

(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
2013 年 10 月 17 日(17.10.2013)



(10) 国際公開番号

WO 2013/153646 A1

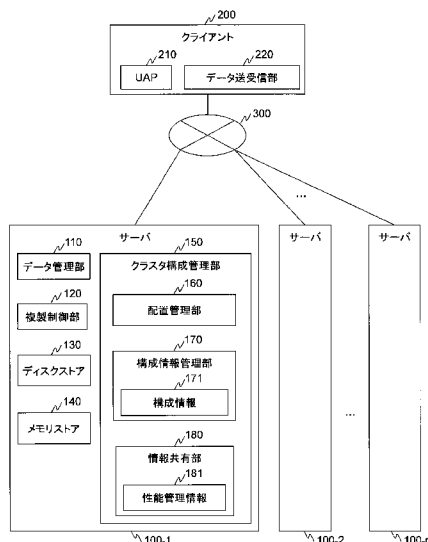
- (51) 国際特許分類:
G06F 12/00 (2006.01) G06F 17/30 (2006.01)
- (21) 国際出願番号: PCT/JP2012/059972
- (22) 国際出願日: 2012 年 4 月 12 日(12.04.2012)
- (25) 国際出願の言語: 日本語
- (26) 国際公開の言語: 日本語
- (71) 出願人 (米国を除く全ての指定国について): 株式会社日立製作所 (HITACHI, LTD.) [JP/JP]; 〒1008280 東京都千代田区丸の内一丁目 6 番 6 号 Tokyo (JP).
- (72) 発明者; および
- (75) 発明者/出願人 (米国についてのみ): 福中 勝博 (FUKUNAKA, Katsuhiko) [JP/JP]; 〒2448555 神奈川県横浜市戸塚区戸塚町 5 0 3 0 番地 株式会社日立製作所 IT プラットフォーム事業本部内 Kanagawa (JP).
- (74) 代理人: 藤井 正弘 (FUJII, Masahiro); 〒1050001 東京都港区虎ノ門一丁目 1 6 番 4 号アーバン虎ノ門ビル 後藤特許事務所 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, QA, RO, RS, RU, RW, 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, RW, SD, SL, SZ, TZ, UG, ZM, ZW), ユーロパ (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), ユーロッパ (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, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR),

[続葉有]

(54) Title: COMPUTER SYSTEM, DATA ALLOCATION MANAGEMENT METHOD, AND PROGRAM

(54) 発明の名称: 計算機システム、データ配置管理方法及びプログラム

[図1]



100-1, 100-2, 100-n Server
110 Data management unit
120 Copy control unit
130 Disk store
140 Memory store
150 Cluster configuration management unit
160 Allocation management unit
170 Configuration information management unit
171 Configuration information
180 Information sharing unit
181 Performance management information
200 Client
220 Data transmission and reception unit

(57) Abstract: The purpose of the invention is to achieve in cloud computing a computer system that automatically sets a KVS configuration taking into account the differences in performance among computers. This computer system is configured by connecting a plurality of computers over a network and uses a database formed from storage areas of the computers to perform a task. Each computer holds performance management information for managing the performance of the computers. Each computer constituting the database has master data and copy data of master data of another computer allocated thereto. The computer system includes a performance information management unit that updates the performance management information, and a cluster configuration management unit that determines, on the basis of the performance management information, a first management range for the master data managed by a first computer and a second computer that holds copy data of the master data managed by the first computer.

