(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2012-506582
(P2012-506582A)
(43)公表日　平成24年3月15日(2012.3.15)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 12/00	(2006.01)	G06F	12/00	545Z	5B065
G06F 3/06	(2006.01)	G06F	12/00	531D	
		G06F	3/06	304F	

審査請求 未請求　予備審査請求 有　（全 19 頁）

(21)出願番号　特願2011-532619 (P2011-532619)
(86)(22)出願日　平成21年10月21日 (2009.10.21)
(85)翻訳文提出日　平成23年6月8日 (2011.6.8)
(86)国際出願番号　PCT/EP2009/063796
(87)国際公開番号　WO2010/046393
(87)国際公開日　平成22年4月29日 (2010.4.29)
(31)優先権主張番号　0802277-4
(32)優先日　平成20年10月24日 (2008.10.24)
(33)優先権主張国　スウェーデン(SE)

(71)出願人　511094244
イーエルテー　イノベイションズ　アーベー
スウェーデン王国　カールスクルーナ　エス－エスエー－３７１　２２，ピー．オー．ボックス　１７７
(74)代理人　110001302
特許業務法人北青山インターナショナル
(72)発明者　メランデル，クリスチャン
スウェーデン王国　ローデビュー　エス－エスエー－３７０　３０，セレヴェーゲン　３

最終頁に続く

(54)【発明の名称】分散データストレージ

(57)【要約】
本発明は、複数のストレージノードを有する分散データストレージシステムに関する。ユニキャストおよびマルチキャスト伝送を使用して、サーバアプリケーションが、ストレージシステム内のデータを読み出し、書き込むことができる。各ストレージノードは、その他のストレージノードのステータスと共に、システム上の読出しおよび書込み操作を監視する。このように、ストレージノードは、システム上のファイルの複製の必要性を検知して、複製プロセスを実行し、このプロセスが、地理的に多様な位置で正しいバージョンを有するファイルの十分な数の複写の記憶を維持する役割を果たす。
【選択図】図１

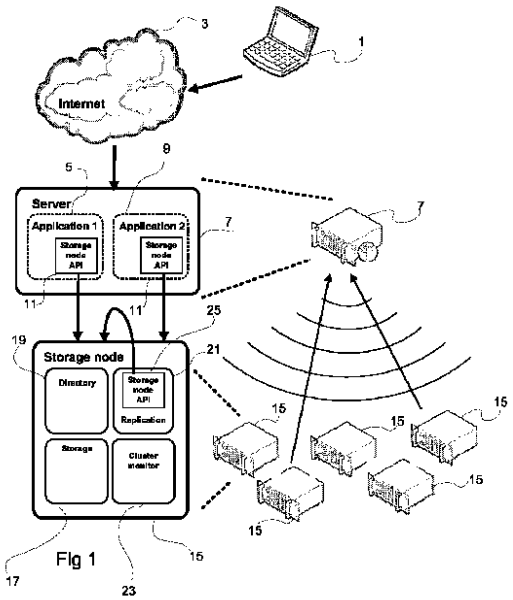

Fig 1

【特許請求の範囲】
【請求項１】
複数のデータストレージノードを有するデータストレージシステム内のデータを維持管理する方法であって、前記データストレージシステムのストレージノードによって利用される方法において、
前記システム内のその他のストレージノードのステータスを、前記データストレージシステムにおいて実行される書込み操作（６５，６７，６９）と共に監視するステップ（５９）と、
前記監視に基づいて、前記データストレージシステムのノード間におけるデータ複製の必要性を示唆する前記データストレージシステムの状態を検知するステップ（５１）と、
そのような状態が検知された場合に、複製プロセスを開始するステップ（５３）とを備え、前記複製プロセスが、複数のストレージノードに、どのストレージノードが特定のデータを格納しているのかを問い合わせるマルチキャストメッセージを送信するステップを含むことを特徴とする方法。
【請求項２】
請求項１に記載の方法において、
監視するステップが、前記システム内のその他のストレージノードからのハートビート信号を聴取するステップ（５９）を含み、複製の必要性を示唆する状態が、ストレージノードの故障であることを特徴とする方法。
【請求項３】
請求項１または２に記載の方法において、
前記データがファイルを含み、前記状態が、ファイル削除またはファイル不一致の一方であることを特徴とする方法。
【請求項４】
請求項１乃至３の何れか１項に記載の方法において、
複製リストが保持され、この複製リストが、複製の必要なファイルを含むことを特徴とする方法。
【請求項５】
請求項４に記載の方法において、
前記複製リストが優先度を含むことを特徴とする方法。
【請求項６】
請求項１乃至５の何れか１項に記載の方法において、
前記複製プロセスが、
前記特定のデータを保有するストレージノードからのレスポンスを受信するステップと、
十分な数のストレージノード上に前記特定のデータが格納されているか否かを判定するステップと、
格納されていない場合に、少なくとも１の追加のストレージノードを選択してそのストレージノードに前記特定のデータを送信するステップとをさらに備えることを特徴とする方法。
【請求項７】
請求項６に記載の方法において、
旧バージョンを含むストレージノード上の前記特定のデータを更新するステップをさらに備えることを特徴とする方法。
【請求項８】
請求項６または７に記載の方法において、
前記システム内のすべてのストレージデバイスの間で、複製すべきファイルのマスターシップの取得をストレージデバイスが試みることから、前記複製プロセスが開始することを特徴とする方法。
【請求項９】

請求項１乃至８の何れか１項に記載の方法において、
監視するステップが、前記データストレージシステムで実行される読出し操作を監視するステップ（６３）をさらに含むことを特徴とする方法。
【請求項１０】
複数のデータストレージノードを有するデータストレージシステム内のデータを維持管理するデータストレージノードであって、
前記システム内のその他のストレージノードのステータスを、前記データストレージシステムにおいて実行される書込み操作と共に監視する手段と、
前記監視に基づいて、前記データストレージシステムのノード間におけるデータ複製の必要性を示唆する前記データストレージシステムの状態を検知する手段と、
そのような状態が検知された場合に、複製プロセスを開始する手段とを備え、前記複製プロセスが、複数のストレージノードに、どのストレージノードが特定のデータを格納しているのかを問い合わせるマルチキャストメッセージを送信することを含むことを特徴とするデータストレージノード。
【請求項１１】
複数のデータストレージノードを有するデータストレージシステムにデータを書き込む方法であって、前記データストレージシステム内のデータにアクセスするアプリケーションを起動するサーバで用いられる方法において、
前記ストレージノードの複数にマルチキャストストレージクエリを送信するステップ（４１）と、
前記ストレージノードのサブセットから、各ストレージノードの地理的な位置に関する地理データを含む複数のレスポンスを受信するステップ（４３）と、
前記レスポンスに基づいて、前記サブセット中の少なくとも２のストレージノードを選択するステップ（４５）と、
選択したストレージノードに、データと、当該データに対応するデータ識別子とを送信するステップ（４７）とを備えることを特徴とする方法。
【請求項１２】
請求項１１に記載の方法において、
前記地理的な位置が、問題のストレージノードの緯度および経度を含むことを特徴とする方法。
【請求項１３】
請求項１２に記載の方法において、
前記レスポンスが、問題のストレージノードのシステム年齢をさらに含むことを特徴とする方法。
【請求項１４】
請求項１２または１３に記載の方法において、
前記レスポンスが、問題のストレージノードのシステム負荷をさらに含むことを特徴とする方法。
【請求項１５】
請求項１２乃至１４の何れか１項に記載の方法において、
前記マルチキャストストレージクエリが、保存されるデータを特定するデータ識別子を含むことを特徴とする方法。
【請求項１６】
請求項１２乃至１５の何れか１項に記載の方法において、
少なくとも３のノードが選択されることを特徴とする方法。
【請求項１７】
請求項１２乃至１６の何れか１項に記載の方法において、
データの格納に成功したストレージノードのリストが、選択したストレージノードに送信されることを特徴とする方法。
【請求項１８】

