

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第6488296号
(P6488296)

(45) 発行日 平成31年3月20日(2019.3.20)

(24) 登録日 平成31年3月1日(2019.3.1)

(51) Int.Cl.	F I
GO 6 F 12/00 (2006.01)	GO 6 F 12/00 5 4 5 A
GO 6 F 13/10 (2006.01)	GO 6 F 12/00 5 1 4 E
	GO 6 F 13/10 3 4 0 A

請求項の数 23 (全 19 頁)

(21) 出願番号	特願2016-529768 (P2016-529768)	(73) 特許権者	510149482
(86) (22) 出願日	平成26年6月10日 (2014.6.10)		ヴィエムウェア インコーポレイテッド
(65) 公表番号	特表2016-530619 (P2016-530619A)		VMware, Inc.
(43) 公表日	平成28年9月29日 (2016.9.29)		アメリカ合衆国 94304 カリフォル
(86) 国際出願番号	PCT/US2014/041761		ニア州 パロ アルト ヒルビュー アベ
(87) 国際公開番号	W02015/030901		ニュー 3401
(87) 国際公開日	平成27年3月5日 (2015.3.5)	(74) 代理人	100105957
審査請求日	平成28年1月21日 (2016.1.21)		弁理士 恩田 誠
審査番号	不服2017-15960 (P2017-15960/J1)	(74) 代理人	100068755
審査請求日	平成29年10月27日 (2017.10.27)		弁理士 恩田 博宣
(31) 優先権主張番号	14/010, 293	(74) 代理人	100142907
(32) 優先日	平成25年8月26日 (2013.8.26)		弁理士 本田 淳
(33) 優先権主張国	米国 (US)		

最終頁に続く

(54) 【発明の名称】 スケーラブル分散ストレージアーキテクチャ

(57) 【特許請求の範囲】

【請求項 1】

コンピュータシステムにおいてオブジェクトへの複数のクライアントによるアクセスを提供する方法であって、

オブジェクトを格納するオブジェクトストアを提供することであって、前記オブジェクトストアは、格納されたオブジェクトに一意識別子をマッピングするフラットな名前空間を使用して、オブジェクトを内部的に追跡し、前記オブジェクトストアは、複数のクライアントをホストする複数のホストコンピュータに収容されるか、または直接取り付けられた複数の物理ストレージデバイスによって支持され、前記複数の物理ストレージデバイスは、第1のホストコンピュータに収容されるか、または直接取り付けられた第1の物理ストレージデバイスと、第2のホストコンピュータに収容されるか、または直接取り付けられた第2の物理ストレージデバイスとを含み、少なくとも1つの格納されたオブジェクトは、前記第1の物理ストレージデバイスおよび前記第2の物理ストレージデバイスの両方によって支持されている、前記オブジェクトストアを提供することと、

前記オブジェクトストアに格納されたオブジェクトに対する複数のクライアントによる同時アクセスをサポートするための前記オブジェクトストア用のファイルシステムインタフェースを提供することであって、前記オブジェクトストアによって内部的に使用される前記フラットな名前空間上に階層化された階層名前空間に対するアブストラクションを複数のクライアントに公開することを含み、前記アブストラクションは、前記階層名前空間へのルートディレクトリおよび1つまたは複数のサブディレクトリを含む、前記ファイル

システムインタフェースを提供することと、

前記ルートディレクトリの前記サブディレクトリとして現れる複数のトップレベルオブジェクトの作成を可能とすることであって、各トップレベルオブジェクトは、指定のクラスタ化ファイルシステムに従って編成された別個の名前空間を有するストレージデバイスの別個のアブストラクションを表し、少なくとも1つの前記トップレベルオブジェクトは、特定のクラスタ化ファイルシステムのファイルシステムオブジェクトを含む、複数のトップレベルオブジェクトの作成を可能とすることと、

前記クラスタ化ファイルシステム内にストアされたファイルへのアクセスを前記クライアントが試みたときに、前記トップレベルオブジェクトを、前記ルートディレクトリの前記アブストラクションのローカルバージョンでのサブディレクトリとして自動マウントすることと

10

を備える方法。

【請求項2】

前記トップレベルオブジェクトは、前記クラスタ化ファイルシステム内にストアされたファイルを表すファイルオブジェクトのためのメタデータをストアするよう構成され、前記ファイルオブジェクトは、前記トップレベルオブジェクトとは別個の前記オブジェクトストア内にストアされる、請求項1に記載の方法。

【請求項3】

ファイルオブジェクトは、前記ファイルの別個の部分を表し、かつ前記ファイルオブジェクトとは別個の前記オブジェクトストア内にストアされた対応するコンポーネントオブジェクトの位置を発見するために前記フラットな名前空間により使用可能な一意識別子をストアするよう構成される、請求項2に記載の方法。

20

【請求項4】

ファイルオブジェクトのために前記トップレベルオブジェクト内にストアされた前記メタデータは少なくとも、前記オブジェクトストア内の前記ファイルオブジェクトの位置を発見するために前記フラットな名前空間により使用可能な一意識別子を備えるファイル記述子を含む、請求項2に記載の方法。

【請求項5】

前記ファイルオブジェクトにより表される前記ファイルは、前記複数のホストコンピュータのうちの1つで実行中の仮想マシンにより使用される仮想ディスクファイルである、請求項4に記載の方法。

30

【請求項6】

前記トップレベルオブジェクトは、前記複数のホストコンピュータのうちのいずれかでインスタンス化されることが可能な複数の仮想マシンのうちの1つによってのみ利用されるファイルオブジェクトのためのメタデータをストアするよう構成される、請求項5に記載の方法。

【請求項7】

前記クラスタ化ファイルシステム内にストアされたファイルが全く前記クライアントによりアクセスされない期間の後、前記トップレベルオブジェクトを自動的にアンマウントすることをさらに備える、請求項1に記載の方法。

40

【請求項8】

前記フラットな名前空間は、メモリ内データベースとして実装され、メモリ内データベースの同期化コピーが前記ホストコンピュータの各々のメモリ内にストアされる、請求項1に記載の方法。

【請求項9】

前記フラットな名前空間が、オブジェクトについて自動生成された一意識別子とのコンフリクトを報告するときに、前記オブジェクトについてのアドミニストレータ特定名称を、前記オブジェクトについての一意識別子に変換することをさらに備える、請求項1に記載の方法。

【請求項10】

50

プロセッサ上で実行されたときに、オブジェクトへの複数のクライアントによるアクセスを提供するオペレーションを実行する命令をストアする、非一時的コンピュータ可読ストレージ媒体であって、前記オペレーションが、

オブジェクトを格納するオブジェクトストアを提供することであって、前記オブジェクトストアは、格納されたオブジェクトに一意識別子をマッピングするフラットな名前空間を使用して、オブジェクトを内部的に追跡し、前記オブジェクトストアは、複数のクライアントをホストする複数のホストコンピュータに収容されるか、または直接取り付けられた複数の物理ストレージデバイスによって支持され、前記複数の物理ストレージデバイスは、第1のホストコンピュータに収容されるか、または直接取り付けられた第1の物理ストレージデバイスと、第2のホストコンピュータに収容されるか、または直接取り付けられた第2の物理ストレージデバイスとを含み、少なくとも1つの格納されたオブジェクトは、前記第1の物理ストレージデバイスおよび前記第2の物理ストレージデバイスの両方によって支持されている、前記オブジェクトストアを提供することと、

前記オブジェクトストアに格納されたオブジェクトに対する複数のクライアントによる同時アクセスをサポートするための前記オブジェクトストア用のファイルシステムインタフェースを提供することであって、前記オブジェクトストアによって内部的に使用される前記フラットな名前空間上に階層化された階層名前空間に対するアブストラクションを複数のクライアントに公開することを含み、前記アブストラクションは、前記階層名前空間へのルートディレクトリおよび1つまたは複数のサブディレクトリを含む、前記ファイルシステムインタフェースを提供することと、

