(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第5803908号
(P5803908)

(45)発行日　平成27年11月4日(2015.11.4)　(24)登録日　平成27年9月11日(2015.9.11)

(51)Int.Cl.		F I		
G06F 12/00	(2006.01)	G06F	12/00	545A
G06F 3/06	(2006.01)	G06F	3/06	301Z
G06F 13/10	(2006.01)	G06F	13/10	340A

請求項の数 6　(全 15 頁)

(21)出願番号　特願2012-512877 (P2012-512877)
(86)(22)出願日　平成23年4月27日(2011.4.27)
(86)国際出願番号　PCT/JP2011/060238
(87)国際公開番号　WO2011/136261
(87)国際公開日　平成23年11月3日(2011.11.3)
審査請求日　平成26年2月13日(2014.2.13)
(31)優先権主張番号　特願2010-103859 (P2010-103859)
(32)優先日　平成22年4月28日(2010.4.28)
(33)優先権主張国　日本国(JP)

(73)特許権者　000004237
日本電気株式会社
東京都港区芝五丁目７番１号
(74)代理人　100106909
弁理士　棚井　澄雄
(74)代理人　100134544
弁理士　森　隆一郎
(74)代理人　100149548
弁理士　松沼　泰史
(74)代理人　100162868
弁理士　伊藤　英輔
(72)発明者　鳥居　隆史
東京都港区芝五丁目７番１号　日本電気株式会社内

最終頁に続く

(54)【発明の名称】ストレージシステム及びストレージシステムの制御方法

(57)【特許請求の範囲】
【請求項１】
複数のストレージノードと管理サーバを具備するストレージシステムにおいて、
前記各ストレージノードは、
データを格納する第１の記憶部と、
ネットワーク的な距離が所定の範囲内にある前記ストレージノードからなるネットワークグループと、前記データに対応するデータ識別情報に基づくストレージグループと、に関するノード情報を格納する第２の記憶部と、を具備し、
前記管理サーバは、
複数の前記ストレージグループの何れに属するかを複数の前記ストレージノードのそれぞれに対して特定し、前記ネットワーク的な距離が所定の範囲内にある複数の前記ネットワークグループの何れに属するかを複数の前記各ストレージノードに対して特定し、複数の前記各ストレージノードに、前記特定した前記ストレージグループの識別情報と前記特定した前記ネットワークグループの識別情報とを割り当て、
前記各ストレージノードは、該ストレージノードが属する前記ネットワークグループ内の全ての前記ストレージノードの情報である当該全ての前記ストレージノードの識別情報と、該ストレージノードが属する前記ストレージグループ内の全ての前記ストレージノードの情報である当該全ての前記ストレージノードの識別情報とを含む前記ノード情報を前記第２の記憶部に記憶し、
前記各ストレージノードは、前記データのアクセス要求を受信したときに、前記データ

のアクセス要求に基づいて、前記ノード情報に含まれる前記ネットワークグループと前記ストレージグループとに関する情報である当該データのアクセスに用いるデータ識別子を分散関数に入力して得られる値と、前記ノード情報とを参照して前記データへのアクセス処理を行う、ストレージシステム。

【請求項２】

前記各ストレージノードは、該ストレージノードが属する前記ネットワークグループ内の全ての前記ストレージノードの情報と、該ストレージノードが属する前記ストレージグループ内の全ての前記ストレージノードの情報とを含む前記ノード情報を前記第２の記憶部に記憶する、請求項１に記載のストレージシステム。

【請求項３】

前記各ストレージノードは、前記データへのアクセス要求に基づいて、当該データのアクセスに用いるデータ識別子を分散関数に入力して得られる値に基づいて前記ストレージグループの識別情報を特定し、当該ストレージグループの識別情報と自ノードと同一のネットワークグループの識別情報が割り当てられた他ストレージノードに対して、前記アクセス要求の対象となるデータのリクエスト要求を行う、

請求項１または請求項２に記載のストレージシステム。

【請求項４】

データを格納する複数のストレージノードを含むストレージシステムにおけるストレージシステムの制御方法であって、

前記データのアクセス要求を受信したときに、ネットワーク的な距離が所定の範囲内にある前記ストレージノードからなるネットワークグループと、前記データに対応するデータ識別情報に基づくストレージグループと、に関するノード情報を参照するステップと、

前記ノード情報に基づいて前記データへのアクセス処理を行うステップと、

を具備し、

前記アクセス要求を受信した前記ストレージノードにおいて、前記データ識別情報に基づいて前記ストレージグループを判定するステップと、

前記ノード情報を参照し、前記アクセス要求を受信した前記ストレージノードが属している前記ネットワークグループ内で前記判定した前記ストレージグループに属する前記ストレージノードを検出するステップと、

前記検出されたストレージノードに前記データへのアクセスを要求するステップと、

前記検出されたストレージノードからの応答に基づいて前記アクセス要求に対して応答するステップと、

を更に具備する、ストレージシステムの制御方法。

【請求項５】

前記検出されたストレージノードにおいて、前記アクセス要求に対応するデータを前記検出されたストレージノードが格納している場合は、前記アクセス要求に対応するデータを、前記アクセス要求を受信した前記ストレージノードに送信するステップと、

前記検出されたストレージノードにおいて、前記アクセス要求に対応するデータを前記検出されたストレージノードが格納していない場合は、前記ノード情報を参照して、前記検出されたストレージノードが属する前記ストレージグループ内の他の前記ストレージノードに対して前記アクセス要求に対応するデータを要求し、前記アクセス要求を受信した前記ストレージノードに前記アクセス要求に対応するデータを送信するステップと、

を更に具備する、請求項４に記載のストレージシステムの制御方法。

【請求項６】

前記検出されたストレージノードにおいて、前記アクセス要求に対応するデータを前記検出されたストレージノードが格納している場合は、前記アクセス要求を受信した前記ストレージノードに前記アクセス要求に対応するデータを送信するステップと、

