(19)日本国特許庁(JP)　(12)公表特許公報(A)　(11)特許出願公表番号
特表2010-512568
(P2010-512568A)
(43)公表日　平成22年4月22日(2010.4.22)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 3/06	(2006.01)	G06F	3/06	302A	5B005
G06F 12/08	(2006.01)	G06F	12/08	557	5B065
		G06F	12/08	541Z	
		G06F	3/06	540	

審査請求 未請求　予備審査請求 有　（全 112 頁）

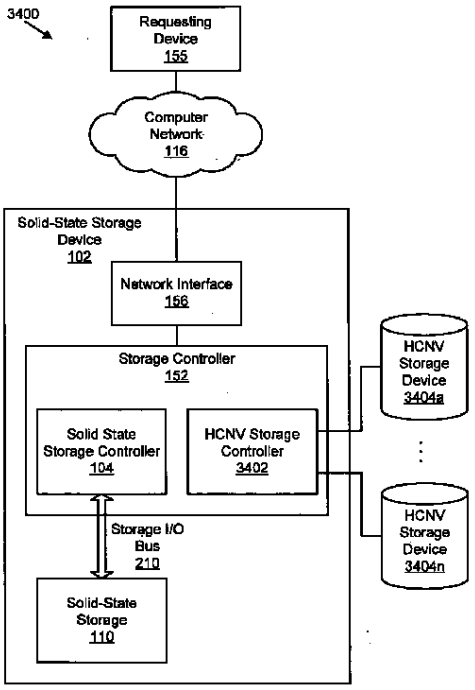
(21)出願番号　特願2009-540309 (P2009-540309)
(86)(22)出願日　平成19年12月6日 (2007.12.6)
(85)翻訳文提出日　平成21年7月31日 (2009.7.31)
(86)国際出願番号　PCT/US2007/025049
(87)国際公開番号　WO2008/070173
(87)国際公開日　平成20年6月12日 (2008.6.12)
(31)優先権主張番号　60/873,111
(32)優先日　平成18年12月6日 (2006.12.6)
(33)優先権主張国　米国 (US)
(31)優先権主張番号　60/974,470
(32)優先日　平成19年9月22日 (2007.9.22)
(33)優先権主張国　米国 (US)

(71)出願人　509157362
フリン，デイビッド
アメリカ合衆国　ユタ州　８４０９３，サンディ，シェイディメドウドライブ　８８５６
(71)出願人　509157351
シュトラッサー，ジョン
アメリカ合衆国　ユタ州　８４０７５，シラキュース，サウス　２３２３
(71)出願人　509157339
サッチャー，ジョナサン
アメリカ合衆国　ユタ州　８４０４３，リーハイ，ノース　２０８０　ウエスト　２２５９

最終頁に続く

(54)【発明の名称】高容量不揮発性ストレージ用のキャッシュとしてのソリッドステートストレージのための装置、システム、及び方法

(57)【要約】
高容量不揮発性ストレージ用のキャッシュとしてのソリッドステートストレージのための装置、システム及び方法が開示されている。その装置、システム及び方法はキャッシュフロントエンドモジュールとキャッシュバックエンドモジュールを具える複数のモジュールで提供される。キャッシュフロントエンドモジュールはストレージ要求に関連するデータ転送を管理する。要求デバイスとソリッドステートストレージとの間のデータ転送は、１又はそれ以上のＨＣＮＶストレージデバイス用のキャッシュとして機能し、データ転送はデータ、メタデータ及びメタデータインデックスのうちの１又はそれ以上を含むことができる。ソリッドステートストレージは不揮発性のソリッドステートデータストレージエレメントのアレイを具えることができる。キャッシュバックエンドモジュールはソリッドステートストレージと、１又はそれ以上のＨＣＮＶストレージデバイスとの間のデータ転送を管理する。
【選択図】　図１５



【特許請求の範囲】
【請求項１】
１又はそれ以上の高容量不揮発性（「ＨＣＮＶ」）ストレージデバイスでデータのストレージを管理するための装置であって、当該装置が、
ストレージ要求に関連するデータ転送を管理し、要求デバイスとソリッドステートストレージとの間の前記データ転送が、１又はそれ以上のＨＣＮＶストレージデバイス用のキャッシュとして機能し、前記データ転送がデータ、メタデータ及びメタデータインデックスのうちの１又はそれ以上を含み、前記ソリッドステートストレージが不揮発性のソリッドステートデータストレージエレメントのアレイを具えるキャッシュフロントエンドモジュールと；
前記ソリッドステートストレージと、前記１又はそれ以上のＨＣＮＶストレージデバイスとの間のデータ転送を管理するキャッシュバックエンドモジュールと；
を具えることを特徴とする装置。
【請求項２】
請求項１に記載の装置において、前記キャッシュフロントエンドモジュール及び前記キャッシュバックエンドモジュールが、前記ソリッドステートストレージを管理するソリッドステートストレージコントローラと共に同一位置に配置されることを特徴とする装置。
【請求項３】
請求項２に記載の装置において、前記キャッシュフロントエンドモジュール、前記キャッシュバックエンドモジュール、及びソリッドステートストレージコントローラが、前記要求デバイスから自律して動作することを特徴とする装置。
【請求項４】
請求項１に記載の装置において、前記ソリッドステートコントローラが、１又はそれ以上の要求デバイスからオブジェクト要求を提供し、前記ソリッドステートストレージ内の前記オブジェクト要求のオブジェクトを管理するオブジェクトストレージコントローラモジュールを更に具えることを特徴とする装置。
【請求項５】
請求項１に記載の装置が、ＲＡＩＤレベルの一致する独立ドライブ冗長アレイ（「ＲＡＩＤ」）の２又はそれ以上のＨＣＮＶストレージデバイス内の前記ソリッドステートストレージにキャッシュされたデータを記憶するＨＣＮＶ　ＲＡＩＤモジュールを更に具え、前記データが全体として要求デバイスに見えることを特徴とする装置。
【請求項６】
請求項１に記載の装置において、前記ソリッドステートストレージ及び前記１又はそれ以上のＨＣＮＶストレージデバイスが、ＲＡＩＤグループとして構成されるハイブリッドストレージデバイスセット内にハイブリッドストレージデバイスを具え、前記ソリッドステートストレージにキャッシュされ、ＨＣＮＶデバイスに後に記憶されるデータセグメントが、ストライプのＮ個のデータセグメントのうちの１つ、又は、前記ストライプのパリティデータセグメントを含み、前記ハイブリッドストレージデバイスがＲＡＩＤストライプのデータセグメントと別個の、１又はそれ以上のクライアントからストレージ要求を受信することを特徴とする装置。
【請求項７】
請求項６に記載の装置において、前記ハイブリッドストレージデバイスが２又はそれ以上のクライアントから２又はそれ以上の同時のストレージ要求を受信する、共有されたフロントエンド分散型ＲＡＩＤグループのストレージデバイスであることを特徴とする装置。
【請求項８】
請求項１に記載の装置において、前記ＨＣＮＶストレージデバイスがハードディスクドライブ（「ＨＤＤ」）、光学ドライブ、及びテープストレージのうちの１つであることを特徴とする装置。
【請求項９】

請求項１に記載の装置において、前記ソリッドステートストレージ及び前記１又はそれ以上のＨＣＮＶストレージデバイスが、ハイブリッドストレージデバイスを具え、当該ハイブリッドストレージデバイスの動作に特有のコードで、１又はそれ以上の要求デバイスをロードする前に、前記１又はそれ以上の要求デバイスに取り付けられた標準デバイスをエミュレートすることによって、前記ハイブリッドストレージデバイスにアクセスを提供する標準デバイスエミュレーションモジュールを更に具え、前記標準デバイスが工業規格ＢＩＯＳによって担持されることを特徴とする装置。

【請求項１０】

請求項１に記載の装置において、前記ソリッドステートストレージデバイスが２又はそれ以上の領域に分割可能であり、１又はそれ以上のパーティションが、前記ＨＣＮＶストレージデバイス用のキャッシュとして機能する前記ソリッドステートストレージと別個の、ソリッドステートストレージとして用いられ得ることを特徴とする装置。

【請求項１１】

請求項１に記載の装置において、１又はそれ以上のクライアントが、前記ソリッドステートストレージデバイス及び前記１又はそれ以上のＨＣＮＶストレージデバイス内に記憶されるファイル又はオブジェクトのうちの１又はそれ以上の状態を管理するために、前記キャッシュフロントエンドモジュール及び前記キャッシュバックエンドモジュールにキャッシュ制御メッセージを送信することを特徴とする装置。

【請求項１２】

請求項１１に記載の装置において、前記キャッシュ制御メッセージが、

前記キャッシュバックエンドモジュールに、前記ソリッドステートストレージ内のオブジェクト又はファイルの一部を留める（ｐｉｎ）制御メッセージ；

前記キャッシュバックエンドモジュールに、前記ソリッドステートストレージ内のオブジェクト又はファイルの一部を解放する（ｕｎｐｉｎ）制御メッセージ；

前記キャッシュバックエンドモジュールに、前記ソリッドステートストレージから前記１又はそれ以上のＨＣＶＮストレージデバイスまでのオブジェクト又はファイルの一部を消去させる制御メッセージ；

前記キャッシュバックエンドモジュールに、前記１又はそれ以上のＨＣＶＮストレージデバイスから前記ソリッドステートストレージに、オブジェクト又はファイルの一部を事前ロードさせる制御メッセージ；及び

前記ソリッドステートストレージに、決められたストレージ空き領域量を空けるために、前記キャッシュバックエンドモジュールに、前記ソリッドステートストレージから前記１又はそれ以上のＨＣＶＮストレージデバイスまでの１又はそれ以上のオブジェクト又はファイルの１又はそれ以上の部分をオフロードさせる制御メッセージ；

のうちの１又はそれ以上を含むことを特徴とする装置。

【請求項１３】

請求項１１に記載の装置において、前記キャッシュ制御メッセージが前記オブジェクト又はファイル用のメタデータ（「キャッシュ制御メタデータ」）を通して通信されることを特徴とする装置。

【請求項１４】

請求項１３に記載の装置において、前記キャッシュ制御メタデータが永続的であることを特徴とする装置。

【請求項１５】

請求項１３に記載の装置において、前記キャッシュ制御メタデータが前記ファイル又はオブジェクトの生成時に設定される属性を通して構築されることを特徴とする装置。

【請求項１６】

請求項１３に記載の装置において、前記キャッシュ制御メタデータがファイル又はオブジェクト管理システムから得られることを特徴とする装置。

【請求項１７】

請求項１に記載の装置が、揮発性キャッシュストレージエレメントを更に具え、前記キ

ャッシュフロントエンドモジュール及び前記キャッシュバックエンドモジュールが、前記揮発性キャッシュストレージエレメントにデータを記憶するステップを更に具え、前記ソリッドステートストレージ及び揮発性キャッシュストレージエレメントに記憶されるデータを管理し、前記バックエンドストレージモジュールが、前記揮発性キャッシュストレージエレメントと、前記ソリッドステートストレージと、前記ＨＣＶＮストレージデバイスとの間のデータ転送を更に管理することを特徴とする装置。

【請求項１８】

請求項１７に記載の装置において、前記ＨＣＶＮストレージデバイスに記憶されるオブジェクト及びファイル用のメタデータ及びインデックスメタデータのうちの１又はそれ以上が、前記ソリッドステートストレージデバイス及び前記揮発性キャッシュストレージエレメント内に維持されることを特徴とする装置。

【請求項１９】

請求項１に記載の装置において、前記ＨＣＶＮストレージデバイスに記憶されるオブジェクト及びファイル用のメタデータ及びインデックスメタデータのうちの１又はそれ以上が、前記ソリッドステートストレージデバイス内に維持されることを特徴とする装置。

【請求項２０】

請求項１に記載の装置において、前記ソリッドステートストレージ及び前記１又はそれ以上のＨＣＮＶストレージデバイスが、ストレージデバイスを具えて、前記ＨＣＮＶストレージデバイスが前記ストレージデバイスに接続されたクライアントの表示から隠されるようにすることを特徴とする装置。

【請求項２１】

１又はそれ以上の高容量不揮発性（「ＨＣＮＶ」）ストレージデバイスでデータのストレージを管理するためのシステムであって，当該システムが、

不揮発性のソリッドステートデータストレージエレメントのアレイを具えるソリッドステートストレージと；

１又はそれ以上のＨＣＮＶストレージデバイスと；

ソリッドステートストレージコントローラと；

ＨＣＮＶストレージデバイスコントローラと；

ストレージ要求に関連するデータ転送を管理し、要求デバイスと前記ソリッドステートストレージとの間の前記データ転送が、前記１又はそれ以上のＨＣＮＶストレージデバイス用のキャッシュとして機能し、前記データ転送がデータ、メタデータ及びメタデータインデックスのうちの１又はそれ以上を含むキャッシュフロントエンドモジュールと；

前記ソリッドステートストレージと、前記１又はそれ以上のＨＣＮＶストレージデバイスとの間のデータ転送を管理するキャッシュバックエンドモジュールと；

を具えるストレージコントローラと；

を具えることを特徴とするシステム。

【請求項２２】

請求項２１に記載のシステムが、前記ストレージコントローラに接続されたネットワークインタフェースを更に具え、当該ネットワークインタフェースがコンピュータネットワークを通して前記要求デバイスと前記ソリッドステートストレージコントローラとの間のデータ転送を促進することを特徴とするシステム。

【請求項２３】

請求項２１に記載のシステムがサーバを更に具え、当該サーバが前記ソリッドステートストレージと、前記１又はそれ以上のＨＣＮＶストレージデバイスと、前記ストレージコントローラとを具えることを特徴とするシステム。

【請求項２４】

請求項２１に記載のシステムにおいて、前記１又はそれ以上のＨＣＮＶストレージデバイスがストレージエリアネットワーク（「ＳＡＮ」）を通して前記ストレージコントローラに接続されることを特徴とするシステム。

【請求項２５】
コンピュータプログラムプロダクトであって、１又はそれ以上の高容量不揮発性（「ＨＣＮＶ」）ストレージデバイスでデータのストレージを管理するための動作を行うのに実行可能な、コンピュータが利用可能なプログラムコードを有するコンピュータ可読媒体を具え、前記コンピュータプログラムプロダクトの動作が、
ストレージ要求に関連するデータ転送を管理するステップであって、要求デバイスとソリッドステートストレージとの間の前記データ転送が、１又はそれ以上のＨＣＮＶストレージデバイス用のキャッシュとして機能し、前記データ転送がデータ、メタデータ及びメタデータインデックスのうちの１又はそれ以上を含み、前記ソリッドステートストレージが不揮発性のソリッドステートデータストレージエレメントのアレイを具えるステップと；
前記ソリッドステートストレージと、前記１又はそれ以上のＨＣＮＶストレージデバイスとの間のデータ転送を管理するステップと；
を含むことを特徴とするコンピュータプログラムプロダクト。
【発明の詳細な説明】
【技術分野】
【０００１】
関連出願の相互引用
この出願は、Ｄａｖｉｄ Ｆｌｙｎｎらによって２００６年１２月６日出願の“Ｅｌｅｍｅｎｔａｌ Ｂｌａｄｅ Ｓｙｓｔｅｍ”と題された米国仮出願番号６０／８７３，１１１、及び、Ｄａｖｉｄ Ｆｌｙｎｎらによって２００７年９月２２日出願の“Ａｐｐａｒａｔｕｓ，Ｓｙｓｔｅｍ，ａｎｄ Ｍｅｔｈｏｄ ｆｏｒ Ｏｂｊｅｃｔ－Ｏｒｉｅｎｔｅｄ Ｓｏｌｉｄ－Ｓｔａｔｅ Ｓｔｏｒａｇｅ”と題された米国仮出願番号６０／９７４，４７０の優先権を主張し、上記仮出願は、引用によって本明細書に組み込まれている。
【０００２】
本発明は、データを管理することに関し、特に、高容量不揮発性ストレージデバイス用キャッシュとしてソリッドステートストレージを用いることに関する。
【背景技術】
【０００３】
一般的にキャッシュは、しばしばアクセスされ、アプリケーション又はオペレーティングシステムの一部として配置されるデータが、ハードディスクドライブ（「ＨＤＤ」）、光学ドライブ、テープストレージ等のような高容量不揮発性（「ＨＣＮＶ」）ストレージデバイスを通してアクセスされなければならない場合よりも更に早い次のアクセスを伴うキャッシュ内に記憶されるので有利である。キャッシュは一般的にはコンピュータ内に含まれる。
【発明の概要】
【０００４】
いくつかのストレージデバイス及びシステムはＨＣＮＶストレージデバイスにキャッシュを具えている。いくつかのＨＣＮＶストレージデバイスは不揮発性ソリッドステートキャッシュを含み、これらはアクセス時間を低減する利益を提供するが、通常は限定されるＨＣＮＶストレージデバイスインタフェースの容量と一致する性能を提供するのみである。マザーボード上に一般的に配置されるいくつかの不揮発性ソリッドステートキャッシュストレージデバイスが存在するが、これらのデバイスはキャッシュコヒーレンスが提供されないようなマルチクライアント環境において用いることができない。ＨＣＮＶデバイスのいくつかのコントローラは更にキャッシュを具える。冗長ＨＣＮＶキャッシュコントローラが複数クライアント間で共有される場合、データが破損されていないことを確認するのに高性能なキャッシュコヒーレンシアルゴリズムが要求される。
【０００５】
一般的に、キャッシュはＤＲＡＭ内に実装され、キャッシュ容量をプレミアムにし、性

能ごとに比較的高い電源を要求する。揮発性キャッシュを支持する電源が失われる場合、キャッシュに記憶されたデータが失われる。一般的にバッテリバックアップが、電源故障の場合にデータ損失を避けるのに用いられるが、バッテリバックアップの停止前に不揮発性メモリに対するキャッシュを消去するかなりの可能性を有する。更に、バッテリバックアップシステムは電力を消費し、冗長性を要求し、信頼性に負の影響を与え、空き領域を消費する。バッテリは標準ベースで更に提供されなければならず、バッテリバックアップは比較的高価である。

【０００６】

前述の考察から、キャッシュとしてソリッドステートストレージを用いてデータを管理する装置、システム及び方法に対するニーズが存在することは明らかである。有利には、このような装置、システム及び方法は少ない電力を消費し、かなり大きな容量を有し、キャッシュに記憶されたデータを維持するのにバッテリバックアップを要求しない不揮発性キャッシュを提供するであろう。

【０００７】

本発明は当該技術分野の現在の状態に応じて、特に、データストレージを管理するための現在利用可能なシステムによって未だ十分に解決されない当該技術分野における問題及びニーズに応じて開発されてきた。従って、本発明は、当該技術分野の上述の欠点の大部分又は総てを克服する１又はそれ以上の高容量不揮発性（「ＨＣＮＶ」）ストレージデバイスでデータのストレージを管理するための装置、システム及び方法を提供するように開発されてきた。

【０００８】

一実施例においては、装置はキャッシュフロントエンドモジュールと、キャッシュバックエンドモジュールとを具える複数のモジュールで提供される。キャッシュフロントエンドモジュールはストレージ要求に関連するデータ転送を管理する。データ転送は要求デバイスと、ソリッドステートストレージとの間で、１又はそれ以上のＨＣＮＶストレージデバイス用のキャッシュとして機能し、データ転送はデータ、メタデータ及びメタデータインデックスのうちの１又はそれ以上を含みうる。ソリッドステートストレージは不揮発性のソリッドステートデータストレージエレメントのアレイを具えることができる。キャッシュバックエンドモジュールはソリッドステートストレージと、１又はそれ以上のＨＣＮＶストレージデバイスとの間のデータ転送を管理する。

【０００９】

装置の一実施例においては、キャッシュフロントエンドモジュール及びキャッシュバックエンドモジュールはソリッドステートストレージを管理するソリッドステートストレージコントローラと共に同一位置に配置される。更なる実施例においては、キャッシュフロントエンドモジュール、キャッシュバックエンドモジュール，及びソリッドステートストレージコントローラは要求デバイスから自律して動作できる。

【００１０】

一実施例においては、装置はＲＡＩＤレベルの一致する独立ドライブ冗長アレイ（「ＲＡＩＤ」）の２又はそれ以上のＨＣＮＶストレージデバイス内のソリッドステートストレージにキャッシュされたデータを記憶するＨＣＮＶ　ＲＡＩＤモジュールを具える。データは全体として要求デバイスに見えうる。別の実施例においては、ソリッドステートストレージ及び１又はそれ以上のＨＣＮＶストレージデバイスは、ＲＡＩＤグループとして構成されるハイブリッドストレージデバイスセット内にハイブリッドストレージデバイスを具えることができる。ソリッドステートストレージにキャッシュされ、ＨＣＮＶデバイスで後に記憶されるデータセグメントは、ストライプのＮ個のデータセグメントのうちの１つ、又はストライプのパリティデータセグメントを含みうる。ハイブリッドストレージデバイスは一般的にＲＡＩＤストライプのデータセグメントと別個に、１又はそれ以上のクライアントからストレージ要求を受信する。更なる実施例においては、ハイブリッドストレージデバイスは２又はそれ以上のクライアントから２又はそれ以上の同時のストレージ要求を受信する、共有されたフロントエンド分散型ＲＡＩＤグループのストレージデバイ

スにできる。

【００１１】

装置の更なる実施例においては、ＨＣＮＶストレージデバイスはハードディスクドライブ（「ＨＤＤ」）、光学ドライブ、又はテープストレージにできる。別の実施例においては、ソリッドステートストレージ及び１又はそれ以上のＨＣＮＶストレージデバイスは、ハイブリッドストレージデバイスにできる。一実施例においては、装置は、ハイブリッドストレージデバイスの動作に特有のコードで、１又はそれ以上の要求デバイスをロードする前に、１又はそれ以上の要求デバイスに取り付けられた標準デバイスをエミュレートすることによって、ハイブリッドストレージデバイスにアクセスを提供する標準デバイスエミュレーションモジュールを更に具える。標準デバイスは一般的には工業規格ＢＩＯＳによって担持されうる。

【００１２】

別の実施例においては、ソリッドステートストレージデバイスは２又はそれ以上の領域に分割可能であり、１又はそれ以上のパーティションが、ＨＣＮＶストレージデバイス用のキャッシュとして機能するソリッドステートストレージと別個の、ソリッドステートストレージとして用いられ得る。更に別の実施例においては、１又はそれ以上のクライアントは、ソリッドステートストレージデバイス及び１又はそれ以上のＨＣＮＶストレージデバイス内に記憶されるファイル又はオブジェクトのうちの１又はそれ以上の状態を管理するために、キャッシュフロントエンドモジュール及びキャッシュバックエンドモジュールにキャッシュ制御メッセージを送信する。

【００１３】

装置の一実施例においては、キャッシュ制御メッセージは１又はそれ以上の制御メッセージを含みうる。制御メッセージの様々な実施例は、キャッシュバックエンドモジュールに、ソリッドステートストレージ内のオブジェクト又はファイルの一部を留める（ｐｉｎ）制御メッセージ、又は、キャッシュバックエンドモジュールに、ソリッドステートストレージ内のオブジェクト又はファイルの一部を解放する（ｕｎｐｉｎ）制御メッセージを含みうる。制御メッセージの他の実施例は、キャッシュバックエンドモジュールに、ソリッドステートストレージから１又はそれ以上のＨＣＶＮストレージデバイスまでのオブジェクト又はファイルの一部を消去させる制御メッセージ、又は、キャッシュバックエンドモジュールに、１又はそれ以上のＨＣＶＮストレージデバイスからソリッドステートストレージに、オブジェクト又はファイルの一部を事前ロードさせる制御メッセージを含みうる。制御メッセージの更なる別の実施例は、ソリッドステートストレージに、決められたストレージ空き領域量を空けるために、キャッシュバックエンドモジュールに、ソリッドステートストレージから１又はそれ以上のＨＣＶＮストレージデバイスまでの１又はそれ以上のオブジェクト又はファイルの１又はそれ以上の部分をオフロードさせる制御メッセージである。一実施例においては、キャッシュ制御メッセージはオブジェクト又はファイル用のメタデータ（「キャッシュ制御メタデータ」）を通して通信される。更なる実施例においては、キャッシュ制御メタデータは永続的にできる。別の実施例においては、キャッシュ制御メタデータはファイル又はオブジェクトの生成時に設定される属性を通して構築されうる。更なる別の実施例においては、キャッシュ制御メタデータはファイル又はオブジェクト管理システムから得うる。

【００１４】

装置の一実施例においては、装置は揮発性キャッシュストレージエレメントを具えることができ、キャッシュフロントエンドモジュール及びキャッシュバックエンドモジュールは、揮発性キャッシュストレージエレメントにデータを記憶し、ソリッドステートストレージ及び揮発性キャッシュストレージエレメントに記憶されるデータを管理できる。バックエンドストレージモジュールは、揮発性キャッシュストレージエレメントと、ソリッドステートストレージと、ＨＣＶＮストレージデバイスとの間のデータ転送を更に管理できる。更なる実施例においては、ＨＣＶＮストレージデバイスに記憶されるオブジェクト及びファイル用のメタデータ及び／又はインデックスメタデータは、ソリッドステートスト

レージデバイス及び揮発性キャッシュストレージエレメント内に維持されうる。

【００１５】

装置の更なる実施例においては、ＨＣＶＮストレージデバイスに記憶されるオブジェクト及びファイル用のメタデータ及び／又はインデックスメタデータは、ソリッドステートストレージデバイス内に維持されうる。別の実施例においては、ソリッドステートストレージ及び１又はそれ以上のＨＣＮＶストレージデバイスが、ストレージデバイスを具え、ＨＣＮＶストレージデバイスがストレージデバイスに接続されたクライアントの表示から隠されるようにできる。

【００１６】

本発明のシステムが更に提供される。システムはモジュール及び装置に関する上述した実施例を実質的に具える。一実施例においては、システムは不揮発性のソリッドステートデータストレージエレメントのアレイを具えるソリッドステートストレージを具える。システムは、１又はそれ以上のＨＣＮＶストレージデバイスと、ストレージコントローラとを更に具える。一実施例においては、ストレージコントローラは、ソリッドステートストレージコントローラと、ＨＣＮＶストレージデバイスコントローラとを具えることができる。ストレージコントローラはキャッシュフロントエンドモジュールと、キャッシュバックエンドモジュールとを更に具えることができる。キャッシュフロントエンドモジュールは、ストレージ要求に関連するデータ転送を管理する。データ転送は一般的に、要求デバイスとソリッドステートストレージとの間で１又はそれ以上のＨＣＮＶストレージデバイス用のキャッシュとして機能している。データ転送はデータ、メタデータ及びメタデータインデックスのうちの１又はそれ以上を含みうる。キャッシュバックエンドモジュールは、ソリッドステートストレージと、１又はそれ以上のＨＣＮＶストレージデバイスとの間のデータ転送を管理する。

【００１７】

一実施例においては、システムはストレージコントローラに接続されたネットワークインタフェースを更に具え、当該ネットワークインタフェースがコンピュータネットワークを通して要求デバイスとソリッドステートストレージコントローラとの間のデータ転送を促進する。別の実施例においては、システムはソリッドステートストレージと、１又はそれ以上のＨＣＮＶストレージデバイスと、ストレージコントローラとを具えるサーバを具える。更なる別の実施例においては、１又はそれ以上のＨＣＮＶストレージデバイスはストレージエリアネットワーク（「ＳＡＮ」）を通してストレージコントローラに接続される。

【００１８】

本発明の方法は、複数のホスト間でデバイスを共有するために更に提供される。開示された実施例における方法は、説明した装置及びシステムの動作に関する、上に提供した機能を実行するのに必要なステップを実質的に含む。一実施例においては、方法は、ストレージ要求に関連するデータ転送を管理するステップを含み、データ転送は、要求デバイスとソリッドステートストレージとの間で、１又はそれ以上のＨＣＮＶストレージデバイス用のキャッシュとして機能している。データ転送はデータ、メタデータ及びメタデータインデックスのうちの１又はそれ以上を含みうる。ソリッドステートストレージは、不揮発性のソリッドステートデータストレージエレメントのアレイを具えることができる、方法は、ソリッドステートストレージと、１又はそれ以上のＨＣＮＶストレージデバイスとの間のデータ転送を管理するステップを更に具えることができる。

【００１９】

本明細書全体にわたる、特徴、利点または同様の単語の引用は、本発明で実現される特徴および利点の総てが、本発明のいずれかの１つの実施例に存在すべき、または、存在することを意味するものではない。むしろ、特徴及び利点に関する用語は、ある実施例に関連して記載される特定の特徴、利点または特性が、本発明の少なくとも１つの実施例に含まれることを意味すると理解されたい。従って、特徴、利点および同様の用語の記載は、本明細書全体にわたって、必須ではないが、同一の実施例を意味する。

【００２０】
更に、記載された本発明の特徴、利点および特性は、１又はそれ以上の実施例においてあらゆる適宜な方法で組み合わせることができる。関連分野の当業者は、特定の実施例の１又はそれ以上の特定の特徴または利点がなくとも、本発明が実施可能であることが分かるであろう。その他の例においては、更なる特徴および利点は、本発明の総ての実施例において示されていない特定の実施例で分かるであろう。
【００２１】
本発明の特徴および利点は、以下の説明および添付の請求の範囲から完全に明らかになり、また、以下に記載する本発明の実施形態によって理解されるであろう。
【図面の簡単な説明】
【００２２】
本発明の利点が容易に理解されるように、簡単に上述した本発明のより具体的な説明は、添付の図面に例示した特定の実施例を参照することによってなされるであろう。これらの図面は、本発明の一般的な実施例のみを示し、従って、本発明の範囲を限定するものではないと理解すべきであり、本発明は、添付の図面の利用を通じて、更に具体的かつ詳細に記載および説明されるであろう。
【００２３】
【図１Ａ】図１Ａは、本発明によるソリッドステートストレージデバイス中のデータ管理のためのシステムの一実施例を示した概略ブロック図である。
【図１Ｂ】図１Ｂは、本発明によるストレージデバイス中のオブジェクト管理のためのシステムの一実施例を示した概略ブロック図である。
【図１Ｃ】図１Ｃは、本発明によるインサーバストレージエリアネットワーク用のシステムの一実施例を示した概略ブロック図である。
【図２Ａ】図２Ａは、本発明によるストレージデバイスのオブジェクト管理のための装置の一実施例を示した概略ブロック図である。
【図２Ｂ】図２Ｂは、本発明によるソリッドステートストレージデバイス中のソリッドステートストレージデバイスコントローラの一実施例を示した概略ブロック図である。、
【図３】図３は、本発明によるソリッドステートストレージデバイス内の書き込みデータパイプラインおよび読み出しデータパイプラインを有するソリッドステートストレージコントローラを示す概略ブロック図である。
【図４Ａ】図４Ａは、本発明によるソリッドステートストレージコントローラ内ののバンクインタリーブコントローラの一実施例を示した概略ブロック図である。
【図４Ｂ】図４Ｂは、本発明によるソリッドステートストレージコントローラの内のバンクインタリーブコントローラの代替的な実施例
【図５Ａ】図５Ａは、本発明によるデータパイプラインを用いたソリッドステートストレージデバイス内のデータを管理するための方法の一実施例を示した概略フローチャート図である。
【図５Ｂ】図５Ｂは、本発明によるインサーバＳＡＮのための方法の一実施例を示した概略フローチャート図である。
【図６】図６は、本発明によるデータパイプラインを用いたソリッドステートストレージデバイス内のデータを管理するための方法の別の実施例を示した概略フローチャート図である。
【図７】図７は、本発明によるバンクインタリーブを用いたソリッドステートストレージデバイス内のデータを管理するための方法の一実施例を示した概略フローチャート図である。
【図８】図８は、本発明によるソリッドステートストレージデバイス内のガベージコレクション用の装置の一実施例を示した概略ブロック図である。
【図９】図９は、本発明によるソリッドステートストレージデバイス内のガベージコレクション用の方法の一実施例を示した概略フローチャート図である。
【図１０】図１０は、本発明による進行型ＲＡＩＤ、フロントエンド分散型ＲＡＩＤ、及

び共有されたフロントエンド分散型ＲＡＩＤ用のシステムの一実施例を示した概略ブロック図である。
【図１１】図１１は、本発明によるフロントエンド分散型ＲＡＩＤ用の装置の一実施例を示した概略ブロック図である。
【図１２】図１２は、本発明によるフロントエンド分散型ＲＡＩＤ用の方法の一実施例を示した概略フローチャート図である。
【図１３】図１３は、本発明による共有されたフロントエンド分散型ＲＡＩＤ用の装置の一実施例を示した概略ブロック図である。
【図１４】図１４は、本発明による共有されたフロントエンド分散型ＲＡＩＤ用の方法の一実施例を示した概略フローチャート図である。
【図１５】図１５は、本発明による高容量不揮発性ストレージデバイス用のキャッシュとしてソリッドステートストレージを有するシステムの一実施例を示した概略ブロック図である。
【図１６】図１６は、本発明による高容量不揮発性ストレージデバイス用のキャッシュとしてソリッドステートストレージを有する装置の一実施例を示した概略ブロック図である。
【図１７】図１７は、本発明による高容量不揮発性ストレージデバイス用のキャッシュとしてソリッドステートストレージを有する方法の一実施例を示した概略フローチャート図である。
【発明を実施するための形態】
【００２４】
本明細書に記載された多数の機能ユニットは、その実装独立性をより具体的に強調するために、モジュールとして表示した。例えば、モジュールは、カスタムＶＬＳＩ回路またはゲートアレイ、論理チップのような標準規格の半導体、トランジスタまたは他の別個の構成要素を具えるハードウェア回路として実装可能である。更に、モジュールは、フィールドプログラマブルゲートアレイ、プログラマブル配列論理回路、プログラマブル論理デバイスなどのような、プログラム可能なハードウェアデバイスに実装可能である。
【００２５】
更に、モジュールは、様々なプロセッサによって実行されるソフトウェアに実装可能である。識別モジュールの実行可能コードは、例えばオブジェクト、手続きまたは関数として構成される、計算機命令の１又はそれ以上の物理又は論理ブロックを含む。にもかかわらず、実行可能な識別モジュールは、物理的に一緒に割り当てられる必要はなく、論理的に互いに接続される場合、モジュールを構成し、モジュールのための規定された目的を達成する、異なる位置に記憶される離隔命令を具えることができる。
【００２６】
実際に、モジュールの実行可能コードは、単一の命令または多数の命令にでき、複数の異なるコードセグメントにわたり、異なるプログラム間で、および、複数のメモリデバイスにわたってさえ割り当てることができる。同様に、動作データは識別され、本明細書でモジュール内部に例示されており、適宜な形態で組み込まれ、あらゆる適宜な種類のデータ構造内で構成される。動作データは、単一のデータセットとして集められ、異なるストレージデバイスを含む異なる位置にわたって割り当てられ、少なくとも部分的に、システムまたはネットワックの単なる電気信号として存在しうる。モジュールまたはモジュールの一部がソフトウェアに実装される場合、ソフトウェア部分は、１又はそれ以上のコンピュータ可読メディアに記憶される。
【００２７】
本明細書全体における「一実施例（ｏｎｅ　ｅｍｂｏｄｉｍｅｎｔ）」、「ある実施例（ａｎ　ｅｍｂｏｄｉｍｅｎｔ）」または同様の用語の引用は、実施例に関連して記載される特定の特徴、構造または特性が本発明の少なくとも１つの実施例に含まれることを意味する。従って、本明細書全体における「一実施例においては（ｉｎ　ｏｎｅ　ｅｍｂｏｄｉｍｅｎｔ）」、「ある実施例においては（ｉｎ　ａｎ　ｅｍｂｏｄｉｍｅｎｔ）」お

よび同様の用語のフレーズの出現は、必ずではないが、同一の実施例を総て引用している。

【００２８】

信号担持媒体については、信号を生成し、信号が生成され、または、ディジタル処理装置で機械が読み出し可能な命令のプログラムの実行をさせることが可能なあらゆる形態をとることができる。信号担持媒体は、伝送線路、コンパクトディスク、ディジタルビデオディスク、磁気テープ、ベルヌーイドライブ、磁気ディスク、パンチカード、フラッシュメモリ、集積回路または他のディジタル処理装置のメモリデバイスに組み込まれうる。

【００２９】

更に、記載された本発明の特徴、構造または特性は、１又はそれ以上の実施例において適宜な方法で組み合わせることができる。以下の説明において、例えば、プログラミング、ソフトウェアモジュール、ユーザの選択項目、ネットワークトランザクション、データベースクエリ、データベース構造、ハードウェアモジュール、ハードウェア回路、ハードウェアチップ等のような、多数の特定の詳細な説明が提供されて、本発明の実施例を完全な理解を提供する。しかしながら、関連分野の当業者は、本発明が１又はそれ以上の特定の詳細な説明なしで、あるいは、他の方法、構成、材料などを用いて実施できることが分かるであろう。別の例においては、既知の構造、材料または動作は、本発明の態様があいまいになるのを避けるために、示していないか、または、詳細に記載していない。

【００３０】

本明細書に含まれる概略的フローチャートは、論理フローチャート図として一般的に説明されている。このように、示された順番および示されたステップは、本方法の一実施例を示す。他のステップ及び方法は、例示された方法の関数、論理または１又はそれ以上のステップに対する効果、またはその一部において同等であると考えられる。更に、使用したフォーマットおよび符号は、方法の論理的なステップを説明するために提供されたものであり、この方法の範囲を限定するものではないと理解されたい。様々な矢印および線がフローチャート図で用いられているが、対応する方法の範囲を限定するものではないことを理解されたい。実際に、いくつかの矢印または他の結合子は、この方法の論理フローを示すためにのみ用いられている。例えば、矢印は、示された方法の列挙されたステップ間における不特定期間の待機又はモニタリング時間を示す。更に、特定の方法が生じる順番は、示された対応するステップの順番に厳密に準拠してもしなくてもよい。

【００３１】

ソリッドステートストレージシステム

【００３２】

図１Ａは、本発明によるソリッドステートストレージデバイスのデータ管理用のシステム１００の一実施例を示した概略ブロック図である。システム１００は、以下に示すように、ソリッドステートストレージデバイス１０２、ソリッドステートストレージコントローラ１０４、書き込みデータパイプライン１０６、読み出しデータパイプライン１０８、ソリッドステートストレージ１１０、コンピュータ１１２、クライアント１１４およびコンピュータネットワーク１１６を具える。

【００３３】

システム１００は、少なくとも１のソリッドステートストレージデバイス１０２を具える。別の実施例においては、システム１００は、２又はそれ以上のソリッドステートストレージ１０２を具える。各ソリッドステートストレージデバイス１０２は、フラッシュメモリ、ナノランダムアクセスメモリ（「ナノＲＡＭまたはＮＲＡＭ」）、磁気抵抗ＲＡＭ（「ＭＲＡＭ」）、ダイナミックＲＡＭ（「ＤＲＡＭ」）、相変化ＲＡＭ（「ＰＲＡＭ」）などのような、不揮発性ソリッドステートストレージ１１０を具える。ソリッドステートストレージデバイス１０２は、図２および図３で更に詳細に示す。ソリッドステートストレージデバイス１０２は、コンピュータネットワーク１１６を通って、クライアント１１４に接続したコンピュータ１１２内に示されている。一実施例においては、ソリッドステートストレージデバイス１０２は、コンピュータ１１２の内部にあり、ＰＣＩ（ｐｅｒ

ｉｐｈｅｒａｌ ｃｏｍｐｏｎｅｎｔ ｉｎｔｅｒｃｏｎｎｅｃｔ：周辺装置相互接続規格）エクスプレス（「ＰＣＩ－ｅ」）バス、シリアルアドバンスドテクノロージアタッチメント（「シリアルＡＴＡ」）バスなどのような、システムバスを用いて接続される。別の実施例においては、ソリッドステートストレージデバイス１０２は、コンピュータ１１２の外部にあり、ユニバーサルシリアルバス（「ＵＳＢ」）接続、ＩＥＥＥ（Ｉｎｓｔｉｔｕｔｅ ｏｆ Ｅｌｅｃｔｒｉｃａｌ ａｎｄ Ｅｌｅｃｔｒｏｎｉｃｓ Ｅｎｇｉｎｅｅｒｓ：米国電気電子技術者協会）１３９４バス（「ＦｉｒｅＷｉｒｅ」）などを用いて接続される。他の実施例においては、ソリッドステートストレージデバイス１０２は、インフィニバンドまたはＰＣＩエクスプレスアドバンスドスイッチング（「ＰＣＩｅ－ＡＳ」）などのような、外部の電気式または光学式バス拡張機能又はバス型ネットワークソリューションを用い、ＰＣＩ（ｐｅｒｉｐｈｅｒａｌ ｃｏｍｐｏｎｅｎｔ ｉｎｔｅｒｃｏｎｎｅｃｔ：周辺装置相互接続規格）エクスプレスバスを用いたコンピュータ１１２に接続される。

【００３４】

様々な実施例において、ソリッドステートストレージ１０２は、デュアルインラインメモリモジュール（「ＤＩＭＭ」）、ドータカードまたはマイクロモジュールの形態でもよい。別の実施例においては、ソリッドステートストレージデバイス１０２は、ラック取付型ブレード内のエレメントである。別の実施例においては、ソリッドステートストレージデバイス１０２は、より高位のアセンブリ（例えば、マザーボード、ラップトップ、グラフィックプロセッサ）に直接統合したパッケージ内に含まれる。別の実施例においては、ソリッドステートストレージデバイス１０２を具える個々の構成は、中位の実装なしにより高位のアセンブリに直接統合される。

【００３５】

ソリッドステートストレージデバイス１０２は、１又はそれ以上のソリッドステートストレージコントローラ１０４を具え、各々は、書き込みデータパイプライン１０６および読み出しデータパイプライン１０８を具え、各々は、ソリッドステートストレージ１１０を具え、これらは、図２および図３で以下に詳細に示す。

【００３６】

システム１００は、ソリッドステートストレージデバイス１０２に接続した１又はそれ以上のコンピュータ１１２を具える。コンピュータ１１２は、ホスト、サーバ、ストレージエリアネットワーク（「ＳＡＮ」）のストレージコントローラ、ワークステーション、パーソナルコンピュータ、ラップトップコンピュータ、ハンドヘルドコンピュータ、スーパコンピュータ、コンピュータクラスタ、ネットワークスイッチ、ルータまたは機器、データベースまたはストレージ機器、データ取得システムまたはデータキャプチャシステム、診断システム、試験システム、ロボット、携帯型電子機器、ワイヤレスデバイスなどにできる。別の実施例においては、コンピュータ１１２は、クライアントでもよく、ソリッドステートストレージデバイス１０２は、コンピュータ１１２から送信されるデータ要求を提供するために自律動作する。この実施例においては、コンピュータ１１２およびソリッドステートストレージデバイス１０２は、コンピュータネットワーク、システムバス、または、コンピュータ１１２と自律型ソリッドステートストレージデバイス１０２との間の接続に適切なその他の通信手段を用いて、接続可能である。

【００３７】

一実施例においては、システム１００は、１又はそれ以上のコンピュータネットワーク１１６を通して、１又はそれ以上のコンピュータ１１２に接続された１又はそれ以上のクライアント１１４を具える。クライアント１１４は、ホスト、サーバ、ＳＡＮのストレージコントローラ、ワークステーション、パーソナルコンピュータ、ラップトップコンピュータ、ハンドヘルドコンピュータ、スーパコンピュータ、コンピュータクラスタ、ネットワークスイッチ、ルータまたは機器、データベースまたはストレージ機器、データ取得またはデータキャプチャシステム、診断システム、試験システム、ロボット、携帯電子機器、ワイヤレスデバイスなどにできる。コンピュータネットワーク１１６は、インターネッ

ト、ワイドエリアネットワーク（「ＷＡＮ」）、メトロポリタンエリアネットワーク（「ＭＡＮ」）、ローカルエリアネットワーク（「ＬＡＮ」）、トークンリング、ワイヤレスネットワーク、ファイバチャネルネットワーク、ＳＡＮ、ネットワーク接続ストレージ（「ＮＡＳ」）、ＥＳＣＯＮ等、または、いずれかのネットワークの組み合わせを具えることができる。更に、コンピュータネットワーク１１６は、イーサネット（登録商標）、トークンリング、ＷｉＦｉ、ＷｉＭａｘなどのような、ＩＥＥＥ８０２ファミリのネットワーク技術からのネットワークを具えることができる。

【００３８】

コンピュータネットワーク１１６は、サーバ、スイッチ、ルータ、ケーブル、無線、および、コンピュータ１１２とクライアント１１４とのネットワーク構築を容易にするのに用いられるその他の設備を具える。一実施例においては、システム１００は、コンピュータネットワーク１１６にわたってピアとして通信する多数のコンピュータ１１２を具える。別の実施例においては、システム１００は、コンピュータ１１６にわたってピアとして通信する複数のソリッドステートストレージデバイス１０２を具える。関連分野の当業者は、１又はそれ以上のコンピュータネットワーク１１６と、１又はそれ以上のクライアント１１４間の単一または冗長接続を有する関連設備、または、１又はそれ以上のソリッドステートストレージデバイス１０２を有する他のコンピュータ、または、１又はそれ以上のコンピュータ１１２に接続された１又はそれ以上のソリッドステートストレージデバイス１０２とを具える他のコンピュータネットワーク１１６を認識するであろう。一実施例においては、システム１００は、コンピュータ１１２なしで、コンピュータネットワーク１１８を通ってクライアント１１６に接続した２又はそれ以上のソリッドステートストレージデバイス１０２を具える。

【００３９】

ストレージコントローラ管理オブジェクト

【００４０】

図１Ｂは、本発明によるストレージデバイスのオブジェクト管理用のシステム１０１の一実施例を示す概略ブロック図である。システム１０１は、各々がストレージコントローラ１５２および１又はそれ以上のデータストレージデバイス１５４を有する１又はそれ以上のストレージデバイス１５０と、１又はそれ以上の要求デバイス１５５とを具える。ストレージデバイス１５２は、互いにネットワークで接続され、１又はそれ以上の要求デバイス１５５に結合される。要求デバイス１５５は、ストレージデバイス１５０ａにオブジェクト要求を送信する。オブジェクト要求は、オブジェクトを生成する要求、オブジェクトにデータを書き込む要求、オブジェクトからデータを読み出す要求、オブジェクトを削除する要求、オブジェクトを動作確認する要求、オブジェクトをコピーする要求などができる。当分野の当業者は、その他のオブジェクト要求もわかるであろう。

【００４１】

一実施例においては、ストレージコントローラ１５２およびデータストレージデバイス１５４は別個のデバイスである。別の実施例においては、ストレージコントローラ１５２およびデータストレージデバイス１５４は、１つのストレージデバイス１５０に統合される。別の実施例においては、データストレージデバイス１５４は、ソリッドステートストレージ１１０であり、ストレージコントローラは、ソリッドステートストレージデバイスコントローラ２０２である。他の実施例においては、データストレージデバイス１５４は、ハードディスクドライブ、光学ドライブ、テープストレージ等にできる。別の実施例においては、ストレージデバイス１５０は、異なる種類の２又はそれ以上のデータストレージデバイス１５４を具えることができる。

【００４２】

一実施例においては、データストレージデバイス１５４は、ソリッドステートストレージ１１０であり、ソリッドステートストレージエレメント２１６、２１８、２２０のアレイとして配置される。別の実施例においては、ソリッドステートストレージ１１０は２又はそれ以上のバンク２１４ａ－ｎに配置される。ソリッドステートストレージ１１０は、

図２Ｂで以下により詳細に示す。

【００４３】

これらのストレージデバイス１５０ａ－ｎは、互いにネットワークで接続され、分散ストレージデバイスとして機能する。要求デバイス１５５に接続されたストレージデバイス１５０ａは、分散ストレージデバイスに対するオブジェクト要求を制御する。一実施例においては、ストレージデバイス１５０および付随するストレージコントローラ１５２は、オブジェクトを管理し、分散オブジェクトファイルシステムとして要求デバイス１５５に存在する。この構成において、並行オブジェクトファイルサーバは、分散オブジェクトファイルシステム型の例である。別の実施例においては、ストレージデバイス１５０および付随のストレージコントローラ１５２は、オブジェクトを管理し、分散オブジェクトファイルサーバとして要求デバイス１５５に存在する。この構成において、並行オブジェクトファイルサーバは、分散オブジェクトファイルサーバ型の例である。これらの実施例およびその他の実施例において、要求デバイス１５５は、オブジェクトを排他的に管理し、または、ストレージデバイス１５０と併用してオブジェクトを管理することに関与することができ、これは一般的に、ストレージデバイス１５０の能力を制限せず、他のクライアント１１４のためのオブジェクトを完全に管理する。縮退型の場合には、各々の分散ストレージデバイス、分散オブジェクトファイルシステムおよび分散オブジェクトファイルサーバは、単一のデバイスとして別個に動作できる。ネットワーク接続されたストレージデバイス１５０ａ－ｎは、分散ストレージデバイス、分散オブジェクトファイルシステム、分散オブジェクトファイルサーバ、および、１又はそれ以上の要求デバイス１５５に対して構成された１又はそれ以上の上記能力の表現を有するその組み合わせとして、動作できる。例えば、ストレージデバイス１５０は、第１の要求デバイス１５５ａ用の分散ストレージデバイスとして動作するように構成でき、一方、要求デバイス１５５ｂ用の分散ストレージデバイスおよび分散オブジェクトファイルシステムとして動作する。システム１０１が、１つのストレージデバイス１５０ａを具える場合、ストレージデバイス１５０ａのストレージコントローラ１５２ａは、オブジェクトを管理し、オブジェクトファイルシステムまたはオブジェクトファイルサーバとして要求デバイス１５５に存在してもよい。

【００４４】

ストレージデバイス１５０が、分散ストレージデバイスとして互いにネットワークで接続される一実施例において、ストレージデバイス１５０は、１又はそれ以上の分散ストレージコントローラ１５２によって管理された独立ドライブ冗長アレイ（「ＲＡＩＤ」）として動作する。例えば、オブジェクトのデータセグメントを書き込む要求は、ＲＡＩＤレベルに応じて、データセグメントがパリティストライプを有するデータストレージデバイス１５４ａ－ｎにわたってストライプ化される結果となる。このような配置の利点の１つは、単一ストレージデバイス１５０がストレージコントローラ１５２、データストレージデバイス１５４またはストレージデバイス１５０のその他の構成のいずれかに故障を有するときに、このようなオブジェクト管理システムが利用し続けることができることである。

【００４５】

冗長ネットワークを用いてストレージデバイス１５０と要求デバイス１５５とを相互接続するとき、オブジェクト管理システムは、ネットワークの１つが動作している限り、ネットワーク故障の存在下で利用し続けることができる。単一のストレージデバイス１５０ａを有するシステム１０１は、更に、複数のデータストレージデバイス１５４ａを具えることができ、ストレージデバイス１５０ａのストレージコントローラ１５２ａは、ＲＡＩＤコントローラとして動作し、ストレージデバイス１５０ａのデータストレージデバイス１５４ａにわたってデータセグメントをストライプ化し、ＲＡＩＤレベルに応じて、パリティストライプを具えることができる。

