

(19) 日本国特許庁 (JP)

(12) 特 許 公 報 (B2)

(11) 特許番号

特許第5553364号
(P5553364)

(45) 発行日 平成26年7月16日 (2014. 7. 16)

(24) 登録日 平成26年6月6日 (2014. 6. 6)

(51) Int. Cl.	F I
GO 6 F 12/00 (2006. 01)	GO 6 F 12/00 5 4 5 Z
GO 6 F 3/06 (2006. 01)	GO 6 F 12/00 5 3 1 D
	GO 6 F 3/06 3 0 4 F

請求項の数 17 (全 16 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)
(65) 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)
(86) 国際出願番号 PCT/EP2009/063796
(87) 国際公開番号 W02010/046393
(87) 国際公開日 平成22年4月29日 (2010. 4. 29)
審査請求日 平成24年7月13日 (2012. 7. 13)
(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)

(73) 特許権者 511094244
コンピュータ ユーザー
スウェーデン王国 カールスクリーナ 3
7 1 3 3, オストラヴィツガタン 3
6
(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 3 7 0 3 0, セレヴェーゲン
3
(72) 発明者 ベルンボ, ステファン
スウェーデン王国 カールスクリーナ エ
スエスエー 3 7 1 3 6, アルキメス
タレガタン 4 6 アー

最終頁に続く

(54) 【発明の名称】 分散データストレージ

(57) 【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内のストレージノードが故障していることを示す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、

各データエンティティについてのホストリストが、データエンティティを格納している前記データストレージシステム内のストレージノードのサブセットを特定する情報を有するものであり、それら複数のホストリストの情報に基づいて、故障したストレージノードにも格納されていたデータエンティティを判定するステップと、

前記データストレージシステムの前記状態を検知した場合に、前記故障したストレージノードに格納されていたデータエンティティの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。

10

20

【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）を含む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む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。 10

【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、 20

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、 30

前記システム内のすべてのストレージデバイス間で、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、前記複製プロセスを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、 40

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内のストレージノードが故障していることを示すデータストレージシステムの状態を検知する手段と、

各データエンティティについてのホストリストが、データエンティティを格納している前記データストレージシステム内のストレージノードのサブセットを特定する情報を有するものであり、それら複数のホストリストの情報に基づいて、故障したストレージノードにも格納されていたデータエンティティを判定する手段と、 50

前記データストレージシステムの前記状態を検知した場合に、前記故障したストレージノードに格納されていたデータエンティティの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方法。

【請求項 1 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のマルチキャストストレージクエリを送信するステップ（4 1）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（4 3）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（4 5）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であるホストリストとを送信するステップ（4 7）とを備えることを特徴とする方法。

【請求項 1 2】

請求項 1 1 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 1 3】

請求項 1 2 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 1 4】

請求項 1 2 または 1 3 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 1 5】

請求項 1 2 乃至 1 4 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 1 6】

請求項 1 2 乃至 1 5 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 1 7】

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の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であるホスト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。

【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。

10

【0003】

そのようなシステムに関連する一つの課題は、単純で、しかも強力な信頼性のあるデータの書き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き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き込み操作と共に、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。

20

【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。

30

【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。

40

【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

50

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。

10

【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。

20

【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。

30

【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】

40

【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。

50

【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。

10

【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、典型的には書き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への書き込み情報は、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。

20

【0023】

各API 11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API 11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。

30

【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API 11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。

40

【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

50

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。

10

【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。

20

【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。

30

【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。

40

【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】

50

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して行うことができる。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。

10

20

30

40

50

【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。

10

【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。

20

【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。

30

【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。

40

【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実際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。

50

【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顺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。

10

【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よって依存する。図示の例では、各ファイルにつき、3つの複製が保存されている。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製が

20

【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。

30

【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。

40

【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

50

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。

10

【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。

20

30

【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。

40

【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、複製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

50

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。

10

【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】

20

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ことみなされて、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。