前記検出されたストレージノードにおいて、前記アクセス要求に対応するデータを前記検出されたストレージノードが格納していない場合は、前記ノード情報を参照して、前記検出されたストレージノードが属する前記ストレージグループ内の前記アクセス要求に対

応するデータを格納している他の前記ストレージノードを検索し、前記検索されたストレージノードを、前記アクセス要求を受信した前記ストレージノードに通知するステップと、
を更に具備する、請求項４に記載のストレージシステムの制御方法。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、大量のデータを保持するストレージシステム及びストレージシステムの制御方法に関する。
本願は、２０１０年４月２８日に、日本に出願された特願２０１０－１０３８５９号に基づき優先権を主張し、その内容をここに援用する。
【背景技術】
【０００２】
ネット上で提供されるデータ量の増加に伴って、大量のデータを保持するストレージが必要とされている。例えば、ウェブ検索サービスを提供する企業では、複数のサーバを並列に並べた分散ストレージ技術を採用している。この分散ストレージ技術は、数千のノード（「ピア」ともいう）にデータを分散して配置し、全体として一つの大きなストレージを構成する技術である。また、分散ストレージ技術は、高価なストレージ専用装置ではなく、比較的安価なサーバを複数並べることによって大容量のストレージを実現することができる技術として、扱うデータ量が増大しているエンタープライズやキャリアの事業分野でも注目されている技術である。一部のペタ（10^{15}）バイトを超えた大容量のデータを格納する事業分野では、ストレージ専用装置においてデータを格納することができる容量がボトルネックとなってしまうため、大量のデータ格納を実現するための方法として、分散ストレージ技術を使用するしか解がないというケースも出始めている。
【０００３】
しかし、分散ストレージ技術においては、データが複数のノードに分散している。このため、データにアクセスしようとするクライアントは、まず、データを持っているノードの位置を知る必要がある。従って、近年注目されている分散ストレージ技術においては、データを保持しているノードの位置を知るための方法が、技術的なポイントとなっている。
【０００４】
データを保持しているノードの位置を知るための一つの方法として、データの位置情報を管理するメタサーバを設けるメタサーバ方式がある。このメタサーバ方式では、ストレージシステムの構成が大規模になることに伴って、データを格納しているノードの位置を検出するメタサーバの性能がボトルネックになることがある。
【０００５】
そのため、データを保持しているノードの位置を知るための別の方法として、分散関数（例えば、ハッシュ関数）を用いてデータの位置を求める方法が注目されている。これは、以前からＰ２Ｐ（Ｐｅｅｒ　ｔｏ　Ｐｅｅｒ：ピアツーピア）ネットワークの分野で用いられている分散ハッシュテーブル（Ｄｉｓｔｒｉｂｕｔｅｄ　Ｈａｓｈ　Ｔａｂｌｅ：ＤＨＴ）の技術を利用しており、キーバリューストレージ（Ｋｅｙ－Ｖａｌｕｅストレージ：ＫＶＳ）と呼ばれている。このＫＶＳは、データにアクセスする識別子をキー（Ｋｅｙ）、データをバリュー（Ｖａｌｕｅ）とし、キー（Ｋｅｙ）を分散関数にかけて、すなわち、キーを分散関数の入力値として、データを格納しているノード（以下、「データ格納ノード」という）の位置を算術的に求める。
【０００６】
ＫＶＳのメタサーバ方式との違いは、メタサーバ方式が全てのデータの位置情報をもたなければならないのに対し、ＫＶＳでは、全てのクライアントで分散関数とノードリストとを共有すればよいだけなので、分散関数とノードリストとを共有する際のコストが小さく、性能的なボトルネックがないことである。ＫＶＳを用いれば、メタサーバ方式のよう

なメタサーバの性能によるボトルネックがなく、ストレージシステムが大規模になった場合でも、性能拡張性（スケーラビリティ）のある大容量ストレージを実現することができる（特許文献１～３参照）。
【先行技術文献】
【特許文献】
【０００７】
【特許文献１】特開２００９－２８９１６１号公報
【特許文献２】特開２００９－１５１４０３号公報
【特許文献３】特開２００８－２６９１４１号公報
【発明の概要】
【発明が解決しようとする課題】
【０００８】
既存のＫＶＳ技術では、分散関数としてハッシュ関数（あるいは、それに類似の分散関数）を用いて、算術的にデータ格納ノードが決定されている。また、ＤＨＴ技術では、分散ハッシュテーブルのルーティングにより、オーバーレイルーティングが行われている。このため、既存のＫＶＳ技術を用いたストレージシステムでは、大きく分けて、例えば、以下のような２つの点がみられる。
【０００９】
まず、１つ目の点は、既存のＫＶＳ技術ではデータ配置の自由度が高いとはいえないことがある点である。ＫＶＳ技術は、大規模のストレージシステムで利用されるため、必然的に各ストレージが大規模ネットワークに分散して配置されることとなる。すると、データにアクセスするノード（クライアント）とデータを保持しているデータ格納ノードとのネットワーク上の距離が遠くなるケースがあり得る。クライアントとデータ格納ノードとの間の遅延が大きければ、ストレージシステムにおける処理の速度も遅くなる。このため、クライアントとデータ格納ノードとをできるだけ近くに配置することによってストレージシステムの処理速度を向上することが求められている。そのためには、データ配置を自由に変更できることが求められる。
また、データを分散させて保持するのではなく、逆にデータを予め定めたデータ格納ノードに集めて、空いているデータ格納ノードを作ることによって、ストレージシステムの省電力化を実現することも考えられる。このようなときにも、データ配置を自由に変更することができれば、ストレージシステムの省電力モードによる制御も可能となる。しかしながら、既存のＫＶＳ技術では、データ格納ノードがハッシュ関数によって決まってしまうため、例えば、あるデータを特定のデータ格納ノードに配置しようとしても、データ配置を自由に制御することができない可能性がある。
【００１０】
そして、２つ目の点は、分散ハッシュテーブルのオーバーレイルーティングが実際のネットワークトポロジーと連動していないため、効率的でないルーティングがなされて、結果的に性能が高いものではないことがある点である。極端な例では、東京から大阪に行くのにアメリカを経由するというようなこともあり得る。
【００１１】
従来、スケーラブルなストレージシステムにおいて、データ配置の自由度を向上させることができるストレージシステムの制御方法が求められていた。
【課題を解決するための手段】
【００１２】
本発明は、例えば、以下の側面を有しても良い。ただし、以下の記述は本発明を限定するものではない。
第１の側面はストレージシステムの制御方法であって、データを格納する複数のストレージノードによって構成されたストレージシステムの制御方法において、該ストレージシステムに含まれる前記複数のストレージノードを、該ストレージシステム内においてネットワーク的な距離が予め定められた距離の範囲内であるストレージノードからなる第１の

