

(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2019-517063
(P2019-517063A)

(43) 公表日 令和1年6月20日(2019.6.20)

(51) Int.Cl.		F I		テーマコード (参考)
G06F 3/06	(2006.01)	G06F 3/06	301A	
G06F 13/14	(2006.01)	G06F 13/14	330E	

審査請求 未請求 予備審査請求 未請求 (全 33 頁)

(21) 出願番号 特願2018-557857 (P2018-557857)
 (86) (22) 出願日 平成29年5月4日 (2017.5.4)
 (85) 翻訳文提出日 平成30年12月25日 (2018.12.25)
 (86) 国際出願番号 PCT/US2017/031162
 (87) 国際公開番号 W02017/192917
 (87) 国際公開日 平成29年11月9日 (2017.11.9)
 (31) 優先権主張番号 15/146,681
 (32) 優先日 平成28年5月4日 (2016.5.4)
 (33) 優先権主張国 米国 (US)

(71) 出願人 511175211
 ビュア 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 940
 41-2055, 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

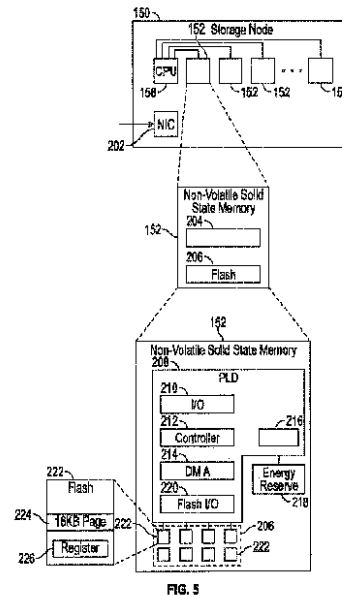
最終頁に続く

(54) 【発明の名称】 ストレージ・クラスタ

(57) 【要約】

ストレージ・システム内で処理能力を管理する方法が提供される。この方法は、複数のブレードを提供することであって、第1のサブセットのそれぞれは、ストレージ・ノードおよびストレージ・メモリを有し、第2の異なるサブセットのそれぞれは、コンピュータ専用ノードを有する、提供することを含む。この方法は、複数のブレードにまたがって、少なくとも1つのコンピュータ専用ノードを含む複数のノードにオーソリティを分配することであって、各オーソリティは、ある範囲のユーザ・データの所有権を有する、分配することを含む。

【選択図】 図5



【特許請求の範囲】**【請求項 1】**

複数のブレードを提供することであって、第 1 のサブセットのそれぞれは、ストレージ・ノードおよびストレージ・メモリを有し、第 2 の異なるサブセットのそれぞれは、コンピュータ専用ノードを有する、提供することと、

前記複数のブレードにまたがって、少なくとも 1 つのコンピュータ専用ノードを含む複数のノードにオーソリティを分配することであって、各オーソリティは、ある範囲のユーザ・データのオーナーシップを有する、分配することと

を含む、ストレージ・システム内で処理能力を管理する方法。

【請求項 2】

コンピュータ専用ノードを有する別のブレードを追加することと、

前記複数のブレードおよび前記別のブレードにまたがって前記オーソリティを再分配することと

をさらに含む、請求項 1 に記載の方法。

【請求項 3】

前記オーソリティの前記分配は、

前記ストレージ・システムへの前記複数のブレードのうちのさらなる 1 つの追加にตอบสนองして、前記複数のブレードのうちの 1 つまたは複数から前記複数のブレードのうちの前記さらなる 1 つへ 1 つまたは複数のオーソリティを移動すること

を含む、請求項 1 に記載の方法。

【請求項 4】

前記オーソリティの前記分配は、前記複数のブレードのそれぞれについて、前記オーソリティにまたがるコンピューティング・リソースの平衡化に従う、請求項 1 に記載の方法。

【請求項 5】

サービスの複数のクラスのそれぞれの入出力処理が個々のサービス・クラスに基づいて 1 つまたは複数のストレージ・ノードまたはコンピュータ・ノードに割り当てられるようにするために、前記複数のブレードにまたがって外部 I/O (入出力) 処理のコンピューティング・タスクを分配すること

をさらに含む、請求項 1 に記載の方法。

【請求項 6】

前記複数のブレードは、ランダム・アクセス・メモリ (RAM) の量、プロセッサ速度、またはプロセッサ・コアの個数を含む処理特性の第 1 のセットを有する第 1 のブレードと、RAM の量、プロセッサ速度、またはプロセッサ・コアの個数を含む処理特性の第 2 のセットを有する第 2 のブレードとを含み、

前記第 1 のブレードのオーソリティおよび前記第 2 のブレードのオーソリティから使用可能な前記処理特性を平衡化するために、前記第 1 のブレードより多数のオーソリティが、前記第 2 のブレードに分配される

請求項 1 に記載の方法。

【請求項 7】

アプリケーション・レイヤ内の 1 つまたは複数のアプリケーションのコンピューティング・タスクは、前記複数のブレードのうちの少なくとも 1 つにまたがって分配され、

前記オーソリティの前記分配は、前記複数のブレードのそれぞれで使用可能な処理能力の相対量に比例する

請求項 1 に記載の方法。

【請求項 8】

プロセッサによって実行される時に、前記プロセッサに、

複数のブレードを提供することであって、第 1 のサブセットのそれぞれは、ストレージ・ノードおよびストレージ・メモリを有し、第 2 の異なるサブセットのそれぞれは、コンピュータ専用ノードを有する、提供することと、

10

20

30

40

50

前記複数のブレードにまたがって、少なくとも1つのコンピュータ専用ノードを含む複数のノードにオーソリティを分配することであって、各オーソリティは、ある範囲のユーザ・データのオーナーシップを有する、分配することと

を含む方法を実行させる命令を有する有形の非一時的コンピュータ可読媒体。

【請求項9】

前記方法は、

コンピュータ専用ノードを有する別のブレードを追加することと、

前記複数のブレードおよび前記別のブレードにまたがって前記オーソリティを再分配することと

をさらに含む、請求項8に記載のコンピュータ可読媒体。

10

【請求項10】

前記オーソリティの前記分配は、

前記ストレージ・システムへの前記複数のブレードのうちのさらなる1つの追加に応答して、前記複数のブレードのうちの1つまたは複数から前記複数のブレードのうちの前記さらなる1つへ1つまたは複数のオーソリティを移動すること

を含む、請求項8に記載のコンピュータ可読媒体。

【請求項11】

前記オーソリティの前記分配は、前記複数のブレードのそれぞれについて、前記オーソリティにまたがるコンピューティング・リソースの平衡化に従う、請求項8に記載のコンピュータ可読媒体。

20

【請求項12】

前記方法は、

サービスの複数のクラスのそれぞれの入出力処理が個々のサービス・クラスに基づいて1つまたは複数のストレージ・ノードまたはコンピュータ・ノードに割り当てられるようにするために、前記複数のブレードにまたがって外部I/O（入出力）処理のコンピューティング・タスクを分配すること

をさらに含む、請求項8に記載のコンピュータ可読媒体。

【請求項13】

前記複数のブレードは、ランダム・アクセス・メモリ（RAM）の量、プロセッサ速度、またはプロセッサ・コアの個数を含む処理特性の第1のセットを有する第1のブレードと、RAMの量、プロセッサ速度、またはプロセッサ・コアの個数を含む処理特性の第2のセットを有する第2のブレードとを含み、

30

前記第1のブレードのオーソリティおよび前記第2のブレードのオーソリティから使用可能な前記処理特性を平衡化するために、前記第1のブレードより多数のオーソリティが、前記第2のブレードに分配される

請求項8に記載のコンピュータ可読媒体。

【請求項14】

ストレージ・システムであって、

複数のブレードであって、第1のサブセットのそれぞれは、ストレージ・ノードおよびストレージ・メモリを有し、第2の異なるサブセットのそれぞれは、コンピュータ専用ノードを有する、複数のブレード

40

を含み、前記複数のブレードは、前記ストレージ・システムを形成し、オーソリティは、前記複数のブレードにまたがって、少なくとも1つのコンピュータ専用ノードを含む複数のノードに分配され、各オーソリティは、ある範囲のユーザ・データのオーナーシップを有する

ストレージ・システム。

【請求項15】

コンピュータ専用ノードを有する別のブレードを追加し、

前記複数のブレードおよび前記別のブレードにまたがって前記オーソリティを再分配する

50

ように構成される、請求項 14 に記載のストレージ・システム。

【請求項 16】

前記複数のブレードは、前記ストレージ・システムへの前記複数のブレードのうちのさらなる 1 つの追加に応答して、前記複数のブレードのうちの 1 つまたは複数から前記複数のブレードのうちの前記さらなる 1 つへ 1 つまたは複数のオーソリティを移動することによって前記オーソリティを分配するように構成される、請求項 14 に記載のストレージ・システム。

【請求項 17】

前記オーソリティは、前記複数のブレードのそれぞれについて、前記オーソリティにまたがるコンピューティング・リソースの平衡化に従って分配される、請求項 14 に記載のストレージ・システム。

10

【請求項 18】

外部 I/O (入出力) 処理のコンピューティング・タスクは、サービスの複数のクラスのそれぞれの入出力処理が個々のサービス・クラスに基づいて 1 つまたは複数のストレージ・ノードまたはコンピュータ・ノードに割り当てられるようにするために、前記複数のブレードにまたがって分配される

請求項 14 に記載のストレージ・システム。

【請求項 19】

前記複数のブレードは、ランダム・アクセス・メモリ (RAM) の量、プロセッサ速度、またはプロセッサ・コアの個数を含む処理特性の第 1 のセットを有する第 1 のブレードと、RAM の量、プロセッサ速度、またはプロセッサ・コアの個数を含む処理特性の第 2 のセットを有する第 2 のブレードとを含み、

20

前記第 1 のブレードのオーソリティおよび前記第 2 のブレードのオーソリティから使用可能な前記処理特性を平衡化するために、前記第 1 のブレードより多数のオーソリティが、前記第 2 のブレードに分配される

請求項 14 に記載のストレージ・システム。

【請求項 20】

アプリケーション・レイヤ内の 1 つまたは複数のアプリケーションのコンピューティング・タスクは、前記複数のブレードのうちの少なくとも 1 つにまたがって分配され、

前記オーソリティの前記分配は、前記複数のブレードのそれぞれで使用可能な処理能力の相対量に比例する

30

請求項 14 に記載のストレージ・システム。

【発明の詳細な説明】

【背景技術】

【0001】

[0001] フラッシュなどのソリッド・ステート・メモリは、現在、大量のデータの記憶のために、集合的に回転媒体と称する従来のハード・ディスク・ドライブ (HDD)、書込可能 CD (コンパクト・ディスク) ドライブ、または書込可能 DVD (デジタル多用途ディスク) ドライブと、テープ・ドライブとを増補しまたは置換するためにソリッド・ステート・ドライブ (SSD) 内で使用されている。フラッシュおよび他のソリッド・ステート・メモリは、回転媒体とは異なる特性を有する。それでも、多くのソリッド・ステート・ドライブは、互換性の理由からハード・ディスク・ドライブ標準規格に従うように設計され、これは、フラッシュおよび他のソリッド・ステート・メモリの、機能強化された特徴の提供または独自の態様の利用を困難にする。

40

【0002】

[0002] 諸実施形態が生じるのは、この文脈内である。

【発明の概要】

【課題を解決するための手段】

【0003】

[0003] いくつかの実施形態では、ストレージ・システム内で処理能力を管理する方法

50

が提供される。この方法は、複数のブレードを提供することであって、ブレードの第1のサブセットのそれぞれは、ストレージ・ノードおよびユーザ・データを記憶するストレージ・メモリを有し、ブレードの第2の異なるサブセットのそれぞれは、コンピューティング動作のメモリを有することのできるコンピュータ・ノード（コンピュータ専用ノードと呼ばれる場合がある）を有する、提供することを含む。この方法は、複数のブレードにまたがって、少なくとも1つのコンピュータ専用ノードを含む複数のノードにオーソリティを分配することであって、各オーソリティは、ある範囲のユーザ・データの所有権（オーナーシップ）を有する、分配することを含む。

【0004】

[0004] いくつかの実施形態では、プロセッサによって実行される時に、プロセッサに方法を実行させる命令をその上に有する有形の非一時的コンピュータ可読媒体が提供される。この方法は、複数のブレードを提供することであって、第1のサブセットのそれぞれは、ストレージ・ノードおよびストレージ・メモリを有し、第2の異なるサブセットのそれぞれは、コンピュータ専用ノードを有する、提供することを含む。この方法は、複数のブレードにまたがって、少なくとも1つのコンピュータ専用ノードを含む複数のノードにオーソリティを分配することであって、各オーソリティは、ある範囲のユーザ・データの所有権を有する、分配することを含む。

10

【0005】

[0005] いくつかの実施形態では、ストレージ・システムが提供される。このシステムは、複数のブレードであって、第1のサブセットのそれぞれは、ストレージ・ノードおよびストレージ・メモリを有し、第2の異なるサブセットのそれぞれは、コンピュータ専用ノードを有する、複数のブレードを含む。このシステムは、ストレージ・システムを形成する複数のブレードを含み、オーソリティは、複数のブレードにまたがって、少なくとも1つのコンピュータ専用ノードを含む複数のノードに分配され、各オーソリティは、ある範囲のユーザ・データの所有権を有する。

20

【0006】

[0006] 諸実施形態の他の態様および利点は、説明される実施形態の原理を例として示す添付図面に関連して解釈される以下の詳細な説明から明白になる。

【0007】

[0007] 説明される実施形態およびその利点は、添付図面に関連して解釈される以下の説明を参照することによって最もよく理解され得る。これらの図面は、説明される実施形態の趣旨および範囲から逸脱せずに当業者によって説明される実施形態に対して行われ得る形態および詳細における変更を決して限定しない。

30

【図面の簡単な説明】

【0008】

【図1】 [0008] いくつかの実施形態による、ネットワーク・アタッチト・ストレージを提供するための複数のストレージ・ノードおよび各ストレージ・ノードに結合された内部ストレージを有するストレージ・クラスタを示す透視図である。

