(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5937599号
(P5937599)
(45)発行日　平成28年6月22日(2016.6.22)　(24)登録日　平成28年5月20日(2016.5.20)

(51)Int.Cl.　F I
G06F 3/06 (2006.01)　G06F 3/06 305C
G06F 3/06 540

請求項の数 7　(全 32 頁)

(21)出願番号　特願2013-531648(P2013-531648)
(86)(22)出願日　平成23年9月20日(2011.9.20)
(65)公表番号　特表2013-539134(P2013-539134A)
(43)公表日　平成25年10月17日(2013.10.17)
(86)国際出願番号　PCT/US2011/052276
(87)国際公開番号　WO2012/044492
(87)国際公開日　平成24年4月5日(2012.4.5)
審査請求日　平成26年9月12日(2014.9.12)
(31)優先権主張番号　12/896,669
(32)優先日　平成22年10月1日(2010.10.1)
(33)優先権主張国　米国(US)

(73)特許権者　513076589
ピュア・ストレージ・インコーポレイテッド
アメリカ合衆国・９４０４１・カリフォルニア州・マウンテン　ビュー・カストロ　ストリート・６５０・スイート・２２０
(74)代理人　100064621
弁理士　山川　政樹
(74)代理人　100098394
弁理士　山川　茂樹
(72)発明者　コルグローヴ，ジョン
アメリカ合衆国・９４０２４・カリフォルニア州・ロス　アルトス・ヴィスタ　グランデ　アヴェニュ・７２２

最終頁に続く

(54)【発明の名称】動的構成のＲＡＩＤアレイにおける再構成読み込み

(57)【特許請求の範囲】
【請求項１】
独立ディスク冗長アレイ（ＲＡＩＤ）における複数のストレージデバイスを備えたデータストレージサブシステムと、ストレージコントローラを備え、
前記ストレージコントローラは、
第１のＲＡＩＤストライプを、前記複数のストレージデバイスに書き込むように構成され、前記複数のストレージデバイスの特定のストレージデバイスに、前記特定のストレージデバイスに保存された前記第１のＲＡＩＤストライプの第１のＲＡＩＤデータを保護するためのデバイス内保護データの第１の量を配置することを含み、
第２のＲＡＩＤストライプを書き込むための前記複数のストレージデバイスのストレージデバイスのサブセットを選択するように構成され、前記サブセットのストレージデバイスの量は、前記複数のストレージデバイスのストレージデバイスの量よりも小さく、前記サブセットは、前記特定のストレージデバイスを含み、
第２のＲＡＩＤストライプを、前記サブセットに書き込むように構成され、前記特定のストレージデバイスに、前記特定のストレージデバイスに保存された前記第２のＲＡＩＤストライプの第２のＲＡＩＤデータを保護するためのデバイス内保護データの第２の量を配置することを含み、前記デバイス内保護データの第２の量は、前記デバイス内保護データの第１の量よりも大きい、
コンピュータシステム。
【請求項２】

前記複数のストレージデバイスは、ソリッドステートストレージデバイスである、
ことを特徴とする請求項１に記載のコンピュータシステム。
【請求項３】
前記第１のＲＡＩＤストライプは、Ｌ＋ｘのレイアウトである第１のＲＡＩＤレイアウトを有し、前記第２のＲＡＩＤストライプは、Ｍ＋ｙのレイアウトである第２のＲＡＩＤレイアウトを有し、Ｌ、ｘ、Ｍ及びｙは整数であり、（１）ＬはＭに等しくない、及び（２）ｘはｙに等しくない、のいずれか又は両方である、
ことを特徴とする請求項１に記載のコンピュータシステム。
【請求項４】
前記複数のストレージデバイスにおける少なくとも１つのストレージデバイスは、前記特定のストレージデバイスにおける前記デバイス内保護データの第１の量とは異なるデバイス内保護データの量を有する、
ことを特徴とする請求項１に記載のコンピュータシステム。
【請求項５】
前記ストレージデバイスのサブセットにおける少なくとも１つのストレージデバイスは、前記特定のストレージデバイスにおける前記デバイス内保護データの第２の量とは異なるデバイス内保護データの量を有する、
ことを特徴とする請求項１に記載のコンピュータシステム。
【請求項６】
独立ディスク冗長アレイ（ＲＡＩＤ）における複数のストレージデバイスを構成するための方法であって、
第１のＲＡＩＤストライプを、前記複数のストレージデバイスに書き込むステップであって、前記複数のストレージデバイスの特定のストレージデバイスに、前記特定のストレージデバイスに保存された前記第１のＲＡＩＤストライプの第１のＲＡＩＤデータを保護するためのデバイス内保護データの第１の量を配置することを含むステップと、
第２のＲＡＩＤストライプを書き込むための前記複数のストレージデバイスのストレージデバイスのサブセットを選択するステップであって、前記サブセットのストレージデバイスの量は、前記複数のストレージデバイスのストレージデバイスの量よりも小さく、前記サブセットは、前記特定のストレージデバイスを含むステップと、
第２のＲＡＩＤストライプを、前記サブセットに書き込むステップであって、前記特定のストレージデバイスに、前記特定のストレージデバイスに保存された前記第２のＲＡＩＤストライプの第２のＲＡＩＤデータを保護するためのデバイス内保護データの第２の量を配置することを含み、前記デバイス内保護データの第２の量は、前記デバイス内保護データの第１の量よりも大きい、ステップと、
を含むことを特徴とする方法。
【請求項７】
独立ディスク冗長アレイ（ＲＡＩＤ）における複数のストレージデバイスを構成ためのプログラム命令を記憶するコンピュータ可読記憶媒体であって、前記プログラム命令は、
第１のＲＡＩＤストライプを、前記複数のストレージデバイスに書き込むように実行可能であり、前記複数のストレージデバイスの特定のストレージデバイスに、前記特定のストレージデバイスに保存された前記第１のＲＡＩＤストライプの第１のＲＡＩＤデータを保護するためのデバイス内保護データの第１の量を配置することを含み、
第２のＲＡＩＤストライプを書き込むための前記複数のストレージデバイスのストレージデバイスのサブセットを選択するように実行可能であり、前記サブセットのストレージデバイスの量は、前記複数のストレージデバイスのストレージデバイスの量よりも小さく、前記サブセットは、前記特定のストレージデバイスを含み、
第２のＲＡＩＤストライプを、前記サブセットに書き込むように実行可能であり、前記特定のストレージデバイスに、前記特定のストレージデバイスに保存された前記第２のＲＡＩＤストライプの第２のＲＡＩＤデータを保護するためのデバイス内保護データの第２の量を配置することを含み、前記デバイス内保護データの第２の量は、前記デバイス内保

護データの第１の量よりも大きい、
コンピュータ可読記憶媒体。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、コンピュータネットワークに関し、より詳細には、複数のソリッドステートストレージデバイス間でデータを効率的に分散させる発明に関する。
【背景技術】
【０００２】
コンピュータのメモリストレージ及びデータ帯域幅が増すと、企業が日々管理するデータの量及び複雑性も増す。通常、データセンターなどの大規模分散型ストレージシステムは、多くの業務を実行する。分散型ストレージシステムは、１又はそれ以上のネットワークにより相互接続されたクライアントコンピュータに結合することができる。分散型ストレージシステムのいずれかの部分が不良を起こし、又は利用できなくなった場合には、企業活動が損なわれ、又は完全に停止する恐れがある。従って、分散型ストレージシステムは、データの利用可能性及び高パフォーマンス機能のための高い標準を維持すると予想される。本明細書で使用するストレージディスクは、ストレージ技術のタイプによってはディスクを含まないものもあるので、ストレージデバイスと呼ぶことができる。
【０００３】
多くの場合、ストレージデバイスは、データ損失からの保護のために、エラー検出機構及びエラー訂正機構を含む。多くの場合、これらの機構は、デバイスにより生成されてデバイス自体に記憶されるエラー訂正符号の形をとる。また、分散型ストレージシステムは、分散アルゴリズムを利用して、一群のストレージデバイス間でデータを分散させることもできる。一般に、これらのアルゴリズムは、中央ディレクトリに依拠せずにデータオブジェクトをストレージデバイスにマッピングする。このようなアルゴリズムの例に、レプリケーション・アンダー・スケーラブル・ハッシング（ＲＵＳＨ）及びコントロールド・レプリケーション・アンダー・スケーラブル・ハッシング（ＣＲＵＳＨ）がある。分散型ストレージシステム内の複数のクライアントは、中央ディレクトリを伴わずに複数のサーバ上のデータオブジェクトに同時にアクセスすることができる。また、記憶されているメタデータの量を低減することもできる。しかしながら、容量、入力／出力（Ｉ／Ｏ）特性及び信頼性問題が異なる複数のストレージディスク間でデータを分散させるという困難なタスクが依然として残る。ストレージデバイス自体と同様に、これらのアルゴリズムも、（ＲＡＩＤ５及びＲＡＩＤ６などの）ＲＡＩＤタイプアルゴリズム又はリードソロモン符号などのエラー検出及び訂正アルゴリズムを含むことができる。
【０００４】
動的に追加及び除去できる複数のストレージデバイス間でデータを分散させるために使用する方法は、選択したストレージデバイスに関連する技術及び機構によって決まる。例えば、上述したアルゴリズムは、ハードディスクドライブ（ＨＤＤ）を利用するシステムに合わせて開発されたものである。ＨＤＤは、各々が磁気媒体で被覆された１又はそれ以上の回転ディスクを含む。これらのディスクは、毎日数時間にわたり毎分数千回転の速さで回転する。また、この回転ディスク上への磁気読み込み／書き込み装置の位置付けには、磁気アクチュエータが関与する。これらのアクチュエータは、摩擦、摩耗、振動及び機械的不均衡の影響を受けやすく、結果的に信頼性問題が生じる。上述したデータ分散アルゴリズムは、ＨＤＤのこれらの特性及び挙動に基づくものである。
【０００５】
別のタイプの記憶ディスクの例に、ソリッドステートディスク（ＳＳＤ）がある。ソリッドステートディスクは、ソリッドステートドライブと呼ぶこともできる。ＳＳＤは、ＨＤＤインターフェイスをエミュレートできるが、ＨＤＤで見られるような電気機械デバイスではなく固体メモリを利用して永続データを記憶する。例えば、ＳＳＤは、一群のフラッシュメモリを含むことができる。可動部品又は機械的遅延がなければ、ＳＳＤのアクセ

ス時間及びレイテンシは、ＨＤＤよりも短くなり得る。しかしながら、通常、ＳＳＤの書き込みレイテンシはかなり長い。入力／出力（Ｉ／Ｏ）特性が異なることに加え、ＳＳＤの故障モードもＨＤＤとは異なる。従って、記憶のためにＳＳＤを備えたシステムでは、ＨＤＤに合わせて開発された分散型データ配置アルゴリズムを利用しながら高パフォーマンス及び高信頼性を実現できない場合がある。

【発明の概要】

【発明が解決しようとする課題】

【０００６】

上記に鑑み、複数のソリッドステートストレージデバイス間でデータを効率的に分散し、エラーを検出して訂正するためのシステム及び方法が望まれている。

【課題を解決するための手段】

【０００７】

複数のソリッドステートストレージデバイス間でデータを効率的に分散して管理するためのコンピュータシステム及び方法の様々な実施形態を開示する。

【０００８】

１つの実施形態では、コンピュータシステムが、ネットワークを介して読み込み及び書き込み要求を受け取るように結合された１又はそれ以上のデータストレージアレイにネットワークを介して読み込み及び書き込み要求を伝達するように構成された複数のクライアントコンピュータを備える。複数のストレージデバイス上に複数の記憶位置を有する（単複の）データストレージアレイを企図する。様々な実施形態では、このストレージデバイスが、データを記憶して保護するための独立ドライブ冗長アレイ（ＲＡＩＤ）構成で構成される。データストレージデバイスは、フラッシュメモリセルなどの、データ記憶のための固体メモリ技術を含むことができる。データストレージサブシステムは、ストレージコントローラをさらに備え、このストレージコントローラは、ストレージデバイスの第１のサブセットを、第１の冗長データセットを含む第１のＲＡＩＤレイアウトで使用するように構成するように構成される。このコントローラは、ストレージデバイスの第２のサブセットを、第２の冗長データセットを含む第２のＲＡＩＤレイアウトで使用するようにさらに構成する。また、このコントローラは、第１のサブセット又は第２のサブセットのいずれにも含まれていない追加のデバイスを、第１のＲＡＩＤレイアウト及び第２のＲＡＩＤレイアウトの両方のための冗長データを記憶するように構成する。

【０００９】

第１のＲＡＩＤレイアウト又は第２のＲＡＩＤレイアウトの特定のストレージデバイスを対象とする所与の読み込み要求を受け取ったことに応答して、コントローラが、特定のストレージデバイスが非エラーに関する相対的に遅い読み込み応答を示していると判断したことに応答して、所与の読み込み要求に対応する再構成読み込みを開始するように構成される実施形態も企図する。また、第１のサブセット及び第２のサブセット内の各デバイスは、デバイス内冗長データを記憶するように構成することができ、第１の冗長データセット及び第２の冗長データセットは、いずれもデバイス間冗長データを含む。さらに、様々な実施形態では、第１のＲＡＩＤレイアウトが、Ｌ＋ｘのレイアウトであり、第２のＲＡＩＤレイアウトが、Ｍ＋ｙのレイアウトであり、Ｌ、ｘ、Ｍ及びｙは整数であり、（１）ＬはＭに等しくない、及び（２）ｘはｙに等しくない、のいずれか又は両方である。

【００１０】

以下の説明及び添付図面を検討すると、これらの及びその他の実施形態が明らかになるであろう。

【図面の簡単な説明】

【００１１】

【図１】ネットワークアーキテクチャの１つの実施形態を示す汎用ブロック図である。

【図２】動的デバイス内冗長スキームの１つの実施形態の汎用ブロック図である。

【図３】データストレージサブシステムにおいてデバイス内保護を調整する方法の１つの実施形態を示す汎用フロー図である。

【図４】ストレージサブシステムの１つの実施形態の汎用ブロック図である。
【図５】デバイスユニットの１つの実施形態の汎用ブロック図である。
【図６】状態テーブルの１つの実施形態を示す汎用ブロック図である。
【図７】フレキシブルなＲＡＩＤデータレイアウトアーキテクチャの１つの実施形態を示す汎用ブロック図である。
【図８】フレキシブルなＲＡＩＤデータレイアウトアーキテクチャの別の実施形態を示す汎用ブロック図である。
【図９】データストレージサブシステム内のレイアウトを動的に決定する方法の１つの実施形態を示す汎用フロー図である。
【図１０】フレキシブルなＲＡＩＤデータレイアウトアーキテクチャのさらに別の実施形態を示す汎用ブロック図である。
【図１１Ａ】デバイスレイアウトの１つの実施形態を示す図である。
【図１１Ｂ】セグメントの１つの実施形態を示す図である。
【図１１Ｃ】異なるページタイプ内のデータストレージ構成の１つの実施形態を示す汎用ブロック図である。
【図１２】ハイブリッドＲＡＩＤデータレイアウトの１つの実施形態を示す汎用ブロック図である。
【図１３】データストレージサブシステム内で代替のＲＡＩＤ構成を選択する方法の１つの実施形態を示す汎用フロー図である。
【発明を実施するための形態】
【００１２】
本発明は様々な修正及び代替形態が可能であるが、図面には特定の実施形態を一例として示し、本明細書ではこれらについて詳細に説明する。しかしながら、図面及びこれらに対する詳細な説明は、開示する特定の形態に本発明を限定することを意図するものではなく、むしろ添付の特許請求の範囲によって定められる本発明の思想及び範囲内にある全ての修正物、同等物及び代替物を含むことを意図するものであると理解されたい。
【００１３】
以下の説明では、本発明を完全に理解できるように数多くの具体的な詳細を示す。しかしながら、当業者であれば、これらの具体的な詳細を伴わずに本発明を実施できると認識すべきである。いくつかの例では、本発明を曖昧にしないように、周知の回路、構造、信号、コンピュータプログラム命令及び技術については詳細に示していない。
【００１４】
図１を参照すると、ネットワークアーキテクチャ１００の１つの実施形態の汎用ブロック図を示している。後述するように、ネットワークアーキテクチャ１００の１つの実施形態は、ネットワーク１８０を介して互いに、及びデータストレージアレイ１２０ａ～１２０ｂに相互接続されたクライアントコンピュータシステム１１０ａ～１１０ｂを含む。ネットワーク１８０は、スイッチ１４０を介して第２のネットワーク１９０に結合することができる。このネットワーク１９０を介して、クライアントコンピュータシステム１１０ｃが、クライアントコンピュータシステム１１０ａ～１１０ｂ及びデータストレージアレイ１２０ａ～１２０ｂに結合される。また、ネットワーク１９０は、スイッチ１５０を介してインターネット１６０又はその他の外部ネットワークに結合することもできる。
【００１５】
なお、代替の実施形態では、クライアントコンピュータ及びサーバ、スイッチ、ネットワーク、データストレージアレイ及びデータストレージデバイスの数及びタイプが、図１に示すものに限定されない。１又はそれ以上のクライアントは、様々な時点でオフライン動作することができる。また、動作中、ユーザがネットワークアーキテクチャ１００への接続、切断及び再接続を行うと、個々のクライアントコンピュータの接続タイプが変化することもある。図１に示す構成要素の各々のさらなる説明を手短に行う。まず、データストレージアレイ１２０ａ～１２０ｂにより提供される機能のいくつかの概要について説明する。