グループと、データを格納しているストレージノードの位置の情報を共有しているストレージノードからなる第２のグループと、にグループ化し、前記第２のグループ毎に該第２のグループを識別する論理的な空間上の識別子を割り当て、データ識別子を分散関数の入力値として前記論理的な空間上の位置を算出し、前記算出された位置に対応する識別子が割り当てられた前記第２のグループに属する前記ストレージノードに該データ識別子に対応するデータを格納する。

【００１３】

また、上記複数のストレージノードのそれぞれは、１つ以上の前記第１のグループと１つ以上の前記第２のグループとに必ず属し、該ストレージノードが属している前記第２のグループ内の他の全てのストレージノードのリストと、該ストレージノードが属している前記第１のグループ内の他の全てのストレージノードのリストとを、記憶してもよい。

【００１４】

また、上記ストレージシステム内でデータのアクセス要求をするストレージノードは、データ識別子を分散関数の入力値として、前記論理的な空間上の位置を算出し、該算出された位置に対応する識別子が割り当てられた前記第２のグループを選択し、該ストレージノードに記憶しているノードリストから、該ストレージノードが属している前記第１のグループ内で前記選択した第２のグループに属している他のストレージノードを検索し、前記検索した他のストレージノードにデータのアクセス要求を出力してもよい。

【００１５】

また、上記ストレージシステム内でデータのアクセス要求を受けたストレージノードは、前記アクセス要求を受けたデータを該ストレージノード内に格納している場合は、該データを要求した前記ストレージノードに要求されたデータを出力し、前記アクセス要求を受けたデータを該ストレージノード内に格納していない場合は、該ストレージノードに記憶しているノードリストから、該ストレージノードが属している前記第２のグループ内で要求されたデータを格納している他のストレージノードを検索し、前記検索した他のストレージノードに、前記データのアクセス要求を転送してもよい。

【００１６】

また、上記ストレージシステム内でデータのアクセス要求を受けたストレージノードは、前記アクセス要求を受けたデータを該ストレージノード内に格納している場合は、該データを要求した前記ストレージノードに要求されたデータを出力し、前記アクセス要求を受けたデータを該ストレージノード内に格納していない場合は、該ストレージノードに記憶しているノードリストから、該ストレージノードが属している前記第２のグループ内で要求されたデータを格納している他のストレージノードを検索し、前記検索した他のストレージノードを、前記アクセス要求をするストレージノードに通知してもよい。

【発明の効果】

【００１７】

上記側面によれば、例えば、スケーラブルなストレージシステムにおいて、データ配置の自由度を向上させることができるという効果が得られる。

【図面の簡単な説明】

【００１８】

【図１】一実施形態によるストレージシステムの概略構成を示したブロック図である。

【図２】従来のストレージシステムにおける既存のＫＶＳ技術を説明する図である。

【図３】一実施形態のストレージシステムにおける論理的なノード構成の一例を示した図である。

【図４】一実施形態のストレージシステムにおけるＫＶＳ技術を説明する図である。

【図５】一実施形態のストレージシステムにおける物理構成の一例を示した図である。

【図６】一実施形態のストレージシステムにおけるデータアクセス処理の流れを示したシーケンス図である。

【発明を実施するための形態】

【００１９】

以下、一実施形態について、図面を参照して説明する。図１は、本実施形態によるストレージシステムの概略構成を示したブロック図である。図１において、ストレージシステム３は、データ格納ノードである複数のストレージノード１を備えている。各ストレージノード１には、１つもしくは複数のＩＤが付与されている。そして、これら複数のストレージノード１が、管理サーバ１０によってストレージシステム３内でマッピングされて、１つのグローバルネームスペースとして構成されている。なお、ストレージシステム３内の各ストレージノード１が設置されている物理的な位置は１カ所ではなく、複数の場所に設置されている各ストレージノード１が、例えば、ネットワークなどによって接続されることによって、１つのグローバルネームスペースを構成している。

【００２０】

クライアント２は、ストレージシステム３のデータにアクセスするノードである。そして、クライアント２は、ストレージシステム３を、１つの大きなストレージとみなしてアクセスする。

【００２１】

ここで、既存のＫＶＳ技術について説明する。図２は、従来のストレージシステムにおける既存のＫＶＳ技術を説明する図である。図２では、それぞれ「ａ」、「ｂ」、「ｃ」、「ｄ」というＩＤが付与されたデータ格納ノードが、論理的な空間の円周上にマッピングされている場合を示している。既存のＫＶＳ技術では、データ識別子であるＫｅｙを、分散関数Ｆにかけて Ｆ（Ｋｅｙ）が求められる。そして、この円周上で、Ｆ（Ｋｅｙ）の位置から右回りに最も近いＩＤを持つデータ格納ノードに、Ｆ（Ｋｅｙ）に対応するデータが保持される。図２においては、ａ＜Ｆ（Ｋｅｙ）≦ｂを満たすＦ（Ｋｅｙ）に対応するデータが、ＩＤとして「ｂ」が付与されたデータ格納ノードに格納されることを示している。