【図2】 [0009] いくつかの実施形態でストレージ・リソースとして図1のストレージ・クラスタのうちの一つまたは複数を使用することのできるエンタープライズ・コンピューティング・システムを示すシステム図である。

40

【図3】 [0010] いくつかの実施形態による、図1のストレージ・クラスタ内での使用に適する、異なる容量を有する複数のストレージ・ノードおよび不揮発性ソリッド・ステート・ストレージを示すブロック図である。

【図4】 [0011] いくつかの実施形態による、複数のストレージ・ノードを結合する相互接続スイッチ（interconnect switch）を示すブロック図である。

【図5】 [0012] いくつかの実施形態による、ストレージ・ノードの内容および不揮発性ソリッド・ステート・ストレージ・ユニットの内容を示す複数レベルのブロック図である。

【図6】 [0013] ハイブリッド・ブレードおよび一つまたは複数のコンピュータ・ブレード

50

ド (compute blade) にまたがって分配された、データを所有するオーソリティを有する、図 1 ~ 図 5 のストレージ・クラスタの実施形態を使用するストレージ・システムを示す図である。

【図 7】 [0014] ハイブリッド・ブレードおよびコンピュート・ブレードにまたがって外部入出力処理用のフロントフェーシング・ティア (front-facing tier)、オーソリティ用のオーソリティ・ティア、およびストレージ・メモリ用のストレージ・ティアに分配される処理能力を示す、図 6 のストレージ・システムを示す図である。

【図 8】 [0015] いくつかの実施形態によるストレージ・クラスタ、ストレージ・ノード、および / または不揮発性ソリッド・ステート・ストレージの実施形態上でまたはこれによって実践され得る、ストレージ・システム内で処理能力を管理する方法を示す流れ図である。

10

【図 9】 [0016] いくつかの実施形態によるストレージ・クラスタ、ストレージ・ノード、および / または不揮発性ソリッド・ステート・ストレージの実施形態上でまたはこれによって実践され得る、ブレードの追加時にストレージ・システム内で処理能力を管理する方法を示す流れ図である。

【図 10】 [0017] 本明細書で説明される実施形態を実施することのできる例示的なコンピューティング・デバイスを示す図である。

【発明を実施するための形態】

【0009】

[0018] 以下に説明される実施形態は、1つまたは複数のユーザ・システムもしくはクライアント・システム、またはストレージ・クラスタの外部の他のソースから発するユーザ・データなどのユーザ・データを記憶するストレージ・クラスタを説明するものである。ストレージ・クラスタは、イレージャ・コーディングおよびメタデータの冗長コピーを使用して、シャーシ内に収容されたストレージ・ノードにまたがってユーザ・データを分配する。イレージャ・コーディングは、ディスク、ストレージ・ノード、または地理的位置などの異なる位置のセットにまたがってデータが記憶される、データ保護またはデータ再構成の方法を指す。フラッシュ・メモリは、諸実施形態と一体化され得るソリッド・ステート・メモリの 1 タイプであるが、諸実施形態は、他のタイプのソリッド・ステート・メモリまたは非ソリッド・ステート・メモリを含む他の記憶媒体に拡張され得る。ストレージ位置および作業負荷の制御は、クラスタ化されたピアツーピア・システム内のストレージ位置にまたがって分配される。様々なストレージ・ノードの間の通信の調停、ストレージ・ノードが使用不能になった時の検出、および様々なストレージ・ノードにまたがる I/O (入出力) の平衡化などのタスクは、すべてが分配されて処理される。データは、いくつかの実施形態でデータ復元をサポートする、データ・フラグメントまたはストライプ単位で複数のストレージ・ノードにまたがって配置されまたは分配される。データの所有権は、入出力パターンとは独立にクラスタ内で再割当され得る。以下でより詳細に説明されるこのアーキテクチャは、システムが動作状態のままでありながらクラスタ内のストレージ・ノードが障害を発生することを可能にする。というのは、データが、他のストレージ・ノードから再構成され、したがって、入出力動作に関して使用可能のままになることができるからである。様々な実施形態では、ストレージ・ノードがクラスタ・ノード、ブレード、またはサーバと呼ばれる場合がある。ストレージ・クラスタのいくつかの実施形態は、ストレージ・メモリを有するハイブリッド・ブレードと、ストレージ・メモリを有しないコンピュート・ブレードとを有する。それぞれがある範囲のユーザ・データの所有権を有するオーソリティは、各オーソリティから使用可能な処理能力を平衡化するためまたはポリシ、合意、もしくはマルチテナント・サービスに従って処理能力を分配するために、ハイブリッド・ブレードまたはハイブリッド・ブレードおよびコンピュート・ブレードにまたがって分配される。

20

30

40

【0010】

[0019] ストレージ・クラスタは、シャーシすなわち 1 つまたは複数のストレージ・ノードを収容するエンクロージャ内に含まれる。配電バスなど、各ストレージ・ノードに電

50

力を供給する機構と、ストレージ・ノード間の通信を可能にする通信バスなどの通信機構とが、シャーシ内に含まれる。ストレージ・クラスタは、いくつかの実施形態によれば、1つの位置で独立システムとして走行することができる。一実施形態では、シャーシは、独立にイネーブルされまたはディスエーブルされることが可能な、配電バスと通信バスとの両方の少なくとも2つのインスタンスを含む。内部通信バスは、イーサネット・バスとすることができるが、Peripheral Component Interconnect (PCI) Express、InfiniBand、その他などの他の技術が、同等に適する。シャーシは、直接またはスイッチを介する複数のシャーシの間の通信とクライアント・システムとの通信とを可能にする外部通信バス用のポートを提供する。外部通信は、イーサネット、InfiniBand、Fibre Channel、その他などの技術を使用することができる。いくつかの実施形態では、外部通信バスは、シャーシ間通信とクライアント通信とに異なる通信バス技術を使用する。スイッチが、シャーシ内またはシャーシの間に展開される場合に、スイッチは、複数のプロトコルまたは技術の間の変換として働くことができる。複数のシャーシが、ストレージ・クラスタを定義するために接続される時に、ストレージ・クラスタは、プロプライエタリ・インターフェースまたは、network file system (NFS)、common internet file system (CIFS)、small computer system interface (SCSI)、もしくはハイパーテキスト転送プロトコル (HTTP) などの標準インターフェースのいずれかを使用してクライアントによってアクセスされ得る。クライアント・プロトコルからの変換は、スイッチで、シャーシ外部通信バスで、または各ストレージ・ノード内で行われ得る。

【0011】

[0020] 各ストレージ・ノードは、1つまたは複数のストレージ・サーバとすることができ、各ストレージ・サーバは、ストレージ・ユニットと呼ばれる場合がある1つまたは複数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る。一実施形態は、各ストレージ・ノード内の単一のストレージ・サーバと、1つと8つとの間の不揮発性ソリッド・ステート・メモリ・ユニットとを含むが、この一例は、限定的であることを意図されたものではない。ストレージ・サーバは、プロセッサ、ダイナミック・ランダム・アクセス・メモリ (DRAM)、内部通信バス用のインターフェース、および電力バスのそれぞれのための配電用のインターフェースを含むことができる。ストレージ・ノードの内部では、インターフェースおよびストレージ・ユニットが、通信バス、たとえば、いくつかの実施形態ではPCI Express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・ノード通信バスを介して内部通信バス・インターフェースに直接にアクセスし、または、バス・インターフェースにアクセスするようにストレージ・ノードに要求することができる。不揮発性ソリッド・ステート・メモリ・ユニットは、組込み中央処理装置 (CPU)、ソリッド・ステート・ストレージ・コントローラ、およびある量、たとえばいくつかの実施形態では2テラバイト (TB) ~ 32TBの間のソリッド・ステート・マス・ストレージを含む。DRAMなどの組込み揮発性記憶媒体およびエネルギー貯蔵装置が、不揮発性ソリッド・ステート・メモリ・ユニット内に含まれる。いくつかの実施形態では、エネルギー貯蔵装置は、電力消失の場合にDRAM内容のサブセットを安定した記憶媒体に転送することを可能にする、キャパシタ、スーパーキャパシタ、またはバッテリーである。いくつかの実施形態では、不揮発性ソリッド・ステート・メモリ・ユニットは、DRAMを置換し、縮小された電力維持装置を可能にする、相変化メモリまたは磁気抵抗ランダム・アクセス・メモリ (MRAM) などのストレージ・クラス・メモリを用いて構成される。

【0012】

[0021] ストレージ・ノードおよ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うちの一つが、ストレージ・クラスタ内で先を見越してデータを再構築する能力である。ストレージ・ノードおよび不揮発性ソリッド・ステート・ストレージは、ストレージ・クラスタ内のストレージ・ノードまたは不揮発性ソリッド・ステート・ストレージが

到達不能である時を、そのストレージ・ノードまたは不揮発性ソリッド・ステート・ストレージを用いてデータを読み取る試みがあるかどうかとは独立に判定することができる。その後、ストレージ・ノードおよび不揮発性ソリッド・ステート・ストレージは、少なくとも部分的に新しい位置でデータを復元し、再構築するために協力する。これは、ストレージ・クラスタを使用するクライアント・システムから開始された読取アクセスのためにデータが必要になるまで待つことなく、システムがデータを再構築するという点で、先を見越した再構築を構成する。ストレージ・メモリおよびその動作の上記およびさらなる詳細は、以下で議論する。

【0013】

[0022] 図1は、いくつかの実施形態による、ネットワーク・アタッチト・ストレージまたはストレージ・エリア・ネットワークを提供するために複数のストレージ・ノード150および各ストレージ・ノードに結合される内部ソリッド・ステート・メモリを有するストレージ・クラスタ160の透視図である。ネットワーク・アタッチト・ストレージ、ストレージ・エリア・ネットワーク、もしくはストレージ・クラスタ、または他のストレージ・メモリは、物理構成要素とそれによって提供されるストレージ・メモリの量との両方の柔軟で再構成可能な配置で、それぞれが1つまたは複数のストレージ・ノード150を有する1つまたは複数のストレージ・クラスタ160を含むことができる。ストレージ・クラスタ160は、ラックにおさまるように設計され、1つまたは複数のラックが、ストレージ・メモリに関して望まれる通りにセット・アップされ、投入され得る。ストレージ・クラスタ160は、複数のスロット142を有するシャーシ138を有する。シャーシ138が、ハウジング、エンクロージャ、またはラック・ユニットと呼ばれる場合があることを了解されたい。一実施形態では、シャーシ138が14個のスロット142を有するが、他の個数のスロットが、たやすく案出される。たとえば、いくつかの実施形態は、4個のスロット、8個のスロット、16個のスロット、32個のスロット、または他の適切な個数のスロットを有する。各スロット142は、いくつかの実施形態では1つのストレージ・ノード150に対処することができる。シャーシ138は、ラックにシャーシ138を取り付けるのに利用され得るフラップ148を含む。ファン144は、ストレージ・ノード150およびその構成要素の冷却のための空気循環を提供するが、他の冷却構成要素が使用され得、あるいは、冷却構成要素のない実施形態が案出され得る。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を一緒に結合し、メモリへの通信のためにネットワークに結合する。図1に示された実施形態では、スイッチ・ファブリック146およびファン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図示されているが、スイッチ・ファブリック146およびファン144の右側のスロット142は、例示のために、空であり、ストレージ・ノード150の挿入に使用可能である。この構成は一例であり、1つまたは複数のストレージ・ノード150が、様々なさらなる配置でスロット142を占有することができる。ストレージ・ノード配置は、いくつかの実施形態では順次または隣接である必要がない。ストレージ・ノード150は、活線挿抜可能であり、これは、ストレージ・ノード150が、システムを停止させまたはその電源を切ることなく、シャーシ138内のスロット142に挿入されまたはスロット142から除去されることが可能であることを意味する。スロット142からのストレージ・ノード150の挿入または除去の際に、システムは、その変化を認識し、適合するために自動的に再構成する。再構成は、いくつかの実施形態では、冗長性を回復することおよび/またはデータもしくは負荷を再平衡化することを含む。

【0014】

[0023] 各ストレージ・ノード150は、複数の構成要素を有することができる。この図に示された実施形態では、ストレージ・ノード150は、CPU 156すなわちプロセッサと、CPU 156に結合されたメモリ154と、CPU 156に結合された不揮発性ソリッド・ステート・ストレージ152とを投入されたプリント回路基板158を含むが、さらなる実施形態では、他の実装および/または構成要素が使用され得る。メモリ154は、CPU 156によって実行される命令および/またはCPU 156によ

って操作されるデータを有する。以下でさらに説明するように、不揮発性ソリッド・ステート・ストレージ 152 は、フラッシュまたは、さらなる実施形態では他のタイプのソリッド・ステート・メモリを含む。

【0015】

[0024] 図2は、ストレージ・リソース108として図1のストレージ・ノード、ストレージ・クラスタ、および/または不揮発性ソリッド・ステート・ストレージのうちの1つまたは複数を使用することのできるエンタープライズ・コンピューティング・システム102のシステム図である。たとえば、図2のフラッシュ・ストレージ128は、いくつかの実施形態で図1のストレージ・ノード、ストレージ・クラスタ、および/または不揮発性ソリッド・ステート・ストレージを一体化することができる。エンタープライズ・コンピューティング・システム102は、処理リソース104、ネットワーク・リソース106、およびフラッシュ・ストレージ128を含むストレージ・リソース108を有する。フラッシュ・コントローラ130およびフラッシュ・メモリ132は、フラッシュ・ストレージ128内に含まれる。様々な実施形態では、フラッシュ・ストレージ128は、1つまたは複数のストレージ・ノードまたはストレージ・クラスタを含むことができ、フラッシュ・コントローラ130はCPUを含み、フラッシュ・メモリ132は、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。いくつかの実施形態では、フラッシュ・メモリ132は、異なるタイプのフラッシュ・メモリまたは同一のタイプのフラッシュ・メモリを含むことができる。エンタープライズ・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の展開に適する環境を示すが、フラッシュ・ストレージ128は、より大型もしくはより小型の他のコンピューティング・システムもしくはコンピューティング・デバイス内で、またはより少数もしくは追加のリソースを有するエンタープライズ・コンピューティング・システム102の変形形態で使用され得る。エンタープライズ・コンピューティング・システム102は、サービスを提供しまたは利用するために、インターネットなどのネットワーク140に結合され得る。たとえば、エンタープライズ・コンピューティング・システム102は、クラウド・サービス、物理コンピューティング・リソース、または仮想コンピューティング・サービスを提供することができる。

