

(19) 日本国特許庁 (JP)

(12) 特 許 公 報 (B2)

(11) 特許番号

特許第6197666号
(P6197666)

(45) 発行日 平成29年9月20日 (2017.9.20)

(24) 登録日 平成29年9月1日 (2017.9.1)

(51) Int.Cl.

F I

G 0 6 F 12/00 (2006.01)

G 0 6 F 12/00 5 3 1 D

G 0 6 F 12/00 5 4 5 A

請求項の数 7 (全 14 頁)

(21) 出願番号 特願2014-12449 (P2014-12449)
 (22) 出願日 平成26年1月27日 (2014.1.27)
 (65) 公開番号 特開2015-141451 (P2015-141451A)
 (43) 公開日 平成27年8月3日 (2015.8.3)
 審査請求日 平成28年10月4日 (2016.10.4)

(73) 特許権者 000005223
 富士通株式会社
 神奈川県川崎市中原区上小田中4丁目1番
 1号
 (74) 代理人 100089118
 弁理士 酒井 宏明
 (72) 発明者 小沢 年弘
 神奈川県川崎市中原区上小田中4丁目1番
 1号 富士通株式会社内
 審査官 後藤 彰

最終頁に続く

(54) 【発明の名称】 記憶装置、複製方法及び複製プログラム

(57) 【特許請求の範囲】

【請求項 1】

複数の記憶ノードを用いてデータの複数のレプリカを記憶する記憶装置において、
 第1のレプリカの記憶ノードへの配置について、データを記憶ノードに分散して記憶させる配置を決定する第1の決定部と、
 第2のレプリカの記憶ノードへの配置について、データを同一の記憶ノードに連続して記憶させ、かつ、第1のレプリカを記憶する記憶ノードと異なる記憶ノードに記憶させる配置を決定する第2の決定部と、
 を備えたことを特徴とする記憶装置。

【請求項 2】

第1のレプリカ及び第2のレプリカの配置先の記憶ノードを識別する識別子を記憶する配置先記憶部と、
 データの最終送付先の記憶ノードを識別する識別子を記憶する最終送付先記憶部とをさらに備え、
 前記第2の決定部は、前記配置先記憶部及び前記最終送付先記憶部に基づいて第2のレプリカの記憶ノードへの配置を決定することを特徴とする請求項1に記載の記憶装置。

【請求項 3】

前記第2の決定部は、ハッシュ値を算出して第2のレプリカの配置先を選定し、選定した配置先が前記最終送付先と異なる場合には、前記ハッシュ値を含むハッシュ範囲を複数に分割することによって前記ハッシュ値に基づく第2のレプリカの配置先を前記最終送付

10

20

先と同じにすることを特徴とする請求項 2 に記載の記憶装置。

【請求項 4】

前記第 2 の決定部は、前記ハッシュ範囲の分割を所定の回数繰り返しても第 2 のレプリカの配置先を前記最終送付先と同じにすることができない場合に、前記ハッシュ値に基づいて選定した配置先への第 2 のレプリカの配置を決定し、前記最終送付先を該配置先に更新することを特徴とする請求項 3 に記載の記憶装置。

【請求項 5】

同一記憶ノードから読み出したいデータであるか否かを判定し、判定結果に基づいてデータを読み出す記憶ノードを決定する読出部をさらに備えたことを特徴とする請求項 1 ~ 4 のいずれか 1 つに記載の記憶装置。

10

【請求項 6】

複数の記憶ノードを用いてデータの複数のレプリカを記憶する記憶装置による複製方法において、

第 1 のレプリカの記憶ノードへの配置について、データを記憶ノードに分散して記憶させる配置を決定し、

第 2 のレプリカの記憶ノードへの配置について、データを同一の記憶ノードに連続して記憶させ、かつ、第 1 のレプリカを記憶する記憶ノードと異なる記憶ノードに記憶させる配を決定する

処理を含むことを特徴とする複製方法。

【請求項 7】

20

複数の記憶ノードを用いてデータの複数のレプリカを記憶する分散ストレージが有するコンピュータで実行される複製プログラムにおいて、

第 1 のレプリカの記憶ノードへの配置について、データを記憶ノードに分散して記憶させる配置を決定し、

第 2 のレプリカの記憶ノードへの配置について、データを同一の記憶ノードに連続して記憶させ、かつ、第 1 のレプリカを記憶する記憶ノードと異なる記憶ノードに記憶させる配を決定する

処理を前記コンピュータに実行させることを特徴とする複製プログラム。

【発明の詳細な説明】

【技術分野】

30

【0001】

本発明は、記憶装置、複製方法及び複製プログラムに関する。

【背景技術】

【0002】

複数のサーバを記憶ノードとしてデータを分散して記憶するデータ分散方法の代表的な方法に、コンシステントハッシュ法がある。コンシステントハッシュ法では、予め用意されたハッシュ関数により、各サーバ及びデータに対してハッシュ値が求められる。例えば、サーバの IP アドレスやデータの名前から数値がハッシュ値として計算される。また、ハッシュ関数の値域はリング形態のハッシュ空間で表される。

【0003】

40

図 8 は、リング形態のハッシュ空間の一例を示す図である。図 8 において、あるサーバは、リング上のある範囲のハッシュ値を持つデータを担当する。例えば、「サーバ S 0 0」は、黒塗りされた範囲のデータを担当し、「サーバ S 1 1」は、斜線で模様づけされた範囲のデータを担当する。

