(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2012 年 12 月 6 日(06.12.2012)

WIPO | PCT

(10) 国際公開番号
WO 2012/164735 A1

(51) 国際特許分類:
G06F 12/00 (2006.01)

(21) 国際出願番号: PCT/JP2011/062759

(22) 国際出願日: 2011 年 6 月 3 日(03.06.2011)

(25) 国際出願の言語: 日本語

(26) 国際公開の言語: 日本語

(71) 出願人 (米国を除く全ての指定国について): 富士通株式会社(FUJITSU LIMITED) [JP/JP]; 〒2118588 神奈川県川崎市中原区上小田中４丁目１番１号 Kanagawa (JP).

(72) 発明者; および

(75) 発明者/出願人 (米国についてのみ): 村田 美穂(MURATA, Miho) [JP/JP]; 〒2118588 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 Kanagawa (JP). 槌本 裕一(TSUCHIMOTO, Yuichi) [JP/JP]; 〒2118588 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 Kanagawa (JP).

(74) 代理人: 服部 毅巖(HATTORI, Kiyoshi); 〒1920082 東京都八王子市東町９番８号 八王子東町センタービル 服部特許事務所 Tokyo (JP).

(81) 指定国 (表示のない限り、全ての種類の国内保護が可能): 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, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

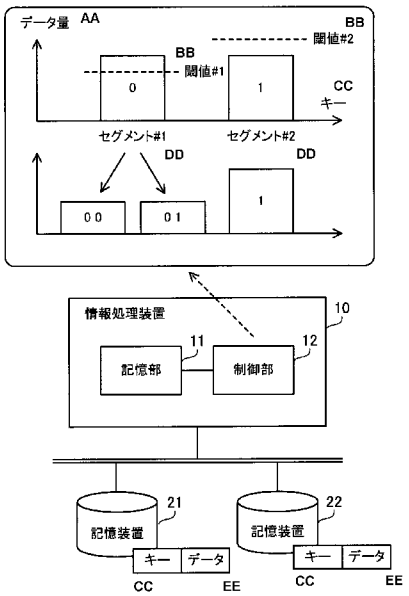
(84) 指定国 (表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), ヨーロッパ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT,

[続葉有]

(54) **Title**: INFORMATION PROCESSING DEVICE, DATA MANAGEMENT METHOD, AND DATA MANAGEMENT PROGRAM

(54) 発明の名称 : 情報処理装置、データ管理方法およびデータ管理プログラム

[図1]



10 INFORMATION PROCESSING DEVICE
11 STORAGE UNIT
12 CONTROL UNIT
21, 22 STORAGE DEVICE
AA DATA AMOUNT
BB THRESHOLD VALUE
CC KEY
DD SEGMENT
EE DATA

(57) **Abstract**: In order to prevent the concentration of a data management load, an information processing device (10) is used in an information processing system wherein a set of keys is divided into multiple segments, and data which is stored in association with a key is managed in units of segments. A storage unit (11) stores segment information indicating the relationship between the key and the segment for at least segments #1 and #2 of the multiple segments. A control unit (12) references the segment information and determines whether to divide segment #1 into two or more segments on the basis of the amount of data stored in association with the key belonging to segment #1 and on the basis of a threshold value #1. The threshold value #1 takes a different value than a threshold value #2 which is used to determine whether to divide segment #2.

(57) 要約: データ管理の負荷が集中することを抑制する。 情報処理装置（１０）は、キーの集合を複数のセグメントに分割し、キーと対応付けて記憶されるデータをセグメント単位で管理する情報処理システムに用いられる。記憶部（１１）は、複数のセグメントのうち少なくともセグメント＃１，＃２について、キーとセグメントとの関係を示すセグメント情報を記憶する。制御部（１２）は、セグメント情報を参照し、セグメント＃１に属するキーと対応付けて記憶されているデータの量と閾値＃１に基づいて、セグメント＃１を２以上のセグメントに分割するか判定する。閾値＃１は、セグメント＃２を分割するか判定するために用いられる閾値＃２とは異なる値をとる。

WO 2012/164735 A1

NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

添付公開書類:

— 国際調査報告（条約第 21 条(3)）

明　細　書

発明の名称 ：

情報処理装置、データ管理方法およびデータ管理プログラム

技術分野

[0001] 本発明は情報処理装置、データ管理方法およびデータ管理プログラムに関する。

背景技術

[0002] 情報処理システムにおけるデータ管理方法の１つとして、キーとデータとを対応付けて１またはそれ以上の記憶装置に保存しておく方法が考えられる。アプリケーションソフトウェアなどのソフトウェアは、例えば、キーを指定して、データの読み出し（リード）や書き込み（ライト）などのデータ操作を行う。データを複数の記憶装置に分散して保存する場合、例えば、キーとハッシュ関数を用いて、データの保存先の記憶装置を決定し、また、所望のデータが保存されている記憶装置を検索する方法が考えられる。

[0003] また、データを効率的に管理できるように、キーの集合を複数のセグメントに分割し、セグメント単位でデータを管理する方法が考えられる。例えば、セグメント単位で、データの保存先の記憶装置を決定することが考えられる。キーの集合を複数のセグメントに分割してキーをセグメントに割り当てる方法として、例えば、ＰＨＴ（Prefix Hash Tree）と呼ばれる２分木を用いる方法が提案されている。

[0004] ＰＨＴの各ノードには、ルートノードからの深さに応じた長さの２進数がラベルとして付与される。ルートノードは、ノード０とノード１を子ノードとしてもつ。ノード０は、ノード００とノード０１を子ノードとしてもつ。キーを２進数で表記したときのプレフィックス（キーの先頭から数ビットの部分）とラベルとを比較することで、各キーが何れか１つの葉ノード（子ノードをもたないノード）に割り当てられる。そして、葉ノードに相当するセグメント単位で、キーおよびデータが管理される。ＰＨＴの深さ（セグメン

トの分割の程度）は、例えば、管理するデータの量に応じて決定される。

先行技術文献

特許文献

[0005] 非特許文献1：Yatin Chawathe, Sriram Ramabhadran, Sylvia Ratnasamy, Anthony LaMarca, Scott Shenker and Joseph Hellerstein, "A Case Study in Building Layered DHT Applications", Proc. of the ACM SIGCOMM 2005 conference on Applications, technologies, architectures and protocols for computer communications, August 22-26, 2005.

発明の概要

発明が解決しようとする課題

[0006] ところで、情報処理システムを運用するに従い、記憶装置にデータが書き込まれて、管理するデータの量が増加することがある。そこで、データをセグメント単位で管理するとき、データ量が大きくなったセグメントを、更に複数のセグメントに動的に分割することが考えられる。しかし、セグメントを分割するか否かを判定する基準の設定方法によっては、複数のセグメントの分割が同時期に集中して発生してしまう場合があるという問題がある。セグメントの分割が同時期に集中すると、データの移動や管理情報の更新などの処理によって情報処理システムの負荷が増大し、当該時期に要求されたデータ操作への応答が遅延するなどの性能低下が生じるおそれがある。

[0007] 一つの側面では、本発明は、データ管理の負荷が集中することを抑制できる情報処理装置、データ管理方法およびデータ管理プログラムを提供することを目的とする。

課題を解決するための手段

[0008] 一実施態様では、キーの集合を複数のセグメントに分割し、キーと対応付けて記憶装置に記憶されるデータをセグメント単位で管理する情報処理システムに用いられる情報処理装置が提供される。情報処理装置は、記憶部と制御部を有する。記憶部は、複数のセグメントのうち少なくとも第１および第

２のセグメントについて、キーとセグメントとの関係を示すセグメント情報を記憶する。制御部は、セグメント情報を参照し、第１のセグメントに属するキーと対応付けて記憶されているデータの量と第１の閾値とに基づいて、第１のセグメントを２以上のセグメントに分割するか判定する。第１の閾値は、第２のセグメントを分割するか判定するために用いられる第２の閾値とは異なる値をとる。

[0009] また、一実施態様では、キーの集合を複数のセグメントに分割し、キーと対応付けて記憶装置に記憶されるデータをセグメント単位で管理する情報処理システムが行うデータ管理方法が提供される。データ管理方法では、複数のセグメントのうちの第１のセグメントに属するキーと対応付けて記憶装置に記憶されているデータの量を算出する。算出したデータの量と第１のセグメントに応じた第１の閾値とに基づいて、第１のセグメントを２以上のセグメントに分割するか判定する。第１の閾値は、複数のセグメントのうちの第２のセグメントを分割するか判定するために用いられる第２の閾値とは異なる値をとる。

[0010] また、一実施態様では、キーの集合を複数のセグメントに分割し、キーと対応付けて記憶装置に記憶されるデータをセグメント単位で管理する情報処理システムに用いられるコンピュータに実行させるデータ管理プログラムが提供される。データ管理プログラムは、コンピュータに、以下の処理を実行させる。複数のセグメントのうちの第１のセグメントに属するキーと対応付けて記憶装置に記憶されているデータの量を算出する。算出したデータの量と第１のセグメントに応じた第１の閾値とに基づいて、第１のセグメントを２以上のセグメントに分割するか判定する。第１の閾値は、複数のセグメントのうちの第２のセグメントを分割するか判定するために用いられる第２の閾値とは異なる値をとる。

発明の効果

[0011] 一実施態様によれば、データ管理の負荷が集中することを抑制できる。

本発明の上記および他の目的、特徴および利点は本発明の例として好まし

い実施の形態を表す添付の図面と関連した以下の説明により明らかになるであろう。

図面の簡単な説明

[0012] [図1]第１の実施の形態の情報処理装置を示す図である。

[図2]第２の実施の形態の情報処理システムを示す図である。

[図3]サーバ装置のハードウェア例を示すブロック図である。

[図4]リクエスト処理の例を示す図である。

[図5]セグメント管理ツリーの例を示す図である。

[図6]サーバ装置へのセグメントの配置方法の例を示す図である。

[図7]サーバ装置のソフトウェア例を示すブロック図である。

[図8]リクエスト処理を示すフローチャートである。

[図9]セグメント分割を示すフローチャートである。

[図10]セグメントの閾値の例を示す図である。

[図11]リクエスト処理およびセグメント分割を示すシーケンス図である。

発明を実施するための形態

[0013] 以下、本実施の形態を図面を参照して説明する。

［第１の実施の形態］

図１は、第１の実施の形態の情報処理装置を示す図である。情報処理装置１０は、キーの集合を複数のセグメントに分割し、キーと対応付けて記憶装置２１，２２に記憶されるデータをセグメント単位で管理する情報処理システムに用いられる。記憶装置２１，２２の一方または両方は、情報処理装置１０が備える記憶装置であってもよい。情報処理装置１０は、記憶部１１および制御部１２を有する。

[0014] 記憶部１１は、セグメント情報を記憶する。セグメント情報は、複数のセグメントのうち少なくともセグメント＃１，＃２について、キーとセグメントとの関係を示す。セグメント情報は、トライ（Trie）木やプレフィックス木と呼ばれる木構造によってセグメントを表現したものであってもよい。記憶部１１は、ＲＡＭ（Random Access Memory）やＨＤＤ（Hard Dirk Drive）

などの記憶装置であってもよい。

[0015] 制御部１２は、記憶部１１のセグメント情報を参照し、セグメント＃１に属するキーと対応付けて記憶されているデータの量と閾値＃１に基づいて、セグメント＃１を２以上のセグメントに分割するか判定する。例えば、データ量と閾値＃１を比較し、データ量が閾値＃１を超えるときセグメント＃１を分割すると判定する。データ量は、セグメント＃１が配置されている記憶装置にアクセスして確認できる。閾値＃１は、セグメント＃２を分割するか判定するために用いる閾値＃２とは異なる値をとる。例えば、閾値＃１は、セグメント＃１の識別情報から所定の算出式に従って算出され、閾値＃２は、セグメント＃２の識別情報から所定の算出式に従って算出される。制御部１２は、ＣＰＵ（Central Processing Unit）およびＲＡＭを用いて実行されるプログラムとして実現してもよい。

[0016] なお、セグメント情報が複数のセグメントを木構造で表現したものである場合、制御部１２は、少なくとも同じ階層レベルのセグメント間では、互いに異なる値になるように閾値を算出してもよい。また、記憶装置２１，２２のうちデータを格納する記憶装置がセグメント単位で選択される場合、制御部１２は、セグメント＃１の分割によって、記憶装置２１，２２へのセグメントの配置を変更してもよい。例えば、セグメント＃１が記憶装置２１に配置されていたとき、セグメント＃１から分割された２以上のセグメントの少なくとも１つを、記憶装置２１から記憶装置２２に配置変更してもよい。

[0017] 第１の情報処理装置１０によれば、複数のセグメントのうちセグメント＃１に属するキーと対応付けて記憶装置２１，２２の何れかに記憶されているデータの量が算出される。算出されたデータの量とセグメント＃１に応じた閾値＃１とに基づいて、セグメント＃１を２以上のセグメントに分割するか判定される。閾値＃１は、複数のセグメントのうちセグメント＃２を分割するか判定するために用いられる閾値＃２とは異なる値をとる。