【００１６】
ネットワークアーキテクチャ１００では、データストレージアレイ１２０ａ～１２０ｂの各々を、クライアントコンピュータシステム１１０ａ～１１０ｃなどの異なるサーバ及びコンピュータ間のデータの共有に使用することができる。また、データストレージアレイ１２０ａ～１２０ｂを、ディスクのミラーリング、バックアップ及び復元、保存データの保管及び検索、並びにストレージデバイス間のデータ移行に使用することもできる。代替の実施形態では、クラスタを形成するために、１又はそれ以上のクライアントコンピュータシステム１１０ａ～１１０ｃを、高速ローカルエリアネットワーク（ＬＡＮ）を介して互いにリンクさせることができる。互いにリンクされた１又はそれ以上のノードはクラスタを形成し、これによりデータストレージアレイ１２０ａ～１２０ｂの１つに存在するクラスタ共有ボリュームなどのストレージリソースを共有することができる。

【００１７】
データストレージアレイ１２０ａ～１２０ｂの各々は、データ記憶のためのストレージサブシステム１７０を含む。ストレージサブシステム１７０は、複数のストレージデバイス１７６ａ～１７６ｍを含むことができる。これらのストレージデバイス１７６ａ～１７６ｍは、クライアントコンピュータシステム１１０ａ～１１０ｃにデータ記憶サービスを提供することができる。ストレージデバイス１７６ａ～１７６ｍの各々は、読み込み及び書き込み要求を受け取るとともに、各々をアレイ内の行及び列としてアドレス指定可能な複数のデータ記憶位置を含むように構成することができる。１つの実施形態では、ストレージデバイス１７６ａ～１７６ｍ内のデータ記憶位置を、論理的で冗長なストレージコンテナ又はＲＡＩＤアレイ（低価格／独立ディスク冗長アレイ）内に配置することができる。しかしながら、ストレージデバイス１７６ａ～１７６ｍは、ディスクを含まないこともある。１つの実施形態では、ストレージデバイス１７６ａ～１７６ｍの各々が、従来のハードディスクドライブ（ＨＤＤ）とは異なる技術をデータ記憶に利用することができる。例えば、ストレージデバイス１７６ａ～１７６ｍの１又はそれ以上は、永続データを記憶するための固体メモリから成るストレージを含み、又はこれにさらに結合することができる。他の実施形態では、ストレージデバイス１７６ａ～１７６ｍの１又はそれ以上が、スピン注入法、磁気抵抗ランダムアクセスメモリ（ＭＲＡＭ）法、又はその他の記憶技術を利用するストレージを含み、又はこのようなストレージにさらに結合することができる。これらの異なる記憶技術により、ストレージデバイス間で異なる信頼性特性が生じ得る。

【００１８】
ストレージデバイス１７６ａ～１７６ｍの各々において使用される技術及び機構のタイプにより、データオブジェクトマッピング、並びにエラー検出及び訂正に使用するアルゴリズムを決定することができる。これらのアルゴリズムで使用されるロジックを、基本オペレーティングシステム（ＯＳ）１１６、ファイルシステム１４０、ストレージサブシステムコントローラ１７４内の１又はそれ以上のグローバルＲＡＩＤエンジン１７８、及びストレージデバイス１７６ａ～１７６ｍの各々における制御ロジックのうちの１又はそれ以上に含めることができる。

【００１９】
１つの実施形態では、含まれる固体メモリが、ソリッドステートドライブ（ＳＳＤ）技術を含む。通常、ＳＳＤ技術は、フラッシュメモリセルを利用する。当業で周知のように、フラッシュメモリセルは、フローティングゲート内に捕捉され蓄積された電子の範囲に基づく二進値を保持する。完全に消去されたフラッシュメモリセルは、フローティングゲート内に電子を全く又は最低数しか蓄積していない。消去されたフラッシュメモリセルには、シングルレベルセル（ＳＬＣ）フラッシュの二進１などの特定の二進値が関連付けられる。マルチレベルセル（ＭＬＣ）フラッシュでは、消去されたフラッシュメモリセルに二進値１１が関連付けられる。フラッシュメモリセル内の制御ゲートに所与の閾値電圧よりも高い電圧を印加した後、このフラッシュメモリセルは、フローティングゲート内に所与の範囲の電子を捕捉する。従って、プログラムされた（書き込まれた）フラッシュメモリセルには、ＳＬＣフラッシュの二進０などの別の特定の二進値が関連付けられる。ＭＬ

Ｃフラッシュセルでは、制御ゲートに印加された電圧に応じて、プログラムされたメモリセルに複数の二進値の１つを関連付けることができる。

【００２０】

一般的に言えば、ＳＳＤ技術では、読み込みアクセスレイテンシタイムがＨＤＤ技術よりも短い。しかしながら、ＳＳＤの書き込みパフォーマンスは、ＳＳＤ内の未使用のプログラマブルブロックの利用可能性によって大きく影響を受ける。ＳＳＤの書き込みパフォーマンスは、ＳＳＤの読み込みパフォーマンスに比べて大幅に遅いので、同様のレイテンシを予想する一部の機能又は動作に関する問題が生じることがある。また、ＨＤＤ技術とＳＳＤ技術の間の技術及び機構の違いにより、データストレージデバイス１７６ａ～１７６ｍの信頼性特性に違いが生じることがある。

【００２１】

様々な実施形態では、ＳＳＤ内のフラッシュセルに新たなデータが書き込まれる前に、一般にこのフラッシュセルを消去しなければならない。また、様々なフラッシュ技術における消去動作は、ブロック単位で行わなければならない。従って、ブロック（消去セグメント又は消去ブロック）内のフラッシュメモリセルは、全てまとめて消去される。フラッシュ消去ブロックは、複数のページを含むことができる。例えば、１ページのサイズが４キロバイト（ＫＢ）であり、１ブロックが６４ページ、すなわち２５６ＫＢを含むことができる。フラッシュデバイスでは、読み込み動作に比べて消去動作のレイテンシの方が相対的に高いことがあり、これにより対応する書き込み動作のレイテンシが増すことがある。フラッシュ技術のプログラミング又は読み込みは、消去ブロックサイズよりも低い粒度レベルで行うことができる。例えば、フラッシュセルには、バイトサイズ、単語サイズ又はその他のサイズでプログラム又は読み込みを行うことができる。

【００２２】

フラッシュセルには、反復的な消去プログラム動作後に摩耗が生じる。この場合、この摩耗は、ＭＬＣフラッシュセルの基板とフローティングゲートの間の誘電酸化物層に注入され捕捉される電荷によるものである。１つの例では、ＭＬＣフラッシュセルが、１０，０００～１００，０００サイクルなどの、消去及びプログラム動作を受ける回数限界を有することができる。また、ＳＳＤには、別のフラッシュセルの消去又はプログラム中に隣接する又は近隣のフラッシュセルに偶発的状態変化を引き起こすプログラムディスターブエラーが生じることがある。さらに、ＳＳＤは、別のフラッシュセルの読み込み中に近隣のフラッシュセルの偶発的状態変化が生じるリードディスターブエラーも含む。

【００２３】

１又はそれ以上のストレージデバイス１７６ａ～１７６ｍの各々の特性が分かると、より効率的なデータオブジェクトマッピング、並びにエラー検出及び訂正を行うことができる。１つの実施形態では、ストレージコントローラ１７４内のグローバルＲＡＩＤエンジン１７８が、ストレージデバイス１７６ａ～１７６ｍに関して、Ｉ／Ｏ要求の応答時間に一貫性がないこと、対応するアクセスに対するデータが誤っていること、エラー率及びアクセス率のうちの少なくとも１つ又はそれ以上を検出することができる。グローバルＲＡＩＤエンジン１７８は、少なくともこれらの特性に応答して、ストレージデバイス１７６ａ～１７６ｍ内の対応するストレージデバイスグループにいずれのＲＡＩＤデータレイアウトアーキテクチャを利用すべきかを判断することができる。また、グローバルＲＡＩＤエンジン１７８は、ストレージデバイス１７６ａ～１７６ｍの特性に基づいて、デバイス内冗長スキーム及びデバイス間ＲＡＩＤデータレイアウトを動的に変更することができる。

【００２４】

図１に、１つの実施形態による、説明した特徴が可能なシステムの例を示す。さらなる詳細については以下で示す。以下、図１を参照しながら、ネットワークアーキテクチャ１００の構成要素についてさらに説明する。

【００２５】

ネットワークアーキテクチャの構成要素

繰り返すが、図示のように、ネットワークアーキテクチャ１００は、ネットワーク１８０及び１９０を介して互いに及びデータストレージアレイ１２０ａ～１２０ｂに相互接続されたクライアントコンピュータシステム１１０ａ～１１０ｃを含む。ネットワーク１８０及び１９０は、無線接続、直接ローカルエリアネットワーク（ＬＡＮ）接続、ストレージエリアネットワーク（ＳＡＮ）、インターネットなどの広域ネットワーク（ＷＡＮ）接続及びルータなどを含む様々な技術を含むことができる。ネットワーク１８０及び１９０は、１又はそれ以上のＬＡＮを含むことができ、これらは無線であってもよい。ネットワーク１８０及び１９０は、リモートダイレクトメモリアクセス（ＲＤＭＡ）ハードウェア及び／又はソフトウェア、伝送制御プロトコル／インターネットプロトコル（ＴＣＰ／ＩＰ）ハードウェア及び／又はソフトウェア、ルータ、リピータ、スイッチ及び／又はグリッドなどをさらに含むことができる。ネットワーク１８０及び１９０内では、イーサネット（登録商標）、ファイバチャネル、ファイバチャネルオーバーイーサネット（ＦＣｏＥ）及びｉＳＣＳＩなどのプロトコルを使用することができる。スイッチ１４０は、ネットワーク１８０及び１９０の両方に関連するプロトコルを利用することができる。ネットワーク１９０は、伝送制御プロトコル（ＴＣＰ）及びインターネットプロトコル（ＩＰ）、すなわちＴＣＰ／ＩＰなどの、インターネット１６０に使用される通信プロトコルの組と整合することができる。スイッチ１５０は、ＴＣＰ／ＩＰスイッチとすることができる。

【００２６】

クライアントコンピュータシステム１１０ａ～１１０ｃは、デスクトップパソコン（ＰＣ）、ワークステーション、ラップトップ、ハンドヘルドコンピュータ、サーバ、サーバファーム、携帯情報端末（ＰＤＡ）及びスマートフォンなどのあらゆる数の固定又はモバイルコンピュータを表す。一般的に言えば、クライアントコンピュータシステム１１０ａ～１１０ｃは、１又はそれ以上のプロセッサコアを備えた１又はそれ以上のプロセッサを含む。各プロセッサコアは、所定の汎用命令セットに従って命令を実行するための回路を含む。例えば、ｘ８６命令セットアーキテクチャを選択することができる。或いは、Ａｌｐｈａ（登録商標）、ＰｏｗｅｒＰＣ（登録商標）、ＳＰＡＲＣ（登録商標）又はその他のいずれの汎用命令セットアーキテクチャを選択してもよい。プロセッサコアは、データ及びコンピュータプログラム命令を求めてキャッシュメモリサブシステムにアクセスすることができる。キャッシュサブシステムは、ランダムアクセスメモリ（ＲＡＭ）及びストレージデバイスを含む記憶階層に結合することができる。

【００２７】

クライアントコンピュータシステム内の各プロセッサコア及び記憶階層は、ネットワークインターフェイスにさらに接続することができる。クライアントコンピュータシステム１１０ａ～１１０ｃの各々は、ハードウェア構成要素に加え、記憶階層内に記憶された基本オペレーティングシステム（ＯＳ）を含むことができる。この基本ＯＳは、例えば、ＭＳ－ＤＯＳ（登録商標）、ＭＳ－ＷＩＮＤＯＷＳ（登録商標）、ＯＳ／２（登録商標）、ＵＮＩＸ（登録商標）、Ｌｉｎｕｘ（登録商標）、Ｓｏｌａｒｉｓ（登録商標）、又は別の公知のオペレーティングシステムなどの様々な特定のオペレーティングシステムのいずれかを表すことができる。従って、基本ＯＳは、エンドユーザに様々なサービスを提供するとともに、様々なプログラムの実行をサポートするソフトウェアフレームワークを提供することができる。また、クライアントコンピュータシステム１１０ａ～１１０ｃの各々は、高水準バーチャルマシン（ＶＭ）をサポートするために使用されるハイパーバイザを含むことができる。当業者には周知のように、デスクトップ及びサーバ内で仮想化を使用して、ＯＳなどのソフトウェアをシステムのハードウェアから完全に又は部分的に分離することができる。仮想化により、データストレージアレイ１２０ａ～１２０ｂの各々におけるストレージデバイス１７６ａ～１７６ｍに対応する（論理装置番号（ＬＵＮ）などの）論理記憶エンティティなどのように、各々が独自のリソースを有する同じ機械上で複数のＯＳが実行されているという錯覚をエンドユーザに与えることができる。

【００２８】

データストレージアレイ１２０ａ～１２０ｂの各々は、クライアントコンピュータシス

テム１１０ａ～１１０ｃなどの異なるサーバ間のデータの共有に使用することができる。データストレージアレイ１２０ａ～１２０ｂの各々は、データを記憶するためのストレージサブシステム１７０を含む。ストレージサブシステム１７０は、複数のストレージデバイス１７６ａ～１７６ｍを含むことができる。これらのストレージデバイス１７６ａ～１７６ｍの各々は、ＳＳＤとすることができる。コントローラ１７４は、受け取った読み込み／書き込み要求を処理するためのロジックを含むことができる。例えば、少なくともコントローラ１７４において、手短に上述したアルゴリズムを実行することができる。受け取った書き込み要求などの動作のバッチ処理には、ランダムアクセスメモリ（ＲＡＭ）１７２を使用することができる。

【００２９】

記憶媒体１３０に記憶された基本ＯＳ１３２、ファイルシステム１３４、いずれかのＯＳドライバ（図示せず）及びその他のソフトウェアは、ファイル及びＬＵＮへのアクセスを可能にする機能を提供し、これらの機能を管理することができる。基本ＯＳ１３４及びＯＳドライバは、記憶媒体１３０上に記憶された、受け取った要求に対応する１又はそれ以上のメモリアクセス動作をストレージサブシステム１７０内で行うようにプロセッサ１２２により実行可能なプログラム命令を含むことができる。