【0004】

そして、データの配置については、データ名のハッシュ値からリング上の位置が求まり、さらに担当するサーバが決まり、決まったサーバに第 1 番目の複製（1stレプリカ）が配置される。第 2 番目のレプリカ（2ndレプリカ）は、例えばリングを時計回りに回った次の領域を担当するサーバ（「データ 1」の場合は「サーバ S 1 1」）に配置される。

【0005】

50

2ndレプリカの配置の仕方はリングの時計回りの次の領域以外の選び方もあるが、サーバが一つ故障してもデータにアクセスできることが複数のレプリカを持つ目的であるから、2つのレプリカを同一のサーバに置かないように2ndレプリカの配置を選ぶ必要がある。例えば、1stレプリカ用にはデータの名前からハッシュ値を求め、2ndレプリカ用には違う値からハッシュ値を求めて同じリング上で配置サーバを決めたとすると、2つのハッシュ値が同一の担当範囲にはいり、レプリカが同一のサーバに配置されるかもしれない。

【0006】

この場合、2つのレプリカが配置されたサーバが故障すると2つのレプリカへのアクセスが不可能となり、データの消失を招くことになる。なお、データの名前からハッシュ値を求める以外のハッシュ値の求め方としては、例えば、名前を逆さにする、あるいはデータの値自体からハッシュ値を求めるなどの方法がある。

10

【0007】

また、複数のレプリカを記憶する分散データ管理システムで、データの値から得られる値に基づいて選択したノードにデータを書き込むと共に、メタデータから得られる値に基づいて選択したノードにデータを書き込む従来技術がある（例えば、特許文献1参照）。

【0008】

また、分散システム内の指定された範囲内にあるデータ蓄積域にデータ要求側が要求する可能性のあるデータのレプリカを事前に生成、蓄積しておき、データ要求側から要求を受信すると、レプリカを用いて応答する従来技術がある（例えば、特許文献2参照）。また、ストレージノードの容量、データ記憶コスト、場所などの特徴に基づいてレプリカの配置を決定する従来技術がある（例えば、特許文献3参照）。

20

【先行技術文献】

【特許文献】

【0009】

【特許文献1】特開2010-271798号公報

【特許文献2】特開2007-133503号公報

【特許文献3】特表2012-524947号公報

【発明の概要】

【発明が解決しようとする課題】

【0010】

30

コンシステントハッシュ方式は、データの名前をハッシュ関数によりハッシュすることで、ほぼ均等にデータが分配されるという特性を持つ。レプリカも、リング上で次の領域を担当するサーバに置かれるので、同様にほぼ均等に分配されることになる。

【0011】

一方、データの利用を考えると、一連のデータをなるべく多くのサーバに配置したい場合と、なるべく少数のサーバに配置したい場合がある。例えば、動画を分割して、データA、データB、・・・のようにしたとき、動画として再生したい場合は、順番に読まれるので、なるべく1つのサーバにまとめてデータを置きたい。一方、動画の解析（色の分布を調べるなど）の場合は、並列に読み出して高速に処理するためにデータをなるべく多くのサーバに配置した方が有利になる。

40

【0012】

このように、データはその使われ方によって、なるべく多くのサーバに分散しておいた方がよい場合と、なるべく少数のサーバにまとめた方がよい場合とがある。しかしながら、一般にデータがどう使われるかは、処理によって変わるので、予めどちらかだけのデータ配置にすべきかを決めることはできないという問題がある。また、同一データのレプリカは同一サーバに記憶させてはいけないことは必須の要件である。

【0013】

本発明は、1つの側面では、一連のデータを多くのサーバに配置したい場合と、少数のサーバに配置したい場合の両方に対応すると共に、同一データのレプリカを同一サーバに記憶させない記憶装置、複製方法及び複製プログラムを提供することを目的とする。

50

【課題を解決するための手段】

【0014】

本願の開示する記憶装置は、1つの態様において、複数の記憶ノードを用いてデータの複数のレプリカを記憶する記憶装置であり、第1の決定部と第2の決定部とを備える。第1の決定部は、データが記憶ノードに分散して記憶されるように第1のレプリカの記憶ノードへの配置を決定する。第2の決定部は、データが同一の記憶ノードに連続して記憶され、かつ、第1のレプリカを記憶する記憶ノードと異なる記憶ノードに記憶されるように第2のレプリカの記憶ノードへの配置を決定する。

【発明の効果】

【0015】

10

1実施態様によれば、一連のデータを多くのサーバに配置したい場合と、少数のサーバに配置したい場合の両方に対応すると共に、同一データのレプリカを同一サーバに記憶させないようにすることができる。

【図面の簡単な説明】

【0016】

【図1】図1は、実施例に係る分散ストレージの構成を示す図である。

【図2】図2は、実施例に係る分散ストレージによるレプリカ配置を説明するための図である。

【図3】図3は、複製装置の機能構成を示す図である。

【図4】図4は、第2配置先決定部がデータ分散テーブルの行を増やす場合の例を示す図である。

20

【図5】図5は、2ndレプリカの配置処理のフローを示すフローチャートである。

【図6】図6は、読み出し処理のフローを示すフローチャートである。

【図7】図7は、実施例に係る複製プログラムを実行するコンピュータのハードウェア構成を示す図である。

【図8】図8は、リング形態のハッシュ空間の一例を示す図である。

【発明を実施するための形態】

【0017】

以下に、本願の開示する記憶装置、複製方法及び複製プログラムの実施例を図面に基づいて詳細に説明する。なお、この実施例は開示の技術を限定するものではない。

30

【実施例】

【0018】

まず、実施例に係る分散ストレージの構成について説明する。図1は、実施例に係る分散ストレージの構成を示す図である。図1に示すように、分散ストレージ1は、複数の複製装置2と、複数のサーバ3とを有する。