前記ルートディレクトリの前記サブディレクトリとして現れる複数のトップレベルオブジェクトの作成を可能とすることであって、各トップレベルオブジェクトは、指定のクラスタ化ファイルシステムに従って編成された別個の名前空間を有するストレージデバイスの別個のアブストラクションを表し、少なくとも1つの前記トップレベルオブジェクトは、特定のクラスタ化ファイルシステムのファイルシステムオブジェクトを含む、複数のトップレベルオブジェクトの作成を可能とすることと、

前記オペレーションは、前記クラスタ化ファイルシステム内にストアされたファイルへのアクセスを前記クライアントが試みたときに、前記トップレベルオブジェクトを、前記ルートディレクトリの前記アブストラクションのローカルバージョンでのサブディレクトリとして自動マウントすることと、

を備える、非一時的コンピュータ可読ストレージ媒体。

【請求項11】

前記トップレベルオブジェクトは、前記クラスタ化ファイルシステム内にストアされたファイルを表すファイルオブジェクトのためのメタデータをストアするよう構成され、前記ファイルオブジェクトは、前記トップレベルオブジェクトとは別個の前記オブジェクトストア内にストアされる、請求項10に記載のコンピュータ可読ストレージ媒体。

【請求項12】

ファイルオブジェクトは、前記ファイルの別個の部分を表し、前記ファイルオブジェクトとは別個の前記オブジェクトストア内にストアされた対応するコンポーネントオブジェクトの位置を発見するために前記フラットな名前空間により使用可能な一意識別子をストアするよう構成される、請求項11に記載のコンピュータ可読ストレージ媒体。

【請求項13】

ファイルオブジェクトのために前記トップレベルオブジェクト内にストアされた前記メタデータは少なくとも、前記オブジェクトストア内の前記ファイルオブジェクトの位置を発見するために前記フラットな名前空間により使用可能な一意識別子を備えるファイル記述子を含む、請求項11に記載のコンピュータ可読ストレージ媒体。

【請求項14】

前記ファイルオブジェクトにより表される前記ファイルは、前記複数のホストコンピュータのうちの1つで実行中の仮想マシンにより使用される仮想ディスクファイルである、請求項11に記載のコンピュータ可読ストレージ媒体。

【請求項 15】

前記トップレベルオブジェクトは、前記複数のホストコンピュータのうちのいずれかでインスタンス化されることが可能な複数の仮想マシンのうちの1つによってのみ利用されるファイルオブジェクトのためのメタデータをストアするよう構成される、請求項14に記載のコンピュータ可読ストレージ媒体。

【請求項 16】

前記フラットな名前空間は、メモリ内データベースとして実装され、メモリ内データベースの同期化コピーが前記ホストコンピュータの各々のメモリ内にストアされる、請求項10に記載のコンピュータ可読ストレージ媒体。

【請求項 17】

プロセッサと、
モジュールをホストするメモリであって、前記プロセッサ上で実行されたときに、オブジェクトへの複数のクライアントによるアクセスを提供するオペレーションを実行するメモリと

を備えるコンピュータシステムであって、前記オペレーションが、

オブジェクトを格納するオブジェクトストアを提供することであって、前記オブジェクトストアは、格納されたオブジェクトに一意識別子をマッピングするフラットな名前空間を使用して、オブジェクトを内部的に追跡し、前記オブジェクトストアは、複数のクライアントをホストする複数のホストコンピュータに収容されるか、または直接取り付けられた複数の物理ストレージデバイスによって支持され、前記複数の物理ストレージデバイスは、第1のホストコンピュータに収容されるか、または直接取り付けられた第1の物理ストレージデバイスと、第2のホストコンピュータに収容されるか、または直接取り付けられた第2の物理ストレージデバイスとを含み、少なくとも1つの格納されたオブジェクトは、前記第1の物理ストレージデバイスおよび前記第2の物理ストレージデバイスの両方によって支持されている、前記オブジェクトストアを提供することと、

前記オブジェクトストアに格納されたオブジェクトに対する複数のクライアントによる同時アクセスをサポートするための前記オブジェクトストア用のファイルシステムインタフェースを提供することであって、前記オブジェクトストアによって内部的に使用される前記フラットな名前空間上に階層化された階層名前空間に対するアブストラクションを複数のクライアントに公開することを含み、前記アブストラクションは、前記階層名前空間へのルートディレクトリおよび1つまたは複数のサブディレクトリを含む、前記ファイルシステムインタフェースを提供することと、

前記ルートディレクトリの前記サブディレクトリとして現れる複数のトップレベルオブジェクトの作成を可能とすることであって、各トップレベルオブジェクトは、指定のクラスタ化ファイルシステムに従って編成された別個の名前空間を有するストレージデバイスの別個のアブストラクションを表し、少なくとも1つの前記トップレベルオブジェクトは、特定のクラスタ化ファイルシステムのファイルシステムオブジェクトを含む、複数のトップレベルオブジェクトの作成を可能とすることと、

前記オペレーションは、前記クラスタ化ファイルシステム内にストアされたファイルへのアクセスを前記クライアントが試みたときに、前記トップレベルオブジェクトを、前記ルートディレクトリの前記アブストラクションのローカルバージョンでのサブディレクトリとして自動マウントすることと

を備える、コンピュータシステム。

【請求項 18】

前記トップレベルオブジェクトは、前記クラスタ化ファイルシステム内にストアされたファイルを表すファイルオブジェクトのためのメタデータをストアするよう構成され、前記ファイルオブジェクトは、前記トップレベルオブジェクトとは別個の前記オブジェクトストア内にストアされる、請求項17に記載のコンピュータシステム。

【請求項 19】

ファイルオブジェクトは、前記ファイルの別個の部分を表し、前記ファイルオブジェク

10

20

30

40

50

トとは別個の前記オブジェクトストア内にストアされた対応するコンポーネントオブジェクトの位置を発見するために前記フラットな名前空間により使用可能な一意識別子をストアするよう構成される、請求項 18 に記載のコンピュータシステム。

【請求項 20】

ファイルオブジェクトのために前記トップレベルオブジェクト内にストアされた前記メタデータは少なくとも、前記オブジェクトストア内の前記ファイルオブジェクトの位置を発見するために前記フラットな名前空間により使用可能な一意識別子を備えるファイル記述子を含む、請求項 18 に記載のコンピュータシステム。

【請求項 21】

前記ファイルオブジェクトにより表される前記ファイルは、前記複数のホストコンピュータのうちの 1 つで実行中の仮想マシンにより使用される仮想ディスクファイルである、請求項 20 に記載のコンピュータシステム。

10

【請求項 22】

前記トップレベルオブジェクトは、前記複数のホストコンピュータのうちのいずれかでインスタンス化されることが可能な複数の仮想マシンのうちの 1 つによってのみ利用されるファイルオブジェクトのためのメタデータをストアするよう構成される、請求項 21 に記載のコンピュータシステム。

【請求項 23】

前記フラットな名前空間は、メモリ内データベースとして実装され、メモリ内データベースの同期化コピーが前記ホストコンピュータの各々のメモリ内にストアされる、請求項 17 に記載のコンピュータシステム。

20

【発明の詳細な説明】

【背景技術】

【0001】

分散システムにより、ネットワーク内の多数のクライアントが共有リソースのプールへアクセスすることが可能である。例えば、分散ストレージシステムにより、ホストコンピュータのクラスタが、各ホストコンピュータ内に位置するか、または、それに取り付けられたローカルディスク（例えば、SSD（Solid State Drive）、PCI（Peripheral Component Interconnect）ベースのフラッシュストレージ、SATA（Serial AT Attachment）、または SAS（Serial Attached SCSI）磁気ディスク）を集約してストレージの単一または共有のプールを作成することが可能となる。このストレージのプール（本開示では時に「データストア」または「ストア」とも称される）は、クラスタ内の全てのホストコンピュータによりアクセス可能であり、ストレージエンティティの単一名前空間として存在し得る（ファイルの場合の階層ファイルシステム名前空間、オブジェクトの場合の一意識別子のフラットな名前空間、等のように）。ホストコンピュータ上で生成された仮想マシン等のストレージクライアントは、データストアを使用し得、例えば、仮想マシンによりその動作中にアクセスされる仮想ディスクを格納する。データストアを形成する共有ローカルディスクは異なる性能特性（例えば、容量、毎秒入力／出力または IOPS（Input／Output Per Second）能力、等）を有し得るため、仮想ディスクまたはその一部を格納するためのこのような共有ローカルディスクの使用量は、各所与の仮想マシンのニーズに基づいて仮想マシン間で分散され得る。

