

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7171452号
(P7171452)

(45)発行日 令和4年11月15日(2022.11.15)

(24)登録日 令和4年11月7日(2022.11.7)

(51)国際特許分類

F I

G 0 6 F 16/182 (2019.01)

G 0 6 F 16/182

G 0 6 F 3/06 (2006.01)

G 0 6 F 3/06 3 0 5 C

G 0 6 F 11/10 (2006.01)

G 0 6 F 11/10 6 1 2

G 0 6 F 16/11 (2019.01)

G 0 6 F 16/11

請求項の数 17 (全26頁)

(21)出願番号	特願2019-6709(P2019-6709)	(73)特許権者	390019839
(22)出願日	平成31年1月18日(2019.1.18)		三星電子株式会社
(65)公開番号	特開2019-128959(P2019-128959 A)		Samsung Electronics Co., Ltd.
(43)公開日	令和1年8月1日(2019.8.1)		大韓民国京畿道水原市靈通区三星路129
審査請求日	令和4年1月18日(2022.1.18)		129, Samsung-ro, Yeongtong-gu, Suwon-si, Gyeonggi-do, Republic of Korea
(31)優先権主張番号	15/876028	(74)代理人	110000051弁理士法人共生国際特許事務所
(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)	(72)発明者	ビチュマニ, レクハ
(31)優先権主張番号	62/635311		アメリカ合衆国 ヴァージニア州 22033, フェアファクス, ウェイソーン
(32)優先日	平成30年2月26日(2018.2.26)		
(33)優先権主張国・地域又は機関	米国(US)		
(31)優先権主張番号	15/967302		
	最終頁に続く		最終頁に続く

(54)【発明の名称】 キーバリュエータ信頼性システム、そのデータ格納方法、及び、その方法を具現するコンピュータコードを含む非一時的なコンピュータ読出し可能な媒体

(57)【特許請求の範囲】

【請求項1】

単一の論理ユニットとして1つの信頼性グループにグループ化されて仮想装置管理レイヤによって管理される1つ以上のストレージ装置を含むキーバリュエータ信頼性システムにおけるデータ格納方法であって、

前記仮想装置管理レイヤにより、前記データが前記データを格納するために客体複製を含む信頼性メカニズムに対応する閾値を満足することを決定する段階と、

前記仮想装置管理レイヤにより、メタデータに含まれて前記信頼性メカニズムを識別する信頼性メカニズム識別子によって識別される前記信頼性メカニズムを使用する段階と、

前記仮想装置管理レイヤにより、前記信頼性メカニズムに応じて前記データを格納する段階と、を有し、

前記仮想装置管理レイヤにより、前記データを格納する段階は、

K Vバリュエータを選択する段階と、

前記K Vバリュエータに対応するキーをハッシングするためのハッシュを計算する段階と、

前記K Vバリュエータに対応するキー客体のレプリカを格納するために前記1つ以上のストレージ装置の中のストレージ装置の部分集合を決定する段階と、

同一の使用者キー名の下で前記ストレージ装置の部分集合に前記K Vバリュエータに対応するアップデートされたバリュエータを書込む段階と、を含むことを特徴とする方法。

【請求項2】

前記閾値は、前記データの客体サイズ、前記データの処理能力考慮 (throughput)

ut consideration)、前記データの読出し／書込み温度(read/write temperature)、及び前記1つ以上のストレージ装置の基本(underlying)消去コーディング能力の中の1つ以上に基づくことを特徴とする請求項1に記載の方法。

【請求項3】

前記仮想装置管理レイヤにより、ブルームフィルタ又はキャッシュを使用して前記信頼性メカニズムに対する前記データをテストする段階を更に含むことを特徴とする請求項1に記載の方法。

【請求項4】

前記仮想装置管理レイヤにより、前記信頼性メカニズム、前記データを格納する前記1つ以上のストレージ装置に対するチェックサムを含む前記メタデータ、前記データを格納する前記1つ以上のストレージ装置に格納された前記データのバリュー客体のサイズ、及び前記1つ以上のストレージ装置の中のどこに前記データが格納されているのかを示す前記1つ以上のストレージ装置のパリティグループのメンバーの位置を記録するための前記メタデータを前記データに対応するキーと共に挿入する段階を更に含むことを特徴とする請求項1に記載の方法。

【請求項5】

前記信頼性メカニズムは、パッキング(packaging)を含み、
前記仮想装置管理レイヤにより、前記データを格納する段階は、

前記信頼性グループの前記1つ以上のストレージ装置の中の同一数の第1のそれぞれの1つ以上のストレージ装置にそれぞれ格納された1つ以上のキー客体を選択する段階と、

前記1つ以上のキー客体に対応する1つ以上のバリュー客体を取り出す段階と、

前記1つ以上のバリュー客体の仮想バリューサイズが同一になるように前記1つ以上のバリュー客体の中の最大のバリューサイズを有しないバリュー客体の終端に仮想ゼロをパディングする段階と、

前記1つ以上のキー客体から1つ以上のパリティ客体を生成する段階と、

前記1つ以上のキー客体を前記第1のそれぞれの1つ以上のストレージ装置に書込む段階と、

前記1つ以上のパリティ客体を前記1つ以上のストレージ装置の中の同一数の第2のそれぞれの1つ以上のストレージ装置に書込む段階と、を含み、

前記第2のそれぞれの1つ以上のストレージ装置は、前記第1のそれぞれの1つ以上のストレージ装置から峻別(distinct)され、

前記1つ以上のキー客体の数と前記1つ以上のパリティ客体の数との和は、前記1つ以上のストレージ装置の数と同一であることを特徴とする請求項1に記載の方法。

【請求項6】

前記信頼性メカニズムは、消去コーディングを使用するパッキングを含み、

前記1つ以上のストレージ装置は、MDS(maximum distance separable)消去コーディングで構成されることを特徴とする請求項5に記載の方法。

【請求項7】

前記信頼性メカニズムは、再生成消去コーディングを使用するパッキングを含み、

前記1つ以上のストレージ装置は、再生成消去コーディングで構成されることを特徴とする請求項5に記載の方法。

【請求項8】

前記信頼性メカニズムは、スプリットिंगを含み、

前記仮想装置管理レイヤにより、前記データを格納する段階は、

前記KVバリューを1つ以上の同一サイズの客体に分割する段階と、

前記1つ以上の同一サイズの客体から1つ以上のパリティ客体を生成する段階と、

前記ハッシュに基づいて前記1つ以上のストレージ装置の中の前記KVバリューが位置する主要ストレージ装置を決定する段階と、

前記主要ストレージ装置から始めて連続する順序で、1つ以上のバリュー客体を前記1

10

20

30

40

50

つ以上のストレージ装置の中の同一数の第1のそれぞれの1つ以上のストレージ装置に書込み、1つ以上のパリティ客体を前記1つ以上のストレージ装置の中の同一数の第2のそれぞれの1つ以上のストレージ装置に書込む段階と、を含み、

前記第2のそれぞれの1つ以上のストレージ装置は、前記第1のそれぞれの1つ以上のストレージ装置から峻別 (distinct) され、

前記1つ以上のバリュー客体の数と前記1つ以上のパリティ客体の数との和は、前記1つ以上のストレージ装置の数と同一であることを特徴とする請求項1に記載の方法。

【請求項9】

前記信頼性メカニズムは、消去コーディングを使用するスプリッティングを含み、

前記1つ以上のストレージ装置は、MDS (maximum distance separable) 消去コーディングで構成されることを特徴とする請求項8に記載の方法。

【請求項10】

前記信頼性メカニズムは、再生成消去コーディングを使用するスプリッティングを含み、

前記1つ以上のストレージ装置は、再生成消去コーディングで構成され、

前記仮想装置管理レイヤにより、前記データを格納する段階は、

前記再生成消去コーディングを使用して前記1つ以上の同一サイズの客体を1つ以上のサブパケットに分割する段階と、

前記1つ以上のパリティ客体を1つ以上のパリティサブパケットに分割する段階と、を更に含むことを特徴とする請求項8に記載の方法。

【請求項11】

信頼性メカニズムに基づいてデータを格納するためのデータ信頼性システムであって、ステートレスデータ保護を使用して仮想装置として構成される1つ以上のストレージ装置と、

客体複製を含む前記信頼性メカニズムに応じて前記1つ以上のストレージ装置を前記仮想装置として管理して前記1つ以上のストレージ装置にデータを格納するように構成された仮想装置管理レイヤと、を備え、

前記仮想装置管理レイヤは、前記データが前記データを格納するための前記信頼性メカニズムに対応する閾値を満足することを決定し、メタデータに含まれて前記信頼性メカニズムを識別する信頼性メカニズム識別子によって識別される前記信頼性メカニズムを使用し、前記信頼性メカニズムに応じて前記データを格納するように構成され、

前記仮想装置管理レイヤは、KVバリューを選択し、前記KVバリューに対応するキーをハッシングするためのハッシュを計算し、前記KVバリューに対応するキー客体のレプリカを格納するために前記1つ以上のストレージ装置の中のストレージ装置の部分集合を決定し、同一の使用者キー名の下で前記ストレージ装置の部分集合に前記KVバリューに対応するアップデートされたバリューを書込むことによって、前記データを格納するように構成されることを特徴とするデータ信頼性システム。

【請求項12】

前記信頼性メカニズムは、パッキングを含み、

前記仮想装置管理レイヤは、前記1つ以上のストレージ装置の中の同一数の第1のそれぞれの1つ以上のストレージ装置に格納された1つ以上のキー客体を選択し、前記1つ以上のキー客体に対応する1つ以上のバリュー客体を取り出し、前記1つ以上のバリュー客体の仮想バリューサイズが同一になるように前記1つ以上のバリュー客体の中の最大のバリューサイズを有しないバリュー客体の終端に仮想ゼロをパディングし、前記1つ以上のキー客体から1つ以上のパリティ客体を生成し、前記1つ以上のキー客体を前記第1のそれぞれの1つ以上のストレージ装置に書込み、前記1つ以上のパリティ客体を前記1つ以上のストレージ装置の中の同一数の第2のそれぞれの1つ以上のストレージ装置に書込むことによって、前記データを格納するように構成され、

前記第2のそれぞれの1つ以上のストレージ装置は、前記第1のそれぞれの1つ以上のストレージ装置から峻別され、

前記1つ以上のキー客体の数と前記1つ以上のパリティ客体の数との和は、前記1つ以

10

20

30

40

50

上のストレージ装置の数と同一であることを特徴とする請求項 1 1 に記載のデータ信頼性システム。

【請求項 1 3】

前記信頼性メカニズムは、スプリットングを含み、

前記仮想装置管理レイヤは、前記 K V バリユーを 1 つ以上の同一サイズの客体に分割し、前記 1 つ以上の同一サイズの客体から 1 つ以上のパリティ客体を生成し、前記ハッシュに基づいて前記 1 つ以上のストレージ装置の中の前記 K V バリユーが位置する主要ストレージ装置を決定し、前記主要ストレージ装置から始めて連続する順序で、1 つ以上のバリユー客体を前記 1 つ以上のストレージ装置の中の同一数の第 1 のそれぞれの 1 つ以上のストレージ装置に書込み、1 つ以上のパリティ客体を前記 1 つ以上のストレージ装置の中の同一数の第 2 のそれぞれの 1 つ以上のストレージ装置に書込むことによって、前記データを格納するように構成され、