【0019】

複製装置2は、分散ストレージ1のユーザが分散ストレージ1のアクセスに使用するクライアント5からインターネット6を介してデータアクセス要求を受け付ける装置である。複製装置2は、データの書き込み要求をクライアント5から受け付けると、データの2つのレプリカを異なるサーバ3に転送する。

40

【0020】

また、複製装置2は、データの読み出し要求をクライアント5から受け付けると、データのレプリカを記憶する2つのサーバ3のいずれかからデータを読み出してクライアント5に送信する。なお、複製装置2とクライアント5は1対1に対応する。

【0021】

サーバ3は、データのレプリカをHDD(Hard Disk Drive)に記憶する記憶ノードである。なお、複数の複製装置2は、複数のサーバ3とLAN(Local Area Network)4を介して接続されるが、他のネットワークを介して接続されてもよい。また、複製装置2間は通信が可能である。

【0022】

50

次に、実施例に係る分散ストレージ 1 によるレプリカ配置について説明する。図 2 は、実施例に係る分散ストレージ 1 によるレプリカ配置を説明するための図である。図 2 に示すように、複製装置 2 はレプリカの配置に用いられるデータ分散テーブル 2 2 と、2ndレプリカの最終送付先のサーバ 3 の識別子を記憶する最終送付先記憶部 2 4 とを有する。

【0023】

複製装置 2 は、1stレプリカをデータ分散テーブル 2 2 の 1st列に識別子が記憶されたいずれかのサーバ 3 からデータのハッシュ値に基づいて選択されたサーバ 3 に 1stレプリカを配置する。また、複製装置 2 は、2ndレプリカを最終送付先記憶部 2 4 が識別子を記憶するサーバ 3 に配置し、データ分散テーブル 2 2 の 2nd列に送付先のサーバ 3 の識別子を記憶する。

10

【0024】

ただし、複製装置 2 は、最終送付先記憶部 2 4 が識別子を記憶するサーバ 3 が 1stレプリカを配置したサーバ 3 と同じである場合には、ハッシュ空間を示すリングで最終送付先の次のサーバ 3 に 2ndレプリカを配置する。そして、複製装置 2 は、2ndレプリカを配置したサーバ 3 の識別子で最終送付先記憶部 2 4 を更新し、データ分散テーブル 2 2 の 2nd列に送付先のサーバ 3 の識別子を記憶する。

【0025】

例えば、データから計算されたハッシュ値が「a b c」である場合には、複製装置 2 は、1stレプリカをデータ分散テーブル 2 2 の 1st列に識別子「C」が記憶された「サーバ C」に配置する。そして、複製装置 2 は、2ndレプリカを最終送付先記憶部 2 4 に識別子「D」が記憶された「サーバ D」に配置し、データ分散テーブル 2 2 の 2nd列に送付先のサーバ 3 の識別子「D」を記憶する。

20

【0026】

このように、複製装置 2 は、データのハッシュ値に基づいて選択されたサーバ 3 に 1stレプリカを配置する。したがって、複製装置 2 は、1stレプリカを用いて一連のデータを多くのサーバ 3 に分散して配置することができる。このため、例えば色の分布を調べるなどの動画解析で、分散ストレージ 1 は、データを並列に読み出すことができ、高速にデータを読み出すことができる。

【0027】

また、複製装置 2 は、データの最終送付先に 2ndレプリカを配置する。したがって、複製装置 2 は、2ndレプリカを用いて一連のデータを同一のサーバ 3 に配置することができる。このため、例えば複数のファイルに分割して分散ストレージ 1 に格納した動画を再生する場合など、分散ストレージ 1 は、一連のファイルを同一サーバ 3 から高速に読み出すことができる。

30

【0028】

また、複製装置 2 は、最終送付先が 1stレプリカの配置先と同じである場合には、ハッシュ空間を示すリングで最終送付先の次のサーバ 3 に 2ndレプリカを配置する。したがって、複製装置 2 は、1stレプリカと 2ndレプリカを異なるサーバ 3 に配置することができる。

【0029】

次に、複製装置 2 の機能構成について説明する。図 3 は、複製装置 2 の機能構成を示す図である。図 3 に示すように、複製装置 2 は、受信部 2 1 と、データ分散テーブル 2 2 と、第 1 配置先決定部 2 3 と、最終送付先記憶部 2 4 と、第 2 配置先決定部 2 5 と、転送部 2 6 と、読出部 2 7 と、テーブル同期部 2 8 とを有する。

40

【0030】

受信部 2 1 は、クライアント 5 から分散ストレージ 1 へのアクセス要求を受信し、データの書き込み要求であればアクセス要求を第 1 配置先決定部 2 3 及び第 2 配置先決定部 2 5 に渡し、データの読み出し要求であればアクセス要求を読出部 2 7 に渡す。

【0031】

データ分散テーブル 2 2 は、レプリカの配置先の決定に用いられるテーブルであり、ハ

50

ハッシュ範囲と各レプリカの配置先のサーバ3の識別子を対応させて記憶する。図2に示した例では、データ分散テーブル22は、ハッシュ範囲毎に、1stと2ndを記憶する。1stは、1stレプリカの配置先のサーバ3の識別子を示し、2ndは、2ndレプリカの配置先のサーバ3の識別子を示す。1stは、ハッシュ範囲から静的に決定される配置先を示し、2ndは、ハッシュ範囲、最終送付先などから動的に決定される配置先を示す。2ndの初期値は、空白である。