【００４６】

一実施例においては、１又はそれ以上のストレージデバイス１５０ａ－ｎは、ソリッドステートストレージデバイスコントローラ２０２およびソリッドステートストレージ１１

０を有するソリッドステートストレージデバイス１０２である場合、ソリッドステートストレージデバイス１０２は、ＤＩＭＭ構成、ドータカード、マイクロモジュールなどに構成され、コンピュータ１１２内に常駐できる。コンピュータ１１２は、サーバ、または、互いにネットワーク接続され、分散ＲＡＩＤコントローラとして動作するソリッドステートストレージデバイスコントローラ１０２を有する同様のデバイスでもよい。有利には、ストレージデバイス１０２は、ＰＣＩ－ｅ、ＰＣＩｅ－ＡＳ、インフィニバンドまたはその他の高性能バス、スイッチングバス、ネットワーク接続バスまたはネットワークを用いて接続されてもよく、ソリッドステートストレージ１１０ａ－ｎにわたってデータセグメントを自律的にストライプ化する単一又は分散ソリッドステートストレージコントローラ２０２を有する極めて小型の、高性能ＲＡＩＤストレージシステムを提供することができる。

【００４７】

一実施例においては、ストレージデバイス１５０と通信するために要求デバイス１５５によって用いられる同一ネットワークは、ピアストレージデバイス１５０ｂ－ｎと通信しＲＡＩＤ機能を得るために、ピアストレージデバイス１５０ａによって用いられる。別の実施例においては、別個のネットワークは、ＲＡＩＤするために、ストレージデバイス１５０間で用いられうる。別の実施例においては、要求デバイス１５５は、ストレージデバイス１５０に冗長要求を送信することによってＲＡＩＤ処理に関与することができる。例えば、要求デバイス１５５は、第１のオブジェクト書き込み要求を第１のストレージデバイス１５０ａに送信し、同一データセグメントを有する第２のオブジェクト書き込み要求を第２のストレージデバイス１５０ｂに送信し、単純なミラーリングを行うことができる。

【００４８】

ストレージデバイス１０２内でオブジェクト処理するための能力を用いて、ストレージコントローラ１５２は、１つのＲＡＩＤレベルを用いて、１つのデータセグメントまたはオブジェクトを記憶する能力を個別に有し、一方、別のデータセグメントまたはオブジェクトは、異なるＲＡＩＤレベルを用いて、または、ＲＡＩＤストライブをせずに、記憶される。これらの複数のＲＡＩＤグループは、ストレージデバイス１５０内の複数のパーティションに関連していてもよい。ＲＡＩＤ０、ＲＡＩＤ１、ＲＡＩＤ５、ＲＡＩＤ６および複合ＲＡＩＤタイプ１０、５０、６０は、データストレージデバイス１５４ａ－ｎを具える様々なＲＡＩＤグループにわたり同時にサポート可能である。当分野の当業者は、同時にサポート可能なその他のＲＡＩＤタイプおよび構成が分かるであろう。

【００４９】

更に、ストレージコントローラ１５２は、ＲＡＩＤコントローラとして自律的に動作するので、ＲＡＩＤコントローラは、進行型ＲＡＩＤを行うことができ、１つのＲＡＩＤレベルを有するデータストレージデバイス１５４にわたってストライプ化したオブジェクトまたはオブジェクトの一部を、要求デバイス１５５が影響を受け、関与し、または、ＲＡＩＤレベルの変化を検出さえすることなく、別のＲＡＩＤレベルに変換可能である。好ましい実施例においては、１つのレベルから別のレベルへとＲＡＩＤ構成を進行させることは、オブジェクトまたはパケットベースでさえも自律的に行われ、ストレージデバイス１５０の１つまたはストレージコントローラ１５２において動作する分散ＲＡＩＤ制御モジュールによって起動される。一般的に、ＲＡＩＤの進行は、ＲＡＩＤ１などのより高性能でより低効率なストレージ構成から、ＲＡＩＤ５などのより低性能でより高い効率のストレージ構成への進行であり、変換は、アクセス頻度に基づいて動的に開始される。なお、ＲＡＩＤ５からＲＡＩＤ１へ構成を進行させることも可能であることは分かるであろう。ＲＡＩＤ進行を開始するためのその他のプロセスは、ストレージシステム管理サーバ要求など、クライアントまたは外部エージェントから構成または要求されうる。当分野の当業者は、自律的にオブジェクトを管理するストレージコントローラ１５２を有するストレージデバイス１０２の特徴および利点が分かるであろう。

【００５０】

インサーバＳＡＮを用いたソリッドステートストレージデバイス

【００５１】

図１Ｃは、本発明によるインサーバストレージエリアネットワーク（「ＳＡＮ」）用システム１０３の一実施例を示した概略ブロック図である。システム１０３は、通常、サーバ（「サーバ１１２」）として設定されるコンピュータ１１２を具える。各サーバ１１２は、１又はそれ以上のストレージデバイス１５０を具え、サーバ１１２およびストレージデバイス１５０は、共有されたネットワークインタフェース１５６に接続される。各ストレージデバイス１５０は、ストレージコントローラ１５２および関連するデータストレージデバイス１５４を具える。システム１０３は、サーバ１１２の内部または外部にあるクライアント１１４、１１４ａ、１１４ｂを具える。クライアント１１４、１１４ａ、１１４ｂは、上述したものと実質的に同様の、１又はそれ以上のコンピュータネットワーク１１６を介して、各サーバ１１２および各ストレージデバイス１５０と通信することができる。

【００５２】

ストレージデバイス１５０は、ＤＡＳモジュール１５８、ＮＡＳモジュール１６０、ストレージ通信モジュール１６２、インサーバＳＡＮモジュール１６４、共通インタフェースモジュール１６６、プロキシモジュール１７０、仮想バスモジュール１７２、フロントエンドＲＡＩＤモジュール１７４およびバックエンドＲＡＩＤモジュール１７６を具え、これらは以下に示す。モジュール１５８－１７６はストレージデバイス１５０内に示される一方、各モジュール１５８－１７６の全てまたは一部は、ストレージデバイス１５０、サーバ１１２、ストレージコントローラ１５２またはその他のロケーションにあってもよい。

【００５３】

インサーバＳＡＮと併せて使用されるサーバ１１２は、サーバとして機能するコンピュータである。このサーバ１１２は、ファイルサーバ機能など、少なくとも１つのサーバ機能を具えるが、その他のサーバ機能も具えてもよい。サーバ１１２は、いくつかのファームの一部であってもよく、他のクライアント１１４に応答する。他の実施例においては、サーバ１１２は、パーソナルコンピュータ、ワークステーションまたはストレージデバイス１５０を収容する他のコンピュータでもよい。サーバ１１２は、ダイレクト接続ストレージ（「ＤＡＳ」）、ＳＡＮ接続ストレージ、または、ネットワーク接続ストレージ（「ＮＡＳ」）としてのサーバ１１２の１又はそれ以上のストレージデバイス１５０にアクセスできる。インサーバＳＡＮまたはＮＡＳに関与するストレージコントローラ１５０は、サーバ１１２の内部または外部にあってもよい。

【００５４】

一実施例においては、インサーバＳＡＮ装置は、少なくとも１つのデータストレージデバイス１５４の少なくとも一部を構成するＤＡＳモジュール１５８を具え、この少なくとも１のデータストレージデバイス１５４は、少なくとも１のクライアント１１４からサーバ１１２にストレージ要求を送るように、サーバ１１２に接続したＤＡＳデバイスとしてサーバ１１２のストレージコントローラ１５２によって制御される。一実施例においては、第１のデータストレージデバイス１５４ａは、第１のサーバ１１２に対するＤＡＳとして設定され、さらに、第１のサーバ１１２ａに対するインサーバＳＡＮストレージデバイスとしても設定される。別の実施例においては、第１のデータストレージデバイス１５４ａは、１つのパーティションは、ＤＡＳであり、他方が、インサーバＳＡＮとなるようにパーティション化される。別の実施例においては、第１のデータストレージデバイス１５４ａ内の少なくともストレージスペースの一部は第１のサーバ１１２ａに対するＤＡＳとして設定され、第１のデータストレージデバイス１５４ａのストレージスペースの同一部分は、第１のサーバ１１２ａに対するインサーバＳＡＮとして設定される。

【００５５】

別の実施例においては、インサーバＳＡＮ装置は、少なくとも１つのクライアント１１４用のＮＡＳデバイスとしてストレージコントローラ１５２を構成してクライアント１１

４からのファイル要求に応答する、ＮＡＳモジュール１６０を具える。さらに、ストレージコントローラ１５２は、第１のサーバ１１２ａ用のインサーバＳＡＮデバイスとして設定可能である。ストレージデバイス１５０は、ストレージデバイス１５０のあるサーバ１１２から独立して、共有されたネットワークインタフェース１５５を介して、コンピュータネットワーク１１６に直接接続できる。

【００５６】

一の構成要素形態において、インサーバＳＡＮ装置は、第１のサーバ１１２ａ内に第１のストレージコントローラ１５２ａを具え、第１のストレージコントローラ１５２ａが少なくとも１つのストレージデバイス１５４ａを制御する。第１のサーバ１１２ａは、第１のサーバ１１２ａおよび第１のストレージコントローラ１５２ａによって共有されるネットワークインタフェース１５６を具える。インサーバＳＡＮ装置は、第１のストレージコントローラ１５２ａと第１のサーバ１１２ａの外部にある少なくとも１つのデバイスとの間の通信を促進するストレージ通信モジュール１６２を具え、第１のストレージコントローラ１５２ａと前記外部デバイスとの間の通信は、第１のサーバ１１２ａから独立している。ストレージ通信モジュール１６２によって、第１のストレージコントローラ１５２ａは、外部通信用ネットワークインタフェース１５６ａに独立してアクセス可能になる。一実施例においては、ストレージ通信モジュール１６２は、ネットワークインタフェース１５６ａのスイッチにアクセスし、第１のストレージコントローラ１５２ａと外部デバイスとの間のネットワークトラフィックを管理（ｄｉｒｅｃｔ）する。

【００５７】

さらに、インサーバＳＡＮ装置は、ネットワークプロトコルおよびバスプロトコルの一方又は両方を用いてストレージ要求に応答するインサーバＳＡＮモジュール１６４を具える。インサーバＳＡＮモジュール１６４は、第１のサーバ１１２ａから独立したストレージ要求に応答し、サービス要求は、内部または外部クライアント１１４、１１４ａから受信する。

【００５８】

一実施例においては、第１のサーバ１１２ａの外部にあるデバイスは、第２のストレージコントローラ１５２ｂである。第２のストレージコントローラは、少なくとも１のデータストレージデバイス１５４ｂを制御する。インサーバＳＡＮモジュール１６４は、第１のサーバ１１２ａから独立して、第１および第２のストレージコントローラ１５２ａ、１５２ｂ間において、ネットワークインタフェース１５６ａを介した通信を使用して、ストレージ要求に応答する。第２のストレージコントローラ１５２ｂは、第２のサーバ１１２ｂまたはその他のデバイス内にあってもよい。

【００５９】

別の実施例においては、第１のサーバ１１２ａの外部にあるデバイスは、クライアント１１４であり、ストレージ要求は、外部クライアント１１４からきて、第１のストレージコントローラは、ＳＡＮの少なくとも１部として設定され、インサーバＳＡＮモジュール１６４は、第１のサーバ１１２ａから独立してネットワークインタフェース１５６ａを介してストレージ要求に応答する。外部クライアント１１４は、第２のサーバ１１２ｂ内でもよく、第２サーバ１１２ｂの外部にあってもよい。一実施例においては、インサーバＳＡＮモジュール１６４は、第１のサーバ１１２が利用できない場合であっても外部クライアント１１４からストレージ要求に応答することができる。

【００６０】

別の実施例においては、ストレージ要求を発するクライアント１１４ａは、第１のサーバ１１２ａの内部にあり、第１のストレージコントローラ１５２ａは、ＳＡＮの少なくとも一部として設定され、インサーバＳＡＮモジュール１６４は、１又はそれ以上のネットワークインタフェース１５６ａおよびシステムバスを介してストレージ要求に応答する。

【００６１】

従来のＳＡＮ設定は、サーバ１１２から遠隔にあるストレージデバイスが、ダイレクト接続ストレージ（「ＤＡＳ」）としてサーバ１１２内にあるようにアクセスでき、ストレ

ージデバイスは、ブロックストレージデバイスとしてみえた。一般的に、ＳＡＮとして接続されたストレージデバイスは、ファイバーチャネル、インターネットスモールコンピュータシステムインタフェース（「ｉＳＣＳＩ」）、ハイパーＳＣＳＩ、ファイバーコネクティビティ（「ＦＩＣＯＮ」）、イーサネット（登録商標）を介したアドバンステクノロージアタッチメント（「ＡＴＡ」）など、ＳＡＮプロトコルを要求する。インサーバＳＡＮは、サーバ１１２内部にストレージコントローラ１５２を具え、ネットワークプロトコルおよび／またはバスプロトコルを用いて、ストレージコントローラ１５２ａと、リモートストレージコントローラ１５２ｂまたは外部クライアント１１４と、の間のネットワーク接続を可能にしている。

【００６２】

一般的に、ＳＡＮプロトコルは、ネットワークプロトコルのフォームであり、より多くのネットワークプロトコルは、ストレージコントローラ１５０ａおよび関連するデータストレージデバイス１５４ａが、ＳＡＮとして設定され、外部クライアント１１４または第２のストレージコントローラ１５２ｂと通信可能なＩｎｆｉｎｉｂａｎｄなどとして存在する。別の実施例においては、第１のストレージコントローラ１５２ａは、イーサネット（登録商標）を使用して、外部クライアント１１４または第２のストレージコントローラ１５２ｂと通信可能である。

【００６３】

ストレージコントローラ１５２は、内部ストレージコントローラ１５２またはクライアント１１４ａとバスを介して通信できる。例えば、ストレージコントローラ１５２は、ＰＣＩエクスプレス入力／出力ビジュアリゼーション（「ＰＣＩｅ－ＩＯＶ」）をサポートできるＰＣＩ－ｅを用いてバスを介して通信できる。他の新規（ｅｍｅｒｇｉｎｇ）バスプロトコルによって、システムバスがコンピュータまたはサーバ１１２の外部に拡張可能となり、ストレージコントローラ１５２ａをＳＡＮとして設定可能になる。このようなバスプロトコルの１つは、ＰＣＩｅ－ＡＳである。本発明は、単純なＳＡＮプロトコルに限定されないが、ストレージ要求に応答する新規ネットワークおよびバスプロトコルの利点を得ることができる。クライアント１１４または外部ストレージコントローラ１５２ｂの形態の外部デバイスは、拡張システムバスまたはコンピュータネットワーク１１６を介して通信できる．本明細書で使用されるように、ストレージ要求は、データ書き込み、データ読み取り、データ消去、データクエリなどの要求を具え、オブジェクトデータ、メタデータおよび管理要求ならびにブロックデータ要求を具えることができる。

【００６４】

従来のサーバ１１２は、通常、サーバ１１２内のデバイスへのアクセスを制御するルートコンプレックスを有する。通常、サーバ１１２のルートコンプレックスは、ネットワークインタフェース１５６を有し、ネットワークインタフェース１５６を介したあらゆる通信は、サーバ１１２によって制御される。しかしながら、インサーバＳＡＮ装置の好適な実施例において、ストレージコントローラ１５２は、独立してネットワークインタフェース１５６にアクセス可能であり、クライアント１１４は、ＳＡＮを形成する第１のサーバ１１２ａ内の１又はそれ以上のストレージコントローラ１５２ａと直接通信することができるか、または、１又はそれ以上の第１のストレージコントローラ１５２ａは、第２のストレージコントローラ１５２ｂまたはその他のリモートストレージコントローラ１５２と互いにネットワーク接続されＳＡＮを形成する。好適な実施例においては、第１のサーバ１１２ａから遠隔にあるデバイスは、１つの共有されたネットワークアドレスを介して第１のサーバ１１２ａまたは第１のストレージコントローラ１５２ａにアクセスできる。一実施例においては、インサーバＳＡＮ装置は、ネットワークインタフェース１５６、ストレージコントローラ１５２、およびサーバ１１２を構成する共通インタフェースモジュール１６６を具え、サーバ１１２およびストレージコントローラ１５２は、共有されたネットワークアドレスを用いてアクセス可能である。

【００６５】

他の実施例においては、サーバ１１２は、２またはそれ以上のネットワークインタフェ

ース１５６を具える。例えば、サーバ１１２は、１つのネットワークインタフェース１５６を介して通信可能であり、ストレージデバイス１５０は、別のインタフェースを介して通信できる。別の例においては、サーバ１１２は、各々がネットーワークインタフェースを有する複数のストレージデバイス１５０を具える。当分野の当業者であれば、１又はそれ以上のストレージデバイス１５０および１又はそれ以上のネットワークインタフェース１５６を有するサーバ１１２のその他の設定であって、１又はそれ以上のストレージデバイス１５０がサーバ１１２から独立してネットワークインタフェース１５６にアクセスすることを理解されたい。当分野の当業者であれば、どのように各種設定が、ネットワーク冗長性をサポートし、有用性を改善するために拡張可能であることを理解されたい。

【００６６】

有利なことに、インサーバＳＡＮ装置は、複雑性と、従来のＳＡＮの費用を大幅に省いている。例えば、一般的なＳＡＮは、外部ストレージコントローラ１５２および関連したデータストレージデバイス１５４を有するサーバ１１２を必要とする。このことは、ラックの更なる空間を塞ぎ、ケーブルやスイッチなどを要求する。ケーブル、スイッチング、従来のＳＡＮを設定するのに必要なその他のオーバヘッドは、空間を必要とし、帯域幅を低下させ、費用がかかる。インサーバＳＡＮ装置によって、ストレージコントローラ１５２および関連ストレージ１５４は、サーバ１１２フォームファクタにおいて適合し、従って、必要な空間を低減し費用を減少させる。さらに、インサーバＳＡＮによって、内部および外部の高速データバスを介して比較的に高速な通信を使用した接続が可能になる。

【００６７】

一実施例においては、ストレージデバイス１５０は、ソリッドステートストレージデバイス１０２であり、ストレージコントローラ１５２はソリッドステートストレージコントローラ１０４であり、データストレージデバイス１５４はソリッドステートストレージ１１０である。本明細書で記載するようにソリッドステートストレージデバイス１０２のスピードにより、この実施例は有利である。さらに、ソリッドステートストレージデバイス１０２は、サーバ１１２内で適合可能であり必要とする空間の小さいＤＩＭＭ内で設定可能である。

【００６８】

サーバ１１２内の１又はそれ以上の内部クライアント１１４ａは、さらに、サーバのネットインタフェース１５６を介してコンピュータネットワーク１１６に接続可能であり、クライアントの接続は、通常、サーバ１１２によって制御される。これにはいくつかの利点がある。クライアント１１４ａは、直接ストレージデバイス１５０に、ローカルまたはリモートでアクセスでき、メモリクライアント１１４ａとストレージデバイス１５０との間においてローカルまたはリモートダイレクトメモリアクセス（「ＤＭＡ」、「ＲＤＭＡ」）データ移動を開始できる。

【００６９】

別の実施例においては、サーバ１１２の内部または外部にあるクライアント１１４、１１４ａは、１又はそれ以上のネットワーク１１６を介して、クライアント１１４に対するファイルサーバとして機能し、インサーバＳＡＮ、外部ＳＡＮおよびハイブリッドＳＡＮの一部として関与する、ローカルで接続されたソリッドステートストレージデバイス１０２を使用する。ストレージデバイス１５０は、ＤＡＳ、インサーバＳＡＮ、ＳＡＮ、ＮＡＳなど、および、これらの組み合わせに同時に関与することができる。さらに、各ストレージデバイスをパーティション化することで、第１のパーティションが、ＤＡＳとしてストレージデバイス１５０を利用可能にし、第２のパーティションが、インサーバＳＡＮの構成要素としてストレージデバイス１５０を利用可能にし、第３のパーティションがＮＡＳとしてストレージパーティション１５０を利用可能にし、第４のパーティションがＳＡＮなどの構成要素としてストレージデバイス１５０を利用可能にする。同様に、ストレージデバイス１５０は、安全およびアクセス制御要求に矛盾がないようにパーティション化できる。当分野の当業者であれば、ストレージデバイス、仮想ストレージデバイス、ストレージネットワーク、仮想ストレージネットワーク、プライベートストレージ、共有され

たストレージ、パラレルファイルシステム、パラレルオブジェクトファイルシステム、ブロックストレージデバイス、オブジェクトストレージデバイス、ストレージアプライアンス、ネットワークアプライアンスなどの、多数の組み合わせおよび順列が、構築およびサポートされることを理解されたい。

【００７０】

さらに、コンピュータネットワーク１１６に直接接続されることによって、ストレージデバイス１５０は、互いに通信可能であり、インサーバＳＡＮとして機能することができる。サーバ１１２内のクライアント１１４ａと、コンピュータネットワーク１１６を介して接続されたクライアント１１４は、ＳＡＮとしてのストレージデバイス１５０にアクセスできる。ストレージデバイス１５０をサーバ１１２に移動し、ＳＡＮとしてストレージデバイス１５０を設定する能力を得ることで、サーバ１１２／ストレージデバイス１５０の組み合わせは、専用のストレージコントローラ、ファイバチャネルネットワークおよび他の設備のための従来のＳＡＮの必要性を排除する。インサーバＳＡＮシステム１０３は、ストレージデバイス１５０が電力、冷房、管理および物理的空間などの共通リソースをクライアント１１４およびコンピュータ１１２と共有することができる利点を有する。例えば、ストレージデバイス１５０は、サーバ１１２の空スロットを満たし、ＳＡＮまたはＮＡＳの実行能力、信頼性および有用性を全てに付与する。当分野の当業者であれば、インサーバＳＡＮシステム１０３のその他の特徴および利点を理解されたい。

【００７１】

他の設定においては、複数のインサーバＳＡＮストレージデバイス１５０ａを１つのサーバ１１２ａインフラストラクチャ内に配置することができる。一実施例においては、サーバ１１２ａは、外部ネットワークインタフェース１５６、外部クライアント１１４、１１４ｂまたは外部ストレージデバイス１５０ｂなしで、ＰＣＩエクスプレスＩＯＶを用いて連結された１又はそれ以上の内部ブレード化サーバクライアント１１４ａからなる。

【００７２】

さらに、１又はそれ以上のコンピュータネットワーク１１６を介して、インサーバＳＡＮストレージデバイス１５０は、ピア（ｐｅｅｒ）ストレージデバイス１５０と通信可能であり、これらのピアストレージデバイス１５０は、コンピュータ１１２に位置しているか（図１Ａ）、または、ＳＡＮおよびインサーバＳＡＮの両方の全ての能力を有するハイブリッドＳＡＮを形成するコンピュータ１１２なしで、コンピュータネットワーク１１６に直接接続されている。このフレキシビリティは、各種の可能性のあるソリッドステートストレージネットワークインプリメンテーションの間で拡張性および移動を単純化する利点を有する。当分野の当業者であれば、その他の組み合わせ、設定、インプリメンテーション、およびソリッドステートコントローラ１０４を配置および相互接続する構造を理解されたい。

【００７３】

ネットワークインタフェース１５６ａは、サーバ１１２ａ内で機能する１つのエージェントのみによって制御可能であり、前記エージェント内で機能するリンク設定モジュール１６８は、ネットワークインタフェース１５６ａから外部ストレージデバイス１５０ｂおよびクライアント１１４、１１４ｂを介して、内部クライアント１１４ａとストレージデバイス１５０ａ／第１のストレージコントローラ１５２ａとの間の通信パスをセットアップ可能である。好適な実施例においては、一旦通信パスが構築されると、個々の内部ストレージデバイス１５０ａおよび内部クライアント１１４ａは、自身のコマンドキューを構築し管理することができ、直接、および、ネットワークインタフェース１５６ａを制御するプロキシまたはエージェントから独立したＲＤＭＡを介して、いずれかの方向で、ネットワークインタフェース１５６ａから外部ストレージデバイス１５０ｂおよびクライアント１１４、１１４ｂを介し、コマンド及びデータの両方を転送することができる。一実施例においては、リンク設定モジュール１６８は、ハードウェアのスタートアップまたは初期化など、初期化プロセス中に通信リンクを構築する。

【００７４】

別の実施例においては、プロキシモジュール１７０は、第１のサーバ１１２ａを介してストレージ要求に応答するのに使用されるコマンドの少なくとも１部を指令し、少なくともデータ、および、ストレージ要求と関連したその他の可能性のあるコマンドは、第１のサーバから独立して、第１のコントローラと外部ストレージデバイスとの間で通信される。別の実施例においては、ストレージデバイス１５０ａおよびクライアント１１４ａに変わって、プロキシモジュール１７０がコマンド又はデータを送る。

【００７５】

一実施例においては、第１のサーバ１１２ａは、第１のサーバ１１２ａ内の１又はそれ以上のサーバを具え、仮想バスモジュール１７２を具え、これによって、第１のサーバ１１２ａ内の１又はそれ以上のサーバが、別の仮想バスを介して、１又はそれ以上のストレージコントローラ１５２ａに独立してアクセス可能となる。ＩＯＶをサポートするネットワークインタフェース１５６ａによって、１又はそれ以上のサーバおよび１又はそれ以上のストレージコントローラは、１又はそれ以上のネットワークインタフェース１５６ａを独立して制御可能になる。

【００７６】

各種実施例においては、インサーバＳＡＮ装置によって、２又はそれ以上のストレージデバイス１５０は、ＲＡＩＤ内に設定可能である。一実施例においては、インサーバＳＡＮ装置は、ＲＡＩＤとして２又はそれ以上のストレージコントローラ１５２を構成するフロントエンドＲＡＩＤモジュール１７４を具える。クライアント１１４、１１４ａからのストレージ要求がデータを記憶する要求を具える場合、フロントエンドＲＡＩＤモジュール１７４は、特定の実装ＲＡＩＤレベルと一致するＲＡＩＤにデータを書き込むことによってストレージ要求に応答する。第２のストレージコントローラ１５２は、第１のサーバ１１２ａ内または第１のサーバ１１２ａの外部に配置可能である。フロントエンドＲＡＩＤモジュール１７４によって、ストレージコントローラ１５２のＲＡＩＤが許容され、ストレージコントローラ１５２は、ストレージ要求を送るクライアント１１４、１１４ａに対してビジブル（ｖｉｓｉｂｌｅ）である。これによって、ストライプおよびパリティ情報は、マスタとして設計されたストレージコントローラ１５２またはクライアント１１４、１１４ａによって管理可能となる。

【００７７】

別の実施例においては、インサーバＳＡＮ装置は、バックエンドＲＡＩＤモジュール１７６を具え、このモジュール１７６は、ＲＡＩＤとしてストレージコントローラによって制御される２又はそれ以上のデータストレージデバイス１５４を設定する。クライアントからのストレージ要求が、データを記憶する要求を具えている場合、バックエンドＲＡＩＤモジュール１７６は、実装されたＲＡＩＤレベルと一致するＲＡＩＤにデータを書き込むことによってストレージ要求に応答し、ＲＡＩＤとして設定されたストレージデバイス１５４は、第１のストレージコントローラ１５２によって制御される１つのデータストレージデバイス１５４としてクライアント１１４、１１４ａによってアクセスされる。このＲＡＩＤ実装によって、ＲＡＩＤがデータストレージデバイス１５４にアクセスするあらゆるクライアント１１４、１１４ａに透明（ｔｒａｎｓｐａｒｅｎｔ）であるように、ストレージコントローラ１５２によって制御されるデータストレージデバイス１５４のＲＡＩＤが可能となる。別の実施例においては、フロントエンドＲＡＩＤとバックエンドＲＡＩＤの両方が実装され、複数レベルのＲＡＩＤを有する。当分野の当業者であれば、本明細書に記載されるように、ソリッドステートストレージコントローラ１０４と関連するソリッドステートストレージ１１０と一致するストレージデバイス１５２をＲＡＩＤするその他の方法を理解されたい。

【００７８】

ストレージコントローラ管理オブジェクト用装置

【００７９】

図２Ａは、本発明によるストレージデバイス内のオブジェクト管理のための装置２００の実施例を示す概略的ブロック図である。この装置２００は、オブジェクト要求受信モジ

ュール２６０、構文解析モジュール２６２、コマンド実行モジュール２６４、オブジェクトインデックスモジュール２６６、オブジェクト要求待ち行列モジュール２６８、メッセージモジュール２７０を有するパケッタイザ３０２、および、オブジェクトインデックス復元モジュール２７２を含むストレージコントローラ１５２を具え、これらを以下に示す。

【００８０】

ストレージコントローラ１５２は、図１Ｂのシステム１０２に関連して記載されたストレージコントローラ１５２と実質的に同様で、図２と関連して記載されたソリッドステートストレージデバイスコントローラ２０２にできる。この装置２００は、１又はそれ以上の要求デバイス１５５からオブジェクト要求を受信するオブジェクト要求受信モジュール２６０を具える。例えば、記憶オブジェクトデータ要求のために、ストレージコントローラ１５２は、ストレージコントローラ１５２に接続したデータストレージデバイス１５４内のデータパケットとしてデータセグメントを記憶する。オブジェクト要求は、一般的に、ストレージコントローラによって管理されたオブジェクト用の１又はそれ以上のオブジェクトデータパケットに記憶された、または、記憶されるべきデータセグメントで指令される。オブジェクト要求はオブジェクトを生成し、ストレージコントローラ１５２が、ローカルまたはリモートダイレクトメモリアクセス（「ＤＭＡ」、「ＲＤＭＡ」）転送を用いることができる、後のオブジェクト要求を通したデータで後に充填することを要求できる。

【００８１】

一実施例においては、オブジェクト要求は、既に生成されたオブジェクトにオブジェクトの総てまたは一部を書き込む、書き込み要求である。一例においては、書き込み要求は、オブジェクトのデータセグメント用である。オブジェクトのその他のデータセグメントは、ストレージデバイス１５０またはその他のストレージデバイス１５２に書き込まれてもよい。別の例においては、書き込み要求は、完全なオブジェクト用である。別の例においては、オブジェクト要求は、ストレージコントローラ１５２によって管理されたデータセグメントからのデータを読み出すことである。さらなる別の実施例においては、オブジェクト要求は、データセグメント又はオブジェクトを削除する削除要求である。

【００８２】

有利には、ストレージコントローラ１５２は、既存のオブジェクトに新規オブジェクトを書き込むか、または、データを加える以上の書き込み要求を受入可能である。例えば、オブジェクト要求受信モジュール２６０が受信する書き込み要求は、ストレージコントローラ１５２によって記録したデータの前にデータを加え、記憶したデータにデータを挿入し、または、データのセグメントを置換する要求を具えることができる。ストレージコントローラ１５２によって維持されるオブジェクトインデックスは、その他のストレージコントローラで利用可能ではないが、サーバや他のコンピュータのファイルシステムのストレージコントローラの外側でのみ現在利用可能である、これらの複雑な書き込み動作に要求されるフレキシビリティを提供する。

【００８３】

装置２００は、オブジェクト要求を１又はそれ以上のコマンドに構文解析する構文解析モジュール２６２を具える。一般的に、構文解析モジュール２６２は、オブジェクト要求を１又はそれ以上のバッファに構文解析する。例えば、オブジェクト要求の１又はそれ以上のコマンドは、コマンドバッファに構文解析できる。一般的に、構文解析モジュール２６２は、オブジェクト要求を調整し、オブジェクト要求の情報は、ストレージコントローラ１５２によって理解され実行される。当分野の当業者は、１又はそれ以上のコマンドにオブジェクト要求を構文解析する構文解析モジュール２６２のその他の機能が分かるであろう。

【００８４】

装置２００は、オブジェクト要求から構文解析されたコマンドを実行するコマンド実行モジュール２６４を具える。一実施例においては、コマンド実行モジュール２６４は、１

つのコマンドを実行する。別の実施例においては、コマンド実行モジュール２６４は、複数のコマンドを実行する。一般的に、コマンド実行モジュール２６４は、書き込みコマンドのようなオブジェクト要求から構文解析されたコマンドを解釈し、次いで、サブコマンドを生成して、待ち行列に入れて実行する。例えば、オブジェクトから構文解析された書き込みコマンドは、ストレージコントローラ１５２に命令し、複数のデータセグメントを記憶する。更に、オブジェクト要求は、暗号化、圧縮など必要な属性を具えていてもよい。コマンド実行モジュール２６４は、ストレージコントローラ１５２に命令して、データセグメントを圧縮し、データセグメントを暗号化し、１又はそれ以上のデータパケットおよび各データパケット用の関連するヘッダを生成し、メディア暗号化キーでデータパケットを暗号化し、エラー修正コードを加え、そして、特定の位置にデータパケットを記憶することができる。特定の位置およびその他のサブコマンドのデータパケットを記憶することは、その他のより低位のサブコマンドに分解される。当分野の当業者は、コマンド実行モジュール２６４が、オブジェクト要求から構文解析される１又はそれ以上のコマンドを実行可能であるその他の方法が分かるであろう。

【００８５】

装置２００は、オブジェクトインデックスモジュール２６６を具え、このモジュール２６６は、オブジェクトを生成するか、または、オブジェクトのデータセグメントを記憶するストレージコントローラ１５２に応答して、オブジェクトインデックスのオブジェクトエントリを生成する。一般的に、ストレージコントローラ１５２は、データセグメントからデータパケットを生成し、データパケットが記憶された位置は、データセグメントが記憶されたときに、割り当てられる。データセグメントとともに、又はオブジェクト要求の一部として受信されたオブジェクトメタデータは、同様の方法で記憶できる。

【００８６】

オブジェクトインデックスモジュール２６６は、データパケットが記憶され、データパケットの物理アドレスが割り当てられるときに、オブジェクトインデックスにオブジェクトエントリを生成する。オブジェクトエントリは、オブジェクトの論理識別子と、ストレージコントローラ１５２が１又はそれ以上のデータパケットおよびあらゆるオブジェクトメタデータパケットを記憶したところに対応する１又はそれ以上の物理アドレスと、の間におけるマッピングを具える。別の実施例においては、オブジェクトインデックスのエントリは、オブジェクトのデータパケットが記憶される前に生成される。例えば、データパケットが早期に記憶すべき物理アドレスをストレージコントローラ１５２が決定する場合、オブジェクトインデックスモジュール２６６は、オブジェクトインデックス内のエントリを早期に生成できる。

【００８７】

一般的には、オブジェクト要求またはオブジェクト要求のグループが、読み出し－変更－書き込み動作中、オブジェクトまたはデータセグメントが変更されることが生じた場合、オブジェクトインデックスモジュール２６６は、変更したオブジェクトに対応するオブジェクトインデックス内のエントリを更新する。一実施例においては、オブジェクトインデックスは、変更されたオブジェクト用のオブジェクトインデックス中に、新規オブジェクトおよび新規エントリを生成する。一般的に、オブジェクトの一部のみが変更された場合、オブジェクトは、変更されたデータパケットおよび未変化のいくつかのデータパケットを具える。この場合、新規エントリは、もともと書き込まれていた未変化のデータパケットと、新規位置に書き込まれた変更されたオブジェクトと、に対するマッピングを具える。

【００８８】

別の実施例においては、オブジェクト要求受信モジュール２６０は、データブロックまたはその他のオブジェクトエレメントを消去するコマンドを具えたオブジェクト要求を受信した場合、ストレージコントローラ１５２は、オブジェクトに対する基準、オブジェクトとの関係、および、消去したデータブロックのサイズを含む情報を含む消去パケットのような、少なくとも１つのパケットを記憶することができる。更に、それは、更に、消去

オブジェクトエレメントがゼロで充填されていることを示してもよい。従って、消去オブジェクト要求は、消去され、メモリ／ストレージのセル中でゼロを用いて実際に記憶された適宜なメモリ／ストレージの一部を実際に有している、実際のメモリまたはストレージをエミュレートするのに用いることができる。

【００８９】

有利には、データセグメントとオブジェクトのメタデータとの間のマッピングを示すエントリを有するオブジェクトインデックスを生成することで、ストレージコントローラ１５２が、オブジェクトを自律的に処理して管理するのを可能にする。この能力によって、ストレージデバイス１５０内にデータを記憶する大きなフレキシビリティを可能にする。オブジェクトのインデックスエントリが一度生成されると、オブジェクトに関する次のオブジェクト要求が、ストレージコントローラ１５２によって効率的に供給される。

【００９０】

一実施例においては、ストレージコントローラ１５２は、構文解析モジュール２６２によって構文解析する前に、オブジェクト要求受信モジュール２６０によって受信される１又はそれ以上のオブジェクト要求を待ち行列に入れるオブジェクト要求待ち行列モジュールを具える。オブジェクト要求待ち行列モジュール２６８は、オブジェクト要求が受信されるときと、オブジェクト要求が実行されるときとの間においてフレキシビリティを可能にする。

【００９１】

別の実施例においては、ストレージコントローラ１５２は、１又はそれ以上のデータセグメントから１又はそれ以上のデータパケットを生成するパケッタイザ３０２を具え、ここで、データパケットがデータストレージデバイス１５４内のストレージ用にサイズ調整される。パケッタイザ３０２は、図３でより詳細に以下に示す。パケッタイザ３０２は、一実施例においては、各パケットのヘッダを生成するメッセージモジュール２７０を具える。このヘッダは、パケット識別子とパケット長を具える。パケット識別子は、パケットが形成されたオブジェクトにパケットを関連させる。

【００９２】

一実施例においては、各パケットは、パケット識別子がオブジェクトと、パケット内に含まれるオブジェクトエレメントのオブジェクト内の関係性を識別するのに十分な情報を含むという点で独立型のパケット識別子を含む。しかしながら、より効率的な好ましい実施例は、コンテナにパケットを記憶することである。

【００９３】

コンテナは、より効率的なパケットのストレージを容易にし、オブジェクトとデータパケットの間の関係、メタデータパケット、コンテナ内に記憶されるオブジェクトに関連したその他のパケットとを構築するのを助けるデータ構造である。ストレージコントローラ１５２は、一般的に、同様の方法で、オブジェクトおよびデータセグメントの一部として受信されるオブジェクトメタデータを処理することに留意されたい。一般的に、「パケット」は、データを含むデータパケット、メタデータを含むメタデータパケット、または、別のパケットタイプの別のパケットを意味する。オブジェクトは、１又はそれ以上のコンテナに記憶でき、コンテナは、一般的に、１程度の固有オブジェクト用のパケットを具える。オブジェクトは、複数のコンテナ間に配置される。一般的に、コンテナは、単一の論理消去ブロック（ストレージ部分）内に記憶され、論理消去ブロック間を一般的に分割しない。

【００９４】

一例においては、コンテナは、２又はそれ以上の論理／仮想ページ間で分割されてもよい。コンテナは、オブジェクトにコンテナを関連させるコンテナラベルによって識別される。コンテナは、多くのパケットに対してゼロを含み、一般的に、コンテナ内のパケットは、１つのオブジェクトからなる。パケットは、オブジェクト属性エレメント、オブジェクトデータエレメント、オブジェクトインデックスエレメントなど、多数のオブジェクトエレメントタイプにできる。２以上のオブジェクトエレメントタイプを含むハイブリッド

パケットを生成できる。各パケットは、同一エレメントタイプの多数のエレメントに対してゼロを含む。コンテナ内の各パケットは、一般的に、オブジェクトとの関連性を識別する固有の識別子を含む。

【００９５】

各パケットは、１のコンテナに関連している。好ましい実施例においては、コンテナは、消去ブロックに限定されず、各消去ブロックの開始点またはその付近で、コンテナパケットを見つけることができる。このことは、破損したパケットヘッダを有する消去ブロックに対するデータ損失を制限することを支援する。この実施例においては、オブジェクトインデックスが使用不能であり、消去ブロック内のパケットヘッダが破損している場合、パケットヘッダから消去ブロックエンドまでの内容は、次のパケットの位置を決定する信頼性のある機構がない可能性が高いので、損失してしまう。別の実施例においては、より信頼性のある方法は、ページ境界に限定されるコンテナを有することである。この実施例は、より多くのヘッダオーバヘッドを必要とする。別の実施例においては、コンテナは、ページ及び消去ブロック境界にわたり動くことができる。これは、必要とするヘッダオーバヘッドは少なくなるが、パケットヘッダが破損している場合、データのより大きな部分が損失される。これらのいくつかの実施例に関して、いくつかのタイプのＲＡＩＤが、さらなるデータの完全性を保証するために用いられることが予測される。

【００９６】

一実施例においては、装置２００は、データストレージデバイス１５４に記憶されるパケットヘッダからの情報を用いて、オブジェクトインデックスのエントリを復元するオブジェクトインデックス復元モジュール２７２を具える。一実施例においては、オブジェクトインデックス復元モジュール２７２は、各パケットが属するオブジェクトを判定するヘッダと、オブジェクトにおけるデータまたはメタデータが属する場所を判定する配列情報とを読み出すことによって、オブジェクトインデックスのエントリを復元する。オブジェクトインデックス復元モジュール２７２は、各パケットおよびタイムスタンプの物理アドレス情報、または、パケットの物理的位置とオブジェクト識別子とデータセグメント配列との間のマッピングを生成する配列情報を用いる。タイムスタンプまたは配列情報はオブジェクトインデックス復元モジュール２７２によって用いられて、インデックスに対して行われる配列の変化を再生し、これによって、一般的に最新状態を復元する。

【００９７】

別の実施例においては、オブジェクトインデックス復元モジュール２７２は、パケットの物理位置と、オブジェクト識別子と、各パケットのシーケンス番号と、を識別して、オブジェクトインデックス内のエントリを復元するコンテナパケット情報とともにパケットヘッダ情報を用いて、パケットを配置する。一実施例においては、消去ブロックは、タイムスタンプされるか、または、パケットが書き込まれるときにシーケンス番号が付与され、タイムスタンプまたは消去ブロックの配列情報は、コンテナヘッダおよびパケットヘッダから集められた情報とともに用いられ、オブジェクトインデックスを復元する。別の実施例においては、消去ブロックが回復されるときに、タイムスタンプまたは配列情報が書き込まれる。

【００９８】

オブジェクトインデックスが揮発性メモリに記憶されるとき、エラー、電力損失、または、ストレージコントローラ１５２がオブジェクトインデックスを保存せずにシャットダウンさせる問題が、オブジェクトインデックスが復元できない場合に、問題になるであろう。オブジェクトインデックス復元モジュール２７２によって、オブジェクトインデックスが揮発性メモリに記憶され、高速アクセスなど揮発性メモリの利点を得ることが可能になる。オブジェクトインデックス復元モジュール２７２によって、ストレージデバイス１５０の外部にあるデバイスに依存せず自律的に、オブジェクトインデックスの迅速な復元が可能となる。

【００９９】

一実施例においては、揮発性メモリのオブジェクトインデックスは、データストレージ

デバイス１５４に一時的に記憶される。特定の実施例においては、オブジェクトインデックスまたは「インデックスメタデータ」は、ソリッドステートストレージ１１０に定期的に記憶される。別の実施例においては、インデックスメタデータが、パケットを記憶するソリッドステートストレージ１１０ａ－１１０ｎ－１と別個のソリッドステートストレージ１１０ｎに記憶される。インデックスメタデータは、データおよび要求デバイス１５５から送信されるオブジェクトメタデータから別個に管理され、ストレージコントローラ１５２／ソリッドステートストレージデバイスコントローラ２０２によって管理される。オブジェクトからの他のデータおよびメタデータと別個にインデックスを管理して記憶することによって、ストレージコントローラ１５２／ソリッドステートストレージデバイスコントローラ２０２がオブジェクトメタデータを不必要に処理することなく、効率的データフローが可能となる。

【０１００】

一実施例においては、オブジェクト要求受信モジュール２６０によって受信されるオブジェクト要求が書き込み要求を具える場合、ストレージコントローラ１５２は、ローカルまたはリモートダイレクトメモリアクセス（「ＤＭＡ」、「ＲＤＭＡ」）動作として、要求デバイス１５５のメモリからオブジェクトの１又はそれ以上のデータセグメントを受信する。好適な例において、ストレージコントローラ１５２は、１又はそれ以上のＤＭＡまたはＲＤＭＡ動作において要求デバイス１５５のメモリからデータを引き出す。別の例においては、要求デバイス１５５は、１又はそれ以上のＤＭＡまたはＲＤＭＡ動作において、データセグメントをストレージコントローラ１５２へ押し出す。別の実施例においては、オブジェクト要求は読み出し要求を具え、ストレージコントローラ１５２は、１又はそれ以上のＤＭＡまたはＲＤＭＡ動作において、オブジェクトの１又はそれ以上のデータセグメントを要求デバイス１５５のメモリに送信する。好適な例においては、ストレージコントローラ１５２は、１又はそれ以上のＤＭＡまたはＲＤＭＡ動作において、データを要求デバイス１５５のメモリに押し出す。別の実施例においては、要求デバイス１５５は、１又はそれ以上のＤＭＡまたはＲＤＭＡ動作において、ストレージコントローラ１５２からデータを引き出す。別の例においては、ストレージコントローラ１５２は、１又はそれ以上のＤＭＡまたはＲＤＭＡ動作において、要求デバイス１５５のメモリからオブジェクトコマンド要求セットを引き出す。別の実施例においては、要求デバイス１５５は、１又はそれ以上のＤＭＡまたはＲＤＭＡ動作において、ストレージコントローラ１５２にオブジェクトコマンド要求セットを押し出す。

【０１０１】

一実施例においては、ストレージコントローラ１５２は、ブロックストレージをエミュレートし、要求デバイス１５５とストレージコントローラ１５２との間で通信されるオブジェクトは、１又はそれ以上のデータブロックを具える。一実施例においては、要求デバイス１５５は、ドライバを具え、ストレージデバイス１５０はブロックストレージデバイスとして存在する。例えば、要求デバイス１５２は、要求デバイス１５５が記憶されたデータブロックを所望する位置の物理アドレスとともに所定サイズのデータのブロックを送信することができる。ストレージコントローラ１５２は、データブロックを受信し、データブロックで送信される物理ブロックアドレスを用いるか、または、オブジェクト識別子としての物理ブロックアドレスの変換を用いる。次いで、ストレージコントローラ１５２は、データブロックをパケット化してデータブロックを任意に記憶することによって、オブジェクトまたはオブジェクトのデータセグメントとしてデータブロックを記憶する。次いで、オブジェクトインデックスモジュール２６６は、物理的ブロックベースのオブジェクト識別子と、ストレージコントローラ１５２がデータブロックのデータを具えるデータパケットを記憶した実際の物理位置とを用いて、オブジェクトインデックス内にエントリを生成する。

【０１０２】

別の実施例においては、ストレージコントローラ１５２は、ブロックオブジェクトを受け入れることでブロックストレージをエミュレートする。ブロックオブジェクトは、ブロ

ック構造中に１又はそれ以上のデータブロックを具えることができる。一実施例においては、ストレージコントローラ１５２は、その他のオブジェクトとしてブロックオブジェクトを処理する。別の実施例においては、オブジェクトは、完全なブロックデバイス、ブロックデバイスのパーティション、又は、トラック、セクタ、チャネルなどを含むブロックデバイスの他の論理若しくは物理サブエレメントを表す。特筆すべきことは、ブロックデバイスＲＡＩＤグループを、進行型ＲＡＩＤなどの異なるＲＡＩＤ構造をサポートするオブジェクトに再マッピングする能力である。当分野の当業者は、従来または将来のブロックデバイスをオブジェクトへの他のマッピングが分かるであろう。

【０１０３】

ソリッドステートストレージデバイス

【０１０４】

図２Ｂは、本発明によるソリッドステートデバイス１０２内に書き込みデータパイプライン１０６および読み出しデータパイプライン１０８を具えるソリッドステートデバイスコントローラ２０２の一実施例２０１を示す概略ブロック図である。ソリッドステートストレージデバイスコントローラ２０２は、各々がソリッドステートストレージ１１０を制御する多数のソリッドステートストレージコントローラ０－Ｎ１０４ａ－ｎを具えることができる。示した実施例においては、２つのソリッドステートコントローラが示され、ソリッドステートコントローラ０　１０４ａおよびソリッドステートストレージコントローラＮ　１０４ｎは、それぞれ、ソリッドステートストレージ１１０ａ－ｎを制御する。示された実施例においては、ソリッドステートストレージコントローラ０　１０４ａは、データチャネルを制御し、接続したソリッドステートストレージ１１０ａはデータを記憶する。ソリッドステートストレージコントローラＮ　１０４ｎは、記憶したデータと関連するインデックスメタデータチャネルを制御し、関連するソリッドステートストレージ１１０ｎはインデックスメタデータを記憶する。代替の実施例においては、ソリッドステートストレージデバイスコントローラ２０２は、単一のソリッドステートストレージ１１０ａを有する単一のソリッドステートコントローラ１０４ａを具える。別の実施例においては、複数のソリッドステートストレージコントローラ１０４ａ－ｎおよび関連するソリッドステートストレージ１１０ａ－ｎがある。一実施例においては、付随のソリッドステートストレージ１１０ａ－１１０ｎ－１に接続した１又はそれ以上のソリッドステートコントローラ１０４ａ－１０４ｎ－１は、データを制御し、付随のソリッドステートストレージ１１０ｎに接続したソリッドステートストレージコントローラ１０４ｎは、インデックスメタデータを制御する。

【０１０５】

一実施例においては、少なくとも１つのソリッドステートコントローラ１０４は、フィールドプログラマブルゲートアレイ（「ＦＰＧＡ」）であり、コントローラ機能は、ＦＰＧＡにプログラムされる。特定の実施例においては、ＦＰＧＡはＸｉｌｉｎｘ（登録商標）ＦＰＧＡである。別の実施例においては、ソリッドステートストレージコントローラ１０４は、特定用途向け集積回路（「ＡＳＩＣ」）またはカスタム論理回路ソリューションのような特別に設計された構成を具える。各ソリッドステートストレージコントローラ１０４は、一般的に、書き込みデータパイプライン１０６および読み出しデータパイプライン１０８を具えており、これらは、図３で更に説明する。別の実施例においては、少なくとも１つのソリッドステートストレージコントローラ１０４は、結合型ＦＰＧＡ、ＡＳＩＣおよびカスタム論理回路構成から構成される。

【０１０６】

ソリッドステートストレージ

【０１０７】

ソリッドステートストレージ１１０は、バンク２１４内で構成され、双方向ストレージ入力／出力（「Ｉ／Ｏ」）バス２１０を通って並列にアクセスできる、非揮発性ソリッドステートストレージエレメント２１６、２１８、２２０のアレイである。一実施例においては、ストレージＩ／Ｏバス２１０は、どの場合でも単方向通信が可能である。例えば、

