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

特許第6903005号
(P6903005)

(45)発行日　令和3年7月14日(2021.7.14)　(24)登録日　令和3年6月24日(2021.6.24)

(51)Int.Cl.		F I		
G06F 11/10	(2006.01)	G06F	11/10	660
G06F 11/20	(2006.01)	G06F	11/10	680
G06F 3/08	(2006.01)	G06F	11/20	694
G06F 13/10	(2006.01)	G06F	3/08	H
G06F 3/06	(2006.01)	G06F	13/10	340A

請求項の数 13　(全 23 頁)　最終頁に続く

(21)出願番号　特願2017-516635(P2017-516635)
(86)(22)出願日　平成27年2月27日(2015.2.27)
(65)公表番号　特表2017-519320(P2017-519320A)
(43)公表日　平成29年7月13日(2017.7.13)
(86)国際出願番号　PCT/US2015/018169
(87)国際公開番号　WO2015/187218
(87)国際公開日　平成27年12月10日(2015.12.10)
審査請求日　平成30年2月23日(2018.2.23)
(31)優先権主張番号　14/296,151
(32)優先日　平成26年6月4日(2014.6.4)
(33)優先権主張国・地域又は機関
米国(US)
(31)優先権主張番号　14/491,552
(32)優先日　平成26年9月19日(2014.9.19)
(33)優先権主張国・地域又は機関
米国(US)

(73)特許権者　511175211
ピュア　ストレージ,　インコーポレイテッド
アメリカ合衆国　カリフォルニア　９４０４１－２０５５,　マウンテン　ビュー,　カストロ　ストリート　６５０,　スイート　４００
(74)代理人　100079108
弁理士　稲葉　良幸
(74)代理人　100109346
弁理士　大貫　敏史
(74)代理人　100117189
弁理士　江口　昭彦
(74)代理人　100134120
弁理士　内藤　和彦

最終頁に続く

(54)【発明の名称】ストレージクラスター

(57)【特許請求の範囲】
【請求項１】
単一シャーシ内の複数のストレージノードにおいて、
前記複数のストレージノードは、ストレージクラスターとして一緒に通信するように構成され、
前記複数のストレージノードの各々は、ユーザデータ記憶のための不揮発性ソリッドステートメモリを有し、及び
前記複数のストレージノードは、前記複数のストレージノードの２つが失われても、前記複数のストレージノードが、消去コードを使用してユーザデータを読み取る能力を維持するように、前記複数のストレージノード全体にわたりユーザデータ及びユーザデータに関連したメタデータを配布するように構成され、
前記単一シャーシは、前記複数のストレージノードを結合する配電及び内部通信バスを伴うエンクロージャであり、かつ、
前記単一シャーシは、前記複数のストレージノードを外部通信バスに結合し、
前記複数のストレージノードは、２つの消去コードスキームに従いユーザデータを配布するように構成され、及び前記２つの消去コードスキームは、前記複数のストレージノードに共存し、
前記複数のストレージノードの２つが失われた後に一方の前記消去コードスキームの前記消去コードを使用してユーザデータを回復し、かつ、前記回復されたユーザデータ及び前記２つの消去コードスキームとは異なる消去コードスキームの消去コードを、残りの前

記複数のストレージノードに書き込むように構成される、
複数のストレージノード。
【請求項２】
前記不揮発性ソリッドステートメモリは、フラッシュメモリのアレイを含む、請求項１に記載の複数のストレージノード。
【請求項３】
前記不揮発性ソリッドステートメモリは、
コントローラ、
前記コントローラに結合された揮発性メモリ、及び
前記揮発性メモリに結合されたエネルギー貯蔵器、
を備え、前記不揮発性ソリッドステートメモリは、停電が検出されると、前記揮発性メモリ内のデータを前記フラッシュメモリのアレイに転送する、請求項２に記載の複数のストレージノード。
【請求項４】
前記複数のストレージノードの各々は、質問を監視し及びそれに応答するように構成され、質問に対する応答がないことは、前記複数のストレージノードの１つが欠陥であることを示す、請求項１に記載の複数のストレージノード。
【請求項５】
前記単一シャーシは、複数のスロットを含み、それらの複数のスロットのうちの各スロットは、前記複数のストレージノードのうちの１つのストレージノードを収容するように構成される、請求項１に記載の複数のストレージノード。
【請求項６】
単一シャーシ内に複数のストレージノードを備え、
前記複数のストレージノードの各々は、ユーザデータ記憶のための不揮発性ソリッドステートメモリを有し、及び
前記複数のストレージノードは、前記複数のストレージノードの２つが欠陥である状態で、前記複数のストレージノードが、消去コードを経て、ユーザデータにアクセスできるように、前記複数のストレージノード全体にわたりユーザデータ及びユーザデータに関連したメタデータを配布するように構成され、
前記単一シャーシは、前記複数のストレージノードを包囲し、かつ、
前記単一シャーシは、前記複数のストレージノード間に通信を与える通信バス及び前記複数のストレージノードに電力を供給する配電システムを備え、
前記複数のストレージノードは、前記複数のストレージノードにわたり第１の消去コードスキームに従って書かれたデータを読み取るように構成され、
前記複数のストレージノードは、前記複数のストレージノードにわたり第２の消去コードスキームに従って書かれたデータを読み取るように構成され、前記第１の消去コードスキームに従って書かれたデータは、前記複数のストレージノードにおいて、前記第２の消去コードスキームに従って書かれたデータと共存し、
前記複数のストレージノードの２つが失われた後に前記第１の消去コードスキームの前記消去コードを使用してユーザデータを回復し、かつ、前記回復されたユーザデータ及び前記第１及び前記第２の消去コードスキームとは異なる消去コードスキームの消去コードを、残りの前記複数のストレージノードに書き込むように構成される、
ストレージクラスター。
【請求項７】
前記複数のストレージノードのうちの各ストレージノードは、フラッシュメモリアレイに結合されたプロセッサを有するプリント回路板を含む、請求項６に記載のストレージクラスター。
【請求項８】
前記単一シャーシは、複数のスロットを含み、それらの複数のスロットのうちの各スロットは、前記複数のストレージノードのうちの１つのストレージノードを収容するように

構成される、請求項６に記載のストレージクラスター。
【請求項９】
前記複数のストレージノードの各々は、ユーザデータを読み取る試みとは独立して、前記複数のストレージノードの１つの欠陥を決定するように構成される、請求項６に記載のストレージクラスター。
【請求項１０】
前記複数のストレージノードは、ユーザデータの異なる部分にアクセスするために異なる形式の消去コードを適用するように構成される、請求項６に記載のストレージクラスター。
【請求項１１】
不揮発性ソリッドステートメモリを有する複数のストレージノードのユーザデータにアクセスするための方法において、
消去コードを通して複数のストレージノード全体にわたってユーザデータを配布し、前記複数のストレージノードは、それらストレージノードをクラスターとして結合する単一シャーシ内に収容され、
前記複数のストレージノードのうちの２つが到達不能であることを決定し、及び
前記複数のストレージノードの残りから、消去コードを経て、ユーザデータにアクセスし、プロセッサが少なくとも１つの方法オペレーションを遂行する、ことを含み、
前記単一シャーシは、前記複数のストレージノード間に通信を与える通信相互接続を含み、
前記単一シャーシは、前記複数のストレージノードを結合する配電バスを含み、
第１の形式の消去コードを使用して前記複数のストレージノードの残りにわたりユーザデータを読み取り、及び
第２の形式の消去コードを使用して前記複数のストレージノードの残りにわたりユーザデータを書き込む、
ことを更に含み、
前記消去コードは、前記複数のストレージノードに共存する２つの異なる消去コードスキームを含み、
前記複数のストレージノードの２つが失われた後に一方の前記消去コードスキームの前記消去コードを使用してユーザデータを回復し、かつ、前記回復されたユーザデータ及び前記２つの消去コードスキームとは異なる消去コードスキームの消去コードを、残りの前記複数のストレージノードに書き込むこと含む、
方法。
【請求項１２】
前記複数のストレージノードのうちの２つが到達不能であることを決定するのは、心臓鼓動の欠如、質問に対する応答の欠如、又は時間切れの１つに基づく、請求項１１に記載の方法。
【請求項１３】
複数の消去コードスキームのどれをユーザデータの配布に適用するか、前記複数のストレージノードにわたって協働的に決定することを更に含む、請求項１１に記載の方法。
【発明の詳細な説明】
【背景技術】
【０００１】
スピンメディアとして全体的に知られている従来のハードディスクドライブ（ＨＤＤ）や書き込み可能なＣＤ（コンパクトディスク）や書き込み可能なＤＶＤ（デジタル多様性ディスク）ドライブ、及びテープドライブを増強するか又はそれに置き換えて大量のデータを記憶するために、現在、フラッシュのようなソリッドステートメモリがソリッドステートドライブ（ＳＳＤ）に使用されている。フラッシュ及び他のソリッドステートメモリは、スピンメディアとは異なる特性を有する。又、互換性の理由で多数のソリッドステートドライブがハードディスクドライブの規格に適合するように設計されるが、これは、フ

ラッシュ及び他のソリッドステートメモリの特徴の向上を与えたり又はその独特の観点の利点を取り入れたりすることを困難にしている。

【０００２】

このような状況の中で実施形態が生成される。

【発明の概要】

【０００３】

ある実施形態では、単一シャーシにおける複数のストレージノードが提供される。単一シャーシにおける複数のストレージノードは、ストレージクラスターとして一緒に通信するように構成される。複数のストレージノードの各々は、ユーザデータストレージとして不揮発性ソリッドステートメモリを含む。複数のストレージノードは、ユーザデータ及びユーザデータに関連したメタデータを複数のストレージノード全体にわたって配布して、複数のストレージノードのうちの２つが失われても、複数のストレージノードが消去コードを使用してユーザデータを読み取る能力を維持するように構成される。シャーシは、配電バス、高速通信バス、並びにそれら配電及び通信バスを使用する１つ以上のストレージノードを設置する能力を含む。更に、不揮発性ソリッドステートメモリを有する複数のストレージノードにおいてユーザデータにアクセスするための方法も提供される。

