(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第6947670号
(P6947670)

(45)発行日　令和3年10月13日(2021.10.13)　(24)登録日　令和3年9月21日(2021.9.21)

(51)Int.Cl.			F I		
G06F	3/06	(2006.01)	G06F	3/06	305Z
G06F	3/08	(2006.01)	G06F	3/06	305C
			G06F	3/06	304F
			G06F	3/08	H
			G06F	3/06	304N

請求項の数 20　(全 19 頁)

(21)出願番号　特願2018-52379(P2018-52379)
(22)出願日　平成30年3月20日(2018.3.20)
(65)公開番号　特開2018-156656(P2018-156656A)
(43)公開日　平成30年10月4日(2018.10.4)
審査請求日　令和3年2月17日(2021.2.17)
(31)優先権主張番号　62/474,039
(32)優先日　平成29年3月20日(2017.3.20)
(33)優先権主張国・地域又は機関
米国(US)
(31)優先権主張番号　62/561,625
(32)優先日　平成29年9月21日(2017.9.21)
(33)優先権主張国・地域又は機関
米国(US)

(73)特許権者　390019839
三星電子株式会社
Ｓａｍｓｕｎｇ　Ｅｌｅｃｔｒｏｎｉｃｓ　Ｃｏ．，Ｌｔｄ．
大韓民国京畿道水原市霊通区三星路１２９
１２９，Ｓａｍｓｕｎｇ－ｒｏ，Ｙｅｏｎｇｔｏｎｇ－ｇｕ，Ｓｕｗｏｎ－ｓｉ，Ｇｙｅｏｎｇｇｉ－ｄｏ，Ｒｅｐｕｂｌｉｃ　ｏｆ　Ｋｏｒｅａ
(74)代理人　110000051
特許業務法人共生国際特許事務所
(72)発明者　キ　ヤン　ソク
アメリカ合衆国，　９４３０３，　カリフォルニア州，　パロ　アルト，　アルテア　ウォーク　８７３

最終頁に続く

(54)【発明の名称】仮想装置階層を利用した複数のメモリ装置を含む仮想装置に対する客体の格納及び読み取り方法とこれを用いたストレージ装置

(57)【特許請求の範囲】

【請求項１】

各々が許容可能な最小サイズの値を有し、ステートレスのデータ保護を利用する仮想装置として構成された複数からなるメモリ装置と、

第１客体及び第２客体のそれぞれのサイズに応じて、前記第１客体及び前記第２客体の各々に第１データ保護及び第２データ保護の中の１つを適用することにより、各々がそれぞれのサイズを有する前記第１客体及び前記第２客体を格納するように、前記仮想装置を管理するように構成された仮想装置階層と、を備え、

前記仮想装置階層は、前記第１データ保護を大型客体に適用し、

前記第２データ保護を小型客体に適用し、

前記第１データ保護又は前記第２データ保護のいずれかを中型客体に適用するように構成されることを特徴とするストレージ装置。

【請求項２】

前記メモリ装置は、１つ以上のデータ装置及び１つ以上のパリティ装置として構成されることを特徴とする請求項１に記載のストレージ装置。

【請求項３】

前記第１データ保護は、消去コーディングを含み、

前記第２データ保護は、複製を含むことを特徴とする請求項２に記載のストレージ装置。

【請求項４】

前記消去コーディングは、前記第１客体及び前記第２客体の中の該当客体が大型客体として分類される場合、データ保護のために利用されることを特徴とする請求項３に記載のストレージ装置。

【請求項５】

前記第１客体及び前記第２客体の中の該当客体は、以下の不等式を満足する場合、前記大型客体として分類されることを特徴とする請求項４に記載のストレージ装置。

（（Ｐ＋１）×Ｏ　＞　（Ｓ＋Ｐ）×ｍ　ＡＮＤ　Ｏ　＞＝　Ｓ×ｍ）

（ここで、Ｏは客体のサイズ、Ｐはパリティ装置の個数、Ｓはデータ装置の個数、ｍは前記複数からなるメモリ装置のそれぞれの許容可能な最小サイズの値の中の最大値を示す。）

【請求項６】

前記複製は、前記第１客体及び前記第２客体の中の該当客体が小型客体として分類される場合、データ保護のために利用されることを特徴とする請求項３に記載のストレージ装置。

【請求項７】

前記第１客体及び前記第２客体の中の該当客体は、以下の不等式を満足する場合、前記小型客体として分類されることを特徴とする請求項６に記載のストレージ装置。

（（Ｐ＋１）×Ｏ　＝＜　（Ｓ＋Ｐ）×ｍ）

（ここで、Ｏは客体のサイズ、Ｐはパリティ装置の個数、Ｓはデータ装置の個数、ｍは前記複数からなるメモリ装置のそれぞれの許容可能な最小サイズの値の中の最大値を示す。）

【請求項８】

前記消去コーディング又は前記複製のいずれかは、前記第１客体及び前記第２客体の中の該当客体が大型客体でも小型客体でもなく中型客体として分類される場合、前記複数からなるメモリ装置で利用可能な１つ以上の空間、前記仮想装置に格納された前記第１客体又は前記第２客体に対するアクセス時間、及び前記第１客体又は前記第２客体がアクセスされる頻度に基づいて、データ保護のために利用されることを特徴とする請求項３に記載のストレージ装置。

【請求項９】

前記第１客体及び前記第２客体の中の該当客体は、以下の不等式を満足する場合、前記中型客体として分類されることを特徴とする請求項８に記載のストレージ装置。

（（Ｐ＋１）×Ｏ　＞　（Ｓ＋Ｐ）×ｍ）　＞　Ｓ×ｍ　＞　Ｏ）

（ここで、Ｏは客体のサイズ、Ｐはパリティ装置の個数、Ｓはデータ装置の個数、ｍは前記複数からなるメモリ装置のそれぞれの許容可能な最小サイズの値の中の最大値を示す。）

【請求項１０】

前記パリティ装置は、前記メモリ装置の固定されたサブセットであることを特徴とする請求項２に記載のストレージ装置。

【請求項１１】

前記パリティ装置は、前記メモリ装置の可変されるサブセットを含み、

前記複数からなるメモリ装置の各々は、データ装置及びパリティ装置の両方として動作するように構成されることを特徴とする請求項２に記載のストレージ装置。

【請求項１２】

前記メモリ装置は、ソリッドステートドライブを含むことを特徴とする請求項１に記載のストレージ装置。

【請求項１３】

仮想装置階層を利用して、各々が許容可能な最小サイズの値を有する複数からなるメモリ装置を含む仮想装置に複数からなる客体を格納する方法であって、

前記方法は、前記仮想装置階層によって、格納される客体を受信する段階と、

前記仮想装置階層によって、Ｏ、Ｐ、Ｓ、及びｍに基づき前記客体が大型であるか又は

小型であるかを判定する段階と、を有し、
ここで、Ｏは客体のサイズ、Ｐはパリティ装置の個数、Ｓはデータ装置の個数、ｍは前記客体を大型又は小型に分類するための前記複数からなるメモリ装置のそれぞれの許容可能な最小サイズの値の中の最大値を示し、
前記客体が大型として分類される場合、消去コーディングのためのチャンクのサイズ及び前記客体のデータチャンクのパディング量を決定する段階と、
各々が前記消去コーディングのためのチャンクのサイズを有する複数からなるデータチャンク内に前記客体を分類する段階と、
前記消去コーディングを用いてＰ個のパリティチャンクを計算する段階と、
前記データチャンク及び前記パリティチャンクを格納するメモリ装置を決定する段階と、
前記メモリ装置に前記データチャンク及び前記パリティチャンクを書き込む段階と、を含み、
前記客体が小型として分類される場合、データ及び複製版のためのメモリ装置を決定する段階と、
前記メモリ装置に前記データ及び前記複製版を書き込む段階と、を含むことを特徴とする方法。
【請求項１４】
前記客体は、大型でも小型でもない場合、中型として分類され、
複製又は前記消去コーディングは、前記複数からなるメモリ装置で利用可能な１つ以上の空間、前記仮想装置に格納された前記客体に対するアクセス時間、及び前記客体がアクセスされる頻度に基づいて適用されることを特徴とする請求項１３に記載の方法。
【請求項１５】
前記客体の中の第１客体及び第２客体のそれぞれに対応するパリティチャンクは、前記メモリ装置の固定されたサブセットに格納されることを特徴とする請求項１３に記載の方法。
【請求項１６】
前記客体の中の第１客体及び第２客体のそれぞれに対応するパリティチャンクは、前記メモリ装置の可変されるサブセットに格納されることを特徴とする請求項１３に記載の方法。
【請求項１７】
前記客体の中の第１客体及び第２客体のそれぞれに対応するデータ及び複製版は、前記メモリ装置の中の異なる装置に格納されることを特徴とする請求項１３に記載の方法。
【請求項１８】
前記複数からなるデータチャンクの中の少なくとも１つは、ゼロでパディングされることを特徴とする請求項１３に記載の方法。
【請求項１９】
仮想装置階層によって、各々が許容可能な最小サイズの値を有する複数からなるメモリ装置を含む仮想装置から、キーと共に、Ｏ、Ｐ、Ｓ、及びｍに基づき小型、中型、又は大型の客体を読み取る方法であって、
ここで、Ｏは客体のサイズ、Ｐはパリティ装置の個数、Ｓはデータ装置の個数、ｍは前記複数からなるメモリ装置のそれぞれの許容可能な最小サイズの値の中の最大値を示し、
前記方法は、前記仮想装置階層によって、データ読み取り要請に対応するキーを受信する段階と、
前記仮想装置階層によって、前記メモリ装置の全てに読み取り要請を送信する段階と、
前記仮想装置階層によって、前記メモリ装置から応答を受信する段階と、を有し、
前記客体が大型又は中型である場合、前記仮想装置階層によって、データチャンク及びパリティチャンクを受信し、消去コーディングを利用して前記客体を復元し、
前記客体が小型である場合、前記受信された応答は、前記客体又は前記客体の複製版であることを特徴とする方法。