【００３０】

データストレージアレイ１２０ａ～１２０ｂの各々は、ネットワークインターフェイス１２４を使用してネットワーク１８０に接続することができる。１つの実施形態では、クライアントコンピュータシステム１１０ａ～１１０ｃと同様に、ネットワークインターフェイス１２４の機能をネットワークアダプタカード上に含めることができる。ネットワークインターフェイス１２４の機能は、ハードウェア及びソフトウェアの両方を使用して実装することができる。ネットワークインターフェイス１２４のネットワークカードによる実装上には、ランダムアクセスメモリ（ＲＡＭ）及び読み取り専用メモリ（ＲＯＭ）の両方を含めることができる。１又はそれ以上の特定用途向け集積回路（ＡＳＩＣ）を使用して、ネットワークインターフェイス１２４の機能を提供することができる。

【００３１】

１つの実施形態では、ユーザデータ及び対応するエラー訂正符号（ＥＣＣ）情報のデータレイアウトを最適化しようと努めるデータストレージモデルを作成することができる。１つの実施形態では、このモデルが、ストレージシステム内のストレージデバイスの特性に少なくとも部分的に基づく。例えば、ソリッドステートストレージ技術を利用するストレージシステムでは、特定のデバイスの特性を使用してこのストレージデバイスのためのモデルを作成するとともに、このモデルが、対応するデータストレージ構成アルゴリズムを通知する機能を果たすようにすることもできる。例えば、使用中の特定のストレージデバイスが、時間とともに信頼性の変化を示す場合、データストレージ構成を動的に変更する上でこのような特性を考慮することができる。

【００３２】

一般的に言えば、コンピューティングシステムのために開発されるあらゆるモデルは不完全である。多くの場合、現実のシステムで所与のシステムを完全にモデル化するには、考慮すべき変数が単純に多すぎる。場合によっては、完全ではないが価値のあるモデルを開発することが可能な場合もある。以下でより詳細に説明するように、基礎を成すデバイスの特性に基づいてストレージシステムをモデル化する実施形態を説明する。様々な実施形態では、システムがどのように挙動し得るかに関するいくつかの予測に基づいてデータストレージ構成の選択が行われる。デバイスの挙動によっては、デバイスの特性の理解に基づいて、他のデバイスの挙動より予測しやすいものもある。しかしながら、デバイスの動作は時間とともに変化することがあり、これに応じて、選択したデータレイアウトも変化することがある。本明細書で使用するデバイスの特性とは、全体としてのデバイスの特性、チップ又はその他の構成要素などのデバイスの下位部分の特性、消去ブロックの特性、又はデバイスに関する他のあらゆる特性のことを意味することができる。

【００３３】

デバイス内冗長性

ここで図２を参照すると、動的デバイス内冗長スキームの１つの実施形態を示す汎用ブロック図を示している。当業者には周知のように、ストレージデバイス内の潜在的なセクタエラーの影響を低減するように複数のデバイス内冗長スキームの１つを選択することができる。通常、「セクタ」という用語は、ディスク上の所与のトラック内のセグメントなどの、ＨＤＤ上における基本記憶単位を意味する。ここでは、「セクタ」という用語は、ＳＳＤ上における基本的な割り当て単位（アロケーションユニット）を意味することもできる。

【００３４】

ＳＳＤ内のアロケーションユニットは、ＳＳＤ内の１又はそれ以上の消去ブロックを含むことができる。図２を参照すると、ユーザデータ２１０は、エンドユーザが修正及びアクセスすべき記憶データ、並びにデバイス間エラー訂正符号（ＥＣＣ）データの両方を意味することができる。デバイス間ＥＣＣデータは、ユーザデータを保持する他のストレージデバイス上の１又はそれ以上のページから生成されたパリティ情報とすることができる。例えば、デバイス間ＥＣＣデータは、ＲＡＩＤデータレイアウトアーキテクチャで使用されるパリティ情報とすることができる。ユーザデータ２１０は、ストレージデバイス１７６ａ～１７６ｋのうちの１又はそれ以上に含まれる１又はそれ以上のページ内に記憶することができる。１つの実施形態では、ストレージデバイス１７６ａ～１７６ｋの各々がＳＳＤである。

【００３５】

ＳＳＤ内の消去ブロックは、複数のページを含むことができる。上述したように、１つの実施形態では、１ページが４ＫＢのデータ記憶空間を含むことができる。１消去ブロックは、６４ページ、すなわち２５６ＫＢを含むことができる。他の実施形態では、消去ブロックが１メガバイト（ＭＢ）もの大きさであり、２５６ページを含むことができる。アロケーションユニットのサイズは、アロケーションユニットのオーバーヘッドトラッキングを低減するために、十分に大きなサイズの比較的少ないユニットが与えられるように選択することができる。１つの実施形態では、１又はそれ以上の状態テーブルが、アロケーションユニットの状態（割り当て済み、空き、消去済み、エラー）、摩耗レベル、及びアロケーションユニット内で発生した総エラー数（訂正可能及び／又は訂正不能）を維持することができる。様々な実施形態では、アロケーションユニットのサイズを、所与のデバイスが利用できるアロケーションユニットの数とアロケーションユニットを維持するオーバーヘッドとのバランスをとるように選択することができる。例えば、１つの実施形態では、アロケーションユニットのサイズを、ＳＳＤの総記憶容量の１／１００パーセントとなるように選択することができる。ページ、消去ブロック及びその他の単位構成に関する他の量のデータ記憶空間も可能であり、企図される。

【００３６】

ストレージデバイス内の所与のセクタ又はその他の記憶単位にアクセスできない場合、潜在的セクタエラー（ＬＳＥ）が発生する。この所与のセクタでは、読み込み又は書き込み動作を完了できないことがある。また、訂正不能なエラー訂正符号（ＥＣＣ）エラーが存在することもある。ＬＳＥは、所与のセクタがアクセスを受けるまで検出されないエラーである。従って、この所与のセクタに以前に記憶したあらゆるデータが失われる恐れがある。ストレージデバイスの不具合後のＲＡＩＤ再構成中に１つのＬＳＥが生じた場合、これによりデータ損失に至る恐れがある。ＳＳＤでは、デバイスの使用年数、デバイスサイズ、アクセス率、ストレージの圧縮度、及び以前の訂正可能及び訂正不能なエラーの発生といった統計値のうちの少なくとも１つから、別のＬＳＥの発生確率が増加することがある。所与のストレージデバイス内のＬＳＥ及びデータ損失を防ぐために、この所与のストレージデバイス内で多様なデバイス内冗長スキームの１つを使用することができる。

【００３７】

デバイス内冗長スキームは、所与のストレージデバイス内で、パリティ情報などのＥＣＣ情報を利用する。このデバイス内冗長スキーム及びそのＥＣＣ情報は、所与のデバイス

に対応し、所与のデバイス内に維持することができるが、デバイス自体が内部的に生成して維持できるＥＣＣとは異なる。一般的に言えば、デバイスの内部的に生成され維持されるＥＣＣは、このデバイスを含むシステムからは見えない。所与のストレージデバイスに含まれるデバイス内ＥＣＣ情報を使用して、この所与のストレージデバイス内におけるデータ記憶の信頼性を高めることができる。このデバイス内ＥＣＣ情報は、ＲＡＩＤデータレイアウトアーキテクチャで利用されるパリティ情報などの、別のストレージデバイスに含めることができる他のＥＣＣ情報に加えられる。

【００３８】

極めて効果的なデバイス内冗長スキームは、所与のＲＡＩＤデータレイアウトの信頼性を十分に高めて、パリティ情報を保持するために使用されるデバイスの数を減少させることができる。例えば、各デバイス上のデータを保護するためにデバイス内冗長性が加わった場合、ダブルパリティＲＡＩＤレイアウトをシングルパリティＲＡＩＤレイアウトに置き換えることができる。一定レベルの記憶効率を得るためには、デバイス内冗長スキームにおける冗長性を高めると、所与のストレージデバイスの信頼性が高まる。しかしながら、このようにして冗長性を高めると、この所与のストレージデバイスの入力／出力（Ｉ／Ｏ）性能に関する不利点も増える可能性がある。

【００３９】

１つの実施形態では、デバイス内冗長スキームが、デバイスを、ユーザデータを記憶するための一群の場所に分割する。例えば、ストライプ２５０ａ～２５０ｃによって示すような、ＲＡＩＤレイアウト内のストライプに対応するデバイス内の一群の場所に分割することができる。ユーザデータ又はデバイス間ＲＡＩＤ冗長性情報は、データ２１０によって示すように、ストレージデバイス１７６ａ～１７６ｋの各々の１又はそれ以上のページに記憶することができる。各ストレージデバイス内では、１又はそれ以上のページにデバイス内エラー回復データ２２０を記憶することができる。本明細書では、デバイス内エラー回復データ２２０をデバイス内冗長データ２２０と呼ぶことができる。当業者であれば周知のように、デバイス内冗長データ２２０は、データ２１０内の情報の選択した一部に対してある関数を実行することにより取得することができる。ＸＯＲベースの演算を使用して、デバイス内冗長データ２２０に記憶すべきパリティ情報を導出することができる。デバイス内冗長スキームの他の例としては、シングルパリティチェック（ＳＰＣ）、最大距離分離（ＭＤＳ）消去符号、インタリーブパリティチェック符号（ＩＰＣ）、ハイブリッドＳＰＣ及びＭＤＳ符号（ＭＤＳ＋ＳＰＣ）、及びカラム対角パリティ（ＣＤＰ）が挙げられる。これらのスキームは、データ２２０を計算する方法に応じて、もたらされる信頼性及びオーバーヘッドの点で異なる。このシステムは、上述の冗長性情報に加え、デバイス上の領域のチェックサム値を計算するように構成することもできる。例えば、デバイスに情報が書き込まれた時にチェックサムを計算することができる。このチェックサムは、システムによって記憶される。システムは、デバイスから情報が読み戻された時に再びチェックサムを計算し、これを最初に記憶した値と比較することができる。２つのチェックサムが異なる場合、情報が正しく読み込まれておらず、システムは、他のスキームを使用してデータを回復することができる。チェックサム機能の例には、巡回冗長検査（ＣＲＣ）、ＭＤ５及びＳＨＡ－１がある。

【００４０】

ストライプ２５０ａ～２５０ｃに示すように、所与のストライプ内にデータ２１０を記憶するために使用する幅、すなわちページ数は、ストレージデバイス１７６ａ～１７６ｋの各々において同じものとなり得る。しかしながら、ストライプ２５０ｂ～２５０ｃに示すように、所与のストライプ内にデバイス内冗長データ２２０を記憶するために使用する幅、すなわちページ数は、ストレージデバイス１７６ａ～１７６ｋの各々において同じでない場合がある。１つの実施形態では、所与のストレージデバイスの特性又は挙動の変化により、対応するデバイス内冗長データ２２０を記憶するために使用する幅を少なくとも部分的に決定することができる。例えば、上述したように、フラッシュセルには、ページをプログラムすること又は読み込むことによって近隣のページに支障を来たし、これらの

近隣のページ内にエラーを引き起こし得るプログラムディスターブエラー及びリードディスターブエラーが生じる。ストレージデバイスが古くなってより多くのエラーが生じている場合、対応するデバイス内冗長データ２２０の量が増加することがある。例えば、ストライプ２５０ｂの書き込み動作前に、ストレージデバイス１７６ａ～１７６ｋの各々の特性をモニタし、これを使用してエラー率の増加を予測することができる。ストレージデバイス１７６ｃ及び１７６ｊのエラーの増加が予測されることが検出される可能性がある。これに応答して、ストレージデバイス１７６ｃ及び１７６ｊのデバイス内冗長データ２２０の量が増加することがある。図２のストライプ２５０ａ及び２５０ｂの例では、ストライプ２５０ａ及び２５０ｂのストレージデバイス１７６ｃ及び１７６ｊの記憶されている保護データの量の増加を確認することができる。例えば、この時点で、ストレージデバイス１７６ｃ及び１７６ｊをシングルパリティで保護するのではなく、これらのデバイスをダブルパリティ又はトリプルパリティで保護することができる。なお、デバイス１７６ｃ及び１７６ｊのデバイス内保護の量を増加させても、同じストライプ内の他のデバイスにおける量を対応して増加させる必要はない。むしろ、ストライプのデータは、各デバイスにおいて望むように異なるレベルの保護を有することができる。

【００４１】

様々な実施形態では、所与のレベルのデータ保護の増減を選択的に行うことができる。例えば、１つの実施形態では、上記の例におけるストレージデバイス１７６ｃ及び１７６ｊなどの、より多くのエラーを生成することが検出されたストレージデバイスに対してしか保護の増加を行わなくてよい。別の実施形態では、ストレージデバイス１７６ｃ及び１７６ｊがより多くのエラーを生成することが検出された場合、ストレージデバイス１７６ａ～１７６ｋの各々に対して保護の増加を行うことができる。１つの実施形態では、デバイス１７６ｋなどのパリティデバイス上におけるデバイス内保護の量を増加させるために、ストライプ内で保護されているデータの量を減少させることが必要となる場合がある。例えば、所与のストライプのパリティデバイス上に記憶されたデバイス内データの量を増加させると、このデバイスが記憶するストライプ内のデータのためのパリティデータの量が必然的に減少する。このパリティデータの量が、ストライプ内のデータを全て保護するために必要な量未満に減少した場合、パリティ保護を継続することが望ましい場合にはストライプ内のデータを減少させなければならない。ストライプ内に記憶されるデータの量の減少させる代替案として、パリティデータを記憶するための異なるデバイスを選択することもできる。様々な選択肢が可能であり、企図される。また、本明細書に記載する図２及びその他の図には、（１７６ｋなどの）別個のパリティデバイスを示していることがあるが、様々な実施形態では、パリティを単一のデバイスに記憶するのではなく複数のデバイスに分散させることができる。従って、別個のパリティデバイスの図示は、一般に説明を簡単にするための論理的な表現と見なすことができる。

【００４２】

ここで図３を参照すると、データストレージサブシステムにおけるデバイス内保護を調整する方法３００の１つの実施形態を示している。一般に、ネットワークアーキテクチャ１００及びデータストレージアレイ１２０ａ～１２０ｂ内で具体化される構成要素は、方法３００に従って動作することができる。この実施形態のステップを順番に示す。しかしながら、ステップによっては、図示のものとは異なる順序で行なうことができるもの、同時に行うことができるもの、他のステップと組み合わせることができるもの、及び別の実施形態には存在しないものもある。

【００４３】

ブロック３０２において、ストレージデバイスにユーザデータを記憶するための第１の空き容量を決定する。このユーザデータは、データ２１０に関して上述したように、ＲＡＩＤアーキテクチャで使用されるエンドユーザアプリケーション又はデバイス間パリティ情報内で使用されるデータとすることができる。この第１の空き容量は、上述したようなストレージデバイス内の１又はそれ以上のページを含むことができる。１つの実施形態では、ストレージコントローラ１７４内のグローバルＲＡＩＤエンジン１７８が、ストレー

ジデバイス１７６ａ～１７６ｍの各々から挙動統計値を受け取る。グローバルＲＡＩＤエンジン１７８は、ストレージデバイス１７６ａ～１７６ｍの２又はそれ以上を含む所与のデバイスグループについて、ＲＡＩＤデータレイアウト及びこれらの２又はそれ以上のストレージデバイスの各々に維持すべき最初のデバイス内冗長性の量を決定することができる。ブロック３０４において、ＲＡＩＤエンジン１７８は、対応するデバイス内保護データをストレージデバイスに記憶するための第２の空き容量を決定することができる。この第２の空き容量は、ストレージデバイス内の１又はそれ以上のページを含むことができる。デバイス内保護データは、上述したデバイス内冗長データ２２０に対応することができる。