【０００４】

ここに述べる実施形態の他の観点及び効果は、それら実施形態の原理を一例として示す添付図面を参照した以下の詳細な説明から明らかとなるであろう。

【０００５】

ここに述べる実施形態及びその効果は、添付図面に関連した以下の説明を参照することにより最も良く理解されよう。これらの図面は、ここに述べる実施形態に対して、ここに述べる実施形態の精神及び範囲から逸脱せずに当業者によりなされる形態及び細部の変更を何ら限定するものではない。

【図面の簡単な説明】

【０００６】

【図１】ある実施形態により、複数のストレージノードと、ネットワーク取り付け型ストレージを形成するために各ストレージノードに結合された内部ストレージとを伴うストレージクラスターの斜視図である。

【図２】ある実施形態により、図１のストレージクラスターの１つ以上をストレージリソースとして使用できる企業用コンピューティングシステムのシステム図である。

【図３】ある実施形態により、図１のストレージクラスターに使用するのに適した複数のストレージノード及び異なる能力を伴う不揮発性ソリッドステートストレージを示す図である。

【図４】ある実施形態により、複数のストレージノードを結合する相互接続スイッチを示すブロック図である。

【図５】ある実施形態により、ストレージノードのコンテンツと、不揮発性ソリッドステートストレージユニットの１つのコンテンツとを示す多レベルブロック図である。

【図６】ある実施形態により、ストレージクラスター、ストレージノード及び／又は不揮発性ソリッドステートストレージの実施形態において又は実施形態により実施できるストレージクラスターの動作方法のフローチャートである。

【図７】ここに述べる実施形態を具現化する規範的なコンピューティング装置を示す図である。

【発明を実施するための形態】

【０００７】

以下の実施形態は、１つ以上のユーザ又はクライアントシステム或いは他の外部ソースから発生されるユーザデータのようなユーザデータを記憶するストレージクラスターについて述べる。このストレージクラスターは、消去コードと、メタデータの冗長コピーとを使用して、シャーシ内に収容されたストレージノードにわたりユーザデータを配布する。消去コードとは、ディスク、ストレージノード又は地理的位置のような１組の異なる位置

にわたりデータが記憶されるデータ保護又は再構成の方法を指す。フラッシュメモリは、実施形態と一体化される一形式のソリッドステートメモリであるが、実施形態は、他の形式のソリッドステートメモリ、又は非ソリッドステートメモリを含む他のストレージ媒体へ拡張することができる。ストレージ位置及びワークロードの制御は、クラスター化されたピア・ツー・ピアシステムにおけるストレージ位置にわたって分散される。種々のストレージノード間の通信を仲裁し、ストレージノードが利用できなくなったときを検出し、そして種々のストレージノードにわたってＩ／Ｏ（入力及び出力）のバランスを取るようなタスクは、全て、分散ベースで取り扱われる。データは、ある実施形態では、データの回復をサポートするデータフラグメント又はストライプで複数のストレージノードにわたって広げられ又は配布される。入力及び出力パターンとは独立してクラスター内でデータの所有権を再指定することができる。以下に詳細に述べるこのアーキテクチャーは、システムが動作したまま、クラスター内のストレージノードがフェイルするのを許す。というのは、データは、他のストレージノードから再構成することができ、従って、入力及び出力動作に利用可能に保たれるからである。種々の実施形態では、ストレージノードは、クラスターノード、ブレード又はサーバーとも称される。

【０００８】

ストレージクラスターは、シャーシ、即ち１つ以上のストレージノードを収容するエンクロージャ内に収容される。各ストレージノードに電力を供給するメカニズム、例えば、配電バス、及び通信メカニズム、例えば、ストレージノード間の通信を可能にする通信バスは、シャーシ内に含まれる。ストレージクラスターは、ある実施形態によれば、１つの位置において独立システムとして動作することができる。１つの実施形態では、シャーシは、独立してイネーブル又はディスエイブルされる配電及び通信バスの両方の、少なくとも２つのインスタンスを収容する。内部通信バスは、イーサネット（登録商標）表バスであるが、ペリフェラル・コンポーネント・インターコネクト（ＰＣＩ）エクスプレス、インフィニバンド、等の他の技術も、等しく適当である。シャーシは、複数のシャーシ間の直接的通信又はスイッチを通しての通信、及びクライアントシステムとの通信を可能にするために外部通信バスのポートを形成する。外部通信は、イーサネット、インフィニバンド、ファイバーチャンネル、等の技術を使用する。ある実施形態では、外部通信バスは、シャーシ間及びクライアント通信に対して異なる通信バス技術を使用する。シャーシ内又はシャーシ間にスイッチが配備されている場合には、そのスイッチは、複数のプロトコル又は技術の間を変換するものとして働く。複数のシャーシがストレージクラスターを画成するように接続されるとき、ストレージクラスターは、独占的インターフェイス又は標準的インターフェイス、例えば、ネットワークファイルシステム（ＮＦＳ）、共通インターネットファイルシステム（ＣＩＦＳ）、小型コンピュータシステムインターフェイス（ＳＣＳＩ）又はハイパーテキスト転送プロトコル（ＨＴＴＰ）を使用して、クライアントによりアクセスされる。クライアントプロトコルからの変換は、スイッチ、シャーシ外部通信バス、又は各ストレージノード内で行われる。

【０００９】

各ストレージノードは、１つ以上のストレージサーバーであり、そして各ストレージサーバーは、１つ以上の不揮発性ソリッドステートメモリユニットに接続され、これは、ストレージユニットとも称される。ある実施形態は、各ストレージノード内及び１から８の不揮発性ソリッドステートメモリユニット間に単一のストレージサーバー含むが、この一例は、これに限定されるものではない。ストレージサーバーは、内部通信バスのためのプロセッサ、ダイナミックランダムアクセスメモリ（ＤＲＡＭ）及びインターフェイスと、各電源バスのための配電手段とを含む。ストレージノードの内部では、インターフェイス及びストレージユニットが通信バスを共有し、例えば、ある実施形態では、ＰＣＩエクスプレスを共有する。不揮発性ソリッドステートメモリユニットは、ストレージノード通信バスを通して内部通信バスインターフェイスに直接的にアクセスするか、又はバスインターフェイスにアクセスすることをストレージノードに要求する。不揮発性ソリッドステートメモリユニットは、埋め込み型中央処理ユニット（ＣＰＵ）、ソリッドステートストレ

ージコントローラ、及びある実施形態では、例えば、２－３２テラバイト（ＴＢ）の大型のソリッドステート大量ストレージを収容する。不揮発性ソリッドステートメモリユニットには、ＤＲＡＭのような埋め込み型揮発性ストレージ媒体、及びエネルギー貯蔵器が含まれる。ある実施形態では、エネルギー貯蔵器は、停電の場合にＤＲＡＭコンテンツのサブセットを安定なストレージ媒体に転送できるようにするキャパシタ、スーパーキャパシタ又はバッテリである。ある実施形態では、不揮発性ソリッドステートメモリユニットは、ストレージクラスメモリで構成され、例えば、ＤＲＡＭに取って代わり且つ省電力維持装置を可能にする相変化又は磁気抵抗性ランダムアクセスメモリ（ＭＲＡＭ）で構成される。

【００１０】

ストレージノード及び不揮発性ソリッドステートストレージの多数の特徴の１つは、ストレージクラスターにおいてデータを先見的に再構築する能力である。ストレージノード及び不揮発性ソリッドステートストレージは、そのストレージノード又は不揮発性ソリッドステートストレージに関するデータを読み取る試みがあるかどうかとは独立して、ストレージクラスターのストレージノード又は不揮発性ソリッドステートストレージに到達できないときを決定することができる。ストレージノード及び不揮発性ソリッドステートストレージは、次いで、少なくとも部分的に新しい位置でデータを回復し且つ再構築するように協働する。これは、ストレージクラスターを使用するクライアントシステムから開始される読み取りアクセスのためにデータが必要になるまで待機することなくシステムがデータを再構築するという点で先見的再構築を構成する。ストレージメモリ及びその動作のこれら及び更なる詳細を以下に述べる。

【００１１】

図１は、ある実施形態により、ネットワーク取り付け型ストレージ又はストレージエリアネットワークを形成するために、複数のストレージノード１５０、及び各ストレージノードに結合された内部ソリッドステートメモリを伴うストレージクラスター１６０の斜視図である。ネットワーク取り付け型ストレージ、ストレージエリアネットワーク、ストレージクラスター、又は他のストレージメモリは、物理的コンポーネント及びそれにより与えられる多量のストレージメモリの両方の、柔軟で且つ再構成可能な構成において、１つ以上のストレージノード１５０を各々有する１つ以上のストレージクラスター１６０を含むことができる。ストレージクラスター１６０は、ラックに適合するように設計され、そしてストレージメモリについて述べるように、１つ以上のラックをセットアップしそしてポピュレートすることができる。ストレージクラスター１６０は、複数のスロット１４２をもつシャーシ１３８を有する。シャーシ１３８は、ハウジング、エンクロージャ又はラックユニットとも称される。ある実施形態では、シャーシ１３８は、１４個のスロット１４２を有するが、他の数のスロットも容易に考えられる。例えば、ある実施形態は、４つのスロット、８つのスロット、１６個のスロット、３２個のスロット、又は他の適当な数のスロットを有する。各スロット１４２は、ある実施形態では、１つのストレージノード１５０を収容することができる。シャーシ１３８は、ラックにシャーシ１３８をマウントするのに使用できるフラップ１４８を備えている。ファン１４４は、ストレージノード１５０及びそのコンポーネントを冷却するための空気循環を与えるが、他の冷却コンポーネントを使用することもでき、又は冷却コンポーネントをもたない実施形態も案出される。スイッチファブリック１４６は、シャーシ１３８内のストレージノード１５０を一緒に結合すると共に、メモリへの通信のためにネットワークにも結合する。図１に示す実施形態では、スイッチファブリック１４６及びファン１４４の左側のスロット１４２は、ストレージノード１５０によって占有されて示されており、一方、スイッチファブリック１４６及びファン１４４の右側のスロット１４２は、空であり、例示の目的でストレージノード１５０を挿入するのに使用できる。この構成は一例に過ぎず、更に別の種々の構成では、１つ以上のストレージノード１５０がスロット１４２を占有することができる。ストレージノードの配列は、ある実施形態では、順次又は隣接である必要はない。ストレージノード１５０は、ホットプラグ可能であり、これは、システムを停止又はパワーダウンするこ