【0032】

図2に示した例では、データ分散テーブル22は、ハッシュ範囲「000～aaa」に対応付けて、1stレプリカの配置先のサーバ3の識別子として静的に決定される「A」を記憶し、2ndレプリカの配置先のサーバ3の識別子はこの時点では記憶していない。また、データ分散テーブル22は、ハッシュ範囲「aaa～afa」に対応付けて、1stレプリカの配置先のサーバ3の識別子として静的に決定される「C」を記憶し、2ndレプリカの配置先のサーバ3の識別子として動的に決定された「D」を記憶する。

【0033】

データ分散テーブル22は、複数の複製装置2で同一の情報を記憶する。したがって、データ分散テーブル22の情報は、複製装置2の間で同期がとられる。また、データ分散テーブル22の行数はサーバ数に比べて非常に大きく、例えばサーバ数の1万倍程度である。

【0034】

第1配置先決定部23は、データの書き込み要求を受信部21から受け取り、書き込みデータからハッシュ値を計算する。そして、第1配置先決定部23は、1stレプリカの配置先をハッシュ値及びデータ分散テーブル22に基づいて決定し、決定した配置先を書き込みデータと共に転送部26に渡す。

【0035】

最終送付先記憶部24は、2ndレプリカの最終送付先のサーバ3の識別子を記憶する。複製装置2はクライアント5と1対1に対応するため、最終送付先記憶部24は、クライアント5毎に最終送付先のサーバ3の識別子を記憶する。最終送付先記憶部24の初期値はランダムに決定される。

【0036】

第2配置先決定部25は、データの書き込み要求を受信部21から受け取り、書き込みデータからハッシュ値を計算する。そして、第2配置先決定部25は、2ndレプリカの配置先をハッシュ値、データ分散テーブル22及び最終送付先記憶部24に基づいて決定し、決定した配置先を書き込みデータと共に転送部26に渡す。

【0037】

第2配置先決定部25は、クライアント5から書き込みが行われる一連のデータの2ndレプリカが、なるべく同一のサーバ3に配置されるように配置先を決定する。具体的には、第2配置先決定部25は、データのハッシュ値に対応するデータ分散テーブル22の2ndの列が空白ならば、2ndの列を最終送付先記憶部24が記憶する値に設定し、最終送付先に2ndレプリカを配置する。これで、このクライアント5から前回書かれた2ndレプリカと同じサーバ3に2ndレプリカが配置されることになる。

【0038】

ただし、1st列と最終送付先記憶部24が記憶する値が同一の場合は、そのままでは2つのレプリカが同一のサーバ3に置かれてしまう。そのような場合には、第2配置先決定部25は、2ndレプリカを前回書いたサーバ3に続けて書くことを諦めて、最終送付先記憶部24が記憶する値をハッシュ空間を示すリング上で時計方向に一つ移動した値に変更する。

【0039】

データのハッシュ値に対応するデータ分散テーブル22の2ndの列が最終送付先記憶部24が記憶する値と等しいならば、第2配置先決定部25は、最終送付先記憶部24が示すサーバ3に2ndレプリカを配置する。なお、2ndの列が最終送付先記憶部24が記憶す

10

20

30

40

50

る値と等しいときは、それまでの処理により 1st列と 2nd列の値が等しいことはない。

【 0 0 4 0 】

データのハッシュ値に対応するデータ分散テーブル 2 2 の 2ndの列が空白でもなく、最終送付先記憶部 2 4 が記憶する値と等しくもない場合は、第 2 配置先決定部 2 5 は、このままでは、前回書いたサーバ 3 に 2ndレプリカを続けて書くことができない。そこで、第 2 配置先決定部 2 5 は、データ分散テーブル 2 2 のハッシュ値に対応する行の担当範囲を複数の行に分割することを所定の最大回（例えば 1 0 回）に渡って試みる。

【 0 0 4 1 】

具体的には、第 2 配置先決定部 2 5 は、ハッシュ値に対応する行を n（例えば 1 0 0）行に分け、それまで 2ndレプリカを保持していたサーバ 3（2nd列に書かれているサーバ 3）に増やした行のどの行に対応するデータを保持しているかを聞く。そして、第 2 配置先決定部 2 5 は、2ndレプリカを保持していたサーバ 3 が回答した行の 2nd列に分割前の 2nd列に書かれていた値を設定する。これにより、データのハッシュ値に対応するデータ分散テーブル 2 2 の 2ndの列が空白になり、最終送付先と同じサーバ 3 に 2ndレプリカを配置することができる可能性がある。

【 0 0 4 2 】

ただし、所定の最大回分割を試みても 2nd列が空白にならない場合には、第 2 配置先決定部 2 5 は、2ndレプリカを前回書いたサーバ 3 に続けて書くことを諦める。そして、第 2 配置先決定部 2 5 は、2nd列の示すサーバ 3 にデータを配置すると共に、最終送付先記憶部 2 4 をそのサーバ 3 の識別子に更新する。

【 0 0 4 3 】

このように、第 2 配置先決定部 2 5 は、2ndレプリカの配置先を予め決めておくのではなく、データを蓄積していく過程で決めて行く。したがって、分散ストレージ 1 は、2ndレプリカの配置状況に応じて適切に新たな 2ndレプリカを配置することができる。

【 0 0 4 4 】

図 4 は、第 2 配置先決定部 2 5 がデータ分散テーブル 2 2 の行を増やす場合の例を示す図である。ハッシュ値が「a d a」のデータ Y を書き込み済で、データ分散テーブル 2 2 が図 4（a）に示す状態であるとする。なお、ハッシュ範囲「p p p ~ q q q」は、ハッシュ値が「p p p」以上「q q q」未満を示す。