(57) 要約: クラウドコンピューティングにおいて、計算機間の性能差を考慮して、KVSの構成を自動的に設定する計算機システムを実現する。ネットワークを介して複数の計算機が接続され、各計算機が有する記憶領域から構成されたデータベースを用いて業務を実行する計算機システムであって、各計算機は計算機の性能を管理する性能管理情報を保持し、データベースを構成する各計算機には、マスターデータ及び他の計算機のマスターデータの複製データが配置され、計算機システムは、性能管理情報を更新する性能情報管理部と、性能管理情報に基づいて、第1の計算機が管理するマスターデータの第1の管理範囲、及び、前記第1の計算機が管理するマスターデータの複製データを保持する第2の計算機を決定するクラスタ構成管理部と、を備える。

WO 2013/153646 A1



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：

計算機システム、データ配置管理方法及びプログラム

技術分野

[0001] 本発明は、複数の計算機から構成される分散データベースに関する。特に、自動的に分散データベースを構築するための設定処理に関する。

背景技術

[0002] 近年、Webを用いたアプリケーションを実行する計算システムにおいてデータ量が爆発的に増大しており、複数の計算機にデータを分散させることによって、データへのアクセス性能を向上させるシステムが様々知られている。例えば、RDBMS (Relational DataBase Management System) では、データを所定の範囲毎に分割し、分割されたデータを複数の計算機に配置することによって、システム全体のアクセス性能を向上させる方法が知られている（例えば、特許文献1参照）。

[0003] 特許文献1には、ネットワーク上の複数のコンピュータに配置されたデータベースに格納された各データの更新処理を、ネットワーク上で唯一のオリジナルサイトでのみ実行し、他のレプリカサイトにおいては、オリジナルサイトで実行された更新結果を受信して、自己が保有するレプリカデータに反映すること発明が記載されている。これによって、ネットワーク上の複数のコンピュータが利用するデータの同一性を保持することができる。

[0004] また、キャッシュサーバ等に用いられるシステムとして、データの識別子であるキーと、データ値（バリュー）とから構成されるキャッシュデータを所定の分散方法にしたがって、複数の計算機システムに配置するKVS (Key Value Store) 等のNoSQL (Not only SQL) データベースが知られている。

[0005] KVSでは、データに高速にアクセス可能な揮発性の記憶媒体、例えばメ

モリに、データを格納する構成、データ格納の永続性に優れる不揮発性の記録媒体、例えばSSD (Solid State Disk) やHDD等に、データを格納する構成、又は、前述したものを併用する構成等の種々の構成がとられている。

[0006] 併用する構成では、高速アクセス性重視や格納性重視等の種々の運用方針によって、複数台の計算機のメモリを統合して構成されるメモリストア及び1台以上の計算機の不揮発性記憶媒体から構成されるディスクストアのバランスを種々変更可能となっている。

[0007] メモリストア及びディスクストアには、データ（バリュー）と、データの識別子（キー）とをペアとしたデータが格納される。

[0008] また、KVSでは、複数のサーバからクラスタを構成して、そのクラスタに含まれるサーバにデータを分散して配置することによって並列処理を実現している。具体的には、サーバが管理するデータの範囲である管理範囲（例えば、キーレンジ）毎に、各サーバにデータが格納される。各サーバは、担当する管理範囲に含まれるデータのマスターとして処理を実行する。すなわち、所定のキーを含む読み出し要求に対して、そのキーが含まれる管理範囲のデータを担当するサーバが、キーに対応するデータを読み出すこととなる。

[0009] したがって、KVSでは、スケールアウトによって並列処理の性能を向上させることができる。

[0010] KVSでは、データの信頼性を確保するために、クラスタを構成するサーバが他のサーバが管理するデータの複製データを格納する構成を採るものが知られている。すなわち、各サーバは、所定の管理範囲に含まれるデータを管理するマスターであると同時に、他のサーバが管理するデータの複製データを保持するスレーブでもある。これによって、サーバに障害が発生した場合であっても、その障害サーバがマスターとして管理するデータを、スレーブである他のサーバが、自己が保持する複製データをマスターデータとすることによって処理を継続することができる。

[0011] なお、以下、マスターであるサーバをマスターサーバと記載し、スレーブであ

るサーバをスレーブサーバとも記載する。

[0012] 前述したようにK V Sを構成するサーバには、管理サーバのような特別なサーバが存在しないため単一点障害が存在しない。すなわち、任意のサーバに障害が発生した場合であっても他のサーバが処理を継続できるため、計算機システムが停止することがない。したがって、K V Sは耐障害性も確保される。