30

40

【0002】

このアプローチは、企業に費用効率の高い性能を提供する。例えば、プールされたローカルディスクを使用する分散ストレージは、安価であり、かつ高度にスケーラブルで、比較的 management が簡単である。このような分散ストレージはクラスタ内の市販品ディスクを使用可能であるので、企業は追加のストレージインフラストラクチャに投資する必要がない。しかしながら、生じる 1 つの課題は、クラスタ内の市販品ディスクに亘ってオブジェクトがどこにストアされるかを効率的に追跡する機構の開発、および、必要なときに如何に効率的にそれらにアクセスするか、という点にある。例えば、フラットな名前空間の利用

50

は、オブジェクトをストアおよび取得する単純すぎる効率的手段を提供し得るものの、クライアントの既存のインタフェースに対して互換性のある方法でオブジェクトを編成するのに有用であり得る、または、異なるクライアントの異なるストレージ要件を別様に満たす、オブジェクト間の階層関係を作成するための十分な柔軟性を、それは提供しない。例えば、このようなスケーラブルなオブジェクトストア（例えば、アプリケーション、ストレージ管理ツール、仮想化ハイパーバイザ、等）を利用することができたであろう多くの既存の環境は、オブジェクトストアがストレージインタフェースに基づいてある階層ファイルシステムを提供することを要求し得る。1つの例は、仮想マシンメタデータを階層ファイルシステム内のファイルの形態でストアする、ヴィエムウェア社（VMware, Inc.）のvSphere Hypervisorである。さらに、データストアは、データストアにストアされた同じデータに同時にアクセスし得るホストコンピュータのクラスタ間で共有されるので、データストアにより提供される単一の名前空間の管理に使用される任意のファイルシステムは、同時実行制御のための機構を有する必要がある。現在の分散またはクラスタ化ファイルシステムは、同時実行制御のある形態を通常提供する。しかしながら、それらの固有の設計における制限のために、このような現在のファイルシステムは、それらがサポート可能な同時「クライアント」（例えば、ファイルシステムにアクセスするホストコンピュータまたは仮想マシン）の数における制限を通常有する。データストアを管理するために使用される現在のファイルシステムが、同時にそれにアクセス可能なクライアントの数の限界を有すると、たとえその容量を増大させるために追加の市販品ストレージが容易にデータストアに追加可能であっても、このような追加のストレージへのアクセスのためにさらなるクライアントが追加されはしないので、高度にスケーラブルなデータストアのユーティリティは、頭打ちになる。

【発明の概要】

【課題を解決するための手段】

【0003】

本開示の一実施形態は、オブジェクトストア内にストアされたオブジェクトへの多数のクライアントによる同時アクセスをサポートするよう意図された、オブジェクトストアのためのファイルシステムインタフェースを提供する方法に関する。本方法によれば、オブジェクトストアについての階層名前空間へのルートディレクトリのアブストラクションが、クライアントに露出される。オブジェクトストアは、複数のホストコンピュータに収容されたまたは直接に取り付けられた複数の物理ストレージデバイスにより支持され、ストアされたオブジェクトへと一意識別子をマッピングするフラットな名前空間を使用して、そのストアされたオブジェクトを内部的に追跡する。ルートディレクトリのサブディレクトリとして現れるトップレベルオブジェクトの作成が可能とされ、各トップレベルオブジェクトは、任意の指定ファイルシステムに従って編成され得る別個の名前空間を有するストレージデバイスの別個のアブストラクションを表す。

【0004】

フラットな名前空間を使用してオブジェクトを内部的にストアし、アクセスする一方で、クライアントにより使用され得る階層名前空間を層化することにより、本開示の技術は、市販品ストレージリソースの分散を使用するラージスケーラブルクラスタ化ファイルシステムをサポートするフレームワークをオファーすることが可能である。例えば、究極的にクライアントによりアクセスされ得るファイルオブジェクトを含むファイルシステムを表すトップレベルオブジェクトを作成する能力を提供して、前述のオブジェクトストアは、既存の分散またはクラスタ化ファイルシステムよりも高度なスケーラビリティをオファーする。なぜならばそれは、例えば、オブジェクトストア内の任意の特定のファイルシステムオブジェクトのために構成された任意の特定の分散クラスタ化ファイルシステム（例えば、VMware VMFS（Virtual Machine File System）、NFS（Network File System）、等）の設計に固有の同時クライアント数の如何なる制限にも限定されないからである。すなわち、クラスタ化ファイルシステムの異なる部分は、あるオブジェクトに一方で存在する名前空間のあるサブ空

間内のデータにアクセスするこのようなクライアントのサブセットによってのみ、特定のクライアントからのスケーラビリティ要件が達成される必要があるように、異なるオブジェクトに配置され得るので、オブジェクトストアのトップにおけるクラスタ化ファイルシステムの実装は、スケーラビリティアドバンテージを有する。

【0005】

他の実施形態は、限定するわけではないが、処理ユニットに本開示の方法の1つ以上の態様を実装させ得る命令を含むコンピュータ可読媒体、ならびに、本開示の方法の1つ以上の態様を実装するよう構成された、プロセッサ、メモリおよびモジュールを有するコンピュータシステムを含む。

【図面の簡単な説明】

10

【0006】

【図1】一実施形態に係る、例示的なコンピューティング環境を示す。

【図2】一実施形態に係る、仮想ディスクを表すオブジェクトストア内に編成されたオブジェクトの例示的な階層構造を示す。

【図3】一実施形態に係る、VSAN (Virtual Storage Area Network) モジュールのコンポーネントを示す。

【図4】一実施形態に係る、定義されたストレージポリシーに基づく仮想ディスクオブジェクト作成の方法フロー図を示す。

【図5】一実施形態に係る、VM (Virtual Machine) で生じたI/Oオペレーションの取り扱いを示す。

20

【発明を実施するための形態】

【0007】

図1は、一実施形態に係る、コンピューティング環境100を示す。図示されるように、コンピューティング環境100は、ノードで実行中の仮想マシン (VM) 112に集約オブジェクトストア116を提供するために、クラスタ110のホストサーバまたはノード111に收容されるか、または直接取り付けられた市販品ローカルストレージを利用するソフトウェアベースの「仮想ストレージエリアネットワーク」 (VSAN) 環境である。以下、用語「收容された」または「に收容された」の使用は、收容され、または別様に直接に取り付けられたことの両方を包含するよう使用され得る。ノード111に收容されるか、または別様に直接に取り付けられたローカル市販品ストレージは、ソリッドステートドライブ (SSD) 117および磁気もしくはスピニングディスク118のうちの少なくとも一つの組み合わせを含み得る。ある実施形態において、SSD 117は、I/O性能向上のための磁気ディスク118手前の読み出しキャッシュおよび書き込みバッファのうち少なくともいずれか一方として機能する。下記に詳述されるように、各ノード111は、(例えば、オブジェクトストアにオブジェクトを作成する、等の) ストレージ管理ワークフローを自動化し、オブジェクトストア内のオブジェクトに対し指定された所定のストレージポリシーに基づいて、(例えば、オブジェクトストア内のオブジェクトに対するI/Oオペレーションを取り扱う、等の) オブジェクトストア内のオブジェクトへのアクセスを提供するために、ストレージ管理モジュール (本開示においては、「VSANモジュール」と称される) を含み得る。例えば、VMは、その使用目的に応じてその「仮想ディスク」のための特定のストレージ要件 (例えば、容量、可用性、IOPS等) を有するようアドミニストレータにより最初に構成され得るため、アドミニストレータは、このような可用性、容量およびIOPS等を定義する各VM毎のストレージプロファイルまたはポリシーを定義し得る。さらに後述するように、VSANモジュールは、定義されたポリシーに基づいてオブジェクトストアの物理ストレージリソースにより支持することにより、指定された仮想ディスクのために「オブジェクト」を作成し得る。