【 0 0 4 5 】

このとき、最終送付先記憶部 2 4 の値 = B である複製装置 2 から、ハッシュ値が「a b c」であるデータ X を書き込む場合、データ X のハッシュ値「a b c」を担当するのは、データ分散テーブル 2 2 のハッシュ範囲から 2 行目になる。したがって、図 4（a）の 2 行目の 1st列の値「C」より、1stレプリカはサーバ C に配置される。

【 0 0 4 6 】

データ分散テーブル 2 2 の 2 行目の 2nd列は D であり、空白でも、最終送付先記憶部 2 4 が記憶する値 = B と等しくもないので、第 2 配置先決定部 2 5 は、データ分散テーブル 2 2 の行数を増やす。この例では、第 2 配置先決定部 2 5 は 4 行増やす。

【 0 0 4 7 】

そして、第 2 配置先決定部 2 5 は、元々の 2 行目の 2ndレプリカに対応するデータを持っているサーバ 3 の識別子は「D」なので、行数を増やした部分のどの行に 2nd列の識別子を設定すべきかをサーバ D に聞き、サーバ D が回答した行の 2nd列に「D」を設定する。

【 0 0 4 8 】

図 4（b）は、第 2 配置先決定部 2 5 によるデータ分散テーブル 2 2 のアップデート結果を示す。図 4（b）に示すように、ハッシュ値が「a d a」であるデータ Y は、行を増加させたデータ分散テーブル 2 2 では 5 行目で管理されるので、5 行目の 2ndは「D」になる。一方、その他の増加した行に対応する 2ndレプリカは存在しないので、その他の増加した行に対応する 2nd列は空白となる。

【 0 0 4 9 】

10

20

30

40

50

そして、図4(b)のデータ分散テーブル22でデータXのハッシュ値「abc」を担当するのは、3行目になる。3行目の2nd列は空白であり、第2配置先決定部25は、最終送付先記憶部24が示す、前回2ndレプリカを書き込んだサーバBを2ndレプリカの配置先とし、データ分散テーブル22を図4(c)に示すように更新する。

【0050】

図3に戻って、転送部26は、第1配置先決定部23又は第2配置先決定部25からデータとデータの送信先指定とを受け取り、指定された送信先サーバ3にデータを送信する。

【0051】

読出部27は、受信部21からデータの読み出し要求を受け取り、データを読み出すサーバ3を特定する。データの読み出し要求には同一サーバ3から読み出したいデータであるか否かが指定されており、読出部27は、同一サーバ3から読み出したいデータであるか否かの指定とデータ分散テーブル22に基づいてデータを読み出すサーバ3を特定する。

10

【0052】

すなわち、読出部27は、同一サーバ3から読み出したいデータである場合には、データ分散テーブル22の2nd列を用いてデータを読み出すサーバ3を特定する。一方、同一サーバ3から読み出したいデータでない場合には、読出部27は、データ分散テーブル22の1st列を用いてデータを読み出すサーバ3を特定する。そして、読出部27は、特定したサーバ3からデータを読み出す。

20

【0053】

テーブル同期部28は、他の複製装置2との間でデータ分散テーブル22の情報の同期をとる。テーブル同期部28は、行の増加や2nd列の更新などがあると、他の複製装置2との間で同期を行う。

【0054】

次に、2ndレプリカの配置処理のフローについて説明する。図5は、2ndレプリカの配置処理のフローを示すフローチャートである。図5に示すように、第2配置先決定部25は、まずcountを「1」で初期化する(ステップS1)。ここで、countは、データ分散テーブル22の行を分割する回数を数えるためのカウンタである。

【0055】

30

そして、第2配置先決定部25は、データのハッシュ値を求め、データ分散テーブル22のハッシュ範囲がデータのハッシュ値を含む行のidを特定する(ステップS2)。ここで、idは行の番号である。そして、第2配置先決定部25は、データ分散テーブル22の2nd列のid行の識別子を判定する(ステップS3)。

【0056】

その結果、データ分散テーブル22の2nd列のid行の識別子が最終送付先記憶部24が記憶する値と等しい場合には、転送部26が、データをデータ分散テーブルの2nd列のid行が示すサーバ3に転送する(ステップS7)。

【0057】

また、データ分散テーブル22の2nd列のid行の識別子が空白である場合には、第2配置先決定部25は、データ分散テーブル22のid行の1st列の識別子と最終送付先記憶部24が記憶する値が一致するか否かを判定する(ステップS4)。その結果、一致する場合には、第2配置先決定部25は、最終送付先記憶部24が記憶する値をリング上で次のサーバ3の識別子に変更し(ステップS5)、一致しない場合には、ステップS6に進む。

40

【0058】

そして、第2配置先決定部25は、データ分散テーブル22の2nd列のid行に最終送付先記憶部24が記憶する値を代入する(ステップS6)。そして、転送部26が、データをデータ分散テーブルの2nd列のid行が示すサーバ3に転送する(ステップS7)。

【0059】

50

また、データ分散テーブル 22 の 2nd列の i d 行の識別子が最終送付先記憶部 24 が記憶する値と等しくなく、かつ、空白でもない場合には、第 2 配置先決定部 25 は、count に「1」を加える（ステップ S8）。そして、第 2 配置先決定部 25 は、count の値が所定の最大回数より大きいかなかを判定する（ステップ S9）。