30

【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。

40

【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】

50

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。

10

【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。

20

【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】

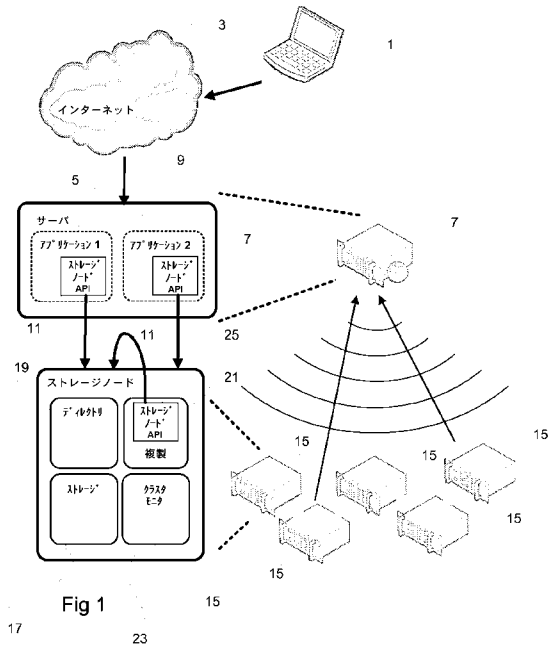
このシステムは、各ストレージノードがストレージノードのセットにわたり保存しているすべてのファイルを維持管理する責任を負うことにより、優れた信頼性を有する自己修復（ストレージノードが故障した場合）自己クリーニング（ファイル不一致またはファイルが削除される場合）システムを提供する。それは容易に拡大縮小可能であり、数多くの様々なアプリケーションのファイルを同時に保存することができる。