[0018] これにより、記憶装置２１，２２にデータが書き込まれてデータ量の大きくなったセグメントを、更に２以上のセグメントに動的に分割することがで

きる。例えば、データの実際の増加量に合わせてセグメントを分割できるため、予め各セグメントのデータ量がどのように増加するか（データの偏り）を把握しておかなくてもよい。よって、記憶装置２１，２２へのセグメントの配置変更などのセグメント管理が容易となる。また、セグメント＃１とセグメント＃２とで異なる閾値を用いて分割の要否を判定するため、セグメント＃１，＃２の分割が同時期に重なって発生する確率を低減できる。よって、セグメントの配置変更やセグメント情報の更新などの処理が集中して情報処理システムの負荷が増大することを抑制でき、情報処理システムの性能低下を抑制できる。

[0019] ［第２の実施の形態］

図２は、第２の実施の形態の情報処理システムを示す図である。第２の実施の形態の情報処理システムは、サーバ装置１００，１００ａ，１００ｂ，１００ｃ、管理装置２００およびクライアント装置３００を含む。サーバ装置１００，１００ａ，１００ｂ，１００ｃおよび管理装置２００は、ネットワーク４１に接続されている。クライアント装置３００は、ネットワーク４２に接続されている。第２の実施の形態の情報処理システムは、いわゆるクラウドシステムとして実現してもよい。

[0020] サーバ装置１００，１００ａ，１００ｂ，１００ｃは、データを分散管理するサーバコンピュータである。サーバ装置１００，１００ａ，１００ｂ，１００ｃは、データを識別するためのキーとデータとを対応付けて記憶し、また、キーからデータが記憶されているサーバ装置を検索するためのセグメント情報を保持している。何れのサーバ装置も、クライアント装置３００からデータ処理のリクエストを受け付けることができる。キーを含むリクエストを受け付けたサーバ装置は、処理対象のデータが記憶されているサーバ装置をセグメント情報に基づいて検索し、検索されたサーバ装置にデータ処理を依頼する。

[0021] 管理装置２００は、ユーザ（例えば、情報処理システムの管理者）が使用するコンピュータである。管理装置２００は、ユーザ操作に基づいて、サー

バ装置１００，１００ａ，１００ｂ，１００ｃを管理する。例えば、管理装置２００は、データを記憶するサーバ装置が情報処理システムに追加されるとき、追加されるサーバ装置の情報をサーバ装置１００，１００ａ，１００ｂ，１００ｃに送信する。また、データの分散配置を調整するための設定情報を、サーバ装置１００，１００ａ，１００ｂ，１００ｃに送信する。

[0022] クライアント装置３００は、ユーザ（例えば、クラウドサービスの利用者）が使用するコンピュータである。クライアント装置３００では、例えば、サーバ装置１００，１００ａ，１００ｂ，１００ｃに記憶されるデータを扱うアプリケーションソフトウェアが実行される。クライアント装置３００は、ネットワーク４１，４２を介して、サーバ装置１００，１００ａ，１００ｂ，１００ｃのうちの任意のサーバ装置に、データ処理のリクエストを送信する。リクエストでは、１つのキーまたはキーの範囲が指定される。データ処理には、データの読み出し（リード）や書き込み（ライト）が含まれる。

[0023] 何れのサーバ装置がクライアント装置３００からリクエストを受信しても、サーバ装置１００，１００ａ，１００ｂ，１００ｃの間でメッセージを伝送することで、要求されたデータ処理が実行される。このように、第２の実施の形態の情報処理システムは、できる限りデータ処理のボトルネックとなる装置が生じないように設計することが可能であり、可用性や応答性能などを向上させることができる。

[0024] 図３は、サーバ装置のハードウェア例を示すブロック図である。サーバ装置１００は、ＣＰＵ１０１、ＲＡＭ１０２、ＨＤＤ１０３、画像信号処理部１０４、入力信号処理部１０５、ディスクドライブ１０６および通信部１０７を有する。上記ユニットは、サーバ装置１００内でバスに接続されている。なお、サーバ装置１００ａ，１００ｂ，１００ｃ、管理装置２００およびクライアント装置３００も、サーバ装置１００と同様のハードウェアによって実現できる。

[0025] ＣＰＵ１０１は、サーバ装置１００における情報処理を制御する演算装置である。ＣＰＵ１０１は、ＨＤＤ１０３に記憶されたプログラムやデータの

少なくとも一部を読み出してＲＡＭ１０２に展開し、プログラムを実行する。なお、サーバ装置１００は、複数の演算装置を備えて、情報処理を分散して実行してもよい。

[0026] ＲＡＭ１０２は、ＣＰＵ１０１が扱うプログラムやデータを一時的に記憶しておく揮発性メモリである。なお、サーバ装置１００は、ＲＡＭ以外の種類のメモリを備えていてもよく、複数個のメモリを備えてもよい。

[0027] ＨＤＤ１０３は、ＯＳ（Operating System）プログラムやアプリケーションプログラムなどのプログラム、および、情報処理に用いられるデータを記憶する不揮発性の記憶装置である。ＨＤＤ１０３は、ＣＰＵ１０１の命令に従って、内蔵の磁気ディスクに対する読み書きを行う。なお、サーバ装置１００は、ＨＤＤ以外の不揮発性の記憶装置（例えば、ＳＳＤ（Solid State Drive））を備えていてもよく、複数の記憶装置を備えてもよい。

[0028] 画像信号処理部１０４は、ＣＰＵ１０１の命令に従って、サーバ装置１００に接続されたディスプレイ３１に画像を出力する。ディスプレイ３１として、例えば、ＣＲＴ（Cathode Ray Tube）ディスプレイや液晶ディスプレイなどを用いることができる。

[0029] 入力信号処理部１０５は、サーバ装置１００に接続された入力デバイス３２から入力信号を取得し、ＣＰＵ１０１に出力する。入力デバイス３２として、例えば、マウスやタッチパネルなどのポインティングデバイスや、キーボードなどを用いることができる。

[0030] ディスクドライブ１０６は、記録媒体３３に記録されたプログラムやデータを読み取る駆動装置である。記録媒体３３として、例えば、フレキシブルディスク（ＦＤ：Flexible Disk）やＨＤＤなどの磁気ディスク、ＣＤ（Compact Disc）やＤＶＤ（Digital Versatile Disc）などの光ディスク、光磁気ディスク（ＭＯ：Magneto-Optical disk）を使用できる。ディスクドライブ１０６は、例えば、ＣＰＵ１０１の命令に従って、記録媒体３３から読み取ったプログラムやデータをＲＡＭ１０２またはＨＤＤ１０３に格納する。

[0031] 通信部１０７は、ネットワーク４１に接続して通信を行う通信インタフェ

ースである。ネットワーク４１への接続方法は、有線でも無線でもよい。すなわち、通信部１０７は、有線通信インタフェースでも無線通信インタフェースでもよい。

[0032] 図４は、リクエスト処理の例を示す図である。図４では、リクエストとして、（Ａ）単一のキーを指定したリード要求、（Ｂ）キーの範囲を指定したリード要求、（Ｃ）ライト要求の例を挙げている。サーバ装置１００ａは、キーとデータの組（ｋｅｙ，ｖａｌｕｅ）として、（２，ｄａｔａ２），（３，ｄａｔａ３）を記憶し、サーバ装置１００ｃは、（４，ｄａｔａ４），（５，ｄａｔａ５）を記憶しているとする。

[0033] （Ａ）サーバ装置１００は、キー２を指定したリード要求ｇｅｔ（２）をクライアント装置３００から受信すると、キー２に対応するデータがサーバ装置１００ａに記憶されていると判定する。そして、キー２に対するリード要求を示すメッセージをサーバ装置１００ａに送信する。サーバ装置１００ａは、ｄａｔａ２をサーバ装置１００に送信する。サーバ装置１００は、ｄａｔａ２をクライアント装置３００に送信する。

[0034] （Ｂ）サーバ装置１００は、キーの範囲２～５を指定したリード要求ｒａｎｇｅ＿ｇｅｔ（２，５）をクライアント装置３００から受信すると、キー２，３に対応するデータがサーバ装置１００ａに記憶され、キー４，５に対応するデータがサーバ装置１００ｃに記憶されていると判定する。そして、キー２，３に対するリード要求を示すメッセージをサーバ装置１００ａに送信し、キー４，５に対するリード要求を示すメッセージをサーバ装置１００ｃに送信する。サーバ装置１００ａは、ｄａｔａ２，３をサーバ装置１００に送信する。サーバ装置１００ｃは、ｄａｔａ４，５をサーバ装置１００に送信する。サーバ装置１００は、ｄａｔａ２～５をクライアント装置３００に送信する。

[0035] （Ｃ）サーバ装置１００は、キー１を指定したライト要求ｓｅｔ（１，ｄａｔａ１）をクライアント装置３００から受信すると、キー１に対応するデータはサーバ装置１００ａに記憶されるものであると判定する。そして、キ

ー１に対するライト要求を示すメッセージをサーバ装置１００ａに送信する。サーバ装置１００ａは、ｄａｔａ１を自身の記憶装置に書き込み、サーバ装置１００に書き込み完了を報告する。サーバ装置１００は、書き込み完了をクライアント装置３００に報告する。

[0036] このように、サーバ装置１００，１００ａ，１００ｂ，１００ｃは、キーとデータとを対応付けて保持し、キー単位で記憶装置からのデータの読み出しや記憶装置へのデータの書き込みを行う。処理対象のキーは、例えば、クライアント装置３００で実行されるアプリケーションソフトウェアによって指定される。よって、サーバ装置１００，１００ａ，１００ｂ，１００ｃのデータ処理の複雑性が抑えられ、負荷が軽減される。

[0037] ライト要求で指定されたキーに対応するデータがまだ存在しない場合、当該ライト要求はデータの追加を意味する。一方、ライト要求で指定されたキーに対応するデータが既に存在する場合、当該ライト要求はデータの上書きを意味する。キーに対応するデータが存在するか否かの確認や上書き可否の判断は、例えば、クライアント装置３００で実行されるアプリケーションソフトウェアの責任で行われる。第２の実施の形態の情報処理システムは、例えば、時刻をキーとするデータ（例えば、ログデータ）の管理に利用できる。

[0038] 図５は、セグメント管理ツリーの例を示す図である。サーバ装置１００，１００ａ，１００ｂ，１００ｃは、セグメント管理ツリーを記述したセグメント情報を有する。セグメント管理ツリーとして、トライ木やプレフィックス木と呼ばれる順序付きツリーを用いることができる。前述の非特許文献１（Yatin Chawathe, Sriram Ramabhadran, Sylvia Ratnasamy, Anthony LaMarca, Scott Shenker and Joseph Hellerstein, "A Case Study in Building Layered DHT Applications"）に記載されたＰＨＴを用いてもよい。

[0039] セグメント管理ツリーでは、キー空間が、セグメントと呼ばれる区間に階層的に分割される。セグメント管理ツリーの各ノードがセグメントに相当する。各ノードには、当該ノードの深さに応じた長さのラベルが付与される。

例えば、ルートノードの左の子ノードにラベル０を付与し、右の子ノードにラベル１を付与する。また、あるノードにラベルＬが付与されているとき、左の子ノードにラベルＬ＋０、右の子ノードにラベルをＬ＋１を付与する。キー空間に含まれる各キーは、当該キーのプレフィックスに一致するラベルが付与された何れか１つの葉ノード（子ノードをもたないノード）に対応付けられる。

[0040] 一例として、キーが５ビットの２進数で表現できる場合を考える。クライアント装置３００から指定されるキーが２進数で表現されていない場合は、２進数に変換すればよい。また、図５に示したように、ラベル０００，００１，０１，１０，１１が付与された５つの葉ノードを含むセグメント管理ツリーを考える。この場合、キー００００００～０００１１は、ラベル０００のノードに対応付けられる。キー００１００～００１１１は、ラベル００１のノードに対応付けられる。キー０１０００～０１１１１は、ラベル０１のノードに対応付けられる。キー１００００～１０１１１は、ラベル１０の葉ノードに対応付けられる。キー１１０００～１１１１１は、ラベル１１の葉ノードに対応付けられる。

[0041] セグメント管理ツリーの葉ノードに相当するセグメントが、サーバ装置１００，１００ａ，１００ｂ，１００ｃに配置される。例えば、ラベル１１のセグメントが、サーバ装置１００に配置される。ラベル０００，１０のセグメントが、サーバ装置１００ａに配置される。ラベル００１，０１のセグメントが、サーバ装置１００ｃに配置される。セグメントを配置するサーバ装置は、例えば、当該セグメントのラベルにハッシュ関数を適用して算出されるハッシュ値に基づいて決定することができる。

