等の高速メモリバスとして実施されてよいデータ通信リンク106を介してRAM111に接続されてよい。RAM111に記憶されるのは、オペレーティングシステム112である。

50

一部の実施態様では、命令 1 1 3 は R A M 1 1 1 に記憶される。命令 1 1 3 は、直接図化フラッシュストレージシステムで動作を実行するためのコンピュータプログラム命令を含んでよい。一実施形態では、直接図化フラッシュストレージシステムは、フラッシュドライブの中のデータブロックを直接的にアドレス指定するフラッシュストレージシステムであり、アドレス変換なしで、フラッシュドライブのストレージコントローラによって実行される。

【0019】

【0077】 実施態様では、ストレージアレイコントローラ 1 0 1 は、データ通信リンク 1 0 5 A ~ 1 0 5 C を介して処理装置 1 0 4 に結合される 1 つ以上のホストバスアダプタ 1 0 3 A ~ 1 0 3 C を含む。実施態様では、ホストバスアダプタ 1 0 3 A ~ 1 0 3 C は、ホストシステム（例えば、ストレージアレイコントローラ）を他のネットワーク及びストレージアレイに接続するコンピュータハードウェアであってよい。一部の例では、ホストバスアダプタ 1 0 3 A ~ 1 0 3 C は、ストレージアレイコントローラ 1 0 1 が S A N に接続できるようにするファイバチャネルアダプタ、ストレージアレイコントローラ 1 0 1 が L A N に接続できるようにするイーサネットアダプタ等であってよい。ホストバスアダプタ 1 0 3 A ~ 1 0 3 C は、例えば P C I e バス等のデータ通信リンク 1 0 5 A ~ 1 0 5 C を介して処理装置 1 0 4 に結合されてよい。

10

【0020】

【0078】 実施態様では、ストレージアレイコントローラ 1 0 1 は、エキスパンダ 1 1 5 に結合されるホストバスアダプタ 1 1 4 を含んでよい。エキスパンダ 1 1 5 は、ホストシステムをより多数のストレージドライブにアタッチするために使用されてよい。エキスパンダ 1 1 5 は、例えば、ホストバスアダプタ 1 1 4 が S A S コントローラとして実施される実施態様において、ホストアダプタ 1 1 4 がストレージドライブにアタッチできるようにするために活用される S A S エキスパンダであってよい。

20

【0021】

【0079】 実施態様では、ストレージアレイコントローラ 1 0 1 は、データ通信リンク 1 0 9 を介して処理装置 1 0 4 に結合されたスイッチ 1 1 6 を含んでよい。スイッチ 1 1 6 は、単一のエンドポイントの中から複数のエンドポイントを作り出し、それによって複数のデバイスが単一のエンドポイントを共用できるようにするコンピュータハードウェアデバイスであってよい。スイッチ 1 1 6 は、例えば P C I e バス（例えば、データバスリンク 1 0 9）に結合され、複数の P C I e 接続ポイントを中央平面に提示する P C I e スイッチであってよい。

30

【0022】

【0080】 実施態様では、ストレージアレイコントローラ 1 0 1 は、他のストレージアレイコントローラにストレージアレイコントローラ 1 0 1 を結合するためのデータ通信リンク 1 0 7 を含む。一部の例では、データ通信リンク 1 0 7 は、クイックパスインターコネクト（Q P I）相互接続であってよい。

【0023】

【0081】 従来のフラッシュドライブを使用する従来のストレージシステムは、従来のストレージシステムの一部であるフラッシュドライブ全体でプロセスを実施してよい。例えば、ストレージシステムのより高いレベルのプロセスは、フラッシュドライブ全体でプロセスを開始し、制御してよい。しかしながら、従来のストレージシステムのフラッシュドライブは、プロセスも実行する独自のストレージコントローラを含んでよい。したがって、従来のストレージシステムの場合、（例えば、ストレージシステムによって開始される）より高いレベルのプロセス及び（例えば、ストレージシステムのストレージコントローラによって開始される）より低いレベルのプロセスは両方とも実行されてよい。

40

【0024】

【0082】 従来のストレージシステムの多様な不備を解決するために、動作はより低いレベルのプロセスによってではなく、より高いレベルのプロセスによって実行されてよい。例えば、フラッシュストレージシステムは、プロセスを提供するストレージコントローラ

50

を含まないフラッシュドライブを含んでよい。したがって、フラッシュストレージシステムのオペレーティングシステム自体がプロセスを開始し、制御してよい。これは、フラッシュドライブの中で直接的にデータブロックをアドレス指定する直接図化フラッシュストレージシステムによって達成され、アドレス変換なしに、フラッシュドライブのストレージコントローラによって実行されてよい。

【0025】

【0083】 フラッシュストレージシステムのオペレーティングシステムは、割当てユニットのリストを識別し、フラッシュストレージシステムの複数のフラッシュドライブ全体で維持してよい。割当てユニットは、完全消去ブロック又は複数の消去ブロックであってよい。オペレーティングシステムは、マップ又はフラッシュストレージシステムのフラッシュドライブのブロックを消去するためにアドレスを直接的にマッピングするアドレス範囲を維持してよい。

10

【0026】

【0084】 フラッシュドライブのブロックを消去するための直接的なマッピングは、データを書き換え、データを消去するために使用されてよい。例えば、動作は、第1のデータが保持され、第2のデータはフラッシュストレージシステムによってもはや使用されていない第1のデータ及び第2のデータを含む1つ以上の割当てユニットに対して実行されてよい。オペレーティングシステムは、第1のデータを他の割当てユニットの中の新しい場所へ書き込み、第2のデータを消去し、割当てユニットを、後続のデータのための使用に利用可能であるとマークするためのプロセスを開始してよい。このようにして、プロセスは、追加のより低いレベルのプロセスがフラッシュドライブのコントローラによって実行されることなく、フラッシュストレージシステムのより高いレベルのオペレーティングシステムによってのみ実行されてよい。

20

【0027】

【0085】 プロセスがフラッシュストレージシステムのオペレーティングシステムによってのみ実行される優位点は、不必要な又は冗長な書き込み動作がプロセスの間に実行されていないので、フラッシュストレージシステムのフラッシュドライブの信頼性の増加を含む。ここで新規性の1つの考えられる点は、フラッシュストレージシステムのオペレーティングシステムでプロセスを開始し、制御する概念である。さらに、プロセスは、複数のフラッシュドライブ全体でオペレーティングシステムによって制御できる。これは、フラッシュドライブのストレージコントローラによって実行されているプロセスとは対照的である。

30

【0028】

【0086】 ストレージシステムは、フェイルオーバーのためにドライブの集合を共用する2つのストレージレイコントローラから成る場合がある、又はストレージシステムは、複数のドライブを活用するストレージサービスを提供する単一のストレージレイコントローラから成るであろう、又はストレージシステムは、それぞれがなんらかの数のドライブ若しくはなんらかの量のフラッシュストレージを有するストレージレイコントローラの分散ネットワークから成り、ネットワークのストレージレイコントローラは、完全なストレージサービスを提供するために協働し、ストレージ割当て及びガベージコレクションを含んだストレージサービスの多様な態様に関して協働する。

40

【0029】

【0087】 図1Cは、一部の実施態様に係るデータストレージのための第3の例のシステム117を示す。システム117（本明細書では「ストレージシステム」とも呼ばれる）は、制限よりはむしろ説明のために多数の要素を含む。システム117が、他の実施態様で同じように又は異なるように構成された同じ要素、より多くの要素、又はより少ない要素を含んでよい場合があることに留意されたい。

【0030】

【0088】 一実施形態では、システム117は、別々にアドレス指定可能な高速書き込みストレージを有するデュアルパリティコンポーネントインターコネクト（『PCI』）

50

フラッシュストレージデバイス 118 を含む。システム 117 は、ストレージコントローラ 119 を含んでよい。一実施形態では、ストレージコントローラ 119 は、CPU、ASIC、FPGA、又は本開示に従って必要な制御構造を実施してよい任意の他の回路網であってよい。一実施形態では、システム 117 は、ストレージデバイスコントローラ 119 の多様なチャンネルに動作可能なように結合された（例えば、フラッシュメモリデバイス 120a から 120n を含む）フラッシュメモリデバイスを含む。フラッシュメモリデバイス 120a から 120n は、フラッシュページ、消去ブロック、及び／又はストレージデバイスコントローラ 119 がフラッシュの多用な態様をプログラムし、取り出すことを可能にするほど十分な制御要素のアドレス指定可能な集合体として、コントローラ 119 に提示されてよい。一実施形態では、ストレージデバイスコントローラ 119 は、ページのデータコンテンツを記憶し、取り出すこと、任意のブロックを配置し、消去すること、フラッシュメモリページ、消去ブロック、及びセルの使用及び再利用に関係する統計を追跡すること、フラッシュメモリの中のエラーコード及び障害を追跡し、予測すること、プログラミングと関連付けられた電圧レベルを制御すること、及びフラッシュセル等のコンテンツを取り出すことを含む動作をフラッシュメモリデバイス 120A～120N に対して実行してよい。

【0031】

【0089】 一実施形態では、システム 117 は、別々にアドレス指定可能な高速書込みデータを記憶するために RAM 121 を含んでよい。一実施形態では、RAM 121 は、1 つ以上の別々の個別素子であってよい。別の実施形態では、RAM 121 は、ストレージデバイスコントローラ 119 又は複数のストレージデバイスコントローラに統合されてよい。RAM 121 は、例えばストレージデバイスコントローラ 119 内の処理装置（例えば、CPU）のための一時プログラムメモリ等、他の目的に活用されてよい。

【0032】

【0090】 一実施形態では、システム 119 は、例えば充電式電池又はコンデンサ等の貯蔵エネルギーデバイス 122 を含んでよい。貯蔵エネルギーデバイス 122 は、RAM のコンテンツをフラッシュメモリに書き込むほど十分な時間の間、RAM（例えば、RAM 121）のなんらかの量、及びフラッシュメモリ（例えば、フラッシュメモリ 120a から 120n）のなんらかの量、ストレージデバイスコントローラ 119 に電力を提供するほど十分なエネルギーを貯蔵してよい。一実施形態では、ストレージデバイスコントローラ 119 は、ストレージデバイスコントローラが外部電力の損失を検出する場合に RAM のコンテンツをフラッシュメモリに書き込んでよい。

【0033】

【0091】 一実施形態では、システム 117 は、2 つのデータ通信リンク 123a、123b を含む。一実施形態では、データ通信リンク 123a、123b は、PCI インタフェースであってよい。別の実施形態では、データ通信リンク 123a、123b は、他の通信規格（例えば、HyperTransport、インフィニバンド等）に基づいてよい。データ通信リンク 123a、123b は、ストレージシステム 117 の他の構成要素からストレージデバイスコントローラ 119 への外部接続を可能にする不揮発性メモリエクスプレス（『NVMe』）又は NVMe オーバーファブリック（『NVMf』）仕様に基いてよい。データ通信リンクが、便宜上本明細書では PCI バスと同義で呼ばれてよいことに留意されたい。

【0034】

【0092】 また、システム 117 は、一方又は両方のデータ通信リンク 123a、123b を介して設けられてよい、又は別々に設けられてよい外部電源（不図示）を含んでもよい。代替実施形態は、RAM 121 のコンテンツを記憶する際の使用専用の別個のフラッシュメモリ（不図示）を含む。ストレージデバイスコントローラ 119 は、アドレス指定可能な高速書込み論理デバイスを含んでよい PCI バス上の論理デバイス、又は PCI メモリとして若しくは永続記憶装置として提示されてよいストレージデバイス 118 の論理アドレス空間の別個の一部を提示してよい。一実施形態では、デバイスの中に記憶するた

めの動作は、RAM 121の中に向けられる。停電時、ストレージデバイスコントローラ119は、長期永続記憶のために、アドレス指定可能な高速書込み論理ストレージと関連付けられた記憶されているコンテンツをフラッシュメモリ（例えば、フラッシュメモリ120aから120n）に書き込んでよい。

【0035】

【0093】 一実施形態では、論理デバイスは、フラッシュメモリデバイス120a～120nのコンテンツの一部又はすべてのなんらかの提示を含んでよく、その提示は、ストレージデバイス118を含んだストレージシステム（例えば、ストレージシステム117）が、フラッシュメモリページを直接的にアドレス指定し、PCIバスを通してストレージデバイスにとって外部であるストレージシステム構成要素から消去ブロックを直接的にプログラムし直すことを可能にする。また、提示は、外部構成要素のうちの1つ以上が、すべてのフラッシュメモリデバイス全体でフラッシュメモリページ、消去ブロック、及びセルの使用及び再利用に関係する統計を追跡することと、フラッシュメモリデバイスの中で及び全体でエラーコード及び障害を追跡し、予測することと、プログラミングと関連付けられた電圧レベルを制御することと、フラッシュセル等のコンテンツを取り出すこととのうちの一部又はすべてを含んだフラッシュメモリの他の態様を制御し、取り出すこととを可能にしてもよい。

【0036】

【0094】 一実施形態では、貯蔵エネルギーデバイス122は、フラッシュメモリデバイス107a～120nに対する進行中の動作の完了を確実にするために十分であってよく、貯蔵エネルギーデバイス122は、それらの動作のために、及びフラッシュメモリへの高速書込みRAMの記憶のためにストレージデバイスコントローラ119及び関連付けられたフラッシュメモリデバイス（例えば、120a～120n）に電力を提供してよい。貯蔵エネルギーデバイス122は、蓄積された統計及びフラッシュメモリデバイス120a～120n及び／又はストレージデバイスコントローラ119によって保たれ、追跡されている他のパラメータを記憶するために使用されてよい。別々のコンデンサ又は貯蔵エネルギーデバイス（例えば、フラッシュメモリデバイス自体に近い、又はフラッシュメモリデバイス自体に埋め込まれたより小さいコンデンサ）は、本明細書に説明される動作のいくつか又はすべてのために使用されてよい。

【0037】

【0095】 例えば経時的に電圧レベルを調整すること、貯蔵エネルギーデバイス122を部分的に放電して対応する放電特徴を測定すること等、多様な方式が、貯蔵エネルギー構成要素の寿命を追跡し、最適化するために使用されてよい。利用可能なエネルギーが経時的に減少する場合、アドレス指定可能な高速書込みストレージの有効利用可能容量は、それが、現在利用可能な貯蔵エネルギーに基づいて安全に書き込むことができることを確実にするために減少されてよい。

【0038】

【0096】 図1Dは、一部の実施態様に係るデータストレージのための第3の例のシステム124を示す。一実施形態では、システム124は、ストレージコントローラ125a、125bを含む。一実施形態では、ストレージコントローラ125a、125bは、デュアルPCIストレージデバイス119a、119b、及び119c、119dにそれぞれ動作可能なように結合される。ストレージコントローラ125a、125bは、なんらかの数のホストコンピュータ127a～127nに（例えば、ストレージネットワーク130を介して）動作可能なように結合されてよい。

【0039】

【0097】 一実施形態では、2つのストレージコントローラ（例えば、125a及び125b）は、例えばSCS）ブロックストレージアレイ、ファイルサーバ、オブジェクトサーバ、データベース、又はデータ解析サービス等、ストレージサービスを提供する。ストレージコントローラ125a、125bは、なんらかの数のネットワークインタフェース（例えば、126a～126d）を通して、ストレージシステム124の外のホストコン

ピュータ 1 2 7 a ~ 1 2 7 n にサービスを提供してよい。ストレージコントローラ 1 2 5 a、1 2 5 b は、完全にストレージシステム 1 2 4 の中で統合サービス又はアプリケーションを提供し、集中型のストレージ及び計算システムを形成してよい。ストレージコントローラ 1 2 5 a、1 2 5 b は、停電、ストレージコントローラ削除、ストレージコントローラ又はストレージシステムの停止、又はストレージシステム 1 2 4 の中の 1 つ以上のソフトウェア構成要素若しくはハードウェア構成要素のなんらかの障害時に動作が失われないことを確実にするために、進行中の動作を記録するためにストレージデバイス 1 1 9 a から 1 1 9 d の中で又は全体で高速書込みメモリを活用してよい。

【0040】

【0098】 一実施形態では、コントローラ 1 2 5 a、1 2 5 b は、一方又は他方の P C I バス 1 2 8 a、1 2 8 b に対する P C I マスタとして動作する。別の実施形態では、1 2 8 a 及び 1 2 8 b は、他の通信規格（例えば、HyperTransport、インフィニバンド等）に基づいてよい。他のストレージシステムの実施形態は、P C I バス 1 2 8 a、1 2 8 b のためのマルチマスタとしてストレージコントローラ 1 2 5 a、1 2 5 b を操作してよい。代わりに、P C I / N V M e / N V M f 交換インフラストラクチャ又はファブリックは、複数のストレージコントローラを接続してよい。いくつかのストレージシステム実施形態は、ストレージデバイスが、ストレージコントローラとだけ通信するよりむしろ、互いに直接的に通信できるようにしてよい。一実施形態では、ストレージデバイスコントローラ 1 1 9 a は、フラッシュメモリデバイスの中に記憶されるデータを合成し、R A M（例えば、図 1 C の R A M 1 2 1）に記憶されているデータから転送するためにストレージコントローラ 1 2 5 a からの指揮下で操作可能であってよい。例えば、R A M コンテンツの再計算されたバージョンは、データの安全性を保証する改善するために、又は再利用のためにアドレス指定可能な高速書込み容量を解放するために、ストレージコントローラが、動作がストレージシステム全体で完全にコミットしたと判断した後、又はデバイスに対する高速書込みメモリが特定の使用容量に達したとき、又は一定量の時間の後に転送されてよい。この機構は、例えばストレージコントローラ 1 2 5 a、1 2 5 b からバス（例えば、1 2 8 a、1 2 8 b）上での第 2 の転送を回避するために使用されてよい。一実施形態では、再計算はデータを圧縮することと、インデックス化又は他のメタデータを接続することと、複数のデータセグメントを互いに結合することと、イレイジャーコード計算を実行すること等を含んでよい。

【0041】

【0099】 一実施形態では、ストレージコントローラ 1 2 5 a、1 2 5 b からの命令下で、ストレージデバイスコントローラ 1 1 9 a、1 1 9 b は、データを計算し、ストレージコントローラ 1 2 5 a、1 2 5 b の関与なしに、R A M（例えば、図 1 C の R A M 1 2 1）に記憶されているデータから他のストレージデバイスに転送するよう作動してよい。この動作は、あるコントローラ 1 2 5 a から別のコントローラ 1 2 5 b に記憶されたデータをミラーリングするために使用されることもあれば、それは圧縮、データ統合、及び／又はストレージコントローラ又はストレージコントローラインタフェース 1 2 9 a、1 2 9 b から P C I バス 1 2 8 a、1 2 8 b に対する負荷を削減するために、イレイジャーコーディング計算及びストレージデバイスへの転送をアンロードするために使用される場合もある。

【0042】

【0100】 ストレージデバイスコントローラ 1 1 9 は、デュアル P C I ストレージデバイス 1 1 8 にとって外部のストレージシステムの他の部分による使用のために高い可用性プリミティブを実施するための機構を含んでよい。例えば、予約プリミティブ又は除外プリミティブは、高可用性ストレージサービスを提供する 2 つのストレージコントローラを有するストレージシステムにおいて、一方のストレージコントローラが、他方のストレージコントローラがストレージデバイスにアクセスする又はストレージデバイスにアクセスし続けることを防ぎ得るように提供されてよい。これは、例えば、一方のコントローラが、他方のコントローラが適切に機能していないことを検出する場合、又は 2 つのストレージ

コントローラ間の相互接続がそれ自体適切に機能していない場合がある場合に使用できるであろう。

【0043】

【0101】 一実施形態では、別々にアドレス指定可能な高速書込みストレージを有するデュアルPCI直接図化ストレージデバイスとの使用のためのストレージシステムは、消去ブロック又は消去ブロックのグループを、ストレージサービスの代わりにデータを記憶するための、又はストレージサービスと関連付けられたメタデータ（例えば、インデックス、ログ等）を記憶するための、又はストレージシステム自体の適切な管理のための割当てユニットとして管理するシステムを含む。サイズが数キロバイトである場合があるフラッシュページは、データが到達すると、又はストレージシステムが長い時間の間隔（例えば、時間の定められた閾値を超えて）データを持続すると、書き込まれてよい。データをより迅速にコミットするために、又はフラッシュメモリデバイスへの書込み数を削減するために、ストレージコントローラは、まず1つ以上のストレージデバイス上の別々にアドレス指定可能な高速書込みストレージにデータを書き込んでよい。

10

【0044】

【0102】 一実施形態では、ストレージコントローラ125a、125bは、ストレージデバイスの年齢及び予想される残りの寿命に従って、又は他の統計に基づいて、ストレージデバイス（例えば、118）の中で及びストレージデバイス（例えば、118）全体で消去ブロックの使用を開始してよい。ストレージコントローラ125a、125bは、フラッシュページ及び消去ブロックの寿命を管理するため及び全体的なシステム性能を管理するためだけではなく、もはや必要とされないページに従ってストレージデバイス間でガベージコレクション及びデータ移行データを開始してもよい。

20

【0045】

【0103】 一実施形態では、ストレージシステム124は、アドレス指定可能な高速書込みストレージヘデータを記憶することの一部として、及び／又は消去ブロックと関連付けられた割当てユニットにデータを書き込むことの一部としてミラーリング方式及び／又は消去コーディング方式を利用してよい。イレイジャーコードは、単一又は複数のストレージデバイス故障に対する冗長性を提供するために、又はフラッシュメモリ動作から若しくはフラッシュメモリセルの劣化から生じるフラッシュメモリページの内部破損から保護するために、消去ブロック若しくは割当てユニットの中で、又は単一のストレージデバイス上のフラッシュメモリデバイスの中で及びフラッシュメモリデバイス全体でだけではなく、ストレージデバイス全体で使用されてよい。多様なレベルのミラーリング及びイレイジャーコーディングが、別々に又は組み合わせられて発生する複数のタイプの故障から回復するために使用されてよい。

30

【0046】

【0104】 図2Aから図2Gに関して示される実施形態は、1つ以上のユーザシステム若しくはクライアントシステム又はストレージクラスタにとって外部の他のソースから生じるユーザデータ等のユーザデータを記憶するストレージクラスタを示す。ストレージクラスタは、イレイジャーコーディング及びメタデータの冗長なコピーを使用し、シャシの中に収容されたストレージノード全体で、又は複数のシャシ全体でユーザデータを分散する。イレイジャーコーディングは、データが、例えばディスク、ストレージノード、又は地理的な位置等の異なる場所の集合全体で記憶されるデータの保護又は再構築の方法を指す。実施形態は、他のタイプのソリッドステートメモリ又は非ソリッドステートメモリを含んだ他の記憶媒体にも拡張され得るが、フラッシュメモリは、実施形態と統合されてよい1つのタイプのソリッドステートメモリである。記憶場所及び作業負荷の制御は、クラスタ化されたピアツーピアシステムでの記憶場所全体で分散される。多様なストレージノード間で通信を仲介すること、ストレージノードがいつ利用できなくなったのかを検出すること、及び多様なストレージノード全体でE/O（入出力）のバランスをとること等のタスクは、すべて分散ベースで処理される。データは、いくつかの実施形態でデータ回復をサポートするデータフラグメント又はストライプ内で複数のストレージノード全体にレイ

40

50

アウト又は分散される。データの所有権は、入力パターン及び出力パターンとは関係なく、クラスタの中で割り当てし直される。データは他のストレージノードから再構築することができ、したがって入力動作及び出力動作に利用可能のままであるので、以下により詳細に説明されるこのアーキテクチャは、クラスタ内のストレージノードが、システムが稼働したままの状態では故障することを可能にする。多様な実施形態では、ストレージノードは、クラスタノード、ブレード、又はサーバと呼ばれる場合がある。

【0047】

【0105】 ストレージクラスタは、シャシ、つまり1つ以上のストレージノードを収容するエンクロージャの中に含まれる場合がある。例えば配電バス等の各ストレージノードに電力を提供するための機構、及びストレージノード間の通信を可能にする通信バス等の通信機構が、シャシの中に含まれる。ストレージクラスタは、いくつかの実施形態によれば、1つの場所の独立したシステムとして実行できる。一実施形態では、シャシは、独立して有効又は無効にされ得る、配電及び通信バスの両方の少なくとも2つのインスタンスを含む。内部通信バスはイーサネットバスであってよいが、例えばP C I e、インフィニバンド、及び他等の他の技術は等しく適している。シャシは、直接的に又はスイッチを通して、複数のシャシ間の及びクライアントシステムとの通信を可能にするための外部通信バス用のポートを提供する。外部通信は、例えばイーサネット、インフィニバンド、ファイバチャネル等の技術を使用してよい。いくつかの実施形態では、外部通信バスは、シャシ間通信及びクライアント通信のための異なる通信バス技術を使用する。スイッチがシャシの中で又はシャシ間に配備される場合、スイッチは、複数のプロトコル又は技術の間の変換の機能を果たしてよい。複数のシャシがストレージクラスタを定義するために接続されるとき、ストレージクラスタは、専用のインタフェース又は例えばネットワークファイルシステム（『N F S』）、共通インターネットファイルシステム（『C I F S』）、スモールコンピュータシステムインタフェース（『S C S I』）、又はハイパーテキスト転送プロトコル（『H T T P』）等の標準的なインタフェースを使用し、クライアントによってアクセスされてよい。クライアントプロトコルからの変換は、スイッチ、シャシ外部通信バスで、又は各ストレージノードの中で起こる場合がある。いくつかの実施形態では、複数のシャシは、アグリゲータスイッチを通して互いに結合又は接続されてよい。結合又は接続されたシャシの一部及び／又はすべては、ストレージクラスタとして指定されてよい。上述されたように、各シャシは複数のブレードを有する場合があります、各ブレードは媒体アクセス制御（『M A C』）アドレスを有するが、ストレージクラスタは、いくつかの実施形態では単一のクラスタI Pアドレス及び単一のM A Cアドレスを有するとして外部ネットワークに提示される。

【0048】

【0106】 各ストレージノードは、1つ以上のストレージサーバであってよく、各ストレージサーバは、ストレージユニット又はストレージデバイスと呼ばれる場合がある、1つ以上の不揮発性ソリッドステートメモリユニットに接続される。一実施形態は、各ストレージノードに、及び1つから8つの不揮発性ソリッドステートメモリユニット間に単一のストレージサーバを含むが、この一例は限定的となることを意図していない。ストレージサーバは、プロセッサ、D R A M、及び内部通信バス及び電力バスのそれぞれの配電のためのインタフェースを含んでよい。いくつかの実施形態では、ストレージノードの内側で、インタフェース及びストレージユニットは、例えばP C I E x p r e s s等の通信バスを共用する。不揮発性ソリッドステートメモリユニットは、ストレージノード通信バスを通して内部通信バスインタフェースに直接的にアクセスしてよい、又はストレージノードにバスインタフェースにアクセスするように要求してよい。不揮発性ソリッドステートメモリユニットは、埋め込みC P U、ソリッドステートストレージコントローラ、及びいくつかの実施形態では、例えば2～32テラバイト（『T B』）の間の大量のソリッドステートマスストレージを含む。例えばD R A M等の埋め込み揮発性ストレージ媒体、及びエネルギー貯蔵装置は、不揮発性ソリッドステートメモリユニットに含まれる。いくつかの実施形態では、エネルギー貯蔵装置は、電力損失時にD R A Mコンテンツの部分集合を

10

20

30

40

50

安定した記憶媒体に転送することを可能にするコンデンサ、超コンデンサ、又は電池である。いくつかの実施形態では、不揮発性ソリッドステートメモリユニットは、例えばDRAMの代わりになり、電力削減ホールドアップ装置(reduced power hold-up apparatus)を可能にする相変化メモリ又は磁気抵抗ランダムアクセスメモリ(『MRAM』)等のストレージクラスメモリで構築される。

【0049】

【0107】 ストレージノード及び不揮発性ソリッドステートストレージの多くの特徴の1つは、ストレージクラス内で先を見越してデータを再構築する能力である。ストレージノード及び不揮発性ソリッドステートストレージは、そのストレージノード又は不揮発性ソリッドステートストレージと関わるデータを読み取る試みがあるかどうかとは無関係に、ストレージノード又はストレージクラス内の不揮発性ソリッドステートストレージがいつ到達不能であるのかを判断できる。ストレージノード及び不揮発性ソリッドステートストレージは、次いで少なくとも部分的に新しい場所のデータを復旧し、再構築するために協調する。これは、システムが、データが、ストレージクラスタを利用するクライアントシステムから開始される読取りアクセスのために必要とされるまで待機することなくデータを再構築するという点で先を見越した再構築を構成する。ストレージメモリ及びその動作のこれらの詳細及び追加の詳細は、以下に説明される。

【0050】

【0108】 図2Aは、いくつかの実施形態に従って、ネットワーク接続ストレージ又はストレージエリアネットワークを提供するために、複数のストレージノード150、及び各ストレージノードに結合された内部ソリッドステートメモリを有する、ストレージクラスタ161の斜視図である。ネットワーク接続ストレージ、ストレージエリアネットワーク、又はストレージクラスタ、又は他のストレージメモリは、それぞれが、物理構成要素及びそれにより提供されるストレージメモリの量の両方の柔軟かつ再構成可能な構成で、1つ以上のストレージノード150を有する1つ以上のストレージクラスタ161を含む場合があるであろう。ストレージクラスタ161はラックに収まるように設計され、1つ以上のラックは、ストレージメモリに対して所望されるようにセットアップされ、装着される場合がある。ストレージクラスタ161は、複数のスロット142を有するシャシ138を有する。シャシ138が、筐体、エンクロージャ、又はラックユニットと呼ばれる場合があることが理解されるべきである。他のスロット数は容易に考案されるが、一実施形態では、シャシ138は14のスロット142を有する。例えば、いくつかの実施形態は、4つのスロット、8つのスロット、16のスロット、32のスロット、又は他の適切な数のスロットを有する。いくつかの実施形態では、各スロット142は、1つのストレージノード150を収容する場合がある。シャシ138は、ラック上にシャシ138を取り付けるために活用できるフラップ148を含む。他の冷却構成要素が使用できるであろうが、ファン144は、ストレージノード150及びその構成要素の冷却のための空気循環を提供する、又は冷却構成要素のない実施形態が考案される場合があるであろう。スイッチファブリック146は、メモリへの通信のためにシャシ138の中のストレージノード150を互いに、及びネットワークに結合する。本明細書に示される実施形態では、スイッチファブリック146及びファン144の左側のスロット142は、ストレージノード150によって占有されて示されている。一方、スイッチファブリック146及びファン144の右側のスロット142は空であり、例示的な目的のためにストレージノード150の挿入に利用可能である。この構成は一例であり、1つ以上のストレージノード150が、多様な追加の構成でスロット142を占有する場合があるであろう。ストレージノードの構成は、いくつかの実施形態では、連続又は隣接している必要はない。ストレージノード150はホットプラグ可能であり、ストレージノード150は、システムを停止する又は電源を切ることなく、シャシ138内のスロット142に挿入できる、又はスロット142から取り外しできることを意味する。ストレージノード150の挿入又はスロット142からの取外し時、システムは、変更を認識し、変更に適応するために自動的に再構成する。いくつかの実施形態では、再構成は、冗長性を復元すること及び／又はデータ若

10

20

30

40

50

しくは負荷のバランスを再び取ることを含む。

【0051】

【0109】 各ストレージノード150は、複数の構成要素を含む場合がある。他の取付物及び／又は構成要素が追加の実施形態で使用されるであろう場合があるが、ここに示される実施形態では、ストレージノード150は、CPU156、つまりプロセッサによって装着されるプリント基板159、CPU156に結合されたメモリ154、及びCU156に結合された不揮発性ソリッドステートストレージ152を含む。メモリ154は、CPU156によって実行される命令、及び／又はCPU156によって影響されるデータを有する。以下にさらに説明されるように、不揮発性ソリッドステートストレージ152は、フラッシュを含む、又は追加の実施形態では他のタイプのソリッドステートメモリを含む。

10

【0052】

【0110】 図2Aを参照すると、ストレージクラス161はスケーラブルであり、不均一なストレージサイズを有する記憶容量が、上述されたように容易に追加されることを意味する。1つ以上のストレージノード150は、各シャシに差し込む又は各シャシから取り外すことができ、ストレージクラスはいくつかの実施形態では自己設定する。プラグインストレージノード150は、配送済み (as delivered) 又は後に付加されるとしてシャシにインストールされるかどうかに関わりなく、異なるサイズを有する場合がある。例えば、一実施形態では、ストレージノード150は、例えば8TB、12TB、16TB、32TB等の4TBの任意の倍数を有する場合がある。追加の実施形態では、ストレージノード150は、他のストレージ量又は記憶容量の任意の倍数を有する場合があるであろう。各ストレージノード150の記憶容量は一斉送信され、データにストライプを付ける (stripe) 方法の決定に影響を与える。最大記憶効率のために、実施形態は、シャシの中の最大で1つ又は最大で2つの不揮発性ソリッドステートストレージユニット152又はストレージノード150の損失のある継続動作の所定の要件を条件に、ストライプで可能な限り幅広く自己設定することができる。

20

【0053】

【0111】 図2Bは、複数のストレージノード150を結合する通信相互接続171A～171F及び配電バス172を示すブロック図である。図2Aを参照し直すと、いくつかの実施形態では、通信相互接続171A～171Fは、スイッチファブリック146に含まれる場合がある、又はスイッチファブリック146で実施される場合がある。複数のストレージクラス161がラックを占有する場合、いくつかの実施形態では、通信相互接続171A～171Fは、ラックスイッチの上部に含まれる場合がある、又はラックスイッチの上部で実施される場合がある。図2Bに示されるように、ストレージクラス161は、単一のシャシ138の中に封入される。外部ポート176は、通信相互接続171A～171Fを通してストレージノード150に結合される。一方、外部ポート174は、ストレージノードに直接的に結合される。外部電力ポート178は、配電バス172に結合される。ストレージノード150は、図2Aに関して説明されるように、変化する量の、及び異なる容量の不揮発性ソリッドステートストレージ152を含む場合がある。さらに、1つ以上のストレージノードは、図2Bに示されるように計算専用のストレージノードであってよい。オーソリティ168は、例えばメモリに記憶されたリスト又は他のデータ構造等、不揮発性ソリッドストレージ152上で実施される。いくつかの実施形態では、オーソリティは、不揮発性ソリッドステートストレージ152の中で記憶され、不揮発性ソリッドステートストレージ152のコントローラ又は他のプロセッサ上で実行中のソフトウェアによってサポートされる。追加の実施形態では、オーソリティ168は、例えばメモリ154に記憶されたリスト又は他のデータ構造として、ストレージノード150で実施され、ストレージノード150のCPU156で実行中のソフトウェアによってサポートされる。オーソリティ168は、いくつかの実施形態では、どのようにして及びどこでデータが不揮発性ソリッドステートストレージ152に記憶されるのかを制御する。この制御は、どのタイプのイレイジャーコーディング方式がデータに適用されるのか、

30

40

50

及びストレージノード150がデータのどの部分を有しているのかを決定するのに役立つ。各オーソリティ168は、不揮発性ソリッドステートストレージ152に割り当てられてよい。各オーソリティは、多様な実施形態では、ファイルシステムによって、ストレージノード150によって、又は不揮発性ソリッドステートストレージ152によってデータに割り当てられる、一連のinode番号、セグメント番号、又は他のデータ識別子を制御してよい。

【0054】

[0112] いくつかの実施形態では、あらゆるデータ、及びあらゆるメタデータは、システム内に冗長性を有する。さらに、あらゆるデータ及びあらゆるメタデータは、オーソリティと呼ばれる場合もあるオーナーを有する。オーソリティが、例えばストレージノードの故障により到達不能である場合、そのデータ又はそのメタデータをどのようにして見つけるのかについての継承の計画がある。多様な実施形態では、オーソリティ168の冗長なコピーがある。オーソリティ168は、いくつかの実施形態では、ストレージノード150及び不揮発性ソリッドステートストレージ152に対する関係性を有する。一連のデータセグメント番号又はデータの他の識別子をカバーする各オーソリティ168は、特定の不揮発性ソリッドステートストレージ152に割り当てられてよい。いくつかの実施形態では、係る範囲のすべてに対するオーソリティ168は、ストレージクラスタの不揮発性ソリッドステートストレージ152上で分散される。各ストレージノード150は、そのストレージノード150の不揮発性ソリッドステートストレージ（複数可）152にアクセスを提供するネットワークポートを有する。データは、セグメント番号と関連付けられるセグメントに記憶される場合があり、そのセグメント番号は、いくつかの実施形態ではRAID（独立した複数のディスクから成る冗長配列）ストライプの構成のためのインダイレクションである。オーソリティ168の割当て及び使用は、このようにしてデータに対するインダイレクションを確立する。インダイレクションは、いくつかの実施形態によれば、間接的に、この場合オーソリティ168を介して、データを参照する能力と呼ばれる場合がある。セグメントは、不揮発性ソリッドステートストレージ152の集合、及びデータを含む場合がある不揮発性ソリッドステートストレージ152の集合の中へのローカル識別子を識別する。いくつかの実施形態では、ローカル識別子は、デバイスへのオフセットであり、複数のセグメントによって連続的に再利用されてよい。他の実施形態では、ローカル識別子は、特定のセグメントにとって一意であり、決して再利用されない。不揮発性ソリッドステートストレージ152のオフセットは、（RAIDストライプの形で）不揮発性ソリッドステートストレージ152への書込み又は不揮発性ソリッドステートストレージ152からの読取りのためにデータの場所を突き止めることに適用される。データは、特定のデータセグメントのためのオーソリティ168を有する不揮発性ソリッドステートストレージ152を含んでよい、又は不揮発性ソリッドステートストレージ152とは異なってよい、不揮発性ソリッドステートストレージ152の複数のユニット全体でストライプを付けられる。

【0055】

[0113] 例えば、データ移動又はデータ再構築中に、データの特定のセグメントの場所が突き止められる変化がある場合、そのデータセグメントのためのオーソリティ168は、その不揮発性ソリッドステートストレージ152又はそのオーソリティ168を有するストレージノード150で相談されるべきである。1つの特定のデータの場所を突き止めるために、実施形態は、データセグメントのハッシュ値を計算する、又はinode番号若しくはデータセグメント番号を適用する。この動作の出力は、その1個の特定のデータのためのオーソリティ168を有する不揮発性ソリッドステートストレージ152を指す。いくつかの実施形態では、この動作に対する2つの段階がある。第1の段階は、例えばセグメント番号、inode番号等のエンティティ識別子（ID）、又はディレクトリ番号をオーソリティ識別子にマッピングする。このマッピングは、ハッシュ又はビットマスク等の計算を含む場合がある。第2の段階は、明示的なマッピングを通して行われる場合がある、特定の不揮発性ソリッドステートストレージ152にオーソリティ識別子をマッ

ピングすることである。動作は繰り返し可能であり、これにより計算が実行されるとき、計算の結果は、オーソリティ168を有する特定の揮発性ソリッドステートストレージ152を繰り返し自在に且つ確実に指す。動作は、入力として到達可能なストレージノードの集合を含む場合がある。到達可能な揮発性ソリッドステートユニットの集合が変化する場合、最適集合が変化する。いくつかの実施形態では、持続値は（つねに真である）現在の割当てであり、計算値は、クラスタがそれに向かって再構成しようと試みるターゲット割当てである。この計算は、到達可能であり、同じクラスタを構成する揮発性ソリッドステートストレージ152の集合の存在下でオーソリティのために最適な揮発性ソリッドステートストレージ152を決定するために使用されてよい。また、計算は、割り当てられた揮発性ソリッドステートストレージが到達不能である場合にもオーソリティが決定され得るように、揮発性ソリッドステートストレージマッピングに対するオーソリティも記録するピア揮発性ソリッドステートストレージ152の順序集合も決定する。特定のオーソリティ168がいくつかの実施形態で利用できない場合、複製の又は代替のオーソリティ168が相談されてよい。

【0056】

【0114】 図2A及び図2Bを参照すると、ストレージノード150上のCPU156の多くのタスクのうちの2つは、書き込みデータを分割し、読取りデータをアセンブルし直すことである。システムが、データが書き込まれると判断するとき、そのデータのオーソリティ168は上記のように位置する。データのセグメントIDがすでに決定されているとき、書き込む要求は、セグメントから決定されたオーソリティ168のホストであると現在判断される揮発性ソリッドステートストレージ152に転送される。揮発性ソリッドステートストレージ152及び対応するオーソリティ168が常駐するストレージノード150のホストCPU156は、次いでデータを分割又はシャード（shard）し、データを多様な揮発性ソリッドステートストレージ152に送出する。送信されたデータは、イレイジャーコーディング方式に従って、データストライプとして書き込まれる。いくつかの実施形態では、データはプルされることが要求され、他の実施形態では、データはプッシュされる。逆に、データが読み取られるとき、データを含んだセグメントIDのためのオーソリティ168は、上述されたように場所を突き止められる。揮発性ソリッドステートストレージ152及び対応するオーソリティ168が常駐するストレージノード150のホストCPU156は、オーソリティによって指される揮発性ソリッドステートストレージ及び対応するストレージノードからデータを要求する。いくつかの実施形態では、データは、データストライプとしてフラッシュストレージから読み取られる。ストレージノード150のホストCPU156は、次いで読み取られたデータをアセンブルし直し、適切なイレイジャーコーディング方式に従って（存在する場合）あらゆるエラーを訂正し、アセンブルし直されたデータをネットワークに転送する。追加の実施形態では、これらのタスクのいくつか又はすべては、揮発性ソリッドステートストレージ152内で処理される場合がある。いくつかの実施形態では、セグメントホストは、ストレージからページを要求し、次いで元の要求をしたストレージノードにデータを送信することによって、データがストレージノード150に送信されることを要求する。

【0057】

【0115】 例えばUNIX様式のファイルシステムにおいて等、いくつかのシステムでは、データは、ファイルシステムのオブジェクトを表すデータ構造を指定するインデックスノードつまりinodeで処理される。オブジェクトは、例えばファイル又はディレクトリである場合があるであろう。メタデータは、他の属性の中で、例えば許可データ及び作成タイムスタンプ等の属性としてオブジェクトに付随してもよい。セグメント番号は、ファイルシステム内の係るオブジェクトのすべて又は一部分に割り当てられる場合があるであろう。他のシステムでは、データセグメントは、他のどこかで割り当てられたセグメント番号で処理される。説明のために、分散の単位はエンティティであり、エンティティはファイル、ディレクトリ、又はセグメントである場合がある。すなわち、エンティティは、ストレージシステムによって記憶されるデータ又はメタデータの単位である。エンティ

ティは、オーソリティと呼ばれる集合にグループ化される。各オーソリティは、オーソリティのエンティティを更新する独占権を有するストレージノードであるオーソリティオーナーを有する。言い換えると、ストレージノードはオーソリティを含み、そのオーソリティは同様にエンティティを含む。

【0058】

【0116】 いくつかの実施形態によれば、セグメントはデータの論理コンテナである。セグメントは、媒体アドレス空間と物理的なフラッシュ場所との間のアドレス空間である。つまり、データセグメント番号はこのアドレス空間内にある。また、セグメントは、高レベルのソフトウェアの関与なしにデータ冗長性を復元する（異なるフラッシュ場所又はデバイスに書き換える）ことを可能にするメタデータを含んでもよい。一実施形態では、セグメントの内部フォーマットは、データの位置を決定するためにクライアントデータ及び中間マッピングを含む。各データセグメントは、適用可能な場合、セグメントをいくつかのデータ及びパリティシャードに分割することによって、例えばメモリ及び他の故障から保護される。データ及びパリティシャードは、イレイジャーコーディング方式に従ってホストCPU156（図2E及び図2Gを参照）に結合された不揮発性ソリッドステートストレージ152全体で分散される、つまりストライプを付けられる。用語セグメントの使用は、いくつかの実施形態では、コンテナ及びセグメントのアドレス空間内でのその位置を指す。用語ストライプの使用は、セグメントと同じシャードの集合を指し、シャードが、いくつかの実施形態に従って冗長性又はパリティ情報とともにどのように分散されるのかを含む。

【0059】

【0117】 一連のアドレス空間変換は、ストレージシステム全体で起こる。上部には、inodeにリンクするディレクトリエントリ（ファイル名）がある。Inodeは、データが論理的に記憶される媒体アドレス空間の中を指す。媒体アドレスは、大きいファイルの負荷を拡散する又は重複排除若しくはスナップショットのようなデータサービスを実施するために一連の間接的な媒体を通してマッピングされてよい。媒体アドレスは、大きいファイルの負荷を拡散する又は重複排除若しくはスナップショットのようなデータサービスを実施するために一連の間接的な媒体を通してマッピングされてよい。セグメントアドレスは、次いで物理的なフラッシュ場所に変換される。物理的なフラッシュ場所は、いくつかの実施形態によればシステム内のフラッシュの量で制限されたアドレス範囲を有する。媒体アドレス及びセグメントアドレスは論理コンテナであり、いくつかの実施形態では、実質的に無限となるように128ビット以上の識別子を使用し、再利用の見込みはシステムの予想寿命よりも長いとして計算される。論理コンテナからのアドレスは、いくつかの実施形態では階層的に割り当てられる。最初に、各不揮発性ソリッドステートストレージユニット152が、アドレス空間の範囲を割り当てられてよい。この割り当てられた範囲内で、不揮発性ソリッドステートストレージ152は、他の不揮発性ソリッドステートストレージ152と同期することなくアドレスを割り当てることができる。

【0060】

【0118】 データ及びメタデータは、変化する作業負荷パターンのために最適化された基本的なストレージレイアウトの集合、及びストレージデバイスごとに記憶される。これらのレイアウトは、複数の冗長性方式、圧縮フォーマット、及びインデックスアルゴリズムを組み込む。これらのレイアウトのいくつかは、オーソリティ及びオーソリティマスタについての情報を記憶する。一方、他は、ファイルメタデータ及びファイルデータを記憶する。冗長性方式は、（例えば、NANDフラッシュチップ等）単一のストレージデバイスの中で破損したビットに耐えるエラー訂正符号、複数のストレージノードの故障に耐えるイレイジャーコード、及びデータセンタ又は地域の故障に耐える複製方式を含む。いくつかの実施形態では、低密度パリティ検査（『LDPC』）符号が、単一のストレージユニットの中で使用される。リードソロモン符号化はストレージクラスタの中で使用され、ミラーリングはいくつかの実施形態ではストレージグリッドの中で使用される。メタデータは、（例えばログ構造化マージツリー等の）順序付けログ構造化インデックスを使用し、

記憶される場合があり、大きいデータは、ログ構造化レイアウトでは記憶されない場合がある。

【0061】

【0119】 エンティティの複数のコピー全体で一貫性を維持するために、ストレージノードは、計算を通して以下の2つのこと、つまり、(1) エンティティを含むオーソリティ、及び(2) オーソリティを含むストレージノードに暗黙のうちに同意する。エンティティのオーソリティに対する割当ては、エンティティをオーソリティに疑似乱数的に割り当てることによって、外部で生成された鍵に基づいてエンティティを範囲に分割することによって、又は単一のエンティティを各オーソリティに入れることによって行われる場合がある。疑似乱数方式の例は、線形ハッシュ法、及びコントロールド・リプリケーション・アンダー・スケーラブル・ハッシング(『CRUSH』)を含んだハッシュのリプリケーション・アンダー・スケーラブル・ハッシング(『RUSH』)ファミリである。いくつかの実施形態では、ノードの集合が変化する場合があるため、疑似乱数割当ては、オーソリティをノードに割り当てるためだけに活用される。オーソリティの集合は変化することができず、したがって任意の主観的な関数がこれらの実施形態で適用されてよい。いくつかの配置方式は、オーソリティをストレージノードに自動的に配置する。一方、他の配置方式は、オーソリティのストレージノードへの明示的なマッピングに依存する。いくつかの実施形態では、疑似乱数方式は、各オーソリティから候補オーソリティオーナーの集合にマッピングするために利用される。CRUSHに関係する疑似乱数データ分散関数は、オーソリティをストレージノードに割り当て、オーソリティが割り当てられる場所のリストを作成してよい。各ストレージノードは、疑似乱数データ分散関数のコピーを有し、オーソリティを分散し、後に発見する又は場所を突き止めるために同じ計算に到達する場合がある。疑似乱数方式のそれぞれは、同じターゲットノードを推断するために、いくつかの実施形態で入力としてストレージノードの到達可能な集合を必要とする。エンティティがオーソリティに入れられると、エンティティは、予想される故障が予想外のデータ損失を生じさせないように物理的デバイスに記憶されてよい。いくつかの実施形態では、リバランスアルゴリズムが、すべてのエンティティのコピーを、同じレイアウト内の、及びマシンの同じ集合上のオーソリティの中に記憶しようと試みる。

【0062】

【0120】 予想される故障の例は、デバイス故障、マシンの盗難、データセンタ火災、及び例えば核事象又は地質学的事象等の地域の災害を含む。種々の故障が、異なるレベルの受容可能なデータ損失につながる。いくつかの実施形態では、ストレージノードの盗難は、システムのセキュリティにも信頼性にも影響を与えない。一方、システム構成によっては、地域の事象は、データ損失なし、数秒若しくは数分の更新の損失、又は完全なデータ損失にもつながる場合があるであろう。

【0063】

【0121】 実施形態では、ストレージの冗長性のためのデータの配置は、データ一貫性のためのオーソリティの配置とは無関係である。いくつかの実施形態では、オーソリティを含むストレージノードは、いずれの永続記憶装置も含まない。代わりに、ストレージノードは、オーソリティを含まない不揮発性ソリッドステートストレージユニットに接続される。ストレージノードと不揮発性ソリッドステートストレージユニットとの間の通信相互接続は、複数の通信技術から成り、不均一な性能特性及びフォールトトレランス特性を有する。いくつかの実施形態では、上述されたように、不揮発性ソリッドステートストレージユニットは、PCIエクスプレスを介してストレージノードに接続され、ストレージノードは、イーサネットバックプレーンを使用し、単一のシャシの中で互いに接続され、シャシは互いに接続されてストレージクラスタを形成する。ストレージクラスタは、いくつかの実施形態では、イーサネット又はファイバチャネルを使用し、クライアントに接続される。複数のストレージクラスタが1つのストレージグリッドの中に構成される場合、複数のストレージクラスタは、インターネット又は例えば「マクロスケール」リンク若しくはインターネットを横断しないプライベートリンク等の他の長距離ネットワーキングリン

クを使用し、接続される。

【0064】

【0122】 オーソリティオーナーは、エンティティを修正する、ある不揮発性ソリッドステートストレージユニットから別の不揮発性ソリッドステートストレージユニットにエンティティを移動させる、並びにエンティティのコピーを付加及び削除する独占権を有する。これは、基本的なデータの冗長性を維持することを可能にする。オーソリティオーナーが故障するとき、廃棄されるとき、又はオーバロードされるとき、オーソリティは新しいストレージノードに転送される。一時障害は、すべての欠陥のないマシンが新しいオーソリティ場所について同意していることを確実にすることを重要にする。一時障害により発生する曖昧性は、リモートシステム管理者による若しくはローカルハードウェア管理者による人手を介して(例えば、故障したマシンをクラスタから物理的に取り外す、又は故障したマシン上でボタンを押すことによって)、例えば Paxos、ホットウォームフェイルオーバー方式等のコンセンサスプロトコルによって自動的に達成できる。いくつかの実施形態では、コンセンサスプロトコルが使用され、フェイルオーバーは自動である。短すぎる期間に多すぎる故障又は複製イベントが発生する場合、システムは、自己保存モードに入り、管理者がいくつかの実施形態に従って介入するまで複製及びデータ移動の活動を休止させる。

10

【0065】

【0123】 オーソリティがストレージノード間で転送され、オーソリティオーナーが自分のオーソリティのエンティティを更新すると、システムはストレージノードと不揮発性ソリッドステートストレージユニットの間でメッセージを転送する。パーシステントメッセージに関して、異なる目的を有するメッセージは異なるタイプである。メッセージのタイプに応じて、システムは異なる順序付け及び耐久性の保証を維持する。パーシステントメッセージが処理されていると、メッセージは一時的に複数の長持ちするストレージハードウェア技術及び長持ちしないストレージハードウェア技術に記憶される。いくつかの実施形態では、メッセージは、RAM、NVRAMに、及びNANDフラッシュデバイスに記憶され、各記憶媒体を効率的に使用するために、さまざまなプロトコルが使用される。待ち時間に敏感なクライアント要求は、複製されたNVRAMで、次いで後にNANDで持続されてよい。一方、背景リバランス動作はNANDに直接的に持続される。

20

【0066】

【0124】 パーシステントメッセージは、送信される前に持続的に記憶される。これは、システムが、故障及び構成要素の交換にも関わらずクライアント要求にサービスを提供し続けることを可能にする。多くのハードウェア構成要素は、システム管理者、製造業者、ハードウェアサプライチェーン、及び継続中のモニタ品質制御インフラに可視である一意の識別子を含むが、インフラストラクチャアドレスの上部で実行中のアプリケーションはアドレスを仮想化する。これらの仮想化されたアドレスは、構成要素の故障及び交換にも関わらずストレージシステムの耐用期間にわたって変化しない。これが、再構成又はクライアント要求の処理の中断なしに、ストレージシステムの各構成要素を経時的に交換できるようにする。つまり、システムは無停止更新をサポートする。

30

【0067】

【0125】 いくつかの実施形態では、仮想化されたアドレスは、十分な冗長性をもって記憶される。連続モニタシステムは、ハードウェア及びソフトウェアのステータス、並びにハードウェアの識別子を相互に関連付ける。これは、欠陥のある構成要素及び製造詳細による故障の検出及び予測を可能にする。また、モニタシステムは、いくつかの実施形態では、構成要素を重大な経路から取り除くことによって故障発生前に影響を受けるデバイスから離すオーソリティ及びエンティティの先を見越した転送を可能にする。

40

【0068】

【0126】 図2Cは、ストレージノード150のコンテンツ、及びストレージノード150の不揮発性ソリッドステートストレージ152のコンテンツを示すマルチレベルブロック図である。データは、いくつかの実施形態では、ネットワークインタフェースコントロ

50

ーラ（『NIC』）202によってストレージノード150に、及びストレージノード150から通信される。各ストレージノード150は、上述されたように、CPU156及び1つ以上の不揮発性ソリッドステートストレージ152を有する。図2Cで1レベル下方に移動すると、各不揮発性ソリッドステートストレージ152は、例えば不揮発性ランダムアクセスメモリ（『NVRAM』）204、及びフラッシュメモリ206等の相対的に高速の不揮発性ソリッドステートメモリを有する。いくつかの実施形態では、NVRAM204は、プログラム／消去サイクル（DRAM、MRAM、PCM）を必要としない構成要素であってよく、メモリが読み取られるよりもはるかにより頻繁に書き込まれることをサポートできるメモリである場合がある。図2Cで別のレベルに下がると、NVRAM204は、エネルギー貯蔵218によってバックアップされる、例えばダイナミックランダムアクセスメモリ（DRAM）216等の高速揮発性メモリとして一実施形態で実施される。エネルギー貯蔵218は、停電の場合には、コンテンツがフラッシュメモリ206に転送されるほど十分に長くDRAM216に電力を供給させておくほど十分な電力を提供する。いくつかの実施形態ではエネルギー貯蔵218は、電力損失の場合に安定した記憶媒体へのDRAM216のコンテンツの転送を可能にするほど十分なエネルギーの適切な供給を供給するコンデンサ、超コンデンサ、電子、又は他のデバイスである。フラッシュメモリ206は、フラッシュダイ222のパッケージ又はフラッシュダイ222のアレイとも呼ばれてよい、複数のフラッシュダイ222として実装される。フラッシュダイ222は、パッケージあたり単一のダイ、パッケージあたり複数のダイ（つまり、マルチチップパッケージ）で、ハイブリッドパッケージで、プリント基板又は他の基板上でむき出しのダイとして、カプセル化されたダイとして等、任意のいくつかの方法でパッケージ化できると理解されることが理解されるべきである。示される実施形態では、不揮発性ソリッドステートストレージ152は、コントローラ212又は他のプロセッサを有し、コントローラ212に結合された入出力（I/O）ポート210を有する。I/Oポート210は、CPU156、及び／又はフラッシュストレージノード150のネットワークインタフェースコントローラ202に結合される。フラッシュ入出力（I/O）ポート220は、フラッシュダイ222に結合され、ダイレクトメモリアccessユニット（DMA）214は、コントローラ212、DRAM216、及びフラッシュダイ222に結合される。示される実施形態では、I/Oポート210、コントローラ212、DMAユニット214、及びフラッシュI/Oポート220は、例えばフィールドプログラマブルゲートアレイ（FPGA）等、プログラム可能論理回路（『PLD』）208上に実装される。本実施形態では、各フラッシュダイ222は、16kB（キロバイト）のページ224として編成されるページ、及びそれを通してデータをフラッシュダイ222に書き込む又はフラッシュダイ222から読み取ることができるレジスタ226を有する。追加の実施形態では、他のタイプのソリッドステートメモリが、フラッシュダイ222の中に示されるフラッシュメモリの代わりに、又はフラッシュメモリに加えて使用される。

【0069】

【0127】 ストレージクラス161は、本明細書に開示される多様な実施形態では、一般にストレージアレイと対比される場合がある。ストレージノード150は、ストレージクラス161を作成する集合体の一部である。各ストレージノード150は、データの
スライス及びデータを提供するために必要とされるコンピューティングを所有する。複数の
ストレージノード150は、データを記憶し、取り出すために協調する。一般にストレ
ージアレイで使用されるストレージメモリ又はストレージデバイスは、データを処理し、
操作することによりあまり関与していない。ストレージアレイ内のストレージメモリ又はスト
レージデバイスは、データを読み取る、書き込む、又は消去するコマンドを受け取る。スト
レージアレイ内のストレージメモリ又はストレージデバイスは、それらが埋め込まれる
より大きいシステム、又はデータが何を意味するのかを認識していない。ストレージアレ
イ内のストレージメモリ又はストレージデバイスは、例えばRAM、ソリッドステートド
ライブ、ハードディスクドライブ等の多様なタイプのストレージメモリを含む場合がある
。本明細書に説明されるストレージユニット152は、同時にアクティブであり、複数の

10

20

30

40

50

目的を果たす複数のインタフェースを有する。いくつかの実施形態では、ストレージノード150の機能性のいくつかは、ストレージユニット152の中にシフトされ、ストレージユニット152をストレージユニット152及びストレージノード150の組合せに変換する。(ストレージデータに対する)計算をストレージユニット152の中に入れることは、この計算をデータ自体により近く配置する。多様なシステム実施形態は、異なる機能を有するストレージノード層の階層を有する。対照的に、ストレージアレイでは、コントローラは、コントローラがシェルフ又はストレージデバイス内で管理するデータのすべてについてのあらゆるものを所有し、知っている。ストレージクラス161では、本明細書に説明されるように、複数のストレージユニット152及び/又はストレージノード150の複数のコントローラが、(例えば、イレイジャーコーディング、データシャード、メタデータ通信及び冗長性、記憶容量拡大又は縮小、データ復旧等のための)多様な方法で協調する。

10

【0070】

【0128】 図2Dは、図2Aから図2Cのストレージノード150及びストレージユニット152の実施形態を使用する、ストレージサーバ環境を示す。このバージョンでは、各ストレージユニット152は、コントローラ212(図2Cを参照)等のプロセッサ、FPGA(フィールドプログラマブルゲートアレイ)、フラッシュメモリ206、及びシャシ138(図2Aを参照)内のPCIe(ペリフェラルコンポーネントインターコネクトエクスプレス)ボード上の(超コンデンサによりバックアップされるDRAM216である、図2B及び図2Cを参照)NVRAM204を有する。ストレージユニット152は、ストレージを含んだ単一の基板として実装されてよく、シャシ内側の最大の許容故障ドメインであってよい。いくつかの実施形態では、最大で2つのストレージユニット152が故障し得、デバイスはデータ損失なしに続行する。

20

【0071】

【0129】 物理ストレージは、いくつかの実施形態でアプリケーションの使用に基づき名付けられた領域に分割される。NVRAM204は、ストレージユニット152DRAM216内の予約メモリの連続ブロックであり、NANDフラッシュによってバックアップされる。NVRAM204は、スプール(例えば、spool_region)として2つのために書き込まれる複数のメモリ領域に論理的に分割される。NVRAM204スプールの中の空間は、独立して各オーソリティ168によって管理される。各デバイスは、ストレージ空間の量を各オーソリティ168に提供する。そのオーソリティ168は、その空間の中の寿命及び割当てをさらに管理する。スプールの例は、分散されたトランザクション又は概念を含む。ストレージユニット152への一次電力が故障するとき、搭載された超コンデンサは、短期間の電力ホールドアップを提供する。このホールドアップ間隔の間、NVRAM204のコンテンツは、フラッシュメモリ206にフラッシュされる。次の電力投入時、NVRAM204のコンテンツは、フラッシュメモリ206から取り戻される。

30

【0072】

【0130】 ストレージユニットコントローラに関して、論理「コントローラ」の責任は、オーソリティ168を含んだブレードのそれぞれ全体で分散される。この論理制御の分散は、ホストコントローラ242、中間層コントローラ244、及びストレージユニットコントローラ(複数可)246として図2Dに示される。パーツは物理的に同じブレード上に配置されてよいが、制御プレーン及びストレージプレーンの管理は、独立して処理される。各オーソリティ168は、実質的には独立したコントローラとしての機能を果たす。各オーソリティ168は、独自のデータ構造及びメタデータ構造、独自のバックグラウンドワーカーを提供し、独自のライフサイクルを維持する。

40

【0073】

【0131】 図2Eは、ブレード252ハードウェアブロック図であり、制御プレーン254、計算プレーン、及びストレージプレーン256、258、並びに図2Dのストレージサーバ環境で図2Aから図2Cのストレージノード150及びストレージユニット152

50

の実施形態を使用し、基本的な物理リソースと対話するオーソリティ 168 を示す。制御プレーン 254 は、ブレード 252 のいずれかで実行するために計算プレーン 256 内の計算リソースを使用できるいくつかのオーソリティ 168 に区分化される。ストレージプレーン 258 はデバイスの集合に区分化され、デバイスのそれぞれがフラッシュ 206 及び NVRAM 204 リソースへのアクセスを提供する。

【0074】

【0132】 図 2 E の計算プレーン及びストレージプレーン 256、258 では、オーソリティ 168 は、基本的な物理リソース（つまり、デバイス）と対話する。オーソリティ 168 の観点から、そのリソースは、物理的デバイスのすべての上でストライプを付けられる。デバイスの観点から、それは、オーソリティがどこでたまたま実行するのかには関係なく、すべてのオーソリティ 168 にリソースを提供する。各オーソリティ 168 は、例えばフラッシュメモリ 206 及び NVRAM 204 のパーティション 260 等、ストレージユニット 152 内のストレージメモリの 1 つ以上のパーティション 260 を割り当てた、又は割り当てられた。各オーソリティ 168 は、ユーザデータを書き込む又は読み取るために、それに属する、それらの割り当てられたパーティション 260 を使用する。オーソリティは、システムの異なる量の物理ストレージと関連付けられる場合がある。例えば、1 つのオーソリティ 168 は、1 つ以上の他のオーソリティ 168 よりも多数のパーティション 260 又はより大きいサイズのパーティション 260 を 1 つ以上のストレージユニット 152 に有する場合があるであろう。

【0075】

【0133】 図 2 F は、いくつかの実施形態に従って、ストレージクラスタのブレード 252 のエラスティシティソフトウェア層を示す。エラスティシティ構造では、エラスティシティソフトウェアは対称的である。つまり、各ブレードの計算モジュール 270 は、図 2 F に示されるプロセスの 3 つの同一の層を実行する。ストレージマネージャ 274 は、ローカルストレージユニット 152 NVRAM 204、及びフラッシュ 206 に記憶されたデータ及びメタデータに対する他のブレード 252 からの読取り要求及び書き込み要求を実行する。オーソリティ 168 は、必要な読取り及び書き込みを、そのストレージユニット 152 上で、対応するデータ又はメタデータが常駐するブレード 252 に発行することによってクライアント要求に応える。エンドポイント 272 は、スイッチファブリック 146 監督ソフトウェアから受け取られたクライアント接続要求を解析し、クライアント接続要求を、達成を担うオーソリティ 168 に中継し、オーソリティ 168 の応答をクライアントに中継する。対称的な 3 層構造は、ストレージシステムの高度の同時並行性を可能にする。エラスティシティは、これらの実施形態で効率的に且つ確実にスケールアウトする。さらに、エラスティシティは、クライアントアクセスパターンとは関係なく、すべてのリソース全体で均等に作業のバランスをとり、通常は、従来の分散ロックで発生するブレード間の調整の必要性の多くを排除することによって同時並行性を最大限にする一意のスケールアウトを実装する。

【0076】

【0134】 まだ図 2 F を参照すると、ブレード 252 の計算モジュール 270 で実行するオーソリティ 168 は、クライアント要求に応えるために必要とされる内部動作を実行する。エラスティシティの 1 つの特徴は、オーソリティ 168 がステートレスであること、つまりオーソリティが、高速アクセスのためにアクティブなデータ及びメタデータを独自のブレード 252 の DRAM にキャッシュに入れるが、オーソリティは、更新がフラッシュ 206 に書き込まれるまで 3 つの別々のブレード 252 上のその NVRAM 204 パーティションのあらゆる更新を記憶する点である。NVRAM 204 に対するすべてのストレージシステムの手書きは、いくつかの実施形態では、3 つの別々のブレード 252 上でのパーティションに対して 3 つある。3 倍にミラーリングされた NVRAM 204 並びにパリティ及びリードソロモン RAID チェックサムによって保護された永続記憶装置を用いると、ストレージシステムは、データ、メタデータ、又はどちらかへのアクセスを失うことなく 2 つのブレード 252 の同時故障を乗り切ることができる。

【0077】

【0135】 オーソリティ168はステートレスであるため、オーソリティはブレード252間で移行できる。各オーソリティ168は一意の識別子を有する。NVRAM204及びフラッシュ206のパーティションは、オーソリティがいくつか実行しているブレード252とではなく、オーソリティ168の識別子と関連付けられる。したがって、オーソリティ168が移行するとき、オーソリティ168は、その新しい場所から同じストレージパーティションを管理し続ける。新しいブレード252がストレージクラスタの実施形態にインストールされるとき、システムは、システムのオーソリティ168による使用のために新しいブレード252のストレージを区画化し、選択されたオーソリティ168を新しいブレード252に移行し、新しいブレード252でエンドポイント272を開始し、それらをスイッチファブリック146のクライアント接続分散アルゴリズムに含めることによって自動的に負荷のバランスを取り直す。

10

【0078】

【0136】 その新しい場所から、移行されたオーソリティ168は、フラッシュ206上でそのNVRAM204パーティションのコンテンツを持続し、他のオーソリティ168からの読取り要求及び書込み要求を処理し、エンドポイント272がそれらに向けるクライアント要求に応える。同様に、ブレード252が故障する又は取り除かれる場合、システムは、システムの残りのブレード252の間でそのオーソリティ168を分散し直す。分散し直されたオーソリティ168は、その新しい場所からその元の機能を実行し続ける。

【0079】

20

【0137】 図2Gは、いくつかの実施形態に係る、オーソリティ168、及びストレージクラスタのブレード252のストレージリソースを示す。各オーソリティ168は、各ブレード252でのフラッシュ206及びNVRAM204のパーティションを独占的に担う。オーソリティ168は、他のオーソリティ168とは無関係にそのパーティションのコンテンツ及び完全性を管理する。オーソリティ168は、入信データを圧縮し、それを一時的にそのNVRAM204パーティションに保存し、次いでそのフラッシュ206パーティション内のストレージのセグメントでデータを強固にし、RAID保護し、持続する。オーソリティ168がフラッシュ206にデータを書き込むと、ストレージマネージャ274は、書込み性能を最適化し、媒体寿命を最大限にするために必要なフラッシュ変換を実行する。背景では、オーソリティ168は、「ガーベージコレクト」する、又はクライアントがデータを上書きすることによって時代遅れのものにするデータによって占有される空間を取り戻す。オーソリティ168のパーティションは共通の要素をもたないので、クライアント及び書込みを実行する、又はバックグラウンド機能を実行する分散されたロックに対する必要性がないことが理解されるべきである。

30

【0080】

【0138】 本明細書に記載される実施形態は、多様なソフトウェアプロトコル、通信プロトコル、及び／又はネットワーキングプロトコルを活用してよい。さらに、ハードウェア及び／又はソフトウェアの構成は、多様なプロトコルに対応するために調整されてよい。例えば、実施形態は、WINDOWS（商標）環境で認証、ディレクトリ、方針、及び他のサービスを提供するデータベースをベースにしたシステムであるアクティブディレクトリを利用してよい。これらの実施形態では、LDAP（ライトウェイトディレクトリアクセスプロトコル）は、例えばアクティブディレクトリ等のディレクトリサービスプロバイダでアイテムを問い合わせし、修正するための一例のアプリケーションプロトコルである。いくつかの実施形態では、ネットワークロックマネージャ（『NLM』）は、システムV様式のアドバイザリーファイルを提供し、ネットワーク上でロックを記録するためにネットワークファイルシステム（『NFS』）と協調して機能する機構として利用される。その1つのバージョンが共通インターネットファイルシステム（『CIFS』）としても知られるサーバメッセージブロック（『SMB』）プロトコルは、本明細書で説明されるストレージシステムと統合されてよい。SMPは、ファイル、プリンタ、及びシリアルポートに共用アクセス、並びにネットワーク上のノード間のさまざまな通信を提供するため

40

50

に通常使用されるアプリケーション層ネットワークプロトコルとして動作する。また、SMBは、認証されたプロセス間通信機構も提供する。AMAZON（商標）S3（シンプルストレージサービス）は、Amazon Web Servicesによって提供されるウェブサービスであり、本明細書に説明されるシステムは、ウェブサービスインタフェース（REST（表現可能な状態の転送）、SOAP（シンプルオブジェクトアクセスプロトコル）、及びBitTorrent）を通してAmazon S3と連動してよい。RESTful API（アプリケーションプログラミングインタフェース）は、一連の小さいモジュールを作成するためにトランザクションを分解する。各モジュールは、トランザクションの特定の基本的な部分をアドレス指定する。特にオブジェクトデータのために、これらの実施形態で提供される制御又は許可は、アクセス制御リスト（『ACL』）の活用を含んでよい。ACLは、オブジェクトに付与される許可のリストであり、ACLは、どの動作が所与のオブジェクトで許可されるのかだけではなく、どのユーザプロセス又はシステムプロセスがオブジェクトへのアクセスを許可されるのかも指定する。システムは、ネットワーク上のコンピュータのために識別及び位置選定システムを提供し、インターネット全体でトラフィックを送る通信プロトコルのために、IPv4だけではなくインターネットプロトコルバージョン6（『IPv6』）も活用する。ネットワーク化されたシステム間でのパケットのルーティングは、イコールコストマルチパスルーティング（『ECMP』）を含んでよい。ECMPは、単一の宛先に転送するネクストホップパケットが、ルーティングメトリック計算で一番上の場所に並ぶ複数の「最善の経路」上で発生する場合があるルーティング戦略である。マルチパスルーティングは、単一ルータに制限されたホップ単位の決定であるため、マルチパスルーティングは、大部分のルーティングプロトコルと併せて使用できる。ソフトウェアは、ソフトウェアアプリケーションの単一インスタンスが複数のカスタマにサービスを提供するアーキテクチャであるマルチテナンシーをサポートしてよい。各カスタマは、テナントと呼ばれてよい。テナントは、アプリケーションのいくつかの部分をカスタマイズする能力を与えられてよいが、いくつかの実施形態では、アプリケーションのコードをカスタマイズしない場合がある。実施形態は、監査ログを維持してよい。監査ログは、コンピューティングシステムにおけるイベントを記録する文書である。どのリソースがアクセスされたのかを文書化することに加えて、監査ログエントリは、通常、宛先アドレス及びソースアドレス、タイムスタンプ、並びに多様な規則とのコンプライアンスのためのユーザログイン情報を含む。実施形態は、例えば暗号化鍵ローテーション等の多様な鍵管理方針をサポートしてよい。さらに、システムは、動的なルートパスワード又はいくつかの変形動的変化パスワードをサポートしてよい。

【0081】

【0139】 図3Aは、本開示のいくつかの実施形態に従って、クラウドサービスプロバイダ302とのデータ通信のために結合されるストレージシステム306の図を説明する。あまり詳細には示されないが、図3Aに示されるストレージシステム306は、図1Aから図1D及び図2Aから図2Gに関して上述されたストレージシステムに類似してよい。いくつかの実施形態では、図3Aに示されるストレージシステム306は、アンバランスなアクティブ／アクティブコントローラを含むストレージシステムとして、バランスアクティブ／アクティブコントローラを含むストレージシステムとして、各コントローラのすべてに満たないリソースが、各コントローラがフェイルオーバをサポートするために使用され得る予備のリソースを有するように活用されるアクティブ／アクティブコントローラを含むストレージシステムとして、完全にアクティブ／アクティブコントローラを含むストレージシステムとして、データセット分離コントローラを含むストレージシステムとして、フロントエンドコントローラ及びバックエンド統合ストレージコントローラを有する二重層アーキテクチャを含むストレージシステムとして、二重コントローラアレイのスケールアウトクラスタを含むストレージシステムとして、及び係る実施形態の組合せとして実施されてよい。

【0082】

【0140】 図3Aに示される例では、ストレージシステム306は、データ通信リンク3

04を介してクラウドサービスプロバイダ302に結合される。データ通信リンク304は、専用のデータ通信リンクとして、例えば広域ネットワーク（『WAN』）若しくはローカルエリアネットワーク（『LAN』）等のデータ通信ネットワーク、又はストレージシステム306とクラウドサービスプロバイダ302との間でデジタル情報をトランスポートできる他のなんらかの機構のうちの1つの使用により提供されるデータ通信経路として実施されてよい。係るデータ通信リンク304は、完全に有線、完全に無線、又は有線データ通信経路及び無線データ通信経路のなんらかの集約であってよい。係る例では、デジタル情報は、1つ以上のデータ通信プロトコルを使用し、データ通信リンク304を介してストレージシステム306とクラウドサービスプロバイダ302との間で交換されてよい。例えば、デジタル情報は、ハンドヘルドデバイス転送プロトコル（『HDT P』）、ハイパーテキスト転送プロトコル（『HTT P』）、インターネットプロトコル（『I P』）、リアルタイム転送プロトコル（『RT P』）、伝送制御プロトコル（『TC P』）、ユーザデータグラムプロトコル（『UD P』）、無線アプリケーションプロトコル（『WAP』）、又は他のプロトコルを使用し、データ通信リンク304を介してストレージシステム306とクラウドサービスプロバイダ302との間で交換されてよい。

【0083】

【0141】 図3Aに示されるクラウドサービスプロバイダ302は、例えばデータ通信リンク304を介したコンピューティングリソースの共用を通してクラウドサービスプロバイダ302のユーザにサービスを提供するシステム及びコンピューティング環境として実施されてよい。クラウドサービスプロバイダ302は、例えばコンピュータネットワーク、サーバ、ストレージ、アプリケーション、及びサービス等の構成可能なコンピューティングリソースの共用プールにオンデマンドアクセスを提供してよい。構成可能なリソースの共用プールは、迅速にセットアップし、最小の管理努力でクラウドサービスプロバイダ302のユーザにリリースされてよい。概して、クラウドサービスプロバイダ302のユーザは、サービスを提供するためにクラウドサービスプロバイダ302によって利用される的確なコンピューティングリソースを認識していない。多くの場合、係るクラウドサービスプロバイダ302は、インターネットを介してアクセス可能であってよいが、当業の読者は、任意のデータ通信リンクを通してユーザにサービスを提供するために共用リソースの使用を概念化する任意のシステムが、クラウドサービスプロバイダ302と見なされてよいことを認識する。

【0084】

【0142】 図3Aに示される例では、クラウドサービスプロバイダ302は、さまざまなサービスをストレージシステム306及びストレージシステム306のユーザに多様なサービスモデルの実施態様を通して提供しようと構成されてよい。例えば、クラウドサービスプロバイダ302は、ストレージシステム306及びストレージシステム306のユーザに、クラウドサービスプロバイダ302が加入者へのサービスとして、例えばバーチャルマシン及び他のリソース等のコンピューティングインフラストラクチャを提供するインフラストラクチャズアサービス（『IaaS』）サービスモデルの実施態様を通してサービスを提供するように構成されてよい。さらに、クラウドサービスプロバイダ302は、ストレージシステム306及びストレージシステム306のユーザに、クラウドサービスプロバイダ302がアプリケーション開発者に開発環境を提供するプラットフォームズアサービス（『PaaS』）サービスモデルの実施態様を通してサービスを提供するように構成されてよい。係る開発環境は、例えばオペレーティングシステム、プログラミング言語実行環境、データベース、ウェブサーバ、又はクラウドプラットフォーム上でソフトウェア解決策を開発し、実行するためにアプリケーション開発者によって利用されてよい他の構成要素を含んでよい。さらに、クラウドサービスプロバイダ302は、ストレージシステム306及びストレージシステム306のユーザに対してアプリケーションを実行するために使用されるプラットフォームだけではなく、クラウドサービスプロバイダ302がアプリケーションソフトウェア、データベースを提供するソフトウェアズアサービス（『SaaS』）サービスモデルの実施態様を通して、ストレージシステム306及

びストレージシステム 306 のユーザにサービスを提供するように構成されてよく、ストレージシステム 306 及びストレージシステム 306 のユーザにオンデマンドソフトウェアを提供し、ローカルコンピュータ上でアプリケーションをインストールし、実行する必要性を排除し、アプリケーションの保守及びサポートを簡略化してよい。クラウドサービスプロバイダ 302 は、ストレージシステム 306 及びストレージシステム 306 のユーザに、アプリケーション、データソース、又は他のリソースに対するアクセスを保護するために使用できる認証サービスをクラウドサービスプロバイダ 302 が提供するオーセンティケーションアズアサービス (authentication as a service) (『AaaS』) サービスモデルの実施態様を通してサービスを提供するようにさらに構成されてよい。また、クラウドサービスプロバイダ 302 は、ストレージシステム 306 及びストレージシステム 306 のユーザに、クラウドサービスプロバイダ 302 がストレージシステム 306 及びストレージシステム 306 のユーザによる使用のために、そのストレージインフラストラクチャへのアクセスを提供するストレージアズアサービスモデルの実施態様を通してサービスを提供するように構成されてもよい。上述されたサービスモデルは、説明のためにだけ含まれ、クラウドサービスプロバイダ 302 によって提供されてよいサービスの制限、又はクラウドサービスプロバイダ 302 によって実装されてよいサービスモデルに関する制限をまったく表さないもので、読者は、クラウドサービスプロバイダ 302 が、追加のサービスモデルの実施態様を通してストレージシステム 306 及びストレージシステム 306 のユーザに追加のサービスを提供するように構成されてよいことを理解する。

【0085】

[0143] 図 3A に示される例では、クラウドサービスプロバイダ 302 は、例えばプライベートクラウドとして、パブリッククラウドとして、又はプライベートクラウドとパブリッククラウドの組合せとして実施されてよい。クラウドサービスプロバイダ 302 がプライベートクラウドとして実施される実施形態では、クラウドサービスプロバイダ 302 は、複数の組織にサービスを提供するよりむしろ単一の組織にサービスを提供することに専念してよい。クラウドサービス 302 がパブリッククラウドとして実施される実施形態では、クラウドサービスプロバイダ 302 は複数の組織にサービスを提供してよい。パブリッククラウド及びプライベートクラウドの配備モデルは異なる場合があり、多様な優位点及び不利な点を伴う場合がある。例えば、パブリッククラウド配備は異なる組織全体でコンピューティングインフラストラクチャの共用を必要とするため、係る配備はセキュリティの懸念がある組織、ミッションクリティカルな作業負荷、アップタイム要件需要等にとっては理想的ではない場合がある。プライベートクラウド配備はこれらの問題のいくつかに取り組むことができるが、プライベートクラウド配備はプライベートクラウドを管理するために自社運用スタッフを必要とする場合がある。さらに代替の実施形態では、クラウドサービスプロバイダ 302 は、プライベートクラウドサービス及びパブリッククラウドサービスのハイブリッドクラウド配備との混合として実施されてよい。

【0086】

[0144] 図 3A には明示的に示されていないが、読者は、追加のハードウェア構成要素及び追加のソフトウェア構成要素が、ストレージシステム 306 及びストレージシステム 306 のユーザに対するクラウドサービスの送達を容易にするために必要である場合があることを理解する。例えば、ストレージシステム 306 は、クラウドストレージゲートウェイに結合されて (又は含んでも) よい。係るクラウドストレージゲートウェイは、例えばストレージシステム 306 を有する施設に位置するハードウェアベース又はソフトウェアベースの機器として実施されてよい。係るクラウドストレージゲートウェイは、ストレージレイ 306 上で実行中のローカルアプリケーションと、ストレージレイ 306 によって活用される遠隔のクラウドベースのストレージとの間のブリッジとして動作し得る。クラウドストレージゲートウェイの使用により、組織は、一時的な iSCSI 又は NAS をクラウドサービスプロバイダ 302 に移動し、組織がそのオンプレミス型ストレージシステムでの場所を節約できるようにする。係るクラウドストレージゲートウェイは、S

C S I コマンド、ファイルサーバコマンド、又は他の適切なコマンドを、クラウドサービスプロバイダ 302 との通信を容易にする R E S T 空間プロトコルに変換できるディスクアレイ、ブロックベースのデバイス、ファイルサーバ、又は他のストレージシステムにエミュレートするように構成されてよい。

【0087】

[0145] ストレージシステム 306 及びストレージシステム 306 のユーザが、クラウドサービスプロバイダ 302 によって提供されるサービスを利用できるようにするために、クラウド移行プロセスは、組織のローカルシステムから（又は別のクラウド環境から）のデータ、アプリケーション、又は他の要素がクラウドサービスプロバイダ 302 に移動される間に起こる場合がある。データ、アプリケーション、又は他の要素をクラウドサービスプロバイダ 302 の環境に無事に移行するためには、クラウド移行ツール等のミドルウェアが、クラウドサービスプロバイダ 302 の環境と組織の環境との間のギャップを埋めるために活用されてよい。また、係るクラウド移行ツールは、データ通信ネットワーク上でクラウドサービスプロバイダ 302 に対する極秘データと関連したセキュリティの懸念に対応するだけではなく、潜在的に高いネットワークコストとクラウドサービスプロバイダ 302 への大量のデータの移行に関連した長い移行時間に対応するように構成されてもよい。ストレージシステム 306 及びストレージシステム 306 のユーザがさらにクラウドサービスプロバイダ 302 によって提供されるサービスを利用できるようにするために、クラウドオーケストレータが使用されて、統合されたプロセス又はワークフローの作成を求めて自動化されたタスクを準備し、調整してよい。係るクラウドオーケストレータは、係る構成要素間で相互接続を管理することだけではなく、それらの構成要素がクラウド構成要素なのか、それともオンプレミス構成要素なのかに関わりなく、多様な構成要素を構成する等のタスクを実行してよい。クラウドオーケストレータは、リンクが正しく構成され、保守されていることを確実にするために、構成要素間の通信及び接続を簡略化できる。

【0088】

[0146] 図 3 A に示される例では、及び上記に簡略に説明されるように、クラウドサービスプロバイダ 302 は、ストレージシステム 306 及びストレージシステム 306 のユーザに対してアプリケーションを実行するために使用されるプラットフォームだけではなく、クラウドサービスプロバイダ 302 がアプリケーションソフトウェア、データベースを提供する SaaS サービスモデルの使用によって、ストレージシステム 306 及びストレージシステム 306 のユーザにサービスを提供するように構成されてよく、ストレージシステム 306 及びストレージシステム 306 のユーザにオンデマンドソフトウェアを提供し、ローカルコンピュータ上にアプリケーションをインストールし、実行する必要性を排除し、このことがアプリケーションの保守及びサポートを簡略化し得る。係るアプリケーションは、本開示の多様な実施形態に従って多くの形をとってよい。例えば、クラウドサービスプロバイダ 302 は、データ解析アプリケーションへのアクセスをストレージシステム 306 及びストレージシステム 306 のユーザに提供するように構成されてよい。係るデータ解析アプリケーションは、例えば、ストレージシステム 306 によって自社サーバと通信されるテレメトリデータを受け取るように構成される場合がある。係るテレメトリデータは、ストレージシステム 306 の多様な動作特性を記述する場合があり、例えばストレージシステム 306 の健康を決定するために、ストレージシステム 306 上で実行している作業負荷を識別するために、ストレージシステム 306 がいつ多様なリソースを使い果たすのかを予測するために、ストレージシステム 306 の動作を改善し得る構成変更、ハードウェア又はソフトウェアのアップグレード、ワークフロー移行、又は他のアクションを推奨するために解析されてよい。

【0089】

[0147] また、クラウドサービスプロバイダ 302 は、ストレージシステム 306 及びストレージシステム 306 のユーザに、仮想化されたコンピューティング環境へのアクセスを提供するように構成されてもよい。係る仮想化されたコンピューティング環境は、例

例えばバーチャルマシン又は他の仮想化コンピュータハードウェアプラットフォーム、仮想ストレージデバイス、仮想化コンピュータネットワークリソース等として実施されてよい。係る仮想化環境の例は、実際のコンピュータをエミュレートするために創られるバーチャルマシン、論理デスクトップを物理マシンから分離する仮想化デスクトップ環境、異なるタイプの具体的なファイルシステムに対する一様なアクセスを可能にする仮想化ファイルシステム、及び多くの他のものを含む場合がある。

【0090】

【0148】 追加の説明のために、図3Bは、本開示のいくつかの実施形態に係るストレージシステム306の図を説明する。あまり詳しく説明されていないが、ストレージシステムは上述された構成要素の多くを含んでよいので、図3Bに示されるストレージシステム306は、図1Aから図1D及び図2Aから図2Gを参照して上述されたストレージシステムに類似してよい。

【0091】

【0149】 図3Bに示されるストレージシステム306は、多くの形で実施されてよいストレージリソース308を含んでよい。例えば、いくつかの実施形態では、ストレージリソース308は、ナノRAM又は基板上に付着されたカーボンナノチューブを利用する別の形の不揮発性ランダムアクセスメモリを含む場合がある。いくつかの実施形態では、ストレージリソース308は、積み重ね可能な交差格子付き(cross-gridded)データアクセスアレイと併せて、ビットストレージがバルク抵抗の変更に基づく3Dクロスポイント不揮発性メモリを含んでよい。いくつかの実施形態では、ストレージリソース308は、単一レベルセル(『SLC』)NANDフラッシュ、マルチレベルセル(『MLC』)NANDフラッシュ、トリプルレベルセル(『TLC』)NANDフラッシュ、クワッドレベルセル(『QLC』)NANDフラッシュ及び他を含んだフラッシュメモリを含んでよい。いくつかの実施形態では、ストレージリソース308は、データが磁気記憶素子を使用して記憶されるスピン移動トルク(『STT』)MRAMを含んだ不揮発性磁気抵抗ランダムアクセスメモリ(『MRAM』)を含んでよい。いくつかの実施形態では、例のストレージリソース308は、セルがいくつかの別個の中間状態を達成できるので、単一のセルに複数のビットを保持する能力を有する場合がある不揮発性相変化メモリ(『PCM』)を含んでよい。いくつかの実施形態では、ストレージリソース308は、光量子情報の記憶及び取出しを可能にする量子メモリを含んでよい。いくつかの実施形態では、例のストレージリソース308は、誘電ソリッドステート材料全体で抵抗を変更することによってデータが記憶される、抵抗変化型メモリ(『ReRAM』)を含んでよい。いくつかの実施形態では、ストレージリソース308は、リッドステート不揮発性メモリが、サブリソグラフィパターン化(sublithographic patterning)技術、セル単位の複数ビット、デバイスの複数の層等のなんらかの組合せを使用し、高密度で製造されてよい、ストレージクラスメモリ(『SCM』)を含んでよい。読者は、他の形のコンピュータメモリ及びストレージデバイスが、DRAM、SRAM、EEPROM、ユニバーサルメモリ、及び多くの他のものを含んだ上述されたストレージシステムによって利用されてよいことを理解する。図3Aに示されるストレージリソース308は、デュアルインラインメモリモジュール(『DIMM』)、不揮発性デュアルインラインメモリモジュール(『NVDIMM』)、M.2、U.2、及び他を含むが、これに限定されるものではない、さまざまなフォームファクタで実施されてよい。

【0092】

【0150】 図3Bに示される例のストレージシステム306は、さまざまなストレージアーキテクチャを実装してよい。例えば、本開示のいくつかの実施形態に係るストレージシステムは、データがブロックで記憶され、各ブロックが基本的に個々のハードドライブの機能を果たすブロックストレージを活用してよい。本開示のいくつかの実施形態に係るストレージシステムは、データがオブジェクトとして管理されるオブジェクトストレージを活用してよい。各オブジェクトは、データ自体、可変量のメタデータ、及びグローバル意識別子を含んでよく、オブジェクトストレージは複数のレベル(例えば、デバイスレベ

10

20

30

40

50

ル、システムレベル、インタフェースレベル)で実装できる。本開示のいくつかの実施形態に係るストレージシステムは、データが階層構造で記憶されるファイルストレージを利用する。係るデータはファイル及びフォルダで保存され、それを記憶するシステムと、同じフォーマットでそれを取り出すシステムの両方に提示される。

【0093】

【0151】 図3Bに示される例のストレージシステム306は、追加のストレージリソースを、スケールアップモデルを使用して加えることができ、追加のストレージリソースを、スケールアウトモデルを使用して加えることができる、又はそのいくつかの組合せによって、ストレージシステムとして実装されてよい。スケールアップモデルでは、追加のストレージは、追加のストレージデバイスを加えることによって加えられてよい。しかしながら、スケールアウトモデルでは、追加のストレージノードはストレージノードのクラスタに加えられてよく、係るストレージノードは追加の処理リソース、追加のネットワーキングリソース等を含む場合がある。

【0094】

【0152】 また、図3Bに示されるストレージシステム306は、ストレージシステム306と、ストレージシステム306の外部にあるコンピューティングデバイスとの間のデータ通信だけではなく、ストレージシステム306の中の構成要素間のデータ通信も容易にする際に有用である場合がある通信リソース310も含む。通信リソース310は、ストレージシステムの外部にあるコンピューティングデバイスだけではなく、ストレージシステムの中の構成要素間のデータ通信も容易にするためにさまざまな異なるプロトコル及びデータ通信ファブリックを活用するように構成されてよい。例えば、通信リソース310は、FCネットワークを介してSCSIコマンドをトランスポートする場合があるFCファブリック及びFCプロトコル等のファイバチャネル(『FC』)技術を含む場合がある。また、通信リソース310は、FCフレームがカプセル化され、イーサネットネットワークを介して伝送されるFCオーバーイーサネット(『FCoE』)技術も含む場合がある。また、通信リソース310は、スイッチドファブリック技術が、チャネルアダプタ間での伝送を容易にするために利用されるインフィニバンド(『IB』)技術を含む場合もある。また、通信リソース310は、PCIエクスプレス(『PCIe』)バスを介してアタッチされた不揮発性記憶媒体がそれを通してアクセスされ得るNVMエクスプレス(『NVMe』)技術及びNVMeオーバーファブリック(『NVMe over F』)技術も含む場合がある。また、通信リソース310は、ストレージシステム306の中のストレージリソース308に、及びストレージシステム306とストレージシステム306の外部にあるコンピューティングデバイスとの間のデータ通信だけではなく、ストレージシステム306の中の構成要素間のデータ通信も容易にするために有用であってよい他の通信リソースにブロックレベルのアクセスを提供するために、シリアルアタッチドSCSI(『SAS』)、ストレージシステム306の中のストレージリソース308をストレージシステム306の中のホストバスアダプタに接続するためのシリアルATA(『SATA』)バスインタフェース、インターネットスモールコンピュータシステムインタフェース(『iSCSI』)技術を利用し、ストレージシステム306の中のストレージリソース308にアクセスするための機構を含む場合もある。

【0095】

【0153】 図3Bに示されるストレージシステム306は、コンピュータプログラム命令を実行し、ストレージシステム306の中で他の計算タスクを実行する際に有用である場合がある処理リソース312も含む。処理リソース312は、1つ以上の中央演算処理装置(『CPU』)だけではなく、いくつかの特定の目的のためにカスタマイズされる1つ以上の特定用途向け集積回路(『ASIC』)を含んでよい。また、処理リソース312は、1つ以上のデジタルシグナルプロセッサ(『DSP』)、1つ以上のフィールドプログラマブルゲートアレイ(『FPGA』)、1つ以上のシステムオンチップ(『SoC』)、又は他の形の処理リソース312も含んでよい。ストレージシステム306は、以下により詳細に説明されるソフトウェアリソース314の実行をサポートすることを含む

が、これに限定されるものではないさまざまなタスクを実行するためにストレージリソース 312 を利用してよい。

【0096】

【0154】 また、図 3 B に示されるストレージシステム 306 は、ストレージシステム 306 の中で処理リソース 312 によって実行されるとき、多様なタスクを実行してよいソフトウェアリソース 314 も含む。ソフトウェアリソース 314 は、ストレージシステム 306 の中で処理リソース 312 によって実行されるとき、ストレージシステムの中に記憶されるデータの完全性を保つための多様なデータ保護技術を実施する際に有用である、例えばコンピュータプログラム命令の 1 つ以上のモジュールを含んでよい。読者は、係るデータ保護技術が、例えばストレージシステムの中のコンピュータハードウェアで実行中のシステムソフトウェアによって、クラウドサービスプロバイダによって、又は他の方法で実施されてよいことを理解する。係るデータ保護技術は、例えば、もはやアクティブに使用されていないデータを、長期の保持のために別個のストレージデバイス又は別個のストレージシステムに移動させる、データアーカイブ技術、ストレージシステムに記憶されるデータが、設備故障又はストレージシステムとのなんらかの他の形の大惨事の場合にデータ損失を回避するためにコピーされ、別の場所に記憶され得るデータバックアップ技術、ストレージシステムに記憶されるデータが、複数のストレージシステムを介してデータがアクセス可能となるように、別のストレージシステムに複製されるデータ複製技術、ストレージシステムの中のデータの状態が時間内に多様な点で取り込まれるデータスナップショット技術、データ及びデータベースの複製のコピーがそれを通して作成される場合があるデータ及びデータベースクローニング技術、及び他のデータ保護技術を含む場合がある。ストレージシステムの故障は、ストレージシステムに記憶されるデータの損失を生じさせない場合があるので、係るデータ保護技術の使用により、ビジネス継続性及び災害復旧の目的が、満たされ得る。

【0097】

【0155】 また、ソフトウェアリソース 314 は、ソフトウェアデファインドストレージ（『SDS』）を実施する際に有用であるソフトウェアを含んでもよい。係る例では、ソフトウェアリソース 314 は、実行されるとき、基本的なハードウェアとは無関係であるデータストレージの方針をベースにしたプロビジョニング及び管理で有用である、コンピュータプログラム命令の 1 つ以上のモジュールを含んでよい。係るソフトウェアリソース 314 は、ストレージハードウェアを管理するソフトウェアからストレージハードウェアを分離するためにストレージ視覚化を実装する上で有用であってよい。

【0098】

【0156】 またソフトウェアリソース 314 は、ストレージシステム 306 内のストレージリソース 308 に向けられる I/O 動作を容易にし、最適化する際に有用であるソフトウェアを含んでもよい。例えば、ソフトウェアリソース 314 は、例えばデータ圧縮、データ重複排除、及び他等の多様なデータ削減技術を実行実施するソフトウェアモジュールを含んでよい。ソフトウェアリソース 314 は、基本的なストレージリソース 308 のよりよい使用を容易にするために I/O 動作をともにインテリジェントにグループ化するソフトウェアモジュール、ストレージシステムの中から移行するためにデータ移行動作を実行するソフトウェアモジュール、及び他の動作を実行するソフトウェアモジュールも含んでよい。係るソフトウェアリソース 314 は、1 つ以上のソフトウェアコンテナとして、又は多くの他の方法で実施されてよい。

【0099】

【0157】 読者は、図 3 B に示される多様な構成要素が、集中型インフラストラクチャとして 1 つ以上の最適化されたコンピューティングパッケージにグループ化されてよいことを理解する。係る集中型インフラストラクチャは、複数のアプリケーションによって共用され、方針主導のプロセスを使用し、集合的に管理される場合がある、コンピュータ、ストレージリソース及びネットワークリソースのプールを含んでよい。係る集中型インフラストラクチャは、ストレージシステム 306 の確立及び動作と関連付けられた多様な

コストも削減しつつ、ストレージシステム 306 の中に多様な構成要素間の互換性の問題を最小限に抑え得る。係る集中型インフラストラクチャは、集中型インフラストラクチャ参照アーキテクチャで、スタンドアロン機器で、ソフトウェア駆動超集中型手法（例えば、超集中型インフラストラクチャ）で、又は他の方法で実施されてよい。