【0060】

その結果、count の値が所定の最大回数より大きくない場合には、第 2 配置先決定部 25 は、データ分散テーブル 22 の i d 行を n 行に分ける（ステップ S10）。ここで、n は例えば 100 である。また、分けて得られる各行の 1st列は元と同じであり、2nd列は空白である。例えば、図 4（b）では、n を 5 として、ハッシュ範囲「aaaa~afaa」の行が 5 行に分けられ、分けられた各行の 1st列は元のハッシュ範囲「aaaa~afaa」の 1st列と同じ「C」となる。

10

【0061】

そして、第 2 配置先決定部 25 は、i d 行に相当するデータを格納しているサーバ 3 に対し、増やしたどの行に対応させてデータを保持しているかを聞き、その行の 2nd列にそのサーバ 3 の識別子を設定する（ステップ S11）。例えば、図 4（b）では、ハッシュ値が「ada」であるデータ Y は、「サーバ D」が格納しているので、第 2 配置先決定部 25 は、増やしたどの行に対応させてデータを保持しているかを「サーバ D」に聞く。そして、第 2 配置先決定部 25 は、「サーバ D」から 5 行目であると知られると、5 行目の 2nd列を「D」に設定する。そして、第 2 配置先決定部 25 は、ステップ S2 に戻る。そして、データ X のハッシュ値「abc」を担当するのは、データ分散テーブル 22 の 3 行目になるので、ステップ S2 において、第 2 配置先決定部 25 は、i d を「3」として特定する。そして、ステップ S6 において、第 2 配置先決定部 25 は、図 4（c）に示すように、データ分散テーブル 22 の 3 行目に、最終送付先記憶部 24 が記憶する「B」を代入する。

20

【0062】

一方、count の値が所定の最大回数より大きい場合には、転送部 26 が、データをデータ分散テーブル 22 の 2nd列の i d 行が示すサーバ 3 に転送する（ステップ S12）。また、第 2 配置先決定部 25 は、最終送付先記憶部 24 の値を転送先のサーバ 3 の識別子に更新する（ステップ S13）。

【0063】

30

このように、第 2 配置先決定部 25 がハッシュ値とデータ分散テーブル 22 と最終送付先記憶部 24 に基づいて 2ndレプリカの配置を決定することで、分散ストレージ 1 は 2ndレプリカを用いて一連のデータをできる限り同一のサーバ 3 に格納することができる。

【0064】

次に、読み出し処理のフローについて説明する。図 6 は、読み出し処理のフローを示すフローチャートである。図 6 に示すように、読出部 27 は、データのハッシュ値を求め、データ分散テーブル 22 の該当行の i d を特定する（ステップ S21）。

【0065】

そして、読出部 27 は、読み出し要求に基づいて、同一サーバ 3 から読み出したいデータであるかなかを判定する（ステップ S22）。その結果、同一サーバ 3 から読み出したいデータである場合には、読出部 27 は、データ分散テーブル 22 の 2nd列の i d 行が示すサーバ 3 からデータを読み出す（ステップ S23）。

40

【0066】

一方、同一サーバ 3 から読み出したいデータでない場合には、読出部 27 は、データ分散テーブル 22 の 1st列の i d 行が示すサーバ 3 からデータを読み出す（ステップ S24）。

【0067】

このように、読出部 27 が読み出し要求に基づいて 1stレプリカ又は 2ndレプリカを読み出すことによって、分散ストレージ 1 はクライアント 5 の要求に適したレプリカを用いて高速にデータを読み出すことができる。

50

【 0 0 6 8 】

上述してきたように、実施例では、第 1 配置先決定部 2 3 がデータが分散して配置されるように 1stレプリカの配置先を決定する。また、第 2 配置先決定部 2 5 がデータが同一のサーバ 3 に連続して記憶され、かつ 1stレプリカと異なるサーバに記憶されるように 2ndレプリカの配置を決定する。したがって、分散ストレージ 1 は、一連のデータを多くのサーバに配置したい場合と、少数のサーバに配置したい場合の両方に対応すると共に、同一データのレプリカを同一サーバに記憶させないようにすることができる。

【 0 0 6 9 】

また、実施例では、データ分散テーブル 2 2 が 1stレプリカと 2ndレプリカの配置先のサーバ 3 の識別子をハッシュ範囲毎に記憶し、最終送付先記憶部 2 4 が 2ndレプリカの最終配布先のサーバ 3 の識別子を記憶する。そして、第 2 配置先決定部 2 5 は、データのハッシュ値、データ分散テーブル 2 2 及び最終送付先記憶部 2 4 に基づいて 2ndレプリカの配置先を決定する。したがって、第 2 配置先決定部 2 5 は、データが同一のサーバ 3 に連続して記憶され、かつ 1stレプリカと異なるサーバに記憶されるように 2ndレプリカの配置を決定することができる。

10

【 0 0 7 0 】

また、実施例では、ハッシュ値に基づく 2ndレプリカの配置先と、最終送付先記憶部 2 4 に基づく最終送付先が異なる場合に、第 2 配置先決定部 2 5 は、ハッシュ範囲を分割して両者が一致するよう試みる。したがって、第 2 配置先決定部 2 5 は、一連のデータができるだけ同一のサーバ 3 に記憶されるように、2ndレプリカを配置することができる。

20

【 0 0 7 1 】

また、実施例では、第 2 配置先決定部 2 5 は、ハッシュ範囲の分割を所定の回数繰り返してもハッシュ値に基づく配置先と最終送付先を一致させることができない場合に、ハッシュ値に基づく配置先に 2ndレプリカを配置する。したがって、データが同一のサーバ 3 だけに過度に集中することを防ぐことができる。