[0042] このように、セグメント単位でデータを配置するサーバ装置を決定することで、キーのローカリティを維持することができ、値の近いキーに対応付けられるデータ同士ができる限り同じサーバ装置に配置されるようになる。このため、範囲指定のリード要求など、範囲指定のデータ処理を効率的に実行できるようになる。セグメント管理ツリーの深さ（セグメントの分割の程度

）は、サーバ装置１００，１００ａ，１００ｂ，１００ｃに記憶されたデータの量に応じて動的に調整される。後述するように、データ量の大きくなったセグメントは、更に複数のセグメントに自動的に分割される。

[0043] なお、以上の説明では、セグメント管理ツリーの例として２分木を用いたが、任意の分岐数のツリーを用いてもよい。葉ノード以外の各ノードがそれぞれｂ個（ｂは２以上の整数）の子ノードを有する場合、各ノードにはｂ進数のラベルを付与すればよい。例えば、ｂ＝３の場合、ルートノードはラベル０，１，２の子ノードを有する。また、ラベル０のノードは、ラベル００，０１，０２の子ノードを有する。キーをセグメントに対応付けるときは、キーをｂ進数に変換して、キーのプレフィックスとラベルを比較すればよい。

[0044] 図６は、サーバ装置へのセグメントの配置方法の例を示す図である。前述の通り、セグメント管理ツリーの葉ノードに対応するセグメントは、当該セグメントのラベルに応じてサーバ装置１００，１００ａ，１００ｂ，１００ｃの何れかに配置される。

[0045] ここでは、ハッシュ関数の値域（ハッシュ値が取り得る値）を０～2^n-1（ｎは自然数）とし、値域をループ状にしたハッシュ値の空間を考える。サーバ装置１００，１００ａ，１００ｂ，１００ｃそれぞれの識別情報（例えば、アドレス）にハッシュ関数を適用して、各サーバ装置に対応するハッシュ値を算出する。また、セグメントのラベルにハッシュ関数を適用して、当該セグメントに対応するハッシュ値を算出する。そして、各サーバ装置のハッシュ値とセグメントのハッシュ値のループ上での位置関係に基づいて、当該セグメントを配置するサーバ装置を決定する。例えば、ループ上で、セグメントのハッシュ値より先にある最小のハッシュ値をもつサーバ装置を選択する。

[0046] 例えば、サーバ装置１００のハッシュ値ｈ（ｓ１）とサーバ装置１００ａのハッシュ値ｈ（ｓ２）とラベル０のハッシュ値ｈ（０）とが、ｈ（ｓ２）＞ｈ（０）＞ｈ（ｓ１）の関係にある場合、ラベル０のセグメントをサーバ

装置１００ａに配置する。また、サーバ装置１００ｂのハッシュ値ｈ（ｓ３）とサーバ装置１００ｃのハッシュ値ｈ（ｓ４）とラベル００１のハッシュ値ｈ（００１）とが、ｈ（ｓ４）＞ｈ（００１）＞ｈ（ｓ３）の関係にある場合、ラベル００１のセグメントをサーバ装置１００ｃに配置する。

[0047] なお、ラベルのＬＳＢ（Least Significant Bit）が０である場合、ＬＳＢからＭＳＢ（Most Significant Bit）に向かって連続する０を削除して、ハッシュ関数を適用する。ただし、ＭＳＢの０は削除しない。例えば、ラベル００，０００はラベル０に変換され、ラベル１０はラベル１に変換される。よって、ラベル０，００，０００のハッシュ値は同一となり、ラベル１，１０のハッシュ値は同一となる。これにより、あるセグメントを２以上のセグメントに分割したとき、分割後の２以上のセグメントの少なくとも１つは、元のセグメントと同じサーバ装置に配置されることになる。

[0048] ただし、サーバ装置１００，１００ａ，１００ｂ，１００ｃへのセグメントの配置方法は、上記方法に限定されない。例えば、セグメントのラベルから算出されたハッシュ値をサーバ装置の数で割って余り（剰余）を求め、剰余に応じてセグメントを配置するサーバ装置を選択する方法が考えられる。例えば、ハッシュ値を４で割り、剰余＝０のときはサーバ装置１００、剰余＝１のときはサーバ装置１００ａ、剰余＝２のときはサーバ装置１００ｂ、剰余＝３のときはサーバ装置１００ｃを選択する。

[0049] 図７は、サーバ装置のソフトウェア例を示すブロック図である。サーバ装置１００は、通信処理部１１０、イベント処理部１２０およびデータ記憶部１３０を有する。通信処理部１１０およびイベント処理部１２０は、例えば、ＣＰＵ１０１およびＲＡＭ１０２を用いて実行されるプログラムとして実現できる。データ記憶部１３０は、例えば、ＲＡＭ１０２やＨＤＤ１０３上の記憶領域として実現できる。サーバ装置１００ａ，１００ｂ，１００ｃも、サーバ装置１００と同様のブロック構成によって実現できる。

[0050] 通信処理部１１０は、リクエスト受信部１１１、セグメント情報記憶部１１２、セグメント管理部１１３およびメッセージ処理部１１４を有する。

リクエスト受信部１１１は、クライアント装置３００からリクエストを受信し、所望のセグメントが配置されたサーバ装置の判定をセグメント管理部１１３に依頼する。また、リクエスト受信部１１１は、読み出されたデータまたは書き込み結果をメッセージ処理部１１４から取得し、レスポンスとしてクライアント装置３００に送信する。

[0051] セグメント情報記憶部１１２は、図５のようなセグメント管理ツリーを記載したセグメント情報を記憶する。セグメント情報は、セグメントの分割に伴って更新される。セグメント情報記憶部１１２は、ハッシュ関数を示す情報を記憶してもよい。なお、セグメント情報記憶部１１２は、ＲＡＭ１０２またはＨＤＤ１０３上の記憶領域として実現できる。

[0052] セグメント管理部１１３は、セグメント情報記憶部１１２に記憶されたセグメント情報に基づいて、リクエストで指定されたキーまたはキーの範囲に対応する１またはそれ以上のセグメントを検索する。そして、セグメントのラベルとハッシュ関数から、検索されたセグメントが配置されている１またはそれ以上のサーバ装置を判定する。セグメント管理部１１３は、検索されたセグメントが自サーバ装置にある場合、イベント処理部１２０にデータ処理を依頼する。検索されたセグメントが他のサーバ装置にある場合、メッセージ処理部１１４に当該他のサーバ装置へのメッセージの送信を依頼する。

[0053] また、セグメント管理部１１３は、イベント処理部１２０からセグメントを分割した旨の報告があると、セグメント情報記憶部１１２に記憶されたセグメント情報を更新する。そして、セグメントの更新を示すメッセージを他のサーバ装置に送信するよう、メッセージ処理部１１４に依頼する。また、セグメント管理部１１３は、メッセージ処理部１１４からの依頼に応じて、セグメント情報記憶部１１２に記憶されたセグメント情報を更新する。なお、セグメント管理部１１３は、セグメントが配置され得るサーバ装置の識別情報を、設定情報として管理装置２００から受信して保持している。

[0054] メッセージ処理部１１４は、他のサーバ装置（サーバ装置１００ａ，１００ｂ，１００ｃ）との間でメッセージを送受信する。メッセージ処理部１１

４は、他のサーバ装置からリード要求やライト要求のメッセージを受信すると、イベント処理部１２０にデータ処理を依頼する。また、セグメント管理部１１３からの依頼に応じて、他のサーバ装置にリード要求やライト要求のメッセージを送信する。また、メッセージ処理部１１４は、セグメント管理部１１３からの依頼に応じて、他のサーバ装置にセグメント更新のメッセージを送信する。また、他のサーバ装置からセグメント更新のメッセージを受信すると、セグメント管理部１１３にセグメント情報の更新を依頼する。

[0055] イベント処理部１２０は、データ処理部１２１、算出式記憶部１２２、閾値算出部１２３および分割判定部１２４を有する。

データ処理部１２１は、セグメント管理部１１３またはメッセージ処理部１１４からの依頼に応じて、データ処理を行う。キーまたはキーの範囲を指定したリード要求である場合、データ処理部１２１は、キーに対応するデータまたはキーの範囲に対応するデータ群をデータ記憶部１３０から読み出す。キーを指定したライト要求である場合、データ処理部１２１は、キーに対応付けてデータをデータ記憶部１３０に書き込む。

[0056] また、データ処理部１２１は、データをデータ記憶部１３０に書き込んだ場合、セグメントの分割の要否を分割判定部１２４に問い合わせる。分割判定部１２４で分割を要すると判定された場合、データ処理部１２１は、分割処理を行う。例えば、データ記憶部１３０に記憶されたデータの一部を他のサーバ装置に移動させる。また、セグメントを分割した旨をセグメント管理部１１３に報告する。

[0057] 算出式記憶部１２２は、セグメントを分割するか判定するための閾値を算出する算出式を記憶する。閾値は、後述するように、分割の要否を判定するセグメントのラベルに基づいて算出される。閾値の算出式は、管理装置２００のユーザが記述し、管理装置２００が算出式記憶部１２２に設定する。なお、算出式記憶部１２２は、ＲＡＭ１０２またはＨＤＤ１０３上の記憶領域として実現できる。

[0058] 閾値算出部１２３は、分割判定部１２４から指定されたセグメントのラベ

ルと、算出式記憶部１２２に記憶された閾値の算出式から、指定されたセグメントの閾値を算出する。そして、算出した閾値を分割判定部１２４に回答する。

[0059] 分割判定部１２４は、データ処理部１２１からの問い合わせに応じて、データの書き込みが行われたセグメントの分割の要否を判定する。分割判定部１２４は、書き込みが行われたセグメントのラベルを指定して、閾値算出部１２３に閾値の算出を依頼する。セグメントのラベルは、例えば、セグメント情報を参照して、ライト要求で指定されたキーから求める。また、分割判定部１２４は、データ記憶部１３０を参照して、書き込みが行われたセグメントのデータ量（例えば、当該セグメントに属するキーに対応付けられたデータの総量）を求める。そして、分割判定部１２４は、データ量が閾値を超える場合、分割を要すると判定し、データ量が閾値以下の場合、分割を要しないと判定する。

[0060] データ記憶部１３０は、キーとデータを対応付けて記憶する。データ記憶部１３０は、記憶領域をセグメント毎に分割して、キーとデータを記憶してもよい。

図８は、リクエスト処理を示すフローチャートである。ここでは、サーバ装置１００が処理を行う場合を考える。サーバ装置１００ａ，１００ｂ，１００ｃも、サーバ装置１００と同様の処理を行う。以下、図８に示す処理をステップ番号に沿って説明する。

[0061] （ステップＳ１１）リクエスト受信部１１１は、クライアント装置３００からリクエストを受信する。リクエストの種類には、１つのキーを指定したリード要求、キーの範囲を指定したリード要求、および、１つのキーを指定したライト要求が含まれる。

[0062] （ステップＳ１２）セグメント管理部１１３は、ステップＳ１１受信されたリクエストがキーの範囲を指定したものか判断する。範囲指定のリクエストである場合、処理をステップＳ１４に進める。それ以外の場合、処理をステップＳ１３に進める。

[0063] （ステップＳ１３）セグメント管理部１１３は、セグメント情報記憶部１１２に記憶されたセグメント情報を参照して、リクエストで指定されたキーの属するセグメントを検索する。セグメントの検索では、例えば、ｂ進数（ｂは２以上の整数）で表現したキーのプレフィックスとラベルを、セグメント管理ツリーのルートノードから葉ノードに向かって比較していき、プレフィックスと一致するラベルをもつ葉ノードを特定する。特定した葉ノードに相当するセグメントが、検索されるセグメントである。

[0064] （ステップＳ１４）セグメント管理部１１３は、セグメント情報を参照して、リクエストで指定された範囲内のキーを少なくとも１つ含むセグメントを全て検索する。複数のセグメントの検索では、例えば、範囲内のキーの最大値と最小値とに共通する最長のプレフィックスを求め、当該プレフィックスをラベルとしてもつノードを特定する。そして、特定したノード以下の複数の葉ノードを特定する。特定した複数の葉ノードに相当する複数のセグメントが、検索されるセグメントである。

[0065] 例えば、キーの範囲が００００１０～００１０１の場合、キーの最小値００００１０と最大値００１０１とに共通する最長のプレフィックスは、００である。よって、ラベル００のノードより下位にある、ラベル０００，００１の葉ノードが特定される。そして、ラベル０００に相当するセグメントとラベル００１に相当するセグメントとが検索される。

[0066] （ステップＳ１５）セグメント管理部１１３は、ステップＳ１３またはステップＳ１４で検索されたセグメントのラベルにハッシュ関数を適用し、ハッシュ値を算出する。そして、ハッシュ値からセグメントが配置されているサーバ装置を特定する。サーバ装置は、例えば、図６に示した方法で特定される。ステップＳ１４で複数のセグメントが検索された場合、複数のサーバ装置が特定されることがある。以下のステップＳ１６～Ｓ１８の処理が、特定されたサーバ装置毎またはセグメント毎に実行される。