【0100】

【0158】 読者は、図 3 B に示されるストレージシステム 306 が、多様なタイプのソフトウェアアプリケーションをサポートするために有用であってよいことを理解する。例えば、ストレージシステム 306 は、係るアプリケーションにストレージリソースを提供することによって、人工知能（『A I』）アプリケーション、データベースアプリケーション、DevOps プロジェクト、電子設計自動化ツール、イベント駆動ソフトウェアアプリケーション、高性能コンピューティングアプリケーション、シミュレーションアプリケーション、高速データキャプチャアプリケーション及び解析アプリケーション、機械学習アプリケーション、媒体生成アプリケーション、媒体供給アプリケーション、ピクチャアーカイブ及び通信システム（『P A C S』）アプリケーション、ソフトウェア開発アプリケーション、仮想現実アプリケーション、拡張現実アプリケーション、及び多くの他のタイプのアプリケーションをサポートする上で有用であってよい。

10

【0101】

【0159】 上述されたストレージシステムは、多種多様なアプリケーションをサポートするために動作してよい。ストレージシステムが、計算リソース、ストレージリソース、及び多種多様な他のリソースを含むという事実を考慮して、ストレージシステムは、例えば A I アプリケーション等、資源集約的であるアプリケーションをサポートするためにうまく適合してよい。係る A I アプリケーションは、デバイスがその環境を認識し、なんらかの目標に対するその成功の確率を最大限にする処置を講じることを可能にしてよい。係る A I アプリケーションの例は、I B M W a t s o n、M i c r o s o f t O x f o r d、G o o g l e D e e p M i n d、B a i d u M i n w a、及び他を含む場合がある。また、上述されたストレージシステムは、例えば機械学習アプリケーション等、資源集約的である他のタイプのアプリケーションをサポートするためにうまく適合してよい。機械学習アプリケーションは、解析モデル構築を自動化するために、多様なタイプのデータ解析を実行してよい。機械学習アプリケーションは、データから反復して学習するアルゴリズムを使用し、コンピュータが明示的にプログラミングされることなく学習することを可能にできる。

20

30

【0102】

【0160】 すでに説明されたりソースに加えて、上述されたストレージシステムは、ときおり視覚処理ユニット（『V P U』）とも呼ばれるグラフィックスプロセッシングユニット（『G P U』）も含んでよい。係る G P U は、表示装置への出力のために意図されたフレームバッファ内での画像の作成を加速するためにメモリを迅速に操作し、改変する特殊化された電子回路として実施されてよい。係る G P U は、ストレージシステムの多くの個別にスケーラブルな構成要素の 1 つとして、を含んだ、上述されるストレージシステムの部分であるコンピューティングデバイスのいずれかの中に含まれてよく、係るストレージシステムの個々にスケーラブルな構成要素の他の例は、ストレージ構成要素、メモリ構成要素、計算構成要素（例えば、C P U、F P G A、A S I C）、ネットワーキング構成要素、ソフトウェア構成要素及び他を含む場合がある。G P U に加えて、上述されたストレージシステムは、ニューラルネットワーク処理の多様な態様での使用のためにニューラルネットワークプロセッサ（『N N P』）も含んでよい。係る N N P は、G P U の代わりに（又は加えて）使用されてよく、個別にスケーラブルであってよい。

40

【0103】

【0161】 上述されたように、本明細書に説明されるストレージシステムは、人工知能アプリケーション、機械学習アプリケーション、ビッグデータ解析アプリケーション、及び多くの他のタイプのアプリケーションをサポートするように構成されてよい。これらの種類のアプリケーションの急成長は、3 つの技術、つまりディープラーニング（D L）、G

50

P Uプロセッサ、及びビッグデータによって動かされている。ディープラーニングは、人間の脳によって刺激を与えられる超並列ニューラルネットワークを利用するコンピューティングモデルである。専門家がソフトウェアを手作りする代わりに、ディープラーニングモデルは、多くの例から学習することによって独自のソフトウェアを書き込む。G P Uは、人間の脳の並行した性質を緩やかに表すアルゴリズムを実行するためにうまく適合している数千のコアを有する近代的なプロセッサである。

【0104】

【0162】 ディープニューラルネットワークの進歩は、データサイエンティストが人工知能（A I）を用いてそのデータを利用するためのアルゴリズム及びツールの新しい波に火をつけた。改善されたアルゴリズム、より大きいデータセット、及び（一連のタスク全体での機械学習のためのオープンソースソフトウェアライブラリを含んだ）多様なフレームワークを用いて、データサイエンティストは、自律運転車両、自然言語処理、及び多くの他のもののような新しい使用例に取り組んでいる。しかしながら、ディープニューラルネットワークの訓練は、高品質の入力データ及び大量の計算の両方を必要とする。G P Uは、同時に大量のデータに対して作用できる超並列プロセッサである。マルチG P Uクラスタに結合されるとき、高スループットパイプラインは、ストレージから計算エンジンに入力データを送るように要求されてよい。ディープラーニングは、単にモデルを構築し、訓練するだけ以上である。データサイエンスチームが成功するために必要な規模、反復、及び実験のために設計されなければならないデータパイプライン全体も存在する。

【0105】

【0163】 データは、近代的なA I及びディープラーニングアルゴリズムの核心である。訓練が開始できる前に、対応しなければならない1つの問題は、正確なA Iモデルを訓練するために重要であるラベル付きデータを収集することを中心に展開する。大量のデータを連続的に収集し、整理し、変形し、ラベルを付け、記憶するためには、本格的なA I配備が必要とされる場合がある。追加の高品質のデータポイントを加えることは、直接的により正確なモデル及びより優れた洞察力になる。データサンプルは、1）外部ソースから訓練システムにデータを取り込み、未処理形式でデータを記憶するステップ、2）データサンプルを適切なラベルにリンクすることを含んだ、データを整理し、訓練に便利なフォーマットで変換するステップ、3）パラメータ及びモデルを調査し、より小さいデータセットで迅速に試験し、生産クラスタの中に押し込むために最も有望なモデルに集中することを反復するステップ、4）新しいサンプルとより古いサンプルの両方を含んだ入力データのランダムバッチを選択するために訓練段階を実行し、モデルパラメータを更新するための計算のために生産C P Uサーバにランダムバッチをフィードするステップ、及び5）ホールドアウトデータでのモデル精度を評価するために、訓練で使用されなかったデータのホールドバック（h o l d b a c k）部分を使用することを含む、評価するステップを含むが、これに限定されるものではない一連の処理ステップを受けてよい。このライフサイクルは、単にニューラルネットワーク又はディープラーニングだけではなく、任意のタイプの並列化機械学習のために適用してよい。例えば、標準的な機械学習フレームワークは、G P Uの代わりにC P Uに頼る場合があるが、データ取込み及び訓練の作業フローは同じである場合がある。読者は、単一の共用ストレージデータハブが、取込み段階、前処理段階、及び訓練段階の中で余分なデータコピーを必要とすることなく、ライフサイクルを通して調整ポイントを作成することを理解する。取り込まれたデータが1つの目的にしか使用されないのはまれであり、共用ストレージは、複数の異なるモデルを訓練する、又は従来の解析をデータに適用するために柔軟性を与える。

【0106】

【0164】 読者は、A Iデータパイプラインの各段階がデータハブ（例えば、ストレージシステム又はストレージシステムの集合体）からの変化する要件を有する場合があることを理解する。スケールアウトストレージシステムは一小さいメタデータから重い、大きいファイルへ、ランダムアクセスパターンから連続アクセスパターンへ、及び低同時並列性から高同時並列性へあらゆる種類のアクセスタイプ及びパターンの妥協のない性能を送

達しなければならない。システムは、構造化されていない作業負荷にサービスを提供する場合があるので、上述されたストレージシステムは、理想的なA Iデータハブとしての機能を果たしてよい。第1の段階では、データは、理想的には、過剰なデータコピーを回避するために、取り込まれ、続く段階が使用する同じデータハブ上に記憶される。次の2つのステップは、任意選択でGPUを含む標準的な計算サーバで実行でき、次いで第4の且つ最後の段階で、完全訓練生産ジョブが強力なCPU加速サーバで実行される。多くの場合、同じデータセットで動作中の実験パイプラインに沿った生産パイプラインがある。さらに、GPU加速サーバは、異なるモデルのために独立して使用できる場合もあれば、分散訓練用の複数のシステムにも及ぶ1つのより大きいモデルで訓練するために互いに結合できる場合もある。共用ストレージ層が遅い場合、次いでデータは段階ごとにローカルストレージにコピーされなければならない、異なるサーバにデータを段階分けする時間の浪費を生じさせる。A I訓練パイプライン用の理想的なデータハブは、すべてのパイプライン段階が同時に動作できるようにするための簡潔性及び性能も有しながら、サーバノード上にローカルに記憶されるデータに類似した性能を発揮する。

【0107】

【0165】 データサイエンティストは、多種多様の方式を通して、訓練されたモデルの有用性、つまりより多くのデータ、より良いデータ、よりスマートな訓練、及び深いモデルを改善するために作業する。多くの場合、同じデータセットを共用し、新しい且つ改善された訓練モデルを作り出すために並行して作業するデータサイエンティストのチームがある。しばしば、同じ共用データセットで同時にこれらの段階の中で作業するデータサイエンティストのチームがある。データ処理、実験、及び本格的な訓練の複数の同時作業負荷は複数のアクセスパターンの要求をストレージ層の上に階層化する。言い換えると、ストレージは、大きいファイル読取りを満たすことができるのではなく、大きい及び小さいファイルの読取り及び書込みの混合物に取り組みなければならない。最後に、複数のデータサイエンティストがデータセット及びモデルを探索する状態では、各ユーザが、独自のやり方でデータを変換し、整理し、使用するための柔軟性を提供するためにデータをそのネイティブフォーマットで記憶することは、きわめて重大であり得る。上述されたストレージシステムは、データセットのための自然な共用ストレージホームに（例えば、RAID 6を使用することによる）データ保護冗長性、及び複数の開発者及び複数の実験に対する共通アクセスポイントとなるために必要な性能を提供し得る。上述されたストレージシステムを使用することは、ローカルな作業のためのデータの部分集合を慎重にコピーする必要性を回避し、エンジニアリングとCPU加速サーバ使用時間の両方を節約し得る。これらのコピーは、未処理データセットとして持続し、増え続ける重荷になり、所望される変換はつねに更新し、変化する。

【0108】

【0166】 読者は、ディープラーニングの成功がなぜ急増しているのかの根本的な理由がより大きいデータセットサイズのモデルの継続的な改善であることを理解する。対照的に、ロジスティック回帰のような旧式の機械学習アルゴリズムは、より小さいデータセットサイズで精度を改善するのを停止する。このようにして、計算リソース及びストレージリソースの分離は、各層の独立したスケールリングも可能にし、両方を共に管理することに固有の複雑さの多くを回避し得る。データセットサイズが大きくなる又は新しいデータセットが検討されるにつれ、スケールアウトストレージシステムは容易に拡大できなければならない。同様に、より多くの同時訓練が必要とされる場合、追加のGPU又は他の計算リソースは、その内部ストレージに対して懸念することなく加えることができる。さらに、上述されたストレージシステムは、ストレージシステムによって提供される無作為な読取り帯域幅、（個々のデータポイントを統合して、より大きいストレージフレンドリファイルを作成するために余分な努力が必要とされないことを意味する）小さいファイル（50 KB）を高速で無作為に読み取るストレージシステムの能力、データセットが大きくなる又はスループット要件が増えるにつれ、ストレージシステムが容量及び性能を拡大縮小できる能力、ストレージシステムのファイル又はオブジェクトをサポートする能力、大きい

ファイル又は小さいファイルのために性能を調整するストレージシステムの能力（つまり、ユーザがファイルシステムをセットアップする必要はない）、生産モデル訓練中でもハードウェア及びソフトウェアの非破壊的な更新をサポートするストレージシステムの能力のために、及び多くの他の理由から、A Iシステムを構築し、操作し、成長させることをより容易にしていよい。

【0109】

【0167】 テキスト、音声、又は画像を含んだ多くのタイプの入力、小さいファイルとしてネイティブに記憶されるので、ストレージ層の小さいファイル性能が重要である場合がある。ストレージ層が小さいファイルをうまく処理しない場合、サンプルを前処理し、より大きいファイルにグループ化するために余分なステップが必要とされる。キャッシュ層としてSSDに頼る、回転盤の上部に構築されたストレージは、必要とされる性能に及ばない場合がある。無作為な入力バッチを用いた訓練がより正確なモデルを生じさせるため、データセット全体は完全な性能でアクセス可能でなければならない。SSDキャッシュは、データの小さい部分集合に高性能を提供するにすぎず、回転するドライブの待ち時間を隠すことには無効である。

【0110】

【0168】 読者は、上述されたストレージシステムがブロックチェーンの（データのタイプの中の）ストレージをサポートするように構成されてよいことを理解する。係るブロックチェーンは、暗号法を使用し、リンクされ、保護されるブロックと呼ばれるレコードの連続的に増え続けるリストとして実施されてよい。ブロックチェーンの各ブロックは、以前のブロックに対するリンクとしてのハッシュポインタ、タイムスタンプ、トランザクションデータ等を含んでよい。ブロックチェーンは、データの修正に耐性があるように設計されてよく、2つの関係者間のトランザクションを、効率的に並びに検証可能且つ恒久的な方法で記録できるオープン分散型台帳としての機能を果たすことができる。これは、ブロックチェーンを、イベント、医療記録、及び例えば、アイデンティティ管理、トランザクション処理、その他等の他の記録管理活動の記録に潜在的に適切にする。

【0111】

【0169】 読者は、いくつかの実施形態では、上述されたストレージシステムが、上述されたアプリケーションをサポートするために他のリソースと対にされてよいことをさらに理解する。例えば、1つのインフラストラクチャは、ディープニューラルネットワーク用のパラメータを訓練するために計算エンジンに相互接続されるディープラーニングアプリケーションを加速するために、グラフィックスプロセッシングユニットでの汎用計算（『GPGPU』）を使用することに特化するサーバ及びワークステーションの形での一次計算を含むことがあるであろう。各システムは、イーサネット外部接続性、インフィニバンド外部接続性、なんらかの他の形の外部接続性、又はそのなんらかの組合せを有してよい。係る例では、GPUは、単一の大きい訓練用にグループ化される場合もあれば、複数のモデルを訓練するために独立して使用される場合もある。また、インフラストラクチャは、例えばNFS、S3等の高性能プロトコルを通してデータにアクセスできる、例えばスケールアウトオールフラッシュファイル又はオブジェクトストアを提供するために上述されたストレージシステム等のストレージシステムを含む場合もあるであろう。また、インフラストラクチャは、例えば、冗長性のためにMLAGポートチャンネルでのポートを介してストレージ及び計算に接続された冗長なトポロジックイーサネットスイッチを含む場合もある。また、インフラストラクチャは、データ取込み、前処理、及びモデルデバッグのために、任意選択でGPUを用いて、ホワイトボックスサーバの形の追加計算を含むことがあるであろう。読者は、追加のインフラストラクチャも可能であることを理解する。

【0112】

【0170】 読者は、上述されたシステムが、サーバノードに配備された分散型ダイレクトアタッチドストレージ（DDAS）解決策を含む場合がある他のシステムに比して上述されたアプリケーションにより適している場合があることを理解する。係るDDAS解決策

は、大きく、あまり連続的ではないアクセスの処理のために構築されてよいが、小さいランダムアクセスはあまり処理できない場合がある。読者は、ストレージシステムがより安全で、よりローカルに且つ内部で管理され、特徴集合及び性能でよりロバストであるオンサイト又はインハウスのインフラストラクチャに含まれてよいのでクラウドベースのリソースの活用により好ましい、又はそれ以外の場合、上述されたアプリケーションをサポートするためのプラットフォームの一部としてクラウドベースリソースの活用により好ましいプラットフォームを上述されたアプリケーションに提供するために上述されたストレージシステムが活用されてよいことを理解する。例えば、IBMのWatson等のプラットフォーム上に構築されたサービスは、企業に、例えば金融取引情報又は識別可能な患者レコード等の個人ユーザ情報を他の機関に分散することを要求する場合がある。したがって、サービスとしてのAIのクラウドベースの提供物は、多様な業務上の理由のためだけでなく、多様な技術的な理由から、例えば上述されたストレージシステム等のストレージシステムによってサポートされるサービスとして内部で管理され、提供されるAIよりも望ましくない場合がある。

10

【0113】

【0171】 読者は、上述されたストレージシステムが、単独で又は他のコンピューティング機械と連携してのどちらかで、他のAI関連ツールをサポートするように構成されてよいことを理解する。例えば、ストレージシステムは、ONNX等のツール、又は異なるAIフレームワークで作成されたモデルを転送することを容易にする他のオープンニューラルネットワーク交換フォーマットを利用してよい。同様に、ストレージシステムは、開発者が、ディープラーニングモデルのプロトタイプを作り、構築し、訓練できるようにする、AmazonのGlueonのようなツールをサポートするように構成されてよい。

20

【0114】

【0172】 読者は、上述されたストレージシステムが、エッジソリューションとして配備されてもよいことをさらに理解する。係るエッジソリューションは、データのソースに近い、ネットワークの端縁でデータ処理を実行することによって、クラウドコンピューティングシステムを最適化するために実施されていてよい。エッジコンピューティングは、集中したポイントからネットワークの論理的な端へアプリケーション、データ、及び計算力（つまり、サービス）をプッシュする場合がある。例えば上述されたストレージシステム等のエッジソリューションを使用することで、計算タスクは、係るストレージシステムによって提供される計算リソースを使用し、実行されてよく、データは、ストレージシステムのストレージリソースを使用するストレージであってよく、クラウドベースのサービスは、（ネットワークリソースを含んだ）ストレージシステムの多様なリソースの使用によりアクセスされてよい。エッジソリューションで計算タスクを実行し、エッジソリューションでデータを記憶し、概してエッジソリューションを利用することによって、高価なクラウドベースリソースの消費は回避され得、実際には、性能の改善は、クラウドベースリソースに対してより大きい依存に対して経験されてよい。

30

【0115】

【0173】 多くのタスクは、エッジソリューションの活用から恩恵を受ける場合があるが、いくつかの特定の用途に係る環境での配備に特に適する場合がある。例えば、ドローン、自律走行車、ロボット及び他のようなデバイスは高速処理を必要とする場合がある—実際には、非常に速いので、データ処理サポートを受けるためにデータをクラウド環境に上げて戻すことは、単に遅すぎる場合がある。同様に、豊富なデータ生成センサの使用により大量の情報を生成する機関車及びガスタービンのような機械は、エッジソリューションの高速データ処理機能から恩恵を受ける場合がある。追加の例として、単に関与する純粋なデータ量のため、クラウドにデータを送信することが（プライバシーの観点、セキュリティの観点、又は財政的な観点からだけでなく）実際的ではない場合があるので、例えば接続されたビデオカメラ等のいくつかのIoTデバイスは、クラウドベースのリソースの活用により適切ではない場合がある。したがって、データの処理、ストレージ、又は通信に依存する多くのタスクは、例えば上述されたストレージシステム等のエッジソリューション

40

50

ンを含むプラットフォームによってより適している場合がある。

【0116】

【0174】 倉庫、流通センタ、又は類似した場所での在庫管理の具体例を検討する。大きい在庫業務、貨物保管業務、出荷業務、注文調達業務、製造作業、又は他の業務は、在庫の棚に大量の在庫、及び大きいデータのファイアホース (fire hose) を生じさせる高解像度デジタルカメラを有する。このデータのすべては、小さなデータのファイアホースヘータ量を削減し得る画像処理システムの中に取り込まれてよい。小さいデータのすべては、ストレージにオンプレミスで記憶されてよい。施設の端縁にあるオンプレミスストレージは、外部レポート、リアルタイム制御、及びクラウドストレージのためにクラウドに結合されてよい。在庫管理は画像処理の結果を用いて実行されてよく、これにより在庫は、棚上で追跡され、補充され、移動され、出荷され、新製品で修正される、又は打ち切られる場合がある／陳腐化した製品が削除される等である。上記の状況は、上述された構成可能な処理システム及びストレージシステムの実施形態の主要な候補である。計算専用のブレードと画像処理に適したオフロードブレードの、おそらくオフロード FPG A 又はオフロードカスタムブレード (複数可) でのディープラーニングとの組合せにより、デジタルカメラのすべてから大量のデータのファイアホースを取り込み、小さいデータのファイアホースを作り出すことができるであろう。小さいデータのすべては次いでストレージノードによって記憶され、ストレージブレードのタイプのどの組合せがデータフローを最もうまく処理しようと、ストレージユニットと動作するであろう。これは、ストレージ及び機能の加速及び統合の一例である。クラウドとの外部通信の必要性、及びクラウド内での外部処理に応じて、並びにネットワーク接続及びクラウドリソースの信頼性に応じて、システムは、爆発的な作業負荷及び可変伝導度信頼性を有するストレージ及び計算の管理のためのサイズに設定されるであろう。また、他の在庫管理の態様に依拠して、システムは、ハイブリッドエッジ／クラウド環境でのスケジューリング及びリソース管理のために構成できるであろう。

【0117】

【0175】 また、上述されたストレージシステムは、ビッグデータ解析での使用のために最適化されてもよい。ビッグデータ解析は、概して、組織がより多くの情報に基づいた経営判断を下すのを支援できる隠れたパターン、未知の相関、市場の傾向、カスタマの好み、及び他の有用な情報を明らかにするために大きく且つ変化に富んだデータセットを調べるプロセスとして説明される場合がある。ビッグデータ解析アプリケーションは、データサイエンティスト、予測モデラー、統計学者、及び他の分析学専門家が、多くの場合従来のビジネスインテリジェンス (BI) 及び分析論プログラムによって活用されないまま残されている他の形のデータを加えた、増え続ける大量の構造化されたトランザクションデータを解析できるようにする。そのプロセスの一部として、例えばインターネットクリックストリームデータ、ウェブサーバログ、ソーシャルメディアコンテンツ、カスタマの電子メール及び調査応答からのテキスト、携帯電話呼詳細記録、IoT センサデータ、並びに他のデータ等の半構造化されたデータ、及び構造化されていないデータが、構造化された形に変換されてよい。ビッグデータ分析は、例えば高性能分析システムを装備した予測モデル、統計アルゴリズム、及び what-if 分析等の要素を有する複雑なアプリケーションを必要とする高度な分析学の形である。

【0118】

【0176】 また、上述されたストレージシステムは、人間の発話に応じてタスクを実行するアプリケーションをサポートしてもよい (システムインタフェースとして実装することを含む)。例えば、ストレージシステムは、例えば Amazon の Alexa、Apple Siri、Google Voice、Samsung Bixby、Microsoft Cortana 及び他等の実行インテリジェントパーソナルアシスタントアプリケーションをサポートしてよい。前文に説明された例は、入力として音声を利用するが、上述のストレージシステムは、聴覚方法又はテキスト方法を介して会話を実施するように構成される、チャットボット、トークボット、チャターボット、又は人工会話エンティティ

をサポートしてもよい。同様に、ストレージシステムは、システム管理者等のユーザが発話を介してストレージシステムと対話できるようにするために係るアプリケーションを実際に実行してよい。本開示に係る実施形態では、係るアプリケーションは、多様なシステム管理動作に対するインタフェースとして活用され得るが、係るアプリケーションは、概して音声対話、音楽再生、やることリストの作成、アラームのセット、ポッドキャストのストリーミング、オーディオブックの再生、並びに天気、交通、及び例えばニュース等の他のリアルタイム情報を提供することができる。

【0119】

【0177】 また、上述されたストレージシステムは、自動運転ストレージのビジョンを達成するためのA Iプラットフォームを実装してもよい。係るA Iプラットフォームは、大量のストレージシステムテレメトリデータポイントを収集し、分析して、楽な管理、分析、及びサポートを可能にすることによってグローバルな予測インテリジェンスを実現させるように構成されてよい。実際に、係るストレージシステムは、作業負荷の展開、相互作用、及び最適化に対するインテリジェントなアドバイスを生成するだけでなく、容量と性能の両方も予測できる場合がある。係るA Iプラットフォームは、すべての入信ストレージシステムテレメトリデータを、問題のフィンガープリントのライブラリに対して走査して、インシデントがカスタマの環境に影響を与える前に、リアルタイムでインシデントを予測し、解決し、性能負荷を予測するために使用される性能に係る数百の変数を取り込むように構成されてよい。

【0120】

【0178】 追加の説明のために、図4は、本開示のいくつかの実施形態に従ってポッドをサポートする複数のストレージシステム(402、404、406)を示すブロック図を説明する。あまり詳細には示されないが、図4に示されるストレージシステム(402、404、406)は、図1Aから図1D、図2Aから図2G、図3Aから図3B、又はその任意の組合せに関して上述されたストレージシステムに類似してよい。実際には、図4に示されるストレージシステム(402、404、406)は、上述されたストレージシステムと同じ構成要素、より少ない構成要素、又は追加の構成要素を含んでよい。

【0121】

【0179】 図4に示される例では、ストレージシステム(402、404、406)のそれぞれは、少なくとも1つのコンピュータプロセッサ(408、410、412)、コンピュータメモリ(414、416、418)、及びコンピュータストレージ(420、422、424)を有するとして示されている。いくつかの実施形態では、コンピュータメモリ(414、416、418)及びコンピュータストレージ(420、422、424)は同じハードウェアデバイスの一部であってよいが、他の実施形態では、コンピュータメモリ(414、416、418)及びコンピュータストレージ(420、422、424)は異なるハードウェアデバイスの一部であってよい。この特定の例でのコンピュータメモリ(414、416、418)とコンピュータストレージ(420、422、424)との相違は、コンピュータメモリ(414、416、418)がコンピュータプロセッサ(408、410、412)に物理的に近接しており、コンピュータプロセッサ(408、410、412)によって実行されるコンピュータプログラム命令を記憶してよく、一方コンピュータストレージ(420、422、424)は、ユーザデータ、ユーザデータを記述するメタデータ等を記憶するための不揮発性ストレージとして実施されるという点であってよい。例えば、図1Aの上記の例を参照すると、特定のストレージシステム(402、404、406)のためのコンピュータプロセッサ(408、410、412)及びコンピュータメモリ(414、416、418)は、コントローラ(110A~110D)の1つ以上の中に常駐してよい。一方、アタッチされたストレージデバイス(171Aから171F)は、特定のストレージシステム(402、404、406)の中のコンピュータストレージ(420、422、424)としての機能を果たしてよい。

【0122】

【0180】 図4に示される例では、示されるストレージシステム(402、404、40

6) は、本開示のいくつかの実施形態に従って、1つ以上のポッド(430、432)にアタッチしてよい。図4に示されるポッド(430、432)のそれぞれは、データセット(426、428)を含む場合がある。例えば、3つのストレージシステム(402、404、406)がアタッチしている第1のポッド(430)は、第1のデータセット(426)を含み、2つのストレージシステム(404、406)がアタッチしている第2のポッド(432)は、第2のデータセット(428)を含む。係る例では、特定のストレージシステムがポッドにアタッチするとき、ポッドのデータセットは、特定のストレージシステムにコピーされ、次いでデータセットが修正されるにつれ、最新に保たれる。ストレージシステムは、ポッドから削除され、データセットは削除されたストレージシステムではもはや最新には保たれない結果となる。図4に示される例では、ポッドにとってアクティブである(それは、故障していない(non-faulted)ポッドの最新の動作中の故障していない部材である)任意のストレージシステムが、ポッドのデータセットを修正する又は読み取る要求を受け取り、処理することができる。

【0123】

【0181】 図4に示される例では、各ポッド(430、432)は、特定のポッド(430、432)と関連付けられるデータセット(426、428)を修正又は読み取るためのアクセス動作の集合だけではなく、管理されたオブジェクト及び管理動作の集合を含んでもよい。係る例では、管理動作は、ストレージシステムのいずれかを通して同等に管理オブジェクトを修正する又は問い合わせしてよい。同様に、データセットを読み取る又は修正するためのアクセス動作は、ストレージシステムのいずれかを通して同等に動作してよい。係る例では、各ストレージシステムは、ストレージシステムによる使用のために記憶され、宣伝されるデータセットの適切な部分集合としてデータセットの別個のコピーを記憶するが、任意の1つのストレージシステムを通して実行され、完了された管理オブジェクト又はデータセットを修正するための動作は、ポッドを問い合わせるための後続の管理オブジェクト、又はデータセットを読み取るための後続のアクセス動作に反映される。

【0124】

【0182】 読者は、ポッドが単にクラスタ化された同期複製されたデータセットよりも多くの機能を実装してよいことを理解する。例えば、ポッドは、それによりデータセットがなんらかの方法で互いから安全に隔離されるテナントを実装するために使用できる。また、ポッドは、各ポッドが別々のアドレスを有するネットワーク(例えばストレージエリアネットワーク、又はインターネットプロトコルネットワーク)上で一意のストレージエンティティとして提示される仮想アレイ又は仮想ストレージシステムを実装するために使用できる。仮想ストレージシステムを実装するマルチストレージシステムポッドの場合、ポッドと関連付けられたすべての物理ストレージシステムは、(例えば、あたかも複数の物理ストレージシステムが、単一のストレージシステムの中への複数のネットワークポートとは異ならないかのよう)に)なんらかの方法で同じストレージシステムとして、それら自体を提示してよい。

【0125】

【0183】 読者は、ポッドが、ボリューム、ファイルシステム、オブジェクト／分析ストア、スナップショット、及び他の管理エンティティの集合体を表す、管理の単位である場合もあり、任意の1つのストレージシステムでの管理変更(例えば、名前変更、プロパティ変更、エクスポート又はポッドのデータセットのなんらかの部分に対する許可を管理すること)がポッドと関連付けられたすべてのアクティブなストレージシステムに自動的に反映されることを理解する。さらに、ポッドはデータ収集及びデータ分析の単位でもある場合があり、性能及び容量の測定基準は、ポッドのためにすべてのアクティブストレージシステム全体で統合する、又はポッドごとに別々にデータの収集及び分析に挑戦するように提示される、又はおそらくそれぞれのアタッチされたストレージシステムの入信コンテンツ及びポッドごとの性能に対する貢献を提示する。

【0126】

【0184】 ポッドのメンバーシップの1つのモデルは、ストレージシステムのリスト、及

びストレージシステムがポッドのために同期中であると見なされるそのリストの部分集合として定義され得る。ストレージシステムは、ポッドがポッドと関連付けられたデータセットの最後の読み取られたコピーのための同一のアイドルコンテンツを有する少なくとも回復の中にある場合に、ポッドのために同期中であると見なされてよい。アイドルコンテンツは、新しい修正の処理なく任意の進行中の修正が完了した後のコンテンツである。これは「クラッシュ回復可能な」一貫性と呼ばれる場合がある。ポッドの回復は、ポッド内で同期したストレージシステムに同時更新を適用する際の差異を調整するプロセスを実施する。回復は、ポッドの多様なメンバーに対して要求されたが、無事に完了したとしていづれの要求者にも信号で知らされなかった同時修正の完了でストレージシステム間のあらゆる不一致を解決する場合がある。ポッドメンバーとして一覧されるが、ポッドのために同期中として一覧されていないストレージシステムは、ポッドから「デタッチされている」として記述できる。ポッドメンバーとして一覧されるストレージシステムは、ポッドのために同期中であり、ポッドのために「オンラインである」ポッドのためのデータをアクティブに供給するために現在利用可能である。

【0127】

【0185】 ポッドの各ストレージシステムメンバーは、それが最後に知っていたどのストレージシステムが同期中であつたのか、及びそれが最後に知っていたどのストレージシステムが、ポッドメンバーの完全な集合を含んでいたのかを含む、メンバーシップの独自のコピーを有してよい。ポッドのためにオンラインであるためには、ストレージシステムは、それ自体をポッドのために同期中と見なさなければならず、それがポッドのために同期中と見なす他のすべてのストレージシステムと通信していなければならない。ストレージシステムが、それが同期中であり、同期中の他のすべてのストレージシステムと通信していることを確信できない場合、次いでストレージシステムは、それが同期中であり、同期中の他のすべてのストレージシステムと通信していることが確信できるまで、ポッドのために新しい入信要求を処理することを停止しなければならない（又は、エラー若しくは例外とともにそれらを完了しなければならない）。第1のストレージシステムは、第2の対にされたストレージシステムがデタッチされるべきであると断定する場合があります、このことにより、第1のストレージシステムは、それが現在リスト中のすべてのシステムと現在同期しているので続行できるようになる。しかし、代わりに、第2のストレージシステムは、第2のストレージシステムが動作を続行している状態で、第1のストレージシステムがデタッチされるべきであると断定することを妨げられなければならない。これは、他の危険の中でも共存できないデータセット、データセット破損、又はアプリケーション破損につながる場合がある「スプリットブレイン」状態を生じさせるであろう。

【0128】

【0186】 対にされたストレージシステムと通信していないときにどのように進むのかを判断する必要がある状況は、ストレージシステムが正常に実行しており、次いで通信が失われたことに気づく間に、それが現在なんらかの以前の障害から回復している間に、それがリブートしている又は一時的な電力損失又は回復された通信停止状態から再開している間に、それが理由のいかに関わらずストレージシステムコントローラのあるセットから別のセットに動作を切り替えている間に、又はこれらの若しくは他の種類のイベントの任意の組合せの間あるいは後に発生する場合がある。実際には、ポッドと関連付けられるストレージシステムがすべての既知のデタッチされていないメンバーと通信できないときはいつでも、ストレージシステムは通信が確立できるまで短時間待機し、オフラインになり、待機を続行する、又はストレージシステムは、通信していないストレージシステムが別の見方を推断することによりスプリットブレインを生じさせるリスクなく、なんらかの手段によって、通信していないストレージシステムをデタッチすることが安全であると判断し、次いで続行する場合があるかのどちらかである。安全なデタッチが十分に迅速に起こる場合、ストレージシステムは、短い遅延以上もほとんどなく、及び残っているオンラインストレージシステムに対する要求を発行する場合があるアプリケーションに対してアプリケーション停止状態を生じさせることなく、ポッドのためにオンラインにとどまることが

できる。

【0129】

【0187】 この状況の一例は、ストレージシステムが、それが古いことを知っている場合があるときである。それは、例えば、第1のストレージシステムが、1つ以上のストレージシステムとすでに関連付けられているポッドに初めて加えられるとき、又は第1のストレージシステムが別のストレージシステムに再接続し、他のストレージシステムがすでに第1のストレージシステムをデタッチ済みとしてマークしていたことを見つけるときに起こる場合がある。この場合、第1のストレージシステムは、単に、それがポッドのために同期中であるストレージシステムのなんらかの他の集合に接続するまで待機するにすぎない。

10

【0130】

【0188】 このモデルは、ストレージシステムがどのようにしてポッドに又は同期中のポッドメンバーリストに加えられる、又はポッドから又は同期中のポッドメンバーリストから削除されるのかについてある程度の検討を要求する。各ストレージシステムはリストの独自のコピーを有するので、及び2つの独立したストレージシステムはまったく同時にそのローカルコピーを更新できないので、及びローカルコピーは、リブートで又は多様な障害の状況で利用できるすべてであるので、一過性の不一致は問題を起こさないことを確実にするために注意しなければならない。例えば、あるストレージシステムがポッドのために同期しており、第2のストレージシステムが加えられる場合、次いで第2のストレージシステムが両方のストレージシステムを最初に同期中と示すために更新されると、次いで障害及び両方のストレージシステムの再起動がある場合、第2は起動し、第1のストレージシステムに接続するために待機する可能性がある。一方、第1は、それが第2のストレージシステムを待機するべきである又は待機するであろうことに気づいていない可能性がある。第2のストレージシステムが次いで第1のストレージシステムと接続できないことに、プロセスを通過してそれをデタッチすることによって対応する場合、次いで、それは第1のストレージシステムが気づいていないプロセスを完了することに成功し、スプリットブレインを生じさせる可能性がある。したがって、ストレージシステムが、それらが通信していない場合に、それらがデタッチプロセスを経ることを選ぶ可能性があるかどうかに関して、不適切に同意しないことがないことを確実にすることが必要である場合がある。

20

【0131】

【0189】 ストレージシステムが、それらが通信していない場合に、それらがデタッチプロセスを経ることを選ぶ可能性があるかどうかに関して不適切に同意しないことがないことを確実にする1つの方法は、ポッドのための同期メンバーリストに新しいストレージシステムを加えるとき、新しいストレージシステムが最初に、それがデタッチされたメンバーであること（及びおそらく、それが同期中のメンバーとして加えられていること）を記憶することを確実にすることである。次いで、既存の同期中のストレージシステムは、新しいストレージシステムがその同じ事実をローカルに記憶する前に、新しいストレージシステムが同期中のポッドメンバーであることをローカルに記憶できる。新しいストレージシステムがその同期ステータスを記憶する前に、リブート又はネットワーク停止状態の集合がある場合、次いで元のストレージシステムは、通信がないために新しいストレージシステムをデタッチしてよいが、新しいストレージシステムは待機する。通信しているストレージシステムをポッドから削除するために、この変更の逆のバージョンが必要とされる可能性がある。最初に、削除中のストレージシステムは、それがもはや同期していないことを記憶し、次いで残るストレージシステムは、削除中のストレージシステムがもはや同期していないことを記憶し、次いですべてのストレージシステムが、そのポッドメンバーシップリストから、削除中のストレージシステムを削除する。実施態様によっては、中間の持続デタッチ状態は必要ではない場合がある。メンバーシップリストのローカルコピーで注意が必要とされるかどうかは、互いをモニタするために、又はそのメンバーシップを確認するためにモデルストレージシステムに依存する場合がある。コンセンサスモデルが両方に使用される場合、又は外部システム（若しくは外部の分散システムあるいはクラス

30

40

50

タ化されたシステム)が、ポッドメンバーシップを記憶し、確証するために使用される場合、次いでローカルに記憶されるメンバーシップリストの不一致は問題にならない場合がある。

【0132】

【0190】 通信が失敗する、又はポッドの1つ若しくは複数のストレージシステムが故障するとき、あるいはストレージシステムが起動し（又は二次コントローラに障害迂回し）、ポッドのために対にされたストレージシステムと通信できず、1つ以上のストレージシステムが1つ以上の対にされたストレージシステムをデタッチすると決定するとき、なんらかのアルゴリズム又は機構が、そうすることが安全であると決定し、デタッチに関してやり通すために利用されなければならない。デタッチを解決する1つの手段は、メンバーシップに過半数（又は定足数）モデルを使用することである。3つのストレージシステムの場合、2つが通信している限り、それらは通信していない第3のストレージシステムをデタッチすることに同意できるが、その第3のストレージシステムは、独力で他の2つのどちらかをデタッチすることを選ぶことはできない。ストレージシステム通信に一貫性がないとき、混乱状態が発生する場合がある。例えば、ストレージシステムAは、ストレージシステムBと通信している可能性があるが、Cとは通信していない。一方、ストレージシステムBはAとCの両方と通信している可能性がある。したがって、A及びBはCをデタッチする、又はB及びCはAをデタッチするであろうが、これを理解するためには、ポッドメンバー間のより多くの通信が必要とされる場合がある。

【0133】

【0191】 ストレージシステムを加え、削除するとき、定足数メンバーシップモデルでは注意する必要がある。例えば、第4のストレージシステムが加えられる場合、次いでストレージシステムの「過半数」はその時点では3つである。（過半数には2つが必要とされる状態での）3つのストレージシステムから（過半数には3つが必要とされる状態での）第4のストレージシステムを含んだポッドへの遷移は、同期リストにストレージシステムを注意深く加えるために上述されたモデルに類似した何かを必要とする場合がある。例えば、第4のストレージシステムはアタッチする状態で開始するが、まだアタッチしておらず、それは定足数を超える票を絶対に引き起こさないであろう。いったんその状態になると、元の3つのポッドメンバーはそれぞれ更新されて、第4のメンバー及び第4のメンバーをデタッチするための3つのストレージシステムの過半数に対する新しい要件を認識する。ストレージシステムをポッドから削除することは、同様に、他のポッドメンバーを更新する前にそのストレージシステムをローカルに記憶された「デタッチ」状態に動かす可能性がある。この異形の方式は、例えばPAXOS又はRAFT等の分散型コンセンサス機構を使用して、任意のメンバーシップの変更を実施する又はデタッチ要求を処理することである。

【0134】

【0192】 メンバーシップ遷移を管理する別の手段は、ポッドメンバーシップを処理するためにストレージシステム自体の外部にある外部システムを使用することである。ポッドのためにオンラインになるために、ストレージシステムは最初に外部ポッドメンバーシップシステムに連絡して、それがポッドのために同期していることを検証しなければならない。ポッドのためにオンラインである任意のストレージシステムは、次いでポッドメンバーシップシステムと通信したままとなるべきであり、ストレージシステムが通信を失う場合には待機する又はオフラインになるべきである。外部ポッドメンバーシップマネージャは、例えばOracle RAC、Linux HA、VERITAS Cluster Server、IBMのHACMP、又は他等の多様なクラスタツールを使用し、きわめて利用可能なクラスタとして実装できるであろう。また、外部ポッドメンバーシップマネージャは、Etcd又はZookeeper等の分散型構成ツール、又はAmazonのDynamoDB等の信頼できる分散型データベースを使用できるであろう。

【0135】

【0193】 図4に示される例では、本開示のいくつかの実施形態に従って、示されるスト

レンジシステム（４０２、４０４、４０６）は、データセット（４２６、４２８）の一部分を読み取る要求を受け取り、データセットの一部分をローカルに読み取る要求を処理してよい。読者は、データセット（４２６、４２８）は、ポッドのすべてのストレージシステム（４０２、４０４、４０６）全体で一貫性があるべきであるので、データセット（４２６、４２８）を修正する要求（例えば、書込み動作）は、ポッドのストレージシステム（４０２、４０４、４０６）間で調整を必要とするが、データセット（４２６、４２８）の一部分を読み取る要求に応えることはストレージシステム（４０２、４０４、４０６）間で類似する調整を必要としないことを理解する。したがって、読取り要求を受け取る特定のストレージシステムは、ポッド内の他のストレージシステムとの同期通信がない状態で、ストレージシステムのストレージデバイスの中に記憶されるデータセット（４２６、４２８）の一部分を読み取ることによってローカルに読取り要求にサービスを提供してよい。複製されたクラスタでの複製されたデータセットのために１つのストレージシステムによって受け取られた読取り要求は、少なくとも、やはり名目上実行中であるクラスタの中で実行中のストレージシステムによって受け取られるとき、大多数のケースでいかなる通信も回避することを期待される。係る読取りは、通常、クラスタ内の他のストレージシステムとの追加の対話が必要とされない状態で、クラスタ化されたデータのローカルコピーから単に読み取ることによって正常に処理されるべきである。

【０１３６】

【０１９４】 読者は、どのストレージシステムが読取り要求を処理するのかに関わりなく、読取り要求が同じ結果を返すように、ストレージシステムが読取り一貫性を確実にするために対策を講じてよいことを理解する。例えば、クラスタ内のストレージシステムの任意の集合によって受け取られる更新の任意の集合に対する結果として生じるクラスタ化されたデータセットコンテンツは、少なくとも更新がアイドルである（すべての以前の修正動作が完了と示され、新しい更新要求が決して受け取られず、処理されない）ときはいつでも、クラスタ全体で一貫性があるべきである。より詳細には、ストレージシステムの集合全体でのクラスタ化されたデータセットのインスタンスは、まだ完了されていない更新の結果としてのみ異なる場合がある。つまり、例えば、そのボリュームブロック範囲で重複する任意の２つの書込み要求、又は書込み要求と重複するスナップショット、コンペアアンドライト（compare-and-write）、又は仮想ブロック範囲コピーの任意の組合せは、データセットのすべてのコピーで一貫した結果を生じさせなければならない。２つの動作は、あたかもそれらが１つのストレージシステムで１つの順序で、及び複製されたクラスタの別のストレージシステムで異なる順序で起こったかのように結果を生じさせるべきではない。

【０１３７】

【０１９５】 さらに、読取り要求は、時間順序一貫にされる場合がある。例えば、１つの読取り要求が複製クラスタで受け取られ、完了され、その読取りの後に次いで重複するアドレス範囲に対する別の読取り要求が続き、一方又は両方の読取りが、形はどうあれ複製クラスタによって受け取られる修正要求と、時間及びボリュームのアドレス範囲で重複する場合（読取り又は修正のいずれかが、同じストレージシステムによって受け取られるのか、それとも複製クラスタ内の異なるストレージシステムによって受け取られるのか）次いで第１の読取りが更新の結果を反映する場合、次いで第２の読取りも、更新に先行したデータをおそらく返すよりむしろ、その更新の結果を反映すべきである。第１の読取りが更新を反映しない場合、次いで第２の読取りは更新を反映する、又は反映しないのかのどちらかである。これは、２つの読取り要求の間で、データセグメントのための「時間」が後方に進むことができないことを確実にする。

【０１３８】

【０１９６】 図４に示される例では、示されるストレージシステム（４０２、４０４、４０６）は、他のストレージシステムの１つ以上とのデータ通信の中断を検出し、特定のストレージシステムがポッドに留まるべきであるかどうかを判断してもよい。他のストレージシステムの１つ以上とのデータ通信の中断は、さまざまな理由から発生する場合がある。

例えば、他のストレージシステムの1つ以上とのデータ通信の中断は、ストレージシステムの1つが故障したため、ネットワーク相互接続が故障したため、又はなんらかの他の理由から発生する場合がある。同期した複製クラスタの重要な態様は、任意の障害処理が、回復不可能な不一致又は応答の任意の不一致を生じさないことを確実にすることである。例えば、ネットワークが2つのストレージシステム間で故障する場合、ストレージシステムのうちの最大で1つがポッドに対する新規に入信するI/O要求を処理し続けることができる。そして、一方のストレージシステムが処理を続行する場合、他方のストレージシステムは、読取り要求を含んだ任意の新しい要求を完了まで処理できない。

【0139】

【0197】 図4に示される例では、示されたストレージシステム(402、404、406)は、特定のストレージシステムが、他のストレージシステムの1つ以上とのデータ通信の中断を検出することに応じて、ポッドに留まるべきであるかどうかを判断してもよい。上述されたように、ポッドの一部として『オンライン』であるためには、ストレージシステムは、それ自体をポッドのために同期中であると思わなければならない、それがポッドのために同期中であると思わす他のすべてのストレージシステムと通信していなければならない。ストレージシステムが、それが同期中であり、同期中である他のすべてのストレージシステムと通信していると確信をもてない場合、次いでストレージシステムは新しい入信要求を処理するのを停止して、データセット(426、428)にアクセスする。したがって、ストレージシステムは、例えば、ストレージシステムが、(1つ以上のテストメッセージを介して)、それがポッドのために同期中であると思わす他のすべてのストレージシステムと通信できるかどうかを判断することによって、それがポッドのために同期中であると思わす他のすべてのストレージシステムも、ストレージシステムがポッドにアタッチされると見なすかどうかを判断することによって、特定のストレージシステムが、それがポッドのために同期中であると思わす他のすべてのストレージシステムと特定のストレージシステムが通信できること、及びそれがポッドのために同期中であると思わす他のすべてのストレージシステムも、ストレージシステムをポッドにアタッチされると見なすことを確認しなければならない両方のステップの組合せを通して、又は他のなんらかの機構を通して、特定のストレージシステムが、ポッドの一部としてオンラインに留まるべきであるかどうかを判断してよい。

【0140】

【0198】 図4に示される例では、示されるストレージシステム(402、404、406)は、特定のストレージシステムがポッド内にとどまるべきであると判断したことに応じて、管理動作及びデータセット動作のために特定のストレージシステム上のデータセットにアクセス可能に保ってよい。ストレージシステムは、例えば、ストレージシステム上に記憶されるデータセット(426、428)のバージョンにアクセスする要求を受け入れ、係る要求を処理することによって、ホスト又は権限を与えられた管理者によって発行されるデータセット(426、428)と関連付けられた管理動作を受け入れ、処理することによって、他のストレージシステムの1つによって発行されるデータセット(426、428)と関連付けられる管理動作を受け入れ、処理することによって、又はなんらかの他の方法で、特定のストレージシステム上のデータセット(426、428)を管理動作及びデータセット動作のためにアクセス可能に保ってよい。

【0141】

【0199】 しかしながら、図4に示される例では、示されるストレージシステム(402、404、406)は、特定のストレージシステムがポッド内に留まるべきではないと判断したことに応じて、特定のストレージシステム上のデータセットを管理動作及びデータセット動作のためにアクセス不可にしてよい。ストレージシステムは、例えばストレージシステムに記憶されるデータセット(426、428)のバージョンにアクセスする要求を拒絶することによって、ホスト又は他の権限を与えられた管理者によって発行されるデータセット(426、428)と関連付けられた管理動作を拒絶することによって、ポッド内の他のストレージシステムの1つによって発行されるデータセット(426、428

）と関連付けられた管理動作を拒絶することによって、又はなんらかの他の方法で、特定のストレージシステム上のデータセット（４２６、４２８）を管理動作及びデータセット動作のためにアクセス不可にしてよい。

【０１４２】

【０２００】 図４に示される例では、示されるストレージシステム（４０２、４０４、４０６）は、他のストレージシステムの１つ以上とのデータ通信の中断は修復されたことを検出し、特定のストレージシステム上のデータセットを管理動作及びデータセット動作のためにアクセス可能にしてもよい。ストレージシステムは、例えば他のストレージシステムの１つ以上からメッセージを受け取ることによって他のストレージシステムの１つ以上とのデータ通信の中断が修復されたことを検出してよい。他のストレージシステムの１つ以上とのデータ通信の中断が修復されたことを検出することに応じて、以前にデタッチされたストレージシステムが、ポッドにアタッチされたままであったストレージシステムと再同期されると、ストレージシステムは、特定のストレージシステム上のデータセット（４２６、４２８）を管理動作及びデータセット動作のためにアクセス可能にしてもよい。

【０１４３】

【０２０１】 図４に示される例では、示されるストレージシステム（４０２、４０４、４０６）は、特定のストレージシステムがもはや管理動作及びデータセット動作を可能にしないようにポッドからオフラインになってよい。示されているストレージシステム（４０２、４０４、４０６）は、特定のストレージシステムがさまざまな理由から管理動作及びデータセット動作をもはや可能にしないように、ポッドからオフラインになってよい。例えば、示されているストレージシステム（４０２、４０４、４０６）は、ストレージシステム自体でのなんらかの障害のため、更新又は他のなんらかの保守がストレージシステムで発生しているため、通信障害のため、又は多くの他の理由からポッドからオフラインになってよい。係る例では、以下に含まれる再同期の項でより詳細に説明されるように、特定のストレージシステムはオフラインになり、特定のストレージシステムが管理動作及びデータセット動作を可能にするように、ポッドでオンラインに戻ったので、示されるストレージシステム（４０２、４０４、４０６）は、データセットに対するすべての更新を含むように特定のストレージシステム上のデータセットをその後更新してよい。

【０１４４】

【０２０２】 図４に示される例では、示されるストレージシステム（４０２、４０４、４０６）は、データセットを非同期で受け取るためにターゲットストレージシステムを識別してもよく、ターゲットストレージシステムは、データセットが同期して複製される複数のストレージシステムの１つではない。係るターゲットストレージシステムは、例えば、同期して複製されたデータセット等を利用するなんらかのストレージシステムとして、バックアップストレージシステムを表してよい。実際には、同期複製は、より良いローカル読取り性能のために、データセットのコピーをサーバのなんらかのラックのより近くに分散するために利用される場合がある。１つの係る場合は、データセンタ又はキャンパスで中心に位置するより大きいストレージシステムに対称で複製されるより小さいトップオブラックストレージシステムであり、それらのより大きいストレージシステムは、信頼性のためにより慎重に管理される、又は非同期複製又はバックアップサービスのために外部ネットワークに接続される。

【０１４５】

【０２０３】 図４に示される例では、示されているストレージシステム（４０２、４０４、４０６）は、他のストレージシステムのいずれかによってターゲットストレージシステムに非同期複製されていないデータセットの一部分を識別し、ターゲットストレージシステムに、他のストレージシステムのいずれかによってターゲットストレージシステムに非同期複製されていないデータセットの一部分を非同期複製してもよく、２つ以上のストレージシステムは、集散的にデータセット全体をターゲットストレージシステムに複製する。係るようにして、特定のデータセットを非同期複製することに関連付けられた作業は、ポッドのメンバーの間で分割されてよく、これによりポッドの各ストレージシステムは、タ

ターゲットストレージシステムにデータセットの部分集合を非同期複製することを担うにすぎない。

【0146】

【0204】 図4に示される例では、示されているストレージシステム(402、404、406)はポッドからデタッチしてもよく、これによりポッドからデタッチする特定のストレージシステムは、データセットがその全体で同期複製されるストレージシステムの集合には含まれなくなる。例えば、図4のストレージシステム(404)が図4に示されるポッド(430)からデタッチした場合、ポッド(430)は、ポッド(430)に含まれるデータセット(426)が全体で同期複製されるであろうストレージシステムとして、ストレージシステム(402、406)を含むにすぎないであろう。係る例では、ストレージシステムをポッドからデタッチすることは、ポッドからデタッチした特定のストレージシステムからデータセットを削除することも含む場合があるであろう。図4のストレージシステム(404)が図4に示されるポッド(430)からデタッチした例を続行すると、ポッド(430)に含まれるデータセット(426)は、削除される、又はそれ以外の場合ストレージシステム(404)から除去される場合があるであろう。

【0147】

【0205】 読者は、さらにサポートできるポッドモデルによって可能にされるいくつかの特有の管理機能があることを理解する。また、ポッドモデル自体が、実施態様により対応される場合があるいくつかの問題を生じさせる。例えば、相互接続が失敗し、ポッドのための別のストレージシステムが仲介を勝ち取ったために、ストレージシステムがポッドのためにオフラインであるが、それ以外の場合実行中であるとき、オフラインストレージシステム上のオフラインのポッドのデータセットにアクセスする願望又は必要性がある場合がある。1つの解決策は、単になんらかのデタッチされたモードでポッドを有効化し、データセットにアクセスできるようにすることである場合がある。しかしながら、その解決策は危険であり、その解決策は、ストレージシステムが通信を取り戻すときに、ポッドのメタデータ及びデータを調整するのをはるかに困難にさせる場合がある。さらに、ホストが、まだオンラインのストレージシステムだけではなくオフラインのストレージシステムにアクセスするための別個の経路もあるであろう。その場合、ホストは、両方のストレージシステムとももはや同期中に保たれていなくても両方のストレージシステムにI/Oを発行する可能性がある。というのは、ホストは、同じ識別子を有するボリュームを報告するターゲットポートを見ており、ホストI/Oドライバは、それが同じボリュームに対する追加の経路を見ると推測するためである。例えばホストがありのまま推測したとしても、両方のストレージシステムに発行された読取り及び書き込みがもはや一致していないので、これは、かなり不利なデータ破損を生じさせる場合がある。この場合の異形として、例えば共用ストレージクラスタ化データベース等のクラスタ化されたアプリケーションでは、あるホスト上で実行しているクラスタ化アプリケーションは、あるストレージシステムを読み取っている又はあるストレージシステムに書き込んでいる可能性があり、別のホストで実行中の同じクラスタ化アプリケーションは、「デタッチされた」ストレージシステムを読み取る又は該ストレージシステムに書き込んでいる可能性があるが、クラスタ化されたアプリケーションの2つのインスタンスは、それらがそれぞれ見ているデータセットが、完了された書き込みに対して完全に一致していると仮定して、互いの間で通信している。それらは一致していないので、その仮定は破られ、アプリケーションのデータセット(例えば、データベース)はすぐに破損されることになる可能性がある。

【0148】

【0206】 これらの問題の両方ともを解決する1つの方法は、オフラインポッド、つまりおそらくオフラインポッドのスナップショットが、ホストI/Oドライバ及びクラスタ化されたアプリケーションが、別のストレージシステムでまだオンラインボリュームと同じであると、コピーされたボリュームを混同しない旨の十分に新しいアイデンティティを有する新しいボリュームで新しいポッドにコピーされることを可能にすることである。各ポッドはデータセットの完全なコピーを維持し、これはクラッシュ一貫性(c r a s h c

onsistent)があるが、おそらく別のストレージシステム上のポッドデータセットのコピーとはわずかに異なるデータセットの完全なコピーを維持するので、並びに各ポッドはポッドコンテンツで動作するために必要とされるすべてのデータ及びメタデータの独立したコピーを有するので、新しいポッド内の新しいボリュームに対して、ポッドの一部の又はすべてのボリューム又はスナップショットの仮想コピーを作るとはわかりやすい問題である。例えば、論理エクステントグラフの実施態様では、必要とされるのは、ポッドのボリューム又はスナップショットと、及び書き込み時にコピーとしてマークされている論理エクステントグラフと関連付けられたコピーされたポッドから論理エクステントグラフを参照する、新しいポッドで新しいボリュームを定義することである。新しいボリュームにコピーされるボリュームスナップショットがどのようにして実装される可能性があるのかと同様に、新しいボリュームは新しいボリュームとして処理されるべきである。ボリュームは、新しいポッド名前空間の中ではあるが、同じ管理名を有してよい。しかし、ボリュームは、異なる基本的な識別子、及び元のボリュームとは異なる論理ユニット識別子を有するべきである。

【0149】

【0207】 一部の場合では、いくつかのインタフェースに提示されるボリュームの分離が、元のボリュームを見る可能性もあるホストネットワークインタフェース又はホストSCSI開始ポートからアクセス不可であることが確定できるように、（例えば、IPネットワークの場合仮想LANを、又はファイバチャネルネットワークの場合仮想SANを作成することによって）仮想ネットワーク分離技術を使用することが可能な場合がある。係る場合、ボリュームのコピーに、元のボリュームとして同じSCSI又は他のストレージ識別子を提供することが安全である場合がある。これは、例えばアプリケーションが、再構成での不当な負担なしに機能するために、ストレージ識別子の特定の集合を見ることを期待する場合に使用できるであろう。

【0150】

【0208】 また、本明細書に説明される技術のいくつかは、障害を処理するための準備を試験するためにアクティブな障害状況の外で使用されるであろう。準備試験（「火災訓練」と呼ばれることもある）は、一般的に災害復旧構成に必要とされ、頻繁且つ繰り返される試験は、災害復旧計画の大部分又はすべての態様が正しく、アプリケーション、データセットに対する任意の最近の変化、又は設備の変化を説明することを確実にするために必要と見なされる。準備試験は、複製を含んだ現在の生産動作に対して無停止であるべきである。多くの場合、実際の動作はアクティブな構成で実際に呼び出すことはできないが、近づく良い方法は、生産データセットのコピーを作るためにストレージ動作を使用し、次いでおそらくそれを仮想ネットワーキングの使用と結合し、災害時に無事に起動しなければならない重要なアプリケーションにとって必要と考えられるすべてのデータを含んだ分離された環境を作り出すことである。同期複製された（又は非同期でも複製された）データセットの係るコピーを、災害復旧準備試験手順を実行し、次いで重要なアプリケーションをそのデータセットで起動して、それが起動し、機能することを確実にすることを期待されるサイト（又はサイトの集合体）の中で利用可能にすることは、それが、災害復旧計画でアプリケーションデータセットの重要な部分が除外されなかったことを確実にするのに役立つので素晴らしいツールである。必要な場合、及び実際的な場合、これは、現実世界の災害復旧引継ぎ状況に可能な限り近づくために、物理マシン又はバーチャルマシンの分離された集合体とおそらく結合された仮想分離ネットワークと結合できるであろう。ポッド（又はポッドの集合）を別のポッドに、ポッドデータセットのある時点の画像としてバーチャルにコピーすることは、元のポッドとは別に単一のサイト（又はいくつかのサイト）への分離を可能にするだけではなく、コピーされたすべての要素を含み、元のポッドに基本的に同様に操作できる分離されたデータセットを直ちに作り出す。さらに、これらは高速動作であり、それらは解体し、所望されるほど頻繁に試験を容易に繰り返すことができる。

【0151】

【0209】 完全な災害復旧試験に向けてさらに近づくためにいくつかの強化を行うことができるであろう。例えば、分離されたネットワークと併せて、SCSI論理ユニットアイデンティティ又は他のタイプのアイデンティティは、試験サーバ、バーチャルマシン、及びアプリケーションが同じ識別子を見るようにターゲットポッドにコピーされるであろう。さらに、サーバの管理環境は、仮想ネットワークの特定の仮想集合からの要求に応じて、スクリプトがオブジェクト名の代替「試験」バージョンとともに試験変形の使用を必要としないように、元のポッド名に対する要求又は動作に応えるように構成されるであろう。追加の強化は、災害引継ぎの場合に引き継ぐホスト側サーバインフラストラクチャが、試験中使用できる場合に使用できる。これは、災害復旧データセンタが、一般に、災害によりそのようにするように命令されるまで使用されない代替のサーバインフラストラクチャで完全に備えられる場合を含む。また、インフラストラクチャが重要ではない動作（例えば、生産データに対して分析を実行すること、又は重要である場合があるが、より重要な機能のために必要とされる場合、休止できるアプリケーション開発機能又は他の機能を単にサポートすること）に使用される可能性がある場合も含む。具体的には、ホストの定義及び構成、並びにそれらを使用するサーバインフラストラクチャは、それらが実際の災害復旧引継ぎイベントのためであり、災害復旧引継ぎ試験の一部として試験され、試験されたボリュームが、データセットのスナップショットを提供するために使用される仮想ポッドコピーからこれらのホスト定義に接続されているので、セットアップできる。関与するストレージシステムの観点から、次いで試験に使用されるこれらのホスト定義及び構成、並びに試験中に使用されるボリュームからホストへの接続構成は、実際の災害引継ぎイベントがトリガされるとき再利用することができ、試験構成と、災害復旧引継ぎの場合に使用される実際の構成との間の構成の相違を大幅に最小限に抑える。

【0152】

【0210】 一部の 경우에는、第1のポッドから及びそれらのボリュームだけを含んだ新しい第2のポッドへボリュームを移動させることが意味をなす場合がある。ポッドメンバーシップ及び高可用性及び復旧特徴は、次いで別々に調整することができ、2つの結果として生じるポッドデータセットの管理は、次いで互いから分離できる。また、一方の方向で行うことができる動作は、他方の方向でも可能であるべきである。ある時点で、元の2つのポッドのそれぞれのボリュームがストレージシステムメンバーシップ及び高可用性並びに復旧特徴及びイベントのためにここで互いを追跡するように、2つのポッドを採取し、それらを1つにマージすることが意味をなす場合がある。両方の動作とも、上述の項で説明された、ポッドのための仲介又は定足数特性を変更するために提案された特徴に頼ることによって、安全に且つ実行中のアプリケーションに対して妥当に最小の中断又は中断なく達成できる。例えば、仲介により、ポッドのメディエータは、ポッドの各ストレージシステムが、第1のメディエータと第2のメディエータの両方に依存するために変更され、それぞれが次いで第2のメディエータだけに依存するために変更されるステップから成るシーケンスを使用し、変更できる。シーケンスの途中で障害が発生する場合、いくつかのストレージシステムは、第1のメディエータと第2のメディエータの両方に依存する場合があるが、復旧及び障害処理により、いくつかのストレージシステムが第1のメディエータだけに依存し、他のストレージシステムが第2のメディエータだけに依存することに決してならない。定足数は、復旧に進むために、第1の定足数モデルと第2の定足数モデルの両方に勝つことに一時的に依存することによって同様に処理できる。これは、障害に直面したポッドの可用性が追加のリソースに依存し、したがって潜在的な可用性を削減する非常に短い期間を生じさせる場合があるが、この期間は非常に短く、可用性の削減は多くの場合、ほとんどない。仲介により、メディエータパラメータの変化が、仲介に使用される鍵の変化にすぎず、使用される仲介サービスが同じである場合、それはいま同じサービスに対する2つの呼出し対そのサービスに対する1つの呼出しに依存し、及びむしろ別々の呼出しよりむしろ2つの別々のサービスにだけ依存するので、次いで可用性の潜在的な削減もより少ない。

【0153】

【0211】 読者は、定足数モデルを変更することがきわめて複雑な場合があることに留意する。第2の定足数モデルにも依存するステップが次いで後続く、ストレージシステムが第2の定足数モデルに参加するが、その第2の定足数モデルで勝つことに依存しない追加のステップが必要となる場合がある。これは、1つのシステムだけが定足数モデルに依存するために変更を処理した場合、過半数は絶対にないので、次いでそれは決して定足数を達成しないという事実を説明するために必要である場合があつてよい。高可用性パラメータ（仲介関係、定足数モデル、引継ぎの優先度）を変更するためにこのモデルを実施させた状態で、ポッドを2つに分割する又は2つのポッドを1つに接合するためにこれらの動作のための安全手順を作成できる。これは、他の1つの機能、つまり2つのポッドが、互換性がある高可用性パラメータを含む場合、第1のポッドにリンクされた第2のポッドがデタッチ関係の処理及び動作、オフライン状態及び同期状態、並びに復旧及び再同期アクションを決定し、推進するために第1のポッドに依存できるように、高可用性のために第2のポッドを第1のポッドにリンクすることを追加することを必要とする場合がある。

【0154】

【0212】 いくつかのボリュームを新規に作成されたポッドに移動させるための動作である、1つのポッドを2つに分割するために、以下、つまり以前には第1のポッドにあったボリュームの集合を移動させる第2のポッドを形成すること、第1のポッドから第2のポッドに高可用性パラメータをコピーして、それらがリンクのために互換性があることを確実にすること、及び高可用性のために第2のポッドを第1のポッドにリンクすることとして説明できる分散動作が形成されてよい。この動作はメッセージとして符号化されてよく、ストレージシステムが、動作がそのストレージシステムで完全に発生する、又は処理が障害により割り込まれる場合まったく発生しないことを確実にするように、ポッド内の各ストレージシステムによって実装されるべきである。いったん2つのポッドのためのすべての同期中のストレージシステムがこの動作を処理すると、ストレージシステムは次いで、第2のポッドがもはや第1のポッドにリンクされないように、第2のポッドを変更する後続の動作を処理する場合がある。ポッドのための高可用性特徴に対する他の変更と同様に、これは、最初にそれぞれの同期中のストレージシステムを、以前のモデル（そのモデルは、高可用性が第1のポッドにリンクされることである）と新しいモデル（そのモデルは、それ自体が現在独立した高可用性である）の両方に依存させることを含む。仲介又は定足数の場合、これは、この変更を処理したストレージシステムが、最初に、第1のポッドのために適宜に達成される仲介又は定足数に依存し、さらに第2のポッドが、仲介又は定足数のための試験を必要とした障害に続いて進む前に、第2のポッドのために達成されている新しい別個の仲介（例えば、新しい仲介鍵）又は定足数に依存することを意味する。定足数モデルを変更する上記説明と同様に、中間ステップは、ストレージシステムが第2のポッドのための定足数に参加し、依存するステップの前に、第2のポッドのための定足数に参加するためにストレージシステムを設定してよい。いったんすべての同期中のストレージシステムが、第1のポッドと第2のポッドの両方のための仲介又は定足数のために新しいパラメータに依存するために変更を処理すると、分割は完了する。

【0155】

【0213】 第2のポッドを第1のポッドに接合することは、本来逆に動作する。第1に、第2のポッドは、ストレージシステムの同一リストを有することによって、及び互換性のある高可用性モデルを有することによって、第1のポッドと互換性があるように調整されなければならない。これは、例えば本書の他の箇所に説明されるストレージシステムを加える若しくは削除するための、又はメディアータ及び定足数モデルを変更するためのステップ等のステップのなんらかの集合を含む場合がある。実施態様に応じて、ストレージシステムの同一リストに達することだけが必要な場合がある。接合することは、それぞれの同期中のストレージシステムで動作を処理して、高可用性のために第2のポッドを第1のポッドにリンクさせることによって進む。その動作を処理する各ストレージシステムは、次いで高可用性のために第1のポッドに、次いで高可用性のために第2のポッドに依存する。第2のポッドのためのすべての同期中のストレージシステムがその動作を処理すると

、ストレージシステムは、次いで第2のポッドと第1のポッドとの間のリンクを排除し、第2のポッドから第1のポッドにボリュームを移行し、第2のポッドを削除するためにそれぞれ後続の動作を処理する。ホスト又はアプリケーションデータセットアクセスは、実施態様が、ホスト若しくはアプリケーションデータセットの修正の適切な方向、又はアイデンティゴとのボリュームに対する読取り動作を可能にする限り、及びアイデンティティがストレージプロトコル又はストレージモデルに対して適宜に保存される限り（例えば、ボリューム及びボリュームにアクセスするためのターゲットポートの使用のための論理ユニット識別子がSCSIの場合に保存される限り）これらの動作を通して保存できる。

【0156】

【0214】 ポッド間でボリュームを移行することが問題を呈する場合がある。ポッドが、同期中のメンバーシップストレージシステムの同一の集合を有する場合、次いで、移行中のボリュームに対する動作を一時的に中断し、それらのボリュームに対する動作に対する制御を新しいポッドのための制御ソフトウェア及び構造に切り替え、次いで動作を再開することが簡単な場合がある。ネットワーク及びポートがポッド間で適切に移行するならば、これは、非常に短い動作の一時中断とは別に、アプリケーションのために連続アップタイムを有するシームレスな移行を可能にする。実施態様に依拠して、一時中断動作は、必要ではない場合もあれば、システムにとって非常に内部であるので、動作の一時中断は影響を及ぼさない場合もある。異なる同期中のメンバーシップ集合を有するポッド間でボリュームをコピーすることは、ますます問題である。コピーのターゲットポッドがソースポッドからの同期中のメンバーの部分集合を有する場合、これはそれほど問題ではない。メンバーシップストレージシステムは、より多くの作業を行う必要なく十分に安全に外すことができる。しかし、ターゲットポッドが同期中のメンバーシップストレージシステムをソースポッド上のボリュームに加える場合、次いで加えられたストレージシステムは、それらが使用できるようになる前にボリュームのコンテンツを含むために同期しなければならない。障害処理が異なり、まだ同期していないメンバーシップストレージシステムからの要求処理がうまくいかない又は転送されなければならない又は読取りが相互接続を横断しなければならないため速くないという点で、同期するまで、これは、コピーされたボリュームをすでに同期しているボリュームとはっきりと異なったままとする。また、内部実装は、同期中で、障害処理のために準備完了しているいくつかのボリュームを処理しなければならない、他は同期していない。

【0157】

【0215】 障害に直面した動作の信頼性に関する他の問題がある。マルチストレージシステムポッド間でボリュームの移行を調整することは、分散動作である。ポッドが障害処理及び回復の単位である場合、及び仲介又は定足数又はどのような手段でもスプリットブレイク状況を回避するために使用される場合、次いで障害処理、回復、仲介、及び定足数のための状態及び構成及び関係の特定の集合を有するあるポッドから別のポッド、次いでポッドのストレージシステムへのボリュームの切替えは、あらゆるボリュームのためのその処理に関係する変更を調整することについて注意深くなくてはならない。動作は、ストレージシステム間でアトミックに分散することはできないが、なんらかの方法で段階分けされなければならない。仲介モデル及び定足数モデルは、本来ポッドに分散トランザクションアトミック性を実装するためのツールを提供するが、これは、実装を増加させることなくポッド間動作には及ばない場合がある。

【0158】

【0216】 同じ第1のストレージシステム及び第2のストレージシステムを共用する2つのポッドの場合の第1のポッドから第2のポッドへのボリュームの単純な移行も検討する。ある時点で、ストレージシステムは、ボリュームが現在第2のポッドにあり、もはや第1のポッドにはないことを定めるために調整する。2つのポッドのためのストレージシステム全体でトランザクションアトミック性のための固有の機構がない場合、次いでネイティブ実装は、2つのポッドからストレージシステムをデタッチするための障害処理を生じさせるネットワーク障害のときに、第1のポッド内のボリュームを第1のストレージシ

テムに残し、第2のポッドを第2のストレージシステムに残す場合があるであろう。ポッドが、どのストレージシステムが他方をデタッチするのに成功するのかを別々に判断する場合、結果は、同じストレージシステムが両方のポッドのために他のストレージシステムをデタッチすることになり、その場合、ボリューム移行回復の結果には一貫性がなければならない、又は異なるストレージシステムが、2つのポッドのために他をデタッチすることになるであろう。第1のストレージシステムが第1のポッドのために第2のストレージシステムをデタッチし、第2のストレージシステムが第2のポッドのために第1のストレージシステムをデタッチする場合、次いで回復により、ボリュームが第1のストレージシステムでの第1のポッドに、及び第2のストレージシステムでの第2のポッドに回復され、ボリュームは次いで実行中であり、両方のストレージシステムのホスト及びストレージアプリケーションにエクスポートされる。代わりに、第2のストレージシステムが、第1のポッドのために第1のストレージシステムをデタッチし、第1のストレージが第2のポッドのために第2のストレージシステムをデタッチする場合、次いで回復により、ボリュームは、第1のストレージシステムによって第2のポッドから廃棄され、ボリュームは、第2のストレージシステムによって第1のポッドから廃棄され、ボリュームが完全に消える結果となる可能性がある。ボリュームが間で移行されているポッドがストレージシステムの異なる集合上にある場合、次いでものごとにはさらに複雑化する場合がある。

【0159】

【0217】 これらの問題の解決策は、ポッドを分割し、接合するための上述された技術とともに中間ポッドを使用することである場合がある。この中間ポッドは、ストレージシステムと関連付けられた可視管理オブジェクトとして絶対に提示されてはならない。このモデルでは、第1のポッドから第2のポッドに移動されるボリュームは、最初に、上述された分割動作を使用し、第1のポッドから新しい中間ポッドに分割される。中間ポッドのためのストレージシステムメンバーは、次いで、必要に応じてストレージシステムを加える又はストレージシステムをポッドから削除することによって、ストレージシステムのメンバーシップに一致するように調整できる。その後、第2のポッドに中間ポッドを接合する場合がある。

【0160】

【0218】 追加の説明のため、図5は、本開示のいくつかの実施形態に従ってポッドをサポートするストレージシステム（402、404、406）によって実行されてよいステップを示すフローチャートを説明する。あまり詳細には示されていないが、図5に示されるストレージシステム（402、404、406）は、図1Aから図1D、図2Aから図2G、図3Aから図3B、図4、又はその任意の組合せに関して上述されたストレージシステムに類似してよい。実際には、図5に示されるストレージシステム（402、404、406）は、上述されたストレージシステムと同じ構成要素、より少ない構成要素、追加の構成要素を含んでよい。

【0161】

【0219】 図5に示される例の方法では、ストレージシステム（402）は、ポッドにアタッチしてよい（508）。ポッドメンバーシップのためのモデルは、ストレージシステムのリスト、及びストレージシステムがポッドのために同期中であると推定されるそのリストの部分集合を含んでよい。ストレージシステムは、それが、少なくともポッドと関連付けられたデータセットの最後の作成コピーの同一のアイドルコンテンツを有する回復の中にある場合、ポッドのために同期している。アイドルコンテンツは、新しい修正の処理がなく任意の進行中の修正が完了した後のコンテンツである。これは「クラッシュ回復可能」一貫性と呼ばれることがある。ポッドメンバーとして一覧されているが、ポッドのために同期中として一覧されていないストレージシステムは、ポッドから「デタッチされている」として記述される場合がある。ポッドメンバーとして一覧されるストレージシステムは、ポッドのために同期中であり、ポッドのために「オンラインである」ポッドのためのデータをアクティブに供給するために現在利用可能である。

【0162】

【0220】 図5に示される例の方法では、ストレージシステム(402)は、例えば用語が上述されたように、オンラインであるポッドの他のストレージシステム(404、406)に記憶されるデータセット(426)の最新バージョンとともにデータセット(426)のそのローカルに記憶されているバージョンを同期させることによって、ポッドにアタッチしてよい(508)。係る例では、ストレージシステム(402)がポッドにアタッチする(508)ために、ポッド内のストレージシステム(402、404、406)のそれぞれの中にローカルに記憶されるポッド定義は、ストレージシステム(402)がポッドにアタッチする(508)ために、更新される必要がある。係る例では、ポッドの各ストレージシステムメンバーは、それが最後に知っていたどのストレージシステムが同期していたのか、及びそれが最後に知っていたどのストレージシステムがポッドメンバーの集合全体を含んでいたのかを含む、メンバーシップの独自のコピーを有する場合がある。

10

【0163】

【0221】 図5に示される例の方法では、ストレージシステム(402)は、データセット(426)の一部分を読み取る要求を受け取ってもよく(510)、ストレージシステム(402)は、データセット(426)の部分を読み取る要求をローカルに処理してよい(512)。読者は、データセット(426)を修正する要求(例えば、書込み動作)がポッド内のストレージシステム(402、404、406)間の調整を必要とするが、データセット(426)は、ポッド内のすべてのストレージシステム(402、404、406)全体で一致している必要があるので、データセット(426)の一部分を読み取る要求に応えることは、ストレージシステム(402、404、406)間の類似する調整を必要としないことを理解する。したがって、読取り要求を受け取る特定のストレージシステム(402)は、ポッド内の他のストレージシステム(404、406)との同期通信がない状態で、ストレージシステムの(402)ストレージデバイスの中に記憶されるデータセット(426)の一部分を読み取ることによって読取り要求にローカルにサービスを提供してよい。複製されたクラスタ内の複製されたデータセットのために1つのストレージシステムによって受け取られる読取り要求は、少なくともやはり名目上実行中であるクラスタの中で実行中であるストレージシステムによって受け取られるとき、圧倒的多数の場合に任意の通信を回避することを期待される。係る読取りは、通常、クラスタ内の他のストレージシステムとの追加の対話が必要とされない状態で、クラスタ化されたデータセットのローカルコピーから単に読み取ることによって正常に処理されるべきである。

20

30

【0164】

【0222】 読者は、読取り要求がどのストレージシステムが読取り要求を処理するのかに関わらずに同じ結果を返すように、ストレージシステムが、読取り一貫性を確実にするための処置を講じてよいことを理解する。例えば、クラスタ内のストレージシステムの任意の集合によって受け取られる更新の任意の集合に対する結果として生じるクラスタ化されたデータセットコンテンツは、クラスタ全体で一貫性があるべきであり、少なくとも任意の時点で更新はアイドルである(すべての以前の修正動作は完了として示されており、新しい更新要求は決して受け取られず、処理されていない)。すなわち、ストレージシステムの集合全体でのクラスタ化されたデータセットのインスタンスは、まだ完了していない更新の結果としてだけ異なる場合がある。つまり、例えば、そのボリュームブロック範囲で重複する任意の2つの書込み要求、又は書込み要求と重複するスナップショット、コンペアアンドライト、又は仮想ブロック範囲コピーの任意の組合せは、データセットのすべてのコピーに対する一貫した結果を生じさせなければならない。2つの動作は、あたかもそれらがあるストレージシステムではある順序で、及び複製クラスタの別のストレージシステムでは異なる順序で起こるかのように結果を生じさせることはできない。

40

【0165】

【0223】 さらに、読取り要求は時間順序が一貫してよい。例えば、1つの読取り要求が複製されたクラスタで受け取られ、完了し、次いでその読取りの後に、複製クラスタによって受け取られる重複するアドレス範囲に対する別の読取り要求が続き、(読取り又は修正のいずれかが、複製クラスタ内の同じストレージシステムによって受け取られるのか、

50

それとも異なるストレージシステムによって受け取られるのかに関わらず) 一方の読取り又は両方の読取りが、複製クラスタによって受け取られる修正要求と、時間及びボリュームアドレス範囲で多少なりとも重複している場合、次いで第1の読取りが更新の結果を反映する場合、次いで第2の読取りも、おそらく更新に先行したデータを返すよりむしろ、その更新の結果を反映するべきである。第1の読取りが更新を反映しない場合、次いで第2の読取りは、更新を反映する、又は反映しないかのどちらかである場合がある。これは、2つの読取り要求の間で、データセグメントの「時間」が決して過去に戻るということがないことを確実にする。

【0166】

【0224】 図5に示される例の方法では、ストレージシステム(402)は、他のストレージシステム(404、406)の1つ以上とのデータ通信の中断を検出する(514)場合もある。他のストレージシステム(404、406)の1つ以上とのデータ通信の中断は、さまざまな理由で発生する場合がある。例えば、他のストレージシステム(404、406)の1つ以上とのデータ通信の中断は、ストレージシステム(402、404、406)のうちの1つが故障したため、ネットワーク相互接続が故障したため、又は他のなんらかの理由により発生する場合がある。同期複製クラスタ化の重要な態様は、いかなる障害処理も、回復不可能な不一致、又は対応の不一致を生じさせないことを確実にすることである。例えば、ネットワークが2つのストレージシステム間で故障する場合、ストレージシステムのうちの多くても1つがポッドのために新規に入信するI/O要求を処理し続ける場合がある。そして、一方のストレージシステムが処理を続行する場合、他方のストレージシステムは、読取り要求を含んだ、新しい要求を完了まで処理することはできない。

【0167】

【0225】 図5に示される例の方法では、ストレージシステム(402)は、特定のストレージシステム(402)がポッドの一部としてオンラインに留まるべきかどうかを判断してもよい(516)。上述されたように、ポッドの一部として『オンライン』であるためには、ストレージシステムは、それ自体をポッドのために同期中であると思なす必要があり、それがポッドのために同期中であると思なす他のすべてのストレージシステムと通信していなければならない。ストレージシステムが、それが同期中であり、同期中である他のすべてのストレージシステムと通信中であることを確信できない場合、次いでストレージシステムはデータセット(426)にアクセスするための新しい入信要求を処理することを停止する場合がある。したがって、ストレージシステム(402)は、特定のストレージシステム(402)がポッドの一部としてオンラインに留まるべきかどうかを、ストレージシステムが、それがポッドのために同期中であると思なすすべての他のストレージシステム(404、406)と(例えば、1つ以上のテストメッセージを介して)通信できるかどうかを判断することによって、それがポッドのために同期中であると思なす他のすべてのストレージシステム(404、406)もストレージシステム(402)がポッドにアタッチされるべきと思なすかどうかを判断することによって、特定のストレージシステム(402)が、それがポッドのために同期中であると思なす他のすべてのストレージシステム(404、406)と通信できること、及びそれがポッドのために同期中であると思なす他のすべてのストレージシステム(404、406)もストレージシステム(402)をポッドにアタッチされるべきであると思なすことを確認しなければならない両方のステップの組合せによって、あるいはなんらかの他の仕組みによって、判断してよい(516)。

【0168】

【0226】 図5に示される例の方法では、ストレージシステム(402)は、特定のストレージシステム(402)がポッドの一部としてオンラインに留まるべきであると肯定的に判断する(518)ことに応えて、特定のストレージシステム(402)上のデータセット(426)を管理動作及びデータセット動作のためにアクセス可能にしておいてもよい(522)。ストレージシステム(402)は、例えば、ストレージシステム(402)

）上に記憶されるデータセット（４２６）のバージョンにアクセスする要求を受け入れ、係る要求を処理することによって、ホスト又は権限を与えられた管理者によって発行されるデータセット（４２６）と関連付けられた管理動作を受け入れ、処理することによって、ポッド内の他のストレージシステム（４０４、４０６）の１つによって、又は他のなんらかの方法で発行されるデータセット（４２６）と関連付けられた管理動作を受け入れ、処理することによって、特定のストレージシステム（４０２）上のデータセット（４２６）を、管理動作及びデータセット動作のためにアクセス可能にしてよい（５２２）。

【０１６９】

【０２２７】 図５に示される例の方法では、ストレージシステム（４０２）は、また、特定のストレージシステムがポッドの一部としてオンラインに留まるべきではない（５２０）と判断したことに応じて、特定のストレージシステム（４０２）上のデータセット（４２６）を管理動作及びデータセット動作のためにアクセス不可にしてもよい（５２４）。ストレージシステム（４０２）は、例えば、ストレージシステム（４０２）に記憶されるデータセット（４２６）のバージョンにアクセスする要求を拒絶することによって、ホスト又は他の権限を与えられた管理者によって発行されるデータセット（４２６）と関連付けられた管理動作を拒絶することによって、ポッド内の他のストレージシステム（４０４、４０６）の１つによって発行されるデータセット（４２６）と関連付けられた管理動作を拒絶することによって、又は他のなんらかの方法で、特定のストレージシステム（４０２）上のデータセット（４２６）を管理動作及びデータセット動作のためにアクセス不可にしてもよい（５２４）。

【０１７０】

【０２２８】 図５に示される例の方法では、ストレージシステム（４０２）は、他のストレージシステム（４０４、４０６）の１つ以上とのデータ通信の中断が修復されたことを検出してもよい（５２６）。ストレージシステム（４０２）は、例えば、他のストレージシステム（４０４、４０６）の１つ以上からメッセージを受け取ることによって、他のストレージシステム（４０４、４０６）の１つ以上とのデータ通信の中断が修復されたことを検出してよい（５２６）。他のストレージシステム（４０４、４０６）の１つ以上とのデータ通信の中断が修復されたことを検出する（５２６）ことに応じて、ストレージシステム（４０２）は、特定のストレージシステム（４０２）上のデータセット（４２６）を管理動作及びデータセット動作のためにアクセス可能にしてよい（５２８）。

【０１７１】

【０２２９】 読者は、図５に示される例が、順序付けは必要とされていないが、多様なアクションが同じ順序の中で発生すると示される実施形態を説明することを理解する。さらに、ストレージシステム（４０２）が、説明されるアクションの部分集合を実施するにすぎない他の実施形態が存在してよい。例えば、ストレージシステム（４０２）は、他のストレージシステム（４０４、４０６）の１つ以上とのデータ通信での中断を検出する（５１４）ステップ、特定のストレージシステム（４０２）がポッドに留まるべきかどうかを判断する（５１６）ステップ、特定のストレージシステム（４０２）上のデータセット（４２６）を管理動作及びデータセット動作のためにアクセス可能にしておく（５２２）ステップ、又は最初にデータセット（４２６）の一部分を読み取る要求を受け取って（５１０）、データセット（４２６）の一部分を読み取る要求をローカルで処理する（５１２）ことなく、特定のストレージシステム（４０２）上のデータセット（４２６）を管理動作及びデータセット動作のためにアクセス不可にする（５２４）ステップを実行してよい。さらに、ストレージシステム（４０２）は、他のストレージシステム（４０４、４０６）の１つ以上とのデータ通信の中断が修復されたことを検出し（５２６）、最初にデータセット（４２６）の一部分を読み取る要求を受け取って（５１０）、データセット（４２６）の一部分を読み取る要求をローカルで処理する（５１２）ことなく、特定のストレージシステム（４０２）上のデータセット（４２６）を管理動作及びデータセット動作のためにアクセス可能にしてよい（５２８）。実際には、本明細書に説明されるステップのいずれも、本明細書に説明される他のステップを実行するための前提条件として、すべての

実施形態で明示的に必要とされていない。

【0172】

【0230】 追加の説明のために、図6は、本開示のいくつかの実施形態に従ってポッドをサポートするストレージシステム(402、404、406)によって実行されてよいステップを示すフローチャートを説明する。あまり詳細には示されないが、図6に示されるストレージシステム(402、404、406)は、図1Aから図1D、図2Aから図2G、図3Aから図3B、図4、又はその任意の組合せに関して上述されたストレージシステムに類似してよい。実際には、図6に示されるストレージシステム(402、404、406)は、上述されたストレージシステムと同じ構成要素、より少ない構成要素、追加の構成要素を含んでよい。

10

【0173】

【0231】 図6に示される例の方法では、ストレージシステム(402、404)の2つ以上は、データセット(426)を非同期で受け取るためのそれぞれターゲットストレージシステム(618)を識別してよい(608)。データセット(426)を非同期で受け取るためのターゲットストレージシステム(618)は、クラウドサービスプロバイダにより提供されるクラウドストレージとして、又は多くの他の方法で、例えば、特定のポッドのメンバーであるストレージシステム(402、404)のどちらとも異なるデータセンタに位置するバックアップストレージシステムとして実施されてよい。読者は、ターゲットストレージシステム(618)は、その全体でデータセット(426)が同期複製される複数のストレージシステム(402、404)の1つではなく、したがってターゲットストレージシステム(618)は、当初、データセット(426)の最新のローカルコピーを含まないことを理解する。

20

【0174】

【0232】 図6に示される例の方法では、ストレージシステム(402、404)の2つ以上は、それぞれ、データセット(426)を含むポッドのメンバーである他のストレージシステムのいずれかによってターゲットストレージ(618)システムに非同期複製されていないデータセット(426)の一部分を識別してよい(610)。係る例では、ストレージシステム(402、404)は、それぞれ、ターゲットストレージシステム(618)に、他のストレージシステムのいずれかによってターゲットストレージシステムに非同期複製されていないデータセット(426)の一部分を非同期複製してよい(612)。第1のストレージシステム(402)が、データセット(426)の第1の部分(例えば、アドレス空間の第1の半分)をターゲットストレージシステム(618)に非同期複製することを担う例を考える。係る例では、第2のストレージシステム(404)は、2つ以上のストレージシステム(402、404)がデータセット(426)全体をターゲットストレージシステム(618)に集合的に複製するように、データセット(426)の第2の部分(例えば、アドレス空間の第2の半分)をターゲットストレージシステム(618)に非同期複製することを担うであろう。

30

【0175】

【0233】 読者は、上述されたように、ポッドの使用により、2つのストレージシステム間の複製関係が、データが非同期複製される関係からデータが同期複製される関係へ切り替えられ得ることを理解する。例えば、ストレージシステムAがストレージシステムBにデータセットを非同期複製するように構成される場合、データセット、メンバーとしてのストレージシステムA、及びメンバーとしてのストレージシステムBを含むポッドを作成することは、データが非同期複製される関係からデータが同期複製される関係へ関係を切り替えることができる。同様に、ポッドを使用することによって、2つのストレージシステム間の複製関係は、データが同期複製される関係から、データが非同期複製される関係へ切り替えられてよい。例えば、データセット、メンバーとしてのストレージシステムA、及びメンバーとしてのストレージシステムBを含むポッドが、(メンバーとしてのストレージシステムAを削除するために又はメンバーとしてのストレージシステムBを削除するために)単にポッドを伸展しないことによって作成される場合、データがストレージシ

40

50

システム間で同期複製される関係は、データが非同期複製される関係にただちに切り替えることができる。係るようにして、ストレージシステムは、非同期複製と同期複製との間で、必要に応じて前後に切り替わってよい。

【0176】

【0234】 この切替えは、実施態様が同期複製と非同期複製の両方のための類似した技術に依存することにより助長される場合がある。例えば、同期複製されたデータセットの再同期が非同期複製に使用されるのと同じ仕組み又は互換性のある仕組みに頼る場合、次いで非同期複製への切替えは、同期中の状態を外し、関係を「恒久回復」モードに類似した状態に残すことと概念的に同一である。同様に、非同期複製から同期複製へ切り替えることは、再同期を完了し、切り替えシステムが同期中のポッドメンバーになるときに「追い付き」、まさに行われるように同期中になることによって概念上作用する場合がある。

10

【0177】

【0235】 代わりに又はさらに、同期複製と非同期複製の両方が、類似する又は同一の共通メタデータ、又は論理エクステント若しくは記憶されたブロックアイデンティティを表し、識別するための共通モデル、又はコンテンツアドレス指定可能な記憶されたブロックを表すための共通モデルに頼る場合、次いでこれらの共通性の態様は、同期複製及び非同期複製に並びに同期複製及び非同期複製から切り替わるとき転送される必要がある場合があるコンテンツを劇的に削減するために利用される場合がある。さらに、データセットが、ストレージシステムAからストレージシステムBに非同期複製され、システムBがさらにそのデータセットをストレージシステムCに非同期複製する場合、次いで共通メタデータモデル、共通論理エクステント若しくはブロックアイデンティティ、又はコンテンツアドレス指定可能記憶ブロックの共通表現は、ストレージシステムAとストレージシステムCとの間の同期複製を可能にするために必要とされるデータ転送を劇的に削減できる。

20

【0178】

【0236】 読者は、上述されたように、ポッドを使用することにより、複製技術が、データ複製以外のタスクを実行するために使用され得ることをさらに理解する。実際には、ポッドは管理オブジェクトの集合を含んでよい。ため、バーチャルマシンを移行することのようなタスクは、本明細書に説明されるポッド及び複製技術を使用し、実施され得る。例えば、バーチャルマシンAがストレージシステムAで実行中である場合、管理オブジェクトとしてのバーチャルマシンA、メンバーとしてのストレージシステムA、及びメンバーとしてのストレージシステムBを含むポッドを作成することによって、バーチャルマシンA及び任意の関連付けられた画像及び定義はストレージシステムBに移行されてよく、そのときに、ポッドが単に破壊される場合がある、メンバーシップが更新される場合がある、又は他の処置が必要に応じて取られ得る。

30

【0179】

【0237】 追加の説明のために、図7は、本開示のいくつかの実施形態に従って2つ以上のストレージシステム(714、724、728)の間で同期複製関係を確立する例の方法を示すフローチャートを説明する。あまり詳細には示されないが、図7に示されるストレージシステム(714、724、728)は、図1Aから図1D、図2Aから図2G、図3Aから図3B、又はその任意の組合せに関して上述されたストレージシステムに類似してよい。実際には、図7に示されるストレージシステム(714、724、728)は、上述されたストレージシステムと同じ構成要素、より少ない構成要素、追加の構成要素を含んでよい。

40

【0180】

【0238】 図7に示される例の方法は、データセット(712)のために、データセット(712)がその全体で同期複製される複数のストレージシステム(714、724、728)を識別すること(702)を含む。図7に示されるデータセット(712)は、例えば、特定のボリュームのコンテンツとして、ボリュームの特定のシャードのコンテンツとして、又は1つ以上のデータ要素の任意の他の集合体として実施されてよい。データセット(712)は、各ストレージシステム(714、724、728)がデータセット(

50

712) のローカルコピーを保持するように、複数のストレージシステム (714、724、728) 全体で同期されてよい。本明細書に説明される例では、係るデータセット (712) は、少なくともアクセスされているクラスタ及び特定のストレージシステムが名目上動作している限り、クラスタ内の任意の一方のストレージシステムが、クラスタ内の任意の他方のストレージシステムよりも実質的により最適に動作しないような性能特性を有するストレージシステム (714、724、728) のいずれかを通してデータセット (712) にアクセスできるように、ストレージシステム (714、724、728) 全体で同期複製される。係るシステムでは、データセット (712) に対する修正は、ストレージシステム (714、724、728) のいずれかでデータセット (712) にアクセスすることが一貫した結果を生じさせるように、各ストレージシステム (714、724、728) に常駐するデータセットのコピーに対して加えられるべきである。例えば、データセットに発行される書込み要求は、すべてのストレージシステム (714、724、728) でサービスを提供されなければならない、又はストレージシステム (714、724、728) のいずれでもサービスを提供されてはならない。同様に、動作のいくつかのグループ (例えば、データセットの中の同じ場所に向けられる2つの書込み動作) は、各ストレージシステム (714、724、728) に常駐するデータセットのコピーが最終的に同一となるように、すべてのストレージシステム (714、724、728) 上で同じ順序で実行されなければならない。データセット (712) に対する修正は、正確に同時に行われる必要はないが、いくつかのアクション (例えば、書込み要求がデータセットに向けられ、すべてのストレージシステムでまだ完了していない書込み要求によってターゲットにされるデータセットの中の場所への読取りアクセスを可能にする旨の肯定応答を発行すること) は、各ストレージシステム (714、724、728) 上のデータセット (712) のコピーが修正されるまで先延ばしにされてよい。

【0181】

[0239] 図7に示される例の方法では、データセット (712) のために、データセット (712) がその全体で同期複製される複数のストレージシステム (714、724、728) を識別すること (702) は、例えば、ポッド定義、又はデータセット (712) を名目上そのデータセット (712) を記憶する1つ以上のストレージシステム (714、724、728) と関連付ける類似したデータ構造を調べることによって実施されてよい。『ポッド』は、該用語がここで及び本願の残りを通して使用されるように、データセット、管理オブジェクト及び管理動作の集合、データセットを修正する又は読み取るためのアクセス動作の集合、及び複数のストレージシステムを表す管理エンティティとして実施されてよい。係る管理動作は、データセットを読み取る又は修正するためのアクセス動作が、ストレージシステムのいずれかを通して同等に動作する、ストレージシステムのいずれかを通して同等に管理オブジェクトを修正する又は問い合わせしてよい。各ストレージシステムは、ストレージシステムによる使用のために記憶され、宣伝されたデータセットの適切な部分集合としてデータセットの別々のコピーを記憶してよく、任意の1つのストレージシステムを通して実行され、完了された管理オブジェクト又はデータセットを修正する動作は、ポッドを問い合わせるために後続の管理オブジェクトで、又はデータセットを読み取るために後続のアクセス動作で反映される。『ポッド』に関する追加の詳細は、参照により本明細書に援用される以前に出願された仮特許出願第62/518,071号で見つけられてよい。係る例では、ポッド定義は、データセット (712)、及びデータセット (712) がその全体で同期複製されるストレージシステム (714、724、728) の集合の少なくとも識別を含んでよい。係るポッドは、対称アクセス、レプリカの柔軟な追加/削除、高可用性データ一貫性、データセット、管理ホストアクセス、アプリケーションクラスタ化等に対する関係性におけるストレージシステム全体での一元的なユーザ管理等を含んだ多くの (おそらく任意選択の) 特性のいくつかをカプセル化してよい。ストレージシステムはポッドに加えられる場合があり、結果的にポッドのデータセット (712) はそのストレージシステムにコピーされ、次いでデータセット (712) が修正されると、最新に保たれる。また、ストレージシステムはポッドから削除される場

10

20

30

40

50

合もあり、結果的にデータセット（712）は削除されたストレージシステムでもはや最新に保たれなくなる。係る例では、ポッド定義又は類似するデータ構造は、ストレージシステムが特定のポッドに加えられ、特定のポッドから削除されると更新されてよい。