前記第 2 のそれぞれの 1 つ以上のストレージ装置は、前記第 1 のそれぞれの 1 つ以上のストレージ装置から峻別 (distinct) され、

前記 1 つ以上のバリユー客体の数と前記 1 つ以上のパリティ客体の数との和は、前記 1 つ以上のストレージ装置の数と同一であることを特徴とする請求項 1 1 に記載のデータ信頼性システム。

【請求項 1 4】

前記信頼性メカニズムは、再生成消去コーディングを使用するスプリットングを含み、

前記 1 つ以上のストレージ装置は、再生成消去コーディングで構成され、

前記仮想装置管理レイヤは、前記再生成消去コーディングを使用して前記 1 つ以上の同一サイズの客体を 1 つ以上のサブパケットに分割し、前記 1 つ以上のパリティ客体を 1 つ以上のパリティサブパケットに分割することによって、前記データを格納するように更に構成されることを特徴とする請求項 1 3 に記載のデータ信頼性システム。

【請求項 1 5】

プロセッサ上で実行される場合、単一の論理ユニットとして 1 つの信頼性グループにグループ化されて仮想装置管理レイヤによって管理される 1 つ以上のストレージ装置を含むキーバリユー信頼性システムにおけるデータ格納方法を具現するコンピュータコードを有する非一時的なコンピュータ読み取り可能な記録媒体であって、

前記方法は、

前記仮想装置管理レイヤにより、前記データが前記データを格納するために客体複製を含む信頼性メカニズムに対応する閾値を満足することを決定する段階と、

前記仮想装置管理レイヤにより、メタデータに含まれて前記信頼性メカニズムを識別する信頼性メカニズム識別子によって識別される前記信頼性メカニズムを使用する段階と、

前記仮想装置管理レイヤにより、前記信頼性メカニズムに応じて前記データを格納する段階と、を有し、

前記仮想装置管理レイヤにより、前記データを格納する段階は、

K V バリユーを選択する段階と、

前記 K V バリユーに対応するキーをハッシングするためのハッシュを計算する段階と、

前記 K V バリユーに対応するキー客体のレプリカを格納するために前記 1 つ以上のストレージ装置の中のストレージ装置の部分集合を決定する段階と、

同一の使用者キー名の下で前記ストレージ装置の部分集合に前記 K V バリユーに対応するアップデートされたバリユーを書込む段階と、を含むことを特徴とする非一時的なコンピュータ読み取り可能な記録媒体。

【請求項 1 6】

前記信頼性メカニズムは、パッキングを含み、

前記仮想装置管理レイヤにより、前記データを格納する段階は、

前記信頼性グループの前記 1 つ以上のストレージ装置の中の同一数の第 1 のそれぞれの 1 つ以上のストレージ装置に格納された 1 つ以上のキー客体を選択する段階と、

前記 1 つ以上のキー客体に対応する 1 つ以上のバリユー客体を取り出す段階と、

前記１つ以上のバリュー客体の仮想バリューサイズが同一になるように前記１つ以上のバリュー客体の中の最大のバリューサイズを有しないバリュー客体の終端に仮想ゼロをパディングする段階と、

前記１つ以上のキー客体から１つ以上のパリティ客体を生成する段階と、

前記１つ以上のキー客体を前記第１のそれぞれの１つ以上のストレージ装置に書込む段階と、

前記１つ以上のパリティ客体を前記１つ以上のストレージ装置の中の同一数の第２のそれぞれの１つ以上のストレージ装置に書込む段階と、を含み、

前記第２のそれぞれの１つ以上のストレージ装置は、前記第１のそれぞれの１つ以上のストレージ装置から峻別（*distinct*）され、

前記１つ以上のキー客体の数と前記１つ以上のパリティ客体の数との和は、前記１つ以上のストレージ装置の数と同一であることを特徴とする請求項１５に記載の非一時的なコンピュータ読み取り可能な記録媒体。

【請求項１７】

前記信頼性メカニズムは、スプリットングを含み、

前記仮想装置管理レイヤにより、前記データを格納する段階は、

前記ＫＶバリューを１つ以上の同一サイズの客体に分割する段階と、

前記１つ以上の同一サイズの客体から１つ以上のパリティ客体を生成する段階と、

前記ハッシュに基づいて前記１つ以上のストレージ装置の中の前記ＫＶバリューが位置する主要ストレージ装置を決定する段階と、

前記主要ストレージ装置から始めて連続した順序で、１つ以上のバリュー客体を前記１つ以上のストレージ装置の中の同一数の第１のそれぞれの１つ以上のストレージ装置に書込み、１つ以上のパリティ客体を前記１つ以上のストレージ装置の中の同一数の第２のそれぞれの１つ以上のストレージ装置に書込む段階と、を含み、

前記第２のそれぞれの１つ以上のストレージ装置は、前記第１のそれぞれの１つ以上のストレージ装置から峻別（*distinct*）され、

前記１つ以上のバリュー客体の数と前記１つ以上のパリティ客体の数との和は、前記１つ以上のストレージ装置の数と同一であることを特徴とする請求項１５に記載の非一時的なコンピュータ読み取り可能な記録媒体。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明に係る開示（以下、単に本発明ともいう）は、一般的にデータストレージシステムに係り、さらに詳細には複数のキーバリューストレージ装置を含むキーバリュー信頼性システムにキーバリューデータを高信頼度の格納ができる信頼性メカニズムを選択する方法に係る。

【背景技術】

【０００２】

消去コーディング（*erasure coding*）のようなデータ信頼性メカニズム（*data reliability mechanisms*）は、複数のストレージ装置を含む多様な設備でストレージ装置の故障及び‘データ汚染’（*data corruptions*）によるデータ損失を解決するのに使用できる。

【０００３】

従来のソリッドステートドライブ（*SSD; solid state drives*）は一般的にブロックインタフェース（*block interface*）のみを使用し、*RAID (redundant array of independent disks)*方式を通じ、消去コーディング（*erasure coding*）又は複製（*replication*）を用いてデータ信頼性を提供できる。

客体フォーマットのサイズが多様になり、且つ構造化されない（*unstructured*）ので、客体とブロック間レベルのインタフェースにおいて効率的なデータ変換（*d*

10

20

30

40

50

ata conversion) が要求されている。さらに、スペース効率性及び高速アクセス時間特性を維持しながら、データ信頼性を確保する必要がある。

【0004】

RAIDのような技法は従来のブロックストレージ装置に対して広く研究されている。しかし、比較的新しいキーバリューストレージ装置は従来のブロック装置と異なるインタフェース及び異なるストレージセマンティクス (storage semantics) を有する。従って、多様な新しいキーバリューストレージ装置 (key-value storage device) にとり、キーバリューデータ及びキーバリューストレージ装置向けに特注されるか、又は特別に採用された新しいデータ信頼性技法が役立つ潜在的な可能性がある。

10

【先行技術文献】

【特許文献】

【0005】

【文献】米国特許第9, 594, 633号公報

米国特許第9, 639, 268号公報

【発明の概要】

【発明が解決しようとする課題】

【0006】

本発明の目的は単一の論理ユニットとして1つの信頼性グループにグループ化され、仮想装置管理レイヤによって管理されるN (但し、Nは整数) 個のストレージ装置を含むキーバリュー信頼性システムのデータ格納方法を提供することにある。

20

【課題を解決するための手段】

【0007】

本発明に係る実施形態は、その信頼性メカニズムが、仮想装置管理レイヤが故障したメモリ装置に存在する全てのキーを復旧し、新しいメモリ装置にコピーできるようにする単一キー復旧手続を各々遂行できるので、本文で記載された実施形態はメモリストレージ分野における改善に貢献できる。

【0008】

(1) 本発明の実施形態によれば、単一の論理ユニットとして1つの信頼性グループにグループ化され、仮想装置管理レイヤによって管理されるN (但し、Nは整数) 個のストレージ装置を含むキーバリュー信頼性システムのデータ格納方法が提供される。

30

前記方法は前記データが前記データを格納するための信頼性メカニズムに対応する閾値を満足するか否かを決定する段階、前記閾値を前記データが満足した場合、前記信頼性メカニズムを選択する段階、及び前記選択された信頼性メカニズムに応じて前記データを格納する段階を含む。

【0009】

(2) 前記閾値は前記データの客体サイズ、前記データの処理能力考慮 (throughput consideration)、前記データの読出し／書込み温度 (read/write temperature) (即ち、読出し／書込み頻度)、及び前記N個のストレージ装置の基本 (underlying) 消去コーディング能力の中で1つ又はそれ以上に基づく。

40

(3) 前記方法は1つ又はそれ以上のブルームフィルタ又はキャッシュを使用して前記信頼性メカニズムに対する前記データをテストする段階をさらに含む。

(4) 前記選択された信頼性メカニズム、前記データを格納する前記N個のストレージ装置の各々に対する1つ又はそれ以上のチェックサム、前記データを格納する前記N個のストレージ装置の各々に格納された前記データのバリュー客体のサイズ、及び前記N個のストレージ装置の中でどこに前記データが格納されているのかを示す前記N個のストレージ装置のパーティグループメンバーの位置を記録するためのメタデータを前記データに対応するキーと共に挿入する段階をさらに含む。

(5) 前記選択された信頼性メカニズムは客体複製を含み、前記データを格納する段階

50

はK Vバリューを選択する段階、前記選択されたK Vバリューに対応するキーをハッシングするためのハッシュを計算する段階、前記K Vバリューに対応するキー客体のレプリカを格納するために前記N個のストレージ装置の中でストレージ装置の部分集合を決定する段階、及び同一の利用者キー名の下で、前記決定されたストレージ装置の部分集合内の各々に前記K Vバリューに対応するアップデートされたバリューを書込む段階を含む。

【0010】

(6) 前記選択された信頼性メカニズムはパッキング (p a c k i n g) を含み、前記データを格納する段階は前記グループ化されてなる信頼性グループの前記N個のストレージ装置の中でk (但し、kは整数) 個のストレージ装置に格納されたk個のキー客体を選択する段階、前記k個のキー客体に対応するk個のバリュー客体を回収する段階、前記k個のバリュー客体全部の仮想サイズが同一になるように、前記k個のバリュー客体の中で最大のバリューサイズを有しないバリュー客体の終端に仮想ゼロをパディングする段階、前記k個のキー客体からr (但し、rは整数) 個のパリティ客体を生成する段階、前記k個のキー客体を前記k個のストレージ装置に書込む段階、及び前記r個のパリティ客体を前記N個のストレージ装置の中でr個のストレージ装置に書込む段階を含み、前記r個のストレージ装置の各々は前記k個のストレージ装置と峻別され (d i s t i n c t)、但し、 $k + r = N$ である。

10

【0011】

(7) 前記選択された信頼性メカニズムは伝統的な (t r a d i t i o n a l) 消去コーディングを使用するパッキングを含み、前記N個のストレージ装置は伝統的な (k、r) MDS (m a x i m u m d i s t a n c e s e p a r a b l e) 消去コーディングで構成される。

20

(8) 前記選択された信頼性メカニズムは再生成消去コーディングを使用するパッキングを含み、前記N個のストレージ装置は (k、r、d) 再生成消去コーディングで構成される。