【請求項２０】
前記キーは、前記複数からなるメモリ装置の中から開始装置又は第１装置を決定するためのハッシュ（キー）を含むことを特徴とする請求項１９に記載の方法。

【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、キーバリューストレージシステムに関し、より詳細には、仮想装置階層を利用した複数のメモリ装置を含む仮想装置に対する客体の格納及び読み取り方法とこれを用いたストレージ装置に関する。
【背景技術】
【０００２】
従来のソリッドステートドライブ（Ｓｏｌｉｄ　Ｓｔａｔｅ　Ｄｒｉｖｅ：ＳＳＤ）は、通常ブロックインターフェースのみを使用し、複数配列の独立ディスク（Ｒｅｄｕｎｄａｎｔ　Ａｒｒａｙ　ｏｆ　Ｉｎｄｅｐｅｎｄｅｎｔ　Ｄｉｓｋ：ＲＡＩＤ）、消去コーディング（ｅｒａｓｕｒｅ　ｃｏｄｉｎｇ）又は複製（ｒｅｐｌｉｃａｔｉｏｎ）を通じてデータ信頼性を提供する。客体（ｏｂｊｅｃｔ）フォーマットの大きさが多様になり非構造化されると共に、客体とブロックレベルのインターフェース間の効果的なデータ変換に対する要求がある。また、空間効率性及び速いアクセス時間（ａｃｃｅｓｓ　ｔｉｍｅ）の特性を維持しつつも、データ信頼性を保証することが好ましい。
【発明の概要】
【発明が解決しようとする課題】
【０００３】
本発明は、上記従来技術に鑑みてなされたものであって、本発明の目的は、仮想装置階層を利用した複数のメモリ装置を含む仮想装置に対する客体の格納及び読み取り方法とこれを用いたストレージ装置を提供することにある。
【課題を解決するための手段】
【０００４】
本発明は、ブロック装置とは異なるキーバリューストレージシステム（例：キーバリューソリッドステートドライブ）に関するものである。
【０００５】
一部の実施形態は、キーバリューソリッドステートドライブに対してデータ信頼性を具現する方法に関するものである。空間オーバーヘッド（ｓｐａｃｅ　ｏｖｅｒｈｅａｄ）に基づいて、複製及び消去コーディングのハイブリッドがキーバリューソリッドステートドライブ群に適用され、これは客体に対するステートレス（ｓｔａｔｅｌｅｓｓ）の可変長消去コードを具現する。
【０００６】
一部の実施形態は、以下の特徴のうちの１つ以上の特徴を有する。１）信頼性は、固定ブロック毎ではなく、可変客体毎に提供される。２）単一のディスク群に対する客体の目標信頼性を具現するために、複製及び消去コーディングを混合して適用する。３）客体に対する好ましい技法の決定において、空間効率性は第１メトリック（ｐｒｉｍａｒｙ　ｍｅｔｒｉｃ）であり、性能は第２メトリックである。４）メカニズムは、複数配列の独立ディスク（ＲＡＩＤ）に類似するステートレスである。５）複製又は消去コーディングのための更なる情報が格納される必要はない。６）客体のサイズに拘らず、アップデートのための読み取り－修正－書き込み（ｒｅａｄ－ｍｏｄｉｆｙ－ｗｒｉｔｅ）動作の必要がない。
【０００７】
一部の実施形態は、キーバリューソリッドステートドライブ群の信頼性を具現する方法を提供する。また、本発明は、ブロック内のデータの一部がアップデートされた場合、ブ

ロック装置に発生する読み取り－修正－書き込み動作を避けることができる。本発明によると、信頼性は、ブロック（例：固定ブロック）毎ではなく、客体（例：可変客体）毎に提供されるためである。

【０００８】

上記目的を達成するためになされた本発明の一態様によるストレージ装置は、ステートレス（ｓｔａｔｅｌｅｓｓ）のデータ保護を利用する仮想装置として構成された複数からなるメモリ装置と、複数からなる客体のそれぞれのサイズに応じて、前記客体の一部に第１データ保護を適用し、前記客体の他の一部に第２データ保護を適用することにより、前記客体を格納するように前記仮想装置を管理する仮想装置階層と、を有する。

【０００９】

前記メモリ装置は、１つ以上のデータ装置及び１つ以上のパリティ（ｐａｒｉｔｙ）装置として構成され得る。

第１データ保護は、消去コーディング（ｅｒａｓｕｒｅ　ｃｏｄｉｎｇ）を含み、前記第２データ保護は、複製（ｒｅｐｌｉｃａｔｉｏｎ）を含み得る。

前記消去コーディングは、前記客体のうちの該当する客体が大型客体として分類される場合、データ保護のために利用され得る。

前記客体のうちの該当客体が不等式（（Ｐ＋１）×Ｏ　＞　（Ｓ＋Ｐ）×ｍ　ＡＮＤ　Ｏ　＞＝　Ｓ×ｍ）を満足する場合、前記大型客体として分類され得る。ここで、Ｏは客体のサイズ、Ｐはパリティ装置の個数、Ｓはデータ装置の個数、ｍは許容可能な最小サイズの値を示す。

前記複製は、前記客体のうちの該当客体が小型客体として分類される場合、データ保護のために利用され得る。

前記客体のうちの該当客体が不等式（（Ｐ＋１）×Ｏ　＝＜　（Ｓ＋Ｐ）×ｍ）を満足する場合、小型客体として分類され得る。ここで、Ｏは客体のサイズ、Ｐはパリティ装置の個数、Ｓはデータ装置の個数、ｍは許容可能な最小サイズの値を示す。

前記消去コーティング又は前記複製のいずれかは、前記客体のうちの該当客体が大型客体としても小型客体としても分類されない場合、性能メトリック（ｐｅｒｆｏｒｍａｎｃｅ　ｍｅｔｒｉｃｓ）及びデータ使用特性に基づいて、データ保護のために利用され得る。

前記複数の客体のうちの該当客体が不等式（（Ｐ＋１）×Ｏ　＞　（Ｓ＋Ｐ）×ｍ）＞　Ｓ×ｍ　＞　Ｏ）を満足する場合、中型客体として分類され得る。ここで、Ｏは客体のサイズ、Ｐはパリティ装置の個数、Ｓはデータ装置の個数、ｍは許容可能な最小サイズの値を示す。

前記パリティ装置は、１つ以上の大型客体を格納する場合、固定され得る。

前記パリティ装置は、１つ以上の大型客体を格納する場合、循環（ｒｏｔａｔｅ）し得る。

前記メモリ装置は、ソリッドステートドライブ（Ｓｏｌｉｄ　Ｓｔａｔｅ　Ｄｒｉｖｅ）を含み得る。

【００１０】

上記目的を達成するためになされた本発明の一態様による仮想装置階層を利用して複数からなるメモリ装置を含む仮想装置に複数からなる客体を格納する方法は、前記仮想装置階層によって、前記客体のうちの該当客体が大型であるか又は小型であるかを判定する段階と、前記客体のうちの該当客体が大型として分類される場合、消去コーディングのためのチャンク（ｃｈｕｎｋ）のサイズ及び前記客体のうちの該当客体の１つ以上のデータチャンクのパディング（ｐａｄｄｉｎｇ）量を決定する段階と、前記消去コーディングを用いてＰ個のパリティチャンクを計算する段階と、前記データチャンク及び前記パリティチャンクを格納するメモリ装置を決定する段階と、前記メモリ装置に前記データチャンク及び前記パリティチャンクを書き込む段階と、前記客体のうちの該当客体が小型として分類される場合、複製を通じて生成されたデータ及び複製版（ｒｅｐｌｉｃａｓ）のためのメモリ装置を決定する段階と、前記メモリ装置に前記データ及び前記複製版を書き込む段階