となく、ストレージノード１５０をシャーシ１３８のスロット１４２に挿入したり、又はスロット１４２から取り外したりできることを意味する。ストレージノード１５０をスロット１４２に挿入したり取り外したりする際に、システムは、変化を認識しそして変化に適応するために自動的に再構成する。再構成とは、ある実施形態では、冗長性の回復及び／又はデータ又は負荷の再バランスを含む。

【００１２】

各ストレージノード１５０は、複数のコンポーネントを有する。ここに示す実施形態では、ストレージノード１５０は、ＣＰＵ１５６、即ちプロセッサによりポピュレートされたプリント回路板１５８、ＣＰＵ１５６に結合されたメモリ１５４、及びＣＰＵ１５６に結合された不揮発性ソリッドステートストレージ１５２を備えているが、更に別の実施形態では、他の取り付け物及び／又はコンポーネントを使用することができる。メモリ１５４は、ＣＰＵ１５６により実行されるインストラクション及び／又はＣＰＵ１５６により操作されるデータを有する。以下に更に説明するように、不揮発性ソリッドステートストレージ１５２は、フラッシュを含み、又は更に別の実施形態では、他の形式のソリッドステートメモリを含む。

【００１３】

図２は、図１のストレージノード、ストレージクラスター及び／又は不揮発性ソリッドステートストレージの１つ以上をストレージリソース１０８として使用できる企業用コンピューティングシステム１０２のシステム図である。例えば、図２のフラッシュストレージ１２８は、ある実施形態では、図１のストレージノード、ストレージクラスター及び／又は不揮発性ソリッドステートストレージを一体化するものである。企業用コンピューティングシステム１０２は、処理リソース１０４と、ネットワークリソース１０６と、フラッシュストレージ１２８を含むストレージリソース１０８とを有する。フラッシュストレージ１２８にはフラッシュコントローラ１３０及びフラッシュメモリ１３２が含まれる。種々の実施形態において、フラッシュストレージ１２８は、１つ以上のストレージノード又はストレージクラスターを、ＣＰＵを含むフラッシュコントローラ１３０、及びストレージノードの不揮発性ソリッドステートストレージを含むフラッシュメモリ１３２と共に含むことができる。ある実施形態では、フラッシュメモリ１３２は、異なる形式のフラッシュメモリ、又は同じ形式のフラッシュメモリを含む。企業用コンピューティングシステム１０２は、フラッシュストレージ１２８の配備に適した環境を示しているが、フラッシュストレージ１２８は、より大きな又はより小さな他のコンピューティングシステム又は装置に、或いはより少数の又は付加的なリソースを伴う種々の企業用コンピューティングシステム１０２に使用することができる。企業用コンピューティングシステム１０２は、サービスを提供し又はサービスを利用するために、インターネットのようなネットワーク１４０に結合される。例えば、企業用コンピューティングシステム１０２は、クラウドサービス、物理的コンピューティングサービス、又はバーチャルコンピューティングサービスを提供することができる。

【００１４】

企業用コンピューティングシステム１０２では、種々のリソースが種々のコントローラにより配置され且つ管理される。処理コントローラ１１０は、プロセッサ１１６及びランダムアクセスメモリ（ＲＡＭ）１１８を含む処理リソース１０４を管理する。ネットワークコントローラ１１２は、ルータ１２０、スイッチ１２２及びサーバー１２４を含むネットワークリソース１０６を管理する。ストレージコントローラ１１４は、ハードドライブ１２６及びフラッシュストレージ１２８を含むストレージリソース１０８を管理する。この実施形態には、他の形式の処理リソース、ネットワークリソース、及びストレージリソースを含ませることができる。ある実施形態では、フラッシュストレージ１２８がハードドライブ１２６に完全に置き換わる。企業用コンピューティングシステム１０２は、種々のリソースを、物理的コンピューティングリソースとして、又はその変形例では、物理的コンピューティングリソースによりサポートされるバーチャルコンピューティングリソースとして、提供し又は割り当てることができる。例えば、種々のリソースは、ソフトウェ

アを実行する１つ以上のサーバーを使用して具現化することができる。ストレージリソース１０８には、ファイル又はデータオブジェクト或いは他の形態のデータが記憶される。

【００１５】

種々の実施形態において、企業用コンピューティングシステム１０２は、ストレージクラスターによってポピュレートされた複数のラックを備え、そしてそれらは、クラスター又はサーバーファームのような単一の物理的位置に配置される。他の実施形態では、複数のラックを、種々の都市、州又は国々のような複数の物理的位置に配置して、ネットワークで接続することができる。各ラック、各ストレージクラスター、各ストレージノード、及び各不揮発性ソリッドステートストレージは、各量のストレージスペースで個々に構成され、そのストレージスペースは、次いで、他とは独立して構成することができる。従って、不揮発性ソリッドステートストレージの各々においてストレージ容量を柔軟に追加し、アップグレードし、差し引きし、回復させ、及び／又は再構成することができる。上述したように、各ストレージノードは、ある実施形態では、１つ以上のサーバーを具現化することができる。

【００１６】

図３は、図１のシャーシに使用するのに適した複数のストレージノード１５０及び異なる能力を伴う不揮発性ソリッドステートストレージ１５２を示すブロック図である。各ストレージノード１５０は、不揮発性ソリッドステートストレージ１５２の１つ以上のユニットを有することができる。各不揮発性ソリッドステートストレージ１５２は、ある実施形態では、ストレージノード１５０又は他のストレージノード１５０における他の不揮発性ソリッドステートストレージ１５２とは異なる容量を含む。或いは又、ストレージノード又は複数のストレージノードにおける全ての不揮発性ソリッドステートストレージ１５２が、同じ容量を有してもよいし、或いは同じ及び／又は異なる容量を組み合わせて有してもよい。この融通性が図３に示されており、図３は、４、８及び３２のＴＢ容量の混合不揮発性ソリッドステートストレージ１５２を有するあるストレージノード１５０；各々３２ＴＢ容量の不揮発性ソリッドステートストレージ１５２を有する別のストレージノード１５０；及び各々８ＴＢ容量の不揮発性ソリッドステートストレージ１５２を有する更に別のストレージノード；の一例を示している。ここに述べる教示によれば、更に別の種々の組み合わせ及び容量が容易に案出される。クラスター化、例えば、ストレージをクラスター化してストレージクラスターを形成する状況では、ストレージノードが、不揮発性ソリッドステートストレージ１５２でもよいし又はそれを含んでもよい。以下に更に述べるように、不揮発性ソリッドステートストレージ１５２は、便利なクラスター化ポイントである。というのは、不揮発性ソリッドステートストレージ１５２は、不揮発性ランダムアクセスメモリ（ＮＶＲＡＭ）コンポーネントを含むからである。

【００１７】

図１及び３を参照すれば、ストレージクラスター１６０は、拡張可能であり、これは、上述したように、非均一なストレージサイズのストレージ容量が容易に追加されることを意味する。１つ以上のストレージノード１５０を各シャーシに差し込んだり取り外したりすることができ、そしてある実施形態では、ストレージクラスターが自己構成を行う。プラグインストレージノード１５０は、納入時にシャーシに設置されるか又は後で追加されるかに関わらず、異なるサイズとすることができる。例えば、ある実施形態では、ストレージノード１５０は、４ＴＢの倍数、例えば、８ＴＢ、１２ＴＢ、１６ＴＢ、３２ＴＢ、等である。更に別の実施形態では、ストレージノード１５０は、他のストレージ量又は容量の倍数である。各ストレージノード１５０のストレージ容量は、ブロードキャストされて、データをどのようにストライプ化するかの判断に影響を及ぼす。最大ストレージ効率のために、ある実施形態では、シャーシ内の不揮発性ソリッドステートストレージユニット１５２又はストレージノード１５０が１つ又は２つまで失われて動作を継続するという所定の要件を受けて、ストライプにおいてできるだけ広く自己構成することができる。

【００１８】

図４は、複数のストレージノード１５０を結合する通信相互接続部１７０及び配電バス

１７２を示すブロック図である。図１に戻ると、通信相互接続部１７０は、ある実施形態では、スイッチファブリック１４６で具現化されるか又はそれに含まれる。複数のストレージクラスター１６０がラックを占有する場合には、通信相互接続部１７０は、ある実施形態では、ラックスイッチの頂部で具現化されるか又はそれに含まれる。図４に示すように、ストレージクラスター１６０は、単一のシャーシ１３８内に包囲される。通信相互接続部１７０を通して外部ポート１７６がストレージノード１５０に結合され、一方、外部ポート１７４がストレージノードに直結される。外部電源ポート１７８が配電バス１７２に結合される。ストレージノード１５０は、図３を参照して述べたように、変化する量及び異なる容量の不揮発性ソリッドステートストレージ１５２を含む。更に、１つ以上のストレージノード１５０は、図４に示したように、計算のみのストレージノードでもよい。オーソリティ(authority)１６８は、不揮発性ソリッドステートストレージ１５２において、例えば、メモリに記憶されたリスト又は他のデータ構造体として具現化される。ある実施形態では、オーソリティは、不揮発性ソリッドステートストレージ１５２内に記憶され、そして不揮発性ソリッドステートストレージ１５２のコントローラ又は他のプロセッサで実行されるソフトウェアによりサポートされる。更に別の実施形態では、オーソリティ１６８は、ストレージノード１５０において、例えば、メモリ１５４に記憶されたリスト又は他のデータ構造体として具現化され、そしてストレージノード１５０のＣＰＵ１５６で実行されるソフトウェアによりサポートされる。オーソリティ１６８は、ある実施形態では、不揮発性ソリッドステートストレージ１５２のどこにどのようにデータを記憶するか制御する。この制御は、どの形式の消去コードスキームがデータに適用されるか及びどのストレージノード１５０がデータのどの部分を有するか決定する上で役立つ。各オーソリティ１６８は、不揮発性ソリッドステートストレージ１５２に指定される。各オーソリティは、種々の実施形態において、ファイルシステム、ストレージノード１５０又は不揮発性ソリッドステートストレージ１５２によりデータに指定されるｉｎｏｄｅ番号、セグメント番号、又は他のデータ識別子の範囲を制御する。