データがソリッドステートストレージ１１０に書き込まれるとき、データは、ソリッドステートストレージ１１０から読み出すことはできない。別の実施例においては、データは、双方向に同時に流れうる。なお、データバスに関して本明細書で用いられるように、双方向が意味するのは、一度に一方向のみにしかデータフローを有さないが、双方向データバス上で一方向のデータフローが停止したときは、データは、双方向データバス上の逆方向にデータをフローすることができるデータ経路である。

【０１０８】

ソリッドステートストレージエレメント（例えば、ＳＳＳ　０．０　２１６ａ）は、一般的に回路基板上のチップ（１又はそれ以上のダイのパッケージ）またはダイとして構成される。示されているように、これらの複数のエレメントが、チップパッケージ、チップパッケージのスタックまたはその他のパッケージエレメントに一緒にパッケージされている場合であっても、ソリッドステートストレージエレメント（例えば、２１６ａ）は、他のソリッドストレージエレメント（例えば、２１８ａ）から独立または準独立的に動作する。示されているように、ソリッドステートストレージ２１６、２１８、２２０の段は、バンク２１４として指定される。示されているように、ソリッドステートストレージ１１０内のｎ×ｍのソリッドステートストレージエレメント２１６、２１８、２２０のアレイにおいて、バンクあたり「ｎ」バンク２１４ａ－ｎおよび「ｍ」ソリッドステートストレージエレメント２１６ａ－ｍ、２１８ａ－ｍ、２２０ａ－ｍが存在していてもよい。一実施例においては、ソリッドステートストレージ１１０ａは、８つのバンク２１４を有するバンク２１４あたり２０のソリッドステートストレージエレメント２１６、２１８、２２０を具え、ソリッドステートストレージ１１０ｎは、１つのバンク２１４を有するバンク２１４あたり、２つのソリッドステートストレージエレメント２１６、２１８を具える。一実施例においては、各ソリッドステートストレージエレメント２１６、２１８、２２０は、シングルレベルセル（「ＳＬＣ」）デバイスからなっている。別の実施例においては、各ソリッドステートストレージエレメント２１６、２１８、２２０は、マルチレベルセル（「ＭＬＣ」）デバイスからなっている。

【０１０９】

一実施例においては、共通ストレージＩ／Ｏバス２１０ａ列（例えば、２１６ｂ、２１８ｂ、２２０ｂ）を共有する複数のバンク用のソリッドステートストレージエレメントは、一緒にパッケージされる。一実施例においては、ソリッドステートストレージエレメント２１６、２１８、２２０は、チップあたり１又はそれ以上のダイを有することができ、１又はそれ以上のチップは、垂直方向にスタックされ、各ダイは別個にアクセス可能である。別の実施例においては、ソリッドステートストレージエレメント（例えば、ＳＳＳ０．０　２１６ａ）は、ダイあたり１又はそれ以上の仮想ダイと、チップあたり１又はそれ以上のダイと、垂直方向にスタックされた１又はそれ以上のチップと、を有し、各仮想ダイは別個にアクセス可能である。別の実施例においては、ソリッドステートストレージエレメントＳＳＳ０．０　２１６ａは、ダイあたり１又はそれ以上の仮想ダイと、チップあたり１又はそれ以上のダイと、を有し、１又はそれ以上のダイの一部または全部が垂直方向にスタックされ、各仮想ダイは別個にアクセス可能である。

【０１１０】

一実施例においては、２つのダイが、グループあたり４つのスタックとともに垂直方向にスタックされ、８つのストレージエレメント（例えば、ＳＳＳ０．０－ＳＳＳ０．８）２１６ａ－２２０ａを形成し、各々は、別のバンク２１４ａ－ｎにある。別の実施例においては、２０のストレージエレメント（例えば、ＳＳＳ０．０－ＳＳＳ２０．０）２１６は、仮想バンク２１４ａを形成し、８の仮想バンクの各々は、２０のストレージエレメント（例えば、ＳＳＳ０．０－ＳＳＳ２０．８）２１６、２１８、２２０を有する。データは、ストレージエレメント（ＳＳＳ０．０－ＳＳＳ０．８）２１６ａ、２１８ａ、２２０ａの特定のグループの総てのストレージエレメントへのストレージＩ／Ｏバス２１０にわたって、ソリッドステートストレージ１１０に送信される。ストレージ制御バス２１２ａは、特定のバンク（例えば、バンク－０　２１４ａ）を選択するために用いられ、総ての

バンク２１４に接続されたストレージＩ／Ｏバス２１０にわたって受信されるデータは、選択されたバンク２１４ａに書き込まれる。

【０１１１】

好ましい実施例においては、ストレージＩ／Ｏバス２１０は、１又はそれ以上の独立Ｉ／Ｏバスからなり（２１０ａ．ａ－ｍ、２１０ｎ．ａ－ｍを具える「ＩＩＯＢａ－ｍ」）、各列内のソリッドステートストレージエレメントは、各ソリッドステートストレージエレメント２１６、２１８、２２０に並行してアクセスする独立Ｉ／Ｏバスの１つを共有するので、総てのバンク２１４は、同時にアクセスされる。例えば、ストレージＩ／Ｏバス２１０の１つのチャネルは、各バンク２１４ａ－ｎの第１のソリッドステートストレージエレメント２１６ａ、２１８ａ、２２０ａに同時にアクセスすることができる。ストレージＩ／Ｏバス２１０の第２のチャネルは、各バンク２１４ａ－ｎの第２のソリッドステートストレージエレメント２１６ｂ、２１８ｂ、２２０ｂに同時にアクセスできる。ソリッドステートストレージ２１６、２１８、２２０の各列は、同時にアクセスされる。一実施例においては、ソリッドステートストレージエレメント２１６、２１８、２２０がマルチレベルで（物理的にスタックされている）ある場合、ソリッドステートストレージエレメント２１６、２１８、２２０の総ての物理レベルは、同時にアクセスされる。本明細書で用いられるように、「同時に」は、ほぼ同時のアクセスを含み、デバイスは、スイッチングノイズを回避するために、わずかに異なる間隔でアクセスされる。「同時に」はこのような構成で用いられて、連続または直列アクセスと区別し、コマンドおよび／またはデータは、交互に別個に送信される。

【０１１２】

一般的に、バンク２１４ａ－ｎは、ストレージ制御バス２１２を用いて別個に選択される。一実施例においては、バンク２１４は、使用可能チップまたは選択チップを用いて選択される。選択チップおよび使用可能チップの両方が使用可能な場合、ストレージ制御バス２１２は、マルチレベルのソリッドステートストレージエレメント２１６、２１８、２２０の１つのレベルを選択することができる。その他の実施例においては、ストレージ制御バス２１２によって他のコマンドが用いられて、マルチレベルのソリッドステートストレージエレメント２１６、２１８、２２０の１つのレベルを別個に選択する。ソリッドステートストレージエレメント２１６、２１８、２２０は、ストレージＩ／Ｏバス２１０およびストレージ制御バス２１２上で送信される制御及びアドレス情報との組み合わせを通じて選択されてもよい。

【０１１３】

一実施例においては、各ソリッドステートストレージエレメント２１６、２１８、２２０は、消去ブロックに分割され、各消去ブロックは、ページに分割される。一般的なページは、２０００バイト（「２ｋＢ」）である。一例においては、ソリッドステートストレージエレメント（例えば、ＳＳＳ０．０）は、２つのレジスタを具え、２つのページをプログラムして、２つのレジスタのソリッドステートストレージエレメント２１６、２１８、２２０が、４ｋＢの容量を有することができる。更に、２０のソリッドステートストレージエレメント２１６、２１８、２２０のバンク２１４は、ストレージＩ／Ｏバス２１０のチャネルを通る同一アドレスでアクセスされる８０ｋＢのページ容量を有する。

【０１１４】

８０ｋＢのソリッドステートストレージエレメント２１６、２１８、２２０のバンク２１４における上記ページのグループは、仮想ページと呼ばれる。同様にバンク２１４ａの各ストレージエレメント２１６ａ－ｍの消去ブロックは、仮想消去ブロックを形成するようにグループ化されてもよい。好ましい実施例においては、ソリッドステートストレージエレメント２１６、２１８、２２０内のページの消去ブロックは、消去コマンドがソリッドステートストレージエレメント２１６、２１８、２２０内で受信されたときに消去される。消去ブロック、ページ、プレーンまたはソリッドステートストレージエレメント２１６、２１８、２２０内のその他の論理および物理部分のサイズおよび数が、技術の進歩に伴い、時宜にわたって変化することが予想されるが、新規構成に一致する多数の実施例が

、可能であり、本明細書の一般的説明と一致すると分かるであろう。

【０１１５】

一般的に、パケットがソリッドステートストレージエレメント２１６、２１８、２２０内の特定位置に書き込まれ、このパケットが特定バンクの特定エレメントの特定消去ブロックに特有である、特定ページ内の位置に書き込まれるように意図されている場合、物理アドレスは、ストレージＩ／Ｏバス２１０上で送信され、パケットがそれに続く。物理アドレスは、パケットをページ内の指定位置へと配向するための、ソリッドステートストレージエレメント２１６、２１８、２２０の情報を十分に保有している。ストレージエレメント（例えば、ＳＳＳ０．０－ＳＳＳ０．Ｎ　２１６ａ、２１８ａ、２２０ａ）の列における総てのストレージエレメントは、ストレージＩ／Ｏバス２１０ａ．ａ内の適宜なバスによって同時にアクセスされ、適宜なページに到達し、ストレージエレメント（ＳＳＳ０．０－ＳＳＳ０．Ｎ　２１６ａ、２１８ａ、２２０ａ）の列において同様にアドレスされたページに、データパケットを書き込むことを回避するため、データパケットが書き込まれるべき正規のページを有するソリッドステートストレージエレメントＳＳＳ０．０　２１６ａを具えるバンク２１４ａは、ストレージ制御バス２１２によって同時に選択される。

【０１１６】

同様に、ストレージＩ／Ｏバス２１２上を移動する読み出しコマンドは、単一のバンク２１４ａと、このバンク２１４ａ内の適宜なページとを選択するために、ストレージ制御バス２１２上の同時コマンドを要求する。好ましい実施例においては、読み出しコマンドはページ全体を読み出し、バンク２１４においては並行して複数のソリッドステートストレージエレメント２１６、２１８、２２０があるため仮想ページ全体は読み出しコマンドを用いて読み出される。しかしながら、読み出しコマンドは、バンクインタリーブに関して以下で説明するように、サブコマンドに分割可能である。更に、仮想ページは、書き込み動作でアクセス可能である。

【０１１７】

消去ブロック消去コマンドは、特定の消去ブロックを消去する特定の消去ブロックアドレスを用いて、ストレージＩ／Ｏバス２１０にわたって、消去ブロックを消去するように送信できる。一般的に、消去ブロック消去コマンドは、仮想消去ブロックを消去するために、ストレージＩ／Ｏバス２１０の並列な経路にわたり送信でき、各々が特定の消去ブロックを消去する特定の消去ブロックアドレスを有する。同時に、特定バンク（例えば、バンク－０　２１４ａ）がストレージ制御バス２１２で選択され、総てのバンク（バンク１－Ｎ　２１４ｂ－ｎ）における同様にアドレスされたブロックが消去されるのを防止する。更に、その他のコマンドは、ストレージＩ／Ｏバス２１０とストレージ制御バス２１２との組み合わせを用いて特定位置に送信できる。当分野の当業者は、双方向ストレージＩ／Ｏバス２１０およびストレージ制御バス２１２を用いて、特定のストレージ位置を選択できるその他の方法が分かるであろう。

【０１１８】

一実施例においては、パケットは、ソリッドステートストレージ１１０に連続に書き込まれる。例えば、パケットは、ストレージエレメント２１６のバンク２１４ａのストレージ書き込みバッファへとストリーミングされ、バッファが一杯の場合、パケットは、指定の仮想ページにプログラムされる。次いで、パケットは、ストレージ書き込みバッファを補充し、一杯の場合、パケットは、次の仮想ページに書き込まれる。次の仮想ページは、同一バンク２１４ａまたは別のバンク（例えば、２１４ｂ）内にあってもよい。このプロセスは、一般的に、仮想消去ブロックが満たされるまで、仮想ページごとに繰り返される。別の実施例においては、ストリーミングは、仮想消去ブロック境界にわたって継続し、このプロセスは、仮想消去ブロックごとに継続する。

【０１１９】

読み出し、変更、書き込み動作において、オブジェクトに関連するデータパケットは、配置され、読み出し動作で読み出される。変更されたオブジェクトのデータセグメントは

、データセグメントが読み出される位置に書き込まれない。その代わり、変更されたデータセグメントは、データパケットに再度変換され、次いで、そのとき書き込まれている仮想ページの次の利用可能位置に書き込まれる。それぞれのデータパケットのオブジェクトインデックスエントリは、変更されたデータセグメントを含むパケットに対し指示するように変更される。変更されていない同一オブジェクトに関連したデータパケットのオブジェクトインデックスの１又はそれ以上のエントリは、未変更データパケットの元の位置に対するポインタを具える。従って、例えば、オブジェクトの前バージョンを維持するために、元のオブジェクトを維持している場合、元のオブジェクトは、もともと書き込まれていた総てのデータパケットに対するオブジェクトインデックスのポインタを有する。この新規オブジェクトは、元のデータパケットに対するオブジェクトインデックスのポインタと、その時に書き込まれている仮想ページにおける変更データパケットに対するポインタとを有する。

【０１２０】

コピー動作において、オブジェクトインデックスは、ソリッドステートストレージ１１０に記憶された多数のパケットにマッピングされた元のオブジェクトのエントリを具える。コピーが生成されるとき、新規オブジェクトが生成され、新規エントリが、元のパケットに新規オブジェクトをマッピングするオブジェクトインデックスに生成される。新規オブジェクトは、ソリッドステートストレージ１１０に書き込まれ、その位置は、オブジェクトインデックスの新規エントリにマッピングされる。新規オブジェクトパケットは、コピーに伝搬されていない元のオブジェクトで変更が生じ、オブジェクトインデックスが損失または破損した場合に、参照される元のオブジェクト内のパケットを識別するために用いられる。

【０１２１】

有益なことに、パケットを連続に書き込むことは、ソリッドステートストレージ１１０のより均等な利用を容易にし、ソリッドストレージデバイスコントローラ２０２がソリッドステートストレージ１１０内のストレージホットスポットおよび様々な仮想ページの利用レベルをモニタすることを可能にする。更に、パケットを連続に書き込むことは、下記で詳細に示すように、強力で有効なガベージコレクションシステムを容易にする。当分野の当業者は、データパケットの連続ストレージのその他の利点が分かるであろう。

【０１２２】

ソリッドステートストレージデバイスコントローラ

【０１２３】

様々な実施例においては、ソリッドステートストレージデバイスコントローラ２０２は、データバス２０４、ローカルバス２０６、バッファコントローラ２０８、バッファ０－Ｎ　２２２ａ－ｎ、マスタコントローラ２２４、ダイレクトメモリアクセス（「ＤＭＡ」）コントローラ２２６、メモリコントローラ２２８、ダイナミックメモリアレイ２３０、スタティックランダムメモリアレイ２３２、管理コントローラ２３４、管理バス２３６、システムバス２４０へのブリッジ２３８、および、種々雑多な論理回路２４２を更に具え、以下にこれを示す。他の実施例においては、システムバス２４０は、１又はそれ以上のネットワークインタフェースカード（「ＮＩＣｓ」）２４４に結合し、そのうちのいくつかは、リモートＤＭＡ（「ＲＤＭＡ」）コントローラ２４６、１又はそれ以上の中央処理装置（「ＣＰＵ」）２４８、１又はそれ以上の外部メモリコントローラ２５０および関連する外部メモリアレイ２５２、１又はそれ以上のストレージコントローラ２５４、ピアコントローラ２５６およびアプリケーション特有のプロセッサ２５８を具え、これらは以下で説明する。システムバス２４０に接続された構成２４４－２５８は、コンピュータ１１２に配置されてもよく、その他のデバイスであってもよい。

【０１２４】

一般的に、ソリッドステートストレージコントローラ１０４は、ストレージＩ／Ｏバス２１０を介してソリッドステートストレージ１１０とデータを通信する。ソリッドステートストレージが、バンク２１４内に配列され、各々のバンク２１４が、並列にアクセスさ

れる複数のストレージエレメント２１６、２１８、２２０を具える一般的な実施例においては、ストレージＩ／Ｏバス２１０は、バスのアレイであり、ストレージエレメント２１６、２１８、２２０の各列の１つは、バンク２１４に及んでいる。本明細書で用いられるように、用語「ストレージＩ／Ｏバス」は、１つのストレージＩ／Ｏバス２１０、または、データ独立性バス２０４のアレイを意味する。好ましい実施例においては、ストレージエレメントの列（例えば、２１６ａ、２１８ａ、２２０ａ）にアクセスする各ストレージＩ／Ｏバス２１０は、ストレージエレメントの列２１６ａ、２１８ａ、２２０ａにアクセスされるストレージ部分（例えば、消去ブロック）の論理－物理マッピングを具えることができる。このマッピングによって、第１のストレージ部分が故障している場合、部分的に故障している場合、アクセス不能な場合、または、その他何らかの問題がある場合に、ストレージ部分の物理アドレスにマッピングされる論理アドレスが、異なるストレージ部分に再マッピングできる。再マッピングは、図３の再マッピングモジュール３１４に関連して更に説明する。

【０１２５】

データは、システムバス２４０、ブリッジ２３８、ローカルバス２０６、バッファ２２を介して、また、最終的にデータバス２０４を介して、要求デバイス１５５からソリッドステートストレージコントローラ１０４と通信可能である。データバス２０４は、一般的に、バッファコントローラ２０８を用いて制御される１又はそれ以上のバッファ２２２ａ－ｎに接続される。バッファコントローラ２０８は一般的に、ローカルバス２０６からバッファ２２２への、並びに、データバス２０４を通ってパイプライン入力バッファ３０６および出力バッファ３３０への、データの転送を制御する。一般的に、バッファコントローラ２２２は、どのように、要求デバイスからきたデータがバッファ２２２に一時的に記憶可能で、次いで、データバス２０４上を転送されるかを制御でき、あるいは、その逆も同様に制御でき、異なるクロックドメインにし、データの衝突を防ぐこと等ができる。バッファコントローラ２０８は、一般的にマスタコントローラ２２４とともに作動し、データフローに調和させる。データが到着するときは、データは、システムバス２４０の上に到着し、ブリッジ２３８を介してローカルバス２０６に転送される。

【０１２６】

一般的に、データは、マスタコントローラ２２４およびバッファコントローラ２０８によって指示されるように、ローカルバス２０６から１又はそれ以上のデータバッファ２２２へと転送される。次いで、データは、ソリッドステートコントローラ１０４を介してバッファ２２２からデータバス２０４へと、及び、ＮＡＮＤフラッシュまたはその他のストレージメディアなどのようなソリッドステートストレージ１１０へと流れる。好ましい実施例においては、データと、このデータとともにくる関連したバンド外のメタデータ（「オブジェクトメタデータ」）は、１又はそれ以上のソリッドステートストレージコントローラ１０４ａ－１０４ｎ－１および関連するソリッドステートストレージ１１０ａ－１１０ｎ－１を具える１又はそれ以上のデータチャンネルを用いて通信され、一方、少なくとも１つのチャネル（ソリッドステートストレージコントローラ１０４ｎ、ソリッドステートストレージ１１０ｎ）は、インデックス情報や、ソリッドステートストレージデバイス１０２に内部生成されるその他のメタデータなどのような、バンド内のメタデータ専用である。

【０１２７】

一般的に、ローカルバス２０６は、双方向バスであるか、または、ソリッドステートストレージデバイスコントローラ２０２の内部のデバイス間、および、ソリッドステートストレージデバイス１０２の内部のデバイスと、システムバス２４０に接続したデバイス２４４－２５８との間において、データおよびコマンドの通信を可能にするバスのセットである。ブリッジ２３８は、ローカルバス２０６とシステムバス２４０との間の通信を容易にする。当分野の当業者は、リング構造またはスイッチ式スター構成、ならびに、バス２４０、２０６、２０４、２１０およびブリッジ２３８の機能などのような、その他の実施例が分かるであろう。

【０１２８】
一般的に、システムバス２４０は、コンピュータ１１２のバスであり、または、ソリッドステートストレージデバイス１０２がインストールまたは接続されているその他のデバイスである。一実施例においては、システムバス２４０は、ＰＣＩ－ｅバス、シリアルアドバンステクノロジーアタッチメント（「シリアルＡＴＡ」）バス、パラレルＡＴＡなどでもよい。別の実施例においては、システムバス２４０は、小型コンピュータシステムインタフェース（「ＳＣＳＩ」）、ＦｉｒｅＷｉｒｅ、ファイバチャネル、ＵＳＢ、ＰＣＩｅ－ＡＳなどのような、外部バスである。ソリッドステートストレージデバイス１０２は、デバイスの内側に適用するように、または、外側に接続されたデバイスとして、パッケージされてもよい。

【０１２９】
ソリッドステートストレージデバイスコントローラ２０２は、ソリッドステートストレージ１０２内のより高位の機能を制御するマスタコントローラ２２４を具える。マスタコントローラ２２４は、様々な実施例において、オブジェクト要求およびその他の要求を解釈することによってデータフローを制御し、インデックスの生成を命令し、データに関連したオブジェクト識別子を、関連するデータ、協調するＤＭＡ要求等の物理的な位置にマッピングする。本明細書に記載される機能の多くは、マスタコントローラ２２４によって完全に、又は部分的に制御される。

【０１３０】
一実施例においては、マスタコントローラ２２４は、埋込式コントローラを用いる。別の実施例においては、マスタコントローラ２２４は、ダイナミックメモリアレイ２３０（ダイナミックランダムアクセスメモリ「ＤＲＡＭ」）、スタティックメモリアレイ３２３（スタティックランダムアクセスメモリ「ＳＲＡＭ」）など、ローカルメモリを用いる。一実施例においては、ローカルメモリは、マスタコントローラ２２４を用いて制御される。別の実施例においては、マスタコントローラは、メモリコントローラ２２８を介してローカルメモリにアクセスする。別の実施例においては、マスタコントローラは、Ｌｉｎｕｘサーバを運用し、ＷｏｒｌｄＷｉｄｅＷｅｂ、ハイパーテキストマークアップ言語（「ＨＴＭＬ」）などのような、様々な共通サーバインタフェースをサポートすることができる。別の実施例においては、マスタコントローラ２２４は、ナノプロセッサを用いる。マスタコントローラ２２４は、プログラマブルな若しくは標準の論理回路、または、上記に挙げたコントローラタイプの組み合わせ、を用いて構成されていてもよい。当分野の当業者は、マスタコントローラの多数の実施例が分かるであろう。

【０１３１】
一実施例においては、ストレージデバイス１５２／ソリッドステートストレージデバイスコントローラ２０２は、複数のデータストレージデバイス／ソリッドステートストレージ１１０ａ－ｎを管理する場合、マスタコントローラ２２４は、ソリッドステートストレージコントローラ１０４ａ－ｎのような、内部コントローラ中間の作業負荷を分割する。例えば、マスタコントローラ２２４は、データストレージデバイス（例えば、ソリッドステートストレージ１１０ａ－ｎ）に書き込まれるオブジェクトを分割し、オブジェクトの一部は、接続したデータストレージデバイスの各々に記憶できる。この特徴は、オブジェクトへのより迅速なストレージおよびアクセスを可能にする性能向上である。一実施例においては、マスタコントローラ２２４は、ＦＰＧＡを用いて実装される。別の実施例においては、マスタコントローラ２２４内のファームウェアは、管理バス２３６、ＮＩＣ２４４に接続したネットワークを介したシステムバス２４０、または、システムバス２４０に接続したその他のデバイスを介して更新可能である。

【０１３２】
一実施例においては、オブジェクトを管理するマスタコントローラ２２４は、ブロックストレージをエミュレートし、コンピュータ１０２、または、ストレージデバイス１５２／ソリッドステートストレージ１０２に接続したその他のデバイスは、ストレージデバイス１５２／ソリッドステートストレージデバイス１０２をブロックストレージデバイスと

みなし、ストレージデバイス１５２／ソリッドステートストレージデバイス１０２内の特定の物理アドレスにデータを送信する。次いで、マスタコントローラ２２４は、ブロックを分割し、それをオブジェクトとしてデータブロックに記憶する。更に、マスタコントローラ２２４は、ブロックと、このブロックとともに送信される物理アドレスとを、マスタコントローラ２２４によって決定される実際の位置にマッピングする。このマッピングは、オブジェクトインデックスに記憶される。一般的に、ブロックエミュレーションのために、ブロックデバイスアプリケーションプログラムインタフェース（「ＡＰＩ」）は、コンピュータ１１２、クライアント１１４、または、ブロックストレージデバイスとしてストレージデバイス１５２／ソリッドステートストレージデバイス１０２を用いることを望むその他のデバイス内のドライバに提供される。

【０１３３】

別の実施例においては、マスタコントローラ２２４は、ＮＩＣコントローラ２４４および埋込型ＲＤＭＡコントローラ２４６で調整し、ジャストインタイムのデータおよびコマンドセットのＲＤＭＡ転送を配信する。ＮＩＣコントローラ２４４は、非透過性のポートの背後に隠れ、カスタムドライバを用いることができる。更に、クライアント１１４のドライバは、標準スタックＡＰＩを使用し、ＮＩＣ２４４とともに機能するＩ／Ｏメモリドライバを介してコンピュータネットワーク１１８にアクセスできる。

【０１３４】

一実施例においては、マスタコントローラ２２４も、独立ドライブ冗長アレイ（「ＲＡＩＤ」）コントローラである。データストレージデバイス／ソリッドステートストレージデバイス１０２が、１又はそれ以上の他のデータストレージデバイス／ソリッドステートストレージデバイス１０２とネットワーク接続される場合、マスタコントローラ２２４は、単階層型ＲＡＩＤ、多階層型ＲＡＩＤ、進行型ＲＡＩＤ等用のＲＡＩＤコントローラでもよい。更に、マスタコントローラ２２４によって、いくつかのオブジェクトはＲＡＩＤアレイに記憶でき、他のオブジェクトはＲＡＩＤなしに記憶できる。別の実施例においては、マスタコントローラ２２４は、分散型ＲＡＩＤコントローラエレメントでもよい。別の実施例においては、マスタコントローラ２２４は、多数のＲＡＩＤ、分散型ＲＡＩＤ、他に記載されるその他の機能を具えることができる。

【０１３５】

一実施例においては、マスタコントローラ２２４は、単一または冗長ネットワークマネージャ（例えば、スイッチ）を調整して、ルーティングを構築し、帯域幅使用率、フェイルオーバ等のバランスをとる。別の実施例においては、マスタコントローラ２２４は、統合型アプリケーション特有の論理回路（ローカルバス２０６を介して）と、関連するドライバソフトウェアを調整する。別の実施例においては、マスタコントローラ２２４は、接続されたアプリケーション特有のプロセッサ２５８または論理回路（外部システムバス２４０）および関連するドライバソフトウェアを調整する。別の実施例においては、マスタコントローラ２２４は、リモートのアプリケーション特有の論理回路（コンピュータネットワーク１１８を介して）および関連するドライバソフトウェアを調整する。別の実施例においては、マスタコントローラ２２４は、ローカルバス２０６または外部バス接続ハードディスクドライブ（「ＨＤＤ」）ストレージコントローラを調整する。

【０１３６】

一実施例においては、マスタコントローラ２２４は、１又はそれ以上のストレージコントローラ２５４と通信し、ストレージデバイス／ソリッドステートストレージデバイス１０２は、ＳＣＳＩバス、インターネットＳＣＳＩ（「ｉＳＣＳＩ」）、ファイバチャネル等を介して接続されたストレージデバイスとして存在してもよい。その間、ストレージデバイス／ソリッドステートストレージデバイス１０２は、オブジェクトを自律的に管理することができ、オブジェクトファイルシステムまたは分散型オブジェクトファイルシステムとして存在してもよい。マスタコントローラ２２４は、ピアコントローラ２５６および／またはアプリケーション特有のプロセッサ２５８によってアクセスすることができる。

【０１３７】

別の実施例においては、マスタコントローラ２２４は、自律的に統合された管理コントローラを調整し、ＦＰＧＡコードおよび／またはコントローラソフトウェアを定期的に確認し、（リセット）を実行中にＦＰＧＡコードを確認し、および／または、（リセット）出力中にコントローラソフトウェアを確認し、外部リセット要求をサポートし、監視タイムアウトによるリセット要求をサポートし、電圧、電流、電力、温度及びその他の環境測定値をサポートし、閾値の設定が中断する。別の実施例においては、マスタコントローラ２２４は、ガベージコレクションを管理し、再利用のための消去ブロックを含まない。別の実施例においては、マスタコントローラ２２４は、摩耗レベルを管理する。別の実施例においては、マスタコントローラ２２４によって、データストレージデバイス／ソリッドステートストレージデバイス１０２が、複数の仮想デバイスに分割され、パーティションベースメディアの暗号化が可能となる。さらなる別の実施例においては、マスタコントローラ２２４は、有利な複数ビットＥＣＣ修正部を有するソリッドステートストレージコントローラ１０４をサポートする。当分野の当業者は、ストレージコントローラ１５２において、または、より具体的には、ソリッドステートストレージデバイス１０２において、マスタコントローラ２２４のその他の特性および機能が分かるであろう。

【０１３８】

一実施例においては、ソリッドステートストレージデバイスコントローラ２０２は、ダイナミックランダムメモリアレイ２３０および／またはスタティックランダムメモリアレイ２３２を制御するメモリコントローラ２２８を具える。上述したように、メモリコントローラ２２８は、独立しているか、または、マスタコントローラ２２４と統合されている。一般的に、メモリコントローラ２２８は、ＤＲＡＭ（ダイナミックランダムメモリアレイ２３０）およびＳＲＡＭ（スタティックランダムメモリアレイ２３２）などのような、いくつかのタイプの揮発メモリを制御する。他の例においては、メモリコントローラ２２８は、電気的消去可能なプログラム可能読み出し専用メモリ（「ＥＥＰＲＯＭ」）などのような、その他のメモリタイプを制御できる。他の実施例においては、メモリコントローラ２２８は、２又はそれ以上のメモリタイプを制御し、メモリコントローラ２２８は、１以上のコントローラを具えてもよい。一般的に、メモリコントローラ２２８は、ＳＲＡＭ２３２を補助するＤＲＡＭ２３０による、実行可能な限りの多くのＳＲＡＭ２３２制御する。

【０１３９】

一実施例においては、オブジェクトインデックスは、メモリ２３０、２３２内に記憶され、次いで、ソリッドステートストレージ１１０ｎのチャネルまたはその他の非揮発性メモリに定期的にオフロードされる。当分野の当業者は、ダイナミックメモリアレイ２３０、メモリコントローラ２２８およびスタティックメモリアレイ２３２のその他の使用および設定が分かるであろう。

【０１４０】

一実施例においては、ソリッドステートストレージデバイスコントローラ２０２はＤＭＡコントローラ２２６を具え、このＤＭＡコントローラ２２６は、ストレージデバイス／ソリッドステートストレージデバイス１０２および１又はそれ以上の外部メモリコントローラ、関連する外部メモリアレイ２５２およびＣＰＵ２４８間のＤＭＡ動作を制御する。なお、外部メモリコントローラ２５０および関連する外部メモリアレイ２５２は、ストレージデバイス／ソリッドステートストレージデバイス１０２の外部にあるので外部と呼ぶことに留意されたい。更に、ＤＭＡコントローラ２２６も、ＮＩＣ２４４を介した要求デバイスおよび関連するＲＤＭＡコントローラ２４６を用いてＲＤＭＡ動作を制御することができる。ＤＭＡおよびＲＤＭＡは、以下でより詳細に説明する。

【０１４１】

一実施例においては、ソリッドステートストレージデバイスコントローラ２０２は、管理バス２３６に接続された管理コントローラ２３４を具える。一般的に、管理コントローラ２３４は、環境基準値およびストレージデバイス／ソリッドステートストレージデバイス１０２のステータスを管理する。管理コントローラ２３４は、管理バス２３６を介して

デバイス温度、ファン速度、電力供給設定などをモニタすることができる。管理コントローラは、ＦＰＧＡコードおよびコントローラソフトウェアのストレージ用の、消去可能なプログラム可能読み出し専用メモリ（「ＥＥＰＲＯＭ」）の読み出しおよびプログラミングをサポートすることができる。一般的に、管理バス２３６は、ストレージデバイス／ソリッドステートストレージデバイス１０２内に様々な構成を接続させる。管理コントローラ２３４は、ローカルバス２０６を介して、アラート、中断などと通信でき、システムバス２４０または他のバスに対対する別の接続を具えることができる。一実施例においては、管理バス２３６は、Ｉｎｔｅｒ－Ｉｎｔｅｇｒａｔｅｄ Ｃｉｒｃｕｉｔ（Ｉ２Ｃ）バスである。当分野の当業者は、管理バス２３６によってストレージデバイス／ソリッドステートストレージデバイス１０２の構成に接続された、管理コントローラ２３４のその他の関連する機能および利用が分かるであろう。

【０１４２】

一実施例においては、ソリッドステートストレージデバイスコントローラ２０２は、特定のアプリケーションにカスタマイズ可能な種々雑多な論理回路２４２を具える。一般的に、ソリッドステートデバイスコントローラ２０２またはマスタコントローラ２２４が、ＦＰＧＡまたはその他の構成可能コントローラを用いて構成される場合、特定のアプリケーション、カスタマ要求、ストレージ要求に基づく、カスタム論理回路が含まれる。

【０１４３】

データパイプライン

【０１４４】

図３は、本発明によるソリッドステートストレージデバイス１０２における書き込みデータパイプライン１０６および読み出しデータパイプライン１０８を有する、ソリッドステートストレージコントローラ１０４の一実施例３００を示した概略ブロック図である。実施例３００は、データバス２０４、ローカルバス２０６およびバッファコントローラ２０８を具え、これらは、図２のソリッドステートストレージデバイスコントローラ２０２に関連して記載されたものと実質的に同等である。書き込みデータパイプラインは、パケッタイザ３０２およびエラー修正コード（「ＥＣＣ」）ジェネレータ３０４を具える。他の実施例においては、書き込みデータパイプラインは、入力バッファ３０６、書き込み同期バッファ３０８、書き込みプログラムモジュール３１０、圧縮モジュール３１２、暗号化モジュール３１４、ガベージコレクタバイパス３１６（読み出しデータパイプライン内の一部を有する）、メディア暗号化モジュール３１８、および、書き込みバッファ３２０を具える。読み出しデータパイプライン１０８は、読み出し同期バッファ３２８、ＥＣＣ修正モジュール３２２、デパケッタイザ（ｄｅｐａｃｋｅｔｉｚｅｒ）３２４、アライメントモジュール３２６および出力バッファ３３０を具える。他の実施例においては、読み出しデータパイプライン１０８は、メディア復号化モジュール３３２、ガベージコレクタバイパス３１６の一部、復号化モジュール３３４、解凍モジュール３３６および読み出しプログラムモジュール３３８を具えることができる。ソリッドステートストレージコントローラ１０４は、更に、制御およびステータスレジスタ３４０、制御待ち行列３４２、バンクインタリーブコントローラ３４４、同期バッファ３４６、ストレージバスコントローラ３４８およびマルチプレクサ（「ＭＵＸ」）３５０を具えることができる。ソリッドステートコントローラ１０４の構成および関連する書き込みデータパイプライン１０６および読み出しデータパイプライン１０８は、以下で説明する。他の実施例においては、同期ソリッドステートストレージ１１０を用いることができ、同期バッファ３０８、３２８を省くことができる。

【０１４５】

書き込みデータパイプライン

【０１４６】

書き込みデータパイプライン１０６は、直接または間接的に別の書き込みデータパイプライン１０６のステージを介して、ソリッドステートストレージに書き込まれるデータ又はメタデータセグメントを受信する、パケッタイザ３０２を具え、ソリッドステートスト

レージ１１０用に調整された１又はそれ以上のパケットを生成する。データまたはメタデータセグメントは、一般的に、オブジェクトの一部であるが、オブジェクト全体を具えることもできる。別の実施例においては、データセグメントは、データブロックの一部であるが、データブロック全体を具えることもできる。一般的に、オブジェクトは、コンピュータ１１２、クライアント１１４またはその他のコンピュータ若しくはデバイスから受信し、ソリッドステートストレージデバイス１０２またはコンピュータ１１２にストリーミングされるデータセグメントでソリッドステートストレージデバイス１０２に送信される。更に、データセグメントは、データ区域など、別の名称として周知であるが、本明細書で規定されるように、オブジェクトまたはデータブロックの総てまたは一部を具える。

【０１４７】

各オブジェクトは、１又はそれ以上のパケットとして記憶される。各オブジェクトは、１又はそれ以上のコンテナパケットを有することができる。各パケットはヘッダを含む。ヘッダは、ヘッダ型フィールドを具えることができる。タイプフィールドは、データ、オブジェクト属性、メタデータ、データセグメントデリミタ（マルチパケット）、オブジェクト構造、オブジェクト結合などを具えることができる。ヘッダは、パケットに含まれるデータのバイト数のような、パケットサイズに関する情報を具えることができる。パケットの長さは、パケットタイプによって構築可能である。ヘッダは、オブジェクトに対するパケットの関係性を構築する情報を具えることができる。一例では、オブジェクト内のデータセグメントの位置を識別するために、データパケットヘッダのオフセットを用いることができる。当分野の当業者は、パケッタイザ３０２によってデータに加えられるヘッダに含まれるその他の情報、および、データパケットに加えられるその他の情報が分かるであろう。

【０１４８】

各パケットは、データまたはデータセグメントからのヘッダおよび可能であればデータを具えることができる。各パケットのヘッダは、パケットが属するオブジェクトにパケットに関連させる、関連情報を具える。例えば、ヘッダは、データパケットを形成していたデータセグメント、オブジェクトまたはデータブロックを示すオブジェクト識別子およびオフセットを具えることができる。更に、ヘッダは、パケットを記憶するための、ストレージバスコントローラによって用いられる論理アドレスを具えることができる。また、ヘッダは、パケットに含まれるバイト数のような、パケットの大きさに関する情報を具えることができる。更に、ヘッダは、データセグメントまたはオブジェクトを復元するとき、データセグメントがオブジェクト内のその他のパケットに対して属しているところを識別するシーケンス番号を具えることができる。ヘッダは、ヘッダ型フィールドを具えることができる。タイプフィールドは、データ、オブジェクト属性、メタデータ、データセグメントデリミタ（マルチパケット）、オブジェクト構造、オブジェクト結合などを具えることができる。当分野当業者は、パケッタイザ３０２によってデータ又はメタデータに加えられるヘッダに含まれてもよいその他の情報、および、パケットに加えることができる他の情報が分かるであろう。

【０１４９】

書き込みデータパイプライン１０６は、パケッタイザ３０２から受信する１又はそれ以上のパケット用の１又はそれ以上のエラー修正コード（「ＥＣＣ」）を生成するＥＣＣジェネレータ３０４を具える。ＥＣＣジェネレータ３０４は、一般的に、エラー修正アルゴリズムを用いて、パケットとともに記憶されるＥＣＣを生成する。パケットとともに記憶されたＥＣＣは、一般的に、送信および記憶を通ったデータに入ったエラーを検出し修正するために用いられる。一実施例においては、パケットは、長さＮの符号化されないブロックとしてＥＣＣジェネレータ３０４内にストリーミングされる。長さＳのシンドロームが算出され、追加され、長さＮ＋Ｓの符号化ブロックとして出力される。Ｎ値およびＳ値は、特別な性能、効率およびロバスト性の基準値を得るために選択されるアルゴリズムの特性に依存している。好ましい実施例においては、ＥＣＣブロックとパケットとの間において、固定された関係性はなく、パケットは、１以上のＥＣＣブロックを具えることがで

き、ＥＣＣブロックは、１以上のパケットを具えることができ、第１のパケットは、ＥＣＣブロック内のどこで終結してもよく、同一のＥＣＣブロック内の第１のパケットの末端の後から第２のパケットは開始することができる。好ましい実施例においては、ＥＣＣアルゴリズムは、動的に変更されない。好ましい実施例においては、データパケットとともに記憶されたＥＣＣは、２より大きいビットにおけるエラーを修正するのに十分にロバストである。

【０１５０】

有利には、シングルビット以上の修正またはダブルビット修正でさえ可能なロバストなＥＣＣアルゴリズムを用いることで、ソリッドステートストレージ１１０の寿命を延長させることができる。例えば、ソリッドステートストレージ１１０内のストレージ媒体としてフラッシュメモリが用いられる場合、フラッシュメモリは、消去サイクルにつきエラーなしで、約１００，０００回書き込み可能である。この利用限界は、ロバストなＥＣＣアルゴリズムを用いて拡張することができる。ＥＣＣジェネレータ３０４を有し、ソリッドステートストレージデバイス１０２に搭載されたＥＣＣ修正モジュール３２２を対応させることで、ソリッドステートストレージデバイス１０２は、エラーを内部で修正することができ、シングルビット修正のような、より低ロバスト性のＥＣＣアルゴリズムを使用した場合よりも長い寿命を有することができる。しかしながら、その他の実施例においては、ＥＣＣジェネレータ３０４は、より低ロバスト性のアルゴリズムを用いることができ、シングルビットまたはダブルビットエラーを修正できる。別の実施例においては、ソリッドステートストレージデバイス１１０は、容量を増大するためにマルチレベルセル（「ＭＬＣ」）フラッシュなど信頼性が低いストレージを具えてもよく、ストレージは、よりロバストなＥＣＣアルゴリズムがなく、十分な信頼性がなくてもよい。

【０１５１】

一実施例においては、書き込みパイプライン１０６は、入力バッファ３０６を具え、この入力バッファ３０６は、ソリッドステートストレージ１１０に書き込まれるデータセグメントを受信し、パケッタイザ３０２のような、書き込みデータパイプライン１０６の次のステージ（またはより複雑な書き込みデータパイプライン１０６用のその他のステージ）が準備され、次のデータセグメントを処理されるまでに、入ってくるデータセグメントを記憶する。一般的に、入力バッファ３０６によって、データセグメントが適宜なサイズのデータバッファを用いて書き込みデータパイプライン１０６によって受信、処理される頻度の差異を許容することができる。更に、入力バッファ３０６によって、データバス２０４の動作効率を改善するために、データバス２０４は、書き込みデータパイプライン１０６によって維持できるよりも大きな頻度で、データを書き込みデータパイプライン１０６に転送できる。一般的に、書き込みデータパイプライン１０６は、入力バッファ３０６を具えていないときは、バッファリング機能は、ソリッドステートストレージ１０２内であるが書き込みパイプライン１０６の外側、ネットワークインタフェースカード（「ＮＩＣ」）内のようなコンピュータ１１２内、または、例えば、リモートダイレクトメモリアクセス（「ＲＤＡＭ」）を用いるときの別のデバイスのような、あらゆる場所で実行される。

【０１５２】

別の実施例においては、書き込みデータパイプライン１０６は、ソリッドステートストレージ１１０にパケットを書き込む前に、ＥＣＣジェネレータ３０４から受信したパケットをバッファリングする書き込み同期バッファ３０８を更に具えることができる。書き込み同期バッファ３０８は、ローカルクロックドメインとソリッドステートストレージクロックドメインとの間の境界に配置され、クロックドメイン差を考慮するためのバッファリングを提供する。他の実施例においては、同期型ソリッドステートストレージ１１０を用いることができ、同期用バッファ３０８、３２８を省くことができる。

【０１５３】

一実施例においては、書き込みデータパイプライン１０６は、更に、メディア暗号化モジュール３１８を具えることができ、このモジュール３１８は、パケッタイザ３０２から

直接または間接的に、１又はそれ以上のパケットを受信して、ＥＣＣジェネレータ３０４にパケットを送信する前に、ソリッドステートストレージデバイス１０２に固有の暗号化キーを用いて、１又はそれ以上のパケットを暗号化する。一般的に、ヘッダを含む完全なパケットは暗号化される。別の実施例においては、ヘッダは、暗号化されない。本明細書においては、暗号化キーは、ソリッドステートストレージ１１０を統合する、暗号化保護を必要とする具体例の外部で管理されるシークレット暗号化キーを意味することを理解されたい。メディア暗号化モジュール３１８および関連するメディア暗号化モジュール３３２は、ソリッドステートストレージ１１０に記憶されるデータの安全レベルを提供する。例えば、データが、メディア暗号化モジュール３１８で暗号化されるとき、ソリッドステートストレージ１１０が、異なるソリッドステートストレージコントローラ１０４、ソリッドステートストレージデバイス１０２またはコンピュータ１１２に接続される場合、ソリッドステートストレージ１１０の内容は、一般的に、大きな労力を要さずに、ソリッドステートストレージ１１０にデータを書き込む間に用いられる同一の暗号化キーを使用せずに読み出すことができない。

【０１５４】

一般的な実施例においては、ソリッドステートストレージデバイス１０２は、不揮発性ストレージの暗号化キーを記憶できず、暗号化キーへのいかなる外部からのアクセスもできない。暗号化キーは、初期化中に、ソリッドステートストレージコントローラ１０４に提供される。ソリッドステートストレージデバイス１０２は、暗号化キーとともに用いられる、非シークレット暗号化ナンスを使用および記憶できる。異なるナンスは、パケット毎に記憶できる。データセグメントは、暗号化アルゴリズムによる保護を改善するために、固有のナンスを有する複数のパケット間に分割できる。暗号化キーは、クライアント１１４、コンピュータ１１２、キーマネージャ、または、ソリッドステートストレージコントローラ１０４によって用いられる暗号化キーを管理するその他のデバイスから受信することができる。別の実施例においては、ソリッドステートストレージ１１０は、２又はそれ以上のパーティションを有し、ソリッドステートストレージコントローラ１０４は、まるで２又はそれ以上のソリッドステートストレージコントローラ１０４であるかのように動作し、各々は、ソリッドステートストレージ１１０内の単一のパーティションで動作する。この実施例においては、固有のメディア暗号化キーは、各パーティションで用いることができる。

【０１５５】

別の実施例においては、書き込みデータパイプライン１０６は、更に、パケッタイザ３０２にデータセグメントを送信する前に、直接または間接的に、入力バッファ３０６から受信されるデータまたはデータセグメントを暗号化する暗号化モジュール３１４を具えることができ、データセグメントは、データセグメントともに受信した暗号化キーを用いて暗号化される。暗号化モジュール３１４は、データを暗号化する暗号化モジュール３１８によって用いられる暗号化キーは、ソリッドステートストレージデバイス１０２内に記憶される総てのデータに共通でなくてもよいが、以下で記載するように、データセグメントを受信するとともに受信されるオブジェクトベースで変化するという点でメディア暗号化モジュール３１８とは異なる。例えば、暗号化モジュール３１８によって暗号化されるべきデータセグメント用の暗号化キーは、データセグメントとともに受信され、データセグメントが属するオブジェクトを書き込むコマンドの一部として受信される。ソリッドステートストレージデバイス１０２は、暗号化キーとともに用いられる各オブジェクトパケットにおいて、非シークレット暗号化ナンスを用いて記憶することができる。異なるナンスは、パケット毎に記憶できる。データセグメントは、暗号化アルゴリズムによる保護を改善するために、固有のナンスを有する複数のパケットに分割できる。一実施例においては、メディア暗号化モジュール３１８によって用いられるナンスは、暗号化モジュール３１４によって用いられるものと同一である。

【０１５６】

暗号化キーは、クライアント１１４、コンピュータ１１２、キーマネージャ、または、

データセグメントを暗号化するのに用いられる暗号化キーを保持したその他デバイスから受信することができる。一実施例においては、ソリッドステートストレージデバイス１０２、コンピュータ１１２、クライアント１１４、または、プライベートおよびパブリックキーを安全に転送して保護する、工業規格の方法を実行する能力を有するその他の外部エージェント、のうちの１つから、暗号化キーはソリッドステートストレージコントローラ１０４に転送される。

【０１５７】

一実施例においては、暗号化モジュール３１８は、第１のパケットを暗号化し、第１の暗号化キーはこのパケットとともに受信され、また、第２のパケットを暗号化し、第２の暗号化キーは第２のパケットとともに受信される。別の実施例においては、暗号化モジュール３１８は、第１のパケットを暗号化し、第１の暗号化キーは、このパケットとともに受信され、第２のデータパケットを暗号化せず次のステージに通過させる。有利には、ソリッドステートストレージデバイス１０２の書き込みデータパイプライン１０６に含まれる暗号化モジュール３１８によって、オブジェクトごとまたはセグメントごとのデータ暗号化が、単一のファイルシステムまたはその他のファイルシステムなしで、対応するオブジェクトまたはデータセグメントを記憶するのに用いられる異なる暗号化キーのトラックを、保持できる。各要求デバイス１５５または対応するキーマネージャは、要求デバイス１５５によって送信されるオブジェクトまたはデータセグメントのみを暗号化するのに用いられる暗号化キーを別個に管理することができる。

【０１５８】

別の実施例においては、書き込みデータパイプライン１０６は、パケッタイザ３０２にデータセグメントを送信する前に、メタデータセグメント用のデータを圧縮する圧縮モジュール３１２を具える。圧縮モジュール３１２は、一般的に、セグメントのストレージサイズを低減するのに当分野の当業者に周知な圧縮ルーチンを用いて、データまたはメタデータセグメントを圧縮する。例えば、データセグメントは、５１２のゼロの記号列を具える場合、圧縮モジュール３１２は、５１２のゼロを、５１２のゼロで示されるコード又はトークンで置換でき、コードは、５１２のゼロによって取られる空き領域よりも更に小型である。

【０１５９】

一実施例においては、圧縮モジュール３１２は、第１の圧縮ルーチンを用いて第１のセグメントを圧縮し、圧縮せずに第２のセグメントを通過させる。別の実施例においては、圧縮モジュール３１２は、第１の圧縮ルーチンを用いて第１のセグメントを圧縮し、第２の圧縮ルーチンを用いて第２のセグメントを圧縮する。ソリッドステートストレージデバイス１０２内にフレキシビリティを有することに利点があり、クライアント１１４、または、ソリッドステートストレージデバイス１０２にデータを書き込むその他のデバイスは圧縮ルーチンを各々特定でき、あるいは、一方が圧縮ルーチンを特定することができ、他方がいずれの圧縮ルーチンを特定しなくてもよい。圧縮ルーチンの選択は、オブジェクトごとのタイプまたはオブジェクトクラスベースに基づいたデフォルトの設定に従って選択されてもよい。例えば、特定のオブジェクトの第１のオブジェクトは、デフォルトの圧縮ルーチン設定を取り消すことができ、同一オブジェクトクラスおよびオブジェクトタイプの第２のオブジェクトは、デフォルトの圧縮ルーチンを用いることができ、同一オブジェクトクラスおよびオブジェクトタイプの第３のオブジェクトは、圧縮を利用しなくてもよい。