と、を有する。
【００１１】
前記客体のうちの該当客体は、大型でも小型でもない場合、中型として分類され、前記複製又は前記消去コーディングは、性能メトリック及びデータ使用特性に基づいて適用され得る。
前記客体のうちの少なくとも２つの客体に対応するパリティチャンクは、前記メモリ装置の固定されたサブセット（ｓｕｂｓｅｔ）に格納され得る。
前記客体のうちの異なる客体に対応するパリティチャンクは、前記メモリ装置の固定されたサブセットに格納され得ない。
前記客体のうちの少なくとも２つの客体に対応するデータ及び複製版は、前記メモリ装置のうちの異なる装置に格納され得る。
前記１つ以上のデータチャンクのうちの少なくとも１つは、ゼロでパディングされ得る。
【００１２】
上記目的を達成するためになされた本発明の一態様による仮想装置階層を利用して複数からなるメモリ装置を含む仮想装置からキー（ｋｅｙ）を含む客体を読み取る方法は、前記仮想装置階層によって、前記メモリ装置の全てに読み取り要請を送信する段階と、前記仮想装置階層によって、前記メモリ装置から応答を受信する段階と、を有し、前記客体が大型である場合、前記仮想装置階層によって、データチャンク及びパリティチャンクを受信し、消去コーティングを利用して前記客体を復元し、前記客体が小型である場合、前記データチャンクは、前記客体又は前記客体の複製版である。
【００１３】
前記キーは、前記複数のメモリ装置のうちから開始装置（ｓｔａｒｔ　ｄｅｖｉｃｅ）又は第１装置（ｐｒｉｍａｒｙ　ｄｅｖｉｃｅ）を決定するためのハッシュ（即ち、キーのハッシュ）を含み得る。
【発明の効果】
【００１４】
本発明によれば、空間オーバーヘッドに基づいて、大型客体の場合に消去コーディング、小型客体の場合に複製を用いたステートレスのハイブリッドが利用される。また、中型客体は、アクセスパターン等に基づいて消去コーディングと複製との間の切り換えが行われる。また、チャンクのサイズは客体毎に変わる。これにより、ブロック内のデータの一部がアップデートされた場合にブロック装置に発生する他の客体との空間共有による読み取り－修正－書き込み動作を避けることができ、空間効率性及び速いアクセス時間の特性を維持しながらデータ信頼性を保証することができる。
【図面の簡単な説明】
【００１５】
【図１】本発明の一実施形態によるキーバリューソリッドステートドライブの概略的な構成図である。
【図２】本発明の一実施形態による装置グループを含む仮想装置及び仮想装置における客体のストレージを示す概念図である。
【図３】本発明の一実施形態による仮想装置に客体を書き込む方法のフローチャートである。
【図４】本発明の一実施形態による共有パリティ方法を用いた図２の仮想装置における大型客体のストレージを示す概念図である。
【図５】本発明の一実施形態による専用パリティ方法を用いた図２の仮想装置における大型客体のストレージを示す概念図である。
【図６】本発明の一実施形態による図２の仮想装置における小型客体のストレージを示す概念図である。
【図７】本発明の一実施形態による仮想装置から客体を読み取る方法のフローチャートである。

【図８】本発明の一実施形態による仮想装置から客体を読み取る方法のフローチャートである。
【発明を実施するための形態】
【００１６】
以下、本発明を実施するための形態の具体例を、図面を参照しながら詳細に説明する。本発明は、多様な様相の実施形態を有することができ、本明細書で説明する実施形態に限定して解釈すべきではない。本明細書で説明する実施形態は、本発明が徹底且つ完全になるように例示として提供し、これを通じて本発明が属する技術分野の通常の知識を有する者に、本発明の様相及び特徴を十分に伝える。従って、本発明が属する技術分野の通常の技術者が本発明の様相及び特徴を完全に理解するに当たって、不要なプロセス、構成、及び技法について説明を省略する。特に言及しない限り、図面及び詳細な説明全般に亘って同一の参照符号は、同一の構成要素を示し、同一の構成要素に関する重複説明はしない。図面に示す構成要素、階層、及び領域は、明確性のために誇張することがある。
【００１７】
本明細書で本発明の特定の実施形態を示して説明しても、特許請求の範囲によって定義される本発明の思想や範囲、及びこれらの均等範囲から逸脱することなく、説明する実施形態について、ある変更及び変形が行われ得るということは、本発明が属する技術分野の通常の技術者にとって当然である。例えば、通常の技術者が理解するように、多様な図面に示す実施形態の特徴は、本発明の思想及び範囲から逸脱することなく、結合される。
【００１８】
本明細書において、「第１」、「第２」、「第３」等の用語は、多様な要素、構成、領域、階層、及び／又はセクションを説明するために用いられるが、このような用語によって、要素、構成、領域、階層、及び／又はセクションが制限されるべきではない。このような用語は、何れかの要素、構成、領域、階層、又はセクションを他の要素、構成、領域、階層、又はセクションと区分するために用いられる。従って、以下に記載する第１要素、構成、領域、階層、又はセクションは、本発明の思想及び範囲を逸脱することなく、第２要素、構成、領域、階層、又はセクションと指称することができる。
【００１９】
要素や階層が他の要素や階層「上に」、「に連結された」、又は「に結合された」と言及する場合、これは直接、「他の要素や階層上に」、「他の要素や階層に連結された」、又は「他の要素や階層に結合された」ことが可能である、或いは１つ以上の媒介要素や媒介階層が存在してもよいことが理解される。また、１つの要素や階層が２つの要素や階層の「間」にあると言及する場合、２つの要素や階層の間に１つの要素や階層があるか、又は１つ以上の媒介要素や媒介階層があってもよいということも理解される。
【００２０】
本明細書で用いる用語は、特定の実施形態を説明するための目的のみであって、本発明を制限しようとするものではない。本明細書で用いる単数形態の用語は、文脈上明らかに特に指示がない限り、複数形態の用語も含む。本明細書において、「含む」という用語を用いる場合、このような用語は、言及する特徴、数字、段階、動作、要素、及び／又は構成の存在を明示するが、１つ以上の他の特徴、数字、段階、動作、要素、構成、及び／又はこれらの集合の存在や付加を排除するものではない。本明細書で用いる「及び／又は」、「及び／若しくは」という用語は、挙げられる１つ以上の関連項目の任意の何れか１つ及び全ての組み合わせを含む。要素が挙げられた後に続く「少なくとも１つ」のような表現は、挙げられた要素の全目録を修飾するものであり、目録の個別の要素を修飾するものではない。
【００２１】
本明細書で用いる「実質的に」、「約」、及びこれと類似する用語は、近似（ａｐｐｒｏｘｉｍａｔｉｏｎ）を示す用語として用いられるものであり、程度（ｄｅｇｒｅｅ）を示す用語として用いられるものではなく、本発明が属する技術分野の通常の技術者に認識される測定値又は計算値に内在した偏差を説明するためのものである。また、本発明の実

施形態に関する説明で用いる「し得る及び／又はできる」という表現は、「本発明の１つ以上の実施形態」を言及するものである。本明細書で用いられる「用いる」及び「用いられる」という用語は、それぞれ「利用する」及び「利用される」という用語と同じ意味と見なされる。更に、「例示的な」という用語は、例示又は一例を指称する。

【００２２】

本明細書で説明する本発明の実施形態によると、電子／電気装置及び／又は任意の他の関連装置や構成（例：ホスト、ソリッドステートドライブ、メモリ装置、及び仮想装置階層）は、任意の適切なハードウェア、ファームウェア（例：特定用途向け集積回路）、ソフトウェア、又ははソフトウェア、ファームウェア、及びハードウェアの適切な組み合わせを利用して具現される。例えば、キーバリュー（Ｋｅｙ　Ｖａｌｕｅ：ＫＶ）のソリッドステートドライブ（Ｓｏｌｉｄ　Ｓｔａｔｅ　Ｄｒｉｖｅ：ＳＳＤ、以下「ＳＳＤ」と称する）、ホスト、ＳＳＤ、メモリ装置、及び仮想装置階層などの装置の多様な構成要素は、１つの集積回路（Ｉｎｔｅｇｒａｔｅｄ　Ｃｉｒｃｕｉｔ：ＩＣ）チップ又は個別のＩＣチップ上に形成される。また、このような装置の多様な構成要素は、フレキシブル印刷回路フィルム（ｆｌｅｘｉｂｌｅ　ｐｒｉｎｔｅｄ　ｃｉｒｃｕｉｔ　ｆｉｌｍ）、テープキャリアパッケージ（Ｔａｐｅ　Ｃａｒｒｉｅｒ　Ｐａｃｋａｇｅ：ＴＣＰ）、印刷回路基板（Ｐｒｉｎｔｅｄ　Ｃｉｒｃｕｉｔ　Ｂｏａｒｄ：ＰＣＢ）上に具現されるか又は１つの基板（ｓｕｂｓｔｒａｔｅ）上に形成される。更に、このような装置の多様な構成要素は、１つ以上のコンピューティング装置の１つ以上のプロセッサで実行され、コンピュータプログラムの命令（語）を実行し、本明細書に記載する多様な機能を行うための他のシステム構成要素と相互作用するプロセス又はスレッド（ｔｈｒｅａｄ）である。コンピュータプログラムの命令（語）はメモリに格納され、メモリは、例えばランダムアクセスメモリ（Ｒａｎｄｏｍ　Ａｃｃｅｓｓ　Ｍｅｍｏｒｙ：ＲＡＭ）又はフラッシュメモリ（例：ＮＡＮＤフラッシュメモリ）装置などの標準メモリ装置を用いるコンピューティング装置で具現される。コンピュータプログラムの命令（語）は、例えばＣＤ－ＲＯＭ、フラッシュドライブ等のような他の非一過性のコンピュータ読み取り可能な記録媒体に格納される。また、本発明が属する技術分野の通常の技術者は、本発明の実施形態の思想や範囲を逸脱することなく、多様なコンピューティング装置の機能が単一のコンピューティング装置に結合若しくは統合されるか、又は特定のコンピューティング装置の機能が１つ以上の他のコンピューティング装置に分散されるということを認識する。

【００２３】

本明細書で用いる技術用語及び科学用語を含む全ての用語は、特に定義しない限り、本発明が属する技術分野の通常の知識を有する者が一般的に理解するものと同様の意味を有する。また、通常用いられる辞典に定義されているような用語は、関連技術及び／又は本明細書の文脈上の意味と一致すると解釈されるべきであり、本明細書で明らかに定義しない限り、理想的又は過度に形式的な意味として解釈されるべきではない。

【００２４】

図１は、本発明の一実施形態によるキーバリューＳＳＤ１０の概略的な構成図である。本実施形態によるストレージシステム（又はストレージ装置）は、図１に示すように、１つ以上のキーバリューＳＳＤを含む。但し、本発明が図１に限定されるわけではない。