複数のデータストレージノードを有するデータストレージシステムにデータを書き込むのに適したサーバであって、
前記ストレージノードの複数にマルチキャストストレージクエリを送信する手段と、
前記ストレージノードのサブセットから、各ストレージノードの地理的な位置に関する地理データを含む複数のレスポンスを受信する手段と、
前記レスポンスに基づいて、前記サブセット中の少なくとも２のストレージノードを選択する手段と、
選択したストレージノードに、データと、当該データに対応するデータ識別子とを送信する手段とを備えることを特徴とするサーバ。

【発明の詳細な説明】

【技術分野】

【０００１】

本開示は、複数のデータストレージノードを有するデータストレージシステムにデータを書き込み、維持管理する方法であって、データストレージシステムのストレージノードおよびサーバにおいて用いられる方法に関するものである。さらに、この開示は、そのような方法を実行することができるストレージノードまたはサーバに関する。

【背景技術】

【０００２】

そのような方法は、例えば、米国特許出願公開２００５／０２４６３９３Ａ１に開示されている。この方法は、地理的に異なる位置で複数のストレージセンタを使用するシステムについて開示されている。記憶データに関する情報を維持管理するために、分散オブジェクトストレージマネージャが含まれている。

【０００３】

そのようなシステムに関連する一つの課題は、単純で、しかも強力で信頼性のあるデータの書込みおよび維持管理を如何に達成するかどうかである。

【発明の概要】

【０００４】

したがって、本開示の目的の一つは、システムにおける弱リンクとなり得る集中保守サーバを使用することなく、分散ストレージシステム（distributed storage system）におけるデータの強力な書込みおよび維持管理を実現することである。この目的は、ストレージノード（storage node）で達成される最初に言及した種類の方法であって、データストレージシステムにおいて実行される書込み操作と共に、システム内のその他のストレージノードのステータスを監視するステップと、その監視に基づいて、データストレージシステムのノード間におけるデータ複製の必要性を示唆するデータストレージシステムの状態を検知するステップと、そのような状態が検知された場合に、複製プロセス（replication process）を開始するステップとを備える方法によって達成される。複製プロセスには、複数のストレージノードに、どのストレージノードが特定のデータを格納しているのかを問い合わせるマルチキャストメッセージを送るステップが含まれる。

【０００５】

そのような方法によって、各ストレージノードが全体システムのデータを維持管理することによりアクティブとなり得る。ストレージノードが機能しなくなった場合には、そのデータは、システム内のその他のノードによって復旧することができる。よって、そのようなシステムは自己回復システムとみなすことができる。

【０００６】

監視には、システム内のその他のストレージノードからのハートビート信号を聴取するステップが含まれるものであってもよい。その場合、複製の必要性を示唆する状態は、故障しているストレージノードである。

【０００７】

データにはファイルが含まれる。複製の必要性を示唆する状態は、ファイルの削除またはファイル不一致の何れかであってもよい。

【０００８】
複製を要するファイルを含む複製リストが保持されて、それが優先順位を含むものであってもよい。
【０００９】
複製プロセスは、どのストレージノードが特定のデータを格納しているのかを問い合わせるマルチキャストメッセージ、要求を複数のストレージデバイスに送信するステップと、前記特定のデータを保有するストレージノードからのレスポンスを受信するステップと、十分な数のストレージノードに前記特定のデータが格納されているか否かを判定するステップと、格納されていない場合に、少なくとも１の追加のストレージノードを選択してそのストレージノードに前記特定のデータを送信するステップとを備えるものであってもよい。また、旧バージョンを含むストレージノードの前記特定のデータを更新するようにしてもよい。
【００１０】
また、複製プロセスは、システム内のすべてのストレージノードの中で、複製すべきファイルのマスターシップ（mastership）を取得することをストレージデバイスが試みることから開始するようにしてもよい。
【００１１】
監視には、データストレージシステムで実行される読出し操作の監視がさらに含まれるものであってもよい。
【００１２】
さらに、本開示は、上記方法に対応する、データの維持管理を実行するデータストレージノードに関連するものである。ここで、ストレージノードは、一般に、その方法の動作を実行するための手段を含む。
【００１３】
前記目的は、最初に言及した種類のデータストレージシステムにデータを書き込む方法によっても達成されるものであり、それは、データストレージシステム内のデータにアクセスするアプリケーションを起動するサーバで達成される。その方法は、複数のストレージノードにマルチキャストストレージクエリ（multicast storage query）を送信するステップと、ストレージノードのサブセットから、各サーバの地理的な位置に関連する地理データを含む複数のレスポンスを受信するステップと、前記レスポンスに基づいて、サブセット中の少なくとも２のストレージノードを選択するステップと、選択したストレージノードに、データと当該データに対応するデータ識別子とを送信するステップとを備える。
【００１４】
この方法は、地理的な多様性が効率的な方法で実現されるという点で、強力なデータの書込みを達成する。
【００１５】
地理的な位置は、問題となっているストレージノードの緯度および経度を含むものであってもよく、レスポンスは、問題となっているストレージノードのシステム負荷および／またはシステム年齢（system age）をさらに含むものであってもよい。
【００１６】
マルチキャストストレージクエリは、保存されるデータを特定するデータ識別子を含むものであってもよい。
【００１７】
典型的には、少なくとも３のノードを保存のために選択することができ、データの保存に成功したストレージノードのリストを、選択したストレージノードに送信することができる。
【００１８】
さらに、本開示は、上記方法に対応する、データの書込みを実行するサーバに関連するものである。ここで、サーバは、一般に、前記方法の動作を実行する手段を備える。