30

40

【0008】

仮想化管理プラットフォーム105は、ノード111のクラスタ110と関連付けられる。仮想化管理プラットフォーム105により、アドミニストレータがノード111でのVMの構成及び生成を管理することが可能となる。図1の実施形態に示されているように

50

、各ノード111は、仮想化層またはハイパーバイザ113、V SANモジュール114、および（SSD117およびノード111の磁気ディスク118を含む）ハードウェア119を含む。ハイパーバイザ113を通じて、ノード111は、多数のVM112を立ち上げて実行することができる。ハイパーバイザ113は、部分的に、各VM112についてコンピューティングリソース（例えば、処理能力、ランダムアクセスメモリ等）を適切に割り当てるためにハードウェア119を管理する。さらに、後述するように、各ハイパーバイザ113は、その対応するV SANモジュール114を通じて、仮想ディスク（またはその一部）用のストレージとしての使用のためにハードウェア119に位置するストレージリソース（例えば、SSD117および磁気ディスク118）へのアクセスと、クラスタ110内の任意のノード111に存在する任意のVM112によりアクセスされ得る他の関連ファイルへのアクセスとを提供する。特定の実施形態において、ヴィエムウェア社（VMware, Inc.）のヴィ・スフィア・ハイパーバイザ（vSphere Hypervisor）が、ハイパーバイザ113としてノード111にインストールされ得、ヴィエムウェアのヴィ・センタサーバ（vCenter Server）が仮想化管理プラットフォーム105として使用され得る。

【0009】

一実施形態において、V SANモジュール114はハイパーバイザ113内に「V SAN」デバイスドライバとして実装される。このような実施形態において、V SANモジュール114は、概念的「V SAN」115へのアクセスを提供し、V SAN115を介してアドミニストレータがオブジェクトストア116により支持される多数のトップレベル「デバイス」または名前空間オブジェクトを生成し得る。よくある状況の1つとして、デバイスオブジェクトの作成中、アドミニストレータは、デバイスオブジェクトのための特定のファイルシステムを規定し得る（このようなデバイスオブジェクトは以下、「ファイルシステムオブジェクト」とも称される）。例えば、一実施形態において、各ノード111内の各ハイパーバイザ113は、ブートプロセスの間、V SANモジュール114により露出される概念的グローバル名前空間のための／v san／ルートノードを発見し得る。例えば、V SANモジュール114により露出されるAPIにアクセスすることにより、ハイパーバイザ113は、V SAN115にその時点で存在する全てのトップレベルファイルシステムオブジェクト（または他のタイプのトップレベルデバイスオブジェクト）を決定し得る。VM（または他のクライアント）がファイルシステムオブジェクトのうちの1つにアクセスしようと試みると、ハイパーバイザ113は、そのときのファイルシステムオブジェクトを動的に「自動マウント」し得る。ある実施形態において、ファイルシステムオブジェクト内のオブジェクトへのアクセスが中止またはある期間アイドルとなるとき、ファイルシステムオブジェクトはさらに、周期的に「自動アンマウント」され得る。V SAN115を通じてアクセス可能なファイルシステムオブジェクト（例えば、／v san／fs_name1、等）は、例えば、VMへの同時アクセス中に同時実行制御を提供するよう設計された、ヴィエムウェアの分散またはクラスタ化ファイルシステム、V MFS（Virtual Machine File System）のような、特定のファイルシステムのセマンティクスをエミュレートするよう実装され得る。V SAN115が多数のファイルシステムオブジェクトをサポートするため、任意の特定のクラスタ化ファイルシステムの限定に縛られることなくオブジェクトストア116を通じたストレージリソースの提供が可能である。例えば、多くのクラスタ化ファイルシステム（例えば、V MFS等）は、ある量のノード111をサポートするためにだけスケール可能である。多数のトップレベルファイルシステムオブジェクトサポートを提供することにより、V SAN115は、このようなクラスタ化ファイルシステムのスケーラビリティ制限を克服する。

【0010】

下記の図2の文脈における更なる詳述に示されるように、ファイルシステムオブジェクトは、それ自身、クラスタ110で実行中のVM112によりアクセス可能な多数の仮想ディスク記述子ファイル（例えば、ヴィ・スフィア（vSphere）環境における、v

10

20

30

40

50

mdkファイル、等）へのアクセスを提供し得る。これらの仮想ディスク記述子ファイルは、仮想ディスクのための実際のデータを含むとともに、オブジェクトストア116により別個に支持される仮想ディスク「オブジェクト」への参照を含む。仮想ディスクオブジェクトは、それ自身、階層的または「複合」オブジェクトであり得、階層的または「複合」オブジェクトは、下記に説明されるように、仮想ディスクの最初の作成時にアドミニストレータにより生成される対応するストレージプロファイルまたはポリシーのストレージ要件（例えば、容量、可用性、IOPS等）を反映する「コンポーネント」オブジェクト（これもオブジェクトストア116により別個に支持される）によりさらに構成される。下記にさらに議論されるように、各VSANモジュール114は、（下記にさらに説明される実施形態において、クラスタレベルオブジェクト管理すなわち「CLOM（Cluster Level Object Manager）」サブモジュールを通じて）、他のノード111の他のVSANモジュール114と通信して、オブジェクトストア116内にストアされた種々のオブジェクト間の位置、構成、ポリシーおよび関係を記述するメタデータを含むメモリ内メタデータデータベース（例えば、各ノード111のメモリ内に、別個に、しかし、同期方式により維持された）を作成および維持する。このメモリ内メタデータデータベースは、例えば、アドミニストレータが最初にVMのために仮想ディスクを作成するとき、および、VMが実施中であり仮想ディスク上のI/Oオペレーション（例えば、読み出しまたは書き込み）を行う場合に、ノード111上のVSANモジュール114により利用される。図3の文脈で下記にさらに議論されるように、VSANモジュール114は、（下記にさらに説明される一実施形態において、ドキュメントオブジェクトマネージャすなわち「DOM（Document Object Manager）」サブモジュールを通じて）I/Oオペレーションの対象となる仮想ディスクの一部を支持する実際の物理ローカルストレージを収容するノード（または、ノード群）へのI/Oオペレーションリクエストを適切に送るために、メモリ内データベース内のメタデータを使用してオブジェクトの階層を横断する。

【0011】

図2は、一実施形態に係る、仮想ディスクを表すオブジェクトストア116内に編成されたオブジェクトの例示的な階層構造を示す。上記において先に詳述されたように、1つのノード111にて実行中のVM112は、オブジェクトストア116内に階層的または複合オブジェクト200としてストアされた仮想ディスク上のI/Oオペレーションを実行し得る。ハイパーバイザ113は、VSANモジュール114を通じてVSAN115のアブストラクションとインターフェースすることにより（例えば、一実施形態において先に議論されたように、仮想ディスクオブジェクトに対応するトップレベルファイルシステムオブジェクトを自動マウントすることにより）、VM112に仮想ディスクへのアクセスを提供する。例えば、VSANモジュール114は、メモリ内メタデータデータベースのそのローカルコピーをクエリすることにより、仮想ディスクのための記述子ファイル210（例えば、vmdkファイル）を格納するVSAN115内にストアされた特定のファイルシステムオブジェクト205（例えば、一実施形態におけるVMFSファイルシステムオブジェクト、等）を識別することができる。ファイルシステムオブジェクト205が、仮想化環境をサポートする際に仮想マシン構成ファイル（例えば、vSphere環境における、vmxファイル、等）等の、その目的に沿った各種の他のファイルをストアし得ることが理解されるであろう。ある実施形態において、各ファイルシステムオブジェクトは、特定のVMに対応するそれらの仮想ディスク（例えば、「毎VM」ファイルシステムオブジェクト）のみをサポートするよう構成され得る。

【0012】

記述子ファイル210は、オブジェクトストア116に別個にストアされ、仮想ディスクを概念的に表す（および本開示において時に仮想ディスクオブジェクトとも称され得る）複合オブジェクト200への参照を含む。複合オブジェクト200は、仮想ディスク作成時にアドミニストレータにより生成された対応ストレージプロファイルまたはポリシー内のストレージ要件（例えば、容量、可用性、IOPS、等）またはサービスレベルアグ