【００２５】

本実施形態によると、キーバリューＳＳＤ１０のキーバリューＡＰＩ１５は、従来のブロックマッピングを必要としないユーザーキーバリュー装置ドライバー２０で動作する。

【００２６】

本実施形態によると、キーバリューＳＳＤ１０を含むストレージシステムは、所望の信頼性（例：目標信頼性）を得るために、客体のそれぞれのサイズに応じて、第１データ保護（例：消去コーディング）を客体の一部に適用し、第２データ保護（例：複製）を客体の他の一部に適用するハイブリッドステートレスのデータ保護方法を利用する。このような方法により、信頼性を損なわずに空間効率的なソリューションが提供される。一部の実施形態において、キーバリューＳＳＤ１０自体がハイブリッドステートレスのデータ保護

方法を行うこともできるが、ストレージシステムによりハイブリッドステートレスのデータ保護方法が行われる場合には、例えば複数のドライブ（ＳＳＤ）に亘る管理（例：仮想装置階層の動作）がより容易になる。

【００２７】

本実施形態によると、客体は、空間効率性のために分類（ｃｌａｓｓｉｆｉｃａｔｉｏｎ）され、サイズに基づいて分類され、それぞれのサイズのクラス毎に相違するバックアップの方法が用いられる。

【００２８】

客体に対する消去コーディングの空間オーバーヘッドが客体に対する複製の空間オーバーヘッドよりも小さい場合、該当客体は、大型客体（ｌａｒｇｅ　ｏｂｊｅｃｔ）と見なされる。その場合、消去コーディングがより好ましい。消去コーディングは、空間フットプリント（ｆｏｏｔｐｒｉｎｔ）が少ないためである。言い換えると、客体が不等式（（Ｐ＋１）×Ｏ　＞　（Ｓ＋Ｐ）×ｍ　ＡＮＤ　Ｏ　＞＝　Ｓ×ｍ、ここで、Ｏは客体のサイズである）を満足する場合、大型客体と見なされる。上記の不等式と以下の不等式において、Ｏは客体のサイズ（即ち、客体サイズ）、Ｐはパリティ装置カウント（即ち、仮想装置におけるパリティ装置の個数）、Ｓはデータ装置カウント（即ち、仮想装置におけるデータ装置の個数）を意味し、ｍは許容可能な最小サイズの値（即ち、個々の装置における全ての最小値のサイズのうちの最大値）を意味する。例えば、本実施形態による「許容可能な最小サイズの値」は、任意の装置の最小値のサイズの要件に違反しないシステムにおいて、任意の装置に格納される値のサイズを指称する。それぞれの装置毎に、該当装置が支援する最小の客体のサイズを有する。本実施形態によると、客体は、全ての装置に対して同一のサイズに分割されるため、分割のサイズは、装置が支援する任意の最小のサイズよりも大きくなければならない。ｍよりも小さいサイズの客体を格納しようとする場合、少なくとも１つの装置は、客体を格納できなくなる。

【００２９】

言い換えると、以下の条件の両方満足する場合、客体は大型客体と見なされる。条件１）客体のサイズＯにパリティ装置の個数よりも１つ大きい値（Ｐ＋１）をかけた値が、許容可能な最小サイズの値ｍにデータ装置の個数Ｓとパリティ装置の個数Ｐとの合計Ｓ＋Ｐをかけた値よりも大きい。そして、条件２）客体のサイズＯがデータ装置の個数Ｓに許容可能な最小サイズの値ｍをかけた値よりも大きい。

【００３０】

本実施形態によると、大型客体は消去コーディングされる。即ち、客体は、Ｓ個のチャンク（ｃｈｕｎｋ）（即ち、データチャンク又はＳ個のポーション（ｐｏｒｔｉｏｎ））に分割され、パリティチャンク（即ち、パリティポーション）は、Ｓ個のチャンクを用いて計算される。Ｓ個及びＰ個のチャンクのそれぞれは、本明細書で記載するように、対応する装置に格納される。

【００３１】

客体に対する複製の空間オーバーヘッドが客体に対する消去コーディングの空間オーバーヘッドよりも小さい場合、該当客体は小型客体（ｓｍａｌｌ　ｏｂｊｅｃｔ）と見なされる。その場合、複製が好ましい。複製は、よりよい読み取り性能を提供し、相対的に複雑な消去コーディングよりもアップデートをうまく処理できるためである。これは、アプリケーションのメタデータが小さくなる傾向があるという観測に鑑みても合理的である。言い換えると、客体が不等式（（Ｐ＋１）×Ｏ　＝＜　（Ｓ＋Ｐ）×ｍ）を満足する場合、小型客体と見なされ、複製される。

【００３２】

即ち、客体のサイズＯにパリティ装置の個数よりも１つ大きい値（Ｐ＋１）をかけた値が、データ装置の個数Ｓとパリティ装置の個数Ｐとの合計Ｓ＋Ｐに許容可能な最小サイズの値ｍをかけた値よりも小さい場合、客体は小型客体と見なされる。

【００３３】

客体が小型又は大型として分類される一部のグレー領域（ｇｒａｙ　ａｒｅａ）が存在

する。例えば、客体が不等式（（Ｐ＋１）×Ｏ　＞　（Ｓ＋Ｐ）×ｍ　＞　Ｓ×ｍ　＞　Ｏ）を満足する場合、客体は中型客体（ｍｅｄｉｕｍ　ｏｂｊｅｃｔ）と見なされ、性能メトリック（ｐｅｒｆｏｒｍａｎｃｅ　ｍｅｔｒｉｃｓ）（例：空間対アクセス時間）及び／又はデータ使用特性（例：アップデートの頻度）に基づいて、複製や消去コーディングが用いられる。

【００３４】

言い換えると、客体のサイズＯにパリティ装置の個数よりも１つ大きい値（Ｐ＋１）をかけた値が、データ装置の個数Ｓとパリティ装置の個数Ｐとの合計に許容可能な最小サイズの値ｍをかけた値よりも大きく、データ装置の個数Ｓとパリティ装置の個数Ｐとの合計に許容可能な最小サイズの値ｍをかけた値がデータ装置の個数Ｓに許容可能な最小サイズの値ｍをかけた値よりも大きく、データ装置の個数Ｓに許容可能な最小サイズの値ｍをかけた値が客体のサイズＯよりも大きい場合、客体は中型客体と見なされる。

【００３５】

例えば、性能がより重要であり、客体が頻繁にアップデートされる場合、複製がよりよい選択になる。このような場合、中型客体は小型客体として分類される。例えば、不等式（（Ｐ＋１）×Ｏ　＝＜　（Ｓ＋Ｐ）×ｍ）　ＯＲ　（（Ｐ＋１）×Ｏ　＞　（Ｓ＋Ｐ）×ｍ）　ＡＮＤ　Ｓ×ｍ　＞　Ｏ、即ち（（Ｐ＋１）×Ｏ　＜＝　（Ｓ＋Ｐ）×ｍ　ＯＲ　Ｏ　＜　Ｓ×ｍ）を満足する場合、客体は、本実施形態において小型として分類される。

【００３６】

他の例として、空間効率性がより重要な場合、消去コーディングが用いられる。このような場合、中型客体は大型客体として分類される。例えば、不等式（（Ｐ＋１）×Ｏ　＞　（Ｓ＋Ｐ）×ｍ　ＡＮＤ　Ｏ　＞＝　Ｓ×ｍ）　ＯＲ　（（Ｐ＋１）×Ｏ　＞　（Ｓ＋Ｐ）×ｍ）　＞　Ｓ×ｍ　＞　Ｏ　＝　（（Ｐ＋１）×Ｏ　＞　（Ｓ＋Ｐ）×ｍ）、即ち（（Ｐ＋１）×Ｏ　＞　（Ｓ＋Ｐ）×ｍ）を満足する場合、客体は本実施形態において大型として分類される。

【００３７】

図２は、本発明の一実施形態による装置グループを含む仮想装置及び仮想装置における客体のストレージを示す概念図であり、装置グループＳＳＤ１、ＳＳＤ２、ＳＳＤ３、ＳＳＤ４、ＳＳＤ５、ＳＳＤ６を含む仮想装置２００及び仮想装置２００内の客体（大型客体２０２及び小型客体２０４）のストレージを示す。装置のうち、ＳＳＤ１、ＳＳＤ２、ＳＳＤ３、及びＳＳＤ４はデータ装置として構成され、ＳＳＤ５及びＳＳＤ６はパリティ装置として構成される。図２には、例示の目的として、４つのデータ装置（ＳＳＤ１、ＳＳＤ２、ＳＳＤ３、ＳＳＤ４）及び２つのパリティ装置（ＳＳＤ５、ＳＳＤ）６のみを図示しているが、仮想装置２００内のデータ及びパリティ装置の個数は、これに限定されない。更に、他のＳＳＤがデータ装置及びパリティ装置として構成される。

【００３８】

例えば、仮想装置２００は、Ｓ個のデータ装置及びＰ個のパリティ装置の全体を含み、パリティ装置は、固定（ｆｉｘｅｄ）又は循環する（図２を参照すると、例としてＳ値に４、Ｐ値に２が用いられる）。例えば、パリティ装置が循環する場合、異なる大型客体の全てのパリティチャンク（又はパリティポーション）がパリティ装置のうちの同一の装置に格納されずに、データ装置の一部は、１つ以上の大型客体のためのパリティ装置として用いられる。言い換えると、パリティ装置が固定された場合は、客体に対応する「Ｐ」個のパリティチャンクがメモリ装置の同一の「Ｐ」セットに格納されるのに対し、パリティ装置が循環する場合は、客体に対応する「Ｐ」個のパリティチャンクが必ずしもメモリ装置の同一のセットに格納されない。また、装置は、平面（ｆｌａｔ）又は階層的（ｈｉｅｒａｒｃｈｙ）構成で組織される。多数の装置に亘って拡散（ｓｐｒｅａｄ）又は複製された客体に対する開始装置は、キーのハッシュ値（ｈａｓｈ　ｖａｌｕｅ）によって決定される。