[0067] （ステップＳ１６）セグメント管理部１１３は、データ処理の対象のセグメントが配置されたサーバ装置として、他のサーバ装置がステップＳ１５で

特定されたか判断する。他のサーバ装置が特定された場合、処理をステップＳ１７に進める。自装置（サーバ装置１００）が特定された場合、処理をステップＳ１８に進める。

[0068] （ステップＳ１７）メッセージ処理部１１４は、ステップＳ１５で特定された他のサーバ装置に対して、リード要求またはライト要求を示すメッセージを送信する。メッセージには、クライアント装置３００からのリクエストで指定されたキーまたはキーの範囲のうち、宛先のサーバ装置に記憶されているキーまたはキーの範囲が含まれる。その後、メッセージ処理部１１４は、応答のメッセージを他のサーバ装置から受信する。応答のメッセージには、キーに対応するデータまたは書き込み完了の報告が含まれる。

[0069] （ステップＳ１８）データ処理部１２１は、クライアント装置３００からのリクエストに応じたデータ処理を行う。リード要求の場合、リクエストで指定されたキーまたはキーの範囲のうち、自装置に記憶されているキーまたはキーの範囲に対応するデータをデータ記憶部１３０から読み出す。ライト要求の場合、リクエストに含まれるキーとデータとを対応付けて、データ記憶部１３０に書き込む。

[0070] （ステップＳ１９）リクエスト受信部１１１は、ステップＳ１１で受信したリクエストに対するレスポンスをクライアント装置３００に送信する。リクエストがリード要求である場合、レスポンスとして、ステップＳ１７，Ｓ１８で自装置または他のサーバ装置から読み出されたデータを送信する。リクエストがライト要求である場合、レスポンスとして、ステップＳ１７，Ｓ１８で確認された書き込み完了の報告を送信する。

[0071] 図９は、セグメント分割を示すフローチャートである。ここでは、サーバ装置１００が処理を行う場合を考える。サーバ装置１００ａ，１００ｂ，１００ｃも、サーバ装置１００と同様の処理を行う。以下、図９に示す処理をステップ番号に沿って説明する。

[0072] （ステップＳ２１）分割判定部１２４は、データ処理部１２１がデータ記憶部１３０に書き込んだデータのキーを特定する。そして、セグメント情報

記憶部１１２に記憶されたセグメント情報に基づいて、キーに対応するセグメントを検索する。

[0073] （ステップＳ２２）閾値算出部１２３は、算出式記憶部１２２に記憶された算出式に、ステップＳ２１で検索されたセグメントのラベルを代入し、当該セグメントの閾値を算出する。閾値の算出式としては、例えば、以下に示す数式（１）を用いる。数式（１）において、Ｌはラベル、Ｒは所定のルート閾値、ｂはセグメント管理ツリーの分岐数を示す。また、ｌｅｎｇｔｈ（Ｌ）はラベルのビット数を示し、ｖａｌｕｅ（Ｌ）はラベルＬが表す１０進数表現の値を示す。

[0074] [数1]

$$th(L) = R \times 2^{f} \quad \text{where } f = \frac{\text{value}(L)}{b^{\text{length}(L)}} \qquad (1)$$

[0075] （ステップＳ２３）分割判定部１２４は、データ記憶部１３０を参照して、ステップＳ２１で検索したセグメントのデータ量（例えば、当該セグメントに属するキーに対応付けられているデータの総量）を算出する。そして、データ量がステップＳ２２で算出された閾値を超えているか判断する。データ量が閾値を超える場合、処理をステップＳ２４に進める。データ量が閾値以下の場合、処理を終了する。

[0076] （ステップＳ２４）セグメント管理部１１３は、ステップＳ２１で検索されたセグメントを分割して得られるｂ個の子セグメントを定義する。例えば、セグメントのラベルがＬでありｂ＝２の場合、ラベルＬ＋０，Ｌ＋１の子セグメントを定義する。

[0077] （ステップＳ２５）データ処理部１２１は、データ記憶部１３０から、ステップＳ２１で検索されたセグメントに属するキーを１つ選択する。

（ステップＳ２６）データ処理部１２１は、ステップＳ２４で定義された子セグメントの中から、ステップＳ２５で選択したキーのプレフィックスと一致するラベルをもつ子セグメントを特定する。そして、データ処理部１２１は、ステップＳ２５で選択したキーを特定した子セグメントに割り当てる

。

[0078] （ステップＳ２７）データ処理部１２１は、ステップＳ２５で全てのキーを選択したか判断する。全てのキーを選択した場合、処理をステップＳ２８に進める。未選択のキーが存在する場合、処理をステップＳ２５に進める。

[0079] （ステップＳ２８）セグメント管理部１１３は、ステップＳ２４で定義した子セグメントのラベルにハッシュ関数を適用し、子セグメントそれぞれの配置先のサーバ装置を特定する。データ処理部１２１は、他のサーバ装置に配置される子セグメントのデータ（当該子セグメントに属するキーと対応付けられているデータ）をデータ記憶部１３０から読み出し、他のサーバ装置に送信する。データ送信が完了すると、他のサーバ装置に配置される子セグメントのデータをデータ記憶部１３０から削除する。

[0080] （ステップＳ２９）セグメント管理部１１３は、ステップＳ２４のセグメント分割が反映されるように、セグメント情報記憶部１１２に記憶されているセグメント情報を更新する。メッセージ処理部１１４は、セグメントの更新を他の全てのサーバ装置に通知する。

[0081] 図１０は、セグメントの閾値の例を示す図である。図１０の閾値の例は、図５のセグメント管理ツリーに数式（１）を適用して算出したものである。

各階層レベルの左端ノード（ルートノードおよびラベル０，００，０００のノード）に対応するセグメントには、所定のルート閾値Ｒが設定される。Ｒとして、１００［ＭＢ（メガバイト）］などのデータ量が予め決められている。左端ノード以外のノードに対応するセグメントには、Ｒを2^{f}倍した閾値が設定される。ｆは分数で表現でき、分母の計算にはラベルのビット数が用いられ、分子にはラベルの表す値が用いられる。よって、同一の階層レベルのセグメント間では、ｆが異なる値となり、閾値が異なる値となる。

[0082] ラベル１のセグメントには、ｆ＝１／２であるため、閾値１．４１Ｒが設定される。ラベル０１のセグメントには、ｆ＝１／４であるため、閾値１．１９Ｒが設定される。ラベル１０のセグメントには、ｆ＝２／４であるため、閾値１．４１Ｒが設定される。ラベル１１のセグメントには、ｆ＝３／４

であるため、閾値１．６８Ｒが設定される。

[0083] ここで、サーバ装置１００，１００ａ，１００ｂ，１００ｃがデータを記憶していない状態（データ集合が空の状態）から、一定の速度でキー空間全体に対して均等にデータが追加される場合を考える。例えば、ルート閾値をＲ［ＭＢ］とし、キー空間全体に対してＳ［ＭＢ／秒］（例えば、１［ＭＢ／秒］）の速度でデータが追加されるとする。

[0084] この場合、データ追加を開始してからの経過時間として、ルートノードのセグメントでは、Ｒ／Ｓ［秒］でデータ量が閾値に達する。ラベル０のセグメントでは、２Ｒ／Ｓ［秒］でデータ量が閾値に達する。ラベル１のセグメントでは、２．８２Ｒ／Ｓ［秒］でデータ量が閾値に達する。ラベル００のセグメントでは、４Ｒ／Ｓ［秒］でデータ量が閾値に達する。ラベル０１のセグメントでは、４．７６Ｒ／Ｓ［秒］でデータ量が閾値に達する。ラベル１０のセグメントでは、５．６４Ｒ／Ｓ［秒］でデータ量が閾値に達する。ラベル１１のセグメントでは、６．７２Ｒ／Ｓ［秒］でデータ量が閾値に達する。ラベル０００のセグメントでは、８Ｒ／Ｓ［秒］でデータ量が閾値に達する。ラベル００１のセグメントでは、８．７２Ｒ／Ｓ［秒］でデータ量が閾値に達する。

[0085] このように、数式（１）に従って各セグメントの閾値を算出すると、データがキー空間に対して均等に追加される場合には、上位の階層レベルのセグメントの分割が完了した後に、下位の階層レベルのセグメントの分割が行われる。例えば、ラベル１のセグメントが分割された後に、ラベル００のセグメントが分割される。ラベル１１のセグメントが分割された後に、ラベル０００のセグメントが分割される。また、全てのセグメントの分割が異なるタイミングで行われる。セグメントを分割するか否かは、当該セグメントのラベルから判定できるため、セグメント間で独立に判定できる。

[0086] なお、図１０の例では、各階層レベルの左端ノードにルート閾値Ｒを付与し、左端ノードから右端ノードに向かって閾値を増加させたが、各階層レベルの右端ノードにルート閾値Ｒを付与し、右端ノードから左端ノードに向か

って閾値を増加させてもよい。また、上記の説明では、データ量と閾値とを比較してセグメントを分割するか判定しているが、データ量と閾値とを比較してセグメントを統合するか判定してもよい。例えば、ラベル０００，００１のセグメントのデータ量の合計が閾値以下のとき、当該２つのセグメントを統合すると判定してもよい。そのとき、分割の判定に用いる閾値と統合の判定に用いる閾値とは、異なる算出式から算出してもよい。

[0087] 図１１は、リクエスト処理およびセグメント分割を示すシーケンス図である。クライアント装置３００がサーバ装置１００に対して、データの書き込みのリクエストを送信し、データの書き込みによってセグメント分割が発生する場合を考える。

[0088] （ステップＳ３１）クライアント装置３００は、データの書き込みを示すリクエストをサーバ装置１００に送信する。リクエストには、キーとデータが含まれる。

（ステップＳ３２）サーバ装置１００は、自装置が有するセグメント情報からキーの属するセグメントを検索し、ラベルとハッシュ関数を用いて当該セグメントが配置されているサーバ装置（ここでは、サーバ装置１００ａ）を特定する。そして、ライト要求のメッセージをサーバ装置１００ａに送信する。メッセージには、キーとデータが含まれる。

[0089] （ステップＳ３３）サーバ装置１００ａは、サーバ装置１００から受信したメッセージに含まれるキーとデータとを対応付けて、自装置が有する記憶装置に書き込む。そして、書き込み完了を示すレスポンスをサーバ装置１００に送信する。

[0090] （ステップＳ３４）サーバ装置１００は、サーバ装置１００ａで書き込みが完了したことを確認すると、クライアント装置３００にレスポンスを送信する。

（ステップＳ３５）サーバ装置１００ａは、データの書き込みによって、セグメントのデータ量が閾値を超えたことを検出する。すると、サーバ装置１００ａは、当該セグメントをｂ個（例えば、２個）のセグメントに更に分

割する。

[0091] （ステップＳ３６）サーバ装置１００ａは、ラベルとハッシュ関数を用いて、分割後のセグメントが配置されるサーバ装置（ここでは、サーバ装置１００ａ，１００ｂ）を特定する。そして、サーバ装置１００ｂに配置されるセグメントのキーとデータをサーバ装置１００ｂに送信する。なお、前述の通り、１つのセグメントを２つの子セグメントに分割したとき、一方の子セグメントが元のセグメントと同じサーバ装置に配置され、他の子セグメントが元のセグメントと異なるサーバ装置に配置されるようにすることもできる。

[0092] （ステップＳ３７）サーバ装置１００ｂは、サーバ装置１００ａから受信したキーとデータを対応付けて、自装置が有する記憶装置に書き込む。そして、書き込み完了を示すレスポンスをサーバ装置１００ａに送信する。

[0093] （ステップＳ３８）サーバ装置１００ａは、サーバ装置１００ｂへの複製が完了したセグメントのキーとデータを、自装置が有する記憶装置から削除する。また、サーバ装置１００ａは、セグメントの分割が反映されるように、自装置が有するセグメント情報を更新する。そして、サーバ装置１００ａは、セグメントの更新を示す通知を、他の全てのサーバ装置（サーバ装置１００，１００ｂ，１００ｃ）に送信する。

[0094] （ステップＳ３９）サーバ装置１００，１００ｂ，１００ｃは、サーバ装置１００ａからの通知に基づいて、自装置が有するセグメント情報を更新する。そして、セグメント情報の更新完了を示すレスポンスを、サーバ装置１００ａに送信する。これにより、サーバ装置１００ａによるセグメント情報の更新が、サーバ装置１００，１００ｂ，１００ｃが有するセグメント情報に反映され、セグメント情報が同期された状態となる。

[0095] 第２の実施の形態によれば、サーバ装置１００，１００ａ，１００ｂ，１００ｃはキーとデータを対応付けて記憶し、クライアント装置３００から指定されたキーについてデータ処理を行うので、データ処理の複雑性を低減でき負荷が低くなる。また、データが複数のサーバ装置に分散して記憶される

と共に、何れのサーバ装置もクライアント装置３００からリクエストを受け付けることができるので、ボトルネックとなる装置が生じないように設計可能であり、情報処理システムの可用性が向上する。