【 0 0 7 2 】

なお、実施例では、複製装置 2 について説明したが、複製装置 2 が有する構成をソフトウェアによって実現することで、同様の機能を有する複製プログラムを得ることができる。そこで、複製プログラムを実行するコンピュータについて説明する。なお、コンピュータは、複数の複製プログラムを実行することができる。

30

【 0 0 7 3 】

図 7 は、実施例に係る複製プログラムを実行するコンピュータのハードウェア構成を示す図である。図 7 に示すように、コンピュータ 3 0 は、メインメモリ 3 1 と、C P U (Central Processing Unit) 3 2 と、L A N インタフェース 3 3 と、H D D 3 4 とを有する。また、コンピュータ 3 0 は、スーパー I O (Input Output) 3 5 と、D V I (Digital Visual Interface) 3 6 と、O D D (Optical Disk Drive) 3 7 とを有する。

【 0 0 7 4 】

メインメモリ 3 1 は、プログラムやプログラムの実行途中結果などを記憶するメモリである。C P U 3 2 は、メインメモリ 3 1 からプログラムを読み出して実行する中央処理装置である。C P U 3 2 は、メモリコントローラを有するチップセットを含む。

40

【 0 0 7 5 】

L A N インタフェース 3 3 は、コンピュータ 3 0 を L A N 経由で他のコンピュータに接続するためのインタフェースである。H D D 3 4 は、プログラムやデータを格納するディスク装置であり、スーパー I O 3 5 は、マウスやキーボードなどの入力装置を接続するためのインタフェースである。D V I 3 6 は、液晶表示装置を接続するインタフェースであり、O D D 3 7 は、D V D の読み書きを行う装置である。

【 0 0 7 6 】

L A N インタフェース 3 3 は、P C I エクスプレスにより C P U 3 2 に接続され、H D D 3 4 及び O D D 3 7 は、S A T A (Serial Advanced Technology Attachment) により C P U 3 2 に接続される。スーパー I O 3 5 は、L P C (Low Pin Count) により C

50

P U 3 2 に接続される。

【 0 0 7 7 】

そして、コンピュータ 3 0 において実行される複製プログラムは、D V D に記憶され、O D D 3 7 によって D V D から読み出されてコンピュータ 3 0 にインストールされる。あるいは、複製プログラムは、L A N インタフェース 3 3 を介して接続された他のコンピュータシステムのデータベースなどに記憶され、これらのデータベースから読み出されてコンピュータ 3 0 にインストールされる。そして、インストールされた複製プログラムは、H D D 3 4 に記憶され、メインメモリ 3 1 に読み出されて C P U 3 2 によって実行される。

【 0 0 7 8 】

10

また、実施例では、複製装置が分散ストレージに含まれる場合について説明したが、本発明はこれに限定されるものではなく、例えば、複製プログラムがクライアントに近接するコンピュータやクライアントで実行される場合にも同様に適用することができる。

【 0 0 7 9 】

また、実施例では、ハッシュを用いてレプリカの配置を決定する場合について説明したが、本発明はこれに限定されるものではなく、例えば、順番に配置サーバを交替していく場合など他の方法でレプリカの配置を決定する場合にも同様に適用することができる。

【 0 0 8 0 】

また、実施例では、複製装置とクライアントが 1 対 1 に対応する場合について説明したが、本発明はこれに限定されるものではなく、 m 及び n を任意の正の整数として、複製装置とクライアントが m 対 n に対応する場合にも同様に適用することができる。

20

【符号の説明】

【 0 0 8 1 】

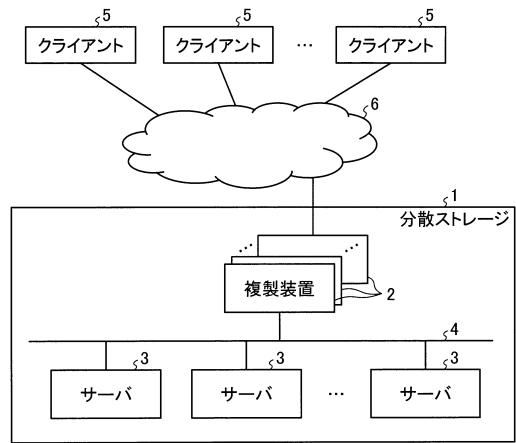
- 1 分散ストレージ
- 2 複製装置
- 3 サーバ
- 4 L A N
- 5 クライアント
- 2 1 受信部
- 2 2 データ分散テーブル
- 2 3 第 1 配置先決定部
- 2 4 最終送付先記憶部
- 2 5 第 2 配置先決定部
- 2 6 転送部
- 2 7 読出部
- 2 8 テーブル同期部
- 3 0 コンピュータ
- 3 1 メインメモリ
- 3 2 C P U
- 3 3 L A N インタフェース
- 3 4 H D D
- 3 5 スーパー I O
- 3 6 D V I
- 3 7 O D D