【００２２】

この既存のＫＶＳ技術によるデータ格納の方法は、クライアントが、分散関数Ｆとデータ格納ノードのリストを共有するのみであるので、クライアントが共有する情報が少なくて済むという利点がある。しかし、データ識別子であるＫｅｙは、一度データに付与された後に変更することができないため、任意のデータ格納ノードにデータを移動することができず、データ配置の自由度がない。

【００２３】

次に、本実施形態のストレージシステム３におけるデータ格納の方法について説明する。図３は、本実施形態のストレージシステム３における論理的なノード構成の一例を示した図である。図３に示すように、本実施形態のストレージシステム３では、ストレージシステム３内の各ストレージノード１を、ストレージグループ４およびネットワークグループ５という２種類のグループにグループ化する。このストレージシステム３内における各ストレージノード１のグループ化は、管理サーバ１０によって行われる。

【００２４】

ストレージグループ４は、ストレージシステム３で用いるＫＶＳ技術に基づいてデータを格納しているストレージノード１の位置の情報を共有するストレージノード１で構成されたグループである。

【００２５】

また、ネットワークグループ５は、ストレージシステム３におけるネットワーク的な距離に基づいて、管理サーバ１０によって決定されたグループであり、ネットワーク的な距離が予め定められた距離の範囲内で比較的近いストレージノード１で構成されたグループである。すなわち、ネットワークグループ５に属する任意の２つのストレージノード１間のネットワーク的な距離は、予め定められた距離の範囲内の距離となる。

【００２６】

ストレージシステム３における各ストレージノード１は、管理サーバ１０によって管理されるいずれかのストレージグループ４に属していると同時に、いずれかのネットワークグループ５に属している。

【００２７】
なお、図３において白抜きの○は、ストレージノード１を示しており、各ストレージノード１を表す○内には、ストレージグループ４における識別番号とネットワークグループ５における識別番号とを表している。より具体的には、各ストレージノード１を表す○内の２桁の符号のうち、左側の符号は、１つのストレージグループ４（図３においては、符号“Ｙ”が付与されたストレージグループ４）におけるストレージノード１の識別番号（１，２，・・・，Ｘ，・・・，ｍ）を表し、右側の符号は、１つのネットワークグループ５（図３においては、符号“Ｘ”が付与されたネットワークグループ５）におけるストレージノード１の識別番号（１，２，・・・，Ｙ，・・・，ｎ）を表している。
【００２８】
次に、本実施形態のストレージシステム３におけるＫＶＳ技術について説明する。図４は、本実施形態のストレージシステム３におけるＫＶＳ技術を説明する図である。図４に示すように、本実施形態のストレージシステム３におけるＫＶＳ技術では、複数のストレージノード１をグループ化することによって、データ配置の自由度を増加させる。そして、ストレージシステム３では、管理サーバ１０によって、各ストレージグループ４に１つもしくは複数のＩＤを付与され、ＩＤを付与されたストレージグループ４を、ストレージシステム３内にマッピングする。図４では、それぞれ「Ａ」、「Ｂ」、「Ｃ」、「Ｄ」というＩＤが付与されたストレージグループ４が、図２に示した既存のＫＶＳ技術と同様に、論理的な空間の円周上にマッピングされている場合を示している。
【００２９】
そして、本実施形態のストレージシステム３におけるＫＶＳ技術では、図２に示した既存のＫＶＳ技術と同様に、データ識別子であるＫｅｙを、分散関数Ｆにかけて Ｆ（Ｋｅｙ）を求める。そして、この円周上で、Ｆ（Ｋｅｙ）の位置から右回りに最も近いＩＤを持つストレージグループ４を、データを保持するストレージグループ４と決定する。続いて、決定されたストレージグループ４内のどのストレージノード１がデータを保持するかを決定し、決定されたストレージノード１に、Ｆ（Ｋｅｙ）に対応したデータを保持する。
【００３０】
次に、本実施形態のストレージシステム３におけるネットワークグループ５について説明する。図５は、本実施形態のストレージシステム３における物理構成の一例を示した図である。上記に述べたとおり、ネットワークグループ５は、ネットワーク的な距離が比較的近いストレージノード１をグループ化したものである。このネットワーク的な距離とは、例えば、ネットワーク経路上のスイッチの段数と考えることができる。より具体的には、図５に示したように、複数のストレージノード１とスイッチ７とから構成される複数のラック６と、それぞれのラック６を束ねる上位スイッチ８からなるストレージシステム３を想定する。この場合、各ラック６がそれぞれのネットワークグループ５に相当する。
【００３１】
ストレージシステム３における各ストレージノード１は、上記に述べたとおり、必ず１つ以上のネットワークグループ５に属していると同時に、必ず１つ以上のストレージグループ４に属している。また、各ストレージノード１が属するストレージグループ４は、ネットワークグループ５内に属しているストレージノード１から、ストレージシステム３における全てのストレージグループ４をたどることができるように、管理サーバ１０によって割り当てられている。つまり、ある１つのストレージノード１が属するネットワークグループ５内の全てのストレージノード１の和集合をとれば、全てのストレージグループ４をカバーすることができるように、各ストレージノード１のストレージグループ４が割り当てられている。
【００３２】
なお、各ストレージグループ４やネットワークグループ５に属するストレージノード１の数は、それぞれのストレージグループ４やネットワークグループ５で異なる数であってもよい。例えば、１つのストレージノード１が、複数のストレージグループ４やネットワークグループ５に属するように、このストレージノード１のストレージグループ４やネッ