【0182】

【0240】 また、図7に示される例の方法は、データセット（712）を同期複製するために使用される複数のストレージシステム（714、724、728）のそれぞれの間で1つ以上のデータ通信リンク（716、718、720）を構成すること（704）も含む。図6に示される例の方法では、ポッド内のストレージシステム（714、724、728）は、高帯域幅データ転送のため、並びにクラスタ通信、ステータス通信、及び管理通信のための両方のために互いと通信しなければならない。これらの別個のタイプの通信は、同じデータ通信リンク（716、718、720）を介する場合もあれば、代替実施形態では、これらの別個のタイプの通信は、別々のデータ通信リンク（716、718、720）を介する場合もあるであろう。デュアルコントローラストレージシステムのクラスタでは、各ストレージシステムの両方のコントローラとも、任意の対にされたストレージシステム（つまり、ポッド内の任意の他のストレージシステム）のための両方のコントローラと通信する名目能力を有するべきである。

【0183】

【0241】 一次／二次コントローラ設計では、アクティブな複製のための全クラスタ通信は、障害が発生するまで一次コントローラ間で実行してよい。係るシステムでは、係るエンティティ間のデータ通信リンクが操作可能であることを検証するために、一次コントローラと二次コントローラとの間、又は別個のストレージシステムの二次コントローラ間でなんらかの通信が発生してよい。他の場合、仮想ネットワークアドレスは、データセンタ間のネットワークリンクのために必要とされる構成を制限するために、又はストレージシステムのクラスタ化された態様の設計を簡略化するために使用される可能性がある。アクティブ／アクティブコントローラ設計では、クラスタ通信は、あるストレージシステムのすべてのアクティブコントローラから任意の対にされたストレージシステムのいくつかの若しくはすべてのアクティブコントローラまで続く可能性がある、又はクラスタ通信は共通スイッチを通してフィルタにかけられる可能性がある、又はクラスタ通信は構成を簡略化するために仮想ネットワークアドレスを使用する可能性がある、又はクラスタ通信はなんらかの組合せを使用する可能性がある。スケールアウト設計では、2つ以上の共通ネットワークスイッチは、データトラフィックを処理するためにストレージシステムの中のすべてのスケールアウトストレージコントローラが、ネットワークスイッチに接続するように使用されてよい。スイッチは、曝露されるネットワークアドレスの数を制限するための技術を使用する可能性もあれば、使用しない可能性もあり、これにより対にされたストレージシステムは、すべてのストレージコントローラのネットワークアドレスで構成される必要はない。

【0184】

【0242】 図7に示される例の方法では、データセット（712）を同期複製するために使用される複数のストレージシステム（714、724、728）のそれぞれの間の1つ以上のデータ通信リンク（716、718、720）を構成すること（704）は、例えば、データ通信ネットワークを介して定められたポートを介して通信するようにストレージシステム（716、718、720）を構成することによって、ストレージシステム（716、724、728）の2つの間のポイントツーポイントデータ通信リンクを介して通信するようにストレージシステム（716、718、720）を構成することによって、又はさまざまな方法で実施されてよい。安全な通信が必要とされる場合、なんらかの形の鍵交換が必要とされる場合もあれば、通信は、例えばSSH（Secure Shell）、SSL、又は公開鍵若しくはディフィーヘルマン鍵共用の回りで構築された他のなんらかのサービス又はプロトコル若しくは妥当な代替策等のなんらかのサービスを通して行われる又はブートされる場合もあるであろう。また、安全な通信は、なんらかの方法でカスタマイズされたアイデンティティと結び付けられるベンダ提供のなんらかのクラウドサービス

を通して仲介される場合もあるであろう。代わりに、例えばバーチャルマシン又は仮想コンテナで実行する等、カスタマ施設で実行するように構成されたサービスは、複製するストレージシステム（716、718、720）間の安全な通信のために必要な鍵交換を仲介するために使用できるであろう。読者は、3つ以上のストレージシステムを含んだポッドが、個々のストレージシステムの大部分又はすべての間で通信リンクを必要とする場合があることを理解する。他の実施形態では追加のデータ通信リンクが存在する場合があるが、図6に示される例では、3つのデータ通信リンク（716、718、720）が示されている。

【0185】

【0243】 読者は、データセット（712）がそれ全体で同期複製されるストレージシステム（714、724、728）間の通信が、いくつかの目的を果たすことを理解する。例えば、1つの目的は、I/O処理の一部として、あるストレージシステム（714、724、728）から別のストレージシステム（714、724、728）にデータを送達することである。例えば、一般的に書き込みを処理することは、書き込みコンテンツ及び書き込みのなんらかの記述をポッドのための任意の対にされたストレージシステムに送達することを必要とする。ストレージシステム（714、724、728）の間のデータ通信によって果たされる別の目的は、ボリューム、ファイル、オブジェクトバケット等を作成すること、拡張すること、削除すること、又は名前を変えることを処理するために、構成変更及び分析データを通信することである場合がある。ストレージシステム（714、724、728）の間のデータ通信によって果たされる別の目的は、ストレージシステム及び相互接続障害を検出し、処理する際に関与する通信を実施することである場合がある。このタイプの通信はタイムクリティカルである場合があり、それが、書き込みトラフィックの大きいバーストがデータセンタ相互接続時突然ダンプされるとき、長いネットワーク待ち行列遅延の後方で抜け出せなくならないことを確実にするために優先される必要がある場合がある。

【0186】

【0244】 読者は、異なるタイプの通信が、多様な組合せで、同じ接続又は異なる接続を使用する場合があり、同じネットワーク又は異なるネットワークを使用する場合があることをさらに理解する。さらに、一部の通信が暗号化され、保護される場合がある。一方、他の通信は暗号化されない可能性がある。一部の 경우에는、データ通信リンクは、あるストレージシステムから別のストレージシステムに（要求自体として直接的に、又はI/O要求が表す動作の論理記述としてのどちらかで）I/O要求を転送するために使用できるであろう。これは、例えばあるストレージシステムがポッドのために最新の同期中のコンテンツを有し、別のストレージシステムが現在ポッドのために最新の同期中のコンテンツを有していない場合に使用できるであろう。係る場合、データ通信リンクが実行中である限り、要求は最新で同期中ではないストレージシステムから、最新で同期中であるストレージシステムに転送できる。

【0187】

【0245】 また、図7に示される例の方法は、複数のストレージシステム（714、724、728）の間に、複数のストレージシステム（714、724、728）のうちの少なくとも1つのためにタイミング情報（710、722、726）を交換すること（706）も含む。図6に示される例の方法では、特定のストレージシステム（714、724、728）のためのタイミング情報（710、722、726）は、例えばストレージシステム（714、724、728）の中のクロックの値として実施されてよい。代替実施形態では、特定のストレージシステム（714、724、728）のためのタイミング情報（710、722、726）は、クロック値のプロキシとしての機能を果たす値として実施されてよい。クロック値のプロキシとしての機能を果たす値は、ストレージシステムの間で交換されるトークンに含められてよい。例えば、特定のストレージシステム（714、724、728）又はストレージコントローラが、特定のときに送信されたとして内部で記録できるシーケンス番号等、クロック値のためのプロキシとしての機能を果たす値

は、実施されてよい。係る例では、トークン（例えば、シーケンス番号）がまた受け取られる場合、関連付けられたクロック値が見つけれ、有効なリースが依然として実施されているのかを判断するための基準として活用される場合がある。図6に示される例の方法では、複数のストレージシステム（714、724、728）の間で複数のストレージシステム（714、724、728）のうちの少なくとも1つのためのタイミング情報（710、722、726）を交換すること（706）は、例えば各ストレージシステム（714、724、728）が周期的に、オンデマンドで、リース確立後所定の時間量の中で、リースが失効するように設定される前の所定の時間量の中で、同期複製関係を開始する若しくは確立し直す試みの一部として、ポッド内のそれぞれの他のストレージシステム（714、724、728）にタイミング情報を送信することによって、又は他のなんらかの方法で実施されてよい。

10

【0188】

【0246】 また、図7に示される例の方法は、複数のストレージシステム（714、724、728）のうちの少なくとも1つのためのタイミング情報（710、722、726）に従って、同期複製リースを確立すること（708）も含み、同期複製リースは、同期複製関係が有効である期間を識別する。図7に示される例の方法では、同期複製関係は、大部分は独立したこれらのストア間でなんらかのデータ（712）を複製するストレージシステム（714、724、728）の集合として形成され、各ストレージシステム（714、724、728）は、ストレージオブジェクトを定義するため、オブジェクトを物理ストレージにマッピングするため、重複排除のため、スナップショットへのコンテンツのマッピングを定義するため等のために関連するデータ構造の独自のコピー及び独自の別々の内部管理を有する。同期複製関係は特定のデータセットにとって特有である場合があり、これにより特定のストレージシステム（714、724、728）は、1つを超える同期複製関係と関連付けられてよく、各同期複製関係は、記述されているデータセットによって差別化され、さらに追加のメンバーストレージシステムの異なる集合から成ってよい。

20

【0189】

【0247】 図7に示される例の方法では、同期複製リースは、さまざまな異なる方法で複数のストレージシステム（714、724、728）のうちの少なくとも1つのためのタイミング情報（710、722、726）に依存して確立されてよい（708）。一実施形態では、ストレージシステムは、クロックを調整するために複数のストレージシステム（714、724、728）のそれぞれのためのタイミング情報（710、722、726）を活用することによって同期複製リースを確立してよい（708）。係る例では、クロックがストレージシステム（714、724、728）のそれぞれについて調整されると、ストレージシステムは、調整されたクロック値を超えて所定の期間延長する同期複製リースを確立してよい（708）。例えば、各ストレージシステム（714、724、728）のためのクロックがXの値となるように調整される場合、ストレージシステム（714、724、728）は、それぞれ、 $X + 2$ 秒まで有効である同期複製リースを確立するように構成されてよい。

30

【0190】

【0248】 代替実施形態では、ストレージシステム（714、724、728）の間でクロックを調整する必要性は、タイミング保証を達成しつつも回避されてよい。係る実施形態では、各ストレージシステム（714、724、728）の中のストレージコントローラは、ローカルの単調増加クロックを有してよい。各コントローラが、それが他のストレージコントローラから受け取った最後のクロック値に沿って、そのクロック値を他のストレージコントローラに送信することによって、同期複製リースは、ストレージコントローラ（例えば、対にされたストレージシステムの一次コントローラと通信しているあるストレージシステムの一次コントローラ）間で確立されてよい（708）。特定のコントローラが、別のコントローラからそのクロック値をまた受け取るとき、特定のコントローラは、その受け取られたクロック値になんらかの同意されたリース間隔を加え、そのローカル

40

50

同期複製リースを確立する（708）ためにそれを使用する。係るようにして、同期複製リースは、別のストレージシステムから受け取られたローカルクロックの値に従って計算されてよい。

【0191】

[0249] 第1のストレージシステム（714）のストレージコントローラが、第2のストレージシステム（724）のストレージコントローラと通信している例を考える。係る例では、第1のストレージシステム（714）のストレージコントローラのための単調増加クロックの値は1000ミリ秒であると仮定する。さらに、第1のストレージシステム（714）のストレージコントローラが、第2のストレージシステム（724）のストレージコントローラに、メッセージが生成されたときのそのクロック値が1000ミリ秒であったことを示すメッセージを送信すると仮定する。係る例では、第1のストレージシステム（714）のストレージコントローラが、第2のストレージシステム（724）のストレージコントローラに、メッセージが生成されたときのそのクロック値が1000ミリ秒であったことを示すメッセージを送信した後の500ミリ秒、第1のストレージシステム（714）のストレージコントローラは、第2のストレージシステム（724）のストレージコントローラから、1）第2のストレージシステム（724）のストレージコントローラの単調増加クロックの値が、メッセージが生成されたとき5000ミリ秒の値であったこと、及び2）第2のストレージシステム（724）によって受け取られた第1のストレージシステム（714）のストレージコントローラの単調増加するクロックの最後の値が1000ミリ秒だったことを示すメッセージを受け取ると仮定する。係る例では、同意されたリース間隔が2000ミリ秒である場合、第1のストレージシステム（714）は、第1のストレージシステム（714）のストレージコントローラのための単調増加クロックが3000ミリ秒の値となるまで有効である同期複製リースを確立する（708）。第1のストレージシステム（714）のストレージコントローラが、第1のストレージシステム（714）のストレージコントローラのための単調増加クロックが3000ミリ秒の値に達するときまでに、第1のストレージシステム（714）のストレージコントローラのための単調増加クロックの更新された値を含むメッセージを第2のストレージシステム（724）のストレージコントローラから受け取らない場合、第1のストレージシステム（714）は、同期複製リースを失効したとして扱い、以下により詳細に説明される多様な処置を講じてよい。読者は、ポッド内の残りのストレージシステム（724、728）の中のストレージコントローラが同様に反応し、同期複製リースの同様な追跡及び更新を実行することを理解する。本来、受取り側コントローラは、ネットワーク及び対にされたコントローラが、その時間間隔中にどこかで実行していたことを保証される場合があり、対にされたコントローラが、それがその時間間隔中のどこかで送信したメッセージを受け取ったことが保証される場合がある。クロックのいずれの調整もなしでは、受取り側コントローラは、その時間間隔のどこで、ネットワーク及び対にされたコントローラが実行していたのかを正確に知ることができず、そのクロック値を送信する際に、又はそのクロック値をまた受け取る際に待ち行列遅延があったかどうかを本当に知ることができない。

【0192】

[0250] 一次コントローラがそのクラスタ通信の一部としてクロックを交換している、それぞれが簡略な一次コントローラを有する2つのストレージシステムから成るポッドでは、各一次コントローラは、それが、いつ、対にされたコントローラが実行していたことを確実に知らないのかに関して制限をかけるために活動リースを使用できる。一次コントローラが不確かになった点で（コントローラの接続の活動リースが失効したとき）、一次コントローラは、それが不確かであること、及び活動リースを再開できるようになる前に適切に同期された接続が確立し直されなければならないことを示すメッセージを送信できる。ネットワークが一方の方向で機能しているが、他方の方向では適切に機能していない場合、これらのメッセージが受け取られてよく、応答は受け取られない場合がある。これは、それ自体の活動リースが失われたメッセージ及び待ち行列遅延の異なる組合せのためにまだ失効していない場合があるため、接続が正常に実行していない旨の、実行中の対に

されたコントローラによる第1の表示であってよい。結果として、係るメッセージが受け取られる場合、それは、独自の活動リースも失効すると見なすべきであり、それは独自のメッセージを送信し始め、接続の同期及び活動リースの再開を調整しようと試みるべきである。それが起こり、クロック交換の新しい集合が成功するまで、どのコントローラもその活動リースが有効であると見なすことはできない。

【0193】

【0251】 このモデルでは、コントローラは、それが再確立メッセージの送信を開始した後リース間隔秒、待機することができ、コントローラは応答を受け取っていない場合、対にされたコントローラがダウンしている、又は対にされたコントローラ自体の接続のためのリースが失効しているかのどちらかであることを保証される場合がある。クロックドリフトの少量のクロックドリフトを取り扱うために、コントローラはリース間隔（つまり、再確立リース）よりもわずかに長く待機する場合がある。コントローラが再確立メッセージを受け取ると、コントローラは、（コントローラは、コントローラの活動リースを送信することが失効したことを知っているの）待機するよりむしろ、再確立リースをただちに失効されると見なすであろうが、メッセージ喪失が、例えば混雑したネットワークスイッチにより引き起こされる一時的な状態であった場合に、多くの場合断念する前に追加のメッセージングを試すことに意味がある。

【0194】

【0252】 代替実施形態では、同期複製リースを確立することに加えて、クラスタメンバーシップリースも、対にされたストレージシステムからのクロック値の受取り時に、又は対にされたストレージシステムと交換されたクロックの受取り時に確立されてよい。係る例では、各ストレージシステムは、あらゆる対にされたストレージシステムとの独自の同期複製リースを及び独自のクラスタメンバーシップリースを有してよい。任意の対との同期複製リースの満了は、処理の休止につながる場合がある。しかしながら、クラスタメンバーシップは、クラスタメンバーシップリースがすべての対と失効するまで再計算することはできない。したがって、クラスタメンバーシップリースの持続時間は、対とのクラスタメンバーシップリースが、そのリンクのための対の同期複製リンクが失効した後まで失効しないことを保証するために、メッセージ及びクロック値の相互作用に基づいて設定されるべきである。読者は、クラスタメンバーシップリースが、ポッド内の各ストレージシステムによって確立することができ、ポッドのメンバーである任意の2つのストレージシステム間の通信リンクと関連付けられてよいことを理解する。さらに、クラスタメンバーシップリースは、同期複製リースの満了後、少なくとも同期複製リースの満了のための期間と同程度に長い期間、延長してよい。クラスタメンバーシップリースは、クロック交換の一部として対にされたストレージシステムから受け取られたクロック値の受取り時に延長されてよく、現在のクロック値からのクラスタメンバーシップリース期間は、少なくとも、交換されたクロック値に基づいて最後の同期複製リース延長のために確立された期間と同程度長くてよい。追加の実施形態では、セッションが最初にいつ交渉されるのかを含んだ追加のクラスタメンバーシップ情報が、接続を介して交換される場合がある。読者は、クラスタメンバーシップリースを利用する実施形態では、各ストレージシステム（又はストレージコントローラ）が、クラスタメンバーシップリースのための独自の値を有してよいことを理解する。係るリースは、クラスタリース満了が、例えばメディエータの競争を通して等、新しいメンバーシップを確立することを可能にし、同期複製リース満了が、新しい要求の処理を強制的に休止させることを所与として、すべてのポッドメンバー全体でのすべての同期複製リースが失効することが保証されるまで失効するべきではない。係る例では、休止は、クラスタメンバーシップ処置を講じることができる前にどこでも実施されることが保証されなければならない。

【0195】

【0253】 読者は、ストレージシステム（714）の1つだけが、データセット（712）のために、データセット（712）がその全体で同期複製される複数のストレージシステム（714、724、728）を識別すること（702）、データセット（712）を

10

20

30

40

50

同期複製するために使用される複数のストレージシステム（７１４、７２４、７２８）のそれぞれの間で１つ以上のデータ通信リンク（７１６、７１８、７２０）を構成すること（７０４）、複数のストレージシステム（７１４、７２４、７２８）の間で複数のストレージシステム（７１４、７２４、７２８）のうちの少なくとも１つのためのタイミング情報（７１０、７２２、７２６）を交換すること（７０６）、及び複数のストレージシステム（７１４、７２４、７２８）のうちの少なくとも１つのためのタイミング情報（７１０、７２２、７２６）に従って、同期複製リースを確立すること（７０８）として示されているが、残りのストレージシステム（７２４、７２８）も係るステップを実施してよいことを理解する。実際には、２つ以上のストレージシステム（７１４、７２４、７２８）の間で同期複製関係を確立することは、２つ以上のストレージシステム（７１４、７２４、

10

【０１９６】

【０２５４】 追加の説明のため、図８は、本開示のいくつかの実施形態に係る２つ以上のストレージシステム（７１４、７２４、７２８）の間で同期複製関係を確立する追加の例の方法を示すフローチャートを説明する。図８に示される例の方法は、データセット（７１２）のために、データセット（７１２）がその全体で同期複製される複数のストレージシステム（７１４、７２４、７２８）を識別すること（７０２）と、データセット（７１２）を同期複製するために使用される複数のストレージシステム（７１４、７２４、７２８）のそれぞれの間で１つ以上のデータ通信リンク（７１６、７１８、７２０）を構成すること（７０４）と、複数のストレージシステム（７１４、７２４、７２８）の間で、複数のストレージシステム（７１４、７２４、７２８）のうちの少なくとも１つのためにタイミング情報（７１０、７２２、７２６）を交換すること（７０６）と、複数のストレージシステム（７１４、７２４、７２８）のうちの少なくとも１つのためのタイミング情報（７１０、７２２、７２６）に従って、同期複製リースを確立することであって、同期複製リースが、同期複製関係が有効である期間を識別する、確立することも含むので図８に示される例の方法は、図４６に示される例の方法に類似する。

20

【０１９７】

【０２５５】 図８に示される例の方法では、複数のストレージシステム（７１４、７２４、７２８）の少なくとも１つのためのタイミング情報（７１０、７２２、７２６）に従って、同期複製リースを確立すること（７０８）は、複数のストレージシステム（７１４、７２４、７２８）の間のクロックを調整すること（８０２）を含む場合がある。図８に示される例の方法では、複数のストレージシステム（７１４、７２４、７２８）間でクロックを調整すること（８０２）は、例えばストレージシステム（７１４、７２４、７２８）間で送信される１つ以上のメッセージ交換により実施されてよい。ストレージシステム（７１４、７２４、７２８）間で送信される１つ以上のメッセージは、例えばそのクロック値が、他のすべてのストレージシステムによって使用されるストレージシステムのクロック値、全ストレージシステムに対するそのクロック値を所定の値に設定するための命令、そのクロック値を更新した旨のストレージシステムからの確認メッセージ等の情報を含んでよい。係る例では、ストレージシステム（７１４、７２４、７２８）は、特定のストレージシステム（例えば、リーダーストレージシステム）のためのクロック値が、他のすべてのストレージシステムによって使用されるべきであり、いくつかの特定の基準（例えば、最高クロック値）を満たすストレージシステムのすべてのからのクロック値は、他のすべてのストレージシステム等、使用されるべきであるように構成されてよい。係る例では、メッセージの交換と関連付けられた送信時間を説明するために、なんらかの所定量の時間が、別のストレージシステムから受け取られるクロック値に加算されてよい。

30

40

【０１９８】

【０２５６】 図８に示される例の方法では、複数のストレージシステム（７１４、７２４、７２８）の少なくとも１つのためのタイミング情報（７１０、７２２、７２６）に従って

50

、同期複製リースを確立すること（708）は、複数のストレージシステム（714、724、728）の間で調整のついていないクロックを交換すること（804）を含む場合がある。複数のストレージシステム（714、724、728）の間で調整がついていないクロックを交換すること（804）は、例えば各ストレージシステム（714、724、728）のストレージコントローラが、上記により詳細に説明されるように、ローカルの単調増加クロックの値を交換することによって実施されてよい。係る例では、各ストレージシステム（714、724、728）は、同意された同期複製リース間隔、及び同期複製リースを確立する（708）ために他のストレージシステム（714、724、728）から受け取られるメッセージングを活用してよい。

【0199】

【0257】 また、図8に示される例の方法は、同期複製リースが失効した後に受け取られたI/O要求の処理を先延ばしにすること（806）も含む。同期複製リースが失効した後にストレージシステムのいずれかによって受け取られたI/O要求は、例えば、新しい同期複製リースが確立されるまで等、同期複製関係を再確立しようと試みるためには十分である所定の時間量、先延ばしにされてよい。係る例では、ストレージシステムは、なんらかのタイプの『ビジー』若しくは一時故障表示で故障することで、又は他のなんらかの方法でI/O要求の処理を先延ばししてよい（806）。

【0200】

【0258】 追加の説明のために、図9は、本開示のいくつかの実施形態に従って2つ以上のストレージシステム（714、724、728）の間の同期複製間関係を確立する追加の例の方法を示すフローチャートを説明する。図9に示される例の方法は、データセット（712）のために、データセット（712）がその全体で同期複製される複数のストレージシステム（714、724、728）を識別すること（702）と、データセット（712）を同期複製するために使用される複数のストレージシステム（714、724、728）のそれぞれの間で1つ以上のデータ通信リンク（716a、716b、718a、718b、720a、720b）を構成すること（704）と、複数のストレージシステム（714、724、728）の間で複数のストレージシステム（714、724、728）のうちの少なくとも1つのためのタイミング情報（710、722、726）を交換すること（706）と、複数のストレージシステム（714、724、728）のうちの少なくとも1つのためのタイミング情報（710、722、726）に従って、同期複製リースを確立すること（708）とであって、同期複製リースが、同期複製関係が有効である期間を識別する、確立することも含むので、図9に示される例の方法は図46に示される例の方法に類似する。

【0201】

【0259】 図9に示される例の方法では、データセット（712）を同期複製するために使用される複数のストレージシステム（714、724、728）のそれぞれの間で1つ以上のデータ通信リンク（716a、716b、718a、718b、720a、720b）を構成すること（704）は、複数のデータ通信タイプのそれぞれのために、データセット（712）を同期複製するために使用される複数のストレージシステム（714、724、728）のそれぞれの間でデータ通信リンク（716a、716b、718a、718b、720a、720b）を構成すること（902）を含む場合がある。図9に示される例の方法では、各ストレージシステムは、ストレージシステムがポッド内の他のストレージシステムに送信する複数のデータ通信タイプを生成するように構成されてよい。例えば、ストレージシステムは、I/O処理の一部であるデータ（例えば、ホストによって発行される書き込み要求の一分としてストレージシステムに書き込まれるデータ）を含む第1のタイプのデータ通信を生成してよく、構成変更（例えば、ボリュームを作成、拡張、削除、又は名前を変えることに応えて生成される情報）を含む第2のタイプのデータ通信を生成するように構成されてよく、ストレージシステムは、ストレージシステム及び相互接続障害を検出し、処理することに関わる通信を含む第3のタイプのデータ通信を生成するように構成されてよい。係る例では、データ通信タイプは、例えば、どのソフトウェ

10

20

30

40

50

アモジュールがメッセージを開始したのかに基づいて、どのハードウェア構成要素がメッセージを開始したのかに基づいて、メッセージを開始させたイベントのタイプに基づいて、及び他の方法で決定されてよい。図9に示される例の方法では、複数のデータ通信タイプのそれぞれのために複数のストレージシステム（714、724、728）のそれぞれの間でデータ通信リンク（716a、716b、718a、718b、720a、720b）を構成すること（902）は、例えば複数のデータ通信タイプのそれぞれに別個の相互接続を使用するようにストレージシステムを構成することによって、複数のデータ通信タイプのそれぞれのために別個のネットワークを使用するようにストレージシステムを構成することによって、又は他の方法で実施されてよい。

【0202】

【0260】 また、図9に示される例の方法は、同期複製リースが失効したことを検出すること（904）も含む。図9に示される例の方法では、同期複製リースが失効したことを検出すること（904）は、例えば特定のストレージシステムが現在のクロック値をリースが有効であった期間に比較することによって実施されてよい。ストレージシステム（714、724、728）が、各ストレージシステム（714、724、728）の中のクロックの値を5000ミリ秒の値に設定するためにクロックを調整し、各ストレージシステム（714、724、728）がそのクロック値を超えて2000ミリ秒のリース間隔の間延長した同期複製リースを確立する（708）ように構成され、これにより各ストレージシステム（714、724、728）の同期複製リースは、特定のストレージシステム（714、724、728）の中のクロックが、7000ミリ秒を超える値に達したときに失効した。係る例では、同期複製リースが失効したことを検出すること（904）は、特定のストレージシステム（714、724、728）の中のクロックが7001ミリ秒以上の値に達したと判断することによって実施されてよい。

【0203】

【0261】 読者は、他のイベントの発生も、各ストレージシステム（714、724、728）に、同期複製リースを失効しているとしてただちに扱わせてよいと理解する。例えば、ストレージシステム（714、724、728）は、ストレージシステム（714、724、728）とポッド内の別のストレージシステム（714、724、728）との間の通信故障を検出すると、同期複製リースを失効しているとしてただちに扱ってよく、ストレージシステム（714、724、728）は、ポッド内の別のストレージシステム（714、724、728）からリース再確立メッセージを受け取ると、同期複製リースを失効しているとして直ちに扱ってよく、ストレージシステム（714、724、728）は、ポッド内の別のストレージシステム（714、724、728）が故障したことを検出すると、同期複製リースを失効しているとしてただちに扱ってよい等々である。係る例では、上述の文章に説明されるイベントのいずれかの発生は、ストレージシステムに、同期複製リースが失効したことを検出させて（904）よい。

【0204】

【0262】 また、図9に示される例の方法は、同期複製関係を再確立すること（906）も含む。図9に示される例の方法では、同期複製関係を再確立すること（906）は、例えば1つ以上の再確立メッセージの使用により実施されてよい。係る再確立メッセージは、例えば、同期複製関係が再確立されるポッドの識別、1つ以上のデータ通信リンクを構成するために必要とされる情報、更新されたタイミング情報等を含む場合がある。係るようにして、ストレージシステム（714、724、728）は、各ストレージシステムが、データセット（712）のために、データセット（712）がその全体で同期複製される複数のストレージシステム（714、724、728）を識別すること（702）と、データセット（712）を同期複製するために使用される複数のストレージシステム（714、724、728）のそれぞれの間で1つ以上のデータ通信リンク（716a、716b、718a、718b、720a、720b）を構成すること（704）と、複数のストレージシステム（714、724、728）の間で複数のストレージシステム（714、724、728）のうちの少なくとも1つのためのタイミング情報（710、722

、726)を交換すること(706)と、複数のストレージシステム(714、724、728)のうちの少なくとも1つのためのタイミング情報(710、722、726)に従って、同期複製リースを確立すること(708)とであって、同期複製リースが、同期複製関係が有効である期間を識別する、確立することのうちの1つ以上を実行することを含むが、これに限定されるものではない、同期複製関係が最初に作成されたのとはほぼ同じ方法で同期複製関係を再確立してよい(906)。

【0205】

[0263] 図9に示される例の方法では、同期複製リースの満了の後に、新しい活動イベントが後に続く又はなんらかの他のアクションが後に続く再確立メッセージが後に続く、イベントのなんらかの集合が続いてよい。データ通信、構成通信、又は他の通信は、遷移中である可能性がある。一方、同期複製リースは失効し、再確立される。実際には、通信は、例えば新しい同期複製リースが確立された後まで受け取られられない場合がある。係る場合、通信は、ポッド、クラスタ、又はネットワークリンクの受胎のある理解に基づいて送信された可能性があり、現在ではその状態のある態様又は別の態様の異なる理解を有するストレージシステム(714、724、728)によって受け取られてよい。したがって、概して、通信がクラスタのなんらかの集合の前に送信された、又はリンク状態が変化する場合、受け取られた通信が廃棄されることを確実にするなんらかの手段があるべきである。通信がクラスタのなんらかの集合の前に送信された、又はリンク状態が変化する場合、受け取られた通信が廃棄されることを確実にする1つの方法は、延長されている作業中の同期複製リースとのリンクを確立又は再確立することに関連付けられるなんらかのセッション識別子(例えば、番号)を確立することである。クラスタ通信が再確立された後、リンクは新しいセッション識別子を得る。この識別子は、データメッセージ、構成メッセージ、又は他の通信メッセージとともに含まれる場合がある。間違ったセッション識別子とともに受け取られる任意のメッセージは、廃棄される、又はセッション識別子不一致を示すエラー応答を生じさせる。

【0206】

[0264] 読者は、ストレージシステム(714、724、728)が、同期複製リースの再確立にどのように対応するのかが、ストレージシステム及びポッドがとる異なる実施形態に基づいて変化する場合があることを理解する。2つのストレージシステムを有する簡略な一次コントローラの場合、コントローラの同期複製リースを受け取ることが失効した後に受け取られる、ストレージシステムで動作(読取り、書込み、ファイル動作、オブジェクト動作、管理動作等)を実行する新しい要求は、その処理を先延ばしさせる、打ち切らせる、又はなんらかの「後で再試行」エラーコードで故障させる場合がある。したがって、実行中の一次ストレージコントローラは、対にされたストレージコントローラの同期複製リースが失効している場合新しい要求を処理していないことが保証される場合があり、それは、いつ独自の再確立リースが失効したのかを確信する場合がある。再確立リースが失効した後、コントローラが、対にされたコントローラをオフラインであると見なし、次いで対にされたコントローラなしでストレージ処理を続行することを含んだ補正処置を検討することが安全である。正確にそれらがどのようなアクションである可能性があるのかは、多種多様な検討及び実施態様の詳細に基づいて異なる場合がある。

【0207】

[0265] 一次コントローラ及び二次コントローラを有するストレージシステムの場合、あるストレージシステムで依然として実行中の一次コントローラは、対にされたストレージシステムの以前の第2のコントローラが引き継いでいる可能性があるとは仮定して、対にされたストレージシステムの以前の第2のコントローラに接続を試みる可能性がある。あるいは、あるストレージシステム上で依然として実行中の一次コントローラは、おそらく最大二次引継ぎ時間であるある特定の時間量待機する可能性がある。二次コントローラが接続し、適当な時間内に新しい同期複製リースとの新しい接続を確立する場合、ポッドは次いで(以下に説明される)一貫性のある状態にそれ自体を回復させ、次いで正常に続行してよい。対にされた二次コントローラが十分に迅速に接続しない場合、次いで依然とし

て実行中の一次コントローラは、例えば、依然として実行中の一次コントローラが対にされたストレージシステムを故障していると思われ、次いで対にされたストレージシステムなしで動作し続けるべきであるかどうかを判断しようと試みる等の追加の処置を講じる場合がある。一次コントローラは、代わりにポッドの中の対にされたストレージシステムで二次コントローラに対するアクティブなリースの接続を保つ可能性がある。その場合、一次対一次の再確立リースの満了は、代わりに、最初にその接続を確立することを必要とするよりむしろ、生き残った一次が、二次の引継ぎを問い合わせるためにその接続を使用することになる可能性がある。2つの一次ストレージコントローラが、ネットワークがそれらの間で機能していない間に実行中である可能性もあるが、ネットワークは一方又は他方の一次コントローラと対にされた二次コントローラとの間で機能している。その場合、ストレージシステムの中での高可用性モニタリングが、一次コントローラから二次コントローラへのフェイルオーバをトリガする条件を自力で検出しない可能性がある。その条件に対する対応は、同期複製を単に再開するために一次から二次へのフェイルオーバをいずれにせよトリガすること、二次を通る一次からの通信トラフィックの経路を定めること、又はまさにあたかも2つのストレージシステム間で通信が完全に故障し、それが発生したのと同じ障害処理を生じさせるかのように動作することを含む。

【0208】

【0266】 複数のコントローラが、（デュアルアクティブ-アクティブコントローラストレージシステムで及びスケールアウトストレージシステムでの両方を含む）ポッドのためにアクティブである場合も、リースは、対にされたストレージシステムの任意の又はすべてのコントローラとの個々のコントローラクラスタ通信によって保たれる可能性がある。この場合、失効した同期複製リースは、ストレージシステム全体でのポッドのための新しい要求処理の休止につながる可能性がある。リースモデルは、それらのクロックの、対にされたストレージシステムでのあらゆる対にされたコントローラとの追加の交換により、ストレージシステム内のすべてのアクティブなコントローラ間でのクロック及び対にされたクロック応答の交換で拡張される場合がある。特定のローカルコントローラのクロックが任意の対にされたコントローラと交換される動作経路がある場合、次いでコントローラは、独立した同期複製リースのために、及びおそらく独立した再確立リースのためにその経路を使用できる。この場合、ストレージシステムの中のローカルコントローラは、互いとの間のローカルリースのために互いの間でクロックをさらに交換してよい。これは、すでにローカルストレージシステムの高可用性機構及びモニタリング機構に組み込まれてよいが、ストレージシステムの高可用性機構に関係するあらゆるタイミングは、活動リース及び再確立リースの持続期間中、又は再確立リース満了と、相互接続障害を処理するために講じられる処置との間の任意の追加の遅延で考慮に入れられるべきである。

【0209】

【0267】 代わりに、ストレージシステム対ストレージシステムのクラスタ通信又はリースプロトコル単独が、少なくとも特定のポッドに対して、個々のマルチコントローラ又はスケールアウトシステムの中で一度に1つの一次コントローラに割り当てられてよい。このサービスは、障害の結果として、又はおそらく負荷不均衡の結果としてコントローラからコントローラに移行してよい。あるいは、クラスタ通信又はリースプロトコルは、クロック交換又は障害状況を分析することの複雑度を制限するためにコントローラの部分集合（例えば、2つ）で実行する可能性がある。各ローカルコントローラは、個々のコントローラが処理休止を達成したことを確実にできるときの潜在的なカスケード遅延を説明するために、ストレージシステム対ストレージシステムのリース、及びリース満了が相応して調整されなければならない後に応答するための時間を処理するコントローラの中のクロックを交換する必要がある場合がある。処理休止に関係するリースについて現在当てにされていない接続は、警告するために依然としてモニタされる可能性がある。

【0210】

【0268】 また、図9に示される例の方法は、データセットのためにI/O処理を引き継ぐと試みること（908）も含む。図9に示される例の方法では、データセット（71

10

20

30

40

50

2) のために I/O 処理を引き継ごうと試みること (908) は、例えばメディアータまで急ぐ (race to) ストレージシステム (714、724、728) によって実施されてよい。特定のストレージシステム (714、724、728) がデータセット (712) のために I/O 処理を無事に引き継ぐ場合、データセット (712) のすべてのアクセスは、同期複製関係が再確立できるまで特定のストレージシステム (714、724、728) によってサービスを提供され、以前の同期複製関係が失効した後に発生したデータセット (712) に対する任意の変更を、次いで他のストレージシステム (714、724、728) 上で転送し、持続する場合がある。係る例では、データセット (712) のために I/O 処理を引き継ごうとする試み (908) は、同期複製リースが失効した後のなんらかの期間の満了後にのみ発生し得る。例えば、(ストレージシステムの 1 つ以上が、データセットのために I/O 処理を引き継ごうと試みることを含む) リンク故障後にどのように始めるのかを決定しようとする試みは、同期複製リースが失効した後の期間まで、つまり、例えば少なくともクロック交換から生じる最大リース時間ほど長く開始しない場合がある。

【0211】

【0269】 読者は、上記に示された例の多くにおいて、ストレージシステム (714) のうちの 1 つしか、上述されたステップを実施するとして示されていないが、2 つ以上のストレージシステム間で同期複製関係を確立することは、2 つ以上のストレージシステムの間での協働及び相互作用を必要とする場合があるので、実際には、ポッド内の (又は形成中のポッド内の) すべてのストレージシステム (714、724、728) が、同時に上述されたステップのうちの 1 つ以上を実施し得ることを理解する。

【0212】

【0270】 追加の説明のために、図 10 は、本開示のいくつかの実施形態に係る 2 つ以上のストレージシステム (1024、1046) の間で同期複製関係を確立する追加の例の方法を示すフローチャートを説明する。図 10 に示される例の方法は、データセット (1022) が 2 つのストレージシステム (1024、1046) 全体だけで同期複製される実施形態を示すが、図 10 に示される例は、データセット (1022) が、2 つの示されるストレージシステム (1024、1046) によって実行されるステップに類似するステップを実行し得る追加のストレージシステム全体で同期複製される実施形態に拡張される場合がある。

【0213】

【0271】 図 10 に示される例の方法は、ストレージシステム (1024) によって、ストレージシステム (1024、1046) と第 2 のストレージシステム (1046) との間に 1 つ以上のデータ通信リンク (1052) を構成すること (1002) を含む。図 10 に示される例の方法では、ストレージシステム (1024) は、例えば、第 2 のストレージシステム (1046) とデータ通信を交換するために使用されるデータ通信ネットワークを介して定義されたポートを識別することによって、第 2 のストレージシステム (1046) とデータ通信を交換するために使用されるポイントツーポイントデータ通信リンクを識別することによって、又はさまざまな方法で、ストレージシステム (1024) と第 2 のストレージシステム (1046) との間に 1 つ以上のデータ通信リンク (1052) を構成してよい (1002)。安全な通信が必要とされる場合、なんらかの形の鍵交換が必要とされる場合もあれば、通信は、例えば SSH (Secure Shell)、SSL、又は公開鍵若しくはディフィーヘルマン鍵共用の回りで構築された他のなんらかのサービス又はプロトコル若しくは妥当な代替策等のなんらかのサービスを通して行われる又はブートされる場合もあるであろう。また、安全な通信は、なんらかの方法でカスタマイズアイデンティティと結び付けられるベンダ提供のなんらかのクラウドサービスを通して仲介される場合もあるであろう。代わりに、例えばバーチャルマシン又は仮想コンテナで実行する等、カスタマ施設で実行するように構成されたサービスは、複製するストレージシステム (1024、1046) 間の安全な通信のために必要な鍵交換を仲介するために使用できるであろう。読者は、3 つ以上のストレージシステムを含んだポッドが、個々の

ストレージシステムの大部分又はすべての間で通信リンクを必要とする場合があることを理解する。図10に示される例の方法では、第2のストレージシステム(1046)は、ストレージシステム(1024)と第2のストレージシステム(1046)との間で1つ以上のデータ通信リンク(1052)を同様に構成してよい(1026)。

【0214】

【0272】 また、図10に示される例の方法は、ストレージシステム(1024)から第2のストレージシステム(1046)に、ストレージシステム(1024)のためのタイミング情報(1048)を送信すること(1004)も含む。ストレージシステム(1024)のためのタイミング情報(1048)は、例えば、ストレージシステム(1024)の中のクロックの値として、クロック値(例えば、ストレージシステム(1024)が最初に記録できるシーケンス番号)の表現として、第2のストレージシステム(1046)の中でクロックのごく最近に受け取られた値として等、実施されてよい。図10に示される例の方法では、ストレージシステム(1024)は、例えば、2つのストレージシステム(1024、1046)間のデータ通信リンク(1052)を介してストレージシステム(1024)から第2のストレージシステム(1046)へ送信される1つ以上のメッセージを介して、ストレージシステム(1024)のためのタイミング情報(1048)を第2のストレージシステム(1046)に送信してよい(1004)。図10に示される例では、第2のストレージシステム(1046)は、第2のストレージシステム(1046)からストレージシステム(1024)に、第2のストレージシステム(1046)のためのタイミング情報(1050)を同様に送信してよい(1030)。

【0215】

【0273】 図10に示される例の方法では、ストレージシステム(1024)から第2のストレージシステム(1046)に、ストレージシステム(1024)のためのタイミング情報(1048)を送信すること(1004)は、ストレージシステム(1024)のクロックの値を送信すること(1006)を含む場合がある。図10に示される例では、ストレージシステム(1024)は、ストレージシステム(1024、1046)の間でクロックを調整する努力の一部として、ストレージシステム(1024)のクロックの値を第2のストレージシステム(1046)に送信してよい(1006)。係る例では、ストレージシステム(1024)は、その値が、2つのストレージシステム(1024、1046)間のデータ通信リンク(1052)を介して第2のストレージシステム(1046)に送信される1つ以上のメッセージを介して送信される(1006)ローカルの単調増加クロックを含んでよい。図10に示される例の方法では、第2のストレージシステム(1046)からストレージシステム(1024)に、第2のストレージシステム(1046)のためのタイミング情報(1050)を送信すること(1030)は、第2のストレージシステム(1046)のクロックの値を送信すること(1032)を同様に含んでよい。

【0216】

【0274】 図10に示される例の方法では、ストレージシステム(1024)から第2のストレージシステム(1046)に、ストレージシステム(1024)のためのタイミング情報(1048)を送信すること(1004)は、第2のストレージシステム(1046)のクロックのごく最近受け取られた値を送信すること(1008)を含む場合もある。図10に示される例の方法では、第2のストレージシステム(1046)のクロックのごく最近受け取られた値を送信すること(1008)は、例えばタイミング保証を達成しながらも、ストレージシステム(1024、1046)間でクロックを調整する必要性を排除するための努力の一部として実施されてよい。係る実施形態では、各ストレージシステム(1024、1046)は、ローカルの単調増加クロックを有してよい。同期複製リリースは、各ストレージシステム(1024、1048)が、それが他のストレージシステム(1024、1048)から受け取った最後のクロック値とともに、他のストレージシステム(1024、1048)にそのクロック値を送信することによってストレージシステム(1024、1048)間で確立されてよい。特定のストレージシステム(1024

、1048)は、別のストレージシステム(1024、1048)からクロック値をまた受け取るとき、特定のストレージシステム(1024、1048)は、受け取られたクロック値になんらかの同意されたリース間隔を加え、同期複製リースを確立するためにそれを使用してよい。図10に示される例の方法では、第2のストレージシステム(1046)からストレージシステム(1024)に、第2のストレージシステム(1046)のためのタイミング情報(1050)を送信すること(1030)は、ストレージシステム(1024)のクロックのごく最近受け取られた値を送信すること(1034)も同様に含んでよい。

【0217】

【0275】 また、図10に示される例の方法は、第2のストレージシステム(1046)からストレージシステム(1024)によって、第2のストレージシステム(1046)のためのタイミング情報(1050)を受け取ること(1010)も含む。図10に示される例の方法では、ストレージシステム(1024)は、2つのストレージアレイ(1024、1046)間のデータ通信リンク(1052)を介して第2のストレージシステム(1046)から送信される1つ以上のメッセージを介して第2のストレージシステム(1046)から、第2のストレージシステム(1046)のためのタイミング情報(1050)を受け取ってよい(1010)。図10に示される例では、第2のストレージシステム(1046)は、ストレージシステム(1024)から、ストレージシステム(1024)のためのタイミング情報を同様に受け取ってよい(1028)。

【0218】

【0276】 また、図10に示される例の方法は、第2のストレージシステム(1046)のためのタイミング情報(1050)に従って、ストレージシステム(1024)でクロック値を設定すること(1012)も含む。図10に示される例では、第2のストレージシステム(1046)のためのタイミング情報(1050)に従って、ストレージシステム(1024)でクロック値を設定すること(1012)は、例えば2つのストレージシステム(1024、1046)間でクロックを調整する努力の一部として実施されてよい。係る例では、2つのストレージシステム(1024、1046)は、例えばそれらのそれぞれのクロック値を、ストレージシステム(1024、1046)間の最高のクロック値よりも高いなんらかの所定量である値に設定するように、それらのそれぞれのクロック値を、ストレージシステム(1024、1046)の対の間の最高のクロック値に等しい値に設定するように、そのそれぞれのクロック値を、各ストレージシステム(1024、1046)のそれぞれのクロック値になんらかの関数を適用することによって生成される値に設定するように、又は他のなんらかの方法で構成されてよい。図10に示される例の方法では、第2のストレージシステム(1046)は、ストレージシステム(1024)のためのタイミング情報(1048)に従って、第2のストレージシステム(1046)のクロック値を同様に設定してよい(1036)。

【0219】

【0277】 また、図10に示される例の方法は、同期複製リースを確立すること(1014)も含む。図10に示される例の方法では、同期複製リースを確立すること(1014)は、例えば、2つのストレージシステム(1024、1046)間の同等のクロック値を超えるなんらかの所定のリース間隔の間延長する同期複製リースを確立することによって、ストレージシステム(1024、1046)のうちの1つと関連付けられた調整がついていないクロック値を超えるなんらかの所定のリース間隔の間延長する同期複製リースを確立することによって、又はなんらかの他の方法で実施されてよい。図19に示される例の方法では、第2のストレージシステム(1046)は、ストレージシステム(1024)のためのタイミング情報(1048)に従って、第2のストレージシステム(1046)のクロック値を同様に設定してよい(1036)。

【0220】

【0278】 また、図10に示される例の方法は、ストレージシステム(1024)によって、同期複製リースが失効したことを検出すること(1016)も含む。図10に示され

10

20

30

40

50

る例の方法では、同期複製リースが失効したことを検出すること（１０１６）は、例えば、ストレージシステム（１０２４）が現在のクロック値を、リースが有効であった期間に比較することによって実施されてよい。ストレージシステム（１０２４、１０４６）が、各ストレージシステム（１０２４、１０４６）の中のクロックの値を５０００ミリ秒の値に調整し、各ストレージシステム（１０２４、１０４６）が、そのクロック値を超える２０００ミリ秒のリース間隔の間延長した同期複製リースを確立する（１０３８）ように構成され、これにより各ストレージシステム（１０２４、１０４６）のための同期複製リースが、特定のストレージシステム（１０２４、１０４６）の中のクロックが１００００ミリ秒を超える値に達したときに失効した例を考える。係る例では、同期複製リースが失効したことを検出すること（１００６）は、ストレージシステム（１０２４）の中のクロックが１０００１ミリ秒以上の値に達したと判断することによって実施されてよい。図１０に示される例の方法では、第２のストレージシステム（１０４６）は、同期複製リースが失効したことを同様に検出してよい（１０４０）。

【０２２１】

【０２７９】 また、図１０に示される例の方法は、ストレージシステム（１０２４）によって、データセット（１０２２）のためにＩ／Ｏ処理を引き継ごうと試みること（１０２０）も含む。図１０に示される例の方法では、データセット（１０２２）のためにＩ／Ｏ処理を引き継ごうと試みること（１０２０）は、例えばメディアータまで急ぐストレージシステム（１０２４）によって実施されてよい。ストレージシステム（１０２４）がデータセット（１０２２）のためのＩ／Ｏ処理を無事に引き継ぐ場合、データセット（１０２２）のすべてのアクセスは、同期複製関係が再確立されるまでストレージシステム（１０２４）によってサービスを提供され、以前の同期複製関係が失効した後に発生したデータセット（１０２２）に対するあらゆる変更は、第２のストレージシステム（１０４６）に転送され、持続される場合がある。図１０に示される例の方法では、第２のストレージシステム（１０４６）は、同様にデータセット（１０２２）のためのＩ／Ｏ処理を引き継ごうと試みてよい（１０４４）。

【０２２２】

【０２８０】 また、図１０に示される例の方法は、ストレージシステム（１０２４）によって、同期複製関係を再確立しようとする試みること（１０１８）も含む。図１０に示される例の方法では、同期複製関係を再確立しようとする試みること（１０１８）は、例えば１つ以上の再確立メッセージを使用することによって実施されてよい。係る再確立メッセージは、例えば、同期複製関係が再確立されるポッドの識別、１つ以上のデータ通信リンクを構成するために必要とされる情報、更新されたタイミング情報等を含む場合がある。係るようにして、ストレージシステム（１０２４）は、同期複製関係が最初に作成されたのと同様と同じように同期複製関係を再確立してよい。図１０に示される例では、第２のストレージシステム（１０４６）は、同様に同期複製関係を再確立しようとする試みてよい（１０４２）。

【０２２３】

【０２８１】 追加の説明のために、図１１は、本開示のいくつかの実施形態に従って複数のストレージシステム（１１３８、１１４０）全体で同期されるデータセット（１１４２）に向けられたＩ／Ｏ動作にサービスを提供する例の方法を示すフローチャートを説明する。あまり詳細には示されないが、図１１に示されるストレージシステム（１１３８、１１４０）は、図１Ａから図１Ｄ、図２Ａから図２Ｇ、図３Ａから図３Ｂ、又はその任意の組合せに関して上述されたストレージシステムに類似してよい。実際には、図１１に示されるストレージシステムは、上述されたストレージシステムと同じ構成要素、より少ない構成要素、追加の構成要素を含んでよい。

【０２２４】

【０２８２】 図１１に示されるデータセット（１１４２）は、例えば特定のボリュームのコンテンツとして、ボリュームの特定の共用のコンテンツとして、又は１つ以上のデータ要素の任意の他の集合体として実施されてよい。データセット（１１４２）は、各ストレージシステム（１１３８、１１４０）がデータセットのローカルコピーを保持するように複

数のストレージシステム（１１３８、１１４０）全体で同期されてよい。本明細書に説明される例では、係るデータセット（１１４２）は、少なくともアクセスされているクラスタ及び特定のストレージシステムが名目上実行している限り、クラスタ内の任意の一方のストレージシステムが、クラスタ内の任意の他方のストレージシステムよりも実質的により最適に動作しないような性能特性を有するストレージシステム（１１３８、１１４０）のいずれかを通してデータセット（１１４２）にアクセスできるように、ストレージシステム（１１３８、１１４０）全体で同期複製される。係るシステムでは、データセット（１１４２）に対する修正は、任意のストレージシステム（１１３８、１１４０）のデータセット（１１４２）にアクセスすることが一貫した結果を生じさせるように、各ストレージシステム（１１３８、１１４０）に常駐するデータセットのコピーに対して加えられるべきである。例えばデータセットに対する書込み要求は、すべてのストレージシステム（１１３８、１１４０）でサービスを提供されなければならない、又は書込みの初めに名目上実行しており、書込みの完了まで名目上実行中のままであったストレージシステム（１１３８、１１４０）のいずれでもサービスを提供されてならない。同様に、動作のいくつかのグループ（例えば、データセットの中の同じ場所に向けられる２つの書込み動作）は、データセットがすべてのストレージシステム（１１３８、１１４０）で最終的に同一となるように、同じ順序で実行されなければならない、又は以下により詳細に説明されるように、すべてのストレージシステム（１１３８、１１４０）で他のステップが講じられなければならない。データセット（１１４２）に対する修正は、正確に同時に行われる必要はないが、いくつかのアクション（例えば、書込み要求がデータセットに向けられ、両方のストレージシステムでまだ完了していない書込み要求によってターゲットにされるデータセットの中の場所への読取りアクセスを可能にする旨の肯定応答を発行すること）は、各ストレージシステム（１１３８、１１４０）上のデータセットのコピーが修正されるまで先延ばしにされてよい。

【０２２５】

【０２８３】 図１１に示される例の方法では、あるストレージシステム（１１４０）の『リーダー』として及び別のストレージシステム（１１３８）の『フォロワー』の名称は、ストレージシステム全体で特定のデータセットを同期複製する目的のために各ストレージシステムのそれぞれの関係を指してよい。係る例では、及び以下により詳細に説明されるように、リーダーストレージシステム（１１４０）は、入信するＩ／Ｏ動作のなんらかの処理を実行し、フォロワーストレージシステム（１１３８）に係る情報を渡すこと、又はフォロワーストレージシステム（１１４０）に求められていない他のタスクを実行することを担ってよい。リーダーストレージシステム（１１４０）は、すべての入信Ｉ／Ｏ動作のためにフォロワーストレージシステム（１１３８）に求められていないタスクを実行することを担ってよい、又は代わりにリーダー－フォロワー関係は、どちらかのストレージシステムによって受け取られるＩ／Ｏ動作の部分集合だけに特有であってよい。例えば、リーダー－フォロワー関係は、第１のボリューム、ボリュームの第１のグループ、論理アドレスの第１のグループ、物理アドレスの第１のグループ、又は他のなんらかの論理デリネータ（*delineator*）又は物理デリネータに向けられるＩ／Ｏ動作に特有であってよい。このようにして、第１のストレージシステムは、ボリューム（又は他のデリネータ）の第１の集合に向けられたＩ／Ｏ動作のためのリーダーストレージシステムとしての機能を果たしてよい。一方、第２のストレージシステムは、ボリューム（又は他のデリネータ）の第２の集合に向けられたＩ／Ｏ動作のためのリーダーストレージシステムとしての機能を果たしてよい。図１１に示される例の方法は、以下により詳細に説明されるように、複数のストレージシステム（１１３８、１１４０）を同期させることが、フォロワーストレージシステム（１１３８）によるデータセット（１１４２）を修正する要求（１１０４）の受取りに応じて実施されてもよいが、複数のストレージシステム（１１３８、１１４０）を同期させることが、リーダーストレージシステム（１１４０）によるデータセット（１１４２）を修正する要求（１１０４）の受取りに応じて発生する実施例を示す。

【０２２６】

【0284】 図11に示される例の方法は、リーダストレージシステム(1140)によるデータセット(1142)を修正する要求(1104)を受け取ること(1106)を含む。データセット(1142)を修正する要求(1104)は、例えばデータセット(1142)に含まれるデータを含むストレージシステム(1140)の中の場所にデータを書き込む要求として、データセット(1142)に含まれるデータを含むボリュームにデータを書き込む要求として、データセット(1142)のスナップショットを撮影する要求として、基本的に仮想範囲コピーとして、データセット(1142)内のデータのなんらかの部分の削除を表すUNMAP動作として、(データセットの中のデータの一部分に対する変更よりむしろ)データセット(1142)の変換を修正するとして、又はデータセット(1142)に含まれるデータのなんらかの部分に対する変更を生じさせるなんらかの他の動作として実施されてよい。図11に示される例の方法では、データセット(1142)を修正する要求(1104)は、例えばバーチャルマシンで実行中であるアプリケーションとして、ストレージシステム(1140)に接続されるコンピューティングデバイスで実行中であるアプリケーションとして、ストレージシステム(1140)にアクセスするように構成されたなんらかの他のエンティティとして実施されてよいホスト(1102)によって発行される。

【0227】

【0285】 また、図11に示される例の方法は、リーダストレージシステム(1140)によって、データセット(1142)に対する修正を記述する情報(1110)を生成すること(1108)も含む。リーダストレージシステム(1140)は、例えば、順序付け対進行中である任意の他の動作を決定することによって、重複する修正の適切な結末(例えば、同じ記憶場所を修正する2つの要求の適切な結末)を決定し、ポッドのすべてのメンバー(例えば、データセットが同期複製されるすべてのストレージシステム)全体で、例えばメタデータの共通要素に対するあらゆる分散状態の変更を計算することによって等、データセット(1142)に対する修正を記述する情報(1110)を生成してよい(1108)。データセット(1142)に対する修正を記述する情報(1110)は、例えば、ストレージシステムにおいて実行されるI/O動作を記述するために使用されるシステムレベル情報として実施されてよい。リーダストレージシステム(1140)は、データセット(1142)を修正する要求(1104)にサービスを提供するために、起こるべきことを理解するのに十分なだけ、データセット(1142)を修正する要求(1104)を処理することによってデータセット(1142)に対する修正を記述する情報(1110)を生成してよい(1108)。例えば、リーダストレージシステム(1140)は、データセット(1142)を修正する他の要求に対する、データセット(1142)を修正する要求(1104)の実行のなんらかの順序付けが必要とされるかどうかを判断する場合がある、又は各ストレージシステム(1138、1140)で同等な結果を生じさせるために、以下により詳細に説明されるように、なんらかの他のステップが講じられなければならない。

【0228】

【0286】 データセット(1142)を修正する要求(1104)が、データセット(1142)内の第1のアドレス範囲からデータセット(1142)内の第2のアドレス範囲にブロックをコピーする要求として実施される例を考える。係る例では、3つの他の書込み動作(書込みA、書込みB、書込みC)は、データセット(1142)の第1のアドレス範囲に向けられると仮定する。係る例では、リーダストレージシステム(1140)は、データセット(1142)内の第1のアドレス範囲からデータセット(1142)内の第2のアドレス範囲にブロックをコピーする前に、書込みA及び書込みBにサービスを提供する(が、書込みCにはサービスを提供しない)場合、フォロワストレージシステム(1138)も、一貫した結果を生じさせるために、データセット(1142)内の第1のアドレス範囲からデータセット(1142)内の第2のアドレス範囲にブロックをコピーする前に、書込みA及び書込みBにサービスを提供しなければならない(が、書込みCにはサービスを提供しない)。したがって、リーダストレージシステム(1140)

が、データセット（１１４２）に対する修正を記述する情報（１１１０）を生成する（１１０８）とき、この例では、リーダーストレージシステム（１１４０）は、フォロワーストレージシステム（１１３８）が、データセット（１１４２）を修正する要求（１１０４）を処理できる前に完了されなければならない他の動作を識別する情報（例えば、書込みＡ及び書込みＢのためのシーケンス番号）を生成できるであろう。

【０２２９】

【０２８７】 ２つの要求（例えば、書込みＡ及び書込みＢ）が、データセット（１１４２）の重複部分に向けられる追加の例を考える。係る例では、フォロワーストレージシステム（１１３８）は書込みＢにサービスを提供し、後で書込みＡにサービスを提供するが、リーダーストレージシステム（１１４０）が書込みＡにサービスを提供し、後で書込みＢにサービスを提供する場合、データセット（１１４２）は、両方のストレージシステム（１１３８、１１４０）全体で一致しないであろう。したがって、リーダーストレージシステム（１１４０）がデータセット（１１４２）に対する修正を記述する情報（１１１０）を生成する（１１０８）とき、この例では、リーダーストレージシステム（１１４０）は、要求が実行されるべきである順序を識別する情報（例えば、書込みＡ及び書込みＢのシーケンス番号）を生成できるであろう。代わりに、リーダーストレージシステム（１１４０）は、各ストレージシステム（１１３８、１１４０）から中間の動作を必要とするデータセット（１１４２）に対する修正を記述する情報（１１１０）を生成するよりむしろ、２つの要求の適切な結末を識別する情報を含む、データセット（１１４２）に対する修正を記述する情報（１１１０）を生成してよい（１１０８）。例えば、書込みＢが書込みＡに論理的に続く（及び書込みＡと重複する）場合、最終結果は、データセット（１１４２）が、書込みＢと重複する書込みＡの部分を含むよりむしろ、書込みＡと重複する書込みＢの部分を含むことでなければならない。係る結末は、メモリで結果をマージし、特定のストレージシステム（１１３８、１１４１０）が書込みＡを実行し、次いでその後書込みＢを実行することを厳格に必要とするよりむしろ、データセット（１１４２）への係るマージの結果を書き込むことによって容易にされるであろう。読者は、より微妙なケースがスナップショット及び仮想アドレス範囲コピーに関係することを理解する。

【０２３０】

【０２８８】 読者は、任意の動作に対する正しい結果が、動作の受取り確認が可能である前に回復可能である点にコミットされなければならないことをさらに理解する。しかし、複数の動作をともにコミットできる場合もあれば、回復が正しさを確実にするのであるならば、動作を部分的にコミットできる場合もある。例えば、スナップショットは、予想される書込みＡ及び書込みＢに対する記録された依存関係とローカルにコミットできるであろうが、Ａ又はＢ自体はコミットされていない可能性がある。スナップショットの受取りを確認することができず、紛失したＩ／Ｏを別のアレイから回復できない場合、回復が結局スナップショットを取り消すことになる可能性がある。また、書込みＢが書込みＡと重複する場合、リーダーストレージシステムはＢにＡの後となるように『命令』してよいが、Ａは実際には廃棄され、次いでＡを書き込む動作は単にＢを待機するであろう。Ａ、ＢとＣ、Ｄとの間のスナップショットと結合された書込みＡ、書込みＢ、書込みＣ、及び書込みＤは、回復がアレイ全体でスナップショット不一致を生じさせることがない限り、及び肯定応答が、より早期の動作が、それが回復可能であることを保証される点まで持続される前に後の動作を完了しない限り、いくつかの又はすべての部分をともにコミットする及び／又は受取りを確認することができるであろう。

【０２３１】

【０２８９】 また、図１１に示される例の方法は、リーダーストレージシステム（１１４０）からフォロワーストレージシステム（１１３８）へ、データセット（１１４２）に対する修正を記述する情報（１１１０）を送信すること（１１１２）も含む。リーダーストレージシステム（１１４０）からフォロワーストレージシステム（１１３８）へ、データセット（１１４２）に対する修正を記述する情報（１１１０）を送信すること（１１１２）は、例えばリーダーストレージシステム（１１４０）が１つ以上のメッセージをフォロワ

ーストレージシステム（１１３８）に送信することによって実施されてよい。また、リーダーストレージシステム（１１４０）は、同じメッセージで又は１つ以上の異なるメッセージで、データセット（１１４２）を修正する要求（１１０４）に対するＩ／Ｏペイロード（１１１４）を送信してもよい。Ｉ／Ｏペイロード（１１１４）は、例えばデータセット（１１４２）を修正する要求（１１０４）が、データセット（１１４２）にデータを書き込む要求として実施されるとき、フォロワーストレージシステム（１１３８）の中のストレージに書き込まれるデータとして実施されてよい。係る例では、データセット（１１４２）を修正する要求（１１０４）が、リーダーストレージシステム（１１４０）によって受け取られた（１１０６）ため、フォロワーストレージシステム（１１３８）は、データセット（１１４２）を修正する要求（１１０４）と関連付けられたＩ／Ｏペイロード（１１１４）をまだ受け取っていない。図１１に示される例では、データセット（１１４２）に対する修正を記述する情報（１１１０）及びデータセット（１１４２）を修正する要求（１１０４）と関連付けられるＩ／Ｏペイロード（１１４２）は、リーダーストレージシステム（１１４０）をフォロワーストレージシステム（１１３８）に結合する１つ以上のデータ通信ネットワークを介して、リーダーストレージシステム（１１４０）をフォロワーストレージシステム（１１３８）に結合する１つ以上の専用データ通信リンク（例えば、Ｉ／Ｏペイロードを送信するための第１のリンク及びデータセットに対する修正を記述する情報を送信するための第２のリンク）を介して、又はなんらかの他の機構を介してリーダーストレージシステム（１１４０）からフォロワーストレージシステム（１１３８）に送信されてよい（１１１２）。

【０２３２】

【０２９０】 また、図１１に示される例の方法は、フォロワーストレージシステム（１１３８）によって、データセット（１１４２）に対する修正を記述する情報（１１１０）を受け取ること（１１１６）も含む。フォロワーストレージシステム（１１３８）は、例えばリーダーストレージシステム（１１４０）からフォロワーストレージシステム（１１３８）に送信される１つ以上のメッセージを介して、データセット（１１４２）に対する修正を記述する情報（１１１０）及びリーダーストレージシステム（１１４０）からのＩ／Ｏペイロード（１１１４）を受け取ってよい（１１１６）。１つ以上のメッセージは、２つのストレージシステム（１１３８、１１４０）間の１つ以上の専用データ通信リンクを介して、リーダーストレージシステム（１１４０）が、ＲＤＭＡもしくは類似する機構を使用し、フォロワーストレージシステム（１１３８）上の所定の記憶場所（例えば、待ち行列の場所）にメッセージを書き込むことによって、又は他の方法で、リーダーストレージシステム（１１４０）からフォロワーストレージシステム（１１３８）に送信されてよい。

【０２３３】

【０２９１】 一実施形態では、フォロワーストレージシステム（１１３８）は、通信機構としてＳＣＳＩ要求（送信者から受信者への書込み、又は受信者から送信者への読取り）の使用の使用により、リーダーストレージシステム（１１４０）からデータセット（１１４２）に対する修正を記述する情報（１１１０）及びＩ／Ｏペイロード（１１１４）を受け取ってよい（１１１６）。係る実施形態では、ＳＣＳＩ書込み要求は、（どのようなデータ及びメタデータであれ含む）送信されることが意図され、特別な疑似デバイスに又は特別に構成されたＳＣＩＣネットワークを介して、又は任意の他の同意されたアドレス指定機構を通して送達されてよい情報を符号化するために使用される。あるいは、代わりに、モデルは、特別なデバイス、特別に構成されたＳＣＳＩネットワーク、又は他の同意された機構も使用し、受信者から送信者へオープンＳＣＳＩ読取り要求の集合を発行できる。データ及びメタデータを含んだ符号化された情報は、これらのオープンＳＣＳＩ要求の１つ以上に対する応答として受信者に送達される。係るモデルは、多くの場合データセンタ間の「ダークファイバ」ストレージネットワークとして配備されるファイバチャネルＳＣＳＩネットワークを介して実装できる。また、係るモデルは、ホストからリモートアレイへのマルチパッシング及びバルクアレイ間（array-to-array）通信用に同じネットワーク回線の使用も可能にする。

【0234】

[0292] また、図11に示される例の方法は、フォロワーストレージシステム(1138)によって、データセット(1142)を修正する要求(1104)を処理すること(1118)も含む。図11に示される例の方法では、フォロワーストレージシステム(1138)は、リーダーストレージシステム(1140)から受け取られたI/Oペイロード(1114)だけではなく、データセット(1142)に対する修正を記述する情報(1110)にも従って、フォロワーストレージシステム(1138)に含まれる1つ以上のストレージデバイス(例えば、NVRAMデバイス、SSD、HDD)のコンテンツを修正することによって、データセット(1142)を修正する要求(1104)を処理してよい(1118)。データセット(1142)を修正する要求(1104)が、データセット(1142)に含まれるボリュームに向けられる書込み動作として実施され、データセット(1142)に対する修正を記述する情報(1110)が、書込み動作が、以前に発行された書込み動作が処理された後にのみ実行できることを示す例を考える。係る例では、データセット(1142)を修正する要求(1104)を処理すること(1118)は、フォロワーストレージシステム(1138)が最初に、以前に発行された書込み動作がフォロワーストレージシステム(1138)で処理されたことを検証し、その後書込み動作と関連付けられたI/Oペイロード(1114)をフォロワーストレージシステム(1138)に含まれる1つ以上のストレージデバイスに書き込むことによって実施されてよい。係る例では、データセット(1142)を修正する要求(1104)は、例えばI/Oペイロード(1114)がフォロワーストレージシステム(1138)の中の永続記憶装置にコミットされているとき、完了され、無事に処理されたと見なされてよい。

10

20

【0235】

[0293] また、図11に示される例の方法は、フォロワーストレージシステム(1138)によってリーダーストレージシステム(1140)に対して、データセット(1142)を修正する要求(1104)の完了の受取りを確認すること(1120)も含む。図11に示される例の方法では、フォロワーストレージシステム(1138)によってリーダーストレージシステム(1140)に対して、データセット(1142)を修正する要求(1104)の完了の受取りを確認すること(1120)は、フォロワーストレージシステム(1138)が、リーダーストレージシステム(1140)に肯定応答(1122)メッセージを送信することによって実施されてよい。係るメッセージは、例えばフォロワーストレージシステム(1138)によるデータセット(1142)を修正する要求(1104)の完了の受取りを確認する(1120)際に有用な任意の追加の情報だけではなく、完了されたデータセット(1142)を修正する特定の要求(1104)を識別する情報も含んでよい。図11に示される例の方法では、リーダーストレージシステム(1140)に対してデータセット(1142)を修正する要求(1104)の完了の受取りを確認すること(1120)は、フォロワーストレージシステム(1138)が、リーダーストレージシステム(1140)に肯定応答(1122)メッセージを発行することによって示される。

30

【0236】

[0294] また、図11に示される例の方法は、リーダーストレージシステム(1140)によって、データセット(1142)を修正する要求(1104)を処理すること(1124)も含む。図11に示される例の方法では、リーダーストレージシステム(1140)は、データセット(1142)を修正する要求(1104)の一部として受け取られたI/Oペイロード(1114)だけではなく、データセット(1142)に対する修正を記述する情報(1110)に従ってリーダーストレージシステム(1140)に含まれる1つ以上のストレージデバイス(例えば、NVRAMデバイス、SSD、HDD)のコンテンツも修正することによって、データセット(1142)を修正する要求(1104)を処理してよい(1124)。データセット(1142)を修正する要求(1104)が、データセット(1142)に含まれるボリュームに向けられる書込み動作として実施され、データセット(1142)に対する修正を記述する情報(1110)が、書込み動

40

50

作が、以前に発行された書込み動作が処理された後にのみ実行できることを示す例を考える。係る例では、データセット（１１４２）を修正する要求（１１０４）を処理すること（１１２４）は、リーダストレージシステム（１１４０）が最初に、以前に発行された書込み動作がリーダストレージシステム（１１４０）によって処理されたことを検証し、その後書込み動作と関連付けられたＩ／Ｏペイロード（１１１４）をリーダストレージシステム（１１４０）に含まれる１つ以上のストレージデバイスに書き込むことによって実施されてよい。係る例では、データセット（１１４２）を修正する要求（１１０４）は、例えばＩ／Ｏペイロード（１１１４）がリーダストレージシステム（１１４０）の中の永続記憶装置にコミットされているとき、完了され、無事に処理されたと見なされてよい。

10

【０２３７】

【０２９５】 また、図１１に示される例の方法は、フォロワーストレージシステム（１１３８）から、フォロワーストレージシステム（１１３８）がデータセット（１１３６）を修正する要求（１１０４）を処理した旨の表示を受け取ること（１１２６）も含む。この例では、フォロワーストレージシステム（１１３８）がデータセット（１１３６）を修正する要求（１１０４）を処理した旨の表示は、フォロワーストレージシステム（１１３８）からリーダストレージシステム（１１４０）に送信される肯定応答（１１２２）メッセージとして実施される。読者は、上述されたステップの多くが、特定の順序で発生するとして示され、説明されるが、特定の順序は実際には必要とされないことを理解する。実際には、フォロワーストレージシステム（１１３８）及びリーダストレージシステム（１１４０）は、独立したストレージシステムであるため、各ストレージシステムは、上述されたステップのいくつかを並行して実行してよい。例えば、フォロワーストレージシステム（１１３８）は、データセット（１１４２）に対する修正を記述する情報（１１１０）を受け取ってよい（１１１６）、データセット（１１４２）を修正する要求（１１０４）を処理してよい（１１１８）、又はリーダストレージシステム（１１４０）がデータセット（１１４２）を修正する要求（１１０４）を処理する（１１２４）前にデータセット（１１４２）を修正する要求（１１０４）の完了の受取りを確認してよい（１１２０）。代わりに、リーダストレージシステム（１１４０）は、フォロワーストレージシステム（１１３８）が、データセット（１１４２）に対する修正を記述する情報（１１１０）を受け取る（１１１６）前にデータセット（１１４２）を修正する要求（１１０４）を処理した（１１２４）、データセット（１１４２）を修正する要求（１１０４）を処理した（１１１８）、又はデータセット（１１４２）を修正する要求（１１０４）の完了の受取りを確認した（１１２０）可能性がある。

20

30

【０２３８】

【０２９６】 また、図１１に示される例の方法は、リーダストレージシステム（１１４０）によって、データセット（１１４２）を修正する要求（１１０４）の完了の受取りを確認すること（１１３４）も含む。図１１に示される例の方法では、データセット（１１４２）を修正する要求（１１０４）の完了の受取りを確認すること（１１３４）は、リーダストレージシステム（１１４０）からホスト（１１０２）に送信される１つ以上の肯定応答（１１３６）メッセージを使用することにより、又はなんらかの他の適切な機構を介して実施されてよい。図１１に示される例の方法では、リーダストレージシステム（１１４０）は、データセット（１１４２）を修正する要求（１１０４）が、データセット（１１４２）を修正する要求（１１０４）の完了の受取りを確認する（１１３４）前にフォロワーストレージシステム（１１３８）によって処理された（１１１８）かどうかを判断してよい（１１２８）。リーダストレージシステム（１１４０）は、例えば、リーダストレージシステム（１１４０）が、フォロワーストレージシステム（１１３８）から、肯定応答メッセージ又はデータセット（１１４２）を修正する要求（１１０４）がフォロワーストレージシステム（１１３８）によって処理された（１１１８）ことを示す他のメッセージを受け取ったかどうかを判断することによって、データセット（１１４２）を修正する要求（１１０４）が、フォロワーストレージシステム（１１３８）によって処理さ

40

50

れた(1118)かどうかを判断してよい(1128)。係る例では、リーダストレージシステム(1140)が、データセット(1142)を修正する要求(1104)がフォロワーストレージシステム(1138)によって処理(1118)され、リーダストレージシステム(1138)によっても処理された(1118)と肯定的に判断する(1130)場合、リーダストレージシステム(1140)は、データセット(1142)を修正する要求(1104)を開始したホスト(1102)に対して、データセット(1142)を修正する要求(1104)の完了の受取りを確認する(1134)ことによって進んでよい。しかしながら、リーダストレージシステム(1140)が、データセット(1142)を修正する要求(1104)が、フォロワーストレージシステム(1138)によってまだ処理(1118)されていない(1132)、又はリーダストレージシステム(1138)によって処理され(1124)ていないと判断する場合、リーダストレージシステム(1140)は、データセット(1142)を修正する要求(1104)が、データセット(1142)がその全体で同期複製されるすべてのストレージシステム(1138、1140)上で無事に処理されたときだけ、データセット(1142)を修正する要求(1104)の完了の受取りを確認してよい(1134)ので、リーダストレージシステム(1140)は、データセット(1142)を修正する要求(1104)を開始したホスト(1102)に対して、データセット(1142)を修正する要求(1104)の完了の受取りをまだ確認していない(1134)場合がある。

【0239】

【0297】 読者は、図11に示される例の方法では、リーダストレージシステム(1140)からフォロワーストレージシステム(1138)へ、データセット(1142)に対する修正を記述する情報(1110)を送信すること(1112)、及びフォロワーストレージシステム(1138)によってリーダストレージシステム(1140)に対して、データセット(1142)を修正する要求(1104)の完了の受取りを確認すること(1120)が、単一のラウンドトリップメッセージングを使用し、実施される場合がある。単一のラウンドトリップメッセージングは、例えばデータ相互接続としてファイバチャネルの使用により使用されてよい。通常、SCSIプロトコルが、ファイバチャネルとともに使用される。いくつかのより旧式の複製技術は、ファイバチャネルネットワークを介してSCSIトランザクションとして基本的にデータを複製するために構築され得るため、係る相互接続はデータセンタ間で一般的にセットアップされる。また、従来ファイバチャネルSCSIインフラストラクチャは、イーサネット及びTCP/IPに基づいたネットワークよりもより少ないオーバーヘッド及びより低い待ち時間を有していた。さらに、データセンタがファイバチャネルを使用し、ストレージアレイを遮断するために内部で接続されるとき、ファイバチャネルネットワークは、あるデータセンタのホストが、ローカルストレージアレイが故障すると、遠隔のデータセンタでのストレージアレイへのアクセスに切り替わることができるように、他のデータセンタまで伸ばされてもよい。

【0240】

【0298】 SCSIは、たとえそれが通常ブロック指向ボリュームにデータを記憶し、取り出すための(又はテープのための)ブロックストレージプロトコルとの使用のために設計されていても、一般的な通信機構として使用できるであろう。例えば、SCSI READ又はSCSI WRITEは、対にされたストレージシステムのストレージコントローラ間でメッセージデータを送達する又は取り出すために使用できるであろう。SCSI WRITEの典型的な実施態様は、2つのメッセージラウンドトリップを必要とする。つまり、SCSI開始プログラムは、SCSI WRITE動作を記述するSCSI CDBを送信し、SCSIターゲットはそのCDBを受け取り、SCSIターゲットは、SCSI開始プログラムに「受信準備完了」メッセージを送信する。SCSI開始プログラムは、次いでSCSIターゲットにデータを送信し、SCSI WRITEが完了すると、SCSIターゲットは、成功完了でSCSI開始プログラムに応答する。他方、SCSI READ要求は、1つのラウンドトリップしか必要としない。つまり、SCSI開始プログラムは、SCSI READ動作を記述するSCSI CDBを送信し、SCSIターゲ

ットはそのCDBを受け取り、データ、及び次いで成功完了とともに応答する。結果として、距離に関して、SCSI READは、SCSI WRITEとして距離関係の待ち時間の半分を生じさせる。このため、データ通信受信機がSCSI READ要求を使用してメッセージを受信する方が、メッセージの送信者がSCSI WRITE要求を使用してデータを送信するよりも速い場合がある。SCSI READを使用することは、単に、メッセージ送信者がSCSIターゲットとして動作し、メッセージ受信者がSCSI開始プログラムとして動作することを必要とする。メッセージ受信者は、任意のメッセージ送信者にいくつかのSCSI CDB READ要求を送信してよく、メッセージ送信者は、メッセージデータが利用可能であるとき未決のCDB READ要求の1つに応答するであろう。SCSIサブシステムは、READ要求が未決であるのが長すぎる（例えば、10秒）場合、タイムアウトする場合があるので、READ要求は、たとえ送信されるメッセージがなくても数秒以内に応答されるべきである。

【0241】

【0299】 SCSIテープ要求は、InterNational Committee on Information Technology StandrdsのT10 Technical CommitteeのSCSI Stream Commands規格に説明されるように、可変サイズのメッセージデータを返すためにより柔軟である場合がある可変応答データをサポートする。また、SCSI規格は、単一ラウンドトリップSCSI WRITEコマンドを可能にできるであろうSCSI WRITE要求のための即時モードもサポートする。読者は、以下に説明される実施形態の多くが単一ラウンドトリップメッセージングも活用することを理解する。

【0242】

【0300】 追加の説明のために、図12は、本開示のいくつかの実施形態に従って、複数のストレージシステム（1138、1140、1150）全体で同期されるデータセット（1142）に向けられたI/O動作にサービスを提供するための追加の例の方法を示すフローチャートを説明する。あまり詳細には示されないが、図11に示されるストレージシステム（1138、1140、1150）は、図1Aから図1D、図2Aから図2G、図3Aから図3B、又はその任意の組合せに関して上述されたストレージシステムに類似してよい。実際には、図11に示されるストレージシステムは、上述されたストレージシステムと同じ構成要素、より少ない構成要素、追加の構成要素を含んでよい。図12に示される例の方法は、図11に示される例の方法に類似する。図12に示される例の方法も、リーダーストレージシステム（1140）によるデータセット（1142）を修正する要求（1104）を受け取ること（1106）と、ストレージシステム（1140）によって、データセット（1142）に対する修正を記述する情報（1110）を生成すること（1108）と、リーダーストレージシステム（1140）からフォローストレージシステム（1138）へ、データセット（1142）に対する修正を記述する情報（1110）を送信すること（1112）と、フォローストレージシステム（1138）によって、データセット（1142）に対する修正を記述する情報（1110）を受け取ること（1116）と、フォローストレージシステム（1138）によって、データセット（1142）を修正する要求（1104）を処理すること（1118）と、フォローストレージシステム（1138）によってリーダーストレージシステム（1140）に対して、データセット（1142）を修正する要求（1104）の完了の受取りを確認すること（1120）と、リーダーストレージシステム（1140）によって、データセット（1142）を修正する要求（1104）を処理すること（1124）と、リーダーストレージシステム（1140）によって、データセット（1142）を修正する要求（1104）の完了の受取りを確認すること（1134）も含むので、図11に示される例の方法に類似する。

【0243】

【0301】 しかしながら、図12に示される例の方法は、データセット（1142）が、ストレージシステムの1つがリーダーストレージシステム（1140）であり、残りのス

トレージシステムがフォローストレージシステム（１１３８、１１５０）である、３つのストレージシステム全体で同期複製される実施形態を示すので、図１１に示される例の方法とは異なる。係る例では、追加のフォローストレージシステム（１１５０）は、リーダーストレージシステム（１１４０）から、データセット（１１４２）に対する修正を記述する情報（１１１０）を受け取る（１１４２）、データセット（１１４２）に対する修正を記述する情報（１１１０）に従ってデータセット（１１４２）を修正する要求（１１０４）を処理する（１１４２）、リーダーストレージシステム（１１４０）に対して、肯定応答（１１４８）メッセージ又は他の適切な機構等を使用することによりデータセット（１１４２）を修正する要求（１１０４）の完了の受取りを確認する（１１４６）ことができるので、図１１に示されたフォローストレージシステム（１１３８）と同じステップの多くを実施する。

10

【０２４４】

【０３０２】 図１２に示される例の方法では、データセット（１１４２）に対する修正を記述する情報（１１１０）は、データセット（１１４２）を修正する要求（１１０４）のための順序付け情報（１１５２）を含む場合がある。図１２に示される例の方法では、データセット（１１４２）を修正する要求（１１０４）のための順序付け情報（１１５２）は、動作（例えば、データセットを修正する要求）と、リーダーストレージシステム（１１４０）によって記述できる共通メタデータ更新との関係の記述を、データセットを修正する別々の要求の間、及びおそらくデータセットを修正する要求と多様なメタデータ変更との間の相互依存性の集合として表す場合がある。これらの相互依存性は、データセットを修正するある要求がなんらかの方法で依存するプレカーソル（precursors）の集合として、データセットを修正するその要求が完了するには真でなければならないプレディケートとして説明される場合がある。

20

【０２４５】

【０３０３】 待ち行列プレディケートは、データセットを修正するその要求が完了するには真でなければならないプレディケートの一例である。待ち行列プレディケートは、データセットを修正する以前の要求が完了するまで、データセットを修正する特定の要求が完了しない場合があることを明記する場合がある。待ち行列プレディケートは、例えば重複する書込みタイプ動作のために使用できる。係る例では、リーダーストレージシステム（１１４０）は、第２の書込みタイプ動作が論理的に第１の係る動作に従うことを宣言することができ、これにより第２の書込みタイプ動作は、第１の書込みタイプ動作が完了するまで完了できない。実施態様に応じて、第２の書込みタイプ動作は、第１の係る書込みタイプ動作が永続的であることが確実にされるまで永続的にもされ得ない（２つの動作はともに永続的にすることができる）。また、待ち行列プレディケートは、不完全なプレカーソル動作の既知の集合（例えば、書込み型動作の集合）がそれぞれ、スナップショットが完了できる前に完了しなければならないことを宣言することによって、スナップショット動作及び仮想ブロック範囲コピー動作にも使用でき、追加の動作は（スナップショットが完了する前に）スナップショットに続くとして識別されるので、これらの動作のそれぞれは、スナップショット動作自体が完了することを前提する場合がある。また、このプレディケートは、それらの続く動作が、スナップショットに含まれるよりむしろボリュームのスナップショット後の画像に適用することも示す場合があるであろう。

30

40

【０２４６】

【０３０４】 スナップショットに使用できるであろう代替プレディケートは、あらゆるスナップショットに識別子を割り当て、特定のスナップショットに含むことができるすべての修正動作をその識別子と関連付けるためである。その結果、スナップショットは、含まれる修正動作のすべてが完了するとき完了できる。これは、カウントプレディケートで行うことができる。データセットがその全体で同期複製される各ストレージシステムは、最後のスナップショット以降又は他のなんらかのまれな動作以降、時間と関連付けられた独自の動作のカウントを実施できる（あるいは、複数のリーダーストレージシステムを実施する実施形態の場合、特定のリーダーストレージシステムによって編成されるそれらの動作

50

を用いて、カウントは、そのリーダストレージシステムによって、それが制御するデータセットの部分について確立できる）。スナップショット動作自体は、次いでスナップショットがそれ自体永続的にされる、又は完了済みとして信号で知らせることができる前に、受け取られ、永続的にされる動作のその数に依存するカウントプレディケートを含む場合がある。（スナップショット完了前に）スナップショットに続くべきである修正動作は、スナップショットに依存する待ち行列プレディケートを所与として先延ばしにされる場合があるか、スナップショットアイデンティティが、修正動作がスナップショットから除外されるべきである旨の表示として使用される場合があるかのどちらかである。仮想ブロック範囲コピー（SCSI EXTENDED COPY又は類似動作）は、待ち行列プレディケートを使用する場合もあれば、仮想ブロック範囲コピーはカウントプレディケート及びスナップショット又は類似する識別子を使用する場合もある。カウントプレディケート及びスナップショット又は仮想コピー識別子を用いると、各仮想ブロック範囲コピーは、たとえコピー動作が1つ又は2つのボリュームの2つの小さい領域しかカバーしなくても、新しい仮想スナップショット又は仮想コピー識別子を確立する可能性がある。上述された例では、データセット（1142）を修正する要求（1104）は、データセット（1142）のスナップショットを撮影する要求を含む場合があり、したがってデータセット（1142）を修正する要求（1104）のための順序付け情報（1152）は、データセット（1142）のスナップショットを撮影する前に完了されなければならないデータセットを修正する1つ以上の他の要求の識別を含む場合がある。

【0247】

【0305】 図12に示される例の方法では、データセット（1142）に対する修正を記述する情報（1110）は、データセット（1142）を修正する要求（1104）と関連付けられた共通メタデータ情報（1154）を含む場合がある。データセット（1142）を修正する要求（1104）と関連付けられた共通のメタデータ情報（1154）は、データセット（1142）が全体で同期複製されるストレージシステム（1138、1140、1150）内のデータセット（1142）と関連付けられる共通メタデータを確実にするために使用されてよい。これに関連する共通メタデータは、例えば1つ以上の要求（例えば、ホストによって発行された1つ以上の書込み要求）によってデータセット（1142）に記憶されたコンテンツ以外の任意のデータとして実施されてよい。共通メタデータは、特にその共通メタデータが、記憶されているコンテンツがどのようにして管理され、回復され、再同期され、スナップショットを撮影される、又は同期複製されるのかに関する場合、同期複製実施態様がなんらかの方法で、データセット（1142）が全体で同期複製されるストレージシステム（1138、1140、1150）全体で一貫性を保つデータを含んでよい。読者は、2つ以上の修正動作が同じ共通メタデータに依存してよく、修正動作自体の順序付けは不必要であるが、2回よりむしろ1回、共通メタデータの一貫性がある適用が必要であることを理解する。共通メタデータに対する複数の依存関係を処理する1つの方法は、リーダストレージシステムからインスタンス化され、記述される別々の動作でメタデータを定めることである。次いで、その共通メタデータに依存する2つの修正動作は、その修正動作に依存する待ち行列プレディケートを与えられる場合がある。共通メタデータに対する複数の依存関係を処理する別の方法は、共通メタデータを2つの動作のうちの第1の動作と関連付け、第2の動作を第1の動作に依存させることである。変形形態は、第2の動作を第1の動作の共通メタデータ態様だけに依存させ、これにより第1の動作のその部分だけが、第2の動作を処理できる前に永続的にされなければならない。共通メタデータに対する複数の依存関係を処理するさらに別の方法は、その共通メタデータに依存するすべての動作記述に共通メタデータを含めることである。これは、共通メタデータを適用することが、例えば単に識別子を共通メタデータにアタッチするだけで等、冪である場合にうまく機能する。その識別子が既に処理されている場合、それは無視できる。一部の 경우에는、識別子は、共通メタデータの部分と関連付けられる可能性がある。

【0248】

【0306】 図12に示される例の方法では、フォロワーストレージシステムが、データセット(1142)を修正する要求(1104)処理した旨の表示を受け取ること(1126)は、フォロワーストレージシステム(1138、1150)のそれぞれから、フォロワーストレージシステム(1138、1150)が、データセット(1142)を修正する要求(1104)処理した旨の表示を受け取ること(1156)を含む場合がある。この例では、各フォロワーストレージシステム(1138、1150)が、データセット(1136)を修正する要求(1104)処理した旨の表示は、各フォロワーストレージシステム(1138、1150)からリーダーストレージシステム(1140)へ送信された別個の肯定応答(1122、1148)メッセージとして実施される。読者は、上述されたステップの多くが特定の順序で発生するとして示され、説明されるが、特定の順序は実際には必要とされないことを理解する。実際には、フォロワーストレージシステム(1138、1150)及びリーダーストレージシステム(1140)は、独立したストレージシステムであるため、各ストレージシステムは並行して上述されたステップのいくつかを実行している場合がある。例えば、フォロワーストレージシステム(1138、1150)の1つ以上は、データセット(1142)に対する修正を記述する情報(1110)を受け取ってよい(1116、1142)、データセット(1142)を修正する要求(1104)を処理してよい(1118、1144)、又はリーダーストレージシステム(1140)が、データセット(1142)を修正する要求(1104)を処理する前に、データセット(1142)を修正する要求(1104)の完了の受取りを確認してよい(1120、1146)。代わりに、フォロワーストレージシステム(1138、1150)の1つ以上が、データセット(1142)に対する修正を記述する情報(1110)を受け取った(1116、1142)、データセット(1142)を修正する要求(1104)を処理した(1118、1144)、又はデータセット(1142)を修正する要求(1104)の完了の受取りを確認した(1120、1146)前に、リーダーストレージシステム(1140)がデータセット(1142)を修正する要求(1104)を処理した(1124)可能性がある。

【0249】

【0307】 また、図12に示される例の方法は、リーダーストレージシステム(1140)によって、データセット(1142)を修正する要求(1104)が、データセット(1142)を修正する要求(1104)の完了の受取りを確認する(1134)前にフォロワーストレージシステム(1138、1150)のそれぞれによって処理された(1118、1144)かどうかを判断すること(1158)も含む。リーダーストレージシステム(1140)は、例えば、リーダーストレージシステム(1140)が、フォロワーストレージシステム(1138、1150)のそれぞれから、データセット(1142)を修正する要求(1104)が、フォロワーストレージシステム(1138、1150)のそれぞれによって処理された(1118、1144)ことを示す肯定応答メッセージ又は他のメッセージを受け取ったかどうかを判断することによって、データセット(1142)を修正する要求(1104)が、フォロワーストレージシステム(1138、1150)のそれぞれによって処理された(1118、1144)かどうかを判断してよい(1158)。係る例では、リーダーストレージシステム(1140)が、データセット(1142)を修正する要求(1104)がフォロワーストレージシステム(1138、1150)のそれぞれによって処理され(1118、1144)、リーダーストレージシステム(1138)によっても処理された(1124)と肯定的に(1162)に判断する場合、リーダーストレージシステム(1140)は、データセット(1142)を修正する要求(1104)を開始したホスト(1102)に対してデータセット(1142)を修正する要求(1104)の完了の受取りを確認する(1134)ことによって進んでよい。しかしながら、リーダーストレージシステム(1140)が、データセット(1142)を修正する要求(1104)が、フォロワーストレージシステム(1138、1150)の少なくとも1つによって処理され(1118、1144)ていない(1160)、又はリーダーストレージシステム(1138)によって処理され(1124)ていないと判断

10

20

30

40

50

する場合、リーダーストレージシステム（１１４０）は、データセット（１１４２）を修正する要求（１１０４）が、データセット（１１４２）がその全体で同期複製されるすべてのストレージシステム（１１３８、１１４０、１１５０）で無事に処理されたときにだけ、データセット（１１４２）を修正する要求（１１０４）を開始したホスト（１１０２）に対してデータセット（１１４２）を修正する要求（１１０４）の完了の受取りを確認してよい（１１３４）ので、リーダーストレージシステム（１１４０）は、データセット（１１４２）を修正する要求（１１０４）を開始したホスト（１１０２）に対してデータセット（１１４２）を修正する要求（１１０４）の完了の受取りをまだ確認して（１１３４）いない場合がある。

【０２５０】

【０３０８】 読者は、図１２に示される例の方法が、データセット（１１４２）が、ストレージシステムのうちの１つがリーダーストレージシステム（１１４０）であり、残りのストレージシステムがフォローストレージシステム（１１３８、１１５０）である３つのストレージシステム全体で同期複製される実施形態を示すが、他の実施形態は、さらに追加のストレージシステムを含んでよいことを理解する。係る他の実施形態では、追加のフォローストレージシステムは、図１２に示されるフォローストレージシステム（１１３８、１１５０）と同じように動作してよい。

【０２５１】

【０３０９】 追加の説明のために、図１３は、本開示のいくつかの実施形態に従って複数のストレージシステム（１１３８、１１４０）全体で同期されるデータセット（１１４２）に向けられるＩ／Ｏ動作にサービスを提供する例の方法を示すフローチャートを説明する。あまり詳細には示されないが、図１３に示されるストレージシステム（１１３８、１１４０）は、図１Ａから図１Ｄ、図２Ａから図２Ｇ、図３Ａから図３Ｂ、又はその任意の組合せに関して上述されたストレージシステムに類似してよい。実際には、図１３に示されるストレージシステムは、上述されたストレージシステムと同じ構成要素、より少ない構成要素、追加の構成要素を含んでよい。

【０２５２】

【０３１０】 図１３に示される例の方法は、フォローストレージシステム（１１３８）によって、データセット（１１４２）を修正する要求（１１０４）を受け取ること（１３０２）を含む。データセット（１１４２）を修正する要求（１１０４）は、例えばデータセット（１１４２）に含まれるデータを含むストレージシステム（１１３８）の中の場所にデータを書き込む要求として、データセット（１１４２）に含まれるデータを含むボリュームにデータを書き込む要求として、又はデータセット（１１４２）に含まれるデータのなんらかの部分に対する変更を生じさせるなんらかの他の動作として実施されてよい。図１３に示される例の方法では、データセット（１１４２）を修正する要求（１１０４）は、例えばバーチャルマシンで実行中であるアプリケーションとして、ストレージシステム（１１３８）に接続されるコンピューティングデバイスで実行中であるアプリケーションとして、又はストレージシステム（１１３８）にアクセスするように構成されたなんらかの他のエンティティとして実施されてよいホスト（１１０２）によって発行される。

【０２５３】

【０３１１】 また、図１３に示される例の方法は、フォローストレージシステム（１１３８）からリーダーストレージシステム（１１４０）に、データセット（１１４２）を修正する要求（１１０４）の論理記述（１３０６）を送信すること（１３０４）も含む。図１３に示される例の方法では、データセット（１１４２）を修正する要求（１１０４）の論理記述（１３０６）は、リーダーストレージシステム（１１３８）によって理解される方法でフォーマットされてよく、データセット（１１４２）を修正する要求（１１０４）で要求される動作（例えば、読取りタイプ動作、スナップショットタイプ動作）のタイプを記述する情報、Ｉ／Ｏペイロードが置かれている場所を記述する情報、Ｉ／Ｏペイロードのサイズを記述する情報、又は他のなんらかの情報を含んでよい。代替実施形態では、フォローストレージシステム（１１３８）は、リーダーストレージシステム（１１４０）

にデータセット（１１４２）を修正する要求（１１０４）（のすべてを含んだ）なんらかの部分単に転送してよい。

【０２５４】

【０３１２】 また、図１３に示される例の方法は、リーダストレージシステム（１１４０）によって、データセット（１１４２）に対する修正を記述する情報（１３１０）を生成すること（１３０８）も含む。リーダストレージシステム（１１４０）は、例えば順序付け対進行中である任意の他の動作を決定すること、ポッドのすべてのメンバー（例えば、データセットがその全体で同期複製されるすべてのストレージシステム）全体で例えばメタデータの共通の要素に対する分散された状態の変化を計算すること等によって、データセット（１１４２）に対する修正を記述する情報（１３１０）を生成してよい（１３０８）。データセット（１１４２）に対する修正を記述する情報（１３１０）は、例えばストレージシステムによって実行されるＩ／Ｏ動作を記述するために使用されるシステムレベルの情報として実施されてよい。リーダストレージシステム（１１４０）は、データセット（１１４２）を修正する要求（１１０４）にサービスを提供するために、起こるべきことを理解するのに十分なだけ、データセット（１１４２）を修正する要求（１１０４）を処理することによってデータセット（１１４２）に対する修正を記述する情報（１３１０）を生成してよい（１３０８）。例えば、リーダストレージシステム（１１４０）は、各ストレージシステム（１１３８、１１４０）で同等な結果を生じさせるために、データセット（１１４２）を修正する他の要求に対する、データセット（１１４２）を修正する要求（１１０４）の実行のなんらかの順序付けが必要とされるかどうかを判断する場合がある。