30

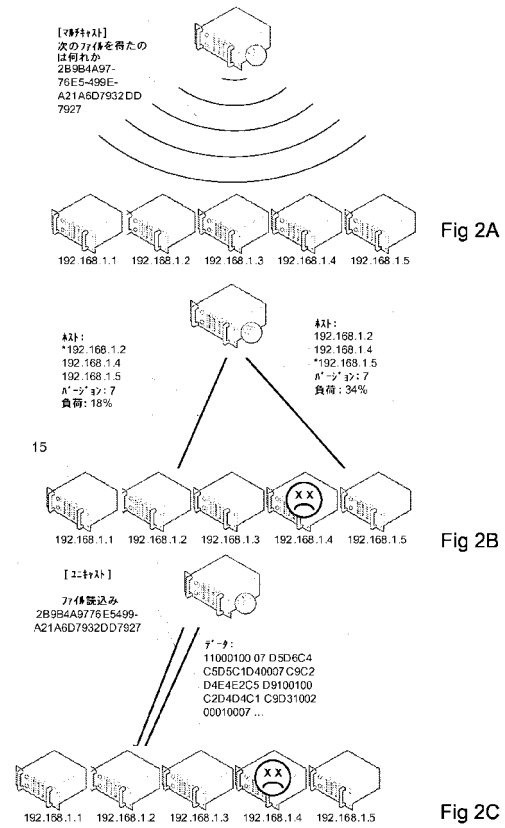
【0076】

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。

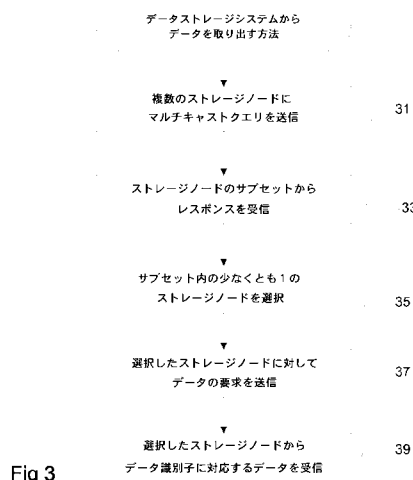
【図 1】



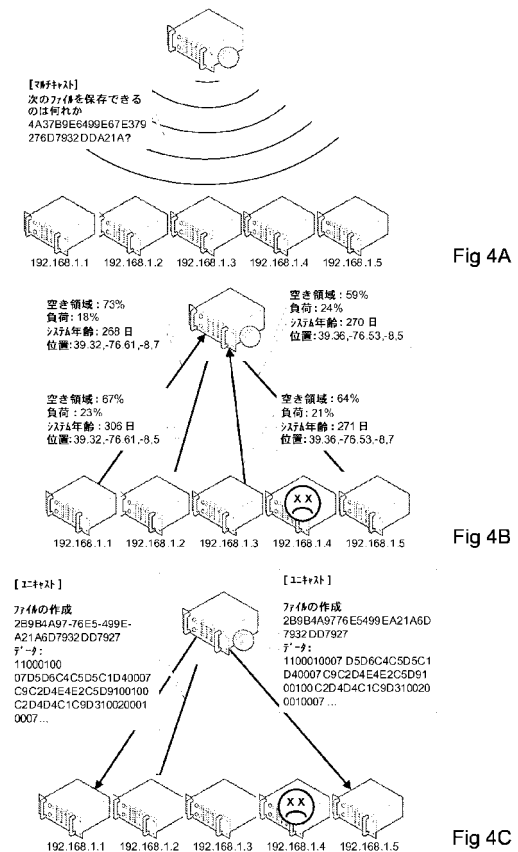
【図 2】



【図 3】



【図 4】



【図 5】

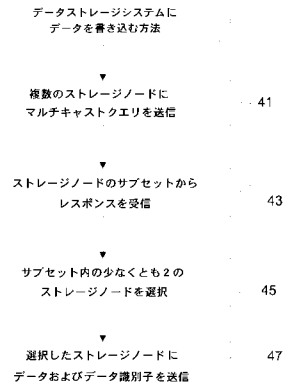


Fig 5

【図 7】

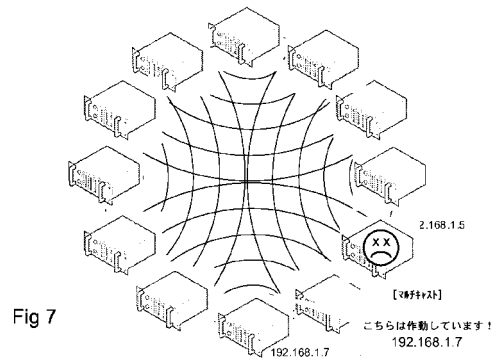


Fig 7

【図 6】

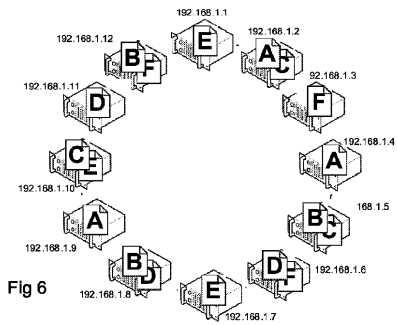


Fig 6

【图 8】

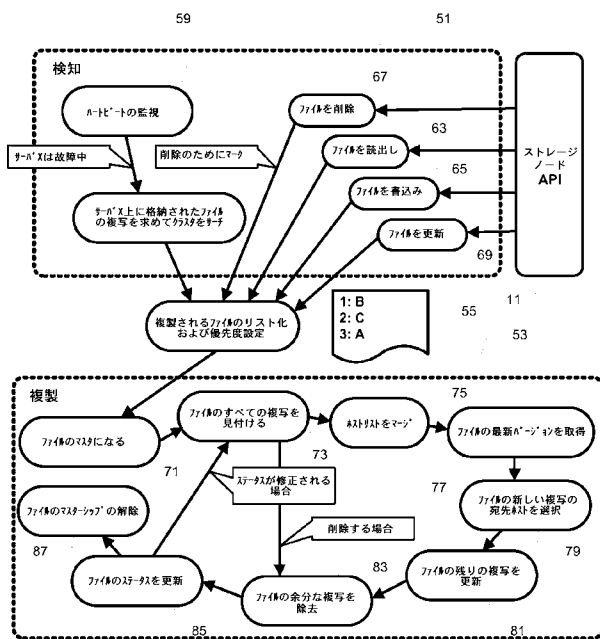


Fig 8

フロントページの続き

- (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 エスーエスエー 371 32, エストラ ケペマンズガ
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 エスーエスエー 371 43, ニームスヴェーゲン 1
セー

審査官 加内 慎也

- (56)参考文献 特開平11-249874 (JP, A)
特開平06-348527 (JP, A)
特開2003-223286 (JP, A)
特開2007-058275 (JP, A)
特開2003-248607 (JP, A)
特開2003-271316 (JP, A)
特開2008-250767 (JP, A)
特開2003-030012 (JP, A)
国際公開第2008/102195 (WO, A1)
米国特許第07266556 (US, B1)

- (58)調査した分野(Int.Cl., DB名)
G06F 12/00
G06F 3/06