【0012】

(9) 前記選択された信頼性メカニズムはスプリットティングを含み、前記データを格納する段階はK Vバリューを選択する段階、前記K Vバリューをk (但し、kは整数) 個の同一サイズの客体に分割する段階、前記k個の同一サイズの客体からr (但し、rは整数) 個のパリティ客体を生成する段階、前記選択されたK Vバリューに対応するキーをハッシングするためのハッシュを計算する段階、前記ハッシュに基づいて前記N個のストレージ装置の中で前記K Vバリューが位置する主要ストレージ装置を決定する段階、及び前記主要ストレージ装置から始めて連続する順序に前記k個の同一サイズの客体及び前記r個のパリティ客体の各々を前記N個のストレージ装置に書込む段階を含み、但し、 $k + r = N$ である。

30

【0013】

(10) 前記選択された信頼性メカニズムは伝統的な消去コーディングを使用するスプリットティングを含み、前記N個のストレージ装置は伝統的な (k、r) MDS (m a x i m u m d i s t a n c e s e p a r a b l e) 消去コーディングで構成される。

(11) 前記選択された信頼性メカニズムは再生成消去コーディングを使用するパッキングを含み、前記N個のストレージ装置は (k、r、d) 再生成消去コーディングで構成され、前記データを格納する段階は前記再生成消去コーディングを使用して前記k個の同一サイズのバリュー客体をm (但し、mは整数) 個のサブパケットに分割する段階、及び前記r個のパリティ客体の各々をm個のパリティサブパケットに分割する段階をさらに含む。

40

【0014】

(12) 本発明の他の実施形態によれば、選択された信頼性メカニズムに基づいてデータを格納するデータ信頼性システムが提供される。前記データ信頼性システムは選択された信頼性メカニズムに基づいてデータを格納するデータ信頼性システムにおいて、テートレスデータ保護を使用して仮想装置として構成されるN (但し、Nは整数) 個のストレージ

50

ジ装置、及び前記選択された信頼性メカニズムに応じて、前記N個のストレージ装置を前記仮想装置として管理して前記N個のストレージ装置の中で選択されたストレージ装置にデータを格納するように構成された仮想装置管理レイヤを含み、前記仮想装置管理レイヤは前記データが前記データを格納するための信頼性メカニズムに対応する閾値を満足するか否かを決定し、前記閾値を前記データが満足した場合、前記信頼性メカニズムを選択し、前記選択された信頼性メカニズムに応じて前記データを格納するように構成される。

【0015】

(13) 前記選択された信頼性メカニズムは客体複製を含み、前記仮想装置管理レイヤはKVバリューを選択し、前記選択されたKVバリューに対応するキーをハッシングするためのハッシュを計算し、前記KVバリューに対応するキー客体のレプリカを格納するために前記N個のストレージ装置の中でストレージ装置の部分集合を決定し、同一のユーザーキー名の下で前記決定されたストレージ装置の部分集合内の各々に前記KVバリューに対応するアップデートされたバリューを書込むことによって、前記データを格納するように構成される。

10

(14) 前記選択された信頼性メカニズムはパッキングを含み、前記仮想装置管理レイヤは前記N個のストレージ装置の中でk（但し、kは整数）個のストレージ装置に格納されたk個のキー客体を選択し、前記k個のキー客体に対応するk個のバリュー客体を回収し、前記k個のバリュー客体全ての仮想バリューサイズが同一になるように、前記k個のバリュー客体の中で最大のバリューサイズを有しないバリュー客体の終端に仮想ゼロをパディングし、前記k個のキー客体からr（但し、rは整数）個のパリティ客体を生成し、前記k個のキー客体を前記k個のストレージ装置に書き込み、前記r個のパリティ客体を前記N個のストレージ装置の中でr個のストレージ装置に書き込むことによって、前記データを格納するように構成され、前記r個のストレージ装置の各々は前記k個のストレージ装置の各々と峻別され、但し、 $k + r = N$ である。

20

【0016】

(15) 前記選択された信頼性メカニズムはスプリットティングを含み、前記仮想装置管理レイヤはKVバリューを選択し、前記KVバリューをk（但し、kは整数）個の同一サイズのバリュー客体に分割し、前記k個の同一サイズのバリュー客体からr個のパリティ客体を生成し、但し、rは整数、前記選択されたKVバリューに対応するキーをハッシングするためのハッシュを計算し、前記ハッシュに基づいて前記N個のストレージ装置の中で前記KVバリューが位置する主要ストレージ装置を決定し、前記主要ストレージ装置から始めて連続する順序に前記k個の同一サイズのバリュー客体及び前記r個のパリティ客体の各々を前記N個のストレージ装置に書き込むことによって、前記データを格納するように構成され、但し、 $k + r = N$ である。

30

(16) 前記選択された信頼性メカニズムは再生成消去コーディングを使用するスプリットティングを含み、前記N個のストレージ装置は(k、r、d)再生成消去コーディングで構成され、前記仮想装置管理レイヤは前記再生成消去コーディングを使用して前記k個の同一サイズの客体をm（但し、mは整数）個のサブパケットに分割し、前記r個のパリティ客体の各々をm個のパリティサブパケットに分割することによって、前記データを格納するようにさらに構成される。

40

【0017】

(17) 本発明のその他の実施形態によれば、プロセッサ上で実行される場合、単一の論理ユニットとして1つの信頼性グループにグループ化され、仮想装置管理レイヤによって管理されるN個のストレージ装置を含むキーバリュー信頼性システムにおいてデータを格納する方法を具現するコンピュータコードを含む非一時的なコンピュータ読み出し可能である媒体が提供される。前記方法は前記データが前記データを格納するための信頼性メカニズムに対応する閾値を満足するか否かを決定する段階、前記閾値を前記データが満足した場合、前記信頼性メカニズムを選択する段階、及び前記選択された信頼性メカニズムに応じてデータを格納する段階を含む。

【0018】

50

(18) 前記選択された信頼性メカニズムは客体複製を含み、前記データを格納する段階はKVバリューを選択する段階、前記選択されたKVバリューに対応するキーをハッシングするためのハッシュを計算する段階、前記KVバリューに対応するキー客体のレプリカを格納するために前記N個のストレージ装置の中でストレージ装置の部分集合を選択する段階、及び同一の利用者キー名の下で、前記KVバリューに対応するアップデートされたバリューを前記決定されたストレージ装置の部分集合内の各々に書込む段階を含む。

(19) 前記選択された信頼性メカニズムはパッキングを含み、前記データを格納する段階は前記信頼性グループの前記N個のストレージ装置の中でk（但し、kは整数）個のストレージ装置に格納されたk個のキー客体を選択する段階、前記k個のキー客体に対応するk個のバリュー客体を回収する段階、前記k個のバリュー客体全ての仮想バリューサイズが同一になるように前記k個のバリュー客体の中で最大のバリューサイズを有しないバリュー客体の終端に仮想ゼロをパディングする段階、前記k個のキー客体からr（但し、rは整数）個のパリティ客体を生成する段階、前記k個のキー客体を前記k個のストレージ装置に書込む段階、及び前記r個のパリティ客体を前記N個のストレージ装置の中でr個のストレージ装置に書込む段階を含み、前記r個のストレージ装置の各々は前記k個のストレージ装置と峻別され、 $k + r = N$ である。

(20) 前記選択された信頼性メカニズムはスプリットティングを含み、前記データを格納する段階はKVバリューを選択する段階、前記KVバリューをk（但し、kは整数）個の同一サイズのバリュー客体に分割する段階、前記k個の同一サイズのバリュー客体からr（但し、rは整数）個のパリティ客体を生成する段階、前記選択されたKVバリューに対応するキーをハッシングするためのハッシュを計算する段階、前記ハッシュに基づいて、前記N個のストレージ装置の中で前記KVバリューが位置する主要ストレージ装置を選択する段階、及び前記主要ストレージ装置から始めて連続した順序に、前記k個の同一サイズのバリュー客体及び前記r個のパリティ客体の各々を前記N個のストレージ装置に書込む段階を含み、 $k + r = N$ である。

【発明の効果】

【0019】

本発明の実施形態に係る信頼性メカニズムによれば、仮想装置管理レイヤが故障したメモリ装置に存在する全てのキーを復旧し、新しいメモリ装置にコピーできるようにする単一キー復旧手順を各々遂行できるので、メモリストレージの信頼性を向上させることができる。

【図面の簡単な説明】

【0020】

上述された思想及び／又はその他の思想は、添付された図面と組み合わせた以下の実施形態の詳細な記述からさらに明確になり且つ真価が分かるであろう。

【図1】本発明の実施形態に係る、選択された信頼性メカニズムに基づいてキーバリューデータを格納するキーバリュー信頼性システムを示すブロック図である。

【図2】本発明の実施形態に係る、キーバリューペアのデータのサイズに対応するサイズ閾値に基づいてキーバリュー信頼性システムによって使用される信頼性メカニズムの選択を示す順序図である。

【図3】本発明の実施形態に係る、伝統的な消去コーディングを使用するK-客体（k、r）消去コーディング、即ち多重客体“パッキング”の信頼性メカニズムに応じてキーバリューデータを格納するように構成されたKVストレージ装置のグループを示すブロック図である。

【図4】本発明の実施形態に係る、伝統的な消去コーディングを使用するK-客体（k、r）消去コーディング、即ち多重客体“パッキング”の信頼性メカニズムに応じるバリュー客体及びパリティ客体の格納を示すブロック図である。

【図5】本発明の実施形態に係る、伝統的な消去コーディングを使用する、“単一客体（Single_Object）（k、r）消去コーディング”、即ち“スプリットティング（Splitting）”の信頼性メカニズムに応じてキーバリューデータを格納するように構

成されたKVストレージ装置のグループを示すブロック図である。

【図6】本発明の実施形態に係る、再生成消去コーディングを使用する、“単一客体（Single_Object）（k、r、d）消去コーディング”、即ち“スプリッティング（Splitting）”の信頼性メカニズムに応じてキーバリュースデータを格納するように構成されたKVストレージ装置のグループを示すブロック図である。

【発明を実施するための形態】

【0021】

本発明の多様な特徴及び本発明を達成する方法は添付された図面及び実施形態に対する以下の詳細な記述を参照してさらに詳細に理解されることができる。以下で、実施形態は添付された図面を参照してさらに詳細に記述する。類似な参照番号は全体で類似な構成要素を指す。しかし、本発明が多様な他の形態に具現されることができ、ここに図解した実施形態に制限されることと理解されてはならない。むしろ、このような実施形態は本発明の全般的な理解を助けるための例示として提供され、当業者に本発明の技術的特徴及び様相を十分に伝達できるものである。従って、当業者が本発明の特徴及び様相を完全に理解するのに不必要なプロセス、要素、及び技法は記述しないことがある。異なりに言及されない限り、添付された図面及び記載された詳細な記述の全般に掛けて、類似な参照番号は類似な要素を指称し、それらの記述は繰り返さない。図面で、要素、階層、及び領域の相対的なサイズは明確性のために誇張されることができる。