[0096] また、キー空間を複数のセグメントに分割し、セグメント単位でデータがサーバ装置に配置されるので、値の近いキーに対応するデータ同士が同一のサーバ装置に記憶される確率が高くなり、キーの範囲を指定したデータ処理を効率的に実行できる。また、セグメント分割の要否は、セグメントのラベルに基づいてセグメント毎に独立に判定できるので、各サーバ装置が他のサーバ装置とネゴシエーションを行わずに判定でき、スケーラビリティを向上できる。また、セグメント分割の要否を判定するための閾値が、セグメント間で異なるように算出されるので、複数のセグメントの分割が同時期に集中する確率を低減することができ、情報処理システムの負荷を軽減できる。

[0097] なお、前述のように、第２の実施の形態のデータ管理は、コンピュータとしてのサーバ装置１００，１００ａ，１００ｂ，１００ｃに、それぞれデータ管理プログラムを実行させることで実現できる。プログラムは、コンピュータ読み取り可能な記録媒体（例えば、記録媒体３３）に記録しておくことができる。記録媒体として、例えば、磁気ディスク、光ディスク、光磁気ディスク、半導体メモリなどを使用できる。磁気ディスクには、ＦＤおよびＨＤＤが含まれる。光ディスクには、ＣＤ、ＣＤ－Ｒ（Recordable）／ＲＷ（Rewritable）、ＤＶＤおよびＤＶＤ－Ｒ／ＲＷが含まれる。

[0098] プログラムを流通させる場合、例えば、当該プログラムを記録した可搬記録媒体が提供される。また、プログラムを他のコンピュータの記憶装置に格納しておき、ネットワーク４１経由でプログラムを配布することもできる。コンピュータは、例えば、可搬記録媒体に記録されたプログラムまたは他のコンピュータから受信したプログラムを記憶装置（例えば、ＨＤＤ１０３）に格納し、当該記憶装置からプログラムを読み込んで実行する。ただし、可搬記録媒体から読み込んだプログラムを直接実行してもよく、他のコンピュータからネットワーク４１を介して受信したプログラムを直接実行してもよ

い。

[0099] 上記については単に本発明の原理を示すものである。更に、多数の変形や変更が当業者にとって可能であり、本発明は上記に示し、説明した正確な構成および応用例に限定されるものではなく、対応する全ての変形例および均等物は、添付の請求項およびその均等物による本発明の範囲とみなされる。

符号の説明

[0100] １０ 情報処理装置

１１ 記憶部

１２ 制御部

２１，２２ 記憶装置

請求の範囲

[請求項1] キーの集合を複数のセグメントに分割し、キーと対応付けて記憶装置に記憶されるデータをセグメント単位で管理する情報処理システムに用いられる情報処理装置であって、

前記複数のセグメントのうち少なくとも第１および第２のセグメントについて、キーとセグメントとの関係を示すセグメント情報を記憶する記憶部と、

前記セグメント情報を参照し、前記第１のセグメントに属するキーと対応付けて記憶されているデータの量と第１の閾値とに基づいて、前記第１のセグメントを２以上のセグメントに分割するか判定する制御部と、を有し、

前記第１の閾値は、前記第２のセグメントを分割するか判定するために用いられる第２の閾値とは異なる値をとる、情報処理装置。

[請求項2] 前記セグメント情報は、前記第１および第２のセグメントの識別情報を含み、

前記第１の閾値は、前記第１のセグメントの識別情報に基づいて算出され、前記第２の閾値は、前記第２のセグメントの識別情報に基づいて算出される、

請求の範囲第１項記載の情報処理装置。

[請求項3] 前記セグメント情報は、前記複数のセグメントを階層構造で表現し、

前記複数のセグメントそれぞれの閾値は、少なくとも同じ階層レベルのセグメント間では異なる値になるように算出される、

請求の範囲第１項または第２項記載の情報処理装置。

[請求項4] 前記情報処理システムは、複数の記憶装置を有し、前記複数の記憶装置の中からデータを格納する記憶装置をセグメント単位で選択し、

前記制御部は、前記第１のセグメントから分割された前記２以上のセグメントの少なくとも１つについて、データを格納する記憶装置が

変更されるよう制御する、

請求の範囲第１項乃至第３項の何れか一項に記載の情報処理装置。

［請求項5］ キーの集合を複数のセグメントに分割し、キーと対応付けて記憶装置に記憶されるデータをセグメント単位で管理する情報処理システムが行うデータ管理方法であって、

前記複数のセグメントのうちの第１のセグメントに属するキーと対応付けて前記記憶装置に記憶されているデータの量を算出し、

算出したデータの量と前記第１のセグメントに応じた第１の閾値とに基づいて、前記第１のセグメントを２以上のセグメントに分割するか判定し、

前記第１の閾値は、前記複数のセグメントのうちの第２のセグメントを分割するか判定するために用いられる第２の閾値とは異なる値をとる、データ管理方法。

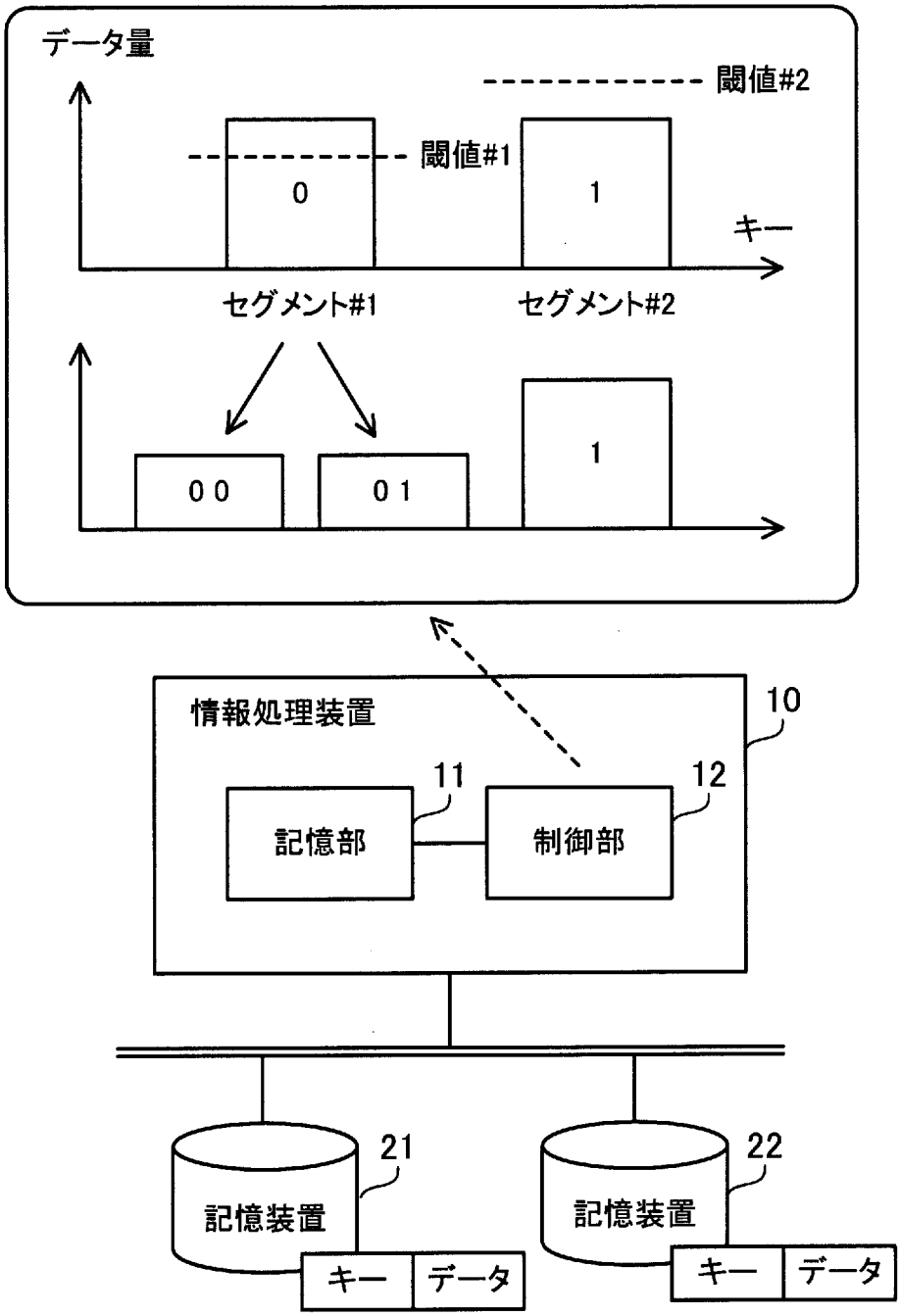
［請求項6］ キーの集合を複数のセグメントに分割し、キーと対応付けて記憶装置に記憶されるデータをセグメント単位で管理する情報処理システムに用いられるコンピュータに実行させるデータ管理プログラムであって、前記コンピュータに、

前記複数のセグメントのうちの第１のセグメントに属するキーと対応付けて前記記憶装置に記憶されているデータの量を算出し、

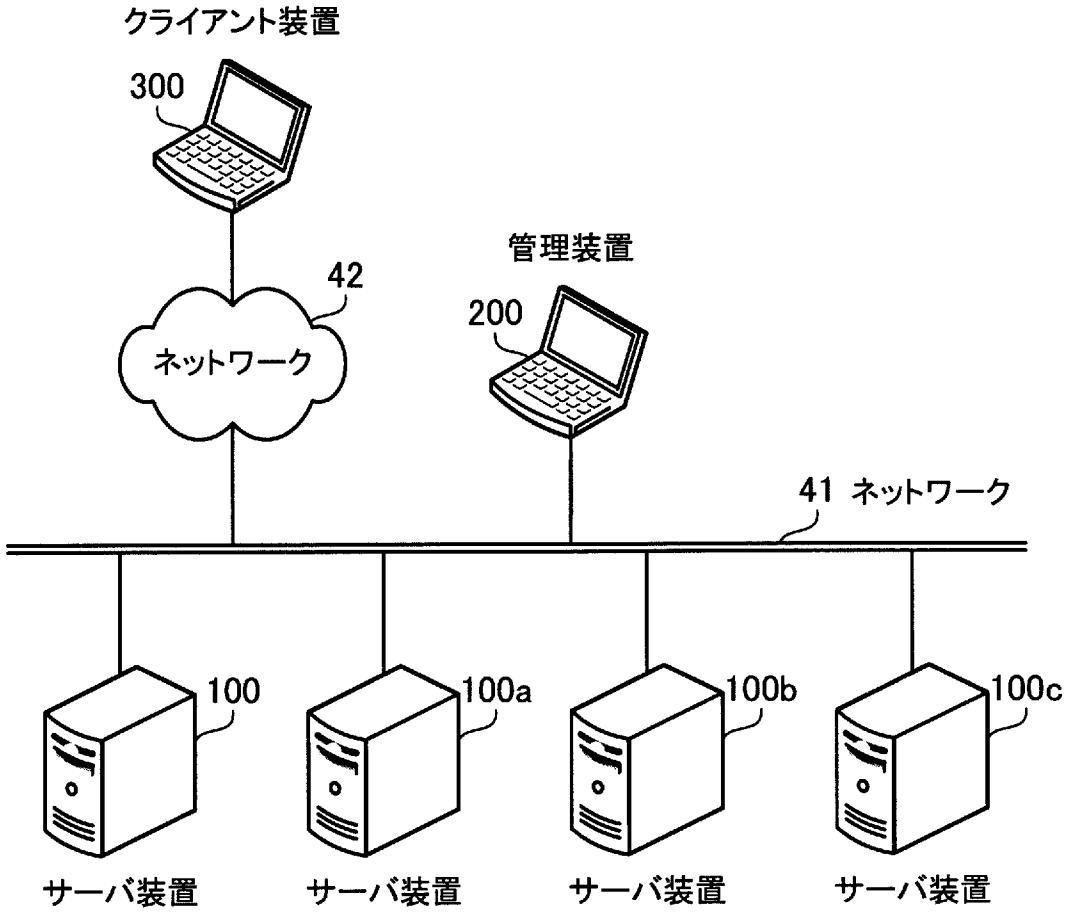
算出したデータの量と前記第１のセグメントに応じた第１の閾値とに基づいて、前記第１のセグメントを２以上のセグメントに分割するか判定する、処理を実行させ、