【００１９】

データの各断片及びメタデータの各断片は、ある実施形態では、システムにおいて冗長性を有する。更に、データの各断片及びメタデータの各断片は、オーソリティとも称されるオーナー(owner)を有する。そのオーソリティが、例えば、ストレージノードの欠陥により到達不能である場合には、そのデータ又はそのメタデータをどのように見つけるかについての継承のプランがある。種々の実施形態において、オーソリティ１６８の冗長コピーがある。オーソリティ１６８は、ある実施形態では、ストレージノード１５０及び不揮発性ソリッドステートストレージ１５２との関係を有している。ある範囲のデータセグメント番号又は他のデータ識別子をカバーする各オーソリティ１６８は、特定の不揮発性ソリッドステートストレージ１５２に指定される。ある実施形態では、そのような全ての範囲に対するオーソリティ１６８は、ストレージクラスターの不揮発性ソリッドステートストレージ１５２にわたって分散される。各ストレージノード１５０は、そのストレージノード１５０の不揮発性ソリッドステートストレージ１５２へのアクセスを与えるネットワークポートを有する。データは、セグメント番号に関連したセグメントに記憶され、そしてそのセグメント番号は、ある実施形態では、ＲＡＩＤ（独立ディスクの冗長アレイ）ストライプの構成に対する間接参照(indirection)である。従って、オーソリティ１６８の指定及び使用は、データの間接参照を確立する。間接参照は、ある実施形態によれば、このケースではオーソリティ１６８を経て、データを間接的に参照する能力とも称される。セグメントは、不揮発性ソリッドステートストレージ１５２のセットを識別し、そしてローカル識別子は、データを収容する不揮発性ソリッドステートストレージ１５２のセットに対するものである。ある実施形態では、ローカル識別子は、装置に対するオフセットであり、複数のセグメントにより順次に再使用される。他の実施形態では、ローカル識別子は、特定のセグメントに対して独特のものであり、決して再使用されない。不揮発性ソリッドステートストレージ１５２のオフセットは、不揮発性ソリッドステートストレージ１５２への書き込み又はそこからの読み取りのための位置データに適用される（ＲＡＩＤス

トライプの形態）。データは、不揮発性ソリッドステートストレージ１５２の複数のユニットにわたってストライプ化され、これは、特定のデータセグメントに対してオーソリティ１６８を有する不揮発性ソリッドステートストレージ１５２を含んでもよいし又はそれとは異なるものでもよい。

【００２０】

例えば、データの移動中又はデータの再構成中に、データの特定セグメントが位置する場所に変化がある場合には、そのオーソリティ１６８を有する不揮発性ソリッドステートストレージ１５２又はストレージノード１５０において、そのデータセグメントについてオーソリティ１６８と協議しなければならない。特定のデータ断片を探索するために、この実施形態では、データセグメントのハッシュ値が計算されるか、或いはｉｎｏｄｅ番号又はデータセグメント番号が適用される。この動作の出力は、その特定のデータ断片に対してオーソリティ１６８を有する不揮発性ソリッドステートストレージ１５２を指す。ある実施形態では、この動作に対して２つのステージがある。第１のステージは、エンティティ識別子（ＩＤ）、例えば、セグメント番号、ｉｎｏｄｅ番号、又はディレクトリ番号をオーソリティ識別子へマップする。このマッピングは、ハッシュ又はビットマスクのような計算を含む。第２のステージは、オーソリティ識別子を特定の不揮発性ソリッドステートストレージ１５２へマップすることであり、これは、明示的マッピングを通して行われる。この動作は繰り返すことができ、計算が行われたとき、計算結果が、オーソリティ１６８を有する特定の不揮発性ソリッドステートストレージ１５２を繰り返し且つ確実に指すようにする。この動作は、到達可能なストレージノードのセットを入力として含む。到達可能な不揮発性ソリッドステートストレージユニットのセットが変化する場合には、最適なセットも変化する。ある実施形態では、持続値は、現在指定値（常に真）であり、そして計算値は、クラスターが再構成を試みるところのターゲット指定値である。この計算は、到達可能で且つ同じクラスターを構成する不揮発性ソリッドステートストレージ１５２のセットの存在中にオーソリティに対して最適な不揮発性ソリッドステートストレージ１５２を決定するのに使用される。又、この計算は、指定の不揮発性ソリッドステートストレージが到達不能であってもオーソリティが決定されるようにオーソリティを不揮発性ソリッドステートストレージのマッピングに記録するピアな不揮発性ソリッドステートストレージ１５２の順序付けされたセットも決定する。ある実施形態では、特定のオーソリティ１６８が利用できない場合に、複写又は代用オーソリティ１６８と協議する。

【００２１】

図１から４を参照すれば、ストレージノード１５０におけるＣＰＵ１５６の多数のタスクのうちの２つは、書き込みデータを分解し及び読み取りデータを再組み立てすることである。データを書き込むべきであることをシステムが決定すると、そのデータに対してオーソリティ１６８が上述したように探索される。データのセグメントＩＤが既に決定されているときには、そのセグメントから決定されたオーソリティ１６８のホストであると現在決定された不揮発性ソリッドステートストレージ１５２へ書き込み要求が転送される。不揮発性ソリッドステートストレージ１５２及びそれに対応するオーソリティ１６８が存在するストレージノード１５０のホストＣＰＵ１５６は、次いで、データを分解又は破片化し、そしてデータを種々の不揮発性ソリッドステートストレージ１５２へ送出させる。その送出されたデータは、消去コードスキームに従ってデータストライプとして書き込まれる。ある実施形態では、データをプルすることが要求され、そして他の実施形態では、データがプッシュされる。逆に、データを読み取るときには、データを含むセグメントＩＤに対するオーソリティ１６８が上述したように探索される。不揮発性ソリッドステートストレージ１５２及びそれに対応するオーソリティ１６８が存在するストレージノード１５０のホストＣＰＵ１５６は、オーソリティにより指摘された不揮発性ソリッドステートストレージ及びそれに対応するストレージノードからデータを要求する。ある実施形態では、データは、フラッシュストレージからデータストライプとして読み取られる。ストレージノード１５０のホストＣＰＵ１５６は、次いで、読み取りデータを再組み立てし、エラー（もしあれば）を適当な消去コードスキームに基づいて修正し、そしてその再組み立

てされたデータをネットワークへ転送する。更に別の実施形態では、これらタスクの幾つか又は全部が不揮発性ソリッドステートストレージ１５２において取り扱われる。ある実施形態では、セグメントホストは、ストレージからページを要求し、次いで、最初に要求を発したストレージノードへデータを送信することにより、ストレージノード１５０へのデータの送信を要求する。

【００２２】

あるシステム、例えば、ＵＮＩＸスタイルのファイルシステムでは、データがインデックスノード又はｉｎｏｄｅで取り扱われ、これは、ファイルシステムにおいてオブジェクトを表すデータ構造を特定する。オブジェクトは、例えば、ファイル又はディレクトリである。メタデータは、オブジェクトを、他の属性の中でも、許可データ及び生成タイムスタンプのような属性として付随する。セグメント番号は、ファイルシステムにおけるそのようなオブジェクトの全部又は一部分に指定される。他のシステムでは、データセグメントが、どこかで指定されたセグメント番号で取り扱われる。説明上、配布の単位は、エンティティであり、そしてエンティティは、ファイル、ディレクトリ、又はセグメントである。即ち、エンティティは、ストレージシステムにより記憶されるデータ又はメタデータの単位である。エンティティは、オーソリティと称されるセットにグループ分けされる。各オーソリティは、オーソリティオーナーを有し、これは、オーソリティにおけるエンティティを更新する排他的権利を有するストレージノードである。換言すれば、ストレージノードは、オーソリティを含み、次いで、オーソリティは、エンティティを含む。

【００２３】

セグメントは、ある実施形態によれば、データの論理的コンテナである。又、セグメントは、媒体アドレススペースと物理的フラッシュ位置との間のアドレススペースであり、即ち、このアドレススペースにはデータセグメント番号がある。又、セグメントは、メタデータも含み、これは、高レベルソフトウェアに関与せずにデータ冗長性を回復（異なるフラッシュ位置又は装置への再書き込み）させることができる。ある実施形態では、セグメントの内部フォーマットは、クライアントデータと、そのデータの位置を決定するための媒体マッピングとを含む。各データセグメントは、そのセグメントを多数のデータ及びパリティシャード（該当する場合）へ分断することにより、例えば、メモリ及び他の欠陥から保護される。データ及びパリティシャードは、消去コードスキームによりホストＣＰＵ１５６（図５）に結合された不揮発性ソリッドステートストレージ１５２にわたって配布され、即ちストライプ化される。期間セグメントの使用は、ある実施形態では、セグメントのアドレススペースにおけるコンテナ及びその場所を指す。期間ストライプの使用は、セグメントと同じシャードセットを指し、ある実施形態によれば、シャードが冗長性又はパリティ情報と共にどのように配布されるかを含む。

【００２４】

一連のアドレススペース変換がストレージシステム全体にわたって行われる。最上部には、ディレクトリエンティティ（ファイル名）があって、ｉｎｏｄｅにリンクしている。ｉｎｏｄｅは、データが論理的に記憶された媒体アドレススペースを指す。媒体アドレスは、一連の間接的媒体を通してマップされて、大きなファイルの負荷を分散させるか、或いは複写除外又はスナップショットのようなデータサービスを具現化する。媒体アドレスは、一連の間接的媒体を通してマップされて、大きなファイルの負荷を分散させるか、或いは複写除外又はスナップショットのようなデータサービスを具現化する。セグメントアドレスは、次いで、物理的フラッシュ位置に変換される。物理的フラッシュ位置は、ある実施形態では、システムにおけるフラッシュの量により限定されたアドレス範囲を有する。媒体アドレス及びセグメントアドレスは、論理的コンテナであり、ある実施形態では、１２８ビット以上の識別子を使用して、実際上無限であるようにし、再使用の見込みは、システムの予想寿命より長いと計算される。論理的コンテナからのアドレスは、ある実施形態では、ハイアラーキー形態で割り当てられる。最初に、各不揮発性ソリッドステートストレージ１５２には、ある範囲のアドレススペースが指定される。この指定範囲内で、不揮発性ソリッドステートストレージ１５２は、他の不揮発性ソリッドステートストレー