【０１６０】

一実施例においては、書き込みデータパイプライン１０６は、ガベージコレクションシステムのデータバイパスの一部として、読み出しデータパイプライン１０８からデータセグメントを受信するガベージコレクションバイパス３１６を具える。ガベージコレクションシステムは、一般的に、パケットが削除用に標識又は変更され、変更されたデータが異なる位置に記憶されるので、もはや有効でないパケットを一般的に標識する。いくつかの場所では、ガベージコレクションシステムは、ストレージの特定のセクションが回復可能

かを判定する。この判定は、利用可能なストレージ容量の不足、閾値に達した無効として標識されたデータの割合、有効なデータの統合、閾値に達し、又はデータ配置に基づき性能を改善したストレージセクションのエラー検出率等によってもよい。多くの因子は、ガベージコレクションアルゴリズムによって考慮され、ストレージセクションを回復すべきときを判定する。

【０１６１】

ストレージセクションが回復のために標識されると、一般的に、そのセクション中の有効なパケットは再配置されなければならない。ガベージコレクタバイパス３１６によって、パケットは、読み出しデータパイプライン１０８に読み出され、次いで、ソリッドステートストレージコントローラ１０４の外にルーティングされずに、書き込みデータパイプライン１０６に直接転送される。好ましい実施例においては、ガベージコレクタバイパス３１６は、ソリッドステートストレージデバイス１０２内で動作する自律型ガベージコレクタシステムの一部である。これによって、ソリッドステートストレージデバイス１０２は、データを管理することが可能となり、データは、ソリッドステートストレージ１１０にわたって系統的に分散され、性能、データ信頼性を改善し、ソリッドステートストレージ１１０のいずれか１の位置または領域の過剰使用または使用不足を回避し、ソリッドステートストレージ１１０の寿命を延長させる。

【０１６２】

ガベージコレクタバイパス３１６は、セグメントを書き込みデータパイプライン１０６に挿入し、その他のセグメントは、クライアント１１６またはその他のデバイスによって書き込まれるように調整する。示された実施例においては、ガベージコレクタバイパス３１６は、書き込みデータパイプライン１０６内のパケッタイザ３０２の前および読み出しデータパイプライン１０８内のデパケッタイザ３２４の後にあるが、読み出しおよび書き込みデータパイプライン１０６、１０８内のどこかに配置されていてもよい。ガベージコレクタバイパス３１６は、書き込みパイプライン１０６のフラッシュ中に用いて、ソリッドステートストレージ１１０内のストレージ効率を改善するために仮想ページの残りを満たし、これにより、ガベージコレクションの頻度を低減することができる。

【０１６３】

一実施例においては、書き込みデータパイプライン１０６は、有効な書き込み動作用のデータをバッファリングする書き込みバッファ３２０を具える。一般的に、書き込みバッファ３２０は、ソリッドステートストレージ１１０内の少なくとも１つの仮想ページを満たすのに十分なパケット容量を具える。これによって、書き込み動作は、中断せずに、ソリッドステートストレージ１１０にデータのページ全体を送信することができる。書き込みデータパイプライン１０６の書き込みバッファ３２０と、読み出しデータパイプライン１０８内のバッファとを、ソリッドステートストレージ１１０内のストレージ書き込みバッファと同一、または、大きい容量にサイズ変更することによって、書き込みおよび読み出しデータは、複数コマンドの代わりに、単一の書き込みコマンドが、ソリッドステートストレージ１１０にデータの仮想ページ全体を送信するように生成されるので、より効率的になる。

【０１６４】

書き込みバッファ３２０が満たされる際に、ソリッドステートストレージ１１０は、その他の読み出し動作に用いられてもよい。このことが有利であるのは、データがストレージ書き込みバッファに書き込まれ、ストレージ書き込みバッファ内に流れるデータがストールする際に、書き込みバッファがより小さいか、書き込みバッファのないその他のソリッドステートデバイスが、ソリッドステートストレージをタイアップできるからである。読み出し動作は、ストレージ書き込みバッファ全体が満たされ、プログラム化されるまで、ブロックされる。書き込みバッファのないまたは書き込みバッファが小さいシステムに対する別のアプローチは、読み出し可能にするために満たされないストレージ書き込みバッファを消去することである。更に、ページを満たすのに複数の書き込み／プログラムサイクルが必要となるので、これは非効率的である。

【０１６５】
仮想ページよりも大きくサイズ変更された書き込みバッファ３２０を有する、示された実施例について、多数のサブコマンドを具える単一の書き込みコマンドは、次いで、単一のプログラムコマンドが続き、各ソリッドステートストレージエレメント２１６、２１８、２２０におけるストレージ書き込みバッファから、各ソリッドステートストレージエレメント２１６、２１８、２２０内の指定されたページに、データのページを転送する。この技術は、データ信頼性および耐久性を低下させることで周知である部分的なページプログラミングを除去し、バッファが満たされるとき読み出しコマンドおよびその他のコマンドのための指定バンクを解放させる利点がある。

【０１６６】
一実施例においては、書き込みバッファ３２０は、ピンポンバッファであり、このバッファの片面が満たされ、次いで、適宜な時間での転送が指定され、ピンポンバッファのもう一方の面が満たされる。別の実施例においては、書き込みバッファ３２０は、データセグメントの仮想ページよりも大きい容量を有する先入れ先出し（「ＦＩＦＯ」）レジスタを具える。当分野の当業者は、データの仮想ページがデータをソリッドステートストレージ１１０に書き込む前に記憶させることができるその他の書き込みバッファ３２０構成が分かるであろう。

【０１６７】
別の実施例においては、書き込みバッファ３２０は、仮想ページよりも小さなサイズであり、情報のより少ないページが、ソリッドステートストレージ１１０内のストレージ書き込みバッファに書き込まれる。その実施例においては、書き込みデータパイプライン１０６内のストールが読み出し動作を止めるのを防ぐために、データは、１つの位置からガベージコレクションプロセスの一部としての別の位置へと移動する必要があるデータが、ガベージコレクションシステムを用いて待ち行列に入れられる。書き込みデータパイプライン１０６にけるデータストールの場合、データは、ガベージコレクタバイパス３１６を介して、書き込みバッファ３２０、次いで、ソリッドステートストレージ１１０内のストレージ書き込みバッファへと供給可能であり、データをプログラミングする前に、仮想ページのページを満たす。この方法では、書き込みデータパイプライン１０６内のデータストールは、ソリッドステートストレージデバイス１０６から読み出しをストールしない。

【０１６８】
別の実施例においては、書き込みデータパイプライン１０６は、書き込みデータパイプライン１０６内の１又はそれ以上のユーザ定義可能な機能を有する書き込みプログラムモジュール３１０を具える。書き込みプログラムモジュール３１０によって、ユーザは書き込みデータパイプライン１０６をカスタマイズすることができる。ユーザは、特定のデータ要求またはアプリケーションに基づいて書き込みデータパイプライン１０６をカスタマイズすることができる。ソリッドステーストレージコントローラ１０４はＦＰＧＡである場合、ユーザは、カスタムコマンドおよび機能を有する書き込みデータパイプライン１０６を比較的容易にプログラムすることができる。更に、ユーザは、書き込みプログラムモジュール３１０を使用し、ＡＳＩＣを用いたカスタム機能を具えるが、しかしながら、ＡＳＩＣをカスタマイズすることは、ＦＰＧＡを用いるよりも難しい。書き込みプログラムモジュール３１０は、バッファおよびバイパス機構を具えることができ、第１のデータセグメントが書き込みプログラムモジュール３１０で実行され、第２のデータセグメントは、書き込みデータパイプライン１０６を介して継続できる。別の実施例においては、書き込みプログラムモジュール３１０は、ソフトウェアを介してプログラム可能なプロセッサコアを具えることができる。

【０１６９】
書き込みプログラムモジュール３１０は、入力バッファ３０６と圧縮モジュール３１２との間に示されているが、書き込みプログラムモジュール３１０は、書き込みデータパイプライン１０６内のあらゆる位置に存在でき、様々なステージ３０２－３２０間に配置可能であることに留意されたい。更に、別個にプログラミングおよび動作される様々なステ

ージ３０２－３２０間に配置される、複数の書き込みプログラムモジュール３１０があってもよい。更に、ステージ３０２－３２０の順番は変更可能である。当分野の当業者は、特定のユーザ要求に基づいてステージ３０２－３２０の順番に対する作動可能な変更が分かるであろう。

【０１７０】

読み出しデータパイプライン

【０１７１】

読み出しデータパイプライン１０８は、要求されたパケットの各ＥＣＣブロックで記憶されるＥＣＣを用いることによって、データエラーがＥＣＣブロックにある場合に、ソリッドステートストレージ１１０から受信される要求されたパケットを判定するＥＣＣ修正モジュール３２２を具える。次いで、ＥＣＣ修正モジュール３２２は、エラーが存在し、エラーがＥＣＣを用いて修正可能である場合、要求されたパケットにおけるエラーを修正する。例えば、ＥＣＣが、６ビットのエラーを検出できるが、３ビットのエラーしか修正できない場合、ＥＣＣ修正モジュール３２２は、エラー中の３ビットまで、要求されたパケットのＥＣＣブロックを修正する。ＥＣＣ修正モジュール３２２は、エラー中のビットを、正しい１またはゼロ状態に変更することで、エラー中のビットを修正し、要求されたデータパケットがソリッドステートストレージ１１０に書き込まれ、ＥＣＣがパケット用に生成された場合と同一になる。

【０１７２】

ＥＣＣが修正できるよりも多く、要求されたパケットがエラー中のビットを含んでいると、ＥＣＣ修正モジュール３２２が判定した場合、ＥＣＣ修正モジュール３２２は、要求されたパケットの破損したＥＣＣブロック中のエラーを修正できず、中断信号を送信する。一実施例においては、ＥＣＣ修正モジュール３２２は、要求されたパケットがエラー状態にあることを示すメッセージを有する中断信号を送信する。メッセージは、ＥＣＣ修正モジュール３２２がエラーを修正できず、エラーを修正するためのＥＣＣ修正モジュール３２２の能力がないことが示される情報を含む。別の実施例においては、ＥＣＣ修正モジュール３２２は、要求されたパケットの破損ＥＣＣブロックを、中断信号および／またはメッセージとともに送信する。

【０１７３】

好ましい実施例においては、ＥＣＣ修正モジュール３２２によって修正不可能な要求されたパケットの、破損したＥＣＣブロックまたは破損したＥＣＣブロックの一部は、マスタコントローラ２２４によって読み出され、修正され、ＥＣＣ修正モジュール３２２に戻されて、読み出しデータパイプライン１０８によって更に処理される。一実施例においては、要求されたパケットのまたは破損したＥＣＣブロックの一部は、データを要求しているデバイスに送信される。要求デバイス１５５は、ＥＣＣブロックを修正することができ、バックアップまたはミラーコピーなど、別のコピーを用いてデータを置換することができ、次いで、要求されたデータパケットの置換データを用いるか、それを読み出しデータパイプライン１０８に戻すことができる。要求デバイス１５５は、エラー中の要求されたパケットのヘッダ情報を用いて、破損した要求されたパケットを置換するか、または、パケットが属するオブジェクトを置換するかを必要とするデータを識別する。別の好ましい実施例においては、ソリッドステートストレージコントローラ１０４は、いくつかのタイプのＲＡＩＤを用いてデータを記憶し、破損したデータを回復することができる。別の実施例においては、ＥＣＣ修正モジュール３２２は、中断信号および／またはメッセージを送信し、受信するデバイスは、要求されたデータパケットに関連した読み出し動作ができない。当分野の当業者は、要求されたパケットの１又はそれ以上のＥＣＣブロックが破損され、ＥＣＣ修正モジュール３２２がエラーを修正できないことを判定したＥＣＣ修正モジュール３２２の結果として、とるべきその他の選択肢や動作が分かるであろう。

【０１７４】

読み出しデータパイプライン１０８は、デパケッタイザ３２４を具え、このデパケッタイザ３２４は、直接または間接的に、ＥＣＣ修正モジュール３２２から要求されたパケッ

トのＥＣＣブロックを受信し、１又はそれ以上のパケットヘッダをチェック及び除去する。デパケッタイザ３２４はヘッダ内のパケット識別子、データ長、データ位置等をチェックすることによってパケットヘッダを確認する。一実施例においては、ヘッダは、読み出しデータパイプライン１０８に配信されたパケットが、要求されたパケットであることを確認するために使用可能なハッシュコードを具える。更に、デパケッタイザ３２４は、パケッタイザ３０２によって加えられた要求されたパケットからヘッダを除去する。デパケッタイザ３２４は、所定のパケット上で動作しないように命令できるが、修正されずに先に通過させる。例としては、復元プロセスの途中で要求されるコンテナラベルであってもよく、ヘッダ情報は、オブジェクトインデックス復元モジュール２７２によって要求される。さらなる例としては、ソリッドステートストレージデバイス１０２内で用いられるように決められた様々なタイプのパケットの転送を含む。別の実施例においては、デパケッタイザ３２４の動作は、パケットタイプに応じたものであってもよい。

【０１７５】

読み出しデータパイプライン１０８は、デパケッタイザ３２４からデータを受信し、所望されないデータを除去するアライメントモジュール３２６を具える。一実施例においては、ソリッドステートストレージ１１０に送信される読み出しコマンドは、パケットのデータを回復させる。データを要求しているデバイスは、回復したパケットの総てのデータを必要としなくてもよく、アライメントモジュール３２６は、所望されないデータを除去する。回復したページ内の総てのデータが要求されているデータである場合、アライメントモジュール３２６は、いずれのデータも除去しない。

【０１７６】

アライメントモジュール３２６は、次のステージにデータセグメントを転送する前に、データセグメントを要求しているデバイスと互換性のある形態で、オブジェクトのデータセグメントとしてデータを再フォーマットする。一般的に、データは、読み出しデータパイプライン１０８によって処理される場合、データセグメントまたはパケットのサイズは、様々なステージで変更される。アライメントモジュール３２６は、受信したデータを用いて、データを、要求デバイス１５５に送信し、応答を形成するのに加えられるのに適したデータセグメントにフォーマットする。例えば、第１のデータパケットの一部からのデータは、第２のデータパケットの一部からのデータと組み合わせることができる。データセグメントが、要求デバイスによって要求されるデータよりも長い場合、アライメントモジュール３２６は、所望されないデータを廃棄できる。

【０１７７】

一実施例においては、読み出しデータパイプライン１０８は、読み出しデータパイプライン１０８によって処理される前に、ソリッドステートストレージ１１０から読み出される１又はそれ以上の要求されたパケットをバッファリングする読み出し同期バッファ３２８を具える。読み出し同期バッファ３２８は、ソリッドステートストレージクロックドメインと、ローカルバスクロックドメインとの間の境界にあり、クロックドメインの差を考慮するバッファリングを提供する。

【０１７８】

別の実施例においては、読み出しデータパイプライン１０８は、出力バッファ３３０を具え、この出力バッファ３３０は、アライメントモジュール３２６からの要求されたパケットを受信し、要求デバイスに送信される前にパケットを記憶する。出力バッファ３３０は、データセグメントが、読み出しデータパイプライン１０８のステージから受信されるときと、データセグメントが、ソリッドステートストレージコントローラ１０４のその他の部分または要求デバイスに送信されるときと、の間の差を考慮している。更に、出力バッファ３３０によって、データバス２０４は、データバス２０４の動作効率を改善するために、読み出しデータパイプライン１０８によって維持されうるよりも高い頻度で、読み出しデータパイプライン１０８からのデータを受信できる。

【０１７９】

一実施例においては、読み出しデータパイプライン１０８は、メディア復号化モジュー

ル３３２を具え、このメディア復号化モジュール３３２は、ＥＣＣ修正モジュール３２２からの１又はそれ以上の暗号化された要求されたパケットを受信し、１又はそれ以上の要求されたパケットをデパケッタイザ３２４に送信する前に、ソリッドステートストレージデバイス１０２に固有の暗号化キーを用いて１又はそれ以上の要求されたパケットを復号化する。一般的に、メディア復号化モジュール３３２によってデータを復号化するのに用いられる暗号化キーは、メディア復号化モジュール３１８によって用いられる暗号化キーと同一である。別の実施例においては、ソリッドステートストレージ１１０は、２又はそれ以上のパーティションを具えていてもよく、ソリッドステートストレージコントローラ１０４は、２又はそれ以上のソリッドステートストレージコントローラ１０４であるかのように動作し、各々は、ソリッドステートストレージ１１０内の単一のパーティションで動作する。この実施例においては、固有のメディア暗号化キーは、各パーティションで用いられてもよい。

【０１８０】

別の実施例においては、読み出しデータパイプライン１０８は、復号化モジュール３３４を具え、この復号化モジュールは、データセグメントを出力バッファ３３０に送信する前に、デパケッタイザ３２４によってフォーマットされるデータセグメントを復号化する。データセグメントは、読み出し要求と共に受信された暗号化キーを用いて復号化され、読み出し要求は、読み出し同期バッファ３２８によって受信される要求されたパケットの回復を開始させる。復号化モジュール３３４は、第１のパケットの読み出し要求とともに受信される暗号化キーを有する第１のパケットを復号化でき、次いで、異なる暗号化キーを有する第２のパケットを復号化することができ、または、復号化せずに読み出しデータパイプライン１０８の次のステージに第２のパケットを通過させることができる。一般的に、復号化モジュール３３４は、メディア復号化モジュール３３２が要求されたパケットを復号化するのに用いるものとは異なる暗号化キーを用いてデータセグメントを復号化する。パケットが非シークレット暗号化ナンスで記憶されていた場合は、ナンスを、暗号化キーとともに用いて、データパケットを復号化する。暗号化キーは、クライアント１１４、コンピュータ１１２、キーマネージャ、または、ソリッドステートストレージコントローラ１０４によって用いられる暗号化キーを管理するその他のデバイスから、受信できる。

【０１８１】

別の実施例においては、読み出しデータパイプライン１０８は、デパケッタイザ３２４によってフォーマットされたデータセグメントを解凍する解凍モジュール３３６を具える。好ましい実施例においては、解凍モジュール３３６は、パケットヘッダおよびコンテナラベルの１又は双方に記憶された圧縮情報を用いて、圧縮モジュール３１２によってデータを圧縮するのに用いられるのと相補的なルーチンを選択する。別の実施例においては、解凍モジュール３３６によって用いられる解凍ルーチンは、解凍しようとするデータセグメントを要求しているデバイスによって命令される。別の実施例においては、解凍モジュール３３６は、オブジェクト毎のタイプまたはオブジェクトクラスベースのデフォルトの設定によって、解凍ルーチンを選択する。第１のオブジェクトの第１のパケットは、デフォルトの解凍ルーチンを無効にすることができ、同一オブジェクトクラスおよびオブジェクトタイプの第２のオブジェクトの第２のパケットは、デフォルトの解凍ルーチンを用いることができ、同一オブジェクトクラスおよびオブジェクトタイプの第３のオブジェクトの第３のパケットは、解凍を利用しなくてもよい。

【０１８２】

別の実施例においては、読み出しデータパイプライン１０８は、読み出しデータパイプライン１０８内の１又はそれ以上のユーザ定義可能な機能を含む、読み出しプログラムモジュール３３８を具える。読み出しプログラムモジュール３３８は、書き込みプログラムモジュール３１０と同様の特性を有し、ユーザは、読み出しデータパイプライン１０８にカスタム機能を提供することができる。読み出しプログラムモジュール３３８は、図３に示されるように配置可能であり、または、読み出しデータパイプライン１０８内で別のポ

ジションに配置可能であり、あるいは、読み出しデータパイプライン１０８内の複数の位置の複数の部分を具えてもよい。更に、別個に動作する読み出しデータパイプライン１０８内の複数の位置内に、複数の読み出しプログラムモジュール３３８があってもよい。当分野の当業者は、読み出しデータパイプライン１０８内の読み出しプログラムモジュール３３８のその他の形態が分かるであろう。書き込みデータパイプライン１０６と同様に、読み出しデータパイプライン１０８のステージは、再配置可能であり、当分野の当業者は、読み出しデータパイプライン１０８内のその他の順番が分かるであろう。

【０１８３】

ソリッドステートストレージコントローラ１０４は、制御およびステータスレジスタ３４０と、対応する制御待ち行列３４２を具える。制御およびステータスレジスタ３４０および制御待ち行列３４２は、書き込みおよび読み出しデータパイプライン１０６、１０８内で処理されるデータに関連した、制御およびシーケンスコマンドならびにサブコマンドを容易にする。例えば、パケッタイザ３０２のデータセグメントは、ＥＣＣジェネレータに関連した制御待ち行列３４２において、１又はそれ以上の対応する制御コマンドまたは命令を有することができる。データセグメントが、パケット化されるとき、命令またはコマンドのいくつかは、パケッタイザ３０２内で実行されてもよい。その他のコマンドまたは命令は、データセグメントから生成された、新規に形成されたデータパケットが、次のステージへ通過するとき、制御およびステータスレジスタ３４０を介して次の制御待ち行列３４２へ通過する。

【０１８４】

コマンドまたは命令は、制御待ち行列３４２に同時にロードされてもよく、パケットは、書き込みデータパイプライン１０６に転送され、各パイプラインステージは、各々のパケットがそのステージで実行されるとき、適宜なコマンドまたは命令を引き出す。同様に、コマンドまたは命令は、制御待ち行列３４２に同時にロード可能であり、パケットは、読み出しデータパイプライン１０８から要求されており、各パイプラインステージは、各々のパケットがそのステージで実行されるときに、適宜なコマンドまたは命令を引き出す。当分野の当業者は、制御およびステータスレジスタ３４０、ならびに、制御待ち行列３４２のその他の特性および機能が分かるであろう。

【０１８５】

ソリッドステートストレージコントローラ１０４および／またはソリッドステートストレージデバイス１０２は更に、バンクインタリーブコントローラ３４４、同期バッファ３４６、ストレージバスコントローラ３４８およびマルチプレクサ（「ＭＵＸ」）３５０を具えることができ、これらは、図４Ａおよび図４Ｂに関連して記載される。

【０１８６】

バンクインタリーブ

【０１８７】

図４Ａは、本発明によるソリッドステートストレージコントローラ１０４内のバンクインタリーブコントローラ３４４の一実施例４００を示した概略ブロック図である。バンクインタリーブコントローラ３４４は、制御およびステータスレジスタ３４０に接続され、ＭＵＸ３５０、ストレージコントローラ３４８および同期バッファ３４６を介してストレージＩ／Ｏバス２１０およびストレージ制御バス２１２に接続され、これらは以下で説明される。バンクインタリーブコントローラは、ソリッドステートストレージ１１０内のバンク２１４に対し、読み出しエージェント４０２、書き込みエージェント４０４、消去エージェント４０６、管理エージェント４０８、読み出し待ち行列４１０ａ－ｎ、書き込み待ち行列４１２ａ－ｎ、消去待ち行列４１４ａ－ｎ、及び管理待ち行列４１６ａ－ｎを、並びに、バンクコントローラ４１８ａ－ｎ、バスアービタ４２０、及びステータスＭＵＸ４２２を具え、これらは以下で説明する。ストレージバスコントローラ３４８は、再マッピングモジュール４３０を有するマッピングモジュール４２４、ステータスキャプチャモジュール４２６、および、ＮＡＮＤバスコントローラ４２８を具え、これらは以下で説明する。

【０１８８】
バンクインタリーブコントローラ３４４は、バンクインタリーブコントローラ３４４において、１又はそれ以上のコマンドを２又はそれ以上の待ち行列に命令し、ソリッドステートストレージ１１０のバンク２１４間で、上記待ち行列に記憶されたコマンドの実行を調整し、第１のタイプのコマンドは１つのバンク２１４ａで実行され、第２のタイプのコマンドは第２のバンク２１４ｂで実行されるようにする。１又はそれ以上のコマンドは、コマンドタイプによって待ち行列に分けられる。ソリッドステートストレージ１１０の各バンク２１４は、バンクインタリーブコントローラ３４４内の対応する待ち行列のセットを有し、各々の待ち行列のセットは、各コマンドタイプ用の待ち行列を具える。
【０１８９】
バンクインタリーブコントローラ３４４は、ソリッドステートストレージ１１０のバンク２１４間で、待ち行列に記憶されたコマンドの実行を調整する。例えば、第１のタイプのコマンドは、１つのバンク２１４ａで実行され、第２位のタイプのコマンドは、第２のバンク２１４ｂで実行される。一般的に、コマンドタイプおよび待ち行列タイプは、読み出しおよび書き込みコマンド、ならびに、待ち行列４１０、４１２を具えるが、更に、ストレージメディア特有のその他のコマンドおよび待ち行列を具えることができる。例えば、図４Ａに示される実施例において、消去および管理待ち行列４１４、４１６が含まれ、フラッシュメモリ、ＮＲＡＭ、ＭＲＡＭ、ＤＲＡＭ、ＰＲＡＭなどに好適である。
【０１９０】
その他のタイプのソリッドステートストレージ１１０のために、その他のタイプのコマンドおよび対応する待ち行列も、本発明の範囲から逸脱することなく含むことができる。ＦＰＧＡソリッドステートストレージコントローラ１０４のフレキシブルな性質は、ストレージメディアにおけるフレキシビリティを与える。フラッシュメモリが、別のソリッドステートストレージタイプに変更された場合、バンクインタリーブコントローラ３４４、ストレージバスコントローラ３４８、および、ＭＵＸ３５０は、データパイプライン１０６、１０８、および、その他のソリッドステートストレージコントローラ１０４の機能に重大な影響を有さないメディアタイプを収容できるように変更可能である。
【０１９１】
図４Ａに示された実施例においては、バンクインタリーブコントローラ３４４は、各バンク２１４用の、ソリッドステートストレージ１１０からのデータを読み出すための読み出し待ち行列４１０、ソリッドステートストレージ１１０に対する書き込みコマンド用の書き込み待ち行列４１２、ソリッドステートストレージ中の消去ブロックを消去するための消去待ち行列４１４、管理コマンド用の管理待ち行列４１６を具える。更に、バンクインタリーブコントローラ３４４は、対応する読み出し、書き込み、消去および管理エージェント４０２、４０４、４０６、４０８を具える。別の実施例においては、制御およびステータスレジスタ３４０ならびに制御待ち行列３４２、または、同様の構成は、バンクインタリーブコントローラ３４４なしで、ソリッドステートストレージ１１０のバンク２１４に送信されるデータ用コマンドを待ち行列に入れる。
【０１９２】
エージェント４０２、４０４、４０６、４０８は、一実施例においては、特定のバンク２１４ａ用に定められた適宜なタイプのコマンドを、バンク２１４ａ用の正しい待ち行列に命令する。例えば、読み出しエージェント４０２は、バンク－１　２１４ｂ用の読み出しコマンドを受信し、読み出しコマンドをバンク－１読み出し待ち行列４１０ｂに命令できる。書き込みエージェント４０４は、ソリッドステートストレージ１１０のバンク－０　２１４ａ内の位置にデータを書き込むための書き込みコマンドを受信でき、次いで、書き込みコマンドをバンク－０書き込み待ち行列４１２ａに送信する。同様に、消去エージェント４０６は、バンク－１　２１４ｂの消去ブロックを消去する消去コマンドを受信でき、次いで、消去コマンドをバンク－１消去待ち行列４１４ｂに送信する。管理エージェント４０８は、一般的に、バンク－０　２１４ａなど、バンク２１４の構成レジスタを読み出すためのリセットコマンドまたは要求のような、管理コマンド、ステータス要求等を

受信する。管理エージェント４０８は、管理コマンドをバンク－０管理待ち行列４１６ａに送信する。

【０１９３】

更に、エージェント４０２、４０４、４０６、４０８は、一般的に、待ち行列４１０、４１２、４１４、４１６の状態をモニタし、待ち行列４１０、４１２、４１４、４１６が完全、ほぼ完全、機能しない等である場合にステータス、中断またはその他のメッセージを送信する。一実施例においては、エージェント４０２、４０４、４０６、４０８は、コマンドを受信し、対応するサブコマンドを生成する。一実施例においては、エージェント４０２、４０４、４０６、４０８は、制御及びステータスレジスタ３４０を介したコマンドを受信し、待ち行列４１０、４１２、４１４、４１６に転送される、対応するサブコマンドを生成する。当分野の当業者は、エージェント４０２、４０４、４０６、４０８のその他の機能が分かるであろう。

【０１９４】

待ち行列４１０、４１２、４１４、４１６は、一般的に、コマンドを受信し、ソリッドステートストレージバンク２１４に送信されることが要求されるまでコマンドを記憶する。一般的な実施例においては、待ち行列４１０、４１２、４１４、４１６は、先入れ先出し（「ＦＩＦＯ」）レジスタまたはＦＩＦＯとして動作する同様の構成である。別の実施例においては、待ち行列４１０、４１２、４１４、４１６は、データにマッチする順番、重要度の順番またはその他の基準でコマンドを記憶する。

【０１９５】

バンクコントローラ４１８は、一般的に、待ち行列４１０、４１２、４１４、４１６からのコマンドを受信し、適宜なサブコマンドを生成する。例えば、バンク－０書き込み待ち行列４１２ａは、データパケットのページをバンク－０　２１４ａに書き込むコマンドを受信することができる。バンク－０コントローラ４１８ａは、適宜な時間で書き込みコマンドを受信し、１又はそれ以上の書き込みサブコマンドを生成することができ、書き込みバッファ３２０に記憶される各データパケットは、バンク－０　２１４ａに書き込まれる。例えば、バンク－０コントローラ４１８ａは、バンク０　２１４ａおよびソリッドステートストレージアレイ２１６の状態を有効にするコマンドを生成し、１又はそれ以上のデータパケットを書き込む適宜な位置を選択し、ソリッドステートストレージメモリアレイ２１６内の入力バッファをクリアし、１又はそれ以上のデータパケットを入力バッファに転送し、入力バッファを選択した位置にプログラミングし、データが正しくプログラミングされたことを検証し、プログラムに破損がある場合、マスタコントローラの中断するステップ、同一の物理位置への書き込みを再試行するステップ、異なる物理位置への書き込みを再試行するステップのうちの１又はそれ以上を行う。更に、例示の書き込みコマンドとともに、ストレージバスコントローラ３４８は、１又はそれ以上のコマンドを、ストレージＩ／Ｏバス２１０ａ用の第１の物理アドレスにマッピングされ、ストレージＩ／Ｏバス２１０ｂ用の第２の物理アドレスにマッピングされ、その他更に以下に述べるようなコマンドの論理アドレスを有する、ストレージＩ／Ｏバス２１０ａ－ｎの各々に乗算する。

【０１９６】

一般的に、バスアービタ４２０は、バンクコントローラ４１８間から選択され、バンクコントローラ４１８内の出力待ち行列からサブコマンドを引き出し、バンク２１４の性能を最適化する順番にストレージバスコントローラ３４８に転送する。別の実施例においては、バスアービタ４２０は、高位中断信号に応答し、標準的な選択基準を変更することができる。別の実施例においては、マスタコントローラ２２４は、制御およびステータスレジスタ３４０を介してバスアービタ４２０を制御することができる。当分野の当業者は、バスアービタ４２０が、バンクコントローラ４１８からソリッドステートストレージ１１０へのコマンドのシーケンスを制御しインタリーブすることができるその他の手段が分かるであろう。

【０１９７】

バスアービタ４２０は、一般的に、適宜なコマンドと、コマンドタイプが必要とされるときは、バンクコントローラ４１８の対応するデータと、の選択を調整し、コマンド及びデータをストレージバウコントローラ３４８に送信する。一般的に、バスアービタ４２０は、更に、適宜なバンク２１４を選択するのに、ストレージ制御バス２１２にコマンドを送信する。フラッシュメモリ、または、非同期式双方向直列ストレージＩ／Ｏバス２１０を有するその他のソリッドステートストレージ１１０の場合、１つのコマンド（制御情報）またはデータセットのみが、同時に送信可能である。例えば、書き込みコマンドまたはデータが、ストレージＩ／Ｏバス２１０のソリッドステートストレージ１１０に送信された場合、読み出しコマンド、読み出されたデータ、消去コマンド、管理コマンドまたはその他のステータスコマンドは、ストレージＩ／Ｏバス２１０に送信できない。例えば、データがストレージＩ／Ｏバス２１０から読み出された場合、データはソリッドステートストレージ１１０に書き込むことができない。

【０１９８】

例えば、バンク－０の書き込み動作中、バスアービタ４２０は、バンク－０コントローラ４１８ａを選択し、このコントローラ４１８ａは、ストレージバスコントローラ３４８に次のシーケンスを実行させる待ち行列のトップに書き込みコマンドまたは書き込みサブコマンドの配列を有することができる。バスアービタ４２０は、書き込みコマンドをストレージバスコントローラ３４８に転送し、書き込みコマンドを、ストレージ制御バス２１２を介してバンク－０　２１４ａを選択し、バンク－０　２１４ａに関連したソリッドステートストレージエレメント１１０の入力バッファをクリアするコマンドを送信し、バンク－０　２１４ａに関連したソリッドステートストレージエレメント２１６、２１８、２２０のステータスを有効にするコマンドを送信することによって設定する。次いで、ストレージバスコントローラ３４８は、ストレージＩ／Ｏバス２１０に書き込みサブコマンドを送信し、論理消去ブロックアドレスからマッピングされるときに、個々の物理消去ソリッドステートストレージエレメント２１６ａ－ｍ用の論理消去ブロックのアドレスを含む、物理アドレスを具える。次いで、ストレージバスコントローラ３４８は、ＭＵＸ３５０を介したストレージＩ／Ｏバス２１０に、書き込み同期バッファ３０８を介した書き込みバッファ３２０を多重化し、書き込みデータを適宜なページにストリーミングする。ページが完全である場合、次いで、ストレージバスコントローラ３４８は、バンク－０　２１４ａに関連するソリッドステートストレージエレメント２１６ａ－ｍに、入力バッファを、ソリッドステートストレージエレメント２１６ａ－ｍ内のメモリセルにプログラミングさせる。最終的に、ストレージバスコントローラ３４８は、状態を有効にして、ページが正しくプログラムされていることを確認する。

【０１９９】

読み出し動作は、上述の書き込み例と同様である。読み出し動作中、一般的に、バスアービタ４２０またはバンクインタリーブコントローラ３４４のその他の構成は、データと、対応するステータス情報を受信し、データを読み出しデータパイプライン１０８に送信し、ステータス情報を制御およびステータスレジスタ３４０に送信する。一般的に、バスアービタ４２０からストレージバスコントローラ３４８に転送される読み出しデータコマンドは、ＭＵＸ３５０に、ストレージＩ／Ｏバス２１０の読み出しデータを読み出しデータパイプライン１０８にゲート制御し、ステータスＭＵＸ４２２を介してステータス情報を適宜な制御及びステータスレジスタ３４０に送信する。

【０２００】

バスアービタ４２０は、様々なコマンドタイプとデータアクセスモードとを調整し、適宜なコマンドタイプまたは対応するデータのみが、所定時間に、バスに存在している。バスアービタ４２０が、書き込みコマンドを選択しており、書き込みサブコマンドと対応するデータとが、ソリッドステートストレージ１１０に書き込まれている場合、バスアービタ４２０は、ストレージＩ／Ｏバス２１０のその他コマンドタイプを許可しない。有利には、バスアービタ４２０は、バンク２１４ステータスに関して受信したステータス情報に沿った、予測したコマンド実行時間のようなタイミング情報を用い、バスのアイドル時間

を最小化または除去する目的として、バスの様々なコマンドの実行を調整する。

【０２０１】

バスアービタ４２０を介したマスタコントローラ２２４は、一般的に、ステータス情報とともに、待ち行列４１０、４１２、４１４、４１６に記憶されたコマンドの予測完了時間を用いて、コマンドに関連したサブコマンドが１つのバンク２１４ａで実行されているとき、他のコマンドの他のサブコマンドが、他のバンク２１４ｂ－ｎで実行される。１つのコマンドが、２１４ａで完全に実行されるとき、バスアービタ４２０は、別のコマンドをバンク２１４ａに命令する。バスアービタ４２０は更に、待ち行列４１０、４１２、４１４、４１６に記憶されたコマンドと、待ち行列４１０、４１２、４１４、４１６に記憶されていない他のコマンドとを調整することができる。

【０２０２】

例えば、消去コマンドは、ソリッドステートストレージ１１０内の消去ブロックのグループを消去するように送信可能である。消去コマンドは、書き込みまたは読み出しコマンドを実行するよりも１０乃至１０００倍以上の時間がかかり、プログラムコマンドを実行するよりも１０乃至１００倍以上の時間がかかる。Ｎバンク２１４のために、バンクインタリーブコントローラ３４４は、消去コマンドをＮコマンドに分割することができ、各々は、バンク２１４ａの仮想消去ブロックを消去する。バンク－０　２１４ａが消去コマンドを実行している間、バスアービタ４２０は、その他のバンク２１４ｂ－ｎで実行されるその他のコマンドを選択することができる。更に、バスアービタ４２０は、ストレージバスコントローラ３４８、マスタコントローラ２２４等の他の構成とともに動作することができ、バス間のコマンド実行を調整する。バンクインタリーブコントローラ３４４のバスアービタ４２０、バンクコントローラ４１８、待ち行列４１０、４１２、４１４、４１６およびエージェント４０２、４０４、４０６、４０８を用いてコマンド実行を調整することで、バンクインタリーブ機能なしで、その他のソリッドステートストレージシステムを介して、大幅に性能を向上させることができる。

【０２０３】

一実施例においては、ソリッドステートストレージコントローラ１０４は、ソリッドステートストレージ１１０のストレージエレメント２１６、２１８、２２０の総てに働く１つのバンクインタリーブコントローラ３４４を具える。別の実施例においては、ソリッドステートコントローラ１０４は、ストレージエレメント２１６ａ－ｍ、２１８ａ－ｍ、２２０ａ－ｍの各列用のバンクインタリーブコントローラ３４４を具える。例えば、１つのバンクインタリーブコントローラ３４４は、ストレージエレメントＳＳＳ０．０－ＳＳＳ０．Ｎ２１６ａ、２１８ａ、２２０ａの一列に働き、第２のバンクインタリーブコントローラ３４４は、ストレージエレメントＳＳＳ１．０－ＳＳＳ１．Ｎ２１６ｂ、２１８ｂ、２２０ｂなどの第２の列に働く。

【０２０４】

図４Ｂは、本発明によるソリッドステートストレージコントローラ内のバンクインタリーブコントローラの代替的な実施例４０１を示した概略ブロック図である。図４Ｂに示された実施例に示された構成２１０、２１２、３４０、３４６、３４８、３５０、４０２－４３０は、各バンク２１４が、単一の待ち行列４３２ａ－ｎを具えて、バンク（例えば、バンク０　２１４ａ）用の読み出しコマンド、書き込みコマンド、消去コマンド、管理コマンドが、バンク２１４ａの単一の待ち行列４３２ａに対して命令されることを除いては図４Ａに関連して示されたバンクインタリーブ装置４００と実質的に同じである。一実施例においては、待ち行列４３２は、ＦＩＦＯである。別の実施例においては、待ち行列４３２は、待ち行列４３２が記憶されている順番とは別の順番で、待ち行列４３２から引き出されたコマンドを有することができる。別の代替的な実施例においては（図示せず）、読み出しエージェント４０２、書き込みエージェント４０４、消去エージェント４０６および管理エージェント４０８は、適宜な待ち行列４３２ａ－ｎに対する単一のエージェント割当コマンドに組み込むことができる。

【０２０５】

別の代替的な実施例においては（図示せず）、コマンドは、１つの待ち行列に記憶され、これらのコマンドは、記憶された方法以外の順番で待ち行列から引き出し可能であり、バンクインタリーブコントローラ３４４は、１つのバンク２１４ａでコマンドを実行でき、他のコマンドは、残りのバンク２１４ｂ－ｎで実行される。当分野の当業者であれは、他のコマンドが他のバンク２１４ｂ－ｎで実行されている間に、１つのバンク２１４ａでコマンドを実行できる、他の待ち行列構成およびタイプが容易に分かるであろう。

【０２０６】

ストレージ特有の構成

【０２０７】

ソリッドストレージコントローラ１０４は、同期バッファ３４６を具え、この同期バッファ３４６は、ソリッドステートストレージ１１０から送信されて受信されたコマンドおよびステータスメッセージをバッファリングする。同期バッファ３４６は、ソリッドステートストレージクロックドメインとローカルバスクロックドメインとの間の境界に配置され、クロックドメインの差を考慮するバッファリングを提供する。同期バッファ３４６、書き込み同期バッファ３０８および読み出し同期バッファ３２８は独立にでき、一緒に動作させて、データ、コマンド、ステータスメッセージなどをバッファリングすることができる。好ましい実施例においては、同期バッファ３４６は、クロックドメインを交差する信号が最小数のところに配置される。当分野の当業者は、クロックドメイン間の同期が、設計の実装のいくつかの態様を最適にするために、ソリッドステートストレージデバイス１０２内の他の位置に任意に移動可能であることが分かるであろう。

【０２０８】

ソリッドステートストレージコントローラ１０４は、ストレージバスコントローラ３４８を具え、このストレージバスコントローラ３４８は、ソリッドステートストレージ１１０に送信され、読み出されるデータに対するコマンドと、ソリッドステートストレージ１１０のタイプに基づいてソリッドステートストレージ１１０から受信したステータスメッセージとを解釈し翻訳する。例えば、ストレージバスコントローラ３４８は、性能特性の異なるストレージ、製造者の異なるストレージなど、異なるタイプのストレージに対する異なるタイミング要求を具えることができる。更に、ストレージバスコントローラ３４８は、ストレージ制御バス２１２に制御コマンドを送信する。

【０２０９】

好ましい実施例においては、ソリッドステートストレージコントローラ１０４は、ＭＵＸ３５０を具え、ＭＵＸ３５０は、マルチプレクサ３５０ａ－ｎのアレイを具え、各マルチプレクサは、ソリッドステートストレージアレイ１１０の列の専用となる。例えば、マルチプレクサ３５０ａは、ソリッドステートストレージエレメント２１６ａ、２１８ａ、２２０ａに関連している。ＭＵＸ３５０は、書き込みデータパイプライン１０６からのデータと、ストレージＩ／Ｏバス２１０を介したストレージバスコントローラ３４８からソリッドステートストレージ１１０へのコマンドをルーティングし、ストレージＩ／Ｏバス２１０を介したソリッドステートストレージ１１０から、読み出しデータパイプライン１０６およびストレージバスコントローラ３４８、同期バッファ３４６およびバンクインタリーブコントローラ３４４を介した制御およびステータスレジスタ３４０へのデータおよびステータスメッセージをルーティングする。

【０２１０】

好ましい実施例においては、ソリッドステートストレージコントローラ１０４は、ソリッドステートストレージエレメントの各列（例えば、ＳＳＳ０．１　２１６ａ、ＳＳＳ０．２　２１８ａ、ＳＳＳ０．Ｎ　２２０ａ）用のＭＵＸ３５０を具える。ＭＵＸ３５０は、書き込みデータパイプライン１０６からのデータと、ストレージＩ／Ｏバス２１０を介してソリッドステートストレージ１１０へ送信されるコマンドを組合せ、コマンドから読み出しデータパイプライン１０８によって処理されるようにデータを分割する。書き込みバッファ３２０に記憶されたパケットは、書き込みバッファ３２０の外部から、ソリッドステートストレージエレメントの各列（ＳＳＳｘ．０からＳＳＳｘ．Ｎ２１６、２１８、

２２０）用の書き込み同期バッファ３０８を介し、ソリッドステートストレージエレメントの各列（ＳＳＳｘ．０からＳＳＳｘ．Ｎ２１６、２１８、２２０）用のＭＵＸ３５０までのバスに命令される。コマンドおよび読み出しデータは、ストレージＩ／Ｏバス２１０からＭＵＸ３５０によって受信される。更に、ＭＵＸ３５０は、ステータスメッセージをストレージバスコントローラ３４８に命令する。

【０２１１】

ストレージバスコントローラ３４８は、マッピングモジュール４２４を具える。マッピングモジュール４２４は、消去ブロックの１又はそれ以上の物理アドレスに消去ブロックの論理アドレスをマッピングする。例えば、ブロック２１４ａにつき２０のストレージエレメント（例えば、ＳＳＳ０．０からＳＳＳ　Ｍ．０　２１６）のアレイを有するソリッドステートストレージ１１０は、ストレージエレメントにつき１の物理アドレスで、消去ブロックの２０の物理アドレスにマッピングされる特定の消去ブロック用の論理アドレスを有することができる。ストレージエレメントが並列にアクセスされるので、ストレージエレメント２１６ａ、２１８ａ、２２０ａの列の各ストレージエレメントにおける同一位置の消去ブロックは、物理アドレスを共有する。列中の総ての消去ブロック（例えば、ストレージエレメントＳＳＳ　０．０、０．１、．．．、０．Ｎ　２１６ａ、２１８ａ、２２０ａにおいて）の代わりに、１の消去ブロックを選択するために、１つのバンク（この場合、バンク－０　２１４ａ）が選択される。

【０２１２】

このような消去ブロック用の論理－物理マッピングは、１つの消去ブロックが損傷またはアクセス不能になった場合、マッピングが別の消去ブロックにマッピングされるように変更可能であるので、有益である。これによって、１つのエレメントの消去ブロックが失敗したとき、仮想消去ブロック全体を失う損失を緩和する。再マッピングモジュール４３０は、消去ブロックの論理アドレスのマッピングを、仮想消去ブロックの１又はそれ以上の物理アドレス（ストレージエレメントのアレイにわたって配置される）に変更する。例えば、仮想消去ブロック１は、ストレージエレメントＳＳＳ　０．０　２１６ａの消去ブロック１、ストレージエレメントＳＳＳ１．０　２１６ｂの消去ブロック１、．．．、ストレージエレメントＭ．０　２１６ｍにマッピング可能であり、仮想消去ブロック２は、ストレージエレメントＳＳＳ０．１　２１８ａの消去ブロック２、ストレージエレメントＳＳＳ１．１　２１８ｂの消去ブロック２、．．．、ストレージエレメントＭ．１　２１８ｍにマッピング可能等である。

【０２１３】

ストレージエレメントＳＳＳ０．０　２１６ａの消去ブロック１が損傷し、摩耗などによるエラーが生じるか、何らかの理由で用いられない場合、再マッピングモジュールは、仮想消去ブロック１の消去ブロックをポインティングした論理アドレス用に、論理－物理マッピングを変更することができる。ストレージエレメントＳＳＳ０．０　２１６ａのスペア消去ブロック（消去ブロック２２１と呼ぶ）が利用可能であり、その時点でマッピングされていない場合、再マッピングモジュールは、ストレージエレメントＳＳＳ０．０　２１６ａの消去ブロック２２１をポインティングするために、仮想消去ブロック１の消去ブロックのマッピングを変更でき、ストレージエレメントＳＳＳ１．０　２１６ｂの消去ブロック１、ストレージエレメントＳＳＳ２．０（図示せず）の消去ブロック１、．．．、ストレージエレメントＭ．０　２１６ｍの消去ブロック１のポインティングを継続する。マッピングモジュール４２４または再マッピングモジュール４３０は、上述した順番（ストレージエレメントの消去ブロック１に対する仮想消去ブロック１、ストレージエレメントの消去ブロック２に対する仮想消去ブロック２）で消去ブロックをマッピングでき、あるいは、その他の基準に基づいた別の順番でストレージエレメント２１６、２１８、２２０の消去ブロックをマッピングできる。

【０２１４】

一実施例においては、消去ブロックは、アクセス時間によってグループ分け可能である。データを特有の消去ブロックのページにプログラミング（書き込み）するなど、コマン

ドを実行する時間を意味する、アクセス時間によるグループ分けは、コマンド完了を平均にすることができ、仮想消去ブロックの消去ブロックにわたって実行されるコマンドは、最も遅い消去ブロックによって制限されることはない。他の実施例においては、消去ブロックは、摩耗レベル、健常などによってグループ分け可能である。当分野の当業者は、消去ブロックをマッピングまたは再マッピングするときに考慮すべきその他のファクタが分かるであろう。

【０２１５】

一実施例においては、ストレージバスコントローラ３４８は、ステータスキャプチャモジュール４２６を具え、このステータスキャプチャモジュール４２６は、ソリッドステートストレージ１１０からのステータスメッセージを受信し、ステータスメッセージをステータスＭＵＸ４２２に送信する。別の実施例においては、ソリッドステートストレージ１１０がフラッシュメモリである場合、ストレージバスコントローラ３４８は、ＮＡＮＤバスコントローラ４２８を具える。ＮＡＮＤバスコントローラ４２８は、読み出しおよび書き込みデータパイプライン１０６、１０８からソリッドステートストレージ１１０内の正しい位置にコマンドを命令し、フラッシュメモリ等の特性に基づいてコマンド実行のタイミングを調整する。ソリッドステートストレージ１１０が別のタイプのソリッドステートストレージ型である場合、ＮＡＮＤバスコントローラ４２８は、ストレージタイプに特有のバスコントローラによって置換できる。当分野の当業者は、ＮＡＮＤバスコントローラ４２８のその他の機能が分かるであろう。

【０２１６】

フローチャート

【０２１７】

図５Ａは、本発明によるデータパイプラインを用いてソリッドステートストレージデバイス１０２におけるデータ管理のための方法５００の一実施例を示した概略フローチャート図である。この方法５００は開始して（５０２）、入力バッファ３０６が、ソリッドステートストレージ１１０に書き込まれる１又はそれ以上のデータセグメントを受信する（５０４）。１又はそれ以上のデータセグメントは、一般的に、オブジェクトの少なくとも一部を具えるが、オブジェクト全体であってもよい。パケッタイザ３０２は、オブジェクトと共に１又はそれ以上のオブジェクト特有のパケットを生成できる。パケッタイザ３０２は、各パケットにヘッダを加え、一般的に、各パケットは、パケット長およびオブジェクト内のパケット用のシーケンス番号を具える。パケッタイザ３０２は、入力バッファ３０６内に記憶される１又はそれ以上のデータまたはメタデータセグメントを受信し（５０４）、ソリッドステートストレージ１１０用にサイズを調整された１又はそれ以上のパケットを生成することによって１又はそれ以上のデータまたはメタデータセグメントをパケット化し（５０６）、各パケットは、１又はそれ以上のセグメントから１つのヘッダおよびデータを具えている。