【０２５５】

【０３１３】 データセット（１１４２）を修正する要求（１１０４）が、データセット（１１４２）内の第１のアドレス範囲からデータセット（１１４２）内の第２のアドレス範囲にブロックをコピーする要求として実施される例を考える。係る例では、３つの他の書込み動作（書込みＡ、書込みＢ、書込みＣ）は、データセット（１１４２）の第１のアドレス範囲に向けられると仮定する。係る例では、リーダストレージシステム（１１４０）は、データセット（１１４２）内の第１のアドレス範囲からデータセット（１１４２）内の第２のアドレス範囲にブロックをコピーする前に、書込みＡ及び書込みＢにサービスを提供する（が、書込みＣにはサービスを提供しない）場合、フォロワーストレージシステム（１１３８）も、一貫した結果を生じさせるために、データセット（１１４２）内の第１のアドレス範囲からデータセット（１１４２）内の第２のアドレス範囲にブロックをコピーする前に、書込みＡ及び書込みＢに命令しなければならない（が、書込みＣには命令しない）。したがって、リーダストレージシステム（１１４０）が、データセット（１１４２）に対する修正を記述する情報（１３１０）を生成する（１３０８）とき、この例では、リーダストレージシステム（１１４０）は、フォロワーストレージシステム（１１３８）が、データセット（１１４２）を修正する要求（１１０４）を処理できる前に完了されなければならない他の動作を識別する情報（例えば、書込みＡ及び書込みＢのためのシーケンス番号）を生成できるであろう。

【０２５６】

【０３１４】 読者は、任意の動作に対する正しい結果が、動作の受取り確認が可能である前に回復可能である点にコミットされなければならないことをさらに理解する。しかし、複数の動作をともにコミットできる場合もあれば、回復が正しさを確実にするのであるならば、動作を部分的にコミットできる場合もある。例えば、スナップショットは、予想される書込みＡ及びＢに対する記録された依存関係とローカルにコミットできるであろうが、Ａ又はＢ自体はコミットされていない可能性がある。スナップショットの受取りを確認することができず、紛失したＩ／Ｏを別のアレイから回復できない場合、回復が結局スナップショットを取り消すことになる可能性がある。また、書込みＢが書込みＡと重複する場合、リーダストレージシステムはＢにＡの後となるように『命令』してよいが、Ａは実際には廃棄され、次いでＡを書き込む動作は単にＢを待機するであろう。Ａ、ＢとＣ、Ｄ

との間のスナップショットと結合された書込み A、B、C、及び D は、回復がアレイ全体でスナップショット不一致を生じさせることがない限り、及び肯定応答が、より早期の動作が、それが回復可能であることを保証される点まで持続される前に後の動作を完了しない限り、いくつかの又はすべての部分をともにコミットする及び／又は受取りを確認することができるであろう。

【0257】

【0315】 また、図 13 に示される例の方法は、リーダーストレージシステム (1140) からフォロワーストレージシステム (1338) にデータセット (1142) に対する修正を記述する情報 (1310) を送信すること (1312) も含む。リーダーストレージシステム (1140) からフォロワーストレージシステム (1338) にデータセット (1142) に対する修正を記述する情報 (1310) を送信すること (1312) は、例えば、リーダーストレージシステム (1140) がフォロワーストレージシステム (1138) に 1 つ以上のメッセージを送信することによって実施されてよい。しかしながら、フォロワーストレージシステム (1138) が、データセット (1142) を修正する要求 (1104) の元の受取人であったという事実を考慮して、リーダーストレージシステム (1140) は、データセット (1142) を修正する要求 (1104) の I/O ペイロードを送信する必要はない場合がある。したがって、フォロワーストレージシステム (1138) は、データセット (1142) を修正する要求 (1104) から I/O ペイロードを抽出してよく、フォロワーストレージシステム (1138) は、I/O ペイロードを、データセット (1142) を修正する要求 (1104) と関連付けられた 1 つ以上の他のメッセージの一部として受け取ってよく、フォロワーストレージシステム (1138) は、I/O ペイロードが既知の場所 (例えば、RDMA 又は RDMA のようなアクセスを介してアクセスされたフォロワーストレージシステム (1138) 内のバッファ) で、又はなんらかの他の方法でホストによって記憶されているので、I/O ペイロードにアクセスできてよい。

【0258】

【0316】 また、図 13 に示される例の方法は、リーダーストレージシステム (1140) によって、データセット (1142) を修正する要求 (1104) を処理すること (1318) も含む。図 13 に示される例の方法では、リーダーストレージシステム (1140) は、例えば、フォロワーストレージシステム (1138) から受け取られた I/O ペイロードだけではなく、データセット (1142) に対する修正を記述する情報 (1110) にも従ってリーダーストレージシステム (1140) に含まれる 1 つ以上のストレージデバイス (例えば、NVRAM デバイス、SSD、HDD) のコンテンツを修正することによって、データセット (1142) を修正する要求 (1104) を処理してよい (1318)。データセット (1142) を修正する要求 (1104) が、データセット (1142) に含まれるボリュームに向けられる書込み動作として実施され、データセット (1142) に対する修正を記述する情報 (1110) が、書込み動作が依然に発行された書込み動作が処理された後にだけ実行できることを示す例を考える。係る例では、データセット (1142) を修正する要求 (1104) を処理すること (1318) は、リーダーストレージシステム (1140) が最初に、以前に発行された書込み動作がリーダーストレージシステム (1140) で処理されたことを検証し、その後書込み動作と関連付けられた I/O ペイロードをリーダーストレージシステム (1140) に含まれる 1 つ以上のストレージデバイスに書き込むことによって実施されてよい。係る例では、データセット (1142) を修正する要求 (1104) は、例えば I/O ペイロードがリーダーストレージシステム (1140) の中の永続記憶装置にコミットされているとき、完了され、無事に処理されたと見なされてよい。

【0259】

【0317】 また、図 13 に示される例は、リーダーストレージシステム (1140) によってフォロワーストレージシステム (1138) に対して、データセット (1142) を修正する要求 (1104) の完了の受取りを確認すること (1320) も含む。図 13 に

示される例では、リーダーストレージシステム（１１４０）は、例えばリーダーストレージシステム（１１４０）からフォロワーストレージシステム（１１３８）に送信される１つ以上の肯定応答（１３２２）メッセージを使用することにより、又は他のなんらかの適切な機構を介してデータセット（１１４２）を修正する要求（１１０４）の完了の受取りを確認してよい（１３２０）。

【０２６０】

[0318] また、図１３に示される例の方法は、リーダーストレージシステム（１１４０）から、データセット（１１４２）に対する修正を記述する情報（１３１０）を受け取ること（１３１４）も含む。フォロワーストレージシステム（１１３８）は、例えばリーダーストレージシステム（１１４０）からフォロワーストレージシステム（１１３８）に送信される１つ以上のメッセージを介して、リーダーストレージシステム（１１４０）からデータセット（１１４２）に対する修正を記述する情報（１１１０）を受け取ってよい（１３１４）。１つ以上のメッセージは、リーダーストレージシステム（１１４０）が、RDMA又は類似した機構を使用し、又は他の方法でフォロワーストレージシステム（１１３８）上の所定の記憶場所（例えば、待ち行列の場所）にメッセージを書き込むことによって、２つのストレージシステム（１１３８、１１４０）の間の１つ以上の専用のデータ通信リンクを介してリーダーストレージシステム（１１４０）からフォロワーストレージシステム（１１３８）に送信されてよい。しかしながら、読者は、図１３に示される例の方法では、フォロワーストレージシステム（１１３８）は、フォロワーストレージシステム（１１３８）によって受け取られたデータセット（１１４２）を修正する要求（１１０４）から係るＩ／Ｏペイロードを抽出することができる、フォロワーストレージシステム（１１３８）は、ホスト（１１０２）から受け取られた１つ以上の他のメッセージから係るＩ／Ｏペイロードを抽出することができる、フォロワーストレージシステム（１１３８）は、フォロワーストレージシステム（１１３８）がホスト（１１０２）によって発行されたデータセット（１１４２）を修正する要求（１１０４）のターゲットであったという事実のためになんらかの他の方法でＩ／Ｏペイロードを入手できるので、リーダーストレージシステム（１１４０）は、データセット（１１４２）を修正する要求（１１０４）と関連付けられたＩ／Ｏペイロードをフォロワーストレージシステム（１１３８）に送信する必要はないことを理解する。

【０２６１】

[0319] 一実施形態では、フォロワーストレージシステム（１１３８）は、通信機構としてSCSI要求（送信者から受信者への書込み、又は受信者から送信者への読取り）の使用の使用により、リーダーストレージシステム（１１４０）からデータセット（１１４２）に対する修正を記述する情報（１１１０）を受け取ってよい（１３１４）。係る実施形態では、SCSI書込み要求は、（どのようなデータ及びメタデータであれ含む）送信することを意図し、特別な疑似デバイスに又は特別に構成されたSCICネットワークを介して、又は任意の他の同意されたアドレス指定機構を通して送達されてよい情報を符号化するために使用される。あるいは、代わりに、モデルは、特別なデバイス、特別に構成されたSCSIネットワーク、又は他の同意された機構も使用し、受信者から送信者へオープンSCSI読取り要求の集合を発行できる。データ及びメタデータを含んだ符号化された情報は、これらのオープンSCSI要求の１つ以上に対する応答として受信者に送達される。係るモデルは、多くの場合データセンタ間の「ダークファイバ」ストレージネットワークとして配備されるファイバチャネルSCSIネットワークを介して実装できる。また、係るモデルは、ホストからリモートアレイへのマルチパッシング及びバルクアレイ間通信用に同じネットワーク回線の使用も可能にする。

【０２６２】

[0320] また、図１３に示される例の方法は、フォロワーストレージシステム（１１３８）によって、データセット（１１４２）を修正する要求（１１０４）を処理すること（１３１６）も含む。図１３に示される例の方法では、フォロワーストレージシステム（１１３８）は、データセット（１１４２）に対する修正を記述する情報（１１１０）に従っ

てフォロワーストレージシステム（１１３８）に含まれる１つ以上のストレージデバイス（例えば、NVRAMデバイス、SSD、HDD）のコンテンツを修正することによって、データセット（１１４２）を修正する要求（１１０４）を処理してよい（１３１６）。データセット（１１４２）を修正する要求（１１０４）が、データセット（１１４２）に含まれるボリュームに向けられる書込み動作として実施され、データセット（１１４２）に対する修正を記述する情報（１１１０）が、書込み動作が、以前に発行された書込み動作が処理された後にだけ実行できることを示す例を考える。係る例では、データセット（１１４２）を修正する要求（１１０４）を処理すること（１３１６）は、フォロワーストレージシステム（１１３８）が最初に、以前に発行された書込み動作がフォロワーストレージシステム（１１３８）で処理されたことを検証し、その後書込み動作と関連付けられたI/Oペイロードをフォロワーストレージシステム（１１３８）に含まれる１つ以上のストレージデバイスに書き込むことによって実施されてよい。係る例では、データセット（１１４２）を修正する要求（１１０４）は、例えばデータセット（１１４２）を修正する要求（１１０４）と関連付けられたI/Oペイロードがフォロワーストレージシステム（１１３８）の中の永続記憶装置にコミットされているとき、完了され、無事に処理されたと見なされてよい。

【０２６３】

【0321】 また、図１３に示される例の方法は、リーダーストレージシステム（１１４０）から、リーダーストレージシステム（１１４０）が、データセット（１１４２）を修正する要求（１１０４）を処理した旨の表示を受け取ること（１３２４）も含む。この例では、リーダーストレージシステム（１１４０）が、データセット（１１４２）を修正する要求（１１０４）を処理した旨の表示は、リーダーストレージシステム（１１４０）からフォロワーストレージシステム（１１３８）に送信された肯定応答（１３２２）メッセージとして実施される。読者は、上述されたステップの多くが特定の順序で発生するとして示され、説明されるが、特定の順序は実際には必要とされないことを理解する。実際には、フォロワーストレージシステム（１１３８）及びリーダーストレージシステム（１１４０）は、独立したストレージシステムであるため、各ストレージシステムは並行して上述されたステップのいくつかを実行している場合がある。例えば、フォロワーストレージシステム（１１３８）は、リーダーストレージシステム（１１４０）から、リーダーストレージシステム（１１４０）が、データセット（１１４２）を修正する要求（１１０４）を処理する（１３１６）前にデータセット（１１４２）を修正する要求（１１０４）を処理した旨の表示を受け取ってよい（１３２４）。同様に、フォロワーストレージシステム（１１３８）は、リーダーストレージシステム（１１４０）から、リーダーストレージシステム（１１４０）が、リーダーストレージシステム（１１４０）からデータセット（１１４２）に対する修正を記述する情報（１１１０）を受け取る（１３１４）前に、データセット（１１４２）を修正する要求（１１０４）を処理した旨の表示を受け取ってよい（１３２４）。

【０２６４】

【0322】 また、図１３に示される例は、フォロワーストレージシステム（１１３８）によって、データセット（１１４２）を修正する要求（１１０４）の完了の受取りを確認すること（１３２６）も含む。データセット（１１４２）を修正する要求（１１０４）の完了の受取りを確認すること（１３２６）は、例えばフォロワーストレージシステム（１１３８）が、データセット（１１４２）を修正する要求（１１０４）を発行したホスト（１１０２）に肯定応答（１３２８）メッセージを発行することによって実施されてよい。図１３に示される例の方法では、フォロワーストレージシステム（１１３８）は、データセット（１１４２）を修正する要求（１１０４）が、データセット（１１４２）を修正する要求（１１０４）の完了の受取りを確認する（１３２８）前に、リーダーストレージシステム（１１４０）によって処理された（１３１８）かどうかを判断してよい。フォロワーストレージシステム（１１３８）は、例えば、フォロワーストレージシステム（１１３８）が、肯定応答メッセージ、又はデータセット（１１４２）を修正する要求（１１０４）が

リーダーストレージシステム（１１４０）によって処理された（１３１８）ことを示す他のメッセージをリーダーストレージシステム（１１４０）から受け取ったかどうかを判断することによって、データセット（１１４２）を修正する要求（１１０４）がリーダーストレージシステム（１１４０）によって処理された（１３１８）かどうかを判断してよい。係る例では、フォロワーストレージシステム（１１３８）が、データセット（１１４２）を修正する要求（１１０４）がリーダーストレージシステム（１１４０）によって処理された（１３１８）と肯定的に判断し、フォロワーストレージシステム（１１３８）もデータセット（１１４２）を修正する要求（１１０４）を処理した（１３１６）場合、フォロワーストレージシステム（１３１８）は、データセット（１１４２）を修正する要求（１１０４）を開始したホスト（１１０２）にデータセット（１１４２）を修正する要求（１１０４）の完了の受取りを確認する（１３２６）によって進んでよい。しかしながら、リーダーストレージシステム（１１４０）が、データセット（１１４２）を修正する要求（１１０４）がリーダーストレージシステム（１１４０）によって処理されていない（１３１８）、又はフォロワーストレージシステム（１１３８）がデータセット（１１４２）を修正する要求（１１０４）をまだ処理していない（１３１６）と判断する場合、フォロワーストレージシステム（１１３８）は、データセット（１１４２）を修正する要求（１１０４）が、データセット（１１４２）がその全体で同期複製されるすべてのストレージシステム（１１３８、１１４０）で無事に処理されたときにだけ、データセット（１１４２）を修正する要求（１１０４）を開始したホスト（１１０２）に対してデータセット（１１４２）を修正する要求（１１０４）の完了の受取りを確認してよいので、フォロワーストレージシステム（１１３８）は、データセット（１１４２）を修正する要求（１１０４）を開始したホスト（１１０２）に対してデータセット（１１４２）を修正する要求（１１０４）の完了の受取りをまだ確認して（１３２６）いない場合がある。

【０２６５】

【０３２３】 追加の説明のため、図１４は、本開示のいくつかの実施形態に従って複数のストレージシステム（１１３８、１１４０、１３３４）全体で同期されるデータセット（１１４２）に向けられるＩ／Ｏ動作にサービスを提供する例の方法を示すフローチャートを説明する。あまり詳細には示されないが、図１３に示されるストレージシステム（１１３８、１１４０、１３３４）は、図１Ａから図１Ｄ、図２Ａから図２Ｇ、図３Ａから図３Ｂ、又はその任意の組合せに関して上述されたストレージシステムに類似してよい。実際には、図１３に示されるストレージシステムは、上述されたストレージシステムと同じ構成要素、より少ない構成要素、追加の構成要素を含んでよい。

【０２６６】

【０３２４】 図１４に示される例の方法は、図１３に示される例の方法に類似してよい。図４に示される例の方法は、フォロワーストレージシステム（１１３８）によって、データセット（１１４２）を修正する要求（１１０４）を受け取ること（１３０２）と、フォロワーストレージシステム（１１３８）からリーダーストレージシステム（１１４０）に、データセット（１１４２）を修正する要求（１１０４）の論理記述（１３０６）を送信すること（１３０４）と、リーダーストレージシステム（１１４０）によって、データセット（１１４２）に対する修正を記述する情報（１３１０）を生成すること（１３０８）と、リーダーストレージシステム（１１４０）によって、データセット（１１４２）を修正する要求（１１０４）を処理すること（１３１８）と、リーダーストレージシステム（１１４０）によってフォロワーストレージシステム（１１３８）に対して、データセット（１１４２）を修正する要求（１１０４）の完了の受取りを確認すること（１３２０）と、リーダーストレージシステム（１１４０）から、データセット（１１４２）に対する修正を記述する情報（１３１０）を受け取ること（１３１４）と、フォロワーストレージシステム（１１３８）によって、データセット（１１４２）を修正する要求（１１０４）を処理すること（１３１６）と、リーダーストレージシステム（１１４０）から、リーダーストレージシステム（１１４０）が、データセット（１１４２）を修正する要求（１１０４）を処理した旨の表示を受け取ること（１３２４）と、フォロワーストレージシステム

(1138)によって、データセット(1142)を修正する要求(1104)の完了の受取りを確認すること(1326)も含むので、図13に示される例の方法に類似してよい。

【0267】

[0325] しかしながら、図14に示される例の方法は、データセット(1142)が、ストレージシステムのうちの1つがリーダーストレージシステム(1140)であり、残りのストレージシステムがフォロワーストレージシステム(1138、1334)である、3つのストレージシステム全体で同期複製される実施形態を示すので、図13に示される例の方法とは異なる。係る例では、追加のフォロワーストレージシステム(1134)は、リーダーストレージシステム(1140)から、データセット(1142)に対する修正を記述する情報(1110)を受け取る(1142)、及びデータセット(1142)9に対する修正を記述する情報(1110)に従ってデータセット(1142)を修正する要求(1104)を処理する(1142)ことができるので、図13に示されたフォロワーストレージシステム(1138)と同じステップの多くを実施する。

【0268】

[0326] 図14に示される例の方法では、リーダーストレージシステム(1140)は、フォロワーストレージシステム(1138、1134)のすべてにデータセット(1142)に対する修正を記述する情報(1110)を送信できる(1338)。図14に示される例の方法では、追加のフォロワーストレージシステム(1334)は、データセット(1142)を修正する要求(1104)を受け取った(1302)フォロワーストレージシステム(1138)にデータセット(1142)を修正する要求(1104)の完了の受取りを確認する(1330)こともできる。図14に示される例の方法では、追加のフォロワーストレージシステム(1334)は、例えば追加のフォロワーストレージシステム(1334)から、データセット(1142)を修正する要求(1104)を受け取った(1302)フォロワーストレージシステム(1138)に送信される1つ以上の肯定応答(1332)メッセージの使用により、又はなんらかの他の適切な機構を介して、データセット(1142)を修正する要求(1104)を受け取った(1302)フォロワーストレージシステム(1138)に対してデータセット(1142)を修正する要求(1104)の完了の受取りを確認できる(1330)。

【0269】

[0327] 図14に示される例では、データセット(1142)を修正する要求(1104)を受け取った(1302)フォロワーストレージシステム(1138)も、他のすべてのフォロワーストレージシステム(1134)がデータセット(1142)を修正する要求(1104)を処理した旨の表示を受け取ってもよい(1336)。この例では、他のすべてのフォロワーストレージシステム(1134)がデータセット(1142)を修正する要求(1104)を処理した旨の表示は、他のフォロワーストレージシステム(1134)から、データセット(1142)を修正する要求(1104)を受け取った(1302)フォロワーストレージシステム(1138)に送信された肯定応答(1332)メッセージとして実施される。読者は、上述されたステップの多くが特定の順序で発生するとして示され、説明されるが、特定の順序は実際には必要とされないことを理解する。実際には、フォロワーストレージシステム(1138、1334)及びリーダーストレージシステム(1140)は、それぞれ独立したストレージシステムであるため、各ストレージシステムは並行して上述されたステップのいくつかを実行している場合がある。例えば、フォロワーストレージシステム(1138)は、リーダーストレージシステム(1140)から、リーダーストレージシステム(1140)が、データセット(1142)を修正する要求(1104)を処理する(1316)前にデータセット(1142)を修正する要求(1104)を処理した旨の表示を受け取ってよい(1324)。さらに、フォロワーストレージシステム(1138)は、他のすべてのフォロワーストレージシステム(1334)が、リーダーストレージシステム(1140)がデータセット(1142)を修正する要求(1104)を処理した旨の表示を受け取る(1324)前に、データセ

ット（１１４２）を修正する要求（１１０４）を処理した旨の表示を受け取ってよい（１３３６）。代わりに、フォロワーストレージシステム（１１３８）は、他のすべてのストレージシステム（１３３４）が、データセット（１１４２）を修正する要求（１１０４）を処理する（１３１６）前に、データセット（１１４２）を修正する要求（１１０４）を処理した旨の表示を受け取ってよい（１３３６）。同様に、フォロワーストレージシステム（１１３８）は、リーダーストレージシステム（１１４０）から、リーダーストレージシステム（１１４０）が、リーダーストレージシステム（１１４０）からデータセット（１１４２）に対する修正を記述する情報（１１０）を受け取る（１３１４）前に、データセット（１１４２）を修正する要求（１１０４）を処理した旨の表示を受け取ってよい（１３２４）。さらに、フォロワーストレージシステム（１１３８）は、他のすべてのフォロワーストレージシステム（１３３４）が、リーダーストレージシステム（１１４０）からデータセット（１１４２）に対する修正を記述する情報（１１１０）を受け取る（１３１４）前に、データセット（１１４２）を修正する要求（１１０４）を処理した旨の表示を受け取ってよい（１３３６）。

10

【０２７０】

【０３２８】 図１４に明示的に示されていないが、フォロワーストレージシステム（１１３８）は、データセット（１１４２）を修正する要求（１１０４）が、リーダーストレージシステム（１１４０）によって処理され（１３１８）、データセット（１１４２）を修正する要求（１１０４）の完了の受取りを確認する（１３２８）前に、他のすべてのフォロワーストレージシステム（１３３４）によっても処理された（１１４４）かどうかを判断してよい。フォロワーストレージシステム（１１３８）は、例えば、フォロワーストレージシステム（１１３８）が、データセット（１１４２）を修正する要求（１１０４）が各ストレージシステム（１１４０、１３３４）によって処理された（１３１８、１１４４）ことを示す肯定応答メッセージをリーダーストレージシステム（１１４０）及び他のすべてのフォロワーストレージシステム（１３３４）から受け取ったかどうかを判断することによって、データセット（１１４２）を修正する要求（１１０４）が、リーダーストレージシステム（１１４０）によって処理され（１３１８）、他のすべてのフォロワーストレージシステム（１３３４）によっても処理された（１１４４）かどうかを判断してよい。係る例では、フォロワーストレージシステム（１１３８）が、データセット（１１４２）を修正する要求（１１０４）が、リーダーストレージシステム（１１４０）、他のすべてのフォロワーストレージシステム（１３３４）、及びフォロワーストレージシステム（１１３８）によって処理されたと肯定的に判断する場合、フォロワーストレージシステム（１１３８）は、データセット（１１４２）を修正する要求（１１０４）を開始したホスト（１１０２）に対して、データセット（１１４２）を修正する要求（１１０４）の完了の受取りを確認する（１３２６）ことによって進んでよい。しかしながら、リーダーストレージシステム（１１４０）が、データセット（１１４２）を修正する要求（１１０４）が、リーダーストレージシステム（１１４０）、他のすべてのフォロワーストレージシステム（１３３４）、又はフォロワーストレージシステム（１１３８）の少なくとも１つによって処理されていないと判断する場合、フォロワーストレージシステム（１１３８）は、データセット（１１４２）を修正する要求（１１０４）が、データセット（１１４２）がその全体で同期複製されるすべてのストレージシステム（１１３８、１１４０、１３３４）で無事に処理されたときにだけ、データセット（１１４２）を修正する要求（１１０４）を開始したホスト（１１０２）に対して、データセット（１１４２）を修正する要求（１１０４）の完了の受取りを確認してよい（１１３４）ので、フォロワーストレージシステム（１１３８）は、データセット（１１４２）を修正する要求（１１０４）を開始したホスト（１１０２）に対してデータセット（１１４２）を修正する要求（１１０４）の完了の受取りをまだ確認して（１３２６）いない場合がある。

20

30

40

【０２７１】

【０３２９】 図１４に明示的に示されていないが、いくつかの実施形態では、１つ又はストレージシステム（１１３８、１１４０、１３３４）で実行中の任意の同時重複読取りをア

50

ンロックしようと努力して、データセット（１１２２）を修正する要求（１１０４）を受け取った（１３０２）フォロワーストレージシステム（１１３８）は、修正動作がどこでも完了したことを信号で知らせるために、リーダーストレージシステム（１１４０）に及び他のフォロワーストレージシステム（１３３４）にメッセージを送り返す場合がある。代わりに、データセット（１１２２）を修正する要求（１１０４）を受け取ったフォロワーストレージシステム（１１３８）は、そのメッセージをリーダーストレージシステム（１１３８）に送信し、リーダーストレージシステム（１１３８）は、完了を伝搬し、どこでも読取りから障害物を取り除くためにメッセージを送信できるであろう。

【０２７２】

【０３３０】 読者は、図１４に示される例の方法が、データセット（１１４２）が、ストレージシステムの１つがリーダーストレージシステム（１１４０）であり、残りのストレージシステムがフォロワーストレージシステム（１１３８、１３３４）である３つのストレージシステム全体で同期複製される実施形態を示すが、他の実施形態が追加のストレージシステムを含んでもよいことを理解する。係る他の実施形態では、追加のフォロワーストレージシステムは、図１４に示される他のフォロワーストレージシステム（１３３４）と同様に動作してよい。

【０２７３】

【０３３１】 また、読者は、図１２に示される例だけが、データセット（１１４２）に対する修正を記述する情報（１３１０）がデータセット（１１４２）を修正する要求（１１０４）のための順序付け情報（１１５２）、データセット（１１４２）を修正する要求（１１０４）と関連付けられた共通メタデータ情報（１１５４）、及びデータセット（１１４２）を修正する要求（１１０４）と関連付けられたＩ／Ｏペイロード（１１１４）を含む実施形態を明示的に示すが、データセット（１１４２）に対する修正を記述する情報（１３１０）が、残りの図に示される例に係る情報のすべて（又は部分集合）を含む場合があることも理解する。さらに、データセット（１１４２）を修正する要求（１１０４）が、データセット（１１４２）のスナップショットを撮影する要求を含む実施形態では、データセット（１１４２）に対する修正を記述する情報（１３１０）は、上述された図のそれぞれでデータセット（１１４２）のスナップショットのコンテンツに含まれるデータセット（１１４２）を修正する１つ以上の他の要求の識別を含む場合もある。

【０２７４】

【０３３２】 読者は、データセット（１１４２）に対する修正を記述する情報（１３１０）が、スナップショットを撮影する前に完了されなければならないデータセット（１１４２）を修正する１つ以上の他の要求を識別する情報を含むよりむしろ、データセット（１１４２）のスナップショットのコンテンツに含まれるデータセット（１１４２）を修正する１つ以上の他の要求の識別を含んだ結果として、いくつかの状況に対応できることを理解する。ある状況とは、アトミック動作がスナップショットを実行し、同じアトミック更新で最後のいくつかの書き込みを完了できる、つまり最後のいくつかの書き込みはスナップショットの「前に」は完了しない状況である。別の状況とは、書き込み完了時に、書き込みが含まれるのであるならば、及びスナップショット自体が、すべての書き込みがすべての同期中ストレージシステムによって完了されるまで、完了したと見なされないのであるならば、書き込みが、スナップショットポイントが撮影された後に実際に完了できるであろう状況である。最後に、スナップショットが受け取られる前に、完了したとして要求者に示されていなかった書き込みは、回復アクションの結果として含まれる、又はスナップショットから省略されることがあるであろう。基本的には、回復は、結果に一貫性があり、完了したとして信号で知らされた動作に関係するいかなる保証にも違反しないならば、受け取られた動作の詳細な履歴を書き換えることができる。

【０２７５】

【０３３３】 追加の説明のために、図１５は、本開示のいくつかの実施形態に従ってデータセットを同期複製するストレージシステム間で仲介するための例の方法を示すフローチャートを説明する。図１５に示される例の方法は、データセット（１５１２）は、２つのス

ストレージシステム（１５１４、１５２４）だけ全体で同期複製される実施形態を示すが、図１５に示される例は、データセット（１５１２）が追加のストレージシステム全体で同期複製される実施形態に拡張できる。

【０２７６】

【0334】 以下の例では、ポッドのためのストレージシステム（１５１４、１５２４）の集合の間の仲介は、ストレージシステムが、通信が通信障害又はなんらかの他の種類のシステム障害のために失われる場合がある、対にされたシステムとの失われた通信を解決できるようにする。以下に説明されるように、仲介の解決策は、定足数、及びストレージシステムのどれがポッドデータセットに向けられるＩ／Ｏ動作の処理、及び例えばメディアータ等に対する競争を続行すべきなのかを決定する外部制御システムの使用を含んでよい。しかしながら、仲介の優位点は、それが定足数プロトコルよりも簡略であり、仲介が、共通の構成である、同期複製されたストレージシステムのための２つのストレージシステム構成とうまく機能する点である。さらに、仲介は、外部制御システム及び競争される場合がある多くの他のタイプのリソースよりもよりロバストであり、構成するのが容易である場合がある。

【０２７７】

【0335】 図１５に示されるように、データセット（１５１２）を同期複製している複数のストレージシステム（１５１４、１５２４）は、一仲介サービス（１５００）は、ストレージシステム間の通信障害の場合に、ストレージシステムがオフラインになる場合に、又はなんらかの他のトリガイメントに起因して、どのストレージシステムがデータセットにサービスを提供し続けるのかを決定し得る一ネットワーク（１５５４）を介して仲介サービス（１５００）と通信してよい。ストレージシステムが互いと通信できない場合、ストレージシステムは同期複製されたデータセットを維持できない場合があり、任意の受け取られたデータセットを修正する要求は、それ以外の場合、データセットが非同期になるであろうため、サービス提供不可となるであろうため、仲介は有利である。この例では、データセットを同期複製しているストレージシステムのための仲介サービスは、ストレージシステム（１５１４、１５２４）にとっては外部である仲介サービス（１５００）によって提供されてよい。この例では、示されている２つのストレージシステム（１５１４、１５２４）しかないが、一般的には、なんらかの他の数の２つ以上のストレージシステムが、データセットを同期複製している同期中リストの一部であってよい。具体的には、第１のストレージシステム（１５１４）は、例えば第２のストレージシステム（１５２４）に対する通信リンク（１５１６）の損失等のトリガイメントを検出した場合、第１のストレージシステム（１５１４）は、それがデータセットの複製に関して同期しているストレージシステムを指定する同期中リストから通信していないストレージシステムを削除するタスクを無事に引き継ぐことができるかどうかを判断するために、外部仲介サービス（１５００）に連絡してよい。他の場合では、第１のストレージシステム（１５１４）は、外部仲介サービス（１５００）に連絡し、それ、つまり第１のストレージシステム（１５００）が、第２のストレージシステムによって同期中リストから削除された可能性があるかと判断してよい。これらの例では、通常の状態では、ストレージシステム（１５１４、１５２４）は、正常に動作するために及びデータセット（１５１２）の同期複製を維持するために仲介サービス（１５００）からいずれの情報も必要としないため、ストレージシステム（１５１４、１５２４）は、外部仲介サービス（１５００）と連続通信している必要はない。言い換えると、この例では、仲介サービス（１５００）は、同期中リストのメンバーシップ管理において積極的な役割を有しておらず、さらに、仲介サービス（１５００）は、同期中リストのストレージシステム（１５１４、１５２４）の正常な動作さえ認識していない場合がある。代わりに、仲介サービス（１５００）は、同期中リストのメンバーシップを決定するために、又はストレージシステムが別のストレージシステムをデタッチするために作用できるかどうかを判断するために、ストレージシステム（１５１４、１５２４）によって使用される持続する情報を単に提供してよい。

【０２７８】

10

20

30

40

50

【0336】 一部の例では、仲介サービス（1500）は、例えばストレージシステム（1514、1524）が互いに通信することを妨げる通信リンク故障等のトリガイベントに応じて、1つ以上のストレージシステム（1514、1524）によって連絡される場合があるが、各ストレージシステム（1514、1524）は、ストレージシステム（1514、1524）間で使用される通信チャネルとは異なる通信チャネルを介して仲介サービス（1500）と通信できる場合がある。結果的に、ストレージシステム（1514、1524）は互いに通信できない場合があるが、それでもストレージシステム（1514、1524）は依然として仲介サービス（1500）と通信してよく、ストレージシステム（1514、1524）は、どのストレージシステムがデータストレージ要求にサービスを提供するために進み得るのかを決定するために仲介サービス（1500）を使用してよい。さらに、仲介サービス（1500）から仲介を勝ち取るストレージシステムは、別のストレージシステムをデタッチし、データセット（1512）を同期複製し続けてよい。ストレージシステムを示す同期中リストを更新してよい。一部の例では、仲介サービス（1500）は、例えば、要求者ストレージシステムを含み、別のストレージシステムを除外するメンバーシップリストを設定する要求等、多様なタイプの要求を処理してよい。この例では、仲介サービス（1500）が現在要求者をメンバーとして一覧する場合、要求は無事に完了し、仲介サービス（1500）が現在要求者をメンバーとして一覧しない場合、要求は失敗する。このようにして、2つのストレージシステム（1514、1524）が、それぞれほぼ同時に要求を行っており、要求が他方を除外するために働く場合、次いで受け取られた第1の要求は成功する場合があり、仲介サービスは、第1の要求に従って他方のストレージシステムを除外するためにメンバーシップリストを設定する。受け取られた第2の要求は、メンバーシップリストがそれを除外するように設定されているため失敗する場合がある。メンバーシップリストを記憶する共用リソースに対する相互に排他的なアクセスは、一度にただ1つのシステムしかメンバーシップリストを設定することを許されないことを確実にするのに役立つ。

【0279】

【0337】 別の例では、仲介は、値が、ポッドメンバーシップパーティション識別子を示して、メンバーシップがポッドからストレージシステムのなんらかの集合を仕切った、又は削除したことを主張するために定義され得る、特定の識別子に基づく場合がある。『ポッド』は、該用語がここで及び本願の残りを通して使用されるように、データセット、管理オブジェクト及び管理動作の集合、データセットを修正する又は読み取るためのアクセス動作の集合、及び複数のストレージシステムを表す管理エンティティとして実施されてよい。係る管理動作は、データセットを読み取る又は修正するためのアクセス動作が、ストレージシステムのいずれかを通して同等に動作する、ストレージシステムのいずれかを通して同等に管理オブジェクトを修正する又は問い合わせしてよい。各ストレージシステムは、ストレージシステムによる使用のために記憶され、宣伝されたデータセットの適切な部分集合としてデータセットの別々のコピーを記憶してよく、任意の1つのストレージシステムを通して実行され、完了された管理オブジェクト又はデータセットを修正する動作は、ポッドを問い合わせるために後続の管理オブジェクトで、又はデータセットを読み取るために後続のアクセス動作で反映される。『ポッド』に関する追加の詳細は、参照により本明細書に援用される以前に出願された仮特許出願第62／518,071号で見つけられてよい。

【0280】

【0338】 パーティション識別子は、ポッドメンバーシップリストを記憶する所与のストレージシステムに加えて、所与のストレージシステムに記憶されたローカル情報であってよい。互いに適切に通信中であり、同期中であるシステムは、同じパーティション識別子を有してよく、ストレージシステムがポッドに加えられると、次いで現在のパーティション識別子はポッドデータコンテンツとともにコピーされてよい。この例では、ストレージシステムのある集合がストレージシステムの別の集合と通信していないとき、各集合からの1つのストレージシステムは、新しく一意のパーティション識別子を考え出し、最初に

共用リソースに対するロックを取得するストレージシステムに対して成功する特定の動作を使用することによって、仲介サービス（1500）によって維持される共用リソースにパーティション識別子を設定しようと試みる場合があり—共用リソースに対するロックを取得できなかった一別のストレージシステムは、特定の動作を実行する試みに失敗する。一実施態様では、アトミックなコンペアアンドセット（compare-and-set）動作が使用されてよく、仲介サービス（1500）によって記憶された最後のパーティション識別子値が、パーティション識別子を新しい値に変更する許可を有するためにストレージシステムによって提供されてよい。この例では、コンペアアンドセット動作は、現在のパーティション識別子を認識しているストレージシステムに対して成功する場合がある—パーティション識別子値を最初に設定するストレージシステムは、現在のパーティション識別子値を認識するストレージシステムであろう。さらに、ウェブサービスプロトコルで利用可能である場合がある条件付きストア（conditional store）又はPUT動作は、この例で説明されるように、パーティション識別子を設定するために機能してよい。例えばSCSI環境において等、他の場合では、コンペアアンドライト動作が使用されてよい。さらに他の場合では、仲介サービス（1500）は、ストレージシステムから要求を受け取ることによってコンペアアンドセット動作を実行してよく、要求は、古いパーティション識別子値及び新しいパーティション識別子値も示し、現在記憶されている値が古いパーティション識別子に等しいとき、且つそのときに限り、仲介サービス（1500）は、記憶されているパーティション識別子を新しいパーティション識別子に変更する。

【0281】

【0339】 このようにして、パーティション識別子に基づいた仲介は、所与のストレージシステムが、一貫性のあるポッドメンバーの区切られた集合の中に含まれるかどうかを判断するためにストレージシステムによって使用されてよい情報を持続するために使用されてよい。一部の場合では、パーティション識別子は、ストレージシステム又はネットワーク相互接続のどちらかの障害に起因する自然発生的なデタッチの場合にだけ変化してよい。これらの例では、制御された方法でポッドのためにそれ自体をオフラインにするストレージシステムは、それ自体を同期中のポッドメンバーとして削除するために他のストレージシステムと通信してよく、このようにして仲介された新しいパターン識別子の形成を必要としない。さらに、それ自体を同期中のポッドのメンバーとして削除するストレージシステムは、次いで仲介された新しいパーティション識別子を必要としない制御された方法で、同期中のポッドメンバーとしてそれ自体を加え直してよい。さらに、新しいストレージシステムは、ストレージシステムが同期中のポッドメンバーと通信しているのであるならば同期中のポッドに加えられてよく、新しいストレージシステムは、仲介された新しいパーティション識別子を必要としない制御された方法でそれら自体を加えてよい。

【0282】

【0340】 結果的に、仲介されたパーティション識別子機構の優位点は、仲介サービス（1500）が、障害、又はストレージシステムの少なくとも1つの集合が、同期中のポッドメンバーシップリストから1つ以上の通信していないストレージシステムを削除しようと試みることによって反応する他のトリガイイベントがあるときにだけ必要とされてよく、通信していないストレージシステムは同を行おうと試みる場合があるが、逆になる点である。別の優位点は、仲介サービス（1500）が、決して絶対的に信頼し得ず、同期中のポッドメンバーによって提供される全体的なストレージサービスの可用性にほとんど影響を与え得ない点である。例えば、2つの同期複製されたストレージシステムがそれぞれ1年に1度故障する場合、次いで仲介サービス（1500）が正確な瞬間に利用できない場合を除いて、2つのストレージシステムの第1のストレージシステムは故障し、第2のストレージシステムは、第1のストレージシステムを削除するために無事に仲介すべきである。要するに、仲介サービス（1500）が作動しており、少なくとも99%の時間利用可能である場合、仲介サービス（1500）が、必要時に利用できない確率は甚だしく低くなる。この例では、仲介サービス（1500）が重大なときに利用できないであろう可

能性は100の中の1（1%以下）にすぎないであろう—これは、年に1度の機能停止を1世紀に1度の機能停止に削減できる。しかしながら、仲介サービス（1500）の非有用性を見込みを減らすためには、仲介サービス（1500）は、仲介サービスが概して利用できない場合に管理者に警告するために周期的にモニタされる場合があり、仲介サービス（1500）は、特定のストレージシステムが利用できなくなる場合、アラートを生成するためにストレージシステムをモニタする場合がある。

【0283】

【0341】 別の例では、ポッドのための同期中メンバーと関連付けられたパーティション識別子を使用することの代替策として、仲介サービス（1500）は、一回限りの仲介競争のターゲットを提供してよい。具体的には、ポッドのための同期中のメンバーストレージシステムが、1つのストレージシステムが他によってデタッチされ得る可能性を可能にする必要があるたびに、仲介競争のターゲットが確立されてよい。例えば、仲介値の表の中の同意された鍵は新しい値に一回限り設定されてよく、仲介を勝ち取るためには、ストレージシステムは、同意された鍵を、他の別に競争しているストレージシステムが使用しないであろう一意の値に設定する。仲介競争より前に、同意された鍵は存在しない場合もあれば、同意された鍵が存在する場合には、鍵は例えばUNSET、又はヌル値等のなんらかの同意されたプレカーソル値に設定されてよい。この例では、鍵を特定の値に設定する動作は、鍵が存在しない場合、鍵がUNSET状態にある場合、又は鍵が現在値に等しい値に設定されている場合に成功する—それ以外の場合、鍵を設定するための動作は失敗する。ストレージシステムの集合が仲介を勝ち取ると、ストレージシステムの残りの集合は将来の仲介のために新しい鍵を定義してよい。この例では、ストレージシステムは、ストレージシステムが仲介競争に勝った可能性があることを知る前に、ストレージシステムが、それが障害を起こし、回復する又はリポートする場合、値を再び使用し得るように、ストレージシステムが仲介競争の前に使用する値を記録してよい。2つ以上のストレージシステムが通信中であり、通信していないストレージシステムのなんらかの他の集合に対して互いに競争している場合、この値は、それらの他の通信中のストレージシステムに共用されてよく、これによりそれらの任意の1つは仲介競争を続行し、おそらくなんらかの追加の一連の障害の後に第2の仲介競争に参加し得る。例えば、第2の仲介競争のターゲットのための一意の値のために競争する前に、正確さが第1の仲介競争のターゲットのために競争する又は第1の仲介競争のターゲットを確認することが必要である場合がある。特に、このシーケンスは、第2の仲介競争のターゲットが、第1の仲介競争のターゲットを共用するすべてのストレージシステムに確実に配布され、すべてのストレージシステムが、それが確実に配布されたことを認識させられるまで必要である場合がある。その時点では、第2の仲介ターゲットのために競争する前に、最初に第1の仲介ターゲットのために競争する継続的な必要性はない場合がある。

【0284】

【0342】 一部の例では、仲介サービス（1500）は、仲介されているストレージシステムの組織又は所有者以外の組織によって提供されるコンピュータシステムで管理される場合がある。例えば、ベンダが2つのストレージシステムをカスタマに販売する場合、ベンダは、ベンダ所有の又はベンダ管理のデータセンタで提供されるサーバでメディアータのホストとなる場合もあれば、ベンダはサービスのホストとなるためにクラウドサービスプロバイダと契約する場合もある。また、ベンダは、仲介サービスが十分に信頼でき、カスタマの障害ゾーンとは異なっていることを確実にしてもよい。ある事例では、他のクラウドサービスプロバイダを除外することなく、仲介サービスはAmazon Web Services（商標）でホストされてよく、仲介サービスは、信頼できるデータベースサービスのためにDynamoDBで実装されてよく、DynamoDBはウェブAPIデータベース更新として条件付きストアプリミティブに対するサポートを提供してよい。一部の例では、仲介サービスは、さらに信頼性を高めるために、複数のクラウドサービスプロバイダ領域又は障害領域全体で動作するように実装されてよい。ベンダを使用して仲介サービスを提供することの1つの優位点は、仲介サービスが構成するのが簡単である点

10

20

30

40

50

である。さらに、ポッドの作成中、ストレージシステムは、仲介サービスから暗号トークンを入手し、パーティション識別子及びポッドメンバーシップリストを記憶することに加えて暗号トークンを記憶してよい。暗号トークンは、ポッドのための一意の仲介サービス情報を安全に通信するために使用され得る。

【0285】

【0343】 一部の 경우에는、仲介サービス(1500)は、ストレージシステムが仲介を試みるときに利用できない場合があり、以下の方法が、少なくとも最終的には、係るサービス機能停止から回復するプロセスを提供する。例えば、ストレージシステムの第1の集合が、仲介サービスを通してストレージシステムの第2の集合をデタッチしようと試みるが、ストレージシステムの第1の集合が仲介サービス(1500)と通信できない場合、次いでストレージシステムの第1の集合は、デタッチ動作を完了できず、ポッドにサービスを提供し続けることができない。一部の 경우에는、ストレージシステムの2つの集合が互いとなんとか再接続し、これによりすべての同期中のストレージシステムが一ただし、仲介サービス(1500)は依然として利用できない状態で一再び通信している場合、ストレージシステムの2つの集合はポッドを同期させ、ポッドにサービスを提供することを再開してよい。しかしながら、この例では、パーティション識別子を変更するために、又はなんであれ仲介と関連付けられた他のプロパティを変更するために、1つ以上の要求が仲介サービス(1500)に送信された可能性があり、ストレージシステムのどれも、要求が受け取られたのか、それとも受け取られず、処理されなかったのかを確信し得ず、確認応答が失われた可能性がある。結果として、障害を有するストレージシステム又はネットワーク相互接続の集合がある場合、次いでストレージシステムは、仲介サービス(1500)がオンラインに戻る場合、及び戻るとき、パーティション識別子についてどの値をアサートするのか確信がもてない場合がある。係る状況では、ポッドのサービスが、すべての同期中のストレージシステムがオンラインに戻り、通信を再開するとき、又は同期中のストレージシステムが仲介サービス(1500)に再接続できるときのどちらかに再開することが好ましい。一実施態様では、すべての同期中のストレージシステムが再接続するとき、同期中のストレージシステムはすべて、仲介サービス(1500)に送信された可能性がある既知のパーティション識別子値を交換する。例えば、2つのストレージシステムがそれぞれパーティション識別子値を変更しようと試み、あるストレージシステムがパーティション識別子を、例えば1749137481890に変更しようと試み、別のストレージシステムが、パーティション識別子を、例えば87927401839に変更しようと試み、仲介サービス(1500)によって受取りが確認されたと知られている最後の値が79223402936であった場合、次いで仲介サービス(1500)は、現在、これらの3つのパーティション識別子の値のいずれかを記憶してよい。結果として、仲介パーティション識別子を新しい値に変更しようとする将来の試みは、変更を行うための権限を獲得しようとしてこれらの3つのパーティション識別子のいずれか又はすべてを供給する場合がある。さらに、パーティション識別子値を変更しようとする第4の試みも障害に遭遇し、後にさらに別の仲介を試みる任意のストレージシステムによって覚えられ必要がある場合がある第4の値を生じさせる場合がある。さらに、任意のストレージシステムが仲介サービス(1500)パーティション識別子値を無事に変更する場合、そのストレージシステムは、任意の同期中のストレージシステムから、及び将来同期中になる任意のストレージシステムからより古いパーティション識別子値をパージする場合がある。

【0286】

【0344】 別の例では、仲介サービス(1500)は、潜在的な将来の各競争に対して準備された一意の鍵に基づいて仲介してよい。係る場合、同期中のストレージシステムは、新しい鍵を使用することに同意してよい。すべての同期中のストレージシステムが新しい鍵を受け取り、記録するまで、新しい鍵が同時にすべてのストレージシステムでアトミックに設定されない場合があることを所与とすれば、すべてのストレージシステムは、任意の以前の仲介試行で各ストレージシステムが設定しようとしたその古い鍵及び値を保持するべきである。この例では、任意の早期の競争されなかった鍵及び任意の早期の鍵／値

10

20

30

40

50

の仲介試行はポッドのためのすべての同期中のストレージシステム間で回覧され、将来の仲介試行に使用する新しい鍵とともに、係る各ストレージシステムで記録されてよい。（新しい鍵を含まない）以前の競争されなかった鍵ごとに、この交換は、すべてのシステムがその鍵に対する競争で使用し得る単一の同意された値を選択してもよい。ポッドのためのすべての同期中のシステムがこれらの仲介の鍵及び値（及び任意の将来の競争のための新しい同意された鍵）のすべてを受け取り、記録した後、ポッド内のストレージシステムは、次いで単一の新しい鍵を選択して、より古い鍵を廃棄することに同意してよい。2つ以上のストレージシステムが同じ仲介鍵を異なる値に設定することを試みた可能性があり、すべての係る値が記録されてよいことに留意されたい。過去の仲介試行のためのこれらのすべての仲介鍵及び鍵／値対を交換する又は受け取るプロセスの間に障害がある場合、次いで他は受け取り、記録した可能性があるが、いくつかのストレージシステムは新しい仲介の鍵及び値を受け取らず、記録していない可能性がある。仲介サービス（1500）が、ポッドのためのすべての同期中のストレージシステムが互いと再接続できる前に利用可能になる場合、次いでポッドのためのストレージシステムの部分集合は、ポッドから別のストレージシステムをデタッチするために仲介サービス（1500）を使用しようと試みる場合がある。仲介を勝ち取るために、ストレージシステムは、すべての記録された鍵をその記録された値に設定しようと試み、それがうまくいく場合、次いで新しい鍵を一意的に設定しようと試みる場合がある。同じ鍵に対して2つ以上の値が記録された場合、次いでそのステップは、それらの値のいずれか1つを設定することが成功する場合、成功する。第1のステップ（以前の鍵を設定すること）が失敗する場合、又は第2のステップ（新しい鍵を新しい一意的な値に設定すること）が失敗する場合、次いで仲介に対するその試みに参加するストレージシステムはオフラインになる（それが新しい鍵について設定しようと試みた値を保持する）場合がある。両方のステップとも成功する場合、次いで通信中のストレージシステムは、通信していないストレージシステムをデタッチし、ポッドを提供し続けてよい。すべての過去の鍵及び値を交換することの代替策として、ストレージシステムは、ポッドのための他のストレージシステムからの鍵及び値の交換なしに、それが試す鍵及び値だけを記録してよい。次いで、同期中のストレージシステムが、（どれも仲介サービスと対話することに成功しなかった）ポッドのための他の同期中のストレージシステムと再接続する場合、同期中のストレージシステムは、1つの新しい仲介鍵を交換し、次いでそれらがともに同意された新しい鍵を受け取り、記録した旨の肯定応答を交換してよい。障害が肯定応答の交換を妨げる場合、次いで新しい鍵を受け取ったことがなかったストレージシステムによる（現在利用可能な仲介サービスに対する）仲介に対する将来の試行は、その以前の鍵及び値を再アサートしようと試みる場合がある。新しい鍵を受け取ったが、ポッドのためのすべてのストレージシステムが鍵を受け取った旨の表示を受け取らなかったストレージシステムは、新しい鍵のための値、最初に以前の鍵、次に新しい鍵をアサートするだけでなく、その以前の仲介鍵をアサートしてよい。その将来の仲介試行はまだ失敗する場合があるが、その結果ストレージシステムは再び他の同期中のストレージシステムに再接続する場合があり、別の鍵につながる新しい鍵を再び不完全に交換する場合がある。これが別の鍵を加える。鍵が、新しい鍵の一連の不完全な交換で経時的に増大するにつれ、ストレージシステムによる将来の仲介試行は、ストレージシステムがすべての鍵の値を無事にアサートする、又はストレージシステムが鍵のアサートの失敗に遭遇し、その時点でストレージシステムが鍵をアサートするのを停止し、オフラインになるまで、それらの鍵が記録された順序で、ストレージシステムがそれらの鍵のために以前にアサートした任意の値とともに、その鍵のそれぞれを再アサートしてよい。

【0287】

【0345】 別の例では、新しい仲介サービスは、現在の仲介サービスが利用できないときに構成されてよい。例えば、ポッドのためのすべての同期中のストレージシステムが互いと通信中であるが、現在の仲介サービスと通信していない場合、次いでポッドは、新しい仲介サービスで構成されてよい。これは、新しい鍵又は新しい仲介値を選択する以前のアルゴリズムに類似しているが、新しい鍵は、単に同じサービスと関連付けられた別の鍵で

10

20

30

40

50

あるよりむしろ新しい仲介サービスを使用するようにさらに構成される。さらに、この動作中に障害がある場合、以前のアルゴリズムと同様に、いくつかのシステムはより古い鍵のために競争する場合があります、したがって古い鍵と新しい仲介サービスでの新しい鍵を知っているシステムは、新しいメディーターサービスで新しい鍵のために競争する場合があります。以前の仲介サービスが恒久的に利用できない場合、次いですべての同期中のストレージシステムは、最終的に互いと再接続し、ポッドサービスが安全に再開できる前に、新しい仲介サービス、及び新しい仲介サービスと関連付けられた任意の鍵及び値の交換を完了すべきである。

【0288】

【0346】 別の例では、障害を解決するためのモデルは、一方のストレージシステムを他方のストレージシステムよりも支持するための優先規則を実施するためであってよい。この例では、好ましいストレージシステムが実行中である場合、ストレージシステムは実行中のままとなり、それが通信していない任意のストレージシステムをデタッチする。さらに、優先システムとの実績のある通信をしていない任意の他のシステムはそれ自体をオフラインにする。この例では、優先ストレージシステムが再接続するストレージシステムをデタッチしており、次いで再接続するストレージシステムが、それがポッドにサービスを提供することを再開できる前にポッドのためにそれを同期させるために最初に再同期されなければならないのに対し、非優先ストレージシステムは、最終的に優先ストレージシステムと再接続するとき、次いで優先ストレージシステムが再接続するストレージシステムをまだデタッチしていなかった場合、2つのストレージシステムは、両方のストレージシステムとも同期中である状態を回復し、該状態から再開してよい。優先ストレージシステムを有することは高可用性を提供するために有用ではない場合があるが、同期複製、特に非対称同期複製の他の用途には有用である場合がある。例えば、データセンタ又はキャンパス内の中心的大型ストレージシステムから、例えばトップオブブラック構成の等、アプリケーションサーバのより近くで実行中のより小型の（おそらくあまり管理されていない）ストレージシステムにミラーリングする例を挙げる。この場合、ネットワーク故障の場合又はトップオブブラックストレージシステムの故障時、集中管理されたストレージシステムが故障する場合にポッドのためのサービスを完全に停止させる一方、より大型のより集中管理されたストレージシステムの方をつねに支持することが有益である場合がある。係るトップオブブラックストレージシステムは、読取り性能を高めるため又はデータセンタストレージネットワークに対する負荷を削減するためだけに使用される可能性があるが、非同期複製又は他のデータ管理サービスが集中管理されたシステムだけで実行中である場合、トップオブブラックストレージシステムが単独で続行することを可能にするよりも、中心的大型ストレージシステムにトラフィックをルート変更する又はサービス提供を停止し、技術サポートを要求する方が好ましい場合がある。さらに、優先規則はより複雑である場合がある—おそらく優先ストレージシステム又は必須ストレージシステムに頼るなんらかの数の追加のストレージシステムとともに、2つ以上の結合された係る「優先」ストレージシステムがある場合がある。この例では、すべての優先ストレージシステム又は必須ストレージシステムが実行中である場合、ポッドはオンラインであり、そのうちのいくつかは実行中ではない場合、ポッドはダウンしている。これは、定足数のサイズが投票するメンバーの数と同じである定足数モデルに類似しているが、それは、すべてに満たない投票メンバーを可能にする一般化された定足数モデルよりも実装するのがより容易である。

【0289】

【0347】 別の例では、ポッドが3つ以上のストレージシステム全体で広げられるとき有用である場合がある機構の組合せが使用されてよい。一例では、優先規則はメタデータと結合されてよい。トップオブブラック例では、データセンタ又はキャンパス内のより大型の中心的大型ストレージシステムはそれ自体、第2の場所の大型ストレージシステムに同期複製される可能性がある。その場合、トップオブブラックストレージシステムは、絶対に単独で再開し得ず、2つの場所のより大型の中心的大型ストレージシステムのいずれかに優先権を与えてよい。その場合の2つのより大型のストレージシステムは、互いの間で仲介する

10

20

30

40

50

ように構成される可能性があり、オンラインに留まる2つのより大型のストレージシステムのどちらかに接続できる任意のより小型のストレージシステムはそのポッドにサービスを提供し続ける場合があり、2つの大型のストレージシステムのどちらにも接続できない（又は、ポッドのためにオフラインである一方にしか接続できない）任意のより小型のストレージシステムは、ポッドにサービスを提供するのを停止する場合がある。さらに、優先モデルは定足数ベースのモデルと結合されてもよい。例えば3つの場所の3つの大型ストレージシステムは、互いの中で定足数モデルを使用する可能性があり、より小型の衛星ストレージシステム又はトップオブブラックストレージシステムはいかなる票も欠き、それらが、オンラインであるより大型の同期中のストレージシステム的一方に接続できる場合のみ機能する。

【0290】

【0348】 機構を結合する別の例では、仲介は定足数モデルと結合されてよい。例えば、通常2つのストレージシステムが通信していない第3のストレージシステムを安全にデタッチできることを確実にするために、互いの中で投票する3つのストレージシステムがあってよい。一方、1つのストレージシステムは、2つの他のストレージシステムを独力で絶対にデタッチすることはできない。しかしながら、2つのストレージシステムが第3のストレージシステムを無事にデタッチした後に、構成はいま、それらが同期中であることに同意し、第3のストレージシステムがデタッチされる旨の事実に参加する2つのストレージシステムに少なくなる。その場合、残っている2つのストレージシステムは、（例えばクラウドサービスを用いた）仲介を使用して、追加のストレージシステム又はネットワークの障害を処理することに同意してよい。この仲介及び定足数の組合せは、さらに拡張され得る。例えば、4つのストレージシステム間で広げられるポッドでは、任意の3つが第4のストレージシステムをデタッチすることができるが、2つの同期中のストレージシステムが互いと通信しているが、他の2つのストレージシステムに通信していない場合、それらは両方とも同期中であると現在考え、次いでそれらは他の2つを安全にデタッチするために仲介を使用するであろう。5つのストレージシステムポッド構成でさえ、4つのストレージシステムが第5のストレージシステムをデタッチするために投票する場合、次いで残りの4つは、それらが等しい2つの半分に分割される場合仲介を使用でき、ポッドが2つのストレージシステムに少なくなると、それらは連続する障害を解決するために仲介を使用できる。5つから3つは、3つの間で定足数を使用し、次いで2つへの低下を可能にする可能性があり、残った2つのストレージシステムは、追加の故障がある場合、再び仲介を使用する。この一般的なマルチモード定足数仲介機構は、対称的なストレージシステムの間の定足数も、独自の仲介も処理できない追加の数の状況を処理できる。この組合せは、障害のある又はときおり到達不能なメディアータを確実に使用できる（又は、カスタマが彼らを完全に信用していない場合があるクラウドメディアータの場合）事例の数を増やし得る。さらに、この組合せは、仲介が単独である結果、第1のストレージシステムが、単に第1のストレージシステムにしか影響を及ぼさないネットワーク障害時に、無事に第2のストレージシステム及び第3のストレージシステムをデタッチする可能性がある、3つのストレージシステムポッドの場合をより良く処理する。また、この組合せは、3つから2つ、次いで1つへの例で説明されるように、一度に1つのストレージシステムに影響を及ぼす一連の障害をより良く処理し得る。これらの組合せは、同期中であること及びデタッチ動作が特定の状態を生じさせるために機能する一言い換えると、デタッチ済みから同期中に移動することがプロセスであり、一連の定足数／メディアータ関係での各段階は、あらゆる点で、すべてのオンライン／同期中ストレージシステムがポッドのための現在の持続性のある状態に参加していることを確実にするため、システムはステータフルである。これは、単に通信中のクラスタノードの過半数を再び有することが、動作を再開するために十分であると予想される、なんらかの他のクラスタモデルにおいてとは異なる。しかしながら、優先モデルは、依然として加えられてよく、衛星ストレージシステム又はトップオブブラックストレージシステムは、仲介にも定足数にも絶対に関与せず、それらが、仲介又は定足数に参加するオンラインストレージシステムに接続できる場合にだけ

10

20

30

40

50

ポッドにサービスを提供する。

【0291】

【0349】 一部の例では、仲介サービス（1500）又は外部ポッドメンバーシップマネージャは、同期複製されたストレージシステム（1514、1524）のための障害ゾーンとは異なる障害ゾーンに位置する場合がある。例えば、2つのストレージシステムポッド（1501）を用いる場合、2つのストレージシステム（1514、1524）が、例えば物理的な場所—一方が都市にあり、他方が都市の郊外にある、又は一方が1つの送電網又はインターネットアクセスポイントに接続されたデータセンタ内にあり、他方が異なる送電網又はインターネットアクセスポイントに接続された別のデータセンタ内にある—によって別個の障害ゾーンに分けられる場合、次いで、2つのストレージシステム以外のなんらかの他の障害ゾーンにすることが概して好ましい。一例として、仲介サービス（1500）は、都市の拡張された市街地の異なる部分にある場合がある、又は異なる送電網若しくはインターネットアクセスポイントに接続される場合がある。しかしながら、同期複製されたストレージシステムは、より良いストレージ信頼性を提供するために同じデータセンタの中にある場合もあり、この場合、ネットワーク、電力、及び冷却のゾーンが考慮に入れられてよい。

【0292】

【0350】 図15に示される例の方法は、トリガイイベントを検出することに応じて第1のストレージシステム（1514）によって、仲介サービス（1500）から仲介を要求すること（1502）を含む。この例では、トリガイイベントは、第1のストレージシステム（1514）と第2のストレージシステム（1524）との間のデータ通信リンク（1516）の通信障害である場合があり、障害を検出することは、ハードウェア故障がことに基づく、伝送の受取りを確認できないことに基づく、又はリトライ努力の失敗に基づく、又は他のなんらかの方法による場合がある。他の場合では、トリガイイベントは、同期複製リースの満了である場合があり、仲介を要求することは、接続を同期させること、及び活動リースの再開を調整しようと試みることの一部である場合がある。係るリースは、最初に、さまざまな異なる方法で複数のストレージシステムの少なくとも1つのためのタイミング情報に従って確立されてよい。例えば、ストレージシステムは、複数のストレージシステムのそれぞれがクロックを調整又は交換するためにタイミング情報を活用することによって同期複製リースを確立してよい。係る例では、いったんクロックがストレージシステムのそれぞれに調整されると、ストレージシステムは、調整又は交換されたクロック値を超えた所定の期間の間延長する同期複製リースを確立してよい。例えば、ストレージシステムごとのクロックが時間Xで調整される場合、ストレージシステムは、それぞれX+2秒有効である同期複製リースを確立するように構成されてよい。クロックを調整又は交換するための追加の説明は、その全体が参照により本明細書に援用される、米国仮特許第62／518,071号の中で見つけられてよい。

【0293】

【0351】 さらに、トリガイイベントを検出することに応じて第1のストレージシステム（1514）によって、仲介サービス（1500）から仲介を要求すること（1502）は、第1のストレージシステム（1514）のコントローラが、トリガイイベントを検出し、ネットワーク（1554）を介して仲介サービス（1500）に要求（1560）を送信することによって実施されてよい。一部の例では、仲介サービス（1500）は—複数のコンピュータシステムに対して—例えば値を記憶するための特定のデータベース入力等の相互に排他的なアクセスをリソースに提供する第3者のサービスであってよい。例えば、仲介サービス（1500）は、クラウドサービスプロバイダによって提供される、データベースを修正する要求を発行するホストコンピュータによって又はリソースに相互に排他的なアクセスを提供するなんらかの第3者サービスによって提供されるデータベースサービスによって提供されてよく、リソースは、特定のクライアントからの要求に基づいた特定の修正を示すことができるストレージ、状態機械、又は他のなんらかのタイプのリソースであってよい。この例では、仲介に対する要求（1560）を送信した後、第1のスト

レージシステム（１５１４）は、正の仲介結果（１５０３Ｂ）又は負の仲介結果又は応答の欠如（１５０３Ｃ）を示す仲介サービス（１５００）からの表示を待機する（１５０３Ａ）。第１のストレージシステム（１５１４）が負の仲介結果を受け取る又は仲介結果を受け取らない（１５０３Ｃ）場合、及び待機する閾値時間量を超えられていない場合、第１のストレージシステム（１５１４）は、より多くの時間待機し続けてよい（１５０６）。しかしながら、待機時間量が閾値量を超える場合、次いで第１のストレージシステム（１５１４）は、別のコンピュータシステムが仲介を勝ち取ったと判断し、それ自体をオフラインにすることによって続行してよい（１５０６）。一部の例では、上述されたように、仲介に対する要求は、やはりポッド（１５０１）を維持するストレージシステムの別のストレージシステムから受け取られたコンペアアンドセット動作のターゲットであってよい共用リソース（１５５２）の値を設定しようと試みるアトミックコンペアアンドセット動作として仲介サービス（１５００）によって受け取られてよく、共用リソース（１５５２）を無事に設定するストレージシステムが仲介を勝ち取る。

【０２９４】

【０３５２】 また、図１５の例は、第２のストレージシステム（１５２４）が、トリガイベントを検出することに応じて仲介サービス（１５００）から仲介を要求する（１５１０）も含む。リガイベントを検出することに応じて仲介サービス（１５００）から仲介を要求する（１５１０）は、トリガイベントに応じて、第１のストレージシステム（１５１４）に対する仲介を要求すること（１５０２）の実施態様と同様に実装されてよい。しかしながら、この例では、第２のストレージシステム（１５２４）は、仲介サービスに要求（１５６２）を送信したことに応じて一第１のストレージシステム（１５１４）の仲介成功とは逆に一失敗メッセージ、又は仲介に対する要求（１５６２）が成功しなかった旨のなんらかの表示を受け取ってよい。

【０２９５】

【０３５３】 図１５の例の方法は、万一正の仲介結果の表示（１５６４）が第１のコンピュータシステム（１５１４）によって受け取られる場合には、仲介サービス（１５００）からの正の仲介結果の表示（１５６４）に応じて、第１のコンピュータシステム（１５１４）が一第２のストレージシステム（１５２４）の代わりに一第１のストレージシステム（１５１４）及び第２のストレージシステム（１５２４）全体で同期複製されるデータセット（１５１２）に向けられるデータストレージ要求を処理すること（１５０４）によって続行する。ポッド（１５０１）を実装するデータセット（１５１２）の同期複製は、データセット（１５１２）に向けられるデータストレージ要求を受け取り、処理することに加えて、その全体が参照により本明細書に援用される米国仮特許第６２／４７０，１７２号及び第６２／５１８，０７１号の図８Ａ及び図８Ｂに関して説明されるように実施されてよい。この例では、図１５に関して上述されたように、正の仲介結果の表示（１５６４）に応じて、第１のストレージシステム（１５１４）は、仲介を勝ち取るストレージシステムと見なされてよく、第１のストレージシステム（１５１４）は、通信が失われたストレージシステムをデタッチしてよい。しかしながら、他の例では、仲介は、説明された仲介の他の方法又は仲介の方法の組合せのいずれかに従って実施されてよい。

【０２９６】

【０３５４】 一部の例では、データセット（１５１２）を同期複製する複数のストレージシステムの間どのストレージシステムが仲介を勝ち取るのかの優先を定義することは、複数のストレージシステムのそれぞれに遅延値を指定することによって実施されてよい。例えば、第１のストレージシステム（１５１４）が優先ストレージシステムとして指定される場合、次いで第１のストレージシステム（１５１４）は仲介サービスから仲介に対する要求を行う前にゼロ（０）の遅延値を割り当てられてよい。しかしながら、非優先ストレージシステムの場合、遅延値は、例えば３秒、又は概して優先ストレージシステムが、単に同期複製されたストレージシステム間の通信損失のために仲介を勝ち取ることになるであろうなんらかの他の値等、ゼロより大きくなるように割り当てられてよい。

【０２９７】

【0355】 追加の説明のために、図16は、本開示のいくつかの実施形態に従ってデータセットを同期複製するストレージシステム間で仲介するための例の方法を示すフローチャートを説明する。図16に示される例の方法は、トリガイベントを検出することに応じて第1のストレージシステム(1514)によって、仲介サービス(1500)から仲介を要求すること(1502)、及び仲介サービス(1500)からの正の仲介経過の表示(1564)に応じて、第1のコンピュータシステム(1514)が—第2のストレージシステム(1524)の代わりに—第1のストレージシステム(1514)及び第2のストレージシステム(1524)全体で同期複製されるデータセット(1512)に向けられたデータストレージ要求を処理すること(1504)も含むので、図16に示される例の方法は図15に示される例の方法に類似している。

10

【0298】

【0356】 しかしながら、図16に示される例の方法は、正の注記結果の表示(1564)に応じて、データセット(1512)を同期複製する複数のストレージシステム(1514、1524)から第2のストレージシステム(1524)をデタッチすること(1602)をさらに含む。別のストレージシステムをデタッチすること(1602)は、データセット(1512)を複製しているストレージシステムの同期中リストからもはや通信していないストレージシステム(1524)を削除することによって、仲介サービス(1500)から正の仲介結果の表示を受け取るストレージシステムで実施されてよく、同期中リストからの削除により、仲介を勝ち取るストレージシステム(1514)は、後に受け取られたデータセットを修正する要求のためにデタッチされたストレージシステムを同期させることを試みない。この例では、2つのストレージシステム(1514、1524)があるが、他の例では、他の数量のストレージシステムが意図される。

20

【0299】

【0357】 追加の説明のために、図17は、本開示のいくつかの実施形態に従ってデータセットを同期複製するストレージシステム間で仲介するための例の方法を示すフローチャートを説明する。図17に示される例の方法も、トリガイベントを検出することに応じて第1のストレージシステム(1514)によって、仲介サービス(1500)から仲介を要求すること(1502)を含むので、図17に示される例の方法は図15に示される例の方法に類似しており、この例ではトリガイベントは通信障害である。

【0300】

30

【0358】 しかしながら、図17に示される例の方法は、第2のストレージシステム(1524)によって実行される任意の活動又はアクションを含まないため、図17に示される例の方法は、図15に示される例の方法とは異なる。この違いは、複数のストレージシステムの間でストレージシステムが故障する、又はそれ以外の場合反応しないこと、及び1つ以上の他のストレージシステムが、同期複製されたデータセット(1512)に向けられたデータストレージ要求にサービスを提供し続けるために仲介サービス(1500)からの仲介を要求することを可能にする。

【0301】

【0359】 図17に示される例の方法は、第1のストレージシステム(1514)と第2のストレージシステム(1524)との間の通信に障害を検出すること(1702)を含み、第1のストレージシステム(1514)及び第2のストレージシステム(1524)は、データセット(1512)を同期複製するストレージシステムに含まれる。通信障害を検出すること(1702)は、図15に関して上述されるように実施されてよい。

40

【0302】

【0360】 さらに、図17に示される例の方法は、正の仲介結果の表示(1564)に応じて、データセット(1512)を同期複製する複数のストレージシステム(1514、1524)から第2のストレージシステム(1524)をデタッチすること(1704)も含む。第2のストレージシステム(1524)をデタッチすること(1704)は、図16に関して説明された第2のストレージシステム(1524)をデタッチすること(1602)と同様に実施されてよい。

50

【0303】

【0361】 追加の説明のため、図18は、本開示のいくつかの実施形態に従ってデータセットを同期複製するストレージシステムの回復のための例の方法を示すフローチャートを説明する。図18に示される例の方法は、データセット(1812)が、そのそれぞれが1つ以上のデータ通信リンク(1816、1818、1820)を介して互いに独立して結合されてよい、2つのストレージシステム(1814、1824、1828)だけで同期複製される実施形態を示すが、図18に示される例は、データセット(1812)が追加のストレージシステム全体で同期複製される実施形態に拡張できる。

【0304】

【0362】 データセット(1842)を同期複製している複数のストレージシステム(1814、1824、1828)は、ホスト(1802)コンピューティングデバイスから要求(1804)を受け取り、処理するための通常の動作中に互いと通信してよい。しかしながら、一部の例では、ストレージシステム(1814、1824、1828)の1つ以上は、失敗する、再起動する、アップグレードする、又はそれ以外の場合利用できない場合がある。この状況での回復は、障害又はなんらかの他のサービス機能停止が、同期中のストレージシステムのうちの少なくとも1つに中断させ、おそらく飛行中(in-flight)動作の文脈を失わせた後に同期中のポッドメンバーストレージシステムを一貫させるプロセスである。『ポッド』は、該用語がここで及び本願の残りを通して使用されるように、データセット、管理オブジェクト及び管理動作の集合、データセットを修正する又は読み取るためのアクセス動作の集合、及び複数のストレージシステムを表す管理エンティティとして実施されてよい。係る管理動作は、データセットを読み取る又は修正するためのアクセス動作が、ストレージシステムのいずれかを通して同等に動作する、ストレージシステムのいずれかを通して同等に管理オブジェクトを修正する又は問い合わせしてよい。各ストレージシステムは、ストレージシステムによる使用のために記憶され、宣伝されたデータセットの適切な部分集合としてデータセットの別々のコピーを記憶してよく、任意の1つのストレージシステムを通して実行され、完了された管理オブジェクト又はデータセットを修正する動作は、ポッドを問い合わせるために後続の管理オブジェクトで、又はデータセットを読み取るために後続のアクセス動作で反映される。『ポッド』に関する追加の詳細は、参照により本明細書に援用される以前に出願された仮特許出願第62/518,071号で見つけられてよい。この例では、示されている3つのストレージシステム(1814、1824、1828)しかないが、概して任意の数のストレージシステムが、データセット(1812)を同期複製している同期中リストの一部であってよい。

【0305】

【0363】 ポッドのメンバーである任意の1つ以上のストレージシステムが中断されると、次いで任意の残りのストレージシステム又は早期に動作を再開する任意のストレージシステムは、(それらがもはや同期中にならないように)それらをデタッチする、又は先に進む前にそれらを待機し、一貫性を確実にするために回復アクションに関与するかのどちらかであってよい。機能停止が十分に短く、回復が十分に迅速である場合、次いでストレージシステムの外部の又はアプリケーション自体を停止させるように障害を起こさない(fault)ストレージシステム上で実行中であるオペレーティングシステム及びアプリケーションは、ストレージ動作処理で一時的な遅延を経験する場合があるが、サービス停止は経験しない場合がある。SCSI及び他のストレージプロトコルは、一時ストレージコントローラ又はインタフェースターゲットコントローラの機能停止により動作が失われた場合、代替ターゲットストレージインタフェースを含むリトライをサポートし、特にSCSIは、ストレージコントローラが回復に関与する間使用できるであろう開始プログラムリトライを要求するBUSYステータスをサポートする。

【0306】

【0364】 一般的には、回復の目標の1つは、進行中の分散動作の予想外の途絶からの任意の不一致を処理すること、及び同期中のポッドメンバーストレージシステムを十分に同

10

20

30

40

50

一にすることにより不一致を解決することである。その時点で、ポッドサービスを提供することは安全に再開できる。十分に同一は、少なくともポッドに記憶されるコンテンツを含み、他の場合では、十分に同一は持続する予約の状態を含んでよい。また、十分に同一は、スナップショットに一貫性がある一完了した修正動作、同時に起こる修正動作、又はより最近に受け取られた修正動作に関しても正しいか、又は一貫して削除されるかのどちらかであることを確実にすることも含んでよい。実施態様に応じて、一貫性をもたせられるべきである他のメタデータがある場合がある。複製ソースから非同期複製ターゲット又はスナップショットベースの複製ターゲットへのコンテンツの移動を追跡又は最適化するために使用されるメタデータがある場合、次いでそれは、複製ソースがポッドのあるメンバーストレージシステムから別のメンバーストレージシステムに円滑に切り替わることを可能にするために一貫性をもたせられる必要がある可能性がある。また、ボリュームの存在及びプロパティも回復される必要がある場合があり、おそらくアプリケーション又は開始するホストシステムに関係する定義も回復される必要がある場合がある。これらのプロパティの多くは、それらがどのように実装されるのかに応じて、標準的なデータベーストランザクション回復技術を使用し、回復されてよい。

【0307】

【0365】 一部の例では、管理メタデータが、ブロックベースのストレージシステムのコンテンツに対して修正動作を実施するストレージシステムで十分に同一であることを確実にすることを超えて、回復は、それらの修正がポッド全体で、及びブロックストレージセマンティックス（COMPARE AND WRITE 及び XDWRITE READ 等の動作の順序、同時並列性、一貫性、アトミック性）を適切に考慮して一貫して適用又は廃棄されることを確実にしなければならない。根本的には、本実施態様は、回復中に、どの動作がポッドのための他のすべての同期中のストレージシステムには適用されなかった可能性がある、ポッドのための少なくとも1つの同期中のストレージシステムに適用された可能性があるのかを知ることができること、及びそれらをどこでも適用する又は取り消すことのどちらかに頼っている。どちらのアクションも一貫性一どこでも適用する又はどこでも取り消す一を生じさせ、すべての動作全体で答えが一樣でなければならない固有の理由はない。取消しは、ポッドのための少なくとも1つの同期中のストレージシステムが動作を適用しなかった場合に許されてよい。一般に、多くの場合、ポッドのための1つ以上の同期中のストレージシステム上にあるいくつかの又はすべての更新を取り消すよりも、ポッドのための任意の同期中のストレージシステムで発見されたが、ポッドのためのすべての同期中のストレージシステム上にはないすべての更新を適用することについて推論の方が簡略である。効率的であるためには、他のシステムでは適用されなかった可能性があるいくつかのシステムで何が適用されたのかを知ることが、概してストレージシステムが未処理データ以外の何かを記録することを必要とする（それ以外の場合、すべてのデータは、比較されなければならない可能性があり、それは桁外れに多大な時間を要するであろう）。以下に説明されるのは、ストレージシステム回復を可能にし得る係の情報記録するための実施態様に関する追加の詳細である。

【0308】

【0366】 一貫性を確実にするための情報を持続的に追跡するための2つの例は、（1）ボリュームのコンテンツが、ポッドのための同期中のストレージシステム全体で異なる可能性があることを識別すること、及び（2）ポッドのためのすべての同期中のシステム全体で普遍的に適用されなかった可能性がある動作の集合体を識別することを含む。第1の例は、ミラーリングのための従来のモデルである。書き込まれている（多くの場合、リストとして又はなんらかの粒度を有するボリュームの論理空間をカバーするビットマップとして）論理領域の追跡マップを保ち、回復中にそのリストを使用して、どの領域があるコピーと別のコピーとの間で異なる可能性があるのかに留意する。追跡マップは、追跡マップの回復が、障害時に流動的であった任意のボリューム領域をカバーすることを保証されるように、ボリュームデータの書き込み前又は中にいくつか又はすべてのミラーに書き込まれる（又は別々に書き込まれる）。この第1の変形形態での回復は、概してコンテンツを

あるコピーから別のコピーにコピーして、それらが同じであることを確認することから成る。

【0309】

【0367】 この場合は、単に同期複製されたストレージシステム間のボリュームコンテンツの潜在的な差異として追跡することはより困難又は高価である場合があるので一動作追跡に基づく一持続性追跡の第2の例は、ポッド内のボリュームの中及び間での大きいボリューム範囲の仮想コピーを同期複製することをサポートするストレージシステムで有用である場合がある（ただし、コンテンツアドレス指定可能ストレージシステムでの追跡及び回復を説明する項を参照すること）。また、簡略なコンテンツ追跡は、例えば非同期複製の形式を駆動するエクステント及びより大きい粒度の識別子を有するコンテンツ追跡グラフにおいて等、同期複製がより複雑な情報を追跡しなければならない、非同期複製ソースが、ポッド内のある同期中のシステムから別のストレージシステムに移行される又は障害を起こされる（*faulted over*）ストレージシステムではあまりうまく機能しない可能性がある。コンテンツの代わりに動作が追跡されるとき、回復は、どこでも完了していない可能性がある動作を識別することを含む。係る動作が識別されると、まさにそれらは例えばリーダー定義の順序付け又はプレディケート又はインターロック例外を通して等の技術を使用する通常の実行時中であるべきなので、任意の順序付け一貫性問題が解決されるはずである。インターロック例外は後述され、プレディケートに関して、動作と共通メタデータ更新との関係の記述は、別々の修正動作間の相互依存性の集合として説明される場合がある—これらの相互依存性は、ある動作がなんらかの方法で依存するプレカーソルの集合として記述される場合があり、プレカーソルの集合は、動作が完了するために真でなければならないプレディケートと見なされる場合がある。この例を続行するために、識別された動作を所与として、動作は次いで再適用されてよい。動作について記録される情報は、ポッドメンバーストレージシステム全体で一貫性がなければならない任意のメタデータ変化を含むべきであり、この記録された情報は、コピーされ、適用される場合がある。さらに、プレディケートは、それらがリーダーとフォロワーとの間の同時並行性に対する制限を広めるために使用される場合、それらのプレディケートが、ストレージシステムが情報を持続する順序を動かすならば持続される情報は多様なもっともらしい結論を暗示するので、保護される必要はない可能性がある。

【0310】

【0368】 米国仮特許出願第62／470，172号及び米国仮特許出願第62／518，071号、つまりその全体として本明細書に援用される参照の中でより徹底的に説明されるように、同期中のストレージシステムの集合は、データ一貫性を提供するための対称的なI/Oモデルを実施してよい。対称的なI/Oモデルでは、複数のストレージシステムは、ポッドの中にデータセットを維持してよく、I/O動作を受け取るメンバーストレージシステムは、ポッド内の他のすべてのストレージシステムでのI/O動作の処理と同時に、ローカルにI/O動作を処理してよい—受け取るストレージシステムは、他のストレージシステム上でI/O動作の処理を開始してよい。しかしながら、一部の 경우에는、複数のストレージシステムは、重複するメモリ領域に書き込む独立したI/O動作を受け取ってよい。例えば、第1の書込みが第1のストレージシステムに入る場合、次いで第1のストレージシステムは第1の書込みを第2のストレージシステムに送信しつつも、第1の書込みをローカルに持続し始めてよい—ほぼ同時に第1の書込みと重複するボリューム領域に対する第2の書込みは、第2のストレージシステムで受け取られ、第2のストレージシステムは、第2の書込みを第1のストレージシステムに送信しつつも、第2の書込みをローカルに持続し始める。この状況では、なんらかの時点で、第1のストレージシステム、第2のストレージシステムのどちらか、又は両方のストレージシステムは、同時に起こる重複があることに気づく場合がある。さらに、この状況では、第1の書込みは、第2のストレージシステムが第1の書込みを持続し、成功表示で応答し、且つ第1のストレージシステムが無事に第1の書込みを持続するまで第1のストレージシステムで完了できない—第2のストレージシステムは、第2の書込みと同様の状況にある。両方のストレージシ

システムとも第1の書込みと第2の書込みの両方にアクセスできるため、どちらのストレージシステムも同時に起こる重複を検出し得、1つのストレージシステムが同時に起こる重複を検出すると、ストレージシステムは、本明細書では「インターロック例外」と呼ばれる例外をトリガしてよい。状況が追加のストレージシステムまで拡大されるとき、1つの解決策は、2つ又はおそらくそれ以上のストレージシステム、書込み動作が優先する同意に到達するためのインターロック例外に関与するストレージシステムを含む。

【0311】

[0369] 重複する書込み要求の場合でのような別の例では、複製を中断し、最終的な回復につながったイベントの時点での時間及びボリュームアドレス範囲において重複していた書込みタイプ要求（例えば、WRITE要求、WRITE SAME要求、UNMAP要求、又は組合せ）は、同期中のストレージシステム間で一貫性なく完了した可能性がある。この状況がどのようにして処理されるのかは、通常の動作中のI/O経路の実装に依存する場合がある。この例では、以下にさらに時間で重複していた第1の書込み及び第2の書込みが説明され、それぞれが、どちらも完了したとして信号で知られる前に、ポッドのためのあるストレージシステムによって又は別のストレージシステムによって受け取られていた。この例は、2つずつを順番に考慮することによって2つ以上の書込みに、並びに第1の書込み及び第2の書込みが複数のストレージシステムで完了した可能性があることを考慮することによって、並びに第1の書込み、第2の書込み、及び第3の書込み（又は追加の書込み）が3つ以上のストレージシステムで一貫性なく完了した可能性があることを考慮することによって2つ以上のストレージシステムに容易に拡張される。説明された技術は、これらの場合に容易に拡張される。インターロック例外に基づいた対称的なI/Oベースのストレージシステム実施態様では、第1の書込みだけが1つのストレージシステムで完了した可能性がある。一方、2つの重複する書込みの第2の書込みは第2のストレージシステムで完了した可能性がある。この場合は、範囲が各書込みの間で重複することに気づくことによって、及びどのストレージシステムも代替の重複書込みを含まないことに気づくことによって検出できる。2つの書込みが完全に重複する（一方が他方を完全にカバーする）場合、次いで2つの書込みのうちの一方は他方のストレージシステムにコピーされ、そのボリュームアドレス範囲のためにそのストレージシステムのコンテンツを置換するために適用されるにすぎない場合がある。書込みが部分的にしか重複していない場合、次いで部分的に重複するコンテンツは、一方のストレージシステムから他方にコピーされ（適用され）る場合がある。一方、重複しない部分は、コンテンツが両方のストレージシステムで統一され、最新にされるように各ストレージシステム間でコピーできる。プレディケート、又はリーダーが、ある書込みが別の書込みに先行することを宣言するためのなんらかの他の手段を有するリーダーベースのシステムでは、書込みを実行しているストレージシステムは、重なり合うようにうまく持続してよい、又は2つをともに持続してよい。別の場合には、実施態様は2つの書込みを別々に及び順不同で持続してよく、順序付けのプレディケートは単に完了シグナリングを制御するためだけに使用される。実施態様が順不同の書込み処理を可能にする場合、次いで先行する例は、どのようにして一貫性を回復できるのかを説明する。ストレージシステムが通常動作中の持続性の順序付けを施行する場合には、次いで回復は、第2のストレージシステムでの第1の書込み及び第2の書込みを除き、第1のストレージシステムでは第1の書込みしか見ない可能性がある。その場合、第2の書込みは、回復の一部として第2のストレージシステムから第1のストレージシステムにコピーできる。

【0312】

[0370] 別の例では、スナップショットも回復され得る。リーダー決定のいくつかの修正がスナップショットに含まれるべきであり、他は含まれるべきではない、修正と同時に起こるスナップショットの場合等の一部の場面では、記録された情報は、特定の書込みがスナップショットの中に含まれるべきかどうかについての情報を含む可能性がある。そのモデルでは、リーダーがスナップショットに含めることを決定したすべてが結局回復後のスナップショットに含まなければならないことを確実にすることは必要ではない場合が

10

20

30

40

50

ある。ポッドのための1つの同期中のストレージシステムがスナップショットの存在を記録し、ポッドのための同期中のストレージシステムが、スナップショットでの包含のために順序付けられた書込みを記録しなかった場合、次いでその書込みを含めずにスナップショットを一樣に適用しても、ポッドのためのすべての同期中ストレージシステム全体で完全に一貫性があるスナップショットコンテンツが生じる。この不一致は、一度も完了されたとして信号で知らされたことがなかった同時に起こる書込み及びスナップショットの場合にのみ起こるべきであり、したがって包含保証は保証されない。つまり、リーダーがプレディケートを割り当て、順序付けすることは、回復順序一貫性のためよりむしろ実行時一貫性のためにだけ必要となる場合がある。回復識別子が、スナップショットでの包含のための書込みを識別するが、回復が書込みを位置決めしない場合、スナップショット動作自体が、実施態様に応じてスナップショットを安全に無視する可能性がある。スナップショットについての同じ議論は、SCSI EXTENDED COPY及び類似した動作によるボリュームアドレス範囲の仮想コピーにも適用する。つまり、リーダーは、ソースアドレス範囲に対するどの書込みが論理的にコピーに先行する可能性があるのか、及びターゲットアドレス範囲に対するどの書込みがアドレス範囲コピーに論理的に先行又は後続する可能性があるのかを定義する。しかしながら、回復中、同じ議論は、スナップショットと同様に適用する。つまり、結果がポッドのための同期中のストレージシステム全体で一貫性があり、どこでも完了した修正を後退させず、データセット消費者が読み取った可能性がある修正を逆にしないのであるならば、ボリューム範囲コピーと同時に起こる書込みは、同時に起こる書込み又はボリューム範囲コピーのどちらかを見逃す場合があるであろう。

10

20

【0313】

【0371】 さらにスナップショットの回復を説明するこの例に関して、任意のストレージシステムがCOMPARE AND WRITEのための書込みを適用した場合、次いで比較は、ポッドのための1つの同期中のストレージシステムで成功したに違いなく、実行時一貫性は、比較がポッドのためのすべての同期中のストレージシステムで成功すべきであったことを意味すべきであったため、任意の係るストレージシステムが書込みを適用した場合、それは、回復前にそれを適用していなかった任意の他のポッドのための同期中のストレージシステムにコピーされ、適用される場合がある。さらにまだ、XDWRITE READ又はXPWRITE要求（又は既存のデータと新しいデータとの間の類似した算術変換演算）の回復は、あるストレージシステムから生じる読取り後に変換の結果を送達することによって動作する場合もあれば、それは、変換する書込みに先行する任意の順序付けデータが、ポッドのための同期中のストレージシステムの全体で一貫性があることが確実にできる場合、及びどの係るストレージシステムがまだ変換する書込みを適用していなかったのかを確実に判断できる場合、変換する書込みをデータとともに他のストレージシステムに送達することによって動作できる。

30

【0314】

【0372】 別の例として、メタデータの回復が実施されてよい。この場合には、回復は、ポッドのための同期中のストレージシステムの間でのメタデータの一貫性のある回復も生じさせるべきであり、そのメタデータは、ポッド全体で一貫していると予想される。このメタデータが動作と含まれるのであるならば、これらは、それらの動作によって説明されるコンテンツ更新とともに適用できる。このデータが既存のメタデータとどのようにしてマージされるのかは、メタデータ及び実施態様に依存する。非同期複製を駆動するためのより長期の変更追跡情報は、多くの場合近傍として非常に簡略にマージできる、又はそれ以外の場合、関係する修正が識別される。

40

【0315】

【0373】 別の例として、動作追跡のための最近の活動を記録することは、障害又は回復につながった他のタイプのサービス中断時にポッド内の同期中のストレージシステムで進行中であつた動作を識別するための多様な方法で実施されてよい。例えば、1つのモデルは、ポッドの中の同期中の各ストレージシステムに対する修正の回復情報を、（更新が高

50

速ジャーナルデバイスを通して段階分けされる場合にうまく機能できる) 任意の修正でアトミックに、又は動作が発生する前に進行中である動作についての情報を記録することによってのどちらかで記録するものである。記録された回復情報は、例えば元の要求に基づいた、又は動作を記述することの一部としてリーダーによって割り当てられるなんらかの識別子に基づいた等の論理動作識別子を含んでよく、どのようなレベルの動作記述も回復が動作するために必要とされる場合がある。同時に起こるスナップショットのコンテンツに含まれる書込みのためにストレージシステムによって記録された情報は、書込みが適用されるボリュームのコンテンツにだけではなくスナップショットにも、書込みが含まれることを示すべきである。いくつかのストレージシステム実施態様では、スナップショットのコンテンツは、より新しいスナップショットの特定の重複コンテンツによって置き換えられられない、又はライブでボリュームに後書き込まれる特定の重複コンテンツによって置き換えられない限り、ボリュームのコンテンツに自動的に含まれる。時間で及びボリュームアドレスで重複する2つの同時に起こる書込みタイプ要求(例えばWRITE要求、WRITE SAME要求、又はUNMAP要求、又は組合せ)は、リーダーが、第2の書込みがポッドのための任意の同期中のストレージシステムによって持続できる前に、第1の書込みがポッドのためのすべての同期中のストレージシステムに最初に持続されることを確実にするように、リーダーによって明示的に順序付けされてよい。これは、不一致が発生しないことを簡単に確実にする。さらに、ボリュームに対する同時に起こる重複書込みは非常にまれであるので、これは許容可能である場合がある。その場合、第2の書込みのために任意の回復するストレージシステムに関するレコードがある場合、次いで第1の書込みは、それが回復を必要とするべきではないようにどこでも完了したに違いない。代わりに、プレディケートは、リーダーが、ストレージシステムが第2の書込みの前に第1の書込みを、順序付けることを必要とすることによって記述されてよい。ストレージシステムは、次いで両方の書込みをともに実行してよく、これによりそれらはともに持続する又はともに持続できないかのどちらかとなるように保証される。別の場合には、ストレージシステムは、第1の書込みの持続性が保証された後、第1の書込み、次いで第2の書込みを持続してよい。COMPARE AND WRITE要求、XDWRITE要求、AD要求、又はXPWRITE要求は、それぞれが動作を実行するとき、プレカーソルコンテンツがすべてのストレージシステムで同一となるように順序付けされるべきである。代わりに、1つのストレージシステムは結果を計算し、すべてのストレージシステムに通常書込みタイプ要求として要求を送達する可能性がある。さらに、これらの動作を回復可能にすることに関して、どの動作がどこでも完了したのかを追跡することは、その新近性を軽視することを可能にしてよく、完了した動作のための動作回復分析を生じさせる記録された情報は、次いで回復によって廃棄されるか、又は効率的に無視されるかのどちらかの場合がある。

【0316】

【0374】 別の例では、完了した動作を捨てることが実施されてよい。記録された情報のクリアを処理するための1つの例は、動作が、ポッドのためのすべての同期中のストレージシステムで処理されたと知られた後、それをすべてのストレージシステム全体でクリアすることである。これは、要求を受け取り、完了を信号で知らせたストレージシステムに、完了が信号で知らされた後にポッドのためのすべてのストレージシステムにメッセージを送信させ、各ストレージシステムがそれらを捨てることを可能にすることによって実施できる。回復は、次いで回復に関与するポッドのためのすべての同期中のストレージシステム全体で捨てられなかったすべての記録された動作について問い合わせることを含む。代わりに、これらのメッセージは、それらが周期的に(例えば、50msごと)、又はなんらかの数の動作(例えば、10から100ごと)の後に起こるようにバッチ処理されるであろう。より完全に完了した動作が潜在的に不完全として報告されるので、このバッチ処理プロセスは、いくぶん増加した回復回数を犠牲にしてメッセージトラフィックを大幅に削減し得る。さらに、(例として)リーダーベースの実施態様では、リーダーは、どの動作が完了しているのかを認識させられ、リーダーはクリアメッセージを送出できるで

10

20

30

40

50

あろう。

【0317】

〔0375〕 別の例では、スライドウィンドウが実装されてよい。係る例は、リーダー及びフォロワーに基づいた実施態様でうまく機能してよく、リーダーは、動作又は動作の集合体にシーケンス番号を付けてよい。このようにして、リーダーがあるシーケンス番号までのすべての動作が完了したことを判断することに応じて、リーダーはすべての同期中のストレージシステムに、そのシーケンス番号までのすべての動作が完了していることを示すメッセージを送信してよい。また、シーケンス番号は任意の番号でもあり、これにより任意の番号と関連付けられたすべての動作が完了したとき、すべてのそれらの動作が完了したことを示すためにメッセージが送信される。シーケンス番号をベースにしたモデルを用いると、回復は、最後に完了したシーケンス番号より大きいシーケンス番号と関連付けられた任意の動作中のストレージシステムに対するすべての動作について問い合わせるであろう。リーダーがない対称的な実施態様では、ポッドに対する要求を受け取る各ストレージシステムは、独自のスライドウィンドウ及びスライドウィンドウアイデンティティ空間を定義できるであろう。その場合、回復は、そのスライドウィンドウアイデンティティが、完了している最後のアイデンティティの後である任意のスライドウィンドウアイデンティティ空間と関連付けられる任意の同期中のストレージウィンドウでのすべての動作について問い合わせることを含んでよく、すべての先行する識別子の動作も完了している。

10

【0318】

〔0376〕 別の例では、チェックポイントが実装されてよい。チェックポイントモデルでは、プレカーソル動作の一樣な集合の完了に依存し、すべての連続する動作が次いで依存する特別な動作が、リーダーによって挿入されてよい。各ストレージシステムは、次いで持続又は完了されたすべてのプレカーソル動作に応じてチェックポイントを持続してよい。連続チェックポイントは、以前のチェックポイントが、ポッドのためのすべての同期中のストレージシステムで持続されたとして信号で知らされた後のあるときに開始されてよい。連続チェックポイントは、すべてのプレカーソル動作がポッド全体で持続された後のあるときまでこのようにして開始されないであろう。それ以外の場合、以前のチェックポイントは、完了しないであろう。このモデルでは、回復は、以前から最後のチェックポイントの後に続くすべての同期中のストレージシステムですべての動作について問い合わせることを含んでよい。これは、ポッドのための任意の同期中のストレージシステムに既知の第2から最後のチェックポイントまでを識別することによって、又は各ストレージシステムに第2から最後の持続されたチェックポイント以降のすべての動作を報告するように頼むことによって達成されるであろう。代わりに、回復は、すべての同期中のストレージシステムで完了したと知られている最後のチェックポイントを検索することを含んでよく、任意の同期中のストレージシステムで続くすべての動作について問い合わせることを含んでよい。チェックポイントがすべての同期中のストレージシステムで完了した場合、次いでそのチェックポイントの前のすべての動作は、どこでも明確に持続されていた。