10

20

【0016】

[0025] エンタープライズ・コンピューティング・システム102内では、様々なリソースが配置され、様々なコントローラによって管理される。処理コントローラ110は、処理リソース104を管理し、処理リソース104は、プロセッサ116およびランダム・アクセス・メモリ(RAM)118を含む。ネットワーク・コントローラ112は、ネットワーク・リソース106を管理し、ネットワーク・リソース106は、ルータ120、スイッチ122、およびサーバ124を含む。ストレージ・コントローラ114は、ストレージ・リソース108を管理し、ストレージ・リソース108は、ハード・ドライブ126およびフラッシュ・ストレージ128を含む。他のタイプの処理リソース、ネットワーク・リソース、およびストレージ・リソースが、実施形態と共に含まれ得る。いくつかの実施形態では、フラッシュ・ストレージ128が、ハード・ドライブ126を完全に置換する。エンタープライズ・コンピューティング・システム102は、様々なリソースを物理コンピューティング・リソースとして、または変形形態では物理コンピューティング・リソースによってサポートされる仮想コンピューティング・リソースとして、提供しまたは割り振ることができる。たとえば、様々なリソースは、ソフトウェアを実行する1つまたは複数のサーバを使用して実施され得る。ファイルもしくはデータ・オブジェクトまたは他の形のデータは、ストレージ・リソース108内に記憶される。

30

40

【0017】

[0026] 様々な実施形態では、エンタープライズ・コンピューティング・システム102は、ストレージ・クラスタを投入された複数のラックを含むことができ、これらは、クラスタまたはサーバ・ファーム内などの単一の物理位置に配置され得る。他の実施形態で

50

は、複数のラックが、ネットワークによって接続された、様々な都市、州、または国などの複数の物理位置に配置され得る。ラックのそれぞれ、ストレージ・クラスタのそれぞれ、ストレージ・ノードのそれぞれ、および不揮発性ソリッド・ステート・ストレージのそれぞれは、他のものとは独立に再構成可能であるストレージ空間のそれぞれの量を伴って個別に構成され得る。したがって、ストレージ容量は、不揮発性ソリッド・ステート・ストレージのそれぞれで柔軟に追加され、アップグレードされ、取り去られ、復元され、かつ/または再構成され得る。前に言及したように、各ストレージ・ノードは、いくつかの実施形態で1つまたは複数のサーバを実施することができる。

【 0 0 1 8 】

[0027] 図 3 は、図 1 のシャーシ内での使用に適する、異なる容量を有する複数のストレージ・ノード 1 5 0 および不揮発性ソリッド・ステート・ストレージ 1 5 2 を示すブロック図である。各ストレージ・ノード 1 5 0 は、不揮発性ソリッド・ステート・ストレージ 1 5 2 の 1 つまたは複数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 1 5 2 は、いくつかの実施形態で、ストレージ・ノード 1 5 0 上または他のストレージ・ノード 1 5 0 内の他の不揮発性ソリッド・ステート・ストレージ 1 5 2 とは異なる容量を含むことができる。代替案では、1 つのストレージ・ノード上または複数のストレージ・ノード上の不揮発性ソリッド・ステート・ストレージ 1 5 2 のすべてが、同一の容量または同一容量および/もしくは異なる容量の組合せを有することができる。この柔軟性が、図 3 に示されており、図 3 は、4 T B 容量、8 T B 容量、および 3 2 T B 容量の混合された不揮発性ソリッド・ステート・ストレージ 1 5 2 を有する 1 つのストレージ・ノード 1 5 0 と、それぞれ 3 2 T B 容量の不揮発性ソリッド・ステート・ストレージ 1 5 2 を有するもう 1 つのストレージ・ノード 1 5 0 と、それぞれ 8 T B 容量の不揮発性ソリッド・ステート・ストレージ 1 5 2 を有するさらに別のストレージ・ノード 1 5 0 との例を示す。様々なさらなる組合せおよび容量が、本明細書の教示に従ってたやすく案出される。クラスタリング、たとえばストレージ・クラスタを形成するためのストレージのクラスタリングの文脈では、ストレージ・ノードは、不揮発性ソリッド・ステート・ストレージ 1 5 2 とされるか、これを含むことができる。不揮発性ソリッド・ステート・ストレージ 1 5 2 は、以下でさらに説明するように、不揮発性ソリッド・ステート・ストレージ 1 5 2 が不揮発性ランダム・アクセス・メモリ (N V R A M) 構成要素を含み得るので、便利なクラスタリング・ポイントである。

【 0 0 1 9 】

[0028] 図 1 および図 3 を参照すると、ストレージ・クラスタ 1 6 0 はスケーラブルであり、これは、上で説明したように、不均一なストレージ・サイズを有するストレージ容量がたやすく追加されることを意味する。1 つまたは複数のストレージ・ノード 1 5 0 が、各シャーシに差し込まれ、除去され得、ストレージ・クラスタは、いくつかの実施形態で自己構成する。プラグイン・ストレージ・ノード 1 5 0 は、配達時にシャーシに設置済みであれ後に追加されるのであれ、異なるサイズを有することができる。たとえば、一実施形態では、ストレージ・ノード 1 5 0 は、4 T B の任意の倍数、たとえば 8 T B 、 1 2 T B 、 1 6 T B 、 3 2 T B など有することができる。さらなる実施形態では、ストレージ・ノード 1 5 0 は、他のストレージ量またはストレージ容量の任意の倍数を有することができる。各ストレージ・ノード 1 5 0 のストレージ容量は、ブロードキャストされ、データをどのようにストライピングするべきかの判断に影響する。最大のストレージ効率のために、一実施形態は、シャーシ内の 1 つまでまたは 2 つまでの不揮発性ソリッド・ステート・ストレージ・ユニット 1 5 2 またはストレージ・ノード 1 5 0 の消失を伴う継続動作という所定の要件の支配下で、ストライプにおいてできる限り幅広く自己構成することができる。

【 0 0 2 0 】

[0029] 図 4 は、複数のストレージ・ノード 1 5 0 を結合する、通信相互接続 1 7 0 および配電バス 1 7 2 を示すブロック図である。戻って図 1 を参照すると、通信相互接続 1 7 0 は、いくつかの実施形態でスイッチ・ファブリック 1 4 6 内に含まれまたはこれを用

いて実施され得る。複数のストレージ・クラスタ160がラックを占有する場合に、通信相互接続170は、いくつかの実施形態で、ラック・スイッチの最上部に含まれ、またはこれを用いて実施され得る。図4に示されているように、ストレージ・クラスタ160は、単一のシャーシ138内に閉じこめられる。外部ポート176は、通信相互接続170を介してストレージ・ノード150に結合され、外部ポート174は、ストレージ・ノードに直接に結合される。外部電力ポート178は、配電バス172に結合される。ストレージ・ノード150は、図3を参照して説明したように、不揮発性ソリッド・ステート・ストレージ152の変化する量および異なる容量を含むことができる。さらに、1つまたは複数のストレージ・ノード150は、図4に示されているようにコンピュータ専用のストレージ・ノードとされ得る。オーソリティ168が、たとえばメモリ内に記憶されるリストまたは他のデータ構造として、不揮発性ソリッド・ステート・ストレージ152上で実施される。いくつかの実施形態では、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、不揮発性ソリッド・ステート・ストレージ152のコントローラまたは他のプロセッサ上で実行するソフトウェアによってサポートされる。さらなる実施形態では、オーソリティ168は、たとえばメモリ154内に記憶されるリストまたは他のデータ構造として、ストレージ・ノード150上で実施され、ストレージ・ノード150のCPU156上で実行するソフトウェアによってサポートされる。オーソリティ168は、いくつかの実施形態で、データが不揮発性ソリッド・ステート・ストレージ152内でどのようにどこに記憶されるのかを制御する。この制御は、どのタイプのイレージャ・コーディング方式がデータに適用されるのかと、どのストレージ・ノード150がデータのどの部分を有するのかと、を判定するのを援助する。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に割り当てられ得る。各オーソリティは、様々な実施形態で、inode番号、セグメント番号、または、ファイル・システム、ストレージ・ノード150、もしくは不揮発性ソリッド・ステート・ストレージ152によってデータに割り当てられる他のデータ識別子の範囲を制御することができる。

【0021】

[0030] データのすべての片およびメタデータのすべての片は、いくつかの実施形態で、システム内で冗長性を有する。さらに、データのすべての片およびメタデータのすべての片は、オーソリティと呼ばれる場合もある所有者を有する。そのオーソリティが、たとえばストレージ・ノードの障害を介して、到達不能である場合には、そのデータまたはメタデータをどのようにして見つけるべきかに関する継続のプランがある。様々な実施形態では、オーソリティ168の冗長なコピーがある。オーソリティ168は、いくつかの実施形態ではストレージ・ノード150および不揮発性ソリッド・ステート・ストレージ152に対する関係を有する。データ・セグメント番号またはデータの他の識別子の範囲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に割り当てられ得る。いくつかの実施形態では、そのような範囲のすべてに関するオーソリティ168が、ストレージ・クラスタの不揮発性ソリッド・ステート・ストレージ152にわたって分配される。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へのアクセスを提供するネットワーク・ポートを有する。データは、セグメント内に記憶され得、このセグメントは、セグメント番号を関連付けられ、そのセグメント番号は、いくつかの実施形態ではRAID (redundant array of independent disk) ストラタイプの構成の間接参照である。したがって、オーソリティ168の割り当ておよび使用は、データへの間接参照を確立する。間接参照は、いくつかの実施形態によれば、間接的に、この場合にはオーソリティ168を介して、データを参照する能力と呼ばれる場合がある。セグメントは、不揮発性ソリッド・ステート・ストレージ152のセットと、データを含む可能性がある不揮発性ソリッド・ステート・ストレージ152のセットへのローカル識別子とを識別する。いくつかの実施形態では、ローカル識別子は、デバイスへのオフセットであり、複数のセグメントによって順次再利用され得る。他の実施形態では、ローカル識別子は、特性のセグメントに関して一意であり、絶対に再利用されない。不揮発性ソリ

10

20

30

40

50

ッド・ステート・ストレージ 152 内のオフセットは、不揮発性ソリッド・ステート・ストレージ 152 (RAID ストライプの形の) に書き込むかこれから読み取るためにデータを突き止めることに適用される。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またがってストライピングされ、この不揮発性ソリッド・ステート・ストレージ 152 は、特定のデータ・セグメントのオーソリティ 168 を有する不揮発性ソリッド・ステート・ストレージ 152 を含み、またはこれとは異なるものとされ得る。

【0022】

[0031] たとえばデータ移動中またはデータ再構成中に、データの特定のセグメントが配置される場所に変化がある場合には、その不揮発性ソリッド・ステート・ストレージ 152 またはそのオーソリティ 168 を有するストレージ・ノード 150 にある、そのデータ・セグメントのオーソリティ 168 が相談されなければならない。データの特定の片を突き止めるために、諸実施形態は、データ・セグメントのハッシュ値を計算し、または `inode` 番号もしくはデータ・セグメント番号を適用する。この動作の出力は、データのその特定の片のオーソリティ 168 を有する不揮発性ソリッド・ステート・ストレージ 152 をポイントする。いくつかの実施形態では、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)、たとえば、セグメント番号、`inode` 番号、またはディレクトリ番号をオーソリティ識別子にマッピングする。このマッピングは、ハッシュまたはビット・マスクなどの計算を含むことができる。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 にマッピングすることであり、これは、明示的なマッピングを介して行われ得る。動作は、反復可能であり、その結果、計算が実行される時に、計算の結果は、そのオーソリティ 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 を繰り返して信頼できる形でポイントするようになる。動作は、入力として到達可能なストレージ・ノードのセットを含むことができる。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、最適セットが変化する。いくつかの実施形態では、永續される値は、現在の割当 (常に真である) であり、計算される値は、クラスタがそれに向かう再構成を試みるターゲット割当である。この計算は、到達可能であり同一のクラスタを構成する不揮発性ソリッド・ステート・ストレージ 152 のセットの存在下でオーソリティの最適不揮発性ソリッド・ステート・ストレージ 152 を判定するのに使用され得る。この計算は、割り当てられた不揮発性ソリッド・ステート・ストレージが到達不能である場合であってもオーソリティが判定され得るようにするために、オーソリティを不揮発性ソリッド・ステート・ストレージ・マッピングに記録もするピア不揮発性ソリッド・ステート・ストレージ 152 の順序集合をも判定する。いくつかの実施形態では、特定のオーソリティ 168 が使用不能である場合に、複製オーソリティ 168 または代理オーソリティ 168 が相談され得る。

【0023】

[0032] 図 1 ~ 図 4 を参照すると、ストレージ・ノード 150 上の CPU 156 の多数のタスクのうち 2 つは、書込データを分解することおよび読取データを再アセンブルすることである。システムが、データが書き込まれなければならないと判定した時に、そのデータのオーソリティ 168 が、上記のように突き止められる。データのセグメント ID が既に判定されている時には、書き込む要求が、セグメントから判定されたオーソリティ 168 のホストであると現在判定されている不揮発性ソリッド・ステート・ストレージ 152 に転送される。不揮発性ソリッド・ステート・ストレージ 152 および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 CPU 156 は、データを分解しまたはシャードし、様々な不揮発性ソリッド・ステート・ストレージ 152 にデータを送信する。送信されたデータは、イレージャ・コーディング方式に従ってデータ・ストライプとして書き込まれる。いくつかの実施形態では、データがプルされることが要求され、他の実施形態では、データはプッシュされる。逆に、データが読み取られる時には、データを含むセグメント ID のオーソリティ 168 が、上で説明したように突き止