ジ１５２と同期せずに、アドレスを割り当てることができる。

【００２５】

データ及びメタデータは、変化するワークロードパターン及びストレージ装置に対して最適化された基礎的なストレージレイアウトのセットにより記憶される。これらのレイアウトは、複数の冗長性スキーム、圧縮フォーマット及びインデックスアルゴリズムを合体する。これらのレイアウトの幾つかは、オーソリティ及びオーソリティマスターに関する情報を記憶し、一方、他のレイアウトは、ファイルメタデータ及びファイルデータを記憶する。冗長性スキームは、単一のストレージ装置（ＮＡＮＤフラッシュチップのような）内の崩壊ビットを許容するエラー修正コード、複数のストレージノードの欠陥を許容する消去コード、及びデータセンター又は領域欠陥を許容する複写スキームを含む。ある実施形態では、低密度のパリティチェック（ＬＤＰＣ）コードが単一のストレージユニット内で使用される。ある実施形態では、リード・ソロモンエンコーディングがストレージクラスター内で使用され、そしてミラーリングがストレージグリッド内で使用される。メタデータは、順序付けされたログ構造化インデックス（例えば、Log Structured Merge Tree）を使用して記憶され、そしてログ構造化レイアウトには大きなデータが記憶されない。

【００２６】

エンティティの複数のコピーにわたって一貫性を維持するため、ストレージノードは、計算を通して２つのことに暗示的に合意する。（１）エンティティを含むオーソリティ、及び（２）オーソリティを含むストレージノード。オーソリティへのエンティティの指定は、エンティティをオーソリティに擬似ランダムに指定するか、エンティティを、外部で発生されるキーに基づいて範囲に分割するか、又は単一エンティティを各オーソリティに配置することにより行うことができる。擬似ランダムスキームは、例えば、リニアハッシュ、及びハッシュのレプリケーション・アンダー・スケーラブル・ハッシュ（ＲＵＳＨ）ファミリーであり、これは、コントロールド・レプリケーション・アンダー・スケーラブル・ハッシュ（ＣＲＵＳＨ）を含む。ある実施形態では、擬似ランダム指定は、ノードのセットが変化し得るために、オーソリティをノードに指定することにしか利用されない。オーソリティのセットは変化せず、従って、これらの実施形態では、主観的機能が適用される。ある配置スキームは、ストレージノードにオーソリティを自動的に配置するが、他の配置スキームは、ストレージノードへのオーソリティの明示的マッピングに依存する。ある実施形態では、擬似ランダムスキームは、各オーソリティから候補オーソリティオーナーのセットへマップするのに使用される。ＣＲＵＳＨに関連した擬似ランダムデータ配布機能は、オーソリティをストレージノードに指定し、そしてオーソリティがどこに指定されているかのリストを生成する。各ストレージノードは、擬似ランダムデータ配布機能のコピーを有し、そしてオーソリティを配布し及び後で発見又は探索するために同じ計算に到着する。擬似ランダムスキームの各々は、ある実施形態では、同じターゲットノードで終らすためにストレージノードの到達可能なセットを入力として要求する。あるエンティティがオーソリティに配置されると、そのエンティティは、予想される欠陥が予期せぬデータロスを招くことがないように物理的な装置に記憶される。ある実施形態では、再バランスアルゴリズムが、全てのエンティティのコピーを、同じマシンセットにおいて且つ同じレイアウトでオーソリティ内に記憶するように試みる。

【００２７】

予想される欠陥は、例えば、装置の欠陥、マシン盗難、データセンターの火災、並びに地域の大災害、例えば、原子力事故や地質学的事象を含む。異なる欠陥は、異なるレベルの許容データロスを招く。ある実施形態では、ストレージノード盗難は、システムのセキュリティにも信頼性にも影響がないが、システムの構成によっては、地域の出来事がノーロスデータ(no loss of data)、数秒又は数分のロスト・アップデート、又は完全なデータロスを招くことがある。

【００２８】

これらの実施形態では、ストレージ冗長性のためのデータの配置は、データ一貫性のためのオーソリティの配置とは独立している。ある実施形態では、オーソリティを含むスト

レージノードは、持続性ストレージを含まない。むしろ、ストレージノードは、オーソリティを含まない不揮発性ソリッドステートストレージに接続される。ストレージノードと不揮発性ソリッドステートストレージユニットとの間の通信相互接続部は、複数の通信技術より成り、そして非均一な性能及び欠陥許容特性を有する。ある実施形態では、上述したように、不揮発性ソリッドステートストレージユニットは、ＰＣＩエクスプレスを経てストレージノードに接続され、ストレージノードは、イーサネットバックプレーンを使用して単一のシャーシ内で一緒に接続され、そしてシャーシは、ストレージクラスターを形成するように一緒に接続される。ストレージクラスターは、ある実施形態では、イーサネット又はファイバーチャンネルを使用してクライアントに接続される。複数のストレージクラスターがストレージグリッドへと構成される場合には、複数のストレージクラスターは、インターネット又は他の長距離ネットワークリンク、例えば、インターネットを横断しない「メトロスケール」リンク又はプライベートリンクを使用して、接続される。

【００２９】

オーソリティオーナーは、エンティティを変更し、エンティティをある不揮発性ソリッドステートストレージユニットから別の不揮発性ソリッドステートストレージユニットへ移行し、及びエンティティのコピーを追加及び除去するための排他的権利を有する。これは、基礎的データの冗長性の維持を許す。オーソリティオーナーが失敗するか、退役させられるか、又は過負荷となったときに、オーソリティは、新たなストレージノードへ移行される。過渡的な欠陥は、全ての非欠陥マシンが新たなオーソリティ位置に合意することを保証するには、些細なことではない。過渡的な欠陥のために生じる曖昧さは、コンセンサスプロトコル、例えば、Ｐａｘｏｓ、ホット・ウォームフェイルオーバースキームにより、又はリモートシステムアドミニストレータによる手動での介在を経て、又はローカルハードウェアアドミニストレータにより（例えば、クラスターから欠陥マシンを物理的に除去するか又は欠陥マシンのボタンを押すことにより）、自動的に解消される。ある実施形態では、コンセンサスプロトコルが使用され、そしてフェイルオーバーは自動である。あまりに短い期間内にあまりに多数の欠陥や複写事象が生じる場合、ある実施形態では、システムが自己保存モードに入り、そしてアドミニストレータの介在まで複写及びデータ移動アクティビティを停止する。

【００３０】

オーソリティはストレージノード間に転送されそしてオーソリティはそれらのオーソリティに更新エンティティを所有するので、システムは、ストレージノードと不揮発性ソリッドステートストレージユニットとの間にメッセージを転送する。持続メッセージに関しては、異なる目的のメッセージは、異なる形式のものである。メッセージの形式に基づいて、システムは、異なる順序及び耐久性保証を維持する。持続メッセージが処理されるときに、メッセージは、複数の耐久性及び非耐久性ストレージハードウェア技術で一時的に記憶される。ある実施形態では、メッセージは、ＲＡＭ、ＮＶＲＡＭ及びＮＡＮＤフラッシュ装置に記憶され、そして各ストレージ媒体を効率的に使用するために種々のプロトコルが使用される。レイテンシーに敏感なクライアントの要求は、複写ＮＶＲＡＭ、その後、ＮＡＮＤにおいて持続されるが、バックグランド再バランス動作は、ＮＡＮＤへ直接的に持続される。

【００３１】

持続メッセージは、複写されるまで持続的に記憶される。これは、システムが、欠陥及びコンポーネントの交換にも関わらず、クライアントの要求にサービスし続けられるようにする。多くのハードウェアコンポーネントは、システムアドミニストレータ、製造者、ハードウェア供給チェーン、及び進行中監視クオリティコントロールインフラストラクチャーに見える独特の識別子を含むが、インフラストラクチャーアドレスの最上部で実行されるアプリケーションは、アドレスをバーチャル化する。これらのバーチャル化されたアドレスは、コンポーネントの欠陥及び交換に関わらず、ストレージシステムの寿命にわたって変化しない。これは、クライアント要求処理の再構成又は中断を伴わずにストレージシステムの各コンポーネントを時間と共に交換できるようにする。