【図面の簡単な説明】
【００１９】
【図１】図１は、分散データストレージシステムを示している。
【図２】図２Ａ乃至図２Ｃは、データ読出しプロセスを示している。
【図３】図３は、データ読出しプロセスを示している。
【図４】図４Ａ乃至図４Ｃは、データ書込みプロセスを示している。
【図５】図５は、データ書込みプロセスを示している。
【図６】図６は、数多くのデータストレージノードにおいて数多くのファイルが格納されている状況を概略的に示している。
【図７】図７は、ハートビート信号の送信を示している。
【図８】図８は、データ維持管理プロセスの概要である。
【発明を実施するための形態】
【００２０】
本開示は、複数のストレージノードを含む分散データストレージシステムに関連するものである。このシステムとそれが使用されるコンテキストの構成の概要が図１に示されている。
【００２１】
ユーザコンピュータ１は、インターネット３を介して、サーバ７上で作動するアプリケーション５にアクセスする。ユーザコンテキストは、図示のように、それ自体良く知られた標準的なクライアント・サーバ構成である。しかしながら、開示されるデータストレージシステムがその他の構成においても有用となり得ることに留意されたい。
【００２２】
図示の例では、２つのアプリケーション５，９がサーバ７上で起動する。当然のことながら、アプリケーションの数は、これと異なるものであってもよい。各アプリケーションは、分散データストレージシステム１３に対するインタフェースを提供して、サーバ上で起動するアプリケーションからの要求、典型的には書込み要求および読出し要求をサポートするＡＰＩ（Application Programming Interface）１１を有する。アプリケーションの観点からすれば、データストレージシステム１３からの読出し情報またはデータストレージシステム１３への書込み情報は、その他の種類のストレージソリューション、例えば、ファイルサーバまたは単なるハードドライブを使用するものとは異なるように見える必要はない。
【００２３】
各ＡＰＩ１１は、データストレージシステム１３のストレージノード１５と通信し、ストレージノードは、互いに通信する。それら通信は、ＴＣＰ（Transmission Control Protocol）およびＵＤＰ（User Datagram Protocol）に基づくものである。これらの概念は当業者によく知られているものであるので、これ以上は説明しない。
【００２４】
なお、同じサーバ７上の異なるＡＰＩ１１が、ストレージノード１５の異なるセットにアクセスするようにしてもよい。また、各ストレージノード１５にアクセスするサーバ７が複数存在するものであってもよい。しかしながら、これは、後述するように、ストレージノードが操作する方法に、それ程大きな影響を与えるものではない。
【００２５】
分散データストレージシステムの構成要素は、ストレージノード１５、およびストレージノード１５にアクセスするサーバ７内のＡＰＩ１１である。したがって、本開示は、サーバ７およびストレージノード１５で実行される方法に関連する。それら方法は、サーバおよびストレージノード上で起動されるソフトウェアの実行として主としてそれぞれ具現化され、分散データストレージシステム全体の操作およびプロパティを共に決定している。
【００２６】
ストレージノード１５は、典型的には、数多くの機能ブロックが提供されているファイ

ルサーバによって具現化することができる。ストレージノードは、記憶媒体１７を備えることができ、それは、典型的には、数多くのハードドライブから構成される。そのハードドライブは、任意には、ＲＡＩＤ（Redundant Array of Independent Disk）システムとして構成される。しかしながら、その他のタイプの記憶媒体も同様に考えられる。

【００２７】

ストレージノード１５は、ディレクトリ１９を含むようにしてもよい。そのディレクトリは、後述するように、ホストリストとして、データエンティティ／ストレージノードの関係のリストを備える。

【００２８】

ホストリストに加えて、各ストレージノードは、ストレージノードのセットまたはグループ内のすべてのストレージノードのＩＰアドレスが含まれるノードリストをさらに含有する。グループ内のストレージノードの数は、２，３のストレージノードから、数百のストレージノードにまで及ぶことがある。ノードリストは、さらにバージョン番号を備えるものであってもよい。

【００２９】

また、ストレージノード１５は、複製ブロック２１およびクラスタ監視ブロック２３を含むものであってもよい。複製ブロック２１は、ストレージノードＡＰＩ２５を含み、詳細は後述するように、複製プロセスの必要性を特定して複製プロセスを行うための機能を実行するように構成されている。複製プロセスが、説明する読出し操作および書込み操作の間にサーバ７により行なわれる動作に大いに対応する動作を含むので、複製ブロック２１のストレージノードＡＰＩ２５は、サーバ７のストレージノードＡＰＩ１１のコードに大いに対応するコードを含むものであってもよい。例えば、複製中に実行される書込み操作は、サーバ７によって実行される書込み操作に大いに対応する。クラスタ監視ブロック２３は、後でより詳細に述べるように、データストレージシステム１３内のその他のストレージノードの監視を実行するように構成されている。

【００３０】

分散データストレージシステムのストレージノード１５は、同じ階層レベルに存在すると考えることができる。保存されたデータエンティティのディレクトリを維持してデータ整合性等を監視する役割を果たすマスタ・ストレージノードを指定する必要性は全くない。その代わりに、ストレージノード１５はすべて同等とみなすことができ、ときには、システム内のその他のストレージデバイスに対して、データ管理操作を実行することもある。この同等性は、システムを強力なものとする。あるストレージノードが機能しないときは、システム内のその他のストレージノードが、機能不全のノードを覆い隠して、信頼性のあるデータストレージを確保することとなる。

【００３１】

以下に、システムの操作を、データの読出し、データの書込み、およびデータの維持管理の順に説明することとする。これら方法が、非常に良く相互に働いたとしても、それらが互いとは無関係に原理上は実行され得ることに留意されたい。すなわち、例えば、本開示のデータ書込み方法が使用されなくても、データ読出し方法が優れた特性を提供する場合もあるし、その逆の場合もまた同様である。

【００３２】

ここで、図２Ａ乃至図２Ｃおよび図３を参照して、読出し方法を説明する。図３は、その方法を示すフローチャートである。

【００３３】

読出しは、システムのその他の機能と同様に、複数のストレージノードと同時に通信するマルチキャスト通信を利用する。ここでは、マルチキャストまたはＩＰマルチキャストによって、マルチキャストアプリケーションのために確保されるＩＰアドレスにメッセージを送信することによって達成されるポイントツーマルチポイント通信が意図されている。

【００３４】

例えば、メッセージ、典型的には要求は、そのようなＩＰアドレス（例えば、２４４．０．０．１）に送信される。また、数多くの受信側のサーバは、そのＩＰアドレスへの加入者として登録されている。受信側のサーバの各々は、自身のＩＰアドレスを有する。ネットワークの切替装置が２４４．０．０．１に向けられたメッセージを受け取ると、切替装置は加入者として登録された各サーバのＩＰアドレスにメッセージを転送する。