トワークグループ５を割り当てることもできる。
【００３３】
各ストレージノード１は、自ストレージノード１が属しているストレージグループ４内の他の全てのストレージノード１のリストと、ネットワークグループ５内の他の全てのストレージノード１のリストとを、ノードリストとして記憶している。それぞれのノードリストには、各ストレージノード１が属しているストレージグループ４およびネットワークグループ５のＩＤと、各ストレージノード１のアドレス（位置）情報と、各ストレージノード１が格納しているデータ（例えば、データの一覧など）の情報が含まれている。
【００３４】
次に、本実施形態のストレージシステム３において各ストレージノード１が記憶しているノードリストの数について説明する。上記に述べたとおり、全てのストレージノード１は、自ストレージノード１が属するネットワークグループ５内の全てのストレージノード１のリストと、自ストレージノード１が属するストレージグループ４内の全てのストレージノード１のリストとを自ストレージノード１のメモリ上に記憶している。ストレージシステム３において各ストレージノード１が記憶しているストレージノード１のリストの総数は、従来のストレージシステムに比べて非常に少ない数となるため、ストレージシステム内におけるメモリ容量の削減と、メンテナンスコストの削減とを実現することができる。
【００３５】
より具体的には、例えば、１０００台のストレージノードで構成されたストレージシステムを考える。従来のストレージシステムにおいて全てのストレージノードのノードリストを記憶する場合、各ストレージノードは、１０００個のリストをノードリストとして記憶する必要がある。
【００３６】
これに対して、本実施形態のストレージシステム３では、自ストレージノード１が属するネットワークグループ５内の全てのストレージノード１の個数のリストと、自ストレージノード１が属するストレージグループ４内の全てのストレージノード１の個数のリストとを、ノードリストとして記憶するのみである。例えば、１０００台のストレージノード１が、Ｎグループのストレージグループ４、およびＭグループのネットワークグループ５にそれぞれグループ化され、各ストレージノード１が、それぞれ、１つのストレージグループ４および１つのネットワークグループ５に属している場合を想定する。この場合、各ストレージノード１は、自ストレージノード１が属するネットワークグループ５内のＮ個のストレージノード１のリストと、ストレージグループ４内のＭ個のストレージノード１のリストのみを記憶するのみであるため、各ストレージノード１が記憶するノードリストは、Ｎ＋Ｍ－１個のリストとなる。ここで、－１個としたのは、図３からもわかるように、自ストレージノード１のリストは、ストレージグループ４とネットワークグループ５とで重複しており、この重複を回避するためである。より具体的には、ストレージグループ４が１００グループ、ネットワークグループ５が１０グループであった場合には、１００＋１０－１＝１０９個のリストを各ストレージノード１が記憶するのみとなる。
【００３７】
これは、従来のストレージシステムにおいて各ストレージノード１が１０００個のリストを記憶していたのに対して、本実施形態のストレージシステム３のストレージノード１が記憶するリストの数は、約１０分の１の数であり、各ストレージノード１内でノードリストの記憶に使用するメモリ容量の削減を実現していることとなる。
【００３８】
また、一般的に、ストレージシステムでは、データにアクセスできない時間を極力減らすため、ストレージシステムの死活監視を定期的に行っている。このストレージシステムの死活監視では、ストレージシステム内の各ストレージノードのエラーをなるべく早く検出する必要があり、ノードリストに含まれているストレージノードが正常に稼動しているか否かという稼働状況をチェックすることによって行われている。もし、ストレージシス

テム内のいずれかのストレージノードにエラーが発生している、ストレージシステムのネットワークが不通になっている、などの原因によって、ストレージシステムが正常に稼動していない場合には、ノードリストの変更が必要となる。この稼働状況のチェックにかかるコストは、ノードリストに含まれるリストの数に比例して大きくなるため、リストの数が多くなると、ストレージシステム全体のスケーラビリティを大きく損なう要因となってしまう。そのため、ノードリストに含まれるリストの数を少なく保つことは、スケーラブルなストレージシステムにとっては重要な項目である。本実施形態のストレージシステム３では、各ストレージノード１がノードリストに記憶しているストレージノード１のリストの数が少ないため、メンテナンスコストの削減を実現することができる。

【００３９】

次に、本実施形態のストレージシステム３において各ストレージノード１がデータを保持しているストレージノード１の検索方法について説明する。図６は、本実施形態のストレージシステム３におけるデータアクセス処理の流れを示したシーケンス図である。図６では、図３に示した符号“Ｘ”が付与されたネットワークグループ５（以下、「ネットワークグループＮＧ＿Ｘ」という）に属する識別番号“１”が付与されたストレージノード１（以下、「ストレージノードＸ１」という）がクライアント２となってデータにアクセスする場合について説明する。

【００４０】

まず、ストレージノードＸ１は、データ識別子（Ｋｅｙ）を分散関数Ｆにかけて Ｆ（Ｋｅｙ）を求める（ステップＳ１０）。そして、Ｆ（Ｋｅｙ）に対応するデータを保持しているストレージノード１が属しているストレージグループ４（以下、「ストレージグループＳＧ＿Ｙ」という）を求める（ステップＳ２０）。例えば、「Ａ」、「Ｂ」をストレージグループ４のＩＤとすると、Ａ＜Ｆ（Ｋｅｙ）≦Ｂを満たす「Ｂ」がＩＤとして付与されたストレージグループ４が求められる（図４参照）。

【００４１】

そして、自ストレージノードＸ１が属するネットワークグループＮＧ＿Ｘ内で、ストレージグループＳＧ＿Ｙに属するストレージノード１をノードリストから求める（ステップＳ３０）。図６では、ネットワークグループＮＧ＿Ｘ内でストレージグループＳＧ＿Ｙに属するストレージノードＸＹ（図３参照）が求められたものとする。そして、ストレージノードＸ１は、ストレージノードＸＹにデータの要求（リクエスト）を送信する（ステップＳ４０）。