【０２１８】

一般的に、第１のパケットは、パケットが生成されたオブジェクトを識別するオブジェクト識別子を具える。第２のパケットは、ソリッドステートストレージデバイス１０２によって用いられる情報を有するヘッダを具え、第２のパケットは、第１のパケット中で識別されたオブジェクトに関連することができ、オフセット情報は、オブジェクト内の第２のパケットおよびデータを配置する。このソリッドステートストレージデバイスコントローラ２０２は、バンク２１４と、パケットがストリーミングされる物理領域を管理する。

【０２１９】

ＥＣＣジェネレータ３０４は、パケッタイザ３０２からのパケットを受信し、データパケット用のＥＣＣを生成する（５０８）。一般的に、パケットとＥＣＣブロックとの間には固定した関連性はない。ＥＣＣブロックは、１又はそれ以上のパケットを具えることができる。パケットは、１又はそれ以上のＥＣＣブロックを具えることができる。パケットは、ＥＣＣブロック内のいずれの場所でも開始および終了することができる。パケットは、第１のＥＣＣブロック内のいずれかの場所からスタートすることができ、次のＥＣＣブ

ロックのいずれかの場所で終了することができる。

【０２２０】

書き込み同期バッファ３０８は、ＥＣＣブロックをソリッドステートストレージ１１０に書き込む前に、対応するＥＣＣブロック内に配置されるように、パケットをバッファリングし（５１０）、次いで、ソリッドステートストレージコントローラ１０４は、クロックドメイン差を考慮した適宜の時間でデータを書き込み（５１２）、方法５００は終了する（５１４）。書き込み同期バッファ３０８は、ローカルクロックドメインとソリッドステートストレージ１１０のクロックドメインとの間の境界に配置される。この方法５００は、便宜上、１又はそれ以上のデータセグメントを受信するステップと、１又はデータパケットを書き込むステップと、を記載しているが、一般的に、データセグメントのストリームが受信され、グループであることに留意されたい。一般的に、ソリッドステートストレージ１１０の完全な仮想ページを具える多数のＥＣＣブロックが、ソリッドステートストレージ１１０に書き込まれる。一般的に、パケッタイザ３０２は、あるサイズのデータセグメントを受信し、別のサイズのパケットを生成する。これは必要に応じて、データ若しくはメタデータセグメント、または、データ若しくはメタデータセグメントの一部が、セグメントのデータ総てがパケットにキャプチャされるように、データパケットを形成するために組み合わされることを要求する。

【０２２１】

図５Ｂは本発明によるインサーバＳＡＮのための方法の一実施例を示した概略フローチャート図である。方法５００は開始し（５５２）、ストレージ通信モジュール１６２は、第１のストレージコントローラ１５２ａと、第１のサーバ１１２ａの外部にある少なくとも１のデバイスとの間の通信を促進する（５５４）。第１のストレージコントローラ１５２ａと外部デバイスとの間の通信は、第１のサーバ１１２ａから独立している。第１のストレージコントローラ１１２ａは第１のサーバ１１２ａ内にあり、第１のストレージコントローラ１５２は少なくとも１のストレージデバイス１５４ａを制御する。第１のサーバ１１２ａは、第１のサーバ１１２ａと第１のストレージコントローラ１５２ａと共に同一位置に配置されるネットワークインタフェース１５６ａを具える。インサーバＳＡＮモジュール１６４はストレージ要求を提供し（５５６）、方法５０１は終了する（５５８）。インサーバＳＡＮモジュールはネットワークプロトコル及び／又はバスプロトコルを用いてストレージ要求を提供する（５５６）。インサーバＳＡＮモジュール１６４は第１のサーバ１１２ａと別個のストレージ要求を提供し（５５６）、サービス要求はクライアント１１４、１１４ａから受信される。

【０２２２】

図６は、本発明によるデータパイプラインを用いてソリッドステートストレージデバイス１０２内のデータを管理する方法６００の別の実施例を示した概略フローチャート図である。この方法６００は開始（６０２）し、入力バッファ３０６が、ソリッドステートストレージ１１０に書き込まれる１又はそれ以上のデータまたはメタデータセグメントを受信する（６０４）。パケッタイザ３０２は、ヘッダを各パケットに加え、各パケットは、オブジェクト内のパケット長を一般的に具える。パケッタイザ３０２は、入力バッファ３０６に記憶された１又はそれ以上のセグメントを受信し（６０４）、ソリッドステートストレージ１１０用にサイズを調整された１又はそれ以上のパケットを生成することによって、１又はそれ以上のセグメントをパケット化し（６０６）、各パケットは、ヘッダと、１又はそれ以上のセグメントからのデータとを具える。

【０２２３】

ＥＣＣジェネレータ３０４は、パケッタイザ３０２からのパケットを受信し、パケット用の１又はそれ以上のＥＣＣブロックを生成する（６０８）。書き込み同期バッファ３０８は、ＥＣＣブロックをソリッドステートストレージ１１０に書き込む前に、対応するＥＣＣブロック内に配置されるようにパケットをバッファリングし（６１０）、次いで、ソリッドステートストレージコントローラ１０４は、クロックドメイン差を考慮して適宜な時間でデータを書き込む（６１２）。データが、ソリッドステートストレージ１１０から

要求されるとき、１又はそれ以上のデータパケットを具えるＥＣＣは、読み出し同期バッファ３２８に読み出され、バッファリングされる（６１４）。パケットのＥＣＣブロックは、ストレージＩ／Ｏバス２１０を介して受信される。ストレージＩ／Ｏバス２１０は双方向であるので、データが読み出されるとき、書き込み動作、コマンド動作が中止される。

【０２２４】

ＥＣＣ修正モジュール３２２は、読み出し同期バッファ３２８で保持される要求されたパケットのＥＣＣブロックを受信し、必要に応じて各ＥＣＣブロック内のエラーを修正する（６１６）。ＥＣＣ修正モジュール３２２は、ＥＣＣブロックに１又はそれ以上のエラーが存在し、そのエラーがＥＣＣシンドロームを用いて修正可能であるか判定し、ＥＣＣ修正モジュール３２２が、ＥＣＣブロックのエラーを修正する（６１６）。ＥＣＣ修正モジュール３２２が、検出されたエラーがＥＣＣを用いて修正できないと判定した場合は、ＥＣＣ修正モジュール３２２は、中断信号を送信する。

【０２２５】

デパケッタイザ３２４は、ＥＣＣ修正モジュール３２２がエラーを修正した後、要求されたパケットを受信し（６１８）、各パケットのパケットヘッダをチェックおよび除去することによってパケットを脱パケット化する（６１８）。アライメントモジュール３２６は、脱パケット化後、パケットを受信し、所望されないデータを除去し、セグメントまたはオブジェクトを要求するデバイスと互換性のある形態においてオブジェクトのデータまたはメタデータセグメントとしてデータパケットを再フォーマットする（６２０）。出力バッファ３３０は、脱パケット化後要求されるパケットを受信し、要求デバイスに送信する前にパケットをバッファリングし（６２２）、この方法６００は終了する（６２４）。

【０２２６】

図７は、本発明によるバンクインタリーブを用いて、ソリッドステートストレージデバイス１０２のデータを管理する方法７００に関する一実施例を示した概略フローチャート図である。この方法６００は開始し（６０２）、バンクインタリーブコントローラ３４４は、１又はそれ以上のコマンドを２又はそれ以上の待ち行列４１０、４１２、４１４、４１６に命令する（６０４）。一般的に、エージェント４０２、４０４、４０６、４０８は、コマンドタイプによってコマンドを待ち行列４１０、４１２、４１４、４１６に命令する。待ち行列４１０、４１２、４１４、４１６の各セットは、各コマンドタイプの待ち行列を具える。バンクインタリーブコントローラ３４４は、バンク２１４間で、待ち行列４１０、４１２、４１４、４１６に記憶されたコマンドの実行を調整し（６０６）、第１のタイプのコマンドは、１つのバンク２１４ａで実行され、第２のタイプのコマンドは、第２のバンク２１４ｂで実行され、この方法６００は終了する（６０８）。

【０２２７】

ストレージ空き領域回復

【０２２８】

図８は、本発明によるソリッドステートストレージデバイス１０２のガベージコレクションのための装置８００に関する一実施例を示した概略ブロック図である。この装置８００は、シーケンシャルストレージモジュール８０２、ストレージ部分選択モジュール８０４、データ回復モジュール８０６、および、ストレージ部分回復モジュール８０８を具え、これらは以下で説明される。他の実施例においては、この装置８００は、ガベージ標識モジュール８１０および消去モジュール８１２を具える。

【０２２９】

装置８００は、ストレージ部分内のページにデータパケットを順々に書き込むシーケンシャルストレージモジュール８０２を具える。パケットは、パケットが新規パケットまたは変更されたパケットであるかにかかわらず順々に記憶される。変更されたパケットは本実施例においては一般的に、既に記憶された位置に戻って書き込まれない。一実施例においては、シーケンシャルストレージモジュール８０２は、ストレージ部分のページ内の第１の位置にパケットを書き込み、次いで、ページの次の位置にパケットを書き込み、それ

から、次、次と、ページが満たされるまで続けられる。次いで、シーケンシャルストレージモジュール８０２は、ストレージ部分内の次のページを満たし始める。これは、ストレージ部分が満たされるまで続く。

【０２３０】

好ましい実施例においては、シーケンシャルストレージモジュール８０２は、バンク（バンク－０　２１４ａ）のストレージエレメント（例えば、ＳＳＳ０．０乃至ＳＳＳＭ．０　２１６）内のストレージ書き込みバッファにパケットを書き込むことを開始する。ストレージ書き込みバッファが完全な場合、ソリッドステートストレージコントローラ１０４は、ストレージ書き込みバッファのデータを、バンク２１４ａのストレージエレメント２１６内の指定されたページにプログラミングできる。次いで、別のバンク（例えば、バンク－１　２１４ｂ）が選択され、シーケンシャルストレージモジュール８０２は、バンク２１４ｂのストレージエレメント２１８のストレージ書き込みバッファにパケットを書き込むことを開始し、第１のバンク－０は、指定されたページをプログラミングする。このバンク２１４ｂのストレージ書き込みバッファが完全な場合、ストレージ書き込みバッファの内容は、各ストレージエレメント２１８の別の指定ページにプログラミングされる。１つのバンク２１４ａをページにプログラミングする間に、別のバンク２１４ｂのストレージ書き込みバッファが満たされうるので、この処理は有効である。

【０２３１】

ストレージ部分は、ソリッドステートストレージデバイス１０２のソリッドステートストレージ１１０の一部を具える。一般的に、ストレージ部分は消去ブロックである。フラッシュメモリについては、消去ブロックの消去動作は、各セルを充電することによって、１を消去ブロックの全ビットに書き込む。これは、位置が総て１として開始するプログラム動作と比較して、冗長的なプロセスであり、データが書き込まれるとき、いくつかのビットは、ゼロで書き込まれるセルを放電することによってゼロに変更される。しかしながら、ソリッドステートストレージ１１０がフラッシュメモリでないか、あるいは、消去サイクルが、読み出しまたはプログラム等の、その他の動作として同様の時間を費やすフラッシュメモリを有する場合は、ストレージ部分は、消去されることが要求されない。

【０２３２】

本明細書で用いられるように、ストレージ部分は、消去ブロックに対する領域と同等であるが、消去されても消去されなくてもよい。本明細書で消去ブロックが用いられる場合、消去ブロックは、ストレージエレメント（例えば、ＳＳＳ０．０　２１６ａ）内の指定されたサイズの特定エリアを指し、一般的に、所定量のページを具える。「消去ブロック」は、フラッシュメモリと共に用いられるとき、一般的に、書き込まれる前に消去されるストレージ部分である。「消去ブロック」は、「ソリッドステートストレージ」とともに用いられる場合、消去されても消去されなくてもよい。本明細書で用いられるように、消去ブロックは、１つの消去ブロック、または、ストレージエレメント（例えば、ＳＳＳ０．０乃至ＳＳＳＭ．０　２１６ａ－ｎ）の各列の消去ブロックを有する消去ブロックのグループ、を具えることができ、消去ブロックは、仮想消去ブロックも意味する。仮想消去ブロックと関連した論理構成を指す場合、消去ブロックは、論理消去ブロック（「ＬＥＢ」）も意味する。

【０２３３】

一般的に、パケットは、処理の順番によって、順々に記憶される。一実施例においては、書き込みデータパイプライン１０６が用いられる場合、シーケンシャルストレージモジュール８０２は、パケットが書き込みデータパイプライン１０６からくる順番でパケットを記憶する。有効データは、下記で説明するように、回復動作中にストレージ部分から回復されるので、上記順番は、別のストレージ部分から読み出される有効データのパケットと混合した、要求デバイス１５５から到達するデータセグメントの結果でもよい。回復された有効なデータパケットを書き込みデータパイプライン１０６に再ルーティングすることは、図３のソリッドステートストレージコントローラ１０４に関連して上述したように、ガベージコレクタバイパス３１６を具えてもよい。

【０２３４】
装置８００は、回復用ストレージ部分を選択するストレージ部分選択モジュール８０４を具える。回復用ストレージ部分を選択することで、データを書き込むシーケンシャルストレージモジュール８０２によってストレージ部分を再利用することができ、従って、回復されたストレージ部分をストレージプールに加えるか、または、ストレージ部分がない、信頼性がない、リフレッシュすべき、あるいは、ストレージプールから一時的または永続的にストレージ部分を外すと判定した後、ストレージ部分から有効データを回復する。別の実施例においては、ストレージ部分選択モジュール８０４は、無効データの多いストレージ部分または消去ブロックを識別することによって、回復用ストレージ部分を選択する。

【０２３５】
別の実施例においては、ストレージ部分選択モジュール８０４は、摩耗の少ないストレージ部分または消去ブロックを識別することによって回復用ストレージ部分を選択する。例えば、摩耗の少ないストレージ部分または消去ブロックを識別することは、無効データの少ない、消去サイクルの数が少ない、ビットエラー率の低い、または、プログラムカウントの低い（回数の低い、バッファ中のデータのページは、ストレージ部分のページに書き込まれ、プログラムカウントは、デバイスが製造されたときから、ストレージ部分が最後に消去されたときから、その他の任意の事象から、および、これらの組み合わせから、測定されてもよい）ストレージ部分を識別することを含む。このストレージ部分選択モジュール８０４は、摩耗量の少ないストレージ部分を判定するために、上記パラメータまたは他のパラメータの組み合わせを用いることができる。摩耗量の低いストレージ部分を判定することによる、回復用ストレージ部分の選択は、用いられているストレージ部分を見つけるのに所望され、摩耗レベルなどに合わせて回復可能である。

【０２３６】
別の実施例においては、ストレージ部分選択モジュール８０４は、摩耗量の多いストレージ部分または消去ブロックを識別することによって回復用ストレージ部分を選択する。例えば、摩耗量の多いストレージ部分または消去ブロックを識別することは、消去サイクル数の多いストレージ部分、ビットエラー率の高いストレージ部分、回復不能ＥＣＣブロックを有するストレージ部分、または、プログラムカウントの高いストレージ部分を識別することを含む。更に、ストレージ部分選択モジュール８０４は、摩耗量の高いストレージ部分を判定する上記または他のパラメータの組み合わせを用いることができる。摩耗量の多いストレージ部分を判定することによって回復用ストレージ部分を選択することは、過剰に用いられているストレージ部分を見つけるのに所望され、消去サイクルなどを用いてストレージ部分をリフレッシュすることによって回復可能であり、または、利用不能としてサービスからストレージ部分を廃棄することができる。

【０２３７】
装置８００は、データ回復モジュール８０６を具え、このデータ回復モジュール８０６は、回復用に選択されたストレージ部分から有効なデータパケットを読み出し、他のデータパケットを有する有効なデータパケットを待ち行列に入れてシーケンシャルストレージモジュール８０２によって順々に書き込み、シーケンシャルストレージモジュール８０２によって書き込まれた有効データの新規物理アドレスを有するインデックスを更新する。一般的に、このインデックスは、データオブジェクトから由来するパケットがソリッドステートストレージ１１０に記憶される物理アドレスに、オブジェクトのデータオブジェクト識別子をマッピングするオブジェクトインデックスである。

【０２３８】
一実施例においては、装置８００は、ストレージ部分回復モジュール８０８を具え、このストレージ部分回復モジュール８０８は、使用または再使用のためのストレージ部分を準備し、データ回復モジュール８０６がストレージ部分から有効データをコピーするのが完了した後、データパケットを順々に書き込むために、利用可能なストレージ部分をシーケンシャルストレージモジュール８０２に標識する。別の実施例においては、装置８００

は、データを記憶するために利用不可能なとき、回復用に選択されたストレージ部分を標識するストレージ部分回復モジュール８０８を具える。一般的に、これは、ストレージ部分または消去ブロックが、信頼性のあるデータストレージ用に用いられる状況ではないと、摩耗量の多いストレージ部分または消去ブロックを識別するストレージ部分選択モジュール８０４によるものである。

【０２３９】

一実施例においては、装置８００は、ソリッドストレージデバイス１０２のソリッドストレージデバイスコントローラ２０２中にある。別の実施例においては、装置８００は、ソリッドステートストレージデバイスコントローラ２０２を制御する。別の実施例においては、装置８００の一部が、ソリッドステートストレージデバイスコントローラ２０２中にある。別の実施例においては、データ回復モジュール８０６によって更新されたオブジェクトインデックスは更に、ソリッドステートストレージデバイスコントローラ２０２中に配置される。

【０２４０】

一実施例においては、ストレージ部分は、消去ブロックであり、装置８００は、データ回復モジュール８０６が、選択された消去ブロックから有効なデータパケットをコピーした後で、ストレージ部分回復モジュール８０８が、利用可能として消去ブロックを標識する前に、回復用に選択された消去ブロックを消去する消去モジュール８１０を具える。読み出しまたは書き込み動作よりも大幅に長い時間を要する消去動作を有するフラッシュメモリおよびその他のソリッドストレートストレージのために、新規データを書き込むことを可能にする前にデータブロックを消去することは、効率的な動作として所望される。ソリッドステートストレージ１１０がバンク２１４中に配置される場合、消去モジュール８１０による消去動作は、１つのバンクで実行され、その他のバンクは、読み出し、書き込みまたはその他の動作を実行する。

【０２４１】

一実施例においては、装置８００は、ガベージ標識モジュール８１２を具え、このモジュール８１２は、データパケットがもはや有効ではないことを示す動作に応じて、ストレージ部分中のデータパケットが無効であると識別する。例えば、データパケットが削除される場合、ガベージ標識モジュール８１２は、データパケットが無効であると識別できる。読み出し－修正－書き込み動作は、データパケットが無効であるとして識別される別の方法である。一実施例においては、ガベージ標識モジュール８１２は、インデックスを更新することによって、データパケットが無効であると識別することができる。別の実施例においては、ガベージ標識モジュール８１２は、無効なデータパケットが削除されたことを示す別のデータパケットを記憶することによって、データパケットが無効であることを識別できる。このことが有利であるのは、ソリッドステートストレージ１１０中に、データパケットが削除されたという情報を記憶することによって、オブジェクトインデックス復元モジュール２６２または同様のモジュールが、無効データパケットが削除されたことを示すエントリを有するオブジェクトインデックスを復元することができるからである。

【０２４２】

一実施例においては、装置８００は、性能全体を改善するために、消去コマンドに続いて、データの仮想ページの残りの部分を満たすのに使用でき、消去コマンドは、書き込みパイプライン１０６が空になり、総てのパケットが、非揮発性ソリッドステートストレージ１１０内に永続的に書き込まれるまで、書き込みパイプライン１０６にデータが流れるのを中止する。これは、必要とされるガベージコレクションの量、ストレージ部分を消去するのにかかる時間、仮想ページをプログラムするのに必要な時間を低減することができる利点を有する。例えば、ソリッドステートストレージ１００の仮想ページ内に書き込むために１つの小さなパケットのみが準備されるときに、消去コマンドが受信される。ほぼ空の仮想ページをプログラミングすることが、浪費された空き領域をすぐに回復する必要を生じさせ、収集された不必要なガベージにして、消去され、回復され、および、シーケンシャルストレージモジュール８０２によって書き込む利用可能な空き領域のプールに戻

されるストレージ部分内の有効データにする。

【０２４３】

無効なデータパケットを実際に消去するよりも、データパケットを無効にして標識することが有効であるのは、上述したように、フラッシュメモリおよび他の同様のストレージについては、消去動作が大幅に時間をとるからである。装置８００に記載されるように、ガベージコレクションシステムが、ソリッドステートストレージ１１０内で自律的に動作できることが、読み出し、書き込みおよびその他のより迅速な動作から消去動作を分ける方法を提供することができ、ソリッドステートストレージデバイス１０２は、他の多くのソリッドステートストレージシステムまたはデータストレージデバイスよりも迅速に動作できる。

【０２４４】

図９は、本発明によるストレージ回復のための方法９００の一実施例を示す概略フローチャート図である。方法９００は開始し（９０２）、シーケンシャルストレージモジュール８０２は、ストレージ部分にデータパケットを順々に書き込む（９０４）。ストレージ部分は、ソリッドステートストレージデバイス１０２内のソリッドステートストレージ１１０の一部である。一般的に、ストレージ部分は消去ブロックである。データパケットは、オブジェクト由来であり、データパケットは、処理の順番によって順々に記憶される。

【０２４５】

ストレージ部分選択モジュール８０４は、回復用のストレージ部分を選択し（９０６）、データ回復モジュール８０６は、回復用に選択されたストレージ部分から有効なデータパケットを読み出す（９０８）。一般的に、有効なデータパケットは、消去または削除用に標識されていないデータパケットであるか、標識されているその他の無効データであり、有効または「優良」データとみなされる。データ回復モジュール８０６は、シーケンシャルストレージモジュール８０２によって順々に書き込まれるようにスケジューリングされた他のデータパケットを有する有効なデータパケットを待ち行列に入れる（９１０）。データ回復モジュール８０６は、シーケンシャルストレージモジュール８０２によって書き込まれる有効データの新規物理アドレスを有するインデックスを更新する（９１２）。このインデックスは、オブジェクト識別子に対するデータパケットの物理アドレスのマッピングを具える。データパケットは、ソリッドステートストレージ１１０中に記憶される際に記憶されるもの、およびデータパケットに対応するオブジェクト識別子である。

【０２４６】

データ回復モジュール８０６は、ストレージ部分からの有効データをコピー完了後、ストレージ部分回復モジュール８０８は、データパケットを順々に書き込むシーケンシャルストレージモジュール８０２が利用可能であると、回復用に選択されたストレージ部分を標識し（９１４）、方法９００は終了する（９１６）。

【０２４７】

空データセグメント指令

【０２４８】

一般的に、データがもはや有用できないとき、そのデータは消去可能である。多くのファイルシステムでは、消去コマンドは、ファイルシステム中のディレクトリエントリを削除し、データを含むストレージデバイスの所定位置にあるデータを代わりに残している。一般的に、データストレージデバイスは、この種の消去動作に含まれない。データを消去する別の方法は、ゼロ、１またはその他のヌルデータキャラクタをデータストレージデバイスに書き込み、実際に消去されたファイルに置換することである。しかしながら、これが非効率なのは、データを送信することが上書きされることであるため、有効なバンド幅が用いられるからである。更に、ストレージデバイス中の空き領域は、無効データを上書きするのに用いられるデータによって占有される。

【０２４９】

ここで記載されるソリッドステートストレージデバイス１０２のようないくつかのストレージデバイスは、ランダムアクセスストレージデバイスではないので、既に記憶された

データを更新することは既存のデータを上書きすることではない。上記デバイス上で、１の記号列またはゼロの記号列を有するデータを上書きする試みは、既存のデータを上書きする所望の意図を完了することなく、有用な空き領域を埋める。ソリッドステートストレージデバイス１０２のような、これらの非ランダムアクセスデバイスについては、クライアント１１４は、一般的に、データを消去するためにデータを上書きする能力を有しない。

【０２５０】

反復したキャラクタ列またはキャラクタ列を受信するときに、受信したデータは、高度に圧縮可能であるが、一般的に、圧縮は、ストレージデバイスに送信する前に、ファイルシステムによってなされる。一般的なストレージデバイスは、圧縮データと非圧縮データとを区別できない。更に、ストレージデバイスは、消去されたファイルを読み出すコマンドを受信することができ、ストレージデバイスは、ゼロ、１またはヌルキャラクタの記号列を要求デバイスに送信することができる。更に、バンド幅は、消去されたファイルを示すデータを送信するのに要求される。

【０２５１】

上述した説明から、ストレージデバイスが、空データセグメントまたは反復したキャラクタ若しくはキャラクタ列を有するデータを表すデータセグメントトークンを記憶することができるように、データを消去すべきという指令を受信するストレージデバイスのための装置、システムおよび方法に対するニーズが存在することは明らかである。装置、システムおよび方法は、更に、既存のデータを消去することができ、結果として用いられたストレージ空き領域は、小さなデータセグメントトークンを具える。従来技術の欠点の一部または全部を克服する装置、システムおよび方法が示される。

【０２５２】

図１０は、本発明によるトークン指令を生成する装置を有するシステム１０００の一実施例を示した概略ブロック図である。この装置は、トークン指令生成モジュール１００２、トークン指令送信モジュール１００４、読み出し受信モジュール１００６、読み出し要求送信モジュール１００８、読み出しトークン指令受信モジュール１０１０、要求クライアント応答モジュール１０１２およびデータセグメント再生成モジュール１０１４を具え、これらは以下で説明する。一実施例においては、この装置は、ストレージコントローラ１５２およびデータストレージデバイス１５４を有するストレージデバイス１５０に接続したサーバ１１２内にあり、これらは、上述したものと実質的に同等である。

【０２５３】

一実施例においては、この装置は、トークン指令を有するストレージ要求を生成するトークン指令生成モジュール１００２を具える。トークン指令は、ストレージデバイス１５０上のデータセグメントトークンを記憶する要求を具える。トークン指令は、ストレージデバイス１５０に送信されて、データセグメントトークンが所定位置に送信されていなかった場合にデータセグメントとして記憶される、反復する同一キャラクタの配列、又は反復する同一キャラクタ列の配列に置換されるように意図される。一実施例においては、反復する同一キャラクタの配列は、データセグメントが空であることを示唆する。例えば、反復する同一キャラクタの配列は、ゼロまたは１であってもよく、ゼロまたは１で埋められたデータはセグメントは、空として解釈される。

【０２５４】

トークン指令は、少なくともデータセグメント識別子およびデータセグメント長を具える。データセグメント識別子は、一般的に、オブジェクトＩＤ、ファイル名、あるいは、ストレージデバイスに反復する同一キャラクタまたはキャラクタ列を記憶しようとする、ファイルシステム、アプリケーション、サーバ１１２などに対する既知のその他の識別子である。データセグメント長は、反復する同一キャラクタまたはキャラクタ列の配列によって求められるストレージ空き領域である。データセグメントトークンおよびトークン指令は、一般的に、反復する同一キャラクタの配列のような、データセグメントのデータを具えていない。

【０２５５】
しかしながら、トークン指令は、少なくとも１インスタンスの反復する同一キャラクタまたはキャラクタ列のような、データセグメントトークンを形成する他の関連情報を具える。更に、トークン指令は、データセグメント位置、ファイルシステムからのアドレス、データセグメントに対応するデータストレージデバイスの位置などのような、メタデータを具えることができる。当分野の当業者は、トークン指令とともに含むことができるその他の情報が分かるであろう。一実施例においては、指令生成モジュール１００２は、トークン指令とともにデータセグメントトークンを生成する。
【０２５６】
一実施例においては、トークン指令生成モジュール１００２は、ストレージデバイス１５０に既存のデータを上書きする要求に応答して、トークン指令および安全消去コマンドの双方を生成する。既存のデータは、トークン指令におけるデータセグメント識別子と同一のデータセグメント識別子を有するストレージデバイスで識別されたデータを具える。一般的に、データを上書きする要求は、無効またはガベージとしてデータを単に標識するか、データまたは他の一般的な削除動作に対するポインタを削除するのに十分ではないが、データは、回復不能なように上書きされるように要求される場合に送信される。例えば、データを上書きする要求は、データが機密情報と考えられ、安全上の理由から破壊されなければならない場合に要求される。
【０２５７】
安全消去コマンドは、ストレージデバイス１５０が既存のデータを上書きするように命令し、既存のデータは回復不能となる。次いで、ストレージデバイス１５０は、既存のデータを上書き、回復、消去するとともに、データセグメントトークンを生成する。結果として、既存のデータは、回復不能になり、データセグメントトークンは、ストレージデバイス１５０に記憶され、一般的に、既存のデータよりも大幅に少ないストレージ空き領域を取る。
【０２５８】
さらなる実施例においては、装置は、消去確認モジュール１０１６を具え、この消去確認モジュール１０１６は、ストレージデバイスの既存のデータが、キャラクタとともに上書きされ、既存のデータが回復不能であることの確認を受信する。この確認は、要求デバイスまたはクライアント１１４に転送され、既存のデータが回復不能である条件におかれたことを確認するために用いられうる。他の実施例においては、安全消去コマンドは、ストレージデバイス１５０が、特有のキャラクタまたはキャラクタ列を有する既存のデータを上書きするように命令でき、実行コマンドは、複数回実行することができる。当分野の当業者は、既存のデータが回復不能であることを確証するために、１又はそれ以上の安全消去コマンドを構成する他の方法が分かるであろう。
【０２５９】
データは、暗号化可能であり、次いで、ストレージデバイス１５０に記憶され、データを記憶すると共にストレージデバイス１５０によって受信される暗号化キーを用いて暗号化がなされる。既存のデータが、記憶される前に受信された暗号化キーで暗号化される場合、別の実施例においては、トークン指令生成モジュール１００２は、既存のデータを上書きするための要求を受信するのに応じて、トークン指令を生成するとともに暗号化消去コマンドを生成する。暗号化消去コマンドは、既存のデータを記憶するのに用いられる暗号化キーを消去し、暗号化キーは回復不能になる。
【０２６０】
一実施例においては、暗号化キーを消去することは、要求デバイスから暗号化キーを消去することを具える。別の実施例においては、暗号化キーを消去することは、サーバ、キーの保管場所、キーが記憶されるその他の位置から暗号化キーを消去することを具える。暗号化キーを消去することは、暗号化キーは、いずれの方法でも回復することができないように、他のデータまたはキャラクタの配列を用いて暗号化キーを置換することを含むことができる。一般的に、暗号化ルーチンが、既存のデータを復号化する試みを邪魔するの

に十分ロバストである既存のデータを暗号化するのに用いられた場合、暗号化キーを消去することは、ストレージデバイス１５０の既存のデータを回復不能にする。既存のデータを上書きする要求は、データを安全上の理由から上書きする安全消去指令、データを消去するためのデータを上書きする要求、または、反復する同一キャラクタまたはキャラクタ列で既存のデータを置換しようとする要求などでもよい。一実施例においては、安全消去指令は、デバイスに、暗号化キーを安全に消去させ、既存のデータを安全に消去させることを両方行う。一実施例においては、暗号化キーの消去によって、ガベージコレクションプロセスがストレージ空き領域回復プロセスの一部としてデータを消去するまで、ストレージデバイスのデータの安全消去を延期することが可能となる。当分野の当業者は、暗号化キーを消去するその他の方法、および、既存のデータを上書きする要求を受信するその他の方法が分かるであろう。

【０２６１】

一実施例においては、トークン指令は、データセグメントトークンを具え、トークン指令送信モジュール１００４は、トークン指令とともにデータセグメントトークンを送信する。別の実施例においては、トークン指令は、データセグメントトークンを具えずに、データセグメントトークンを生成するためのストレージデバイス１５０用のコマンドを具える。この実施例においては、トークン指令送信モジュール１００４は、データセグメントトークンを生成するコマンドを有するトークン指令を送信し、データセグメントトークンを送信しない。

【０２６２】

この装置は、ストレージデバイス１５０にトークン指令を送信するトークン指令送信モジュール１００４を具える。一般的に、トークン指令送信モジュール１００４は、ストレージ要求の一部としてトークン指令を送信する。ストレージ要求は、オブジェクト要求の形態、データ要求、または当分野の当業者に周知の他の形態であってもよい。トークン指令生成モジュール１００２が安全消去指令を生成する場合、トークン指令送信モジュール１００４は、安全消去指令をストレージデバイス１５０に送信する。トークン指令生成モジュール１００２は、消去暗号化キーコマンドを生成する場合、必要がある場合には、コマンドを実行する別のデバイスに消去暗号化キーコマンドが送信される。

【０２６３】

一実施例においては、トークン指令送信モジュール１００４は、データセグメントトークンなしでトークン指令を送信する。この実施例においては、トークン指令は、データセグメントトークを生成するための、ストレージデバイス１５０に対する命令及び情報を具える。別の実施例においては、トークン指令送信モジュール１００４は、データセグメントトークンを具えるトークン指令を送信する。この実施例においては、ストレージデバイス１５０は、トークン指令と共に受信されたデータセグメントが、データセグメントを表し、適宜なアクションをとり、データセグメントトークンを記憶し、データセグメントトークンは、一般的なデータとしてデータセグメントトークンを単に記憶するのではないデータセグメントを表すことを認識できる。

【０２６４】

特定の実施例においては、装置は、ストレージデバイス１５０からデータセグメントを読み出すストレージ要求を受信する読み出し受信モジュール１００６と、ストレージデバイス１５０にストレージ要求を送信する読み出し要求送信モジュール１００８とを具える。一般的に、ストレージ要求は、外部クライアント１１４のような要求クライアント１１４、サーバ１１２で動くアプリケーションまたはファイルサーバのような、サーバ１１２内部のクライアント１１４等から受信される。当分野の当業者は、読み出し受信モジュール１００６がストレージ要求を受信できる要求クライアント１１４として機能するその他のデバイスおよびソフトウェアが分かるであろう。

【０２６５】

ストレージ要求は、トークン指令送信モジュール１００４によってストレージデバイス１５０に送信されるトークン指令に記憶されるように要求されたデータセグメントトーク

ンに対応する、データセグメントを読み出すための要求を具える。一実施例においては、要求クライアント１１４は、データセグメントがデータセグメントトークンの形態で記憶されたことを認識しない。別の実施例においては、要求デバイスは、データセグメントが、データセグメントトークンとして記憶されたことを認識するが、データセグメントトークンに記憶された情報を認識しない。

【０２６６】

特定の実施例においては、装置は更に、ストレージデバイスから要求されたデータセグメントトークンに対応するメッセージを受信する読み出しトークン指令受信モジュール１０１０を具え、このメッセージは、少なくともデータセグメント識別子およびデータセグメント長を具える。このメッセージは、一般的に、データセグメントのデータを含まない。更に、このメッセージは、データセグメント位置または反復する同一キャラクタ若しくはキャラクタ列のような、データセグメントトークンに記憶されたその他の情報を具えることができる。特定の実施例においては、この装置は、ストレージデバイス１５０から受信したメッセージから構築された要求クライアント１１３に、応答を送信する要求クライアント応答モジュール１０１２を具える。

【０２６７】

一実施例においては、読み出しトークン指令受信モジュール１０１０も、メッセージにおいて、既存のデータがキャラクタで上書きされたことの確認を受信し、既存のデータは回復不能となり、既存のデータは、ストレージデバイスに以前に記憶されており、メッセージで受信されたデータセグメントトークンからの同一データセグメント識別子として参照される。この確認は、データセグメントを読み出すためにストレージ要求から別個にストレージデバイス１５０から受信されてもよい。

【０２６８】

別の実施例においては、要求クライアント１１４は、データセグメントを要求する場合、この装置は、メッセージ中に含まれる情報を用いて、データセグメントのデータを復元するデータセグメント再生成モジュール１０１４を具える。この場合、要求クライアントに送信される応答は、復元されたデータセグメントを含む。別の実施例においては、要求クライアントに送信される応答は、ストレージデバイス１５０から受信したメッセージに含まれる情報を具える。要求クライアント１１４は、次いで、データセグメントを復元し、その他の方法において、この情報を用いることができる。別の実施例においては、メッセージは、データセグメントトークンを具える。データセグメントトークンは、要求クライアント１１４に転送する前にデータセグメントを復元するために、データセグメント再生成モジュール１０１４によって用いられ、あるいは、要求クライアント応答モジュール１０１２が、データセグメントトークンを単純に転送することができる。

【０２６９】

一実施例においては、トークン指令を有するストレージ要求は、更に、ストレージデバイス１５０にストレージ空き領域を割り当てる要求を具え、要求された、割り当てられた空き領域は、データセグメント長とほぼ同一のストレージ空き領域量である。別の実施例においては、ストレージ空き領域を割り当てる要求は、データセグメント長とは異なるストレージ空き領域量用である。例えば、ストレージデバイス１５０がソリッドステートストレージデバイス１０２である場合、ソリッドステートストレージデバイス１０２は、ハードドライブまたはその他の低価格の長期ストレージに接続でき、ソリッドステートストレージ１１０は、長期ストレージ用キャッシュとして構成される。ストレージを割り当てる要求は、ソリッドステートストレージデバイス１０２に、ソリッドステートストレージデバイス１０２にデータを書き込む準備をする際、長期ストレージに対するキャッシュの一部を消去させる。当分野の当業者は、所望されるストレージ空き領域を割り当てる要求であるその他の条件が分かるであろう。

【０２７０】

一実施例においては、装置は、読み出し受信モジュール１００６、読み出し要求送信モジュール１００８、読み出しトークン指令受信モジュール１０１０および要求クライアン

ト応答モジュール１０１２で提供され、これらは、上述したものと実質的に同等なものにできる。その実施例においては、モジュール１００６－１０１２は、トークン生成モジュール１００２またはトークン指令送信モジュール１００４を具える装置から独立していてもよい。一実施例においては、この装置は、データセグメント再生成モジュール１０１４を具え、これは、上述したデータセグメント再生成モジュール１０１４と実質的に同じである。

【０２７１】

図１１は、本発明によるトークン指令を生成し送信する方法１１００の一実施例を示した概略フローチャート図である。この方法１１００は開始し（１１０２）、トークン指令生成モジュール１００２は、トークン指令を有するストレージ要求を生成し（１１０４）、トークン指令は、ストレージデバイス１５０にデータセグメントトークンを記憶する要求を具える。トークン指令送信モジュール１００４は、ストレージデバイス１５０にトークン指令を送信し（１１０６）、方法１１００は終了する（１１０８）。一実施例においては、ストレージ要求は、データセグメントトークンを記憶するトークン指令を具え、ストレージ要求は、データセグメントのデータを実質的に含まない。別の実施例においては、ストレージ要求は、データセグメントからのデータを含む。好ましい実施例においては、ソフトウェアアプリケーションは、データセグメントの生成を回避するためにトークン指令を用いてストレージ要求を生成する。別の実施例においては、ソフトウェアアプリケーションは、トークン指令の生成を要求する。

【０２７２】

図１２は、本発明によるデータセグメントトークンを読み出す方法１２００の一実施例を示した概略フローチャート図である。この方法１２００は、開始し（１２０２）、読み出し受信モジュール１００６は、要求クライアント１１４から、ストレージデバイス１５０からのデータセグメントを読み出すストレージ要求を受信する（１２０４）。読み出し要求送信モジュール１００８は、ストレージデバイス１５０にストレージ要求を送信する（１２０６）。

【０２７３】

読み出しトークン指令受信モジュール１００８は、ストレージデバイス１５０から、要求されたデータセグメントトークンに対応するメッセージを受信し（１２０８）、このメッセージは、少なくともデータセグメント識別子およびデータセグメント長を具える。このメッセージは、データセグメントのデータを実質的に含まない。要求クライアント応答モジュール１０１２は、ストレージデバイス１５０から受信されたメッセージから構築された要求クライアントに対する応答を送信し（１２１０）、この方法１２００は、終了する（１２１２）。

【０２７４】

図１３は、本発明によるデータセグメントトークンを管理するための装置を有するシステム１３００の一実施例を示した概略ブロック図である。このシステム１３００は、書き込み要求受信モジュール１３０２およびデータセグメントトークンストレージモジュール１３０４を有する装置を具え、様々な実施例においては、トークン指令生成モジュール１３０６、読み出し要求受信モジュール１３０８、読み出しデータセグメントトークンモジュール１３１０、送信データセグメントトークンモジュール１３１４および送信データセグメントモジュール１３１６を有する読み出し要求応答モジュール１３１２、復元データセグメントモジュール１３１８、消去確認モジュール１３２２を有する安全消去モジュール１３２０、ストレージ空き領域割り当てモジュール１３２４、を具え、これらは、以下で説明する。システム１３００は、ストレージコントローラ１５２およびデータストレージデバイス１５４を有するストレージデバイス１５０を具え、これらは、上述したものと実質的に同じである。システム１３００は、要求デバイス１３２６を具え、これは以下に記載されているように、ストレージデバイス１５０と接続している。

【０２７５】

示された実施例においては、モジュール１３０２－１３２４は、ストレージデバイス１

５０またはストレージコントローラ１５２に含まれる。別の実施例においては、１又はそれ以上のモジュール１３０２－１３２４の少なくとも一部分は、ストレージデバイス１５０の外側に配置される。更なる実施例においては、要求デバイス１３２６は、ドライバ、ソフトウェアまたは１又はそれ以上のモジュール１３０２－１３２４のその他の機能の形態において、モジュール１３０２－１３２４の一部を具える。例えば、トークン生成モジュール１３０６および復元データセグメントモジュール１３１８は、要求デバイス１３２６中に示されている。当分野の当業者は、モジュール１３０２－１３２４の機能を分散し、実装するその他の方法が分かるであろう。

【０２７６】

この装置は、要求デバイス１３２６からのストレージ要求を受信する書き込み要求受信モジュール１３０２を具え、ストレージ要求は、ストレージデバイス１５０内にデータセグメントを記憶する要求を具える。データセグメントは、反復する同一キャラクタ又はキャラクタ列の配列を具える。一般的に、反復する同一キャラクタの配列は、データセグメントが空であることを意味する。これは特に、反復する同一キャラクタの配列が１又はゼロである場合に正しい。この装置は、ストレージデバイス１５０内にデータセグメントトークンを記憶するデータセグメントトークンストレージモジュール１３０４を具える。データセグメントトークンは、少なくともデータセグメント識別子およびデータセグメント長を具える。データセグメントトークンは、データセグメントからの実際のデータを実質的に含まない。

【０２７７】

データセグメントトークンは、多数の形態で記憶可能である。一実施例においては、インデックス中のエントリを具えており、このインデックスは、ストレージデバイス１５０に記憶された情報およびデータに対応する。例えば、このインデックスは、図２に示された装置２００に関連して上述したように、オブジェクトインデックスであってもよい。更に、このインデックスは、ファイルシステムインデックス、ブロックストレージインデックス、または、当分野の当業者に周知なその他のインデックスであってもよい。別の実施例においては、データセグメントトークンは、ストレージデバイス１５０に記憶されるメタデータを具え、あるいは、その形態である。別の実施例においては、セグメントトークンが、ストレージデバイスにメタデータとして記憶され、インデックス中のエントリを具える。当分野の当業者は、データセグメントトークンを記憶する他の方法が分かるであろう。

【０２７８】

一実施例においては、ストレージ要求は、データセグメントトークンを記憶するトークン指令を具え、このストレージ要求は、データセグメントのデータを実質的に含まない。トークン指令は、データセグメントトークンを生成するデータセグメントトークンまたはコマンドを具えることができる。トークン指令は、データセグメントトークンを具えていない場合、データトークンストレージモジュール１３０４は、トークン指令中の情報からデータセグメントトークンを生成する。トークン指令がデータセグメントトークンを具える場合は、データセグメントトークンストレージモジュール１３０４が、トークン指令中のデータセグメント識別子によって識別されるデータセグメントを表すデータ構造としてデータセグメントを認識し、適宜にデータセグメントトークンを記憶する。

【０２７９】

一般的に、データセグメントトークンストレージモジュール１３０４が、データセグメントトークンを認識する場合、データセグメントトークンは、ストレージデバイス１５０に記憶される他のデータとは何らかの方法で区別される。例えば、要求デバイス１３２６は、データを圧縮のみし、圧縮オブジェクト、ファイルまたはセグメントを送信するが、ストレージデバイス１５０は、他のストレージ要求によって受信された他のデータから圧縮データセグメントを区別できない。

【０２８０】

データセグメントトークンストレージモジュール１３０４は、受信されたデータセグメ

ントトークンがデータセグメントトークンであることを認識する場合、データセグメントトークンストレージモジュール１３０４は、読み出すとき、データセグメントトークンが、データセグメントトークンではなくデータセグメントとして存在するような方法で、データセグメントトークンを記憶することができる。当分野の当業者は、受信したデータセグメントトークンが、データセグメントではなくデータセグメントトークンであると認識した後に、データセグメントトークンストレージモジュール１３０４がデータセグメントトークンを記憶することができるような他の方法が分かるであろう。

【０２８１】

別の実施例においては、ストレージ要求は、データセグメントからのデータを具える。この実施例においては、この装置は、データセグメントからデータセグメントトークンを生成するトークン生成モジュール１３０６を具え、データセグメントトークンは、データセグメントを記憶するストレージ要求に応答して生成される。更なる実施例においては、トークン生成モジュール１３０６は、可能なドライバ形態において、要求デバイス１３２６中に常駐する。

【０２８２】

一実施例においては、この装置は、既存のデータが回復不能であるように、キャラクタを用いて既存のデータを上書きする安全消去モジュール１３２０を具え、既存のデータは、ストレージ要求で識別されたデータセグメントと同一のデータセグメント識別子で識別された、ストレージデバイスで既に記憶されたデータセグメントのデータを具える。この実施例においては、データセグメントトークンは、データセグメント識別子およびデータセグメント長とともに記憶され、データセグメントトークンに記憶された同一のデータセグメント識別子によって識別される既存のデータは、既存のデータを上書きすることで消去される。一般的に、既存のキャラクタは、ゼロ、１、または、その他のキャラクタ列によって上書きされ、データは破壊され、回復不能となる。

【０２８３】

さらなる実施例においては、安全消去モジュールは、既存のデータが上書きされることを示すメッセージを送信する消去確認モジュール１３２２を更に具える。一般的に、メッセージは、要求デバイス１３２６に送信される。消去確認メッセージは、安全消去モジュール１３２０が既存のデータを上書きした後に送信される。このメッセージは、ストレージ要求として同一のトランザクションにおいて、あるいは、異なるトランザクションにおいて、送信されてもよい。

【０２８４】

別の実施例においては、安全消去モジュール１３２０は、ストレージ空き領域回復動作中に、既存のデータを上書きする。例えば、ストレージデバイス１５０がソリッドステートストレージデバイス１０２である場合、上述したように、ストレージ空き領域回復動作は、図８に示した装置８００に関して記載されたガベージコレクションに関連してもよい。しかしながら、既存のデータを上書きする要求を含むストレージ空き領域回復動作は、一般的に促進され、既存のデータが記憶されるストレージ位置が、確認メッセージが消去確認モジュール１３２２によって送信される前に、必ず回復される。一実施例においては、既存のデータは、安全消去が要求されていることを示すように、標識され、あるいは、識別される。一般的に、確認メッセージは、消去を標識した既存のデータが上書きされて回復不能になるまで、送信されない。別の実施例においては、安全消去モジュール１３２０は、後のストレージ空き領域回復のために無効であることを既存のデータに標識するのみである。別の実施例においては、安全消去は、既存のデータが無効であることを示すインデックスを更新し、後のストレージ空き領域回復中にデータが上書きされるまで、このデータへのアクセスを阻止する。

【０２８５】

一実施例においては、安全消去モジュール１３２０は、データセグメントが記憶される毎に、既存のデータを上書きする。別の実施例においては、既存のデータを上書きする要求を特に具え、安全消去モジュール１３２０は、既存のデータを上書きする要求に応答し

て、既存のデータを上書きする。別の実施例においては、安全消去モジュール１３２０は、既存のデータが消去されたという確認に関するメタデータ情報を記憶し、次の読み出しが消去を指示することができる。

【０２８６】

安全消去が受信されない場合の他の実施例において、既存のデータが削除される。一実施例においては、データを削除することは、インデックスエントリ、アドレスなどを削除することを具える。好ましい実施例においては、データセグメントトークンが記憶されるとき、対応する既存のデータは、無効あるいはストレージ回復の準備ができていると標識される。このデータは、ストレージ回復動作またはガベージコレクション動作などにおいて後で回復可能である。

【０２８７】

特定の実施例においては、この装置は、データセグメントを読み出すためのストレージ要求を受信する読み出し要求受信モジュール１３０８、ストレージ要求によって要求されるデータセグメントに対応するデータセグメントトークンを読み出す読み出しデータセグメントトークンモジュール１３１０、および、要求デバイス１３２６に対する応答を送信する読み出し要求応答モジュール１３１２を具える。この応答は、要求されたデータセグメントに対応するデータセグメントトークンを用いて生成される。

【０２８８】

一実施例において、データセグメントを読み出すストレージ要求は、ストレージ要求に関連しており、ストレージ要求が成功したことを確認するのに役立つ。別の実施例においては、データセグメントを読み出す要求は、ストレージ要求から独立しており、ストレージ要求を生成した要求デバイス１３２６または別の別個の要求デバイス１３２６によって開始されてもよい。

【０２８９】

一実施例においては、要求デバイスが、実際のデータの位置にあるデータセグメントトークンからの情報を受信する場合、読み出し要求応答モジュール１３１２は、要求デバイス１３２６へのメッセージに応答して送信する送信データセグメントトークンモジュール１３１４を具える。このメッセージは、少なくともデータセグメント識別子およびデータセグメント長を具えるが、更に、データセグメント位置、少なくとも１インスタンスの反復する同一キャラクタ又はキャラクタ列、あるいは、その他の関連情報を具えてもよい。一般的に、メッセージは、データセグメントトークンに含まれるもの以外に、データセグメントの実際のデータを具えていない。

【０２９０】

別の実施例においては、要求デバイス１３２６は、データセグメントを受信するとされる場合、この装置は、データセグメントトークンを用いて、データセグメントのデータを復元する復元データセグメントモジュール１３１８を具える。読み出し要求応答モジュール１３１２は、更に、要求デバイス１３２６に復元された、要求されたデータセグメントを送信する、送信データセグメントモジュール１３１６を具える。別の実施例において、復元データセグメントモジュール１３１８は、可能であれば、ドライバの形態において、要求デバイス１３２６に常駐し、送信データセグメントトークモジュール１３１４は、データセグメントトークン情報を有するメッセージを要求デバイス１３２６に送信する。復元データセグメントモジュール１３１８は、要求デバイス１３２６で、メッセージから要求されたデータセグメントを復元する。

【０２９１】

一実施例においては、システム１３００は、読み出し要求受信モジュール１３０８、読み出しデータセグメントトークンモジュール１３１０、読み出し要求応答モジュール１３１２を含む別の装置を具え、これらは、上述したものと実質的に同一である。この装置は、書き込み要求受信モジュール１３０２およびデータセグメントトークンストレージモジュール１３０４を具える装置から独立していてもよい。一実施例においては、読み出し要求応答モジュール１３１２は、送信データセグメントトークンモジュール１３１４および