められる。不揮発性ソリッド・ステート・ストレージ 152 および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 CPU 156 は、不揮発性ソリッド・ステート・ストレージとそのオーソリティによってポイントされる対応するストレージ・ノードとにデータを要求する。いくつかの実施形態では、データは、データ・ストライプとしてフラッシュ・ストレージから読み取られる。次に、ストレージ・ノード 150 のホスト CPU 156 は、読み取られたデータを再アセンブルし、適当なイレージャ・コーディング方式に従ってすべてのエラー（存在する場合に）を訂正し、再アセンブルされたデータをネットワークに転送する。さらなる実施形態では、これらのタスクの一部またはすべてが、不揮発性ソリッド・ステート・ストレージ 152 内で処理され得る。いくつかの実施形態では、セグメント・ホストは、ストレージにページを要求することと、その後、元々の要求を行っているストレージ・ノードにデータを送信することとによって、データがストレージ・ノード 150 に送信されることを要求する。

10

【0024】

[0033] 一部のシステム、たとえば UNIX スタイルのファイル・システムでは、データは、インデックス・ノードすなわち `inode` を用いて処理され、`inode` は、ファイル・システム内のオブジェクトを表現するデータ構造を指定する。オブジェクトは、たとえばファイルまたはディレクトリとすることができる。メタデータが、他の属性の中でもパMISSION・データおよび作成タイムスタンプなどの属性としてオブジェクトに付随することができる。セグメント番号は、ファイル・システム内のそのようなオブジェクトのすべてまたは一部に割り当てられ得る。他のシステムでは、データ・セグメントは、他所で割り当てられたセグメント番号を用いて処理される。議論において、分配の単位はエンティティであり、エンティティは、ファイル、ディレクトリ、またはセグメントとすることができる。すなわち、エンティティは、ストレージ・システムによって記憶されるデータまたはメタデータの単位である。エンティティは、オーソリティと呼ばれるセットにグループ化される。各オーソリティは、オーソリティ内のエンティティを更新するための排他的権利を有するストレージ・ノードであるオーソリティ所有者を有する。言い換えると、ストレージ・ノードはオーソリティを含み、そのオーソリティはエンティティを含む。

20

【0025】

[0034] いくつかの実施形態によれば、セグメントは、データの論理コンテナである。セグメントは、媒体アドレス空間と物理フラッシュ位置との間のアドレス空間である、すなわち、データ・セグメント番号は、このアドレス空間内にある。セグメントは、メタデータをも含むことができ、メタデータは、より上のレベルのソフトウェアの介入なしでデータ冗長性を回復する（異なるフラッシュ位置またはデバイスに再書込される）ことを可能にする。一実施形態では、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判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、セグメントを複数のデータおよび適用可能な場合にパリティ・シャードに分解することによって、たとえばメモリ障害および他の障害から保護される。データおよびパリティ・シャードは、イレージャ・コーディング方式に従って、ホスト CPU 156 に結合された不揮発性ソリッド・ステート・ストレージ 152 にまたがって分配される、すなわちストライピングされる（図 5 を参照されたい）。用語セグメントの使用は、いくつかの実施形態で、コンテナと、セグメントのアドレス空間内でのその場所とを指す。いくつかの実施形態によれば、用語ストライプの使用は、シャードの、セグメントと同一のセットを指し、シャードが冗長性情報またはパリティ情報と一緒にどのように分配されるのかを含む。

30

40

【0026】

[0035] 一連のアドレス空間変換が、ストレージ・システム全体にまたがって行われる。最上部には、`inode` にリンクするディレクトリ・エントリ（ファイル名）がある。`inode` は、データが論理的に記憶される媒体アドレス空間をポイントする。媒体アドレスは、大きいファイルの負荷を拡散するために一連の間接媒体を介してマッピングされ

50

、または複製もしくはスナップショットなどのデータ・サービスを実施することができる。媒体アドレスは、大きいファイルの負荷を拡散するために一連の間接媒体を介してマッピングされ、または複製もしくはスナップショットなどのデータ・サービスを実施することができる。その後、セグメント・アドレスが、物理フラッシュ位置に変換される。物理フラッシュ位置は、いくつかの実施形態によれば、システム内のフラッシュの量によって境界を定められるアドレス範囲を有する。媒体アドレスおよびセグメント・アドレスは、論理コンテナであり、いくつかの実施形態では、実用上無限になるために128ビット以上の識別子を使用し、再利用の見込みは、システムの期待される寿命より長いものとして計算される。論理コンテナからのアドレスは、いくつかの実施形態では階層的な形で割り振られる。当初に、各不揮発性ソリッド・ステート・ストレージ152は、アドレス空間の範囲を割り当てられ得る。この割り当てられた範囲内で、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との同期化なしにアドレスを割り振ることができる。

10

20

30

40

50

【0027】

[0036] データおよびメタデータは、変化する作業負荷パターンおよびストレージ・デバイスに関して最適化される、基礎になるストレージ・レイアウトのセットによって記憶される。これらのレイアウトは、複数の冗長性方式、圧縮フォーマット、およびインデックス・アルゴリズムを組み込む。これらのレイアウトの一部は、オーソリティおよびオーソリティ・マスタに関する情報を記憶し、他のレイアウトは、ファイル・メタデータおよびファイル・データを記憶する。冗長性方式は、単一のストレージ・デバイス(NANDフラッシュ・チップなど)内の破壊されたビットを許容する誤り訂正符号と、複数のストレージ・ノードの障害を許容するイレージャ符号と、データセンタ障害または地域の障害を許容する複製方式とを含む。いくつかの実施形態では、低密度パリティ検査(LDPC)符号が、単一のストレージ・ユニット内で使用される。いくつかの実施形態では、リード・ソロモン符号化が、ストレージ・クラスタ内で使用され、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、順序付きログ構造化インデックス(Log Structured Merge Treeなど)を使用して記憶され得、大きいデータは、ログ構造化レイアウト内では記憶されない場合がある。

【0028】

[0037] あるエンティティの複数のコピーにまたがって一貫性を維持するために、ストレージ・ノードは、計算を介して2つのことすなわち(1)そのエンティティを含むオーソリティおよび(2)そのオーソリティを含むストレージ・ノードについて暗黙のうちに合意する。オーソリティへのエンティティの割当は、エンティティをオーソリティに擬似ランダムに割り当てることによって、外部で作られる鍵に基づいてエンティティを範囲に分割することによって、または単一のエンティティを各オーソリティ内に配置することによって、行われ得る。擬似ランダム方式の例は、線形ハッシュ化と、Controlled Replication Under Scalable Hashing(CRUSH)を含む、ハッシュのReplication Under Scalable Hashing(RUSH)ファミリとである。いくつかの実施形態では、擬似ランダム割当は、ノードのセットが変化する可能性があるので、ノードへのオーソリティの割当だけに関して利用される。オーソリティのセットは、変化することができず、したがって、この実施形態では、任意の全射関数が適用され得る。一部の配置方式は、オーソリティをストレージ・ノード上に自動的に配置するが、他の配置方式は、ストレージ・ノードへのオーソリティの明示的なマッピングに頼る。いくつかの実施形態では、擬似ランダム方式が、各オーソリティから候補オーソリティ所有者のセットにマッピングするのに利用される。CRUSHに係る擬似ランダム・データ分配関数は、オーソリティをストレージ・ノードに割り当て、オーソリティがどこに割り当てられるのかのリストを作成することができる。各ストレージ・ノードは、擬似ランダム・データ分配関数のコピーを有し、分配の同一の計算と、その後のオーソリティの発見または突き止めに達することができる。擬似ランダム方式のそれぞれは、いくつかの実施形態で、同一のターゲット・ノードを推断

するために、入力としてストレージ・ノードの到達可能なセットを必要とする。エンティティがオーソリティ内に配置された後に、そのエンティティは、物理デバイス上で記憶され得、その結果、予期される障害が予期されないデータ消失につながるものがなくなる。いくつかの実施形態では、再平衡化アルゴリズムが、マシンの同一のセット上の同一のレイアウトのオーソリティ内のすべてのエントリのコピーを記憶することを試みる。

【0029】

[0038] 予期される障害の例は、デバイス障害、マシンの盗難、データセンタ火災、および、原子力事故または地質学的イベントなどの地域災害を含む。異なる障害は、異なるレベルの許容可能なデータ消失につながる。いくつかの実施形態では、盗まれたストレージ・ノードは、システムのセキュリティにも信頼性にも影響しないが、システム構成に応じて、地域イベントが、データの消失なし、数秒もしくは数分の失われた更新、または完全なデータ消失にさえつながる可能性がある。

10

【0030】

[0039] 諸実施形態では、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である。いくつかの実施形態では、オーソリティを含むストレージ・ノードは、永続ストレージを全く含まない。その代わりに、それらのストレージ・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・ユニットに接続される。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は、複数の通信技術からなり、不均一な性能特性およびフォールト・トレランス特性を有する。いくつかの実施形態では、上で言及したように、不揮発性ソリッド・ステート・ストレージ・ユニットは、PCI expressを介してストレージ・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、シャーシは、ストレージ・クラスタを形成するために一緒に接続される。ストレージ・クラスタは、いくつかの実施形態ではイーサネットまたはファイバ・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・クラスタが、ストレージ・グリッドに構成される場合に、複数のストレージ・クラスタは、インターネットまたは、「メトロ・スケール」リンクもしくはインターネットをトラバースしないプライベート・リンクなどの他の長距離ネットワークング・リンクを使用して接続される。

20

【0031】

[0040] オーソリティ所有者は、エンティティを変更し、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・ユニットへエンティティを移植し、エンティティのコピーを追加し、除去する、排他的な権利を有する。これは、基礎になるデータの冗長性を維持することを可能にする。オーソリティ所有者が障害を発生し、任を解かれようとしており、または過負荷である時に、オーソリティは、新しいストレージ・ユニットに転送される。過渡的障害は、すべての欠陥のないマシンが新しいオーソリティ位置に合意することを保証することを非自明にする。過渡的障害に起因して生じる曖昧さは、Paxos、ホット・ウォーム・フェイルオーバー(hot-warm failover)方式などのコンセンサス・プロトコルによって、リモート・システム管理者による手動介入を介して、またはローカル・ハードウェア管理者によって(障害を発生したマシンをクラスタから物理的に除去すること、または障害を発生したマシンのボタンを押すことなどによって)、自動的に達成され得る。いくつかの実施形態では、コンセンサス・プロトコルが使用され、フェイルオーバーは自動的である。いくつかの実施形態によれば、多すぎる障害または複製イベントが短すぎる時間期間内に発生する場合には、システムは、自己保存モードに入り、管理者が介入するまで複製アクティビティおよびデータ移動アクティビティを停止する。

30

40

【0032】

[0041] オーソリティがストレージ・ノードの間で転送され、オーソリティ所有者がそのオーソリティ内のエンティティを更新する時に、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・ユニットとの間でメッセージを転送する。永続

50

メッセージに関して、異なる目的を有するメッセージは、異なるタイプを有する。メッセージのタイプに応じて、システムは、異なる順序付け保証および異なる耐久性保証を維持する。永続メッセージが処理されつつある時に、それらのメッセージは、複数の耐久性のあるストレージ・ハードウェア技術および耐久性のないストレージ・ハードウェア技術内で一時的に記憶される。いくつかの実施形態では、メッセージは、RAM内、NVRAM内、およびNANDフラッシュ・デバイス上で記憶され、様々なプロトコルが、各記憶媒体を効率的に利用するために使用される。待ち時間に敏感なクライアント要求は、複製されたNVRAM内で、その後NAND内で永続化され得るが、バックグラウンドの再平衡化動作は、NANDに直接に永続化される。

【0033】

[0042] 永続メッセージは、送信される前に永続的に記憶される。これは、システムが、障害および構成要素交換にもかかわらずクライアント要求のために働き続けることを可能にする。多数のハードウェア構成要素が、システム管理者、製造業者、ハードウェア・サプライ・チェーン、および進行中の品質管理監視インフラストラクチャに可視の一意識別子を含むが、そのインフラストラクチャ・アドレスの上で走行するアプリケーションは、アドレスを仮想化する。これらの仮想化されたアドレスは、構成要素の障害および交換にかかわらず、ストレージ・システムの寿命の間に変化しない。これは、ストレージ・システムの各構成要素が、再構成またはクライアント要求処理の混乱を伴わずに経時的に交換されることを可能にする。

【0034】

[0043] いくつかの実施形態では、仮想化されたアドレスは、十分な冗長性を伴って記憶される。連続監視システムは、ハードウェア状況、ソフトウェア状況、およびハードウェア識別子を相関させる。これは、欠陥のある構成要素および製造詳細に起因する障害の検出および予測を可能にする。いくつかの実施形態で、監視システムは、構成要素をクリティカル・パスから除去することによって、障害が発生する前の、影響を受けるデバイスからのオーソリティおよびエンティティの先を見越した転送をも可能にする。

【0035】

[0044] 図5は、ストレージ・ノード150の内容およびストレージ・ノード150の不揮発性ソリッド・ステート・ストレージ152の内容を示す複数レベルのブロック図である。いくつかの実施形態で、データは、ネットワーク・インターフェース・コントローラ(NIC)202によってストレージ・ノード150との間で通信される。各ストレージ・ノード150は、上で議論したように、CPU 156と1つまたは複数の不揮発性ソリッド・ステート・ストレージ152とを有する。図5内で1レベル下に移動すると、各不揮発性ソリッド・ステート・ストレージ152は、不揮発性ランダム・アクセス・メモリ(NVRAM)204およびフラッシュ・メモリ206などの相対的に高速の不揮発性ソリッド・ステート・メモリを有する。いくつかの実施形態では、NVRAM 204は、プログラム/消去サイクルを必要としない構成要素(DRAM、MRAM、PCM)とすることができ、メモリが読み取られるよりはるかに頻繁に書き込まれることをサポートできるメモリとすることができる。図5内の別のレベルに下に移動すると、一実施形態では、NVRAM 204は、エネルギー貯蔵218によってバック・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などの高速揮発性メモリとして実施される。エネルギー貯蔵218は、電源障害の場合に、DRAM 216を内容がフラッシュ・メモリ206に転送されるのに十分に長く電力を供給される状態に保つのに十分な電力を供給する。いくつかの実施形態で、エネルギー貯蔵218は、電力消失の場合にDRAM 216の内容を安定した記憶媒体に転送することを可能にするのに十分なエネルギーの適切な供給を供給する、キャパシタ、スーパーキャパシタ、バッテリー、または他のデバイスである。フラッシュ・メモリ206は、複数のフラッシュ・ダイ222として実施され、複数のフラッシュ・ダイ222は、フラッシュ・ダイ222のパッケージまたはフラッシュ・ダイ222のアレイと呼ばれる場合がある。フラッシュ・ダイ222が、1パッケージあたり単一のダイを用いて、1パッケージあたり複数のダイ(すなわ