【００３２】
ある実施形態では、バーチャル化されたアドレスは、充分な冗長性で記憶される。連続的監視システムは、ハードウェア及びソフトウェア状態とハードウェア識別子を相関させる。これは、欠陥コンポーネント及び製造細部による欠陥の検出及び予想を許す。又、監視システムは、ある実施形態では、重要な経路からコンポーネントを除去することにより欠陥が生じるまで影響のある装置からのオーソリティ及びエンティティの先見的転送も可能にする。
【００３３】
図５は、ストレージノード１５０のコンテンツ及びストレージノード１５０の不揮発性ソリッドステートストレージ１５２のコンテンツを示す多レベルブロック図である。ある実施形態では、データは、ネットワークインターフェイスコントローラ（ＮＩＣ）２０２によりストレージノード１５０へ及びストレージノード１５０から通信される。上述したように、各ストレージノード１５０は、ＣＰＵ１５６及び１つ以上の不揮発性ソリッドステートストレージ１５２を有する。図５において１レベル下方に移動すると、各不揮発性ソリッドステートストレージ１５２は、比較的高速の不揮発性ソリッドステートメモリ、例えば、不揮発性ランダムアクセスメモリ（ＮＶＲＡＭ）２０４、及びフラッシュメモリ２０６を有する。ある実施形態では、ＮＶＲＡＭ２０４は、プログラム／消去サイクルを要求しないコンポーネント（ＤＲＡＭ、ＭＲＡＭ、ＰＣＭ）であり、且つメモリの読み取りより遥かに頻繁に書き込みをサポートできるメモリである。図５において別のレベル下方に移動すると、ＮＶＲＡＭ２０４は、ある実施形態では、エネルギー貯蔵器２１８によりバックアップされるダイナミックランダムアクセスメモリ（ＤＲＡＭ）２１６のような高速揮発性メモリとして具現化される。エネルギー貯蔵器２１８は、停電の際にフラッシュメモリ２０６にコンテンツを転送するに充分な長さでＤＲＡＭ２１６を通電状態に保持するに充分な電力を供給する。ある実施形態では、エネルギー貯蔵器２１８は、停電時に安定な記憶媒体にＤＲＡＭ２１６のコンテンツを転送できるに充分な適当なエネルギー供給を与えるキャパシタ、スーパーキャパシタ、バッテリ、又は他の装置である。フラッシュメモリ２０６は、複数のフラッシュダイ２２２として具現化され、これは、フラッシュダイ２２２のパッケージ又はフラッシュダイ２２２のアレイとも称される。フラッシュダイ２２２は、パッケージ当たり１つのダイ、パッケージ当たり複数のダイ（即ち、マルチチップパッケージ）、ハイブリッドパッケージ、プリント回路板又は他の基板上の裸ダイ、カプセル化ダイ、等の多数の仕方でパッケージできることが明らかである。ここに示す実施形態では、不揮発性ソリッドステートストレージ１５２は、コントローラ２１２又はプロセッサと、コントローラ２１２に結合された入力／出力（Ｉ／Ｏ）ポート２１０とを有する。Ｉ／Ｏポート２１０は、フラッシュストレージノード１５０のＣＰＵ１５６及び／又はネットワークインターフェイスコントローラ２０２に結合される。フラッシュ入力／出力（Ｉ／Ｏ）ポート２２０は、フラッシュダイ２２２に結合され、そしてダイレクトメモリアクセスユニット（ＤＭＡ）２１４は、コントローラ２１２、ＤＲＡＭ２１６及びフラッシュダイ２２２に結合される。ここに示す実施形態では、Ｉ／Ｏポート２１０、コントローラ２１２、ＤＭＡユニット２１４及びフラッシュＩ／Ｏポート２２０は、プログラマブルロジック装置（ＰＬＤ）２０８、例えば、フィールドプログラマブルゲートアレイ（ＦＰＧＡ）において具現化される。この実施形態では、各フラッシュダイ２２２は、１６ｋＢ（キロバイト）ページ２２４として編成されたページと、フラッシュダイ２２２へデータを書き込むか又はそこから読み取るところのレジスタ２２６とを有する。更に別の実施形態では、フラッシュダイ２２２内に示されたフラッシュメモリに代って、又はそれに加えて、他の形式のソリッドステートメモリが使用される。
【００３４】
図６は、ストレージクラスターを操作する方法のフローチャートである。この方法は、ここに述べるストレージクラスター及びストレージノードの種々の実施形態において又は種々の実施形態により実施することができる。この方法の種々のステップは、ストレージクラスターのプロセッサ又はストレージノードのプロセッサのようなプロセッサによって

遂行することができる。この方法は、その一部分又は全体を、ソフトウェア、ハードウェア、ファームウェア、又はその組み合わせで実施することができる。この方法は、ユーザデータを消去コードと共に配布するアクション６０２で開始される。例えば、ユーザデータは、１つ以上の消去コードスキームを使用してストレージクラスターのストレージノードにわたって配布することができる。ストレージクラスターのストレージノードには２つ（又はある実施形態ではそれ以上）の消去コードスキームが共存できる。ある実施形態では、各ストレージノードは、データを書き込むときに複数の消去コードスキームのどれを適用するか決定し、そしてデータを読み取るときにどの消去コードスキームを適用するか決定することができる。これらは、同じ消去コードスキームでもよいし又は異なる消去コードスキームでもよい。

【００３５】

この方法は、アクション６０４へ進み、ストレージクラスターのストレージノードがチェックされる。ある実施形態では、ストレージノードが心臓鼓動についてチェックされ、各ストレージノードは、心臓鼓動として働くメッセージを周期的に発生する。別の実施形態では、チェックは、ストレージノードに質問する形態であり、質問に対して応答がないことは、ストレージノードが欠陥であることを指示する。判断アクション６０６において、２つのストレージノードが到達不能であるかどうか決定される。例えば、ストレージノードの２つがもはや心臓鼓動を発生しないか、又はストレージノードの２つが質問に応答しないか、又はそれらの組み合わせであるか、或いは他の指示である場合には、他のストレージノードの１つは、ストレージノードの２つが到達不能であると決定することができる。このような状態でない場合には、フローがアクション６０２へ戻り、例えば、ユーザデータが到達するときにストレージのためにユーザデータをストレージノードに書き込む等の、ユーザデータの配布を継続する。ストレージノードの２つが到達不能であると決定された場合には、フローがアクション６０８へ続く。

【００３６】

判断アクション６０８において、消去コードを使用して残りのストレージノードにおいてユーザデータにアクセスする。ユーザデータは、ある実施形態では、ストレージクラスターの外部の１人以上のユーザ又はクライアントシステム或いは他のソースから発生するデータを指すことが明らかである。ある実施形態では、消去コードの形式は、二重冗長性を含み、このケースでは、２つの欠陥ストレージノードがある状態で、残りのストレージノードが読み取り可能なユーザデータを有する。消去コード形式は、コードワードのうちの２ビットのロスを許すエラー修正コードを含み、ストレージノードの２つが失われてもデータを回復できるようにストレージノードにわたってデータが配布される。判断アクション６１０では、データを再構築すべきかどうか決定される。データを再構築すべきでない場合には、フローがアクション６０２へ戻り、ユーザデータを消去コードと共に配布することを継続する。データを再構築すべき場合には、フローがアクション６１２へ分岐する。ある実施形態では、データを再構築する判断は、２つのストレージノードが到達不能になった後に行われるが、他の実施形態では、データを再構築する判断は、１つのストレージノードが到達不能になった後に行われてもよい。データを再構築する判断に考慮される種々のメカニズムは、エラー修正カウント、エラー修正レート、読み取り欠陥、書き込み欠陥、心臓鼓動のロス、質問に対する応答欠陥、等を含む。図６の方法に対する適当な変更は、これら及び更に別の実施形態について容易に理解されよう。

【００３７】

アクション６１２において、消去コードを使用してデータが回復される。これは、アクション６０８に関して上述した消去コードの例によるものである。より詳細には、データは、残りのストレージノードから、例えば、エラー修正コードを使用して回復されるか、又は残りのストレージノードから読み取られるか、の適当な方である。２つ以上の形式の消去コードがストレージノードに共存する場合には、２つ以上の形式の消去コードを使用してデータを回復することができる。判断アクション６１４において、データを並列に読み取らねばならないかどうか決定される。ある実施形態では、２つ以上のデータ経路があ

り（例えば、データの二重冗長性）、データは、２つの経路にわたって並列に読み取ることができる。データを並列に読み取るべきでない場合には、フローがアクション６１８へ分岐する。データを並列に読み取るべきである場合には、フローがアクション６１６へ分岐し、結果の競争となる。次いで、競争の勝者が、回復されたデータとして使用される。

【００３８】

アクション６１８において、再構築のため消去コードスキームが決定される。例えば、ある実施形態では、各ストレージノードは、ストレージユニットにわたってデータを書き込むときに２つ以上の消去コードスキームのどれを適用するか判断することができる。ある実施形態では、ストレージノードは、消去コードスキームを決定するように協働する。これは、どのストレージノードが特定データセグメントのための消去コードスキームについての役割を果たすか決定することにより、又はその役割を果たすためのストレージノードを指定することにより、行うことができる。ある実施形態では、証言、投票、又は判断ロジック、等の種々のメカニズムを使用して、このアクションを達成する。不揮発性ソリッドステートストレージは、証言者（ある実施形態では）又は投票者（ある実施形態では）として働き、オーソリティのあるコピーが欠陥となった場合に、不揮発性ソリッドステートストレージの残りの機能及びオーソリティの残りのコピーが欠陥オーソリティのコンテンツを決定できるようにする。アクション６２０において、回復されたデータが、消去コードと共に、残りのストレージノードにわたって書き込まれる。例えば、再構築のために決定される消去コードスキームは、データを回復する際に、即ちデータを読み取る際に適用される消去コードスキームとは異なる。より詳細には、２つのストレージノードが失われることは、ある消去コードスキームを残りのストレージノードにもはや適用できず、そしてストレージノードは、残りのストレージノードに適用できる消去コードスキームへ切り換えられることを意味する。

【００３９】

ここに述べる方法は、従来の汎用コンピュータシステムのようなデジタル処理システムで遂行されることが明らかである。或いは又、１つの機能のみを遂行するように設計又はプログラムされた特殊目的コンピュータが使用されてもよい。図７は、ここに述べる実施形態を具現化する規範的コンピューティング装置を示す図である。図７のコンピューティング装置は、ある実施形態により、ストレージノード又は不揮発性ソリッドステートストレージのための機能の実施形態を遂行するのに使用される。このコンピューティング装置は、中央処理ユニット（ＣＰＵ）７０１を備え、これは、バス７０５を通して、メモリ７０３及び大量ストレージ装置７０７に結合される。大量ストレージ装置７０７は、ある実施形態では、ローカル又はリモートであるディスクドライブのような持続性データストレージ装置を表わす。大量ストレージ装置７０７は、ある実施形態では、バックアップストレージを具現化するものである。メモリ７０３は、リードオンリメモリ、ランダムアクセスメモリ、等を含む。コンピューティング装置に存在するアプリケーションは、ある実施形態では、メモリ７０３又は大量ストレージ装置７０７のようなコンピュータ読み取り可能な媒体に記憶されるか又はそれを経てアクセスされる。又、アプリケーションは、コンピューティング装置のネットワークモデム又は他のネットワークインターフェイスを経てアクセスされる変調電子信号の形態でもよい。ＣＰＵ７０１は、ある実施形態では、汎用プロセッサ、特殊目的プロセッサ、又は特別にプログラムされたロジック装置で実施されることが明らかである。