【００４４】

ブロック３０６において、所与のデバイスグループに含まれる各ストレージデバイス内の第１の空き容量にデータを書き込む。１つの実施形態では、ユーザデータ及びデバイス間パリティ情報が、いずれも所与のデバイスグループに含まれる複数のストレージデバイスにわたる単一のＲＡＩＤストライプとして書き込まれる。再び図２を参照して分かるように、対応する書き込まれるデータの幅は、各ストレージデバイスにおいて同じである。ブロック３０８において、ＥＣＣアルゴリズム、ＸＯＲベースのアルゴリズム、又はその他のあらゆる適当なアルゴリズムにより、デバイス内保護データを生成する。また、システムは、正しく取り出されなかったデータを識別しやすくするためにチェックサムを生成することができる。ブロック３１０において、生成されたデバイス内保護データを、ストレージデバイス内の第２の空き容量に書き込む。

【００４５】

ブロック３１２において、ＲＡＩＤエンジン１７８は、１又はそれ以上のストレージデバイスの挙動をモニタすることができる。１つの実施形態では、ＲＡＩＤエンジン１７８が、対応するストレージデバイスのモデルを含み、このモデルに入力すべき挙動統計値をストレージデバイスから受け取ることができる。このモデルは、ストレージデバイスの既知の特性を利用することにより、ストレージデバイスの挙動を予測することができる。例えば、このモデルは、所与のストレージデバイスのエラー率の増加が近いことを予測することができる。ＲＡＩＤエンジン１７８は、信頼性に影響を与える所与のストレージデバイスの特性を検出した（条件付きブロック３１４）場合、ブロック３１６において、データ及び対応するデバイス内冗長データを記憶するための第１及び第２の空き容量を調整することができる。例えば、ＲＡＩＤエンジンは、少なくともデバイスの使用年数、アクセス率及びエラー率などの上述した統計値をモニタすることができる。再び図２を参照して分かるように、ＲＡＩＤエンジン１７８は、ストレージデバイス１７６ｃ及び１７６ｊのエラー数が増加したことを検出することができる。或いは、ＲＡＩＤエンジンは、ストレージデバイス１７６ｃ及び１７６ｊのエラー数の増加を予測することができる。従って、ＲＡＩＤエンジン１７８は、第２のストライプ２５０ｂを書き込む前に、ストレージデバイス１７６ａ～１７６ｋの各々にデータ２１０及びデータ２２０を記憶するために使用するページ数を調整することができる。同様に、ＲＡＩＤエンジン１７８は、ストレージデバイス１７６ｂの信頼性が低下していることを検出することができる。従って、ＲＡＩＤエンジン１７８は、第３のストライプ２５０ｃを書き込む前に、ストレージデバイス１７６ａ～１７６ｋの各々にデータ２１０及びデータ２２０を記憶するために使用するページ数を再び調整することができる。

【００４６】

ストレージデバイス特性のモニタリング

ここで図４を参照すると、ストレージサブシステムの１つの実施形態の汎用ブロック図を示している。１又はそれ以上のデバイスグループ１７３ａ～１７３ｍの１つにおいて、１又はそれ以上のストレージデバイス１７６ａ～１７６ｍの各々を分割することができる。同様に、他のデバイスを含む他のデバイスグループも存在し得る。各ストレージデバイスの対応する動作キュー及び状態テーブルを、デバイスユニット４００ａ～４００ｗの１つに含めることができる。これらのデバイスユニットを、ＲＡＭ１７２に記憶することが

できる。デバイスグループ１７３ａ～１７３ｍの各々に関し、対応するＲＡＩＤエンジン１７８ａ～１７８ｍを含めることができる。各ＲＡＩＤエンジン１７８は、対応するデバイスグループ内のストレージデバイスの各々の統計値を追跡するモニタ４１０を含むことができる。データレイアウトロジック４２０は、対応するストレージデバイス内の、ユーザデータ、デバイス間冗長データ及びデバイス内冗長データに対して割り当てるべき空き容量を決定することができる。ストレージコントローラ１７４は、ウェアレベリング、ガベージコレクション、Ｉ／Ｏスケジューリング、重複除外、並びに着信及び発信パケットのプロトコル変換といったタスクのうちの少なくとも１つを行うためのその他の制御ロジック４３０を含むことができる。

【００４７】

ここで図５を参照すると、デバイスユニットの１つの実施形態の汎用ブロック図を示している。デバイスユニットは、デバイスキュー５１０及びテーブル５２０を含むことができる。デバイスキュー５１０は、読み込みキュー５１２、書き込みキュー５１４、及びその他の動作キュー５１６などの１又はそれ以上の他のキューを含むことができる。各キューは、１又はそれ以上の対応する要求５３０ａ～５３０ｄを記憶するための複数のエントリを含むことができる。例えば、対応するＳＳＤのデバイスユニットは、少なくとも読み込み要求、書き込み要求、トリム要求及び消去要求などを記憶するためのキューを含むことができる。テーブル５２０は、状態データ又は統計値５３０を記憶するための複数のエントリを各々が含む１又はそれ以上の状態テーブル５２２ａ～５２２ｂを含むことができる。また、この図及びその他の図には、キュー及びテーブルが特定数のエントリを含むように示しているが、必ずしもエントリ自体が互いに対応するわけではない。さらに、キュー、テーブル及びエントリの数は、図示のものとは異なることがあり、また互いに異なることもある。

【００４８】

ここで図６を参照すると、所与のデバイスに対応する状態テーブルの１つの実施形態を示す汎用ブロック図を示している。１つの実施形態では、このようなテーブルが、ＳＳＤなどの所与のストレージデバイスの状態情報、エラー情報及び摩耗率情報に対応するデータを含むことができる。この情報に、対応するＲＡＩＤエンジンがアクセスすることにより、このＲＡＩＤエンジンは、デバイス間保護及びデバイス内保護の両方に使用するデータ記憶及びスキームのために割り当てられた空間を動的に変更できるようになる。１つの実施形態では、この情報が、デバイスの使用年数６０２、エラー率６０４、デバイス６０６上で検出された総エラー数、回復可能なエラー数６０８、回復不能なエラー数６１０、デバイスのアクセス率６１２、記憶されたデータの使用年数６１４、及び割り当て空間６１６ａ～６１６ｎの１又はそれ以上の割り当て状態、のうちの少なくとも１つ又はそれ以上を含むことができる。これらの割り当て状態は、使用中、空き、及びエラーなどを含むことができる。

【００４９】

フレキシブルなＲＡＩＤレイアウト

ここで図７を参照すると、フレキシブルなＲＡＩＤデータレイアウトアーキテクチャの１つの実施形態を示す汎用ブロック図を示している。ＲＡＩＤエンジンは、ストレージデバイス１７６ａ～１７６ｋに使用すべき保護レベルを決定することができる。例えば、ＲＡＩＤエンジンは、ストレージデバイス１７６ａ～１７６ｋにＲＡＩＤダブルパリティを利用すると決定することができる。デバイス間冗長データ２４０は、対応するユーザデータから生成されたＲＡＩＤダブルパリティ値を表すことができる。１つの実施形態では、ストレージデバイス１７６ｊ及び１７６ｋが、ダブルパリティ情報を記憶することができる。他のレベルのＲＡＩＤパリティ保護も可能であり、企図されると理解されたい。また、他の実施形態では、ダブルパリティ情報を、各ＲＡＩＤストライプのストレージデバイス１７６ｊ及び１７６ｋに記憶するのではなく、この記憶をストレージデバイス間で循環させることもできる。ダブルパリティ情報がストレージデバイス１７６ｊ及び１７６ｋに記憶されるように示しているのは、図示及び説明を簡単にするためである。

【００５０】
ここで図８を参照すると、フレキシブルなＲＡＩＤデータレイアウトアーキテクチャの別の１つの実施形態を示す汎用ブロック図を示している。図７に示す例と同様に、ストレージデバイス１７６ａ～１７６ｋにはダブルパリティを使用することができる。この例では、ＲＡＩＤダブルパリティについて説明するが、ＲＡＩＤデータレイアウトアーキテクチャではあらゆる量の冗長性を選択することができる。
【００５１】
動作中、ＲＡＩＤエンジン１７８は、ストレージデバイス１７６ａ～１７６ｋの特性をモニタして、デバイスが最初の又はその他の所与の信頼性レベルよりも高い信頼性レベルを示していると判断することができる。これに応答して、ＲＡＩＤエンジン１７８は、ＲＡＩＤ保護をＲＡＩＤダブルパリティからＲＡＩＤシングルパリティに変更することができる。他のＲＡＩＤデータレイアウトアーキテクチャでは、サポートされている冗長性の量を別様に減少させることができる。他の実施形態では、ストレージデバイス１７６ａ～１７６ｋのモニタ及び保護レベルの変更を、ストレージコントローラ１７４内の他のロジックによって行うことができる。
【００５２】
引き続き上記の例を参照すると、所与のＲＡＩＤストライプ上で実行される後続の書き込み動作に関しては、シングルパリティ情報のみを生成して記憶することができる。例えば、サポートされている冗長性の量を変更した後の書き込み動作に関しては、後続のＲＡＩＤストライプ内でストレージデバイス１７６ｋを使用しなくてもよい。また、ストレージデバイス１７６ｋに記憶されたデータを無効にし、これによりストレージを解放することもできる。その後、ストレージデバイス１７６ｋの解放されたデータに対応するページを、他の用途に再割り当てすることができる。このパリティ保護の量を低減し、パリティ保護データを記憶するために以前に使用されていた空間を解放するプロセスのことを、「パリティシュレッディング」と呼ぶことができる。さらに、ストレージデバイス１７６ｋがＳＳＤである実施形態では、ストライプ２５０ａ内のページを書き換える前に、ストレージデバイス１７６ｋ内で１又はそれ以上の消去動作を行うことができる。
【００５３】
上記のパリティシュレッディングの例を引き続き参照すると、パリティシュレッディング後にストライプ２５０ａ内のストレージデバイス１７６ｋの再割り当てされたページに記憶されたデータは、ストライプ２５０ａに対応しない他のＲＡＩＤストライプのユーザデータ又は対応するＲＡＩＤシングルパリティ情報を保持することができる。例えば、ストライプ２５０ａ内のストレージデバイス１７６ａ～１７６ｊに記憶されたデータは、パリティシュレッディング前に実行される１又はそれ以上の書き込み動作に対応することができる。ストライプ２５０ａ内のストレージデバイス１７６ｋに記憶されたデータは、パリティシュレッディング後に実行される１又はそれ以上の書き込み動作に対応することができる。同様に、ストライプ２５０ｂ内のストレージデバイス１７６ａ～１７６ｊに記憶されたデータは、パリティシュレッディング前に実行される１又はそれ以上の書き込み動作に対応することができる。ストライプ２５０ｂ内のストレージデバイス１７６ｋのページは、解放後に消去して、サポートされている冗長性の量を変更した後に実行される１又はそれ以上の書き込み動作に対応するデータに書き換えることができる。なお、冗長性情報が複数のストレージデバイスを循環する場合、このスキームはより一層効果的となり得る。このような実施形態では、シュレッディングにより解放される空間も、同様にストレージデバイスにわたって分散される。
【００５４】
再び図８を参照すると、ストライプ２５０ｃ内のストレージデバイス１７６ｋに示す割り当て解除されたページは、パリティシュレッディング前にＲＡＩＤダブルパリティ情報を記憶していた可能性のある記憶位置を表す。しかしながら、現時点でこれらのページは無効であり、未だ再割り当てされていない。上記の例では、ストレージデバイス１７６ｋ内のページを解放して再割り当てする方法及びタイミングが、ＳＳＤの特定の特性によっ

て決まる。これらの特性の例として、１又はそれ以上のページを再プログラムする（書き換える）前に少なくとも消去ブロック全体を消去することが挙げられる。図８からわかるように、パリティをシュレッドする際に、デバイス全体をシュレッドする必要はない。むしろ、個々のストライプに関して、望むようにパリティをシュレッドすることができる。同様に、ストライプのパリティ保護を増加させることもでき、別のデバイス上に記憶された保護データをストライプに追加することもできる。

【００５５】

ここで図９を参照すると、ＲＡＩＤレイアウトを動的に決定する方法の１つの実施形態を示している。一般に、ネットワークアーキテクチャ１００及びデータストレージアレイ１２０ａ～１２０ｂ内で具体化される構成要素は、方法９００に従って動作することができる。図９には、２つのプロセス９１０及び９２０を示している。これらのプロセスの各々は、同時に又は所与の順序で動作することができる。さらに、この実施形態のステップを順番に示す。しかしながら、ステップによっては、図示のものとは異なる順序で行なうことができるもの、同時に行うことができるもの、他のステップと組み合わせることができるもの、及び別の実施形態には存在しないものもある。ブロック９１０は、ストレージ制御システムが、システム内のストレージデバイスの特性及び挙動をモニタするプロセスを示す（ブロック９１２）。例えば、図６で説明したような特性を観察及び／又は記録することができる。信頼性の変化などの特定の状態が検出された場合（判定ブロック９１４）、記憶したデータに使用する保護の量を変更することができる（ブロック９１６）。例えば、所与のデバイスの使用年数が比較的短い場合には、デバイスの信頼性が分からないことがある（例えば、デバイスが「初期故障」を起こして比較的早い時期に機能しなくなることもある）。従って、ＲＡＩＤストライプ毎に１又はそれ以上の追加のストレージデバイスを使用して、パリティ情報を記憶することができる。その後の時点で、時間とともにこれらのデバイスが信頼できると判明した時に、この追加の保護を除去することができる。様々な実施形態では、デバイスのエラー率に関する特性を維持することができる。例えば、訂正可能な及び／又は訂正不能なエラーに関する特性を維持し、これらを使用して所与のデバイスの信頼性に関する判断を行うことができる。ストレージコントローラは、この情報に基づいて、デバイス又はストライプの様々な保護レベルを動的に変更することができる。

【００５６】

図９のブロック９２０は、ストレージのストライプ又はその他の部分を割り当てる時点で（判定ブロック９２２）、データに使用すべきレイアウト及び保護レベルに関する決定を行なう（ブロック９２４）ことができるプロセスを大まかに示すものである。なお、この時にブロック９１０のプロセスを実施することもできる。或いは、既にプロセス９１０によって保護レベルが決定され記憶されている場合もある。この場合、ブロック９２４の決定は、この記憶されているデータに基づくことができる。１つの実施形態では、所与のレイアウトを決定すると、このレイアウトに使用すべき特定のデバイスをデバイスグループから選択することができる（ブロック９２５）。例えば、１つの実施形態では、２０個のデバイスからなるグループを使用することができる。５＋２のレイアウトを決定した場合、この２０個のデバイスからなるグループから任意の７つのデバイスを使用するように選択することができる。また、選択した５＋２のレイアウトによるその後の書き込みにおいて、この同じ７つのデバイスを使用する必要はない。ストライプのレイアウト、保護レベル及びデバイスを決定した後、このストライプに書き込みを行うことができる（ブロック９２６）。

【００５７】

様々な実施形態では、ＲＵＳＨアルゴリズムを利用して、所与のストライプのデータ及び冗長性情報がいずれのデバイス上に存在するようになるかを判断することができる。例えば、ＲＵＳＨアルゴリズムを使用して、ストレージデバイス１７６ａ～１７６ｋの所与のストライプの８＋２のＲＡＩＤレイアウトに利用すべき特定のデバイスを選択することができる。一般的に言えば、本明細書では、一般にＭ＋Ｎのレイアウトは、所与のデータ