10

20

30

40

50

ち、マルチチップ・パッケージ)を用いて、ハイブリッド・パッケージ内で、プリント回路基板または他の基板の裸のダイとして、カプセル化されたダイとして、その他、任意の個数の形でパッケージ化され得ることを了解されたい。図示の実施形態では、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または他のプロセッサと、コントローラ212に結合された入出力(I/O)ポート210とを有する。I/Oポート210は、CPU156および/またはストレージ・ノード150のネットワーク・インターフェース・コントローラ202に結合される。フラッシュ入出力(I/O)ポート220が、フラッシュ・ダイ222に結合され、直接メモリ・アクセス・ユニット(DMA)214が、コントローラ212、DRAM216、およびフラッシュ・ダイ222に結合される。図示の実施形態では、I/Oポート210、コントローラ212、DMAユニット214、およびフラッシュI/Oポート220は、プログラマブル論理デバイス(PLD)208、たとえばフィールド・プログラマブル・ゲート・アレイ(FPGA)上で実施される。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、それを介してデータがフラッシュ・ダイ222に書き込まれまたは読み取られ得るレジスタ226とを有する。さらなる実施形態では、他のタイプのソリッド・ステート・メモリが、フラッシュ・ダイ222内に示されたフラッシュ・メモリの代わりにまたはそれに加えて使用される。

10

【0036】

[0045] ストレージ・クラスタ160は、本明細書で開示される様々な実施形態で、一般にストレージ・アレイと対比され得る。ストレージ・ノード150は、ストレージ・クラスタ160を作成するコレクションの一部である。各ストレージ・ノード150は、データを提供するのに必要なデータのスライスおよびコンピューティングを所有する。複数のストレージ・ノード150が、データを記憶し、取り出すために協力する。ストレージ・メモリまたはストレージ・デバイスは、一般にストレージ・アレイ内で使用される時に、データの処理および操作との関連がより少ない。ストレージ・アレイ内のストレージ・メモリまたはストレージ・デバイスは、データを読み取り、書き込み、または消去するコマンドを受け取る。ストレージ・アレイ内のストレージ・メモリまたはストレージ・デバイスは、それらがその中に組み込まれる、より大きいシステム、またはデータの意味を知らない。ストレージ・アレイ内のストレージ・メモリまたはストレージ・デバイスは、RAM、ソリッド・ステート・ドライブ、ハード・ディスク・ドライブ、その他など、様々なタイプのストレージ・メモリを含むことができる。本明細書で説明される不揮発性ソリッド・ステート・ストレージ152は、同時にアクティブになる、複数の目的のために働く複数のインターフェースを有する。いくつかの実施形態では、ストレージ・ノード150の機能性の一部が、不揮発性ソリッド・ステート・ストレージ152にシフトされ、不揮発性ソリッド・ステート・ストレージ152を不揮発性ソリッド・ステート・ストレージ152とストレージ・ノード150との組合せに変形する。コンピューティング(ストレージ・データに関する)を不揮発性ソリッド・ステート・ストレージ152内に配置することは、このコンピューティングをデータ自体のより近くに配置する。様々なシステム実施形態は、異なる能力を有するストレージ・ノード・レイヤの階層を有する。対照的に、ストレージ・アレイでは、コントローラが、棚またはストレージ・デバイス内で管理するデータのすべてに関するあらゆるものを所有し、知っている。ストレージ・クラスタ160内では、本明細書で説明するように、複数の不揮発性ソリッド・ステート・ストレージ・ユニット152および/またはストレージ・ノード150内の複数のコントローラが、様々な形で協力する(たとえば、イレージャ・コーディング、データ・シャーディング、メタデータの通信および冗長性、記憶容量の増減、データ復元、その他のため)。

20

30

40

【0037】

[0046] 図6は、ハイブリッド・ブレード602および1つまたは複数のコンピュータ・ブレード604にまたがって分配された、データを所有するオーソリティ168を有する、図1~図5のストレージ・クラスタ160の実施形態を使用するストレージ・システムの図である。ブレードは、回路網、プロセッサ、および関連するハードウェアを有する

50

物理構成である。各ハイブリッド・ブレード602は、ユーザ・データを記憶するストレージ・メモリを有し、このストレージ・メモリは、この実施形態ではフラッシュ・メモリ206であるが、さらなる実施形態では他のタイプのストレージ・メモリとすることができ、各ハイブリッド・ブレード602は、CPU 156およびDRAM 216を含む処理リソースを有する。各コンピュータ・ブレード604は、CPU 156およびDRAM 216を含む処理リソースを有するが、ハイブリッド・ブレード602とは異なって、ユーザ・データを記憶するストレージ・メモリを有しない。すなわち、ハイブリッド・ブレード602とは異なって、コンピュータ・ブレード604は、コンピュータ・ブレード604自体の上ではユーザ・データを記憶しない。ハイブリッド・ブレード602およびコンピュータ・ブレード604は、プログラム・メモリ用のROM（読取専用メモリ）など、他のタイプのメモリを有することができ、あるいは、プログラム・メモリおよび動作パラメータ用のDRAM 216または他のタイプのRAMすなわちシステム・メモリを使用することができる。各ブレード602、604は、ネットワーク・モジュール606（たとえば、図5のネットワーク・インターフェース・コントローラ202を参照されたい）を有し、ブレード602、604は、一緒に結合されてストレージ・クラスタ160を形成する。すべてのユーザ・データは、図1～図5を参照して説明したように、ハイブリッド・ブレード602上のストレージ・メモリ内に記憶される。ストレージ・システム内での様々な個数のハイブリッド・ブレード602またはハイブリッド・ブレード602とコンピュータ・ブレード604との混合の使用は、ストレージ・メモリの量および処理能力をシステムおよびクライアントの必要に合わせて調整することを可能にし、システム性能のこれらの態様のそれぞれまたは両方の交換およびアップグレードを可能にする。

【0038】

[0047] 図1～図5内のストレージ・クラスタ160の実施形態と同様に、各ブレード602、604は、ブレード602、604のすべてがストレージ・クラスタ160の動作の少なくとも一部に参加するという意味で、ストレージ・ノード150をホスティングしまたはストレージ・ノード150とすることができる。ノードは、ストレージ・システム内の論理構成であり、ストレージ・システム内の挙動、インテリジェンス、プロトコル、その他の責任を負う。ノードは、ブレード内に存在し、物理ブレード内の物理リソースを利用することができる。ハイブリッド・ブレード602は、ストレージ・メモリを有するストレージ・ノードであり、またはそのようなノードをその中に有し、コンピュータ・ブレード604は、ストレージ・メモリを有しないコンピュータ専用ノードであり、またはそのようなノードをその中に有する。すなわち、ハイブリッド・ブレード602上のストレージ・ノード150は、ハイブリッド・ブレード602上のコンピューティング・リソースとストレージ・メモリとの両方を使用することができ、ユーザ・データの読み書きおよび他のストレージ・ノード・タスクの実行の際に他のハイブリッド・ブレード602上のストレージ・メモリを使用することができる。コンピュータ・ブレード604上のコンピュータ専用ノードは、コンピュータ・ブレード604上のコンピューティング・リソースを使用することができるが、コンピュータ・ブレード604にストレージ・メモリが欠けているので、ハイブリッド・ブレード602上のストレージ・メモリを使用することができる。コンピュータ・ブレード604上のコンピュータ専用ノードは、コンピュータ・ブレード604上のコンピューティング・リソース（ローカル・メモリ、たとえばROMおよびDRAM 216を含む）を使用するが、ストレージ・ノード150によって実行されるストレージ・タスクではなくコンピューティング・タスクを実行し、したがって、ストレージ・ノード150と同一の形でハイブリッド・ブレード604のいずれかのストレージ・メモリを使用することはしない。たとえば診断、修理、またはストレージ・ノード150の通常のタスクの外部の他の目的もしくは機能のために、コンピュータ専用ノードがハイブリッド・ブレード602上のストレージ・メモリにアクセスする可能性があるアプリケーションがある可能性がある。本明細書で説明されるコンピュータ専用ノードは、いくつかの実施形態でコンピュータ・ノードと呼ばれる場合がある。オーソリティ1

68は、任意のストレージ・ノード内に存在することができ、したがって、ハイブリッド・ブレード602および/またはコンピュータ・ブレード604内に存在することができる。各ストレージ・ノードは、1つまたは複数のオーソリティ168を保持することができる。各オーソリティ168は、すべての他のオーソリティ168によって所有されるユーザ・データの範囲とオーバーラップしない範囲のユーザ・データを所有し、他のオーソリティ168とは独立に、その範囲のユーザ・データのイレージャ・コーディングおよび配置を選択し、制御する。図6に示された例では、左端のハイブリッド・ブレード602は4つのオーソリティ168を有し、右端のハイブリッド・ブレード602は4つのオーソリティ168を有し、左端のコンピュータ・ブレード604は2つのオーソリティ168を有し、右端のコンピュータ・ブレード604は4つのオーソリティ168を有する。これは例にすぎず、ブレード602、604のそれぞれは、様々な個数のオーソリティ168を有することができ、この個数がブレード602、604のそれぞれにおいて同一である必要はない。

【0039】

[0048] オーソリティ168は、様々な個数および方向で、あるブレード602、604から別のブレード602、604に移動され得る。図6のこの例では、オーソリティ168のうちの1つが、左端のハイブリッド・ブレード602から左端のコンピュータ・ブレード604に移動されるが、その代わりに、右端のコンピュータ・ブレード604（またはシステム内の任意の他のコンピュータ・ブレード604もしくは別のハイブリッド・ブレード602）に移動され得る。オーソリティ168のうちの1つは、右端のハイブリッド・ブレード602から左端のコンピュータ・ブレード604に移動される。コンピュータ・ブレード604上のオーソリティは、同様に、別のコンピュータ・ブレード604またはハイブリッド・ブレード602などに移動され得る。

【0040】

[0049] オーソリティ168を突き止め、かつ/または移動する様々な機構が、本教示に従って当業者によって開発され得る。たとえば、オーソリティ168が、様々なブレード602、604内のDRAM 216内に示されている。いくつかの実施形態では、様々なパラメータ、マップ、アカウントイング、レコード、ポインタ、その他、および/またはオーソリティ168を実施するデータ構造は、ブレード602、604のうちの1つのDRAM 216内に存在し、あるブレード602、604のDRAM 216から別のブレード602、604のDRAM 216にこの情報をコピーすることによって移動され得る。オーソリティ168のアクションを実行するために実行されるソフトウェア・コードは、DRAM 216内に存在し、同様に移動され得る。代替案では、ソフトウェア・コードは、不揮発性メモリなどの別のメモリまたはブレード602、604のファームウェア内に存在し、CPU 156によって実行されるが、マルチスレッディング・システム内の1つもしくは複数のパラメータまたは1つもしくは複数の実行スレッドに従ってアクティブ化されまたは非アクティブ化され得る。一実施形態では、オーソリティ168の様々なパラメータは、あるブレード602、604から別のブレード602、604に移動され、ブレード602、604のそれぞれのメモリ内のソフトウェア・コードは、ブレード602内のメモリ内に存在するオーソリティ168のパラメータに従って動作する。

【0041】

[0050] ブレードおよび602、604は、異なる量のコンピューティング能力もしくは処理能力、処理特性、またはコンピューティング・リソースを有することができる。たとえば、ある製品の異なるモデルまたは異なるバージョンが提供され得、あるいは、より後のバージョンが、より新しい、より高速の、より高密度のプロセッサもしくはメモリ、またはより多数のプロセッサ・コア608などを有することができる。一例では、左端のコンピュータ・ブレード604内に示されているように、あるCPU 156が、4つのコア608を有し、右端のコンピュータ・ブレード604内に示されているように、別のCPU 156が、8つのコア608を有する。一方のCPU 156は、別のCPU

156より高速のクロック速度を有することができる。一方のDRAM 216は、別のDRAM 216より多数のメガバイト、ギガバイト、もしくはテラバイト、またはより高速のアクセス時間を有することができる。これらの要因は、ハイブリッド・ブレード602およびコンピュート・ブレード604に影響する可能性がある。複数のハイブリッド・ブレード602を有するストレージ・クラスタ160に単一のコンピュート・ブレード604を追加することさえもが、システムの性能を高めることができ、複数のコンピュート・ブレード602の追加は、性能をさらに高めることができる。ストレージ・リソースとコンピュート・リソースとの両方を有する、ある個数のハイブリッド・ブレード602と、コンピュート専用ノードを有する、別の個数のコンピュート・ブレード604とを有する異種システムは、ハイブリッド・ブレード602だけを有する同種システムとは異な

10

って平衡化され得る。諸実施形態は、コンピューティング能力またはコンピューティング・リソースを同一のブレード604上でのストレージにささげる必要がないコンピュート専用ノード上のコンピューティング能力またはコンピューティング・リソースを利用することができる。処理制限されているストレージ・システムにあまりに多数のオーソリティ168を追加することが、性能を低下させる可能性が高いことに留意する価値がある。処理能力720をも追加する（たとえば、1つまたは複数のハイブリッド・ブレード602および/またはコンピュート・ブレード604を追加することによって）のと同時にオーソリティ168を追加する（たとえば、データのより多くの総量処理するために）ことは、性能を所与のレベルに保つようにシステムをスケールアップする。他の例が、たやすく考案される。