【0022】

多様な実施形態が実施形態及び／又は中間構造の概略図である断面図を参照して本文で記述する。このように、例えば製造技術及び／又は許容誤差の結果として図解された形態からの変形が予想されなければならない。さらに、本文で記載された特定な構造的又は機能的な記述は本発明の思想に係る実施形態を記述するための単純な図解である。即ち、本文に記載された実施形態は領域の特定に図解された形態に制限されることと解釈されてはならず、例えば製造の結果である形状の偏差を含むべきである。例えば、長方形に図解されたインプラント領域は、伝統的に、インプラントから非インプラント領域への二進変化ではなく、それらのエッジでラウンドされたか、又は曲線の特徴及び／又はインプラント濃度の勾配を有することができる。同様に、インプラントによって形成された埋め込み領域は埋め込み領域とインプラントが発生する表面との間の領域で一部のインプラントを惹起することができる。即ち、図面に図解された領域は本質的に概略であり、それらの形状が装置の領域の実際形状を図解することを意図しなく、これを制限することではない。追加的に、当業者が認識することができるように、記載された実施形態は本発明の思想又は範囲から逸脱しなく、多様な他の方式に変形されることができる。

【0023】

以下の詳細な記述で、説明の便宜及び多様な実施形態の理解を助けるために、多数の特定記述が提供される。しかし、多様な実施形態が詳細な記述無しで又は1つ又はそれ以上の均等な代替物と共に具現されることができる。他の例で、広く公知された構造及び装置は多様な実施形態を不必要に曖昧にしないようにブロック図の形態に図示される。

【0024】

“第1（first）”、“第2（second）”、“第3（third）”等のような用語は多様な要素、構成、領域、階層、及び／又は区域を記述するために本文で使用されるが、このような要素、構成、領域、階層、及び／又は区域はこのような用語に限定されないことがよく理解されるべきである。このような用語は1つの要素、構成、領域、階層、又は区域を他の1つの要素、構成、領域、階層、又は区域から区分するためのみに使用される。即ち、以下で記載される第1要素、構成、領域、階層、又は区域は本発明の思想及び範囲から逸脱しなく、第2要素、構成、領域、階層、又は区域と称されることができる。

【0025】

“～の下（beneath、below、lower、under）”、“～の上（above、upper）”等のような空間的に相対的な用語（spatially relative terms）は図面で図解された他の1つの要素又は特徴と1つの要素又は特徴

10

20

30

40

50

との関連性を容易に説明するために本文で使用されることができる。空間的に相対的な用語は図面に図示された指向性に追加的に動作又は使用で装置の他の指向性 (orientations) を含むことと意図されることがよく理解される。例えば、図面で装置が裏返されている場合、他の要素又は特徴の“下 (below or beneath or under)”として記述した要素は他の要素又は特徴の“上 (above)”に向かうことになる。即ち、“下 (below、under)”の例示的な用語は上及び下の方向を全て含むことができる。装置が他の方向 (例えば、90°回転するか、又は他の方向) を指向することができ、本文で使用される空間的に相対的な記述はこれによって解釈されなければならない。類似に、第1部分が第2部分“上 (on)”に整列されたことと記述した場合、これは重力方向に基づいてその上部面に対する限定無しで、第1部分が第2部分の上部面又は下部面に整列されたことを示す。

10

【0026】

要素、階層、領域、又は構成が他の要素、階層、領域、又は構成と“連結された (on、connected to、又はcoupled to)”ことと指称される場合、他の要素、階層、領域、又は構成と直接的に (directly) 連結させるか、又は1つ又はそれ以上の中間要素、階層、領域、又は構成が存在することができる。しかし、“直接的に連結された”の用語は1つの構成要素が中間構成無しで他の構成要素と直接的に連結されることを指す。一方、“～間 (between、immediately between)”又は“隣接 (adjacent to又はdirectly adjacent to)”のように構成の間の関係を記述する他の表現は類似に解釈されることができる。追加的に、要素又は階層が2つの要素又は階層の間 (between) であることと指称される場合、要素又は構成の間にただ要素又は階層のみが存在するか、又は1つ又はそれ以上の中間要素又は階層がさらに存在することができることと理解されることができる。

20

【0027】

本文で使用される用語はただ特定の実施形態を記述するための例示的なものであり、本発明がこれに限定されることではない。本文で使用されるように、脈絡で明確に異なりに指称されない限り、単数用語は複数の形態を含むことと意図される。“含む”の用語が詳細な記述で使用される場合、列挙された特徴、整数、段階、動作、要素、及び／又は構成の存在を定義するが、1つ又はそれ以上の他の特徴、整数、段階、動作、要素、構成、及び／又はそれらのグループの存在又は追加を排除しない。本文で使用されるように、“及び／又は”の用語は関連されて列挙された目録の1つ又はそれ以上の全ての組合又は一部を含む。

30

【0028】

本文で使用されるように、“実質的な (substantial)”、“約 (about、approximately)”の用語及び類似な用語は近似の用語として使用され、程度の用語 (terms of degree) として使用されなく、当業者によって認識されることができる測定されるか、又は計算された値で固有な偏差を記述することと意図される。本文で使用される“約 (about、approximately)”の用語は言及された値を含み、特定の量の測定と連関された誤差及び問題の測定値 (例えば、測定システムの限界) を考慮した、当業者によって決定された特定の値に対する許容可能な偏差内であることを意味する。例えば、“約 (about)”は1つ又はそれ以上の標準偏差内であることを、又は言及された値の±30%、20%、10%、5%内であることを意味することができる。本文で使用されるように、“使用 (use、using、及びused)”の用語は“活用 (utilize、utilizing、及びutilized)”と同意語として看做されることができる。また“例示的な (exemplary)”の用語は“例 (example) 又は図解 (illustration)”を指称することと意図される。

40

【0029】

特定の実施形態が異なりに具現される場合、特定のプロセス順序が記載された順序と異なりに遂行されることができる。例えば、2つの連続して記述されたプロセスは実質的に同時に遂行されるか、又は記載された順序と反対の順序に遂行されることができる。

50

【0030】

本文に記載された本発明の実施形態に係る電気又は電子装置及び／又は他の連関された装置又は構成は適切なハードウェア、ファームウェア（例えば、アプリケーションー特定集積回路（ASIC；application-specific integrated circuit）、ソフトウェア、又はソフトウェア、ファームウェア、及びハードウェアの組合を使用して具現されることができる。例えば、このような装置の多様な構成は1つの集積回路（IC；integrated circuit）チップ又は別のICチップに形成されることができる。さらに、このような装置の多様な構成は軟性印刷回路フィルム（flexible printed circuit film）、テープキャリアパッケージ（TCP；tape carrier package）、印刷回路基板（PCB；printed circuit board）に具現できるか、又は1つの基板（substrate）上に形成されることができる。さらに、このような装置の多様な構成は本文で記述した多様な機能を遂行するために他のシステム構成と通信し、コンピュータプログラム命令語を実行する、1つ又はそれ以上のコンピューティング装置での、1つ又はそれ以上のプロセッサで駆動するプロセス又はスレッドである。コンピュータプログラム命令語はRAM（random access memory）のような標準メモリ装置を使用するコンピューティング装置に具現されることができるメモリに格納される。コンピュータプログラム命令語は、例えばCD-ROM、フラッシュドライブ等のような他の非一時的なコンピュータ読出し可能な媒体に格納されることができる。また、当業者は本発明の例示的な実施形態の思想及び様相からの逸脱しなく、多様なコンピューティング装置の機能が単一コンピューティング装置に組合又は集積されることができるか、又は特定コンピューティング装置の機能が1つ又はそれ以上の他のコンピューティング装置に分散されることを理解することができる。

【0031】

異なりに定義されない限り、本文で使用される技術的／科学的用語を含む全ての用語は本発明が属する技術分野での当業者によって共通的に理解されることと同一の意味を有する。共用辞書に定義されたことのような用語は連関された技術の脈絡及び／又は本発明の詳細な記述でそれらの意味と一貫された意味を有することと解釈されるべきであり、本文に明示的に定義されない限り、理想的であるか、又はあまりにも形式的な意味に解釈されてはならない。

【0032】

以下で記述するように、本開示発明の実施形態は単一の論理ユニットにグループ化された複数のキーバリュー（KV；key-value）ストレージ装置で構成されたキーバリュー信頼性システム（key-value reliability system）にキーバリューデータを信頼できるように格納する方法を提供する。さらに、本発明の実施形態はドライブを管理し、キーバリュー（KV；key-value）ペアの格納を制御するステートレスハイブリッド信頼性管理者（stateless hybrid reliability manager）を提供する。ステートレスハイブリッド信頼性管理者は、客体複製（Object Replication）；K-客体（k、r）消去コーディング-パッキング（K-Object（k、r）erasure coding-Packing）；単一客体（k、r）消去コーディング-スプリッティング（Single Object（k、r）erasure coding-Splitting）；K-客体（k、r、d）再生成コーディング-パッキング（K-Object（k、r、d）regeneration coding-Packing）；単一客体（k、r、d）再生成コーディング-スプリッティング（Single Object（k、r、d）regeneration coding-Splitting）を含む複数のプラグ可能な（pluggable）信頼性メカニズム／技法／具現に依存する。

【0033】

複数のプラグ可能な信頼性メカニズムに依存したステートレスハイブリッド信頼性管理者が装置を管理でき、KVペアの格納を制御でき、信頼性メカニズムの選択に対して開示

された方法が異なるサイズのKVペアの効率的な格納、回収、及び復旧を確保できるので、記載された実施形態はメモリストレージ（例えば、キーバリューストレージ装置内にキーバリューストレージデータの格納）の性能を向上できる。

【0034】

図1は本発明の実施形態に係る、選択された信頼性メカニズムに基づいてキーバリューストレージデータを格納するキーバリューストレージ信頼性システムを示すブロック図である。

【0035】

図1を参照すれば、上述されたように、多様な新しいキーバリューストレージ（KV）「ストレージ装置」{別名、「メモリ装置」「ドライブ」「KV-SSD」}(130)にとり、キーバリューストレージデータ及びKVストレージ装置130向けに特注される(tailored)か、又は特別に採用された新しいデータ信頼性メカニズムが役立つ潜在的な可能性がある。従って、このようなKVストレージ装置130のためのハイブリッドキーバリューストレージ信頼性システムは1つ又はそれ以上のプラグ可能信頼性メカニズム(pluggable reliability mechanisms)に応じてハイブリッド信頼性メカニズムを使用してKVストレージ装置130を管理し、それらにKVペアの格納を制御するステートレスハイブリッド信頼性管理者/仮想装置管理者レイヤ/仮想装置管理レイヤ120を含む。

たとえば、本文に記述されたKVストレージ装置を指すためにSSD(solid-state drives)が一般的に使用されるが、他のストレージ装置が本発明の実施形態に係って使用され得る。本発明の実施形態に係る仮想装置管理レイヤ120のデザイン及び動作は以下で記述される。

【0036】

現在の実施形態で、仮想装置管理レイヤ120はキーバリューストレージ信頼性システムにキーバリューストレージデータ/KVペア170を信頼性があるように格納する方法を可能にできる。キーバリューストレージ信頼性システムは単一の論理ユニット(single logical unit)にグループ化された複数のKVストレージ装置130を含む。単一の論理ユニットは信頼性グループ140と称される。

【0037】

信頼性グループ140のKVストレージ装置130はキーバリューストレージデータ170に対応する、消去コーディングされたデータ及び/又は複製されたデータの各々のチャンクを格納する。信頼性グループ140のKVストレージ装置130は、キーバリューストレージ動作が仮想装置管理レイヤ120を通じて指向される(bedirected to)、見かけ上は単一仮想装置110として露出される。