【００４２】

続いて、リクエストを受信したストレージノードＸＹは、要求されたデータが、自ストレージノードＸＹが保持しているデータであるか否かを検索する（ステップＳ５０）。要求されたデータが自ストレージノードＸＹの保持しているデータである場合、ストレージノードＸＹは、ストレージノードＸ１からのリクエストに応答し、要求されたデータをストレージノードＸ１に送信する（ステップＳ６０）。そして、ストレージノードＸ１が、ストレージノードＸＹから送信されてきたデータを受信することによって、ストレージノードＸ１によるデータのアクセスを完了する。

【００４３】

また、ステップＳ５０において、要求されたデータが自ストレージノードＸＹの保持しているデータでない場合、ストレージノードＸＹは、自ストレージノードＸＹが属しているストレージグループＳＧ＿Ｙ内の他のストレージノード１に、要求されたデータが分散されていると判断する。そして、ストレージノードＸＹは、ノードリストから要求されたデータを格納しているストレージグループＳＧ＿Ｙ内の他のストレージノード１を求め、ストレージノードＸ１からのリクエストを、ストレージグループＳＧ＿Ｙ内の他のストレージノード１に転送する（ステップＳ６１）。図６では、ストレージグループＳＧ＿Ｙに属するストレージノード２Ｙ（図３参照）にリクエストが転送された場合を示している。なお、ストレージノードＸＹがストレージグループ４内の他のストレージノード１にリクエストを転送する方法は、ストレージグループ４内におけるデータ分散方法に依存する。

このデータ分散方法に関しては、後述する。

【００４４】

そして、リクエストが転送されたストレージノード２Ｙは、自ストレージノード２Ｙが保持しているデータから、要求されたデータを検索する（ステップＳ７０）。そして、ストレージノード２Ｙは、ストレージノードＸＹから転送されてきたストレージノードＸ１からのリクエストに応答し、要求されたデータをストレージノードＸＹに送信する（ステップＳ８０）。そして、ストレージノードＸＹは、ストレージノード２Ｙからのリクエストの応答とデータとを、同じネットワークグループＮＧ＿Ｘ内のストレージノードＸ１に転送する（ステップＳ８１）。そして、ストレージノードＸ１が、ストレージノードＸＹから転送されてきたデータを受信することによって、ストレージノードＸ１によるデータのアクセスを完了する。

【００４５】

なお、図６の説明においては、ステップＳ５０において、リクエストを受信したストレージノード１が要求されたデータを保持していない場合に、要求されたデータを格納しているストレージノード１にリクエストを転送する例を説明したが、ストレージノード１が他のストレージノード１にリクエストを転送せず、要求元のストレージノード１に他のストレージノード１を通知する方法とすることもできる。より具体的には、まず、ステップＳ５０において、ストレージノードＸＹは、ノードリストから要求されたデータを格納しているストレージグループＳＧ＿Ｙ内の他のストレージノード１であるストレージノード２Ｙを求める。そして、ステップＳ６０において、ストレージノードＸＹが要求されたデータをストレージノードＸ１に送信する代わりに、要求したデータがストレージグループＳＧ＿Ｙ内のストレージノード２Ｙに格納されていること通知する。そして、ストレージノードＸ１は、通知されたストレージノード２Ｙに直接データの要求（リクエスト）を送信（再送信）し、ストレージノード２Ｙから送信されてきた要求したデータを受信する。このように、ストレージノードＸ１が、ストレージノード２Ｙにリクエストを送信し直すことによって、ストレージノード２Ｙから送信されてきたデータを直接受信することができる。

【００４６】

次に、本実施形態のストレージシステム３におけるストレージグループ４内のデータ分散方法について、２つの方法を説明する。まず、１つ目の方法は、ストレージグループ４内の全てのデータ配置を管理する１つのストレージノード１（以下、「メタサーバ」という）を決め、そのメタサーバがストレージグループ４内の全てのデータ配置を集中して管理する集中メタサーバ方式である。この集中メタサーバ方式は、データ配置を集中して管理するため、データの移動や複製の管理が容易である。

【００４７】

この集中メタサーバ方式では、クライアント２がストレージシステム３内のデータをアクセスする際に、必ずメタサーバに対しての問い合わせが行われる。しかし、ストレージシステム３において、データを移動する頻度はそれほど多くないため、例えば、クライアント２からアクセスされるデータが特定のデータに集中している（局所性がある）可能性がある。この場合には、各ストレージノード１が、自ストレージノード１内でアクセスされたデータの位置情報を、例えば、キャッシュメモリなどに一時記憶（キャッシュ）しておくことによって、メタサーバに対しての問い合わせを行わずに、ストレージシステム３内のデータをアクセスすることもできる。

【００４８】

なお、ストレージシステム３内のデータをアクセスする際に、メタサーバに対しての問い合わせを行う場合には、メタサーバの性能がボトルネックになる可能性がある。また、ストレージシステム３の構成が大規模である場合には、ストレージシステム３内のストレージグループ４がネットワークに分散しているため、ネットワークトラフィックが増えてしまう可能性がある。この場合においても、各ストレージノード１におけるキャッシュの機能と組み合わせることによって、上記に述べた欠点を解決することができる可能性が高

い。

【００４９】

また、本実施形態のストレージシステム３におけるストレージグループ４内のデータ分散方法の２つ目の方法は、ストレージグループ４間のデータ配置と同様に、分散関数によってストレージノード１を決定するハッシュ方式である。このハッシュ方式では、ハッシュ値の範囲と対応するストレージノード１とが組となったハッシュテーブルに基づいて、ハッシュ値からストレージノード１を求める。この方法は、従来のハッシュテーブルを用いた方法と同様である。ただし、本実施形態のストレージシステム３においては、データ配置の自由度を高めるために、ハッシュ値の範囲を細かい単位（最も小さくなるハッシュ値）に分割していることが異なる。