10

20

30

40

50

リーメント (SLA (Service Level Agreement)) に適する仮想ディスクのストレージ編成または構成 (本開示において仮想ディスク「ブループリント」と時に称される) を記述するメタデータを格納する。例えば、図2の実施形態において、複合オブジェクト200は、仮想ディスクの2つのミラーコピー (例えば、ミラー) が RAID (Redundant Arrays of Inexpensive Disks) 0構成内でさらにそれぞれストライプ化される RAID 1構成を記述する仮想ディスクブループリント215を含む。したがって、複合オブジェクト225は、各仮想ディスクミラー内の各ストライプ (例えば、仮想ディスクのデータパーティション) に対応する、多数の「リーフ」または「コンポーネント」オブジェクト220_xに対する参照を含み得る。各コンポーネントオブジェクト220のための (例えば、各ストライプのための) メモリ内メタデータベースにおいてVSANモジュール114によりアクセス可能なメタデータは、ストライプ (およびこのような物理リソース内のストライプの位置) を実際に格納する物理ストレージリソース (例えば、磁気ディスク118、等) を収容するクラスタ110内の特定のノード111_xに対するマッピングを提供するか、または、別様に同特定のノード111_xを識別する。

【0013】

図3は、一実施形態に係る、VSANモジュール114の構成要素を示す。前述のように、ある実施形態において、VSANモジュール114は、ハイパーバイザ113へのVSAN115のアブストラクションの露出するデバイスドライバとして実行し得る。VSANモジュール114の各種のサブモジュールは、異なる責務を取り扱い、このような責務に応じてユーザ空間315またはカーネル空間320のいずれかにおいて動作し得る。図3の実施形態に示されるように、VSANモジュール114は、ユーザ空間315で動作するクラスタレベルオブジェクト管理 (CLOM: (CLOM: cluster level object management)) サブモジュール325を含む。CLOMサブモジュール325は、アドミニストレータによる仮想ディスクの作成中に、仮想ディスクブループリントを生成し、このような仮想ディスクブループリントのために作成されたオブジェクトが、アドミニストレータにより設定されたストレージプロファイルまたはポリシー要件を満たすよう構成されることを確実にする。 (例えば、仮想ディスクのための) オブジェクト作成中にアクセスされることに加えて、CLOMサブモジュール325は、 (例えば、仮想ディスクブループリントまたはオブジェクトストア116内の実際の物理ストレージへの仮想ディスクブループリントのマッピングを動的に修正または別様に更新するために) オブジェクトに関するストレージプロファイルまたはポリシーへのアドミニストレータによりなされた変更の際、または、クラスタまたは作業負荷に対する変更が、現在のストレージプロファイルまたはポリシーに準拠しないオブジェクトを生じるときに、アクセスされ得る。

【0014】

一実施形態において、アドミニストレータが、仮想ディスクオブジェクト200等の複合オブジェクトのストレージプロファイルまたはポリシーを作成すると、CLOMサブモジュール325は、各種の発見的アルゴリズムおよび分散アルゴリズムのうち少なくとも一方を適用して、ストレージポリシーを満たすか、または別様にストレージポリシーに適合するクラスタ110内の構成 (例えば、ミラーリングを介した所望の冗長性およびストライピングを介したアクセス性能を達成するためのRAID構成: 負荷バランスを達成するために、どのノードのローカルストレージが、仮想ディスクのある部分/パーティション/ストライプを格納すべきか、等) を記述する仮想ディスクブループリント215を生成する。例えば、CLOMサブモジュール325は、一実施形態において、仮想ディスクが最初にアドミニストレータにより作成されたときに、図2における仮想ディスクオブジェクト200のRAID 1/RAID 0構成を記述するブループリント215の生成を担当する。上記において説明されたように、ストレージポリシーは、容量、IOPS、可用性および信頼性についての要件を規定し得る。ストレージポリシーはまた、作業負荷特徴 (例えば、ランダムまたはシーケンシャルアクセス、I/Oリクエストサイズ、

10

20

30

40

50

キャッシュサイズ、期待キャッシュヒット率、等)を規定し得る。さらに、アドミニストレータはまた、あるノード111(またはノード111に収容されたローカルディスク)を優先的に使用するためにV SANモジュール114への親和度を規定し得る。例えば、VMの新たな仮想ディスクの提供時に、アドミニストレータは、仮想ディスクが400GBの予備容量、150読み出しIOPSの予約、300書き込みIOPSの予約および99.99%の所望の可用性を有することを規定する仮想ディスクのストレージポリシーまたはプロファイルを生成し得る。生成されたストレージポリシーの受信時に、CLOMサブモジュール325は、生成されたストレージポリシーに適する複合オブジェクト(例えば、仮想ディスクオブジェクト)の仮想ディスクブループリントを生成する目的で、そのV SANモジュール114により維持されるメモリ内メタデータデータベースを調べてクラスタ110の現在の状態を判定する。下記においてさらに説明されるように、CLOMサブモジュール325は、その対応する分散オブジェクトマネージャ(DOM)サブモジュール340にブループリントを伝達し、分散オブジェクトマネージャ(DOM)サブモジュール340は、オブジェクト空間116と相互作用して、例えば、クラスタ110の各種ノード111内の物理ストレージ位置に複合オブジェクトのコンポーネントオブジェクト(例えば、ストライプ)を割り当てるか、または別様にマッピングすることにより、ブループリントを実施する。

10

【0015】

CLOMサブモジュール325およびDOMサブモジュール340に加えて、図3にさらに示されるように、V SANモジュール114はまた、クラスタ110の状態についての情報をV SANモジュール114の他のサブモジュールに提供するために、上述のメモリ内メタデータデータベースを維持するとともに、クラスタ110内の各ノード111のステータス、アクセス可能性および可視性をモニタリングすることによりクラスタ110の全般的な「ヘルス」の追跡も行なう、クラスタモニタリング、メンバシップおよびディレクトリサービス(CMMD S(Cluster Monitoring, Membership, and Directory Services))サブモジュール335を含み得る。メモリ内メタデータデータベースは、各種ノード111、ノード111内に収容されたストレージリソース(SSD、磁気ディスク、等)およびその特性/能力、ノード111の現在の状態、および、それらの対応するストレージリソース、ノード111間のネットワーク経路、等のV SAN環境の物理インベントリを維持するディレクトリサービスとして機能する。先に説明されたように、物理インベントリを維持することに加えて、メモリ内メタデータデータベースはさらに、オブジェクトストア116にストアされるオブジェクトについてのメタデータのカatalog(例えば、どのような複合オブジェクト・コンポーネントオブジェクトが存在するか、どのようなコンポーネントオブジェクトがどのような複合オブジェクトに属するか、どのノードがいずれのオブジェクトへのアクセスを制御する「コーディネータ」または「オーナー」として機能するか、各オブジェクトについてのサービス品質要件、オブジェクト構成、オブジェクトの物理ストレージ位置へのマッピング、等)を提供する。先に説明されたように、V SANモジュール114内の他のサブモジュールは、更新目的でCMMD Sサブモジュール335(図3において接続線で表される)にアクセスして、クラスタトポロジおよびオブジェクト構成における変更を学習し得る。例えば、先に説明されたように、仮想ディスク作成中、CLOMサブモジュール325は、仮想ディスクブループリントを生成するためにメモリ内メタデータデータベースにアクセスし、実行中のVM112からのI/Oオペレーションに対処する目的で、DOMサブモジュール340は、メモリ内メタデータデータベースにアクセスして、対応する複合オブジェクト(仮想ディスクオブジェクト)のコンポーネントオブジェクト(例えば、ストライプ)を格納するノード111、および、I/Oオペレーションを満たすためにそれらのノードが到達可能な経路を決定する。