[0013] なお、スレーブサーバの台数、すなわち、複製データの格納先となるサーバの台数は、計算機システムによって任意に設定することができる。

[0014] K V Sなどで用いられるデータの分散配置方法としては、C o n s i s t e n t H a s h i n g法、R a n g e法及びL i s t法等の種々の方法が用いられる。

[0015] 例えば、C o n s i s t e n t H a s h i n g法では、まず、キーに対するハッシュ値を算出し、算出したハッシュ値をサーバの台数で除算した余りを算出する。その余りがサーバの識別番号に一致するサーバにデータが配置されるようになっている。

先行技術文献

特許文献

[0016] 特許文献1：特開2002-297428号公報

発明の概要

発明が解決しようとする課題

[0017] 従来のオンプレミス（例えば、同一企業内でのシステム運用）では、同一性能のサーバを用いて分散K V Sを構築及び運用することが一般的である。しかし、クラウドコンピューティングでは、性能の異なるサーバを用いて分散K V Sを構築及び運用する必要がある。この場合、各サーバの性能差を考慮しないとシステムの性能が低下する可能性がある。

[0018] C o n s i s t e n t H a s h i n g法を用いて分散K V Sを構築する場合、従来のシステムでは等間隔にデータが分散配置されるが、クラウドコ

ンピューティングでは、サーバの性能差を考慮して、各サーバが担当するデータの割当量、すなわち、管理範囲を決定する必要がある。また、スレーブサーバを設定する場合には、マスタサーバの性能とスレーブサーバの性能とを考慮する必要がある。

[0019] 例えば、メモリ容量の小さいサーバが、メモリ容量の大きいサーバの複製データを保持するように設定された場合、メモリ容量の小さいサーバのメモリ内に複製データが全て格納できない。したがって、HDD等の記憶装置に一部の複製データを格納する必要があるため、システム全体としてのアクセス性能が低下する。

[0020] また、メモリ容量の大きいサーバが、メモリ容量の小さいサーバの複製データを保持するように設定された場合、メモリ容量の大きいサーバのメモリ使用量が小さいため、システム全体としてメモリ使用効率が低下する。

[0021] 前述の理由によって、従来技術では、クラウドコンピューティングを用いて分散KVSを構築する場合、管理者が手動でデータを分散配置するための設定（管理範囲及びスレーブサーバを設定）する必要がある。

[0022] 本発明は、前述した課題に鑑みてなされた発明である。すなわち、クラウドコンピューティングを用いて分散KVSを構築する場合に、サーバの性能に基づいて、自動的にデータの分散配置の設定することを目的の一つとする。

課題を解決するための手段

[0023] 本願において開示される発明の代表的な一例を示せば以下の通りである。すなわち、ネットワークを介して複数の計算機が接続され、前記複数の計算機が有する記憶領域から構成されたデータベースを用いて業務を実行する計算機システムであって、前記複数の計算機の各々は、プロセッサと、前記プロセッサに接続される記憶装置と、前記ネットワークを介して他の前記計算機と通信するためのネットワークインタフェースとを有し、前記複数の計算機の各々の性能を管理する性能管理情報を保持し、前記データベースには、キー及びデータ値から構成されるデータが格納され、前記複数の計算機の各

々には、前記複数の計算機の各々に割り当てる前記キーの範囲を示す管理範囲を決定するための分散アルゴリズムに基づいて、前記複数の計算機の各々がマスタとして管理するマスタデータが配置され、他の前記複数の計算機のうちの一つの計算機が管理する前記マスタデータの複製データが配置され、前記計算機システムは、前記複数の計算機の各々から性能に関する情報を取得し、取得された情報に基づいて性能管理情報を更新する性能情報管理部と、前記性能管理情報に基づいて、前記複数の計算機の各々が管理する前記マスタデータの前記管理範囲、及び、前記複数の計算機の各々が管理するマスタデータの複製データを保持する副計算機を決定するクラスタ構成管理部と、を備えることを特徴とする。