前記第１の閾値は、前記複数のセグメントのうちの第２のセグメントを分割するか判定するために用いられる第２の閾値とは異なる値をとる、データ管理プログラム。

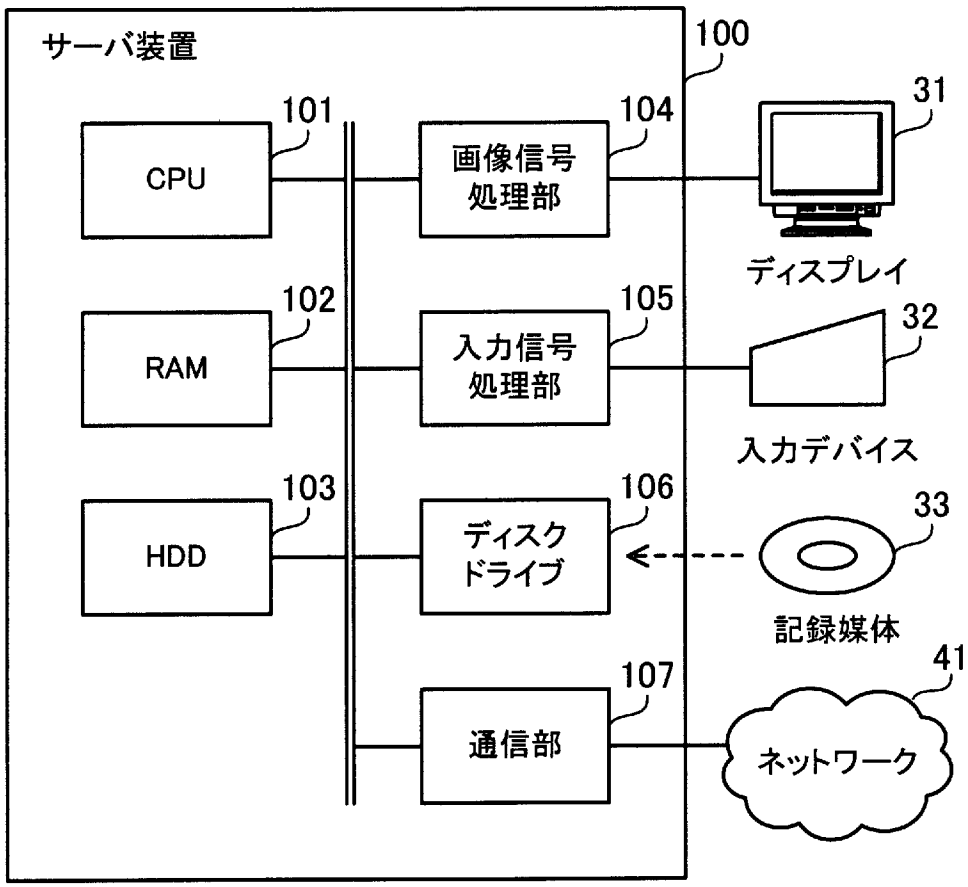
[図1]


データ量
閾値#2
閾値#1
0
1
キー
セグメント#1
セグメント#2
00
01
1
情報処理装置
10
11
12
記憶部
制御部
21
22
記憶装置
記憶装置
キー
データ
キー
データ

[図2]

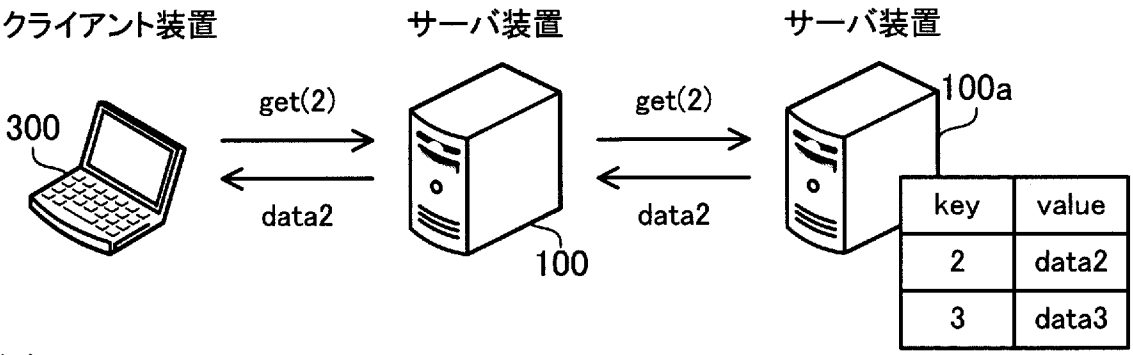

クライアント装置
300
42
ネットワーク
管理装置
200
41 ネットワーク
100
100a
100b
100c
サーバ装置
サーバ装置
サーバ装置
サーバ装置

[図3]

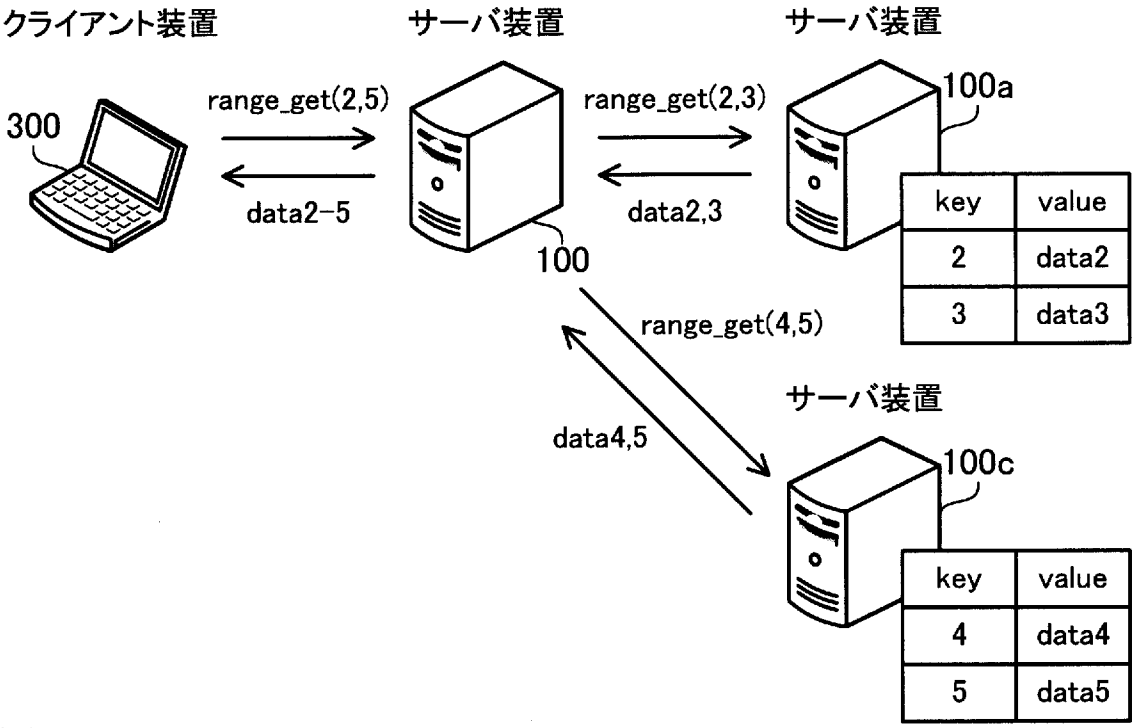

サーバ装置
100
101
CPU
102
RAM
103
HDD
104
画像信号処理部
105
入力信号処理部
106
ディスクドライブ
107
通信部
31
ディスプレイ
32
入力デバイス
33
記録媒体
41
ネットワーク

[図4]

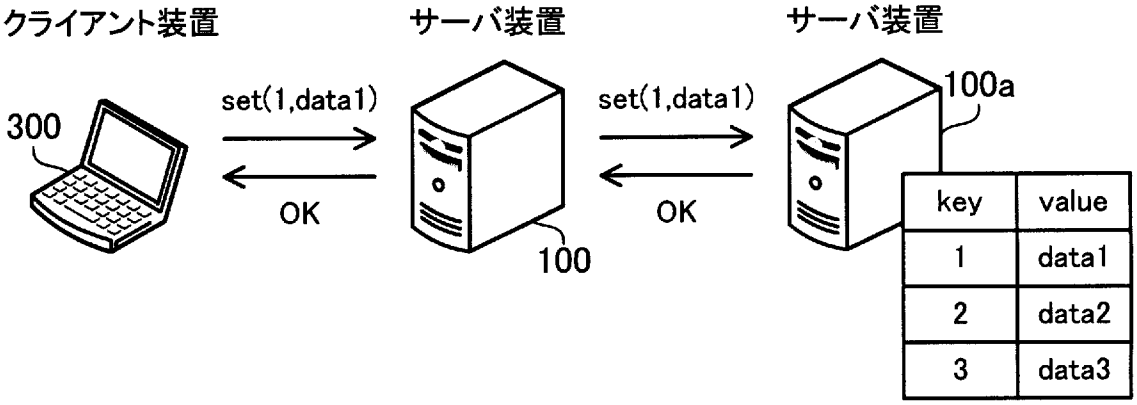
(A)


クライアント装置
サーバ装置
サーバ装置
300
get(2)
data2
get(2)
data2
100
100a
key
value
2
data2
3
data3

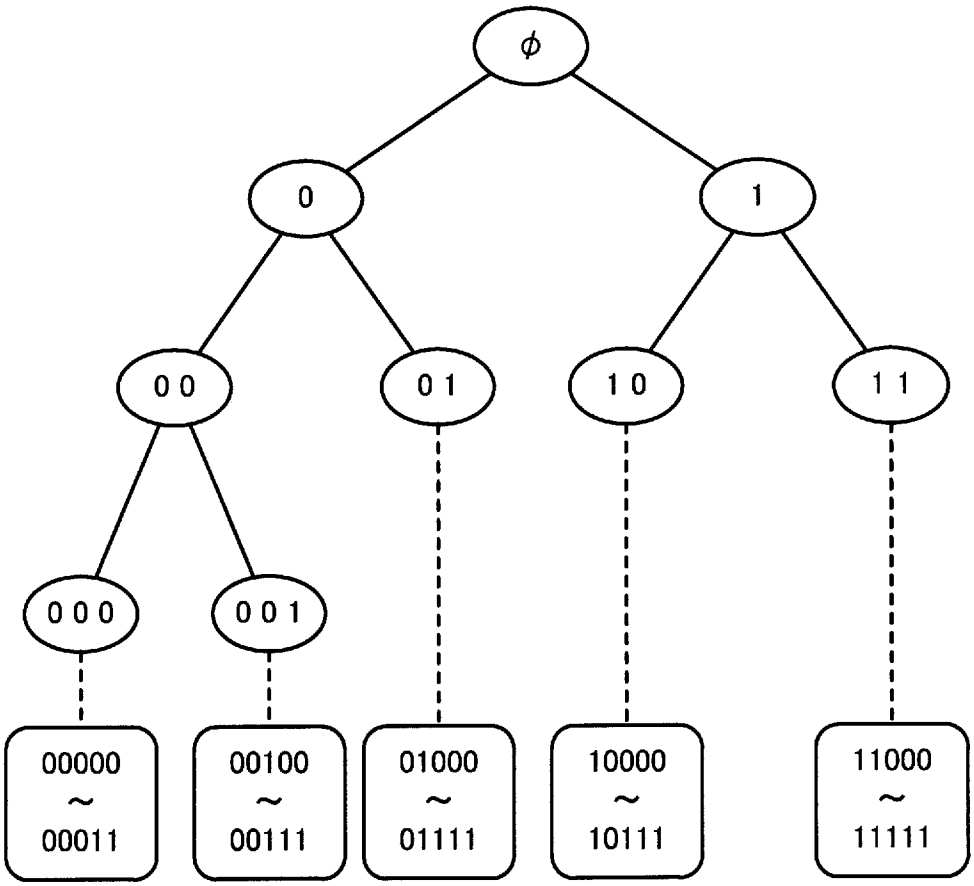
(B)


クライアント装置
サーバ装置
サーバ装置
300
range_get(2,5)
data2-5
range_get(2,3)
data2,3
100
100a
key
value
2
data2
3
data3
range_get(4,5)
data4,5
サーバ装置
100c
key
value
4
data4
5
data5

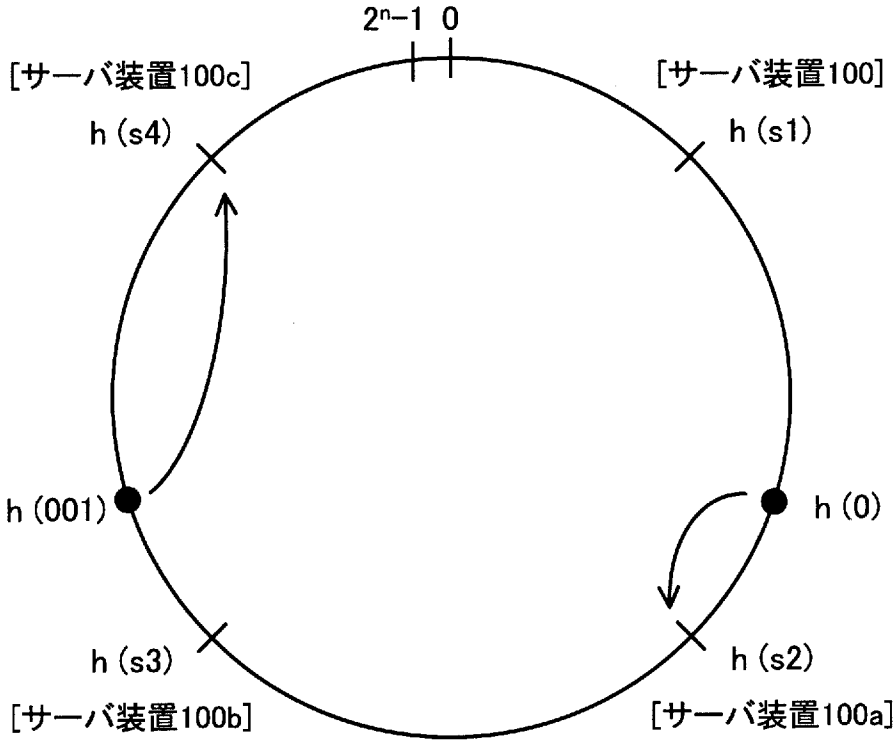
(C)