【００４０】

ディスプレイ７１１は、バス７０５を経て、ＣＰＵ７０１、メモリ７０３及び大量ストレージ装置７０７と通信する。ディスプレイ７１１は、ここに述べるシステムに関連した視覚ツール又はレポートを表示するよう構成される。入力／出力装置７０９は、コマンド選択の情報をＣＰＵ７０１へ通信するためにバス５０５に結合される。外部装置への及び外部装置からのデータは、入力／出力装置７０９を経て通信されることが明らかである。ＣＰＵ７０１は、図１から６を参照して述べた機能を可能にするためにここに述べる機能を実行するよう定義される。この機能を実施するコードは、ある実施形態では、ＣＰＵ７

０１のようなプロセッサにより実行するためにメモリ７０３又は大量ストレージ装置７０７内に記憶される。コンピューティング装置のオペレーティングシステムは、ＭＳ－ＷＩＮＤＯＷＳ™、ＵＮＩＸ™、ＬＩＮＵＸ™、ｉＯＳ™、ＣｅｎｔＯＳ™、Ａｎｄｒｏｉｄ™、Ｒｅｄｈａｔ　ＬＩＮＵＸ™、ｚ／ＯＳ™、又は他の既知のオペレーティングシステムである。又、ここに述べる実施形態は、バーチャル型コンピューティングシステムと一体化できることが明らかである。

【００４１】

ここでは、詳細な説明のための実施形態が開示される。しかしながら、ここに開示される特定の機能的詳細は、実施形態を説明するための単なる代表例に過ぎない。しかしながら、それら実施形態は、多数の別の形態で実施されてもよく、ここに述べる実施形態のみに限定されると解釈してはならない。

【００４２】

第１、第２、等の語は、ここでは、種々のステップ又は計算を説明するために使用されるが、これらのステップ又は計算は、これらの語により限定されてはならないことを理解されたい。これらの語は、あるステップ又は計算を別のものと区別するために使用されるに過ぎない。例えば、この開示の範囲から逸脱せずに、第１の計算は、第２の計算と呼ばれてもよく、そして同様に、第２のステップは、第１のステップと呼ばれてもよい。ここで使用する語「及び／又は」及び記号「／」は、列挙される関連アイテムの１つ以上のいずれの及び全ての組み合わせを含む。

【００４３】

ここで使用する単数形“ａ”“an”及び“the”は、文脈がそうでないことを明らかに示さない限り、複数形も含むものとする。更に、語“comprises(備える)”“comprising(備えている)”“includes(含む)”及び／又は“including(含んでいる)”は、ここで使用するとき、述べた特徴、整数、ステップ、オペレーション、エレメント、及び／又はコンポーネントの存在を特定するが、１つ以上の他の特徴、整数、ステップ、オペレーション、エレメント、コンポーネント、及び／又はそのグループの存在又は追加を除外するものではない。それ故、ここで使用する用語は、特定の実施形態を説明するためのものに過ぎず、それに限定されるものではない。

【００４４】

又、別の具現化では、示された機能／行動は、図に示された順序から外れて行われてもよいことに注意されたい。例えば、順次に示された２つの図面は、関与する機能／行動に基づいて、実際には、実質的に同時に実行されてもよいし、又は時々、逆の順序で実行されてもよい。

【００４５】

前記実施形態に留意して、それらの実施形態は、コンピュータシステムに記憶されたデータに関与する種々のコンピュータ実施オペレーションを使用することを理解されたい。それらのオペレーションは、物理量の物理的操作を要求するものである。通常、必ずしもそうでないが、それらの量は、記憶、転送、合成、比較、さもなければ、操作することのできる電気的又は磁気的信号の形態をとる。更に、遂行される操作は、しばしば、発生、識別、決定、又は比較、等の語で呼ばれる。実施形態の一部分を形成するここに述べるオペレーションは、いずれも、有用なマシンオペレーションである。又、それらの実施形態は、それらのオペレーションを遂行するためのデバイス又は装置にも関連している。その装置は、要求された目的に対して特に構成されてもよいし、又は装置は、コンピュータに記憶されたコンピュータプログラムにより選択的にアクチベートされ又は構成される汎用コンピュータでもよい。特に、種々の汎用マシンは、ここでの教示に従って書かれたコンピュータプログラムで使用することもできるし、又は要求されたオペレーションを遂行するために特殊な装置を構成することも便利である。

【００４６】

モジュール、アプリケーション、レイヤ、エージェント、又は他の、方法で動作可能なエンティティは、ハードウェア、ファームウェア、又はプロセッサ実行ソフトウェア、或

いはその組み合わせとして具現化される。ソフトウェアベースの実施形態がここに開示される場合には、コントローラのような物理的マシンにおいてソフトウェアを実施できることが明らかであろう。例えば、コントローラは、第１モジュール及び第２モジュールを含むことができる。コントローラは、例えば、方法、アプリケーション、レイヤ又はエージェントの種々のアクションを遂行するように構成できる。

【００４７】

又、それらの実施形態は、非一時的コンピュータ読み取り可能な媒体上のコンピュータ読み取り可能なコードとして具現化することもできる。コンピュータ読み取り可能な媒体は、コンピュータシステムにより後で読み取られるデータを記憶できるデータストレージ装置である。コンピュータ読み取り可能な媒体は、例えば、ハードドライブ、ネットワーク取り付け型ストレージ（ＮＡＳ）、リードオンリメモリ、ランダムアクセスメモリ、ＣＤ－ＲＯＭ、ＣＤ－Ｒ、ＣＤ－ＲＷ、磁気テープ、並びに他の光学的及び非光学的データストレージ装置を含む。又、コンピュータ読み取り可能な媒体は、ネットワーク結合のコンピュータシステムにわたって分散されて、コンピュータ読み取り可能なコードが分散形態で記憶され且つ実行されるようにする。ここに述べる実施形態は、ハンドヘルド装置、タブレット、マイクロプロセッサシステム、マイクロプロセッサベースの又はプログラム可能な消費者向け電子装置、消費者向け電子装置、ミニコンピュータ、メインフレームコンピュータ、等を含む種々のコンピュータシステム構成で具現化される。又、これら実施形態は、ワイヤベースネットワーク又はワイヤレスネットワークを通してリンクされたリモート処理装置によってタスクが遂行される分散型コンピューティング環境において具現化することもできる。

【００４８】

方法の操作は、特定の順序で説明したが、説明した操作と操作との間に他の操作が遂行されてもよく、説明した操作は、それらが若干異なる時間に生じるように調整されてもよく、又は説明した操作は、処理に関連した種々の間隔で処理操作を行えるシステムにおいて分散されてもよいことを理解されたい。

【００４９】

種々の実施形態において、ここに述べる方法及びメカニズムの１つ以上の部分がクラウドコンピューティング環境の一部分を形成してもよい。そのような実施形態では、リソースが１つ以上の種々のモデルに従ってサービスとしてインターネットを経て与えられる。そのようなモデルは、インフラストラクチャーをサービス（ＩａａＳ）として、プラットホームをサービス（ＰａａＳ）として、及びソフトウェアをサービス（ＳａａＳ）として含む。ＩａａＳでは、コンピュータインフラストラクチャーがサービスとして配信される。そのようなケースでは、コンピューティング装置は、一般的に、サービスプロバイダーにより所有されそして操作される。ＰａａＳモデルでは、ソフトウェア解決策を開発するためにデベロッパーにより使用されるソフトウェアツール及びその基礎的な装置は、サービスとして提供され、そしてサービスプロバイダーによりホストされる。ＳａａＳは、典型的に、サービスプロバイダーのライセンスソフトウェアをオンデマンドのサービスとして含む。サービスプロバイダーは、ソフトウェアをホストするか、又はソフトウェアを所与の期間中顧客に配備する。前記モデルの多数の組み合わせが考えられ、そして意図される。

【００５０】

種々のユニット、回路又は他のコンポーネントは、１つ又は複数のタスクを遂行するように「構成される」として説明され又は請求される。この点に関して、句「構成される」は、ユニット／回路／コンポーネントがオペレーション中に１つ又は複数のタスクを遂行する構造体（例えば、回路）を含むことを示すことで構造体を暗示するのに使用される。従って、ユニット／回路／コンポーネントは、指定のユニット／回路／コンポーネントが現在動作していない（例えば、オンでない）ときでもタスクを遂行するように構成されると言うことができる。「構成される」言語と共に使用されるユニット／回路／コンポーネントは、ハードウェア、例えば、回路、オペレーションを実施するために実行可能なプロ

グラムインストラクションを記憶するメモリ、等を含む。ユニット／回路／コンポーネントが１つ以上のタスクを遂行するように「構成される」と表すことは、そのユニット／回路／コンポーネントに対して３５ Ｕ.Ｓ.Ｃ.１１２を引用しないことが明確に意図される。更に、「構成される」は、問題のタスク（１つ又は複数）を遂行できるやり方でソフトウェア及び／又はファームウェア（例えば、ＦＰＧＡ、又はソフトウェアを実行する汎用プロセッサ）を動作することで操作される包括的構造体（例えば、包括的回路）を含む。又、「構成される」は、１つ以上のタスクを実施又は遂行するための装置（例えば、集積回路）を製造するように製造プロセス（例えば、半導体製造ファシリティ）を適応させることも含む。

【００５１】

以上の記載は、説明上、特定の実施形態を参照して述べた。しかしながら、上述した議論は、余すところのないものではなく、又、本発明を、ここに開示する正確な形態に限定するものでもない。前記教示に鑑み多数の変更及び変形が考えられる。前記実施形態は、それら実施形態の原理及びその実際的な応用を最良に説明するために選択され且つ記載されたもので、従って、当業者であれば、それら実施形態及び種々の変更を、意図される特定の用途に適するように最良に利用することができるであろう。従って、ここに示す実施形態は、例示と考えられ、これに限定されるものではなく、そして本発明は、ここに示す詳細に限定されず、特許請求の範囲内及びその等効物の中で変更可能である。