【００５０】

例えば、あるデータを別のストレージノード１に移動するときには、ハッシュテーブル上で移動するデータのハッシュ値が対応するストレージノード１を変えることによって、データを別のストレージノード１に移動することができる。このデータの移動方法は、従来のハッシュテーブルを用いたデータの移動においても同様である。しかし、従来のハッシュテーブルでは、ハッシュ値の範囲を分割していないため、データの移動を行う際に、ハッシュ値の範囲に含まれる全てのデータを移動する必要があり、データの移動におけるコストがかかりすぎる場合がある。これに対して、本実施形態のストレージシステム３では、ハッシュ値の範囲を細かい単位に分割しているため、分割したハッシュ値の範囲に含まれるデータのみを移動することができ、従来のハッシュテーブルを用いたデータの移動のように、全てのデータを移動する必要がない。

【００５１】

なお、本実施形態のストレージシステム３では、ハッシュ値の範囲を細かい単位に分割することによって、移動するデータの量を少なくすることができるが、ハッシュ値の範囲の分割によってハッシュテーブルが大きくなる。しかし、ハッシュテーブルのサイズに上限を設け、ハッシュテーブルのサイズが上限に達したときに、アクセスの少ないハッシュ範囲を隣接するハッシュ範囲と融合することによって、ハッシュテーブルのサイズを小さく（圧縮）することができる。

【００５２】

次に、本実施形態のストレージシステム３において、データを移動させる場合の一例を説明する。データの移動は、同一のストレージグループ４内のストレージノード１間で行う。例えば、図３に示したノード構成の一例では、符号“Ｙ”が付与されたストレージグループ４に属する識別番号“１”が付与されたストレージノード１（１Ｙ）から“ｍ”が付与されたストレージノード１（ｍＹ）までのいずれか１つのストレージノード１にデータを移動する。このように、本実施形態のストレージシステム３では、同一のストレージグループ４内のいずれか１つのストレージノード１にデータを移動することができる。

【００５３】

上記に述べたとおり、本実施形態によるストレージシステム３によれば、同一のストレージグループ４内のストレージノード１間でデータを移動させることによって、データ配置の自由度を向上させることができる。また、データ配置の変更を行う理由としては、ストレージシステム３のデータにアクセスするクライアント２とアクセスされるデータを保持しているストレージノード１とにおけるネットワーク上の距離が遠いことが考えられるが、本実施形態によるストレージシステム３では、ネットワークトポロジーに応じてネットワークグループ５を設定しているため、ストレージグループ４内でクライアント２とネットワークグループ５が同じストレージノード１を必ず見つけることができる。このように、ストレージグループ４内でデータを移動することができる自由度があれば、十分にストレージシステム３の効率を向上することができる。そして、データ配置の自由度が向上することによって、省電力モードによる制御も可能となり、ストレージシステム３の省電力化を実現することもできる。

【００５４】

なお、ストレージシステム内の任意のストレージノードにデータを移動することができれば、データ配置の自由度が最大となるが、その場合には、データ配置が変更されたことをストレージシステム内の全てのストレージノードが把握する必要がある。しかし、これは、ストレージシステムの構成が大規模になるにしたがって、全てのストレージノードがデータ配置の変更を把握するためのコストが上昇し、結果としてストレージシステム全体のスケーラビリティを大きく損なう要因となる。それに対して本実施形態のストレージシステム３における制御方法であれば、データ配置を変更したストレージグループ４内のストレージノード１のみがデータ配置の変更を把握していればよい。そして、データ配置の変更をしていない他のストレージグループ４内のストレージノード１にとっては、データ配置を変更したストレージグループ４であってもストレージグループとしての変更がされていないため、データ配置の変更を把握する必要がない。従って、データ配置の変更を把握するためのコストを低く抑えることができ、結果としてストレージシステム全体のスケーラビリティが向上することとなる。そして、これは、ストレージシステム３のデータにアクセスするクライアント２とアクセスされるデータを保持しているストレージノード１とにおけるネットワーク上の距離を近くするという目的を、十分に達成しているということができる。

【００５５】

上記実施形態におけるストレージシステムの制御方法では、スケーラブルなストレージシステムにおいて、ストレージグループとネットワークグループという直交したグループでストレージノードをグルーピングする。ストレージグループの中ではネットワークトポロジーを考慮したデータ配置を行えるようにして、データ配置の自由度を向上させることができる。これにより、柔軟なデータ配置をすることができるため、例えば、処理ノードとデータとを近くに配置することによって、ストレージシステムの効率を向上することができる。また、上記実施形態のストレージシステムの制御方法では、データ配置の自由度が向上するため、例えば、データを予め定めたストレージノードに集めることによって、ストレージシステムの低消費電力化を実現する省電力の制御を行うことや、不要なトラフィックを防いでアクセス速度を向上させることができる。

【００５６】

以上、一実施形態について、図面を参照して説明してきたが、具体的な構成はこの実施形態に限定されるものではなく、種々の変更も含まれる。

【００５７】

なお、上記実施の形態におけるストレージノード１は、ＣＰＵ（Ｃｅｎｔｒａｌ　Ｐｒｏｃｅｓｓｉｎｇ　Ｕｎｉｔ）、メモリ、ハードディスク、ネットワークインタフェース、などを有してもよい。上記ストレージノード１は、一般に広く用いられるサーバコンピュータ、パソコン、などであってもよい。また、上記の実施形態はソフトウェア、もしくは、ハードウェアによって実装されてもよい。