【００３５】

原則的には、１つのサーバだけをマルチキャストアドレスへの加入者として登録することもでき、その場合、ポイントツーポイント通信が達成される。しかしながら、本開示のコンテキストでは、マルチキャストスキームが使用されるので、そのような通信もマルチキャスト通信とみなすこととする。

【００３６】

また、ユニキャスト通信も単一の受信者との通信に関して使用される。

【００３７】

図２Ａおよび図３に関して、データストレージシステムからデータを取り出す方法は、複数のストレージノード１５にマルチキャストクエリを送信するステップ３１を含む。図示の例では、５つのストレージノードが存在し、その各々がＩＰ（Internet Protocol）アドレス１９２．１６８．１．１，１９２．１６８．１．２などを有する。ストレージノードの数は、言うまでもなく一例に過ぎない。クエリは、データ識別子“２Ｂ９Ｂ４Ａ９７－７６Ｅ５－４９９Ｅ－Ａ２１Ａ６Ｄ７９３２ＤＤ７９２７”を含み、それは、非常によく知られた、例えば、汎用一意識別子、ＵＵＩＤであってもよい。

【００３８】

ストレージノードは、識別子に対応するデータを探してそれら自身をスキャンする。そのようなデータが見つかる場合には、ストレージノードは、レスポンスを送信し、そのレスポンスが、図２Ｂに示すように、サーバ７により受信される（ステップ３３）。図示のように、レスポンスは、任意には、ストレージノードが関連データのコピーを有することを示す情報に加えて更なる情報を含むことができる。具体的には、レスポンスは、データを含むその他のストレージノードに関するストレージノードディレクトリからの情報と、データのどのバージョンがストレージノードに含まれているのかに関する情報と、現時点でどの負荷をストレージノードが受けているのかに関する情報とを含むことができる。

【００３９】

レスポンスに基づいて、サーバは、データが取り出される１またはそれ以上のストレージノードを選択し（ステップ３５）、図２Ｃに示すように、そのストレージノードにデータのユニキャスト要求を送信する（ステップ３７）。

【００４０】

データの要求に応じて、１または複数のストレージノードは、ユニキャストにより関連データをサーバに送信し、サーバはそのデータを受信する（ステップ３９）。図示の例では、１のストレージノードのみが選択されている。これで十分であるが、一貫性検査を可能にする２セットのデータを受け取るために、複数のストレージノードを選択することも可能である。データの転送が失敗した場合、サーバは取り出しのために別のストレージノードを選ぶことができる。

【００４１】

ストレージノードの選択は、良好な全体システム性能を達成するために、幾つかの要因を考慮したアルゴリズムに基づくものとすることができる。典型的には、その他の概念も十分に考えられるが、最新データバージョンおよび最低の負荷を有するストレージノードが選択されることとなる。

【００４２】

任意には、サーバが、関連するすべてのストレージノードに、どのノードがどのバージョンのデータを含むのかを示すリストを送信することによって、操作を終えるようにしてもよい。この情報に基づいて、ストレージノードは、説明する複製プロセスによって、適切にデータを自身で維持管理することができる。

【００４３】
図４Ａ乃至図４Ｃおよび図５は、分散データストレージシステムのためのデータ書込みプロセスを示している。
【００４４】
図４Ａおよび図５に関して、上記方法は、サーバがマルチキャストストレージクエリを複数のストレージノードに送信するステップ（４１）を備える。ストレージクエリは、データ識別子を含み、基本的には、受信側のストレージノードがこのファイルを格納することができるかという問い合わせから構成される。任意には、ストレージノードが、この名称のファイルを既に有しているかどうかをそれらの内部ディレクトリで確認し、サーバがファイルの名称を変更するようなことが万一発生した場合には、サーバ７に通知することができる。
【００４５】
何れにしても、ストレージノードの少なくともサブセット（一部）が、サーバ７にユニキャスト送信によるレスポンスを与えることとなる。典型的には、予め設定された最低限の空きディスク容量を有するストレージノードは、クエリに答えることとなる。サーバ７は、各サーバの地理的位置に関する地理データを含むレスポンスを受信する（ステップ４３）。例えば、図４Ｂに示すように、地理データには、各サーバの緯度、経度および高度が含まれる。しかしながら、郵便番号等のようなその他のタイプの地理データも考えられる。
【００４６】
地理データに加えて、ストレージノード選択プロセスへの入力に役立つ更なる情報を与えるようにしてもよい。図示の例においては、各ストレージノードの空き容量が、ストレージノードのシステム年齢の指示およびストレージノードが現在受けている負荷の指示と共に与えられている。
【００４７】
受信したレスポンスに基づいて、サーバは、データ保存用として、サブセットのうち、少なくとも２のストレージノード、典型的な実施形態では、３のストレージノードを選択する（ステップ４５）。ストレージノードの選択は、異なるデータを考慮したアルゴリズムによって行なわれる。その選択は、ある種の地理的な多様性を達成するために行なわれる。同じラックにあるファイルサーバのみをストレージノードとして選択することは、望ましくは、少なくとも回避すべきである。典型的には、大きな地理的な多様性は、異なる大陸上のストレージノードを選択することによっても達成することができる。地理的な多様性に加えて、その他のパラメータを選択アルゴリズムに含ませるようにしてもよい。最低限の地理的な多様性が達成される限りは、空き容量、システム年齢および現在の負荷も考慮に入れるようにしてもよい。
【００４８】
ストレージノードを選択したら、保存するデータおよびそれに対応するデータ識別子が、典型的にはユニキャスト送信を使用して、選択した各ノードに送信される。
【００４９】
任意には、書込み操作に成功した各ストレージノードがサーバに確認応答を送信することによって、操作を終わらせるようにしてもよい。その後、サーバは、どのノードが書込みに成功したデータを有し、どのノードが有していないのかを示すリストを、関連するすべてのストレージノードに送る。この情報に基づいて、ストレージノードは、説明する複製プロセスによって、適切にデータを自身で維持管理することができる。例えば、１つのストレージノードの書込みが失敗した場合には、そのファイルについて所望数の保存ストレージノードを達成するために、もう１つのストレージノードにファイルを複製する必要がある。
【００５０】
データ書込み方法は、優れた地理的な多様性が提供されるように、それ自体、非常に強力な方法でサーバ７内のＡＰＩがデータを格納することを可能にする。