20

30

40

【0016】

上記において説明されたように、DOMサブモジュール340は、I/Oオペレーションの取り扱いの間、および、オブジェクト作成の間、DOMサブモジュール340が実行

50

される特定のノード111のローカルストレージにストアされたオブジェクトストア116内のそれらのコンポーネントオブジェクト、および、そのノード111が現在「コーディネータ」または「オーナー」として指定され続けているある他の複合オブジェクトへのアクセスの制御およびそのオペレーションの取り扱いを実施する。例えば、VMからのI/Oオペレーションの取り扱い時に、ある実施形態における複合オブジェクトの階層的性質のため、目的の複合オブジェクト（例えば、I/Oオペレーションの対象となる仮想ディスクオブジェクト）のためのコーディネータとして機能するDOMサブモジュール340は、第2のノード111（またはノード群）内の異なるDOMサブモジュール340とのネットワークを介してさらなる通信を必要とし得る。第2のノード111（またはノード群）は、第2のノード111のローカルストレージにストアされ、かつI/Oオペレーションの対象である仮想ディスクの一部分である仮想ディスクオブジェクトの特定のコンポーネントオブジェクト（例えば、ストライプ等）についてのコーディネータとして機能する。I/Oオペレーションを発行するVMが、仮想ディスクオブジェクトのコーディネータとも異なるノード111上に存在すると、VMを実行中のノードのDOMサブモジュール340は、コーディネータのDOMサブモジュール340ともネットワークを介して通信する必要があるであろう。ある実施形態において、I/Oオペレーションを発行するVMが、I/Oオペレーションの対象となる仮想ディスクオブジェクトのコーディネータと異なるノード上に存在すると、2つのノードの2つのDOMサブモジュール340は、VMを実行中のノードに対して仮想ディスクオブジェクトのコーディネータの役割を変更するために互いに通信し得る（例えば、それにより、VMを実行中のノードと仮想ディスクオブジェクトについてのコーディネータとして機能するノードとの間のI/Oオペレーションを調整するのに必要なネットワーク通信の量が低減される）。

【0017】

DOMサブモジュール340もまた同様に、オブジェクト作成中に互いに通信する。例えば、仮想ディスク作成中にCLOMモジュール325により生成された仮想ディスクブループリントは、どのノード111が仮想ディスクオブジェクトおよびその対応するコンポーネントオブジェクト（ストライプ等）についてのコーディネータとして機能すべきかを指定する情報を含み得る。このような指定されたノードについてのDOMサブモジュール340の各々には、オブジェクトに関するメタデータでメモリ内メタデータデータベースを更新する目的で、それらの各オブジェクトを作成し、このようなオブジェクトにローカルストレージを（必要に応じ）割り当て、それらのオブジェクトをそれらの対応するCMMDSサブモジュール335に通知するためのリクエストが発行される（例えば、実施形態に応じて、仮想ディスクオブジェクトについてのコーディネータとして指定されたDOMサブモジュール340により、または、仮想ディスクブループリントを生成するノードのDOMサブモジュール340により、等）。このようなリクエストを実施するために、DOMサブモジュール340は、そのノード111のローカルSSDおよび磁気ディスクとの通信を実際に行なうVSANモジュール114内のコンポーネントとして機能するログ構造化オブジェクトマネージャ（LSOM（Log Structured Object Manager））サブモジュール350と相互作用する。コンポーネントオブジェクトについてのローカルストレージの割り当て（およびそのノードがコーディネータとして機能する複合オブジェクトについてのポリシーおよび構成等の他のメタデータのストア、等）に加えて、LSOMサブモジュール350はさらに、例えば、ストレージリソースが輻輳したかどうかを報告するために、そのノード111のローカルストレージへのI/Oオペレーションの流れをモニタリングする。

【0018】

図3はまた、論理エンドポイント（例えば、ノード、オブジェクト、等）間の任意サイズのデータグラムを提供する高信頼データグラムトランスポート（RDT（Reliable Datagram Transport））サブモジュール345を示し、エンドポイントは潜在的に、多数の経路に渡り得る。一実施形態において、基礎となるトランスポートはTCP（Transmission Control Protocol）であ

10

20

30

40

50

る。代替的に、RDMA (Remote Direct Memory Access) 等の他のトランスポートが使用され得る。RDTサブモジュール345は、例えば、上記において先に説明されたようにオブジェクト作成またはI/Oオペレーション取り扱いのために、DOMサブモジュール340が互いに通信する時に、使用される。ある実施形態において、RDTモジュール345は、CMMD Sモジュール335と相互作用することにより、メモリ内メタデータデータベース内の最新の位置情報を維持し、リンクヘルスステータスに基づいて接続を作成、除去または再確立するために、論理エンドポイントのアドレスを動的に解決する。例えば、CMMD Sモジュール335がリンクがヘルシーでないと報告すると、RDTサブモジュール345は、よりよい状況でのリンクを好み、接続を切り得る。

10

【0019】

図4は、一実施形態に係る、定義されたストレージポリシーに基づく仮想ディスクオブジェクト作成の方法フロー図を示す。例えば、ステップ400において、アドミニストレータは、容量、可用性およびIOPS要件（例えば、定義されたストレージポリシー）を有する仮想ディスクを作成するために、仮想管理プラットフォーム105のユーザインタフェースと相互作用し得る。一実施形態において、仮想管理プラットフォーム105は、ステップ405において「マスター」ノード111に仮想ディスクについてのオブジェクトを作成するようリクエストし得る。ステップ410において、このようなマスターノード111は、VSANモジュール内のそのCLOMサブモジュール325を通じて仮想ディスクブループリントを生成し得る。先に説明されたように、CLOMサブモジュール35は、CMMS (Computerized Maintenance Management System) サブモジュール335のメモリ内メタデータデータベースへのコンサルティングにより判定されたように、クラスタ110のステータスに基づいて仮想ディスクオブジェクト（例えば、複合オブジェクト）の作成のために仮想ディスクブループリントを生成する。仮想ディスクブループリントは、仮想ディスクオブジェクトのコーディネータまたはオーナーとして機能すべき特定のノードを識別し得る。ステップ415において、マスターノード111のDOMサブモジュール340は、識別されたノードのDOMサブモジュール340に仮想ディスクオブジェクトを作成するようリクエストし得る。ステップ420において、識別されたノードのDOMサブモジュール340は、例えば、その対応するLSOMサブモジュール350と通信することにより、リクエストを受信し、仮想ディスクオブジェクトを作成し、仮想ディスクオブジェクトを記述するメタデータをそのローカルストレージに持続的に格納する。ステップ425において、DOMサブモジュール340は、仮想ディスクオブジェクトブループリントに基づいて、仮想ディスクブループリント内の任意のコンポーネントオブジェクトについてのコーディネータまたはオーナーとして機能するよう指定されているクラスタ110内のそれらの他のノードを識別する。DOMサブモジュール340は、コンポーネントオブジェクトについてのコーディネータとして機能する他のノードのDOMサブモジュール340と（例えば、そのRTP (Real-time Transport Protocol) サブモジュール345を使用して）通信し、それらのローカルストレージ内にこのようなコンポーネントオブジェクトを支持するデータを格納する。このようなDOMサブモジュール340が、仮想ディスクオブジェクトのコーディネータのDOMサブモジュール340から各コンポーネントオブジェクトを作成するリクエストを受信すると、DOMサブモジュール340は、今度はステップ430において、コンポーネントオブジェクト（およびその関連するメタデータ）についてのローカルストレージを割り当てるために各モジュール350と通信する。ひとたびこのようなコンポーネントオブジェクトが作成されると、ステップ435において、それらのDOMサブモジュール340は、そのCMMSサブモジュール335のメモリ内メタデータデータベースにコンポーネントの作成を通知する。ステップ440において、仮想ディスクオブジェクトのコーディネータのDOMサブモジュール340はまた、メモリ内メタデータデータベースを更新するために、その作成をそのCMMD Sサブモジュール335に通知し、最終的にはアドミニストレータに確認を送信する（例えば