【0038】

仮想装置110は仮想装置管理レイヤ120としてステートレスハイブリッド信頼性管理者(stateless hybrid reliability manager)を含む。即ち、仮想装置管理レイヤ120はステートレス方式(即ち、キーバリューストレージと装置との間のマッピングを維持することを必要としない)により動作する。

【0039】

従って、仮想装置110はN個のKVストレージ装置130(但し、Nは整数)(例えば、KV-SSD130-1、130-2、130-3、130-4、…、130-N)を通じてキーバリューストレージデータ170を格納し、仮想装置管理レイヤ120を通じてKVストレージ装置130にキーバリューストレージデータ170を格納する。即ち、仮想装置管理レイヤ120はKVストレージ装置130を管理し、それにKVペアを格納することを制御する。

【0040】

他の実施形態で、キーバリューストレージ信頼性システムはキーバリューストレージデータ170のキーと関連されたメタデータ及び/又はデータを選択的に格納して動作速度を向上させるキャッシュをさらに含み得る。信頼性メカニズムはKVペアに対応するメタデータを含むようにKVペアのバリューストレージを添付できる。即ち、キー及びバリューストレージの両方はメタデータ識別子“MetaID”に対応する情報を添付されて、そのKVペアに特有の追加的なメタデータを格納

できる。メタデータはチェックサム (checksum)、データを格納するのに使用される信頼性メカニズムを識別するための信頼性メカニズム識別子、消去コード識別子、客
体サイズ、パリティグループ番号の位置等を含む。

【0041】

他の実施形態で、キーバリューストレージ信頼性システムは以下で記述する信頼性メカニズムに対応するブルームフィルタ (bloom filters) をさらに含み得る。ブルームフィルタは対応する信頼性メカニズムを使用して格納されたキーを格納でき、それによって、キーバリューストレージシステムを読み出し動作において支援できる。従って、キーバリューストレージシステムの1つ又はそれ以上のブルームフィルタ又はキャッシュは、現存する信頼性メカニズムに対してキーの速やかなテストを可能にできる。

10

【0042】

本文で記述する信頼性メカニズムの各々は、KVストレージ装置130の中で対応する1つの数字のキーモジュロ (key modulo) に対する同一のハッシュ関数を使用してキーバリューストレージデータ170に対応するKVペアの第1のコピー又はチャンクを先ず格納できる。即ち、プラグ可能な信頼性メカニズムの各々に対して、信頼性メカニズムは使用者キーと同一のキーを使用して少なくとも第1のコピー／チャンクを格納できる。

【0043】

上述されたように、本発明の実施形態は複数のKVストレージ装置130にキーバリューストレージデータ170の信頼性ある格納を保障する複数のプラグ可能な信頼性メカニズム (multiple pluggable reliability mechanisms) を提供する。従って、仮想装置管理レイヤ120は信頼性メカニズムを必要とし、信頼性メカニズムの中で何れが使用されるかを決定できる。

20

【0044】

信頼性メカニズムは仮想装置110のセットアップ中に設定されるバリュースize 閾値 (value-size thresholds)、及び／又は、客
体読み出し／書き込み頻度 (object read/write frequency) などの政策に基づき得る。従って、仮想装置管理レイヤ120はシステムの指定された政策に基づいて適切な信頼性メカニズムを選択することができる。

【0045】

本発明の実施形態の5つの信頼性メカニズムに対して、仮想装置管理レイヤ120によって信頼性メカニズムがどのように動作するか、そして信頼性メカニズムが何時適切に使用され、選択されるかを以下で記述する。この5つの信頼性メカニズムは、客
体複製 (Object Replication)、K客
体 (k, r) 消去コーディングパッキング (K-Object (k, r) erasure coding-Packing)、単
一客
体 (k, r) 消去コーディングスプリッティング (Single Object (k, r) erasure coding-Splitting)、K客
体 (k, r, d) 再生成コーディングパッキング (K-Object (k, r, d) regeneration coding-Packing)、及び単
一客
体 (k, r, d) 再生成コーディングスプリッティング (Single Object (k, r, d) regeneration coding-Splitting) と称され得る。

30

40

【0046】

図2は本発明の実施形態に係る、KVペアのデータのサイズに対応するサイズ閾値に基づいたキーバリューストレージシステムによって使用される信頼性メカニズムの選択を示す順序図 (200) である。

【0047】

図2を参照すれば、サイズ閾値に基づいた支援される信頼性メカニズムの全体 (例えば、5つの上述された信頼性メカニズム) に対して、仮想装置管理レイヤ120はデータ (例えば、キーバリューストレージデータ170) のバリュースize を決定でき、バリュースize が各々の信頼性メカニズムに対応する与えられた閾値 (t_i) より小さいか否かを決定でき、バリュースize 閾値の要求を満足する第1信頼性メカニズムを選択できる。

50

【0048】

例えば、S210で、仮想装置管理レイヤ120はサイズ閾値基盤の支援されている“n”(nは整数)個の信頼性メカニズムを受信する。S220で、仮想装置管理レイヤ120は1からnの順に1回に1つずつ、信頼性メカニズムの各々を単純に検討する。S230で、支援されるべき信頼性メカニズムの各々の検討で、仮想装置管理レイヤ120はデータのバリューサイズが各々の信頼性メカニズムに対応する閾値 t_i より小さいか否かを決定する。

【0049】

S240で、データのバリューサイズより大きいか、或いは同一である閾値 t_i を有する信頼性メカニズムが発見された場合、仮想装置管理レイヤ120はその信頼性メカニズムの使用を選択する。S250で、S240で使用される信頼性メカニズムが決定されたか、又はS220の最終反復でバリューサイズを満足する適切な閾値 t_i を有する信頼性メカニズムがN個の信頼性メカニズムの中で無いと決定された場合、仮想装置管理レイヤ120は使用されるべき信頼性メカニズムの決定を終了する。

【0050】

現在の実施形態で、“n”は本文に記載された実施形態の5つの異なる信頼性メカニズムに依拠して5に等しい。比較的に非常に小さいキーバリュー(即ち、バリューサイズが比較的小さい場合)に対して、仮想装置管理レイヤ120は客体複製(Object Replication)の信頼性メカニズムの使用を選択する。若干大きいキーバリューに対して、仮想装置管理レイヤ120はパッキング、そしてその次に、スプリッティング(Splitting)(例えば、順に)の信頼性メカニズムを選択する一方、各々に対して伝統的な消去コーディングを使用する。しかし、さらに大きいキーバリューに対して、仮想装置管理レイヤ120はパッキング、その次にスプリッティング(Splitting)を選択する一方、伝統的な消去コーディングの代わりに再生成消去コーディング(regeneration erasure coding)を使用する。

【0051】

本発明の実施形態で、使用のための信頼性メカニズムの選択は、客体の客体サイズ、客体に対する処理能力要求条件、対応するキーバリューペアの読出し／書込み温度(read/write temperature)、即ち読出し／書込み頻度、複数のKVストレージ装置の基本8 underlying)コーディング能力、及び／又はキーがホットであるか、又はコールドであるかに対する検出の中で1つ又はそれ以上に基づき得る。例えば、“ホット(hot)”キーは、それらのバリューサイズに拘わらず、客体複製(Object Replication)の信頼性メカニズムを使用する反面に、“コールド(cold)”キーはそれらのバリューサイズに依拠して消去コーディング方式の信頼性メカニズムの中で1つに適用される。他の実施形態として、客体複製(Object Replication)の信頼性メカニズムを使用するか否かに対する決定はサイズ及び読出し／書込み温度の両方に基づき得る。従って、客体読出し／書込み温度、即ち客体読出し／書込み頻度に対応する閾値が、図2の順序図200でサイズに対応する閾値の代わりに、信頼性メカニズムを決定するのに使用され得る。

【0052】

5つの信頼性メカニズムの各々の動作は以下で記述する。

【0053】

図1を再び参照すれば、以前に言及されたように、KVペアのバリューサイズが比較的小さい場合、“客体複製(Object Replication)”の信頼性メカニズムが仮想装置管理レイヤ120による選択として適切である。客体複製(Object Replication)は客体毎(例えば、キーバリューデータ/KVペア170毎)に適用される。客体複製(Object Replication)の信頼性メカニズムは、高いストレージオーバーヘッド(storage overhead)を有するが、読出し及び復旧費用が低いので、非常に小さいバリューサイズに対して適切である。

【0054】

客体複製 (Object Replication) の信頼性メカニズムは頻繁なアップデートを有するキーバリュー (例えば、キーバリューデータ 170) に対しても適切であり、従って読み出し及び書き込み頻度に基づいて選択される。

【0055】

客体複製 (Object Replication) の間に、書き込みが発生する毎に、キーバリューデータ 170 は 1 つ又はそれ以上の追加的な KV ストレージ装置 130 に複製される。キーバリューデータ 170 の主要コピー (primary copy) はキーモジュロ (N) のハッシュによって指定された KV ストレージ装置 130 の中の 1 つに配置される。キーバリューデータ 170 の主要コピーのレプリカ (複製、replicas) は循環的な方式により、順次連続した KV ストレージ装置 130、又は直接隣接する (immediately adjacent) KV ストレージ装置 130 に配置される。

10

【0056】

仮想装置管理レイヤ 120 又は使用者はキーバリューデータ 170 のコピーをどのくらい多く生成するかを決定する。例えば、仮想装置管理レイヤ 120 を使用する分散されたシステムは 3 ウェイ複製 (3-way replication) を選択し、3 ウェイ複製を既定方式に (default) 設定できる。しかし、システムの利用者は客体のレプリカの数を選択された規定値 (default) より多く或いは少なく構成することができる。

【0057】

従って、例えば、3 ウェイ複製が使用され、主要 (primary) KV ストレージ装置 130-2 がデータ (例えば、キーバリューデータ 170) の主要コピーを含む場合、仮想装置管理レイヤ 120 はデータの主要コピーのレプリカを後続複製 KV ストレージ装置 130-3、130-4 に格納し、データの全てのコピーは同一である。即ち、データのコピーはデータの主要コピーを含む KV ストレージ装置 130-2 に後続する 2 つの (又はそれ以上の) 直接後続する KV ストレージ装置 130-3、130-4 (例えば、循環的な方式により) に格納される。

20

【0058】

データのコピーは複製 KV ストレージ装置 130-3、130-4 において、主要 KV ストレージ装置 130-2 と同一のキー名 (key name) / 同一の利用者キー (user key) 下で格納される。データの全てのコピーは複製されたキーバリューデータ 170 を識別するための識別子及びチェックサムを含む。

30

【0059】

従って、特定の KV ストレージ装置 130 が故障している場合 (例えば、KV ストレージ装置 130-3 が故障している場合)、故障している KV ストレージ装置 130-3 の直接前後の KV ストレージ装置 130 (例えば、KV ストレージ装置 130-3 の直接前後の KV ストレージ装置 130-2、130-4) に対してキー名に関して復旧メカニズム (recovery mechanism) を使用してバリューを復旧することによって、複製されたキーが復旧されることを保障する。