30

40

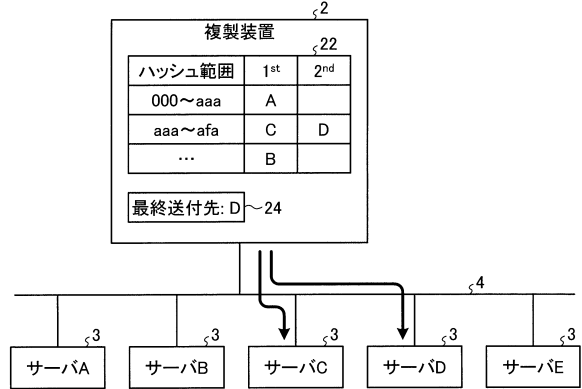
【図 1】

実施例に係る分散ストレージの構成を示す図



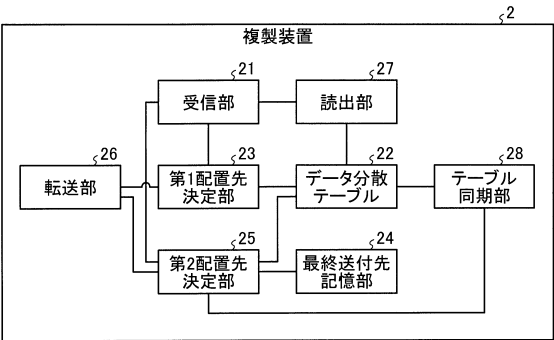
【図 2】

実施例に係る分散ストレージによるレプリカ配置を説明するための図



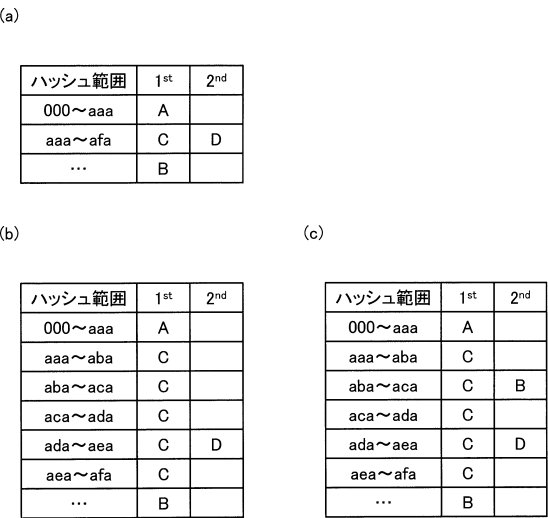
【図 3】

複製装置の機能構成を示す図

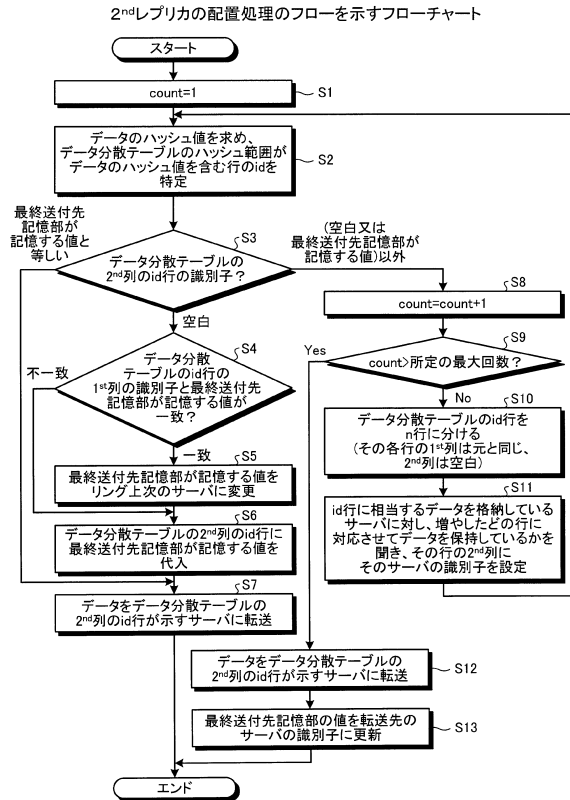


【図 4】

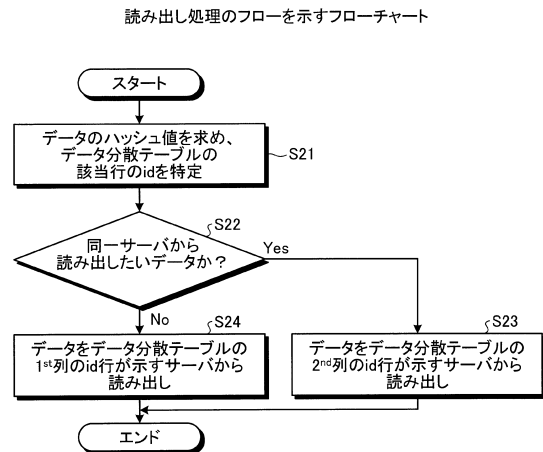
第2配置先決定部がデータ分散テーブルの行を増やす場合の例を示す図



【 図 5 】

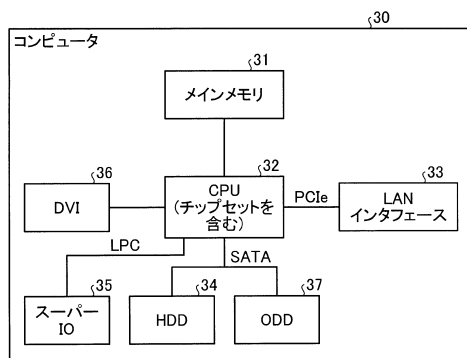


【 図 6 】



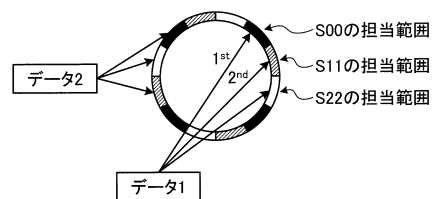
【 図 7 】

実施例に係る複製プログラムを実行するコンピュータのハードウェア構成を示す図



【 図 8 】

リング形態のハッシュ空間の一例を示す図



フロントページの続き

(56)参考文献 国際公開第2011/027775(WO,A1)
特開2013-182575(JP,A)
国際公開第2011/136261(WO,A1)

(58)調査した分野(Int.Cl.,DB名)
G06F 12/00