／または送信データセグメントモジュール１３１６を具え、この装置は、復元データセグメントモジュール１３１８を具えてもよく、モジュール１３１４、１３１６、１３１８が上述したものと実質的に同じである。

【０２９２】

図１４は、本発明によるデータセグメントトークンを記憶する方法１４００の一実施例を示す概略フローチャート図である。この方法１４００は、開始し（１４０２）、書き込み要求受信モジュール１３０２が、要求デバイス１３２６からのストレージ要求を受信し（１４０４）、上記ストレージ要求は、ストレージデバイス１５０にデータセグメントを記憶する要求を具える。データセグメントは、反復する同一キャラクタ又はキャラクタ列の配列を具える。データセグメントトークンストレージモジュール１３０４は、ストレージデバイス１５０にデータセグメントトークンを記憶し（１４０６）、方法１４００は終了する（１４０８）。データセグメントトークンは、少なくともデータセグメント識別子およびデータセグメント長を具え、データセグメントトークンは、ほとんどの部分に関してデータセグメントからのデータを具えていない。

【０２９３】

図１５は、本発明によるデータセグメントトークンを読み出す方法１５００の一実施例を示した概略フローチャート図である。この方法１５００は開始し（１５０２）、読み出し要求受信モジュール１３０８が、ストレージデバイス１５０からデータセグメントを読み出すストレージ要求を受信する（１５０４）。データセグメントは、データセグメントトークンによってストレージデバイスに表現され、このデータセグメントは、反復する反復する同一キャラクタ又はキャラクタ列の配列を具える。このデータセグメントトークンは、少なくともデータセグメント識別子およびデータセグメント長を具え、データセグメントトークンは、データセグメントからのデータを具えていない。読み出しデータセグメントトークンモジュール１３１０は、ストレージ要求において要求されたデータセグメントに対応するデータセグメントを読み出し（１５０６）、読み出し要求応答モジュール１３１２は、要求デバイス１５０に応答を送信し（１５０８）、この方法１５００は終了する（１５１０）。この応答は、要求されたデータセグメントに対応するデータセグメントトークンを用いて生成される。

【０２９４】

進行型ＲＡＩＤ

【０２９５】

独立ドライブ冗長アレイ（「ＲＡＩＤ」）は様々な目的を達成するために多くの方法で構成される。本明細書に記載のように、ドライブはデータ用の大容量ストレージデバイスである。ドライブ又はストレージデバイスは、ソリッドステートストレージ１１０、ハードディスクドライブ（「ＨＤＤ」）、テープドライブ、光学ドライブ又は当該技術分野の当業者に公知のその他の大容量ストレージデバイスにできる。一実施例においては、ドライブは仮想容量としてアクセスされる大容量ストレージデバイスの一部分を具える。別の実施例においては、ドライブは仮想容量として共にアクセス可能で、かつ、ＲＡＩＤとして、「ただのディスク／ドライブの束」（「ＪＢＯＤ」）」として等でストレージエリアネットワーク（「ＳＡＮ」）に構成される２又はそれ以上のデータストレージデバイスを具える。一般的には、ドライブはストレージコントローラ１５２を通して単一ユニット又は仮想容量としてアクセスされる。好ましい実施例においては、ストレージコントローラ１５２はソリッドステートストレージコントローラ１０４を具える。当該技術分野の当業者は、ＲＡＩＤに構成されうる大容量ストレージデバイスの形態における、他のドライブの形態が分かるであろう。本明細書に記載のそれらの実施例においては、ドライブ及びストレージデバイス１５０は互換的に用いられる。

【０２９６】

従来的には、様々なＲＡＩＤ構成はＲＡＩＤレベルと呼ばれる。１の基本ＲＡＩＤ構成はストレージデバイス１５０のミラーコピーを生成するＲＡＩＤレベル０である。ＲＡＩＤ０の利点は、１又はそれ以上のストレージデバイス１５０のデータの完全なコピーが１

又はそれ以上のストレージデバイス１５０のミラーコピーで更に利用可能であり、プライマリドライブ又はミラーリングドライブのデータの読み出しが比較的高速であることである。ＲＡＩＤ０は更にプライマリストレージデバイス１５０における故障の場合に、データのバックアップコピーを更に提供する。ＲＡＩＤ０の欠点は、書き込みデータが二度書き込まれなければならないため、書き込みが比較的低速であることである。

【０２９７】

別の従来のＲＡＩＤ構成はＲＡＩＤレベル１である。ＲＡＩＤ１においては、ＲＡＩＤに書き込まれるデータがストレージデバイス１５０のセットにおけるＮ個のストレージデバイス１５０に対応するＮ個のデータセグメントに分割される。Ｎ個のデータセグメントは「ストライプ」を形成する。複数のストレージデバイス１５０にわたってデータをストライピングすることによって、単一のストレージデバイス１５０がＮ個のデータセグメントを含むデータを保存できるよりも高速に、Ｎ個のデータセグメントを記憶するのに並行してストレージデバイス１５０が動くため、性能が向上する。しかしながら、データが複数のストレージデバイス１５０にわたって分散し、複数のストレージデバイス１５０のアクセス時間が総ての所望のデータを含むストレージデバイス１５０からデータの読み出しよりも一般的に低速であるため、データを読み出すのは比較的低速である。更には、ＲＡＩＤ１は故障の保護を提供しない。

【０２９８】

普及しているＲＡＩＤ構成は、Ｎ個のストレージデバイス１５０にわたるＮ個のデータセグメントのストライピングするステップと、Ｎ＋１ストレージデバイス１５０でパリティデータセグメントを記憶するステップとを具えるＲＡＩＤレベル５である。そのＲＡＩＤは、ストレージデバイス１５０の単一の故障を許容するため、ＲＡＩＤ５は故障耐性を提供する。例えば、ストレージデバイス１５０が故障した場合、ストライプのデータセグメントの欠損は、他の利用可能なデータセグメント及びストライプ用に特に算出されたパリティデータセグメントを用いて生成されうる。更に、ＲＡＩＤ５は一般的には、ＲＡＩＤされたストレージデバイス１５０のセットの各ストレージデバイス１５０がデータの完全なコピーではなく、ストライプのデータセグメント又はパリティデータセグメントのみを記憶することを要求するため、ＲＡＩＤ０より少ないストレージ空き領域を用いる。ＲＡＩＤ１と同様、ＲＡＩＤ５はデータの書き込みが比較的高速であるが、データを読み出すのが比較的低速である。しかしながら、パリティデータセグメントはストライプのＮ個のデータセグメントから各ストライプ用に算出しなければならないため、一般的な従来のＲＡＩＤ５用のデータの書き込みは、ＲＡＩＤ１よりも低速である。

【０２９９】

別の普及しているＲＡＩＤ構成は二重分散パリティを具えるＲＡＩＤレベル６である。ＲＡＩＤ６においては、２のストレージデバイス１５０がパリティミラーデバイス（例えば、１６０２ａ、１６０２ｂ）として割り当てられる。ストライプ用の各パリティデータセグメントは別個に算出されて、ストレージデバイスセットのいずれかの２のストレージデバイス１５０の欠損が残りの利用可能なデータセグメント及び／又はパリティデータセグメントを用いて回復可能である。ＲＡＩＤ６はＲＡＩＤ５と同様の性能の利点及び欠点を有する。

【０３００】

入れ子構造ＲＡＩＤは、高信頼性が要求される場合に故障耐性を増大させるのに更に用いられうる。例えば各々がＲＡＩＤ５として構成される２のストレージデバイスセットがＲＡＩＤ０構成でミラーリングされうる。生じた構成はＲＡＩＤ５０と呼ばれる。ＲＡＩＤ６が各々のミラーリングセットで用いられる場合、その構成はＲＡＩＤ６０と呼ばれる。入れ子構造ＲＡＩＤ構成は基礎をなすＲＡＩＤグループと同様の性能の問題を一般的に有している。

【０３０１】

前述の記載から、故障耐性の利益と、ＲＡＩＤ０、ＲＡＩＤ５、ＲＡＩＤ６のような従来の故障耐性のあるＲＡＩＤレベルよりも高速なデータ書き込み等を提供する一方、ＲＡ

ＩＤ１、ＲＡＩＤ５、ＲＡＩＤ６等のような従来のストライピングのＲＡＩＤレベルよりも高速なデータ読み出しを更に提供する進行型ＲＡＩＤ用の装置、システム及び方法に対するニーズが存在することは明らかであろう。有益には、このような装置、システム及び方法は、パリティデータセグメントが、ストレージ統合動作前又はストレージ統合動作の一部のように算出されるの要求されるまで、Ｎ個のデータセグメントをパリティミラーストレージデバイス１６０２に書き込み、ＲＡＩＤ０システムの利点を提供する。

【０３０２】

図１０は本発明による進行型ＲＡＩＤ用のシステム１６００の一実施例を示す概略ブロック図である。システム１６００は１又はそれ以上のクライアント１１４によりコンピュータネットワーク１１６を通してアクセス可能なＮ個のストレージデバイス１５０と、Ｍ個のパリティミラーストレージデバイス１６０２とを具える。Ｎ個のストレージデバイス１５０及びパリティミラーストレージデバイス１６０２は、１又はそれ以上のサーバ１１２に配置できる。ストレージデバイス１５０、サーバ１１２、コンピュータネットワーク１１６、及びクライアント１１４は実質的に上述のものと同様である。パリティミラーデバイス１６０２は一般的にＮ個のストレージデバイス１５０と同様または同一であり、一般的にストライプ用のパリティミラーストレージデバイス１６０２として指定される。

【０３０３】

一実施例においては、Ｎ個のストレージデバイス１５０及びＭ個のパリティミラーストレージデバイス１６０２は、１のサーバ１１２に含まれるか、それを通してアクセス可能であり、システムバス、ＳＡＮ、又は他のコンピュータネットワーク１１６を用いて互いにネットワーク化できる。別の実施例においては、Ｎ個のストレージデバイス１５０及びＭ個のパリティミラーストレージデバイス１６０２は複数のサーバ１１２ａ－ｎ＋ｍに配置され、それを通してアクセス可能にできる。例えば、ストレージデバイス１５０及びパリティミラーストレージデバイス１６０２は、図１Ｃのシステム１０３及び図５Ｂの方法１０５に関連して上述したようなインサーバＳＡＮの一部にできる。

【０３０４】

一実施例においては、パリティミラーストレージデバイス１６０２は進行型ＲＡＩＤに記憶されたストライプの総てのパリティデータセグメントを記憶する。別の好ましい実施例においては、進行型ＲＡＩＤに割り当てられたストレージデバイスセット１６０４のストレージデバイス１５０は、特定のストライプ用のパリティミラーストレージデバイス１６０２であるように割り当てられ、その割り当てはローテーションし、パリティデータセグメントは各ストライプについて、Ｎ＋Ｍ個のストレージデバイス１５０の間でローテーションする。この実施例は、単一のストレージデバイス１５０を割り当てて、各ストライプ用のパリティミラーストレージデバイス１６０２にするのを介して、性能の利点を提供する。パリティミラーストレージデバイス１６０２をローテーションさせることによって、パリティデータセグメントを算出及び記憶することに関するオーバヘッドは、分散される。

【０３０５】

一実施例においては、ストレージデバイス１５０は各々が関連するソリッドステートストレージ１１０及びソリッドステートストレージコントローラ１０４を有するソリッドステートストレージデバイス１０２である。別の実施例においては、各ストレージデバイス１５０はソリッドステートストレージコントローラ１０４を具え、関連するソリッドステートストレージ１１０はテープストレージ又はハードディスクドライブのような低価格、低性能のストレージ用のキャッシュとして動作する。別の実施例においては、１又はそれ以上のサーバ１１２はストレージ要求を進行型ＲＡＩＤに送信する１又はそれ以上のクライアント１１４を具える。当該技術分野の当業者は、進行型ＲＡＩＤ用に構成されうるＮ個のストレージデバイス１５０と、１又はそれ以上のパリティミラーストレージデバイス１６０２とを有する他のシステム構成が分かるであろう。

【０３０６】

図１７は、本発明による進行型ＲＡＩＤ用の装置１７００の一実施例を示す概略ブロッ

ク図である。様々な実施例においては、装置１７００はストレージ要求受信モジュール１７０２、ストライピングモジュール１７０４、パリティミラーモジュール１７０６、パリティ進行モジュール１７０８、パリティ変更モジュール１７１０、ミラーリング設定モジュール１７１２、更新モジュール１７１４、直接クライアント応答モジュール１７１８を有するミラーリング復元モジュール１７１６、事前統合モジュール１７２０、事後統合モジュール１７２２、データ再構築モジュール１７２４、及びパリティ再構築モジュール１７２６を具え、以下に述べられている。モジュール１７０２－１７２６はサーバ１１２に示されるが、モジュール１７０２－１７２６の機能の一部又は総ては複数のサーバ１１２、ストレージコントローラ１５２、ストレージデバイス１５０、クライアント１１４に更に分散されうる。

【０３０７】

装置１７００はデータを記憶する要求を受信するストレージ要求受信モジュール１７０２を具え、データはファイル又はオブジェクトのデータである。一実施例においては、ストレージ要求はオブジェクト要求である。別の実施例においては、ストレージ要求はブロックストレージ要求である。一実施例においては、ストレージ要求はデータを含まず、ストレージデバイス１５０及びパリティミラーストレージデバイス１６０２によって、クライアント１１４又は他のソースからのＤＭＡ又はＲＤＭＡデータに対して用いられうるコマンドを含む。別の実施例においては、ストレージ要求はストレージ要求の結果として記憶されるべきデータを含む。別の実施例においては、ストレージ要求はストレージデバイスセット１６０４に記憶されるデータを有することが可能な１のコマンドを含む。別の実施例においては、ストレージ要求は複数のコマンドを含む。当該技術分野の当業者は進行型ＲＡＩＤに好適なデータを記憶する他のストレージ要求が分かるであろう。

【０３０８】

データは装置１７００にアクセス可能な位置に記憶される。一実施例においては、データはクライアント１１４又はサーバによって用いられるＲＡＭのようなランダムアクセスメモリ（「ＲＡＭ」）で利用可能である。別の実施例においては、データはハードディスクドライブ、テープストレージ、又は他の大容量ストレージデバイスに記憶される。一実施例においては、データはオブジェクトとして、又は、ファイルとして構成される。別の実施例においては、オブジェクト又はファイルの一部分であるデータブロックとして構成される。当該技術分野の当業者はストレージ要求の対象であるデータの他の形式及び位置が分かるであろう。

【０３０９】

装置１７００はデータ用のストライプパターンを算出するストライピングモジュール１７０４を具える。ストライプパターンは１又はそれ以上のストライプを含み、各ストライプはＮ個のデータセグメントのセットを具える。一般的に、ストライプのデータセグメントの数はどのくらいの数のストレージデバイス１５０がＲＡＩＤグループに割り当てられるかに依存する。例えば、ＲＡＩＤ５が用いられた場合、１のストレージデバイス１５０はパリティミラーストレージデバイス１６０２ａとして割り当てられて、特定のストライプ用のパリティデータを記憶する。４の他のストレージデバイス１５０ａ、１５０ｂ、１５０ｃ、１５０ｄがＲＡＩＤグループに割り当てられた場合、ストライプはパリティデータセグメントに加えた、４のデータセグメントを有する。ストライピングモジュール１７０４はＮ個のデータセグメントを、Ｎ個のストレージデバイス１５０ａ－ｎに対するＮ個のストライプに書き込み、Ｎ個のデータセグメントの各々は、ストライプに割り当てられたストレージデバイス１５０のセット１６０４内の別個のストレージデバイス１５０ａ、１５０ｂ、．．．、１５０ｎに書き込まれる。当該技術分野の当業者は、特定のＲＡＩＤレベルのＲＡＩＤグループに割り当てられうるストレージデバイス１５０の様々な組合せ、及びストライピングパターンを生成し、ストライプごとにＮ個のデータセグメントにデータを分割する方法が分かるであろう。

【０３１０】

装置１７００は、ストライプのＮ個のデータセグメントのセットを、ストレージデバイ

スセット１６０４内の１又はそれ以上のパリティミラーストレージデバイス１６０２に書き込むパリティミラーモジュール１７０６を具え、パリティミラーストレージデバイス１６０２がＮ個のストレージデバイス１５０に加えて存在する。Ｎ個のデータセグメントはその後パリティデータセグメントの将来の算出のために利用できる。パリティデータセグメントをすぐに算出するのではなく、パリティミラーモジュール１７０６はＮ個のデータセグメントのセットをパリティミラーストレージデバイス１６０２にコピーし、一般的にＮ個のデータセグメントを記憶するよりも少ない時間を必要とする。Ｎ個のデータセグメントが一度パリティミラーストレージデバイス１６０２に記憶された場合、Ｎ個のデータセグメントは、Ｎ個のストレージデバイス１５０のうちの１つが利用不可能になる場合に、データを復元するために読み出し、使用するために利用可能である。Ｎ個のデータセグメントの総てが１のストレージデバイス（例えば、１６０２ａ）から一緒に利用できるために、データの読み出しは更にＲＡＩＤ０構成の利点を有する。１以上のパリティミラーストレージデバイス（例えば、１６０２ａ、１６０２ｂ）のために、パリティミラーモジュール１７０６はＮ個のデータセグメントを各パリティミラーストレージデバイス１６０２ａ、１６０２ｂにコピーする。

【０３１１】

装置１７００は、ストレージ統合動作に応じてストライプ用の１又はそれ以上のパリティデータセグメントを算出するパリティ進行モジュール１７０８を具える。Ｎ個のデータセグメントから算出される１又はそれ以上のパリティデータセグメントは、パリティミラーストレージデバイス１６０２に記憶される。パリティ進行モジュール１７０８は、１又はそれ以上のパリティミラーストレージデバイス１６０２の各々に、パリティデータセグメントを記憶する。ストレージ統合動作は１又はそれ以上のパリティミラーストレージデバイス１６０２のうちの少なくとも１つの、少なくともストレージ空き領域又はデータ又は両方を回復するように導かれる。例えば、ストレージ統合動作は図８及び９の装置８００及び方法９００に関連して上述したような、ソリッドステートストレージデバイス１０２のデータガベージコレクションにできる。ストレージ統合動作はストレージ空き領域を増加させるためにデータを統合する、ハードディスクドライブ用のデフラグ動作又は他の同様の動作を更に含みうる。本明細書で用いられるように、ストレージ統合動作はデータを回復するための動作を含み、例えば、ストレージデバイス１５０が利用不可能である場合、エラー、又は、パリティミラーストレージデバイス１６０２からデータを読み出す他の理由から回復させる。別の実施例においては、パリティ生成モジュール１７０８は、パリティミラーストレージデバイス１６０２が混雑が少ない場合には、パリティデータセグメントを単に算出する。

【０３１２】

有利には、ストライプのパリティデータセグメントの算出及び記憶を遅らせることにより、パリティミラーストレージデバイス１６０２のＮ個のデータセグメントは、データセグメントを読み出すため、データを回復するため、より多くのストレージ空き領域がパリティミラーストレージデバイス１６０２に必要となるまでにデータを再構築するために、又はストレージ統合動作に対する他の理由のために利用可能である。パリティ進行モジュール１７０８はその時に、ストレージ要求受信モジュール１７０２、ストライピングモジュール１７０４、又はパリティミラーモジュール１７０６から自律して、バックグラウンド動作として実行できる。当該技術分野の当業者は、進行型ＲＡＩＤ動作の一部としてパリティデータセグメントの算出を遅らせる他の理由が簡単に分かるであろう。

【０３１３】

一実施例においては、データを記憶する要求を受信し、ストライプパターンを算出してＮ個のデータセグメントをＮ個のストレージデバイスに書き込み、Ｎ個のデータセグメントのセットをパリティミラーストレージデバイスに書き込み、パリティデータセグメントを算出するモジュール１７０２－１７０８の機能の一部又は総てが、ストレージデバイスセット１６０４のストレージデバイス１５０、クライアント１１４及びサードパーティのＲＡＩＤ管理デバイスに生じる。サードパーティのＲＡＩＤ管理デバイスはサーバ１１４

又はその他のコンピュータにできる。

【０３１４】

一実施例においては、装置１７００は、各ストライプのために、ストレージデバイスセット１６０４内のストレージデバイス１５０のどちらがストライプ用の１又はそれ以上のパリティミラーストレージデバイス１６０２になるように割り当てられるかを交互に代替するパリティ代替モジュール１７１０を具える。図１０のシステム１６００に関連して上述したように、ストライプ用のパリティミラーストレージデバイスのためにどちらのストレージデバイス１５０が用いられるかをローテーションすることによって、様々なパリティデータセグメントの作業領域算出が、ストレージデバイスセット１６０４のストレージデバイス１５０の間で分散される。

【０３１５】

別の実施例においては、ストレージデバイスセット１６０４は第１のストレージデバイスセットであり、装置１７００は第１のストレージセット１６０４に加えて１又はそれ以上のストレージデバイスセットを生成するミラーリング設定モジュール１７１２を具えて、１又はそれ以上の追加のストレージデバイスセットの各々が、１又はそれ以上の追加のストレージセットの各々のＮ個のストレージデバイス１５０にＮ個のデータセグメントを書き込む、少なくとも１の付随するストライピングモジュール１７０４を具えるようにする。関連する実施例においては、１又はそれ以上の追加のストレージデバイスセットの各々は、Ｎ個のデータセグメントのセットを記憶するための付随したパリティミラーモジュール１７０６と、１又はそれ以上のパリティデータセグメントを算出するためのパリティ進行モジュール１７０８とを具える。ミラーリング設定モジュール１７１２が１又はそれ以上のミラーリングストレージデバイスセットを生成する場合、ＲＡＩＤはＲＡＩＤ５０のような入れ子構造ＲＡＩＤにできる。この実施例においては、ＲＡＩＤレベルはＲＡＩＤ１０から進行され、データはＲＡＩＤ５０又はＲＡＩＤ６０までストライピング及びミラーリングされ、パリティデータセグメントは各ストレージデバイスセット１６０４用に算出及び記憶される。

【０３１６】

一実施例においては、装置１７００は更新モジュール１７１４を具える。更新モジュール１７１４はパリティミラーストレージデバイス１６０２のＮ個のデータセグメントがパリティデータセグメントに進行されなかった場合に、一般的に用いられる。更新モジュール１７１４は更新されたデータセグメントを受信し、更新されたデータセグメントはＮ個のストレージデバイス１５０に記憶されるＮ個のデータセグメントの既存のデータセグメントに対応する。更新モジュール１７１４は更新されたデータセグメントを、既存のデータセグメントが記憶されるストライプのストレージデバイス１５０に、及び、ストライプの１又はそれ以上のパリティミラーストレージデバイス１６０２にコピーする。更新モジュール１７１４はＮ個のストレージデバイス１５０ａ－ｎのストレージデバイス１５０に記憶された既存のデータセグメントを、更新されたデータセグメントと置換し、１又はそれ以上のパリティミラーストレージデバイス１６０２に記憶された対応する既存のデータセグメントを更新されたデータセグメントと置換する。

【０３１７】

一実施例においては、データセグメントを置換するステップは、データセグメントをストレージデバイス１５０に書き込むステップと、次いで後のガベージコレクションのために、対応するデータセグメントを無効と標識するステップとを含む。この実施例のある例が、図８及び９に関連して上述されたソリッドステートストレージ１１０及びガベージコレクション装置のために記載されている。別の実施例においては、データセグメントを置換するステップは更新されたデータセグメントで既存のデータセグメントを上書きするステップを含む。

【０３１８】

一実施例においては、ストレージデバイス１６０４のセットは第１のストレージデバイスセットであり、装置１７００は第１のストレージセット１６０４のストレージデバイス

１５０に記憶されたデータセグメントを回復するミラーリング復元モジュール１７１６を具え、第１のストレージセット１６０４のストレージデバイス１５０は利用不可能になる。データセグメントはデータセグメントのコピーを含むミラーストレージデバイスから回復される。ミラーストレージデバイスはＮ個のデータセグメントのコピーを記憶する１又はそれ以上のストレージデバイス１５０のセットの１つを具える。

【０３１９】

更なる実施例においては、ミラーリング復元モジュール１７１６は、データセグメントを読み出す、クライアント１１４からの読み出し要求に応じてデータセグメントを回復する。別の関連する実施例においては、ミラーリング復元モジュール１７１６は、ミラーストレージデバイスからクライアント１１４に要求されたデータセグメントを送信する直接クライアント応答モジュール１７１８を更に具える。この実施例においては、要求されたデータセグメントはクライアント１１４にコピーされ、クライアント１１４はデータセグメントをクライアント１１４に送信する前にデータセグメントが回復されるまでクライアント１１４は待つ必要はない。

【０３２０】

一実施例においては、装置１７００は、データセグメントを読み出す要求に応じて、ストレージセット１６０４のストレージデバイス１５０に記憶されるデータセグメントを回復する事前統合復元モジュールを具える。その実施例においては、ストレージデバイス１５０は利用不可能であり、パリティ進行モジュール１７０８が１又はそれ以上のパリティミラーストレージデバイス１６０２に１又はそれ以上のパリティデータセグメントを生成する前に、データセグメントはパリティミラーストレージデバイス１６０２から回復される。

【０３２１】

別の実施例においては、装置１７００はストレージセットのストレージデバイス１５０に記憶されるデータセグメントを回復する事後統合復元モジュール１７２４を具える。その実施例においては、ストレージデバイス１５０は利用不可能であり、パリティ進行モジュール１７０８が１又はそれ以上のパリティデータセグメントを生成した後に、パリティミラーストレージデバイス１５０の１又はそれ以上に記憶される１又はそれ以上のパリティデータセグメントを用いて、データセグメントは回復される。例えば、事後統合復元モジュール１７２４はパリティデータセグメント及び利用可能なＮ個のストレージデバイス１５０の利用可能なデータセグメントを用いて、欠損したデータセグメントを再生成する。

【０３２２】

一実施例においては、装置１７００は再構築動作中の置換ストレージデバイスの回復したデータセグメントを記憶するデータ再構築モジュール１７２４を具え、回復したデータセグメントは利用不可能なストレージデバイス１５０に記憶された利用不可能なデータセグメントと合致する。利用不可能なストレージデバイス１５０はストレージデバイスセット１６０２のＮ個のストレージデバイス１５０のうちの１つである。一般的には、再構築動作は利用不可能なデータセグメントを記憶するストレージデバイス１５０の故障後に生じる。再構築動作はデータセグメントを置換ストレージデバイスに復元することであり、利用不可能なストレージデバイス１５０に既に記憶されたデータセグメントと合致させる。

【０３２３】

データセグメントは、いくつかのソースからの再構築動作で回復されうる。例えば、合致するデータセグメントがパリティミラーストレージデバイス１６０２に常駐する場合、進行の前にパリティミラーストレージデバイス１６０２からデータセグメントが回復されうる。別の例においては、データセグメントは利用不可能なデータセグメントのコピーを含むミラーストレージデバイスから回復されうる。一般的には、回復したデータセグメントが１又はそれ以上のパリティミラーストレージデバイス１６０２に常駐しない場合、データセグメントはミラーストレージデバイスから回復されるが、合致するデータセグメン

トがミラーストレージデバイスで利用可能である場合でさえも、ミラーストレージデバイスから回復されうる。

【０３２４】

別の例においては、回復したデータセグメントがパリティミラーストレージデバイス１６０４又はミラーストレージデバイスに常駐しない場合、再生成されるデータセグメントは１又はそれ以上のパリティデータセグメント及びＮ個のデータセグメントの利用可能なデータセグメントから再生成される。一般的には、欠損したデータセグメントは、いくつかの形態で別のストレージデバイス１５０に存在しない場合のみにのみ再生成される。

【０３２５】

別の実施例においては、装置１７００はパリティ再構築動作中に置換ストレージデバイスで回復したパリティデータセグメントを再構築するパリティ再構築モジュール１７２６を具え、回復したパリティデータセグメントは、利用不可能なパリティミラーストレージデバイスに記憶された、利用不可能なパリティデータセグメントに合致させる。利用不可能なパリティミラーストレージデバイスは、１又はそれ以上のパリティミラーストレージデバイス１６０２のうちの１つである。パリティ再構築動作は置換ストレージデバイスにパリティデータセグメントを復元して、利用不可能なパリティミラーストレージデバイスに前に記憶されたパリティデータセグメントに合致させる。

【０３２６】

再構築動作中に回復したパリティデータセグメントを再生成するために、再構築用に用いられたデータは様々なソースを元にできる。一例においては、回復したパリティデータセグメントは、ストライプのミラーコピーを記憶するストレージデバイス１５０の第２のセット内の、パリティミラーストレージデバイス１６０２に記憶されたパリティデータセグメントを用いて回復される。ミラーコピーが利用可能な場合、回復したパリティデータセグメントは再算出する必要がないため、ミラーリングパリティデータセグメントを用いることが所望される。別の例においては、Ｎ個のデータセグメントがＮ個のストレージデバイスで利用可能な場合、回復したパリティデータセグメントは、Ｎ個のストレージデバイス１５０のうちの１つに記憶されたＮ個のデータセグメントから再生成される。一般的には、単一の故障が、パリティミラーストレージデバイス１６０２が再構築される際に生じた場合、Ｎ個のデータセグメントはＮ個のストレージデバイス１５０で利用可能となる。

【０３２７】

別の例においては、Ｎ個のデータセグメントのうちの１又はそれ以上が第１のストレージデバイスセット１６０４のＮ個のストレージデバイス１５０から利用不可能であり、合致するパリティデータセグメントがストレージデバイス１５０の第２のセットで利用可能でない場合、回復したパリティデータセグメントは、Ｎ個のデータセグメントのコピーを記憶するストレージデバイス１５０の第２のセットの１又はそれ以上のストレージデバイス１５０から再生成される。更に別の例においては、回復したパリティデータセグメントは、ストレージデバイス１５０の１又はそれ以上のセット内の位置に拘らず、利用可能なデータセグメント及び不合致のパリティデータセグメントから再生成される。

【０３２８】

パリティミラーストレージデバイスがストレージデバイスセット１６０４のストレージデバイス１５０間で代替された場合、一般的にはデータ再構築モジュール１７２４及びパリティ再構築モジュール１７２６は、再構築されるストレージデバイス１５０でデータセグメント及びパリティデータセグメントを再構築すると共に動作する。第２のパリティミラーストレージデバイス１６０２ｂが利用可能である場合、データ再構築モジュール１７２４及びパリティ再構築モジュール１７２６は、ストレージデバイスセット１６０４の２のストレージデバイス１５０、１６０２の故障後に、２のストレージデバイスを再構築できる。パリティミラーストレージデバイス１６０２がパリティミラーデータセグメントを生成するように進行されない場合、パリティミラーストレージデバイス１６０２が進行され、ストライプ用のパリティデータセグメントが算出されて記憶され、パリティデータセ

グメントを算出するのに用いられるパリティミラーストレージデバイス１６０２のＮ個のデータセグメントが消去された場合よりも、データセグメント又はストレージデバイス１５０の回復は迅速である。

【０３２９】

図１８は本発明による進行型ＲＡＩＤを用いてデータセグメントを更新するための装置１８００の一実施例を示した概略ブロック図である。一般的には、装置１８００はＲＡＩＤグループに属し、パリティミラーストレージデバイス１６０２のうちの１又はそれ以上が進行され、パリティデータセグメントを含み、パリティデータセグメントを生成するのに用いられるＮ個のデータセグメントを含まない。装置１８００は更新受信モジュール１８０２、更新コピーモジュール１８０４、パリティ更新モジュール１８０６を具え、以下に述べられている。装置１８００のモジュール１８０２－１８０６は、サーバ１１２に示され、ストレージデバイス１５０、クライアント１１４、又はデバイスのいずれかの組合せに存在せず、いくつかのデバイス間で分散されうる。

【０３３０】

ストライプ、データセグメント、ストレージデバイス１５０、ストレージデバイスセット１６０４、パリティデータセグメント、及び１又はそれ以上のパリティミラーストレージデバイス１６０２は図１７の装置１７００に関連して上述したようなストライプと実質的に同様である。装置１８００は更新されたデータセグメントを受信する更新受信モジュール１８０２を具え、更新されたデータセグメントは既存のストライプの既存のデータセグメントに対応する。別の実施例においては、更新受信モジュール１８０２は複数の更新を更に受信でき、一緒に又は別個に更新を処理できる。

【０３３１】

装置１８００は、対応する既存のデータセグメントが記憶されるストレージデバイス１５０に、及び、既存のストライプに対応する１又はそれ以上のパリティミラーストレージデバイス１６０２に、更新されたデータセグメントをコピーする更新コピーモジュール１８０４を具える。別の実施例においては、更新コピーモジュール１８０４はパリティミラーストレージデバイス１６０２に、又は、既存のデータセグメントを記憶するストレージデバイス１５０に更新されたデータセグメントをコピーし、その後、更新されたデータセグメントのコピーが他のデバイス１６０２、１５０に転送されるのを確認する。

【０３３２】

装置１８００はストレージ統合動作に応じて既存のストライプの１又はそれ以上のパリティミラーストレージデバイス用の１又はそれ以上の更新されたパリティデータセグメントを算出するパリティ更新モジュール１８０６を具える。ストレージ統合動作は図１７の装置１７００に関連して上述したようなストレージ統合動作と同様である。ストレージ統合動作は１又はそれ以上の更新されたパリティデータセグメントを有する１又はそれ以上のパリティミラーストレージデバイス１６０２の少なくともストレージ空き領域及び／又はデータを回復するように導かれる。１又はそれ以上のパリティデータセグメントを更新するのを待機することによって、ストレージ空き領域を統合することがより便利に又は必要になるまで、更新が延期できる。

【０３３３】

一実施例においては、更新されたパリティデータセグメントは、既存のパリティデータセグメント、更新されたデータセグメント、及び既存のデータセグメントから算出される。一実施例においては、既存のデータセグメントは更新されたパリティデータセグメントの生成用の既存のデータセグメントを読み出す前に、所定の位置に維持される。この実施例の利点は、パリティミラーストレージデバイス１６０２、又は、更新されたパリティデータセグメントが生成される他の位置に既存のデータセグメントをコピーすることに関連するオーバヘッドが、必要になるまで延期されうることである。この実施例の欠点は、既存のデータセグメントを維持するストレージデバイス１５０が故障した場合に、更新されたパリティデータセグメントが生成されうる前に、既存のデータセグメントが回復されなければならないことである。

【０３３４】
別の実施例においては、既存のデータセグメントが記憶されるＮ個のストレージデバイス１５０ａ－ｎのストレージデバイス１５０が、更新されたデータセグメントのコピーを受信する場合に、既存のデータセグメントはデーターミラーストレージデバイス１６０２にコピーされる。既存のデータセグメントは次いでストレージ統合動作までに記憶される。別の実施例においては、既存のデータセグメントはＮ個のストレージデバイス１５０ａ－ｎのストレージデバイス１５０のストレージ統合動作に応じてデーターミラーストレージデバイス１６０２にコピーされ、更新されたパリティデータセグメントの算出をトリガするストレージ統合動作の前に、ストレージ統合動作が生じる場合に、既存のデータセグメントが記憶される。後者の実施例は、既存のデータセグメントが記憶されるストレージデバイス１５０か、パリティミラーストレージデバイス１６０２かのいずれかのストレージ統合動作によって要求されるまで、既存のデータセグメントがコピーされないため、有利である。

【０３３５】
一実施例においては、更新されたパリティデータセグメントは、既存のパリティデータセグメント、更新されたデータセグメント、及びデルタデータセグメントから算出され、デルタデータセグメントは更新されたデータセグメントと、既存のデータセグメントとの間の差として生成される。一般的には、デルタデータセグメントを生成するステップは、パリティデータセグメントを更新する際の部分的解決又は介在ステップである。デルタデータセグメントを生成するステップは、高度に圧縮可能であり、送信前に圧縮されうるために有利である。

【０３３６】
一実施例においては、デルタデータセグメントは、更新されたパリティデータセグメントの生成用のデルタデータセグメントを読み出す前に、既存のデータセグメントを記憶するストレージデバイスに記憶される。別の実施例においては、既存のデータセグメントが記憶されるストレージデバイス１５０が、更新されたデータセグメントのコピーを受信する場合に、デルタデータセグメントはデーターミラーストレージデバイス１６０２にコピーされる。別の実施例においては、デルタデータセグメントは、既存のデータセグメントが記憶されるストレージデバイス１５０のストレージ統合動作に応じて、データーミラーストレージデバイス１６０２にコピーされる。既存のデータセグメントをコピーするのと同様に、後者の実施例は、既存のデータセグメントを記憶するストレージデバイス１５０のストレージ統合動作、又は、更新されたパリティデータセグメントの算出をトリガする別のストレージ統合動作のうちの早い方まで、デルタデータファイルが移動しないため、有利である。

【０３３７】
様々な実施例においては、モジュール１８０２、１８０４、１８０６の動作部分の総て、すなわち、更新されたデータセグメントを受信するステップ、更新されたデータセグメントをコピーするステップ、及び更新されたパリティデータセグメントを算出するステップは、ストレージデバイスセット１６０４のストレージデバイス１５０、クライアント１１４、又はサードパーティのＲＡＩＤ管理デバイスで生じる。別の実施例においては、ストレージ統合動作は更新受信モジュール１８０２及び更新コピーモジュール１８０４の動作から自律して導かれる。

【０３３８】
図１９は、本発明による進行型ＲＡＩＤを用いたデータを管理するための方法１９００の一実施例を示した概略フローチャート図である。方法１９００は開始し（１９０２）、ストレージ要求受信モジュール１７０２はデータを記憶する要求を受信し（１９０４）、データがファイル又はオブジェクトのデータとなる。ストライピングモジュール１７０４はデータ用のストライプパターンを算出し、Ｎ個のデータセグメントをＮ個のストレージデバイス１５０に書き込む（１９０６）。ストライプパターンは１又はそれ以上のストライプを含む。各ストライプはＮ個のデータセグメントのセットを含み、Ｎ個のデータセグ

メントの各々が、ストライプに割り当てられるストレージデバイス１６０４のセット内の別個のストレージデバイス１５０に書き込まれる。

【０３３９】

パリティミラーモジュール１７０６は、ストレージデバイス１６０４のセット内の１又はそれ以上のパリティミラーストレージデバイス１６０２に、ストライプのＮ個のデータセグメントのセットを書き込む（１９０８）。１又はそれ以上のパリティミラーストレージデバイスは、Ｎ個のストレージデバイス１５０ａ－ｎに追加した状態にある。パリティ生成モジュール１７０８は待ち状態のストレージ統合動作があるかどうかを判定する（１９１０）。パリティ生成モジュール１７０８が待ち状態のストレージ統合動作はないと判定した（１９１０）場合、方法１９００は戻り待ち状態のストレージ統合動作があるかどうかを再度判定する（１９１０）。別の実施例においては、ストレージ要求受信モジュール１７０２、ストライピングモジュール１７０４、及びパリティミラーモジュール１７０６はストレージ要求を受信し続け、ストライピングパターンを算出し、データセグメントを記憶する。

【０３４０】

パリティ生成モジュール１７０８が、待ち状態のストレージ統合動作がないと判定した場合、パリティ生成モジュール１７０８はストライプ用のパリティデータセグメントを算出する（１９１４）。パリティデータセグメントはパリティミラーストレージデバイス１６０２に記憶されたＮ個のデータセグメントから算出される。パリティ生成モジュール１７０８はパリティミラーストレージデバイス１６０２にパリティデータセグメントを記憶し（１９１２）、方法１９００は終了する（１９１６）。ストレージ統合動作は、Ｎ個のデータセグメントを記憶する要求を受信し（１９０４）、Ｎ個のデータセグメントをＮ個のストレージデバイスに書き込み（１９０６）、あるいは、Ｎ個のデータセグメントを１又はそれ以上のパリティミラーストレージデバイスに書き込むことから自律して導かれる。ストレージ統合動作は少なくともパリティミラーストレージデバイス１６０２の少なくともストレージ空き領域又はデータを回復するように導かれる。

【０３４１】

図２０は、本発明による進行型ＲＡＩＤを用いてデータセグメントを更新するための方法２０００の一実施例を示す概略フローチャート図である。方法２０００は開始し（２００２）、更新受信モジュール１８０２は更新されたデータセグメントを受信し（２００４）、更新されたデータセグメントが既存のストライプの既存のデータセグメントに対応する。更新コピーモジュール１８０４は対応する既存のデータセグメントが記憶されたストレージデバイス１５０に、及び既存のストライプに対応する１又はそれ以上のパリティミラーストレージデバイス１６０２に、更新されたデータセグメントをコピー（２００６）する。

【０３４２】

パリティ更新モジュール１８０６はストレージ統合動作が待ち状態かどうかを判定する（２００８）。パリティ更新モジュール１８０６が待ち状態のストレージ統合動作がないと判定した（２００８）場合、パリティ更新モジュール１８０６はストレージ統合動作を待つ。一実施例においては、方法２０００は戻って、他の更新されたデータセグメントを受信し（２００４）、更新されたデータセグメントをコピーする（２００６）。パリティ更新モジュール１８０６が待ち状態のストレージ統合動作はないと判定した（２００８）場合、パリティ更新モジュール１８０６は、既存のストライプの１又はそれ以上のパリティミラーストレージデバイス用の１又はそれ以上の更新されたパリティデータセグメントを算出し（２０１０）、方法２０００は終了する（２０１２）。

【０３４３】

フロントエンド分散型ＲＡＩＤ

【０３４４】

従来のＲＡＩＤシステムはデータを受信し、データ用のストライピングパターンを算出し、データをデータセグメントに分割し、パリティストライプを算出し、ストレージデバ

イスのデータを記憶し、データセグメントを更新する、等をするのに機能するＲＡＩＤコントローラで構成される。いくつかのＲＡＩＤコントローラによって、いくつかの機能が分散可能となるが、ＲＡＩＤコントローラにより管理されるストレージデバイスは、ＲＡＩＤでストライピングされるデータを記憶するために直接にはクライアント１１４と通信しない。代わりに、ストレージ要求及びＲＡＩＤするためのデータはストレージコントローラを通過する。

【０３４５】

ＲＡＩＤに記憶されるべきデータの総てに接触するためにＲＡＩＤコントローラを要求することは、データフローのボトルネックを生成するため非効率的である。このことは読み出し－変更－書き込み処理中は特に当てはまり、バンド幅及びＲＡＩＤグループのドライブの総ての性能が消費される一方、サブセットだけが実際には更新される。更にはＲＡＩＤコントローラによって管理されるデータ用に指定されたストレージデバイスの領域は、一般的にはＲＡＩＤグループ専用であり、別個にアクセスできない。クライアントによるストレージデバイス１５０へのアクセスは、一般的にストレージデバイス１５０をパーティショニングすることによって得られる。パーティショニングが用いられる場合、汎用ストレージ用のアクセス可能なパーティションはＲＡＩＤ用に用いられず、ＲＡＩＤグループに割り当てられたパーティションは汎用データストレージ用にアクセス可能ではない。包括的に利用を最適化するためにパーティションをオーバスクライブする方式は、複雑であり管理するのがより難しい。更に、あるＲＡＩＤグループ用に割り当てられたストレージ空き領域は、１のＲＡＩＤコントローラがマスタとして指定されない限り、１以上のＲＡＩＤコントローラによってアクセスされず、マスタＲＡＩＤコントローラが停止、非機能性等でない限り、他のＲＡＩＤコントローラがスレイブとして動作する。

【０３４６】

一般的なＲＡＩＤコントローラはＲＡＩＤグループのストレージデバイス１５０の外側にパリティデータセグメントを更に生成する。パリティデータセグメントは一般的に生成されて、その後ストレージのためにストレージデバイス１５０に送信され、ＲＡＩＤコントローラの算出能力を要求するためこれは非効率になりうる。更に、パリティデータセグメントの位置及び更新を追跡することは、ストレージデバイス１５０で自律して行う代わりに、ＲＡＩＤコントローラで行わなければならない。

【０３４７】

別個のＲＡＩＤコントローラがオフラインである場合にデータが利用可能のままであることを保証することが必要である場合、ＲＡＩＤコントローラは一般的には、ドライブに及び互いに相互接続、及び／又は完全なセットしてミラーリングされ、データ利用可能性を管理するのを高価及び難易にし、ストレージサブシステムの信頼性を劇的に低下させる。

【０３４８】

必要となるのは、データセグメントごと、オブジェクトごと、ファイルごと、又は同様のベースでＲＡＩＤを可能にし、ＲＡＩＤコントローラ及び、クライアントとストレージデバイスとの間に位置している対のＲＡＩＤコントローラの必要性を除去する、フロントエンド分散型ＲＡＩＤ用のシステム、装置及び方法である。このようなシステム、装置及び方法において、ＲＡＩＤグループは、あるデータセグメント、オブジェクト、又はファイルのために生成され、あるＲＡＩＤコントローラによってストレージデバイスの１のグループ内で管理されうる一方、第２のＲＡＩＤグループは第１のＲＡＩＤグループの同一のストレージデバイスの一部の周りにある別のデータセグメント、オブジェクト、又はファイルのために生成されうる。ＲＡＩＤ制御機能はクライアント１１４、サードパーティのＲＡＩＤ管理デバイスの間で、又はストレージデバイス１５０の間で分散されうる。フロントエンド分散型ＲＡＩＤシステム、装置、及び方法はＲＡＩＤグループのストレージデバイス１５０にコマンドを更に送信でき、ダイレクトメモリアクセス（「ＤＭＡ」）又はリモートＤＭＡ（「ＲＤＭＡ」）を通してデータに直接的にアクセスし、コピーすることをストレージデバイス１５０に許容できる。

【０３４９】
図１０は本発明によるフロントエンド分散型ＲＡＩＤに対しアクセスされうるシステム１６００の一実施例を示す概略ブロック図である。進行型ＲＡＩＤに関連する図１０に示した構成に対する上の記載が、フロントエンド分散型ＲＡＩＤに更に適用可能である。フロントエンド分散型ＲＡＩＤに関し、ストレージデバイスセット１６０４はＲＡＩＤグループを形成し、自律的にかつ、ネットワーク１１６又は１又はそれ以上の冗長ネットワーク１１６上でクライアント１１４からストレージ要求を別個に受信及び提供できるストレージデバイス１５０を具える。
【０３５０】
ストレージデバイスセット１６０４内のストレージデバイス１５０の間から、１又はそれ以上がストライプ用のパリティミラーストレージデバイス１６０２として指定される。一般的には１又はそれ以上のパリティミラーストレージデバイス１６０２は、他のストレージデバイス１５０と実質的に同様に機能する。指定されたパリティミラーストレージデバイス１６０２がストレージデバイスセット１６０４のストレージデバイス１５０の間で代替する場合の一般的な構成においては、パリティミラーストレージデバイス１６０２はノンパリティミラーストレージデバイスとして更に動作しなければならないために、他のストレージデバイス１５０と同様の特性を実質的に有している。同様の特性はＲＡＩＤグループ内の動作、及び、上述したような別個のクライアント１１４の通信に対する自律動作に関して存在する。様々な実施例においては、ストレージデバイスセット１６０４のストレージデバイス１５０は、述べられたＲＡＩＤ環境内で機能することに関連しない他の態様においては異なっていてもよい。
【０３５１】
ストレージデバイスセット１６０４のストレージデバイス１５０は、１又はそれ以上のサーバ１１２内にグループ化されたスタンドアロンにでき、各々がサーバ１１２に常駐でき、１又はそれ以上のサーバ１１２を通してアクセスされうる等ができる。１又はそれ以上のクライアント１１４は、１又はそれ以上のストレージデバイス１５０を具えるサーバ１１２に常駐でき、別個のサーバ１１２に常駐でき、１又はそれ以上のコンピュータネットワーク１１６を通してストレージデバイス１５０をアクセスするコンピュータ、ワークステーション、ラップトップ等に常駐できる等ができる。
【０３５２】
一実施例においては、ネットワーク１１６はシステムバスを具え、ストレージデバイスセット１６０４のストレージデバイス１５０、１６０２の１又はそれ以上は、システムバスを用いて通信する。例えば、システムバスはＰＣＩ－ｅバス、シリアルアドバンスドテクノロージアタッチメント（「シリアルＡＴＡ」）バス、パラレルＡＴＡ等にできる。別の実施例においては、システムバスは小型コンピュータシステムインタフェース（「ＳＣＳＩ」）、ＦｉｒｅＷｉｒｅ、ファイバチャネル、ＵＳＢ、ＰＣＩｅ－ＡＳ、インフィニバンド等のような外部バスである。当該技術分野の当業者は、自律的であり、１又はそれ以上のネットワーク１１６上のクライアント１１４からストレージ要求を別個に受信及び提供することができるストレージデバイス１５０を有する他のシステム１６００の構成が分かるであろう。
【０３５３】
図１１は本発明によるフロントエンド分散型ＲＡＩＤ用の装置２１００の一実施例を示す概略ブロック図である。様々な実施例においては、装置２１００はストレージ要求受信モジュール２１０２、ストライピング連結モジュール２１０４、パリティミラー連結モジュール２１０６、ストレージ要求送信モジュール２１０８、フロントエンドパリティ生成モジュール２１１０、パリティ代替モジュール２１１２、データセグメント回復モジュール２１１４、データ再構築モジュール２１１６、パリティ再構築モジュール２１１８、及びピアツーピア通信モジュール２１２０を具え、以下に述べられている。様々な実施例においては、装置２１００はソリッドステートストレージデバイス１０２のようなストレージデバイス１５０、ソリッドステートストレージコントローラ１０４のようなストレージ

デバイスコントローラ１５２、サーバ１１２、サードパーティのＲＡＩＤ管理デバイス等に含まれ、１以上の構成の間で分散されうる。

【０３５４】

ストレージデバイスセット１６０４にデータを記憶するストレージ要求を受信するストレージ要求受信モジュール２１０２を具える。データはファイル又はオブジェクトの一部分にでき、あるいは、完全なファイル又はオブジェクトにできる。ファイルは任意の情報のブロック、又は情報を記憶するためのリソースを含み、コンピュータプログラムで利用できる。ファイルはプロセッサによってアクセスされるいずれかのデータ構造を含みうる。ファイルは、データベース、テキストの記号列、コンピュータコード等を含みうる。オブジェクトは一般的にはオブジェクト指向型プログラミング用のデータ構造であり、データを有する又は有しない構造を含みうる。一実施例においては、オブジェクトはファイルのサブセットである。別の実施例においては、オブジェクトはファイルから独立している。いずれのケースにおいても、オブジェクト又はファイルは、完全なデータのセット、データ構造、コンピュータコード、及びストレージデバイスで記憶されうる他の情報を含むように本明細書において規定されている。