【0060】

客体複製 (Object Replication) の信頼性メカニズムを要約すれば、仮想装置管理レイヤ 120 はキーバリューデータ 170 を受信し、キー客体をハッシュしてキー客体のレプリカを格納するのに使用される KV ストレージ装置 130 を決定する。仮想装置管理レイヤ 120 は、以後に、同一の利用者キー名 (例えば、適切な Metadata フィールド) の下で、アップデートされたバリューを選択された KV ストレージ装置 130 (例えば、選択された KV ストレージ装置 130-2、130-3、130-4) に書き込む。

40

【0061】

図 3 は本発明の実施形態に係る、伝統的な消去コーディング (traditional erasure coding) を使用する K-客体 (k, r) 消去コーディング (K-Object (k, r) erasure coding) 又は多重客体 “パッキング” (m

50

multiple object “Packing”）の信頼性メカニズムに応じてキーバリューストデータを格納するように構成されたKVストレージ装置のグループを示すブロック図である。

【0062】

図3を参照すれば、伝統的な消去コーディングを使用するパッキングの信頼性メカニズムは、チャンクに分割されることが（例えば、より良いデータ処理能力（throughput）を得るために）適切ではない小さいバリュースイズを有するデータに対して選択される。例えば、伝統的な消去コーディングを使用するパッキングの信頼性メカニズムは、以前に記述された客体複製（Object Replication）の信頼性メカニズムの選択に帰着したバリュースイズよりは大きいバリュースイズ（しかし、依然として比較的小さい）を有するデータに対して仮想装置管理レイヤ120によって選択される。

10

【0063】

伝統的な消去コーディングを使用するパッキングは伝統的な（k、r）MDS（maximum distance separable）消去コーディングでもって構成され、体系的なMDSコード（systemic MDS code）と共に使用される。例として、該消去コードは基本的に（by default）、（4、2）リードソロモンコード（Reed-Solomon Code）であり、その理由は、（4、2）リードソロモンコードが比較的広く研究されており、それに対応する高速具現ライブラリが容易に利用可能であるからである。

【0064】

伝統的な消去コーディングを使用するパッキングの信頼性メカニズムを使用することで、同一のパリティグループ／消去コードグループ340を形成するN個の相異なるKVストレージ装置330-1、・・・、330-Nの一部であるk個の相異なるKVストレージ装置330のキューからのk個のキー／キー客体350が選択され、消去コーディングされてパッキングされる（kは整数である）。

20

【0065】

例えば、仮想装置管理レイヤ120は各KVストレージ装置330（例えば、図1の信頼性グループ140の各KVストレージ装置130）に対して最近に書込まれたキー客体350のバッファを維持して、仮想装置管理レイヤ120がk個の相異なるKVストレージ装置からk個のキー客体350を選択して消去コーディングの実行を可能にし、これによってKVペアに対応するk個のキー客体350をパッキングする。

30

【0066】

現在の実施形態で、仮想装置管理レイヤ120は4つの異なるKVストレージ装置330-1、330-3、330-4、330-Nから各々4つのキー客体350x、350y、350b、350cを選択する（現在の実施形態で、k=4）。

【0067】

図4は本発明の実施形態に係る、伝統的な消去コーディングを使用するK-客体（k、r）消去コーディング、又は多重客体“パッキング”の信頼性メカニズムに応じるバリューストデータ及びパリティ客体の格納を示すブロック図である。

【0068】

図3及び図4を参照すれば、キー客体350は「キーモジュールnのハッシュ」番目のKVストレージ装置330に位置される。即ち、各キー客体350に対してキーモジュールnの各々のハッシュが遂行されて、その特定のKVストレージ装置330のキューに伝送される。現在の実施形態で、第iキー（Key_i）350-iはハッシュされ、KV-SSD1（330-1）に位置され、第jキー（Key_j）350-jはハッシュされ、KV-SSD2（330-2）に位置され、第kキー（Key_k）350-kはハッシュされ、KV-SSD4（330-4）に位置される。

40

【0069】

格納された各々のバリューストデータ450の使用者バリュースト長さ／バリューストサイズ462は、書込まれた使用者バリュースト長さ／バリューストサイズと同一である。しかし、消去コーディ

50

ングを可能にする一貫性のために、使用者バリュ／バリュ客体 450 はそれらに添付された“0”充填 (filling) / 仮想ゼロ / 仮想ゼロパディング 464 を有することによって全て同一のサイズを有すると看做される。

即ち、相異なるバリュ客体 450 の各々の使用者バリュサイズ 462 が変化する可能性がある (即ち、バリュ客体 450 は可変する長さを有するか、又は可変長さキーバリュである) ので、仮想ゼロパディング 464 の方法を具現することによって (即ち、コーディングのために仮想ゼロパディング 464 のゼロをバリュ客体 450 にパディングする一方、パディングされたゼロを含むバリュ客体 450 を示すデータを実際に再書込むことを避けることによって)、パリティ客体 460 はパリティグループ 340 内で最大サイズのバリュ客体 470 と同一のサイズを有する。

10

従って、現在の実施形態で、バリュ客体 450 “Val x”、“Val y”、及び “Val b” は KV ストレージ装置に実際に格納されない仮想ゼロでもってパディングされ、それによって最大サイズのバリュ客体 470 “Val c” と同一のサイズを有することと看做される。従って、パリティ客体 460 が計算される。

【0070】

k 個のキー客体 350 のコーディングの後に、仮想装置管理レイヤ 120 は k 個のキー客体 350 に対応する k 個のバリュ / k 個のバリュ客体 450 から r 個のパリティ客体 460 を計算する。この時、r は整数であり、 $k + r = N$ であり、N はパリティグループ 340 内の KV ストレージ装置 330 (例えば、図 1 の信頼性グループ 140 の N 個の KV ストレージ装置 130) の数である。

20

【0071】

仮想装置管理レイヤ 120 は r 個のパリティ客体 460 をパリティグループ 340 内の残る r 個の相異なる KV ストレージ装置 330 (即ち、k 個のキー客体 350 が選択され、消去コーディングされる k 個の KV ストレージ装置 330 を含むキューから r 個の KV ストレージ装置が区分される) に格納する。従って、k 個のキー客体 350 及び r 個のパリティ客体 460 の各々は N 個のストレージ装置 330 の中で相異なる各々に格納され、それに対応するデータはパリティグループ 340 の N 個の KV ストレージ装置 330 の各々に均等に分散される。

【0072】

伝統的な消去コーディングを使用するパッキングの信頼性メカニズムに対して、読出し及び書込みは比較的簡単であるが、パリティの再計算及び復旧はそれほど簡単ではない。パリティの復旧及び再計算に対して (例えば、アップデートの場合)、どのキー客体 350 が同一のパリティグループ 340 に一緒にグループ化されているかを知り、それによってパリティの計算を可能にするために、キー客体 350 のグループに関する情報は、各バリュ客体 450 の実際のバリュサイズ 462 (即ち、仮想ゼロパディング 464 を除いたバリュサイズ 462) と共に、KV ストレージ装置 330 (例えば、図 1 の KV ストレージ装置 130) の各々にメタデータ客体として格納され得る。従って、現在の実施形態で、追加的なメタデータがキー客体 350 (例えば、図 1 の信頼性グループ 140 に位置されたキー客体 350)、キー客体 350 に対応するバリュ客体 450 の各々の元の長さ、及びキー客体 350 のコーディング順序に KV ストレージ装置 130 を格納するのに使用され得る。

30

40

【0073】

例えば、メタデータ客体バリュは信頼性グループ 140 のキー客体 350 の全体を表示し (indicate)、またバリュ客体 450 のバリュサイズ 462 を表示するフィールドを含み、パリティ客体キー (即ち、仮想ゼロパディング 464 のゼロを含むバリュ客体 450)、パリティ客体 460 のバリュサイズ 462、及び r 個のパリティ客体 460 が格納された対応する r 個の KV ストレージ装置 330 の識別のための装置 ID を示す相異なるフィールドを含み得る。

【0074】

データは使用者キーを使用して格納される。メタデータは使用者キー及び “Meta da

50

ta”と称するMetaIDインジケータ（表示部、indicator）を使用して形成された内部キーに格納される。さらに、バリュー客体450がどこで終了されたか、仮想ゼロパディング464のゼロがどこで開始するかを決定することによって、バリュー客体450が再生成される場合に正確な再構成のためにバリューサイズ462はメタデータに格納されて仮想ゼロパディング464の位置（即ち、ゼロが何処に追加されたか）の認識を可能にする。

【0075】

KVストレージ装置330の中で1つが故障であれば、データ及びメタデータの全てが同一のKVストレージ装置330に格納される可能性があるため、データ及びメタデータが消失する可能性が潜在的にあり、それによって、復旧が不可能になる可能性である。しかし、このような状況を防止するために、メタデータ客体バリューは、以前に言及された客体複製（Object Replication）の信頼性メカニズムをメタデータ客体バリューに対して具現できる仮想装置管理レイヤ120の“客体複製エンジン（Object Replication Engine）”を使用して複製され得る。

【0076】

追加的に、メタデータ客体バリューは信頼性グループ140の全ての客体に対して同一であるため、KVストレージ装置330が客体連結（object linking）を支援する場合、同一のメタデータ客体バリューは同一のKVストレージ装置330に共通に位置する多重のキー名（key names）に連結されることができる。さらに、集団書込み（batch writing）が支援される場合、客体バリューはさらに良い処理能力のために共にバッチ処理（batched）され得る。

【0077】

現在の実施形態に係る伝統的な消去コーディングを使用するパッキングの信頼性メカニズムを要約すれば、仮想装置管理レイヤ120はバッファを介してk個の異なるKVストレージ装置330からk個の最近に格納されたキー客体350を選択できる。

仮想装置管理レイヤ120は、以後に、各々のキー客体350に対応するバリュー客体450（パリティグループ440の最大サイズのバリュー客体470を除く）を回収し、仮想ゼロパディング350でもってパディングしてバリュー客体450を同一のサイズ（例えば、最大サイズのバリュー客体470のサイズ）に生成できる。

仮想装置管理レイヤ120は、以後に、MDSコードプロセスを使用してk個のキー客体350からr個のパリティ客体を生成できる。

仮想装置管理レイヤ120は、以後に、r個のパリティ客体460をN個のKVストレージ装置330の中でキー客体350が選択されたk個のKVストレージ装置330と相異なるr個のKVストレージ装置330に書込みできる。この時、k+rはNと同一である。

仮想装置管理レイヤ120は、以後に、上述された情報を示すメタデータ客体を生成できる。

最後に、仮想装置管理レイヤ120は使用者キー及びメタデータ識別子から形成されたキーと共にキー客体350及びパリティ客体460をN個のKVストレージ装置330（例えば、複製エンジンと類似である）に書込みできる。

【0078】

図5は本発明の実施形態に係る、伝統的な消去コーディングを使用する単一客体（k、r）消去コーディング（Single Object（k、r）erasure coding）、又は“スプリットティング（Splitting）”の信頼性メカニズムに応じてキーバリューデータを格納するように構成されたKVストレージ装置のグループを示すブロック図である。

【0079】

図5を参照すれば、以前に言及された客体複製（Object Replication）及び伝統的な消去コーディングを使用するパッキングの信頼性メカニズムに適合するバリューサイズより大きいバリューサイズを有するバリューに対して、仮想装置管理レイ

10

20

30

40

50

ヤ 1 2 0 は、伝統的な消去コーディングを使用する“単一客体 (Single Object) (k, r) 消去コーディング”、即ち“スプリットティング (splitting)”の信頼性メカニズムを選択できる。