【００５１】
書込みおよび読出し操作に加えて、サーバ７内のＡＰＩは、ファイルを削除する操作およびファイルを更新する操作を実行することができる。それらのプロセスは、データ維持管理プロセスに関連させて以下に説明することとする。
【００５２】
データ維持管理プロセスの目的は、故障していない十分な数のストレージノードがそれぞれ最新バージョンの各ファイルを確実に保存するようにすることである。また、それは、削除されたファイルがどのストレージノードにも保存さないようにする機能を提供することができる。その維持管理は、ストレージノード自体によって実行される。よって、データストレージの維持管理を担う専用の“マスタ”の必要はない。“マスタ”は、システムの弱点となり得るため、上記維持管理により、信頼性を改善することができる。
【００５３】
図６は、数多くのデータストレージノードにおいて数多くのファイルが格納されている状況を概略的に示している。図示の例では、１９２．１６８．１．１から１９２．１６８．１．１２まで連続的に付されたＩＰアドレスを有する１２のノードが、説明目的に描かれている。しかしながら、言うまでもなく、ＩＰアドレスの番号は同じ範囲である必要性は全くない。ノードは、説明を単に簡単にするために循環的順序で配置されている。すなわち、ノードは特定の順序を有する必要はない。各ノードは、単純化のために、文字Ａ－Ｆによって特定される１または２のファイルを格納する。
【００５４】
図８に関して、データを維持管理する方法は、データストレージシステムのノード間におけるデータの複製の必要性を示唆するデータストレージシステムの状態を検知するプロセス５１と、複製プロセス５３とを備える。検出プロセス５１の結果として、複製の必要性が確認されたファイルのリスト５５が得られる。このリストは、複製の様々な必要性の優先順位に関するデータをさらに含むものであってもよい。このリストに基づいて、複製プロセス５３が行なわれる。
【００５５】
分散ストレージの強さは、各ファイルの正しいバージョンの十分な数の複製がシステム内に保存されているかどうかに依存する。図示の例では、各ファイルにつき、３つの複写が保存されている。しかしながら、例えば、アドレス１９２．１６８．１．５を有するストレージノードが故障した場合、ファイル“Ｂ”および“Ｃ”について保存された複写が所望数に満たなくなる。
【００５６】
すなわち、複製の必要性をもたらす１つの事象は、システム内のストレージノードの故障である。
【００５７】
システム内の各ストレージノードは、システム内のその他のストレージノードの状態を監視することができる。これは、図７に示すように、一定間隔で各ストレージノードに所謂ハートビート信号を出力させることにより、実行することができる。図示の例では、アドレス１９２．１６８．１．７を有するストレージノードが、システム内のその他のストレージノードに対して、正常に機能していることを示すマルチキャスト信号５７を発している。この信号は、ハートビートを監視するステップ５９（図８を参照）を実行するシステム内の動作中のその他のすべてのストレージノードまたはそのサブセットによって受信される。しかしながら、アドレス１９２．１６８．１．５を有するストレージノードにあっては、故障しているため、ハートビート信号を発することはない。したがって、その他のストレージノードは、このノードからハートビート信号が長い時間発せられていないこと、すなわち問題となっているストレージノードが故障していることに気付くこととなる。
【００５８】
ハートビート信号は、ストレージノードのアドレスに加えて、そのノードリストバージ

ョン番号を含むようにしてもよい。その後、別のストレージノードが、ハートビート信号を聴取して、送信側のストレージノードがより最近のバージョンのノードリストを有していることを見出したときは、そのノードリストを転送するように、その送信側のストレージノードに要求することができる。これは、単に、ストレージノードを加えるか、取り除いて、新しいバージョンのノードリストをストレージノードの１つに送ることにより、ストレージノードの追加および削除を得ることができることを意味している。その後、このノードリストは、システム内のその他のすべてのストレージノードに広められることとなる。

【００５９】

再び図８に関して、各ストレージノードは、故障中のストレージノードによって保存されたファイルを求めて内部ディレクトリを探索する（ステップ６１）。ファイル“Ｂ”および“Ｃ”を自身で保存するストレージノードは、故障しているストレージノードを見付けることとなり、よって、そのリスト５５上の対応ファイルを追加することができる。

【００６０】

また、検知プロセスは、ファイルを複製する必要性を示唆するその他の状態を明らかにすることもできる。典型的には、そのような状態としては、不一致の場合、すなわち、１またはそれ以上のストレージノードが旧バージョンのファイルを有している場合があり得る。削除操作は、ファイルの現実の物理的削除を実行するため、複製プロセスも示唆する。その後、サーバの削除操作は、ストレージノードが問題のファイルに削除フラグをセットすることを単に必要とする。したがって、各ノードは、データストレージシステム内で実行される読出しおよび書込み操作を監視することができる。読出しおよび書込み操作の最後にサーバ７によって提供される情報は、１のストレージノードが旧バージョンのファイルを含むこと（読出し操作の場合）、あるいはストレージノードが書込み操作に成功しなかったことを、それぞれ示すことができる。どちらの場合においても、維持管理プロセスの全体のオブジェクトが完了するように、複製によるデータの維持管理の必要性が存在する。

【００６１】

基本的な読出しおよび書込み操作６３、６５に加えて、少なくとも２の追加プロセス、すなわち、以下に簡単に説明する削除プロセス６７および更新プロセス６９が、複製の必要性が存在することを示す情報を与えることができる。

【００６２】

削除プロセスは、サーバ７（図１を参照）によって開始される。読出しプロセスと同様に、サーバは、特定のデータ識別子を持つデータをどのストレージノードが有しているのかを見つけ出すために、すべてのストレージノードに対して、マルチキャストによりクエリを送信する。ストレージノードは、関連する識別子を持つデータを求めて自身をスキャンし、問題となるデータを有する場合にユニキャスト送信により応答する。応答は、データを保有するその他のストレージノードの、ストレージノードディレクトリからのリストを含むようにしてもよい。その後、サーバ７は、削除されるファイルを格納していると考えられるストレージノードに、ユニキャスト要求を送信する。各ストレージノードは、ファイルに関するフラグであって、そのファイルが削除されるべきであることを示すフラグをセットする。その後、そのファイルが複製リストに加えられて、確認応答がサーバに送られる。その後、複製プロセスは、後述するように、ファイルを物理的に削除する。

【００６３】

更新プロセスは、削除プロセスの機能と類似のサーチ機能と、書込みプロセスで実行される機能と類似の書込み機能とを備える。サーバは、特定のデータ識別子を持つデータをどのストレージノードが有しているのかを見つけ出すために、すべてのストレージノードに対して、マルチキャストによりクエリを送信する。ストレージノードは、関連する識別子を持つデータを求めて自身をスキャンし、問題となるデータを有する場合にユニキャスト送信により応答することができる。応答は、データを保有するその他のストレージノードの、ストレージノードディレクトリからのリストを含むようにしてもよい。その後、サ

ーバ７は、ユニキャスト要求を送信して、データを更新するようにストレージノードに命じる。その要求は、当然のことながら、更新データを含んでいる。データを更新するストレージノードは、確認応答をサーバに送信する。サーバは、データの更新に成功したストレージノードと成功しなかったストレージノードとを有するリストを含むユニキャスト送信を送ることにより応答する。また、このリストは維持管理プロセスで使用することができる。