20

30

【0319】

〔0377〕 別の例では、論理エクステンツの複製された有向非巡回グラフに基づくポッドの回復が実施されてよい。しかしながら、係る実施態様を記述する前に、論理エクステンツの有向非巡回グラフを使用するストレージシステムが最初に説明される。

40

【0320】

〔0378〕 ストレージシステムは、論理エクステンツを含んだ有向非巡回グラフに基づいて実装されてよい。このモデルでは、論理エクステンツは2つのページ、つまりなんらかの方法で記憶されているなんらかの量のデータを参照するリーフ（leaf）論理エクステンツ、及びその他のリーフ論理エクステンツ又は複合論理エクステンツを参照する複合論理エクステンツに分類される場合がある。

【0321】

〔0379〕 リーフエクステンツは、さまざまな方法でデータを参照できる。リーフエクステンツは、記憶されたデータの単一の範囲（例えば、64キロバイトのデータ）を直接的

50

に指す場合もあれば、リーフエクステントは記憶されたデータに対する参照の集合体（例えば、物理的に記憶されたブロックに対する範囲と関連付けられたなんらかの数の仮想ブロックをマッピングするコンテンツの1メガバイト「範囲」）である場合もある。後者の場合、これらのブロックは、なんらかのアイデンティティを使用し、参照されてよく、エクステントの範囲内のいくつかのブロックは、何にもマッピングされ得ない。また、その後者の場合、これらのブロックリファレンスは一意である必要はなく、なんらかの数のボリュームの中及び全体でのなんらかの数の論理エクステントの中の仮想ブロックからの複数のマッピングが同じ物理的に記憶されるブロックにマッピングすることを可能にする。記憶されたブロックリファレンスの代わりに、論理エクステントは簡略なパターンを符号化できるであろう。例えば、同一のバイトの文字列であるブロックは、ブロックが同一バイトの繰り返されるパターンであることを単に符号化できるであろう。

10

【0322】

【0380】 複合論理エクステントは、それぞれ複合論理エクステントのサブレンジから基本的なリーフ論理エクステント又は複合論理エクステントにコンテンツの論理範囲をマッピングする複数のマップを含む、なんらかの仮想サイズを有するコンテンツの論理範囲である場合がある。複合論理エクステントのコンテンツに関係する要求を変換することは、次いで複合論理エクステントのコンテキストの中で要求のためのコンテンツ範囲を採取すること、その要求がどの基本的なリーフ論理エクステント又は複合論理エクステントにマッピングするのかを決定すること、及びそれらの基本的なリーフ論理エクステント又は複合論理エクステントの中のコンテンツの適切な範囲に適用するために要求を変換することを伴う。

20

【0323】

【0381】 ボリューム、つまりファイル又は他のタイプのストレージオブジェクトは、複合論理エクステントとして記述できる。したがって、（説明の大部分では、単にボリュームと呼ばれる）これらの提示されているストレージオブジェクトは、このエクステントモデルを使用し、編成できる。

【0324】

【0382】 実施態様に応じて、リーフ論理エクステント又は複合論理エクステントは、複数の他の複合論理エクステントから参照され、実質的に、ボリュームの中及び全体でのコンテンツのより大きい集合体の安価な重複を可能にするであろう。したがって、論理エクステントは、基本的には参照の非巡回グラフの中で配置される場合があり、それぞれがリーフ論理エクステントで終了する。これは、ボリュームのコピーを作成するために、ボリュームのスナップショットを作成するために、又はX T E N D E D C O P Y又は類似するタイプの動作の一部としてボリュームの中で及びボリューム間でサポートする仮想範囲コピーの一部として使用できる。

30

【0325】

【0383】 実施態様は、各論理エクステントに、それに名前を付けるために使用できるアイデンティティを提供してよい。複合論理エクステントの中の参照は、論理エクステント識別子及びそれぞれの係る論理エクステントアイデンティティに対応する論理的なサブレンジを含んだリストになるので、これは参照を簡略化する。また、論理エクステントの中では、それぞれの記憶されたデータブロックリファレンスはそれに名前を付けるために使用されるなんらかのアイデンティティに基づいてもよい。

40

【0326】

【0384】 エクステントのこれらの重複使用をサポートするために、追加機能、つまりコピーオンライト論理エクステントを加えることができる。修正動作がコピーオンライトのリーフ論理エクステント又は複合論理エクステントに影響を及ぼすとき、論理エクステントはコピーされ、コピーは新しい参照であり、（実施態様に応じて）おそらく新しいアイデンティティを有する。コピーは、どのような修正が修正動作から生じて、基本的なリーフ論理エクステント又は複合論理エクステントに関係するすべての参照又はアイデンティティを保持する。例えば、W R I T E、W R I T E S A M E、X D W R I T E R E A

50

D、XPWRITE、又はCOMPARE AND WRITE要求は、ストレージシステムに新しいブロックを記憶（又は、既存の記憶されているブロックを識別するために重複排除技法を使用）してよく、ブロックの新しい集合に対するアイデンティティを参照又は記憶するために対応するリーフ論理エクステントを修正し、おそらくブロックの以前の集合のための参照及び記憶されているアイデンティティを置き換える。代わりに、UNMAP要求は、1つ以上のブロックリファレンスを削除するためにリーフ論理エクステントを修正してよい。両方のタイプの事例では、リーフ論理エクステントが修正される。リーフ論理エクステントがコピーオンライトである場合、次いで古いエクステントから影響を受けていないブロックリファレンスをコピーしてから、修正動作に基づいたブロックリファレンスを置換又は削除することによって形成される新しいリーフ論理エクステントが作成される。

10

【0327】

【0385】リーフ論理エクステントの場所を突き止めるために使用された複合論理エクステントは、次いで新しいリーフ論理エクステントリファレンス又はコピーされ、修正されたリーフ論理エクステントと関連付けられたアイデンティティを、以前のリーフ論理エクステントの代替物として記憶するために修正されてよい。その複合論理エクステントがコピーオンライトである場合、次いで新しい複合論理エクステントは新しい参照として又は新しいアイデンティティとともに作成され、その基本的な論理エクステントに対する任意の影響を受けていない参照又はアイデンティティは、その新しい複合論理エクステントにコピーされ、以前のリーフ論理エクステントリファレンス又はアイデンティティは、新しい論理エクステントリファレンス又はアイデンティティで置換される。

20

【0328】

【0386】このプロセスは、修正動作を処理するために使用される非巡回グラフを通る探索経路に基づいて、参照されたエクステントから参照する複合エクステントにさらに過去に遡って続行し、すべてのコピーオンライト論理エクステントがコピーされ、修正され、置換される。

【0329】

【0387】これらのコピーされたリーフ論理エクステント及び複合論理エクステントは、次いでコピーオンライトであるという特徴を除外する場合があります、これにより追加の修正は追加のコピーを生じさせない。例えば、第1のときに、コピーオンライト「親」複合エクステントの中のなんらかの基本的な論理エクステントが修正され、その基本的な論理エクステントはコピーされ、修正され、コピーは次いで、親複合論理エクステントのコピーされ、置換されたインスタンスの中に書き込まれる新しいアイデンティティを有する。しかしながら、第2のときに、なんらかの他の基本的な論理エクステントがコピーされ、修正され、他の基本的な論理エクステントコピーの新しいアイデンティティは、親複合論理エクステントに書き込まれ、親は、次いで追加のコピーなく定位置で修正され、親複合論理エクステントに対する参照の代わりに必要な置き換える場合がある。

30

【0330】

【0388】現在のリーフ論理エクステントがない、ボリュームの又は複合論理エクステントの新しい領域に対する修正動作は、それらの修正の結果を記憶するために新しいリーフ論理エクステントを作成してよい。その新しい論理エクステントが既存のコピーオンライト論理エクステントから参照される場合、次いでその既存のコピーオンライト複合論理エクステントは、新しい論理コンテンツを参照し、既存のリーフ論理エクステントを修正するためのシーケンスに類似するコピー、修正、及び置換の別の動作のシーケンスを生じさせる。

40

【0331】

【0389】親複合論理エクステントが（実施態様に基づいて）新しい修正動作のために作成するための新しいリーフ論理エクステントを含む関連付けられたアドレス範囲をカバーするほど十分に大きく成長できない場合、次いで親複合論理エクステントは、さらに再び新しい参照又は新しいアイデンティティである単一の「孫」複合論理エクステントから次

50

いで参照される2つ以上の新しい複合論理エクステントにコピーされてよい。その孫論理エクステントがそれ自体、コピーオンライトである別の複合論理エクステントを通して見つけれられる場合、次いでその別の複合論理エクステントは、上述の段落で説明されるのと類似した方法でコピーされ、修正され、置換される。このコピーオンライトモデルは、論理エクステントのこれらの有向非巡回グラフに基づいて、ストレージシステム実施態様の中でスナップショット、ボリュームコピー、及び仮想ボリュームアドレス範囲コピーを実装することの一部として使用できる。それ以外の場合書込み可能なボリュームの読取り専用コピーとしてスナップショットを作成するために、ボリュームと関連付けられた論理エクステントのグラフは、マークされたコピーオンライトであり、元の複合論理エクステントに対する参照はスナップショットによって保持される。ボリュームに対する修正動作は、次いで必要に応じて論理エクステントコピーを作成し、それらの修正動作の結果及び元のコンテンツを保持するスナップショットを記憶するボリュームを生じさせる。ボリュームコピーは、元のボリュームとコピーされたボリュームの両方ともコンテンツを修正し、独自のコピーされた論理エクステントグラフ及びサブグラフを生じさせることを除き、類似している。

【0332】

【0390】 仮想ボリュームアドレス範囲コピーは、（ブロックリファレンスに対する変更がコピーオンライトリーフ論理エクステントを修正しない限り、それ自体、コピーオンライト技術を使用することを伴わない）リーフ論理エクステントの中及び間でブロックリファレンスをコピーすることによって動作する場合がある。代わりに、仮想ボリュームアドレス範囲コピーは、リーフ論理エクステント又は複合論理エクステントに対する参照を重複させる場合があり、これは、より大きなアドレス範囲のボリュームアドレス範囲コピーにはうまく機能する。さらに、これは、グラフが単にリファレンスツリーよりむしろ参照の有向非巡回グラフになることを可能にする。重複された論理エクステントリファレンスと関連付けられたコピーオンライト技術は、仮想アドレス範囲コピーのソース又はターゲットに対する修正動作が、ボリュームアドレス範囲コピー動作の直後に同じ論理エクステントを共用するターゲット又はソースに影響を及ぼすことなくそれらの修正を記憶するために新しい論理エクステントの作成を生じさせることを確実にするために使用できる。

【0333】

【0391】 また、ポッドのための入力／出力動作も、論理エクステントの有向非巡回グラフを複製することに基づいて実施されてよい。例えば、ポッドの中の各ストレージシステムは、論理エクステントのプライベートグラフを実装することができ、これによりポッドのための1つのストレージシステムに関するグラフは、ポッドのための任意の第2のストレージシステム上のグラフに特定の関係を有さない。しかしながら、ポッド内のストレージシステム間でグラフを同期させることには価値がある。これは、再同期のため、及び例えば非同期複製又はリモートストレージシステムに対するスナップショットベースの複製等の特徴を調整するために有用である場合がある。さらに、それは、スナップショットの配布を処理するため及びコピー関係処理のためのなんらかのオーバーヘッドを削減するためにも有用である場合がある。係るモデルでは、ポッドのためのすべての同期中のストレージシステム全体で同期中のポッドのコンテンツを保つことは、ポッドのためのすべての同期中のストレージシステム全体ですべてのボリュームのためにリーフ論理エクステント及び複合論理エクステントのグラフを同期中に保つこと、及びすべての論理エクステントのコンテンツが同期中であることを確実にすることと、基本的に同じである。同期中であるためには、リーフ論理エクステント及び複合論理エクステントを一致させることは同じアイデンティティを有するべきである、又はマッピング可能なアイデンティティを有するべきであるかのどちらかである。マッピングは、中間マッピングテーブルのなんらかの集合を伴う場合もあれば、なんらかの他のタイプのアイデンティティ変換を伴う場合もあるであろう。一部の場合では、リーフ論理エクステントによってマッピングされたブロックのアイデンティティも同期中に保つことができるであろう。

【0334】

【0392】 ポッドごとに単一リーダーを有するリーダー及びフォロワーに基づいたポッド実施態様では、リーダーは、論理エクステントグラフに対する任意の変更を決定することを任される場合がある。新しいリーフ論理エクステント又は複合論理エクステントが作成される場合、それはアイデンティティを与えられる場合がある。既存のリーフ論理エクステント又は複合論理エクステントが修正を有する新しい論理エクステントを形成するためにコピーされる場合、新しい論理エクステントは、修正のなんらかの集合を有する以前の論理エクステントのコピーとして記述される場合がある。既存の論理エクステントが分割される場合、分割は、結果として生じる新しいアイデンティティとともに記述される場合がある。論理エクステントが、なんらかの追加の複合論理エクステントから基本的な論理エクステントとして参照される場合、その参照は、その基本的な論理エクステントを参照するための複合論理エクステントに対する変更として記述される場合がある。

10

【0335】

【0393】 ポッドでの修正動作は、このようにして（コンテンツを拡張するために新しい論理エクステントが作成される場合、又はスナップショット、ボリュームコピー、及びボリュームアドレス範囲コピーに関係するコピーオンライト状態を処理するために、論理エクステントがコピーされ、修正され、置換される場合）論理エクステントグラフに対する修正の記述を配布すること、及びリーフ論理エクステントのコンテンツに対する修正のための記述及びコンテンツを配布することを含む。有向非巡回グラフの形のメタデータを使用することからくる追加の利点は、上述されたように、物理ストレージ内の記憶されたデータを修正するE/O動作が、物理ストレージに記憶されたデータに対応するメタデータの修正によりユーザレベルで物理ストレージに記憶されたデータを修正することなく一効果を与えられ得る点である。物理ストレージがソリッドステートドライブであってよいストレージシステムの開示された実施形態では、フラッシュメモリに対する修正に付随する磨耗は、フラッシュメモリの読取り、消去、又は書込みを通しての代わりに、E/O動作によってターゲットにされるデータを表すメタデータの修正により効果を与えられるE/O動作のために回避又は削減され得る。さらに、仮想化ストレージシステムでは、上述されたメタデータは、仮想アドレス、つまり論理アドレスと物理アドレス、つまり実アドレスとの関係性を処理するために使用されてよい言い換ええると、記憶されたデータのメタデータ表現は、それがフラッシュメモリ上の磨耗を削減する又は最小限に抑える点でフラッシュフレンドリ（flash-friendly）と見なされてよい仮想化ストレージシステムを可能にする。

20

30

【0336】

【0394】 リーダーストレージシステムは、ポッドデータセットのそのローカルコピー及びローカルストレージシステムのメタデータとの関連でこれらの記述を実装するために、独自のローカル動作を実行してよい。さらに、同期中のフォロワーは、ポッドデータセットのそれらの別々のローカルコピー及びその別々のローカルストレージシステムのメタデータとの関連でこれらの記述を実装するために、独自の別々のローカル動作を実行する。リーダーとフォロワーの両方が完了すると、結果は互換性のあるリーフ論理エクステントコンテンツを有する論理エクステントの互換性のあるグラフである。論理エクステントのこれらのグラフは、次いで上記の例に説明されたように一種の「共通メタデータ」になる。この共通メタデータは、修正動作と必須共通メタデータとの間の依存関係として記述される場合がある。グラフへの変換は、後続の修正動作との待ち行列プレディケート関係を有する別々の動作として記述できる。代わりに、各修正動作は、ポッド全体で完了するとまだ知られていない特定の同じグラフ変換に頼る各修正動作は、それが頼る任意のグラフ変換の部分を含む場合がある。既に存在している「新しい」リーフ論理エクステント又は複合論理エクステントを識別する動作記述を処理することは、その部分はすでになんらかの早期の動作の処理で処理されていたので、新しい論理エクステントを作成することを回避する場合があり、代わりにリーフ論理エクステント又は複合論理エクステントのコンテンツを変更する動作処理部分しか実装できない。変換が互いに互換性があることを確実にすることはリーダーの役割である。例えば、ポッドを受け取る2つの書込みで開始できる

40

50

。第1の書込みは、複合論理エクステントAを複合論理エクステントBとして形成されたコピーで置換し、リーフ論理エクステントCをリーフ論理エクステントDとしてのコピーで及び第2の書込みのためのコンテンツを記憶するための修正で置換し、さらにリーフ論理エクステントDを複合論理エクステントBに書き込む。一方で、第2の書込みは同じコピー、及び複合論理エクステントAの複合論理エクステントBとの置換を暗示するが、異なるリーフ論理エクステントEをコピーし、第2の書込みのコンテンツを記憶するために修正される論理エクステントFで置換し、さらに論理エクステントFを論理コンテンツBに書き込む。その場合には、第1の書込みのための記述は、AのBとの置換及びCのDとの置換、並びにDの複合論理エクステントBへの書込み、及び第1の書込みのコンテンツのリーフエクステントBへの書込みを含む場合があり、第2の書込みの記述は、リーフエクステントFに書き込まれる第2の書込みのコンテンツとともに、AのBとの置換及びEのFとの置換、並びにFの複合論理エクステントBへの書込みを含む場合がある。リーダー又は任意のフォロワーは、次いで任意の順序で第1の書込み又は第2の書込みを処理する場合があり、最終結果は、BがAをコピーし、置換し、DがCをコピーし、置換し、FがEをコピーし、置換し、D及びFが複合論理エクステントBに書き込まれることである。Bを形成するためのAの第2のコピーは、Bがすでに存在していることを認識することによって回避できる。このようにして、リーダーは、ポッドのための同期中のストレージシステム全体で論理エクステントグラフのために互換性のある共通メタデータを維持することを確実にする場合がある。

【0337】

【0395】 論理エクステントの有向非巡回グラフを使用するストレージシステムの実施態様を所与として、論理エクステントの複製された有向非巡回グラフに基づいたポッドの回復が実施されてよい。具体的には、この例では、ポッドでの回復は、複製されたエクステントグラフに基づいてよく、次いでリーフ論理エクステントのコンテンツを回復することだけではなく、これらのグラフの一貫性を回復することも伴う。回復の本実施態様では、動作は、ポッドのためのすべてのストレージシステム全体で完了したと知られていないすべてのリーフ論理エクステントコンテンツ修正だけではなく、ポッドのためのすべての同期中のストレージシステムで完了したと知られていないグラフ変換に対しても問合せすることを含んでよい。係る問い合わせることは、なんらかの調整されたチェックポイント以降の動作に基づいているであろう、又は単に完了したと知られていない動作であり、各ストレージシステムは、完了したとしてまだ信号で知らせていない通常動作中の動作のリストを保つ。この例では、グラフ変換は簡単である。グラフ変換は新しい事柄を作成し、古い事柄を新しい事柄にコピーし、古い事柄を2つ以上の分割された新しい事柄にコピーする場合もあれば、それらは他のエクステントへのその参照を修正するために複合エクステントを修正する。任意の論理エクステントを作成又は置換する任意の同期中のストレージシステムで見つけられた任意の記憶されている動作記述は、その論理エクステントをまだ有していない任意の他のストレージシステムでコピーされ、実行される場合がある。関与されたリーフ論理エクステント又は複合論理エクステントが適切に回復されている限り、リーフ論理エクステント又は複合論理エクステントに対する修正を説明する動作は、それらの修正を、まだそれらを適用していなかった任意の同期中のストレージシステムに適用できる。

【0338】

【0396】 さらにこの例では、ポッドの回復は、以下を含んでよい。

- ・ポッドのためのすべての同期中のストレージシステムで完了したと知られていなかったリーフ論理エクステント及び複合論理エクステントの作成、ならびにもしあれば、それらのプレカーソルのリーフ論理エクステント及び複合論理エクステントについてすべての同期中のストレージシステムを問い合わせることと、
- ・ポッドのためのすべての同期中のストレージシステムで完了したと知られていなかったリーフ論理エクステントに対する修正動作についてすべての同期中のストレージシステムを問い合わせることと、

・既存のリーフ論理エクステント及び複合論理エクステントに対する新しい参照として論理アドレス範囲コピー動作について問い合わせることと、

・リーフ論理エクステントに対する、完了したと知られていない修正を識別することであって、そのリーフ論理エクステントが、やはり回復を必要とする場合がある代替リーフ論理エクステントのためのソースである—これにより、すでにそれをコピーしていなかった任意の同期中のストレージシステムでリーフ論理エクステントコピーが回復される前に、すべての同期中のストレージシステムに対するそのリーフ論理エクステントに対して修正できるように—識別することと、

・すべてのリーフ論理エクステント及び複合論理エクステントのコピー動作を完了することと、

・新しい論理エクステントリファレンスに名前を付けること、リーフ論理エクステントコンテンツを更新すること、又は論理エクステントリファレンスを削除することを含んだ、リーフ論理エクステント及び複合論理エクステントにすべての追加更新を適用することと、

・すべての必要なアクションが完了したと判断することであって、その時点で回復の追加の態様が進むことができる、判断することと

を含んでよい。

【0339】

【0397】 別の例では、論理エクステントグラフを使用する代替策として、ストレージは複製されたコンテンツアドレス指定可能ストアに基づいて実装されてよい。コンテンツアドレス指定可能ストアでは、データのブロックごとに（例えば、512バイト、4096バイト、8192バイトごとに、又は16384バイトごとにも）、（フィンガープリントと呼ばれる場合もある）一意のハッシュ値が、ブロックコンテンツに基づいて計算され、これによりボリューム又はボリュームのエクステント範囲は、特定のハッシュ値を有するブロックに対する参照のリストとして記述できる。同じハッシュ値を有するブロックに対する参照に基づいた同期複製されたストレージシステム実施態様では、複製は、第1のストレージシステムがブロックを受け取り、それらのブロックのフィンガープリントを計算し、それらのフィンガープリントのためのブロックリファレンスを識別し、参照されたブロックに対するボリュームブロックのマッピングに対する更新として1つ又は複数の追加のストレージシステムに対する変更を送達することを伴う場合があるであろう。ブロックが第1のストレージシステムによってすでに記憶されていることがわかる場合、そのストレージシステムは（参照が同じハッシュ値を使用するため、又は参照のための識別子が同一であるか、容易にマッピングできるかのどちらかであるため）追加のストレージシステムのそれぞれで参照に名前を付けるためにその参照を使用できる。代わりに、ブロックが、第1のストレージシステムによって見つけれられない場合、次いで第1のストレージシステムのコンテンツは、そのブロックコンテンツと関連付けられたハッシュ値又はアイデンティティとともに、動作記述の一部として他のストレージシステムに送達されてよい。さらに、各同期中のストレージシステムのボリューム記述は、次いで新しいブロックリファレンスで更新される。係るストアでの回復は、次いでボリュームのための最近更新されたブロックリファレンスを比較することを含んでよい。ブロックリファレンスがポッドのための異なる同期中のストレージシステム間で異なる場合、次いで各参照の1つのバージョンは、それらを一貫性があるものにするために他のストレージシステムにコピーされる場合がある。1つのシステム上のブロックリファレンスが存在しない場合、次いでそれはその参照のためのブロックを記憶するなんらかのストレージシステムからコピーされる。仮想コピー動作は、仮想コピー動作を実施することの一部として参照をコピーすることによって係るブロック又はハッシュリファレンスストアでサポートされる場合がある。

【0340】

【0398】 システム回復のための特定の実施態様に関して、図18に示される例の方法は、データセット（1812）を同期複製する複数のストレージシステム（1814、1824、1828）の中の少なくとも1つのストレージシステムによって、データセット（1812）を修正する要求（1804）を受け取ること（1842）を含む。データセッ

10

20

30

40

50

ト（１８１２）を修正する要求（１８０４）を受け取ること（１８４２）は、データセット（１８１２）を修正する要求（１８０４）を受け取ること（１８０６）と同様に実施されてよい。

【０３４１】

【0399】 また、図１８に示される例の方法は、データセット（１８１２）を修正する要求（１８０４）が、データセット（１８１２）を同期複製する複数のストレージシステム（１８１４、１８２４、１８２８）のすべてのストレージシステムで適用されたかどうかを示す回復情報（１８５２）を生成すること（１８４４）も含む。データセット（１８１２）を修正する要求（１８０４）が、データセット（１８１２）を同期複製する複数のストレージシステム（１８１４、１８２４、１８２８）のすべてのストレージシステムで適用されたかどうかを示す回復情報（１８５２）を生成すること（１８４４）は、異なる追跡に基づいた回復、重複書込みの回復を含んだ動作追跡に基づいた回復、スナップショットの回復、メタデータ及び共通メタデータの回復、完了した動作を捨てること、スライドウィンドウを使用すること、及びチェックポイントを使用することを含んだ動作追跡のための最近の活動を記録することに基づいた回復、論理エクステンツの複製された有向非巡回グラフに基づいたポッドの回復、及び複製されたコンテンツアドレス指定可能なストアでの追跡及び回復を含んだ、上述されたさまざまな技術を使用し、実装されてよい。要するに、回復情報を生成するために多様な技術が使用されてよく、回復情報は、複数のストレージシステム（１８１４、１８２４、１８２８）の中のどのストレージシステム、データセット（１８１２）を修正する要求（１８０４）を示す。

【０３４２】

【0400】 また、図１８に示される例の方法は、システム障害に応じて、修正する要求が、データセット（１８１２）を同期複製する複数のストレージシステム（１８１４、１８２４、１８２８）のすべてのストレージシステムで適用されたかどうかを示す回復情報（１８５２）に従って回復アクションを適用すること（１８４６）も含む。回復アクションは、データセット（１８１２）を修正する要求（１８０４）を適用していなかったすべてのストレージシステムでデータセット（１８１２）を修正する要求（１８０４）を適用することによって実施されてよい。回復情報（１８５２）は、複数のストレージシステム（１８１４、１８２４、１８２８）の中のどのストレージシステムが、ごく最近受け取られた要求（１８０４）を含んだ、同期複製されたデータセット（１８１２）を修正する１つ以上の要求を適用したのか、それとも適用しなかったのかを示す追跡情報を含んでよい。しかしながら、他の場合では、回復アクションは、要求（１８０４）の適用を完了した、部分的に完了したストレージシステムの集合でのデータセット（１８１２）を修正する要求（１８０４）の適用を取り消す、又はアンドゥすることによって実施されてよい。概して、デフォルト回復アクションは、データセット（１８１２）を修正する他の未決の要求に加えて、無事に要求（１８０４）を完了しなかった各ストレージシステムを識別し、要求（１８０４）を適用することであってよい。回復アクションの他の実施態様は、差異追跡に基づいた回復、重複書込みの回復を含んだ動作追跡に基づいた回復、スナップショットの回復、メタデータ及び共通メタデータの回復、完了した動作を捨てること、スライドウィンドウを使用すること、及びチェックポイントを使用することを含んだ動作追跡の最近の活動を記録することに基づいた回復、論理エクステンツの複製された有向非巡回グラフに基づいたポッドの回復、及び複製されたコンテンツアドレス指定可能なストアでの追跡及び回復の記述に関して上述される。

【０３４３】

【0401】 追加の説明のために、図１９は、本開示のいくつかの実施形態に従ってデータセットを同期複製するストレージシステムのための回復の例の方法を示すフローチャートを説明する。図１９に示される例の方法も、データセット（１８１２）を同期複製する複数のストレージシステム（１８１４、１８２４、１８２８）の中の少なくとも１つのストレージシステムによって、データセット（１８１２）を修正する要求（１８０４）を受け取ること（１８４２）と、データセット（１８０４）を修正する要求（１８０４）が、デ

ータセット（１８１２）を同期複製する複数のストレージシステム（１８１４、１８２４、１８２８）のすべてのストレージシステムで適用されたかどうかを示す回復情報（１８５２）を生成すること（１８４４）と、システム障害に応じて、修正する要求が、データセット（１８１２）を同期複製する複数のストレージシステム（１８１４、１８２４、１８２８）のすべてのストレージシステムで適用されたかどうかを示す回復情報（１８５２）に従って回復アクションを適用すること（１８４６）とを含むので、図１９に示される例の方法は図１８に示される例の方法に類似している。

【０３４４】

【０４０２】 しかしながら、図１９に示される例の方法は、回復情報を生成すること（１８４４）が、処理されたと確認された動作について複数のストレージシステムの他のストレージシステムを問い合わせること（１９０２）と、動作が完了したと確認されていないストレージシステムの集合を決定すること（１９０４）と、回復アクションを適用すること（１８４６）が、ストレージシステムの集合に対して、完了したと確認されていない動作を完了すること（１９０６）を含むことをさらに指定する。

【０３４５】

【０４０３】 完了された又は処理されたと確認された動作について複数のストレージシステムの他のストレージシステムを問い合わせること（１９０２）は、ポッド内のボリュームの中及び間のボリューム範囲の仮想コピーを同期複製することをサポートするストレージシステムのための動作追跡に関して上述されたように実施されてよい。具体的には、動作がポッドのためのすべての同期中のストレージシステムで完了されたと確認された後に、すべてのストレージシステム全体で完了した動作を捨てることに関して上述されたように、要求を受け取り、完了を信号で知らせたストレージシステムに、完了が信号で知らされた後にポッドのためのすべてのストレージシステムにメッセージを送信させ、各ストレージシステムがそれらを捨てることを可能にすることによって実施されてよい。回復は、次いで回復に関与するポッドのためのすべての同期中のストレージシステム全体で捨てられなかったすべての記録された動作について問い合わせることを伴う。

【０３４６】

【０４０４】 動作が完了したと確認されていないストレージシステムの集合を決定すること（１９０４）は、他のストレージシステムを問い合わせること（１９０２）からの結果に基づいて実施されてよく、ストレージシステムの集合は、問い合わせること（１９０２）が捨てられていなかった動作のリストを含んだ１つ以上のストレージシステムによってポピュレートされる。

【０３４７】

【０４０５】 ストレージシステムの集合に対して、完了したと確認されていない動作を完了すること（１９０６）は、本明細書に説明されるように、ストレージシステムの集合に動作を再発行し、完了していない動作ごとに、対応する要求に従ってデータセットに対する修正を記述する情報を送信し（１８１２）、本明細書に説明されるステップを完了することによって実施されてよい。

【０３４８】

【０４０６】 追加の説明のために、図２０は、本開示のいくつかの実施形態に従ってデータセットを同期複製するストレージシステムのための回復の例の方法を示すフローチャートを説明する。図２０に示される例の方法も、データセット（１８１２）を同期複製する複数のストレージシステム（１８１４、１８２４、１８２８）の中の少なくとも１つのストレージシステムによって、データセット（１８１２）を修正する要求（１８０４）を受け取ること（１８４２）と、データセット（１８１２）を修正する要求（１８０４）がデータセット（１８１２）を同期複製する複数のストレージシステム（１８１４、１８２４、１８２８）のすべてのストレージシステムで適用されたかどうかを示す回復情報（１８５２）を生成すること（１８４４）と、システム障害に応じて、修正する要求が、データセット（１８１２）を同期複製する複数のストレージシステム（１８１４、１８２４、１８２８）のすべてのストレージシステムで適用されたかどうかを示す回復情報（１８５２）

に従って回復アクションを適用すること（1846）を含むので、図20に示される例の方法は図18に示される例の方法に類似している。

【0349】

【0407】 しかしながら、図20に示される例の方法は、回復情報を生成すること（1844）が、複数のストレージシステムでデータセット（1812）を修正する要求（2004）を適用することに向かう進行を追跡することによって、回復情報（1852）を生成すること（2002）を含むことをさらに指定する。

【0350】

【0408】 複数のストレージシステムでデータセット（1812）を修正する要求（1804）を適用することに向かう進捗を追跡することによって、回復情報（1852）を生成すること（2002）は、処理又は完了されたと確認される動作を決定するためにチェックポイントングを使用することによって、上述されたように実施されてよい。このようにして、生成された（2002）回復情報（1852）は、どのストレージシステムが、データセット（1812）を修正する要求（1804）を処理又は完了していないのかを示してよい。

【0351】

【0409】 データセット（1812）を修正する要求（1804）を適用すること（2004）は、上述されたように実施されてよい要求（1804）を再発行するための1つ以上のストレージシステムを識別するために回復情報（1852）を使用することと、データセット（1812）を修正する要求（1804）のために、要求（1804）に従ってデータセットに対する修正を記述する情報を送信することと、上述されたステップを完了することによって実施されてよい。

【0352】

【0410】 データセット（1812）を修正する要求を適用しなかったストレージシステムでデータセット（1812）を修正する要求（1804）をアンドウする（2006）ことは、要求（1804）が、要求（1804）が処理又は完了された1つ以上のストレージシステムを識別するために回復情報（1852）を使用することによって実施されてよい。さらに、要求をアンドウすること（2006）は、要求（1804）が完了したストレージシステムごとに、データセット（1812）を修正する各要求に対応する変更のログを各ストレージシステムで維持することに依存する場合があります。データセット（1812）を修正する各要求は、識別子とさらに関連付けられてよい。また、ログは、要求識別子ごとに、メタデータ表現のバージョンを関連付けてもよい。要求識別子を適用する前に、データセットの状態を表す有向非巡回グラフを含むメタデータ表現のバージョンを関連付けてもよい。一部の例では、係るバージョン情報は、スナップショットに対応してよい。上述されたように、データセットの可視化表現を所与として、及びデータセットを修正する対応する要求によって上書きされたデータに加えて、特定の要求に対応するデータセットのメタデータ表現に対する相違だけが記憶されることを所与として、ログのストレージ用件は最小限に抑えられるべきである。このようにして、ストレージシステムのコントローラは、ログを使用し、要求（1804）の適用前の事前情報にデータセットの応対を復元し、要求（1804）の適用前の事前状態にメタデータ表現の現在の状態を定義してよい。

【0353】

【0411】 追加の説明のために、図21は、本開示のいくつかの実施形態に従ってデータセットを同期複製するストレージシステムの再同期の例の方法を示すフローチャートを説明する。あまり詳細には示されないが、図21に示されるストレージシステム（2114、2124、2128）は、図1Aから図1D、図2Aから図2G、図3Aから図3B、又はその任意の組合せに関して上述されたストレージシステムに類似してよい。実際には、図21に示されるストレージシステム（2114、2124、2128）は、上述されたストレージシステムと同じ構成要素、より少ない構成要素、追加の構成要素を含んでよい。

10

20

30

40

50

【0354】

【0412】 図21に示される例のストレージシステム構成は、データセット(2112)、及びデータセット(2112)が同期複製されてよい、複数のストレージシステム(2114、2124、2128)を含む。図21に示されるデータセット(2112)は、例えば、特定のボリュームのコンテンツとして、ボリュームの特定のシャードのコンテンツとして、又は1つ以上のデータ要素の任意の他の集合体として実施されてよい。データセット(2112)は、各ストレージシステム(2114、2124、2128)がデータセット(2112)のローカルコピーを保持するように、複数のストレージシステム(2114、2124、2128)全体で同期されてよい。本明細書に説明される例では、係るデータセット(2112)は、少なくともアクセスされているクラスタ及び特定のストレージシステムが名目上実行している限り、クラスタ内の任意の一方のストレージシステムが、クラスタ内の任意の他方のストレージシステムよりも実質的により最適に動作しないような性能特性を有するストレージシステム(2114、2124、2128)のいずれかを通してデータセット(2112)にアクセスできるように、ストレージシステム(2114、2124、2128)全体で同期複製される。係るシステムでは、データセット(2112)に対する修正は、ストレージシステム(2114、2124、2128)のいずれかでデータセット(2112)にアクセスすることが一貫した結果を生じさせるように、各ストレージシステム(2114、2124、2128)に常駐するデータセットのコピーに対して加えられるべきである。例えば、データセットに発行される書込み要求は、すべてのストレージシステム(2114、2124、2128)でサービスを提供されなければならない、又はストレージシステム(2114、2124、2128)のいずれでもサービスを提供されてはならない。

【0355】

【0413】 さらに、データセット(2112)の場合、データセット(2112)がその全体で同期複製される複数のストレージシステム(2114、2124、2128)は、例えばデータセット(2112)を、そのデータセット(2112)を名目上記憶する1つ以上のストレージシステム(2114、2124、2128)と関連付けるポッド定義又は類似したデータ構造を調べることによって実施されてよい。係る例では、ポッド定義は、データセット(2112)の少なくとも1つの識別、及びデータセット(2112)がその全体で同期複製されるストレージシステム(2114、2124、2128)の集合を含んでよい。係るポッドは、対称アクセス、レプリカの柔軟性がある追加／削除、高可用性データ一貫性、データセットに対する関係におけるストレージシステム全体での一様なユーザ管理、管理ホストアクセス、アプリケーションクラスタ化等を含んだいくつかの(おそらく任意選択の)データ一貫性の一部をカプセル化してよい。ストレージシステムは、ポッドに加えることができ、ポッドのデータセット(2112)は、そのストレージシステムにコピーされ、次いでデータセット(2112)が修正されるにつれ、最新に保たれる。また、ストレージシステムは、ポッドから削除することもでき、データセット(2112)は、もはや削除されたストレージシステムで最新に保たれない。係る例では、ポッド定義又は類似するデータ構造は、ストレージシステムが特定のポッドに加えられ、特定のポッドから削除されるにつれ更新されてよい。

【0356】

【0414】 また、図21に示される例のストレージシステムは、データセット(2112)を同期複製するために使用される複数のストレージシステム(2114、2124、2128)のそれぞれの間に1つ以上のデータ通信リンク(2116、2118、2120)も含む。図21に示される例の方法では、ポッド内のストレージシステム(2114、2124、2128)は、ともに高帯域幅データ転送、並びにクラスタ通信、ステータス通信、及び管理通信のために互いと通信する。これらの別個のタイプの通信は、同じデータ通信リンク(2116、2118、2120)を介している、又は代替実施形態では、これらの別個のタイプの通信は、別々のデータ通信リンク(2116、2118、2120)を介するであろう。

10

20

30

40

50

【0357】

【0415】 データセットを同期複製するストレージシステムを実装するための追加の詳細は、その全体として参照により含まれる米国仮特許第62／470，172号及び大62／518，071号の中で見つけられてよい。

【0358】

【0416】 図21に示されるように、複数のストレージシステム（2114、2124、2128）は、データセット（2112）を同期複製しており、ホストコンピューティングデバイスからI/O要求を受け取り、処理するための通常の動作中互いと通信してよい。しかしながら、一部の例では、ストレージシステム（2114、2124、2128）の1つ以上は、故障する、再起動する、アップグレードする、又はそれ以外の場合利用できない場合があり、その結果、1つ以上のストレージシステム（2114、2124、2128）は同期外れになる場合がある。通常の動作を再開するために、同期中のストレージシステム及び同期していないストレージシステムは、回復動作及び再同期を経る一回復は、その全体として本明細書に参照により含まれる出願参照番号第15／696，418号により詳細に説明され、再同期は以下に説明される。

10

【0359】

【0417】 ポッドに加えられるストレージシステムの初期同期—又は、ポッドからデタッチされたストレージシステムの後続の再同期—は、そのストレージシステムがポッドサービスを提供する際のアクティブな使用のためにオンラインにされる前に、ポッドのための同期中のストレージシステムから初期化されていない、つまり同期していないストレージシステムにすべてのコンテンツ、又はすべての紛失したコンテンツをコピーすることを含む。係る初期同期は、ポッドの拡張部として導入されるストレージシステムごとに実行されてよい。

20

【0360】

【0418】 ポッドに加えられたストレージシステムに対するコンテンツの初期同期と、ポッドのための同期中のストレージシステムに比してイベントのなんらかの集合を通して同期しなくなったストレージシステムを再同期することとの相違は、概念上きわめて類似している。例えば再同期の場合、同期中のポッドメンバーストレージシステムと同期していないポッドメンバーとの間で異なる場合があるすべてのブロックは、同期していないポッドメンバーがポッドのための同期中のポッドメンバーストレージシステムとしてオンラインに戻ることができる前に最新にされる。初期の同期では、これはすべてのブロックを更新することを含んでよく、結果的に、それはすべてのブロックが異なる場合がある再同期に概念上類似している。言い換えると、初期同期は、任意のボリュームが初期状態から修正される前、又は任意のボリュームが作成される又はポッドに加えられる前に、ポッドの始まりでデタッチされたストレージシステムを再度アタッチすることに同等と見なされてよい。

30

【0361】

【0419】 概して、再同期は、デタッチされたポッドをポッドが同期しており、オンラインに戻ることができる点まで戻すために少なくとも2つの事柄、つまり（a）ポッドがデタッチされ、同期中のポッドメンバーによって保持されていなかった頃に、デタッチされたポッドで持続されていた任意の変更を取り消すこと、上書きすること、又はそれ以外の場合置換すること、及び（b）ポッドのためのコンテンツ及び共通メタデータと一致するためにアタッチするストレージシステムを更新することを達成する。オンラインに戻されるためには、ストレージシステムの再アタッチは、同期複製を再有効化すること、対称的な同期複製を再有効化すること、及び再アタッチされたストレージシステムでポッドのための動作の受取り及び処理を再有効化することを含んでよい。ポッドのための動作は、読取り、データ修正動作、又は管理動作を含んでよい。

40

【0362】

【0420】 ストレージシステムをデタッチするプロセスでは、なんらかの数の動作がポッドのために進行中である可能性がある。さらに、それらの動作のいくつかは、デタッチさ

50

れたストレージシステムだけで持続した可能性があり、他の動作は、デタッチが処理された直後には同期中のままであったストレージシステムだけで持続した可能性があり、他の動作は、デタッチされたストレージシステムと同期中のままであったストレージシステムの両方で持続していた可能性がある。この例では、ポッドのための同期中の状態は、デタッチされたストレージシステムだけで持続された動作を記録していないため、ストレージシステムのデタッチ以降のポッドのための同期中のコンテンツ及び共通メタデータに対する任意の更新は、それらの更新を含まないであろう。これが一更新をアンドゥすることによって明示的に、又は再同期プロセスの一部としてそのコンテンツを上書きすることによって暗示的に、のどちらかで一これらの更新が取り消されるべきであろう理由である。同期中のストレージシステム自体では、これらは、デタッチされたストレージシステムの再アタッチを開始する前に説明される2つのリスト、つまり（a）デタッチ時に同期中未決動作リストと呼ばれることがあり、進行中であり、再アタッチするストレージシステムがポッドからデタッチされ、ポッドからのデタッチの後の任意の持続時間、同期中にとどまった動作のリスト、及び（b）再アタッチするストレージシステムがポッドからデタッチされたときの時間帯の間のコンテンツ又は共通メタデータに対する変更のリストがある場合がある。さらに、ポッド及びストレージシステムの実施態様に応じて、同期中のストレージシステムと関連付けられた2つのリストは、単一のリスト、つまり再アタッチするストレージシステム上にないと知られていないコンテンツによって表されてよい。複数のストレージシステムがデタッチされ、特にそれらのストレージシステムが異なるポッドでは、各デタッチ以降の変更を追跡することは、別々のリストを生じさせる一及びそれらのリストがどのようにして記述されるのかは、あるポッドの実施態様から別のポッドの実施態様で大幅に異なる場合がある。一部の 경우에는、デタッチの時点から変更を追跡し、それらの変更をアタッチするストレージシステムにコピーすることを超えた追加の問題は、再同期中に受け取られる新しい修正動作が、アタッチするストレージシステムに適用されることを確実にすることである。概念上、この問題は、データをコピーする動作、及びポッドによって受け取られる修正動作の処理が、結果が、アタッチの最後で、及びアタッチするストレージシステムがポッドのために同期中であると見なす前に正しく最新となるように、マージされてよいことを確実にすることとして記述されてよい。

【0363】

【0421】 単一の変更されたコンテンツ再同期に関して、再同期のための1つのモデルは、同期中のストレージシステムとアタッチするストレージシステムとの間で異なる場合があるブロックの完全なリストーデタッチされたブロックのリストを生成すること、及び修正動作がフォロワーストレージシステムに対して起こるにつれ、任意の修正動作の複製を開始することである。異なる場合があるブロックの完全なリストは、同期中のストレージシステムからのデタッチ時の同期中の未決動作からのもの、アタッチするストレージシステムからのデタッチの時点での未決動作、及びデタッチ以来変化したと知られていたブロックを含んでよい。修正動作は、説明されるように、それらの修正コンテンツを記憶する場合があり、再同期は、デタッチされたブロックのリストからブロックの範囲の場所を突き止め、同期中のストレージシステムからアタッチするストレージシステムへ、セクションでこれらのブロックをコピーすることにより進んでよい。この例では、特定のセクションをコピーしている間に、コピーされているセクションと重複する、入信する修正動作は、コピーの間阻止される場合もあれば、セクションがコピーされた後にそれらの修正動作を適用するための準備がなされてよい。この解決策は、例えばEXTENDED_COPY動作の仮想化された実施態様等、仮想ブロック範囲コピー動作によって問題を生じさせる場合がある。さらに、コピーのためのソース範囲は、まだ再同期されていない場合があり、しかもターゲット範囲はすでに再同期されている場合がある。つまり、仮想ブロック範囲コピー動作の簡単な実施態様は、データが、仮想ブロック範囲コピー動作が受け取られるときに既知ではないため（実施態様に応じて）正しいデータをターゲット範囲にコピーできない、又は再同期動作自体が、ターゲット範囲が、それがその最終的な形式で絶対に再同期されなかったとき、ターゲット範囲が正しく同期されたと推定した可能性がある

10

20

30

40

50

ため、ターゲット範囲を正しく再同期できない場合がある。しかしながら、この問題にはいくつかの解決策がある。1つの解決策は、再同期中に仮想ブロック範囲コピー動作を許可しないことである。これは、仮想ブロック範囲コピー動作—クライアントがファイルコピー動作及びバーチャルマシンクローン又は以降動作を操作することを含んだ—の共通使用が、通常、読取り要求及び書込み要求のシーケンスを通してそれ自体直接的にコンテンツをコピーすることによって仮想ブロック範囲コピーに回答するため、多くの場合に機能し得る。別の解決策は、任意の仮想アドレス範囲コピー動作のターゲットアドレス範囲を上書きし、次いでソースデータが利用可能になるときに上書きを説明しながら、コピー動作を実行する修正動作ではなく、不完全な仮想範囲コピー動作を記憶にとどめることである。再同期のターゲットが、コピーのためのソースデータが正しくないことを知らない場合があることを所与として、すべての係る動作は、完全なコピーが完了するまで延期されなければならない可能性がある。再同期のターゲットが、どの領域がまだコピーされていないのかを認識させられる、又はいつ再同期がボリュームの特定の領域を処理することを完了したのかを認識し得る場合、最適化が起こり得る。

【0364】

【0422】 ストレージシステムを再同期することの別の態様は、更新されたブロック追跡である場合がある。例えば、ストレージシステムがデタッチされている間に修正されるすべての個々のブロックのリストを保つ（次いで、それらを個別に再同期する）ことは、延長された機能停止期間が多数のブロックを生じさせる場合があるため、一部の 경우에는 実際的ではない場合がある—いくつかのストレージシステムは、不連続ブロックの大きい集合体を非常に効率的に読み取ることができない。結果的に、一部の 場合には、例えばボリュームの1MB範囲等、追跡範囲を開始することが追跡されたメタデータの量を削減するためにはより実際的である場合がある。このコース気質の追跡は、より短期の動作追跡に遅れて更新される場合があり、数分ダウンしているのか、数時間ダウンしているのか、数日ダウンしているのか、又は数週間ダウンしているのかに関わりなく、任意の時代遅れのストレージシステムの再同期を処理するために必要とされる限り、保護されてよい。機械的なスピニングストレージとは対照的に、ソリッドステートストレージを用いると、ボリュームの、又はボリュームの集合体の、又は完全なポッドのどの個々のブロックなのかを追跡することは、変化したそれらの個々のブロックだけを再同期することがきわめて実際的であるように、きわめて実際的である場合がある。概して、無作為な読取り及び書込みペナルティはほとんどなく、マルチレベルマップから読み取ることに対するペナルティは少ししかなく、結果的に短期間（例えば、100ミリ秒から10秒の範囲内、又は数百の動作ごとに又は数千の動作ごとに）にわたる動作としてのきめの細かい活動を、すべての修正されたブロックに名前を付けるきめの細かいマップにマージすることは相対的に容易である。さらに、最近の活動のリストは、最近ジャーナルデバイス（高書込み帯域幅及び高上書き速度をサポートすることを目的とした多様な特色のNVRAM等の高速書込みストレージ）に記録されているコンテンツ修正をカバーするリストであってよいが、それらの修正についてのメタデータはおそらく実際のコンテンツよりも長い期間ジャーナルに保たれる。この例では、すべての活動のマージされたリストは、各ビットがブロック又はブロックの小さいグループを表すビットマップである場合もあれば、マージされたリストは、ボリューム単位で例えばBツリー等のツリー構造に編成されたブロック番号のリスト又はブロック範囲のリストである場合もある。近傍のブロック番号は、あるブロック番号から別のブロック番号への差異として記憶されてよい。ため、ブロック番号の係るリストは容易に圧縮されてよい。

【0365】

【0423】 また、ストレージシステムを再同期させることは、シーケンス番号を追跡することによるブロック追跡を含んでもよい。例えば、いくつかのストレージシステムは、通常動作中—すべての修正について—それぞれのシーケンス番号をそれぞれの修正と関連付ける。係る場合には、デタッチの頃にデタッチされたアレイに複製されなかった可能性がある任意のコンテンツを含んだ、デタッチ以来修正されたすべてのコンテンツを見つける

ためには、ポッドからデタッチされたストレージシステムと同期したと知られている最後のシーケンス番号が、ポッドのための同期中のストレージシステムを問い合わせるために必要とされるすべてである。

【0366】

【0424】 また、ストレージシステムを再同期させることは、スナップショットとして変更を追跡することを含んでもよい。例えば、スナップショットは、過去のなんらかのとき以来の変更を追跡するために使用されてよく、ストレージシステムは、完了したと知られていないコンテンツを除外することによって、デタッチのときにスナップショットを製造してよい。代わりに、スナップショットは、規則正しく、又はなんらかの周期性をもって作成されてよく、スナップショット作成の時点は、どのスナップショットがデタッチされたストレージシステムを再同期させるための基礎となり得るのかを判断するために、デタッチの時点に比較されてよい。変形形態として、デタッチの前にポッド全体で作成されたスナップショットは、ポッドのための同期中のストレージシステムとデタッチされたストレージシステムの両方に存在するべきであり、再同期のために多様な方法で使用されてよい。例えば、再アタッチされているストレージシステムのコンテンツは、デタッチに先行するその最後に同期されたスナップショットに戻され、ポッド内の現在同期中のコンテンツと一致するためにその時点から前に進められてよい。概して、スナップショットは、以前のスナップショットに対する差異を示す又は現在のコンテンツに対する差異を示す。スナップショットのこれらの特徴を使用し、再アタッチするストレージシステムにコンテンツを再同期させることは、再アタッチの時点と前回の完全なデタッチ前の同期スナップショットの時点との差異を複製することを含んでよい。一部の 경우에는、再同期は、フォールバックとしてスナップショットベースのモデルを使用してよい。例えば、(約数分の機能停止期間等の) 短い機能停止期間は、きめの細かい追跡、又はストレージシステムがデタッチしたとき以来に発生した記録再生動作を通して処理されてよく、より長い機能停止期間は、数分ごとに撮影されるスナップショットに戻ることによって処理されてよい。閾値分数は、デフォルト値である場合もあれば、ユーザ又は管理者により指定される場合もある。相対的に不定期に起こるスナップショットは、低い長期オーバーヘッドを有する場合があるが、再起動されるためにより多くのデータを生成する場合があるため、係る構成は、实际的であってよい。例えば一部の 경우에는、デタッチの5分前に撮影されたスナップショットは、最大5分間分のコンテンツ修正を転送するのに対して10秒の機能停止期間は、記録された動作をリプレイすることによって処理されてよく、再同期は10秒以下で起こり得る。他の場合には、機能停止期間後の再同期は、例えば短期マップの蓄積サイズに対する制限によって等、蓄積された変更に基づく場合がある。

【0367】

【0425】 一部の 경우에는、再同期は同期複製に基づく場合がある。例えば、上述されたスナップショットをベースにした再同期モデルは、別の再同期モデルもサポートしてよい。つまり、非同期複製又は周期的な複製をサポートするストレージシステムは、再同期中にコンテンツを複製するためにスナップショット機構を使用してよい。非同期複製モデル又は周期的な複製モデルは、短い機能停止期間の間に時代遅れのデータを潜在的にコピーしてよく、周期的な複製モデルはスナップショット又はチェックポイントの差分表現に基づいてよく、差分表現は機能停止期間を自動的に処理する。非同期複製に関しては、上記説明と同様に延長された機能停止期間のためのバックアップとしてのスナップショット又はチェックポイントに対する依存がある場合があり、結果として実施態様を結合する、又は再同期のために係る利用可能な非同期複製実施態様又は周期複製実施態様を活用することが实际的である場合がある。ただし、1つの問題は、非同期複製モデル又は周期複製モデルが複製ターゲットを最後まで最新に又は完全に同期中にさせるように構成され得ない点である。結果として、係る再同期実施態様を用いると、新しい飛行中動作は、飛行中動作が適用される場合があり、これによりアタッチするストレージシステムに対するすべての修正はポッドに対して最新であるように追跡され得る。

【0368】

〔0426〕 一部の 경우에는、再同期は、多段階再同期を含むように実施されてよい。例えば、第1の段階で、なんらかの点までのコンテンツが、ポッドのための同期中のストレージシステムからポッドのためのアタッチするストレージシステムへ複製されてよい。この例では、第2のスナップショットはアタッチ中に撮影されてよく、デタッチ前に同期されたと知られている最後のスナップショットであった第1のスナップショットとの間の差異、及び第1のスナップショットと、アタッチするストレージシステムに複製される第2のスナップショットの差異。係る機構は、アタッチするストレージシステムを、それがアタッチ前よりもより密接に同期中にさせ得るが、それはまだ最新ではない場合がある。したがって、第3のスナップショットが作成されてよく、第3のスナップショットと第2のスナップショットとの差異が決定され、次いでアタッチするストレージシステムに複製されてよい。この第3のスナップショット、及び決定された差異は、第2のスナップショットまで複製されたコンテンツ及び現在のコンテンツの差異の部分を構成してよい。さらに、最新になって数秒以内に得るために、追加のスナップショットが撮影され、複製されてよいことが考えられる。この時点で、最後のスナップショットが複製されるまで修正動作は休止される場合がある—それによって、アタッチするストレージシステムをポッドのために最新にさせる。他の場合には、1つ以上のスナップショットを、最終的な再同期スナップショット後に受け取られる修正動作が、それらが複製されたスナップショットコンテンツとマージできるように処理されるなんらかのモードに、複製後に切り替えることが可能である。係る実施態様は、アタッチするストレージシステムにそれらの修正動作を追跡させ、スナップショット複製が完了した後に—又は、スナップショット複製が特定の修正動作により影響を受けた特定のボリューム領域を同期させたと知られた後に—修正動作を適用することを含んでよい。この基本的なコンテンツがコピーされたと知られるまですべての動作を追跡することにより多数の追跡された動作が生じる場合があるので、本実施態様は追加のオーバーヘッドを有する場合がある。代替策は、例えば、特定の共通のメタデータ又はあるブロック範囲から別のブロック範囲への拡張されたコピー動作に依存する書込み、及び再同期が、そのコンテンツ又は共通メタデータ情報の処理を優先させる要求等、最近受け取られた動作に関係するコンテンツを検討することである。このようにして、係るプロセスによりコピーされたと知られるコンテンツに結びつけられた任意の受け取られた動作は、次いで追跡構造をはるかに迅速にリリースさせてよい。

〔0369〕

〔0427〕 一部の 경우에는、再同期は、論理エクステントの有向非巡回グラフを使用するために実施されてよい。上述されたように、複製されたストレージシステムは、論理エクステントの有向非巡回グラフに基づいてよい。係るストレージシステムでは、再同期のプロセスは、ポッドのための同期中のストレージシステムからポッドのためのアタッチするストレージシステムへ論理エクステントグラフを複製することが期待されてよい—すべてのリーフ論理エクステントコンテンツを含み、同期中のポッドメンバーとしてアタッチするストレージシステムを有効化する前にグラフが同期され、同期されたままであることを確実にする。このモデルでの再同期は、アタッチのためのターゲットストレージシステムに、ボリュームごとに—ファイル又はオブジェクトをベースにしたストレージシステムでのファイル又はオブジェクトごとに—トップレベルのエクステントアイデンティティを取り出させることによって進んでよい。アタッチターゲットにすでに既知である任意の論理エクステントアイデンティティは最新と見なされてよいが、任意の未知の複合論理エクステントは取り出され、次いでそれぞれがアタッチターゲットにとってすでに既知であるか、又はアタッチターゲットにとって未知であるかのどちらかである基本的なリーフ論理エクステント又は複合論理エクステントに分解されてよい。さらに、任意の未知のリーフ論理エクステントはコンテンツを取り出す場合もあれば、ブロックがターゲットストレージシステムによってすでに記憶されているかどうかを判断するために記憶されているブロックのアイデンティティを取り出す場合もある—次いで認識されていないブロックは同期中のストレージシステムから取り出される。しかしながら、ストレージシステムのデタッチの頃のなんらかの数のエクステントは同じアイデンティティであるが、異なるコンテンツ

10

20

30

40

50

を有する場合があるため、論理エクステントを読み取り専用とマークした動作だけが、動作を修正した結果として新しい論理エクステントを形成し得るため、係る手法は、必ずしも再同期につながらない場合がある。さらに、進行中の修正動作は、デタッチにつながる障害の間、異なるストレージシステムで異なるように完了した可能性があり、それらの修正動作が非読み取り専用論理エクステントであった場合、次いでそれらの論理エクステントは2つのストレージシステムで同じアイデンティティを有するが、異なるコンテンツを有する場合がある。しかしながら、いくつかの解決策は含む、ポッドのためのストレージシステムの1つの集合が別のストレージシステムをデタッチするとき、ストレージシステムの集合は、進行中の修正動作と関連付けられたリーフ論理エクステント及び複合論理エクステントをマークし、それらのリーフ論理エクステント及び複合論理エクステントを、デタッチされたストレージシステムを含む将来の再アタッチ動作と関連付けてよい。同様に、ポッドのための再アタッチストレージシステムは、それが知っていた、進行中の論理エクステントと関連付けられたリーフ論理エクステント及び複合論理エクステントを識別してよい。結果として、（リーフエクステントのための）そのコンテンツ、又は（複合論理エクステントのための）その参照が、任意の未知のリーフ論理エクステント又は複合論理エクステントを転送することに加えて、転送される必要がある論理エクステントの2つのセット。代わりに、調整されたスナップショットは、複製されたポッドの中で周期的に撮影されてよく、再アタッチ動作のターゲットは、最後の調整されたスナップショット後に作成された論理エクステントが再同期中に廃棄又は無視されることを確実にしてよい。さらに別の代替策として、ストレージシステムがポッドからデタッチされる期間中、残りの同期中のストレージシステムは、すべての完了した動作からのコンテンツを表し、スナップショットよりも後に起こるポッドコンテンツに適用するためのすべての潜在的に進行中の動作をリプレイするスナップショットを製造してよい。この結果、デタッチされたストレージシステムにすでに複製されていない任意のコンテンツは、デタッチされたストレージシステムが絶対に受け取った可能性がない新しい論理エクステントアイデンティティを与えられている。

【0370】

【0428】 再起動実施態様に直面する場合がある別の問題は、エクステントグラフをベースにした同期複製を完全に同期させ、ライブで実行中にすることである。例えば、再同期は、例えば、最初に、アタッチの始まりで作成されたスナップショット等のより最近のスナップショットを転送することによって、ターゲットストレージシステムに、上述されたようにそれを同期中のストレージシステムから取り出させることによって、進んでよく、ターゲットは、それが有していないリーフ論理エクステント及び複合論理エクステントを増分的に要求する。このプロセスは、デタッチ時の進行中の動作を説明することを含んでよく、このプロセスの最後で、そのより最近のスナップショットまでのコンテンツは、ポッドのための同期中のストレージシステムと、アタッチするストレージシステムとの間で同期される。さらに、このプロセスは、ターゲットストレージシステムを同期中のストレージシステムにより近づけるために、別のスナップショットで、及びおそらく追加のスナップショットで繰り返されてよい。しかしながら、なんらかの時点で、ライブデータも転送される必要がある場合があり、これを行うためには、ライブの修正動作の複製が最後の再同期スナップショットの後にアタッチストレージシステムへの転送のために有効化されてよく、これによりスナップショットに含まれないすべての修正動作は、アタッチするストレージシステムに送達されてよい。本実施態様は、スナップショットに含まれるリーフ論理エクステント及び複合論理エクステントに対する修正を記述する動作を生じさせ、これらの記述は、（指定されたコンテンツを有する）新しいリーフ論理エクステント及び複合論理エクステントの作成、又は既存のリーフ論理エクステント及び複合論理エクステントの、新しいアイデンティティを有するそれらのエクステントの修正されたコピーとの置換を含んでよい。動作の記述が新しい論理エクステントを作成する、又はアタッチするストレージシステムにとってすでに既知の論理エクステントを置換する場合では、動作は、アタッチするストレージシステムが同期中であるかのように正常に処理されてよい。動作

10

20

30

40

50

のための記述が、アタッチするストレージシステムにすでに既知ではない論理エクステン
トの少なくとも1つの置換を含む場合では、その動作は、完了を可能にするために永続的
にされてよいが、動作の完全な処理は、置換されている論理エクステンツが受け取られる
まで遅延されてよい。さらに、係る論理エクステンツコンテンツ転送を待機しているこれ
らの動作と関連付けられたオーバーヘッドを削減するために、アタッチするストレージシ
ステムは、他の論理エクステンツよりも早く取り出されるそれらの論理エクステンツを優先
してよい。この例では、ストレージシステムに係る既存の論理エクステンツを待機するこ
れらの動作をどのように効率的に処理できるのかに従って、ライブ動作を可能にする前に
任意の一連のスナップショット画像を転送するなんの理由もない場合がある。代わりに、
デタッチの時点からの（又はデタッチ前のなんらかの時点からの）状態情報を記述する再
同期スナップショットが転送できるであろう一動作は、やはり上述されたように、同期中
のストレージシステムからアタッチするストレージシステムにスナップショットを転送し
ながら、上述されたように処理される。

【0371】

〔0429〕 一部の 경우에는、再同期実施態様に直面する問題は、再同期中にブロックリフ
ァレンスを保つことである。例えば、同期複製されたストレージシステムでは、特定の書
き込まれたブロック、又は動作と関連付けられたブロックの特定の集合は、そのブロック
又はそのブロック集合の書き込みのための動作記述に含まれる識別を与えられてよい。この
例では、そのブロック又はブロック集合の一部若しくはすべてを置換する新しい書き込みは
、ブロック又はブロック集合のために新しいアイデンティティを供給してよく、この新しい
アイデンティティは、（例えば、SHA-256、又は異なるブロックが同じハッシュ
値を生じさせる適切に非常に小さい可能性を有するなんらかの他の機構を使用すること等
の）ブロックコンテンツの安全なハッシュから構築される場合もあれば、新しいアイデン
ティティは、2つの書き込みが同一のブロックコンテンツを含んでいたかどうかに関わりなく
、一意の方法で単に書き込み自体を識別する場合もある。例えば、新しいアイデンティティ
は、シーケンス番号又はタイムスタンプであってよい。さらに、ブロック又はブロック
集合のための新しいアイデンティティが、書き込み動作の分散された記述で共用され、ブロ
ック又はブロック集合を書き込むことの一部として各ストレージシステムのなんらかのマ
ップに記憶される場合、次いで論理エクステンツは、これらのブロック又はブロック集合
アイデンティティに関してそのコンテンツを記述してよい。係る実施態様では、リーフエ
クステンツの再同期は、すでにアタッチするストレージシステムに記憶されたブロック又は
ブロックセットを同期中のストレージシステムから転送するよりむしろ参照してよい。
本実施態様は、再同期中に転送される総データを削減し得る。例えば、デタッチの頃にア
タッチするストレージシステムにすでに書き込まれていたが、再同期スナップショットに
含まれていなかったデータは、そのアイデンティティとともに記憶された可能性があり、
そのブロック又はブロック集合のアイデンティティはすでに既知であり、記憶されている
ため、再び転送される必要はない場合がある。さらに、なんらかの数の仮想拡張コピー動
作が、ストレージシステムがデタッチされた期間中に2つのリーフ論理エクステンツ間で
ブロックリファレンスのコピーを生じさせた場合、次いでブロック又はブロック集合のアイ
デンティティは、仮想コピーされたブロックが2度転送されていないことを確実にする
ために使用されてよい。

【0372】

〔0430〕 一部の 경우에는、再同期実施態様は、記憶されていたブロックが、ブロックコ
ンテンツの安全なハッシュに基づく場合がある、一意のアイデンティティを有する場合が
あるコンテンツアドレス指定ストアを使用してよい。この例では、再同期は、同期中のス
トレージシステム上のポッドに関係するすべてのブロックアイデンティティのリストを、
ポッド内のボリューム（又はファイル又はオブジェクト）に対するそれらのブロックアイ
デンティティのマッピングとともに、アタッチするストレージシステムに転送することに
よって進んでよい。この場合、アタッチ動作は一ボリュームからコンテンツへマッピング
を変更するライブ動作の処理と統合されてよい一ポッドのための同期中のストレージシ

10

20

30

40

50

テムから、アタッチするストレージシステムが認識していないこれらのブロックを転送することによって進んでよい。さらに、ポッドコンテンツからブロックアイデンティティへのなんらかの早期のバージョンが、ストレージシステムがポッドからデタッチする前から既知である場合、次いでその早期バージョンと現在のバージョンとの差異は、マッピング全体を転送する代わりに転送されてよい。

【0373】

【0431】 図21に示される例の方法で続行すると、例の方法は、同期していないストレージシステム(2124)に記憶される同期していないデータセット(2113)と、1つ以上の同期中のストレージシステム(2114、2128)に記憶される同期中のデータセット(2112)の差異(2152)を識別すること(2102)を含む。同期していないストレージシステム(2124)に記憶される同期していないデータセット(2113)と、1つ以上の同期中のストレージシステム(2114、2128)に記憶される同期中のデータセット(2112)の差異(2152)を識別すること(2102)は、上記に詳細に説明されるように、多様な技術を使用し、実装されてよい。1つの例の技術は、1つ以上のコントローラがブロック追跡を使用して、例えばストレージシステム(2124)等の任意の所与のストレージシステムがデタッチされる間に、及び上記に詳細に説明されるように、修正されるすべての個々のブロックのリストを保つことを含む。他の例の技術は、上記に詳細に説明されるように、他の技術の中でも、ブロック追跡及びスナップショットの組合せ、又はシーケンス番号によるブロック追跡を使用することを含む。

【0374】

【0432】 この例では、ブロック追跡は、同期中のストレージシステムで修正されるすべての個々のブロックのリストを生成するために使用され、ストレージシステム(2124)が再アタッチされるまで、ブロック追跡はストレージシステム(2124)が、デタッチしたと検出されると開始する—これは、次のステップ、つまり以下に説明する同期させること(2104)が開始するときである。さらに、「デタッチされた」ストレージシステムは、ポッドメンバーとして一覧されるが、ポッドのために同期中として一覧されないストレージシステムと見なされてよい—ポッドメンバーとして一覧されるストレージシステムは、ポッドメンバーがオンラインである、又はポッドのためにデータを石斛的に提供するために現在利用できる場合に同期中である。この例では、ポッドの各ストレージシステムメンバーは、ポッドのためのメンバーストレージシステムを示すメンバーシップリストの独自のコピーを有する場合があります、メンバーシップリストは、どのストレージシステムが現在同期中であると知られているのか、及びどのストレージシステムがポッドメンバーの完全な集合に含まれるのかを含む。概して、ポッドのためにオンラインであるためには、所与のストレージシステムのためのメンバーシップリストは、所与のストレージシステムがポッドのために同期中であることを示し、所与のストレージシステムは、ポッドのために同期中であるとして示される所与のストレージシステムのためのメンバーシップリスト内のすべての他のストレージシステムと通信できる。ストレージシステムが、メンバーシップリストによって同期中であるとして示される他のストレージシステムと通信できない場合、次いでストレージシステムは、ストレージシステムが、それが再び同期中であることを検証できるまでデータセットを修正する入信要求を処理するのを停止する(又はエラー若しくは冷害をもって要求を完了する)。特定のストレージシステムは、疑わしいストレージシステムがデタッチされるべきであると判断してよく、これが、同期中であるとして示されるメンバーシップ中のストレージシステムと同期中であることに基づいて、特定のストレージシステムが、動作し続けることを可能にする。さらに、この状況では、複数の隔離されたストレージシステムがI/O要求を処理している「スプリットブレイン」状況を回避するために、疑わしいストレージシステムは、特定のストレージシステムによる継続的な処理から妨げられ、疑わしいストレージシステムは、どのストレージシステムが、ポッドに向けられるI/O要求を処理することを続行するのか、及びどのストレージシステムが、ポッドに向けられるI/O要求を処理することを停止するのかを判断するために仲介サービスを要求することを妨げられる。係る仲介プロセスのための追加の詳細

は、その全体として参照により含まれる、出願参照番号第15／703，559号にさらに説明される。

【0375】

【0433】 また、図21の例の方法は、同期していないストレージシステム(2124)に記憶される同期していないデータセット(2113)と、1つ以上の同期中のストレージシステム(2114、2128)との差異(2152)に従って、同期していないデータセット(2113)を同期中のデータセット(2112)と同期させること(2104)も含む。同期していないデータセット(2113)を同期中のデータセット(2112)と同期させること(2104)は、上述されたように、複数の技術を使用し、実施されてよい。再同期のための1つの例の技術は、ストレージシステム(2124)がデタッチされていた間に生成された修正済みのブロックのリストを所与として、再アタッチされたストレージシステム(2124)上での修正済みのブロックのリストに対応する任意の修正動作を複製することを含む。この例では、修正動作を複製することは、I/O動作を処理することに関して上述されたように、フォローストレージシステム及びリーダーストレージシステムが同期しているときに、フォローストレージシステムがリーダーストレージシステムによって提供されるI/O動作をどのようにして実施するのかに類似して実施されてよい。上述されたように、再同期は修正動作も含んでよく、修正コメントを記憶してよく、再同期はデタッチされたブロックのリストからブロックの範囲の場所を突き止め、それらのブロックを、同期中のストレージシステムからアタッチするストレージシステムまでのセクションにコピーすることによって実施されてもよい。同期を実施すること(2104)のための他の例の技術は、上記により詳細に説明されている。

【0376】

【0434】 また、図21に示される例の方法は、同期していないストレージシステム(2124)に記憶される同期していないデータセット(2113)と、1つ以上の同期中のストレージシステム(2114、2128)に記憶される同期中のデータセット(2112)との間の同期複製関係を再確立すること(2106)も含む。同期していないデータセット(2113)と同期中のデータセット(2112)との間の同期複製関係を再確立すること(2106)は、その全体として参照により含まれる出願参照番号第15／713，153号の図4から図7に関して説明されるように、ポッドが最初に作成されるとき、ストレージシステム間で同期複製関係を最初に確立することと同様に、実施されてよく、データセットのために、データセットがその全体で同期複製される複数のストレージシステムを識別すること(2102)と、データセットを同期複製するために使用される複数のストレージシステムのそれぞれの間で1つ以上のデータ通信リンクを構成すること(2104)と、複数のストレージシステム間で、複数のストレージシステムのうちの少なくとも1つのためのタイミング情報を交換すること(2106)と、複数のストレージシステムの少なくとも1つのためのタイミング情報に従って、同期複製リースを確立すること(2108)とを含み、同期複製リースは、同期複製関係が有効である期間を識別する。

【0377】

【0435】 追加の説明のために、図22は、本開示のいくつかの実施形態に従ってデータセットを同期複製するストレージシステムのための際同期のための追加の例の方法を示すフローチャートを説明する。図22に示される例の方法も、同期していないストレージシステム(2124)に記憶される同期していないデータセット(2113)と、1つ以上の同期中のストレージシステム(2114、2128)に記憶される同期中のデータセット(2112)の差異(2152)を識別すること(2102)と、同期していないストレージシステム(2124)に記憶される同期していないデータセット(2113)と、1つ以上の同期中のストレージシステム(2114、2128)との差異(2152)に従って、同期していないデータセット(2113)を同期中のデータセット(2112)と同期させること(2104)と、同期していないストレージシステム(2124)に記憶される同期していないデータセット(2113)と、1つ以上の同期中のストレージシステム(2114、2128)に記憶される同期中のデータセット(2112)との間の

同期複製関係を再確立すること（２１０６）とを含むので、図２２に示される例の方法は図２１に示される例の方法に類似している。

【０３７８】

【０４３６】 しかしながら、図２２に示される例の方法は、同期していないストレージシステム（２１２４）に記憶される同期していないデータセット（２１１３）と関連付けられたメタデータと、１つ以上の同期中のストレージシステム（２１１４、２１２８）に記憶される同期中のデータセット（２１１２）と関連付けられたメタデータとの差異（２２５２）を識別すること（２２０２）をさらに含む。同期していないストレージシステム（２１２４）に記憶される同期していないデータセット（２１１３）と関連付けられたメタデータと、１つ以上の同期中のストレージシステム（２１１４、２１２８）に記憶される同期中のデータセット（２１１２）と関連付けられたメタデータとの差異（２２５２）を識別すること（２２０２）は、より詳細に上述されたように、多様な技術を使用し、実施されてよい。上述された一例のように、再同期中のストレージシステムがデタッチされた期間の間に発生する修正動作を追跡することに加えて、ストレージシステムは、再同期中のストレージシステムがデタッチされた期間の間に追跡された修正動作を記述するメタデータも追跡し、ログ又はジャーナルデバイスに追跡されたメタデータを記憶してもよい。

【０３７９】

【０４３７】 図２２に示される例の方法は、同期していないストレージシステム（２１２４）に記憶される同期していないデータセット（２１１３）と関連付けられたメタデータと、１つ以上の同期中のストレージシステム（２１１４、２１２８）に記憶される同期中のデータセット（２１１２）と関連付けられたメタデータとの差異（２２５２）に従って、同期していないデータセット（２１１３）と関連付けられたメタデータを同期中のデータセット（２１１２）と関連付けられたメタデータと同期させること（２２０４）をさらに含む—これは多様な技術を使用し、実施されてよい。一例の技術は、１つ以上の同期中のストレージシステム（２１１４、２１２８）が、ストレージシステム（２１２４）がデタッチされた後に発生したメタデータに対する変更のリストを生成し、維持することを含み、メタデータに対する変更のリストは、同期していないストレージシステム（２１２４）を同期中にするために、同期していないストレージシステム（２１２４）でのポッドデータのメタデータ表現を更新するために使用されてよく、同期中であることは、「共通メタデータ」を表すための互換性のあるグラフを含み、メタデータ及び共通メタデータは、出願参照番号第１５／６６，４１８号にさらに説明され、この出願はその全体として参照により含まれる。上述されたように、いくつかの動作は、デタッチされたストレージシステム（２１２４）で持続し、同期していないストレージシステム（２１２４）に記憶される同期していないデータセット（２１１３）と関連付けられたメタデータに対する修正を生じさせ、メタデータに対するこれらの修正は、専用のリストで、又は同期していないストレージシステム（２１２４）に記憶される同期されていないデータセット（２１１３）と、１つ以上の同期中のストレージシステム（２１１４、２１２８）に記憶される同期中のデータセット（２１１２）との差異を識別すること（２１０２）に関して上述された動作の同じリストで説明されてよい。

【０３８０】

【０４３８】 追加の説明のために、図２３は、本開示のいくつかの実施形態に従ってデータセットを同期複製するストレージシステムのための再同期の追加の例の方法を示すフローチャートを説明する。図２３に示される例の方法も、同期していないストレージシステム（２１２４）に記憶される同期していないデータセット（２１１３）と、１つ以上の同期中のストレージシステム（２１１４、２１２８）に記憶される同期中のデータセット（２１１２）の差異（２１５２）を識別すること（２１０２）と、同期していないストレージシステム（２１２４）に記憶される同期していないデータセット（２１１３）と、１つ以上の同期中のストレージシステム（２１１４、２１２８）との差異（２１５２）に従って、同期していないデータセット（２１１３）を同期中のデータセット（２１１２）と同期させること（２１０４）と、同期していないストレージシステム（２１２４）に記憶され

る同期していないデータセット（２１１３）と、１つ以上の同期中のストレージシステム（２１１４、２１２８）に記憶される同期中のデータセット（２１１２）との間の同期複製関係を再確立すること（２１０６）も含むので、図２３に示される例の方法は図２１に示される例の方法に類似している。

【０３８１】

【０４３９】 しかしながら、図２３に示される例の方法は、同期していないストレージシステム（２１２４）に記憶される同期していないデータセット（２１１３）と、１つ以上の同期中のストレージシステム（２１１４、２１２８）に記憶される同期中のデータセット（２１１２）の差異を識別すること（２１０２）が、同期中のデータセット（２１１２）に記憶されるコンテンツとは異なるコンテンツを含む同期していないデータセット（２１１３）の中の１つ以上のブロック（２３５２）を識別すること（２３０２）をさらに含み、これは多様な技術を使用し、実施されてよいことを指定する。１つの例の技術として、上記により詳細に説明されたように、同期中のストレージシステム（２１１４、２１２８）とアタッチするストレージシステム（２１２４）との間で異なる場合があるブロック（２３５２）のリストを生成することを含む一同期中のストレージシステムの１つは、特定のストレージシステムをデタッチすること、又は特定のストレージシステムがデタッチしたことを検出することに応じてブロックのリストの生成を開始してよい。さらに、ストレージシステムをデタッチすることは、その全体として参照により本明細書に援用される出願参照番号第１５／６９６，４１８号により詳細に説明される。

【０３８２】

【０４４０】 図２３に示される例の方法は、同期していないストレージシステム（２１２４）に記憶される同期していないデータセット（２１１３）と、１つ以上の同期中のストレージシステム（２１１４、２１２８）に記憶される同期中のデータセット（２１１２）との差異（２１５２）に従って、同期していないデータセット（２１１３）を同期中のデータセットと同期させること（２１０４）が、同期中のデータセット（２１１３）の中の１つ以上のブロック（２３５２）に一致するために同期していないデータセット（２１１３）の中の１つ以上のブロックを修正すること（２３０４）を含むことをさらに指定する。同期中のデータセット（２１１３）の中の１つ以上のブロック（２３５２）に一致するために同期していないデータセット（２１１３）の中の１つ以上のブロックを修正すること（２３０４）は一上記により詳細に説明されるように一デタッチされたブロックのリストから１つ以上のブロック（２３５２）の中のメモリアドレスの範囲の場所を突き止め、メモリアドレスのそれらの範囲からのコンテンツを、同期中のストレージシステム（２１１４、２１２８）の１つからアタッチするストレージシステム（２１２４）にコピーすることによって実施されてよい。

【０３８３】

【０４４１】 追加の説明のために、図２４は、本開示のいくつかの実施形態に従ってデータセットを同期複製するストレージシステムのための再同期のための追加の例の方法を示すフローチャートを説明する。図２４に示される例の方法も、同期していないストレージシステム（２１２４）に記憶される同期していないデータセット（２１１３）と、１つ以上の同期中のストレージシステム（２１１４、２１２８）に記憶される同期中のデータセット（２１１２）の差異（２１５２）を識別すること（２１０２）と、同期していないストレージシステム（２１２４）に記憶される同期していないデータセット（２１１３）と、１つ以上の同期中のストレージシステム（２１１４、２１２８）との差異（２１５２）に従って、同期していないデータセット（２１１３）を同期中のデータセット（２１１２）と同期させること（２１０４）と、同期していないストレージシステム（２１２４）に記憶される同期していないデータセット（２１１３）と、１つ以上の同期中のストレージシステム（２１１４、２１２８）に記憶される同期中のデータセット（２１１２）との間の同期複製関係を再確立すること（２１０６）も含むので、図２４に示される例の方法は図２１に示される例の方法に類似している。

【０３８４】

【0442】 しかしながら、図 2 4 に示される例の方法は、同期していないストレージシステム（2 1 2 4）に記憶される同期していないデータセット（2 1 1 3）と、1つ以上の同期中のストレージシステム（2 1 1 4、2 1 2 8）に記憶される同期中のデータセット（2 1 1 2）の差異を識別すること（2 1 0 2）が、1つ以上の同期中のストレージシステム（2 1 1 4、2 1 2 8）の少なくとも1つによって、同期していないストレージシステムがデタッチされて以後発生したデータセット（2 1 1 2）に対する1つ以上の修正（2 4 5 2）を識別すること（2 4 0 2）をさらに含むことを指定する。1つ以上の同期中のストレージシステム（2 1 1 4、2 1 2 8）の少なくとも1つによって、同期していないストレージシステムがデタッチされて以後発生したデータセット（2 1 1 2）に対する1つ以上の修正（2 4 5 2）を識別すること（2 4 0 2）は、いくつかの技術によって実施されてよい。1つの例の技術として、同期中のストレージシステム（2 1 2 4、2 1 2 8）の1つ以上は、上述されたブロック追跡技術の1つを実施してよい。他の例の技術は、上記に詳細に説明されるように、他の技術の中でも、ブロック追跡とスナップショットの組合せ、又はシーケンス番号によるブロック追跡を使用することを含む。

【0 3 8 5】

【0443】 追加の説明のために、図 2 5 は、本開示のいくつかの実施形態に従ってデータセットを同期複製するストレージシステムのための再同期の追加の例の方法を示すフローチャートを説明する。図 2 5 に示される例も、同期していないストレージシステム（2 1 2 4）に記憶される同期していないデータセット（2 1 1 3）と、1つ以上の同期中のストレージシステム（2 1 1 4、2 1 2 8）に記憶される同期中のデータセット（2 1 1 2）の差異（2 1 5 2）を識別すること（2 1 0 2）と、同期していないストレージシステム（2 1 2 4）に記憶される同期していないデータセット（2 1 1 3）と、1つ以上の同期中のストレージシステム（2 1 1 4、2 1 2 8）との差異（2 1 5 2）に従って、同期していないデータセット（2 1 1 3）を同期中のデータセット（2 1 1 2）と同期させること（2 1 0 4）と、同期していないストレージシステム（2 1 2 4）に記憶される同期していないデータセット（2 1 1 3）と、1つ以上の同期中のストレージシステム（2 1 1 4、2 1 2 8）に記憶される同期中のデータセット（2 1 1 2）との間の同期複製関係を再確立すること（2 1 0 6）とを含むので、図 2 5 に示される例の方法は図 2 1 に示される例の方法に類似している。

【0 3 8 6】

【0444】 しかしながら、図 2 5 に示される例の方法は、同期していないストレージシステム（2 1 2 4）に記憶される同期していないデータセット（2 1 1 3）と、1つ以上の同期中のストレージシステム（2 1 1 4、2 1 2 8）に記憶される同期中のデータセット（2 1 1 2）の差異を識別すること（2 1 0 2）が、1つ以上の同期中のストレージシステム（2 1 1 4、2 1 2 8）の少なくとも1つによって、同期していないストレージシステム（2 1 2 4）がデタッチされたとき未決であったデータセット（2 1 1 2）に対する1つ以上の修正（2 5 5 2）を識別すること（2 5 0 2）をさらに含むことを指定する。1つ以上の同期中のストレージシステム（2 1 1 4、2 1 2 8）の少なくとも1つによって、同期していないストレージシステム（2 1 2 4）がデタッチされたとき未決であったデータセット（2 1 1 2）に対する1つ以上の修正（2 5 5 2）を識別すること（2 5 0 2）は、より詳細に上述されるように、いくつかの技術を使用することによって実施されてよい。

【0 3 8 7】

【0445】 追加の説明のために、図 2 6 は、本開示のいくつかの実施形態にしたがってデータセットを同期複製するストレージのための再同期のための追加の例の方法を示すフローチャートを説明する。図 2 6 に示される例の方法も、同期していないストレージシステム（2 1 2 4）に記憶される同期していないデータセット（2 1 1 3）と、1つ以上の同期中のストレージシステム（2 1 1 4、2 1 2 8）に記憶される同期中のデータセット（2 1 1 2）の差異（2 1 5 2）を識別すること（2 1 0 2）と、同期していないストレージシステム（2 1 2 4）に記憶される同期していないデータセット（2 1 1 3）と、1つ

以上の同期中のストレージシステム（２１１４、２１２８）との差異（２１５２）に従って、同期していないデータセット（２１１３）を同期中のデータセット（２１１２）と同期させること（２１０４）と、同期していないストレージシステム（２１２４）に記憶される同期していないデータセット（２１１３）と、１つ以上の同期中のストレージシステム（２１１４、２１２８）に記憶される同期中のデータセット（２１１２）との間の同期複製関係を再確立すること（２１０６）とを含むので、図２６に示される例の方法は図２１に示される例の方法に類似している。

【０３８８】

【０４４６】 しかしながら、図２６に示される例の方法は、より詳細に上述されたように、多様な技術を使用し、実施されてよい、同期していないストレージシステム（２１２４）が、同期していないストレージシステム（２１２４）に記憶される同期していないデータセット（２１１３）と、１つ以上の同期中のストレージシステム（２１１４、２１２８）に記憶される同期中のデータセット（２１１２）との間の同期複製関係からデタッチしたことを検出すること（２６０２）をさらに含む。また、図２６に示される例の方法は、より詳細に上述されたように、同期していないストレージシステム（２１２４）がデタッチされて以後発生したデータセット（２１１２）に対する修正（２６５２）を追跡すること（２６０４）も含む。

【０３８９】

【０４４７】 追加の説明のために、図２７は、本開示のいくつかの実施形態に従ってデータセットを同期複製するストレージシステムのための再同期のための追加の例の方法を示すフローチャートを説明する。図２３に示される例の方法も、同期していないストレージシステム（２１２４）に記憶される同期していないデータセット（２１１３）と、１つ以上の同期中のストレージシステム（２１１４、２１２８）に記憶される同期中のデータセット（２１１２）の差異（２１５２）を識別すること（２１０２）と、同期していないストレージシステム（２１２４）に記憶される同期していないデータセット（２１１３）と、１つ以上の同期中のストレージシステム（２１１４、２１２８）との差異（２１５２）に従って、同期していないデータセット（２１１３）を同期中のデータセット（２１１２）と同期させること（２１０４）と、同期していないストレージシステム（２１２４）に記憶される同期していないデータセット（２１１３）と、１つ以上の同期中のストレージシステム（２１１４、２１２８）に記憶される同期中のデータセット（２１１２）との間の同期複製関係を再確立すること（２１０６）とを含むので、図２３に示される例の方法は、図２１に示される例の方法に類似している。

【０３９０】

【０４４８】 しかしながら、図２７に示される例の方法は、同期していないストレージシステム（２１２４）に記憶される同期していないデータセット（２１１３）と、１つ以上の同期中のストレージシステム（２１１４、２１２８）との差異（２１５２）に従って、同期していないデータセット（２１１３）を同期中のデータセット（２１１２）と同期させること（２１０４）が、より詳細に上述されたように、多様な技術を使用し、実装されてよい、１つ以上の同期中のシステム（２１１４、２１２８）の１つ以上のスナップショットを、同期していないストレージシステム（２１２４）に複製すること（２７０２）をさらに含むことを指定する。

【０３９１】

【０４４９】 追加の説明のために、図２８は、本開示のいくつかの実施形態に従ってデータセットを同期複製するストレージシステムのための再同期のための追加の例の方法を示すフローチャートを説明する。図２８に示される例の方法も、同期していないストレージシステム（２１２４）に記憶される同期していないデータセット（２１１３）と、１つ以上の同期中のストレージシステム（２１１４、２１２８）に記憶される同期中のデータセット（２１１２）の差異（２１５２）を識別すること（２１０２）と、同期していないストレージシステム（２１２４）に記憶される同期していないデータセット（２１１３）と、１つ以上の同期中のストレージシステム（２１１４、２１２８）との差異（２１５２）に

従って、同期していないデータセット（２１１３）を同期中のデータセット（２１１２）と同期させること（２１０４）と、同期していないストレージシステム（２１２４）に記憶される同期していないデータセット（２１１３）と、１つ以上の同期中のストレージシステム（２１１４、２１２８）に記憶される同期中のデータセット（２１１２）との間の同期複製関係を再確立すること（２１０６）とを含むので、図２８に示される例の方法は図２１に示される例の方法に類似している。

【０３９２】

【０４５０】 しかしながら、図２８に示される例の方法は、同期していないストレージシステム（２１２４）に記憶される同期していないデータセット（２１１３）と、１つ以上の同期中のストレージシステム（２１１４、２１２８）との差異（２１５２）に従って、同期していないデータセット（２１１３）を同期中のデータセット（２１１２）と同期させること（２１０４）が、より詳細に上述されるように、多様な技術を使用し、実施されてよい、同期していないストレージシステム（２１２４）だけで持続していたデータセット（２１１３）の１つ以上の修正（２８５２）を識別すること（２８０２）と、同期していないストレージシステム（２１２４）だけで持続していたデータセット（２１１３）の１つ以上の修正（２８５２）をアンドウすること（２８０４）をさらに含むことを指定する。

【０３９３】

【０４５１】 追加の説明のために、図２９は、本開示のいくつかの実施形態に従ってデータセットを同期複製するストレージシステムのための再同期の追加の例の方法を示すフローチャートを説明する。図２９に示される例の方法も、同期していないストレージシステム（２１２４）に記憶される同期していないデータセット（２１１３）と、１つ以上の同期中のストレージシステム（２１１４、２１２８）に記憶される同期中のデータセット（２１１２）の差異（２１５２）を識別すること（２１０２）と、同期していないストレージシステム（２１２４）に記憶される同期していないデータセット（２１１３）と、１つ以上の同期中のストレージシステム（２１１４、２１２８）との差異（２１５２）に従って、同期していないデータセット（２１１３）を同期中のデータセット（２１１２）と同期させること（２１０４）と、同期していないストレージシステム（２１２４）に記憶される同期していないデータセット（２１１３）と、１つ以上の同期中のストレージシステム（２１１４、２１２８）に記憶される同期中のデータセット（２１１２）との間の同期複製関係を再確立すること（２１０６）とを含むので、図２９に示される例の方法は図２１に示される例の方法に類似している。

【０３９４】

【０４５２】 しかしながら、図２９に示される例の方法は、指定する。同期していないストレージシステム（２１２４）に記憶される同期していないデータセット（２１１３）と、１つ以上の同期中のストレージシステム（２１１４、２１２８）に記憶される同期中のデータセット（２１１２）との間の同期複製関係を再確立すること（２１０６）が、より詳細に上述されたように、多様な技術を使用し、実装されてよい、同期していないストレージシステム（２１２４）のために、データベース（２１１３）のためのＩ／Ｏ処理を可能にすること（２９０２）をさらに含む。

【０３９５】

【０４５３】 追加の説明のために、図３０は、本開示のいくつかの実施形態に従って同期複製されたストレージシステム（３０１４、３０２４、３０２８）への接続性を管理するための例の方法を示すフローチャートを説明する。あまり詳細には示されないが、図３０に示されるストレージシステム（３０１４、３０２４、３０２８）は、図１Ａから図１Ｄ、図２Ａから図２Ｇ、図３Ａから図３Ｂ、又はその任意の組合せに関して上述されたストレージシステムに類似してよい。実際には、図３０に示されるストレージシステムは、上述されたストレージシステムと同じ構成要素、より少ない構成要素、追加の構成要素を含んでよい。

【０３９６】

【０４５４】 図３０に示される例の方法は、データセット（３０１２）がその全体で同期複

10

20

30

40

50

製される複数のストレージシステム（３０１４、３０２４、３０２８）を識別すること（３００２）を含む。図３０に示されるデータセット（３０１２）は、例えば特定のボリュームのコンテンツとして、ボリュームの特定のシャードのコンテンツとして、又は１つ以上のデータ要素の任意の他の集合体として実施されてよい。データセット（３０１２）は、各ストレージシステム（３０１４、３０２４、３０２８）がデータセット（３０１２）のローカルコピーを保持するように、複数のストレージシステム（３０１４、３０２４、３０２８）全体で同期されてよい。本明細書に説明される例では、係るデータセット（３０１２）は、少なくともアクセスされているクラスタ及び特定のストレージシステムが名目上動作している限り、クラスタ内の任意の一方のストレージシステムが、クラスタ内の任意の他方のストレージシステムよりも実質的により最適に動作しないような性能特性を有するストレージシステム（３０１４、３０２４、３０２８）のいずれかを通してデータセット（３０１２）にアクセスできるように、ストレージシステム（３０１４、３０２４、３０２８）全体で同期複製される。係るシステムでは、データセット（３０１２）に対する修正は、ストレージシステム（３０１４、３０２４、３０２８）のいずれかでデータセット（３０１２）にアクセスすることが一貫した結果を生じさせるように、各ストレージシステム（３０１４、３０２４、３０２８）に常駐するデータセットのコピーに対して加えられるべきである。例えば、データセットに発行される書込み要求は、すべてのストレージシステム（３０１４、３０２４、３０２８）で実行されなければならない、又はストレージシステム（３０１４、３０２４、３０２８）のいずれでも実行されてはならない。同様に、動作のいくつかのグループ（例えば、データセットの中の同じ場所に向けられる２つの書込み動作）は、各ストレージシステム（３０１４、３０２４、３０２８）に常駐するデータセットのコピーが、すべてのストレージシステム（３０１４、３０２４、３０２８）で最終的に同一となるように、同じ順序で、又はそれらがあたかも同じ順序で実行されたかのように、すべてのストレージシステム（３０１４、３０２４、３０２８）で実行されなければならない。データセット（３０１２）に対する修正は、正確に同時に行われる必要はないが、いくつかのアクション（例えば、書込み要求がデータセットに向けられ、すべてのストレージシステムでまだ完了していない書込み要求によってターゲットにされるデータセットの中の場所への読取りアクセスを可能にする旨の肯定応答を発行すること）は、各ストレージシステム（３０１４、３０２４、３０２８）上のデータセット（３０１２）のコピーが修正されるまで先延ばしにされてよい。

【０３９７】

【０４５５】 書込み要求（又は複数のストレージシステム全体で同期複製されるデータセットを修正する他の要求）の処理と対照的に、他のタイプの要求は、係る動作の分散メッセージング待ち時間を増加することなく、要求を受け取ったストレージシステムによってローカルにサービスを提供されてよい。例えば、読取り要求、問合せ要求、又はデータセット（３０１２）の修正を生じさせない他の要求は、通常は、係る動作の分散メッセージング待ち時間を増加することなく、要求を受け取ったストレージシステムによってローカルに処理できる。例えば、ホストが、データセット（３０１２）がその全体で同期分散されるストレージシステム（３０１４、３０２４、３０２８）のクラスタの中の第１のストレージシステム（３０１４）に読取り要求を発行する場合、次いで実施態様は、ストレージシステム（３０１４、３０２４、３０２８）の間に、読取り要求を完了するために必要とされ、頻繁にローカルの複製されていないストレージシステムの待ち時間と同一である読取り待ち時間を生じさせるインラインメッセージングがないことを確実にできる。一部の例では、係る動作（例えば、読取り要求）は、実施態様の中でブロックされ得る。書込み要求（つまり、読取り要求に応じて読み取られるデータセットの部分と重複するデータセットの部分にデータを書き込む要求）、又はストレージシステム（３０１４、３０２４、３０２８）のすべてでまだ完了していない他の形の矛盾した修正動作を構成することによって実施態様の中でブロックされ得る。ブロックは、例えば、１つ以上の同時に起こる修正要求と時間的に重複する複数の読取り要求に対する順序付け要件を保つために必要である場合がある。係るブロックは、同じ書込み又は他の修正動作と重複する、ポッド内の同

じストレージシステム又は別のストレージシステム上の書込み動作又は他の修正動作と同時に起こり、ポッド内の別のストレージシステムで第1の読取りの後に第2の読取りが続く、1つのストレージシステムでの第1の読取りが、第2の読取りのための修正動作の前からコンテンツを戻す一方、第1の読取りのための修正動作の結果を絶対に戻さないことを確実にするために使用できる。ストレージシステムがそれについて知った、及びポッド内のどこでもまだ処理されていない飛行中の修正動作のための重複する読取り要求のブロックは、読取り動作のためのこの逆の時間順が、他のすべての重複する読取り要求もその重複する修正動作から結果を返すことも確定するまで、重複する修正動作からの結果を返す可能性がある読取り要求を先延ばしにすることによって起こらないことを確実にできる。