【０３５５】

ストレージデバイスセット１６０４は、１又はそれ以上のネットワーク１１６上のクライアント１１４からストレージ要求を独立して受信するＲＡＩＤグループを形成する自律型ストレージデバイス１５０を具える。ストレージデバイスセット１６０４内の自律型ストレージデバイス１５０の１又はそれ以上は、ストライプ用のパリティミラーストレージデバイス１６０２として指定される。別のクライアント１１４からの他のストレージ要求は、第２のストレージデバイスセットで記憶され、第２のストレージデバイスセットは、第１のストレージデバイスセット１６０４と同一のストレージデバイス１５０（及び、パリティミラーストレージデバイス１６０２）の１又はそれ以上を具えることができる。双方のストレージデバイスセット１６０４に共通のストレージデバイス１５０は、共通ストレージデバイス１５０内の重複する割り当てられたストレージ空き領域を有することができる。

【０３５６】

装置２１００はデータ用のストライプパターンを算出するストライピング連結モジュール２１０４を具える。ストライプパターンは１又はそれ以上のストライプを含む。各ストライプはＮ個のデータセグメントのセットからなる。ストライプのＮ個のデータセグメントは１又はそれ以上の空データセグメントを更に含みうる。ストライピング連結モジュール２１０４は、ストライプに割り当てられたストレージデバイスセット１６０４のＮ個のストレージデバイス１５０ａ－ｎのうちの１つと、Ｎ個のデータセグメントの各々を連結させる。一実施例においては、ストライピング連結モジュール２１０４は、ストレージ要求を送信するクライアント１１４からのデータセグメントに対応するデータを得るためにストレージデバイス１５０に命令する、ストレージデバイス１５０に送信すべきストレージ要求とともに、データセグメントをストレージデバイス１５０と連結させる。

【０３５７】

別の実施例においては、ストレージ要求はデータセグメントのデータを実質的に含まない。データを実質的に含まないことは、ストレージ要求が一般的に、ストレージ要求の対象となるデータを含まず、データの一部分である、キャラクタ、キャラクタ列等を含みうることを意味する。例えば、データがゼロの配列のような、反復する同一キャラクタの配列を含む場合、ストレージ要求は、データ中に含まれる総てのゼロを含まないゼロの配列を含むという指示を含みうる。当該技術分野の当業者は、大部分のデータを送信しないが、ストレージ要求中の特定のキャラクタ又はキャラクタ列の少量又は単一のインスタンスを許容する他の方法が分かるであろう。ストレージ要求は、Ｎ個のストレージデバイス１５０ａ－ｎがＤＭＡ又はＲＤＭＡ動作等を用いてデータを検索するのを可能にするコマンドを含む。

【０３５８】

別の実施例においては、ストライピング連結モジュール２１０４は、ストレージデバイス１５０へ送信すべきストレージ要求内で、データセグメントのデータを識別することによって、データセグメントをストレージデバイス１５０と連結させる。データセグメントのデータを識別することは、データセグメント識別子、データセグメント位置若しくはアドレス、データセグメント長、又はどのデータがデータセグメントを含むのかをデータセグメントが認識するのを可能にする他の情報を含みうる。

【０３５９】

一実施例においては、ストライピング連結モジュール２１０４はストレージ要求においてデータセグメントをストレージデバイス１５０と連結させて、クライアント１１４がデータセグメントを含むデータをブロードキャストにおいて送信できるようにし、各ストレージデバイス１５０が付随するデータセグメントを記憶し、ストレージデバイス１５０に割り当てられないデータセグメントに対応するデータを廃棄できるようにする。別の実施例においては、ストライピング連結モジュール２１０４は、可能であれば各データセグメントをアドレス付けすることによって、ストレージ要求においてデータセグメントをストレージデバイス１５０と連結させて、クライアント１１４がデータセグメントを含むデータをマルチキャストにおいて送信できるようにし、各ストレージデバイス１５０が付随するデータセグメントを記憶し、ストレージデバイス１５０に割り当てられないデータセグメントに対応するデータを廃棄できるようにする。当該技術分野の当業者は、１又はそれ以上のデータセグメントを１又はそれ以上のストレージデバイスにブロードキャスト、マルチキャスト、ユニキャスト、エニーキャスト等するために、データセグメントをストレージデバイス１５０と連結させるストライピングモジュール２１０４用の他の方法が分かるであろう。

【０３６０】

関連する実施例においては、ストライピング連結モジュール２１０４はストレージ要求においてデータセグメントをストレージデバイス１５０と連結させて、クライアント１１４がストレージ要求をブロードキャスト、マルチキャスト、ユニキャスト等をできるようにし、ストレージデバイス１５０と連結するデータセグメントに属するクライアント１１４からストレージ要求の一部を各ストレージデバイス１５０が受信できるようにし、ストレージデバイス１５０と連結する１又はそれ以上のデータセグメントに属しないストレージ要求部分を廃棄できるようにする。

【０３６１】

別の実施例においては、ストレージ要求受信モジュール２１０２により受信されたストレージ要求は、ストレージ要求の対象となるデータを含み、ストライピング連結モジュール２１０４は、データセグメントを含むストレージデバイス１５０のために、ストレージ要求を与えることによって、データセグメントをストレージデバイス１５０と連結させる。ストライピング連結モジュール２１０４はクライアント１１４、サードパーティのＲＡＩＤ管理デバイス、ストレージデバイス１５０、１６０２等内で動作可能である。

【０３６２】

装置２１００は、Ｎ個のデータセグメントのセットを、ストレージデバイスセット１６０４の１又はそれ以上のパリティミラーストレージデバイス１６０２と連結させるパリティミラー連結モジュール２１０６を具える。１又はそれ以上のパリティミラーストレージデバイス１６０２はＮ個のストレージデバイス１５０ａ－ｎに加えて存在する。一実施例においては、パリティミラー連結モジュール２１０６はＮ個のデータセグメントのセットを、各パリティミラーストレージデバイス１６０２に連結させて、各パリティミラーストレージデバイス１６０２は、ストライプのＮ個のデータセグメントを受信及び記憶して、パリティデータセグメントを生成できるようにする。別の実施例においては、パリティミラー連結モジュール２１０６はストライプのデータセグメントを各パリティミラーストレージデバイス１６０２と連結させて、パリティミラーストレージデバイス１６０２ａ－ｍはＮ個のストレージデバイス１５０ａ－ｎに記憶されるＮ個のデータセグメントに対するミラーとして動作するようになる。

【０３６３】
様々な実施例においては、パリティミラー連結モジュール２１０６は、単一のストレージ要求、複数のストレージ要求、又はＤＭＡ、ＲＤＭＡ、ブロードキャスト、マルチキャスト用のパリティミラーストレージデバイス１６０２を設定し、あるいは、ストレージ要求中にＮ個のデータセグメントを含むストレージ要求のような、ストライピング連結モジュール２１０４に関連して上述したような他の関連技術を用いて、Ｎ個のデータセグメントのセットを１又はそれ以上のパリティミラーストレージデバイス１６０２と連結させる。パリティミラー連結モジュール２１０６は、クライアント１１４、サードパーティのＲＡＩＤ管理デバイス、ストレージデバイス１５０、１６０２等の中で動作可能である。
【０３６４】
装置２１００はストレージデバイスセット１６０４において、１又はそれ以上のストレージ要求を各ストレージデバイス１５０、１６０２に送信するストレージ要求送信モジュール２１０８を具え、各ストレージ要求はストレージデバイス１５０、１６０２に、ストレージ要求を受信するストレージデバイス１５０、１６０２に連結する１又はそれ以上のデータセグメントを記憶するのに十分である。一実施例においては、各ストレージ要求はストレージ要求の対象となるデータを含まない。更なる実施例においては、各ストレージ要求は、ストレージデバイスセット１６０４のＮ個のストレージデバイス１５０及びパリティミラーストレージデバイス１６０２が、ＤＭＡ又はＲＤＭＡを用いて付随するデータセグメントのデータをダウンロードするのを可能にする。別の実施例においては、ストレージ要求は、関連するストレージ要求又はクライアント１１４のブロードキャストからの、付随するデータセグメント用の関連するデータを選定するのに十分な情報を含む。別の実施例においては、ストレージ要求は、付随するデータセグメントのデータを含む。
【０３６５】
一実施例においては、各ストレージ要求はストライプのストレージデバイスセット１６０４の一部であるストレージデバイス１５０、１６０２を識別する。ストレージデバイスセット１６０４のストレージデバイス１５０、１６０４の識別を含むことによって、マスタとして動作するストレージデバイス１５０の故障の場合には、別のストレージデバイス１５０が、ＲＡＩＤされたデータを管理するためのマスタとして引き継ぐことができる。別の実施例においては、ストレージデバイスセット１６０４の識別によって、自律型ストレージデバイス１５０、１６０２は、ストレージデバイスがオフラインの場合にはデータを回復し、置換ストレージデバイスが、クライアントと別個にストレージデバイスセット１６０４内に付加される場合にはデータを再構築することを可能にする。別の実施例においては、ストレージデバイスセット１６０４のストレージデバイス１５０、１６０２の識別は、データセグメント又はストレージ要求の送信用のマルチキャストグループを表わしている。この識別はストレージデバイスセット１６０４内のストレージデバイス１５０、１６０２に記憶されたオブジェクト又はファイル用のメタデータと共に、記憶されうる。
【０３６６】
一実施例においては、パリティミラー連結モジュール２１０６がＮ個のデータセグメントのセットを１又はそれ以上のパリティミラーストレージデバイス１６０２の各々と連結させる場合、装置２１００はクライアント１１４とは別個に、ストライプ用のパリティデータセグメントを算出し、パリティミラーストレージデバイス１６０２にパリティデータセグメントを記憶するフロントエンドパリティ生成モジュール２１１０を具える。パリティデータセグメントは、パリティミラーストレージデバイス１６０２に提供されるＮ個のデータセグメントのセットから算出される。１以上のパリティミラーストレージデバイス１６０２がストレージデバイスセット１６０４に含まれる場合、フロントエンドパリティ生成モジュール２１１０は一般的に様々なパリティデータセグメントを生成して、ストレージデバイスセット１６０４内の２又はそれ以上のストレージデバイス１５０、１６０２は故障可能になり、パリティデータセグメント情報が、利用不可能なデータセグメント又はパリティデータセグメントの回復を可能にする。
【０３６７】

別の実施例においては、フロントエンドパリティ生成モジュール２１１０は、ストレージデバイスセット１６０４のストレージデバイス１５０及び／又はサードパーティのＲＡＩＤ管理デバイス内で動作する場合、パリティデータセグメントを算出する。例えば、ストレージ要求を送信するクライアント１１４と別個のサーバ１１２は、パリティデータセグメントを算出できる。別の実施例においては、フロントエンドパリティ生成モジュール２１１０はパリティミラーストレージデバイス内で動作して、パリティデータセグメントを算出する。例えば、パリティミラーストレージデバイス１６０２内のストレージコントローラ１５２は、ストレージデバイスセット１６０４によって形成されるＲＡＩＤグループ用のマスタストレージコントローラとして動作できる。

【０３６８】

別の実施例においては、フロントエンドパリティ生成モジュール２１１０はパリティデータセグメントを算出し、次いで、算出されたパリティデータセグメントを、ミラーを形成する第２のストレージデバイスのセットの１又はそれ以上の更なるパリティミラーストレージデバイス１６０４に送信する。この実施例は、パリティデータセグメントを算出することに関連するオーバヘッドが、ネットワーク１１６上のデータトラフィックを低減させる更なる利益を有する各ストレージデバイスセット１６０４に対してではなく、一度に消費されるため、有利である。

【０３６９】

一実施例においては、ストレージデバイス１５０が利用不可能であり、要求が、利用不可能なデータセグメントか、利用不可能なデータセグメントを含むデータのいずれかを読み出すように受信される場合、装置２１００はストレージデバイスセット１６０４のストレージデバイス１５０に記憶されたデータセグメントを回復するデータセグメント回復モジュール２１１２を具える。データセグメントは、ストレージデバイスセット１６０４の利用可能なストレージデバイス１５０のデータセグメント、又は、ストレージデバイスセット１６０４の利用可能なストレージデバイス１５０、１６０２のパリティデータセグメント及びデータセグメントの組合せを用いて、あるいは、データセグメントのコピーを含むミラーストレージデバイスから回復される。一般的には、ミラーストレージデバイスはＮ個のデータセグメントのコピーを記憶するミラーストレージデバイスセットの１のストレージデバイス１５０である。データセグメント回復モジュール２１１２は動作して、ストレージデバイス１５０、パリティミラーストレージデバイス１６０２、サードパーティのＲＡＩＤ管理デバイス、ミラーストレージデバイス等から利用不可能なデータセグメントを回復できる。

【０３７０】

別の実施例においては、装置２１００は再構築動作中に、置換ストレージデバイス１５０に回復したデータセグメントを記憶するデータ再構築モジュール２１１４を具える。例えば、ストレージデバイス１５０が故障、同期の損失等のために利用不可能になった場合、データ再構築モジュール２１１４はストレージデバイス１５０を再構築して、利用不可能なストレージデバイス１５０を置換できる。一実施例においては、再構築されるストレージデバイス１５０は、利用可能にされた元のストレージデバイス１５０である。

【０３７１】

回復したデータセグメントは、ストレージデバイスセット１６０４の利用不可能なストレージデバイス１５０に記憶された利用不可能なデータセグメントに合致させる。再構築動作は一般的に、データセグメント及びパリティデータセグメントの１又はそれ以上を置換ストレージデバイス１５０に復元して、利用不可能なストレージデバイス１５０に前に記憶されたデータセグメント及びパリティデータセグメントに合致させる。

【０３７２】

一実施例においては、回復したデータセグメントはストレージデバイスセット１６０２の利用可能なストレージデバイス１５０で利用可能なデータセグメントを用いて、再構築動作によって回復される。別の実施例においては、回復したデータセグメントは、パリティミラーストレージデバイス１６０２の１又はそれ以上からのパリティデータセグメント

と、ストレージデバイスセット１６０４の利用可能なストレージデバイス１５０の利用可能なデータセグメントとの組合せを用いて、再構築動作によって回復される。別の実施例においては、回復したデータセグメントは、パリティミラーストレージデバイス１６０２から読み出された、合致するデータセグメントを用いて、再構築動作により回復される。更なる別の実施例においては、回復したデータセグメントは、ミラーストレージデバイスからの合致するデータセグメントを用いて、再構築動作により回復される。データ再構築モジュール２１１４は動作し、クライアント１１４、サードパーティのＲＡＩＤ管理デバイス、ストレージデバイス１５０、１６０２、ミラーストレージデバイス等からの受信したデータセグメントを記憶するであろう。

【０３７３】

別の実施例においては、装置２１００は再構築動作で、置換ストレージデバイス１６０２に回復したパリティデータセグメントを再構築するパリティ再構築モジュール２１１６を具える。再構築動作は、データ再構築モジュール２１１４に関連して記載した再構築動作と実質的に同様である。パリティ再構築モジュール２１１６が回復したパリティデータセグメントを再構築することを除いては、パリティ再構築モジュール２１１６は、データ再構築モジュール２１１４と同様に動作する。回復したパリティデータセグメントは、ストライプに割り当てられた利用不可能なパリティミラーストレージデバイス１６０２に記憶される利用不可能なパリティデータセグメントに合致される。

【０３７４】

様々な実施例において、ミラーリングストレージデバイスセットのパリティミラーストレージデバイス１６０２に記憶されたパリティデータセグメントをコピーすることによって、ストレージデバイスセット１６０４のパリティミラーストレージデバイス１６０２からパリティデータセグメントをコピーすることによって（利用不可能なパリティデータセグメントと同一の場合）、ストレージデバイスセット１６０４の利用可能なストレージデバイス１５０、１６０２、及びデータセグメントのコピーを含むミラーストレージデバイスに記憶されたＮ個のデータセグメント及びパリティデータセグメントの１又はそれ以上を用いてパリティデータセグメントを生成することによって、等によって回復される。データ再構築モジュール２１１６は動作して、回復したデータセグメントを記憶する一方、クライアント１１４、サードパーティのＲＡＩＤ管理デバイス、ストレージデバイス１５０、ミラーストレージデバイス等に常駐する。

【０３７５】

有利には、装置２１００は本明細書に記載のフロントエンド分散型ＲＡＩＤ動作専用のパーティションに、ストレージデバイス１５０、１６０２のデータを記憶することに限定されない。代わりに、自律型ストレージデバイス（例えば、１５０ａ）は、クライアント１１４からストレージ要求を別個に受信して、ストライピング連結モジュール２１０４、パリティミラー連結モジュール２１０６及びフロントエンドパリティ生成モジュール２１１０によってデータを記憶するのに更に利用可能なストレージデバイス１５０ａの１又はそれ以上の領域のＲＡＩＤされた、又はＲＡＩＤされないデータを記憶できる。

【０３７６】

一実施例においては、ストレージ要求受信モジュール２１０２によって受信され、又は、ストレージ要求送信モジュール２１０８によって送信される１又はそれ以上のストレージ要求が、ストライプのストレージデバイスセット１６０４を具えるストレージデバイス１５０を識別する。有利には、ストレージ要求でストレージデバイスセット１６０４のストレージデバイス１５０を識別することは、バックアップＲＡＩＤコントローラを促進して、マスタコントローラが非機能性である場合に動作する。例えば、ストレージデバイスセット１６０４のストレージデバイス１５０がストレージ要求で識別され、マスタコントローラがパリティミラーストレージデバイス１６０２にあり、利用不可能である場合、別のパリティミラーストレージデバイス１６０２、又は、Ｎ個のストレージデバイス１５０ａ－ｎのうちの別のものが、マスタコントローラになりうる。

【０３７７】

一実施例においては、装置２１００は各ストライプに対して、ストレージデバイスセット１６０４内のどのストレージデバイス１５０がストライプ用のパリティミラーストレージデバイス１６０２として指定されるかを代替するパリティ代替モジュール２１１８を具える。パリティ代替モジュール２１１８の利点は上述のとおりである。別の実施例においては、ストレージデバイスセット１６０４のストレージデバイス１５０はピアのグループを形成し、装置２１００は、ストレージデバイスセット１６０４のストレージデバイス１５０、１６０２内のストレージ要求を送信及び受信するピアツーピア通信モジュール２１２０を具える。ピアツーピア通信モジュール２１２０はピアデバイスをストレージデバイスセット１６０４の外側に有して、ストレージ要求を更に送信及び受信できる。

【０３７８】

好ましい実施例においては、ストレージ要求は装置２１００のモジュール２１０２－２１２０を用いてストレージデバイスセット１６０４のストレージデバイス１５０、１６０２にわたって、オブジェクトのデータをストライピングすることによって、オブジェクトを記憶するオブジェクト要求である。別の実施例においては、ストレージデバイスセット１６０４の自律型ストレージデバイス１５０、１６０２の１又はそれ以上は、少なくとも一部の第１のオブジェクト又はファイル用の第１のＲＡＩＤグループ内に割り当てられ、少なくとも一部の第２のオブジェクト又はファイル用の第２のＲＡＩＤグループ内に割り当てられる。例えば、１のストレージデバイス１５０ａは１又はそれ以上のストライプ用のストレージデバイスセット１６０４のためのマスタＲＡＩＤコントローラにでき、第２のストレージデバイス１５０ｂはストレージデバイスセットのストレージデバイス１５０の一部又は総てを含む、ＲＡＩＤグループ用のマスタＲＡＩＤコントローラにできる。有利には、装置２１００はストレージデバイス１５０、１６０２をグループ化して、様々なクライアント１１４のためにＲＡＩＤグループを形成する際にフレキシビリティを与える。

【０３７９】

図１２は、本発明によるフロントエンド分散型ＲＡＩＤ用の方法の一実施例を示す概略フローチャート図である。方法２２００は開始し（２２０２）、ストレージ要求受信モジュール２１０２はストレージ要求を受信して（２２０４）、ストレージデバイスセット１６０４のＮ個のストレージデバイス１５０ａ－ｎのデータを記憶する。ストライピング連結モジュール２１０４は、データ用のストライプパターンを算出し（２２０６）、Ｎ個のデータセグメントのそれぞれを、Ｎ個のストレージデバイス１５０ａ－ｎの１つと連結させる（２２０８）。

【０３８０】

パリティミラー連結モジュール２１０６は、Ｎ個のデータセグメントのセットを１又はそれ以上のパリティミラーストレージデバイス１６０２と連結させる（２２１０）。ストレージ要求送信モジュール２１０８は１又はそれ以上のストレージ要求をストレージデバイスセット１６０４の各ストレージデバイス１５０、１６０２に送信する（２２１２）。各ストレージ要求は、ストレージデバイス１５０に、ストレージ要求を受信するストレージデバイス１５０に関連する１又はそれ以上のデータセグメントを記憶するのに十分である。データのデータセグメントは次いで、ストレージ要求によって命令されるように、ＤＭＡ、ＲＤＭＡ、ブロードキャスト、マルチキャストを用いてストレージデバイスセット１６０４のストレージデバイス１５０、１６０２に転送される。選択的に、フロントエンドパリティ生成モジュール２１１０はストライプ用のパリティデータセグメントを算出し（２２１４）、方法２２００は終了する（２２１６）。

【０３８１】

共有されたフロントエンド分散型ＲＡＩＤ

【０３８２】

従来のＲＡＩＤはＲＡＩＤ専用の各ディスクの少なくとも一部で、ディスク又は他のストレージデバイスのアレイを利用する。ＲＡＩＤコントローラはＲＡＩＤグループに対するストレージ要求を管理する。冗長システムについては、マスタＲＡＩＤコントローラは

故障又は利用不可能になった場合に、引き継ぐようなバックアップＲＡＩＤコントローラを有している。ＲＡＩＤに記憶された同一データにアクセスするようになる複数クライアントからのストレージ要求は、一般的に出現順に順々に実行される。

【０３８３】

図１０、１１、及び１２にそれぞれ示されたシステム１６００、装置２１００、及び方法２２００に関して上述したように、フロントエンド分散型ＲＡＩＤシステムは、分散型ＲＡＩＤコントローラとして機能できるストレージコントローラ１５２をそれぞれ具えることができる自律型ストレージデバイス１５０を具え、ストレージデバイス１５０はそれぞれ、複数のクライアント１１４を供する複数の重複するＲＡＩＤグループで構成されうる。場合により、２のクライアント１１４は同一データにアクセスするようにできる。ストレージ要求が最初に到達して実行する場合、一般的にはデータの不一致がない。他方、同一データに対する２又はそれ以上のストレージ要求が同時、又は実質的に同時に到達した場合、データは破損するであろう。

【０３８４】

例えば、データがＲＡＩＤグループの４のストレージデバイス１５０に記憶され、ストレージデバイス１５０の１つがパリティミラーストレージデバイス１６０２として割り当てられ、第１のクライアント１１４がストレージ要求をＲＡＩＤコントローラとして動作する第１のストレージコントローラ１５０ａに送信し、第２のクライアント１１４が第２のストレージ要求を第２のＲＡＩＤコントローラとして動作する第２のストレージデバイス１５０ａに送信し、双方のストレージ要求が同一データにアクセスする場合、第１のストレージデバイス１５０ａは、第１のストレージデバイス１５０ａ、次いで、ＲＡＩＤグループの他のストレージデバイス１５０ｂ－ｎのストレージ要求の実行を開始できる。同時に、第２のストレージデバイス１５０ｂの第２のＲＡＩＤコントローラは別のストレージデバイス（例えば、１５０ｂ）、次いで、ＲＡＩＤグループの残りのストレージデバイス１５０ａ、ｃ－ｎの第２のストレージ要求の実行を開始できる。この実行における不合致は、ストレージデバイス１５０の間の物理的な距離、実行時間の差異等によって生じうる。結果として破損したデータとなる。

【０３８５】

必要となるのは、同一データのアクセスを求める同時のストレージ要求を処理する共有されたフロントエンド分散型ＲＡＩＤのためのシステム、装置及び方法である。有益には、システム、装置及び方法はデータへのアクセスを制御し、１のストレージ要求は第２のストレージ要求が実行される前に、実行される。

【０３８６】

図１０は、進行型ＲＡＩＤ及びフロントエンド分散型ＲＡＩＤに加えて、本発明による共有されたフロントエンド分散型ＲＡＩＤに対するシステム１６００として働く一実施例を示す概略ブロック図である。進行型ＲＡＩＤ及びフロントエンド分散型ＲＡＩＤに関連する図１０に示した構成に対する上の記載が、共有されたフロントエンド分散型ＲＡＩＤに更に適用可能である。フロントエンド分散型ＲＡＩＤと同様、ストレージデバイスセット１６０４はＲＡＩＤグループを形成し、自律的にかつ、ネットワーク１１６上でクライアント１１４からストレージ要求を別個に受信及び提供できるストレージデバイス１５０を具える。

【０３８７】

共有されたフロントエンド分散型ＲＡＩＤに関し、システム１６００は２又はそれ以上のクライアント１１４のそれぞれが同一データに関するストレージ要求を送信するような、２又はそれ以上のクライアント１１４を具える。１のストレージ要求が別のストレージ要求の別の到達前に完了しないようにストレージ要求が到達する点で、ストレージ要求は同時である。ストレージデバイスセット１６０４内のストレージデバイス１５０の間から、１又はそれ以上がストライプ用のパリティミラーストレージデバイス１６０２として指定される。一般的には、１又はそれ以上のパリティミラーストレージデバイス１６０２は他のストレージデバイス１５０と実質的に同様に機能する。

【０３８８】
指定されたパリティミラーストレージデバイス１６０２がストレージデバイスセット１６０４のストレージデバイス１５０の間で代替する場合の一般的な構成においては、パリティミラーストレージデバイス１６０２はノンパリティミラーストレージデバイスとして更に動作しなければならないために、他のストレージデバイス１５０と同様の特性を実質的に有している。同様の特性はＲＡＩＤグループ内の動作、及び、上述したような別個のクライアント１１４の通信に対する自律動作に関して存在する。様々な実施例においては、ストレージデバイスセット１６０４のストレージデバイス１５０は、述べられたＲＡＩＤ環境内で機能することに関連しない他の態様においては異なっていてもよい。
【０３８９】
ストレージデバイスセット１６０４のストレージデバイス１５０は、１又はそれ以上のサーバ１１２内にグループ化されたスタンドアロンにでき、各々がサーバ１１２に常駐でき、１又はそれ以上のサーバ１１２を通してアクセスされうる等ができる。１又はそれ以上のクライアント１１４は、１又はそれ以上のストレージデバイス１５０を具えるサーバ１１２に常駐でき、別個のサーバ１１２に常駐でき、コンピュータネットワーク１１６を通してストレージデバイス１５０をアクセスするコンピュータ、ワークステーション、ラップトップ等に常駐できる等ができる。
【０３９０】
一実施例においては、ネットワーク１１６はシステムバスを具え、ストレージデバイスセット１６０４のストレージデバイス１５０、１６０２の１又はそれ以上は、システムバスを用いて通信する。例えば、システムバスはＰＣＩ－ｅバス、シリアルアドバンスドテクノロージアタッチメント（「シリアルＡＴＡ」）バス、パラレルＡＴＡ等にできる。別の実施例においては、システムバスは小型コンピュータシステムインタフェース（「ＳＣＳＩ」）、ＦｉｒｅＷｉｒｅ、ファイバチャネル、ＵＳＢ、ＰＣＩｅ－ＡＳ、インフィニバンド等のような外部バスである。当該技術分野の当業者は、自律的であり、ネットワーク１１６上のクライアント１１４からストレージ要求を別個に受信及び提供することができるストレージデバイス１５０を有する他のシステム１６００の構成が分かるであろう。
【０３９１】
図１３は、本発明による共有されたフロントエンド分散型ＲＡＩＤ用の装置２３００の一実施例を示す概略ブロック図である。装置２３００は、多重ストレージ要求受信モジュール２３０２、ストライピングモジュール２３０４、パリティミラーモジュール２３０６、シーケンサモジュール２３０８、マスタ確認モジュール２３１０、マスタ判定モジュール２３１２、マスタエラーモジュール２３１４、パリティ生成モジュール２３１６、及びパリティ代替モジュール２３１８を具え、以下に述べられている。
【０３９２】
装置２３００は、少なくとも２のクライアント１１４から少なくとも２のストレージ要求を受信する多重ストレージ要求受信モジュール２３０２を具えて、ストレージデバイスセット１６０２のストレージデバイス１５０にデータを記憶する。データはファイル又はオブジェクトのデータを含む。装置に関連するストレージ要求は、各々が共通のデータの少なくとも一部を有し、更に、１のストレージ要求がストレージ要求の到達前に完了されないように到達することによる同時のストレージ要求である。これらの同時のストレージ要求はフロントエンド分散型ＲＡＩＤシステム１６００中の共通データを破損する危険を冒す。一実施例においては、同時のストレージ要求は１のクライアント１１４からにできる。別の実施例においては、同時のストレージ要求は２又はそれ以上のクライアント１１４からになる。
【０３９３】
複数のストレージ要求がストレージデバイスセット１６０２のストレージデバイス１５０に記憶された１又はそれ以上のデータセグメントを更新でき、既に記憶されたデータが、ストレージデバイスセット１６０２のストレージデバイス１５０に記憶されたデータセグメント内にストライピングモジュール２３０４によってストライピングされる。一実施

例においては、ストレージ要求は初めてのデータをＲＡＩＤグループに書き込む。このケースにおいては、データは一般的に他の場所に存在し、１又はそれ以上のクライアント１１４によってアクセスし、あるストレージ要求はデータをＲＡＩＤグループにコピーする一方、別のストレージ要求はデータに同時アクセスする。

【０３９４】

複数のストレージ要求はストレージデバイスセット１６０２のストレージデバイス１５０に記憶される１又はそれ以上のデータセグメントを更新する１の要求と、共通にデータの少なくとも一部分を標的にする１又はそれ以上の読み出し要求を含みうる。更新要求が完全ではない場合に、次いでストレージデバイスセット１６０２のストレージデバイス１５０からの読み出し要求応答は、データを破損する既存の更新データの組合せを構成しうる。

【０３９５】

装置２３００は同時のストレージ要求の各々に対して、データ用のストライプパターンを算出し、ストライプのＮ個のデータセグメントをストレージデバイスセット１６０４内のＮ個のストレージデバイス１５０ａ－ｎに書き込むストライピングモジュール２３０４を具える。ストライプパターンは１又はそれ以上のストライプを含み、各ストライプはＮ個のデータセグメントのセットを含む。Ｎ個のデータセグメントの各々はストレージデバイスセット１６０４内の別個のストレージデバイス１５０に書き込まれ、ストライプに割り当てられる。装置２３００は同時のストレージ要求の各々に対して、ストライプのＮ個のデータセグメントをパリティミラーストレージデバイス１６０２として指定されたストレージデバイスセット１６０４内のストレージデバイス１５０に書き込むパリティミラーモジュール２３０６を具える。パリティミラーストレージデバイス１６０２はＮ個のストレージデバイス１５０ａ－ｎに加えられて存在する。

【０３９６】

ストライピングモジュール２３０４はファイル又はオブジェクトの一部である１又はそれ以上のデータセグメントを読み出すことから、１又はそれ以上のストレージデバイス１５０ａ－ｎの同一性を算出するのに更に用いられる。

【０３９７】

装置２３００は、第２のクライアント１１４からの第２のストレージ要求を実行する前に、第１のクライアント１１４からの第１のストレージ要求の完了を確認するシーケンサモジュール２３０８を具え、少なくとも２の同時のストレージ要求は第１及び第２のストレージ要求を含む。他の実施例においては、シーケンサモジュール２３０８は２又はそれ以上の他の同時のストレージ要求を実行する前に、第１のストレージ要求の完了を確認する。有益には、シーケンサモジュール２３０８は同時のストレージ要求の順番の実行を促進して、データの損失を防ぐ。一実施例においては、シーケンサモジュール２３０８は、総てのストレージ要求がデータ用にアクセスされなければならないマスタコントローラを用いることによって、ロッキングシステム、２相コミットを用いることによって、又は当該技術分野に公知の他の手段を用いることによって、同時のストレージ要求の実行を調整する。シーケンサモジュール２３０８によって用いられる方法の一部は以下に述べられている。

【０３９８】

一実施例においては、シーケンサモジュール２３０８は、第２のストレージ要求の実行前の第１のストレージ要求とともに、ストレージ要求を受信したストレージデバイスセット１６０２のストレージデバイス１５０の各々から受取り文字を受信することによって、同時のストレージ要求の実行前に第１のストレージ要求の完了を確認する。一般的には、受取り文字はストレージ要求の完了を通知する。一実施例においては、ストレージ要求によって影響が与えられるストレージデバイス１５０の各々は、シーケンサモジュール２３０８が第２のストレージ要求の実行を開始する前に、ストレージデバイス１５０の各々から受信された受取り文字に書き込まれる。

【０３９９】

一実施例においては、ストレージ要求の完了は、同一のストレージデバイス１５０ａの待ち状態の第２のストレージ要求の一部の実行前に、単一のストレージデバイス（例えば、１５０ａ）に命令された第１のストレージ要求の一部の完了を含みうる。シーケンサモジュール２３０８は、ストレージデバイス１５０のストレージ要求部分の完了を別個に確認できる。この実施例においては、第１のストレージ要求に関連するデータセグメントを書き込むことは、第１のストレージ要求の総てのデータセグメントが完了されるまで遅らせる必要はない。シーケンサモジュール２３０８はストレージデバイスセット１６０４のストレージデバイス１５０に生じる様々な実行を調整して、データが破損されないことを確認できる。

【０４００】

一実施例においては、ストレージ要求の完了についての受取り文字は、ストライピングモジュール２３０４及びパリティミラーモジュール２３０６がそれぞれ、ストレージ要求に関連するデータセグメントをストレージデバイスセット１６０４のストレージデバイス１５０に書き込んだ後に受信される。別の実施例においては、ストレージ要求の完了についての受取り文字は、ストライピングモジュール２３０４及びパリティミラーモジュール２３０６がそれぞれ、ストレージ要求に関連するデータセグメントをストレージデバイスセット１６０４のストレージデバイス１５０に書き込み、ストレージデバイス１５０、１６０２の各々がデータセグメントが書き込まれたことを確認した後に受信される。

【０４０１】

一実施例においては、シーケンサモジュール２３０８は、最初に到達した同時の要求の中から、ストレージ要求を選択することによって、実行用の第１のストレージ要求を選択する。別の実施例においては、シーケンサモジュール２３０８は、最も早いタイムスタンプを有するストレージ要求を選択することによって、実行用の第１のストレージ要求を選択する。別の実施例においては、シーケンサモジュール２３０８は、いくつかの選択基準を用いてストレージ要求を選択することによって、実行用の第１のストレージ要求を選択する。例えば、シーケンサモジュール２３０８は、高優先度として、要求クライアント１１４によっていくつかの方法で標識されたストレージ要求を選択、好まれるクライアント１１４からのストレージ要求を選択等をすることができる。当該技術分野の当業者は、シーケンサモジュール２３０８がいくつかの選択基準を用いて第１のストレージ要求を選択できる他の方法が分かるであろう。

【０４０２】

一実施例においては、ストレージ要求受信モジュール２３０２、ストライピングモジュール２３０４、パリティミラーモジュール２３０６、及びシーケンサモジュール２３０８は、同時のストレージ要求を制御し提供するマスタコントローラ（図示せず）の一部である。マスタコントローラの総て又は一部は、クライアント１１４、サードパーティのＲＡＩＤ管理デバイス、ストレージデバイスセット１６０４のストレージデバイス１５０、又はストレージデバイス１５０のストレージコントローラ１５２で常駐及び動作可能である。マスタコントローラを用いてデータ用のサービス要求を実行することによって、シーケンサモジュール２３０２はデータに向けられたストレージ要求を認識でき、次いで、同時のストレージ要求を認識でき、ストレージデバイスセットのストレージデバイス１５０に記憶されたデータが破損されない方法で同時のストレージ要求を配列できる。当該技術分野の当業者はデータに向けられたストレージ要求の提供を制御するマスタコントローラの他の実装が分かるであろう。

【０４０３】

別の実施例においては、マスタコントローラは、１又はそれ以上のクライアント１１４から同時のストレージ要求を提供できる２又はそれ以上のマスタコントローラのグループの一部であり、ストレージ要求はストレージデバイスセット１６０４のストレージデバイス１５０に記憶されたデータに向けられる。例えば、マスタコントローラは第１のクライアント１１４用のストレージ要求を提供でき、第２のマスタコントローラは第２のクライアント１１４用のストレージ要求を提供できる。第１及び第２のクライアント１１４は双

方ともストレージデバイスセット１６０４のストレージデバイス１５０に記憶されたデータにアクセスでき、このようにして同時のストレージを可能にする。あるマスタコントローラは１のストレージデバイス１５０ａの一部でできるが、他のマスタコントローラは第２のストレージデバイス１５０ｂの一部にできる。別の実施例においては、第１のマスタコントローラは第１のストレージデバイスセット１６０４ａの一部にでき、第２のマスタコントローラはミラーリングストレージデバイスセット１６０４ｂの一部にできる。

【０４０４】

マスタコントローラが、ストレージデバイスセット１６０４のストレージデバイス１５０にアクセスするマスタコントローラのグループの一部である場合、装置２３００は、受信されたストレージ要求を提供するマスタコントローラが、受信されたストレージ要求の実行前の１又はそれ以上の同時のストレージ要求の実行より先に、ストレージ要求の実行を制御していることを確認するマスタ確認モジュール２３１０を具えることができる。この実施例においては、同時のストレージ要求が他のマスタコントローラによって受信され、サービス要求は他のマスタコントローラによって受信される同時のストレージ要求と共通の少なくともデータの一部分を有する。

【０４０５】

例えば、マスタコントローラはストレージ要求を受信でき、マスタ確認モジュール２３１０は次いでストレージ要求の実行前に他のマスタコントローラをポーリングして、そのマスタコントローラが更にストレージ要求のデータ用のマスタコントローラであることを確認できる。確認の一部は、マスタコントローラが相互に通信する確認を含み、指定されたマスタコントローラがストレージ要求の実行前に確認できるようにする。これは、あるフロントエンドＲＡＩＤがマスタとして指定され、別のものがバックアップである場合に有用にできる。別の例においては、マスタコントローラはストレージ要求を受信して、ファイル又はオブジェクトからデータセグメントを読み出すストレージ要求を受信でき、マスタ確認モジュール２３１０は次いで他のマスタコントローラをポーリングして、更新がファイル又はオブジェクトで進行中ではないことを確認できる。別の例においては、マスタコントローラはマスタ確認モジュールを用いて、ストレージ要求用のデータの制御を取得できる。

【０４０６】

マスタコントローラがストレージ要求の実行用のマスタであると確認する１の方法は、スリーウェイポーリング方式を用いることであり、コントローラが進行するストレージ要求用のマスタであるとみなすために、２のデバイス／コントローラは利用可能でなければならない。その方式はマスタにすべきと争うコントローラに対しサードパーティとなるデバイス（図示せず）を用いて、どちらのコントローラがマスタにすべく割り当てられるかの記録を維持する。このマスタ確認デバイスは別のコントローラにし、サーバにあるクライアントにする等をすることができ、マスタコントローラとして動作できるグループでコントローラと通信できる。マスタ確認モジュール２３１０の一部はその後各コントローラに属するマスタ確認モジュール２３１０の一部を伴い、マスタ確認デバイスに常駐できる。

【０４０７】

一例においては、システム１６００はその各々をマスタにできる、第１のフロントエンド分散型ＲＡＩＤコントローラ（「第１のコントローラ」）及び第２のフロントエンド分散型ＲＡＩＤコントローラ（「第２のコントローラ」）と、別個のマスタ確認デバイスとを具える。第１及び第２のコントローラ及びマスタ確認デバイスは総て互いに通信する。マスタ確認デバイス２３１０は第１のコントローラをマスタコントローラとして、及び第２のコントローラをストレージデバイスセット１６０４のストレージデバイス１５０に記憶されたデータ用のバックアップとして指定でき、マスタ確認デバイス２３１０はコントローラ及びマスタ確認デバイスの、このマスタ情報を記憶できる。通信が第１のコントローラと、第２のコントローラと、マスタ確認デバイスとの間で維持される限り、マスタ確認モジュール２３１０は第１のコントローラがマスタであると確認できる。

【０４０８】
第１のマスタコントローラがストレージ要求を受信し、第２のバックアップコントローラが利用不可能になり、又は、第１のコントローラとマスタ確認デバイスで通信が損失した場合、マスタ確認モジュール２３１０はマスタ確認デバイスと第１のコントローラとの間の通信を通して第１のコントローラが未だマスタであることを確認し、第１のコントローラとマスタ確認デバイスの双方が、第１のコントローラが実際のマスタであることを確認するため、マスタ確認モジュール２３１０によりストレージ要求を進行することが可能となる。第２のバックアップコントにより受信されたストレージ要求はマスタ確認モジュール２３１０を通して、第２のコントローラがマスタではないと認識するため、進行しない。

【０４０９】
一方、第１のマスタコントローラが利用不可能であり、又は、第２のバックアップコントローラ及びマスタ確認デバイスと通信できず、第２のバックアップコントローラがストレージ要求を受信する場合、マスタ確認モジュール２３１０は、第２のコントローラとマスタ確認モジュールの双方が第１のコントローラと通信できないことを認識し、マスタ確認モジュール２３１０は第２のバックアップコントローラをマスタにすべきと指定し、ストレージ要求は進行できる。マスタ指定の変更は、その際第２のコントローラに記録されうる。

【０４１０】
第１のコントローラが動作し、第２のコントローラ及びマスタ確認デバイスで損失した通信のみがある場合、第１のコントローラによって受信されるデータ用のどのストレージ要求も実行されない。通信が復元される場合、第２のコントローラとマスタ確認デバイスの双方が、第２のコントローラをマスタと認識するため、第１のコントローラはストレージ要求をこれ以上実行しないであろう。当然にして、このマスタ指定はリセットされうる。当該技術分野の当業者はマスタコントローラのうちの１つにマスタ指定を割り当て及び再割り当てする、様々な静的及び動的手段が分かるであろう。

【０４１１】
マスタ確認デバイスが利用不可能であり、第１のストレージコントローラがストレージ要求を受信する場合、第１及び第２のコントローラで動作するマスタ確認モジュール２３１０の一部は、第１のコントローラがマスタであり、ストレージ要求が進行できることを確認できる。第２のコントローラがストレージ要求を受信する場合、第１及び第２のコントローラで動作するマスタ確認モジュール２３１０の一部は、第１のコントローラがマスタであり、ストレージ要求が進行できないことを確認できる。他の実施例においては、２以上のコントローラがポーリング方式の一部である。当該技術分野の当業者はマスタ確認モジュール２３１０がストレージ要求の実行の前にコントローラがマスタであることを確認できる他の方法が分かるであろう。

【０４１２】
別の実施例においては、装置２３００はマスタ判定モジュール２３１２を具える。ストレージ要求を送信する前に、マスタ判定モジュール２３１２はマスタコントローラのグループにマスタ判定要求を送信する。マスタコントローラのグループは次いで、どちらのコントローラがストレージ要求用のマスタとして指定されるかを識別し、マスタコントローラを識別する応答をマスタ判定モジュール２３１２に送信する。マスタ判定モジュール２３１２はストレージ要求用のマスタコントローラの識別表示を受信し、要求デバイスに命令して、ストレージ要求を指定されたマスタコントローラに送信する。一実施例においては、マスタ判定モジュール２３１２はクライアント１１４で常駐し動作する。別の実施例においては、マスタ判定モジュール２３１２はサードパーティのＲＡＩＤ管理デバイスで常駐し実行される。別の実施例においては、マスタ判定モジュール２３１２はストレージデバイス１５０に常駐する。別の実施例においては、マスタ判定モジュールは２以上のストレージデバイス１５０間で分散される。

【０４１３】

更なる実施例においては、装置２３００はエラー表示を返すマスタエラーモジュール２３１４を具える。一実施例においては、マスタコントローラによって制御される多重ストレージ要求受信モジュール２３０２が、マスタコントローラによって制御されないストレージ要求を受信する場合、マスタエラーモジュール２３１４はエラー表示を返す。

【０４１４】

別の実施例においては、マスタコントローラがストレージ要求の実行完了の時に、判定されたマスタではもはやないと、マスタ判定モジュール２３１２又はマスタ確認モジュール２３１０が判定する場合に、マスタエラーモジュール２３１４はエラー表示を返す。この実施例は一般的にマスタコントローラがストレージ要求を実行し始め、グループの他のマスタコントローラで通信を損失し、又は、ポーリング方式の場合、他のマスタコントローラ及びマスタ確認デバイスで通信を損失する場合に生じる。別の実施例においては、マスタコントローラによって制御される多重ストレージ要求受信モジュール２３０２が、マスタコントローラによって制御されないストレージ要求を受信する場合、マスタエラーモジュール２３１２はエラー表示を返す。

【０４１５】

別の実施例においては、マスタコントローラは１又はそれ以上のセカンダリマスタコントローラに対するストレージ要求を制御する。セカンダリマスタコントローラはそれぞれ、ストレージデバイスセット１６０４のストレージデバイス１５０、１６０２に記憶されたデータ用のストレージ要求を制御する。別の実施例においては、セカンダリマスタコントローラを制御するマスタコントローラは更に、ストレージデバイスセット１６０４のストレージデバイス１５０、１６０２に記憶されたデータに向けられたストレージ要求用のセカンダリマスタコントローラである。

【０４１６】

別の実施例においては、マスタコントローラは１又はそれ以上のセカンダリマスタコントローラに対するストレージ要求を制御し、各セカンダリマスタコントローラはセカンダリマスタコントローラに固有の、ストレージデバイスセットのストレージデバイス１５０に記憶されるデータ用のストレージ要求を制御する。装置２３００はフレキシブルであり、いずれのマスタコントローラも、セカンダリマスタコントローラとして動作する他のコントローラに対してマスタにできる。いくつかのセカンダリマスタコントローラはストレージデバイスセット１６０４を共有でき、その他のものは異なるストレージデバイスセットを制御できる。他の実施例においては、マスタコントローラはパリティミラーストレージデバイス１６０２、又は、Ｎ個のストレージデバイス１５０ａ－ｎのうちの１つにできる。

【０４１７】

別の実施例においては、マスタコントローラがオフラインであるか、指定されたマスタであると判定できない場合、セカンダリマスタコントローラはマスタコントローラにできる。当該技術分野の当業者は、１又はそれ以上のセカンダリマスタコントローラの間で、マスタ指定を割り当て及び再割り当てするための様々な静的及び動的手段が分かるであろう。

【０４１８】

好ましい実施例においては、装置２３００はストライプ用のパリティデータセグメントを算出し、パリティミラーストレージデバイス１６０２にパリティデータセグメントを記憶するパリティ生成モジュール２３１６を具える。パリティストライプはパリティミラーストレージデバイス１６０２のＮ個のデータセグメントのセットから算出される。この実施例は一般的なＲＡＩＤ５、ＲＡＩＤ６、又はその他のＲＡＩＤレベルであるが、一般的にはＲＡＩＤ０、ＲＡＩＤ１、ＲＡＩＤ１０等には含まれない。

【０４１９】

別の好ましい実施例においては、装置２３００は各ストライプに対して、ストレージデバイスセット１６０４内のストレージデバイス１５０のどちらが、ストライプ用の１又はそれ以上のパリティミラーストレージデバイス１６０２であると割り当てられるかを代替

するパリティ代替モジュール２３１８を具える。ストライプごとにパリティデータセグメントをローテーションすることは、性能を改善する。パリティ代替モジュール２３１８はストライピングモジュール２３０４とともに用いられ、ファイル又はオブジェクトの一部である１又はそれ以上のデータセグメントを読み出し、書き込み、又は更新することから、１又はそれ以上のストレージデバイス１５０ａ－ｎの同一性を計算できる。

【０４２０】

様々なモジュール２３０２－２３１８の機能は、単一のマスタコントローラに一緒に常駐でき、又は、１又はそれ以上のクライアント１１４、サードパーティのＲＡＩＤ管理デバイス、及び１又はそれ以上のストレージデバイス１５０、１６０２の間で分散されうる。当該技術分野の当業者は本明細書に記載の機能が分散される様々な実施例が分かるであろう。

【０４２１】

図１４は、本発明による共有されたフロントエンド分散型ＲＡＩＤ用の方法２４００の実施例を示す概略フローチャート図である。

方法２４００は開始し（２４０２）、多重ストレージ要求受信モジュール２３０２は少なくとも２のクライアント１１４から少なくとも２のストレージ要求を受信して（２４０４）、ストレージデバイスセット１６０４の１又はそれ以上のストレージデバイス１５０のデータを読み出すか記憶する。データは、ファイル又はオブジェクトからのものであり、ストレージ要求はそれぞれ、共通にデータの少なくとも一部分を有し、到達によって同時になり、別の少なくとも２のストレージ要求の到達前にストレージ要求が完了しない。ストライピングモジュール２３０４はデータ用のストライプパターンを算出し（２４０６）、ストライプパターンは１又はそれ以上のストライプを含み、各ストライプはＮ個のデータセグメントのセットを含む。ストライピングモジュール２３０４は更に、ストレージデバイスセット１６０４内のＮ個のストレージデバイス１５０ａ－ｎにストライプのＮ個のデータセグメントを読み出し又は書き込み（２４０８）、Ｎ個のデータセグメントの各々は、別個のストレージデバイス１５０に書き込まれ又はそこから読み出される。

【０４２２】

ストレージ要求が書き込み動作である場合、パリティミラーモジュール２３０６はストライプのＮ個のデータセグメントのセットをストレージデバイスセット１６０４内の１又はそれ以上のパリティミラーストレージデバイス１６０２に書き込み（２４１０）、パリティミラーストレージデバイス１６０２はＮ個のストレージデバイス１５０ａ－ｎに加えた状態になる。パリティミラーモジュール２３０６は更に、パリティミラーデバイス１６０２に記憶されたデータセグメント、又はパリティデータセグメントを読み出すことができる（２４１０）。シーケンサモジュール２３０８は第２のクライアント１１４からの第２のストレージ要求を実行する前に、第１のクライアント１１４からの第１のストレージ要求の完了を確認し（２４１２）、方法２４００は終了する（２４１６）。第１及び第２のストレージ要求は同時のストレージ要求である。

【０４２３】

本発明は、本発明の精神または本質的特徴から逸脱することなく、その他の特定の形態で実施され得る。示された実施例は、例示のみであって、限定ではないと総ての点で見なされるべきである。本発明の範囲は、従って、上述した説明ではなく、添付の請求の範囲によって規定される。請求項と同等の意味および範囲にある総ての変更は、本発明の範囲に包含されるべきである。

【０４２４】

高容量不揮発性ストレージ用のキャッシュとしてのソリッドステートストレージ

【０４２５】

一般的にキャッシュは、しばしばアクセスされ、アプリケーション又はオペレーティングシステムの一部として配置されるデータが、ハードディスクドライブ（「ＨＤＤ」）、光学ドライブ、テープストレージ等のような高容量不揮発性（「ＨＣＮＶ」）ストレージデバイスを通してアクセスされなければならない場合よりも更に早い次のアクセスを伴う