ストライプのＭ個のデータデバイス及びＮ個のパリティデバイスを含むレイアウトを表すことができる。また、上述したように、パリティを完全に特定のデバイス内に配置するのではなく、デバイスにわたって分散させることもできる。従って、８＋２のレイアウトは、１０個のデバイスにわたってストライピングされたデータ及びパリティを含むことができ、これらのデバイスのうちの８つがデータを記憶し、２つがパリティを記憶する。その後のある時点で、１２＋２のレイアウトを選択することができる。このように、書き込みを行う（例えば、ストライプを書き込む）時点で、所望のレイアウト及び保護特性を動的に決定することができる。１つの実施形態では、ストレージデバイス１７６ａ～１７６ｋが、３０個、５０個、又はそれ以上のストレージデバイスなどの、１０個よりも多くのストレージデバイスを含むことができる。しかしながら、８＋２のレイアウトによるストライプでは、ストレージデバイスのうちの１０個しか利用されない。なお、デバイスのうちの任意の１０個を選択し、ストライプを記憶する際に使用するこれらの１０個のデバイスを選択するために、任意の好適なアルゴリズムを使用することができる。例えば、ＣＲＵＳＨアルゴリズムを使用して、所与の８＋２のＲＡＩＤレイアウトのために、ストレージデバイス１７６ａ～１７６ｋのうちのいずれの１０個を利用すべきかを選択することができる。

【００５８】

ストレージデバイス１７６ａ～１７６ｋのために選択される８＋２のＲＡＩＤレイアウトの１つの例では、これらのストレージデバイスのうちの２つを使用して、パリティ情報などのエラー訂正符号（ＥＣＣ）情報を記憶することができる。この情報を使用して、再構成読み込み要求を行うことができる。再び図８を参照すると、この例では、ＲＡＩＤダブルパリティ情報を記憶するためにストレージデバイス１７６ｊ及び１７６ｋを選択することができる。この場合も、パリティ情報を常に同じストレージデバイスに記憶するのではなく、ＲＡＩＤアレイに含まれるストレージデバイス１７６ａ～１７６ｋの各々の間で循環式に記憶することができる。図示及び説明を簡単にするために、ストレージデバイス１７６ｊ及び１７６ｋがＲＡＩＤダブルパリティを記憶しているものとして説明する。

【００５９】

ブロック９２６において、書き込み動作の実行中に、ＲＡＩＤアレイに含まれる複数のストレージデバイスにわたり、メタデータ、ユーザデータ、デバイス内パリティ情報及びデバイス間パリティ情報をＲＡＩＤストライプとして書き込むことができる。ブロック９１２において、ＲＡＩＤエンジン１７８は、ＲＡＩＤアレイ内の１又はそれ以上のストレージデバイスの挙動をモニタすることができる。１つの実施形態では、ＲＡＩＤエンジン１７８が、図４に示すようなモニタ４１０及びデータレイアウトロジック４２０を含むことができる。ＲＡＩＤエンジン１７８は、少なくとも所与のストレージデバイスの使用年数、エラー数及びエラータイプ、最後のデータ割り当て以降に検出された構成変更、所与のデータの使用年数、及びＲＡＩＤアレイ内の記憶空間の現在の使用状況などをモニタすることができる。

【００６０】

ＲＡＩＤエンジン１７８によりモニタされたデータは、図４に示すデバイスユニット４００ａ～４００ｗの１つなどのＲＡＭ１７２に記憶することができる。このデータを記憶するために、図５及び図６に示す例のようなテーブルを使用することができる。対応するＲＡＩＤエンジンに含まれるロジックは、ストレージデバイスの最新の統計値をモニタすることにより、ストレージデバイスの挙動を検出して予測することができる。例えば、このモデルは、所与のストレージデバイスのエラー率の増加が近いことを予測することができる。

【００６１】

（単複の）ストレージデバイスの信頼性が増したことが検出された（条件付きブロック９０８）場合、ブロック９１０において、ＲＡＩＤエンジンは、システム内のデータ保護レベルを低下させることができる。例えば、１つの実施形態では、ストレージサブシステムに記憶されているパリティ情報の量を減少させることができる。上記の例に関して、Ｒ

ＡＩＤエンジンは、対応する８＋２のＲＡＩＤアレイのＲＡＩＤダブルパリティをＲＡＩＤシングルパリティに低下させて、８＋１のＲＡＩＤアレイに変換することができる。他の例では、所与のＲＡＩＤアレイが、ブロック９１６の前に、ＲＡＩＤアーキテクチャにおいてＮレベルの量の冗長性又はパリティを利用していることがある。ブロック９１６において、ＲＡＩＤエンジンは、Ｎ＞１かつ１＜ｍ＜Ｎとする（Ｎ－ｍ）レベルの量の冗長性を利用すると決定することができる。従って、その後の所与のＲＡＩＤストライプの書き込み動作中には、所与のＲＡＩＤストライプ内で、より少ないｍ個のストレージデバイスに書き込みが行われるようになる。

【００６２】

ＲＡＩＤエンジン（又は別の構成要素）は、システム内のデータ保護レベルを低下させるために、上述したようなパリティシュレッディングを行うことができる。その後、ストレージコントローラ１７４は、シュレッディング動作の結果として解放されたページを、その後の書き込み動作で使用されるように再割り当てすることができる。

【００６３】

上述したように、ストレージデバイス１７６ａ～１７６ｋの各々が古くなってデータで満たされると、ＲＡＩＤアレイから追加のパリティ情報を除去することができる。メタデータ、ユーザデータ、対応するデバイス内冗長性情報、及びデバイス間冗長性情報の一部は残存する。８＋２のＲＡＩＤアレイを使用した上記の例に関しては、ストレージデバイス１７６ａ～１７６ｊに記憶された情報が残存する。一方、追加のデバイス間冗長性情報又は追加のパリティ情報は、ＲＡＩＤアレイから除去することができる。例えば、ストレージデバイス１７６ｋに記憶された追加のパリティ情報を、ＲＡＩＤストライプから除去することができる。

【００６４】

上記の例でストレージデバイス１７６ａ～１７６ｊに記憶された情報などの残存した情報は、適所に残存することができる。上記の例におけるストレージデバイス１７６ｋ内の対応するページなどの、追加のパリティ情報を記憶する記憶空間は、その後の書き込み動作のために再利用して再割り当てすることができる。１つの実施形態では、個々の新たな割り当てにより、新たな仮想アドレスが受け取られる。個々の新たな割り当ては、所与のサイズ、所与の配置又は構成を有することができ、所与の（仮想又は物理）記憶空間に適応することができる。１つの実施形態では、ストレージデバイス１７６ａ～１７６ｋの各々及びストレージデバイス内の各割り当てページが、識別情報を含むヘッダを有する。このような識別情報により、所与の構成を変更することなく、解放された追加のパリティ情報のために記憶空間を再利用できるようになる。

【００６５】

ストレージデバイス１７６ａ～１７６ｋの１又はそれ以上がＳＳＤである実施形態では、消去ブロック内の１又はそれ以上のページを再プログラムする前に消去ブロックが消去される。従って、ストレージデバイス１７６ｋがＳＳＤである実施形態では、ストレージデバイス１７６ｋ内の解放されたページを再プログラムする前に、対応する消去ブロックが消去される。元々の８＋２のＲＡＩＤアレイを使用する上記の例に関しては、ページにデータ２１０を再プログラムする前に、ストライプ２５０ａ～２５０ｂ内のストレージデバイス１７６ｋの１又はそれ以上の消去ブロックが消去される。この時、元々の８＋２のＲＡＩＤアレイは８＋１のＲＡＩＤアレイになっており、ストレージデバイス１７６ｊは、パリティシュレッディング前に書き込まれたＲＡＩＤストライプにシングルパリティ情報を提供する。

【００６６】

当業者には周知のように、所与のストレージデバイスの読み込み又は書き込みエラー中には、対応するＲＡＩＤストライプ内のサポートされているデバイス間パリティ情報からデータを再構成することができる。この再構成されたデータをストレージデバイスに書き込むことができる。しかしながら、この再構成されたデータがストレージデバイスに対して書き込みエラーとなった場合、このストレージデバイス上に記憶されている全てのデー

タを、対応するパリティ情報から再生することができる。この再生されたデータは、別の場所に再配置ことができる。フラッシュメモリでは、フラッシュ変換層（ＦＴＬ）が、データの記憶位置をリマップする。また、フラッシュメモリでは、データの再配置が、消去ブロック内の対応するページを再プログラムする前に消去ブロック全体を消去することを含む。マッピングテーブルを消去ブロック対ページの粒度に維持することにより、リマッピングテーブルをよりコンパクトにすることができる。さらに、再配置中には、パリティシュレッディング中に解放された追加のページを使用することができる。

【００６７】

オフセットパリティ

ここで図１０を参照すると、フレキシブルなＲＡＩＤデータレイアウトアーキテクチャのさらに別の実施形態を示す汎用ブロック図を示している。図８に示す汎用ブロック図と同様に、フレキシブルなＲＡＩＤデータレイアウトアーキテクチャを使用することができる。ストレージデバイス１７６ａ～１７６ｋは、複数のストレージデバイスにわたってレイアウトされた複数のＲＡＩＤストライプを含む。ストレージデバイス１７６ａ～１７６ｋの各々は複数のページを含むが、図を簡単にするために、ページ１０１０及びページ１０２０にしかラベルを付けていない。図示の例では、ストレージデバイス１７６ｊ及び１７６ｋがダブルパリティ情報を記憶するダブルパリティＲＡＩＤデータレイアウトが選択されている。

【００６８】

ストレージデバイス１７６ａ～１７６ｋのページの各々は、特定のタイプのデータを記憶する。一部のページは、ユーザデータ２１０及び対応する生成されたデバイス間パリティ情報２４０を記憶する。他のページは、対応する生成されたデバイス内パリティ情報２２０を記憶する。さらに他のページは、メタデータ２４２を記憶する。メタデータ２４２は、ページヘッダ情報、ＲＡＩＤストライプ識別情報、及び１又はそれ以上のＲＡＩＤストライプのログデータなどを含むことができる。ストレージデバイス１７６ａ～１７６ｋのページの各々は、デバイス間パリティ保護及びデバイス内パリティ保護に加え、各所与のページに記憶されたチェックサムなどの追加の保護を含むこともできる。様々な実施形態では、各ストライプの最初の単一のメタデータページを、他のストライプヘッダから再生することができる。或いは、データをデバイス間パリティによって保護できるように、このページがパリティシャード内の異なるオフセットに存在することもできる。「シャード」は、デバイスの一部を表す。従って、パリティシャードは、パリティデータを記憶するデバイスの一部を意味する。

【００６９】

物理層

様々な実施形態では、本明細書で説明するシステムが物理層を含み、これを通じてシステムの他の要素がストレージデバイスと通信することができる。例えば、スケジューリングロジック、ＲＡＩＤロジック及びその他のロジックは、ソフトウェア及び／又はハードウェアのあらゆる好適な組み合わせを含む物理層を介してストレージデバイスと通信することができる。一般に、物理層は、永続ストレージへのアクセスの提供及びデータストレージの完全性に関する機能の実行を含む様々な機能を実行する。

【００７０】

図１１Ａに、５００ＧＢのデバイスの仮想デバイスレイアウトの１つの実施形態を示す。様々な実施形態では、本明細書で説明するストレージデバイスの最初にパーティションテーブル１１０１をフォーマットし、デバイスの最後にパーティションテーブルのコピーをフォーマットすることができる。また、最初と最後のブロックにデバイスヘッダ１１０３を記憶することができる。例えば、フラッシュベースのストレージデバイスでは、最初と最後の消去ブロックにデバイスヘッダを記憶することができる。上述したように、消去ブロックは、典型的には２５６ＫＢ～１ＭＢのフラッシュ構造である。第１の消去ブロック内には、追加の未使用空間を確保することができる（パディング１１０５）。各デバイスには、ログ及び診断情報１１０７を書き込むための第２の消去ブロックを確保すること

ができる。これらの間の残りの消去ブロックは、複数の消去ブロックのアロケーションユニット（ＡＵ）１１０９に分割される。ＡＵサイズは、デバイス毎に十分な数のＡＵが存在して良好な割り当て粒度が得られるように選択することができる。１つの実施形態では、オーバーヘッドを避けるのに十分な多さではあるが、追跡しやすいように多すぎない単位での割り当てを可能にするように、デバイス上に１０，０００個ほどの範囲内のＡＵが存在する。ＡＵ（割り当て済み／空き／消去済み／不良）の状態の追跡は、ＡＵ状態テーブルに維持することができる。ＡＵの摩耗率は、摩耗レベルテーブルに維持することができ、総エラー数は、ＡＵエラーテーブルに維持することができる。

【００７１】

様々な実施形態では、物理層が、（異なるノード上に存在し得る）デバイスの組にわたる各デバイス内の１つのセグメントシャードを含むセグメント内で空間を割り当てる。図１１Ｂに、セグメント及びそのセグメントの様々な識別可能部分の１つの実施形態を、１つの考えられるセグメントレイアウトで示す。図示の実施形態では、複数のデバイスに記憶された単一のセグメントを示している。データデバイス：データ０～データＮ、並びにパリティデバイス：パリティＰ及びパリティＱを示す。１つの実施形態では、各セグメントシャードが、各デバイス上でシャードのサイズが等しくなるように、デバイスに１又はそれ以上のアロケーションユニットを含む。セグメントシャードを示すために、セグメントシャード１１２３を挙げている。図１１Ｂには、Ｉ／Ｏ読み込みサイズ１１２７も示しており、これは１つの実施形態ではページに対応する。Ｉ／Ｏシャードのページパリティの１又はそれ以上のページを含むことができるＩ／Ｏパリティチャンク１１２９も示している。

【００７２】

１つの実施形態では、各セグメントが、以下のパラメータのうちの１又はそれ以上を含むことができる独自の構成を有する。

（１）ＲＡＩＤレベル－セグメント内のデバイス間保護に使用されるＲＡＩＤレベル。これにより、ミラーリング、パリティ、又はＥＣＣ　ＲＡＩＤ、及びどれだけのセグメントシャードがパリティを含むかを決定することができる。

（２）デバイスレイアウトＩ／Ｏシャードサイズ－書き込み中に各デバイスにわたるストライピングに使用するサイズを表す。このサイズは、典型的には２５６ＫＢ～１ＭＢとなり、恐らくは各デバイス上の消去ブロックサイズの倍数となる。図１１Ｂには、例示目的でＩ／Ｏシャードサイズ１１２５を挙げている。

（３）Ｉ／Ｏ読み込みサイズ－論理的読み込みサイズである。各Ｉ／Ｏシャードを一連の論理ページとしてフォーマットすることができる。さらに各ページは、ページ内のデータのヘッダ及びチェックサムを含むことができる。読み込みが発行された場合、この読み込みは、１又はそれ以上の論理ページに対するものとなり、チェックサムを使用して各ページ内のデータを検証することができる。

（４）Ｉ／ＯシャードＲＡＩＤレベル－Ｉ／Ｏシャードは、再生中に見つかった潜在的なエラーを処理すべきシャード内パリティを有する。このパラメータは、どのタイプのパリティがシャード内保護に使用されているか、従ってどれだけのシャード内パリティのコピーが維持されるかを決定する。

（５）Ｉ／Ｏパリティチャンク－様々な実施形態では、ストレージデバイスが、ページ単位でＥＣＣを行うことができる。従って、エラーが見つかった場合、物理ページ全体の不具合を示している可能性がある。Ｉ／Ｏパリティチャンクは、セグメント内の各デバイス上の物理ページサイズの最小公倍数であり、Ｉ／Ｏパリティチャンク又はＩ／Ｏ読み込みサイズのうちの大きい方にＩ／Ｏシャードをストライピングすることにより、シャード内パリティが計算される。ページパリティの１又はそれ以上のページを含めることができる。様々な実施形態では、チェックサム検証が失敗に終わった場合、このパリティを使用してデータを再生することができる。

【００７３】

様々な実施形態では、個々の新たなセグメントに書き込みが行われる際に、そのセグメ

ントのＲＡＩＤ構成が選択される。ＲＡＩＤ構成の選択は、現在のアクティブなノード及びデバイスの組、並びにセグメント内のデータのタイプなどの因子に基づくことができる。例えば、１０個のノード又はデバイスが利用可能な場合、（８＋２）のＲＡＩＤ６の構成を選択し、２つのデバイス又はノード故障に耐えるように、これらのノードにわたってセグメントをストライピングすることができる。その後、ノードが故障した場合、次のセグメントを（７＋２）のＲＡＩＤ６の構成に切り替えることができる。セグメント内では、セグメントシャードの一部がデータを含み、一部が（パリティなどの）ＥＣＣを含む。