【００６４】

再び図８に関して、読出し操作６３、書込み操作６５、削除操作６７および更新操作６９のすべてが、複写の必要性を有することを示すものであってもよい。同じことは、ハートビート監視５９に適用する。このため、全体の検知プロセス５１は、どのファイルを複製しなければならないのかに関するデータを生成する。例えば、読出し操作または更新操作が、特定のストレージノードが旧バージョンのファイルを含むことを明らかにするようにしてもよい。削除プロセスは、特定のファイルに削除フラグをセットすることができる。ハートビート監視は、故障しているストレージノードに格納された数多くのファイルが新たなストレージノードに複製されることを明らかにすることができる。

【００６５】

各ストレージノードは、格納しているすべてのファイルについて複製の必要性を監視して、複製リスト５５を維持管理する。このため、複製リスト５５は、複製する必要性のある数多くのファイルを含んでいる。それらファイルは、各複製の優先度と一致するように順序付けされるものであってもよい。典型的には、３つの異なる優先レベルが存在するものであってもよい。最高レベルは、ストレージノードが最後のオンライン複写を保持するファイルのために確保される。そのようなファイルは、適当なレベルの冗長性が達成されるように、その他のストレージノードに早急に複製する必要がある。中間レベルの優先度は、ストレージノード間でバージョンが一致しないファイルに関するものであってもよい。より低いレベルの優先度は、故障しているストレージノードに格納されるファイルに関するものであってもよい。

【００６６】

ストレージノードは、複製リスト５５上のファイルを、その優先レベルに従って取り扱う。本明細書中で動作ストレージノードと呼ばれるストレージノードについて複製プロセスを説明するが、すべてのストレージノードがこのように作動するものであってもよい。

【００６７】

維持管理プロセスの複製部分５３は、動作ストレージノードが複製しようとするファイルのマスタになることを試みるステップ７１で開始される。動作ストレージノードは、問題のファイルを格納していると知られているその他のストレージノードに対してマスタとなるために、ユニキャスト要求を送信する。ディレクトリ１９（図１を参照）は、どのストレージノードに問い合わせるべきかに関する情報を含むホストリストを提供する。ストレージノードの１つが肯定的に応答しないような場合、例えば、衝突する要求の場合には、当分の間、ファイルがそのリストに戻されて、その代わりに、リスト上の次のファイルで試みがなされる。若しくは、動作ストレージノードは、このファイルのマスタであるとみなされて、その他のストレージノードは、動作ストレージノードが問題のファイルのマスタであることを示すフラグをセットする。

【００６８】

次のステップは、分散ストレージシステムにおいて問題のファイルのすべての複写を見付けることである（ステップ７３）。これは、動作ストレージノードがマルチキャストクエリをすべてのストレージノードに送信してそれらの何れがそのファイルを有しているのかを問い合わせることにより、実行することができる。そのファイルを有するストレージノードは、クエリに対するレスポンスを提示する。そのクエリには、それらがホストリスト（すなわち、各ストレージノードのディレクトリで維持される関連ファイルを保有するストレージノードのリスト）と同様に保持しているそのファイルのバージョンが含まれる。その後、それらのホストリストが動作ストレージノードによってマージされ（ステップ

７５）、その結果、マスタ・ホストリストが、取り出したすべてのホストリストの結合に対応して形成されるものとなる。動作ストレージノードがマスタになることを試みたときに問い合わせられなかった追加のストレージノードが見付かった場合に、追加のストレージノードについて、そのステップをここで繰り返すようにしてもよい。マスタ・ホストリストは、どのバージョンのファイルを様々なストレージノードが保持して全体のストレージシステム内のファイルのステータスを示しているのかに関する情報を含む。

【００６９】

動作ストレージノードが問題のファイルの最新のバージョンを有していない場合には、最新のバージョンを有しているストレージノードの１つからそのファイルが取り出される（ステップ７７）。

【００７０】

その後、動作ストレージノードは、ホストリストを変更すべきかどうか、典型的には、追加的なストレージノードを加えるべきかどうかを決定する（ステップ７９）。その場合、動作ストレージノードは、サーバによって実行されるように、かつ図４Ａ乃至図４Ｃおよび図５に関連して説明したように、書込みプロセスに非常に類似のプロセスを実行することができる。このプロセスの結果として、ファイルが新しいストレージノードに書き込まれる。

【００７１】

バージョン不一致の場合には、動作ストレージノードがその他のストレージノードに格納されるファイルの複写を更新し（ステップ８１）、その結果、格納されたファイルがすべて正しいバージョンを有するものとなる。

【００７２】

格納されたファイルの余分な複写は削除するようにしてもよい（ステップ８３）。複製プロセスが削除操作によって開始される場合、そのプロセスはこのステップに直接ジャンプすることができる。その後、すべてのストレージノードがファイルの削除を受け入れたらすぐに、動作ストレージノードがユニキャストを使用して、すべてのストレージノードに、単に、問題のファイルを物理的に削除するように要求する。ストレージノードは、ファイルが削除されたことを確認応答する。

【００７３】

さらに、ファイルのステータス、すなわちマスタ・ホストリストが更新される。その後、任意には、複製の必要性がもはや存在しないことを確認するために、ステップ７３－８３を繰り返すことも可能である。この繰り返しは、ステップ８５で更新される必要のない一貫したマスタ・ホストリストをもたらすものとなる。

【００７４】

その後、そのファイルの複製プロセスが終了となり、動作ストレージノードが、ホストリスト上のすべてのその他のストレージノードに対応メッセージを送信することにより、ファイルのマスタとしてのステータスを解除することができる（ステップ８７）。

【００７５】

このシステムは、各ストレージノードがストレージノードのセットにわたり保存しているすべてのファイルを維持管理する責任を負うことにより、優れた信頼性を有する自己修復（ストレージノードが故障した場合）自己クリーニング（ファイル不一致またはファイルが削除される場合）システムを提供する。それは容易に拡大縮小可能であり、数多くの様々なアプリケーションのファイルを同時に保存することができる。