【００３９】

また、データ装置は、必要及び／又はユーザーの設計上の選択に応じて、パリティ装置として再構成され、その逆の場合も同様である。例えば、仮想装置２００の装置の個数は、目標信頼性に基づいて設定可能である。消去コーディングにおいて、Ｐ個の欠陥に耐えるための装置の総数は、データ装置の個数Ｓとパリティ装置の個数Ｐとの合計になる。複製において、Ｐ個の欠陥に耐える装置の総数は、Ｐ＋１になる。装置の容量は、互いに同一又は類似する。

【００４０】

本実施形態によると、仮想装置２００内の装置セット、又は仮想装置２００に対応する装置セットは、信頼性管理の単位となるグループを構成する。装置ＳＳＤ１、ＳＳＤ２、ＳＳＤ３、ＳＳＤ４、ＳＳＤ５、ＳＳＤ６のグループは、単一のサーバーやラック（ｒａｃｋ）内に存在するか、又はサーバーやラック全域に亘って存在し、階層アーキテクチャ又は平面アーキテクチャを有するように構造化される。

【００４１】

装置グループを含む仮想装置２００は、仮想装置階層２１０と称される階層によって管理されるため、装置グループは、単一の仮想装置として提供される。仮想装置階層２１０は、ステートレスである。仮想装置階層２１０は、実行の際（ｒｕｎｔｉｍｅ）に、客体の個数、使用可能な容量、及び／又は同類のものなどの装置の最小限のメタデータ情報をキャッシュして維持する。特に、本実施形態による仮想装置階層２１０は、キー情報を維持する必要がない（例：キーに対するマッピングがない）。仮想装置２００の容量は、全ての装置容量のうちの最小容量（例：図２の場合、ＳＳＤ１、ＳＳＤ２、ＳＳＤ３、ＳＳＤ４、ＳＳＤ５、及びＳＳＤ６の容量のうちの最小容量）にグループ内装置の個数をかけた値によって決定される。

【００４２】

仮想装置階層２１０は、各装置が処理可能な最小値のサイズ及び最大値のサイズを知る。仮想装置階層２１０は、仮想装置２００の最小値のサイズ及び最大値のサイズを決定する。例えば、本実施形態によると、個々の装置の全ての最小値のサイズｍ＿ｉのうちの最大値は、仮想装置２００の最小値のサイズｍとして定義され、個々の装置の全ての最大値のサイズＭ＿ｉのうちの最小値は、仮想装置２００の最大値のサイズＭとして定義される。他の実施形態において、仮想装置の最大値のサイズＭは、個々の装置の全ての最大値のサイズＭ＿ｉのうちの最小値にデータ装置の個数Ｓをかけた値によって定義される。

【００４３】

一部の実施形態による仮想装置２００は、本発明の技術分野に属する通常の技術者に知られている任意の適切な消去コーディングアルゴリズムを利用し、使用可能なＭＤＳ（Ｍａｘｉｍｕｍ　Ｄｉｓｔａｎｃｅ　Ｓｅｐａｒａｂｌｅ）アルゴリズム（例：リードソロモン符号（Ｒｅｅｄ－Ｓｏｌｏｍｏｎ　ｃｏｄｅ））を用いる。図２に示すように、大型客体２０２には、「２」のパリティ値を有する消去コード（即ち、消去コーディングアルゴリズム）が適用され、パリティ１及びパリティ２が用いられる。

【００４４】

本実施形態によると、客体（例：大型客体２０２）は、Ｓ個のチャンクに分割され符号化される（データ及びパリティ装置（即ち、Ｓ＋Ｐ個の装置）に同一のサイズで分散される）。例えば、消去コーディングにより符号化された大型客体２０２は、データ１、データ２、データ３、データ４、パリティ１、及びパリティ２に分割される。客体が占める実際のストレージ空間は、本発明の実施形態において帯域（ｂａｎｄ）と指称する。帯域は、消去コーディングの場合、Ｓ＋Ｐ個の装置に亘って存在するが、複製の場合には、Ｐ＋１個の装置に亘って存在する。例えば、複製は、小型客体２０４に適用される。帯域は、客体を完全に含む（即ち、客体全体が帯域内に格納される）。一部の実施形態において、帯域は、客体のＳ個のチャンクが格納されるＳ個の装置に亘って存在する。

【００４５】

客体のサイズが装置の割当単位又は整列単位と合わない場合、帯域内の客体に対して割り当てられた余分の空間がパディングされる（例：「０」でパディングされる）。例えば

、図２は、大型客体２０２のデータ４が「０」でパディングされ、データ４の全てのデータビットを格納するために必要でない余分の空間を占めることを示している。帯域のサイズは、実施形態によって変わる。

【００４６】

図３は、本発明の一実施形態による仮想装置（例：図２、図４～６の仮想装置２００）に客体を書き込む方法のフローチャートである。図４は、本発明の一実施形態による共有パリティの方法を用いた仮想装置２００における大型客体（２４２、２４４）のストレージを示す概念図である。図５は、本発明の一実施形態による専用パリティの方法を用いた仮想装置２００における大型客体（２４２、２４４）のストレージを示す概念図である。図６は、本発明の一実施形態による仮想装置２００の小型客体（客体１（２６２）、客体２（２６４））のストレージを示す概念図である。

【００４７】

図３に示すように、段階３００で、仮想装置階層（例：図２、４～６の仮想装置階層２１０）は、キーを含むサイズＯの客体を仮想装置（例：図２、４～６の仮想装置２００）に書き込めというインストラクション（ｉｎｓｔｒｕｃｔｉｏｎｓ）又は命令（ｃｏｍｍａｎｄ）を（例えば、ホスト装置から）受信する。他の実施形態において、書き込みのインストラクション又は命令は、ホストによって提供された書き込みのインストラクションに応答して、仮想装置階層によって生成される。

【００４８】

段階３０２で、仮想装置階層は、上述のような不等式を用いて客体が大型か否かを判定する。例えば、（（Ｐ＋１）×Ｏ　＞　（Ｓ＋Ｐ）×ｍ　ＡＮＤ　Ｏ　＞＝　Ｓ×ｍ）である場合、客体は大型と見なされる。ここで、Ｏは客体のサイズ、Ｐはパリティ装置の個数、Ｓはデータ装置の個数、ｍは許容可能な最小サイズの値（即ち、個々の装置における全ての最小値のサイズのうちの最大値）を意味する。

【００４９】

段階３０２で客体が大型客体として分類された場合、段階３１２に示すように、仮想装置階層は、消去コーディングのためのデータチャンクのサイズを決定し、１つ以上のデータチャンクに対するパディングの量（例：「０」でパディング）を決定する。その後、客体は、パディングとの整列（ａｌｉｇｎｍｅｎｔ　ｗｉｔｈ　ｐａｄｄｉｎｇ）を考慮して同一サイズのＳ個のチャンクに分割され、段階３１４に示すように、本発明の技術分野に属する通常の技術者に知られている適切な消去コーディングアルゴリズムを利用して、Ｓ個のチャンクからＰ個のコードチャンク（即ち、Ｐ個のパリティチャンク）が生成（例：計算）される。

【００５０】

段階３１６で、仮想装置階層は、分散ポリシー（ｄｉｓｔｒｉｂｕｔｉｏｎ　ｐｏｌｉｃｙ）に基づいて、データチャンク及びパリティチャンクを格納するための装置（即ち、Ｓ個の装置及びＰ個の装置）を決定する。例えば、分散ポリシーは、キーのハッシュ値により客体に対する開始装置を決定すること、並びに／又は固定された装置及び／若しくはスポット（ｓｐｏｔ）にデータ及び／若しくはパリティチャンクを格納することを含む。段階３１８で、データチャンク及びパリティチャンクは、対応する装置に書き込まれる。例えば、Ｓ＋Ｐ個のチャンクは、Ｓ＋Ｐ個の装置（例：図２のＳＳＤ１、ＳＳＤ２、ＳＳＤ３、ＳＳＤ４、ＳＳＤ５、ＳＳＤ６）へのストレージのために分散される。例えば、図４に示す循環するパリティ装置において、キーのハッシュにより決定された装置でデータの書き込みが始まり、それぞれのブロック、即ちチャンク（及びパリティブロック、即ちパリティチャンク）は、第１装置の最初のデータから始まって順に書き込まれる。例えば、図５に示すように、固定されたパリティ装置において、全てのデータ及びパリティブロック（即ち、チャンク）は、予め指定された装置に格納される。その場合、開始装置もまた、予め指定される。小型客体の場合、開始装置及び複製装置もまた、キーをハッシュすることにより決定される。

【００５１】

図４に示すように、パリティ装置は、共有（即ち、循環）される。即ち、格納された大型客体に応じて、データチャンクを格納するためのデータ装置又はパリティチャンクを格納するためのパリティ装置の両方に単一の装置が用いられる。例えば、客体２４２は、データ１、データ２、データ３、（「０」でパディングされた）データ４、パリティ１、及びパリティ２に分割され、客体２４４もまた、データ１、データ２、データ３、（「０」でパディングされた）データ４、パリティ１、及びパリティ２に分割される。図４から分かるように、大型客体２４２のデータ１、データ２、データ３、及びデータ４は、それぞれ仮想装置２００のＳＳＤ１、ＳＳＤ２、ＳＳＤ３、及びＳＳＤ４に格納されるのに対し、大型客体２４４のデータ１、データ２、データ３、及びデータ４は、それぞれ仮想装置２００のＳＳＤ６、ＳＳＤ１、ＳＳＤ２、及びＳＳＤ３に格納される。