発明の効果

[0024] 本発明によれば、計算機間の性能を考慮することによって、各計算機の管理範囲、及び各計算機が管理するマスタデータの複製データを保持する計算機（スレーブサーバ）を自動的に設定することができる。したがって、クラウドコンピューティングにおいて、最適な分散KVSを構築することができる。

図面の簡単な説明

[0025] [図1]本発明の第1の実施例における計算機システムの構成を示すブロック図である。

[図2]本発明の第1の実施例におけるサーバのハードウェア構成の一例を示す説明図である。

[図3]本発明の第1の実施例におけるメモリストアに格納されるデータの一例を示す説明図である。

[図4]本発明の第1の実施例における構成情報の一例を示す説明図である。

[図5]本発明の第1の実施例における性能管理情報の一例を示す説明図である。

[図6]本発明の第1の実施例におけるサーバが実行する処理を説明するフローチャートである。

[図7]本発明の第1の実施例における入力画面の一例を示す説明図である。

[図8]本発明の第1の実施例における下り階段型アルゴリズムの一例を示す説明図である。

[図9]本発明の第1の実施例における山型アルゴリズムの一例を示す説明図である。

[図10]本発明の第1の実施例における配置決定処理の詳細を説明するフローチャートである。

[図11A]本発明の第1の実施例における性能管理情報の更新方法の一例を示す説明図である。

[図11B]本発明の第1の実施例における性能管理情報の更新方法の一例を示す説明図である。

[図11C]本発明の第1の実施例における性能管理情報の更新方法の一例を示す説明図である。

[図12A]本発明の第1の実施例における性能管理情報の更新方法の一例を示す説明図である。

[図12B]本発明の第1の実施例における性能管理情報の更新方法の一例を示す説明図である。

[図12C]本発明の第1の実施例における性能管理情報の更新方法の一例を示す説明図である。

[図13A]本発明の第1の実施例における確認画面の一例を示す説明図である。

[図13B]本発明の第1の実施例における確認画面の一例を示す説明図である。

[図14]本発明の第1の実施例の変形例における山型配列処理を説明するフローチャートである。

[図15]本発明の第2の実施例におけるクラスタ構成変更処理を説明するフローチャートである。

[図16A]本発明の第2の実施例における新規サーバの配置決定処理を説明するフローチャートである。

[図16B]本発明の第2の実施例における新規サーバの配置決定処理を説明する

フローチャートである。

[図17]本発明の第2の実施例における性能管理情報の更新方法の一例を示す説明図である。

[図18]本発明の第2の実施例における性能管理情報の更新方法の一例を示す説明図である。

発明を実施するための形態

[0026] 以下の説明では、Consistent Hashing法を適用した分散KVSを例に説明する。

第1の実施例

[0027] 図1は、本発明の第1の実施例における計算機システムの構成を示すブロック図である。

[0028] 計算機システムは、複数のサーバ100、クライアント200及びネットワーク300から構成される。各サーバ100間又はサーバ100とクライアント200との間は、ネットワーク300を介して互いに接続される。

[0029] ネットワーク300は、LAN、WAN及びSAN等の有線及び無線の種々の構成が考えられる。本発明は、サーバ100及びクライアント200が通信できるものであればどのようなネットワークであってもよい。なお、ネットワーク300には、複数のネットワーク装置（図示省略）が含まれる。ネットワーク装置は、例えば、スイッチ及びゲートウェイなどが含まれる。

[0030] 本実施例では、複数のサーバ100からクラスタを構成し、それらサーバ100が備える記憶領域上にNoSQLデータベースが構築される。本実施例ではNoSQLデータベースとして、KVSを用いるものとする。