10

【0398】

【0456】 図30に示される例の方法では、データセット(3012)のために、データセット(3012)がその全体で同期複製される複数のストレージシステム(3014、3024、3028)を識別すること(3002)は、例えばデータセット(3012)を、そのデータセット(3012)を名目的に記憶する1つ以上のストレージシステム(3014、3024、3028)と関連付けるポッド定義又は類似するデータ構造を調べることによって実施されてよい。『ポッド』は、該用語がここで及び本願の残りを通して使用されるように、データセット、管理オブジェクト及び管理動作の集合、データセットを修正する又は読み取るためのアクセス動作の集合、及び複数のストレージシステムを表す管理エンティティとして実施されてよい。係る管理動作は、データセットを読み取る又は修正するためのアクセス動作が、ストレージシステムのいずれかを通して同等に動作する、ストレージシステムのいずれかを通して同等に管理オブジェクトを修正する又は問い合わせしてよい。各ストレージシステムは、ストレージシステムによる使用のために記憶され、宣伝されたデータセットの適切な部分集合としてデータセットの別々のコピーを記憶してよく、任意の1つのストレージシステムを通して実行され、完了された管理オブジェクト又はデータセットを修正する動作は、ポッドを問い合わせるために後続の管理オブジェクトで、又はデータセットを読み取るために後続のアクセス動作で反映される。『ポッド』に関する追加の詳細は、参照により本明細書に援用される以前に出願された仮特許出願第62/518,071号で見つけられてよい。ストレージシステムはポッドに加えられる場合があり、結果的にポッドのデータセット(3012)は、そのストレージシステムにコピーされ、次いでデータセット(3012)が修正されると、最新に保たれる。また、ストレージシステムは、ポッドから削除される場合もあり、結果的にデータセット(3012)は削除されたストレージシステムでもはや最新に保たれなくなる。係る例では、ポッド定義又は類似するデータ構造は、ストレージシステムが特定のポッドに加えられる、特定のポッドから削除されると更新されてよい。

20

30

【0399】

【0457】 また、図30に示される例の方法は、データセット(3012)に向けられるI/O動作を発行できるホスト(3032)を識別すること(3004)も含む。図30に示されるホスト(3032)は、例えばストレージシステム(3014、3024、3028)にとって外部で実行中のアプリケーションサーバとして、又は1つ以上のデータ通信経路を介してストレージシステム(3014、3024、3028)にアクセス要求(例えば、読取り、書込み)を発行する任意の他のデバイスとして実施されてよい。データセット(3012)に向けられたI/O動作を発行できる特定のホスト(3032)を識別すること(3004)は、例えば、ストレージシステム(3014、3024、3028)の1つ以上が、リスト、又はストレージシステム(3014、3024、3028)がそこからデータセット(3012)に向けられるI/O動作を受け取った各ホストの識別を含む他のデータ構造を保持することによって、リスト、又はデータセット(3012)にアクセスするために必要な適切な許可を有する各ホストを識別する他のデータ構造を調べることによって、又は他のなんらかの方法で実施されてよい。

40

【0400】

50

[0458] また、図30に示される例の方法は、ホスト(3032)と、データセット(3012)がその全体で同期複製される複数のストレージシステム(3014、3024、3028)との間の複数のデータ通信経路(3022、3026、3030)を識別すること(3006)も含む。ホスト(3032)と複数のストレージシステム(3014、3024、3028)との間の各データ通信経路(3022、3026、3030)は、ホスト開始プログラムポートとストレージシステムターゲットポートとの間の、ホスト上のネットワークインタフェースとストレージシステム上のネットワークインタフェースとの間等の関係を表す場合がある。係る例では、いくつかのホスト開始プログラムポート及びいくつかのストレージシステムターゲットポートがある場合があり、ストレージシステムは、それぞれが複数のターゲットポートをホストし得るいくつかのストレージコントローラを含む場合もある。別々のストレージシステム上のターゲットポート又はネットワークインタフェースは、同じポッド内であっても、通常、互いとは別個であるべきである。ターゲットポートは、アクティブ／最適化済み、アクティブ／未最適化、スタンバイ、及びオフラインに関して共通の状態を共用する、ストレージシステムボリュームと関連付けられたポートのグループであるターゲットポートグループを使用し、管理されてよい。ターゲットポートグループは、全体としてストレージシステムと関連付けられるよりむしろ、個々のストレージシステムの各ストレージコントローラと関連付けられてよい。実際には、ターゲットポートグループは、単一のストレージコントローラの中でもターゲットポートの部分集合と関連付けられることを含み、完全に任意である場合がある。また、ストレージシステムは、ターゲットポートグループを構築する又はターゲットポートグループに知らせる際に、ホスト開始プログラム情報も使用できるであろう。ただし、ストレージシステムは、マルチパッシングドライバスタックでの混乱がないことを確実にするために、各ホスト開始プログラムにこの情報を一貫して提供しなければならない。図30に示される例の方法では、ホスト(3032)と、データセット(3012)がその全体で同期複製される複数のストレージシステム(3014、3024、3028)との間の複数のデータ通信経路(3022、3026、3030)を識別すること(3006)は、例えば、以下の段落でより詳細に説明されるSCSI非対称論理ユニットアクセス(『ALUA』)機構の使用により、なんらかの他のネットワーク発見ツールの使用により、又は他のなんらかの方法で実施されてよい。

[0401]

[0459] また、図30に示される例の方法は、ホスト(3032)と、データセット(3012)がその全体で同期複製される複数のストレージシステム(3014、3024、3028)との間の複数のデータ通信経路(3022、3026、3030)の中から1つ以上の最適経路を識別すること(3008)も含む。図30に示されるストレージシステムは、ホスト(3032)と、データセット(3012)がその全体で同期複製される複数のストレージシステム(3014、3024、3028)と、ストレージシステムと関連付けられたストレージ通信エンドポイントの間の複数のデータ通信経路(3022、3026、3030)の中から1つ以上の最適経路を識別してよい(3008)。図30に示される例の方法では、1つ以上の最適経路を識別すること(3008)は、単一の最適経路を識別すること又は複数の最適経路を識別することを含んでよい。例えば、多様な性能スレッシュホールドを満たす任意の経路が識別されてよい、所定数の最も適した経路(例えば、他の利用可能な経路に対して優れた性能を示すそれらの経路)が識別されてよい、所定の割合の最も適した経路が識別されてよい等、(ホストと特定のストレージシステムとの間の経路等の)より多くの最適経路の部分集合が識別されてよい。読者は、ストレージシステムホスト(3014、3024、3028)は、互いからなんらかの距離に位置してよいので、ストレージシステム(3014、3024、3028)は、別々のストレージネットワーク又はストレージネットワークの別々の部分に位置してよいので、又は他のなんらかの理由から、ホスト(3032)が、別のストレージシステムに対してあるストレージシステムにI/O動作を発行することに関連付けられた性能優位点があつてよいことを理解する。例えば、遠いデータセンタ又はキャンパスの中に物理的に位置するス

ストレージシステムにI/O動作を発行するホスト（3032）に比して、ホスト（3032）が、ホスト（3032）と同じデータセンタ又はキャンパスの中に物理的に位置するストレージシステムにI/O動作を発行することに関連付けられた性能優位点がある場合がある。信頼性のために、ホスト（3032）にすべてのストレージシステム（3032）に対する接続性を保持させることは有益であってよいが、性能のために、ホスト（3032）が特定のストレージシステムを通してデータセット（3012）にアクセスすることが好ましい場合がある。読者は、異なるホストがデータセット（3012）にアクセスする場合があるため、あるホストがデータセット（3012）にアクセスするための1つ以上の最適経路が、別のホストがデータセット（3012）にアクセスするための1つ以上の最適経路とは異なる場合があることを理解する。いくつかの実施形態では、2つのストレージシステムが十分に類似しているため、両方のストレージシステムへの経路が最適と見られ得ることが考えられる。例えば、2つのストレージシステムが同じデータセンタ又はキャンパスにあり、ホストとそれらの2つのストレージシステムとの間に豊富なネットワークを有する場合、第3のストレージシステムは十分に遠いので、第3のストレージシステムはフォールバックとして、を除き、使用されるべきではないが、ホストと2つの十分に類似したストレージシステムとの間の経路は、すべて最適経路として識別される（3008）ことに対する候補となってよい。

【0402】

【0460】 図30に示される例の方法では、ホスト（3032）と、データセット（3012）がその全体で同期複製される複数のストレージシステム（3014、3024、3028）との間の複数のデータ通信経路（3022、3026、3030）の中から1つ以上の最適経路を識別すること（3008）は、例えば、SCSI ALUA機構の使用により実施されてよい。SCSI ALUA機構は、コマンド及び問合せの集合として、（SCSIでは「論理ユニットとしても知られる」）ストレージボリュームへの複数のターゲットポートを通したストレージシステムボリュームへの非同期アクセスに対するサポートを記述するSCSI SPC-4技術規格及びSBC-3技術規格に説明される。係る実施形態では、（そのコンテンツが、複数のストレージシステム全体で同期複製されるデータセットを表す場合がある）ボリュームは、1つ以上のターゲットポートグループと関連付けられた複数のSCSIターゲットポートを通してホスト（3032）に一意のIDを報告することができ、これによりホスト（3032）は、1つ以上のSCSIホストポートを通して、そのホストポート及び宣伝されたターゲットポートのいくつかの又はすべての組合せを通してボリュームにアクセスするようにそのI/Oドライバを構成できる。一意のボリュームIDは、同じボリュームにアクセスするすべてのSCSI論理ユニット番号、ホストポート、及びターゲットポート句の組合せを認識するためにホスト（3032）I/Oドライバによって使用される場合がある。ホストI/Oドライバは、次いで識別されたボリュームの状態及びコンテンツを修正するために、それらの組合せ（経路）のいくつか、いずれか、又はすべての下方にSCSIコマンドを発行することができる。障害の結果、ホストは代替経路を下方に要求を再発行する場合があり、性能の懸念により、ホスト（3032）は、複数のポート及び複数のネットワークの相互接続の使用により、改善されたホスト対ストレージシステム帯域幅の利点を得るために複数の経路を多大に利用する場合がある。

【0403】

【0461】 SCSIのためのALUA仕様を用いて、それぞれがそれぞれ状態を割り当てられる場合がある、ボリュームにアクセスできる複数のターゲットポートグループを説明することができる。ターゲットポートグループは、ストレージシステム上の1つ以上のSCSIポートを表す場合がある。マルチコントローラストレージシステムでは、ターゲットポートグループは、1つのコントローラ上のすべてのSCSIターゲットポートを表す可能性がある、又は対称でアクセス可能な同期複製されたストレージシステムを用いると、ターゲットポートグループは、個々のストレージシステム上のすべてのSCSIターゲットポートを表す場合がある、又はターゲットポートは他のなんらかの方法でグループ化

される場合がある。ターゲットポートグループと関連付けることのできる状態は、ポートグループが、I/O（アクティブ／最適化済み）を発行するために好まれるべきか、I/O（アクティブ／非最適化）、スタンバイ（I/Oが、アクティブ／最適化済み又はアクティブ／非最適化に戻る状態の変化があるまで発行できない）を発行するために好まれるべきではないのか、又はターゲットポートは応答しないため等でオフラインとなる場合があるのかを示す。SCSI仕様は、ターゲットポートグループ及び状態を割り当てられたALUAターゲットポートグループの定義を、各ボリュームに特有だけでなく、各要求ホストに（又は各要求ホストポートにも）特有とすることができ、これによりボリュームごとに、ストレージシステムは、ターゲットポートグループ、及び状態を割り当てられたターゲットポートグループの一意の集合を、各ホスト又はそのボリュームにアクセスできるホストポートに提示できる。

10

【0404】

【0462】 対称でアクセス可能な同期複製されたストレージシステムを用いると、ポッド内のすべてのストレージシステムは、ポッド内のすべてのストレージシステムが、ホストにとって、あたかもポッドのためのいくつかの又はすべてのストレージシステムでSCSIターゲットポートを通して同じボリュームを提示する1つのストレージシステムであるかのようにホストに同じボリュームを提示できる。これらの機構は、次いでポッド内のボリュームへアクセスを向ける、及び向け直すために所望されるすべての機能を提供できる。（例えば、ネットワーク又はホストのそのストレージシステムに対する地理的な近接性のために）ポッドのための特定のストレージシステムにとってより良い性能を得るホストの場合、そのホストのホストポートに対するそのストレージシステムのための状態を割り当てられたALUAターゲットポートグループは、アクティブ／最適化済みとして示される場合がある。一方、ポッドのためのその特定のストレージシステムに対しより少ない性能を得る他のホストの場合、その他の他のホストポートに対するストレージシステムのための状態が割り当てられたALUAターゲットポートグループは、アクティブ／非最適化として示される場合がある。このようにして、アクティブ／最適化であると判断されるターゲットポートグループのメンバーは、最適経路（複数可）としてホストに識別されてよい（3008）。

20

【0405】

【0463】 新しいストレージシステムがポッドに加えられる場合、次いで新しいターゲットポートグループがそのポッド内のボリュームごとに、そのボリュームにアクセスするホストポートに加えられ、ターゲットポートグループは、新しいストレージシステムのためのホスト／ストレージシステム近接性に適切な状態を割り当てられる場合がある。なんらかの数のSANレベルイベントの後、ホストは、ボリュームごとに新しいポートを認識し、新しい経路を適切に使用するようにそのドライバを構成できる。ストレージシステムは、ポッドの代わりに、ホストが現在新規に加えられたストレージシステムでSCSIターゲットポートを使用するように適切に構成されていると判断するためにホストアクセス（例えば、REPORT LUNSコマンド及びINQUIRYコマンドを待機する等）がないかモニタできる。ストレージシステムがポッドから削除される場合、次いでポッドに留まる他のストレージシステムは、ポッドのボリュームのための削除されたストレージシステムについて任意のターゲットポート又はターゲットポートグループを任意のホストポートに報告するのを停止できる。さらに、削除されたストレージシステムは、任意のREPORT LUN要求でポッドのボリュームを一覧することを停止することができ、ストレージシステムは、ポッドのボリュームに対するコマンドに応じてボリュームが存在していないことを報告し始めることができる。ボリュームがポッドの中に又はポッドの中から移動され、結果的にボリュームがストレージシステムの拡大又は削減された集合と関連付けられる場合、同じアクションは、ストレージシステムを加える又はポッドからストレージシステムを削除する際に適用されるであろう個々のボリュームに適用できる。障害の取り扱いに関しては、任意の係る経路が利用可能であり、適切に機能している場合には、ホストI/Oドライバは、アクティブ／最適化済みとして割り当てられるターゲットポート

30

40

50

グループ内のターゲットポートを通してそれらのボリュームにアクセスするが、アクティブ／非最適化経路が利用不能であり、適切に機能していない場合、アクティブ／非最適化経路に切り替えることができる。

【0406】

【0464】 また、図30に示される例の方法では、ホスト（3032）と、データセット（3012）がその全体で同期複製される複数のストレージシステム（3014、3024、3028）との間の複数のデータ通信経路（3022、3026、3030）の中から1つ以上の最適経路を識別すること（3008）は、例えばポッド内の特定のインタフェース又はストレージシステムに対するホスト経路が、他の特定のインタフェース又はポッド内のストレージシステムへのホスト経路よりも、より低い待ち時間、より良いスループット、又はより少ない切替えインフラストラクチャを有することを判断するために、タイミング情報又はネットワーク情報を使用することによって自動化された方法で実施されてよい。したがって、係る例では、ホスト（3032）と、データセット（3012）がその全体で同期複製される複数のストレージシステム（3014、3024、3028）との間の複数のデータ通信経路（3022、3026、3030）の中から1つ以上の最適経路を識別すること（3008）は、ホストとストレージシステムの1つとの間の相対的に最も低い待ち時間を示す1つ以上のデータ通信経路を識別することと、所定の閾値以下である、ホストとストレージシステムの1つとの間の待ち行列を示す1つ以上のデータ通信経路を識別すること等を含む。

【0407】

【0465】 例えばIPベースのネットワークでは、pingコマンド及びtracerouteコマンド（又はその基本的なICMPエコー要求を活用すること）が、待ち時間、及び既知のホストネットワークインタフェースとポッド内のストレージシステムとの間のネットワークルートを決定するために使用されてよい。tracerouteファシリティ、又は（ルートがICMP時間超過応答を送信する前のネットワークホップの数を制限するための）徐々に増加するTTLフィールドを有するICMPエコー要求の直接的な使用は、より高い待ち時間を有する特定のネットワークホップがあると判断するために、又は係るネットワークホップがないことを判断するために使用できる。この場合では、高い待ち時間ホップを有するストレージインタフェースルートに対するホストインタフェースは、高待ち時間ホップのないストレージインタフェースルートへのホストインタフェースをサポートして回避される場合がある。代わりに、他に比べてほとんどホップを有さず、低待ち時間を有するネットワークルートがある場合、ネットワークインタフェースを有するストレージシステムが好まれる場合がある。

【0408】

【0466】 ファイバチャネルベースのネットワークでは、そのFibre Channel Working GroupからのStorage Networking Industry AssociationによってサポートされるプラグインライブラリのHBA API仕様は、FCストレージネットワークを精密に計画するために使用できる。また、ファイバチャネルプロトコルのELSエコー機能も、ネットワーク待ち時間を検出するために使用できる。上述されたIPネットワークと同様に、これは、他のホストポートとターゲットポートの組合せよりも低い待ち時間及び少ないネットワークホップを有するターゲットポートネットワークに対して、ホストポートを識別するために使用でき、このことは、ホストごとのアクティブ／最適化済み対アクティブ／非最適についてポッド内のストレージシステムを構成するために、ポッド内のどのストレージシステムがあるホスト又は別のホストにより近くに又はより良く接続されるのかを判断するために使用できる。

【0409】

【0467】 また、図30に示される例の方法は、ホスト（3032）に対して、1つ以上の最適経路の識別を示すこと（3010）も含む。図30に示される例の方法では、ストレージシステム（3014、3024、3028）は、例えばストレージシステム（3014、3024、3028）とホスト（3032）との間で交換される1つ以上のメッセ

ージにより、ホスト（3032）に最適経路の識別を示してよい（3010）。係るメッセージは、上述された機構の多くを使用し、交換されてよく、ポート識別子、ネットワークインタフェース識別子、又はなんらかの他の識別子の使用により最適経路を識別してよい。例えば、ホスト（3032）に対して、1つ以上の最適経路の識別を示すこと（3010）は、SCSE ALUA機構を介してホスト（3032）に対して1つ以上のアクティブ／最適化済み経路を示すこと（434）によって実施されてよい。

【0410】

【0468】 読者は、本明細書に説明されるストレージシステム（3014、3024、3028）が、ポート又はネットワークインタフェースの名前を付けられた集合としてホスト（3032）を定義するためにホスト定義を活用してよく、それらのホスト定義は、追加の情報又は例えばオペレーティングシステム若しくはアプリケーションタイプ若しくは作業負荷の分類等の追加の特徴を含んでよいことを理解する。ホスト定義は、管理の便宜性のために名前を与えられ、ストレージシステムユーザインタフェースの第1級のオブジェクトとして表される場合もあれば、特定のアプリケーション、ユーザ、又はホストベースのデータベース若しくはファイルシステムクラスと関連付けられたすべてのホストを一覧するために、多様な方法とともにグループ化される場合もある。これらのホスト定義は、ストレージシステムがホスト場所についての情報を関連付けるために、又はポッドのためのホスト対ストレージシステムの優先度のための便利な管理オブジェクトとしての機能を果たす場合がある。ストレージシステムが、ポッドごとに1つのホスト定義よりむしろ、ホストごとに1つのホスト定義を管理することが便利な場合がある。ホストと関連付けられた開始プログラムポート及びネットワークインタフェースは、すべてのポッドについておそらく同じであるため、それは便利な場合がある。ポッドが、各ポッドが他のポッドから安全に分離される強力な形の仮想機器として使用される場合、これは当てはまらない場合があるが、これは係る安全なポッド分離に達していない任意の用途又は実施態様のためにセットアップするためには便利且つより容易である場合がある。

【0411】

【0469】 ポッドがポッド内の複数のストレージシステムから同じホストにデータセット又はストレージオブジェクトを供給する場合、並びにそのホストのためのALUA状態、及びホスト対ストレージシステムの優先度が、ポッドのためのすべてのストレージシステム全体で調整されたやり方で管理されなければならない場合、次いでホスト定義はポッド全体で調整又は同期される必要がある場合があるが、ポッドのための大部分の他の管理されているオブジェクトとは異なり、ホストは、ポッドオブジェクトよりむしろ（ネットワークインタフェース及びSCSEターゲットポートは多くの場合ストレージシステムオブジェクトであるため）ストレージシステムオブジェクトであってよい。結果として、定義が調和しない場合があるため、ホストオブジェクトはポッドメンバー間で容易に同期されない場合がある。

【0412】

【0470】 さらに、ホストは、ホスト側開始プログラムとネットワークインタフェースの1つの集合を通してポッドのためのあるストレージシステムに、及びホスト側開始プログラム及びネットワークインタフェースの異なる集合を通してポッドのための別のストレージシステムに相互接続されてよい。さらに、2つの集合の間にはなんらかの重複がある場合がある、又は2つの集合の間には重複がない場合がある。一部の 경우에는、インタフェースが同じホストを表すと判断するためにストレージシステムが使用できるホスト情報がある場合がある。例えば、インタフェースが、同じiSCSI IQNを使用する場合もあれば、多様な開始プログラム又はネットワークインタフェースが同じホストを表すことを示すために、ホスト側ドライバがストレージシステムにホスト情報を供給する場合もある。他の場合には、係る情報がない場合がある。発見可能な情報がない場合は、ホスト定義のパラメータは、代わりに、ホスト名をネットワークエンドポイント、iSCSI IQN、開始プログラムポート等の集合に関連付けるために、ユーザによって又はなんらかのAPI若しくは他のインタフェースによりストレージシステムに供給されてよい。

【0413】

【0471】 ポッドと関連付けられたデータセットの部分が、ホスト定義を通して特定のホストにエクスポートされる（つまり、それがネットワークエンドポイントのリスト、i S C I I Q N、又は開始プログラムポートを通して、ポッドの現在のストレージシステムの独自のネットワークエンドポイント、及び S C S I ターゲットからホスト定義に基づいたホストに提供される）場合、次いで追加のストレージシステムがポッドに加えられると、加えられたストレージシステムのホスト定義を調べることができる。同じホストオブジェクト名の付いたホストも、ホストネットワークエンドポイント、i S C I I Q N、又は開始プログラムポートの重複するリストを有するホストも加えられたストレージシステムに存在しない場合、次いでホスト定義は加えられたストレージシステムにコピーできる。ホストネットワークエンドポイント、i S C S I I Q N、及び開始プログラムポートの同じ名前及び同じ構成を有するホスト定義が加えられたストレージシステムに存在する場合、次いで元のポッドメンバーストレージシステムからのホスト定義、及び加えられたストレージシステムのホスト定義は、それ以降リンクし、調整することができる。同じ名前であるが、ホストネットワークエンドポイント、i S C I I Q N、又は開始プログラムポートの異なる構成を有するホストが、加えられたストレージシステムに存在する場合、次いでホストオブジェクトの限定版が、ポッドのためのストレージシステム間で交換でき、異なるバージョンはストレージシステム限定子で名前を付けられる。例えば、ストレージシステム A は、A : H 1 としてそのホスト定義を提示してよい。一方、ストレージシステム B は、B : H 1 としてそのホスト定義を提示してよい。同は、名前は異なるが、ホストネットワークエンドポイント、i S C S I I Q N、又は開始プログラムポートでなんらかの重複を有するホスト定義にも提供される場合がある。その場合、ホスト定義はストレージシステム間でコピーされない場合があるが、代わりにストレージシステムはローカルのまま留まり、例えばホスト開始プログラム X 及び Y を一覧するホスト定義 A : H 1、並びにホスト開始プログラム Y 及び X を一覧するホスト定義 B : H 2 を生じさせる場合がある。さらに、動作は、これらのホスト定義を同期させるために提供される場合がある。2つのホスト定義が同じ名前、又はホストネットワークエンドポイント、i S C S I I Q N、又は開始プログラムポートの重複する集合を有する場合、次いでユーザは、共通名、次いでともにリンクされる交換されたホストネットワークエンドポイント、i S C S I I Q N、又は開始プログラムポートとともにそれらを統一するための簡略なインタフェースを提供される場合がある。2つの定義間の唯一の衝突が、いくつかのホスト定義が、別のインタフェースに一覧されないホストインタフェースを含むが、少なくとも1つのホストインタフェースが一致することである場合、これらの定義は、ユーザに係る要求を行うのを待機するよりむしろ、自動的にマージされ、リンクできるであろう。

【0414】

【0472】 ポッド内のストレージシステムが、一連の条件（障害、停止等）の結果として別のストレージシステムをデタッチする場合、次いでデタッチされたストレージシステムは、それがポッドのためにオフラインであるが、それ以外の場合はまだ実行中である場合、そのホスト定義に変更を加える場合がある。また、ポッドのためにオンラインに留まるストレージシステムは、そのホスト定義に対しても変更を加える場合がある。結果は一致しないホスト定義である場合がある。デタッチされたストレージシステムが後にポッドに再接続される場合、次いでホスト定義は一致なくなる場合がある。その時点で、ポッドは、各ストレージシステムでの別々の定義を区別するために、そのストレージシステム名接頭辞とともにホスト定義を報告することを再開してよい。

【0415】

【0473】 ホスト定義の別の態様は、ホスト定義が、どのストレージシステムのターゲットポートがアクティブ／最適化済みの A L U A ステータスを生じさせるべきなのか、及びどれがアクティブ／非最適化の A L U A ステータスを生じさせるべきなのかの観点で、何が A L U A 情報のために返されるのかを構成することの一部として、場所又はストレージシステムの優先度を定義してよい点である。また、この状態は、ストレージシステム間で

調整され、リンクされる必要がある場合もある。そうだとしたら、それは調整を要する別の態様である場合がある。また、ポッド内のストレージシステムのためにホスト定義を調整するときに検出される衝突、つまり任意の設定の欠如は、ユーザに場所又はストレージシステムの優先度を設定するのを促す機会も提示してよい。

【0416】

【0474】 読者は、ホスト定義が、実際には、ポッドレベルのオブジェクトよりむしろストレージシステムレベルのオブジェクトである場合があるので、同じホスト定義が、単一のストレージシステムを越えて広がらないポッドに対してだけではなく、複数のストレージシステム間で広がったポッドに対しても使用できることに留意する。純粋にローカルなポッド（つまり別のポッドよりストレージシステムの異なる集合に広がるあるポッド）との関連でのホストの使用は、ホスト（又はホストのリスト）がどのように見られるのかを改変できるであろう。例えば、ローカルポッドとの関連では、ストレージシステムによってホストを限定することは、意味をなさない場合があり、ローカルストレージシステムに対する経路を有さないホストを一覧することも意味をなさない場合がある。この例は、異なるメンバーストレージシステムを有するポッドに拡張できるであろう。例えば、ホスト定義は、（ストレージシステム限定を有用にする）1つのポッドのために、対にされたストレージシステムと衝突する場合がある。一方、ホスト定義は、（ストレージシステム限定を潜在的に不必要にする）異なるポッドのために、異なる対にされたストレージシステムと衝突しない場合がある。類似する問題は、あるポッドが追加のストレージシステムのある集合に広げられ、別のポッドが追加のストレージシステムの別の集合に広げられるときにホストの使用で発生する場合がある。その場合、特定のポッドの問題に関連性のあるストレージシステムのためのターゲットインタフェースだけ、並びにポッドのための関連性のあるストレージシステム上のターゲットインタフェースに対して可視であるホストエンドポイント、iSCSI IQN、及び開始プログラムポートだけ。

【0417】

【0475】 図30に示されるストレージシステム（3014、3024、3028）の1つだけしか、上述されたステップを実行するとして明示的に示されていないが、読者は、ストレージシステム（3014、3024、3028）のそれぞれが、上述されたステップを実行してよいことを理解する。実際には、ストレージシステム（3014、3024、3028）のそれぞれは、ほぼ同時に上述されたステップを実行してよく、これにより最適経路の識別は調整された取組みとなる。例えば、各ストレージシステム（3014、3024、3028）は、それ自体とホストとの間のすべてのデータ通信経路を個々に識別し、それ自体とホストとの間の各データ通信と関連付けられた多様な性能測定基準を収集し、1つ以上の最適経路を識別しようと努力して係る情報を他のストレージシステムと共用してよい。

【0418】

【0476】 追加の説明のために、図31は、本開示のいくつかの実施形態に従って同期複製されたストレージシステム（3014、3024、3028）への接続性を管理するための追加の例の方法を示すフローチャートを説明する。あまり詳細には示されないが、図31に示されるストレージシステム（3014、3024、3028）は、図1Aから図1D、図2Aから図2G、図3Aから図3B、又はその任意の組合せに関して上述されたストレージシステムに類似してよい。実際には、図31に示されるストレージシステムは、上述されたストレージシステムと同じ構成要素、より少ない構成要素、追加の構成要素を含んでよい。

【0419】

【0477】 図31に示される例の方法は、データセット（3012）がその全体で同期複製される複数のストレージシステム（3014、3024、3028）を識別すること（3002）と、データセット（3012）に向けられるI/O動作を発行できるホスト（3032）を識別すること（3004）と、ホスト（3032）と、データセット（3012）がその全体で同期複製される複数のストレージシステム（3014、3024、3

028) との間の複数のデータ通信経路 (3022、3026、3030) を識別すること (3006) と、ホスト (3032) と、データセット (3012) がその全体で同期複製される複数のストレージシステム (3014、3024、3028) との間の複数のデータ通信経路 (3022、3026、3030) の中から1つ以上の最適経路を識別すること (3008) と、ホスト (3032) に対して、1つ以上の最適経路の識別を示すこと (3010) とを含むので、図31に示される例の方法は図30に示される例の方法に類似している。

【0420】

【0478】 また、図31に示される例の方法は、データセット (3012) が、ストレージシステムの更新された集合全体で同期複製されることを検出すること (3102) も含む。図31に示される例の方法では、データセット (3012) が同期複製されるストレージシステムの集合は、さまざまな理由から変化する可能性がある。データセット (3012) が同期複製されるストレージシステムの集合は、例えば1つ以上の適切に機能しているストレージシステムがポッドに加えられる、又はポッドから削除されるために変化する可能性がある。さらに、データセット (3012) が同期複製されるストレージシステムの集合は、例えば1つ以上のストレージシステムが到達不能になり、到達不能又は利用不能になったことに応えて、ポッドからデタッチされるために変化する可能性がある。図31に示される例では、データセット (3012) が、ストレージシステムの更新された集合全体で同期複製されることを検出すること (3102) は、例えば、ポッド定義に対する変更を検出することによって、ストレージシステムが到達不能になった、若しくはそれ以外の場合利用不能になったことを検出することによって、又はなんらかの他の方法で実施されてよい。

【0421】

【0479】 図31に示される例の方法では、データセット (3012) がストレージシステムの更新された集合全体で同期複製される実施形態が示される。係る例では、ストレージシステム (3024) は、データセット (3012) が同期複製される複数のストレージシステム (3014、3024、3028) の1つとして当初識別される (3002) として図30に説明された。しかしながら、図31に示される例は、ストレージシステム (3024) が到達不能になった、又はそれ以外の場合利用不能になった、本明細書では、他のストレージシステム (3014、3028) 間のデータ通信リンク (3016、3018、3020) 及びストレージシステム (3024) とのデータ通信のためにストレージシステム (3024) とホスト (3032) を結合するために使用できるデータ通信経路 (3030) のための (アクティブな接続がないことを表す) 点線の使用により示される実施形態を示す。したがって、データセット (3012) がその全体で同期複製されるストレージシステムの更新された集合は2つのストレージシステム (3014、3028) しか含んでいないのに対して、データセット (3012) がその全体で同期複製されたストレージシステムの初期の集合は、示されているすべてのストレージシステム (3014、3024、3028) を含んでいた。

【0422】

【0480】 また、図31に示される例の方法は、ホスト (3032) とストレージシステム (3014、3028) の更新された集合との間の複数のデータ通信経路を識別すること (3104) も含む。図31に示される例の方法では、ホスト (3032) とストレージシステム (3014、3028) の更新された集合との間の複数のデータ通信経路を識別すること (3104) は、例えば、より詳細に上述される SCSI ALUA 機構の使用により、他のなんらかのネットワーク発見ツールの使用により、又はなんらかの他の方法で実施されてよい。

【0423】

【0481】 また、図31に示される例の方法は、ホスト (3032) とストレージシステムの更新された集合との間の複数のデータ通信経路の中から最適経路の更新された集合を識別すること (3106) も含む。図31に示される例の方法では、最適経路の更新され

10

20

30

40

50

た集合を識別すること（3106）は、単一の最適経路を識別すること、又は複数の最適経路を識別することを含んでよい。多様な性能閾値を満たす任意の経路が識別されてよい、所定数の最も適した経路（例えば、他の利用可能な経路に対して優れた性能を示すそれらの経路）が識別されてよい、所定の割合の最も適した経路が識別されてよい等、例えば、（ホストと特定のストレージシステムとの間の経路等の）より多くの最適経路の部分集合が識別されてよい。読者は、ストレージシステムホスト（3014、3024、3028）は、互いからなんらかの距離に位置してよいので、ストレージシステム（3014、3024、3028）は、別々のストレージネットワーク又はストレージネットワークの別々の部分に位置してよいので、又は他のなんらかの理由から、ホスト（3032）が、別のストレージシステムに対してあるストレージシステムにI/O動作を発行することに
10
関連付けられた性能優位点があってよいことを理解する。例えば、遠いデータセンタ又はキャンパスの中に物理的に位置するストレージシステムにI/O動作を発行するホスト（3032）に比して、ホスト（3032）が、ホスト（3032）と同じデータセンタ又はキャンパスの中に物理的に位置するストレージシステムにI/O動作を発行することに
20
関連付けられた性能優位点がある場合がある。信頼性のために、ホスト（3032）にすべてのストレージシステム（3032）に対する接続性を保持させることは有益であってよいが、性能のために、ホスト（3032）が特定のストレージシステムを通してデータセット（3012）にアクセスすることが好ましい場合がある。読者は、異なるホストがデータセット（3012）にアクセスする場合があるため、あるホストがデータセット（3012）にアクセスするための1つ以上の最適経路が、別のホストがデータセット（3012）にアクセスするための1つ以上の最適経路とは異なる場合があることを理解する。
30
図30に示される例の方法では、ホスト（3032）とストレージシステムの更新された集合との間の複数のデータ通信経路の中から最適経路の更新された集合を識別すること（3106）は、例えば上述されたSCSI ALUA機構の使用により、特定のインタフェース又はポッド内のストレージシステムに対するホスト経路が、他の特定のインタフェース又はポッド内のストレージシステムへのホスト経路よりも、より低い待ち時間、より良いスループット、又はより少ない切替えインフラストラクチャを有することを判断するために、タイミング情報又はネットワーク情報を使用することによって自動化された方法で、又はなんらかの他の方法で実施されてよい。読者は、上述の段落のいくつかが『集合』を参照しているが、係る集合が単一のメンバーを含む場合があり、係る集合がどのよう
40
にして表されるのかに関して特定の制限を課していないことを理解する。

【0424】

【0482】 読者は、上述された性能の優位点のさまざまな情報源があり得ることを理解する。書込みの場合、より遠方のストレージシステムへの書込みを要求するホストは、ストレージシステムからストレージシステムへの複製自体に（いずれにせよ）必要とされる長距離帯域幅に加えて、ホストから遠方のストレージシステムへのネットワークのために長距離帯域幅を必要とする（これは、それ以外の場合必要ではないホストからストレージへの帯域幅を使用する、又はそれは一方の方向だけでのトラフィックが必要であるはずであったときに両方向での書込みコンテンツのためのトラフィックを生じさせる）。さらに、書込みの場合、長距離待ち時間が多大である場合、次いでその待ち時間は4倍又は6倍生じる（遠方のストレージシステムからローカルストレージシステムへの書込みコンテンツの送達に加えられ、ローカルストレージシステムから遠方のストレージシステムへの完了又は類似した表示の送達に加えられ、2段階又は4段階の書込み要求の最後の部分のための遠方のストレージシステムからホストに送信される最終的な完了が加えられるホストから遠方のストレージシステムへの2段階又は4段階の書込みのうちの第1の段階又は3つの段階）。それに対し、ローカルストレージシステムへの書込みの場合、長距離待ち時間は2回だけつまり、ローカルストレージシステムから遠方のストレージシステムへの書込みコンテンツの送達のために1回、及び遠方のストレージシステムからローカルストレージシステムへの完了又は類似する表示の送達のために1回、生じる。読取りの場合、ローカルストレージシステムからの読取りを要求するホストは、多くの場合、長距離帯域幅は
50

まったく消費せず、通常長距離待ち時間ペナルティを生じさせない。

【0425】

[0483] また、図31に示される例の方法は、ホスト(3012)に対し、更新された最適経路の識別を示すこと(3108)も含む。図31に示される例の方法では、ストレージシステム(3014、3024、3028)は、例えばストレージシステム(3014、3024、3028)とホスト(3032)との間で交換される1つ以上のメッセージを通してホスト(3032)への更新された最適経路の識別を示してよい(3018)。係るメッセージは、上述された機構の多くを使用し、交換されてよく、ポート識別子、ネットワークインタフェース識別子、又はなんらかの他の識別子の使用により最適経路を識別してよい。読者は、いくつかの実施形態では、ストレージシステム(3014、3024、3028)が、それによりホスト(3032)に対して更新された最適経路の識別を示してよい(3108)プロセスの一部分は、係る情報をホスト(3012)によって発行されるコマンドに対する応答にピギーバックすることを含む場合がある。例えば、ストレージシステム(3014、3024、3028)の1つは、ホストに対するSCSIユニットの注意を集めてよい(3110)。SCSIユニットの注意は、デバイス(例えば、ストレージシステム)がホスト側のSCSIDライバに、デバイスの動作状態又はファブリック状態が変化したことを知らせることを可能にする仕組みである。別の言い方をすれば、ストレージシステムは、ユニットの注意を集めることによって、ホストに、ホストが状態の変化についてストレージシステムに問い合わせるべきであることを示してよく、該状態の変化により、ホストは、ターゲットポートグループ状態が、アクティブ／最適化済み及びアクティブ／非最適化のターゲットポートグループの異なる集合を示すために変化したことに気づくことができる。係る例では、ターゲット(例えば、ストレージシステム)は、次にホスト側のSCSIDライバにユニットの注意をクリアする前に更新されたALUA状態を要求するように命じるコマンドに対する応答がホスト(3012)に送信されるときに、ホスト(3012)に返される「ユニットの注意」を内部で集める。この仕組みは、ストレージシステムが、ホストにそのALUA状態を所望されるように更新させることを可能にしてよいが、それはオフラインではないなんらかのターゲットポートに発行されるなんらかの種類の将来のSCSI要求に依存する。係る例では、SCSIプロトコルがホスト(3012)によって発行されるコマンド及びターゲット(例えば、ストレージシステム)によって返される応答の形をとるため、SCSI「ユニット注意」の仕組みはターゲットがホスト(3012)に頼んでもいない更新を返す方法を提供するので、最適経路を更新するための情報を転送することは、わずかに遠回りにこの機構でピギーバックする必要がある場合がある。

【0426】

[0484] 追加の説明のために、図32は、本開示のいくつかの実施形態に従って同期複製されたストレージシステム(3014、3024、3028)への接続性を管理するための追加の例の方法を示すフローチャートを説明する。あまり詳細には示されないが、図32に示されるストレージシステム(3014、3024、3028、3210)は、図1Aから図1D、図2Aから図2G、図3Aから図3B、又はその任意の組合せに関して上述されたストレージシステムに類似してよい。実際には、図32に示されるストレージシステムは、上述されたストレージシステムと同じ構成要素、より少ない構成要素、追加の構成要素を含んでよい。

【0427】

[0485] 図32に示される例の方法は、データセット(3012)がその全体で同期複製される複数のストレージシステム(3014、3024、3028)を識別すること(3002)、データセット(3012)に向けられるI/O動作を発行できるホスト(3032)を識別すること(3004)、ホスト(3032)と、データセット(3012)がその全体で同期複製される複数のストレージシステム(3014、3024、3028)との間の複数のデータ通信経路(3022、3026、3030)を識別すること(3006)、ホスト(3032)と、データセット(3012)がその全体で同期複製さ

れる複数のストレージシステム（３０１４、３０２４、３０２８）との間の複数のデータ通信経路（３０２２、３０２６、３０３０）の中から１つ以上の最適経路を識別すること（３００８）、及びホスト（３０３２）に対して、１つ以上の最適経路の識別を示すこと（３０１０）、データセット（３０１２）が、ストレージシステムの更新された集合全体で同期複製されることを検出すること（３１０２）、ホスト（３０３２）とストレージシステム（３０１４、３０２８）の更新された集合との間の複数のデータ通信経路を識別すること（３１０４）、ホスト（３０３２）とストレージシステムの更新された集合との間の複数のデータ通信経路の中から最適経路の更新された集合を識別すること（３１０６）、ホスト（３０１２）に対し、更新された最適経路の識別を示すこと（３１０８）も含むので、図３２に示される例の方法は、図３０及び図３１に示される例の方法に類似している。

10

【０４２８】

【０４８６】 図３２に示される例の方法では、データセット（３０１２）が、ストレージシステムの更新された集合全体で同期複製されることを検出すること（３１０２）は、データセット（３０１２）がその全体で同期複製されていたストレージシステムの元の集合からストレージシステム（３０２４）がデタッチしたことを検出すること（３２０２）を含む場合がある。図３２に示される例の方法では、ストレージシステム（３０２４）は、ストレージシステム（３０２４）がもはや複数のストレージシステム全体でデータセット（３０１２）の同期複製に関与していないとき『デタッチされた』と見なされてよい。特定のストレージシステムは、例えばストレージシステムの中のハードウェア故障のために、ストレージシステムがデータ通信に従事することを妨げるネットワーク故障のために、ストレージシステムに対する電力の損失のために、ストレージシステム上でのソフトウェア衝突のために、又はさまざまな他の理由からデタッチしてよい。図３２に示される例の方法では、ストレージシステム（３０２４）が、データセット（３０１２）がその全体で同期複製されていたストレージシステムの元の集合からデタッチしたことを検出すること（３２０２）は、例えば、ストレージシステムが利用不能に又はそれ以外の場合到達不能になったと判断することによって実施されてよい。図３２に示される例の方法では、ストレージシステム（３０２４）によって使用されるすべてのデータ通信リンク（３０１６、３０１８、３０２０、３２０８）及びデータ通信経路（３０３０）が、ストレージシステム（３０２４）によって使用されるデータ通信リンク（３０１６、３０１８、３０２０、３２０８）及びデータ通信経路（３０３０）が稼働していないことを示すための点線で表されるので、ストレージシステムの１つ（３０２４）は、ストレージシステムがデータ通信に従事するのを妨げるネットワーク故障のためにデタッチされているとして示される。

20

30

【０４２９】

【０４８７】 また、図３２に示される例の方法では、データセット（３０１２）が、ストレージシステムの更新された集合全体で同期複製されることを検出すること（３１０２）は、データセット（３０１２）がその全体で同期複製されていたストレージシステムの元の集合に含まれていなかったストレージシステム（３２１０）が、データセット（３０１２）がその全体で同期複製されているストレージシステムの集合にアタッチしたことを検出すること（３２０４）も含む場合がある。図３２に示される例の方法では、ストレージシステム（３２１０）は、ストレージシステム（３２１０）が複数のストレージシステム全体でのデータセット（３０１２）の同期複製に関与しているとき『アタッチされている』と見なされてよい。特定のストレージシステムは、例えばストレージシステムがポッドに加えられるため、ストレージシステムがストレージシステムの中のハードウェア故障から回復するため、ストレージシステムがネットワーク故障から回復するため、ストレージシステムがストレージシステムに対する電力の損失から回復するため、ストレージシステムがストレージシステム上のソフトウェア衝突から回復するため、又はさまざまな他の理由からアタッチしてよい。図３２に示される例の方法では、上記図のいずれにも含まれていなかったストレージシステム（３２１０）の１つは、データセット（３０１２）がその全体で同期複製されるストレージシステムの集合にアタッチされているとして示され、スト

40

50

レンジシステム（３２１０）は、１つ以上のデータ通信リンク（３２０６）及びデータ通信経路（３２１２）を介してホスト（３０３２）及び他のストレージシステム（３０２４）とのデータ通信のために結合される。読者は、データ通信リンクがストレージシステム（３０２８、３２１０）のいくつかの間に示されていないが、係るデータ通信リンクは実際には存在する場合があるが、ここでは説明の便宜上省略されているにすぎないことを理解する。

【０４３０】

【０４８８】 また、図３２に示される例の方法は、データセット（３０１２）がその全体で同期複製されていたストレージシステムの元の集合に含まれていなかったストレージシステム（３２１０）に対するホストアクセスをモニタすること（３２０６）を含む場合もある。上述されたように、新しいストレージシステムがポッドに加えられる場合、次いで新しいターゲットポートグループがそのポッド内のボリュームごとに、そのボリュームにアクセスするホストポートに加えられ、ターゲットポートグループは、新しいストレージシステムのためのホスト／ストレージシステム近接性に適切な状態を割り当てられる場合がある。なんらかの数のＳＡＮレベルイベントの後、ホストは、ボリュームごとに新しいポートを認識し、新しい経路を適切に使用するようにそのドライバを構成できる。ストレージシステムは、ポッドの代わりに、ホストが現在新規に加えられたストレージシステムでＳＣＳＩターゲットポートを使用するように適切に構成されていると判断するためにホストアクセス（例えば、ＲＥＰＯＲＴ ＬＵＮＳコマンド及びＩＮＱＵＩＲＹコマンドを待機する等）がないかモニタできる。係る例では、ホストは、ホストがそのターゲットポートグループのメンバーにコマンドを発行する準備が完了していることに依存するポッドに対して処置を講じる前に、ホストがターゲットポートグループの新規に加えられたメンバーにコマンドを発行する準備が完了している旨の保証がないかモニタされてよい。これは例えばポッドメンバーの削除を調整するときには有用である場合がある。係る例では、削除されているストレージシステムを使用していると知られている１つ以上のホストが、最近加えられたストレージシステムを使用していることがまだ気づかれていない場合、次いで削除されているメンバーはそれらのホストの１つ以上のために機能するとして知られる最後の残っているストレージシステムである場合、次いで動作が進むことを可能にする前にアラートを発行することが有益である可能性がある（又は動作を即座に妨げることができる）。

【０４３１】

【０４８９】 追加の説明のために、図３３は、本開示のいくつかの実施形態に係る同期複製されたストレージシステム（３０１４、３０２４、３０２８）への接続性を管理するための追加の例の方法を示すフローチャートを説明する。あまり詳細には示されないが、図３３に示されるストレージシステム（３０１４、３０２４、３０２８、３２１０）は、図１Ａから図１Ｄ、図２Ａから図２Ｇ、図３Ａから図３Ｂ、又はその任意の組合せに関して上述されたストレージシステムに類似してよい。実際には、図３３に示されるストレージシステムは、上述されたストレージシステムと同じ構成要素、より少ない構成要素、追加の構成要素を含んでよい。

【０４３２】

【０４９０】 図３３に示される例の方法は、データセット（３０１２）がその全体で同期複製される複数のストレージシステム（３０１４、３０２４、３０２８）を識別すること（３００２）と、データセット（３０１２）に向けられるＩ／Ｏ動作を発行できるホスト（３０３２）を識別すること（３００４）と、ホスト（３０３２）と、データセット（３０１２）がその全体で同期複製される複数のストレージシステム（３０１４、３０２４、３０２８）との間の複数のデータ通信経路（３０２２、３０２６、３０３０）を識別すること（３００６）と、ホスト（３０３２）と、データセット（３０１２）がその全体で同期複製される複数のストレージシステム（３０１４、３０２４、３０２８）との間の複数のデータ通信経路（３０２２、３０２６、３０３０）の中から１つ以上の最適経路を識別すること（３００８）と、ホスト（３０３２）に対して、１つ以上の最適経路の識別を示す

こと（３０１０）とを含むバイもあるので図３０、図３１、及び図３２に示される例の方法に類似している。

【０４３３】

【0491】 また、図３３に示される例の方法は、ホスト（３０３２）と、データセット（３０１２）が同期複製される複数のストレージシステム（３０１４、３０２４、３０２８）との間の少なくとも１つ又は複数のデータ通信経路（３０２２、３０２６、３０３０）に対する変更を検出すること（３３０２）も含む。ホスト（３０３２）と、データセット（３０１２）が同期複製される複数のストレージシステム（３０１４、３０２４、３０２８）との間の少なくとも１つ又は複数のデータ通信経路（３０２２、３０２６、３０３０）に対する変更を検出すること（３３０２）は、例えば、特定のデータ通信経路がもはや操作可能ではないことを検出することによって、特定のデータ通信経路全体での性能（例えば、帯域幅、スループット）が所定の閾値量を超えて変化したと判断することによって、より少ない又は追加のホップが特定のデータ通信経路に導入されたと判断することによって等、実施されてよい。読者は、ホスト（３０３２）と、データセット（３０１２）がその全体で同期複製される複数のストレージシステム（３０１４、３０２４、３０２８）との間の複数のデータ通信経路（３０２２、３０２６、３０３０）に対する変更が、どの特定のデータ通信経路が最適経路として識別されるのかに影響を与える場合があり、したがってストレージシステム（３０１４、３０２４、３０２８）は、ホスト（３０３２）と、データセット（３０１２）がその全体で同期複製される複数のストレージシステム（３０１４、３０２４、３０２８）との間の複数のデータ通信経路（３０２２、３０２６、３０３０）の中から最適経路を識別するステップ（３００８）、及びホスト（３０３２）に対して、１つ以上の最適経路の識別を示すステップ（３０１０）を繰り返す必要がある場合があることを理解する。

【０４３４】

【0492】 また、図３３に示される例の方法は、ホスト（３０３２）への変更を検出すること（３３０４）も含む。ホスト（３０３２）に対する変更は、例えばホスト（３０３２）に対するソフトウェア又はハードウェアのアップグレードの結果として、ホスト（３０３２）に対する電力の損失の結果として、ホスト（３０３２）でのハードウェア又はソフトウェアの故障の結果として、ホスト（３０３２）が移動される結果として、新しいホストが、データセット（３０１２）に向けられるＩ／Ｏ動作を発行するなんらかのアプリケーションの実行をサポートとするために使用される結果として、又はさまざまな他の理由から発生する場合がある。読者は、ホスト（３０３２）に対する変更が、どの特定のデータ通信経路が最適経路として識別されるのかに影響を与える場合があり、したがってストレージシステム（３０１４、３０２４、３０２８）は、ホスト（３０３２）と、データセット（３０１２）がその全体で同期複製される複数のストレージシステム（３０１４、３０２４、３０２８）との間の複数のデータ通信経路（３０２２、３０２６、３０３０）の中から最適経路を識別するステップ（３００８）、及びホスト（３０３２）に対して、１つ以上の最適経路の識別を示すステップ（３０１０）を繰り返す必要がある場合があることを理解する。

【０４３５】

【0493】 追加の説明のために、図３４は、本開示の実施形態に従ってストレージシステム（３４２４、３４２６、３４２８）全体で同期複製されたデータへの接続性を管理する追加の例の方法を示すフローチャートを説明する。あまり詳細には示されないが、図３４に示されるストレージシステム（３４２４、３４２６、３４２８）は、図１Ａから図１Ｄ、図２Ａから図２Ｇ、図３Ａから図３Ｂ、又はその任意の組合せに関して上述されたストレージシステムに類似してよい。実際には、図３４に示されるストレージシステムは、上述されたストレージシステムと同じ構成要素、より少ない構成要素、追加の構成要素を含んでよい。さらに、図３４に示されるストレージシステム（３４２４、３４２６、３４２８）のそれぞれは、１つ以上のデータ通信リンク（３４２０、３４２２）を介して互いに接続され、１つ以上のデータ通信経路（３４１０、３４１２、３４１４）を介してホスト

(3402)に接続されてもよい。

【0436】

【0494】 図34に示される例の方法は、複数のストレージシステム(3424、3426、3428)全体で同期複製されるデータセット(3418)に向けられるI/O動作(3416)を受け取ること(3404)を含む。図34に示される例の方法では、ホスト(3402)は、複数のストレージシステム(3424、3426、3428)全体で同期複製されるデータセット(3418)に向けられるI/O動作(3416)を、例えば、ホストで実行中であるアプリケーションから、ホスト(3402)とのなんらかのユーザとの対話の結果として、又はさまざまな他の方法で受け取る(3404)。複数のストレージシステム(3424、3426、3428)全体で同期複製されるデータセット(3418)に向けられるI/O動作(3416)は、例えばデータセット(3418)ヘデータを書き込む要求として、データセット(3418)からデータを読み取る要求として、データセット(3418)のデータをコピーし、係るコピーを他のどこかに記憶する要求として、データセット(3418)のデータのスナップショットを撮影する要求として等、実施されてよい。

10

【0437】

【0495】 また、図34に示される例の方法は、I/O動作(3416)を受け取るための好ましいストレージシステムとして複数のストレージシステム(3424、3426、3428)のうちの特定のストレージシステム(3426)を識別すること(3406)も含む。図34に示される例の方法では、ホスト(3402)は、例えば、ストレージシステム(3242、3426、3428)のそれぞれにI/O動作を発行するときにホスト(3402)が以前に経験したことがある応答時間を追跡すること(又はそうでなければ、応答時間を記述する情報にアクセスできる)こと、及びI/O動作(3416)を受け取るための好ましいストレージシステムとして最速の応答時間を示したストレージシステム(3426)を選択することによってI/O動作(3416)を受け取るための好ましいストレージシステムとして複数のストレージシステム(3424、3426、3428)のうちの特定のストレージシステム(3426)を識別してよい(3406)。読者は、ホスト(3432)が、I/O動作(3416)を受け取るための好ましいストレージシステムとして特定のストレージシステム(3426)を識別する(3406)のために単独で又は組み合わせて使用され得る他の測定基準(例えば、測定基準に関する信頼性、測定基準に関する有用性、スループット測定基準)を記述する情報を追跡してよい、又は情報にアクセスできることを理解する。代わりに、ホスト(3402)は、構成パラメータとしてシステム管理者から、ストレージレイ自体から、又はなんらかの他の方法で好ましいストレージシステムの識別を受け取るように構成されてよく、これによりI/O動作(3416)を受け取るための好ましいストレージシステムとして特定のストレージシステム(3426)を識別すること(3406)は、なんらかの構成パラメータ又はホスト(3402)の中に記憶された他の構成情報を単に調べることによって実施されてよい。

20

30

【0438】

【0496】 また、図34に示される例の方法は、好ましいストレージシステム(3426)を識別した後に、データセット(3418)に向けられる1つ以上のI/O動作(3416)のために好ましいストレージシステム(3426)に1つ以上のI/O動作(3416)を発行すること(3408)も含む。図34に示される例の方法では、ホスト(3402)は、例えばホスト(3402)と好ましいストレージシステム(3426)との間のデータ通信経路(812)を介してホスト(3402)と好ましいストレージシステム(3426)との間で交換される1つ以上のメッセージを介して好ましいストレージシステム(3426)に対してデータセット(3418)に向けられる1つ以上のI/O動作(3416)を発行してよい(3408)。

40

【0439】

【0497】 追加の説明のために、図35は、本開示の実施形態に従ってストレージシステム(3424、3426、3428)全体で同期複製されたデータへの接続性を管理する

50

ための追加の例の方法を示すフローチャートを説明する。あまり詳細には示されないが、図 3 5 に示されるストレージシステム (3 4 2 4、3 4 2 6、3 4 2 8) は、図 1 A から図 1 D、図 2 A から図 2 G、図 3 A から図 3 B、又はその任意の組合せに関して上述されたストレージシステムに類似してよい。実際には、図 3 5 に示されるストレージシステムは、上述されたストレージシステムと同じ構成要素、より少ない構成要素、追加の構成要素を含んでよい。

【0 4 4 0】

[0498] 図 3 5 に示される例の方法は、複数のストレージシステム (3 4 2 4、3 4 2 6、3 4 2 8) 全体で同期複製されるデータセット (3 4 1 8) に向けられる I/O 動作 (3 4 1 6) を受け取ること (3 4 0 4) と、I/O 動作 (3 4 1 6) を受け取るための好ましいストレージシステムとして複数のストレージシステム (3 4 2 4、3 4 2 6、3 4 2 8) のうちの特定のストレージシステム (3 4 2 6) を識別すること (3 4 0 6) と、好ましいストレージシステム (3 4 2 6) を識別した後に、データセット (3 4 1 8) に向けられる 1 つ以上の I/O 動作 (3 4 1 6) のために好ましいストレージシステム (3 4 2 6) に 1 つ以上の I/O 動作 (3 4 1 6) を発行すること (3 4 0 8) も含むので、図 3 5 に示される例の方法は図 3 4 に示される例の方法に類似している。

【0 4 4 1】

[0499] また、図 3 5 に示される例の方法は、複数のストレージシステム (3 4 2 4、3 4 2 6、3 4 2 8) のうちの複数のストレージシステムのためにそれぞれの応答時間を決定すること (3 5 0 2) も含む。図 3 5 に示される例の方法では、ホスト (3 4 0 2) は、複数のストレージシステム (3 4 2 4、3 4 2 6、3 4 2 8) のうちの複数のストレージシステムのためにそれぞれの応答時間を、例えば類似する I/O 動作にサービスを提供するためにストレージシステム (3 4 2 4、3 4 2 6、3 4 2 8) のそれぞれによって必要とされる時間量を決定することによって、類似する I/O 動作にサービスを提供するために、ストレージシステム (3 4 2 4、3 4 2 6、3 4 2 8) のそれぞれによって必要とされる平均時間量を追跡すること等によって決定する (3 5 0 2)。係る例では、ホスト (3 4 0 2) は、1 つ以上の内部クロックの使用により、1 つ以上のメッセージに付けられるタイムスタンプを調べることによって、又はなんらかの他の方法で係る情報を追跡してよい。

【0 4 4 2】

[0500] 図 3 5 に示される例の方法では、I/O 動作 (3 4 1 6) を受け取るための好ましいストレージシステムとして複数のストレージシステム (3 4 2 4、3 4 2 6、3 4 2 8) のうちの特定のストレージシステム (3 4 2 6) を識別すること (3 4 0 6) は、複数のストレージシステム (3 4 2 4、3 4 2 6、3 4 2 8) のうちの複数のストレージシステムのためのそれぞれの応答時間に従って I/O 動作 (3 4 1 6) を受け取るための好ましいストレージシステムとして複数のストレージシステム (3 4 2 4、3 4 2 6、3 4 2 8) のうちの特定のストレージシステムを識別すること (3 5 0 4) を含む場合がある。図 3 5 に示される例の方法では、ホスト (3 4 0 2) は、例えば、好ましいストレージシステムとして最速の応答時間と関連付けられたストレージシステムを選択することによって、その応答時間が好ましいストレージシステムとしてのサービス閾値の所定の質を満たす任意のストレージシステムを選択することによって、又は他のなんらかの方法で、複数のストレージシステム (3 4 2 4、3 4 2 6、3 4 2 8) のうちの複数のストレージシステムのためのそれぞれの応答時間に従って、I/O 動作 (3 4 1 6) を受け取るための好ましいストレージシステムとして複数のストレージシステム (3 4 2 4、3 4 2 6、3 4 2 8) のうちの特定のストレージシステムを識別してよい (3 5 0 4)。

【0 4 4 3】

[0501] また、図 3 5 に示される例の方法は、ストレージシステム (3 4 2 4、3 4 2 6、3 4 2 8) の少なくとも 1 つのための応答時間の変化を検出すること (3 5 0 6) も含む。図 3 5 に示される例の方法では、ホスト (3 4 0 2) は、ストレージシステムのそれぞれに対して追加試験を実行した結果として、平均応答時間が所定の閾値量を越えて逸

脱したと判断することによって、特定のデータ通信リンクを介してメッセージを交換する能力に対するなんらかの中断を検出することによって、又は他のなんらかの方法で、ストレージシステム（３４２４、３４２６、３４２８）の少なくとも１つのための応答時間の変化を検出する（３５０６）。

【０４４４】

【０５０２】 また、図３５に示される例の方法は、好ましいストレージシステムとして複数のストレージシステム（３４２４、３４２６、３４２８）の異なるストレージシステムを、応答時間の変化に従って選択すること（３５０８）も含む。図３５に示される例の方法では、ホスト（３４０２）は、Ｉ／Ｏ動作（３４１６）を受け取るための好ましいストレージシステムとして複数のストレージシステム（３４２４、３４２６、３４２８）の異なるストレージシステムを、例えば選好ましいストレージシステムとして最速の更新された応答時間と関連付けられたストレージシステムを選択することによって、その更新された応答時間が、好ましいストレージシステムとしてサービス閾値の所定の質を満たす任意のストレージシステムを選択することによって、又は他のなんらかの方法で選択する（３５０８）。

【０４４５】

【０５０３】 追加の説明のために、図３６は、本開示の実施形態に従って、ストレージシステム（３４２４、３４２６、３４２８）全体で同期複製されたデータへの接続性を管理する追加の例の方法を示すフローチャートを説明する。あまり詳細には示されないが、図３６に示されるストレージシステム（３４２４、３４２６、３４２８）は、図１Ａから図１Ｄ、図２Ａから図２Ｇ、図３Ａから図３Ｂ、又はその任意の組合せに関して上述されたストレージシステムに類似してよい。実際には、図３６に示されるストレージシステムは、上述されたストレージシステムと同じ構成要素、より少ない構成要素、追加の構成要素を含んでよい。

【０４４６】

【０５０４】 図３６に示される例の方法は、複数のストレージシステム（３４２４、３４２６、３４２８）全体で同期複製されるデータセット（３４１８）に向けられるＩ／Ｏ動作（３４１６）を受け取ること（３４０４）と、Ｉ／Ｏ動作（３４１６）を受け取るための好ましいストレージシステムとして複数のストレージシステム（３４２４、３４２６、３４２８）のうちの特定のストレージシステム（３４２６）を識別すること（３４０６）と、好ましいストレージシステム（３４２６）を識別した後に、データセット（３４１８）に向けられる１つ以上のＩ／Ｏ動作（３４１６）のために好ましいストレージシステム（３４２６）に１つ以上のＩ／Ｏ動作（３４１６）を発行すること（３４０８）も含むので、図３６に示される例の方法は図３４に示される例の方法に類似している。

【０４４７】

【０５０５】 また、図３６に示される例の方法は、ストレージシステム（３４２８）の１つから、好ましいストレージシステムの識別（３６１０）を受け取ること（３６０２）も含む。図３６に示される例の方法では、ホスト（３４０２）は、ストレージシステム（３４２８）の１つから、ストレージシステム（３４２８）とホスト（３４０２）との間のデータ通信経路（３４１４）を介して交換される１つ以上のメッセージを介して好ましいストレージシステムの識別（３６１０）を受け取ってよい（３６０２）。ホスト（３４０２）は、例えばホストの中に記憶される構成設定値として好ましいストレージシステムの識別（３６１０）を保持してよい。したがって、他の実施形態では、構成設定値は異なる方法で（例えば、システム管理者によって、ホストで実行中の別のソフトウェアモジュールによって設定されてよいが、Ｉ／Ｏ動作（３４１６）を受け取るための好ましいストレージシステムとして複数のストレージシステム（３４２４、３４２６、３４２８）のうちの特定のストレージシステム（３４２６）を識別すること（３４０６）は、構成設定値に従って実施されてよい。

【０４４８】

【０５０６】 また、図３６に示される例の方法は、ホスト（３４０２）がある場所から別の

場所に移動したことを検出すること（３６０４）も含む。図３６に示される例の方法では、ホスト（３４０２）は、例えばホスト（３４０２）が新しいデータ通信相互接続に接続されたと判断することによって、ホスト（３４０２）がラックの中の異なる位置の中に取り付けられている、又は新しいラックに取り付けられていることを検出することによって、又は他のなんらかの方法で、それが場所から別の場所に移動したことを検出してよい（３６０４）。ラックに対する又はデータセンタの中の、又はデータセンタ若しくはキャンパスのネットワークトポロジーに基づいた場所は、ホストとポッドのための特定のストレージシステムとの間の性能に影響を及ぼす場合がある「局所性」の態様である場合がある。ラックに及ぶ又は複数のネットワークに接続される単一のストレージシステムでは、局所性はポッドの中の個々のストレージシステム上の個々のストレージネットワークアダプタにも適用する可能性がある。

10

【０４４９】

【０５０７】 また、図３６に示される例の方法は、データセット（３４１８）に向けられるＩ／Ｏ動作を受け取るための好ましいストレージシステムとして複数のストレージシステムの異なるストレージシステムを識別すること（３６０８）も含む。図３６に示される例の方法では、ホスト（３４０２）は、例えばストレージシステムＮそれぞれと関連付けられた応答時間を再測定し、最速の応答時間を示すストレージシステムを選択することによって、データセット（３４１８）に向けられるＩ／Ｏ動作を受け取るための好ましいストレージシステムとして複数のストレージシステムの異なるストレージシステムを識別してよい（３６０８）。図３６に示される例の方法では、データセット（３４１８）に向けられるＩ／Ｏ動作を受け取るための好ましいストレージシステムとして複数のストレージシステムの異なるストレージシステムを識別すること（３６０８）は、ホスト（３４０２）が移動したことを検出することに応じて実施されてよい。

20

【０４５０】

【０５０８】 また、図３６に示される例の方法は、ホスト（３４０２）に対する構成変更を検出すること（３６０６）も含む。図３６に示される例の方法では、ホスト（３４０２）は、例えば、なんらかのソフトウェアの異なるバージョンがホスト（３４０２）にインストールされていることを検出することによって、ホスト（３４０２）の中のなんらかのハードウェア構成要素が変更された、又は加えられたことを検出すること等によってホスト（３４０２）に対する構成変更を検出してよい（３６０６）。図３６に示される例の方法では、代わりに、データセット（３４１８）に向けられるＩ／Ｏ動作を受け取るための好ましいストレージシステムとして複数のストレージシステムの異なるストレージシステムを識別すること（３６０８）は、ホスト（３４０２）に対する構成変更を検出することに応じて実施されてよい。

30

【０４５１】

【０５０９】 追加の説明のために、図３７は、本開示のいくつかの実施形態に係る仲介サービスのための自動ストレージシステム構成のための例の方法を示すフローチャートを説明する。あまり詳細には示されないが、図３７に示されるストレージシステム（３７００Ａ～３７００Ｎ）は、図１Ａから図１Ｄ、図２Ａから図２Ｇ、図３Ａから図３Ｂ、又はその任意の組合せに関して上述されたストレージシステムに類似してよい。実際には、図３７に示されるストレージシステム（３７００Ａ～３７００Ｎ）は、上述されたストレージシステムと同じ構成要素、より少ない構成要素、追加の構成要素を含んでよい。

40

【０４５２】

【０５１０】 以下の例では、仲介サービスのための自動ストレージシステム構成は、ストレージシステム（３７００Ａ～３７００Ｎ）の集合の中の所与のストレージシステム（３７００Ａ）が、仲介ターゲット又はサービスから仲介を要求するように構成されるかどうかを判断することを含んでよい。所与のストレージシステム（３７００Ａ）が仲介を要求するように構成されていない場合、所与のストレージシステム（３７００Ａ）は、操作可能になる前に、事前に構成された場所から仲介サービスハンドルを要求又は入手するように構成される。ストレージシステムが仲介を要求するように構成されるかどうかの係る判断

50

は、ストレージシステムが最初にオンラインにされる又は起動されるときに起こる場合があり、ストレージシステムは、出荷前に、例えば、構成サービス（３７５１）等の指定された構成サービスから構成情報を要求するように構成されており、構成サービスは独立したコンピュータシステム又は構成サービスを提供するサードパーティコンピューティング環境の中で動作する場合がある。

【０４５３】

【０５１１】 さらに、一部の例では、ストレージシステム（３７００Ａ～３７００Ｎ）の中の所与のストレージシステム（３７００Ａ）が、仲介サービスから仲介を要求するように構成されるかどうかを判断することは、ストレージシステム（３７００Ａ～３７００Ｎ）の中でデータセット同期複製を開始することに応じて実行される。この例では、ストレージシステムがデータセット又はポッドを同期複製するストレージシステムの集合に加えられるのに応じて、現在ポッドのメンバーであるストレージシステムは、仲介サービス及び加えられているストレージシステムに対する仲介競争ターゲットのための又は関係する１つ以上のハンドルを自動的に転送してよい。このようにして、第１のストレージシステムは、構成サービスから仲介ターゲットを受け取り、ポッドに加えられた各ストレージシステムはポッドメンバーからその仲介ターゲットを受け取り、結果的にすべてのメンバーストレージシステムは同じ仲介ターゲットから仲介を要求するように構成される。

【０４５４】

【０５１２】 上述されたように、仲介サービスハンドルは、ポッドが作成されるとき、ポッドが広げられるとき、又はポッド内のストレージシステム間の将来の通信障害の場合には仲介サービスが必要とされ得るように、ポッドが最初に広げられるときに要求されてよい。仲介サービスハンドルは広域ネットワーク上の連絡アドレスであってよく、ストレージシステムクラスタ又はポッドに対する仲介必要性を処理するための鍵のプールを管理するために使用できるトークンを暗号で守ってよい。代わりに、ポッド内の第１のストレージシステムは、第１の又は後続の仲介競争で使用するための周知の仲介サービスと使用するための安全なハンドルを決定してよく、ハンドルは、仲介サービスとの特定の又は必要な対話なく、その第１のストレージシステムによってプライベートに決定され、ハンドルは、次いですでにポッド内にある又はポッドとして他のストレージシステムに通信される。

【０４５５】

【０５１３】 例えば、ストレージシステム間の通信障害等のエラーに応じて仲介サービスを従事させるプロセスストレージシステムは、広域ネットワークを介する連絡アドレスとして示すハンドル、及び仲介のための鍵のプールを管理するために使用できる暗号で安全なトークンを記憶するように構成されてよい—その全体として本願に援用される出願参照番号第１５／７０３，５５９号の中でより詳細に説明される。やはり出願参照番号第１５／７０３，５５９号の中で説明されるのは、データセットを複製するストレージシステムの集合の中のどのストレージシステムが、データセットに向けられたＩ／Ｏ要求にサービスを提供し続けるべきかを判断するための多様な定足数プロトコルの使用である。

【０４５６】

【０５１４】 しかしながら、本願参照番号第１５／７０３，５５９号は、仲介及び定足数プロトコルのための実施態様を説明しているが、本開示の焦点は、仲介サービスのための自動ストレージシステム構成である。言い換えると、ストレージシステムは、構成サービスに連絡して、ストレージシステムがそれ自体を、仲介を通して通信障害に対応するように構成するように、仲介サービスに対するハンドルを要求するために出荷前にセットアップされてよく、ストレージシステムは、より多くのストレージシステムがポッドに加えられるときに処理される仲介を自動的に転送する—管理者等のユーザは、ストレージシステムが仲介を実行するように構成されるためになんの処置を講じる必要もない。

【０４５７】

【０５１５】 図３７に示されるように、データセット（３７５２）を同期複製している複数のストレージシステム（３７００Ａ～３７００Ｎ）は、１つ以上のネットワーク（不図示）を介してそれぞれの他のストレージシステムと、及び仲介サービス（３７０１）と通信

10

20

30

40

50

してよい。仲介サービス（３７０１）は、どのストレージシステムが、ストレージシステム間の通信障害の場合に、ストレージシステムがオフラインになった場合に、又はなんらかの他のトリガイメントに起因してデータセットにサービスを提供し続けるのかを決定してよい。さらに、この例では、ストレージシステムがオフラインになることに応えて、構成サービス（３７５１）は、事前に設定された連絡アドレスの使用により、仲介サービス（３７０１）に対する１つ以上のハンドルを要求するためのストレージシステムによって自動的に到達可能であってよい。一般的に、任意の数のストレージシステムは、データセット（３７５２）を同期複製している同期中リストの一部であってよい。

【０４５８】

【０５１６】 図３７に示される例の方法は、ストレージシステム（３７００Ａ～３７００Ｎ）の中の特定のストレージシステム（３７００Ａ）が、データセット（３７５２）を同期複製するストレージシステム間の仲介のために仲介ターゲットから仲介を要求するように構成されていないと判断すること（３７０２）を含む。ストレージシステム（３７００Ａ～３７００Ｎ）の中の特定のストレージシステム（３７００Ａ）が仲介ターゲットから仲介を要求するように構成されていないと判断すること（３７０２）は、コントローラの中の起動プロセスを含んだストレージシステム（３７００Ａ）によって実装されてよく、起動プロセスは、通信障害、又は通信障害に関係する中断がどのように処理されるのかを構成してよい。例えば、仲介ハンドラ（３７６２）が、例えば仲介ハンドルが決定されたかどうか、受け取れたかどうか、それとも要求されたかどうかを示すステータスフラグ又は条件コードを読み取ることによってチェックしてよい、コントローラ４６２）の起動プロセス。仲介ハンドラ（３７６２）が、仲介ハンドルがすでに構成されたことを検出する場合、次いで起動プロセスのこの部分は完了する。しかしながら、仲介ハンドラ（３７６２）が、仲介ハンドルが構成されていないと検出する場合、仲介ハンドラ（３７６２）は、係るハンドルを決定する又は構成サービス（３７５１）から仲介ハンドルを要求するために進んでよい。仲介サービス（３７０１）のための連絡アドレスは、ストレージシステムが出荷される前に製造メーカによって定義されたシステム設定値又はシステム変数であってよい。

【０４５９】

【０５１７】 また、図３７に示される例の方法は、構成サービス（３７５１）から特定のストレージシステム（３７００Ａ）によって、仲介サービス（３７０１）のための１つ以上のサービスハンドルを示す構成情報を要求すること（３７０４）も含む。特定のストレージシステム（３７００Ａ）によって構成サービス（３７５１）から仲介サービス（３７０１）のための１つ以上のサービスハンドルを示す構成情報を要求すること（３７０４）は、仲介ハンドラ（３７６２）が構成サービス（３７５１）のための記憶されている連絡情報にアクセスすること、及び仲介サービス（３７０１）のための１つ以上のサービスハンドルに対する要求（３７５４）を送信することによって実施されてよい。

【０４６０】

【０５１８】 また、図３７に示される例の方法は、構成サービス（３７５１）から受け取られた１つ以上のサービスハンドルに従って、ストレージシステム（３７００Ｂ～３７００Ｎ）の１つとの通信障害を検出することに応えて、仲介サービスと通信するように仲介ハンドラを構成すること（３７０６）も含む。構成サービス（３７５１）から受け取られた１つ以上のサービスハンドルに従って、ストレージシステム（３７００Ｂ～３７００Ｎ）の１つとの通信障害を検出することに応えて、仲介サービスと通信するように仲介ハンドラを構成すること（３７０６）は、仲介ハンドラ（３７６２）が、通信障害に対応するときに使用された仲介サービス（３７０１）のためにサービスハンドル又は連絡アドレスを定義することによって実施されてよく、サービスハンドル（３７５６）は、要求（３７５４）に応える構成サービス（３７５１）からの応答メッセージの中で指定されてよい。

【０４６１】

【０５１９】 追加の説明のために、図３８は、本開示のいくつかの実施形態に従って仲介サービスのための自動ストレージシステム構成のための例の方法を示すフローチャートを説

10

20

30

40

50

明する。図38に示される例の方法は、ストレージシステム(3700A~3700N)の中の特定のストレージシステム(3700A)が、データセット(3752)を同期複製するストレージシステム間の仲介のために仲介ターゲットから仲介を要求するように構成されていないと判断すること(3702)と、構成サービス(3751)から特定のストレージシステム(3700A)によって、仲介サービス(3701)のための1つ以上のサービスハンドルを示す構成情報を要求すること(3704)と、構成サービス(3751)から受け取られた1つ以上のサービスハンドルに従って、ストレージシステム(3700B~3700N)の1つとの通信障害を検出することに応じて、仲介サービスと通信するように仲介ハンドラを構成すること(3706)も含むので、図38に示される例の方法は図37に示される例の方法に類似している。

10

しかしながら、図38に示される例の方法は、ストレージシステムを、データセット(3752)を同期複製するストレージシステム(3700A~3700N)に加えることに応えて、加えられているストレージシステムに仲介サービスのための1つ以上のハンドルを自動的に転送すること(3802)をさらに含む。加えられたストレージシステムが同期中リストのメンバーになり得るようにストレージシステムをポッドに加えるプロセスが本明細書に説明される。ストレージシステムをポッドに加えるための説明されたプロセスを所与として、仲介サービスのための1つ以上のハンドルを、加えられているストレージシステムに転送すること(3802)は、加えられているシステムに対して、仲介ハンドラによって使用されるサービスハンドルが、コマンドの中で送信されているサービスハンドルであると定義されることを示すコマンドを送信するための説明されたプロセスをさらに指定することによって実施されてよい。

20

【0462】

【0520】 追加の説明のために、図39は、本開示のいくつかの実施形態に従って仲介サービスの自動ストレージシステム構成のための例の方法を示すフローチャートを説明する。図39に示される例の方法は、ストレージシステム(3700A~3700N)の中の特定のストレージシステム(3700A)が、データセット(3752)を同期複製するストレージシステム間の仲介のために仲介ターゲットから仲介を要求するように構成されていないと判断すること(3702)と、構成サービス(3751)から特定のストレージシステム(3700A)によって、仲介サービス(3701)のための1つ以上のサービスハンドルを示す構成情報を要求すること(3704)と、構成サービス(3751)から受け取られた1つ以上のサービスハンドルに従って、ストレージシステム(3700B~3700N)の1つとの通信障害を検出することに応じて、仲介サービスと通信するように仲介ハンドラを構成すること(3706)も含むので、図39に示される例の方法は図37に示される例の方法に類似している。

30

【0463】

【0521】 しかしながら、図39に示される例の方法は、仲介サービス(3701)に対して特定のストレージシステム(3700A)によって、ストレージシステム(3700A~3700N)の間の仲介のための安全な鍵又は無作為抽出された鍵(3952)を提供すること(3902)をさらに含む。仲介サービス(3701)に対して特定のストレージシステム(3700A)によって、ストレージシステム(3700A~3700N)の間の仲介のための安全な鍵又は無作為抽出された鍵(3952)を提供すること(3902)は、その全体として本明細書に援用される出願参照第15/703,559号の中に説明させるように実施されてよい。

40

【0464】

【0522】 追加の説明のために、図40は、本開示のいくつかの実施形態に係る仲介サービスのための自動ストレージシステム構成のための例の方法を示すフローチャートを説明する。あまり詳細には示されないが、図40に示されるストレージシステム(3700A~3700N)は、図1Aから図1D、図2Aから図2G、図3Aから図3B、又はその任意の組合せに関して上述されたストレージシステムに類似してよい。実際には、図40に示されるストレージシステム(3700A~3700N)は、上述されたストレージシ

50

システムと同じ構成要素、より少ない構成要素、追加の構成要素を含んでよい。

【0465】

【0523】 自動ストレージシステムのこの例の実施態様では、出荷又はインストールされる前に、ストレージシステムが構成サービス等の固定された事前に構成された場所から仲介サービスバンドルを要求又は入手するように事前に構成される代わりに一ストレージシステムは、仲介サービスのための固定された連絡情報で事前に構成されてよい。データセットのためにポッドを作成する際に、所与のストレージシステムは、鍵のインスタンスを生成し、鍵を仲介サービスに提供することによって、ポッド内のストレージシステム間で仲介するための仲介サービスの使用を構成してよい。さらに、データセットのためにポッドを作成する際に、所与のストレージシステムは一例えば、ストレージシステムがポッドに加えられると一ポッド内のストレージシステムのそれぞれに対する仲介サービスに提供できる同じ生成された鍵を提供することによって、ポッド内の他のストレージシステムを構成してもよい。生成された鍵を受け取る際に、他のストレージシステムは、仲介がシステム障害に対する適切な対応であると判断する場合、仲介を要求するために生成された鍵を使用してよい。

10

【0466】

【0524】 この例では、仲介サービスは、特定の鍵の使用をロックする単純な要求に応答してよく、仲介サービスは、特定の鍵を受け取る前に特定の鍵で構成されず、仲介サービスは任意の特定のストレージシステム、ポッド、又はカスタマから特定の鍵を受け取るように構成されていない。特定の鍵が所与のストレージシステムによって仲介サービスでロックされている場合、次いで他のストレージシステムは仲介サービスでその特定の鍵をロックすることはできない。

20

【0467】

【0525】 概して、鍵は、アタッカーが鍵の値を予測し、それを事前にロックする可能性があるエクスプロイトコード、つまりハッキングの試みを防ぐために暗号で安全な方式を使用し、生成されてよい、又は名前を付けられてよい。さらに、クラスタ又はポッド内のストレージシステムの集合は、クラスタ又はポッドに対する次の競争を決定する際に使用するための特定の鍵を使用し、各仲介試行後に使用される新規に生成される鍵を交換してよい。別の例では、ストレージシステムがクラスタ又はポッドに加えられた後、ポッド構成の態様の1つは、加えられているストレージシステムに、次の仲介要求に使用するのための現在の鍵を提供することを含んでよい。例えば、ストレージデバイス故障又はネットワーク障害等のなんらかのイベントが仲介をトリガする場合、次いで1つ以上のストレージシステムは、仲介サービスに対する現在の鍵のロックを要求してよい。現在の鍵に対するロックに対するこの要求は、せいぜい1人の要求者に対して成功する場合がある一要求に対する応答が失われる場合、ロックは再び要求される（又は実施態様によっては、問い合わせされる）場合があり、後続の要求は、後続の要求が同じ要求者からくるとして仲介サービスによって識別される場合に成功する場合がある。

30

【0468】

【0526】 一部の例では、特定の仲介鍵の単回使用後、クラスタ又はポッド内の任意の残りのアクティブなストレージシステムは、将来の仲介試行に使用するのために新しい鍵を交換してよい。また、新しい鍵のこの交換は一障害時にメンバーであったストレージシステムが互いの通信を再開するのであるならば一以前の仲介が仲介サービスから成功した回答を受け取ることにより絶対に確認されなかった場合に実行される場合がある。

40

【0469】

【0527】 代替実施態様では、カスタマは、カスタマによって管理されるバーチャルマシンで動作するサービス等の仲介サービスを明示的に構成してよい。さらに別の実施態様では、クラスタ又はポッドを構成する際に中間ステップがある場合があり、それによりストレージシステムは、サービス場所アドレス又はクラスタ若しくはポッドが仲介のために使用する他の連絡情報を決定するために、1つのアドレスで仲介構成サービスに連絡する。本実施態様では、この中間ステップは、クラスタ又はポッドのためのコア構成プロセスの

50

一部であってよく、サービス場所アドレス又は他の連絡情報は、クラスタ又はポッドの中のストレージシステム間で交換され、サービス場所アドレス又は他の連絡情報は、新しいストレージシステムがクラスタ又はポッドに加えられると、新しいストレージシステムに送信される。係る実施態様は、ベンダが特定のカスタマのために特定の仲介サービスインスタンスを調整する状況で有用である場合がある。代わりに、係る実施態様は、地形に基づいた仲介サービスの場所を明示的に突き止めるために、又はカスタマの場所をクラウドサービス可用性又は確実性ゾーンに一致させるために使用されてよい。

【0470】

【0528】 一部の例では、仲介サービスに関して、仲介サービスは、カスタマの対話なしでストレージシステムによって連絡されるように構成されてよく、仲介サービスは、クラスタ又はポッドの中の仲介の特定のインスタンスに対してロックし、同ロックに対して矛盾する要求がなされなかった場合にだけ成功して応答することによって動作し、成功は、1つのストレージシステムが、少なくとも1つの他方のストレージシステムから一方のストレージシステムを隔離した障害を検出後にクラスタ又はポッドのために安全にサービスを再開できることを確実にすることの一部としてクラスタ又はポッド内の1つのストレージシステムによって使用される。仲介サービスは、1つ以上の接続するクライアントから要求を受け取る複数のフロントエンドウェブサーバを提供するサービス等のクラウドベースのサービスであってよく、フロントエンドウェブサーバは、なんらかの数のIPアドレスにマッピングする一複数のネットワークスイッチの後方でさらに視覚化されてよい。特定のDNSホスト名に対して宣伝してよい。一部の例では、複数のホスト名が提供されてよく、フロントエンドサーバは、複数のバックエンドサーバ全体で仲介に対する要求をばらまく、つまり分散させるように構成されてよい。例えば、仲介サービスの場合、ストレージシステムからの特定のロック要求と関連付けられた特定の仲介鍵を所与として、特定の鍵は、次いで受け取られた鍵ごとにロックを実施する複数のバックエンドデータベースのいずれかに対してハッシュ値を計算されて(hashed)よい。さらに、複数のストレージシステムは、競争して同じポッドに対して仲介するために同じ鍵を使用し得るため、ストレージシステムは、フロントエンドウェブサーバそれらが同じ全体的なクラウドベースの仲介サービスと関連付けられる限り一のいずれかに連絡してよく、ロックを実施するために同じバックエンドデータベースに対してハッシュ値を計算してよい。この例では、バックエンドデータベースは、適切な保証を有する分散型トランザクションデータベース(例えば、とりわけDynamoDB(商標))として実装される場合もあれば、バックエンドデータベースは共用ストレージ上の高可用性データベースサーバとして実装される場合もあれば、バックエンドデータベースは高可用性及びデータ冗長性のための適切な機構を有する同期複製されたデータベースサーバとして実装される場合もあれば、バックエンドデータベースは適切な保証された条件付きストアプリミティブを有するオブジェクトストレージモデルを使用し、実装される場合もあれば、バックエンドデータベースはさまざまな他の技術のいずれにより実装されてもよい。

【0471】

【0529】 さらに、一部の例では、仲介サービスは、各鍵一各クラスタの鍵、ポッドの鍵、又はカスタマの鍵、又は他のタイプのドメインに関係する鍵が、実施態様に応じて他の鍵、クラスタ、ポッド、カスタマ、又は他のタイプのドメインから安全に隔離されることを確実にするために実装されてよい。これは、例えば、仲介サービスのための効果的なマルチテナンシーモデルを保証できる。

【0472】

【0530】 図40に示される例の方法は、ストレージシステム(3700A~3700N)のうちの特定のストレージシステム(3700A)によって、データセット(3752)を同期複製するストレージシステム(3700A~3700N)の間の仲介のために仲介サービス(3701)から仲介を要求するようにストレージシステムの1つ以上を構成することを決定すること(4002)を含む。ストレージシステム(3700A~3700N)のうちの特定のストレージシステム(3700A)によって、データセット(37

52)を同期複製するストレージシステム(3700A~3700N)の間の仲介のために仲介サービス(3701)から仲介を要求するようにストレージシステムの1つ以上を構成することを決定すること(4002)は、異なる技術を使用し、実施されてよい。一例では、ストレージシステムを構成することを決定すること(4002)は、1つ以上の別のストレージシステムを含むためのクラスタ又はポッドの拡大に基づいて実施されてよく、新しいストレージシステムがクラスタ又はポッドに加えられる場合、次いでそれぞれの新しいストレージシステムは、それぞれの他の既存のストレージシステムが仲介のために使用するように構成される仲介のための鍵を与えられる。別の例では、ストレージシステムを構成することを決定すること(4002)は、データセット(3752)を同期複製するポッドのメンバーであるストレージシステムごとに、所与のストレージシステムが鍵のインスタンスを与られているかどうかを記述するメタデータを維持することによって実施されてよく、メタデータの初期状態は、ストレージシステムが鍵を有していないことを示す場合がある。例えば、鍵に対する最新の意見の一致がない場合、特定のストレージシステム(3700A)は、コンピュータ的に予測が実現可能ではないであろう鍵を生成する暗号技術を使用し、上述されたように鍵のインスタンスを生成してよい。さらに、特定のストレージシステム(3700A)は、起動中又は周期的に、鍵が生成されていない一つまり、鍵値に対する最新の意見の一致がないと判断する場合があります、それに応じて、鍵のインスタンスを生成する場合があります。さらに、一部の例では、仲介サービスは、互いに同期中であるストレージシステムの中の同期されたデータセット(3752)の、メタデータを含んだ複製されたコンテンツの一部である場合がある。次いで、定足数方針が、どのストレージシステムが互いに同期中のままとなるのかを判断するためのアルゴリズムの全体的な集合の一部として、ストレージシステムの間で使用されてよい。このようにして、ストレージシステム(3700A~3700N)は、仲介サービスが使用される場合、鍵の同じインスタンスを使用してよい。

【0473】

【0531】 また、図40に示される例の方法は、ストレージシステム(3700A~3700N)の1つ以上に仲介サービス(3701)から仲介を要求するために鍵(4052)のインスタンスを提供すること(4004)も含む。ストレージシステム(3700A~3700N)の1つ以上に仲介サービス(3701)から仲介を要求するために鍵(4052)のインスタンスを提供すること(4004)は、特定のストレージシステム(3700A)から、仲介サービスから仲介を要求するように構成されていないと判断された2つ以上のストレージシステムのそれぞれに対し、1つ以上のネットワークポートを使用し、及び1つ以上の通信ネットワーク全体で鍵のインスタンスを送信することによって実施されてよい。一部の例では、1つ以上のストレージシステム(3700B~3700N)に鍵のインスタンスを提供すること(4004)は、ストレージシステム(3700A~3700N)のうちの特定のストレージシステム(3700A)によって、ストレージシステムの1つ以上が、データセット(3752)を同期複製するストレージシステム(3700A~3700N)の間の仲介のために仲介サービス(3701)から仲介を要求するように構成されていないと判断すること(4002)に応じて、実行されてよい。

【0474】

【0532】 一部の実施態様では、正常な状況なら、ポッドは仲介鍵で開始してよく、ポッドが追加のストレージシステムを含むために広げられると、仲介鍵を送信するであろう一部の例では「リーダー」と見なされる場合がある特定のストレージシステムは、ある鍵が一度使用されたことに応じて、後続の仲介競争のために要求される新しい仲介鍵を送信するであろう。言い換えると、上記の実施態様と対照的に、一部の実施態様では、ストレージシステムが仲介鍵を使用するように構成されているかどうかを明示的に判断する代わりに、ストレージシステムは、ポッドの作成に基づいて、ポッドの作成の部分として仲介鍵にアクセスできる場合があるストレージシステムがポッドに加えられることに応じて、仲介鍵は、ポッド内のストレージシステム全体でのデータ及びメタデータの同期の一部として新しいストレージシステムに配布される。さらに、上述されたように、本実施態

様では、仲介鍵を使用すること又は仲介鍵の使用を試みることに応えて、新しい鍵が生成され、ポッド内のストレージシステムに配布されてよい。このようにして、ポッド内のストレージシステムは、ポッドの作成に応じて生成される同じ仲介鍵を使用し、追加のストレージシステムを含むようにポッドを広げ、仲介鍵を使用する又は使用しようと試みるように構成されてよい。上述されたように、仲介鍵を使用するように構成することは、仲介ハンドラによって使用可能な現在の鍵をポッドのストレージシステム全体で同期されている仲介鍵であると指定することを含んでよい。

【0475】

【0533】 また、図40に示される例の方法は、仲介サービス(3701)に対して、鍵(4052)のインスタンスを提供すること(4006)を含み、仲介サービスが、鍵(4052)のインスタンスを提供する所与のストレージシステムに仲介サービスを提供することを含む。仲介サービスが、鍵(4052)のインスタンスを提供する所与のストレージシステムに仲介サービスを提供する、仲介サービス(3701)に対して鍵(4052)のインスタンスを提供すること(4006)は、1つ以上のネットワークポートを使用し、及び1つ以上の通信ネットワーク全体で、特定のストレージシステム(3700A)から仲介サービス(3701)に鍵のインスタンスを送信することによって実施されてよい。一部の場合では、仲介サービスが、鍵(4052)のインスタンスを提供する所与のストレージシステムに仲介サービスを提供する、仲介サービス(3701)に対して鍵(4052)のインスタンスを提供すること(4006)は、ストレージシステム(3700A~3700N)のうちの特定のストレージシステム(3700A)によって、ストレージシステムのうちの1つ以上が、データセット(3752)を同期複製するストレージシステム(3700A~3700N)の間の仲介のために仲介サービス(3701)から仲介を要求するように構成されていないと判断すること(4002)に応じて実行されてよい。

【0476】

【0534】 また、図40に示される例の方法は、ストレージシステム(3700B~3700N)の少なくとも1つとの通信障害を検出することに応じて仲介サービスに仲介サービスの鍵のインスタンスを提供するように仲介ハンドラ(3762)を構成すること(708)も含む。ストレージシステム(3700B~3700N)の少なくとも1つの通信障害を検出することに応じて仲介サービスに仲介サービスの鍵のインスタンスを提供するように仲介ハンドラ(3762)を構成すること(4006)は、仲介ハンドラ(3762)が、通信障害に対応するとき、仲介サービス(3701)に提供される生成された鍵のインスタンスとなるように現在の鍵を定義することによって実施されてよい。

【0477】

【0535】 追加の説明のために、図41は、本開示のいくつかの実施形態に従って、ともにストレージデータの論理ボリューム又は論理ボリュームの一部分を表してよい、メタデータオブジェクトの構造化された集合体として実装されてよいメタデータ表現の図を説明する。メタデータ表現4150、4154、及び4160は、ストレージシステム(4106)の中に記憶されてよく、1つ以上のメタデータ表現は、例えばストレージシステム(4106)の中に記憶されたボリューム又はボリュームの部分等の複数のストレージオブジェクトのそれぞれについて生成及び維持されてよい。

【0478】

【0536】 他のタイプのメタデータオブジェクトの構造化された集合体が考えられるが、この例では、メタデータ表現は、ノードの有向非巡回グラフ(DAG)として構造化されてよく、任意の所与のノードに対する効率的なアクセスを維持するために、DAGは多様な方法に従って構造化され、バランスを取られてよい。例えば、メタデータ表現のためのDAGは、一種のBツリーとして定義され、メタデータ表現の構造に対する変更に応じて相応してバランスがとられてよく、メタデータ表現に対する変更は、メタデータ表現によって表される基本的なデータに対する変更に応じて、又は加えて発生する場合がある。この例では、簡略にするために2つのレベルしかないが、他の例では、メタデータ表現は、

10

20

30

40

50

複数のレベルにわたって及んでよく、数百又は数千のノードを含んでよく、各ノードは他のノードに対する任意の数のリンクを含んでよい。

【0479】

【0537】 さらに、この例では、メタデータ表現のリーフは、ボリューム又はボリュームの部分のために記憶されたデータに対するポインタを含んでよく、論理アドレス、つまりボリューム及びオフセットは、論理アドレスに対応する記憶されたデータを参照する1つ以上のリーフノードに到達するためにメタデータ表現を識別し、メタデータ表現を通してナビゲートするために使用されてよい。例えば、ボリューム(4152)は、メタデータオブジェクトノード(4152、4152A~4152N)を含むメタデータ表現(4150)によって表されてよく、リーフノード(4152A~4152N)は、それぞれのデータオブジェクト(4153A~4153N、4157)に対するポインタを含む。データオブジェクトは、ストレージシステム(4106)の中の任意のサイズのデータの単位であってよい。例えば、データオブジェクト(4153A~4153N、4157)は、それぞれ論理エクステントであってよく、論理エクステントは、例えば1MB、4MB等のなんらかの指定されたサイズ又はなんらかの他のサイズであってよい。

【0480】

【0538】 この例では、スナップショット(4156)は、記憶されたオブジェクトのスナップショット、この場合はボリューム(4152)として作成されてよく、スナップショット(4156)が作成された時点で、スナップショット(4156)のためのメタデータ表現(4154)は、ボリューム(4152)のためのメタデータ表現(4150)のためのメタデータオブジェクトのすべてを含む。さらに、スナップショット(4156)の作成に応じて、メタデータ表現(4154)は、読取り専用となるように指定されてよい。しかしながら、メタデータ表現を共用するボリューム(4152)は修正され続けてよく、スナップショットが作成される瞬間、修正がボリューム(4152)に対応するデータに加えられるので、ボリューム(4152)のためのメタデータ及びスナップショット(4156)は同一であり、修正に応じて、ボリューム(4152)のメタデータ表現及びスナップショット(4156)は、分岐し、異なってくる。

【0481】

【0539】 例えば、ボリューム(4152)を表現するためのメタデータ表現(4150)及びスナップショット(4156)を表現するためのメタデータ表現(4154)を所与として、ストレージシステム(4106)は、特定のデータオブジェクト(4153B)の中に最終的に記憶されるデータに書き込むI/O動作を受け取ってよく、データオブジェクト(4153B)は、リーフノードポインタ(4152B)によって指され、リーフノードポインタ(4152B)は、メタデータ表現(4150、4154)の部分である。書き込み動作に応じて、メタデータ表現(4154)によって参照される読取り専用データオブジェクト(4153A~4153N)は未変更のままとなり、ポインタ(4152B)も未変更のままとなってよい。しかしながら、現在のボリューム(4152)を表すメタデータ表現(4150)は、書き込み動作によって書き込まれるデータを保持するために新しいデータオブジェクトを含むように修正され、修正されたメタデータ表現は、メタデータ表現(4160)で示される。さらに、書き込み動作は、データオブジェクト(4153B)の一部分だけに向けられてよく、結果的に、新しいデータオブジェクト(4157)は、書き込み動作のためのペイロードに加えてデータオブジェクト(4153B)の以前のコンテンツのコピーを含んでよい。

【0482】

【0540】 この例では、書き込み動作を処理することの一部として、ボリューム(4152)のためのメタデータ表現(4160)は、既存のメタデータオブジェクトポインタ(4152B)を削除するために、及び新しいメタデータオブジェクトポインタ(4158)を含めるために修正され、新しいメタデータオブジェクトポインタ(4158)は、新しいデータオブジェクト(4157)を指すように構成され、新しいデータオブジェクト(4157)は、書き込み動作によって書き込まれたデータを記憶する。さらに、ターゲット