20

30

40

50

、マスターノードを介して仮想管理プラットフォーム105に通信し戻す)。

【0020】

図5は、一実施形態に係る、VMから生じたI/Oオペレーションの取り扱いを示す。特定のノードで実行中のVMがその仮想ディスクへのI/Oオペレーションを実施すると、VMのゲストオペレーティングシステムは、ステップ500において、その仮想ディスクについて意図されたI/Oオペレーションリクエストを(ゲストオペレーティングシステムのデバイスドライバを通じて)送信し、ステップ505において、そのリクエストはハイパーバイザ113により受信され、ハイパーバイザ113内のI/Oスタックの各種の層を介してV SANモジュール114のDOMサブモジュール340に最終的には送信されて変換される。ステップ510において、DOMサブモジュール340により受信されたI/Oリクエストは、CMMSSサブモジュール335のメモリ内メタデータデータベースへのアクセスにより仮想ディスクオブジェクトのコーディネータノードを識別するためにDOMサブモジュール340が使用する仮想ディスクを表すオブジェクトについての一意識別子を含む(ある実施形態において、コーディネータノードの識別子の一意識別子へのマッピングをルックアップするためにメモリ内メタデータデータベースへアクセスすることは、仮想ディスクオブジェクトが最初にアクセスされた時にのみ生じ、このようなマッピングは、後続のルックアップが必要とされないように将来のI/Oオペレーションについて持続する)。仮想ディスクオブジェクトについてのコーディネータノードを識別すると、VMを実行中のノードのDOMサブモジュール340は、ステップ515においてI/Oオペレーションを実行するようリクエストするために(例えば、そのRTPサブモジュール345を使用して)コーディネータノードのDOMサブモジュール340と通信する。先に説明されたように、ある実施形態において、VMを実行中のノードと仮想ディスクオブジェクトのコーディネータとして機能するノードとが異なると、2つのDOMサブモジュールは、仮想ディスクオブジェクトのコーディネータの役割を、実行中のVMのノードとなるよう更新するように通信する。コーディネータによるI/Oリクエストを受信すると、ステップ520において、そのDOMサブモジュールは、I/Oオペレーションの対象となる仮想ディスクオブジェクトの特定のコンポーネントオブジェクト(例えば、ストライプ)のそれらのコーディネータノードを(例えば、ある実施形態において、再びメモリ内メタデータデータベースを参照することにより)識別する。例えば、I/OオペレーションがRAID 0構成の多数のストライプ(例えば、多数のコンポーネントオブジェクト)にまたがれば、DOMサブモジュール340は、I/Oオペレーションを分割し、対応するI/Oリクエストを2つのストライプに対応する関連コンポーネントオブジェクトについての各コーディネータノードに適切に送信し得る。ステップ525において、仮想ディスクオブジェクトのコーディネータノードのDOMサブモジュールは、識別されたコンポーネントオブジェクトのコーディネータノードのDOMサブモジュールがI/Oオペレーションリクエストを実行するようリクエストし、ステップ530において、識別されたコンポーネントオブジェクトについてのこのようなコーディネータノードのDOMサブモジュールは、コンポーネントオブジェクトがストアされるローカルストレージリソース内でI/Oオペレーションを実行するために、それらの対応するLSOMサブモジュールと相互作用する。

【0021】

1つ以上の実施形態が理解を明確にするために詳細に記述されてきたが、特許請求の範囲内においてある変更および修正がなされ得ることは明らかである。従って、記述された実施形態は例示的であり制限するものではないと考えられるべきであって、特許請求の範囲は、本開示にて与えられた詳細に制限されるべきではなく、しかし特許請求の範囲およびその均等物内において修正され得る。例えば、前述の多数の実施形態は、仮想マシンをV SANモジュールにより提供される仮想ディスクにアクセスするクライアントとして記述したが、非仮想化ホストサーバおよびそこで実行中の非仮想化アプリケーションのうちの少なくとも一方のクラスタ等の任意のクライアントが、代替実施形態においてV SANモジュールを同様に利用し得ることが理解されるべきである。同様に、V SANモジュール

10

20

30

40

50

ルの代替実施形態が、限定するものではないが、REST (Representational State Transfer) オブジェクト、ファイル、ファイルシステム、ブLOB (バイナリーラージオブジェクト) および他のオブジェクト等の、仮想ディスク以外の高レベルストレージオブジェクトを作成可能であり得る。同様に、VSANモジュール114がハイパーバイザ113内に埋め込まれるとして概して記述されてきたが、代替実施形態は、例えば特別な仮想マシンまたは仮想アプライアンス、別個のアプリケーションまたは分散オブジェクトストアを提供および管理するためにコンピューティングプラットフォームに挿入可能な任意の他の「プラグابل (pluggable)」モジュールまたはドライバとして、ハイパーバイザ113とは別個のVSANモジュールを実装し得る。同様に、前述の実施形態はブループリントを編成する1つの技術としてRAID構成を参照してきたが、他の実施形態において、限定するものではないが、抹消符号および他の類似技術の使用を含む他の構成を利用し得ることが理解されるべきである。本開示の記述はオブジェクト内の参照オブジェクトへの「一意識別子」を用いて説明されてきたが、一意識別子 (ハッシュ、等) を生成する技術において、真に一意的な識別子の生成が必ずしも保証されるわけではなく、それゆえ、ある実施形態が、オブジェクトの識別子が真に一意的ではない例がある場合に名前衝突を取り扱う技術をさらに実装し得ることが理解されるであろう。1つのこのような実施形態において、意図されている一意識別子に加えて、さらなるアドミニストレータまたはユーザ特定識別子も、その作成の間にオブジェクトに割り当てられる (または別様にオブジェクトの一意識別子にマッピングされる)。新たに作成されたオブジェクトと同じ一意識別子を有する既存のオブジェクトのために名前衝突がある場合、アドミニストレータまたはユーザ特定名称は、新たに作成されたオブジェクトのユーザ識別子に変換され得る。

【0022】

本開示に記述された各種の実施形態は、コンピュータシステム内にストアされたデータを伴う各種のコンピュータ実装オペレーションを採用し得る。例えば、これらのオペレーションは、通常は物理量の物理マニピュレーションを要求し得るが、必ずしもではないものの、これらの量は電氣的または磁氣的信号の形態を取り得、そこでは、それら、あるいは、それらの代表値は、ストアされ、伝送され、結合され、比較され、または別様にマニピュレートされることが可能である。さらに、このようなマニピュレーションは、しばしば製造、識別、判定または比較の観点で参照される。1つ以上の実施形態の一部を形成する本開示で記述された任意のオペレーションは、有用なマシンオペレーションであり得る。さらに、1つ以上の実施形態はまた、それらのオペレーションを実行するためのデバイスまたは装置に関する。装置は、特定の要求目的のために特別に構築され得、または、コンピュータにストアされたコンピュータプログラムにより選択的に活性化され、または構成される汎用コンピュータであり得る。特に、各種の汎用マシンが、本開示の教示に従って書かれたコンピュータプログラムと共に使用され得、または、要求されるオペレーションの実行のためにより特別化された装置を構築することが、より都合がよい場合もあり得る。

【0023】

本開示に記述された各種の実施形態は、ハンドヘルドデバイス、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またはプログラマブルコンシューマエレクトロニクス、ミニコンピュータ、メインフレームコンピュータ、等を含む他のコンピュータシステム構成で実践され得る。

【0024】

1つ以上の実施形態は、1つ以上のコンピュータ可読媒体内に具現化された1つ以上のコンピュータプログラムとして、または、1つ以上のコンピュータプログラムモジュールとして実装され得る。用語「コンピュータ可読媒体」は、後にコンピュータシステムに入力され得るデータを格納することが可能な任意のデータストレージデバイスを指し、コンピュータ可読媒体は、コンピュータによりコンピュータプログラムを読み取り可能なようにコンピュータプログラムを具現化するための任意の既存のまたは将来開発される技術に

基づき得る。コンピュータ可読媒体の例には、ハードドライブ、ネットワーク接続ストレージ（NAS：Network Attached Storage）、リードオンリーメモリ、ランダムアクセスメモリ（例えば、フラッシュメモリデバイス）、CD（コンパクトディスク）、CD-ROM、CD-RまたはCD-RW、DVD（デジタルバーサタイルディスク）、磁気テープ、および、他の光学および非光学データストレージデバイスが含まれる。コンピュータ可読媒体はまた、コンピュータ可読コードが分散方式でストアされ、実行されるように、ネットワーク結合コンピュータシステム上に分散され得る。