【００５２】

また、客体２４２のパリティ１及びパリティ２がそれぞれ仮想装置２００のＳＳＤ５及びＳＳＤ６に格納されるのに対し、客体２４４のパリティ１及びパリティ２は、それぞれ仮想装置２００のＳＳＤ４及びＳＳＤ５に格納される。従って、合わせてＳ個のデータ装置及びＰ個のパリティ装置が存在しても、パリティ装置が循環するため、専用のパリティ装置はない。

【００５３】

図４に示す例とは異なり、図５は、専用パリティ装置、即ち仮想装置２００のＳＳＤ５及びＳＳＤ６を利用する具現例を示す。例えば、大型客体２４２及び大型客体２４４の両方のデータ１、データ２、データ３、（「０」でパディングされた）データ４、パリティ１、及びパリティ２は、それぞれ仮想装置２００のＳＳＤ１、ＳＳＤ２、ＳＳＤ３、ＳＳＤ４、ＳＳＤ５、及びＳＳＤ６に格納される。

【００５４】

循環するパリティの実施形態及び小型客体のために、客体に対する開始装置は、キーのハッシュ値によって決定される。例えば、図４の共有パリティ装置において、開始装置は、ハッシュ（キー）％（Ｓ＋Ｐ）によって決定される。その後、後続のデータ及びパリティチャンク（即ち、Ｓ＋Ｐ個のチャンク）が順に（ハッシュ（キー）＋　１）％（Ｓ＋Ｐ）、（ハッシュ（キー）＋　２）％（Ｓ＋Ｐ）、…、（ハッシュ（キー）＋　Ｓ　＋　Ｐ　－　１）％（Ｓ＋Ｐ）に書き込まれる。専用パリティ装置が存在する場合、（Ｓ＋Ｐ）の代わりにＳ個の装置が用いられる。

【００５５】

データ及びパリティチャンクが対応する装置に書き込まれた後、段階３２０で大型客体の書き込みプロセスが完了する。

【００５６】

段階３０２で、客体が大型と判定されない場合、プロセスは段階３０４に進み、客体が小型か否か（即ち、（（Ｐ＋１）×Ｏ　＝＜　（Ｓ＋Ｐ）×ｍ）を満足するか）に関して判定する。客体が小型と判定された場合、仮想装置階層は、複製を開始し、段階３０８で、分散ポリシーに基づき、どの装置を利用してデータ及び複製版を格納するかを決定する。例えば、分散ポリシーは、キーのハッシュ値によって客体に対する開始装置を決定すること、並びに／又は固定された装置及び／若しくはスポットにデータ及び／若しくは複製版を格納することを含む。その後、段階３１０で、データ及び複製版が対応する装置に書き込まれる。

【００５７】

本実施形態によると、パディングとの整列を考慮して、Ｐ＋１個の複製版（１つのデータコピー及びＰ個のパリティコピーを含む）が客体に対して生成され、複製版は、Ｐ＋１個の装置に分散される。例えば、図６に示すように、客体１（２６２）は３回複製され（１つのデータ及び２つの複製版を含む）、複製版はそれぞれ仮想装置２００のＳＳＤ１、ＳＳＤ２、及びＳＳＤ３に格納される。同様に、客体２（２６４）も３回複製され（１つのデータ及び２つの複製版を含む）、複製版はそれぞれ仮想装置２００のＳＳＤ３、ＳＳＤ４、及びＳＳＤ５に格納される。図６に示す例において、仮想装置２００は、合わせて

Ｓ個のデータ装置及びＰ個のパリティ装置を含む。また、客体１（２６２）及び客体２（２６４）の両方が小型客体であるため、図６に示す例において、消去コーディングは用いられない。

【００５８】

キーのハッシュ値を用いて、Ｓ＋Ｐ個の装置のうちの第１装置が選択される。Ｐ個の複製版は、ストレージ組織、性能、及び／又は同類のもの等に基づいて決定的に選択される。例えば、データは第１装置に格納され、複製版は、専用パリティ装置の使用に拘らず、（ハッシュ（キー）＋１）％（Ｓ＋Ｐ）、（ハッシュ（キー）＋２）％（Ｓ＋Ｐ）、…、（ハッシュ（キー）＋Ｐ）％（Ｓ＋Ｐ）、又は異なるノード、ラックに格納される。

【００５９】

再び図３を参照すると、段階３０４で客体が小型ではないと判定された場合、即ち客体が大型でもなく（段階３０２参照）小型でもない（段階３０４参照）場合、客体は中型客体（即ち、（Ｐ＋１）×Ｏ　＞　（Ｓ＋Ｐ）×ｍ　＞　Ｓ×ｍ　＞　Ｏ）と判定され、プロセスは段階３０６に進み、中型客体が小型客体として処理されるか否かを判定する。段階３０６で中型客体が小型客体として処理される場合、プロセスは段階３０８に進んで小型客体のストレージプロセスを開始し、段階３０６で中型客体が大型客体として処理される場合、プロセスは段階３１２に進んで大型客体のストレージプロセスを開始する。

【００６０】

図７及び図８は、本発明の一実施形態による仮想装置（例：図２、図４～６の仮想装置２００）から客体を読み取る方法のフローチャートである。仮想装置階層（例：図２、図４～６の仮想装置階層２１０）は、キー及び値のサイズなどの客体のメタデータを維持しないため、読み取る客体が小型か大型かについて知らない。従って、仮想装置階層は、客体のユーザーキーを用いて、全ての物理装置（即ち、Ｓ＋Ｐ個の装置）に読み取り要請を送信することで、読み取りプロセス７００を開始し、段階７０２に示すように、サブ－読み取り要請を全ての物理装置に送信する。段階７０４で、仮想装置階層は、装置から応答を受信する。ユーザー（例：ホスト）が要請する客体が大型である場合、段階７０６でエラーが存在しないと判定されると、Ｓ＋Ｐ個の全ての装置は、ユーザーキーを含む要請に対してそれぞれの応答を返信する。

【００６１】

例えば、読み取る客体が大型客体であり、エラーが無い場合、全ての装置（即ち、Ｓ＋Ｐ個の装置）は応答する。しかし、Ｎ個の装置にエラーがある場合、Ｓ＋Ｐ－Ｎ個の装置のみが応答する。仮想装置階層が同一サイズの任意のＳ個のチャンクを受信する限り（即ち、データチャンクＳの総数と同一のデータチャンクＳ及びパリティチャンクＰの任意の組み合わせ）、ユーザー客体を復元（ｒｅｂｕｉｌｄ）することができる。言い換えると、装置のパリティ個数（即ち、Ｐと同一の装置の個数）を超える失敗が発生しない限り、大型客体の場合、データが復元される。

【００６２】

受信したチャンクの総数がＳよりも小さいか又はチャンクのサイズが同一でない場合、エラーが存在する。全ての装置がＮＯＮ＿ＥＸＩＳＴエラーを返信するか又は復旧不能なエラーが発生した場合、存在しない客体の読み取りになる。

【００６３】

最初は、仮想装置階層が客体のタイプを知らないため、客体のタイプを「ＮＯＮＥ」として初期化する。段階７０８で判定されるように客体が大型である場合、段階７１８で客体タイプが判定される。段階７１８で客体タイプが「ＮＯＮＥ」である場合、段階７２０で客体タイプは大型に設定される。段階７１８で客体タイプが「ＮＯＮＥ」と判定されない場合、仮想装置階層は、段階７３２で客体タイプが大型であるか否かを判定する。段階７３２で客体タイプが大型でない場合、段階７３４に示すように、エラーと決定される。段階７２０で客体タイプが大型に設定された後、又は段階７３２で客体タイプが大型と判定された場合、仮想装置階層は、段階７２２で全てのデータチャンクを有するか否かを判定する。