20

【0042】

[0051] 図7は、ハイブリッド・ブレード602およびコンピュート・ブレード604にまたがって外部I/O処理用のフロントフェーシング・ティア714、オーソリティ168用のオーソリティ・ティア716、およびストレージ・メモリ（たとえば、フラッシュ・メモリ206または他のタイプのストレージ・メモリ）用のストレージ・ティア718に分配される処理能力720を示す、図6のストレージ・システムの図である。処理能力720の分配とは、作業、処理タスク、コンピューティング、コンピューティング・タスク、処理アクティビティまたはコンピューティング・アクティビティ、I/O処理（外部または内部）、その他が、様々なティア714、716、718内のデバイスおよびプロセスに配置され、専用になれ、割り当てられ、そのために提供され、スケジューリングされ、割り振られ、使用可能になれ、その他が行われることを意味する。たとえば、フロントフェーシング・ティア714への処理能力720の分配は、フロントフェーシング・ティア714内のリソースが、処理能力720の一部を用いて外部I/O処理を実行できることを意味する。オーソリティ・ティア716への処理能力720の分配は、オーソリティ168が、処理能力720の一部を用いてオーソリティ168に固有の責務を実行できることを意味する。ストレージ・ティア718への処理能力720の分配は、ストレージ・ティア718内のデバイスおよびプロセスが、処理能力720の一部を用いてストレージ責務を実行できることを意味する。この例は様々なティアを議論するが、これは、限定的であることを意図されたものではない。というのは、この例が、例示のために利用される一例であるからである。処理能力は、プロセスまたはスレッドを特定のプロセッサに

30

またはその逆に割り当てることによって、プロセスまたはスレッドの優先順位を配置することによって、およびコンピューティング・システム内でたやすく案出されるさらなる形で、分配され得る。

40

【0043】

[0052] 複数のテナント702が、入出力要求を行っており、この入出力要求を、ストレージ・クラスタ160が、外部I/O処理704として処理し、サービスしつつある。様々なポリシおよび合意706、708、710が、ストレージ・システム内の所定の位置にある。システム内にコンピュート・ブレード604がない時にハイブリッド・ブレード602にまたがる、または両方のタイプのブレード602、604がシステム内に存在する時にハイブリッド・ブレード602およびコンピュート・ブレード604にまたがる

50

、処理能力720は、オペレーション・ティア712、たとえばフロントフェーシング・ティア714、オーソリティ・ティア716、およびストレージ・ティア718に分配される。これが、以下で説明するように様々な形、組合せ、およびシナリオにおいて発生する可能性があることを了解されたい。

【0044】

[0053] 外部I/Oに関するクライアントからの要求の受信専用であるフロントフェーシング・ティア714は、I/O要求を復号し、要求がどこに行くのか、すなわち、各要求がどのオーソリティ168に送られるべきなのかを見つけ出す。これは、様々な計算およびマップを必要とし、処理能力720の一部を要する。いくつかの実施形態では、外部I/O処理に関するクライアントからのI/O要求は、任意のストレージ・ノードすなわち、任意のハイブリッド・ブレード602または任意のコンピュート・ブレード604で受信される可能性がある。いくつかの実施形態では、I/O要求は、1つまたは複数の特定のブレード602、604にルーティングされ得る。この実施形態での外部I/O要求処理およびスループットは、フロントフェーシング・ティア714への処理能力720の分配に従って決定される。

10

【0045】

[0054] 次に、フロントフェーシング・ティア714から下って、オーソリティ・ティア716は、オーソリティ168が要求する様々なタスクを実行する。オーソリティ・ティア716でのシステムの挙動は、各オーソリティ168が仮想コントローラまたは仮想プロセッサであるかのようなものであり、これは、処理能力720のさらなる部分を要する。オーソリティ・ティア716へのおよびオーソリティ・ティア716内での処理能力720の分配は、様々な実施形態でオーソリティ168ごとにまたはブレード602、604ごとに行われ得、オーソリティ168にまたがって等しくまたは均等に分配されるか、所与のブレード602、604内のオーソリティ168にまたがって変化することができる。

20

【0046】

[0055] オーソリティ・ティア716の下のストレージ・ティア718は、ストレージ・メモリが責任を負うタスクの世話をし、これは、処理能力720の別の部分を要する。さらに、ストレージ・メモリ用のコンピューティング能力は、ストレージ・ユニット152のそれぞれで、たとえばコントローラ212から使用可能である。処理能力がティア714、716、718のそれぞれにどのように分配され、処理能力が所与のティア714、716、718内でどのように分配されるのかは、柔軟であり、ストレージ・クラスタ160によって、および/またはユーザ、たとえば管理者によって決定され得る。

30

【0047】

[0056] あるシナリオでは、当初にストレージ・クラスタ160内にハイブリッド・ブレード602だけがあり、1つまたは複数のコンピュート・ブレード604が、たとえばアップグレードまたは改善として、追加される。これは、システムが使用可能な処理能力720を増大させる。このシナリオでは、ストレージ・メモリの総量は変化しない(ストレージ・メモリを有するハイブリッド・ブレード602が追加されない)が、システム内のプロセッサの総数および処理能力720の総量は、コンピュート・ブレード604の追加の結果として増やされる。オーソリティ168は、どれほどの処理能力720が各ブレード602、604上で使用可能であるのかに従って、分配されまたは再分配され得る。たとえば、CPU 156のすべてが処理速度およびコア608の個数において同等である場合に、ブレード602、604のそれぞれは、等しい個数のオーソリティ168を受け取ることができる。いくつかの実施形態では、ハイブリッド・ブレード602であれコンピュート・ブレード604であれ、ブレードのうちの1つが、より強力なプロセッサ(すなわち、より多くの処理能力720)を有する場合に、そのブレード602、604は、より多数のオーソリティ168を割り当てられる可能性がある。オーソリティ168を分配する1つの形は、各ブレード602、604の処理能力720の相対的な量に比例して各ブレード602、604にオーソリティを割り当てまたは割り振ることである。

40

50

それを行うことは、各オーソリティ168が、そのオーソリティ168が存在するブレード602、604上でそのオーソリティ168からアクセス可能な処理能力720の同等な量を有するように、オーソリティ168にまたがって処理能力720を平衡化する。これは、ブレード602、604のうちの1つまたは複数に新しいオーソリティ168を追加すること、または1つもしくは複数のオーソリティ168をあるブレード602、604から別のブレード602、604に移動することのいずれかを必然的に伴う可能性がある。図6に示された実施形態および例を戻って参照すると、これは、1つまたは複数のコンピュータ・ブレード604がストレージ・クラスタに追加され、これが1つまたは複数のオーソリティ168の再配置をトリガする時にあてはまる可能性がある。

【0048】

[0057] 関連するシナリオでは、オーソリティは、ブレード602、604のそれぞれで使用可能な、DRAM206（または他のタイプのRAMもしくはメモリ）の量またはDRAM206の性能（たとえば、読取アクセス速度および書込アクセス速度）に比例して、様々なブレード602、604に割り当てられ、分配され、再分配され、または再配置される。より大量のDRAM206を有するブレード602、604は、より少量のDRAM206を有するブレード602、604より多数のオーソリティを受け取りまたは有するはずである。それを行うことは、オーソリティ168が存在するブレード602、604上でオーソリティ168からアクセス可能な同等な量のメモリとしての各オーソリティ168なるように、オーソリティ168にまたがってRAMを平衡化する。

【0049】

[0058] 別のシナリオでは、ストレージ・クラスタ160のブレード602、604のすべての処理能力720の総量の諸部分が、サービス品質(QoS)ポリシー706、サービス水準合意(service level agreement)708、サービス・クラスおよび/またはマルチテナント・サービス710に従ってオペレーション・ティア712に分配される。たとえば、ポリシー、合意、またはサービス・クラスが、より高いレベルの外部I/O処理、たとえば、データ・ストレージへのおよび/またはデータ・ストレージからあるテナント702、サービスのクラス、IPアドレスもしくはIPアドレスの範囲、データの範囲、データのタイプ、その他への、別のものより高いデータ・スループットを提供することである場合に、フロントフェーシング・ティア714および/またはオーソリティ・ティア716内で、他のテナント702、サービスのクラス、その他と比較して、より大量の処理能力720が、そのテナント702、サービスのクラス、その他に割り振られる。外部I/O処理のコンピューティング・タスクは、ブレード602、604にまたがって分配され得、その結果、各テナント702、サービスのクラス、その他のI/O処理が、1つまたは複数のストレージ・ノードに割り当てられるようになり、これらのストレージ・ノードは、個々のテナント、クライアント、アプリケーション、サービス・クラス、その他に基づいて様々な組合せのハイブリッド・ブレード602および/またはコンピュータ・ブレード604上に存在することができる。アプリケーション・レイヤ内のアプリケーション（すなわち、ストレージ・クラスタ160を動作させるソフトウェアとは別個のアプリケーション・ソフトウェア）のコンピューティング・タスクは、様々な組合せの、個々のアプリケーションまたはアプリケーションのグループと個々のブレードまたはブレード602、604のグループとに基づいて、ブレード602、604のうちの1つまたは複数にまたがって分配され得る。たとえば、アプリケーションのあるセットのコンピューティング・タスクが、ブレード602、604のあるグループに割り当てられる可能性があり、アプリケーションの別のセットのコンピューティング・タスクが、ブレード602、604の別のグループに割り当てられる可能性があり、これらが、オーバーラップするグループまたはオーバーラップしないグループである可能性がある。テナント702、サービスのクラス、その他に属するデータのinodeのオーソリティ168は、たとえばオーソリティ168を適当に移動することによって、他のテナント702、サービスのクラス、その他と比較したクロック周波数またはプロセッサ速度によって重みを付けられて、プロセッサ・コア608のより大きい比率を割り当てられ得る。

10

20

30

40

50

システムは、たとえば起動されるスレッドの個数を制御することまたは優先順位に関してスレッドに重みを付けることによって、ストレージ・メモリに対して、オーソリティ 168 のためにどれほどの処理能力 720 が使用可能であるのかを平衡化することができる。テナント 702 およびサービス品質（たとえば、スループット、待ち時間、応答性）に対して保証されるストレージの量は、直交的に（すなわち、独立に）、定常的に、弾力的に、または需要に基づいて、システム内で調整され得る。いくつかの実施形態では、ヒューリスティックが、システムの上記および他の態様の測定および調整に適用され得る。ポリシー、合意、またはテナントに対する変更は、1つまたは複数のオーソリティ 168 の再配置をトリガすることもできる。

【0050】

[0059] 図 8 は、ストレージ・システム内で処理能力を管理する方法の流れ図である。この方法は、本明細書で説明されるストレージ・クラスタおよびストレージ・ノードの様々な実施形態上またはそれによって実践され得る。この方法の様々なステップは、ストレージ・クラスタ内のプロセッサまたはストレージ・ノード内のプロセッサなどのプロセッサによって実行され得る。この方法の一部またはすべては、ソフトウェア、ハードウェア、ファームウェア、またはその組合せで実施され得る。この方法は、アクション 802 で開始し、アクション 802 では、ハイブリッド・ブレードが提供される。各ハイブリッド・ブレードは、ストレージ・ノードおよびストレージ・メモリを含む。アクション 804 では、コンピュート・ブレードが提供される。各コンピュート・ブレードは、コンピュート専用ノードを含み、ストレージ・メモリを含まない。コンピュート・ブレードは、D R A M または他の R A M などのシステム・メモリを有するが、コンピュート・ブレード自体の上にソリッドステート・ストレージ・メモリまたはディスクベースのストレージ・メモリを有しない。アクション 806 では、オーソリティがブレードにまたがって分配される。すなわち、オーソリティは、ハイブリッド・ブレードおよびコンピュート・ブレード上に配置され、これらに移動され、またはこれらの上で他の形で確立される。

【0051】

[0060] アクション 808 では、処理能力が、上で言及した合意および/またはポリシーに従って、ブレードにまたがってティアに、たとえば、フロントフェーシング・ティア、オーソリティ・ティア、およびストレージ・ティアに分配される。合意またはポリシーに含まれるものに応じて、処理能力は、固定された量または可変量でティアのそれぞれに分配され得る。この実施形態では、フロントフェーシング・ティアは、外部 I/O 要求をサービスする、すなわち、外部 I/O 処理用であり、オーソリティ・ティアは、オーソリティへのサービス用であり、ストレージ・ティアは、ストレージ・メモリへのサービス用である。

【0052】

[0061] アクション 810 では、外部 I/O 処理が、フロントフェーシング・ティア内で実行され、内部 I/O 処理（すなわち、ストレージ・システム内の様々なリソースに関する内部 I/O 動作の処理）が、オーソリティ・ティアおよびストレージ・ティア内で実行される。判断アクション 812 では、処理能力をオーソリティ・ティア内で再分配すべきかどうかという質問が尋ねられる。回答が否定である場合には、処理能力をオーソリティ・ティア内で再分配する必要はなく、流れはアクション 810 に戻って分岐して、外部および内部の I/O 処理の実行を継続する。回答が肯定である場合には、処理能力をオーソリティ・ティア内で再分配しなければならず、流れはアクション 814 に進む。これは、たとえば、ストレージ・クラスタへのコンピュート・ブレードの挿入、ハイブリッド・ブレードの挿入、またはポリシー、合意、もしくはマルチテナント・サービスに対する変更によってトリガされ得る。

【0053】

[0062] アクション 814 では、オーソリティが、あるブレードから別のブレードに移動される（たとえば、ハイブリッド・ブレードからコンピュート・ブレードに、ハイブリッド・ブレードから別のハイブリッド・ブレードに、コンピュート・ブレードから別のコ

10

20

30

40

50

ンピュータ・ブレードに、またはコンピュータ・ブレードからハイブリッド・ブレードに
さえ)。いくつかの実施形態では、複数のオーソリティがブレードの間で移動される。変
形形態では、ティアの間またはティアのうちの1つの中の処理能力が、合意もしくはポリ
シの変化またはブレードの挿入に応答して再分配され得る。その後、流れはアクション 8
10に戻って、外部および内部のI/O処理の実行を継続する。変形形態では、流れは、
他所に進んでさらなるアクションを実行することができる。