【００７４】

１つの実施形態では、５種類のセグメントが存在する。これらのセグメントのうちの３つは、ＡＵ状態テーブル、ＡＵエラーテーブル及び摩耗率テーブルに対応する。いくつかの実施形態では、これらの３つのセグメントをミラーリングしてさらに保護することができる。これらの３つのセグメントに加え、ミラーリングを通じてさらに保護することもできるメタデータセグメントが存在する。最後に、クライアントブロック及びログ情報を保持するデータセグメントが存在する。ログ情報は、セグメント内のクライアントブロックに関連する最新情報を含む。データセグメントは、パリティＰ及びパリティＱシャードを使用して、図１１Ｂに示すようなＲＡＩＤ６によって保護される可能性が高い。上記に加え、起動時に全てのセグメントシャードのヘッダからの情報を投入されるメモリ内データ構造としてセグメントテーブルが維持される。いくつかの実施形態では、このテーブルを全てのノード上で完全にキャッシュして、あらゆるノードがストレージアクセスを物理アドレスに変換できるようにすることができる。しかしながら、他の実施形態では、論理基準を取ってデータが記憶されているセグメントレイアウトノードを識別できるセグメントテーブルを各ノードが有することができるオブジェクトストレージモデルを使用することができる。その後、ノード上の正確な記憶位置を識別するための要求がノードに伝えられる。図１１Ｂには、セグメント内のかなりの量の空間を占めるあらゆる（ボリューム、スナップショットの）組み合わせを識別するセグメントテールデータも示す。スナップショットを除去する場合、データスクラバが、このデータに基づいて、ガベージコレクションを行うセグメントの識別を支援することができる。

【００７５】

１つの実施形態では、基本書き込み単位が、セグメント内のデバイスの各々における１つのＩ／Ｏシャードであるｓｅｇｉｏである。ｓｅｇｉｏ内の各論理ページには、ページのチェックサム（「メディア」チェックサムと呼ぶことができる）を含むページヘッダが、データの実際のページサイズが１ページをわずかに下回るようにフォーマットされる。セグメントのパリティシャード内のページについては、データページ内のページチェックサムがパリティページによって保護されるように、ページヘッダが小さくなっている。各Ｉ／Ｏシャードの最終ページは、この場合も小さなヘッダを有して、消去ブロック内の全てのチェックサム及びページデータをページ不具合から保護するパリティページである。ここで言うページサイズとは、１又はそれ以上の物理フラッシュページとすることができるＩ／Ｏ読み込みサイズのことである。セグメントによっては、物理ページよりも小さな読み込みサイズを使用することができる。この方法は、検索情報に対する読み込みをインデックスで駆動することができ、より小さなデータ部分を読み込みながら所望のデータを取得できるメタデータに対して行うことができる。このような場合、物理ページの半分を読み込むということは、より少ないデータをＩ／Ｏバス（及びネットワーク）に結び付けて、より少ないデータを検証する（例えば、チェックサムを行う）ことを意味する。ある実施形態は、物理ページよりも小さい読み込みサイズをサポートするために、消去ブロックの最後に複数のパリティページを含んで、全てのパリティページの全体的サイズがフラッシュページサイズに等しくなるようにすることができる。

【００７６】

消去ブロックの摩耗率が増すにつれ、エラーの可能性は高くなる。摩耗率を追跡することに加え、高いエラー確率が識別される１又は複数の消去ブロック上に、エラーが観察される頻度に関するデータを維持することもできる。消去ブロックによっては、単一のＲＡ

ＩＤ５パリティの代わりに、消去ブロックの最後にダブル又トリプルエラー訂正パリティを保持することを決定できるものもある。この場合、これに応じてｓｅｇｉｏのデータペイロードを低減することができる。全ての消去ブロックではなく、ｓｅｇｉｏ内の不良な消去ブロックのみを低減することが必要となり得る。消去ブロック内のページヘッダを使用して、いずれのページがパリティであり、いずれのページがデータであるかを識別することができる。

【００７７】

ストレージからページが読み込まれる時には常に、ページチェックサムを使用してコンテンツを検証することができる。この検証がうまくいかなかった場合、消去ブロックパリティを使用してデータの再生を試みることができる。これがうまくいかなかった場合、セグメントのデバイス間ＥＣＣを使用してデータを再構成することができる。

【００７８】

データセグメントでは、ペイロード領域を２つの領域に分割することができる。記憶されたクライアントブロックに関する最新情報を含むことができるログデータとしてフォーマットされたページが存在するようになる。ペイロード領域の残りは、クライアントブロックとしてフォーマットされたページを含むことができる。このクライアントブロックデータは、圧縮された形で記憶することができる。数多くの圧縮アルゴリズムが可能であり、企図される。また、様々な実施形態では、Ｉｎｔｅｌ（登録商標）高度暗号化標準命令を使用してチェックサムを生成することができる。また、データと同じページ内に存在して、データの圧縮に使用するアルゴリズムの識別などの、クライアントブロックを読み込むために必要な情報を含むクライアントブロックのヘッダも存在することができる。ガベージコレクションは、ｓｅｇｉｏ内のクライアントブロックヘッダ及びログエントリの両方を利用することができる。また、クライアントブロックは、重複除外、及び解凍データの正しさをチェックするために使用される未圧縮データのチェックサムとすることができるデータハッシュを有することができる。

【００７９】

いくつかの実施形態では、セグメント及びｓｅｇｉｏが、これらを順序付けるために使用される単調増加するＩＤ番号を有することができる。ｓｅｇｉｏへの書き込みの一部として、論理層が、以前のフラッシュに対する依存を記録することができる。起動時には、物理層が、セグメント及びｓｅｇｉｏｓの順序付きリストを構築することができ、ｓｅｇｉｏが別の未完成のｓｅｇｉｏに依存している場合、これをロールバックして書き込まれていないと見なすことができる。

【００８０】

摩耗率テーブル

各デバイスのローカルなセグメントには、各デバイスの摩耗率テーブル（ＷＬＴ）を記憶することができる。この情報を、各セグメントシャードのヘッダに記憶することもできる。１つの実施形態では、摩耗情報が、アロケーションユニットが消去及び再利用された回数を表す整数である。摩耗情報は正確でない場合もあるので、一定量の行動が行われた場合、又は相当期間にわたってシステムがアイドルであった場合に、デバイスに対するテーブルのフラッシュを行なうことができる。ＷＬＴは、新たなＷＬＴセグメントを割り当てる際に古いＷＬＴセグメントをクリーニングすることに関与することもできる。追加の保護層を加えるために、古いコピーを解放する前にこれらを維持することができる。例えば、テーブルマネージャが、ＷＬＴエントリの以前の消去ブロック及び現在の消去ブロックを常に保持し、新たなセグメントを割り当てる場合には、この新たなセグメントの第２の消去ブロックに書き込みを行うまで古いセグメントを解放しないように保証する。

【００８１】

ＡＵ状態テーブル

ＡＵ状態テーブル（ＡＳＴ）は、各ＡＵの状態を追跡する。この状態には、空き、割り当て済み、消去済み、及び不良がある。このＡＳＴを、デバイス上のセグメントに記憶することができる。割り当て済み又は空きへの状態変更は同期更新とすることができ、不良

又は消去済みへの状態変更は非同期更新とすることができる。一般に、このテーブルは十分に小さく、ＮＶＲＡＭに記録できるだけの十分な更新を有することができる。ＡＳＴは、新たなセグメントを割り当てる際に古いＷＬＴセグメントをクリーニングすることに関与することができる。ＡＳＴは、ドライブ上の各ＡＵの第１のブロックをスキャンすることによって完全に回復できるので、古いＡＳＴのコピーを保持する必要はない。

【００８２】

ＡＵエラーテーブル

ＡＵエラーテーブル（ＡＥＴ）を使用して、各ＡＵ内の回復可能エラー及び回復不能エラーの数を追跡することができる。ＡＥＴは、デバイス上のセグメントに記憶され、各フィールドは、２バイトの整数とすることができる。このテーブル全体は、ＡＵ当たり４バイトの比較的小さなものとすることができる。

【００８３】

ここで図１１Ｃを参照すると、異なるページタイプ内のデータストレージ構成の１つの実施形態を示す汎用ブロック図を示している。図示の実施形態では３つのページタイプを示しているが、他のタイプも可能であり、企図される。図示のページタイプは、メタデータ１１５０を含むページ１１１０、ユーザデータ１１６０を含むページ１１２０、及びパリティ情報１１７０（デバイス間又はデバイス内）を含むページ１１３０を含む。ページ１１１０～１１３０の各々は、ヘッダ及び識別情報を含むことができるメタデータ１１４０を含む。また、１１１０～１１３０ページの各々は、対応するチェックサム又はその他のエラー検出及び／又は訂正符号などのページ内エラー回復データ１１４２を含むことができる。このチェックサム値により、所与のデバイスグループ内のストレージデバイス１７６ａ～１７６ｋに記憶されたデータをさらに保護することができる。

【００８４】

さらに、ページ１１３０は、ページ間エラー回復データ１１４４を含むことができる。このデータ１１４４は、他のストレージデバイスに記憶されたページ内データ１１４２から導出されたＥＣＣ情報とすることができる。例えば、再び図１０を参照すると、デバイス間パリティ情報２４０を記憶するストレージデバイス１７６ｊ内の各ページが、ページ間エラー回復データ１１４４を記憶することもできる。データ１１４４は、パリティ、チェックサム、又はストレージデバイス１７６ａ～１７６ｉの１又はそれ以上に記憶されたページ内エラー回復データ１１４２から生成されたその他の値とすることもできる。１つの実施形態では、データ１１４４が、他のストレージデバイスに記憶された１又はそれ以上の他のチェックサム値１１４２から生成されたチェックサム値である。ストレージデバイス１７６ｊの所与のページ内のデータ１１４４を、ストレージデバイス１７６ａ～１７６ｉの１又はそれ以上の対応するページ内のデータ１１４２と位置合わせするために、対応するページにパディング１１４６を追加することができる。

【００８５】

１つの実施形態では、エンドユーザアプリケーションが、ＨＤＤでは５１２バイトのセクタの境界上でＩ／Ｏ動作を実行する。追加の保護を加えるために、８バイトのチェックサムを加えて５２０バイトのセクタを形成することができる。様々な実施形態では、フラッシュメモリベースのシステム内で圧縮及びリマッピングを行って、セクタ境界ではなくバイト境界上にユーザデータを配置可能にすることができる。また、ヘッダの後であってユーザデータの前のページ内にチェックサム（８バイト又は４バイトなど）を配置して、これを圧縮することができる。ページ１１１０～１１３０の各々には、この配置を示している。

【００８６】

エンドユーザアプリケーションが５１２バイトのセクタを読み込む場合、１つの実施形態では２ＫＢ～８ＫＢのサイズの対応するページが、ページの最初に８バイトのチェックサムによる追加の保護を有する。様々な実施形態では、２セクタサイズの非電力のためにこのページをフォーマットしなくてもよい。ページ１１１０～１１２０に示すように、チェックサムをページ内に２、３バイトオフセットさせることができる。このオフセットに

より、ページ１１３０などのパリティページが、パリティページをカバーするチェックサム、及び他のページのチェックサムを保護するためのＥＣＣを記憶できるようになる。

【００８７】

さらに別の保護レベルでは、チェックサム値を計算する際にデータ位置情報を含めることができる。ページ１１１０～１１３０の各々におけるデータ１１４２は、この情報を含むことができる。この情報は、論理アドレス及び物理アドレスの両方を含むことができる。この情報には、セクタ番号、データチャンク及びオフセット番号、トラック番号、平面番号などを含めることもできる。

【００８８】

代替の構成

ここで図１２を参照すると、ハイブリッドＲＡＩＤデータレイアウト１２００の１つの実施形態を示す汎用ブロック図を示している。３つの区分を示しているが、あらゆる数の区分を選択することができる。各区分は、図１に示すデバイスグループ７１３ａ～１７３ｂなどの別個のデバイスグループに対応することができる。各区分は、複数のストレージデバイスを含む。１つの実施形態では、ＣＲＵＳＨアルゴリズムなどのアルゴリズムを利用して、データストレージに使用すべきＲＡＩＤデータレイアウトアーキテクチャにおいていずれのデバイスを使用すべきかを選択することができる。

【００８９】

図示の例には、Ｌ＋１のＲＡＩＤアレイ、Ｍ＋１のＲＡＩＤアレイ及びＮ＋１のＲＡＩＤアレイを示している。様々な実施形態では、Ｌ、Ｍ及びＮが全て異なってもよく、同じであってもよく、又はこれらの組み合わせであってもよい。例えば、区分１にＲＡＩＤアレイ１２１０を示している。他のストレージデバイス１２１２は、区分１内の他のＲＡＩＤアレイの候補である。同様に、ＲＡＩＤアレイ１２２０は、区分２内の所与のＲＡＩＤアレイを示す。他のストレージデバイス１２２２は、区分２内の他のＲＡＩＤアレイの候補である。ＲＡＩＤアレイ１２３０は、区分３内の所与のＲＡＩＤアレイを示す。他のストレージデバイス１２３２は、区分３内の他のＲＡＩＤアレイの候補である。

【００９０】

ＲＡＩＤアレイ１２１０、１２２０及び１２３０の各々では、ストレージデバイスＰ１が、それぞれのＲＡＩＤアレイ内のＲＡＩＤシングルパリティ保護を行う。ストレージデバイスＤ１～ＤＮは、それぞれのＲＡＩＤアレイ内のユーザデータを記憶する。この場合も、ストレージデバイスＤ１～ＤＮとＰ１の間で、ユーザデータ及びＲＡＩＤシングルパリティ情報の記憶を循環させることができる。しかしながら、デバイスＤ１～ＤＮにユーザデータが記憶されたものとして説明する。同様に、図示及び説明を容易にするために、デバイスＰ１にＲＡＩＤシングルパリティ情報が記憶されたものとして説明する。

【００９１】

１又はそれ以上の所与のＲＡＩＤアレイに対してさらなる量の冗長性をサポートするために、３つの区分の各々のうちの１又はそれ以上のストレージデバイスを選択することができる。例えば、区分３内のストレージデバイスＱ１を、ＲＡＩＤアレイ１２１０、１２２０及び１２３０の各々と組み合わせることができる。ストレージデバイスＱ１は、ＲＡＩＤアレイ１２１０、１２２０及び１２３０の各々にＲＡＩＤダブルパリティ情報を提供することができる。この追加のパリティ情報は、アレイ１２１０、１２２０又は１２３０の１つにストライプが書き込まれた時に生成され記憶される。さらに、この追加のパリティ情報は、アレイ１２１０、１２２０及び１２３０の各々のストライプをカバーすることができる。従って、総ストレージデバイス数に対するＲＡＩＤパリティ情報を記憶するストレージデバイスの数の比率が小さくなる。例えば、区分の各々がＮ＋２のＲＡＩＤアレイを使用する場合、総ストレージデバイス数に対するＲＡＩＤパリティ情報を記憶するストレージデバイスの数の比率は、３（２）／（３（Ｎ＋２））、すなわち２／Ｎ＋２である。対照的に、ハイブリッドＲＡＩＤレイアウト１２００の比率は、（３＋１）／（３（Ｎ＋１））、すなわち４／（３（Ｎ＋１））である。