【００７６】

本発明は、特定の開示した実施例に制限されるものではなく、添付の特許請求の範囲内において様々な方法で変更および修正することが可能である。

【図１】

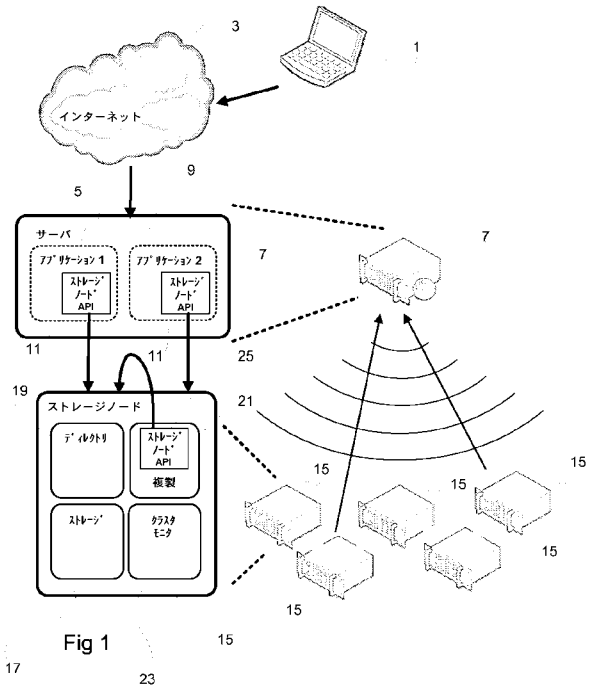

Fig 1

【図２】

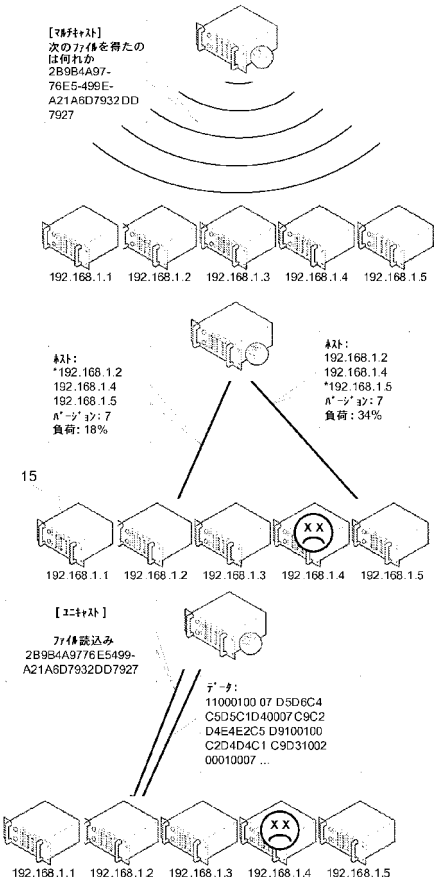

Fig 2A

Fig 2B

Fig 2C

【図３】

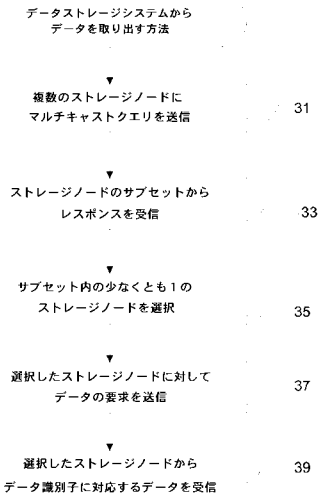

Fig 3

【図４】

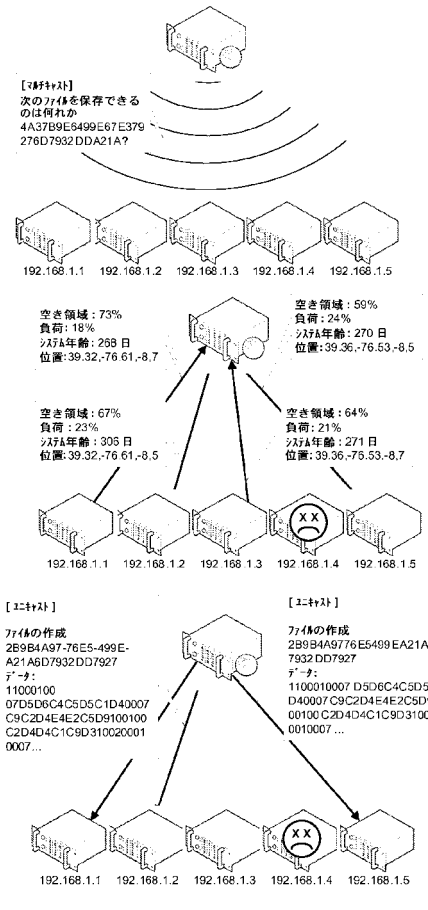

Fig 4A

Fig 4B

Fig 4C

【図５】

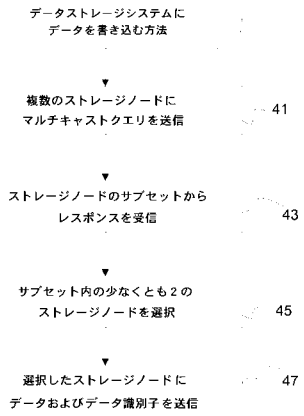

Fig 5

【図６】

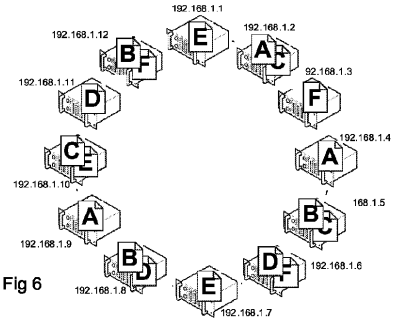

Fig 6

【図７】

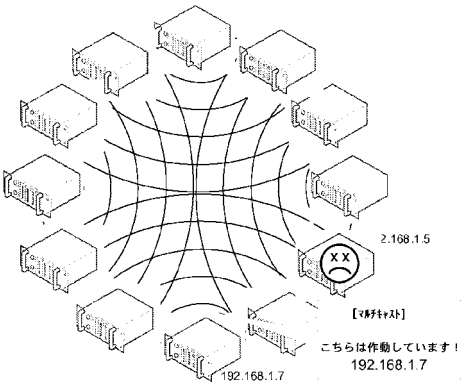

Fig 7

【図８】

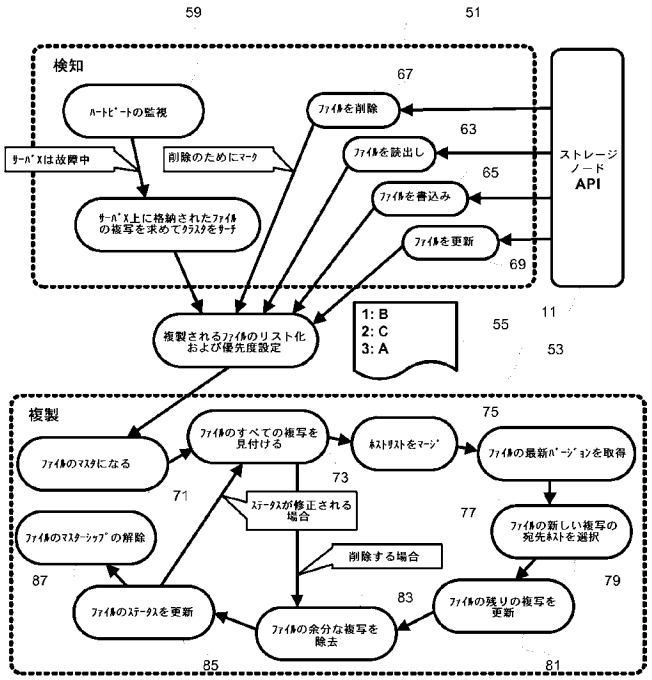

Fig 8

【手続補正書】
【提出日】平成22年6月24日(2010.6.24)
【手続補正１】
【補正対象書類名】特許請求の範囲
【補正対象項目名】全文
【補正方法】変更
【補正の内容】
【特許請求の範囲】
【請求項１】
複数のデータストレージノードを有するデータストレージシステム内のデータを維持管理する方法であって、前記データストレージシステムのストレージノードによって利用される方法において、
前記システム内のその他のストレージノードのステータスを、前記データストレージシステムにおいて実行される書込み操作（６５，６７，６９）と共に監視するステップ（５９）であって、前記ストレージノードが、データエンティティを格納しているストレージノードを含むホストリストに対するアクセスを有するステップと、
前記監視に基づいて、前記データストレージシステムのノード間におけるデータ複製の必要性を示唆する前記データストレージシステムの状態を検知するステップ（５１）と、
そのような状態が検知された場合に、複製プロセスを開始するステップ（５３）とを備え、前記複製プロセスが、複数のストレージノードに、どのストレージノードが特定のデータを格納しているのかを問い合わせるマルチキャストメッセージを送信するステップを含むことを特徴とする方法。
【請求項２】
請求項１に記載の方法において、
監視するステップが、前記システム内のその他のストレージノードからのハートビート信号を聴取するステップ（５９）を含み、複製の必要性を示唆する状態が、ストレージノードの故障であることを特徴とする方法。
【請求項３】
請求項１または２に記載の方法において、
前記データがファイルを含み、前記状態がファイル削除またはファイル不一致の一方であることを特徴とする方法。
【請求項４】
請求項１乃至３の何れか１項に記載の方法において、
複製リストが保持され、この複製リストが、複製の必要なファイルを含むことを特徴とする方法。
【請求項５】
請求項４に記載の方法において、
前記複製リストが優先度を含むことを特徴とする方法。
【請求項６】
請求項１乃至５の何れか１項に記載の方法において、
前記複製プロセスが、
前記特定のデータを保有するストレージノードからのレスポンスを受信するステップと、
十分な数のストレージノード上に前記特定のデータが格納されているか否かを判定するステップと、
格納されていない場合に、少なくとも１の追加のストレージノードを選択してそのストレージノードに前記特定のデータを送信するステップとをさらに備えることを特徴とする方法。
【請求項７】
請求項６に記載の方法において、

旧バージョンを含むストレージノード上の前記特定のデータを更新するステップをさらに備えることを特徴とする方法。

【請求項８】

請求項６または７に記載の方法において、

前記システム内のすべてのストレージデバイスの間で、複製すべきファイルのマスターシップの取得をストレージデバイスが試みることから、前記複製プロセスが開始することを特徴とする方法。

【請求項９】