クライアント装置
サーバ装置
サーバ装置
300
set(1,data1)
OK
set(1,data1)
OK
100
100a
key
value
1
data1
2
data2
3
data3

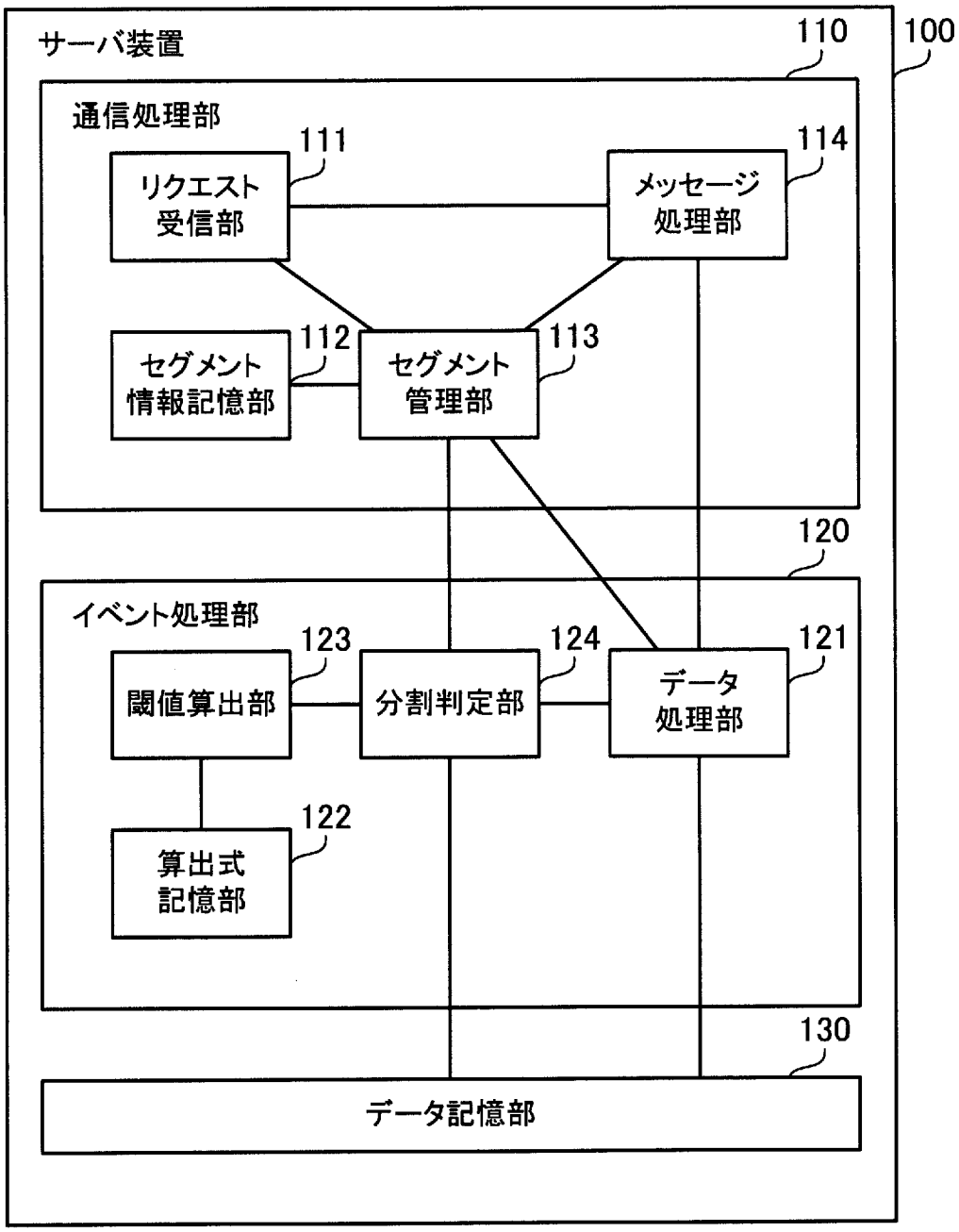
[図5]


φ
0
1
0 0
0 1
1 0
1 1
0 0 0
0 0 1
00000 ～ 00011
00100 ～ 00111
01000 ～ 01111
10000 ～ 10111
11000 ～ 11111

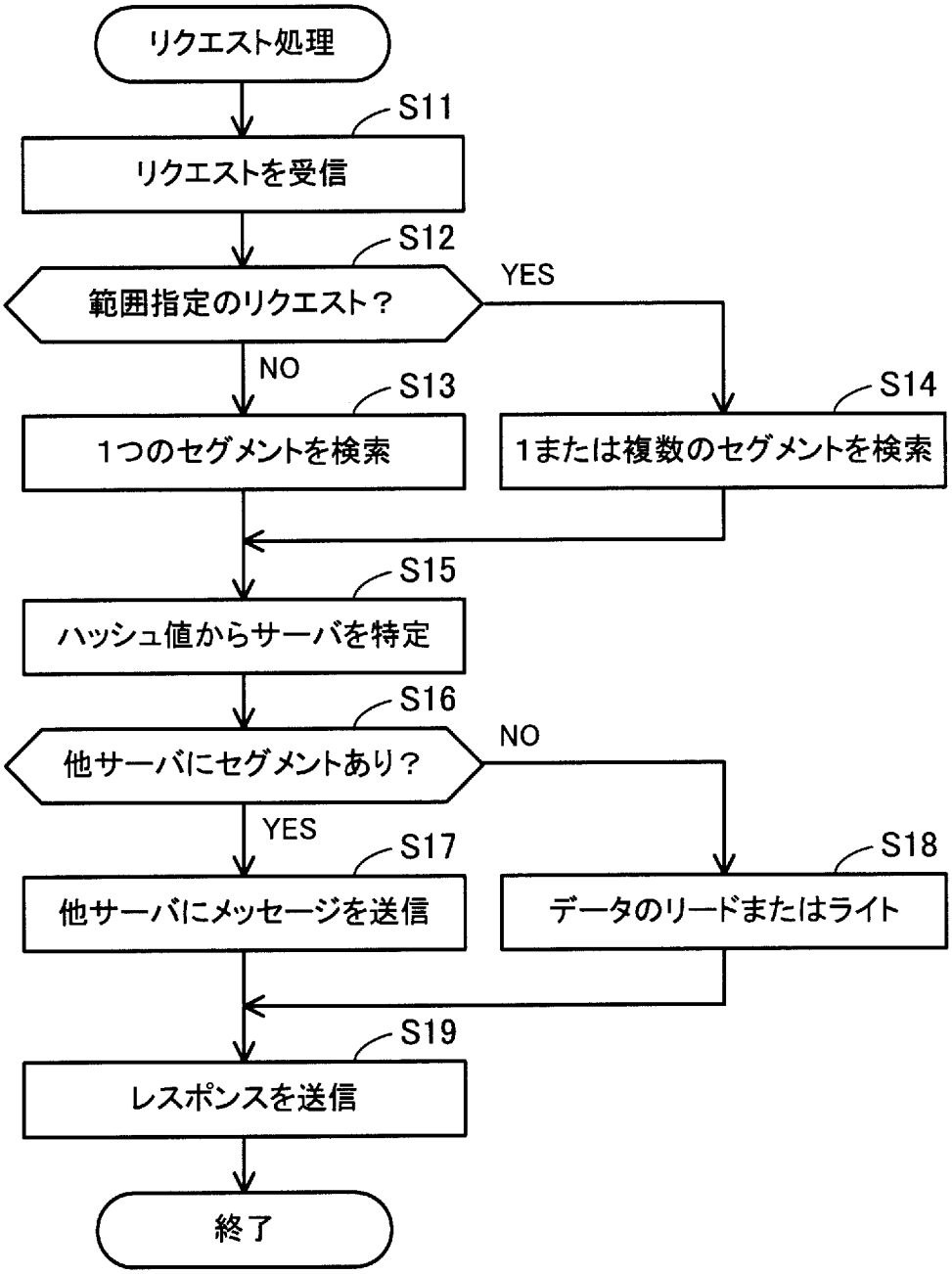
[図6]


2n−1 0
[サーバ装置100c]
h (s4)
[サーバ装置100]
h (s1)
h (001)
h (0)
h (s3)
[サーバ装置100b]
h (s2)
[サーバ装置100a]

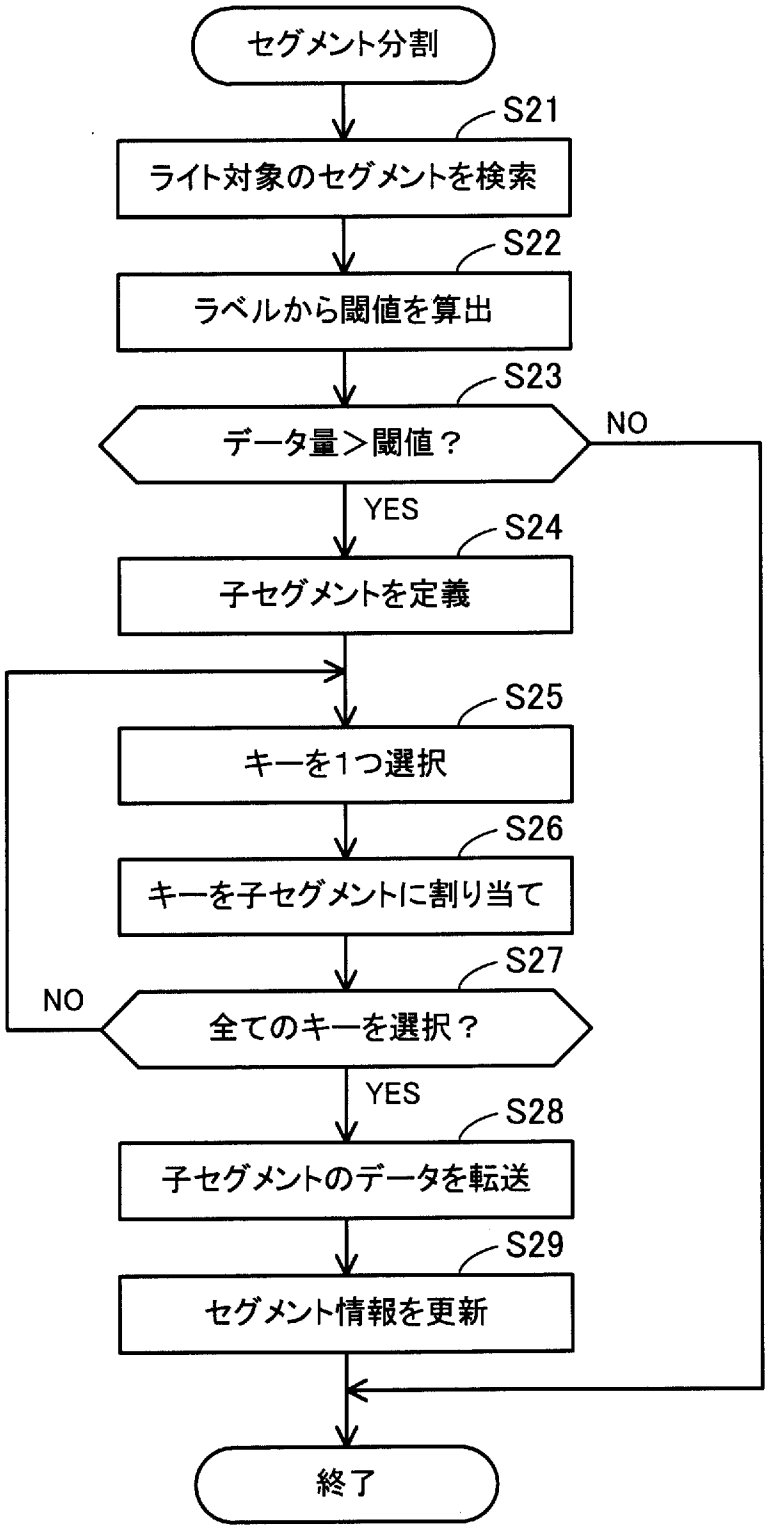
[図7]