キャッシュ内に記憶されるので有利である。キャッシュは一般的にはコンピュータ内に含まれる。

【０４２６】

いくつかのストレージデバイス及びシステムはＨＣＮＶストレージデバイスにキャッシュを具えている。いくつかのＨＣＮＶストレージデバイスは不揮発性ソリッドステートキャッシュを含み、これらはアクセス時間を低減する利益を提供するが、通常は限定されるＨＣＮＶストレージデバイスインタフェースの容量と一致する性能を提供するのみである。マザーボード上に一般的に配置されるいくつかの不揮発性ソリッドステートキャッシュストレージデバイスが存在するが、これらのデバイスはキャッシュコヒーレンスが提供されないようなマルチクライアント環境において用いることができない。ＨＣＮＶデバイスのいくつかのコントローラは更にキャッシュを具える。冗長ＨＣＮＶキャッシュコントローラが複数クライアント間で共有される場合、データが破損されていないことを確認するのに高性能なキャッシュコヒーレンシアルゴリズムが要求される。

【０４２７】

一般的に、キャッシュはＤＲＡＭ内に実装され、キャッシュ容量をプレミアムにし、性能ごとに比較的高い電源を要求する。揮発性キャッシュを支持する電源が失われる場合、キャッシュに記憶されたデータが失われる。一般的にバッテリバックアップが、電源故障の場合にデータ損失を避けるのに用いられるが、バッテリバックアップの停止前に不揮発性メモリに対するキャッシュを消去するかなりの可能性を有する。更に、バッテリバックアップシステムは電力を消費し、冗長性を要求し、信頼性に負の影響を与え、空き領域を消費する。バッテリは標準ベースで更に提供されなければならず、バッテリバックアップは比較的高価である。

【０４２８】

前述の考察から、キャッシュとしてソリッドステートストレージを用いてデータを管理する装置、システム及び方法に対するニーズが存在することは明らかである。有利には、このような装置、システム及び方法は少ない電力を消費し、かなり大きな容量を有し、キャッシュに記憶されたデータを維持するのにバッテリバックアップを要求しない不揮発性キャッシュを提供するであろう。

【０４２９】

図１５は、本発明による高容量不揮発性ストレージデバイス用のキャッシュとしてソリッドステートストレージ１１０を有するシステム３４００の一実施例を示す概略ブロック図である。システム３４００はソリッドステートストレージコントローラ１０４及びＨＣＬＶコントローラ３４０２と、ソリッドステートストレージ１１０と、ネットワークインタフェース１５６とを具えるストレージコントローラ１５２を有するソリッドステートストレージデバイス１０２を具える。システム３４００はコンピュータネットワーク１１６を通してソリッドステートストレージデバイス１０２に接続された要求デバイス１５５と、１又はそれ以上のＨＣＮＶストレージデバイス３４０４ａ－ｎとを具える。当該技術分野の当業者は図１５に示されたシステム３４００が単なる一実施例であって、ソリッドステートストレージ１１０をストレージデバイス用のキャッシュにできる多くのその他の構成が可能であることが分かるであろう。

【０４３０】

システム３４００はネットワークインタフェース１５６とストレージコントローラ１５２とを有するソリッドステートストレージデバイス１０２を具える。別の実施例においては、ネットワークインタフェース１５６はソリッドステートストレージデバイス１０２の外側にある。例えば、ネットワークインタフェース１５６はソリッドステートストレージデバイス１０２を具える又は具えないサーバ１１２内にできる。

【０４３１】

示された実施例においては、ソリッドステートストレージデバイス１０２は、ソリッドステートストレージコントローラ１０４と、高容量不揮発性（「ＨＣＮＶ」）ストレージコントローラ３４０２とを具えるストレージコントローラ１５２を具える。別の実施例に

おいては、ソリッドステートストレージデバイス１０２はストレージコントローラ１５２内にない、ソリッドステートストレージコントローラ１０４及びＨＣＮＶストレージコントローラ３４０２を具える。他の実施例においては、ソリッドステートストレージデバイス１０２はＨＣＮＶストレージコントローラ３４０２を含むソリッドステートストレージコントローラ１０４を具え、その逆もできる。

【０４３２】

示された実施例においては、システム３４００は統合型ソリッドステートストレージ１１０を有し、外部ＨＣＮＶストレージデバイス３４０４ａ－ｎを有するソリッドステートストレージデバイス１０２を具える。別の実施例においては、ストレージコントローラ１５２、１０４、３４０２はソリッドステートストレージ１１０から分離できる。別の実施例においては、コントローラ１５２、１０４、３４０２及びソリッドステートストレージ１１０は、ＨＣＮＶストレージデバイス３４０４に含まれる。ＨＣＮＶストレージデバイス３４０４はネットワークインタフェース１５６を更に具えることができる。当該技術分野の当業者は、多くのその他の構成が可能であることが分かるであろう。ソリッドステートストレージデバイス１０２、ソリッドステートストレージコントローラ１０４、ソリッドステートストレージ１１０、ストレージＩ／Ｏバス２１０、ネットワークインタフェース１５６、コンピュータネットワーク１１６、及び要求デバイス１５５は上述のデバイス及びバスの他の実施例と実質的に同様である。

【０４３３】

一実施例においては、要求デバイス１５５はシステムバスを通してソリッドステートストレージデバイス１０２、ストレージコントローラ１５２、ソリッドステートストレージコントローラ１０４等に接続される。要求デバイス１５５と、ソリッドステートストレージ１１０との間のデータ転送は、システムバスで生じうる。

【０４３４】

ＨＣＮＶストレージデバイス３４０４は一般的には不揮発性ストレージを提供する高容量ストレージデバイスであり、一般的にはソリッドステートストレージ１１０よりもデータを読み出すこと及び書き出すことについては遅い。ＨＣＮＶストレージデバイス３４０４はソリッドステートストレージ１１０よりも単位あたりのストレージ容量が安くなる。ＨＣＮＶストレージデバイス３４０４はハードディスクドライブ（「ＨＤＤ」）、光学ドライブ、テープストレージ等にできる。ＨＣＮＶストレージデバイス３４０４用のキャッシュとしてソリッドステートストレージ１１０を提供することは、一般的にデータアクセス及びストレージの速度を増加させる。当該技術分野の当業者は、ＨＣＮＶストレージデバイス３４０４用のキャッシュとしてのソリッドステートストレージ１１０の利点が分かるであろう。

【０４３５】

一実施例においては、ＨＣＮＶストレージデバイス３４０４はストレージエリアネットワーク（「ＳＡＮ」）を通してストレージコントローラ１５２に接続される。一実施例においては、別個のＳＡＮコントローラがＨＣＮＶストレージデバイス３４０４をストレージコントローラ１５２に接続する。別の実施例においては、ＨＣＮＶストレージコントローラ３４０２又はストレージコントローラ１５２はＳＡＮコントローラとして作用する。当該技術分野の当業者はＨＣＮＶストレージデバイス３４０４がＳＡＮに接続されうる他の方法が分かるであろう。

【０４３６】

図１６は、本発明による高容量不揮発性ストレージデバイス用のキャッシュとしてのソリッドステートストレージを有する装置３５００の一実施例を示した概略ブロック図である。装置３５００はキャッシュフロントエンドモジュール３５０２、キャッシュバックエンドモジュール３５０４、オブジェクトストレージコントローラ３５０６、ＨＣＮＶモジュール３５０８、及び標準デバイスエミュレーションモジュール３５１０を具え、以下に説明されている。装置３５００のモジュール３５０２－３５１０は、ソリッドステートストレージコントローラ１０４と、ＨＣＮＶストレージコントローラ３４０２とを有するス

トレージコントローラ１５２に示されているが、各モジュール３５０２－３５１０の一部及び総てがソリッドステートストレージコントローラ１０４、ＨＣＮＶストレージコントローラ３４０２、サーバ１１２、ＨＣＮＶストレージデバイス３４０４、又はその他の場所に含まれている。

【０４３７】

装置３５００はストレージ要求と関連するデータ転送を管理するキャッシュフロントエンドモジュール３５０２を具え、データ転送は要求デバイス１５５とソリッドステートストレージ１１０との間で、１又はそれ以上のＨＣＮＶストレージデバイス３４０４ａ－ｎ用のキャッシュとして機能している。装置３５００はソリッドステートストレージ１１０と、ＨＣＮＶストレージデバイス３４０４ａ－ｎとの間のデータ転送を管理するキャッシュバックエンドモジュール３５０４を更に具える。データ転送はデータ、メタデータ、及び／又はメタデータインデックスを含みうる。上述のように、ソリッドステートストレージ１１０は不揮発性のソリッドステートデータストレージエレメントのアレイ２１６、２１８、２２０であり、一般的にバンク２１４に配置される。様々な実施例においては、ソリッドステートストレージ１１０はフラッシュメモリ、ナノランダムアクセスメモリ（「ナノＲＡＭ」又は「ＮＲＡＭ」）、磁気抵抗ＲＡＭ（「ＭＲＡＭ」）、ダイナミックＲＡＭ（「ＤＲＡＭ」）、相変化ＲＡＭ（「ＰＲＡＭ」）等にできる。

【０４３８】

一般的に、キャッシュフロントエンドモジュール３５０２、キャッシュバックエンドモジュール３５０４、及びソリッドステートストレージコントローラ１０４は要求デバイス１５５から自律して動作する。例えば、要求デバイス１５５は単一のストレージデバイスとして付随するストレージ１１０、３４０４ａ－ｎとともに、ソリッドステートストレージコントローラ１０４及びＨＣＮＶストレージコントローラ３４０２を有するストレージコントローラ１５２を表示する。別の例においては、要求デバイス１５５はＨＣＮＶストレージデバイス３４０４ａ－ｎを表示でき、ソリッドステートストレージ１１０は非表示にできる。

【０４３９】

一実施例においては、ソリッドステートストレージコントローラ１０４は１又はそれ以上の要求デバイス１５５からオブジェクト要求を提供し、ソリッドステートストレージ１１０内のオブジェクト要求のオブジェクトを管理するオブジェクトストレージコントローラモジュール３５０６を具える。その実施例においては、ソリッドステートコントローラ１０４はオブジェクトストレージコントローラモジュール３５０６とともに、上述のような、特に図２Ａに示した装置に関連して記載されるような、オブジェクト要求を管理する。

【０４４０】

一実施例においては、装置３５００はＲＡＩＤレベルの一致する独立ドライブ冗長アレイ（「ＲＡＩＤ」）の２又はそれ以上のＨＣＮＶストレージデバイス３４０４ａ－ｎ内のソリッドステートストレージ１１０にキャッシュされたデータを記憶するＨＣＮＶ　ＲＡＩＤモジュール３５０８を具える。その実施例においては、データは全体として要求デバイス１５５に見え、ＲＡＩＤすることは要求デバイス１５５から隠されるようになる。例えば、キャッシュフロントエンドモジュール３５０２はソリッドステートストレージ１１０に要求デバイス１５５からデータをキャッシュでき、キャッシュバックエンドモジュール３５０４は、ＨＣＮＶ　ＲＡＩＤモジュール３５０８と提携して、データをストライプし、ＲＡＩＤレベルの一致するＨＣＮＶストレージデバイス３４０４ａ－ｎにデータセグメント及びパリティデータセグメントを記憶できる。当該技術分野の当業者は、要求デバイス１５５からのデータがＨＣＮＶストレージデバイス３４０４ａ－ｎにＲＡＩＤできる他の方法が分かるであろう。

【０４４１】

別の実施例においては、ソリッドステートストレージ１１０及びＨＣＮＶストレージデバイスは、ＲＡＩＤグループとして構成されたハイブリッドストレージデバイスセット内

のハイブリッドストレージデバイスを具える。例えば、ハイブリッドストレージデバイスセットは、フロントエンド分散型ＲＡＩＤストレージデバイスセット１６０４にでき、ハイブリッドストレージデバイスは図１０、１１、及び１２にそれぞれ示されたシステム１６００、装置２１００、及び方法２２００に関連して上述のように設定されたストレージデバイスのストレージデバイス１５０、１６０２にできる。その実施例においては、ソリッドステートストレージ１１０にキャッシュされ、ＨＣＮＶデバイス３４０４で後に記憶されるデータセグメントは、ストライプのＮ個のデータセグメントのうちの１つ、又はストライプのパリティデータセグメントである。フロントエンドＲＡＩＤと同様、ハイブリッドストレージデバイスはＲＡＩＤストライプのデータセグメントと別個に、１又はそれ以上のクライアント１１４からストレージ要求を受信する。

【０４４２】

更なる実施例においては、ハイブリッドストレージデバイスは、２又はそれ以上のクライアント１１４から２又はそれ以上の同時のストレージ要求を受信する共有されたフロントエンド分散型ＲＡＩＤグループ１６０４のストレージデバイス１５０、１６０２であり、図１０、１３、及び１４にそれぞれ示されたシステム１６００、装置２３００、及び方法２４００に関連して述べたような共有されたフロントエンドＲＡＩＤと関連して上述される。有益には、この実施例は、共有された冗長キャッシュが、更なる複雑なコヒーレンスアルゴリズム及びプロトコルを有することなく、コヒーレンスを維持することを保証する。

【０４４３】

別の実施例においては、ソリッドステートストレージ１１０及びＨＣＮＶストレージデバイス３４０４ａ－ｎはハイブリッドストレージデバイスを具え、装置３５００はハイブリッドストレージデバイスの動作に特有のコードで、要求デバイス１５５をロードする前に、１又はそれ以上の要求デバイス１５５に取り付けられた標準デバイスをエミュレートすることによって、ハイブリッドストレージデバイスにアクセスを提供する標準デバイスエミュレーションモジュール３５１０を具える。その実施例においては、標準デバイスは工業規格ＢＩＯＳによって担持される。このブートストラップ動作によって、ハイブリッドデバイスが、ソリッドステートストレージコントローラ１０４、ＨＣＮＶストレージコントローラ３４０４、及び装置３５００のできる限りの他のモジュール３５０２－３５１０に特有のドライバが、要求デバイス１５５にロードできるまで、限られた機能性を有する要求デバイス１５５によって認知、アクセスされることが可能である。

【０４４４】

一実施例においては、ソリッドステートストレージデバイス１１０は２又はそれ以上の領域に分割可能であり、１又はそれ以上のパーティションが、ＨＣＮＶストレージデバイス３４０４用のキャッシュとして機能するソリッドステートストレージと別個の、ソリッドステートストレージとして用いられる。例えば、ソリッドステートストレージ１１０のいくつかのパーティションは汎用データストレージ用のクライアント１１４によってアクセスできる一方、１又はそれ以上のパーティションはＨＣＮＶストレージデバイス３４０４用のキャッシュとして割り当てられる。

【０４４５】

一実施例においては、１以上のクライアント１１４（又は要求デバイス１５５）はキャッシュフロントエンドモジュール３５０２及びキャッシュバックエンドモジュール３５０４にキャッシュ制御メッセージを送信して、ソリッドステートストレージデバイス１１０及び１又はそれ以上のＨＣＮＶストレージデバイス３４０４内に記憶されるファイル又はオブジェクトのうちの１又はそれ以上の状態を管理できる。有益には、ファイルごと、オブジェクトごと、又はデータセグメントベースごとのキャッシュを管理するクライアント１１４／要求デバイス１５５の能力は、ソリッドステートストレージデバイス１１０を共有するようにかなりのフレキシビリティを提供する。

【０４４６】

多数のキャッシュ制御メッセージが許容可能であり、実現可能である。例えば、キャッ

シュ制御メッセージは、キャッシュバックエンドモジュール３５０４に、ソリッドステートストレージ１１０内のオブジェクト又はファイルの一部を留める（ｐｉｎ）制御メッセージを含みうる。別のキャッシュ制御メッセージは、キャッシュバックエンドモジュール３５０４に、ソリッドステートストレージ１１０内のオブジェクト又はファイルの一部を解放する（ｕｐｐｉｎ）制御メッセージを含みうる。別のキャッシュ制御メッセージは、キャッシュバックエンドモジュール３４０４に、ソリッドステートストレージ１１０から１又はそれ以上のＨＣＶＮストレージデバイス３４０４までのオブジェクト又はファイルの一部を消去させる制御メッセージを含みうる。別のキャッシュ制御メッセージは、キャッシュバックエンドモジュール３４０４に、１又はそれ以上のＨＣＶＮストレージデバイス３４０４からソリッドステートストレージ１１０に、オブジェクト又はファイルの一部を事前ロードさせる制御メッセージを含みうる。別のキャッシュ制御メッセージは、ソリッドステートストレージ１１０に、決められたストレージ空き領域量を空けるために、キャッシュバックエンドモジュール３５０４に、ソリッドステートストレージから１又はそれ以上のＨＣＶＮストレージデバイス３４０４までの１又はそれ以上のオブジェクト又はファイルの１又はそれ以上の部分をオフロードさせる制御メッセージを含みうる。当該技術分野の当業者は、他の可能なキャッシュ制御メッセージが分かるであろう。

【０４４７】

一実施例においては、キャッシュ制御メッセージはオブジェクト又はファイル用のメタデータ（「キャッシュ制御メタデータ」）を通して通信される。一実施例においては、キャッシュ制御メタデータは永続的である。別の実施例においては、キャッシュ制御メタデータはファイル又はオブジェクトの生成時に設定される属性を通して構築される。その実施例においては、属性は特定のオブジェクトクラス、デフォルト、特定のファイル型の特徴等に対する関係を通して継承されうる。当該技術分野の当業者は、キャッシュ制御メッセージがメタデータを通して通信されうる他の方法が分かるであろう。

【０４４８】

一実施例においては、システム３４００は揮発性キャッシュストレージエレメントを具える。例えば、ソリッドステートストレージ１１０に加えて、システム３４００は揮発性のいくつかのタイプのランダムアクセスメモリ（「ＲＡＭ」）を更に具えることができる。この実施例においては、

キャッシュフロントエンドモジュール３５０２及びキャッシュバックエンドモジュール３５０４は揮発性キャッシュストレージエレメントにいくつかのデータを記憶し、ソリッドステートストレージ１１０及び揮発性キャッシュストレージエレメントに記憶されたデータを管理し、バックエンドストレージモジュール３５０４は、揮発性キャッシュストレージエレメントと、ソリッドステートストレージと、ＨＣＶＮストレージデバイスとの間のデータ転送を更に管理する。例えば、重要ではない、又は、別のソースから簡単に回復されうるデータは揮発性キャッシュに記憶できる一方、その他のデータは、キャッシュとして機能するソリッドステートストレージ１１０に記憶できる。

【０４４９】

更なる実施例においては、ＨＣＶＮストレージデバイス３４０４に記憶されたオブジェクト及びファイル用のメタデータ及び／又はインデックスメタデータは、ソリッドステートストレージデバイス１１０内及び揮発性キャッシュストレージエレメントに維持される。図２Ａに示された装置２００に関連して上述したように、特定のメタデータは揮発性キャッシュストレージエレメントに記憶でき、揮発性キャッシュストレージエレメント中のデータが損失した場合に、インデックスを再構築するのに用いられうる。一実施例においては、メタデータ及びインデックスメタデータは、ソリッドステートストレージ１１０に記憶され、揮発性キャッシュストレージエレメントは含まれない。当該技術分野の当業者は、キャッシュとして機能するソリッドステートストレージ１１０と共に揮発性キャッシュストレージエレメントを用いる他の利点及び方法が分かるであろう。

【０４５０】

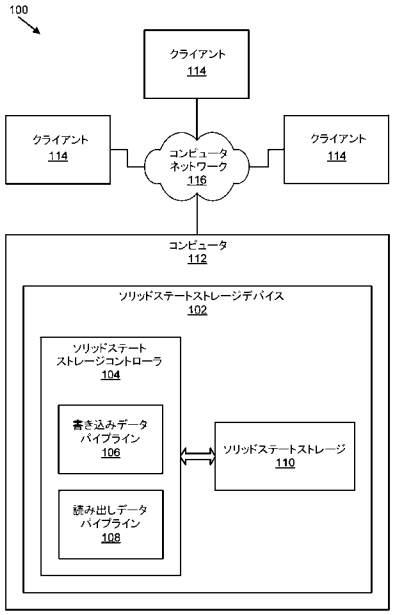
図１７は、本発明による高容量不揮発性ストレージデバイス用のキャッシュとしてソリ

ッドステートストレージを有する方法３６００の一実施例を示した概略フローチャート図である。方法３６００は開始し（３６０２）、キャッシュフロントエンドモジュール３５０２はストレージ要求と関連するデータ転送を管理し（３６０４）、データ転送は、要求デバイス１５５とソリッドステートストレージ１１０との間で、１又はそれ以上のＨＣＮＶストレージデバイス３４０４ａ－ｎ用のキャッシュとして機能している。キャッシュバックエンドモジュール３５０４は、ソリッドステートストレージ１１０と、１又はそれ以上のＨＣＮＶストレージデバイス１１０との間のデータ転送を管理し（３６０６）、方法３６００は終了する（３６０８）。方法３６００は図１０の装置３５００に関連して上述したのと実質的に同様に動作する。

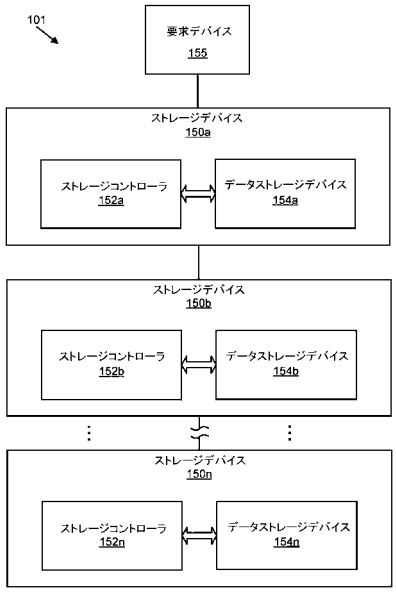
【０４５１】

本発明は、本発明の精神または本質的特徴から逸脱することなく、その他の特定の形態で実施され得る。示された実施例は、例示のみであって、限定ではないと総ての点で見なされるべきである。本発明の範囲は、従って、上述した説明ではなく、添付の請求の範囲によって規定される。請求項と同等の意味および範囲にある総ての変更は、本発明の範囲に包含されるべきである。

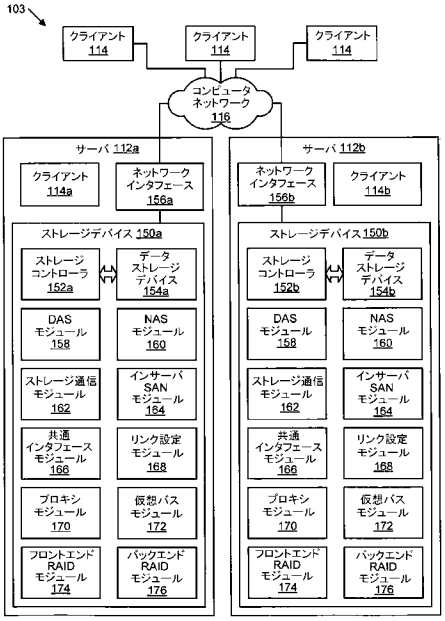
【図１Ａ】



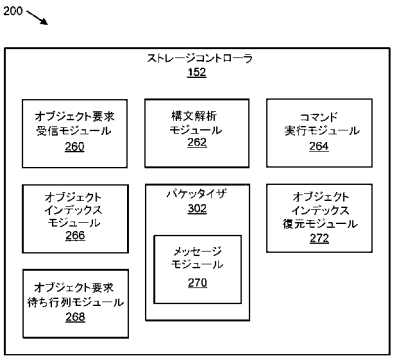
【図１Ｂ】



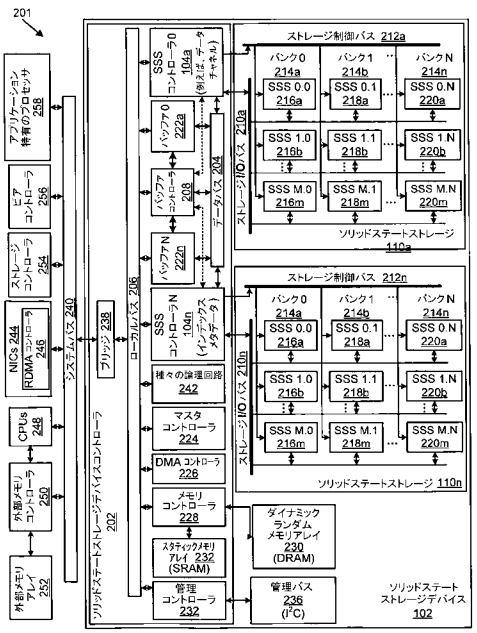
【図１Ｃ】



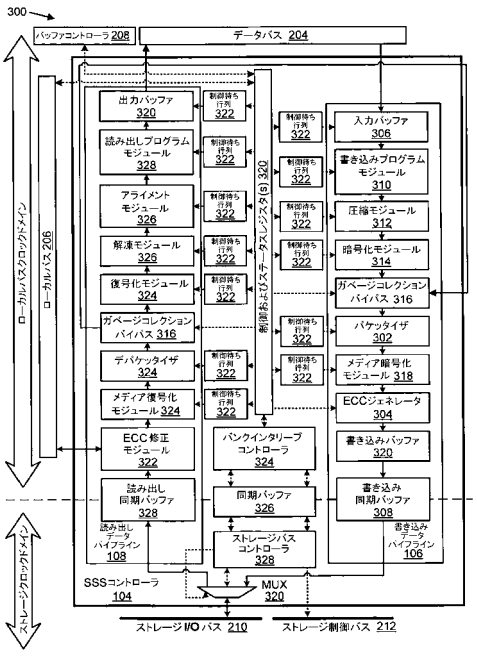
【図２Ａ】



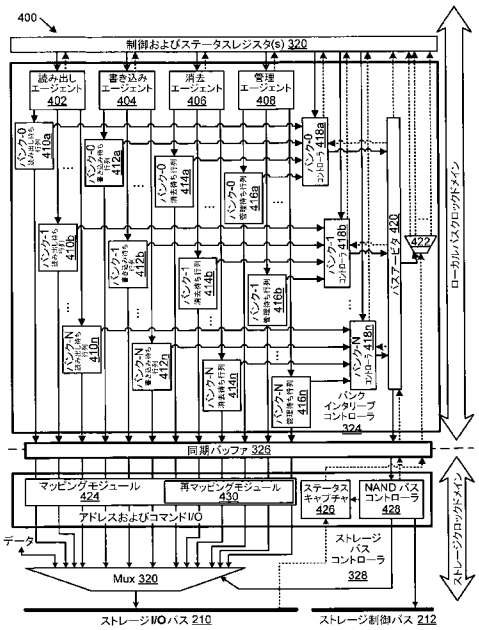
【図２Ｂ】



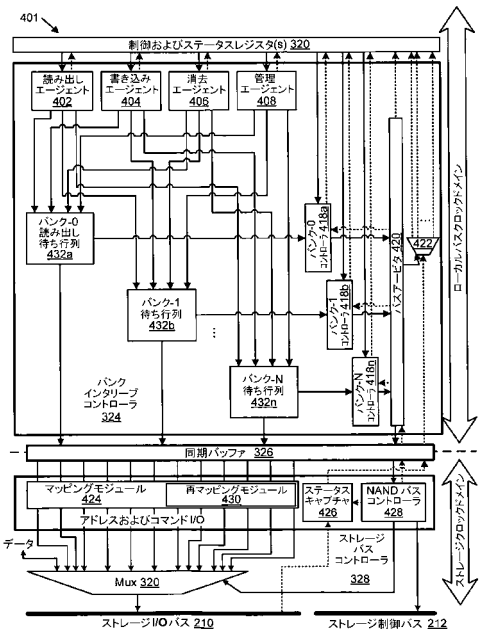
【図３】



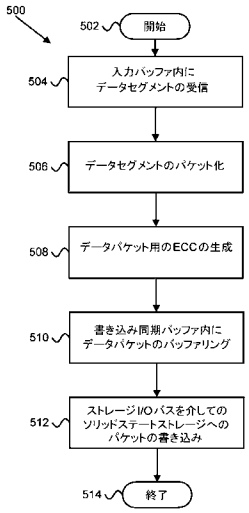
【図４Ａ】



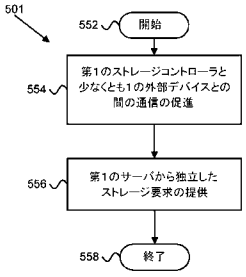
【図４Ｂ】



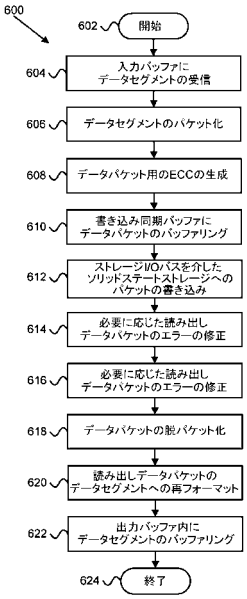
【図５Ａ】



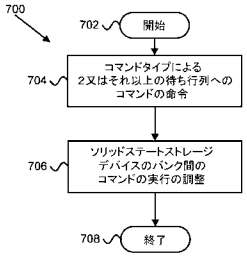
【図５Ｂ】



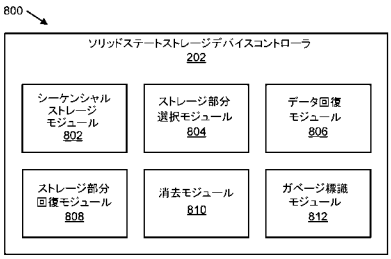
【図６】



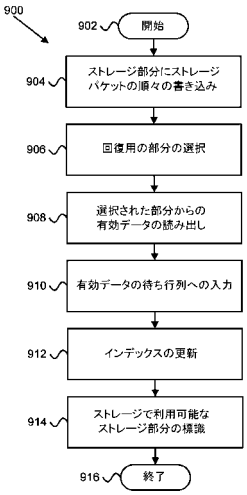
【図７】



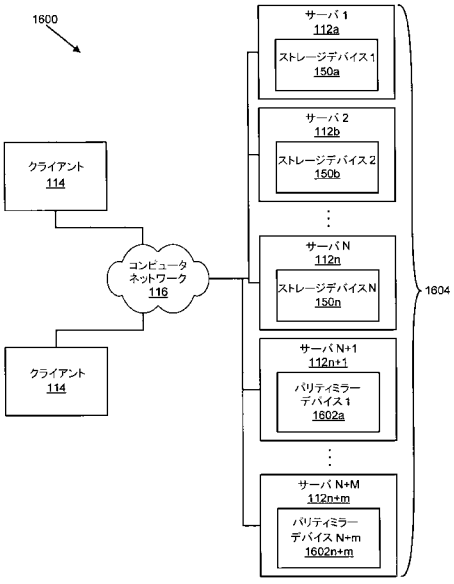
【図８】



【図９】

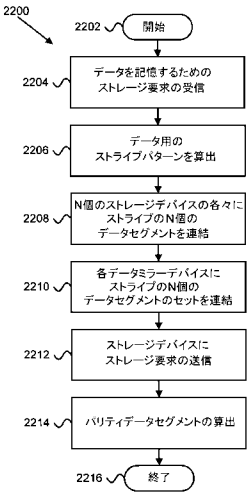


【図１０】

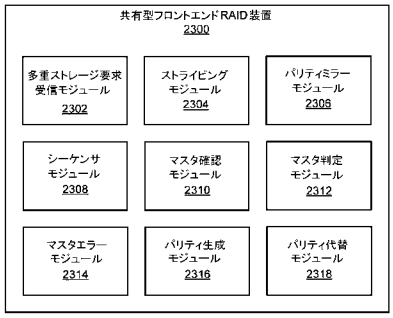


【図１１】

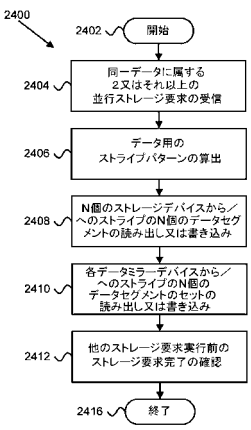
【図１２】



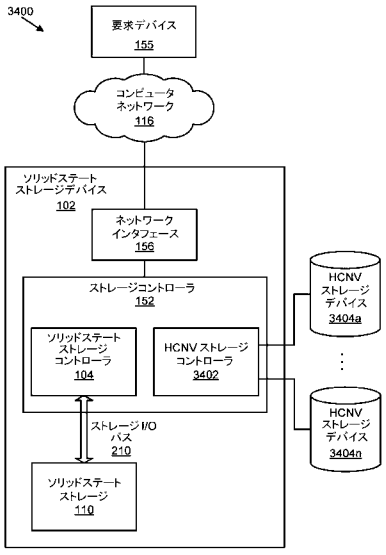
【図１３】



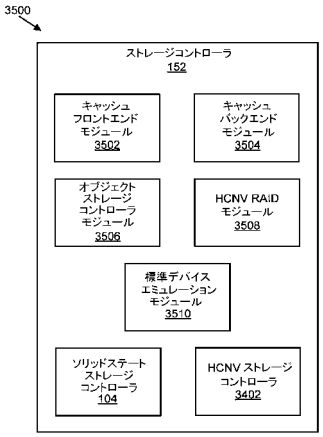
【図１４】



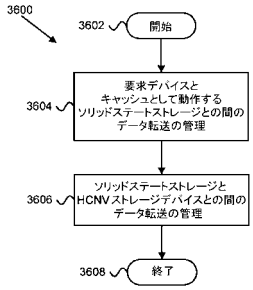
【図１５】



【図１６】



【図１７】



【手続補正書】
【提出日】平成20年10月9日(2008.10.9)
【手続補正１】
【補正対象書類名】特許請求の範囲
【補正対象項目名】全文
【補正方法】変更
【補正の内容】
【特許請求の範囲】
【請求項１】
１又はそれ以上の高容量不揮発性（「ＨＣＮＶ」）ストレージデバイスでデータのストレージを管理するための装置であって、当該装置が、
ストレージ要求に関連するデータ転送を管理し、要求デバイスとソリッドステートストレージとの間の前記データ転送が、１又はそれ以上のＨＣＮＶストレージデバイス用のキャッシュとして機能し、前記データ転送がデータ、メタデータ及びメタデータインデックスのうちの１又はそれ以上を含み、前記ソリッドステートストレージが不揮発性のソリッドステートデータストレージエレメントのアレイを具えるキャッシュフロントエンドモジュールと；
前記ソリッドステートストレージと、前記１又はそれ以上のＨＣＮＶストレージデバイスとの間のデータ転送を管理するキャッシュバックエンドモジュールと；
を具え、ソリッドステートコントローラが、１又はそれ以上の要求デバイスからオブジェクト要求を提供し、前記ソリッドステートストレージ内の前記オブジェクト要求のオブジェクトを管理するオブジェクトストレージコントローラモジュールを更に具えることを特徴とする装置。
【請求項２】
請求項１に記載の装置において、前記キャッシュフロントエンドモジュール及び前記キャッシュバックエンドモジュールが、前記ソリッドステートストレージを管理するソリッドステートストレージコントローラと共に同一位置に配置されることを特徴とする装置。
【請求項３】
請求項１に記載の装置において、前記キャッシュフロントエンドモジュール、前記キャッシュバックエンドモジュール、及びソリッドステートストレージコントローラが、前記要求デバイスから自律して動作することを特徴とする装置。
【請求項４】
請求項１に記載の装置において、前記ソリッドステートコントローラが、現在のデータの直前に記憶されたデータに追加された、前記現在のデータを記憶するログストレージデバイスを具えることを特徴とする装置。
【請求項５】
請求項１に記載の装置が、ＲＡＩＤレベルの一致する独立ドライブ冗長アレイ（「ＲＡＩＤ」）の２又はそれ以上のＨＣＮＶストレージデバイス内の前記ソリッドステートストレージにキャッシュされたデータを記憶するＨＣＮＶ　ＲＡＩＤモジュールを更に具え、前記データが全体として要求デバイスに見えることを特徴とする装置。
【請求項６】
請求項１に記載の装置において、前記ソリッドステートストレージ及び前記１又はそれ以上のＨＣＮＶストレージデバイスが、ＲＡＩＤグループとして構成されるハイブリッドストレージデバイスセット内にハイブリッドストレージデバイスを具え、前記ソリッドステートストレージにキャッシュされ、ＨＣＮＶデバイスに後に記憶されるデータセグメントが、ストライプのＮ個のデータセグメントのうちの１つ、又は、前記ストライプのパリティデータセグメントを含み、前記ハイブリッドストレージデバイスがＲＡＩＤストライプのデータセグメントと別個の、１又はそれ以上のクライアントからストレージ要求を受信することを特徴とする装置。
【請求項７】

請求項６に記載の装置において、前記ハイブリッドストレージデバイスが２又はそれ以上のクライアントから２又はそれ以上の同時のストレージ要求を受信する、共有されたフロントエンド分散型ＲＡＩＤグループのストレージデバイスであることを特徴とする装置。
【請求項８】
請求項１に記載の装置において、前記ＨＣＮＶストレージデバイスがハードディスクドライブ（「ＨＤＤ」）、光学ドライブ、及びテープストレージのうちの１つであることを特徴とする装置。
【請求項９】
請求項１に記載の装置において、前記ソリッドステートストレージ及び前記１又はそれ以上のＨＣＮＶストレージデバイスが、ハイブリッドストレージデバイスを具え、当該ハイブリッドストレージデバイスの動作に特有のコードで、１又はそれ以上の要求デバイスをロードする前に、前記１又はそれ以上の要求デバイスに取り付けられた標準デバイスをエミュレートすることによって、前記ハイブリッドストレージデバイスにアクセスを提供する標準デバイスエミュレーションモジュールを更に具え、前記標準デバイスが工業規格ＢＩＯＳによって担持されることを特徴とする装置。
【請求項１０】
請求項１に記載の装置において、前記ソリッドステートストレージデバイスが２又はそれ以上の領域に分割可能であり、１又はそれ以上のパーティションが、前記ＨＣＮＶストレージデバイス用のキャッシュとして機能する前記ソリッドステートストレージと別個の、ソリッドステートストレージとして用いられ得ることを特徴とする装置。
【請求項１１】
請求項１に記載の装置において、１又はそれ以上のクライアントが、前記ソリッドステートストレージデバイス及び前記１又はそれ以上のＨＣＮＶストレージデバイス内に記憶されるファイル又はオブジェクトのうちの１又はそれ以上の状態を管理するために、前記キャッシュフロントエンドモジュール及び前記キャッシュバックエンドモジュールにキャッシュ制御メッセージを送信することを特徴とする装置。
【請求項１２】
請求項１１に記載の装置において、前記キャッシュ制御メッセージが、
前記キャッシュバックエンドモジュールに、前記ソリッドステートストレージ内のオブジェクト又はファイルの一部を留める（ｐｉｎ）制御メッセージ；
前記キャッシュバックエンドモジュールに、前記ソリッドステートストレージ内のオブジェクト又はファイルの一部を解放する（ｕｎｐｉｎ）制御メッセージ；
前記キャッシュバックエンドモジュールに、前記ソリッドステートストレージから前記１又はそれ以上のＨＣＶＮストレージデバイスまでのオブジェクト又はファイルの一部を消去させる制御メッセージ；
前記キャッシュバックエンドモジュールに、前記１又はそれ以上のＨＣＶＮストレージデバイスから前記ソリッドステートストレージに、オブジェクト又はファイルの一部を事前ロードさせる制御メッセージ；及び
前記ソリッドステートストレージに、決められたストレージ空き領域量を空けるために、前記キャッシュバックエンドモジュールに、前記ソリッドステートストレージから前記１又はそれ以上のＨＣＶＮストレージデバイスまでの１又はそれ以上のオブジェクト又はファイルの１又はそれ以上の部分をオフロードさせる制御メッセージ；
のうちの１又はそれ以上を含むことを特徴とする装置。
【請求項１３】
請求項１１に記載の装置において、前記キャッシュ制御メッセージが前記オブジェクト又はファイル用のメタデータ（「キャッシュ制御メタデータ」）を通して通信されることを特徴とする装置。
【請求項１４】
請求項１３に記載の装置において、前記キャッシュ制御メタデータが永続的であること

を特徴とする装置。
【請求項１５】
請求項１３に記載の装置において、前記キャッシュ制御メタデータが前記ファイル又はオブジェクトの生成時に設定される属性を通して構築されることを特徴とする装置。
【請求項１６】
請求項１３に記載の装置において、前記キャッシュ制御メタデータがファイル又はオブジェクト管理システムから得られることを特徴とする装置。
【請求項１７】
請求項１に記載の装置が、揮発性キャッシュストレージエレメントを更に具え、前記キャッシュフロントエンドモジュール及び前記キャッシュバックエンドモジュールが、前記揮発性キャッシュストレージエレメントにデータを記憶するステップを更に具え、前記ソリッドステートストレージ及び揮発性キャッシュストレージエレメントに記憶されるデータを管理し、前記バックエンドストレージモジュールが、前記揮発性キャッシュストレージエレメントと、前記ソリッドステートストレージと、前記ＨＣＶＮストレージデバイスとの間のデータ転送を更に管理することを特徴とする装置。
【請求項１８】
請求項１７に記載の装置において、前記ＨＣＶＮストレージデバイスに記憶されるオブジェクト及びファイル用のメタデータ及びインデックスメタデータのうちの１又はそれ以上が、前記ソリッドステートストレージデバイス及び前記揮発性キャッシュストレージエレメント内に維持されることを特徴とする装置。
【請求項１９】
請求項１に記載の装置において、前記ＨＣＶＮストレージデバイスに記憶されるオブジェクト及びファイル用のメタデータ及びインデックスメタデータのうちの１又はそれ以上が、前記ソリッドステートストレージデバイス内に維持されることを特徴とする装置。
【請求項２０】
請求項１に記載の装置において、前記ソリッドステートストレージ及び前記１又はそれ以上のＨＣＮＶストレージデバイスが、ストレージデバイスを具えて、前記ＨＣＮＶストレージデバイスが前記ストレージデバイスに接続されたクライアントの表示から隠され、前記要求デバイスに単一のストレージデバイスとして見えるようにすることを特徴とする装置。
【請求項２１】
１又はそれ以上の高容量不揮発性（「ＨＣＮＶ」）ストレージデバイスでデータのストレージを管理するためのシステムであって，当該システムが、
不揮発性のソリッドステートデータストレージエレメントのアレイを具えるソリッドステートストレージと；
１又はそれ以上のＨＣＮＶストレージデバイスと；
ソリッドステートストレージコントローラと；
ＨＣＮＶストレージデバイスコントローラと；
ストレージ要求に関連するデータ転送を管理し、要求デバイスと前記ソリッドステートストレージとの間の前記データ転送が、前記１又はそれ以上のＨＣＮＶストレージデバイス用のキャッシュとして機能し、前記データ転送がデータ、メタデータ及びメタデータインデックスのうちの１又はそれ以上を含むキャッシュフロントエンドモジュールと；
前記ソリッドステートストレージと、前記１又はそれ以上のＨＣＮＶストレージデバイスとの間のデータ転送を管理するキャッシュバックエンドモジュールと；
を具えるストレージコントローラと；
を具え、前記ソリッドステートコントローラが、１又はそれ以上の要求デバイスからオブジェクト要求を提供し、前記ソリッドステートストレージ内の前記オブジェクト要求のオブジェクトを管理するオブジェクトストレージコントローラモジュールを更に具えることを特徴とするシステム。

【請求項２２】
請求項２１に記載のシステムが、前記ストレージコントローラに接続されたネットワークインタフェースを更に具え、当該ネットワークインタフェースがコンピュータネットワークを通して前記要求デバイスと前記ソリッドステートストレージコントローラとの間のデータ転送を促進することを特徴とするシステム。
【請求項２３】
請求項２１に記載のシステムがサーバを更に具え、当該サーバが前記ソリッドステートストレージと、前記１又はそれ以上のＨＣＮＶストレージデバイスと、前記ストレージコントローラとを具えることを特徴とするシステム。
【請求項２４】
請求項２１に記載のシステムにおいて、前記１又はそれ以上のＨＣＮＶストレージデバイスがストレージエリアネットワーク（「ＳＡＮ」）を通して前記ストレージコントローラに接続されることを特徴とするシステム。
【請求項２５】
コンピュータプログラムプロダクトであって、１又はそれ以上の高容量不揮発性（「ＨＣＮＶ」）ストレージデバイスでデータのストレージを管理するための動作を行うのに実行可能な、コンピュータが利用可能なプログラムコードを有するコンピュータ可読媒体を具え、前記コンピュータプログラムプロダクトの動作が、
キャッシュフロントエンドモジュールを用いて、ストレージ要求に関連するデータ転送を管理するステップであって、要求デバイスとソリッドステートストレージとの間の前記データ転送が、１又はそれ以上のＨＣＮＶストレージデバイス用のキャッシュとして機能し、前記データ転送がデータ、メタデータ及びメタデータインデックスのうちの１又はそれ以上を含み、前記ソリッドステートストレージが不揮発性のソリッドステートデータストレージエレメントのアレイを具えるステップと；
キャッシュバックエンドモジュールを用いて、前記ソリッドステートストレージと、前記１又はそれ以上のＨＣＮＶストレージデバイスとの間のデータ転送を管理するステップと；
を含み、ソリッドステートコントローラが、１又はそれ以上の要求デバイスからオブジェクト要求を提供し、前記ソリッドステートストレージ内の前記オブジェクト要求のオブジェクトを管理するオブジェクトストレージコントローラモジュールを更に具えることを特徴とするコンピュータプログラムプロダクト。
【請求項２６】
請求項２０に記載の装置において、前記単一のストレージデバイスが少なくともブロックストレージデバイス、オブジェクトストレージデバイス、及びファイルストレージデバイスのうちのいずれかとして前記要求デバイスに見えるように構成可能であることを特徴とする装置。

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2007/025049

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F12/08 G06F3/06

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 02/01365 A (INTEL CORP [US]; COULSON RICHARD [US]) 3 January 2002 (2002-01-03) abstract page 3, line 1 - page 4, line 23 page 7, line 3 - page 8, line 26 page 10, line 28 - page 11, line 2 page 14, line 13 - line 15 page 14, line 22 - line 27 page 15, line 12 - line 24 page 16, line 2 figures 1,7,8 -/--	1-9, 11-18, 21-23,25

[X] Further documents are listed in the continuation of Box C. [X] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier document but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
17 April 2008	14/05/2008

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Knutsson, Frédéric

Form PCT/ISA/210 (second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2007/025049

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2005/177687 A1 (RAO RAGHAVENDRA J [US]) 11 August 2005 (2005-08-11) abstract paragraph [0008] paragraph [0016] paragraph [0019] - paragraph [0025] figures 1,2	1-8, 11-25
X	"Windows PC Accelerators"[Online] 5 December 2006 (2006-12-05), pages 1-16, XP002476842 Retrieved from the Internet: URL:http://download.microsoft.com/download /9/c/5/9c5b2167-8017-4bae-9fde-d599bac8184 a/perfaccel.doc> [retrieved on 2008-04-16] abstract page 3, line 1 - page 4, line 6 page 4, line 28 - page 5, line 12 page 5, line 20 - page 7, line 15 page 7, line 42 - line 46 page 8, line 38 - page 13, line 2 page 14, line 1 - line 9	1-4, 8-16,19, 21-23,25

Form PCT/ISA/210 (continuation of second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/US2007/025049

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 0201365 A	03-01-2002	AU 7514701 A	08-01-2002
		CN 1527973 A	08-09-2004
		DE 10196380 T0	16-10-2003
		GB 2379538 A	12-03-2003
		JP 3951918 B2	01-08-2007
		JP 2004506256 T	26-02-2004
US 2005177687 A1	11-08-2005	EP 1723528 A1	22-11-2006
		WO 2005078588 A1	25-08-2005

フロントページの続き

(81)指定国 AP(BW,GH,GM,KE,LS,MW,MZ,NA,SD,SL,SZ,TZ,UG,ZM,ZW),EA(AM,AZ,BY,KG,KZ,MD,RU,TJ,TM),EP(AT,BE,BG,CH,CY,CZ,DE,DK,EE,ES,FI,FR,GB,GR,HU,IE,IS,IT,LT,LU,LV,MC,MT,NL,PL,PT,RO,SE,SI,SK,TR),OA(BF,BJ,CF,CG,CI,CM,GA,GN,GQ,GW,ML,MR,NE,SN,TD,TG),AE,AG,AL,AM,AT,AU,AZ,BA,BB,BG,BH,BR,BW,BY,BZ,CA,CH,CN,CO,CR,CU,CZ,DE,DK,DM,DO,DZ,EC,EE,EG,ES,FI,GB,GD,GE,GH,GM,GT,HN,HR,HU,ID,IL,IN,IS,JP,KE,KG,KM,KN,KP,KR,KZ,LA,LC,LK,LR,LS,LT,LU,LY,MA,MD,ME,MG,MK,MN,MW,MX,MY,MZ,NA,NG,NI,NO,NZ,OM,PG,PH,PL,PT,RO,RS,RU,SC,SD,SE,SG,SK,SL,SM,SV,SY,TJ,TM,TN,TR,TT,TZ,UA,UG,US,UZ,VC,VN,ZA,ZM,ZW

（特許庁注：以下のものは登録商標）
１．Ｌｉｎｕｘ

(71)出願人 509157063
ザッペ，マイケル
アメリカ合衆国 コロラド州 ８００３３，ホイートリッジ，シムズストリート ４６１５
(71)出願人 309039325
フュージョン マルチシステムズ，インク．（ディービイエイ フュージョンーアイオー）
アメリカ合衆国 ユタ州 ８４１２１，ソルトレイクシティ，６番フロア，サウス ３０００ イースト ６３５０
(74)代理人 100096024
弁理士 柏原 三枝子
(74)代理人 100125520
弁理士 高橋 剛一
(74)代理人 100155310
弁理士 柴田 雅仁
(74)代理人 100156339
弁理士 米村 道子
(72)発明者 フリン，デイビッド
アメリカ合衆国 ユタ州 ８４０９３，サンディ，シェイディメドウドライブ ８８５６
(72)発明者 シュトラッサー，ジョン
アメリカ合衆国 ユタ州 ８４０７５，シラキュース，サウス ２３２３
(72)発明者 サッチャー，ジョナサン
アメリカ合衆国 ユタ州 ８４０４３，リーハイ，ノース ２０８０ ウエスト ２２５９
(72)発明者 ザッペ，マイケル
アメリカ合衆国 コロラド州 ８００３３，ホイートリッジ，シムズストリート ４６１５

Ｆターム(参考) 5B005 JJ11 KK02 KK15 MM12
5B065 BA05 CC08 CE12 CE26 CH01