請求項１乃至８の何れか１項に記載の方法において、

監視するステップが、前記データストレージシステムで実行される読出し操作を監視するステップ（６３）をさらに含むことを特徴とする方法。

【請求項１０】

複数のデータストレージノードを有するデータストレージシステム内のデータを維持管理するデータストレージノードであって、

前記システム内のその他のストレージノードのステータスを、前記データストレージシステムにおいて実行される書込み操作と共に監視する手段であって、その監視を、データエンティティを格納しているストレージノードが含まれるホストリストに基づいて行う手段と、

前記監視に基づいて、前記データストレージシステムのノード間におけるデータ複製の必要性を示唆する前記データストレージシステムの状態を検知する手段と、

そのような状態が検知された場合に、複製プロセスを開始する手段とを備え、前記複製プロセスが、複数のストレージノードに、どのストレージノードが特定のデータを格納しているのかを問い合わせるマルチキャストメッセージを送信することを含むことを特徴とするデータストレージノード。

【請求項１１】

複数のデータストレージノードを有するデータストレージシステムにデータを書き込む方法であって、前記データストレージシステム内のデータにアクセスするアプリケーションを起動するサーバで用いられる方法において、

前記ストレージノードの複数にマルチキャストストレージクエリを送信するステップ（４１）と、

前記ストレージノードのサブセットから、各ストレージノードの地理的な位置に関する地理データを含む複数のレスポンスを受信するステップ（４３）と、

前記レスポンスに基づいて、前記サブセット中の少なくとも２のストレージノードを選択するステップ（４５）と、

選択したストレージノードに、データと、当該データに対応するデータ識別子と、当該データの格納に成功しているストレージノードのリストとを送信するステップ（４７）とを備えることを特徴とする方法。

【請求項１２】

請求項１１に記載の方法において、

前記地理的な位置が、問題のストレージノードの緯度および経度を含むことを特徴とする方法。

【請求項１３】

請求項１２に記載の方法において、

前記レスポンスが、問題のストレージノードのシステム寿命をさらに含むことを特徴とする方法。

【請求項１４】

請求項１２または１３に記載の方法において、

前記レスポンスが、問題のストレージノードのシステム負荷をさらに含むことを特徴とする方法。

【請求項１５】

請求項１２乃至１４の何れか１項に記載の方法において、
前記マルチキャストストレージクエリが、保存されるデータを特定するデータ識別子を含むことを特徴とする方法。
【請求項１６】
請求項１２乃至１５の何れか１項に記載の方法において、
少なくとも３のノードが選択されることを特徴とする方法。
【請求項１７】
複数のデータストレージノードを有するデータストレージシステムにデータを書き込むのに適したサーバであって、
前記ストレージノードの複数にマルチキャストストレージクエリを送信する手段と、
前記ストレージノードのサブセットから、各ストレージノードの地理的な位置に関する地理データを含む複数のレスポンスを受信する手段と、
前記レスポンスに基づいて、前記サブセット中の少なくとも２のストレージノードを選択する手段と、
選択したストレージノードに、データと、当該データに対応するデータ識別子と、当該データの格納に成功しているストレージノードのリストとを送信する手段とを備えることを特徴とするサーバ。

フロントページの続き

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

(72)発明者 ベルンボ，ステファン
スウェーデン王国 カールスクルーナ エス－エスエー－３７１ ３６，アルキメスタレガタン ４６アー
(72)発明者 ペッテション，グスタフ
スウェーデン王国 カールスクルーナ エス－エスエー－３７１ ３２，エストラ ケペマンスガタン ２９
(72)発明者 パーション，ロジャー
スウェーデン王国 カールスクルーナ エス－エスエー－３７１ ４３，ニームスヴェーゲン １セェー

Ｆターム(参考) 5B065 BA06 CE21 EA35