【符号の説明】

【００５２】

１０２：企業用コンピューティングシステム
１０４：処理リソース
１０６：ネットワークリソース
１０８：ストレージリソース
１１０：処理コントローラ
１１２：ネットワークコントローラ
１１４：ストレージコントローラ
１１６：プロセッサ
１１８：ランダムアクセスメモリ（ＲＡＭ）
１２０：ルータ
１２２：スイッチ
１２４：サーバー
１２６：ハードドライブ
１２８：フラッシュストレージ
１３０：フラッシュコントローラ
１３２：フラッシュメモリ
１３８：シャーシ
１４０：ネットワーク
１４２：スロット
１４４：ファン
１４６：スイッチファブリック
１４８：フラップ
１５０：ストレージノード
１５２：不揮発性ソリッドステートストレージ
１５４：メモリ
１５６：ＣＰＵ
１５８：プリント回路板
１６０：ストレージクラスター
７０１：ＣＰＵ
７０３：メモリ

７０５：バス
７０７：大量ストレージ
７０９：入力／出力装置
７１１：ディスプレイ

【図１】

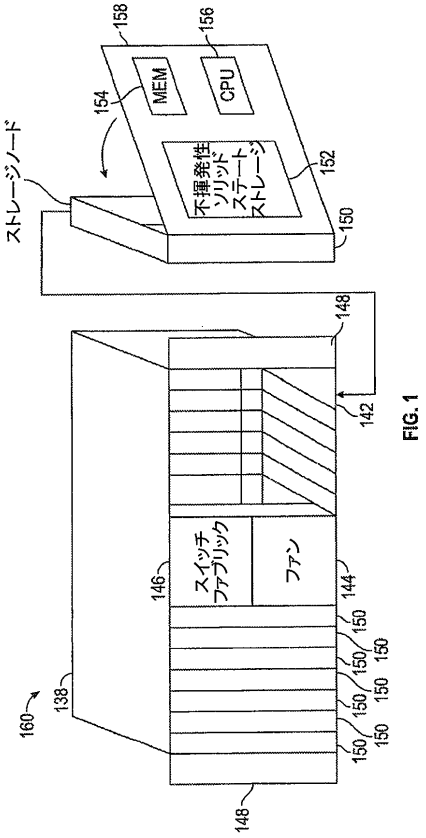

FIG. 1

【図２】

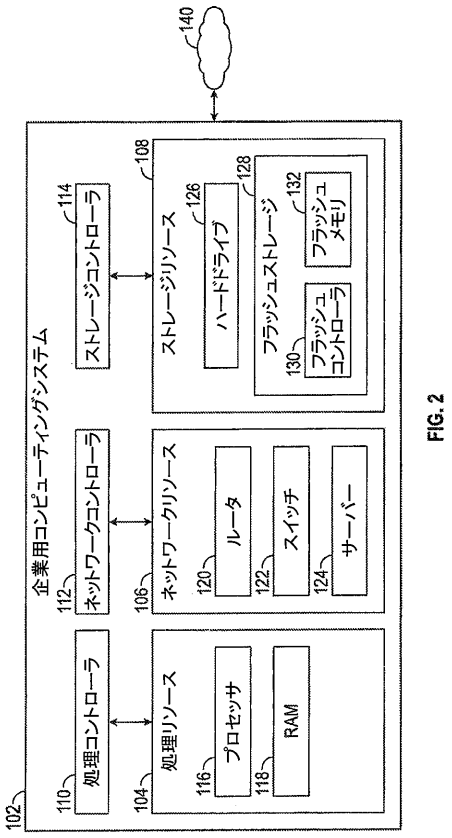

FIG. 2

【図３】

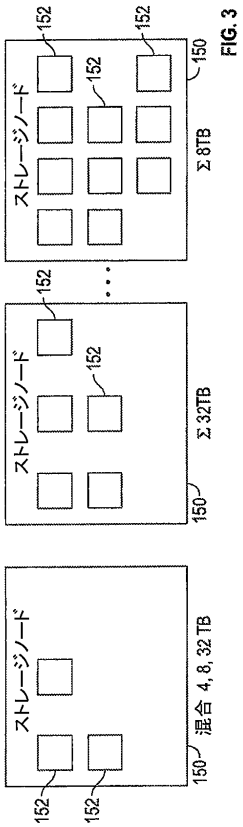

FIG. 3

【図４】

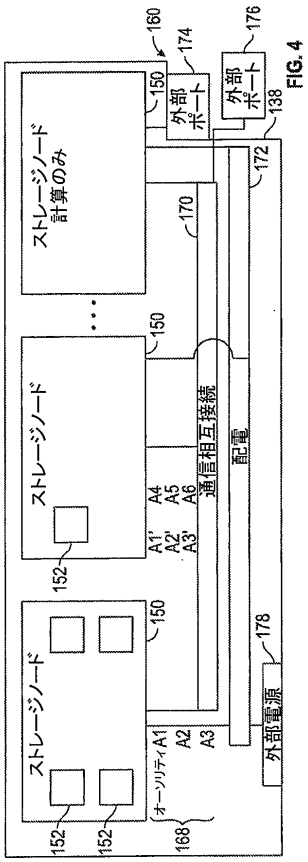

FIG. 4

【図５】

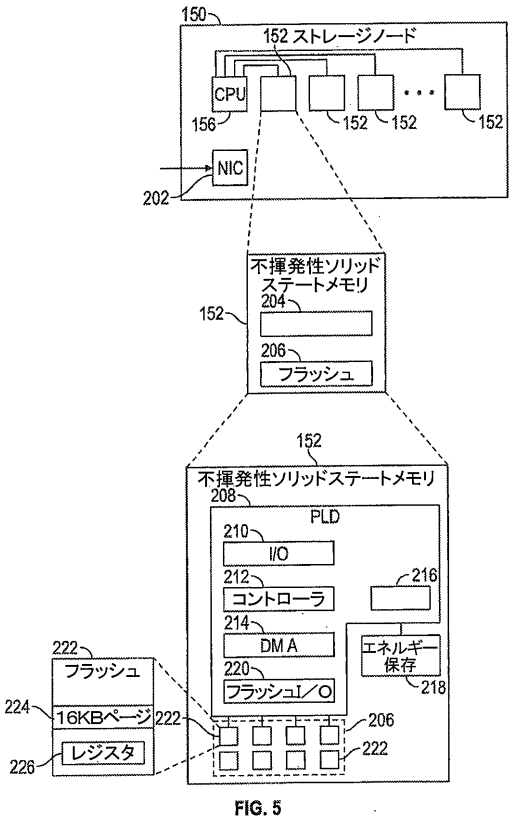

FIG. 5

【図６】

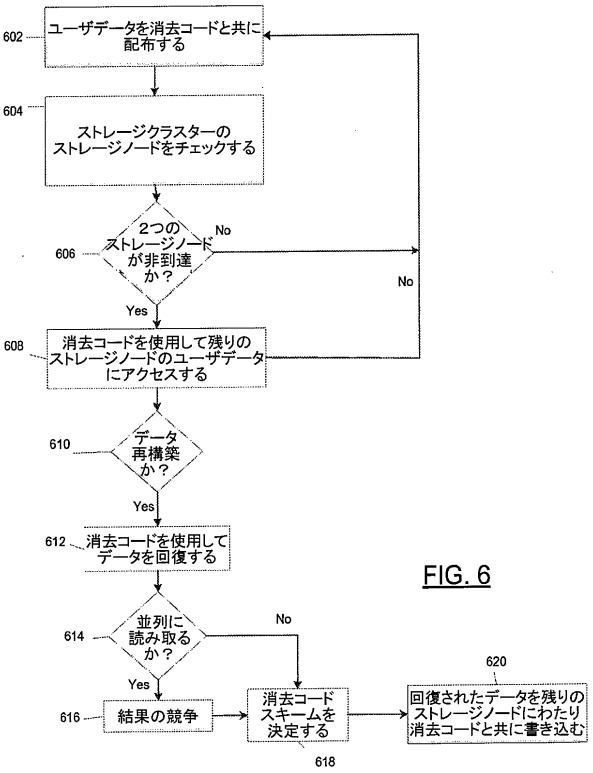

FIG. 6

【図７】

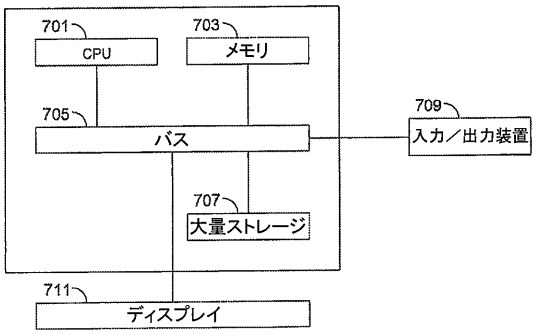

FIG. 7

フロントページの続き

(51)Int.Cl. FI
G06F 3/06 305C

(31)優先権主張番号 14/610,766
(32)優先日 平成27年1月30日(2015.1.30)
(33)優先権主張国・地域又は機関
米国(US)

前置審査

(74)代理人 100126480
弁理士 佐藤 睦
(74)代理人 100121843
弁理士 村井 賢郎
(72)発明者 ヘイズ ジョン
アメリカ合衆国 カリフォルニア州 ９４０４１ マウンテンビュー カストロ ストリート ６５０ スイート ４００
(72)発明者 コルグローヴ ジョン
アメリカ合衆国 カリフォルニア州 ９４０４１ マウンテンビュー カストロ ストリート ６５０ スイート ４００
(72)発明者 リー ロバート
アメリカ合衆国 カリフォルニア州 ９４０４１ マウンテンビュー カストロ ストリート ６５０ スイート ４００
(72)発明者 ヴァイゲル ピーター
アメリカ合衆国 カリフォルニア州 ９４０４１ マウンテンビュー カストロ ストリート ６５０ スイート ４００
(72)発明者 ボテス パール
アメリカ合衆国 カリフォルニア州 ９４０４１ マウンテンビュー カストロ ストリート ６５０ スイート ４００

審査官 漆原 孝治

(56)参考文献 特表２０１３－５４４３８６（ＪＰ，Ａ）
特開２００７－２６５３１４（ＪＰ，Ａ）
特開２００７－２４２０１８（ＪＰ，Ａ）
国際公開第２０１４／０２５８２１（WO，Ａ１）
特開２０１０－１２８８８６（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ １１／１０
Ｇ０６Ｆ ３／０６
Ｇ０６Ｆ ３／０８
Ｇ０６Ｆ １１／２０
Ｇ０６Ｆ １３／１０