データオブジェクトを参照したメタデータオブジェクトポインタ（４１５２Ｂ）の例外はあるが、ボリューム（４１５２）のためのメタデータ表現（４１６０）は、以前のメタデータ表現（４１５０）の中に含まれたすべてのメタデータオブジェクトを含み続け、メタデータオブジェクトポインタ（４１５２Ｂ）は、上書きされたであろう読取り専用データオブジェクト（４１５３Ｂ）を参照し続ける。

【０４８３】

【０５４１】 このようにして、メタデータ表現を使用し、ボリューム又はボリュームの一部は、メタデータオブジェクトを作成することによって、及びデータオブジェクトの実際の重複なしに、スナップショットを撮られると見なされてよい、又はコピーされると見なされてよいデータオブジェクトの重複は、書き込み動作が、メタデータ表現によって参照される読取り専用データオブジェクトの１つに向けられるまで延期されてよい。

10

【０４８４】

【０５４２】 言い換えると、ボリュームを表現するためにメタデータ表現を使用することの優位点は、ボリュームのスナップショット又はコピーが作成され、一定のより古いときに、具体的にはスナップショット又はコピーのためにメタデータオブジェクトを作成し、スナップショットを撮られている又はコピーされているボリュームの既存のメタデータ表現に対するスナップショット又はコピーメタデータオブジェクトのための参照を作成するのに要する時間内に、アクセス可能であってよいという点である。

【０４８５】

【０５４３】 例の使用として、仮想化されたコピーバイリファレンス（*copy-by-reference*）は、ボリュームのスナップショットを作成する際のメタデータ表現の使用に類似するやり方でメタデータ表現を利用してよい。仮想化されたコピーバイリファレンスのためのメタデータ表現は、多くの場合ボリューム全体のためのメタデータ表現の一部分に相当してよい。仮想化されたコピーバイリファレンスの例の実施態様は、仮想化されたストレージシステムのコンテキストの中にあってよく、ボリュームの中及び間の複数のブロック範囲は、記憶されたデータの一元化されたコピーを参照してよい。係る仮想化ストレージシステムでは、上述されたメタデータは、仮想アドレス、つまり論理アドレスと物理アドレス、つまり実アドレスとの関係进行处理するために使用されてよい。言い換えると、記憶されたデータのメタデータ表現は、それがフラッシュメモリ上の摩耗を削減する又は最小限に抑える点でフラッシュフレンドリと見なされてよい仮想化ストレージシステムを可能にする。

20

30

【０４８６】

【０５４４】 一部の例では、論理エクステントは、簡略な集合体として、又は論理エクステントリファレンスの集合として形成されるなんらかのより大規模な論理エクステントの中の論理的に関連するアドレス範囲として、を含んだ多様な方法で結合されてよい。これらのより大きい結合は、多様な種類の論理エクステントアイデンティティを与えられる場合もあり、さらにより大きい論理エクステント又は集合体にさらに結合される場合があるであろう。コピーオンライトステータスは、実施態様に応じて多様な層に及び多様な方法で適用する場合があるであろう。例えば、エクステントの論理的集合体のうちの１つの論理的集合体に適用されたコピーオンライトステータスは、コピーされた集合体に未変更の論理エクステントに対する参照を保持させ、コピーオンライト論理集合体の一部だけしか変更されないときに（必要に応じて任意の未変更の記憶されたデータブロックに対する参照をコピーすることによる）コピーオンライトされた論理エクステントを作成させる可能性がある。

40

【０４８７】

【０５４５】 重複排除、ボリュームスナップショット、又はブロック範囲スナップショットは、記憶されたデータブロックを参照すること、又は論理エクステントを参照すること、又は論理エクステント（若しくは論理エクステントの識別された集合体）をコピーオンライトとしてマークすることの組合せによりこのモデルで実施されてよい。

【０４８８】

50

【0546】 さらに、フラッシュストレージシステムを用いると、記憶されたデータブロックは、多様な方法とともに編成され、グループ化されてよい。集合体は、より大きい消去ブロックの一部であるページの中に書き出されるので、削除又は置換された記憶されたデータブロックの最終的なガベージコレクションは、消去ブロック全体が消去され、再利用のために準備できるように、他のどこかでなんらかのページ数で記憶されたコンテンツを移動させることを伴ってよい。物理フラッシュページを選択し、最終的にそれらを移行し、ガベージコレクションし、次いで再利用のためにフラッシュ消去ブロックを消去するこのプロセスは、論理エクステント、重複排除、圧縮、スナップショット、仮想コピー、又は他のストレージシステム機能も処理しているストレージシステムの態様によって調整、駆動、実行される場合もあれば、されない場合もある。ページを選択し、ページを移行し、消去ブロックをガベージコレクションし、消去するための調整又は駆動されたプロセスは、例えば、用途の数、エージング予測、記憶されたデータを回復するために過去に必要なにされた電圧レベル又はリトライの数に対する調整等、フラッシュメモリデバイスセル、ページ、及び消去ブロックの多様な特徴をさらに考慮に入れる場合がある。また、それらは、ストレージシステムの中のすべてのフラッシュメモリ全体での分析及び予測を考慮に入れてもよい。

10

【0489】

【0547】 ストレージシステムが、論理エクステントを含んだ有向非巡回グラフに基づいて実装されてよいこの例を続行するために、論路エクステントは、2つのタイプ、つまりなんらかの量の記憶されたデータをなんらかの方法で参照するリーフ論理エクステント、及び他のリーフ又は複合論理エクステントを参照する複合論理エクステントに分類できる。

20

【0490】

【0548】 リーフエクステントは、さまざまな方法でデータを参照できる。リーフエクステントは、記憶されたデータの単一の範囲（例えば、64キロバイトのデータ）を直接的に指す場合もあれば、リーフエクステントは記憶されたデータに対する参照の集合体（例えば、物理的に記憶されたブロックに対する範囲と関連付けられたなんらかの数の仮想ブロックをマッピングするコンテンツの1メガバイト「範囲」）である場合もある。後者の場合、これらのブロックは、なんらかのアイデンティティを使用し、参照されてよく、エクステントの範囲内のいくつかのブロックは、何にもマッピングされ得ない。また、その後者の場合、これらのブロックリファレンスは一意である必要はなく、なんらかの数のボリュームの中及び全体でのなんらかの数の論理エクステントの中の仮想ブロックからの複数のマッピングが同じ物理的に記憶されるブロックにマッピングすることを可能にする。記憶されたブロックリファレンスの代わりに、論理エクステントは簡略なパターンを符号化できるであろう。例えば、同一のバイトの文字列であるブロックは、ブロックが同一バイトの繰り返されるパターンであることを単に符号化できるであろう。

30

【0491】

【0549】 複合論理エクステントは、それぞれ複合論理エクステントのサブレンジから基本的なリーフ論理エクステント又は複合論理エクステントにコンテンツの論理範囲をマッピングする複数のマップを含む、なんらかの仮想サイズを有するコンテンツの論理範囲である場合がある。複合論理エクステントのコンテンツに関係する要求を変換することは、次いで複合論理エクステントのコンテキストの中で要求のためのコンテンツ範囲を採取すること、その要求がどの基本的なリーフ論理エクステント又は複合論理エクステントにマッピングするのかを決定すること、及びそれらの基本的なリーフ論理エクステント又は複合論理エクステントの中のコンテンツの適切な範囲に適用するために要求を変換することを伴う。

40

【0492】

【0550】 ボリューム、つまりファイル又は他のタイプのストレージオブジェクトは、複合論理エクステントとして記述できる。したがって、これらの提示されているストレージオブジェクトは、このエクステントモデルを使用し、編成できる。

【0493】

50

【0551】 実施態様に応じて、リーフ論理エクステント又は複合論理エクステントは、複数の他の複合論理エクステントから参照され、実質的に、ボリュームの中及び全体でのコンテンツのより大きい集合体の安価な重複を可能にするであろう。したがって、論理エクステントは、基本的には参照の非巡回グラフの中で配置される場合があり、それぞれがリーフ論理エクステントで終了する。これは、ボリュームのコピーを作成するために、ボリュームのスナップショットを作成するために、又はX T E N D E D C O P Y又は類似するタイプの動作の一部としてボリュームの中で及びボリューム間でサポートする仮想範囲コピーの一部として使用できる。

【0494】

【0552】 実施態様は、各論理エクステントに、それに名前を付けるために使用できるアイデンティティを提供してよい。複合論理エクステントの中の参照は、論理エクステント識別子及びそれぞれの係る論理エクステントアイデンティティに対応する論理的なサブレンジを含んだリストになるので、これは参照を簡略化する。また、論理エクステントの中では、それぞれの記憶されたデータブロックリファレンスはそれに名前を付けるために使用されるなんらかのアイデンティティに基づいてもよい。

【0495】

【0553】 エクステントのこれらの重複使用をサポートするために、追加機能、つまりコピーオンライト論理エクステントを加えることができる。修正動作がコピーオンライトのリーフ論理エクステント又は複合論理エクステントに影響を及ぼすとき、論理エクステントはコピーされ、コピーは新しい参照であり、（実施態様に応じて）おそらく新しいアイデンティティを有する。コピーは、どのような修正が修正動作から生じて、基本的なリーフ論理エクステント又は複合論理エクステントに関係するすべての参照又はアイデンティティを保持する。例えば、W R I T E、W R I T E S A M E、X D W R I T E R E A D、X P W R I T E、又はC O M P A R E A N D W R I T E要求は、ストレージシステムに新しいブロックを記憶（又は、既存の記憶されているブロックを識別するために重複排除技法を使用）してよく、ブロックの新しい集合に対するアイデンティティを参照又は記憶するために対応するリーフ論理エクステントを修正し、おそらくブロックの以前の集合のための参照及び記憶されているアイデンティティを置き換える。代わりに、U N M A P要求は、1つ以上のブロックリファレンスを削除するためにリーフ論理エクステントを修正してよい。両方のタイプの事例では、リーフ論理エクステントが修正される。リーフ論理エクステントがコピーオンライトである場合、次いで古いエクステントから影響を受けていないブロックリファレンスをコピーしてから、修正動作に基づいたブロックリファレンスを置換又は削除することによって形成される新しいリーフ論理エクステントが作成される。

【0496】

【0554】 リーフ論理エクステントの場所を突き止めるために使用された複合論理エクステントは、次いで新しいリーフ論理エクステントリファレンス又はコピーされ、修正されたリーフ論理エクステントと関連付けられたアイデンティティを、以前のリーフ論理エクステントの代替物として記憶するために修正されてよい。その複合論理エクステントがコピーオンライトである場合、次いで新しい複合論理エクステントは新しい参照として又は新しいアイデンティティとともに作成され、その基本的な論理エクステントに対する任意の影響を受けていない参照又はアイデンティティは、その新しい複合論理エクステントにコピーされ、以前のリーフ論理エクステントリファレンス又はアイデンティティは、新しい論理エクステントリファレンス又はアイデンティティで置換される。

【0497】

【0555】 このプロセスは、修正動作を処理するために使用される非巡回グラフを通る検索経路に基づいて、参照されたエクステントから参照する複合エクステントにさらに過去に遡って続行し、すべてのコピーオンライト論理エクステントがコピーされ、修正され、置換される。

【0498】

【0556】 これらのコピーされたリーフ論理エクステント及び複合論理エクステントは、次いでコピーオンライトであるという特徴を除外する場合があります、これにより追加の修正は追加のコピーを生じさせない。例えば、第1のときに、コピーオンライト「親」複合エクステントの中のなんらかの基本的な論理エクステントが修正され、その基本的な論理エクステントはコピーされ、修正され、コピーは次いで、親複合論理エクステントのコピーされ、置換されたインスタンスの中に書き込まれる新しいアイデンティティを有する。しかしながら、第2のときに、なんらかの他の基本的な論理エクステントがコピーされ、修正され、他の基本的な論理エクステントコピーの新しいアイデンティティは、親複合論理エクステントに書き込まれ、親は、次いで追加のコピーなく定位置で修正され、親複合論理エクステントに対する参照の代わりに必要な置き換える場合がある。

10

【0499】

【0557】 現在のリーフ論理エクステントがない、ボリュームの又は複合論理エクステントの新しい領域に対する修正動作は、それらの修正の結果を記憶するために新しいリーフ論理エクステントを作成してよい。その新しい論理エクステントが既存のコピーオンライト論理エクステントから参照される場合、次いでその既存のコピーオンライト複合論理エクステントは、新しい論理コンテンツを参照し、既存のリーフ論理エクステントを修正するためのシーケンスに類似するコピー、修正、及び置換の別の動作のシーケンスを生じさせる。

【0500】

【0558】 親複合論理エクステントが（実施態様に基づいて）新しい修正動作のために作成するための新しいリーフ論理エクステントを含む関連付けられたアドレス範囲をカバーするほど十分に大きく成長できない場合、次いで親複合論理エクステントは、さらに再び新しい参照又は新しいアイデンティティである単一の「孫」複合論理エクステントから次いで参照される2つ以上の新しい複合論理エクステントにコピーされてよい。その孫論理エクステントがそれ自体、コピーオンライトである別の複合論理エクステントを通して見つけられる場合、次いでその別の複合論理エクステントは、上述の段落で説明されるのと類似した方法でコピーされ、修正され、置換される。このコピーオンライトモデルは、論理エクステントのこれらの有向非巡回グラフに基づいて、ストレージシステム実施態様の中でスナップショット、ボリュームコピー、及び仮想ボリュームアドレス範囲コピーを実装することの一部として使用できる。それ以外の場合書込み可能なボリュームの読取り専用コピーとしてスナップショットを作成するために、ボリュームと関連付けられた論理エクステントのグラフは、マークされたコピーオンライトであり、元の複合論理エクステントに対する参照はスナップショットによって保持される。ボリュームに対する修正動作は、次いで必要に応じて論理エクステントコピーを作成し、それらの修正動作の結果及び元のコンテンツを保持するスナップショットを記憶するボリュームを生じさせる。ボリュームコピーは、元のボリュームとコピーされたボリュームの両方ともコンテンツを修正し、独自のコピーされた論理エクステントグラフ及びサブグラフを生じさせることを除き、類似している。

20

30

【0501】

【0559】 仮想ボリュームアドレス範囲コピーは、（ブロックリファレンスに対する変更がコピーオンライトリーフ論理エクステントを修正しない限り、それ自体、コピーオンライト技術を使用することを伴わない）リーフ論理エクステントの中及び間でブロックリファレンスをコピーすることによって動作する場合がある。代わりに、仮想ボリュームアドレス範囲コピーは、リーフ論理エクステント又は複合論理エクステントに対する参照を重複させる場合があり、これは、より大きなアドレス範囲のボリュームアドレス範囲コピーにはうまく機能する。さらに、これは、グラフが単にリファレンスツリーよりむしろ参照の有向非巡回グラフになることを可能にする。重複された論理エクステントリファレンスと関連付けられたコピーオンライト技術は、仮想アドレス範囲コピーのソース又はターゲットに対する修正動作が、ボリュームアドレス範囲コピー動作の直後に同じ論理エクステントを共用するターゲット又はソースに影響を及ぼすことなくそれらの修正を記憶する

40

50

ために新しい論理エクステントの作成を生じさせることを確実にするために使用できる。

【0502】

【0560】 また、ポッドのための入力／出力動作も、論理エクステントの有向非巡回グラフを複製することに基づいて実施されてよい。例えば、ポッドの中の各ストレージシステムは、論理エクステントのプライベートグラフを実装することができ、これによりポッドのための1つのストレージシステムに関するグラフは、ポッドのための任意の第2のストレージシステム上のグラフに特定の関係を有さない。しかしながら、ポッド内のストレージシステム間でグラフを同期させることには価値がある。これは、再同期のため、及び例えば非同期複製又はリモートストレージシステムに対するスナップショットベースの複製等の特徴を調整するために有用である場合がある。さらに、それは、スナップショットの配布を処理するため及びコピー関係処理のためのなんらかのオーバーヘッドを削減するためにも有用である場合がある。係るモデルでは、ポッドのためのすべての同期中のストレージシステム全体で同期中のポッドのコンテンツを保つことは、ポッドのためのすべての同期中のストレージシステム全体ですべてのボリュームのためにリーフ論理エクステント及び複合論理エクステントのグラフを同期中に保つこと、及びすべての論理エクステントのコンテンツが同期中であることを確実にすることと、基本的に同じである。同期中であるためには、リーフ論理エクステント及び複合論理エクステントを一致させることは同じアイデンティティを有するべきである、又はマッピング可能なアイデンティティを有するべきであるかのどちらかである。マッピングは、中間マッピングテーブルのなんらかの集合を伴う場合もあれば、なんらかの他のタイプのアイデンティティ変換を伴う場合もあるであろう。一部の 경우에는、リーフ論理エクステントによってマッピングされたブロックのアイデンティティも同期中に保つことができるであろう。

【0503】

【0561】 ポッドごとに単一リーダーを有するリーダー及びフォロワーに基づいたポッド実施態様では、リーダーは、論理エクステントグラフに対する任意の変更を決定することを任される場合がある。新しいリーフ論理エクステント又は複合論理エクステントが作成される場合、それはアイデンティティを与えられる場合がある。既存のリーフ論理エクステント又は複合論理エクステントが修正を有する新しい論理エクステントを形成するためにコピーされる場合、新しい論理エクステントは、修正のなんらかの集合を有する以前の論理エクステントのコピーとして記述される場合がある。既存の論理エクステントが分割される場合、分割は、結果として生じる新しいアイデンティティとともに記述される場合がある。論理エクステントが、なんらかの追加の複合論理エクステントから基本的な論理エクステントとして参照される場合、その参照は、その基本的な論理エクステントを参照するための複合論理エクステントに対する変更として記述される場合がある。

【0504】

【0562】 ポッドでの修正動作は、このようにして（コンテンツを拡張するために新しい論理エクステントが作成される場合、又はスナップショット、ボリュームコピー、及びボリュームアドレス範囲コピーに関係するコピーオンライト状態を処理するために、論理エクステントがコピーされ、修正され、置換される場合）論理エクステントグラフに対する修正の記述を配布すること、及びリーフ論理エクステントのコンテンツに対する修正のための記述及びコンテンツを配布することを含む。有向非巡回グラフの形のメタデータを使用することからくる追加の利点は、上述されたように、物理ストレージ内の記憶されたデータを修正するI/O動作が、物理ストレージに記憶されたデータに対応するメタデータの修正によりユーザレベルで物理ストレージに記憶されたデータを修正することなく一効果を与えられ得る点である。物理ストレージがソリッドステートドライブであってよいストレージシステムの開示された実施形態では、フラッシュメモリに対する修正に付随する磨耗は、フラッシュメモリの読取り、消去、又は書込みを通しての代わりに、I/O動作によってターゲットにされるデータを表すメタデータの修正により効果を与えられるI/O動作のために回避又は削減され得る。さらに、仮想化ストレージシステムでは、上述されたメタデータは、仮想アドレス、つまり論理アドレスと物理アドレス、つまり実アド

レスとの関係性を処理するために使用されてよい。言い換えると、記憶されたデータのメタデータ表現は、それがフラッシュメモリ上の摩耗を削減する又は最小限に抑える点でフラッシュフレンドリと見なされてよい仮想化ストレージシステムを可能にする。

【0505】

【0563】 リーダーストレージシステムは、ポッドデータセットのそのローカルコピー及びローカルストレージシステムのメタデータとの関連でこれらの記述を実装するために、独自のローカル動作を実行してよい。さらに、同期中のフォロワーは、ポッドデータセットのそれらの別々のローカルコピー及びその別々のローカルストレージシステムのメタデータとの関連でこれらの記述を実装するために、独自の別々のローカル動作を実行する。リーダー動作とフォロワー動作の両方が完了すると、結果は互換性のあるリーフ論理エクステンツコンテンツを有する論理エクステンツの互換性のあるグラフである。論理エクステンツのこれらのグラフは、次いで上記の例に説明されたように一種の「共通メタデータ」になる。この共通メタデータは、修正動作と必須共通メタデータとの間の依存関係として記述される場合がある。グラフへの変換は、1つ以上の他の動作とともに、例えば依存関係等の関係を記述してよいプレディケートの集合又はより多くのプレディケートの中の別々の動作として記述できる。言い換えると、動作間の相互依存性は、1つの動作がなんらかの方法で依存するプレカーソルの集合として記述されてよく、プレカーソルの集合は、動作が完了するために真でなければならないプレディケートと見なされてよい。プレディケートのより完全な説明は、その全体として参照により本明細書に含まれる出願参照番号第15／696，418号の中で見つけられてよい。代わりに、ポッド全体で完了するとまだ知られていない特定の同じグラフ変換に頼る各修正動作は、それが頼る任意のグラフ変換の部分を含む場合がある。既に存在している「新しい」リーフ論理エクステンツ又は複合論理エクステンツを識別する動作記述を処理することは、その部分はすでになんらかの早期の動作の処理で処理されていたので、新しい論理エクステンツを作成することを回避することができ、代わりにリーフ論理エクステンツ又は複合論理エクステンツのコンテンツを変更する動作処理の部分しか実装できない。変換が互いに互換性があることを確実にすることはリーダーの役割である。例えば、ポッドを受け取る2つの書き込みで開始できる。第1の書き込みは、複合論理エクステンツAを複合論理エクステンツBとして形成されたコピーで置換し、リーフ論理エクステンツCをリーフ論理エクステンツDとしてのコピーで及び第2の書き込みのためのコンテンツを記憶するための修正で置換し、さらにリーフ論理エクステンツDを複合論理エクステンツBに書き込む。一方で、第2の書き込みは、複合論理エクステンツAの同じコピー及び複合論理エクステンツBとの置換を暗示するが、異なるリーフ論理エクステンツEをコピーし、第2の書き込みのコンテンツを記憶するために修正される論理エクステンツFで置換し、さらに論理エクステンツFを論理コンテンツBに書き込む。その場合には、第1の書き込みの記述は、AのBとの置換及びCのDとの置換、並びにDの複合論理エクステンツBへの書き込み、及び第1の書き込みのコンテンツのリーフエクステンツBへの書き込みを含む場合があり、第2の書き込みの記述は、リーフエクステンツFに書き込まれる第2の書き込みのコンテンツとともに、AのBとの置換及びEのFとの置換、並びにFの複合論理エクステンツBへの書き込みを含む場合がある。リーダー又は任意のフォロワーは、次いで任意の順序で第1の書き込み又は第2の書き込みを処理でき、最終結果は、BがAをコピーし、置換し、DがCをコピーし、置換し、FがEをコピーし、置換し、D及びFが複合論理エクステンツBに書き込まれることである。Bを形成するためのAの第2のコピーは、Bがすでに存在していることを認識することによって回避できる。このようにして、リーダーは、ポッドのための同期中のストレージシステム全体で論理エクステンツグラフのために互換性のある共通メタデータを維持することを確実にする場合がある。

【0506】

【0564】 論理エクステンツの有向非巡回グラフを使用するストレージシステムの実施態様を所与として、論理エクステンツの複製された有向非巡回グラフに基づいたポッドの回復が実施されてよい。具体的には、この例では、ポッドでの回復は、複製されたエクステ

10

20

30

40

50

ントグラフに基づいてよく、次いでリーフ論理エクステントのコンテンツを回復することだけでなく、これらのグラフの一貫性を回復することも伴う。回復の本実施態様では、動作は、ポッドのためのすべてのストレージシステム全体で完了したと知られていないすべてのリーフ論理エクステントコンテンツ修正だけでなく、ポッドのためのすべての同期中のストレージシステムで完了したと知られていないグラフ変換に対しても問合せすることを含んでよい。係る問い合わせることは、なんらかの調整されたチェックポイント以降の動作に基づいているであろう、又は単に完了したと知られていない動作であり、各ストレージシステムは、完了したとしてまだ信号で知らせていない通常動作中の動作のリストを保つ。この例では、グラフ変換は簡単である。グラフ変換は新しい事柄を作成し、古い事柄を新しい事柄にコピーし、古い事柄を2つ以上の分割された新しい事柄にコピーする場合もあれば、それらは他のエクステントへのその参照を修正するために複合エクステントを修正する。任意の論理エクステントを作成又は置換する任意の同期中のストレージシステムで見つけられた任意の記憶されている動作記述は、その論理エクステントをまだ有していない任意の他のストレージシステムでコピーされ、実行される場合がある。関与されたリーフ論理エクステント又は複合論理エクステントが適切に回復されている限り、リーフ論理エクステント又は複合論理エクステントに対する修正を説明する動作は、それらの修正を、まだそれらを適用していなかった任意の同期中のストレージシステムに適用できる。

【0507】

【0565】 別の例では、論理エクステントグラフを使用する代替策として、ストレージは複製されたコンテンツアドレス指定可能ストアに基づいて実装されてよい。コンテンツアドレス指定可能ストアでは、データのブロックごとに（例えば、512バイト、4096バイト、8192バイトごとに、又は16384バイトごとにも）、（フィンガープリントと呼ばれる場合もある）一意のハッシュ値が、ブロックコンテンツに基づいて計算され、これによりボリューム又はボリュームのエクステント範囲は、特定のハッシュ値を有するブロックに対する参照のリストとして記述できる。同じハッシュ値を有するブロックに対する参照に基づいた同期複製されたストレージシステム実施態様では、複製は、第1のストレージシステムがブロックを受け取り、それらのブロックのフィンガープリントを計算し、それらのフィンガープリントのためのブロックリファレンスを識別し、参照されたブロックに対するボリュームブロックのマッピングに対する更新として1つ又は複数の追加のストレージシステムに対する変更を送達することを伴う場合があるであろう。ブロックが第1のストレージシステムによってすでに記憶されていることがわかる場合、そのストレージシステムは（参照が同じハッシュ値を使用するため、又は参照のための識別子が同一であるか、容易にマッピングできるかのどちらかであるため）追加のストレージシステムのそれぞれで参照に名前を付けるためにその参照を使用できる。代わりに、ブロックが、第1のストレージシステムによって見つけられない場合、次いで第1のストレージシステムのコンテンツは、そのブロックコンテンツと関連付けられたハッシュ値又はアイデンティティとともに、動作記述の一部として他のストレージシステムに送達されてよい。さらに、各同期中のストレージシステムのボリューム記述は、次いで新しいブロックリファレンスで更新される。係るストアでの回復は、次いでボリュームのための最近更新されたブロックリファレンスを比較することを含んでよい。ブロックリファレンスがポッドのための異なる同期中のストレージシステム間で異なる場合、次いで各参照の1つのバージョンは、それらを一貫性があるものにするために他のストレージシステムにコピーされる場合がある。1つのシステム上のブロックリファレンスが存在しない場合、次いでそれはその参照のためのブロックを記憶するなんらかのストレージシステムからコピーされる。仮想コピー動作は、仮想コピー動作を実施することの一部として参照をコピーすることによって係るブロック又はハッシュリファレンスストアでサポートされる場合がある。

【0508】

【0566】 追加の説明のために、図42Aは、本開示のいくつかの実施形態に従ってデータセットを同期複製するストレージシステムの間でメタデータを同期するための例の方法

を示すフローチャートを説明する。あまり詳細には示されないが、図 4 2 A に示されるストレージシステム (4 2 0 0 A) は、図 1 A から図 1 D、図 2 A から図 2 G、図 3 A から図 3 C、又はその任意の組合せに関して上述されたストレージシステムに類似してよい。実際には、図 4 2 A に示されるストレージシステム (4 2 0 0 A) は、上述されたストレージシステムと同じ構成要素、より少ない構成要素、追加の構成要素を含んでよい。

【0 5 0 9】

【0567】 上述されたように、メタデータは、データセットを同期複製しているストレージシステムの間で同期されてよい。係るメタデータは、ポッドの中のストレージオブジェクトの中の仮想アドレスに対する、ポッドの中に記憶されたコンテンツのセグメントのマッピングに関係するポッドの代わりに、ストレージシステムによって記憶される、共通メタデータ又は共用メタデータと呼ばれる場合があり、それらのマッピングに関する情報は、ポッドに関係するストレージ動作のための正しい挙動—又はより良い性能—を確実にするためにポッドのためのメンバーストレージシステム間で同期される。一部の例では、ストレージオブジェクトは、ボリューム又はスナップショットを実装してよい。同期されたメタデータは (a) ポッド内のシステムストレージ間でボリュームコンテンツマッピングを同期させておくための情報、(b) 回復チェックポイントのため又は進行中の書き込み動作のための追跡データ、(c) 非同期複製又は周期複製のためのデータ及びマッピング情報のリモートストレージシステムへの送達に関係する情報を含んでよい。

【0 5 1 0】

【0568】 ポッド内のストレージシステムの間でコンテンツマッピングを同期させたままにするための情報は、スナップショットの効率的な作成を可能にしてよく、このことは同様にしてその後続の更新、スナップショットのコピーを可能にする、又はスナップショット削除は、ポッドメンバーストレージシステム全体で効率的且つ一貫して実行されてよい。

【0 5 1 1】

【0569】 回復チェック委員会とのため又は進行中の書き込み動作のためにデータを追跡することは、効率的な衝突回復、及び部分的に又は完全にポッドのための個々のストレージシステムで適用された可能性があるが、ポッドのための他のストレージシステムでは完全に適用されなかった可能性があるコンテンツ又はボリュームマッピングの効率的な検出を可能にしてよい。

【0 5 1 2】

【0570】 非同期複製又周期複製のためのリモートストレージシステムへのデータ及びマッピング情報の送達に関係する情報は、ポッドのための 2 つ以上のメンバーストレージシステムが、非同期複製又は周期複製を駆動するために使用されるメタデータをマッピングし、識別する際の不一致に対処するための最小の懸念で、複製されたポッドコンテンツのためのサーバとしての機能を果たすことを可能にしてよい。

【0 5 1 3】

【0571】 一部の例では、共用メタデータは、名前が付けられたグループのための若しくはグループの表示、又はポッドのための同期複製されたデータセット全体の部分集合である 1 つ以上のボリューム又は 1 つ以上のストレージオブジェクトの識別子を含んでよい。データセットのボリューム又はストレージオブジェクトの係るは、一貫性グループと呼ばれることがある。一貫性グループは、一貫性のあるスナップショット、非同期複製又は周期複製に使用されるデータセットのボリューム又はストレージシステムの集合を指定するために定義されてよい。一部の例では、一貫性グループは、例えばホスト若しくはホストネットワークポートの特定の集合に接続される、又はアプリケーション若しくはバーチャルマシン若しくはコンテナの特定の集合に接続されるすべてのボリュームを含むことによって動的に計算されてよく、アプリケーション、仮想マシン、若しくはコンテナは、外部サーバシステム上で動作してよい、又はポッドのメンバーであるストレージシステムの 1 つ以上で動作してよい。他の例では、一貫性グループは、データ若しくはデータの集合のタイプのユーザ選択又は動的計算に類似した一貫性グループの仕様に従って定義されてよく、ユーザは、例えばコマンド又は管理コンソールを通して、特定の若しくは名前を付け

10

20

30

40

50

られた一貫性グループが、ホスト若しくはホストネットワークポートの特定の集合に接続されたすべてのボリュームを含むために作成される、又はアプリケーション若しくはバーチャルマシン若しくはコンテナの特定の集合のためのデータを含むために作成されることを指定してよい。

【0514】

【0572】 一貫性グループを使用する例では、一貫性グループの第1の一貫性グループスナップショットは、第1のデータセットスナップショット時の一貫性グループのメンバーであるすべてのボリューム又は他のストレージオブジェクトのためのスナップショットの第1の集合を含んでよく、同じ一貫性グループの第2の一貫性グループスナップショットは、第2のデータセットスナップショット時で一貫性グループのメンバーであるボリューム又は他のストレージオブジェクトのためのスナップショットの第2の集合を含む。他の例では、データセットのスナップショットは、非同期で1つ以上のターゲットストレージシステムに記憶されてよい。同様に、一貫性グループの非同期複製は、一貫性グループのメンバーボリューム及び他のストレージオブジェクトに対する動的変更を説明してよく、非同期複製リンクのソース又はターゲットのどちらかでの一貫性グループの一貫性グループスナップショットは、データセットスナップショットが関係するときの一貫性グループの関係でメンバーであるボリューム及び他のストレージオブジェクトを含む。非同期複製接続のターゲットの場合、データセットスナップショットが関係する時間は、それが受け取られ、ターゲットで一貫性グループスナップショットのときのプロセスにあったときに送信者の動的データセットに依存する。例えば、非同期複製のターゲットが、例えば2000動作後方にあり、それらの動作のいくつかが一貫性グループメンバーの変更であり、係る変更の第1の集合がソースにとって2000を超える動作前であり、変更の第2の集合が最後の2000の範囲内である場合、次いでターゲットでのその時点での一貫性グループスナップショットは、メンバー変更の第1の集合を説明し、変更の第2の集合を説明しない。非同期複製のターゲットの他の使用は、それらの使用のためのボリューム又は他のストレージオブジェクト（及びそのコンテンツ）を決定する際に一貫性グループのためのデータセットの時点の性質を同様に説明してよい。例えば、非同期複製が2000動作後方である同じ場合では、災害復旧フェイルオーバのためのターゲットの使用は、それらがソースで2000動作前であったときのボリューム及び他のストレージオブジェクト（及びそのコンテンツ）を含むデータセットから開始する可能性がある。この説明では、ソースで同時に起こる動作（例えば、書込み、ストレージオブジェクトの作成又は削除、ボリューム又は他のストレージオブジェクト又は一貫性データからの他のデータの包含又は除外に影響を及ぼすプロパティに対する変更、又は進行中であり、同じ時点で完了したとして信号で知らされていない他の動作）は、単一の明確に定義された順序付けを有さない可能性があるため、動作のカウントはソースで同時に起こる動作の任意の可能にされた順序付けに基づいたなんらかの妥当な順序付けを表す必要があるにすぎない。

【0515】

【0573】 一貫性グループを使用する別の例として、一貫性グループスナップショットの複製に基づいた周期複製の場合、それぞれの複製された一貫性グループスナップショットは、各一貫性グループスナップショットがソースで形成されたときのボリューム及び他のストレージオブジェクトを含むであろう。一貫性グループのメンバーシップが、共通、つまり共用のメタデータを使用することによって一貫して保たれることを確実にすることは、障害—又は、複製のソース若しくはデータセットスナップショットを形成するシステムにポッド内のあるストレージシステムから別のストレージシステムに切り替えさせる場合がある—が、それらの一貫性グループスナップショット又は一貫性グループ複製を適切に処理するために必要とされる情報を失わないことを確実にする。さらに、このタイプの処理は、ポッドのメンバーである複数のストレージシステムが非同期複製又は周期複製のためのソースシステムとしての機能を同時に果たすことを可能にしてよい。

【0516】

【0574】 さらに、セグメントのストレージオブジェクトへのマッピングを記述する同期

されたメタデータは、マッピング自体に限定されず、他のストレージシステム情報の中で、例えばシーケンス番号（又は記憶されたデータを識別するためのなんらかの他の値）、タイムスタンプ、ボリューム／スナップショット関係、チェックポイントアイデンティティ、階層を定義するツリー若しくはグラフ、又はマッピング関係の有向グラフ等の追加の情報を含んでよい。

【0517】

【0575】 図42Aに示されるように、データセット（4258）を同期複製している複数のストレージシステム（4200A～4200N）は、ポッドのための同期中リスト内のそれぞれの他のストレージシステム（4200B～4200N）と通信してよい。ストレージシステムは、実行するためのI/O動作を記述するメタデータ及び個々のストレージシステムに記憶されたデータセット（4258）のそれぞれのローカルメタデータ表現に加えられる更新を記述するメタデータを交換してよい。さらに、各ストレージシステム（4200A、4200B、... 4200N）は、ストレージオブジェクト（4256、4260... 4262）のそれぞれのバージョンを記憶してよい。

10

【0518】

【0576】 図42Aに示される例の方法は、ストレージシステム（4200A～4200N）の第1のストレージシステム（4200A）で、データセット（4258）に向けられたI/O動作（4252）を受け取ること（4202）を含む。ストレージシステム（4200A～4200N）の第1のストレージシステム（4200A）で、データセット（4258）に向けられたI/O動作（4252）を受け取ること（4202）は、例えばストレージエリアネットワーク（158）、インターネット、又はホストコンピュータ（4251）がその全体でストレージシステム（4200A）と通信してよい任意のコンピュータネットワーク等のネットワーク全体でパケット又はデータを転送するための1つ以上の通信プロトコルを使用することによって実施されてよい。この例では、ストレージシステム（4200A）は、例えばSCSIポート等のネットワークポートで受け取られたI/O動作（4252）を受け取ってよく、I/O動作（4252）は、ポッド内のストレージシステム（4200A～4200N）全体で同期複製されているデータセット（4258）の一部であるメモリ場所に向けられる書込みコマンドである。

20

【0519】

【0577】 また、図42Aに示される例の方法は、I/O動作（4252）に従って、ストレージオブジェクト（4256）の中の仮想アドレスに対するコンテンツのセグメントのマッピングを記述するメタデータ更新（4254）を決定すること（4204）であって、ストレージオブジェクト（4256）はデータセット（4258）を含む、決定すること（4204）を含む。I/O動作（4252）に従って、ストレージオブジェクト（4256）の中の仮想アドレスに対するコンテンツのセグメントのマッピングを記述するメタデータ更新（4254）を決定すること（4204）であって、ストレージオブジェクト（4256）がデータセット（4258）を含む決定することは、ポッドのストレージシステム（4200A～4200N）全体で同期されるメタデータのコンテンツに関して上述されるように情報を決定する又は識別することによって実施されてよく、I/O動作（4252）からの情報も、メタデータ更新（4254）、係る論理、つまり仮想アドレス、ペイロードサイズ、及びI/O動作（4252）ペイロードが、データセット（4258）の中に以前に記憶されたデータに関して含まれる又は組み込まれる方法を記述する重複排除情報等の他の情報に含まれてもよい。

30

40

【0520】

【0578】 また、図42Aに示される例の方法は、メタデータ更新（4254）に従って第2のストレージシステム上でメタデータ表現を更新するためにメタデータ更新（4254）を第2のストレージシステム（4200B）に送信することによってストレージシステム（4200A～4200N）の第2のストレージシステム（4200B）でメタデータを同期させること（4206）も含む。メタデータ更新（4254）に従って第2のストレージシステム上でメタデータ表現を更新するためにメタデータ更新（4254）を第

50

2のストレージシステム(4200B)に送信することによってストレージシステム(4200A~4200N)の第2のストレージシステム(4200B)でメタデータを同期させること(4206)は、実施されてよい。1つ以上のネットワークポートを使用し、及び1つ以上の通信ネットワーク(不図示)全体で、ポッド内のそれぞれの他のストレージシステム(4200Bから4200N)にメタデータ更新(4254)を送信することによって実施されてよい。それぞれの他のストレージシステム(4200Bから4200N)は、同期されたデータセット(4258)のそれぞれのローカルメタデータ表現を更新するためにメタデータ更新(4254)を受け取ってよい。各ストレージシステム(4200B~4200N)がメタデータ更新(4254)を受け取り、処理した後、すべてのシステムで同期されたデータセット(4258)に対応するメタデータが同期される。

10

【0521】

【0579】 追加の説明のために、図42Bは、本開示のいくつかの実施形態に従ってデータセットを同期複製するストレージシステムの間でメタデータを同期させる例の方法を示すフローチャートを説明する。図42Bに示される例の方法は、ストレージシステム(4200A~4200N)の第1のストレージシステム(4200A)で、データセット(4258)に向けられたI/O動作(4252)を受け取ること(4202)と、I/O動作(4252)に従って、ストレージオブジェクト(4256)の中の仮想アドレスに対するコンテンツのセグメントのマッピングを記述するメタデータ更新(4254)を決定すること(4204)であって、ストレージオブジェクト(4256)はデータセット(4258)を含む、決定することと、メタデータ更新(4254)に従って第2のストレージシステム上でメタデータ表現を更新するためにメタデータ更新(4254)を第2のストレージシステム(4200B)に送信することによってストレージシステム(4200A~4200N)の第2のストレージシステム(4200B)でメタデータを同期させること(4206)も含むので、図42Bに示される例の方法は図42Aに示される例の方法に類似している。

20

【0522】

【0580】 しかしながら、図42Bに示される例の方法は、I/O動作(4252)を第1のストレージシステム(4200A)上のデータセット(4258)に適用すること(4288)と、第1のストレージシステム(4200A)上でI/O動作(4252)を無事に適用することに応じて、第1のストレージシステム(4200A)で及びメタデータ更新(4254)に従って、I/O動作(4252)を適用する前にデータセット(4258)に対応するメタデータのバージョンを更新すること(4290)と、1つ以上の他のI/O動作に対するI/O動作(4252)の順序付けを記述するプレディケートメタデータを決定すること(4292)と、をさらに含む。

30

【0523】

【0581】 I/O動作(4252)を第1のストレージシステム(4200A)上のデータセット(4258)に適用すること(4288)は、図1に関して上述されたように、ストレージシステム(4200A)のコントローラによって実施されてよく、例えばNVRAM等のストレージシステム(4200A)及び例えばフラッシュメモリ等の永続記憶装置又は任意のタイプのソリッドステート不揮発性メモリのメモリ構成要素の1つ以上を使用し、書込み動作を実施するコントローラの記述。

40

【0524】

【0582】 第1のストレージシステム(4200A)で及びメタデータ更新(4254)に従って、I/O動作(4252)を適用する前にデータセット(4258)に対応するメタデータのバージョンを更新すること(4290)は、対応するストレージオブジェクト(4258)、つまりソースボリュームのためのメタデータ表現の一部分を識別し、部分が一部の場ではソースボリューム全体であってよいデータセット(4258)を記憶することによって実施されてよい。さらに、ストレージオブジェクト(4256)のためのメタデータ表現の部分は、I/O動作(4252)のためのメモリアドレスデータを使用して、上述されたメタデータオブジェクトの構造化された集合体を横断し、I/O動作

50

(4252) のためのメモリアドレスのデータオブジェクトに対応するノードを見つけることによって識別されてよい。さらに、ストレージオブジェクト(4256)のためのメタデータ表現の中の1つ以上のノードを参照するメタデータ表現のためのメタデータオブジェクトルートノードが作成されてよく、メタデータオブジェクトルートノードは、ストレージオブジェクト(4256)全体のメタデータ表現の中の1つ以上のノードの部分に指定する場合もあれば、I/O動作(4252)に対応するストレージオブジェクト(4256)全体のためのメタデータ表現の部分だけを参照するための他の表示を指定する場合もある。このようにして、データセットのメタデータ表現は、I/O動作(4252)の無事適用を反映する。

【0525】

【0583】 1つ以上の他のI/O動作に対するI/O動作(4252)の順序付けを記述するプレディケートメタデータを決定すること(4292)は、受け取られた各I/O動作を追跡し、任意の依存関係がI/O動作間に存在するかどうかを判断することによって実施されてよく、係るI/O動作が識別された後、まさにそれらは例えばリーダー定義の順序付け又はプレディケート又はインターロック例外を通して等の技術を使用する通常の実行時中であるべきなので、任意の順序付け一貫性問題は解決されるはずである。インターロック例外は、その全体として本明細書に援用される出願参照番号第15/696,418号の中で説明される。プレディケートに関しては、動作と共通メタデータ更新との関係の説明は、別々の修正動作間の相互依存性の集合として説明されてよい—これらの相互依存性は、ある動作がなんらかの方法で依存するプレカーソルの集合として記述される場合があり、プレカーソルの集合は、動作が完了するために真でなければならないプレディケートと見なされる場合がある。さらに、プレディケートは、それらがリーダーとフォロワーとの間の同時並行性に対する制限を広めるために使用される場合、それらのプレディケートが、ストレージシステムが情報を持続する順序を動かすならば持続される情報は多様なもっともらしい結論を暗示するので、保護される必要はない可能性がある。

【0526】

【0584】 追加の説明のために、図43は、本開示のいくつかの実施形態に従ってデータセットを同期複製するストレージシステムの間アクティブメンバーシップを決定するための例の方法を示すフローチャートを説明する。あまり詳細には示されないが、図43に示されるストレージシステム(4300A~4300N)は、図1Aから図1D、図2Aから図2G、図3Aから図3C、又はその任意の組合せに関して上述されたストレージシステムに類似してよい。実際には、図43に示されるストレージシステム(4300A~4300N)は、上述されたストレージシステムと同じ構成要素、より少ない構成要素、追加の構成要素を含んでよい。

【0527】

【0585】 以下の例では、データセットを同期複製するストレージシステムの間エラーから回復する方法を決定することは、定足数プロトコルに従事させるかどうかを判断することを含んでよい。以下に説明されるように、ストレージシステムクラスタ又はポッド内でのアクティブメンバーシップを決定することは、データセットに向けられるI/O動作にサービスを提供し続けるストレージシステムの部分集合を決定することによって、例えば通信障害又はストレージデバイス故障等のエラーを克服し得、ストレージシステムの別の部分集合は、データセットに向けられたI/O動作にサービスを提供することを続行しない。このようにして、仲介又は定足数の方針のどちらかを通して、1つ以上のストレージシステムは、ストレージシステムがストレージシステムの同期中リストから加えられる又は削除されると、同期複製されたデータセットを修正するI/O動作の履歴を制御すると決定される。

【0528】

【0586】 例えば、ストレージシステム間の通信障害等のエラーに応じて仲介サービスを従事させるプロセスストレージシステムは、広域ネットワークを介する連絡アドレスとして示すハンドル、及び仲介のための鍵のプールを管理するために使用できる暗号で安全

10

20

30

40

50

なトークンを記憶するように構成されてよい—その全体として本願に援用される出願参照番号第14/703, 559号の中でより詳細に説明される。やはり出願参照番号第15/703, 559号の中で説明されるのは、データセットを複製するストレージシステムの集合の中のどのストレージシステムが、データセットに向けられたI/O要求にサービスを提供し続けるべきかを判断するための多様な定足数プロトコルの使用である。

【0529】

【0587】 本願参照番号第14/703, 559号は、仲介及び定足数プロトコルのための実施態様を説明しているが、本開示の焦点は、ストレージシステムクラスタでのアクティブメンバーシップを決定するためにどの技術を実行するのかを判断するための分析である。例えば、いくつかの状況では、1つ以上のストレージシステムが、別の1つ以上のストレージシステムを介して仲介に従事し、仲介を勝ち取ることが可能である—他方の1つ以上のストレージシステムが仲介を勝ち取った場合、I/O動作にサービスを提供するためのストレージシステムの性能は、1つ以上のデータストレージ測定基準又は性能基準に従ってより優れているであろう。言い換えると、仲介を勝ち取る1つ以上のシステムが仲介を勝ち取ることは必ずしも有利ではない。以下に説明されるのは、仲介を勝ち取るべきではない1つ以上のストレージシステムが、より高い性能を示しているストレージシステムにより大きい数量の投票を割り当てることを含み、仲介を勝ち取る、又は特定のホストシステムにより密接に接続される、又はその測定基準がより悪いストレージシステムに割り当てられる投票の数量に比べて他の比較的に良い測定基準を有する状況を回避するための技術である。

【0530】

【0588】 一部の例では、ストレージシステムクラスタでアクティブメンバーシップを決定するために、データセットに向けられたI/O要求にサービスを提供し続けるための1つ以上のストレージシステムの集合を決定するためのデフォルト技術は—ストレージシステム(4300A)が、定足数プロトコルの使用が、I/O要求にサービスを提供し続けるための1つ以上のストレージシステムの集合を決定するための定足数を確立できないであろうと判断又は証明することができない限り—定足数プロトコルを使用することによって実施されてよい。言い換えると、ストレージシステム(4300A)の中のストレージデバイス間の通信障害等のエラーに応じて、ストレージシステム(4300A)のコントローラは、定足数を確立できるかどうかを判断してよい—特定の定足数プロトコルの下で定足数を確立できる場合、次いで定足数プロトコルは、ストレージシステムクラスタ内のアクティブな又は同期中のメンバーシップを決定するために使用され、それ以外の場合、特定の定足数プロトコルの下で定足数を確立できない場合、次いでストレージシステム(4300A)は、ストレージシステムクラスタ内のアクティブな又は同期中のメンバーシップを決定するために仲介に従事してよい。簡略な場合には、同期中リストのメンバーである2つのストレージシステムがある場合、他の単一のストレージシステムが定足数を形成できないため、次いで通信障害に応じて定足数分析は実行されない。一部の例では、所与のストレージシステム又は互いに通信するストレージシステムの集合は、ストレージシステムの別の集合が、ストレージシステムの各集合に対する投票の比較に基づいて定足数を形成し得る可能性を排除することができない限り、ストレージシステム又は互いに通信するストレージシステムは、通信で結合されていない1つ以上のストレージシステムをデタッチする定足数方針に頼る場合がある。言い換えると、互いと通信中である1つ以上のストレージシステムが、定足数を形成できる場合、次いで通信していない1つ以上のストレージシステムは、仲介に訴えることなく定足数方針を使用し、デタッチされてよい。

【0531】

【0589】 一部の例では、ストレージシステム(4300A)のためのコントローラは、互いと通信しているストレージシステムの集合を決定することによって、定足数を確立し得るかどうかを判断してよい—このことは、データセットを同期複製しているストレージシステム(4300A~4300N)の完全な集合に含まれるストレージシステム(4300B~4300N)の同期中リストを参照しているストレージシステム(4300A)

に基づいて、それらが通信していないストレージシステムの集合を決定する働きもする。さらに、ストレージシステム（４３００Ａ）は、すべてのストレージシステムの集合又は最後に知られていた同期中のストレージシステム（４３００Ａ～４３００Ｎ）の集合の中の各ストレージシステムが、所与の定足数プロトコルに関して有する票数の記憶されている表示を参照してよい。このようにして、ストレージシステムの第１の集合のそれぞれに対応するそれぞれの票とともに、互いと通信しているストレージシステムの決定された第１の集合、及びストレージシステムの第２の集合のそれぞれに対応するそれぞれの票とともに、分析を実行しているストレージシステム（４３００Ａ）と通信していないストレージシステムの決定された第２の集合を所与として、ストレージシステム（４３００Ａ）は、ストレージシステムの第１の集合のストレージシステムが、定足数を確立するほど十分な票を有しているのか、及びストレージシステムの第２の集合のストレージシステムが定足数を確立するほど十分な票を有する可能性があるのかを判断してよい。

10

【０５３２】

【０５９０】 言い換えると、通信障害を検出し、通信障害に対応するストレージシステム（４３００Ａ）は、（１）通信していないストレージシステムがおそらく定足数を形成し得るかどう、（２）通信に留まっているストレージシステムがおそらく定足数を形成し得るかどう、及び／又は（３）通信していない任意のストレージシステムが、通信しているストレージシステムが定足数を形成し得るかどうかを一少なくとも、ストレージシステムのどの集合が通信に留まるのか、ストレージシステムのどの集合が通信していないのか、及びストレージシステムの各集合の中の各ストレージシステムに対応するそれぞれの票に応じて一おそらく決定し得るかどうかを判断してよい。

20

【０５３３】

【０５９１】 さらに、ストレージシステム（４３００Ａ）が、別のストレージシステム又はストレージシステムの別の集合が定足数を形成できないであろうことを確実にすることによって、ストレージシステム（４３００Ａ）は、それが、ストレージシステム（４３００Ａ）と通信する１つ以上の他のストレージシステムとともに、仲介を勝ち取る場合、ストレージシステムの１つ以上が、仲介を勝ち取ったストレージシステムと再同期するようになった場合、データセットに不一致がなくなるように、他のストレージシステム又はストレージシステムの他の集合は、同期複製されたデータセットのバージョンを作成できるであろうことを確実にする。

30

【０５３４】

【０５９２】 ある例として、なんらかのシステム障害の前にポッドのメンバーとして同期中リストに属するストレージシステムと関連付けられた偶数の票があつてよい。この例では、同期中のポッドメンバーのストレージシステムの第１の集合が互いに通信中であり、ストレージシステムの第１の集合が定足数を確立するための票の正確に半分に相当する場合、互いと通信していた可能性があるが、ストレージシステムの第１の集合とではない一ストレージシステムの他の集合は、定足数を確立するために必要とされる票の半分以上を構成できないであろう。この例では、ストレージシステムの第１の集合に含まれるストレージシステムが、ストレージシステムの第１の集合も、ストレージシステムの他の集合もおそらく定足数を形成し得ないと判断する場合があります、それに応じてストレージシステムは仲介を開始してよい。

40

【０５３５】

【０５９３】 別の例として、定足数が通信している又は通信していないシステムによって確立し得るかどうのための判断は、複数のそれぞれの障害イベントに応じて繰り返し決定される場合がある一例えば、仲介又は定足数投票に従事することによって応える等、異なる対応は、各障害イベントに応じて実行されてよい。例えば、通信障害等の障害の前、ストレージシステムの集合は、ポッドの同期中のメンバーであつてよい。この例では、定足数を確立するための１票を有する特定のストレージシステムは、ポッドの他のメンバーとの通信を失う場合があります、それに応じて、ポッドの他のメンバーは、定足数を確立するほど十分な票を有することによって、投票を通じて特定のストレージシステムを削除する。

50

特定のストレージシステムを削除するためのこの定足数投票段階の完了後、ポッドの同期中メンバーは、特定のストレージシステムを除いたポッドの同期中システムである一定足数を確立するための合計で4票となるストレージシステムの同期中メンバーリストを生じさせる。この例を続けるためには、上述されたように、障害によりストレージシステムが、合計で2票を有するストレージシステムの集合に属することになる場合、次いでストレージシステムは、定足数は可能ではないと判断する場合があります、仲介を開始する場合がある。

【0536】

【0594】 一部の例では、ポッドの同期中メンバーであるストレージシステムは、ゼロ票を含んだ、異なるそれぞれの数の票を割り当てられる場合がある。例えば、ポッドの同期中メンバーであるストレージシステムの集合の場合、異なる票分布は、(a) 単一票を有するすべてのストレージシステム、(b) 複数票を有するいくつかのストレージシステム及び単一票を有するいくつかのストレージシステム、(c) 複数票を有するいくつかのストレージシステム、複数票を有するいくつかのストレージシステム、及びゼロ票を有するいくつかのストレージシステム、又は(d) 異なる数の表を有する各ストレージシステムを含む。言い換えると、一般的には、ポッドの同期中メンバーである任意の所与のストレージシステムは、ゼロ票を含んだ任意の数の表を割り当てられてよい。ゼロ票を有するストレージシステムの一例は、ポッドのメンバーであるソースストレージシステムからまだポッド内にないターゲットストレージシステムへのデータセットの移行中に発生する場合がある。移行の完了前、ソースストレージシステムはその1つ又は複数の票を制御し、ターゲットストレージシステムはいずれの票も制御せず、移行完了後、ターゲットストレージシステムは、1票又は複数票の制御を与えられてよく、ソースストレージはいかなる票も制御する又は有することがないように更新されてよい。

【0537】

【0595】 一部の例では、同期中のリストは、追加のストレージシステムがポッドに加えられる、又はポッドからデタッチされると、ストレージシステムの集合の中で確立されてよく、各ストレージシステムはポッドのメンバーを示すメタデータを維持してよく、同期中リストはポッドの各メンバーのステータスをさらに示す。ストレージシステムがポッドに加えられると、加えられるストレージシステムは、ポッド内の既存のストレージシステムによって、仲介サービスに連絡するための仲介ハンドルを与えられてよい。さらに、ポッドに対する変更が加えられると、同期中のリストは現在のポッドの中のストレージシステムの現在のメンバーシップを反映するために更新される。ポッド定義及び管理に関する追加の説明は、その全体として参照により本明細書に専用される出願参照番号第62/470,172号及び第62/518,071号の中に見つけられる。さらに、複数のストレージシステムを含むようにポッドが広げられる、つまり拡張されるにつれ、ポッド内のストレージシステムは、特定の仲介サービスから仲介を要求するように構成されてよい。このようにして、ポッド内の各ストレージシステムは、仲介が所与のエラーに対する対応であると判断される場合、同じ仲介サービスから仲介を要求する。仲介サービスにアクセスするためのストレージシステムの構成は、本明細書にさらに説明される。

【0538】

【0596】 図43に示されるように、データセット(4352)を同期複製している複数のストレージシステム(4300A~4300N)は、1つ以上のネットワーク(不図示)を介してそれぞれの他のストレージシステムと、及び仲介サービス(4301)と通信してよい。仲介サービス(4301)は、ストレージシステム間の通信障害の場合、ストレージシステムがオフラインになる場合、又はなんらかの他のトリガイイベントに起因して、どのストレージサービスがデータセットにサービスを提供し続けるのかを決定してよい。概して、任意の数のストレージシステムは、データセット(4352)を同期複製している同期中のリストの一部であってよい。

【0539】

【0597】 図43に示される例の方法は、ストレージシステム(4300A~4300N

）のうちの特定のストレージシステム（４３００Ａ）によって、ストレージシステム（４３００Ｂ～４３００Ｎ）の１つ以上との通信の損失を伴う通信障害を検出すること（４３０２）であって、特定のストレージシステム（４３００Ａ）が、例えば仲介サービス（４３０１）等の仲介ターゲットから仲介を要求するように構成される、検出することを含む。通信障害を検出すること（４３０２）は、特定のストレージシステム（４３００Ａ）のコントローラが、通信リンク（４３５４）又はチャンネルを介して、なんらかの期間のうちに別のストレージシステムから通信（４３００Ｂ～４３００Ｎ）を受け取らないことを含んだ、いくつかの技術を使用し、実施されてよい。別の例では、通信障害を検出すること（４３０２）は、通信チャンネルが正しく動作していないと判断するためにクロック交換プロトコルに従って特定のストレージシステム（４３００Ａ）のコントローラによって実施されてよく、クロックは、本明細書に含まれる他の項でより詳細に説明される。通信障害を検出する（４３０２）のための他の標準的な技術も実施されてよい。

10

【０５４０】

【０５９８】 また、図４３の例の方法は、１つ以上のストレージシステム（４３００Ｂ～４３００Ｎ）の少なくとも１つが、通信障害に応じて、例えば仲介サービス（４３０１）等の仲介ターゲットから仲介を要求するように構成されると判断すること（４３０４）も含む。１つ以上のストレージシステム（４３００Ｂ～４３００Ｎ）の少なくとも１つが、通信障害に応じて、例えば仲介サービス（４３０１）等の仲介ターゲットから仲介を要求するように構成されると判断すること（４３０４）は、ストレージシステム（４３００Ａ）が上述された分析を受けることによって実施されてよい。ストレージシステム（４３００Ａ）と通信する１つ以上のシステムに対応する票の比較、及びストレージシステム（４３００Ａ）と通信していない１つ以上のストレージシステムに対応する票に基づいて、ストレージシステム（４３００Ａ）が、ストレージシステム（４３００Ａ）と通信していない１つ以上のストレージシステムが定足数を形成できるかどうか、及びストレージシステム（４３００Ａ）と通信していないそれらの１つ以上のストレージシステムが、仲介ターゲットとの仲介におそらく従事している可能性があるかどうかを判断できる。

20

【０５４１】

【０５９９】 また、図４３の例の方法は、ストレージシステム（４３００Ｂ～４３００Ｎ）の１つ以上が仲介ターゲットから仲介を要求するように構成されると判断すること（４３０４）に応じて、仲介ターゲットから仲介を要求するかどうかを判断すること（４３０６）も含む。仲介ターゲットから仲介を要求するかどうかを判断すること（４３０６）は、分析、及び通信が失われた１つ以上のストレージシステムの少なくとも１つが、仲介ターゲットから仲介を要求している場合がある旨の判断（４３０４）に基づいて実施されてよい。ストレージシステム（４３０４Ａ）が、１つ以上のストレージシステムの少なくとも１つが、メタデータターゲットから仲介を要求している場合があると判断する場合、次いでストレージシステム（４３００Ａ）は、仲介ターゲットからも仲介を要求する場合がある。それ以外の場合、ストレージシステム（４３００Ａ）、及びストレージシステム（４３００Ａ）と通信中の任意のストレージシステムは、１つ以上のストレージシステムの少なくとも１つをデタッチするために定足数方針に従事する場合もあれば、ストレージシステム（４３００Ａ）及びストレージシステム（４３００Ａ）と通信する任意のストレージシステムは、ストレージシステムの別の集合が、定足数を有する可能性があり、同期複製されたデータセットのコピーを用いて操作するのを停止すると判断してよい（それらのストレージシステムは、それがストレージシステムの同期中リスト及び同期外リストの状態を判断するのに役立つ場合がある複数のストレージシステムと通信が確立できるまで実質的にオフラインになる）。

30

40

【０５４２】

【０６００】 追加の説明のために、図４４は、本開示のいくつかの実施形態に従ってデータベースを同期複製するストレージシステム間のアクティブメンバーシップを決定する例の方法を示すフローチャートを説明する。図４４に示される例の方法は、ストレージシステム（４３００Ａ～４３００Ｎ）のうちの特定のストレージシステム（４３００Ａ）によ

50

って、ストレージシステム（４３００Ｂ～４３００Ｎ）の１つ以上との通信障害を検出すること（４３０２）であって、特定のストレージシステム（４３００Ａ）が仲介ターゲットから仲介を要求するように構成される、検出することと、１つ以上のストレージシステム（４３００Ｂ～４３００Ｎ）の少なくとも１つが、通信障害に応じて、例えば仲介サービス（４３０１）等の仲介ターゲットから仲介を要求するように構成されると判断すること（４３０４）と、ストレージシステム（４３００Ｂ～４３００Ｎ）の１つ以上が、仲介ターゲットから仲介を要求するように構成されると判断すること（４３０４）に応じて、仲介ターゲットから仲介を要求するかどうかをと判断すること（４３０６）も含むので、図４４に示される例の方法は、図４に示される例の方法に類似している。

【０５４３】

【０６０１】 しかしながら、図４４に示される例は、１つ以上のストレージシステムの第１の集合によって、１つ以上のストレージシステムの第１の集合のストレージシステムの間の一貫性のある通信があると判断すること（４４０２）であって、ストレージシステムのうちの各ストレージシステムは、１つ以上のストレージシステムの第１の集合が、１つ以上のストレージシステムの第２の集合をデタッチし得るかどうかを判断する定足数プロトコルの中のゼロ以上の票に対応する、判断することと、１つ以上のストレージシステムの第１の集合によって、１つ以上のストレージシステムの第２の集合の間でのストレージシステムとの通信の欠如を判断すること（４４０４）であって、１つ以上のストレージシステムの第１の集合が定足数を形成できない、判断することと、１つ以上のストレージシステムの第１の集合によって、１つ以上のストレージシステムの第２の集合も定足数を形成できないと判断すること（４４０６）をさらに含む。

【０５４４】

【０６０２】 １つ以上のストレージシステムの第１の集合によって、１つ以上のストレージシステムの第１の集合のストレージシステムの間の一貫性のある通信があると判断すること（４４０２）は、１つ以上のストレージシステムの第１の集合の各ストレージシステムが、１つ以上のストレージシステムの第１の集合の他のすべてのシステムとステータスメッセージを交換することによって実施されてよい。別の例では、所与のストレージシステムは、通信障害が検出された１つ以上のストレージシステムの例外はあるが、互いと一貫性のある通信をするストレージシステムがすべて同期中リストのそれらのストレージシステムであると判断してよい。

【０５４５】

【０６０３】 １つ以上のストレージシステムの第１の集合によって、１つ以上のストレージシステムの第２の集合の間でのストレージシステムとの通信の欠如を判断すること（４４０４）は、図４３に関して上述されたように、同様に通信障害を検出すること（４３０２）によって実施されてよい。

【０５４６】

【０６０４】 １つ以上のストレージシステムの第１の集合によって、１つ以上のストレージシステムの第２の集合も定足数を形成できないと判断すること（４４０６）は、ストレージシステムのある集合が、ストレージシステムの別の集合が定足数を形成できることを確認し得るかどうかに関して上述されたように実施されてよい。

【０５４７】

【０６０５】 追加の説明のために、図４５は、本開示のいくつかの実施形態に従ってデータセットを同期複製するストレージシステムの間のアクティブメンバーシップを決定するための例の方法を示すフローチャートを説明する。図４５に示される例の方法は、ストレージシステム（４３００Ａ～４３００Ｎ）のうちの特定のストレージシステム（４３００Ａ）によって、ストレージシステム（４３００Ｂ～４３００Ｎ）の１つ以上との通信障害を検出すること（４３０２）であって、特定のストレージシステム（４３００Ａ）が、仲介ターゲットから仲介を要求するように構成される、検出することと、１つ以上のストレージシステム（４３００Ｂ～４３００Ｎ）の少なくとも１つが、通信障害に応じて、例えば仲介サービス（４３０１）等の仲介ターゲットから仲介を要求するように構成されると判

断すること（４３０４）と、１つ以上のストレージシステムの第１の集合によって、１つ以上のストレージシステムの第１の集合のストレージシステムの間に一貫性のある通信があると判断すること（４４０２）と、１つ以上のストレージシステムの第１の集合によって、１つ以上のストレージシステムの第２の集合の間でのストレージシステムとの通信の欠如を判断すること（４４０４）と、１つ以上のストレージシステムの第１の集合によって、１つ以上のストレージシステムの第２の集合が定足数を形成するための十分な数又は票をもたないであろうと判断すること（４４０６）と、ストレージシステム（４３００Ｂ～４３００Ｎ）の１つ以上が仲介ターゲットから仲介を要求するように構成されると判断すること（４３０４）に応じて、仲介ターゲットから仲介を要求するかどうかを判断すること（４３０６）も含むので、図４５に示される例の方法は図４４に示される例の方法に類似している。

10

【０５４８】

【０６０６】 しかしながら、図４５に示される例の方法は、１つ以上のストレージシステムの第１の集合によって、１つ以上のストレージシステムの第２の集合の間でのストレージシステムとの通信の欠如を判断すること（４４０４）が、１つ以上のストレージシステムの第１の集合が、データセットを同期複製しているストレージシステムのための票の正確に半分を含んだ、１つ以上のストレージシステムの第１の集合に従って、定足数を形成できないと判断すること（４５０２）をさらに含むことをさらに指定する。

【０５４９】

【０６０７】 １つ以上のストレージシステムの第１の集合が、第１のデータセットを同期複製しているストレージシステムのための票の正確に半分を含んだ、１つ以上のストレージシステムの第１の集合に従って、定足数を形成できないと判断すること（４５０２）は、ストレージシステムのある集合が、ストレージシステムの別の集合が定足数を形成できると確証し得るかどうかに関して上述されたように実施されてよく、１つ以上のストレージシステムの第１の集合が合計定足数票の正確に半分又は半分以上を制御する場合、次いで残りのストレージシステムが票の欠如に対して定足数を確立することは可能ではないであろう。

20

【０５５０】

【０６０８】 追加の説明のために、図４６は、本開示のいくつかの実施形態に従ってデータセットを同期複製するストレージシステムの間でのアクティブメンバーシップを決定する例の方法を示すフローチャートを説明する。図４６に示される例の方法は、ストレージシステム（４３００Ａ～４３００Ｎ）のうちの特定のストレージシステム（４３００Ａ）によって、ストレージシステム（４３００Ｂ～４３００Ｎ）の１つ以上との通信障害を検出すること（４３０２）であって、特定のストレージシステム（４３００Ａ）が、仲介ターゲットから仲介を要求するように構成される、検出することと、１つ以上のストレージシステム（４３００Ｂ～４３００Ｎ）の少なくとも１つが、通信障害に応じて、例えば仲介サービス（４３０１）等の仲介ターゲットから仲介を要求するように構成されると判断すること（４３０４）と、１つ以上のストレージシステムの第１の集合によって、１つ以上のストレージシステムの第１の集合のストレージシステムの間に一貫性のある通信があると判断すること（４４０２）と、１つ以上のストレージシステムの第１の集合によって、１つ以上のストレージシステムの第２の集合の間でのストレージシステムとの通信の欠如を判断すること（４４０４）と、１つ以上のストレージシステムの第１の集合によって、１つ以上のストレージシステムの第２の集合も定足数を形成できないであろうと判断すること（４４０６）と、ストレージシステム（４３００Ｂ～４３００Ｎ）の１つ以上が仲介ターゲットから仲介を要求するように構成されると判断すること（４３０４）に応じて、仲介ターゲットから仲介を要求するかどうかを判断すること（４３０６）も含むので、図４６に示される例の方法は図４４に示される例の方法に類似している。

30

40

【０５５１】

【０６０９】 しかしながら、図４６に示される例の方法は、１つ以上のストレージシステムの第１の集合によって、１つ以上のストレージシステムの第２の集合も定足数を形成でき

50

ないと判断すること（４４０６）が、１つ以上のストレージシステムの第２の集合が、データセットを同期複製するストレージシステムのための票の半分を含んだ１つ以上のストレージシステムの第２の集合に従って定足数を形成できないであろうと判断すること（４６０２）をさらに含むことをさらに指定する。

【０５５２】

【０６１０】 １つ以上のストレージシステムの第２の集合が、データセットを同期複製するストレージシステムのための票の半分を含んだ１つ以上のストレージシステムの第２の集合に従って定足数を形成できないであろうと判断すること（４６０２）は、ストレージシステムのある集合が、ストレージシステムの別の集合が定足数を形成できることを確認し得るかどうかに関して上述されたように実施されてよい。

10

【０５５３】

【０６１１】 追加の説明のために、図４７は、本開示のいくつかの実施形態に従ってデータセットを同期複製するストレージシステムの間のアクティブメンバーシップを決定するための例の方法を示すフローチャートを説明する。あまり詳細には示されないが、図４７に示されるストレージシステム（４３００Ａ～４３００Ｎ）は、図１Ａから図１Ｄ、図２Ａから図２Ｇ、図３Ａから図３Ｃ、又はその任意の組合せに関して上述されたストレージシステムに類似してよい。実際には、図４７に示されるストレージシステム（４３００Ａ～４３００Ｎ）は、上述されたストレージシステムと同じ構成要素、より少ない構成要素、追加の構成要素を含んでよい。

【０５５４】

20

【０６１２】 図４７に示される例の方法は、ストレージシステムのうちの第１のストレージシステムによって、ストレージシステムの第２のストレージシステムとの通信障害を検出すること（４７０２）を含む。ストレージシステムのうちの第１のストレージシステムによって、ストレージシステムの第２のストレージシステムとの通信障害を検出すること（４７０２）は、図４３に関して上述されたように、ストレージシステム（４３００Ａ～４３００Ｎ）のうちの特定のストレージシステム（４３００Ａ）によって、ストレージシステム（４３００Ｂ～４３００Ｎ）の１つ以上との通信障害を検出すること（４３０２）と同様に実施されてよい。

【０５５５】

【０６１３】 また、図４７に示される例の方法は、仲介を勝ち取ることにより、性能基準が、第２のストレージシステムがＩ／Ｏ要求にサービスを提供し続けることに比してより良く満たされるように、第１のストレージシステムと通信しているストレージシステムが、データセット（４３５２）に向けられたＩ／Ｏ要求にサービスを提供し続けることが可能になるであろうことを示す仲介分析（４７５２）を生成すること（４７０４）も含む。仲介分析（４７５２）を生成すること（４７０４）は、ストレージシステム（４３００Ａ）が、第１のストレージシステム、及び第１のストレージシステムと通信している１つ以上のストレージシステムの１つ以上の性能特徴を、第１のストレージシステムと通信していない１つ以上のストレージシステムの１つ以上の対応する性能特徴に対して比較して、ストレージシステムのどの集合がより効果的に、効率的に、又は確実にＩ／Ｏ要求を処理するであろうかを判断することによって実施されてよい。例えば、ストレージシステムは、

30

40

Ｉ／Ｏ要求を処理することに関する性能を示す複数の測定基準を追跡してよく、測定基準は、他の要因の中でもプロセッサ速度、ネットワーク待ち時間、読取り待ち時間、書込み待ち時間に基づいて影響を受ける場合がある。

【０５５６】

【０６１４】 他の例では、仲介分析は、通信していないストレージシステムで定足数を確立し得るかどうかの判断を反映してよく、仲介分析は、ストレージシステム（４３００Ａ）が、図４３に関して上述されたように、仲介又は定足数プロトコルに従事することを示す。

【０５５７】

【０６１５】 また、図４７に示される例の方法は、仲介分析に従って、仲介ターゲット（４３０１）から仲介を要求すること（４７０６）も含む。仲介分析に従って、仲介ターゲッ

50

ト（４３０１）から仲介を要求すること（４７０６）は、要求する仲介に関して上述されたように、及び出願参照番号第１４／７０３，５５９号に関してさらに説明されるように、実施されてよい。

【０５５８】

【０６１６】 追加の説明のために、図４８は、本開示のいくつかの実施形態に従って、データセットを同期複製するストレージシステムの間でメタデータを同期するための例の方法を示すフローチャートを説明する。あまり詳細には示されないが、図４８に示されるストレージシステム（４８００Ａ）は、図１Ａから図１Ｄ、図２Ａから図２Ｇ、図３Ａから図３Ｃ、又はその任意の組合せに関して上述されたストレージシステムに類似してよい。実際には、図４８に示されるストレージシステム（４８００Ａ）は、上述されたストレージシステムと同じ構成要素、より少ない構成要素、追加の構成要素を含んでよい。

10

【０５５９】

【０６１７】 これらの例では、ポッドメンバーシップは、ストレージシステムのリストを使用し、定義されてよく、そのリストの部分集合は、そのポッドのために同期される、つまり同期中となると推測されてよい。一部の場合では、リストの部分集合は、ポッドのためのストレージシステムのあらゆる１つを含んでよく、リストは、すべてのストレージシステムにとって共通であり、ポッドメンバーシップの変更に応えて適用される１つ以上の一貫性プロトコルの使用によりポッド全体で一貫性をもって維持されるメタデータと見なされてよい。『ポッド』は、該用語がここで及び本願の残りを通して使用されるように、データセット、管理オブジェクト及び管理動作の集合、データセットを修正する又は読み取るためのアクセス動作の集合、及び複数のストレージシステムを表す管理エンティティとして実施されてよい。係る管理動作は、データセットを読み取る又は修正するためのアクセス動作が、ストレージシステムのいずれかを通して同等に動作する、ストレージシステムのいずれかを通して同等に管理オブジェクトを修正する又は問い合わせしてよい。各ストレージシステムは、ストレージシステムによる使用のために記憶され、宣伝されたデータセットの適切な部分集合としてデータセットの別々のコピーを記憶してよく、任意の１つのストレージシステムを通して実行され、完了された管理オブジェクト又はデータセットを修正する動作は、ポッドを問い合わせるために後続の管理オブジェクトで、又はデータセットを読み取るために後続のアクセス動作で反映される。『ポッド』に関する追加の詳細は、参照により本明細書に援用される以前に出願された仮特許出願第６２／５１８，０７１号で見つけられてよい。

20

30

【０５６０】

【０６１８】 ストレージシステムは、ポッドがポッドと関連付けられたデータセットの最後の読み取られたコピーのための同一のアイドルコンテンツを有する少なくとも回復の中にある場合に、ポッドのために同期中と見なされてよい。アイドルコンテンツは、任意の進行中の修正が新しい修正の処理なしに完了した後のコンテンツである。一部の場合では、これは「クラッシュ回復可能な」一貫性と呼ばれる場合がある。ポッドの回復は、ポッド内で同期したストレージシステムに同時更新を適用する際の差異を調整するプロセスと見なされてよい。回復は、ポッドの多様なメンバーに対して要求されたが、無事に完了したとしていずれの要求者にも絶対に信号で知らされなかった同時修正の完了でストレージシステム間のあらゆる不一致を解決する場合がある。

40

【０５６１】

【０６１９】 ポッドのためのストレージシステムのリストの使用を所与として、ポッドメンバーとして一覧されるが、ポッドのために同期中として一覧されないストレージシステムは、ポッドからデタッチされていると見なされてよい。逆に、ポッドのためのストレージシステムのリストを使用し、ポッドメンバーとして一覧され、ポッドのためにデータを積極的に提供するために同期中及び現在利用可能としても一覧されるストレージシステムは、ポッドのためにオンラインであると見なされてよい。さらに、ポッドの各ストレージシステムは、それが最後に知ったどのストレージシステムが同期中であったのか、及びそれが最後に知ったどのストレージシステムが、ポッドメンバーの完全な集合を含んでいたの

50

かを含んだ、メンバーシップリストの独自のコピーを有してよい。

【0562】

【0620】 この例では、ポッドのためにオンラインであるために、所与のストレージシステムのためのメンバーシップリストは、所与のストレージシステムがポッドのために同期中であることを示し—所与のストレージシステムは、同期中であるとして示されるメンバーシップリスト内のすべての他のストレージシステムと通信できる。ストレージシステムが、それが同期中であり、同期中として示されるメンバーシップ内の他のすべてのストレージシステムと通信していることを確認できない場合、次いでストレージシステムは、ストレージシステムが、それが同期中であり、同期中として示されるメンバーシップ内の他のすべてのストレージシステムと通信中であることを確認できるまで、新しい入信I/Oコマンド又はポッドに向けられる要求を処理することを停止する。一部の例では、ストレージシステムが、それが同期中であり、同期中として示されるメンバーシップリスト内の他のすべてのストレージシステムと通信していることを確認できない場合、次いで新しいI/Oコマンド又は要求の処理を停止することの代わりに、ストレージシステムはI/Oコマンド又は要求の処理を停止する代わりに、ストレージシステムは、エラー又は例外でI/Oコマンド又は要求を完了する。I/Oコマンド又は要求は、異なるネットワークプロトコルを使用する他のタイプの要求の中でSCSI要求であってよい。一例として、第1のストレージシステムは、メンバーシップリストの中の第2のストレージシステムが、1つ以上の基準に基づいてデタッチされるべきかを判断してよく、第1のストレージシステムが第2のストレージシステムをデタッチした結果は、第1のストレージシステムが、少なくとも第1のストレージシステムが、メンバーシップリストから第2のストレージシステムを削除した後に、メンバーシップリストに留まるストレージシステムのすべてと現在同期中であるため、I/Oコマンドを受け取り、処理し続けることである。しかしながら、他の危険の中でも共存できないデータセット、データセット破損、又はアプリケーション破損につながる「スプリットブレイン」状況を回避するために、第2のストレージシステムは、第2のストレージシステムが—第1のストレージシステムに加えて—ポッドのためのデータセットに向けられたI/Oコマンドを受け取り、処理し続けるように、第1のストレージシステムをデタッチすることから妨げられなければならない。言い換えると、ポッド内の2つの異なるストレージシステムが、それらが無事に互いをデタッチしたと考える場合、次いでスプリットブレイン状況は結果として起こる場合がある。

【0563】

【0621】 同期中として示されるメンバーシップリスト内の別のストレージシステムと通信していないときに進む方法を決定する状況は、ストレージシステムが正常に動作中であり、1つ以上の失われた通信に気づく間に発生する場合がある、ストレージシステムが以前の障害から回復している間に発生する場合がある、ストレージシステムが、どのような理由であれ、ストレージシステムコントローラのある集合からストレージシステムコントローラの別の集合に動作を切り替えているときに発生する場合がある、ストレージシステムの起動中又はネットワークインタフェースが接続若しくは有効化されるときに発生する場合がある、あるいはこれらの及び他の種類のイベントの任意の組合せの間若しくは後に発生する場合がある。言い換えると、ポッドと関連付けられるストレージシステムがメンバーシップリストのすべての既知のデタッチされていないメンバーと通信できないときはいつでも、ストレージシステムは、通信が確立され得るまで、例えばなんらかの所定の時間量、待機する、若しくはオフラインになり、おそらく待機し続けるかのどちらかである場合がある、又はストレージシステムは、スプリットブレイン状況を被る危険なしに通信していないストレージシステムをデタッチすることが安全であると判断する場合がある。さらに、安全なデタッチが十分に迅速に起こる場合、ストレージシステムは、短い遅延にすぎず、ほとんど又は少しも障害のない要求でポッドのために継続的にオンラインのままとなつてよい、又はいくつかの要求はより低いレベルの要求者側動作処理を通して回復できる「ビジー」又は「再び試す」を生じさせてよく、アプリケーション又は他の高レベルの動作に対して悪影響を及ぼさない。

【0564】

【0622】 いくつかの状況では、ポッド内の所与のストレージシステムは、それが、最初に同期中であるポッドに加えられた後、ポッド内の他のストレージシステムに関して古い、又は異なって構成されていると判断する場合があります、所与のストレージシステムは、ポッド内の既存のストレージシステムが、ソフトウェア、ファームウェア、ハードウェア、又はソフトウェア、ファームウェア、若しくは所与のストレージシステムよりも新しい、つまり異なるハードウェアの組合せで構成されてよい。別の例として、所与のストレージシステムは、所与のストレージシステムが別のストレージシステムに再接続するに依て、古い、又は異なって構成されると判断し、他のストレージシステムが所与のストレージシステムをデタッチされたとマークしたと判断する—この場合、所与のストレージシステムは、それが、ポッドのために同期中であるストレージシステムのなんらかの他の集合に接続するまで待機してよい。

10

【0565】

【0623】 これらの例では、ストレージシステムが、ポッドから又は同期中メンバーシップリストから加えられる、又は削除される方法が、一過性の不一致を回避し得るかどうかを判断してよい。例えば、一過性の不一致は、各ストレージシステムがメンバーシップリストのそれぞれのコピーを有しているため、及びポッドの中の2つ以上の独立したストレージシステムが、異なるときにそのそれぞれのメンバーシップリストを更新—少なくとも正確に同時以外のときにそのそれぞれのメンバーシップリストを少なくとも更新—してよいため、及びおそらく他のメンバーシップリストと一貫性がないメンバーシップリストのローカルコピーが、所与のストレージシステムが利用可能にしてよいすべてのメンバーシップ情報であってよいため、発生する場合がある。一例として、第1のストレージシステムがポッドのために同期中であり、第2のストレージシステムが加えられる場合、次いで第2のストレージシステムが、第1のストレージシステムと第2のストレージシステムの両方をそのそれぞれのメンバーシップリストで同期中として一覧するために更新される場合—第1のストレージシステムが、第1のストレージシステム及び第2のストレージシステムをそのそれぞれのメンバーシップリストで同期中として一覧する前に一次いで障害が発生し、第1のストレージシステムと第2のストレージシステムの両方の再起動を引き起こす場合、第2のストレージシステムは起動し、第1のストレージシステムに接続するために待機する場合がある。一方、第1のストレージシステムは、それが第2のストレージシステムを待機しなければならない又は待機できるであろうと認識していない場合がある。この例を続けると、第2のストレージシステムは、次いで第1のストレージシステムをデタッチするプロセスを経ることによって、第1のストレージシステムと接続することができないことに依る場合、次いで第2のストレージシステムは、第1のストレージシステムが認識していないプロセスを完了するのに成功する場合があります、結果的にスプリットブレイン状況を生じさせる。

20

30

【0566】

【0624】 上記例に説明される状況を妨げるための例の技術として、ポッド内のストレージシステムは、個々のストレージシステムが、それらが通信していない場合、それらがデタッチプロセスを経ることを選ぶ可能性があるかどうかに関して不一致とならない方針に従ってよい。個々のストレージシステムが不一致にならないことを確実にする例の技術は、新しいストレージシステムをポッドのための同期中メンバーシップリストに加えるとき、新しいストレージシステムが最初に、新しいストレージシステムがデタッチされたメンバーであることを記憶することを確実にすることである。この時点で、既存の同期中ストレージシステムは、新しいストレージシステムが、新しいストレージシステムが同期中のポッドメンバーであることをローカルに記憶する前に、新しいストレージシステムが同期中のポッドメンバーである表示をローカルに記憶してよい。結果的に、新しいストレージシステムがそれ自体のために同期中ステータスを記憶する前に、リブート又はネットワーク障害又は機能停止の集合がある場合、次いで元のストレージシステム—新しいストレージシステムを加える試みの前に、ポッドの同期中メンバーであるストレージシステム—は

40

50

、非通信のために新しいストレージシステムをデタッチしてよいが、新しいストレージシステムは待機する。

【0567】

【0625】 この例を続けると、メンバーシップの係る変更の逆バージョンが、通信中のストレージシステムをポッドから削除するために必要とされる可能性がある—最初に削除又はデタッチされているストレージシステムが、同期中ではないという表示をローカルに記憶し、ポッドに留まるストレージシステムは、続いて削除されているシステムがもはや同期中ではない旨の表示を記憶する。この時点で、ポッドに留まるストレージシステムと、削除されているストレージシステムの両方は、削除されているストレージシステムをそのそれぞれのメンバーシップリストから削除する。この例では、実施態様に応じて、中間の持続されるデタッチ状態が必要ではない場合がある。

10

【0568】

【0626】 さらに、メンバーシップリストのローカルコピーで注意が必要とされるかどうかは、互いをモニタするため、又はそのメンバーシップを確認するためのモデルストレージシステムの使用次第である場合がある。例えば、両方にコンセンサスモデルが使用される場合、又は外部システム—又は外部分散型システム若しくはクラスタ化システム—が、ポップメンバーシップを記憶し、確認するために使用される場合、次いでローカルに記憶されたメンバーシップリストでの不一致は、重要ではなくなる場合がある。

【0569】

【0627】 自然発生的なメンバーシップ変更を解決するための一部の例のモデルは、定足数、外部ポッドメンバーシップマネージャ、又は既知のリソースに対する競争の使用を含む。これらの例のモデルは、通信故障、ポッド内の1つ以上のストレージシステムが故障すること、又はポッド内の対にされたストレージシステムと通信できない、ストレージシステムが起動すること（又は二次コントローラに障害迂回すること）に応じて使用されてよい。ポッドメンバーシップの変更をトリガする場合があるこれらのイベントを所与として、異なるメンバーシップモデルは、ポッド内のストレージシステムが、1つ以上の対にされたストレージシステムを安全であるやり方でデタッチすることをどのようにして決定するのか、及び1つ以上のストレージシステムをデタッチするときにやり通す方法を定義するために異なる機構を使用してよい。

20

【0570】

【0628】 一部の例では、メンバーシップ変更に関する意見の一致に到達する際に使用される複数のメンバーシップリストがある場合がある。例えば、ストレージシステムの所与のグループの場合、各ストレージシステムが同期中リスト上又は同期外リスト上にある場合があり、各ストレージシステムは、同期中リスト及び同期外リストのそれぞれのローカルコピーを記憶する。この例では、ストレージシステムのグループはストレージシステム{A, B}であってよく、最初にポッドはストレージシステムAを含んでよく、ポッドはストレージシステムAからストレージシステムBに伸ばされる、つまり拡大される。ポッドのこの伸展は、ポッドのためのストレージシステムのメンバーシップを拡大することに同等であり、ストレージシステムA及びストレージシステムBが接続されることを確実にすることによって開始してよい。ストレージシステムA及びストレージシステムBが接続されることを確実にすることは、伸展動作に先行する構成ステップであってよい—しかしながら、ストレージシステムAとストレージシステムB間の単なる接続性はポッドを広げるのではなく、むしろストレージシステムAとストレージシステムB間の接続性はポッドを広げることを可能にする。この例では、ストレージシステムAは—例えば、ボリューム、ポッド、及びストレージシステムを管理するための管理コンソールから—ポッド、つまりストレージシステムAでのポッドの特定のボリュームをストレージシステムBに広げることを示すコマンドを受け取ってよい。ストレージシステムAとストレージシステムBの間の接続性を所与として、初期状態は、ストレージシステムAが{A}を示す同期中リスト及び{B}を示す同期外リストを記憶するとして説明されてよく、エポック識別子はnに等しく、メンバーシップシーケンスはmに等しく、ストレージシステムBは、同期中リ

30

40

50

ストと同期外リストの両方のために空のリストを記憶する。ストレージシステム A が伸展コマンドを受け取ることに応えて、ストレージシステム A はストレージシステム B に、ポッド識別子、エポック識別子 n によって識別されるセッションを示すメッセージを送信してよく、それに応えて、ストレージシステム B はストレージシステム A に通信し返す。さらに、ストレージシステム A 及び B の間の構成レベルハートビートは、ストレージシステムのための同期中リスト及び同期外リストをストレージシステム B に配布してよく、それに応えてストレージシステム B は、それが同期中のメンバーではないと判断し、ストレージシステム A との再同期動作を開始してよく、このことがストレージシステム A とストレージシステム B の両方全体でポッドを同期させる。さらに、再同期に応えて、ストレージシステム A は、更新された同期中リスト $\{A, B\}$ をストレージシステム B に書き込み、次いでストレージシステム B が応答するのを待機してよい。この時点で、ストレージシステム A は、同期中の動作に関してストレージシステム B との通信を開始する準備が完了している—しかしながら、ストレージシステム B は、ストレージシステム B が更新された同期中リスト $\{A, B\}$ をポッドの同期中メンバーとして受け取るまで、係る通信に関与しない。例えば、ストレージシステム A はストレージシステム B とのクロック交換動作を開始することによって通信を始めてよいが、ストレージシステム B は、ストレージシステム B が未決の同期中リスト $\{A, B\}$ を受け取るまでクロック交換動作を開始しない場合がある。クロック交換は、その全体として本明細書に含まれる出願参照番号第 62/470,172 号及び第 62/518,071 号の中により詳細に説明されている。

【0571】

[0629] この例を続けると、伸展解除する (`unstretch`)、つまりストレージシステムをポッド内のメンバーシップから削除するために、メンバーストレージシステムは以下のステップを講じてよい。例えば、ポッドメンバーシップが現在、両方のストレージシステム A 及びストレージシステム B が $\{A, B\}$ の同じ同期中リスト、並びに $\{\}$ の同期外リスト、 n の現在のエポック、及び m の現在のメンバーシップシーケンスを有する $\{A, B\}$ である場合—この状況で、ストレージシステム A は、ポッドを伸展解除してストレージシステム B を除外する要求を受け取る場合がある。伸展解除要求に応えて、ストレージシステム A は、ストレージシステム B に、コミットされたメンバーシップリストを示すメッセージを送信してよく、 $\{A, B\}$ の同期中リスト及び $\{\}$ の同期外リストを示し、 $\{A\}$ の同期中リスト及び $\{\}$ の同期外リスト及び n の現在のエポック及び $(m+1)$ のメンバーシップシーケンスを示す未決のメンバーシップリストを示す。ストレージシステム B は、ストレージシステム A からメッセージを受け取ることに応えて、メッセージの中に示される状態情報を適用し、ストレージシステム A に対して、状態変更が適用されたことを応答する。ストレージシステム A は、状態変更の肯定応答をストレージシステム B から受け取ることに応えて、そのローカル状態情報を更新して $\{A\}$ の同期中リスト及び $\{\}$ の同期外リストのためのコミットされたメンバーシップリスト、 $\{A\}$ の同期中リスト及び $\{\}$ の同期外リスト、及び $(n+1)$ のエポックのための未決メンバーシップリストを示し、ストレージシステム B は、次いでストレージシステム B との通信を停止する。ストレージシステム B は、失われたセッションを検出するが、 $\{A\}$ の同期中リストを有するため、ストレージシステム B は、ストレージシステム A からセッションを再確立することを要求し、ストレージシステム B がもはやポッドのメンバーではないことを示す応答を受け取る。

【0572】

[0630] メンバーシップモデルとして定足数を使用する例では、デタッチ動作を解決するための 1 つの技術は、メンバーシップのために過半数—つまり、定足数—モデルを使用することである。例えば、3 つのストレージシステムを所与として、2 つが通信している限り、通信中の 2 つは、通信していない第 3 のストレージシステムをデタッチすることに同意できる。しかしながら、第 3 のストレージシステムは、通信中の 2 つのストレージシステムのどちらかをデタッチすることを独力で選ぶことはできない。一部の 경우에는、ポッドの中のストレージシステム通信に一貫性がないとき混乱が生じる場合がある。この例

では、ストレージシステム {A, B, C} の場合、ストレージシステム B が両方のストレージシステム A とストレージシステム C と通信している場合があるのに対し、ストレージシステム A はストレージシステム B と通信している場合があるが、ストレージシステム A はストレージシステム C と通信していない場合がある。この状況では、両方のストレージシステム A とストレージシステム B は、ストレージシステム C をデタッチしてよい—又は両方のストレージシステム B とストレージシステム C は、ストレージシステム A をデタッチしてよい—が、メンバーシップを理解するためにはポッドメンバー間のより多くの通信が必要とされる場合がある。

【0573】

【0631】 この例を続行すると、定足数方針、つまり定足数プロトコルは、ストレージシステムを加える、又はストレージシステムをポッドから削除するためのこの状況を解決し得る。例えば、第4のストレージシステムがポッドに加えられる場合、次いでストレージシステムの過半数は3つのストレージシステムとなる。過半数に2つが必要とされる3つのストレージシステムから、過半数に3つが必要とされる4つのストレージシステムを有するポッドへの遷移は、ストレージシステムを同期中リストに慎重に加えるための上述されたモデルに類似する何かを必要とする場合がある。例えば、第4のストレージシステム、例えばストレージシステム D は、アタッチする状態で開始するが、まだアタッチされておらず、それは決して定足数を超える票を生じさせないであろう。そのストレージシステム D がアタッチする状態にあることを所与として、ストレージシステム A、ストレージシステム B、及びストレージシステム C は、それぞれストレージシステム D を認識するように更新され、3つのストレージシステムが任意の特定のストレージシステムをポッドからデタッチする過半数決定に達する新しい要件について更新されてよい。さらに、所与のストレージシステムをポッドから削除することは、ポッド内の他のストレージシステムを更新する前に所与のストレージシステムをデタッチする状態に同様に遷移してよい。一部の例では、定足数モデルでの問題は、共通構成が、正確に2つのストレージシステムを有するポッドである点であり、係る場合では、1つの解決策は、ポッドに対する定足数投票だけに参加するが、それ以外の場合、ポッドのためのデータセットを記憶しないネットワークの中にストレージシステムを加えることである。この場合、係る投票専用メンバーは、一般的に、一連の定足数投票を引き起こさないであろうが、同期中ストレージシステムとして構成されたポッド内のストレージシステムによって引き起こされる投票にのみ参加するであろう。

【0574】

【0632】 メンバーシップモデルとして外部ポッドメンバーシップ方法を使用する例では、1つの技術は、ポッドメンバーシップを扱うためにストレージシステム自体の外部にある外部システムを使用し、メンバーシップ遷移を管理することを含む。例えば、ポッドのメンバーになるためには、将来のストレージシステムは、ポッドメンバーシップシステムに連絡して、ポッドに対するメンバーシップを要求し、将来のストレージシステムがポッドのために同期中であることを検証するように構成される。このモデルでは、ポッドのためにオンライン、つまり同期中である任意のストレージシステムは、ポッドメンバーシップシステムとの通信に留まるべきであり、ポッドメンバーシップシステムとの通信が失われる場合、待機する、つまりオフラインになるべきである。この例では、ポッドメンバーシップシステムは、例えば、Oracle (商標) RAC、Linux HA、VERITAS (商標) Cluster Server、IBM (商標) HACMP、又は他等、多様なクラスタツールを使用する高可用性クラスタとして実装されてよい。他の例では、ポッドメンバーシップシステムは、例えば Etc d (商標) 若しくは Zookeeper (商標) 等の分散型構成ツール、又は例えば Amazon による DynamoDB (商標) 等の信頼性のある分散型データベースを使用し、実装されてよい。さらに、他の例では、ポッドメンバーシップは、例えば RAFT 又は PAXOS 等の分散型コンセンサスアルゴリズムを使用し、決定されてよく、RAFT からの概念に基づいた実施態様は、メンバーシップのための RAFT ベースの内部アルゴリズムを含む場合もあれば、有効で最新の

10

20

30

40

50

メンバーシップを決定するため、及び最新のメンバーシップ情報の現在値を決定するための総合的な解決策の一部として使用され得るログ様式の更新一貫性のための R A F T に着想を得た (R A F T - i n s p i r e d) アルゴリズムを含む場合もある。

【 0 5 7 5 】

【 0633 】 メンバーシップモデルとして既知のリソースに対する競争、つまりレーシングプロトコルを使用する例では、ポッドのためのクラスタマネージャが、他者を除外してなんらかの方法でロックされ得るなんらかのリソースに対するアクセスを要求することにより、又はいくつかの係るリソースの過半数に対するアクセスを要求することによりメンバーシップ変更を解決することによって実施されてよい。例えば、1つの技術は、S C S I 予約又は S C S I 持続性予約等のリソース予約を使用して、1つ以上のネットワーク化された S C S I デバイスに対するロックを得ることである。この例では、これらのネットワーク化されたデバイスの構成された集合の過半数がストレージシステムによってロックできる場合、次いでそのストレージシステムは他のストレージシステムをデタッチしてよい。それ以外の場合、ストレージシステムは、他のストレージシステムをデタッチできないであろう。さらに、オンライン、つまり同期中に留まるために、ストレージシステムはリソースに対するこれらのロックを頻繁に再アサート若しくは試験する、又はリソースに対するこれらのロックをアサートしている、再アサートしている、又は試験しているなんらかの他のストレージシステムと通信中である必要がある場合がある。さらに、さまざまな方法でアサートされ、試験され得るネットワーク化された計算リソースは、同様に使用されてよい。