【産業上の利用可能性】

【００５８】

上記の実施形態は、例えば、大量のデータを保持するストレージシステムに適用が可能である。

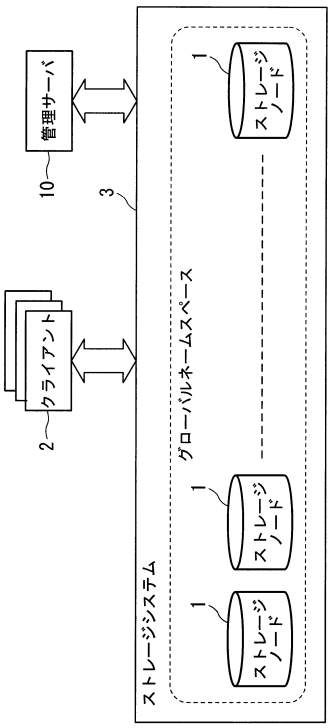
【符号の説明】

【００５９】

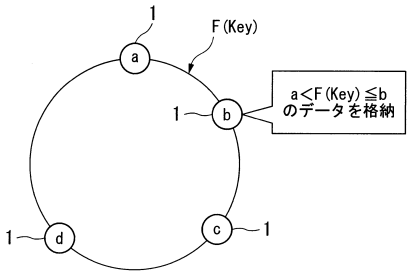
１・・・ストレージノード
２・・・クライアント
３・・・ストレージシステム
４・・・ストレージグループ
５・・・ネットワークグループ
６・・・ラック
７・・・スイッチ
８・・・上位スイッチ

１０・・・管理サーバ

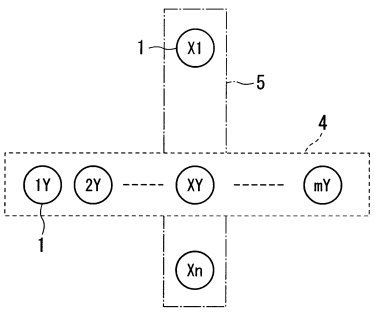
【図１】



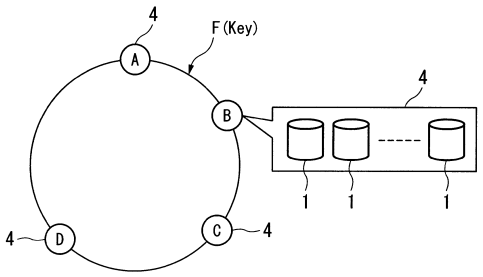
【図２】



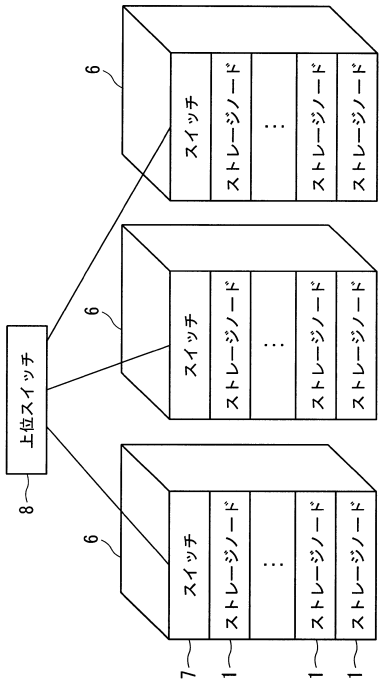
【図３】



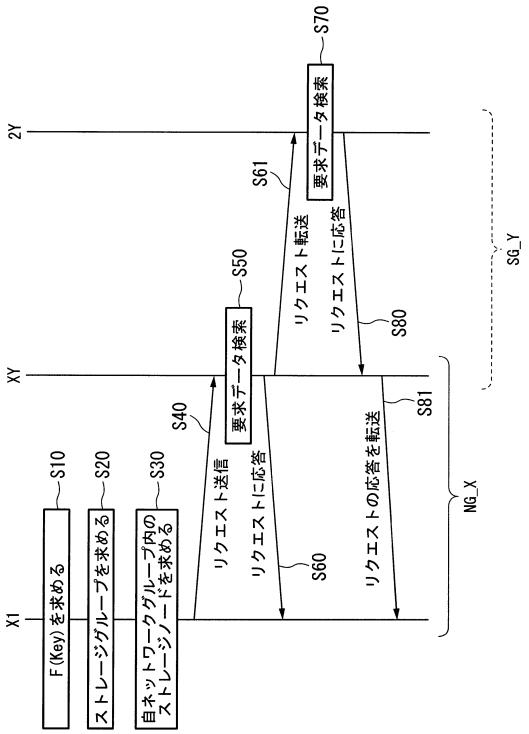
【図４】


4
F(Key)
A
B
C
D
4
4
4
4
1
1
1

【図５】


上位スイッチ
8
6
6
6
スイッチ
ストレージノード
ストレージノード
ストレージノード
7
1
1
1

【図６】


X1
XY
2Y
F(Key)を求める
ストレージグループを求める
自ネットワークグループ内のストレージノードを求める
S10
S20
S30
S40
S50
S60
S61
S70
S80
S81
リクエスト送信
リクエストに応答
要求データ検索
リクエスト転送
リクエストに応答
リクエストの応答を転送
NG_X
SG_Y

フロントページの続き

（出願人による申告）国等の委託研究の成果に係る特許出願（平成２１年度　独立行政法人新エネルギー・産業技術総合開発機構「グリーンネットワーク・システム技術研究開発プロジェクト（グリーンＩＴプロジェクト）／エネルギー利用最適化データセンタ基盤技術の研究開発／サーバの最適構成とクラウド・コンピューティング環境における進化するアーキテクチャーの開発／クラウド・コンピューティング技術」委託研究、産業技術力強化法第１９条の適用を受ける特許出願）

前置審査

審査官　加内　慎也

(56)参考文献　特開２００８－０１１３３０（ＪＰ，Ａ）
特開２００８－１４６１１９（ＪＰ，Ａ）
特開２００７－２８０３０４（ＪＰ，Ａ）
特開２０１０－０７４６０４（ＪＰ，Ａ）
特開２００７－１４８５４５（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１２／００
Ｇ０６Ｆ　　３／０６
Ｇ０６Ｆ　１３／１０