【００９２】

ユーザデータの記憶に使用するストレージデバイスの数を増やすことにより、上記の比率を低減することが可能である。例えば、ストレージデバイスＱ１を利用するのではなく、区分の各々が３Ｎ＋２のＲＡＩＤアレイを利用するようにすることができる。このような場合、総ストレージデバイス数に対するＲＡＩＤパリティ情報を記憶するストレージデバイスの数の比率は、２／（３Ｎ＋２）である。しかしながら、再構成読み込み動作中には、単一のデバイス故障のための再構成読み込み要求を（３Ｎ＋１）個のストレージデバイスが受け取る。対照的に、ハイブリッドＲＡＩＤレイアウト１２００では、単一のデバイス故障のための再構成読み込み要求をＮ個のストレージデバイスしか受け取らない。

【００９３】

なお、３つの区分の各々は、異なるＲＡＩＤデータレイアウトアーキテクチャを利用することができる。所与のＲＡＩＤデータレイアウトアーキテクチャの選択は、総ストレージデバイス数に対するＲＡＩＤパリティ情報を記憶するストレージデバイスの数の所与の比率に基づくことができる。また、この選択は、再構成中に再構成読み込み要求を受け取ることができる所与のストレージデバイス数に基づくこともできる。例えば、ＲＡＩＤアレイ１２１０、１２２０及び１２３０は、Ｌ＋ａ、Ｍ＋ｂ及びＮ＋ｃなどの構成をそれぞれ含むことができる。

【００９４】

区分内のＲＡＩＤアレイの１又はそれ以上に対してさらなる量の冗長性をサポートするために、上記の又はその他の条件に基づいてストレージデバイスＱ１などの１又はそれ以上のストレージデバイスを選択することができる。上記のＲＡＩＤアレイ及びこれらのＲＡＩＤアレイの各々に対して追加の保護を行うストレージデバイスの数Ｑを含む３つの区分の例では、総ストレージデバイス数に対するＲＡＩＤパリティ情報を記憶するストレージデバイスの数の比率は、（ａ＋ｂ＋ｃ＋Ｑ）／（Ｌ＋ａ＋Ｍ＋ｂ＋Ｎ＋ｃ＋Ｑ）である。単一のデバイス故障の場合、上記の例における区分１～３では、再構成読み込み要求を受け取るストレージデバイスの数は、それぞれＬ、Ｍ及びＮである。なお、一般に上記の説明は、図１２の３つの異なる区分について行ったものである。このような実施形態では、所与のレイアウトが特定のデバイスグループに制限されるこの種の「ハード」区分が、１つの区分内の再構成読み込みが別の区分の再構成読み込みと衝突しないことを保証することができる。しかしながら、他の実施形態では、区分が、上述したようなハードではないこともある。むしろ、デバイスのプールを考えた場合、これらのデバイスのいずれからもレイアウトを選択することができる。例えば、デバイスを大きなプールとして扱う場合、（Ｌ＋１、Ｍ＋１、Ｎ＋１）＋１などのレイアウトを構成することが可能である。従って、構成が重複し、再構成読み込みが衝突しかねない可能性がある。プールのサイズに対してＬ、Ｍ及びＮが小さい場合、通常の読み込みに対する再構成読み込みの割合を低く保つことができる。

【００９５】

ここで図１３を参照すると、データストレージサブシステムにおいて代替のＲＡＩＤ構成を選択する方法１３００の１つの実施形態を示している。一般に、ネットワークアーキテクチャ１００及びデータストレージアレイ１２０ａ～１２０ｂ内で具体化される構成要素は、方法１３００に従って動作することができる。この実施形態のステップを順番に示す。しかしながら、ステップによっては、図示のものとは異なる順序で行なうことができるもの、同時に行うことができるもの、他のステップと組み合わせることができるもの、及び別の実施形態には存在しないものもある。

【００９６】

ブロック１３０２において、ストレージコントローラ１７４内のＲＡＩＤエンジン１７８又はその他のロジックが、所与の数のデバイスを使用して、ストレージサブシステムの各区分内のＲＡＩＤアレイにユーザデータを記憶すると決定する。その後、ＲＵＳＨ又はその他のアルゴリズムを使用して、いずれのデバイスを使用すべきかを選択することができる。１つの実施形態では、各区分が、同じ数のストレージデバイスを利用する。他の実施形態では、各区分が、異なる固有の数のストレージデバイスを利用してユーザデータを

記憶することができる。ブロック１３０４において、ストレージコントローラ１７４が、いくつかのストレージデバイスがサブシステムの各区分内に対応するデバイス間エラー回復（パリティ）データを記憶するのをサポートすると決定することができる。この場合も、各区分は、同じ数又は異なる固有の数のストレージデバイスを利用してＲＡＩＤパリティ情報を記憶することができる。

【００９７】

ブロック１３０６において、ストレージコントローラは、Ｑ個のストレージデバイスが追加のデバイス間エラー回復（パリティ）データをサブシステムの区分にわたって記憶するのをサポートすると決定することができる。ブロック１３０８において、選択したストレージデバイスに、ユーザデータ及び対応するＲＡＩＤパリティデータを書き込むことができる。再び図１２を参照すると、区分１のＲＡＩＤアレイ１２１０などの所与のＲＡＩＤアレイに書き込みが行われた時に、１又はそれ以上のビットのパリティ情報を生成して、区分３のストレージデバイスＱ１に記憶することができる。

【００９８】

ストレージコントローラ１７４が、所与の区分内で読み込み再構成を行う状態を検出した（条件付きブロック１３１０）場合、及び所与の区分が、ＲＡＩＤパリティ情報を保持しているストレージデバイスを、いくつかの利用できないストレージデバイスを処理するのに十分な数だけ有している（条件付きブロック１３１２）場合、ブロック１３１４において、この所与の区分内の１又はそれ以上の対応するストレージデバイスを使用して（単複の）再構成読み込み動作を行う。この状態は、デバイス故障によって所与のＲＡＩＤアレイ内のストレージデバイスを利用できないこと、又はこれらのデバイスが所与のパフォーマンスレベル未満で動作していることを含むことができる。この所与のＲＡＩＤアレイは、所与の区分内のＲＡＩＤパリティ情報を記憶している数のストレージデバイスを使用して、最大数の利用できないストレージデバイスを処理することができる。例えば、上記の例の区分１のＲＡＩＤアレイ１２１０がＬ＋ａのＲＡＩＤアレイである場合、このＲＡＩＤアレイ１２１０は、１＜＝ｋ＜＝ａとするｋ個のストレージデバイスが利用できない場合に、区分１内のストレージデバイスのみを利用して読み込み再構成を行うことができる。

【００９９】

所与の区分が、ＲＡＩＤパリティ情報を保持しているストレージデバイスを、いくつかの利用できないストレージデバイスを処理するのに十分な数だけ有していない（条件付きブロック１３１２）場合、及びこのいくつかの利用できないストレージデバイスを処理するのに十分なＱ個のストレージデバイスが存在する（条件付きブロック１３１６）場合、ブロック１３１８において、１又はそれ以上の対応するＱ個のストレージデバイスを使用して再構成読み込み動作を行う。読み込み再構成中には、ユーザデータを記憶している他の区分内の１又はそれ以上のストレージデバイスにアクセスすることができる。これらのストレージデバイスの選択は、１又はそれ以上のＱ個のストレージデバイスに記憶されたパリティ情報の導出方法に基づくことができる。例えば、再び図１２を参照すると、ストレージデバイスＱ１に記憶された対応するＲＡＩＤパリティ情報を生成するために使用された可能性があるという理由で、読み込み再構成中に区分２のストレージデバイスＤ２にアクセスすることができる。いくつかの利用できないストレージデバイスを処理するのに十分な数のＱ個のストレージデバイスが存在しない場合（条件付きブロック１３１６）、ブロック１３２０において、対応するユーザデータを別のソースから読み込むことができ、又はこれらのユーザデータを失われたものと見なすことができる。

【０１００】

なお、上述の実施形態は、ソフトウェアを含むことができる。このような実施形態では、方法及び／又は機構を実装するプログラム命令をコンピュータ可読媒体で搬送し、又はこれに記憶することができる。プログラム命令を記憶するように構成された数多くのタイプの媒体が利用可能であり、これらは、ハードディスク、フロッピー（登録商標）ディスク、ＣＤ－ＲＯＭ、ＤＶＤ、フラッシュメモリ、プログラマブルＲＯＭ（ＰＲＯＭ）、ラ

ンダムアクセスメモリ（ＲＡＭ）及び他の様々な形態の揮発性又は不揮発性ストレージを含む。

【０１０１】

様々な実施形態では、本明細書で説明した方法及び機構の１又はそれ以上の部分が、クラウドコンピューティング環境の一部を成すことができる。このような実施形態では、１又はそれ以上の様々なモデルにより、インターネットを介してリソースをサービスとして提供することができる。このようなモデルとして、インフラストラクチャ・アズ・ア・サービス（ＩａａＳ）、プラットホーム・アズ・ア・サービス（ＰａａＳ）、及びソフトウェア・アズ・ア・サービス（ＳａａＳ）を挙げることができる。ＩａａＳでは、コンピュータインフラストラクチャがサービスとして配信される。このような場合、一般にサービスプロバイダがコンピュータ設備を所有し運営する。ＰａａＳモデルでは、開発者がソフトウェアソリューションを開発するために使用するソフトウェアツール及び基本設備をサービスプロバイダがサービスとして供給しホストすることができる。通常、ＳａａＳは、サービスプロバイダのライセンスソフトウェアをサービスオンデマンドとして含む。サービスプロバイダは、このソフトウェアをホストすることができ、又はこのソフトウェアを一定期間にわたって顧客に展開することができる。上記のモデルの数多くの組み合わせが可能であり、企図される。

【０１０２】

以上、実施形態についてかなり詳細に説明したが、上記開示を完全に理解すると、当業者には数多くの変形及び修正が明らかになるであろう。以下の特許請求の範囲は、このような変形及び修正を全て含むと解釈すべきであることが意図される。