【 0 5 7 6 】

【 0634 】 この例を続けると、ポッドのすべてのストレージシステムメンバーによる引き延ばされた機能停止が、一方のストレージシステムがメンバーポッドとして再開することを可能にし、他のストレージシステムメンバーをデタッチすることを可能にしつつ、適切に処理できることを確実にするためには、上述されたようなネットワークリソースは、なんらかの他のストレージシステムが、再開するストレージシステムポッドメンバーを以前にデタッチしたことがなかったことを検証するために使用されてよい持続性のあるプロパティを有さなければならない。しかしながら、サービスが、ステータス情報又は他のメタデータを持続的に記憶する能力なしにリソース予約だけを提供する場合、次いでリソース要約サービスは、特定のストレージシステムがアクセスを得た後に問い合わせられ、書き込まれる可能性がある、例えばサードパーティのデータベース又はクラウドストレージ等のなんらかの外部に記憶されたデータへのアクセスを得るために使用されてよい書き込まれるデータは、デタッチされたストレージシステムが、それがデタッチされたと判断するために問い合わせる場合がある情報を記録する場合がある。

【 0 5 7 7 】

【 0635 】 一部の例では、レーシングプロトコルは、あるストレージシステムがポッドから別のストレージシステムをデタッチする権限を有するかどうかを決定するサービスである仲介サービスを使用し、実装されてよい。仲介サービスの例の実施態様は、その全体として本明細書に援用される出願番号第 1 5 / 7 0 3 , 5 5 9 号の中にさらに説明される。

【 0 5 7 8 】

【 0636 】 別の例では、ポッドが3つ以上のストレージシステム全体で広げられるとき有用である場合がある機構の組合せが使用されてよい。一例では、優先規則はメタデータと結合されてよい。トップオブブラック例では、データセンタ又はキャンパス内のより大型の中央ストレージシステムはそれ自体、第2の場所の大型ストレージシステムに同期複製される可能性がある。その場合、トップオブブラックストレージシステムは、絶対に単独で再開し得ず、2つの場所のより大型の中央ストレージシステムのいずれかに優先権を与えてよい。その場合の2つのより大型のストレージシステムは、互いの間で仲介するように構成される可能性があり、オンラインに留まる2つのより大型のストレージシステムのどちらかに接続できる任意のより小型のストレージシステムはそのポッドにサービスを提供し続ける場合があり、2つの大型のストレージシステムのどちらにも接続できない（又は、

ポッドのためにオフラインである一方にしか接続できない) 任意のより小型のストレージシステムは、ポッドにサービスを提供するのを停止する場合がある。さらに、優先モデルは定足数ベースのモデルと結合されてもよい。例えば3つの場所の3つの大型ストレージシステムは、互いの間で定足数モデルを使用する可能性があり、より小型の衛星ストレージシステム又はトップオブブラックストレージシステムはいかなる票も欠き、それらが、オンラインであるより大型の同期中のストレージシステムの一方に接続できる場合のみ機能する。

【0579】

[0637] 機構を結合する別の例では、仲介は定足数モデルと結合されてよい。例えば、通常2つのストレージシステムが通信していない第3のストレージシステムを安全にデタッチできることを確実にするために、互いの間で投票する3つのストレージシステムがあってよい。一方、1つのストレージシステムは、2つの他のストレージシステムを独力で絶対にデタッチすることはできない。しかしながら、2つのストレージシステムが第3のストレージシステムを無事にデタッチした後に、構成はいま、それらが同期中であることに同意し、第3のストレージシステムがデタッチされる旨の事実の同意する2つのストレージシステムに少なくなる。その場合、残っている2つのストレージシステムは、(例えばクラウドサービスを用いた) 仲介を使用して、追加のストレージシステム又はネットワークの障害を処理することに同意してよい。この仲介及び定足数の組合せは、さらに拡張され得る。例えば、4つのストレージシステム間で広げられるポッドでは、任意の3つが第4のストレージシステムをデタッチすることができるが、2つの同期中のストレージシステムが互いと通信しているが、他の2つのストレージシステムに通信していない場合、それらは両方とも同期中であると現在考え、次いでそれらは他の2つを安全にデタッチするために仲介を使用するであろう。5つのストレージシステムポッド構成でさえ、4つのストレージシステムが第5のストレージシステムをデタッチするために投票する場合、次いで残りの4つは、それらが等しい2つの半分に分割される場合仲介を使用でき、ポッドが2つのストレージシステムに少なくなると、それらは連続する障害を解決するために仲介を使用できる。5つから3つは、3つの間で定足数を使用し、次いで2つへの低下を可能にする可能性があり、残った2つのストレージシステムは、追加の故障がある場合、再び仲介を使用する。この一般的なマルチモード定足数仲介機構は、対称的なストレージシステムの間の定足数も、独自の仲介も処理できない追加の数の状況の処理できる。この組合せは、障害のある又はときおり到達不能なメディエータを確実に使用できる(又は、カスタマが彼らを完全に信用していない場合があるクラウドメディエータの場合) 事例の数を増やし得る。さらに、この組合せは、仲介が単独である結果、第1のストレージシステムが、単に第1のストレージシステムにしか影響を及ぼさないネットワーク障害時に、無事に第2のストレージシステム及び第3のストレージシステムをデタッチする可能性がある、3つのストレージシステムポッドの場合をより良く処理する。また、この組合せは、3つから2つ、次いで1つへの例で説明されるように、一度に1つのストレージシステムに影響を及ぼす一連の障害をより良く処理し得る。これらの組合せは、同期中であること及びデタッチ動作が特定の状態を生じさせるために機能する一言い換えると、デタッチ済みから同期中に移動することがプロセスであり、一連の定足数/メディエータ関係での各段階は、あらゆる点で、すべてのオンライン/同期中ストレージシステムがポッドのための現在の持続性のある状態に同意していることを確実にするため、システムはステートフルである。これは、単に通信中のクラスタノードの過半数を再び有することが、動作を再開するために十分であると予想される、なんらかの他のクラスタモデルにおいてとは異なるしかしながら、優先モデルは、依然として加えられてよく、衛星ストレージシステム又はトップオブブラックストレージシステムは、仲介にも定足数にも絶対に関与せず、それらが、仲介又は定足数に参加するオンラインストレージシステムに接続できる場合にだけポッドにサービスを提供する。

【0580】

[0638] 図48に示される例の方法は、メンバーシップイベントが、データセット(4

10

20

30

40

50

858)を同期複製するストレージシステム(4800A~4800B)の集合に対するメンバーシップの変更に相当すると判断すること(4802)を含む。メンバーシップイベントが、データセット(4858)を同期複製するストレージシステム(4800A~4800B)の集合に対するメンバーシップの変更に相当すると判断すること(4802)は、異なる技術を使用し、実施されてよい。一例として、ストレージシステム(4800A)は、ポッド(4854)が新しいストレージシステム(4800N)を含むように広げられることを示す、又はポッド(4854)が既存のストレージシステム(4800N)を除外するために伸展解除されることを示すI/Oコマンドを受け取ってよい。別の例として、ストレージシステム(4800A)は、ストレージシステムの集合の特定のストレージシステム(4800N)で、通信が失われたこと、又は通信が指定の閾値を超えて信頼できなくなった又は非効率になったことを検出し、判断してよい。

10

【0581】

【0639】 ストレージシステム(4800A~4800N)の集合のうちのストレージシステム(4800A)で、ポッド(4854)のためのI/Oコマンド、又はデータセット(4858)に向けられたI/O動作(4852)を受け取ることは、例えばストレージエリアネットワーク(158)、インターネット、又はホストコンピュータ(4851)がその全体でストレージシステム(4800A)と通信し得る任意のコンピュータネットワーク等のネットワークを介してパケット又はデータを移動させるための1つ以上の通信プロトコルを使用することによって実施されてよい。一部の場合では、ポッド(4854)のためのI/Oコマンド又はデータセット(4858)に向けられたI/O動作(4852)を受け取ることは、ポッド(4800A)のストレージシステム(4800A~4800N)間の通信相互接続(173)―又は、ストレージシステム(4800A)にとって内部であるなんらかの他の通信チャネル―を使用することによって実施されてよい―I/Oコマンド又は動作は、ストレージシステムコンピューティングリソースに常駐している又は実行中であるアプリケーション又はプロセスから受け取られる。さらに、常駐又は遠隔であるアプリケーションは、ストレージシステム(4800A~4800N)が同期中であり、オンラインであることに依存している機能性を提供し得るファイルシステム、データオブジェクト、データベースを実装する際にストレージシステム(4800A~4800N)を使用してよい―及びこれらのプロトコル又はアプリケーションのいずれかは、同期複製され、対称的にアクセス可能な基本的なストレージ実施態様に作用する分散された実施態様であってよい。この例では、ストレージシステム(4800A)は、例えばSCSIポート等のネットワークポートで受け取られるI/Oコマンド又はI/O動作(4852)を受け取ってよく、I/O動作(4852)は、ポッド内のストレージシステム(4800A~4800N)全体で同期複製されているデータセット(4858)の一部である記憶場所に向けられる書込みコマンドである。

20

30

【0582】

【0640】 また、図48に示される例の方法は、データセット(4858)を同期複製するためのストレージシステムの新しい集合を決定するためにメンバーシップイベントに従って、1つ以上のメンバーシッププロトコルを適用すること(4804)も含む。データセット(4858)を同期複製するためのストレージシステムの新しい集合を決定するためにメンバーシップイベントに従って、1つ以上のメンバーシッププロトコルを適用すること(4804)は、定足数プロトコル、外部ポッドメンバーシップマネージャプロトコル、又はレーシングプロトコルの任意の1つ以上を使用し、上述されたように実施されてよい。

40

【0583】

【0641】 また、図48に示される例の方法は、データセット(4858)に向けられた1つ以上のI/O動作(4852)のために、1つ以上のストレージシステムの新しい集合によって同期複製されたデータセット(4858)に1つ以上のI/O動作(4852)を適用すること(4806)も含む。ストレージシステムの新しい集合によって同期複製されたデータセット(4858)に1つ以上のI/O動作(4852)を適用すること

50

(4806)は、その全体として本明細書に含まれ、データセットに対する任意の変更が、ポッドのすべての同期中のストレージシステムメンバー全体で同期複製されるように、I/O動作を受け取り、処理することを説明する出願参照番号第62/470,172号、及び第62/518,071号の中に説明されるように実施されてよい。

【0584】

【0642】 読者は、上述された方法が、上述されたストレージシステムの任意の組合せによって実施され得ることを理解する。さらに、上述されたストレージシステムのいずれかは、例えばAmazon(商標)Web Services(『AWS』)、Google(商標)Cloud Platform、Microsoft(商標)Azure、又は他等のクラウドサービスプロバイダによって提供されるストレージと対になってよい。したがって、係る例では、特定のポッドのメンバーは、クラウドサービスプロバイダによって提供されるストレージから成るストレージシステムだけではなく、上述されたストレージシステムの1つを含んでよい。同様に、特定のポッドのメンバーは、クラウドサービスプロバイダによって提供されるストレージから成るストレージシステムの論理的表現だけから成ってよい。例えば、ポッドの第1のメンバーは、第1のAWS可用性ゾーンのストレージから成るストレージシステムの論理的表現であってよい。一方、ポッドの第2のメンバーは、第2のAWS可用性ゾーンのストレージから成るストレージシステムの論理的表現であってよい。

10

【0585】

【0643】 クラウドサービスプロバイダによって提供されるストレージから成るストレージシステムにデータセット(又はバーチャルマシン等の他の管理オブジェクト)を同期複製し、本願に説明される他のすべての機能を実行する能力を助長するために、多様なストレージシステム機能を実施するソフトウェアモジュールは、クラウドサービスプロバイダによって提供される処理リソース上で実行されてよい。係るソフトウェアモジュールは、例えば、ブロックデバイス、Amazon(商標)Machine Image(『AMI』)インスタンス等のクラウドサービスプロバイダによってサポートされる1つ以上のバーチャルマシンで実行してよい。代わりに、係るソフトウェアモジュールは、ハードウェアに直接アクセスできるAmazon(商標)EC2ベアメタルインスタンス等のクラウドサービスプロバイダによって提供されるベアメタル環境で代わりに実行してよい。係る実施形態では、Amazon(商標)EC2ベアメタルインスタンスは、ストレージシステムを効果的に形成するために高密度フラッシュドライブと対にされてよい。どちらの実施態様でも、ソフトウェアモジュールは、理想的には、クラウドリソース上で、例えばvSAN(商標)等のVMware(商標)によって提供される仮想化ソフトウェア及びサービス等の他の従来のデータセンササービスと同一場所に配置されるであろう。読者は、他の多くの実施態様も可能であり、本開示の範囲内にあることを理解する。

20

30

【0586】

【0644】 読者は、ポッド内のデータセット又は他の管理オブジェクトがオンプレミスストレージシステムに保持され、ポッドが、そのリソースがクラウドサービスプロバイダによって提供されるストレージシステムを含むように広げられる状況では、データセット又は他の管理オブジェクトは、そのリソースが暗号化されたデータとしてクラウドサービスプロバイダによって提供されるストレージシステムに転送されてよいことを理解する。係るデータは、オンプレミスストレージシステムによって暗号化されてよく、これによりクラウドサービスプロバイダにより提供されるリソースに記憶されるデータは暗号化されるが、クラウドサービスプロバイダは暗号化鍵を有していない。このようにして、クラウドに記憶されたデータは、クラウドは暗号化鍵にアクセスできないのでより安全である場合がある。同様に、データが最初にオンプレミスストレージシステムに書き込まれるとき、ネットワーク暗号化を使用でき、暗号化されたデータはクラウドが続けて暗号化鍵にアクセスできないように、クラウドに転送できるであろう。

40

【0587】

【0645】 クラウドサービスプロバイダによって提供されるストレージから成るストレージ

50

ジシステムの使用により、災害復旧はサービスとして提供され得る。係る例では、データセット、作業負荷、他の管理オブジェクト等は、オンプレミスストレージシステム上に常駐してよく、そのリソースがクラウドサービスプロバイダによって提供されるストレージシステムに同期複製されてよい。災害がオンプレミスストレージシステムに発生する場合、そのリソースがクラウドサービスプロバイダによって提供されるストレージシステムは、データセットに向けられた要求の処理を引き継ぎ、データセットを別のストレージシステムに移行するのを支援する等であってよい。同様に、そのリソースがクラウドサービスプロバイダによって提供されるストレージシステムは、重度使用の期間中又はそれ以外の場合必要に応じて使用されてよいオンデマンドの二次ストレージシステムとしての機能を果たしてよい。読者は、本明細書に説明される機能の多くを開始するユーザインタフェース又は類似した機構が設計され得、これにより災害復旧をサービスとして可能にすることが、ただ1回マウスクリックを実行することほど簡略になることを理解する。

【0588】

【0646】 クラウドサービスプロバイダによって提供されるストレージから成るストレージシステムの使用により、高可用性もサービスとして提供され得る。係る例では、オンプレミスストレージシステムに常駐してよいデータセット、作業負荷、又は他の管理オブジェクトは、そのリソースがクラウドサービスプロバイダによって提供されるストレージシステムに同期複製されてよい。係る例では、例えばAWS Direct Connect等のクラウドに対する専用のネットワーク接続性のため、さまざまな場所からのAWSに対するサブミリ秒の待ち時間を達成できる。したがって、アプリケーションは、前もっての多大な出費なしに伸展クラスタモードで実行することができ、高可用性は、購入され、保守される等の複数の別個に位置するオンプレミスストレージシステムに対する必要性なしに達成され得る。読者は、本明細書に説明される機能の多くを開始するユーザインタフェース又は類似した機構が設計され得、これによりアプリケーションを有効化にすることは、ただ1回マウスクリックを実行することによってクラウドの中に拡大され得ることを理解する。

【0589】

【0647】 クラウドサービスプロバイダによって提供されるストレージから成るストレージシステムの使用により、システム回復もサービスとして提供され得る。係る例では、オンプレミスストレージシステムに常駐してよいデータベース、管理オブジェクト、及び他のエンティティの時点コピーは、そのリソースがクラウドサービスプロバイダによって提供されるストレージシステムに同期複製され得る。係る例では、ストレージシステムを特定の時点に戻して回復する必要性が生じる場合、そのリソースがクラウドサービスプロバイダによって提供されるストレージシステム上に含まれるデータセット及び他の管理オブジェクトの時点コピーは、ストレージシステムを回復するために使用され得る。

【0590】

【0648】 クラウドサービスプロバイダによって提供されるストレージから成るストレージシステムの使用により、オンプレミスストレージシステムに記憶されるデータは、多様なクラウドサービスによる使用のためにクラウドの中にネイティブにパイプで送られてよい(piped)。係る例では、データがオンプレミスストレージシステムに記憶されたようにそのネイティブフォーマットであるデータはクローン化され、多様なクラウドサービスのために使用できるフォーマットに変換されてよい。例えば、データがオンプレミスストレージシステムに記憶されたようにそのネイティブフォーマットであるデータはクローン化され、データ分析問合せがデータに対して実行され得るように、Amazon(商標)Redshiftによって使用されるフォーマットに変換されてよい。同様に、オンプレミスストレージシステムで記憶されていたようにそのネイティブフォーマットであるデータはクローン化され、Amazon(商標)DynamoDB、Amazon(商標)Aurora、又は他のなんらかのクラウドデータベースサービスによって使用されるフォーマットに変換される。係る変換はオンプレミスストレージシステムの外部で発生するため、オンプレミスストレージシステムの中のリソースは、I/O動作にサービスを提

供する際に使用するために保護され、保持されてよい。一方、必要に応じてスパンアップできるクラウドリソースは、オンプレミスストレージシステムがI/O動作の一次サービサーとして動作し、クラウドサービスプロバイダによって提供されるリソースから成るストレージシステムがバックアップストレージシステムのより多くとして動作する実施形態では、特に貴重である場合があるデータ変換を実行するために使用される。実際に、管理オブジェクトはストレージシステム全体で同期され得るため、オンプレミスストレージシステムが、最初に抽出・加工・書込み（『ETL』）パイプラインで必要とされるステップを実施することを担っていた実施形態では、係るパイプラインの構成要素はクラウドにエクスポートされ、クラウド環境で実行してよい。係る技術の使用により、データセットの時点コピー（つまり、スナップショット）を分析サービスに対する入力として使用することを含んだ、サービスとしての分析も提供されてよい。

10

【0591】

【0649】 読者は、アプリケーションが上述されたストレージシステムのいずれかで実行でき、いくつかの実施形態では、係るアプリケーションが一次コントローラ、二次コントローラで、又は同時に両方のコントローラでも実行できることを理解する。係るアプリケーションの例は、背景バッチデータベーススキャンを行うアプリケーション、実行時データの統計分析を行っているアプリケーション等を含む場合がある。

【0592】

【0650】 例の実施形態は、おもに完全機能コンピュータシステムとの関連で説明される。しかしながら、当業の読者は、本開示が、任意の適切なデータ処理システムとの使用のためにコンピュータ可読記憶媒体に配置されたコンピュータプログラム製品で実施され得ることを認識する。係るコンピュータ可読記憶媒体は、磁気媒体、光媒体、又は他の適切な媒体を含んだ、機械可読情報用の任意の記憶媒体であってよい。係る媒体の例は、当業者に思い浮かぶように、ハードドライブ内の磁気ディスク又はディスクレット、光ドライブ用コンパクトディスク、磁気テープ及び他を含む。当業者は、適切なプログラミング手段を有する任意のコンピュータシステムが、コンピュータプログラム製品で実施される方法のステップを実行できることをただちに認識する。また、当業者は、本明細書に説明される例の実施形態のいくつかは、コンピュータハードウェアにインストールされ、実行中のソフトウェアに向けられているが、それにも関わらず、ファームウェアとして又はハードウェアとして実施される代替実施形態が、十分に本開示の範囲内にあることも認識する。

20

【0593】

【0651】 実施形態は、システム、方法、及び／又はコンピュータプログラム製品を含む場合がある。コンピュータプログラム製品は、プロセッサに本開示の態様を実施させるためにその上にコンピュータ可読プログラム命令を有する1つのコンピュータ可読記憶媒体（又は複数の媒体）を含んでよい。

【0594】

【0652】 コンピュータ可読記憶媒体は、命令実行デバイスによって使用するための命令を保持し、記憶できる有形のデバイスである場合がある。コンピュータ可読記憶媒体は、例えば電子ストレージデバイス、磁気ストレージデバイス、光ストレージデバイス、電磁ストレージデバイス、半導体ストレージデバイス、又は上記の任意の適切な組合せであってよいが、これに限定されるものではない。コンピュータ可読記憶媒体のより具体的な例の包括的ではないリストは、以下、つまりポータブルコンピュータディスクレット、ハードディスク、ランダムアクセスメモリ（RAM）、読取り専用メモリ（ROM）、消去可能PROM（EPROM又はフラッシュメモリ）、スタティックランダムアクセスメモリ（SRAM）、読取り専用コンパクトディスク（CD-ROM）、デジタル多用途ディスク（DVD）、メモリスティック、フロッピーディスク、例えばパンチカード又は命令がその上に記録された溝内の隆起構造等の機械的に符号化されたデバイス、及び上記の任意の適切な組合せを含む。コンピュータ可読記憶媒体は、本明細書に使用されるように、本来、例えば電波又は他の自由に伝搬する電磁波、導波管又は他の伝送媒体（例えば、光ファイバケーブルを通過する光パルス）を通過して伝搬する電磁波、又はワイヤを通過して伝送され

40

50

る電気信号等の一次的な信号であるとして解釈されるべきではない。

【0595】

【0653】 本明細書に説明されるコンピュータ可読プログラム命令は、コンピュータ可読記憶媒体からそれぞれのコンピューティングデバイス／処理装置に、又は外部コンピュータ若しくは外部ストレージデバイスに、例えばインターネット、ローカルエリアネットワーク、広域ネットワーク、及び／又は無線ネットワークを介してダウンロードできる。ネットワークは、銅伝送ケーブル、光伝送ファイバ、無線伝送、ルータ、ファイアウォール、スイッチ、ゲートウェイコンピュータ、及び／又はエッジサーバ等を含んでよい。各コンピューティングデバイス／処理装置内のネットワークアダプタカード又はネットワークインタフェースは、ネットワークからコンピュータ可読プログラム命令を受け取り、それぞれのコンピューティングデバイス／処理装置の中のコンピュータ可読記憶媒体での格納のためにコンピュータ可読プログラム命令を転送する。

10

【0596】

【0654】 本開示の動作を実施するためのコンピュータ可読プログラム命令は、アセンブラ命令、インストラクションセットアーキテクチャ（ISA）命令、機械命令、マシン依存命令、マイクロコード、ファームウェア命令、状態設定データ、又は例えばSmalltalk、C++等のオブジェクト指向プログラミング言語、及び例えば「C」プログラミング言語若しくは類似したプログラミング言語等の従来の手続き型プログラミング言語を含んだ1つ以上のプログラミング言語の任意の組合せで書き込まれたソースコード若しくはオブジェクトコードのどちらかであってよい。コンピュータ可読プログラム命令は、完全にユーザのコンピュータ上で、部分的にユーザのコンピュータ上で、スタンドアロンソフトウェアパッケージとして、部分的にユーザのコンピュータ上及び部分的にリモートコンピュータ上で、又は完全にリモートコンピュータ若しくはサーバ上で実行してよい。後者の状況では、リモートコンピュータは、ローカルエリアネットワーク（LAN）又は広域ネットワーク（WAN）を含んだ任意のタイプのネットワークを通してユーザのコンピュータに接続される場合もあれば、接続は（例えば、インターネットサービスプロバイダを使用し、インターネットを通して）外部コンピュータに行われる場合もある。いくつかの実施形態では、例えばプログラマブル論理回路、フィールドプログラマブルゲートアレイ（FPGA）、又はプログラマブルロジックアレイ（PLA）を含んだ電子回路網は、本開示の態様を実行するために、コンピュータ可読プログラム命令の状態情報を利用して電子回路網を個人向けにすることによってコンピュータ可読プログラム命令を実行してよい。

20

30

【0597】

【0655】 本開示の態様は、本開示のいくつかの実施形態に係る方法、装置（システム）及びコンピュータプログラム製品のフローチャート図及び／又はブロック図を参照して本明細書で説明される。フローチャート図及び／又はブロック図の各ブロック、並びにフローチャート図及び／又はブロック図の中のブロックの組合せが、コンピュータ可読プログラム命令によって実装できることが理解される。

【0598】

【0656】 これらのコンピュータ可読プログラム命令は、機械を生産するために、汎用コンピュータ、専用コンピュータ、又は他のプログラム可能データ処理装置のプロセッサに提供されてよく、これによりコンピュータ又は他のプログラム可能データ処理装置のプロセッサを介して実行する命令は、フローチャート及び／又はブロック図の1つ又は複数のブロックに指定される機能／行為を実施するための手段を作り出す。また、これらのコンピュータ可読プログラム命令は、コンピュータ、プログラム可能データ処理装置、及び／又は他のデバイスに特定の方法で機能するように命令できるコンピュータ可読記憶媒体に記憶されてよく、これによりその中に命令が記憶されているコンピュータ可読記憶媒体は、フローチャート及び／又はブロック図の1つ又は複数のブロックに指定される機能／行為の態様を実施する命令を含んだ製造品を含む。

40

【0599】

50

[0657] また、コンピュータ可読プログラム命令は、一連の操作ステップをコンピュータ、他のプログラム可能装置、又は他のデバイスで実行させて、コンピュータ実装プロセスを生じさせるためにコンピュータ、他のプログラム可能データ処理装置、又は他の装置にロードされてもよく、これによりコンピュータ、他のプログラム可能装置、又は他のデバイスで実行する命令は、フローチャート及び／又はブロック図の1つ又は複数のブロックで指定される機能／行為を実施する。

【0600】

[0658] 図に示されるフローチャート及びブロック図は、本開示の態様な実施形態に係るシステム、方法、及びコンピュータプログラム製品の考えられる実施態様のアーキテクチャ、機能性、及び動作を示す。この時点で、フローチャート図又はブロック図の各ブロックは、指定された論理機能（複数可）を実施するための1つ以上の実行可能命令を含むモジュール、セグメント、又は命令の部分を表してよい。いくつかの代替実施態様では、ブロックで注記された機能は、図に注記される順序を外れて起こる場合がある。例えば、連続して示される2つのブロックが、実際には実質的に同時に実行される場合もあれば、関与する機能性に応じてブロックが逆順で実行される場合もある。ブロック図及びフローチャート図の各ブロック、及びブロック図及び／又はフローチャート図のブロックの組合せが、指定された機能若しくは行為を実行する、又は専用ハードウェア及びコンピュータ命令の組合せを実施する専用ハードウェアベースのシステムによって実施される場合があることも留意されたい。

【0601】

[0659] 読者は、本明細書に説明されるステップが、さまざまな方法で実施されてよいこと、及び特定の順序付けが必要とされていないことを理解する。上述の説明から、修正及び変更が、本開示の多様な実施形態でその真の精神から逸脱することなく加えられてよいこともさらに理解される。本明細書の説明は、例示の目的のためだけであり、限定的な意味で解釈されるべきではない。本開示の範囲は、以下の特許請求の範囲の言語によってのみ限定される。

10

20

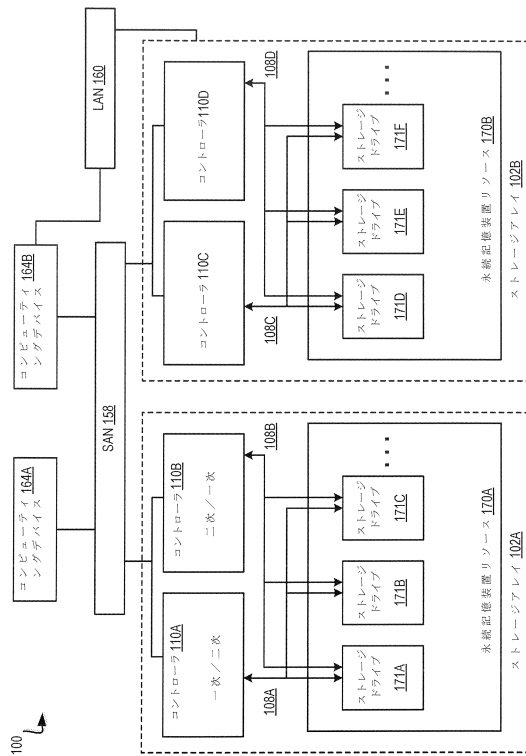
30

40

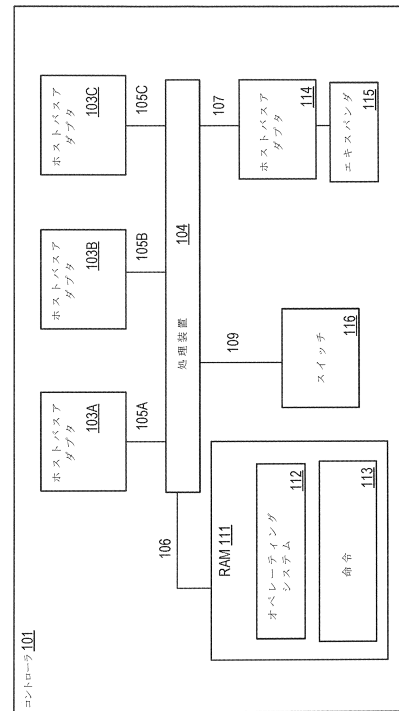
50

【図面】

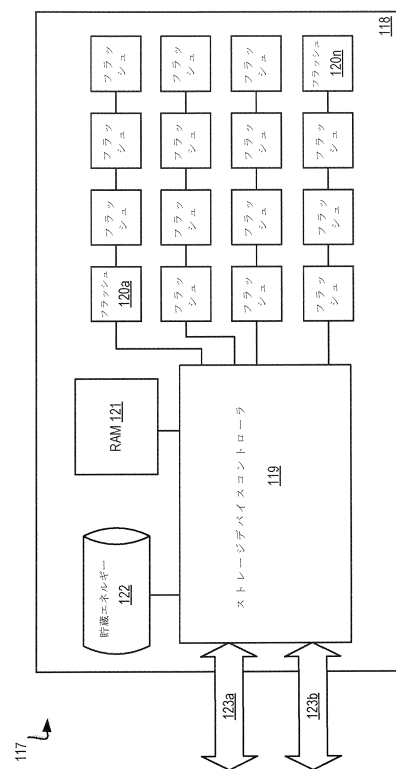
【図 1 A】



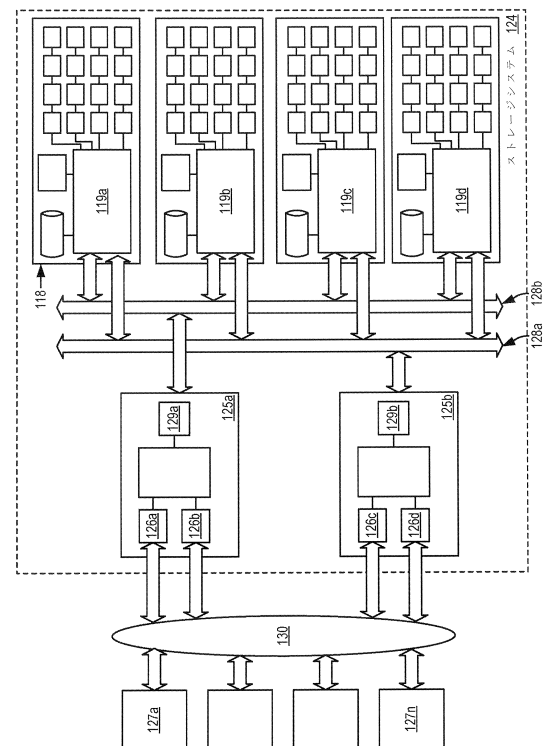
【図 1 B】



【図 1 C】



【図 1 D】



10

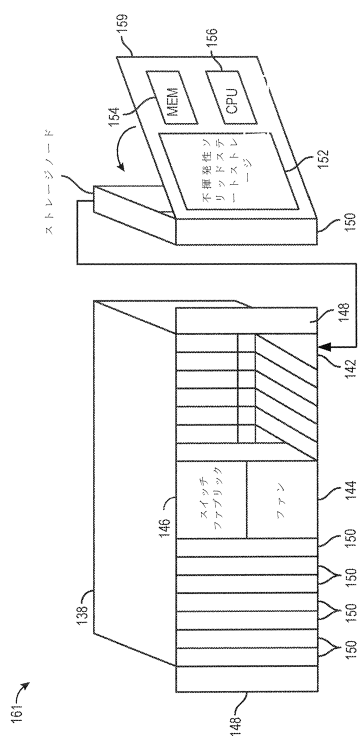
20

30

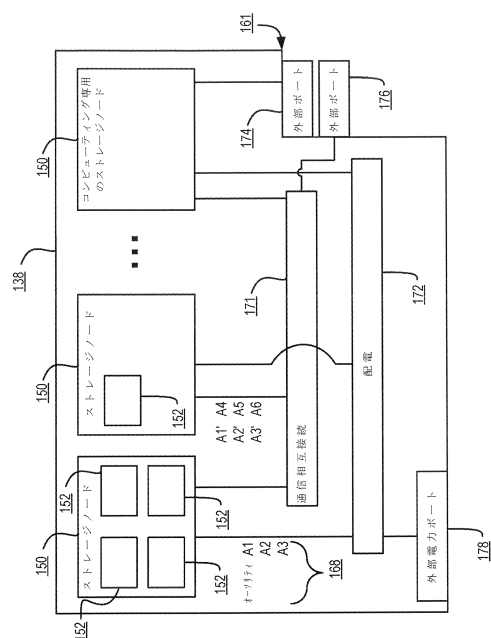
40

50

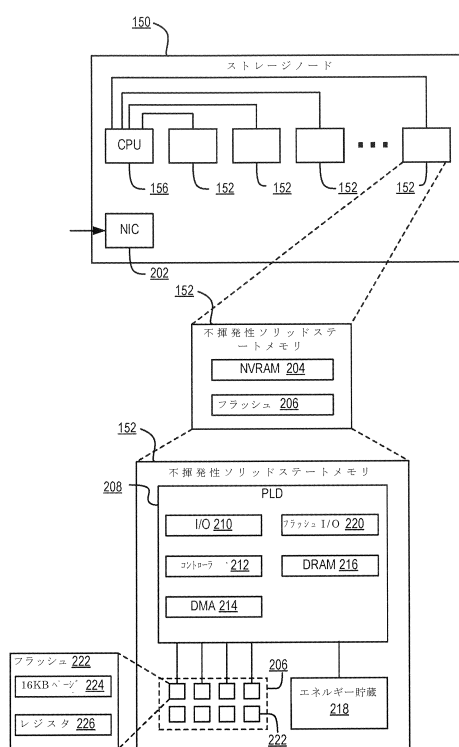
【図 2 A】



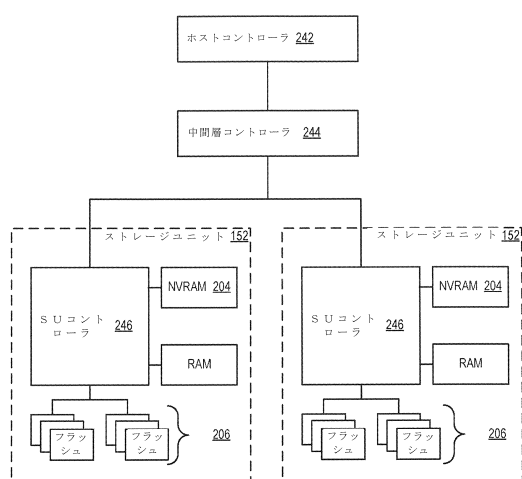
【図 2 B】



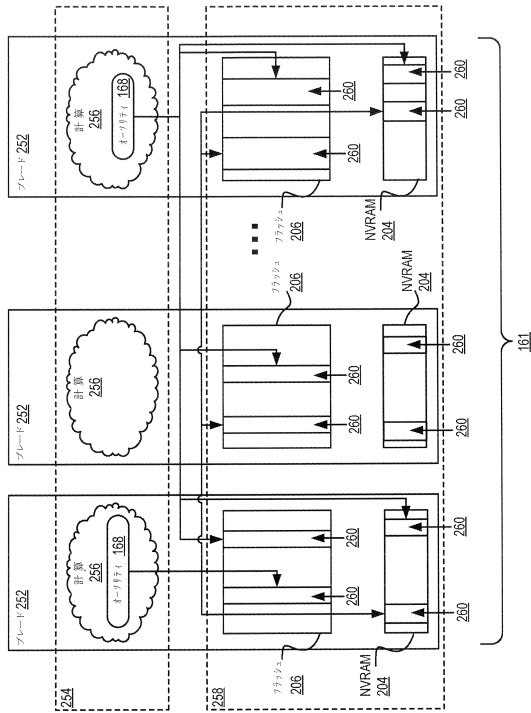
【図 2 C】



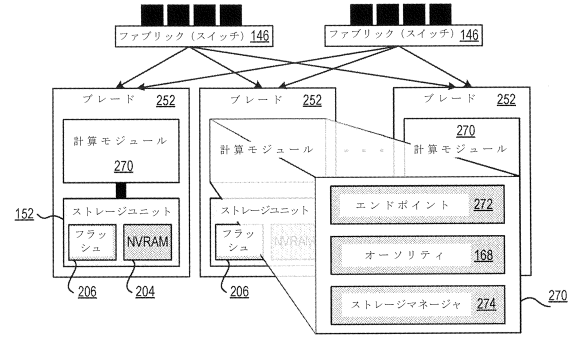
【図 2 D】



【図 2 E】



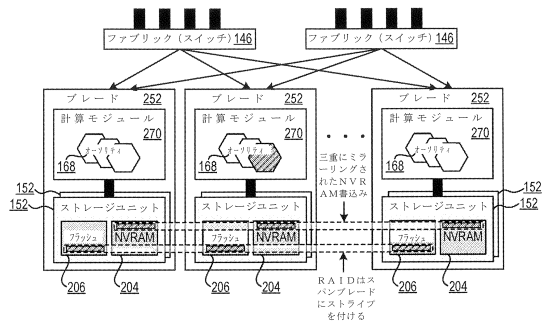
【図 2 F】



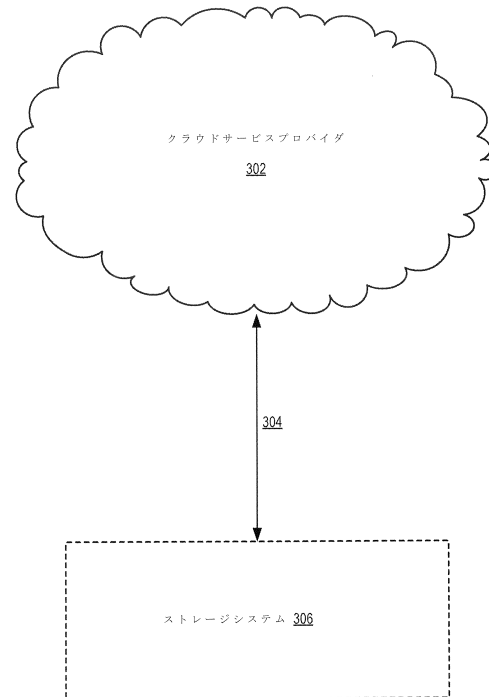
10

20

【図 2 G】



【図 3 A】

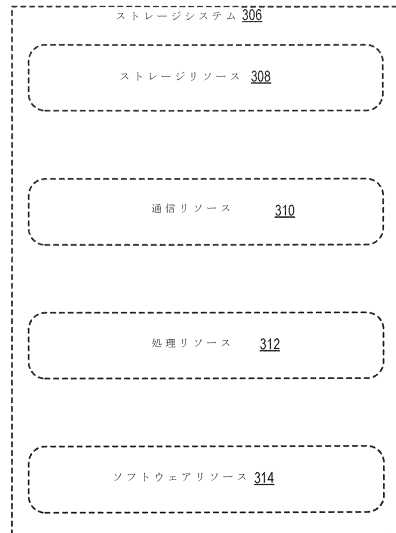


30

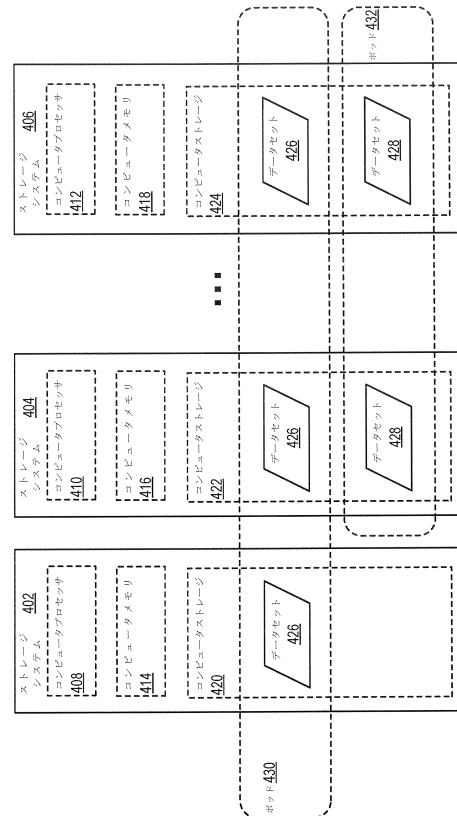
40

50

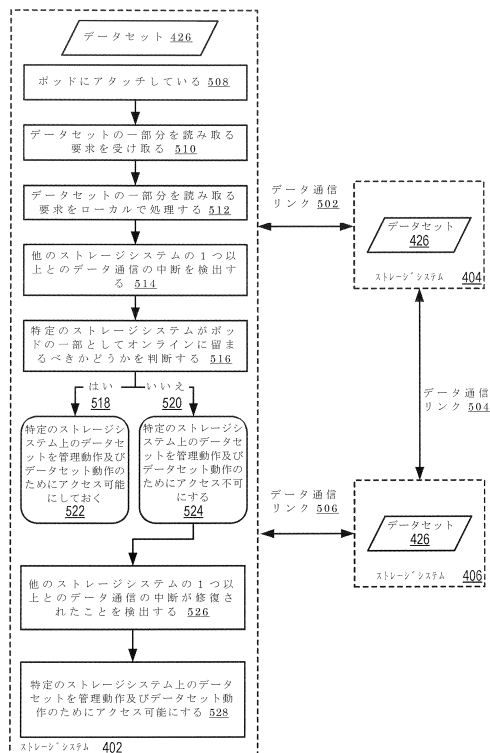
【図 3 B】



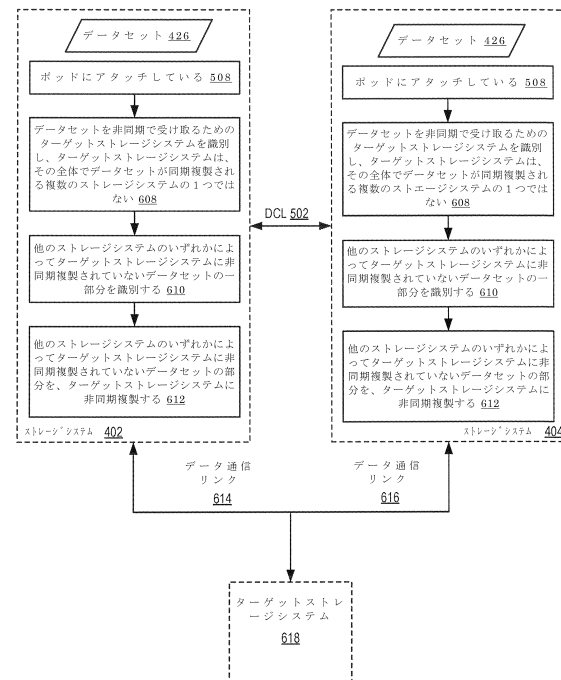
【図 4】



【図 5】



【図 6】



10

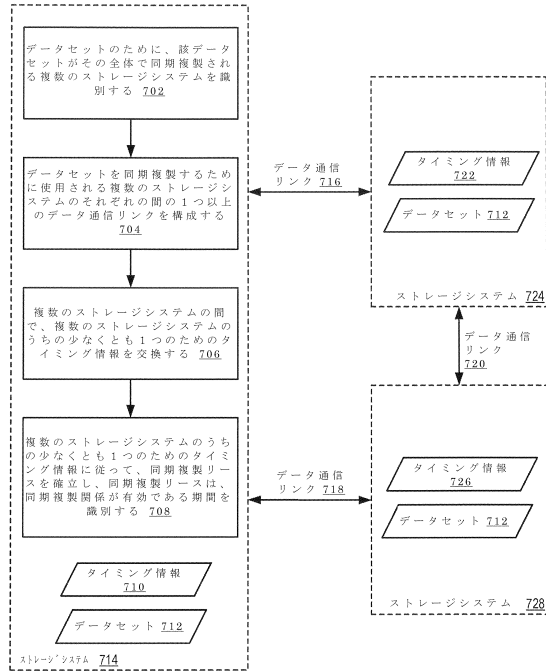
20

30

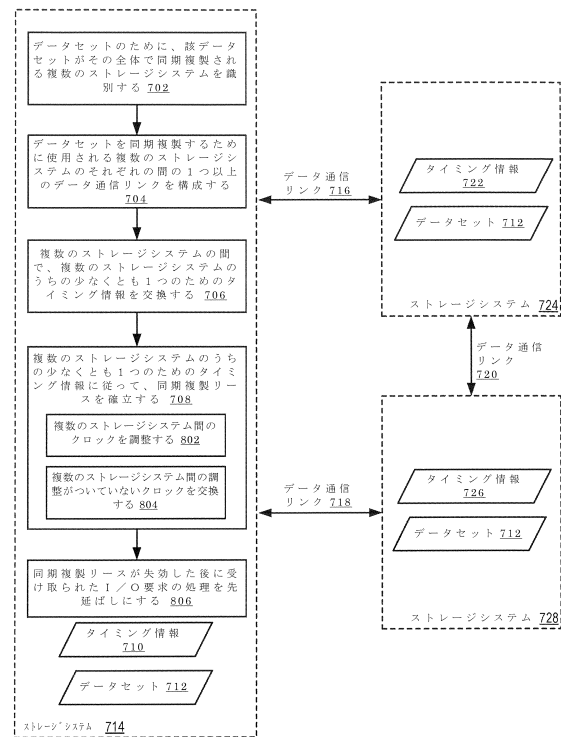
40

50

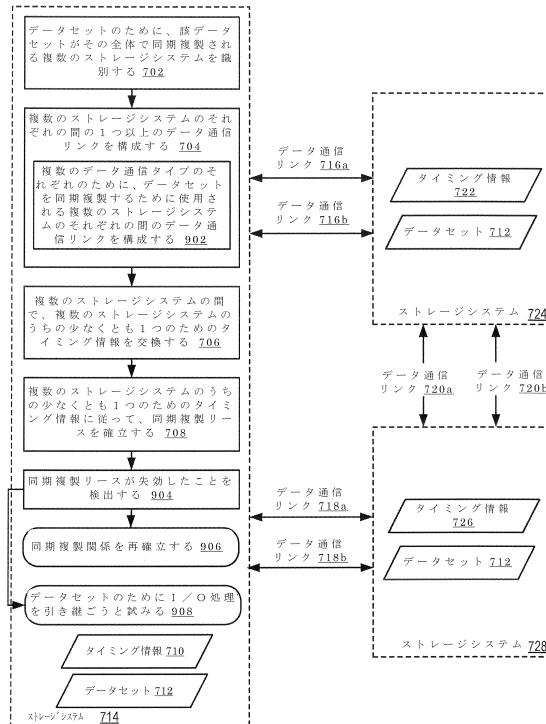
【図 7】



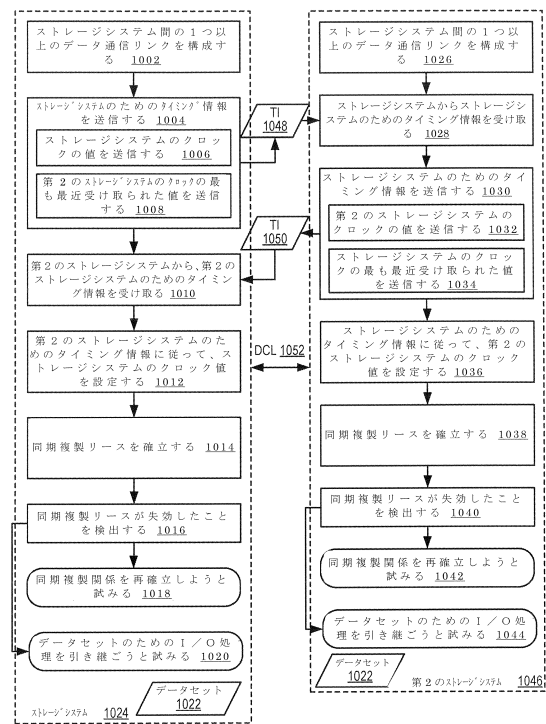
【図 8】



【図 9】



【図 10】



10

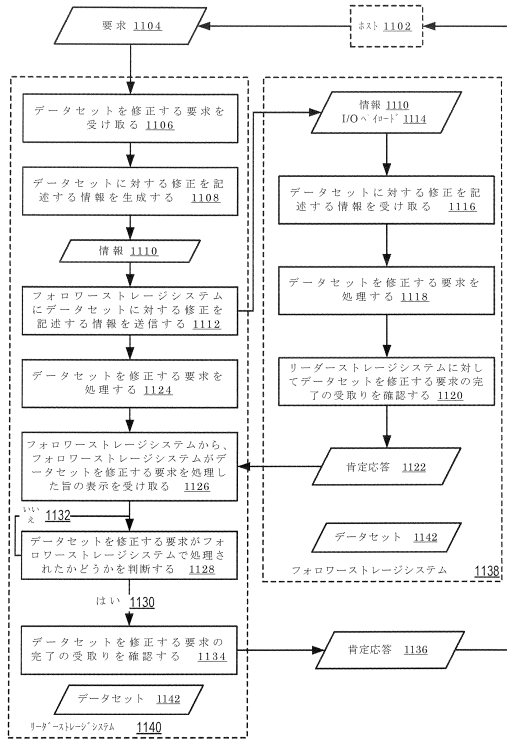
20

30

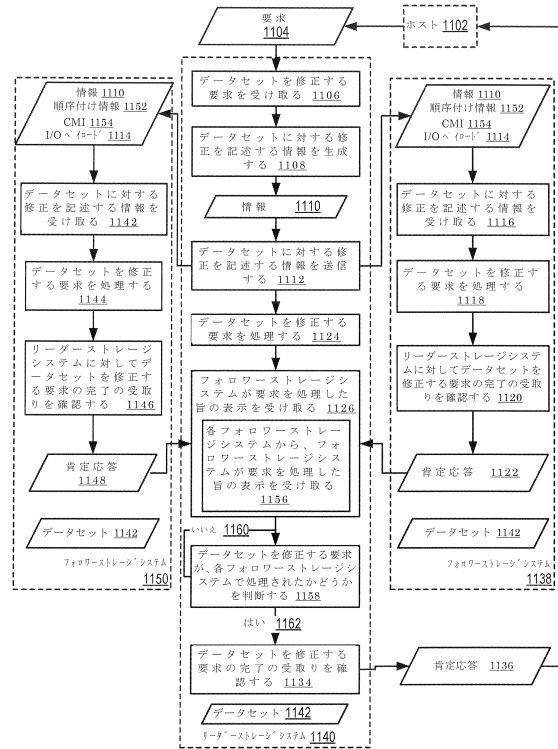
40

50

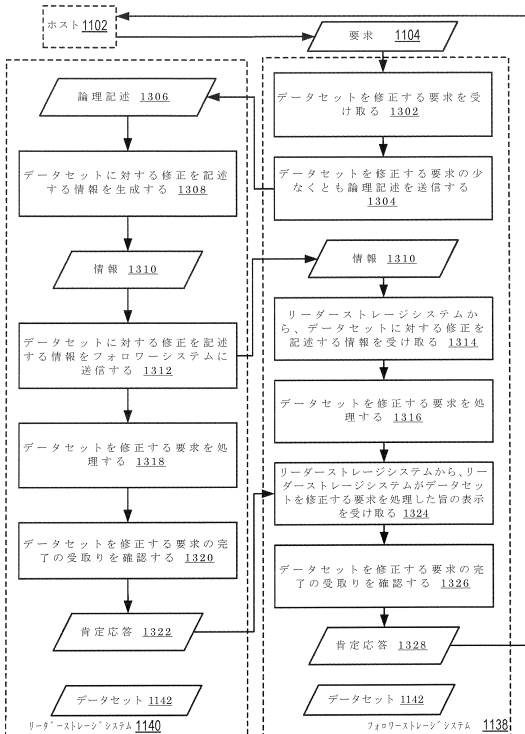
【図 11】



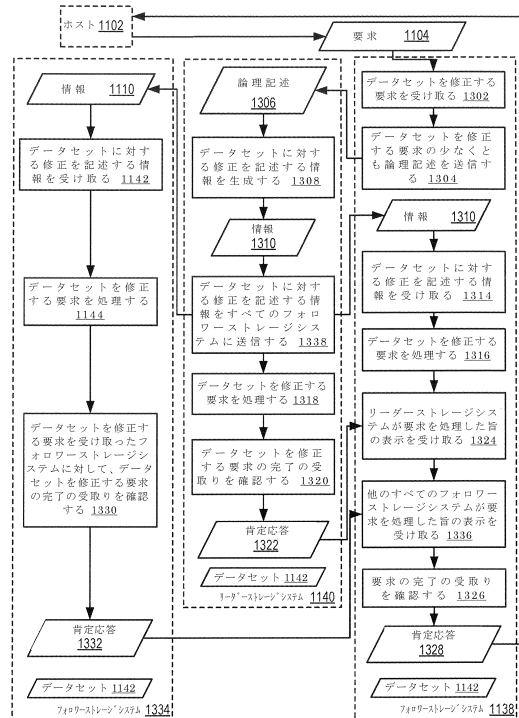
【図 12】



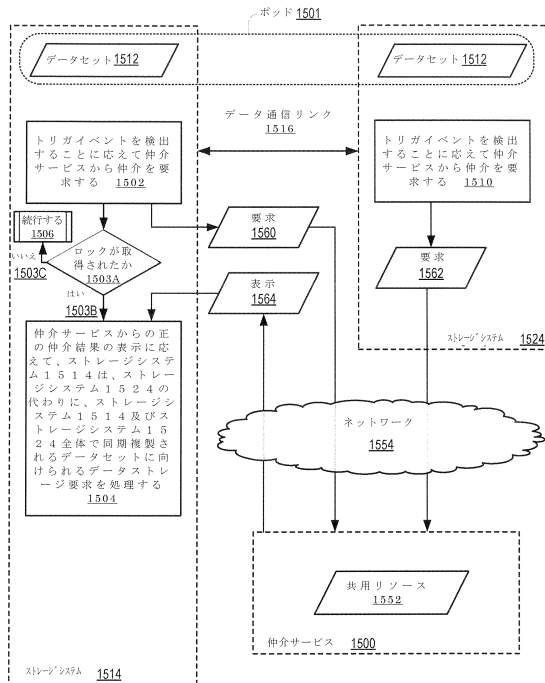
【図 13】



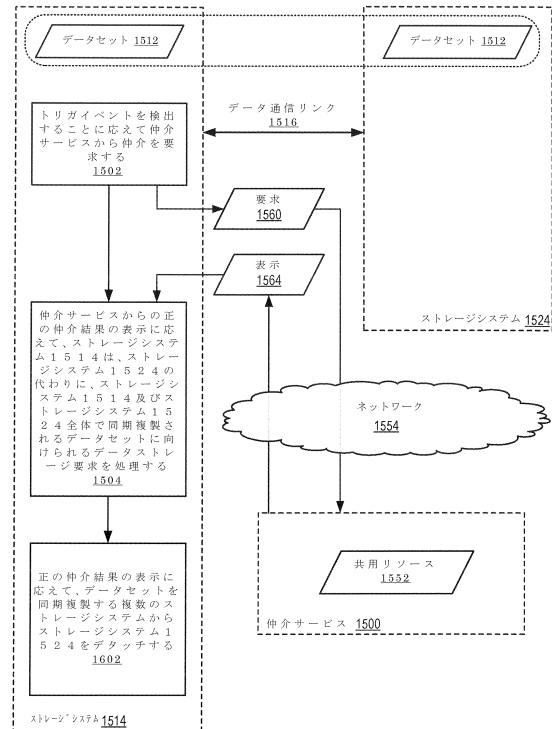
【図 14】



【図 15】



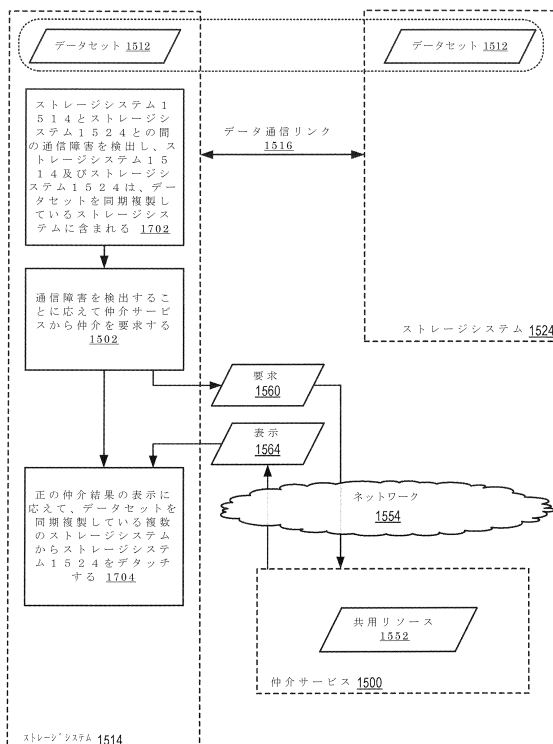
【図 16】



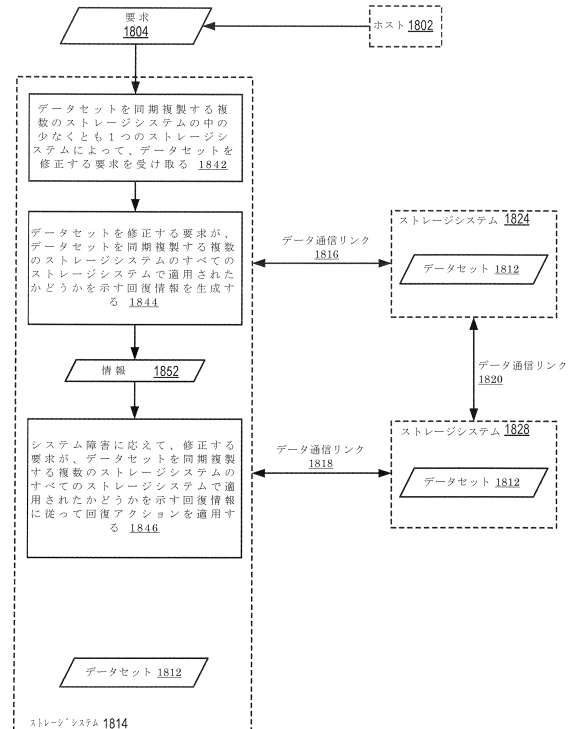
10

20

【圖 17】



【圖 18】

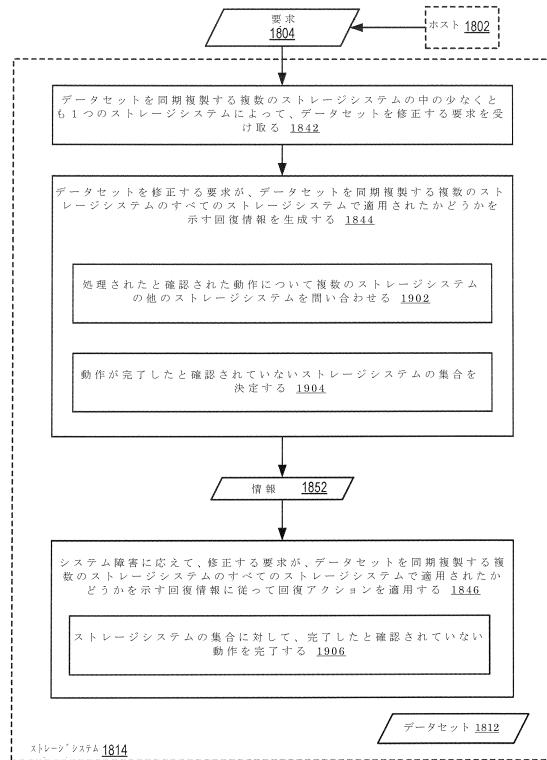


30

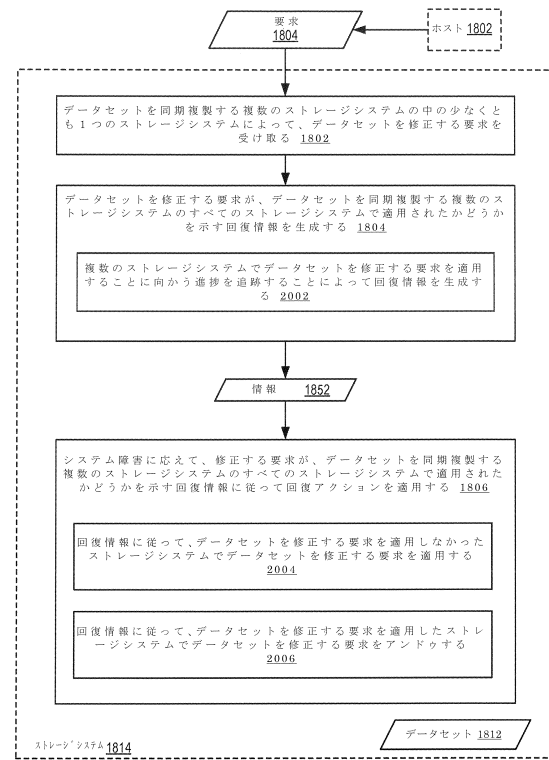
40

50

【図 19】



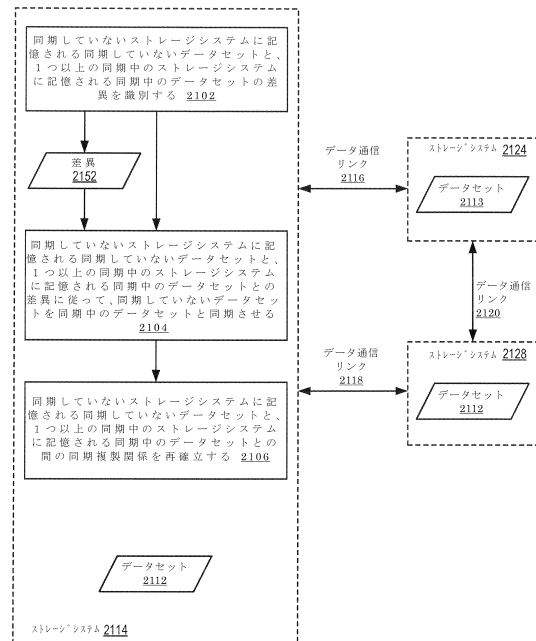
【図 20】



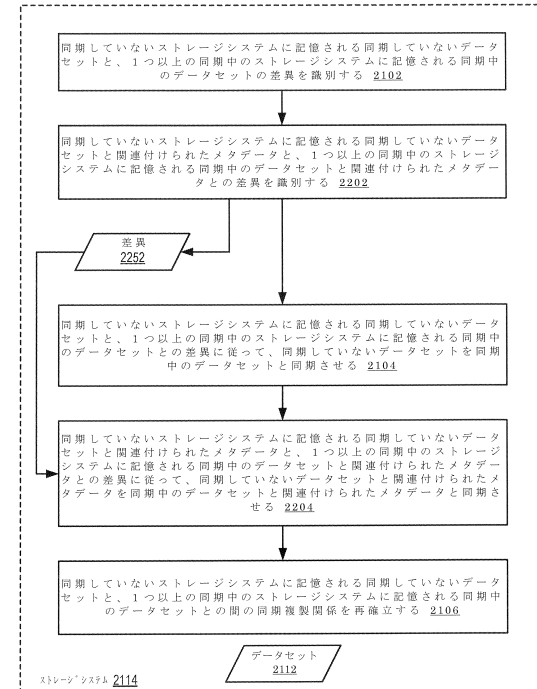
10

20

【図 21】



【図 22】

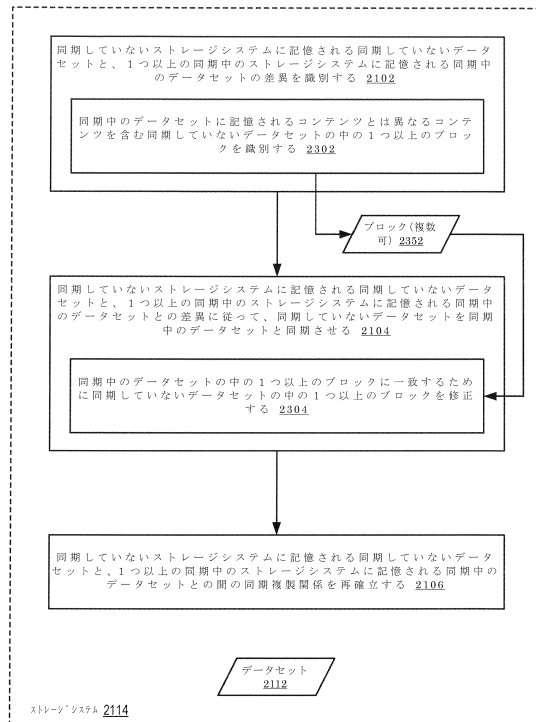


30

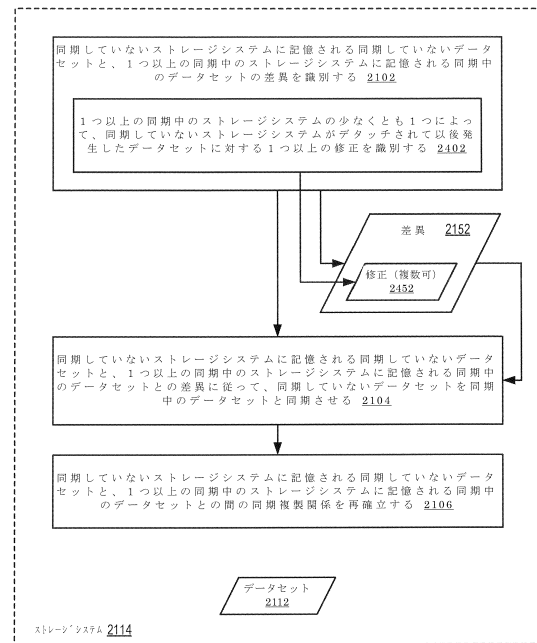
40

50

【図 23】



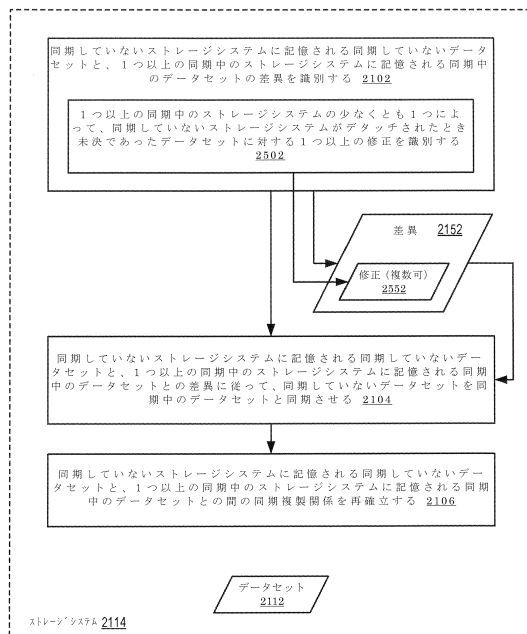
【図 24】



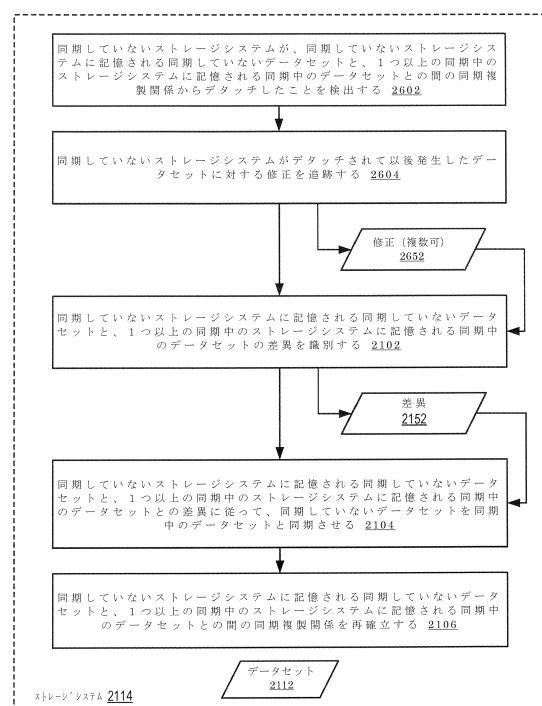
10

20

【図 25】



【図 26】

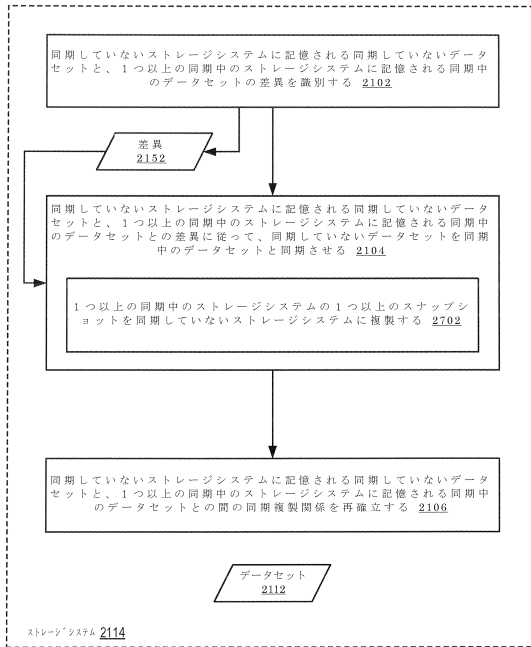


30

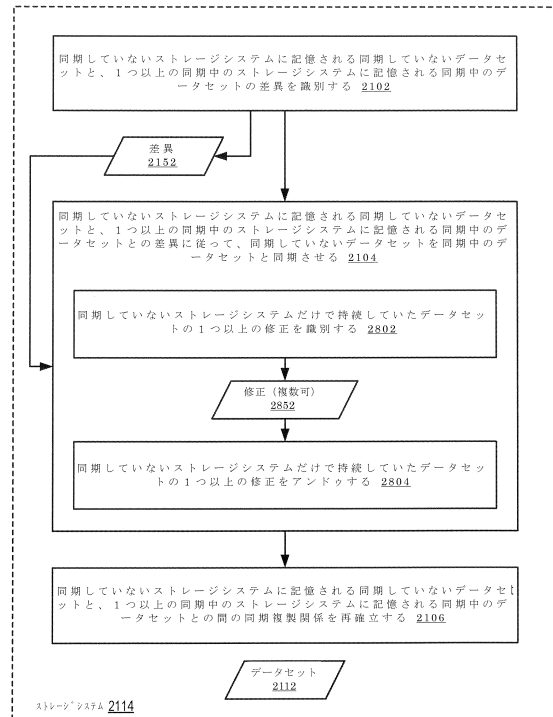
40

50

【図 27】



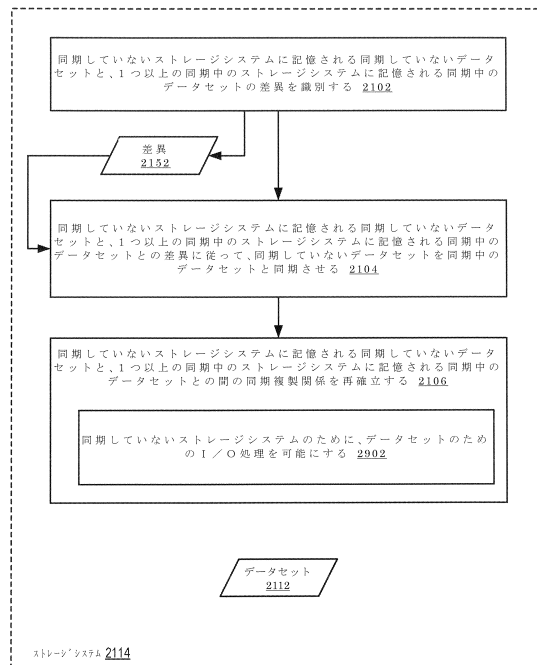
【図 28】



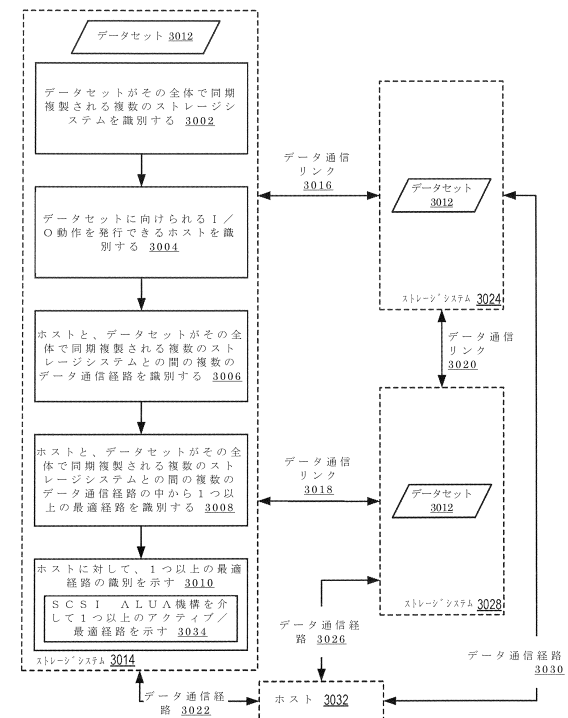
10

20

【図 29】



【図 30】

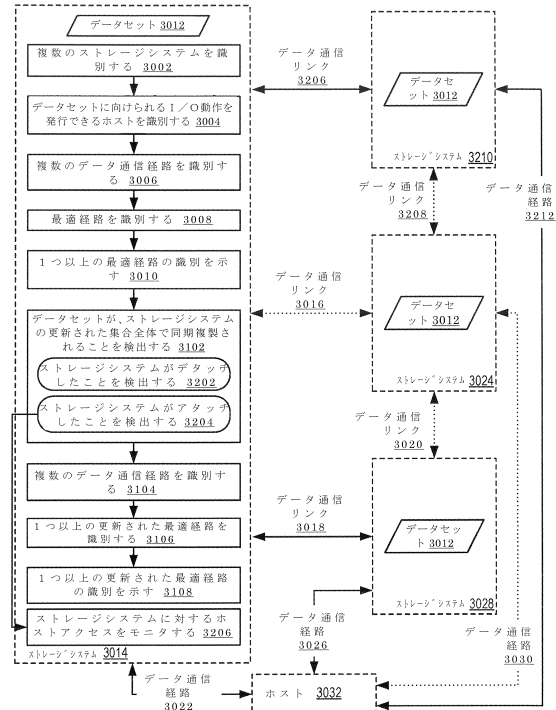


30

40

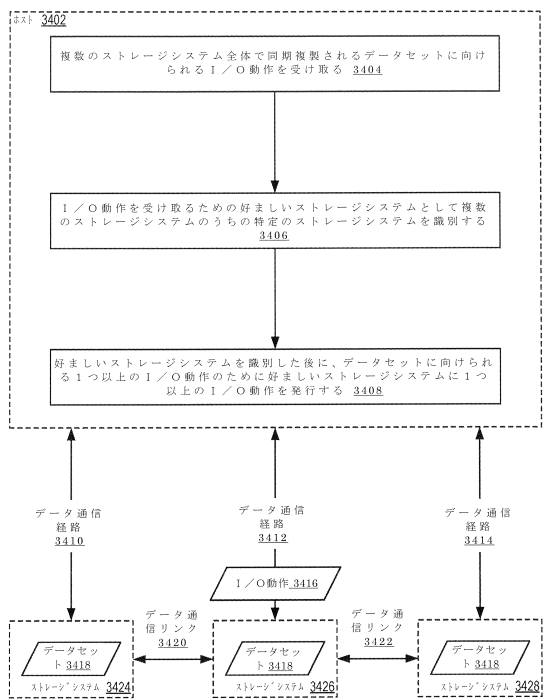
50

【図 3 2】



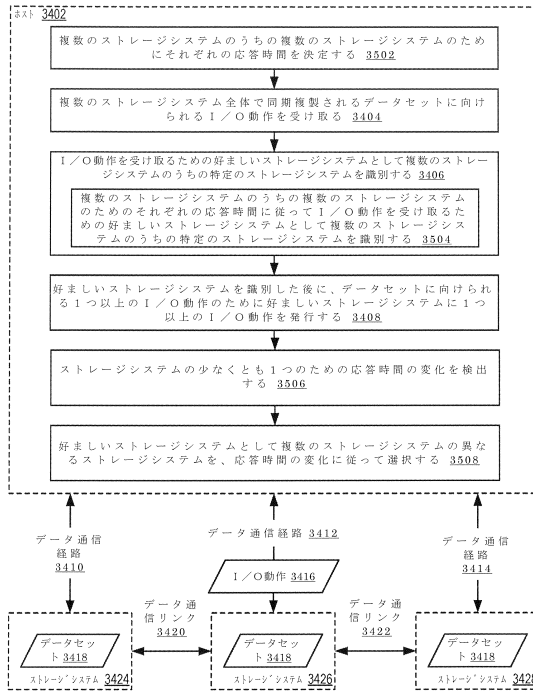
20

【図 3 4】

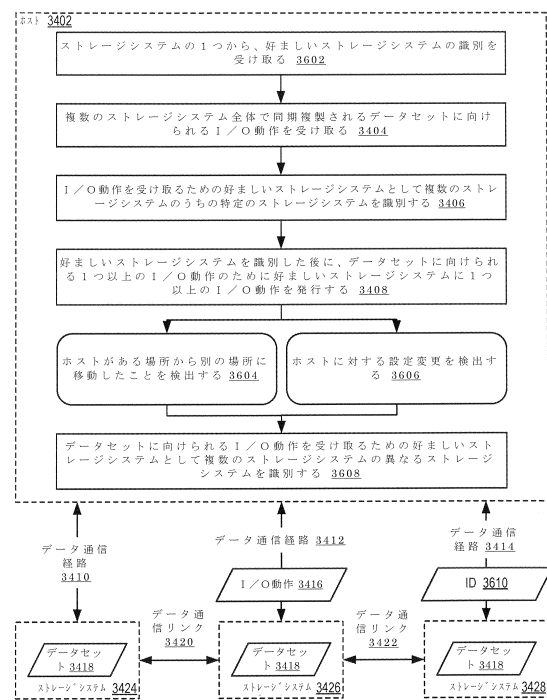


40

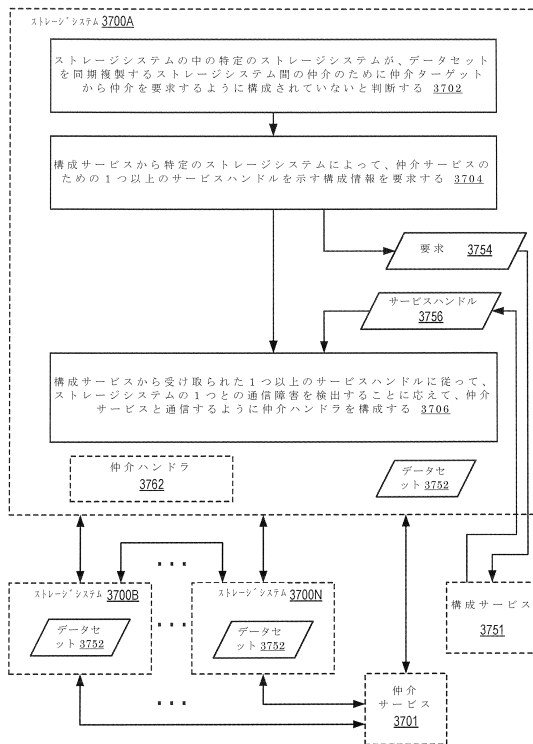
【図 35】



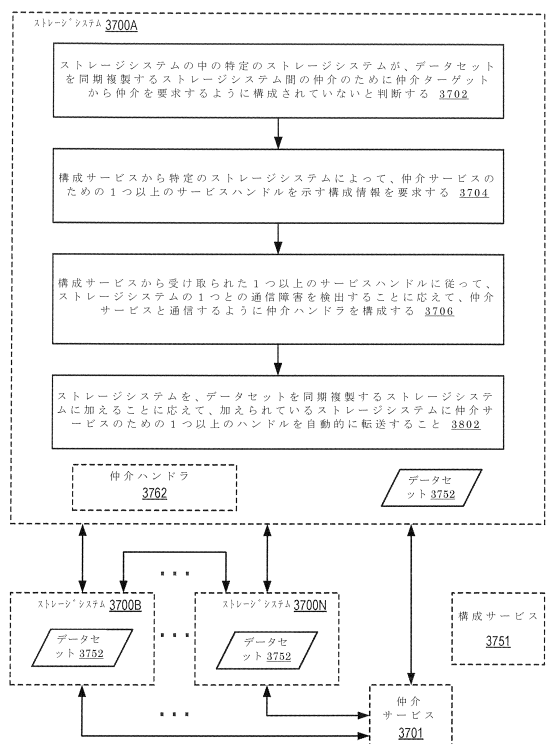
【図 36】



【図 37】



【図 38】



10

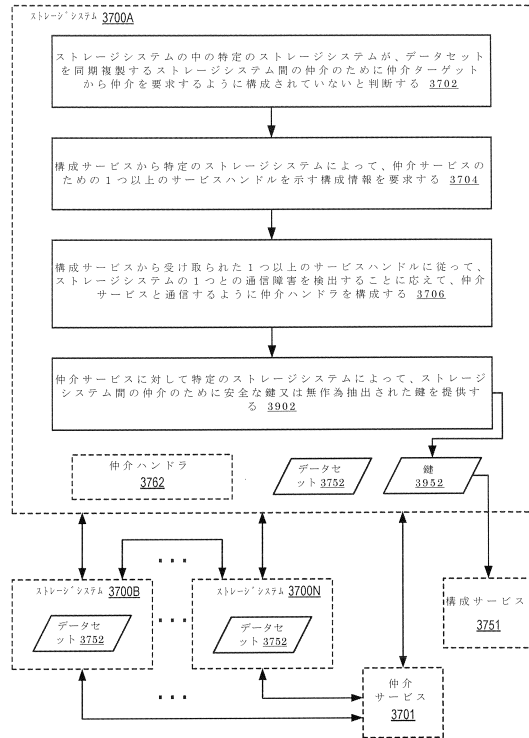
20

30

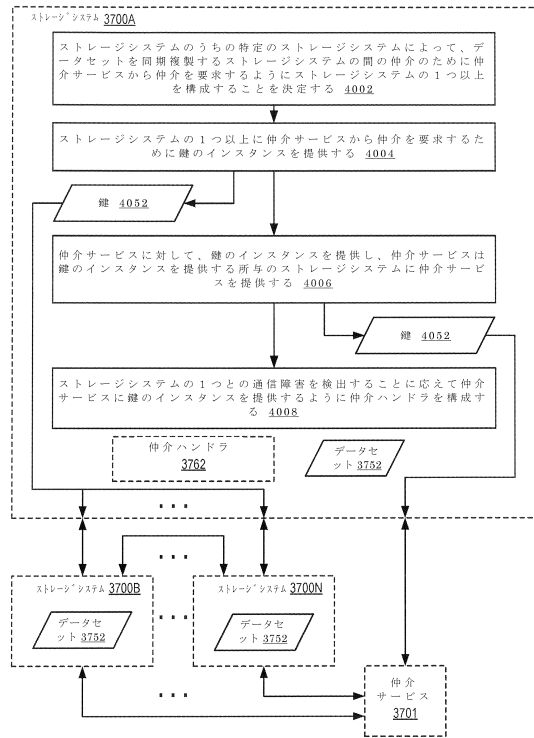
40

50

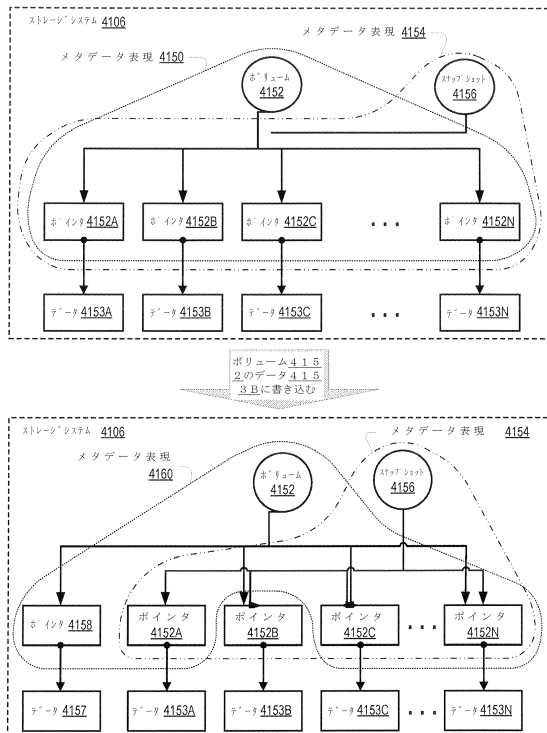
【図 39】



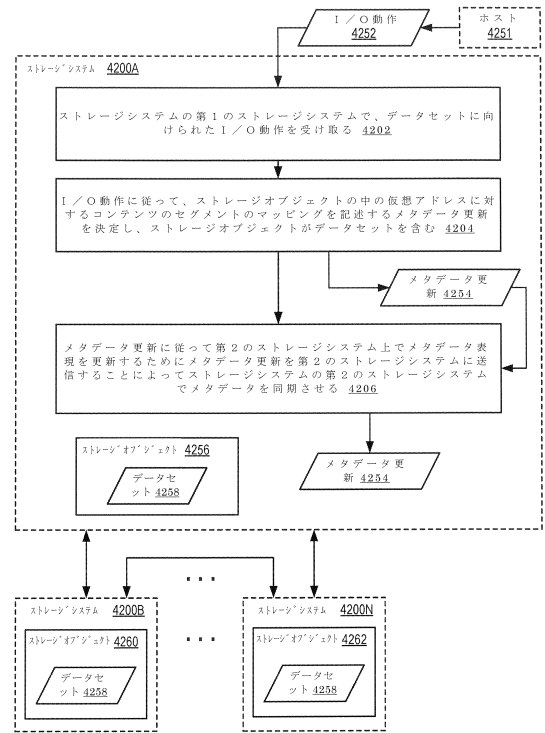
【図 40】



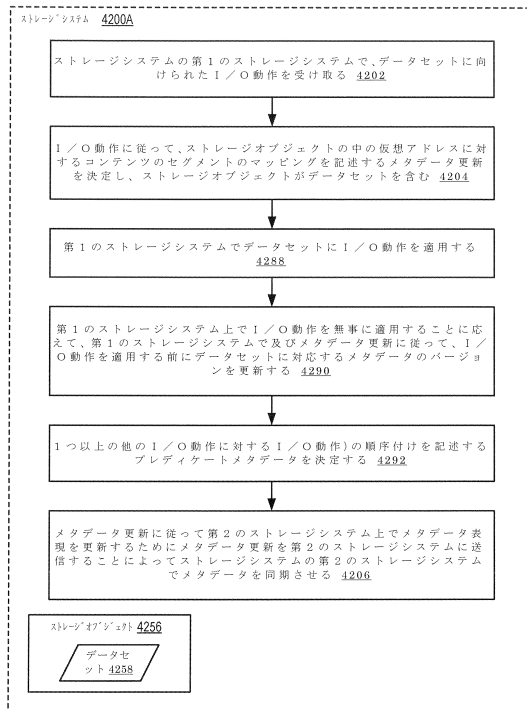
【図 4 1】



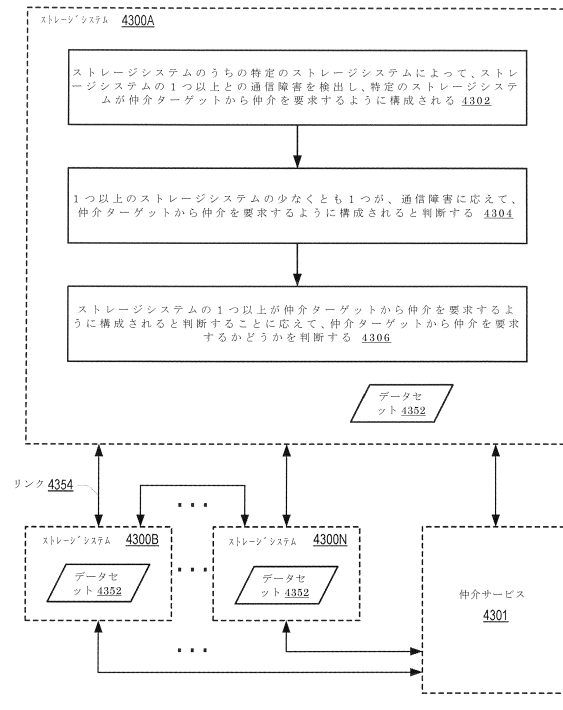
【図 4 2 A】



【図 4 2 B】



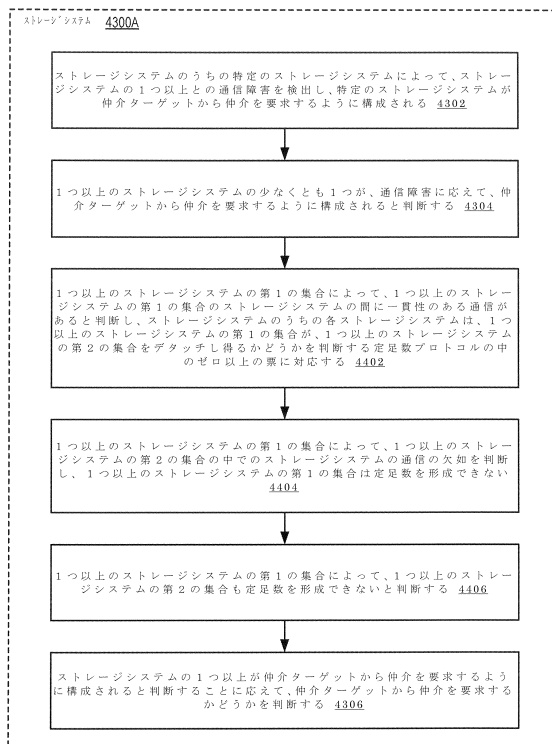
【図 4 3】



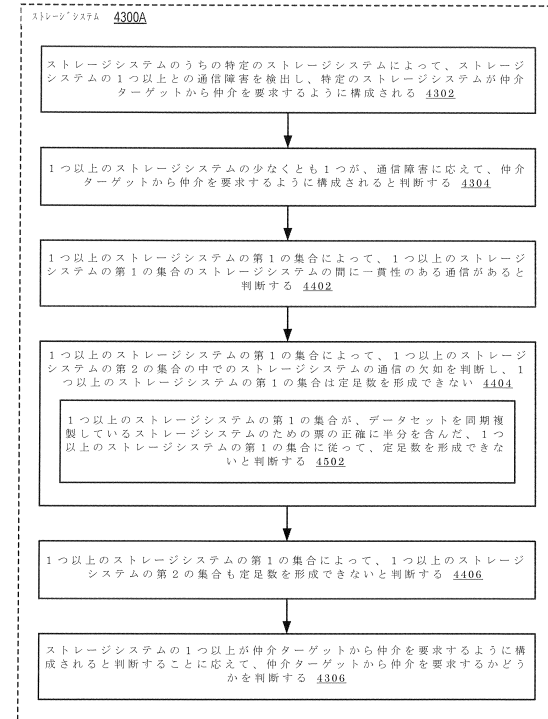
10

20

【図 4 4】



【図 4 5】

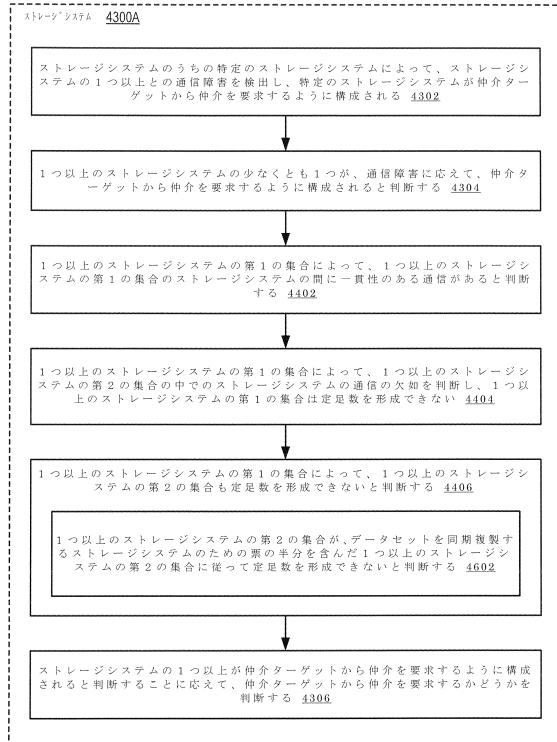


30

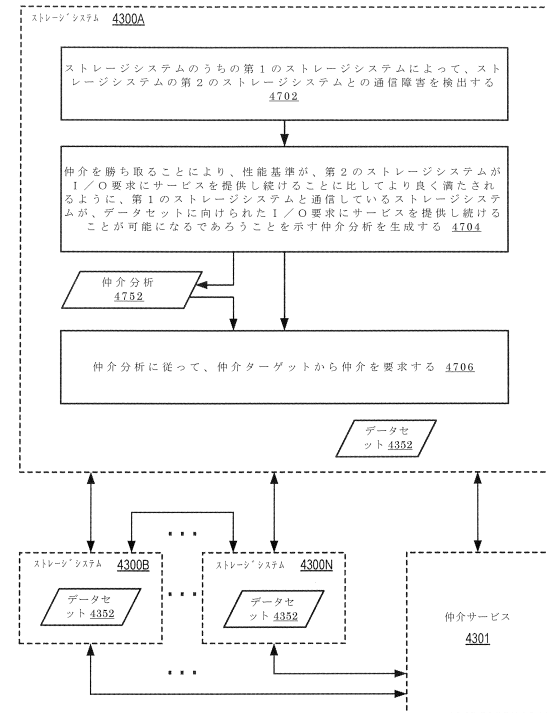
40

50

【図 4 6】



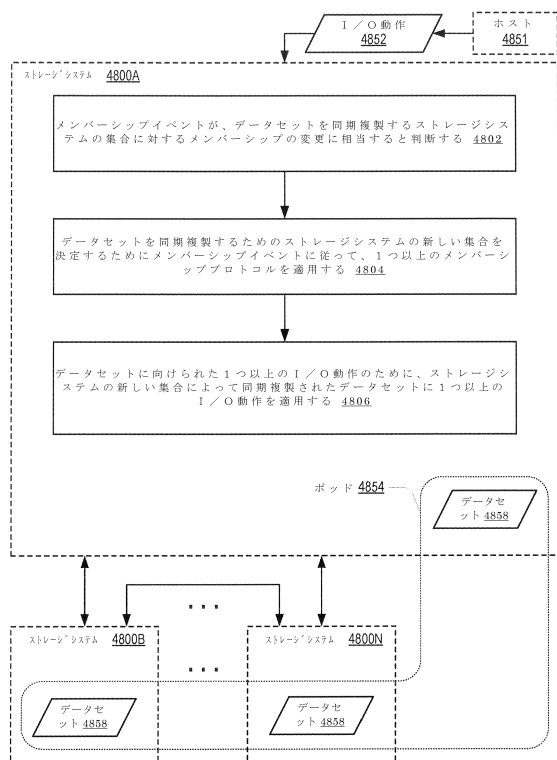
【図 4 7】



10

20

【図 4 8】



30

40

50

フロントページの続き

(33)優先権主張国・地域又は機関

米国(US)

(31)優先権主張番号 62/518,071

(32)優先日 平成29年6月12日(2017.6.12)

(33)優先権主張国・地域又は機関

米国(US)

(31)優先権主張番号 15/842,850

(32)優先日 平成29年12月14日(2017.12.14)

(33)優先権主張国・地域又は機関

米国(US)

(31)優先権主張番号 62/598,989

(32)優先日 平成29年12月14日(2017.12.14)

(33)優先権主張国・地域又は機関

米国(US)

(74)代理人 100126480

弁理士 佐藤 睦

(72)発明者 ボーツ, パー

アメリカ合衆国, カリフォルニア州 94041 マウンテン ビュー, カストロ ストリート 65
0, スイート 260

(72)発明者 コルグローブ, ジョン

アメリカ合衆国, カリフォルニア州 94041 マウンテン ビュー, カストロ ストリート 65
0, スイート 260

(72)発明者 ドリスコル, アラン

アメリカ合衆国, カリフォルニア州 94041 マウンテン ビュー, カストロ ストリート 65
0, スイート 260

(72)発明者 グランウォルド, デイヴィッド

アメリカ合衆国, カリフォルニア州 94041 マウンテン ビュー, カストロ ストリート 65
0, スイート 260

(72)発明者 ホジソン, スティーヴン

アメリカ合衆国, カリフォルニア州 94041 マウンテン ビュー, カストロ ストリート 65
0, スイート 260

(72)発明者 カー, ロナルド

アメリカ合衆国, カリフォルニア州 94041 マウンテン ビュー, カストロ ストリート 65
0, スイート 260

審査官 三橋 竜太郎

(56)参考文献 米国特許第09378219 (US, B1)

特開2016-119062 (JP, A)

特開2005-316684 (JP, A)

米国特許出願公開第2005/0273565 (US, A1)

(58)調査した分野 (Int.Cl., DB名)

G06F 16/00-16/958

G06F 13/10

G06F 3/06