【0054】

[0063] 図9は、いくつかの実施形態によるストレージ・クラスタ、ストレージ・ノ
ード、および/または不揮発性ソリッド・ステート・ストレージの実施形態上ではまたはそれ
によって実践され得る、ブレードの追加時にストレージ・システム内で処理能力を管理す
る方法の流れ図である。この方法は、図8の方法に関し、変形形態では、図8を参照して
説明した方法と組み合わせられまたはその諸部分を置換することができる。この方法は、判
断アクション902で始まり、判断アクション902では、ストレージ・システムにブレ
ードを追加すべきかどうか判定される。回答が否定である場合には、ブレードが追加さ
れてはならないか追加されず、いくつかの実施形態では、この方法は、ある時間期間だけ
待機し、ブレードが追加されるべきかどうかをチェックする。回答が肯定である場合には
、ブレードが追加され、流れは判断アクション904に進む。判断アクション904では
、新たに追加されるブレードがオーソリティ・ティアに参加するかどうか判定される。
たとえば、新たに追加されるブレードが、コンピュータ専用ノードを有するが、オーソリ
ティ・ティアに参加しないことが判断され得る。あるいは、新たに追加されるブレードが
、コンピュータ専用ストレージ・ノードを有し、オーソリティ・ティアへの参加としてオー
ソリティによってまたはオーソリティの代わりに実行されるアクションに参加すること
が判断され得る。回答が否定である場合には、新たに追加されるブレードは、オーソリ
ティ・ティアに参加せず、流れは、判断アクション902に進む。回答が肯定である場合に
は、新たに追加されるブレードは、オーソリティ・ティアに参加し、流れは、判断アクシ
ョン906に進む。

10

20

【0055】

[0064] 判断アクション906では、新しいオーソリティを新たに追加されたブレード
に移動すべきかどうか判定される。回答が肯定である場合には、流れはアクション90
8に進み、アクション908では、新しいオーソリティが、新たに追加されたブレードに
移動されまたは追加される。上で言及したように、1つまたは複数のブレードから1つま
たは複数のさらなるブレードへのオーソリティの移動は、望み通りに処理能力を再分配す
る。流れは、アクション902に戻って分岐して、ブレードが追加されつつありまたは追
加されるのかどうかを調べる。変形形態では、流れは、他所に分岐してさらなるタスクを
実行することができる。判断アクション904が、新たに追加されるブレードがオーソリ
ティ・ティアに参加しないと判定した場合には、新たに追加されるブレードは、移動され
るオーソリティのいずれの受取からも除外され得、あるいは、判断が再訪問され得、この
場合には、新たに追加されるブレードは、移動されるオーソリティのうちの1つまたは複
数を受け取ることができる。オーソリティが移動された後に、流れは、アクション902
に戻って進み、あるいは、変形形態では、他所に分岐してさらなるタスクを実行する。

30

40

【0056】

[0065] 本明細書で説明される方法が、従来の汎用コンピュータ・システムなどのデジ
タル処理システムを用いて実行され得ることを了解されたい。代替案では、1つの機能だ
けを実行するように設計されまたはプログラムされる特殊目的コンピュータが使用され得
る。図10は、本明細書で説明される実施形態を実施することのできる例示的なコンピ
ューティング・デバイスを示す図である。図10のコンピューティング・デバイスは、いく
つかの実施形態によるストレージ・システム内で処理能力を管理する機能性の実施形態を
実行するのに使用され得る。このコンピューティング・デバイスは、バス1005を介し
てメモリ1003およびマス・ストレージ・デバイス1007に結合される中央処理装置
(CPU)1001を含む。マス・ストレージ・デバイス1007は、いくつかの実施形

50

態でローカルまたはリモートとすることができる、フロッピ・ディスク・ドライブまたは固定ディスク・ドライブなどの永続データ・ストレージ・デバイスを表す。メモリ 1003 は、読取専用メモリ、ランダム・アクセス・メモリ、その他を含むことができる。コンピューティング・デバイス上に存在するアプリケーションは、いくつかの実施形態で、メモリ 1003 またはマス・ストレージ・デバイス 1007 などのコンピュータ可読媒体上に記憶されまたはこれを介してアクセスされ得る。アプリケーションは、コンピューティング・デバイスのネットワーク・モデムまたは他のネットワーク・インターフェースを介してアクセスされる変調された変調された電子信号の形であるものとする。CPU 1001 が、いくつかの実施形態で、汎用プロセッサ、専用プロセッサ、または特別にプログラムされた論理デバイス内で実施され得ることを了解されたい。

10

【0057】

[0066] ディスプレイ 1011 は、バス 1005 を介して CPU 1001、メモリ 1003、およびマス・ストレージ・デバイス 1007 と通信している。ディスプレイ 1011 は、本明細書で説明されるシステムに関連する任意の視覚化ツールまたはレポートを表示するように構成される。入出力デバイス 1009 は、コマンド選択内の情報を CPU 1001 に通信するためにバス 1005 に結合される。外部デバイスへおよび外部デバイスからのデータが、入出力デバイス 1009 を介して通信され得ることを了解されたい。CPU 1001 は、図 1 ~ 図 9 を参照して説明した機能性を使用可能にするために本明細書で説明する機能性を実行するように定義され得る。この機能性を実施するコードは、いくつかの実施形態で、CPU 1001 などのプロセッサによる実行のためにメモリ 1003 またはマス・ストレージ・デバイス 1007 内に記憶され得る。コンピューティング・デバイス上のオペレーティング・システムは、MS DOS (商標)、MS-WINDOWS (商標)、OS/2 (商標)、UNIX (商標)、LINUX (商標)、または他の既知のオペレーティング・システムとすることができる。本明細書で説明される実施形態が、物理コンピューティング・リソースを用いて実施される仮想化されたコンピューティング・システムと統合されることも可能であることを了解されたい。

20

【0058】

[0067] 詳細な例示の実施形態が、本明細書で開示される。しかし、本明細書で開示される特定の機能的詳細は、実施形態の説明において単に典型的なものである。しかし、諸実施形態は、多数の代替形態で実施され得、本明細書で示される実施形態のみに限定されると解釈してはならない。

30

【0059】

[0068] 用語第 1、第 2、その他が、様々なステップまたは計算を記述するために本明細書で使用される場合があるが、これらのステップまたは計算が、これらの用語によって限定されてはならないことを理解されたい。これらの用語は、あるステップまたは計算を別のステップまたは計算から区別するためにのみ使用される。たとえば、本開示の範囲から逸脱せずに、第 1 の計算が第 2 の計算と呼ばれ得、同様に、第 2 のステップが第 1 のステップと呼ばれ得る。本明細書で使用される時に、用語「および/または」「および」「/」記号は、関連するリストされた項目のうちの一つまたは複数の任意のすべての組合せを含む。

40

【0060】

[0069] 本明細書で使用される時に、単数形「a」、「an」、および「the」は、文脈がそうではないことを明瞭に示さない限り、複数形をも含むことが意図されている。用語「含む」(「comprises」、「comprising」、「includes」、および/または「including」) は、本明細書で使用される時に、述べられた特徴、整数、ステップ、動作、要素、および/または構成要素の存在を指定するが、一つまたは複数の他の特徴、整数、ステップ、動作、要素、構成要素、および/またはその群の存在または追加を除外しないことをさらに理解されたい。したがって、本明細書で使用される用語法は、特定の実施形態を説明するためのみのものであって、限定的であることは意図されていない。

50

【 0 0 6 1 】

[0070] いくつかの代替実施態様では、注記された機能／行為が、図に示された順序から外れて発生する可能性があることにも留意されたい。たとえば、関係する機能性／行為に依存して、連続して示される2つの図が、実際には実質的に同時に実行される場合があり、あるいは、時には逆の順序で実行される場合がある。

【 0 0 6 2 】

[0071] 上の実施形態を念頭において、諸実施形態が、コンピュータ・システム内に記憶されたデータを用いる様々なコンピュータ実施される動作を使用することができることを理解されたい。これらの動作は、物理量の物理的操作を必要とする動作である。必ずではないが通常、これらの量は、記憶され、転送され、組み合わせられ、比較され、他の形で操作されることが可能な電気信号または磁気信号の形をとる。さらに、実行される操作は、しばしば、作る、識別する、判定する、または比較するなどの言葉で参照される。諸実施形態の一部を形成する、本明細書で説明される動作のいずれもが、有用な機械動作である。諸実施形態は、これらの動作を実行するためのデバイスまたは装置にも関する。装置は、要求される目的のために特に構成され得、あるいは、装置は、コンピュータ内に記憶されたコンピュータ・プログラムによって選択的にアクティブ化されまたは構成される汎用コンピュータとされ得る。具体的には、様々な汎用マシンが、本明細書の教示に従って記述されたコンピュータ・プログラムと共に使用され得、あるいは、要求される動作を実行するように、より特殊化された装置を構成することが、より便利である場合がある。

10

【 0 0 6 3 】

[0072] モジュール、アプリケーション、レイヤ、エージェント、または他の方法実施可能なエンティティは、ハードウェア、ファームウェア、もしくはソフトウェアを実行するプロセッサ、またはその組合せとして実施され得る。ソフトウェアベースの実施形態が本明細書で開示される場合に、そのソフトウェアが、コントローラなどの物理機械内で実施され得ることを了解されたい。たとえば、コントローラが、第1のモジュールおよび第2のモジュールを含む可能性がある。コントローラは、たとえば方法、アプリケーション、レイヤ、またはエージェントの、様々なアクションを実行するように構成され得る。

20

【 0 0 6 4 】

[0073] 諸実施形態は、非一時的コンピュータ可読媒体上のコンピュータ可読コードとして実施されることも可能である。コンピュータ可読媒体は、その後コンピュータ・システムによって読み取られ得るデータを記憶することのできる任意のデータ・ストレージ・デバイスである。コンピュータ可読媒体の例は、ハード・ドライブ、ネットワーク・アタッチト・ストレージ(NAS)、読取専用メモリ、ランダム・アクセス・メモリ、CD-ROM、CD-R、CD-RW、磁気テープ、ならびに他の光学的および非光学的なデータ・ストレージ・デバイスを含む。コンピュータ可読媒体は、コンピュータ可読コードが分散された形で記憶され、実行されるようにするために、ネットワーク結合されたコンピュータ・システムを介して分散もされ得る。本明細書で説明される実施形態は、ハンドヘルド・デバイス、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またはプログラム可能な消費者エレクトロニクス、ミニコンピュータ、メインフレーム・コンピュータ、および類似物を含む様々なコンピュータ・システム構成を用いて実践され得る。諸実施形態は、有線ベースのネットワークまたは無線ネットワークを介してリンクされたりリモート処理デバイスによってタスクが実行される分散コンピューティング環境内でも実践され得る。

30

40

【 0 0 6 5 】

[0074] 方法動作が、特定の順序で説明されたが、他の動作が、説明された動作の間に実行され得、説明された動作が、それらがわずかに異なる時に行われるようにするために調整され得、あるいは、説明された動作が、処理に関連する様々なインターバルでの処理動作の発生を可能にするシステム内で分散され得ることを理解されたい。

【 0 0 6 6 】

[0075] 様々な実施形態では、本明細書で説明される方法および機構の1つまたは複数

50

の部分、クラウドコンピューティング環境の一部を形成することができる。そのような実施形態では、リソースは、1つまたは複数の様々なモデルに従うサービスとしてインターネットを介して提供され得る。そのようなモデルは、インフラストラクチャ・アズ・ア・サービス (IaaS)、プラットフォーム・アズ・ア・サービス (PaaS)、およびソフトウェア・アズ・ア・サービス (SaaS) を含むことができる。IaaSでは、コンピュータ・インフラストラクチャが、サービスとして配信される。その場合に、コンピューティング機器は、一般に、サービス・プロバイダによって所有され、運営される。PaaSモデルでは、ソフトウェア・ソリューションを開発するために開発者によって使用されるソフトウェア・ツールおよび基礎になる機器が、サービス・プロバイダによってサービスとして提供され、ホスティングされ得る。SaaSは、通常、サービス・プロバイダがサービスとしてオン・デマンドでソフトウェアのライセンスを発行することを含む。サービス・プロバイダは、ソフトウェアをホスティングすることができ、あるいは、所与の時間期間の間に顧客にソフトウェアを展開することができる。上記モデルの多数の組合せが、可能であり、企図されている。

10

20

30

40

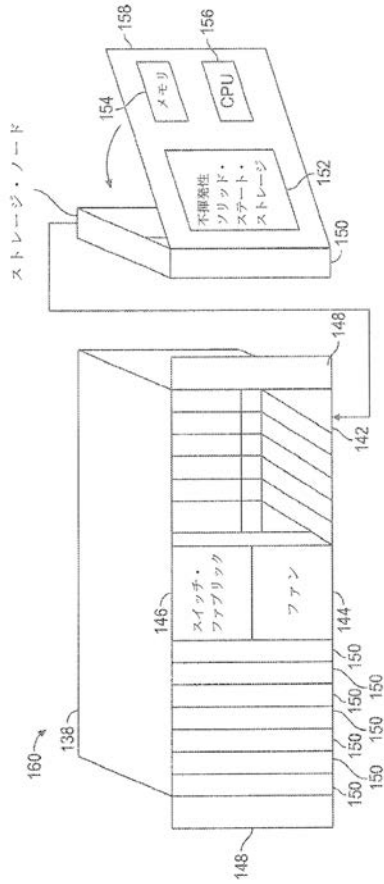
【0067】

[0076] 様々なユニット、回路、または他の構成要素が、1つまたは複数のタスクを実行する「ように構成される」ものとして説明され、または請求される場合がある。そのような文脈では、句「ように構成される」は、ユニット/回路/構成要素が、動作中に1つまたは複数のタスクを実行する構造 (たとえば、回路網) を含むことを示すことによって、構造を暗示するのに使用される。したがって、ユニット/回路/構成要素は、指定されたユニット/回路/構成要素が現在は動作していない (すなわち、オンではない) 時であってもタスクを実行するように構成されると言われ得る。「ように構成される」という言語と共に使用されるユニット/回路/構成要素は、ハードウェア、たとえば、回路、動作を実施するために実行可能なプログラム命令を記憶するメモリ、その他を含む。ユニット/回路/構成要素が1つまたは複数のタスクを実行する「ように構成される」と具陳することは、そのユニット/回路/構成要素に関して米国特許法第112条第6段落に頼らないことが特に意図されている。さらに、「ように構成される」は、問題のタスク (1つまたは複数) を実行することのできる形で動作するためにソフトウェアおよび/またはファームウェアによって操作される包括的構造 (たとえば、包括的回路網) (たとえば、ソフトウェアを実行するFPGAまたは汎用プロセッサ) を含むことができる。「ように構成される」は、1つまたは複数のタスクを実施しまたは実行するように適合されたデバイス (たとえば、集積回路) を製造するために製造プロセス (たとえば、半導体製造施設) を適合させることをも含むことができる。