【0025】

さらに、記述された仮想化方法は概して、仮想マシンが特定のハードウェアシステムに沿ったインターフェースを提示するものとしてきたが、記述された方法は、任意の特定のハードウェアシステムに直接には対応しない仮想化と共に使用され得る。ホスト化実施形態、非ホスト化実施形態、または、その両者間の区別があいまいとなりがちな実施形態として実装される、各種の実施形態に係る仮想化システムが、全て想定される。さらに、各種の仮想化オペレーションが、全体としてまたは部分的にハードウェア内に実装され得る。例えば、ハードウェア実装は、非ディスクデータをセキュアにするためにストレージアクセスリクエストの修正のためのルックアップテーブルを採用し得る。

10

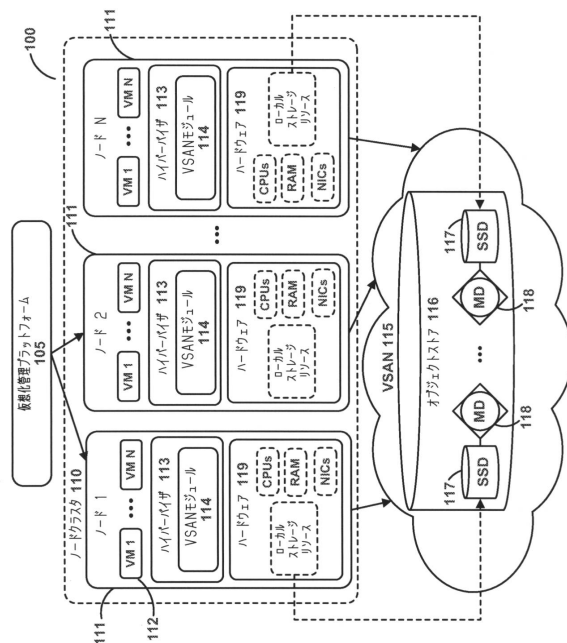
【0026】

仮想化の程度にかかわらず、多くの変形、修正、付加および改良が可能である。仮想化ソフトウェアはそれゆえ、ホスト、コンソールまたは仮想化機能を実行するゲストオペレーティングシステムのコンポーネントを含み得る。複数のインスタンスが、単一のインスタンスとして本開示に説明されるコンポーネント、オペレーションまたは構造のために提供され得る。最後に、各種コンポーネント、オペレーションおよびデータストアの境界は、いくらか恣意的であり、特定のオペレーションは特定の例示的構成の文脈で例示される。機能性の他の割り当てが想定され、1つ以上の実施形態の範囲に収まり得る。一般に、例示の構成において別個のコンポーネントとして表される構造および機能性は、結合された構造またはコンポーネントとして実装され得る。同様に、単一のコンポーネントとして表される構造および機能性は、別個のコンポーネントとして実装され得る。これらのおよび他の変形、修正、付加および改良は、添付の請求項の範囲内に収まり得る。請求項において、要素およびステップの内少なくともいずれか一方は、請求項において明示的に述べられていない限り、オペレーションの如何なる特定の順序も暗示しない。

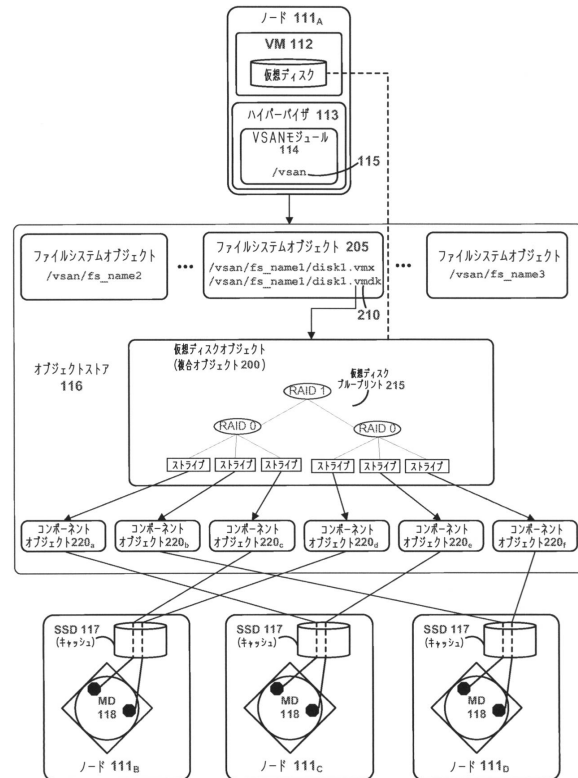
20

30

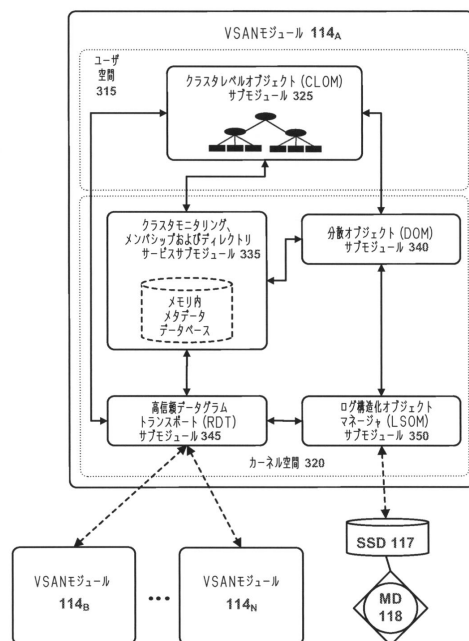
【図 1】



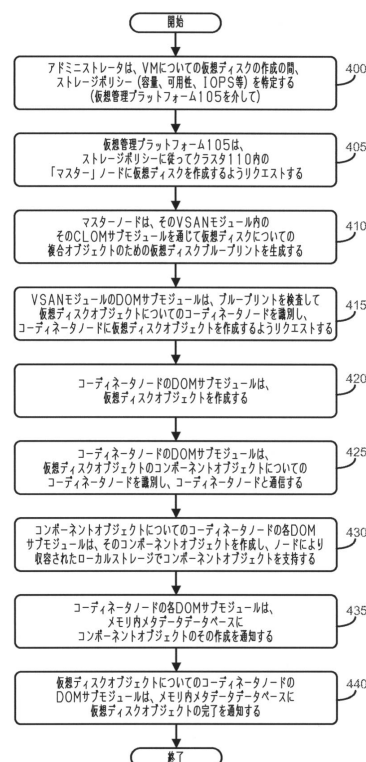
【図 2】



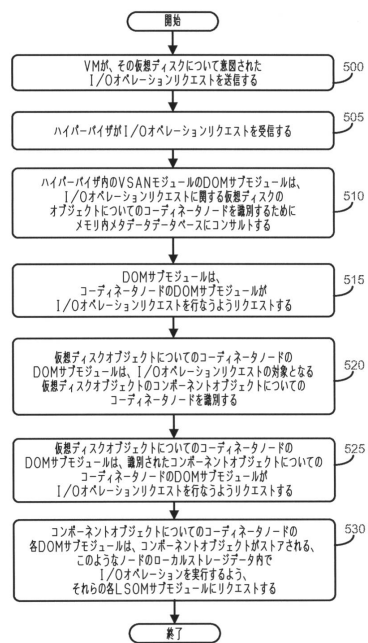
【図 3】



【図 4】



【図 5】



フロントページの続き

- (72)発明者 カラマノリス、クリストス
アメリカ合衆国 94304 カリフォルニア州 パロ アルト ヒルビュー アベニュー 34
01
- (72)発明者 バサニ、ソーム
アメリカ合衆国 94304 カリフォルニア州 パロ アルト ヒルビュー アベニュー 34
01

合議体

審判長 仲間 晃

審判官 辻本 泰隆

審判官 山崎 慎一

- (56)参考文献 米国特許出願公開第2003/0158836 (US, A1)
特開2010-186223 (JP, A)
特開2010-44789 (JP, A)
特開2006-107501 (JP, A)
米国特許出願公開第2005/0097073 (US, A1)
米国特許出願公開第2001/0034733 (US, A1)
特開平10-49423 (JP, A)

(58)調査した分野(Int.Cl., DB名)

G06F12/00

G06F13/10