伝統的な消去コーディングを使用するスプリットティングの信頼性メカニズムは、比較的大きいバリュースイズを有し、k 個の同一サイズのパリユー客体 5 5 0 に分割された場合、良い被処理能力を有する K V パリユー (キーパリユーデータ / K V ペア) 5 7 0 に適合する K V パリユー (キーパリユーデータ / K V ペア) 単位の信頼性メカニズムである。

【0080】

K V パリユー 5 7 0 を分割した後に、実施形態によれば、仮想装置管理レイヤ 1 2 0 は k 個のパリユー客体 5 5 0 の各々に対してチェックサムを計算する。その後に、仮想装置管理レイヤ 1 2 0 は k 個のパリユー客体 5 5 0 の各々の前にメタデータを挿入する。

10

【0081】

伝統的な消去コーディングを使用するスプリットティングは K V パリユー 5 7 0 を複数のさらに小さいバリュースイズに分割すること、その後に K V パリユー 5 7 0 を分割して得られた複数のさらに小さいバリュースイズを k 個の連続した K V ストレージ装置 5 3 0 に亘って分散することを含む。従って、k 個の同一サイズのパリユー客体 5 5 0 のサイズは基本 (underlying) K V ストレージ装置 5 3 0 の各々によって支援され得る。

【0082】

伝統的な消去コーディングを使用するスプリットティングを使用する場合、仮想装置管理レイヤ 1 2 0 は体系的な M D S コード (例えば、仮想装置レイヤ 1 2 0 が、基本 (default) コードとして、(4, 2) リードソロモンコードなどの伝統的な (k, r) M D S 消去コーディングを前提として構成される場合) を使用して生成された r 個のパリティパリユー / 客体 5 6 0 をまた追加できる。以後に、上述された伝統的な消去コーディングを使用するパッキングの信頼性メカニズムと類似な方式により、仮想装置管理レイヤ 1 2 0 は k 個のパリユー客体 5 5 0 及び r 個のパリティ客体 5 6 0 を N 個の K V ストレージ装置 5 3 0 に書き込みできる ($k + r = N$)。

20

【0083】

従って、仮想装置管理レイヤ 1 2 0 は比較的大きい K V パリユー 5 7 0 を k 個のパリユー客体 5 5 0 に分割でき、r 個のパリティ客体 5 6 0 を計算し、追加でき、k 個のパリユー客体 5 5 0 及び r 個のパリティ客体 5 6 0 を ($k + r$) 個の K V ストレージ装置 5 3 0 に格納できる。

30

【0084】

伝統的な消去コーディングを使用するスプリットティングの信頼性メカニズムを使用する場合、K V パリユー 5 7 0 のパリユーに対応するキー 5 8 0 をハッシングした後に、仮想装置管理レイヤ 1 2 0 は対応するパリユー客体を格納する (例えば、k 個のパリユー客体 5 5 0 の中で第 1 のパリユー客体、図 5 の実施形態における D 1 をハッシュマークゼロに格納する) ために、主要 K V ストレージ装置 5 3 0 a (例えば、図 5 に図示された実施形態における、K V - S S D 2) を決定できる。k + r 個のパリユー客体 5 5 0、5 6 0 は同一の使用者キー名の下で、主要 K V ストレージ装置 5 3 0 a 及び (N - 1) 個の連続的な K V ストレージ装置 5 3 0 の各々に書込まれ得る。

40

即ち、図 5 に図示された実施形態で、k 個のパリユー客体 5 5 0 の中で第 1 のパリユー客体は主要 K V ストレージ装置 5 3 0 a “K V - S S D 2” に書込まれ、k 個のパリユー客体 5 5 0 の中で残りは r 個のパリティ客体 5 6 0 と共に K V ストレージ装置 5 3 0 “K V - S S D 3” 乃至 “K V - S S D N” 及び “K V - S S D 1” に循環方式の順に書込まれる (例えば、上述された客体複製の信頼性メカニズムに対して記述されたことと類似な方式である)。

【0085】

伝統的な消去コーディングを使用するスプリットティングの信頼性メカニズムを要約すれば、仮想装置管理レイヤ 1 2 0 は比較的大きい K V パリユー 5 7 0 を k 個の同一サイズの

50

バリュウ客體 550 に分割できる。仮想装置管理レイヤ 120 は、以後に、MDS コード プロセスを使用して k 個のバリュウ客體 550 に対する r 個のパリティ客體 560 を生成 できる。仮想装置管理レイヤ 120 は、以後に、KV バリュウ 570 のバリュウに対応す るキーをハッシングして第 1 のバリュウ客體が位置するべき主要 KV ストレージ装置 53 0 a を決定できる。仮想装置管理レイヤ 120 は、以後に、仮想装置管理レイヤ 120 に よって生成され、循環方式の順に主要 KV ストレージ装置 530 a 及び $(N-1)$ 個の連 続した KV ストレージ装置 530 に対応する適切な MetaID フィールドを含む同一の 使用者キー名の下で、 $(k+r)$ 個のバリュウ客體 550 とパリティ客體 560 を書き込み できる。

【0086】

図 3 及び図 4 を再び参照すれば、他の実施形態によれば、仮想装置管理レイヤ 120 は 再生成消去コーディングを使用する、 K -客體 (k, r, d) 消去コーディング、即ち多 重客體 (multiple__object “パッキング” の信頼性メカニズムを選択できる (例えば、図 2 の順序図 200 に従って)。現在の信頼性メカニズムは仮想装置管理レイ ヤ 120 が k 個の客體を k 個の KV ストレージ装置にパッキングする点で、以前に記述さ れた伝統的な消去コーディングを使用するパッキングの信頼性メカニズムと類似である。 しかし、再生成消去コーディングを使用するパッキングは伝統的な (k, r) 消去コード を使用する代わりに、(k, r, d) 再生成コードを使用する。従って、図 3 及び図 4 は 現在の実施形態に対しては、一般的実施形態として参照され得る。

【0087】

従って、再生成コードが適切であるけれども、客體を分割することが適切ではなく、若 しくは、客體を分割などせずにそのまま維持する (intact) ことがさらに適切であ る場合に、再生成消去コーディングを使用するパッキングが使用され得る。再生成消去コー ディングを使用するパッキングは以前に言及された客體複製、並びに伝統的な消去コー ディングを使用するスプリッティング及びパッキングの信頼性メカニズムのために使用さ れるバリュウサイズより大きいバリュウサイズに対して適切であり得る。

再生成消去コーディングを使用するパッキングは、客體の複数のサブパケットの読出し が全体客體の読出しよりも低い性能に帰着しない場合に使用され得る。再生成消去コー ディングを使用するパッキングは、基本 (underlying) KV ストレージ装置 (例 えば、図 1 の KV ストレージ装置 130、又は図 3 の KV ストレージ装置 330) が復旧 /再構成の間に補助できて、再生成コードを認識可能な (aware) KV ストレージ装 置である場合に、適切であり得る。

【0088】

図 6 は本発明の実施形態に係る、再生成消去コーディングを使用する“単一客體 (Single__Object) (k, r, d) 消去コーディング”、即ち、“スプリッティング (Splitting)” の信頼性メカニズムに応じてキーバリュウデータを格納するように 構成された KV ストレージ装置のグループを示すブロック図である。

【0089】

図 6 を参照すれば、現在の信頼性メカニズムは、伝統的な (k, r) MDS 消去コー ディングの代わりに (k, r, d) 再生成コードを使用する点のみを除外すれば、仮想装置 管理レイヤが、図 5 に図示されたように、伝統的な消去コーディングを使用するスプリッ ティングと類似な方式による動作を可能にし得る。再生成消去コーディングを使用する パッキングと同様に、現在の信頼性メカニズムは基本 (underlying) KV スト レージ装置 630 が復旧 /再構成の間に補助できて、再生成コードを認識可能な KV スト レージ装置である場合に適切であり得る。

【0090】

KV バリュウ客體 670 が以前に記載された信頼性メカニズムと対応する KV バリュウ 客體よりさらに大きいバリュウサイズを有し、KV バリュウ客體 670 の k 個のスプリッ ト (バリュウ客體) 650 の複数のサブパケット 690 の読出しが、スプリット (バリュウ 客體) 650 全体の読出し (例えば、伝統的な消去コーディングを使用するスプリッテ

10

20

30

40

50

イングの信頼性メカニズムで遂行される)よりも低い性能に帰着しない場合に、再生成消去コーディングを使用するスプリッティングが適切であり得る。

【0091】

再生成消去コーディングを使用するスプリッティングの信頼性メカニズムは、KVバリュー客体670がk個の同一サイズバリュー客体(スプリット)650に分割(split)され、各スプリット650が複数のサブパケット690(例えば、現在の実施形態では、スプリット650が1個当たり4つのサブパケット690に分割される)に仮想的にさらに分割(split)され、KVバリュー客体670からの複数のサブパケット690の読出しがKVバリュー客体670の全体の読出しよりも良い処理能力を有する場合に適切な処理量を有し得る程の非常に大きいバリューサイズを有するKVバリュー客体に適合し得るKVバリュー客体(KVペア)単位のメカニズムである。バリューサイズは全ての基本(underlying)KVストレージ装置630によって支援される。

10

【0092】

伝統的な消去コーディングを使用するスプリッティングと類似に、図5に図示されたように、現在信頼性メカニズムの仮想装置管理レイヤ120は体系的な再生成コードを使用してr個のパリティ客体660を追加して、k個のスプリット(バリュー客体)650及びr個のパリティ客体660をN個のKVストレージ装置630に書き込み得る(ここで、 $(k+r)=N$)。しかし、r個のパリティ客体660の各々は複数個のパリティサブパケット692(該複数個の個数は、例えば、スプリット(バリュー客体)650当たりのサブパケット690の個数に対応する)に分割され得る。伝統的な消去コーディングを使用するスプリッティングと異なり、現在の実施形態では、基本(default)コードが(4、2、5)ジグザグコード(zigzag code)であり得る。

20

【0093】

再生成消去コーディングを使用するスプリッティングの信頼性メカニズムを要約すれば、仮想装置レイヤ120は大きいKVバリュー客体670をk個の同一サイズのスプリット(バリュー客体)650に分割できる。仮想装置管理レイヤ120は、以後に、k個の客体650の各々をm個の同一サイズサブパケット690に分割できる(k、mは整数である)。仮想装置管理レイヤ120は、以後に、再生成コーディングプロセスを使用してk個の客体650に対するr個のパリティ客体660を生成して、r個のパリティ客体660の各々は、m個の同一サイズのパリティサブパケット692に分割され得る。

30

仮想装置管理レイヤ120は、以後に、KVバリュー客体670に対応するキーをハッシングして第1スプリット(バリュー客体)D1が位置する主要KVストレージ装置630a(この場合、KV-SSD_2)を決定できる。仮想装置管理レイヤ120は、以後に、仮想装置管理レイヤ120によって生成され、循環方式により主要KVストレージ装置630a(この場合、KV-SSD_2)及び(N-1)個の連続したKVストレージ装置630(この場合、KV-SSD_3~KV-SSD_k+r、KV-SSD_1)に対応する適切なMetaIDフィールドを含み得る同一の使用者キー名の下で、各々がm個のサブパケット690を含むk個のスプリット(バリュー客体)650(D1~Dk)、及び各々がm個のパリティサブパケット692を含むr個のパリティ客体660(Dk+1~Dk+r)を書き込み得る。