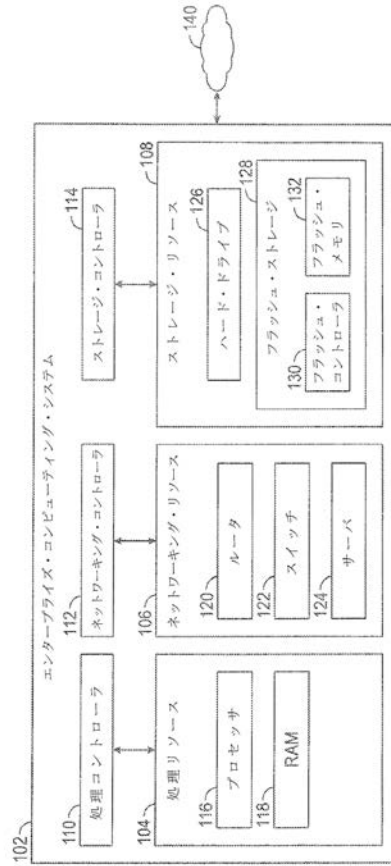
【0068】

[0077] 前述の記述は、説明のために、特定の実施形態を参照して記述された。しかし、上の例示的な議論は、網羅的であることまたは開示された正確な形態に本発明を限定することを意図されたものではない。多数の変更および変形が、上記教示に鑑みて可能である。諸実施形態は、諸実施形態の原理およびその実用的応用を最もよく説明し、これによって、企図される特定の使用に適するものになることができるものとして諸実施形態および様々な変更を当業者が最もよく利用することを可能にするために選択され、説明された。したがって、本実施形態は、制限的ではなく例示的と考えられなければならない、本発明は、本明細書で与えられる詳細に限定されてはならず、添付の特許請求の範囲の範囲および同等物の中で変更され得る。

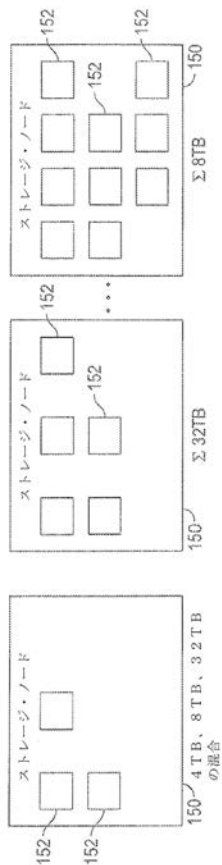
【 図 1 】



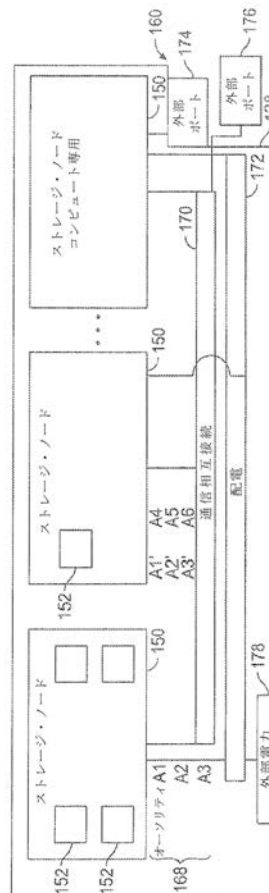
【 図 2 】



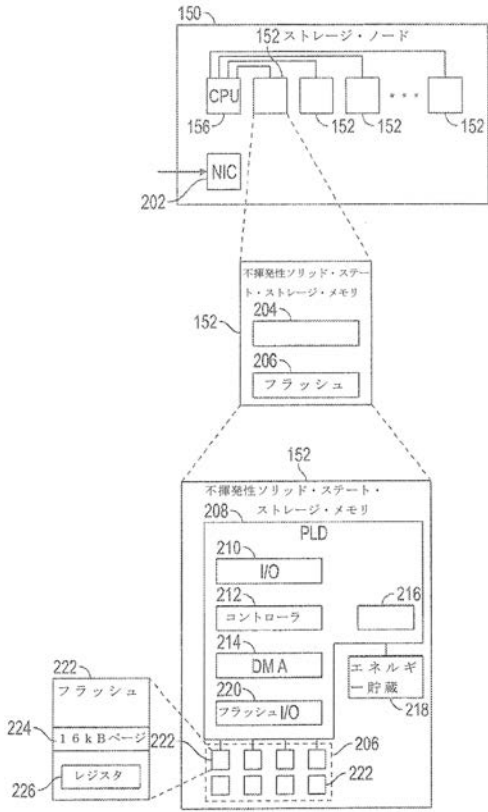
【 図 3 】



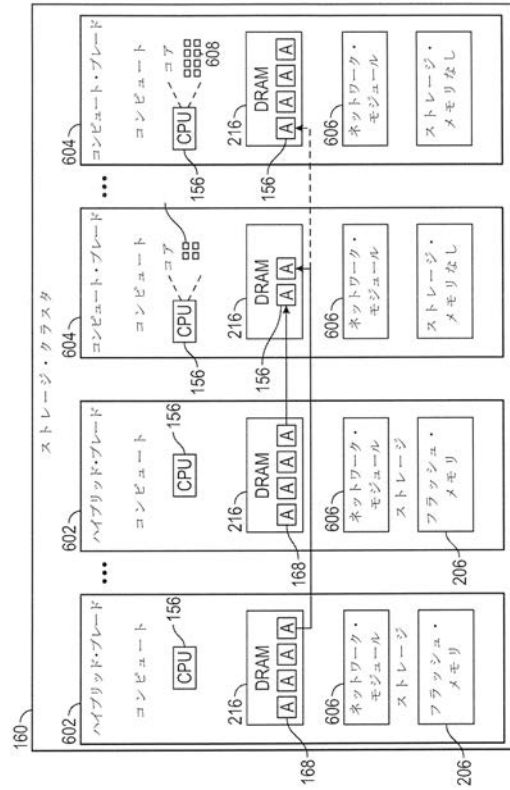
【 図 4 】



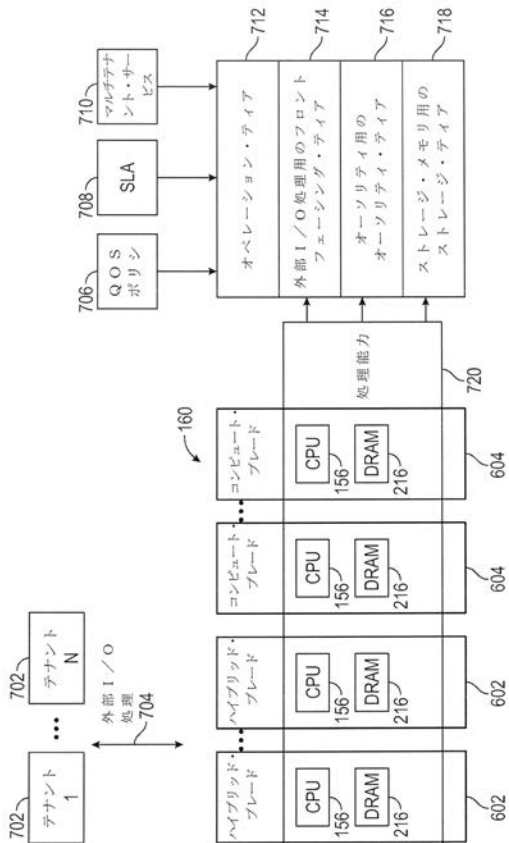
【図5】



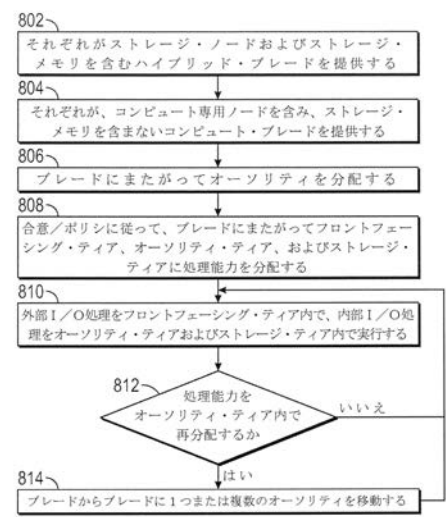
【図6】



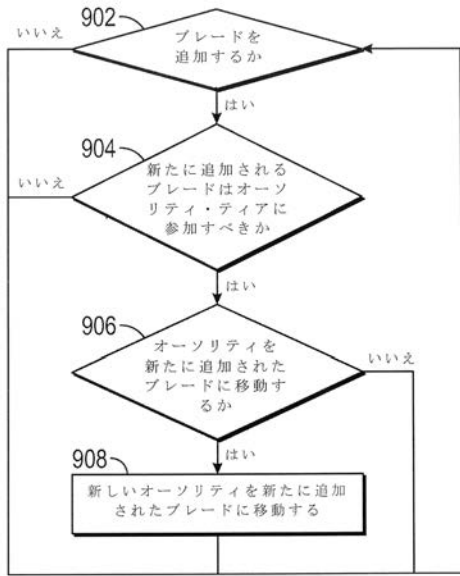
【図7】



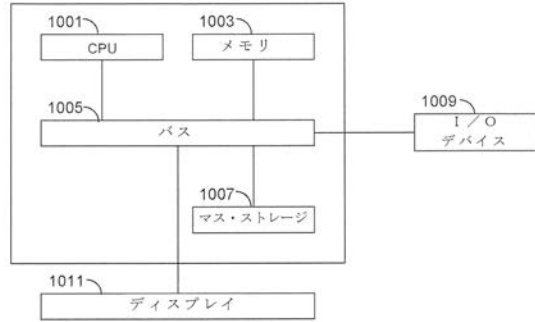
【図8】





【図9】



【図10】



【 国際調査報告 】

INTERNATIONAL SEARCH REPORT		International application No. PCT/US2017/031162
A. CLASSIFICATION OF SUBJECT MATTER G06F 3/06(2006.01)i		
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 3/06; H04L 29/08; G06F 17/30; G06F 11/14; G06F 17/60; G06F 12/00; G06F 9/50		
Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched Korean utility models and applications for utility models Japanese utility models and applications for utility models		
Electronic data base consulted during the international search (name of data base and, where practicable, search terms used) eKOMPASS(KIPO internal) & Keywords: storage, authority, ownership, distribute, balance, move		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2015-0363424 A1 (THE BOEING COMPANY) 17 December 2015 See paragraphs [0008]-[0010], [0043], [0056], [0065], [0109]; and figure 3.	1-20
Y	US 2015-0355974 A1 (PURE STORAGE, INC.) 10 December 2015 See paragraphs [0013], [0016]-[0017], [0024], [0036], [0041]; and figure 4.	1-20
Y	US 2014-0068627 A1 (SILICON GRAPHICS INTERNATIONAL CORP.) 06 March 2014 See paragraphs [0010], [0013], [0026], [0035]; and figure 1.	6-7, 13, 19-20
A	US 2011-0231602 A1 (HAROLD WOODS et al.) 22 September 2011 See paragraphs [0010], [0025]; and figure 3.	1-20
A	WO 2005-048159 A1 (DATIC SYSTEMS INCORPORATED) 26 May 2005 See paragraphs [0004], [0034]; and figure 2.	1-20
<input type="checkbox"/> Further documents are listed in the continuation of Box C. <input checked="" type="checkbox"/> 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 application or patent 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 17 August 2017 (17.08.2017)		Date of mailing of the international search report 17 August 2017 (17.08.2017)
Name and mailing address of the ISA/KR  International Application Division Korean Intellectual Property Office 189 Cheongsu-ro, Seo-gu, Daejeon, 35208, Republic of Korea Facsimile No. +82-42-481-8578		Authorized officer KANG, Hee Gok  Telephone No. +82-42-481-8264

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/US2017/031162

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2015-0363424 A1	17/12/2015	US 9569461 B2	14/02/2017
US 2015-0355974 A1	10/12/2015	AU 2015-269360 A1 EP 3152686 A1 US 9218244 B1 WO 2015-188004 A1	22/12/2016 12/04/2017 22/12/2015 10/12/2015
US 2014-0068627 A1	06/03/2014	US 2016-0335131 A1 US 9424098 B2	17/11/2016 23/08/2016
US 2011-0231602 A1	22/09/2011	None	
WO 2005-048159 A1	26/05/2005	AU 2004-290049 A1 AU 2004-290049 B2 CA 2545359 A1 CA 2545359 C CN 101018808 A CN 101018808 B EP 1687772 A1 EP 1692175 A2 EP 2357238 A2 EP 2357238 A3 IL 175503 A JP 2007-513609 A JP 2012-050446 A JP 4979385 B2 US 2007-0071719 A1 US 2011-0190476 A1 US 2012-0183601 A1 US 2014-0235829 A1 US 7560265 B2 US 8138312 B2 US 8735358 B2 WO 2005-047478 A2 WO 2005-047478 A3	26/05/2005 29/09/2011 26/05/2005 06/11/2012 15/08/2007 23/07/2014 09/08/2006 23/08/2006 17/08/2011 01/02/2012 05/09/2006 31/05/2007 15/03/2012 18/07/2012 29/03/2007 04/08/2011 19/07/2012 21/08/2014 14/07/2009 20/03/2012 27/05/2014 26/05/2005 01/06/2006

フロントページの続き

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

(74)代理人 100126480

弁理士 佐藤 睦

(72)発明者 コルグローブ, ジョン

アメリカ合衆国, カリフォルニア州 9 4 3 0 3, マウンテン ビュー, カストロ ストリート
6 5 0, スイート 2 6 0

(72)発明者 デイビス, ジョン ディー.

アメリカ合衆国, カリフォルニア州 9 4 3 0 3, マウンテン ビュー, カストロ ストリート
6 5 0, スイート 2 6 0

(72)発明者 ヘイズ, ジョン マーティン

アメリカ合衆国, カリフォルニア州 9 4 3 0 3, マウンテン ビュー, カストロ ストリート
6 5 0, スイート 2 6 0

(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 9 4 3 0 3, マウンテン ビュー, カストロ ストリート
6 5 0, スイート 2 6 0