サーバ装置
100
110
通信処理部
111
リクエスト
受信部
114
メッセージ
処理部
112
セグメント
情報記憶部
113
セグメント
管理部
120
イベント処理部
123
閾値算出部
124
分割判定部
121
データ
処理部
122
算出式
記憶部
130
データ記憶部

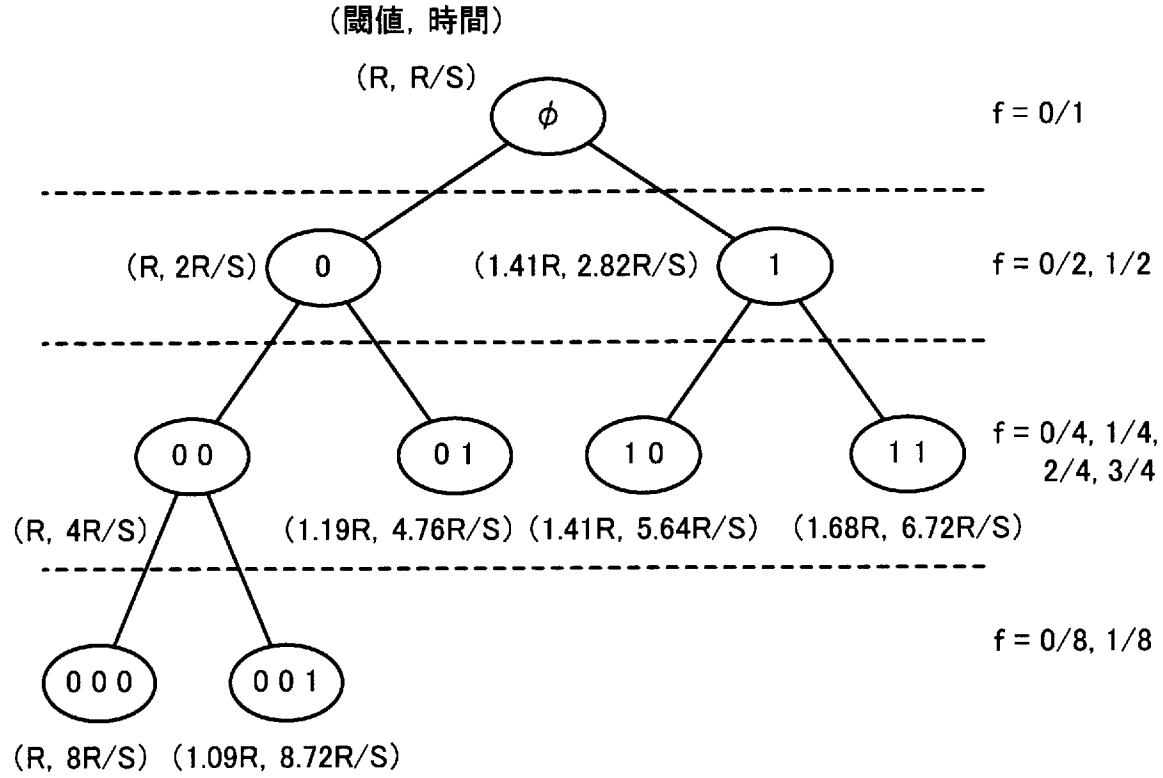
[図8]


リクエスト処理
S11
リクエストを受信
S12
範囲指定のリクエスト？
YES
NO
S13
1つのセグメントを検索
S14
1または複数のセグメントを検索
S15
ハッシュ値からサーバを特定
S16
他サーバにセグメントあり？
NO
YES
S17
他サーバにメッセージを送信
S18
データのリードまたはライト
S19
レスポンスを送信
終了

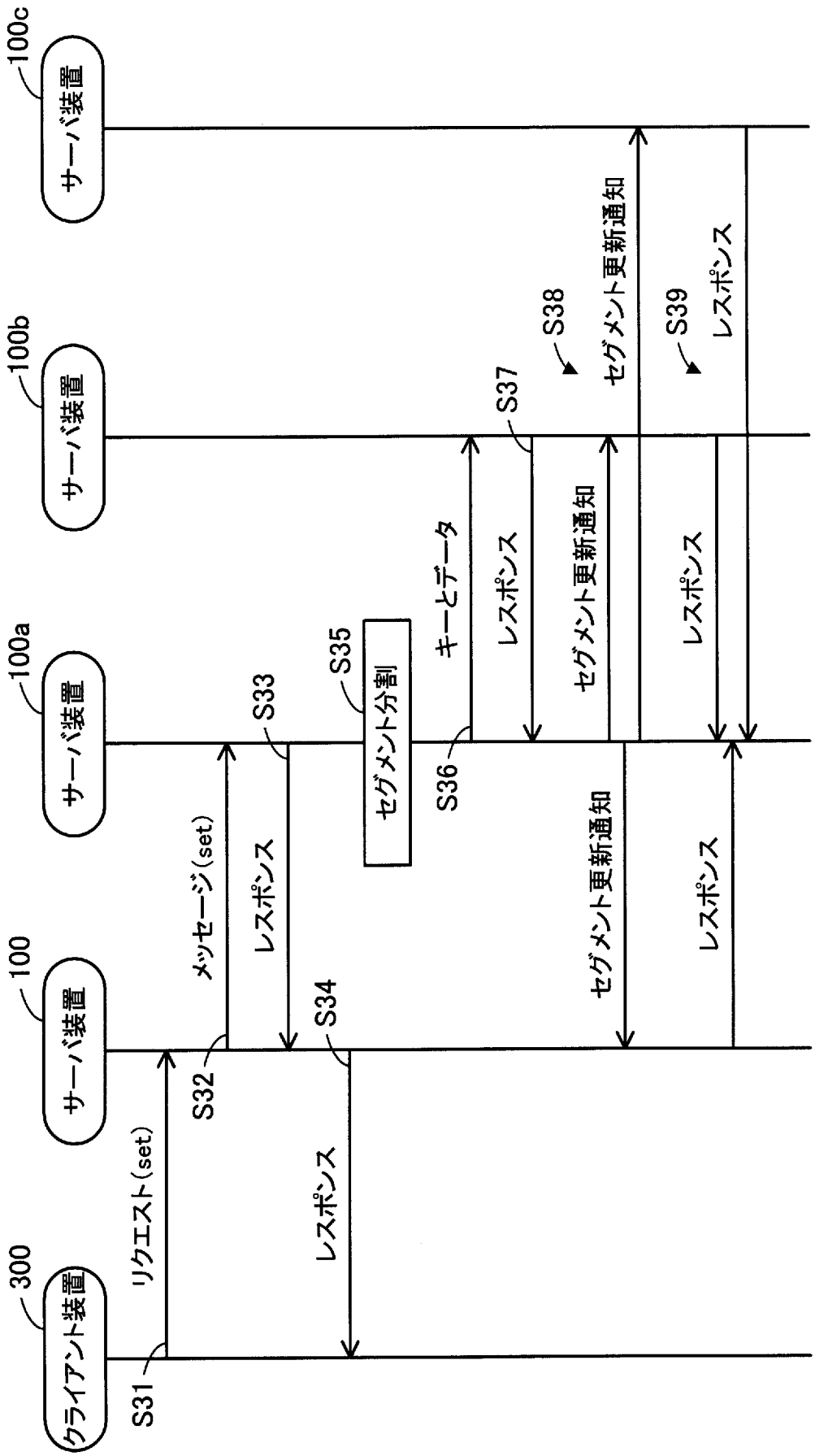
[図9]


セグメント分割
S21
ライト対象のセグメントを検索
S22
ラベルから閾値を算出
S23
データ量＞閾値？
NO
YES
S24
子セグメントを定義
S25
キーを1つ選択
S26
キーを子セグメントに割り当て
S27
NO
全てのキーを選択？
YES
S28
子セグメントのデータを転送
S29
セグメント情報を更新
終了

[図10]


(閾値, 時間)
(R, R/S)
φ
f = 0/1
(R, 2R/S)
0
(1.41R, 2.82R/S)
1
f = 0/2, 1/2
0 0
0 1
1 0
1 1
f = 0/4, 1/4, 2/4, 3/4
(R, 4R/S)
(1.19R, 4.76R/S)
(1.41R, 5.64R/S)
(1.68R, 6.72R/S)
f = 0/8, 1/8
0 0 0
0 0 1
(R, 8R/S)
(1.09R, 8.72R/S)

[図11]


300
クライアント装置
100
サーバ装置
100a
サーバ装置
100b
サーバ装置
100c
サーバ装置
S31
リクエスト(set)
S32
メッセージ(set)
S33
レスポンス
S34
レスポンス
S35
セグメント分割
S36
キーとデータ
S37
レスポンス
S38
セグメント更新通知
S39
レスポンス
セグメント更新通知
レスポンス
セグメント更新通知
レスポンス

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2011/062759

A. CLASSIFICATION OF SUBJECT MATTER
G06F12/00(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06F12/00, G06F13/00, H04L12/56

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
Jitsuyo Shinan Koho 1922-1996 Jitsuyo Shinan Toroku Koho 1996-2011
Kokai Jitsuyo Shinan Koho 1971-2011 Toroku Jitsuyo Shinan Koho 1994-2011

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
JSTPlus(JDreamII), IEEE Xplore

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X A	JP 2009-164940 A (KDDI Corp.), 23 July 2009 (23.07.2009), paragraphs [0030] to [0047]; fig. 1 to 3 (Family: none)	1,4-6 2-3
A	JP 2007-280304 A (Brother Industries, Ltd.), 25 October 2007 (25.10.2007), paragraphs [0030] to [0074], [0183] to [0204] & US 2008/0319956 A1 & WO 2007/119418 A1	1-6
A	JP 2010-79888 A (NEC Laboratories America Inc.), 08 April 2010 (08.04.2010), entire text; all drawings & US 2010/0082636 A & EP 2178003 A2 & CA 2675328 A & CN 101685467 A & CA 2675328 A1	1-6

☒ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
23 June, 2011 (23.06.11)	05 July, 2011 (05.07.11)
Name and mailing address of the ISA/ Japanese Patent Office Facsimile No.	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2011/062759

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	Masatoshi SEKINE, Kaoru SEZAKI, "Adaptive Content Placement for Frequent Data Updates in P2P Network", IEICE Technical Report, 01 March 2007 (01.03.2007), vol.106, no.578, pages 197 to 202	1-6

国際調査報告

国際出願番号　ＰＣＴ／ＪＰ２０１１／０６２７５９

Ａ．　発明の属する分野の分類（国際特許分類（ＩＰＣ））

Int.Cl.　G06F12/00(2006.01)i

Ｂ．　調査を行った分野

調査を行った最小限資料（国際特許分類（ＩＰＣ））

Int.Cl.　G06F12/00, G06F13/00, H04L12/56

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	１９２２－１９９６年
日本国公開実用新案公報	１９７１－２０１１年
日本国実用新案登録公報	１９９６－２０１１年
日本国登録実用新案公報	１９９４－２０１１年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

JSTPlus(JDreamII)
IEEE Xplore

Ｃ．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求項の番号
X A	JP 2009-164940 A（ＫＤＤＩ株式会社）2009.07.23, [0030]-[0047], 第1-3図（ファミリーなし）	1, 4-6 2-3
A	JP 2007-280304 A（ブラザー工業株式会社）2007.10.25, [0030]-[0074], [0183]-[0204] & US 2008/0319956 A1 & WO 2007/119418 A1	1-6

☑　Ｃ欄の続きにも文献が列挙されている。

☐　パテントファミリーに関する別紙を参照。

＊　引用文献のカテゴリー
「Ａ」特に関連のある文献ではなく、一般的技術水準を示すもの
「Ｅ」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
「Ｌ」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
「Ｏ」口頭による開示、使用、展示等に言及する文献
「Ｐ」国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献
「Ｔ」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
「Ｘ」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
「Ｙ」特に関連のある文献であって、当該文献と他の１以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
「＆」同一パテントファミリー文献

国際調査を完了した日 ２３．０６．２０１１	国際調査報告の発送日 ０５．０７．２０１１
国際調査機関の名称及びあて先 日本国特許庁（ＩＳＡ／ＪＰ） 郵便番号１００－８９１５ 東京都千代田区霞が関三丁目４番３号	特許庁審査官（権限のある職員）　5M　４５３５ 北村　学 電話番号　０３－３５８１－１１０１　内線　３５９９

Ｃ（続き）．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求項の番号
A	JP 2010-79888 A (エヌイーシー ラボラトリーズ アメリカ インク) 2010.04.08, 全文, 全図 & US 2010/0082636 A & EP 2178003 A2 & CA 2675328 A & CN 101685467 A & CA 2675328 A1	1-6
A	関根 理敏, 瀬崎 薫, Ｐ２Ｐネットワークにおける高頻度なデータ更新を考慮した適応的コンテンツ配置法, 電子情報通信学会技術研究報告, 2007.03.01, 第106巻 第578号, p. 197～202	1-6