40

【0094】

上述されたことによれば、仮想装置管理レイヤは、キーバリューデータの1つ又はそれ以上の属性に基づいてキーバリューデータの格納に対して信頼性メカニズムのグループから適切な信頼性メカニズムを選択できる。従って、本文に記載された実施形態は、記載された信頼性メカニズムが各々何れも単一キー復旧手続を遂行できるので、メモリストレージの分野における性能改善に貢献する。

メモリ装置の全体が故障した場合、本発明の実施形態の仮想装置管理レイヤは故障したメモリ装置に存在するキーの全てを復旧し、新しいメモリ装置にコピーできる。仮想装置管理レイヤは、信頼性グループ内の故障したメモリ装置に隣接するメモリ装置に存在するキーの全体に対する反復動作を遂行し、信頼性メカニズムが故障であるメモリ装置に存在

50

することと判断したキーに対してキー単位複数動作を遂行することによって、キー全体の復旧及びコピーを達成できる。

【0095】

基本信頼性メカニズム（例えば、基本（underlying）ストレージ装置のサイズ制約に従う）によって支援されるバリュースイズより大きいバリュースイズを有する非常に大きいKVペア（＝KVバリュー）が信頼性管理者によって複数のKVペアに明示的に分割され、信頼性メカニズムが複数のスプリット及びスプリット数の情報をKVバリューに格納されたメタデータと共に格納するので、記載された実施形態はメモリストレージ分野における改善にさらに貢献する。

【0096】

本文に記載された実施形態は、特定の用語を採用した場合でも、その用語は一般的、且つ記述的な意味にのみ使用され、且つ解釈されるべきであり、これに制限する意図はない。一部の実施形態で、本発明が属する技術分野における通常の技術者には明らかなように、特定の実施形態において明瞭に相異なると言及されない限り、特定実施形態と連関されて記述された特徴、属性、及び／又は要素は他の実施形態と連関されて記述された特徴、属性、及び／又は要素と組合わされて又は独立的に使用され得る。従って、当業者は形態又は詳細な記述の多様な変形が、今回の開示に含まれた機能的な均等物と合わせて以下の特許請求の範囲に表明された今回の開示の思想及び範囲から逸脱せずに、具現され得ることが理解されよう。

【符号の説明】

【0097】

- 110 単一仮想装置
- 120 仮想装置管理レイヤ
- 130 KVストレージ装置
- 140 信頼性グループ
- 170 キーバリューデータ、キーバリューデータ／KVペア
- 330、330-1、・・・、330-N KVストレージ装置
- 340 パリティグループ／消去コードグループ
- 350、350-i、350-j、350-y、350-b、350-k、350-c
- キー／キー客体
- 450 バリュー客体
- 460 パリティ客体
- 462 使用者バリュー長さ／バリュースイズ
- 464 “0” 充填／仮想ゼロ／仮想ゼロパディング
- 470 最大サイズのバリュー客体
- 530 ストレージ装置
- 530a 主要KVストレージ装置
- 550 バリュー客体
- 560 パリティ客体
- 570 KVバリュー／キーバリューデータ／KVペア
- 580 キー
- 630 KVストレージ装置
- 630a 主要KVストレージ装置
- 650 スプリット（バリュー客体）
- 660 パリティ客体
- 670 KVバリュー客体
- 690 サブパケット
- 692 パリティサブパケット

10

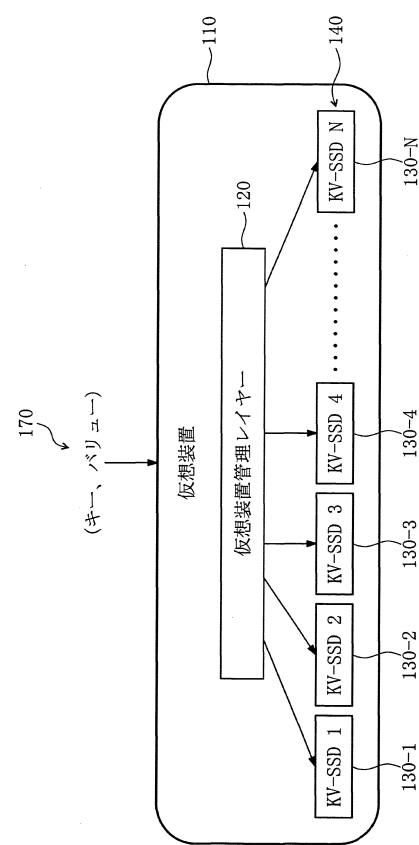
20

30

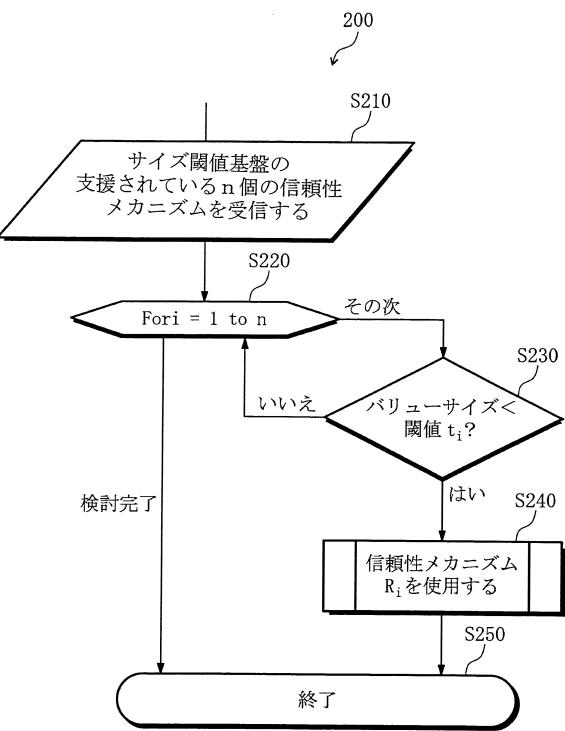
40

50

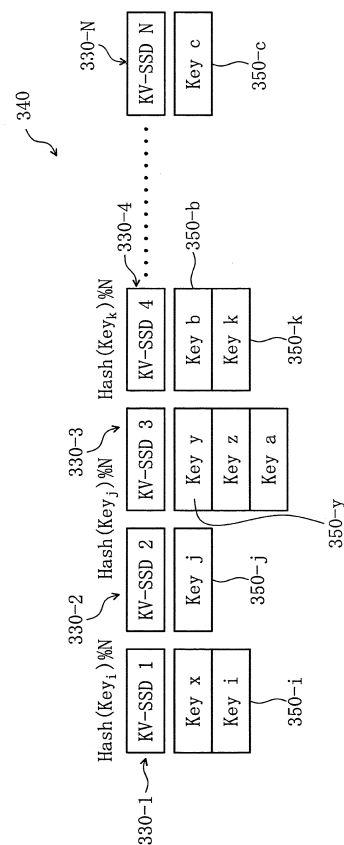
【図面】
【図 1】



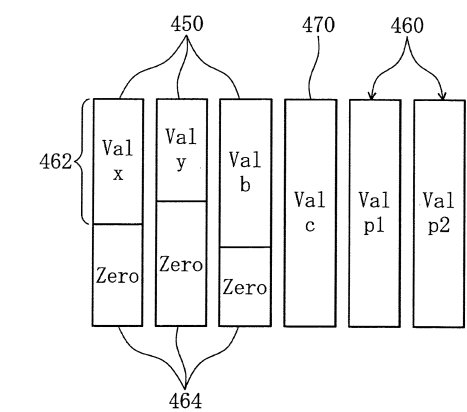
【図 2】



【図 3】



【図 4】



10

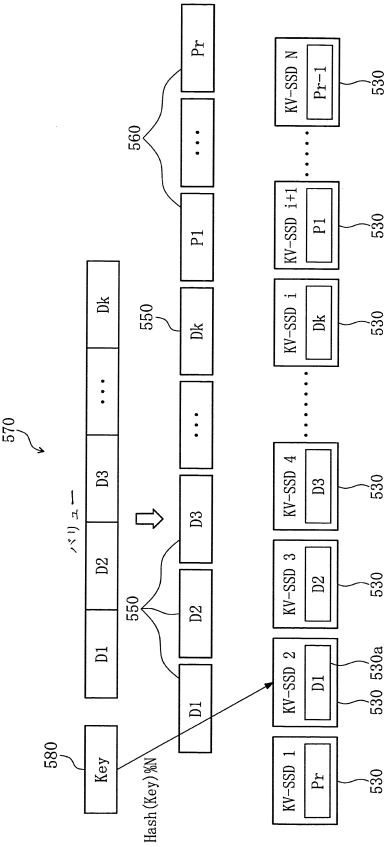
20

30

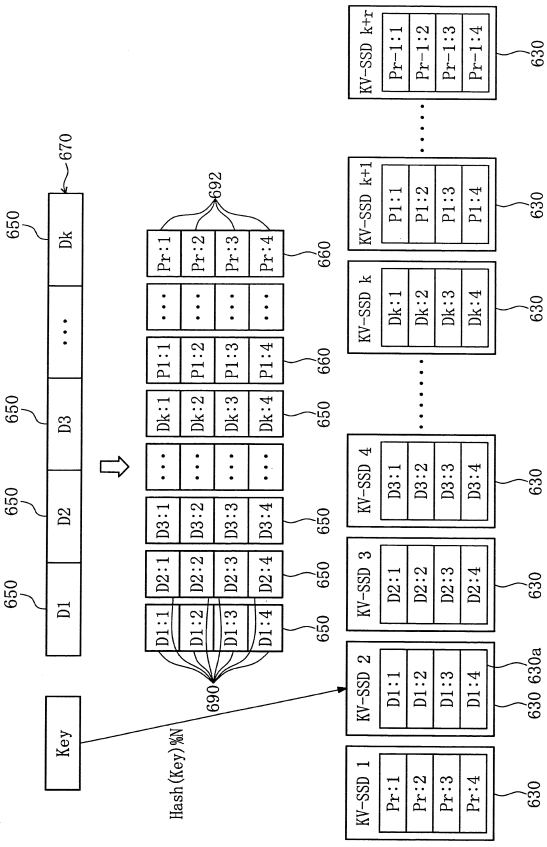
40

50

【図 5】



【図 6】



10

20

30

40

50

フロントページの続き

(32)優先日 平成30年4月30日(2018.4.30)

(33)優先権主張国・地域又は機関
米国(US)

(31)優先権主張番号 62/713479

(32)優先日 平成30年8月1日(2018.8.1)

(33)優先権主張国・地域又は機関
米国(US)

(31)優先権主張番号 16/165655

(32)優先日 平成30年10月19日(2018.10.19)

(33)優先権主張国・地域又は機関
米国(US)

早期審査対象出願

ブレイス, 3871

(72)発明者 奇 亮 ソク

アメリカ合衆国 カリフォルニア州 94303, パロアルト, アルテア ウォーク, 873

審査官 鹿野 博嗣

(56)参考文献 特表2015-519674(JP, A)

特開2013-050836(JP, A)

米国特許出願公開第2016/0006461(US, A1)

(58)調査した分野 (Int.Cl., DB名)

G06F 16/182

G06F 3/06

G06F 11/10

G06F 16/11