【図１】

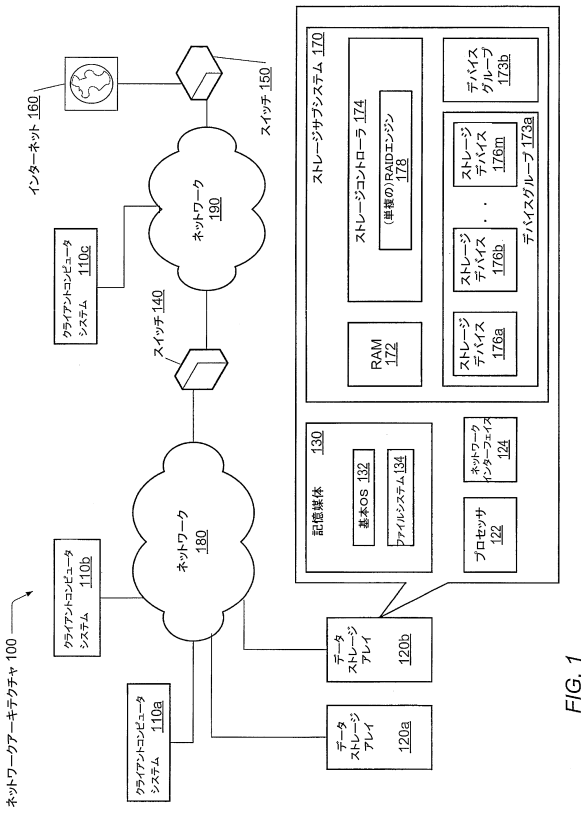

FIG. 1

【図２】

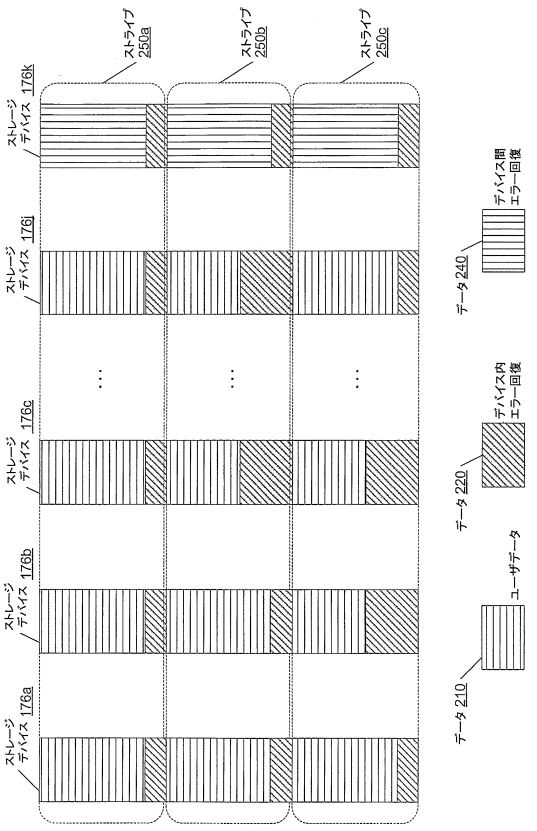

FIG. 2

【図３】

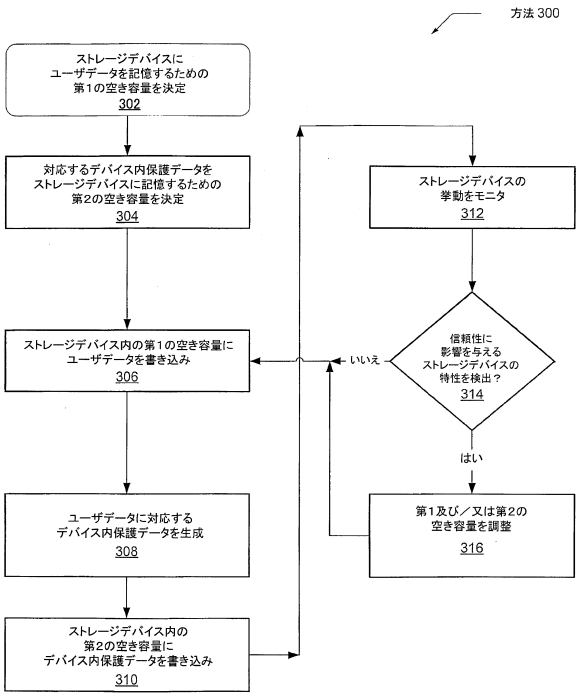

FIG. 3

【図４】

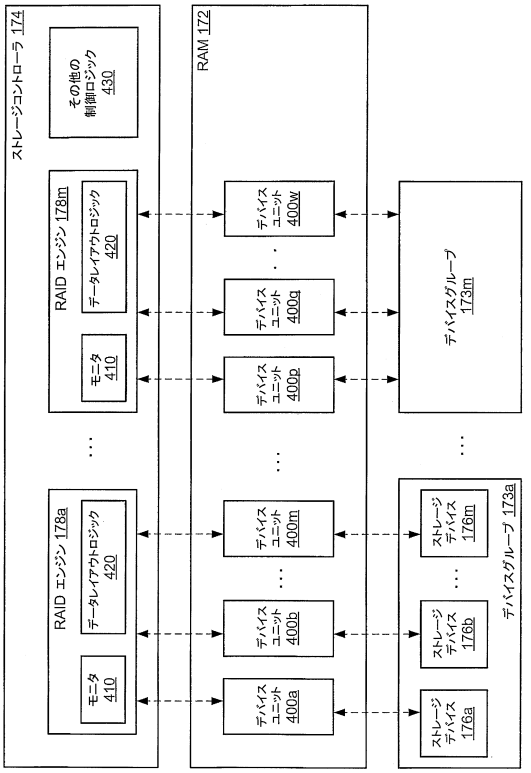

FIG. 4

【図５】

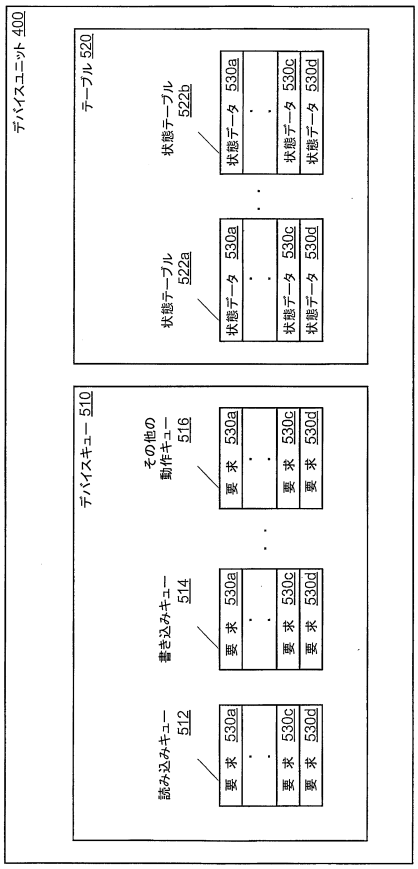

FIG. 5

【図６】

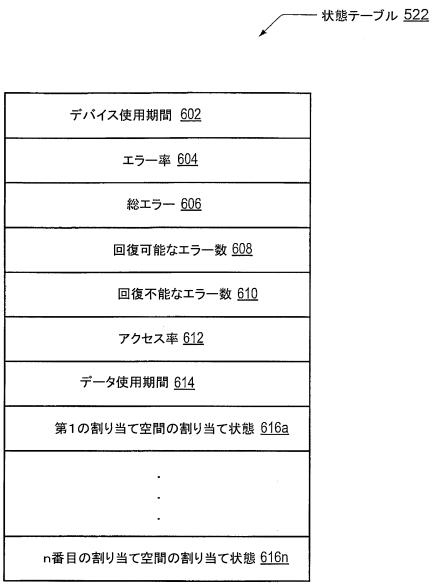

FIG. 6

【図７】

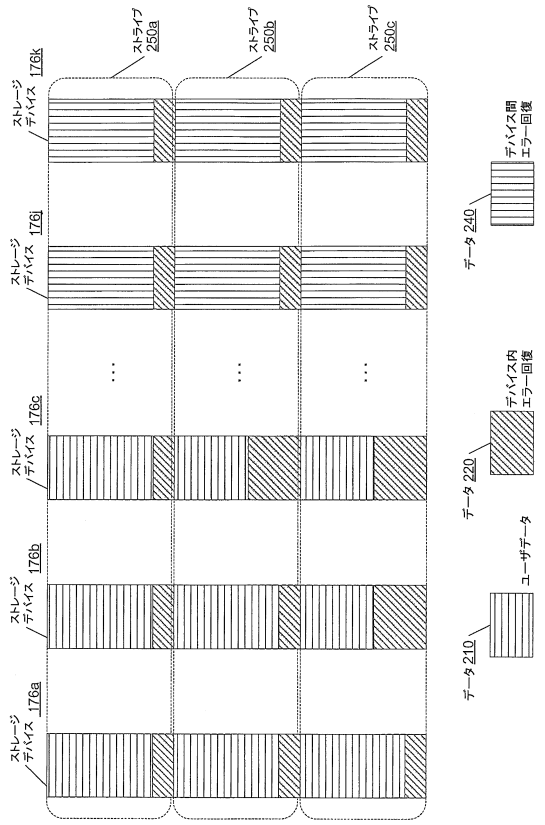

FIG. 7

【図８】

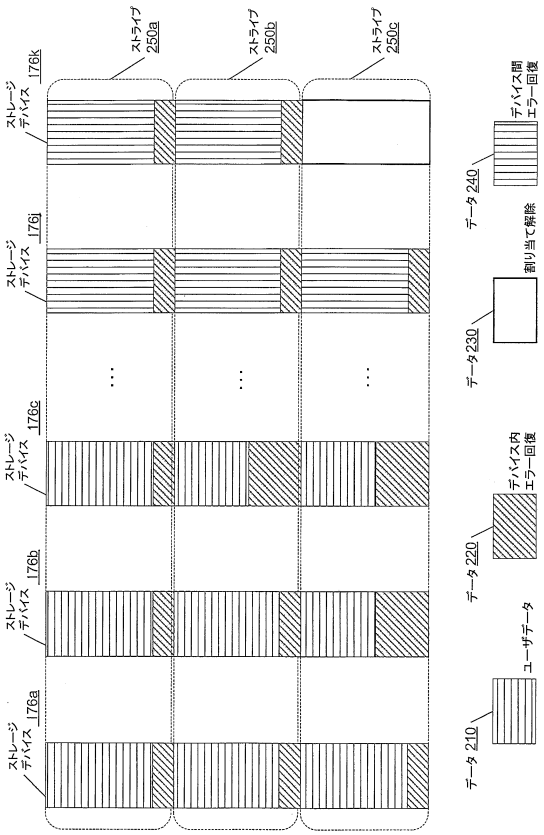

FIG. 8

【図９】

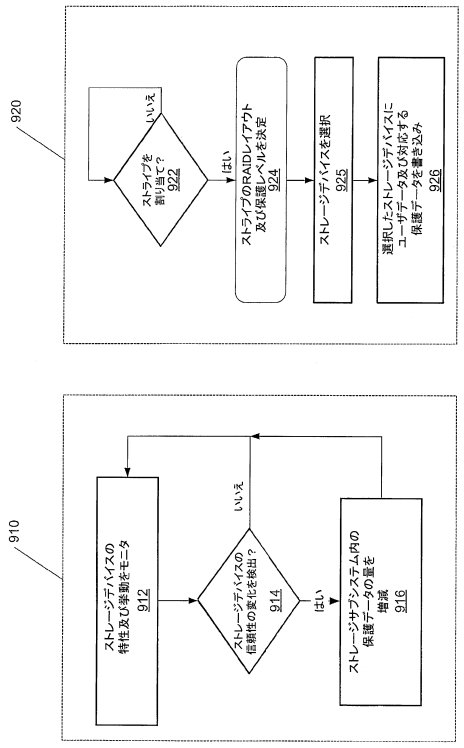

FIG. 9

【図１０】

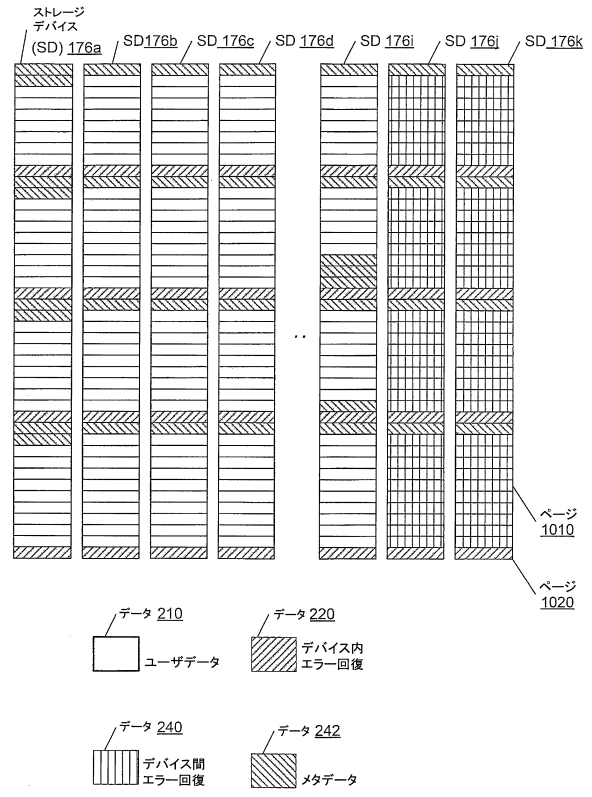

FIG. 10

【図１１Ａ】

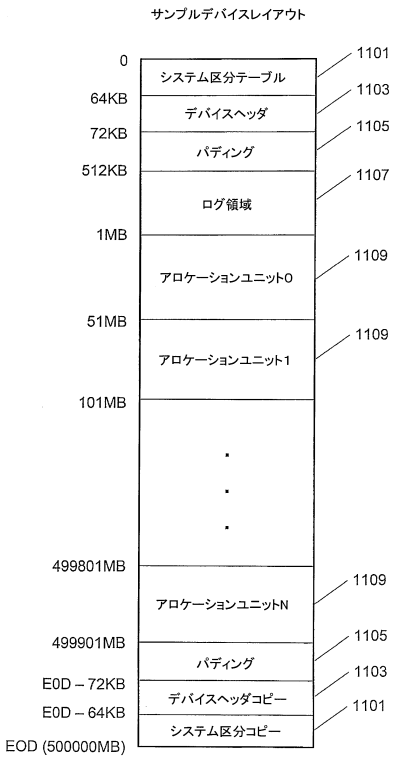

FIG. 11A

【図１１Ｂ】

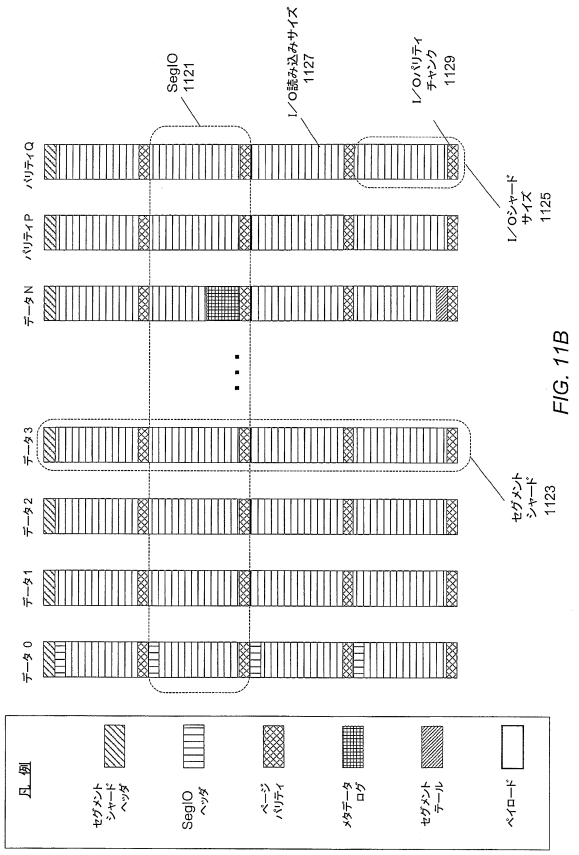

FIG. 11B

【図１１Ｃ】

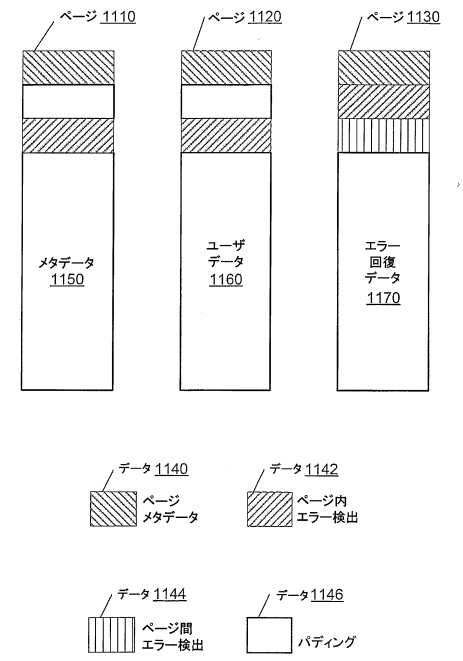

FIG. 11C

【図１２】

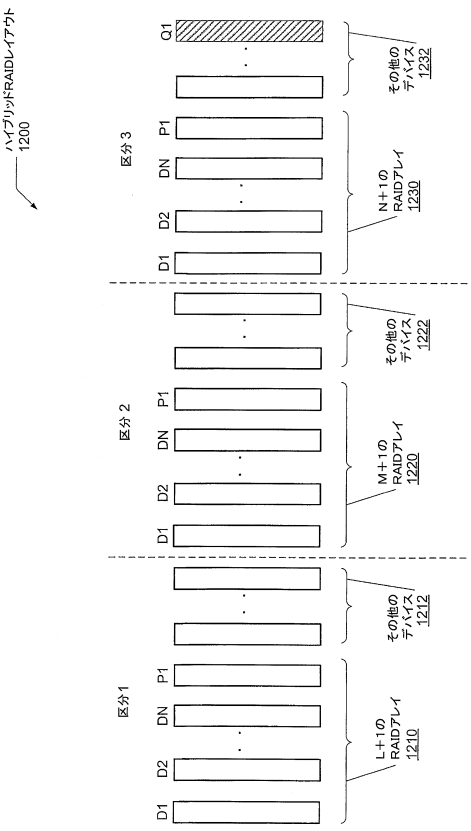

FIG. 12

【図１３】

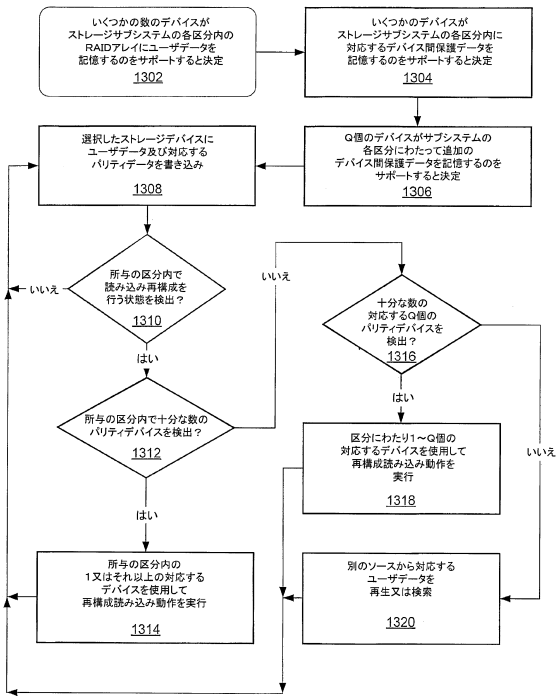
いくつかの数のデバイスが
ストレージサブシステムの各区分内の
RAIDアレイにユーザデータを
記憶するのをサポートすると決定
1302
いくつかのデバイスが
ストレージサブシステムの各区分内に
対応するデバイス間保護データを
記憶するのをサポートすると決定
1304
Q個のデバイスがサブシステムの
各区分にわたって追加の
デバイス間保護データを記憶するのを
サポートすると決定
1306
選択したストレージデバイスに
ユーザデータ及び対応する
パリティデータを書き込み
1308
所与の区分内で
読み込み再構成を
行う状態を検出？
1310
いいえ
はい
所与の区分内で十分な数の
パリティデバイスを検出？
1312
いいえ
はい
十分な数の
対応するQ個の
パリティデバイスを
検出？
1316
はい
いいえ
区分にわたり1～Q個の
対応するデバイスを使用して
再構成読み込み動作を
実行
1318
所与の区分内の
1又はそれ以上の対応する
デバイスを使用して
再構成読み込み動作を実行
1314
別のソースから対応する
ユーザデータを
再生又は検索
1320

FIG. 13

フロントページの続き

(72)発明者　ヘイズ，ジョン
アメリカ合衆国・９４０４１・カリフォルニア州・マウンテン　ビュー・ハイスクール　ウェイ・８００・ナンバー・３３０
(72)発明者　ホン，ボー
アメリカ合衆国・９４０４３・カリフォルニア州・マウンテン　ビュー・ウェスト　ミドルフィールド　ロード・１５５５・アパートメント・９５
(72)発明者　ミラー，イーサン
アメリカ合衆国・９５０６０・カリフォルニア州・サンタクルーズ・カルカー　ドライブ・２０３

審査官　田中　啓介

(56)参考文献　米国特許出願公開第２００９／０２１０７４２（ＵＳ，Ａ１）
特開平０７－２００１９１（ＪＰ，Ａ）
国際公開第２０１０／１０６５７４（WO，Ａ１）
米国特許第０５６５７４３９（ＵＳ，Ａ）
特開２００９－２１７４０８（ＪＰ，Ａ）
特開平０８－２２１８７５（ＪＰ，Ａ）
特開平０７－２６１９４５（ＪＰ，Ａ）
特開平０８－１９４５８７（ＪＰ，Ａ）
米国特許出願公開第２００８／０２２９０１２（ＵＳ，Ａ１）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ３／０６－３／０８
１１／０８－１１／１０