【００６４】
段階７２２で仮想装置階層が全てのデータチャンクが受信されていると判定した場合、段階７３０に示すように、読み取りプロセスが完了する。一方、段階７２２で全てのデータチャンクが受信されていないと判定した場合、仮想装置階層は、段階７２４で全ての装置から応答が受信されているか否かを判定する。段階７２４で全ての装置が応答した場合、仮想装置階層は、段階７２６でデータチャンクを少なくともＳ個（全てのデータチャンク及びパリティチャンクをカウント）有するか否かを判定する。Ｓ個よりも小さい数のチャンクが受信された場合、仮想装置階層は、段階７３４に示すように、エラーと決定する。少なくともＳ個のチャンク（受信された全てのデータチャンク及びパリティチャンクをカウント）が正確に受信された場合、仮想装置階層は、段階７２８で消去コーディングアルゴリズムを用いてＳ個のチャンクで客体を復元し、段階７３０で読み取りプロセスを完了する。例えば、１つ以上の装置が予想外にオフラインであった場合、１つ以上の装置が応答しないこともある。従って、一部の実施形態において、全ての装置が応答しなくても、少なくともＳ個のチャンクが受信される限り、仮想装置階層は、段階７２８に示すように、客体の復元を進める。
【００６５】
仮想装置階層が段階７０８で客体が大型でないと判定した場合、プロセスは段階７１０に進み、客体タイプが「ＮＯＮＥ」か否かを判定する。段階７１０で客体タイプが「ＮＯＮＥ」である場合、段階７１２で客体タイプは小型に設定される。段階７１０で客体タイプが「ＮＯＮＥ」でない場合、段階７１６で客体タイプが小型か否かに関して判定される。その際、客体タイプが小型でない場合、段階７３４に示すように、エラーと決定する。段階７１２で客体タイプが小型に設定されるか、又は段階７１６で仮想装置階層が客体タイプを小型と判定した場合、仮想装置階層は、段階７１４で、受信されたチャンクが有効か否かを判定する。段階７１４で、受信されたチャンクが有効である場合、段階７３０に示すように、読み取りプロセスが完了する。
【００６６】
ユーザー（例：ホスト）が要請した客体が小型である場合、複製版（即ち、第１コピー（ｐｒｉｍａｒｙ　ｃｏｐｙ）及び複製版（ｒｅｐｌｉｃａｓ）のうちの１つ）を有するＰ＋１個の装置は、エラーが無いか否かを返信し、残りの装置は客体が存在しないことを知らせるエラーを返信する。段階７１４で仮想装置階層が任意の有効なチャンクを受信する限り、装置は客体を有する。全ての装置がＮＯＴ＿ＥＸＩＳＴエラーを返信する場合、このような客体は存在しない（又はエラーが存在する）。全ての装置が返信するのではなく、返信した全ての装置がＮＯＴ＿ＥＸＩＳＴを報告する場合、段階７３４に示すように、復旧不能なエラーが発生する。
【００６７】
仮想装置階層が段階７１４でチャンクが有効でないと判定した場合、段階７２４で全ての装置から応答が受信されているか否かに関して判定する。段階７２４で全ての装置から応答が受信されていない場合、仮想装置階層は、段階７０４で全ての装置から応答を得るように進め、図７に示すように、段階７０６でエラーが存在するか否か等を判定するプロセスを引き続き行う。
【００６８】
本実施形態によると、仮想装置階層は、各装置に全ての客体キーを列挙するように要請し、読み取りに失敗した場合に復元（ｒｅｃｏｎｓｔｒｕｃｔｉｏｎ）するために、全てのキーに対する概念的に全体的な順序を有する。仮想装置階層は、キーを順に１つずつ確認する。
【００６９】
客体が大型である場合、仮想装置階層は、ハッシュ（キー）を用いることにより、キーに対する開始装置を決定し、固定されたパリティ装置が用いられない場合は、開始装置の情報に基づいて、どのチャンクが生成されるべきか（データチャンク又はコードチャンク）を決定する。パリティ装置が用いられる場合には、どのチャンクが復元されなければな

らないかが明確である。新たな装置のチャンクは、大型客体の読み取りの場合と同様に、有効なチャンクで復元される。

【００７０】

客体が小型である場合、仮想装置階層は、ハッシュ（キー）を用いることにより、キーに対する第１装置を決定し、第１装置情報に基づいて、どの装置が複製版を有するかを決定する。新たな装置が複製版を有さなければならない場合、客体は新たな装置に書き込まれる。これは、装置の全ての客体を巡回し（ｖｉｓｉｔ）、失敗した装置が復元されるまで繰り返される。

【００７１】

このように、本発明の１つ以上の実施形態によると、空間オーバーヘッドに基づいて消去コーディング及び複製のステートレスのハイブリッドが利用される。また、中型客体は、例えばアクセス（ａｃｃｅｓｓ）パターン等に基づいて、消去コーディングと複製との間の切り換えが行われる。また、チャンクのサイズは、客体毎に変わる。また、他の客体との空間共有による読み取り－修正－書き込み動作はこれ以上必要ではない。

【００７２】

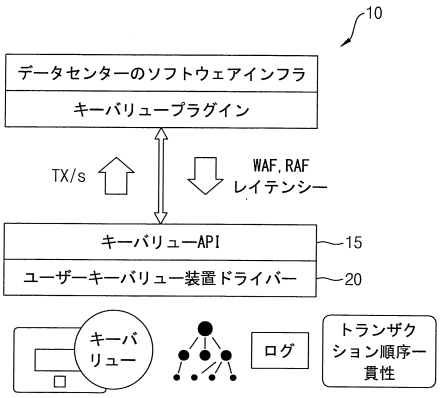
以上、本発明の実施形態について図面を参照しながら詳細に説明したが、本発明は、上述の実施形態に限定されるものではなく、本発明の技術的範囲から逸脱しない範囲内で多様に変更実施することが可能である。

【符号の説明】

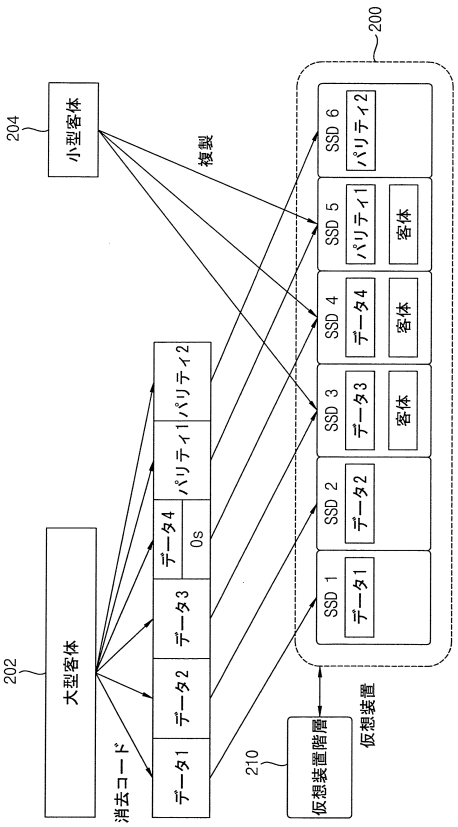
【００７３】

１０　キーバリューＳＳＤ
１５　キーバリューＡＰＩ
２０　ユーザーキーバリュー装置ドライバー
２００　仮想装置
２０２、２４２、２４４　大型客体
２０４　小型客体
２１０　仮想装置階層
２６２、２６４　客体１、客体２（小型客体）

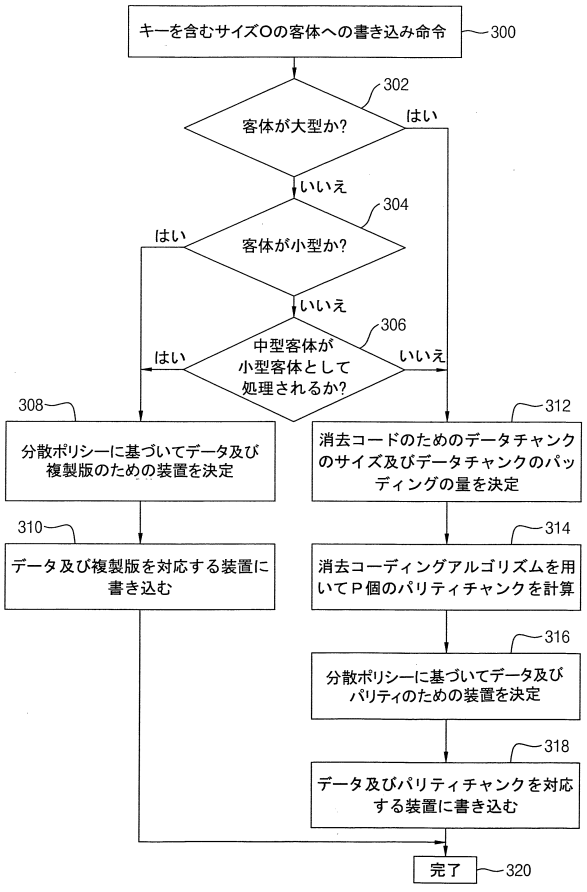
【図１】


10
データセンターのソフトウェアインフラ
キーバリュープラグイン
TX/s
WAF, RAF
レイテンシー
キーバリューAPI
15
ユーザーキーバリュー装置ドライバー
20
キーバリュー
ログ
トランザクション順序一貫性

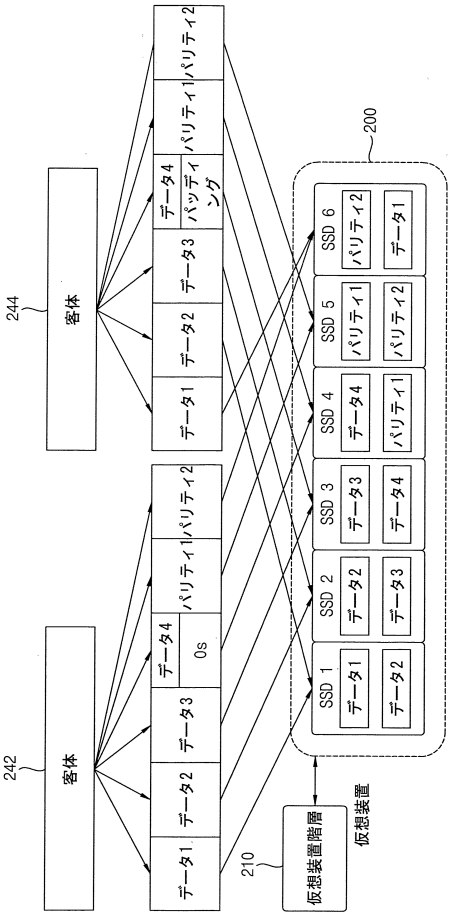
【図２】


200
204
小型客体
複製
SSD 6
パリティ2
SSD 5
パリティ1
客体
SSD 4
データ4
客体
SSD 3
データ3
客体
SSD 2
データ2
SSD 1
データ1
202
大型客体
データ4
0s
データ3
データ2
データ1
パリティ1
パリティ2
消去コード
210
仮想装置階層
仮想装置

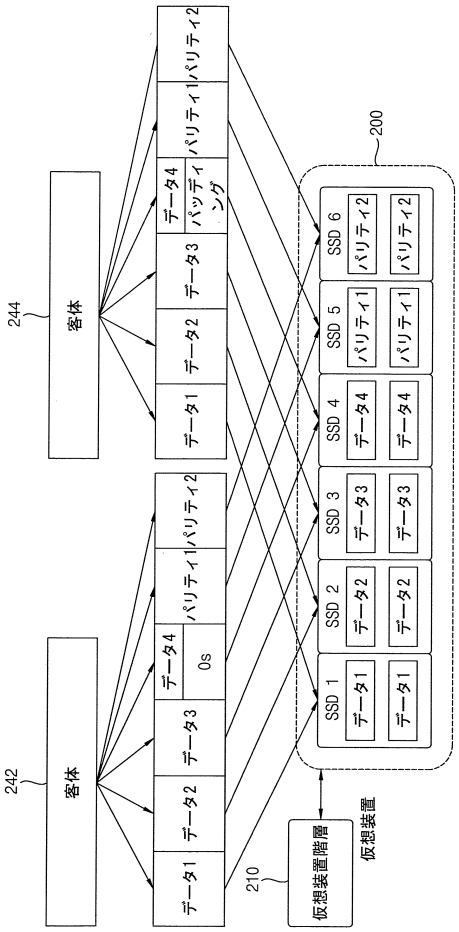
【図３】


キーを含むサイズOの客体への書き込み命令
300
客体が大型か?
302
はい
いいえ
客体が小型か?
304
はい
いいえ
中型客体が小型客体として処理されるか?
306
はい
いいえ
308
分散ポリシーに基づいてデータ及び複製版のための装置を決定
310
データ及び複製版を対応する装置に書き込む
312
消去コードのためのデータチャンクのサイズ及びデータチャンクのパッディングの量を決定
314
消去コーディングアルゴリズムを用いてP個のパリティチャンクを計算
316
分散ポリシーに基づいてデータ及びパリティのための装置を決定
318
データ及びパリティチャンクを対応する装置に書き込む
完了
320

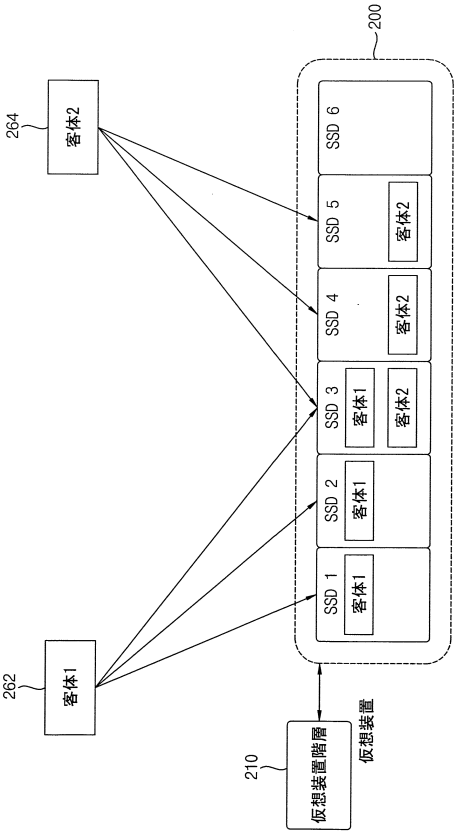
【図４】


200
244
客体
データ4
パッディング
データ3
データ2
データ1
パリティ1
パリティ2
SSD 6
パリティ2
データ1
SSD 5
パリティ1
パリティ2
SSD 4
データ4
パリティ1
SSD 3
データ3
データ4
SSD 2
データ2
データ3
SSD 1
データ1
データ2
242
客体
データ4
0s
データ3
データ2
データ1
パリティ1
パリティ2
210
仮想装置階層
仮想装置

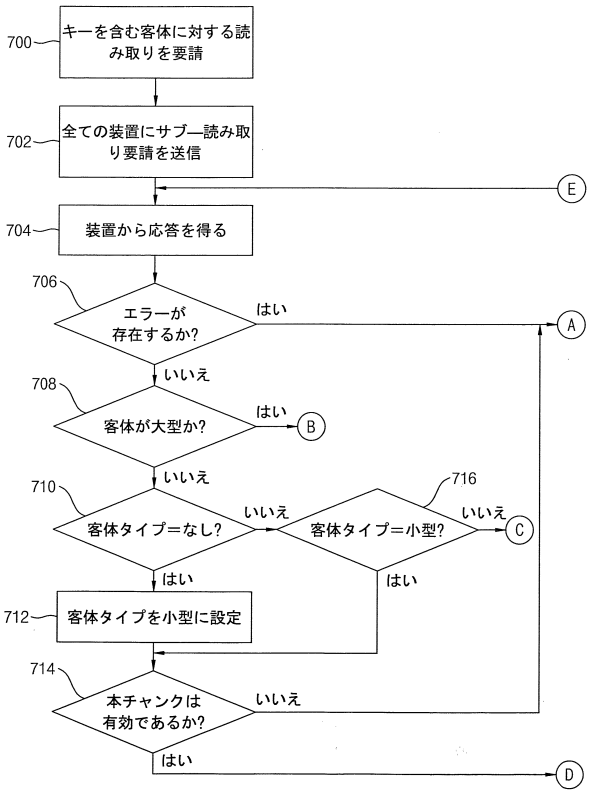
【図５】


244
客体
242
客体
データ1
データ2
データ3
データ4
パッディング
パリティ1
パリティ2
0s
200
SSD 1
SSD 2
SSD 3
SSD 4
SSD 5
SSD 6
210
仮想装置階層
仮想装置

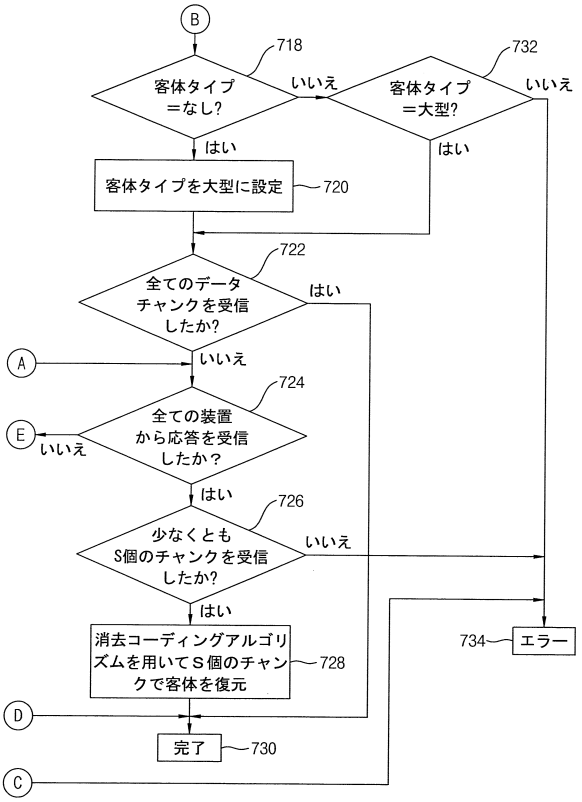
【図６】


264
客体2
262
客体1
200
SSD 1
SSD 2
SSD 3
SSD 4
SSD 5
SSD 6
客体1
客体2
210
仮想装置階層
仮想装置

【図７】


700
キーを含む客体に対する読み取りを要請
702
全ての装置にサブ—読み取り要請を送信
704
装置から応答を得る
706
エラーが存在するか?
はい
いいえ
708
客体が大型か?
はい
いいえ
710
客体タイプ＝なし?
いいえ
716
客体タイプ＝小型?
いいえ
はい
712
客体タイプを小型に設定
714
本チャンクは有効であるか?
いいえ
はい
A
B
C
D
E

【図８】


B
718
客体タイプ＝なし?
いいえ
はい
732
客体タイプ＝大型?
いいえ
はい
720
客体タイプを大型に設定
722
全てのデータチャンクを受信したか?
はい
いいえ
A
724
全ての装置から応答を受信したか？
E
いいえ
はい
726
少なくともS個のチャンクを受信したか?
いいえ
はい
728
消去コーディングアルゴリズムを用いてS個のチャンクで客体を復元
734
エラー
D
730
完了
C

フロントページの続き

(31)優先権主張番号 62/562,219
(32)優先日 平成29年9月22日(2017.9.22)
(33)優先権主張国・地域又は機関
米国(US)
(31)優先権主張番号 15/876,028
(32)優先日 平成30年1月19日(2018.1.19)
(33)優先権主張国・地域又は機関
米国(US)

早期審査対象出願

審査官 松平 英

(56)参考文献 特表２０１５－５１９６７４（ＪＰ，Ａ）
特表２０１５－５２０５８８（ＪＰ，Ａ）
特表２０１６－５００１８３（ＪＰ，Ａ）
国際公開第２０１６／１３７４０２（WO，Ａ１）
米国特許第０８５０４５３５（ＵＳ，Ｂ１）
井上陽治，第１０回 オブジェクトストレージの宿命、容量効率を高める方法とは？，onlilne，日本，ＩＴｍｅｄｉａ，2016年02月10日，［検索日 ２０２１年 5月１０日］、インターネット＜https://www.itmedia.co.jp/enterprise/articles/1602/10/news021.html＞
浅沼，ＨＤＦＳ Ｅｒａｓｕｒｅ Ｃｏｄｉｎｇの紹介とＹａｈｏｏ！ＪＡＰＡＮにおける運用事例，online，日本，Ｙａｈｏｏ！ＪＡＰＡＮ，2017年03月07日，［検索日 ２０２１年 5月１０日］、インターネット＜https://techblog.yahoo.co.jp/infrastructure/hdfs_erasure_coding/＞

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ 3／０６－３／０８
１２／０８－１２／１２８
１３／１０－１３／１４
１３／３８－１３／４２
１６／００－１６／９５８
１７／３０