[0031] サーバ100-1は、所定の管理範囲毎にデータが配置されており、当該管理範囲に含まれるデータを管理するマスタサーバとして稼動する。また、サーバ100-1は、他のサーバ100-2及び／又はサーバ100-nが管理する管理範囲に含まれるデータの複製データを保持しており、スレーブサーバとして稼動する。同様に、サーバ100-2及び100-nのそれぞれも、自己の管理範囲に含まれるデータを管理するマスタサーバとして機能

し、他のサーバ１００がマスタとして管理する管理範囲の複製データを保持するようになっている。

[0032] また、本実施例のクラスタは、計算機システム全体の管理を行う管理サーバとなる唯一のサーバ１００が存在せず、全てのサーバ１００を同等のサーバとして扱う構成である。これによって、一つのサーバ１００に障害が発生しても、他のスレーブサーバが新たなマスタサーバとして処理を継続することができるため、計算機システムを停止することなく処理を継続することができる。

[0033] 本実施例のサーバ１００は、データ管理部１１０、複製制御部１２０、ディスクストア１３０、メモリストア１４０及びクラスタ構成管理部１５０を有する。

[0034] ディスクストア１３０及びメモリストア１４０は、ＫＶＳを構成するデータベースである。ディスクストア１３０及びメモリストア１４０には、キーとバリューとを一組としたデータが格納される。なお、各サーバ１００のディスクストア１３０及びメモリストア１４０には、管理範囲に含まれるデータが格納される。

[0035] メモリストア１４０に対するアクセス性能は、ディスクストア１３０に対するアクセス性能より高いため、通常メモリストア１４０にデータが格納される。一方、ディスクストア１３０には、メモリストア１４０に格納できない容量のデータ及びアクセス頻度が低いデータ等が格納される。

[0036] なお、計算機システムは、マスタデータの格納については、ディスクストア１３０を利用せずにメモリストア１４０のみを利用する構成や、自及び／又は他のサーバ１００の複製データをディスクストア１３０にのみ利用する構成等、種々の運用が適用可能である。特に、分散ＫＶＳシステムは、要求元に対する応答の早さを課題の一つとする。このためメモリストア１４０のみを利用する構成では、マスタ・スレーブの切替え時を始め、全てのデータに対する応答の早さを期待できる。他方、ディスクストア１４０を併用する構成では、サーバ停止時のデータバックアップとしての効果も期待できる。

- [0037] データ管理部 110 は、サーバ 100 が管理するデータに対する各種処理を制御する。データ管理部 110 は、クライアント 200 からの要求を受け付け、その要求に基づいて、データの読み出しや書き込み等の処理を制御する。
- [0038] 複製制御部 120 は、クライアント 200 からのアクセス要求を受信し、受信したアクセス要求をデータ管理部 110 に転送する。また、複製制御部 120 は、受信したアクセス要求に対する処理の結果をクライアント 200 に送信する。また、複製制御部 120 は、サーバ 100 自身が管理する管理範囲のデータを複製し、生成された複製データを他のサーバ 100 に送信する。
- [0039] クラスタ構成管理部 150 は、複数のサーバ 100 から構成されるクラスタを管理する。クラスタに含まれるサーバ 100 を用いて分散 KVS が構築される。クラスタ構成管理部 150 は、複数のモジュールから構成され、具体的には、配置管理部 160、構成情報管理部 170 及び情報共有部 180 を含む。
- [0040] 配置管理部 160 は、各サーバ 100 の管理範囲と、各サーバ 100 に対するスレーブサーバとを決定する。より具体的には、配置管理部 160 は、管理範囲の幅と、スレーブサーバの配置関係とを決定する。
- [0041] ここで、スレーブサーバの配置関係とは、一つのサーバ 100 に対してどのサーバ 100 をスレーブサーバに設定するかを示す情報である。
- [0042] 本実施例では、配置管理部 160 は、マスタサーバとスレーブサーバとの間の性能差が小さくなるようにスレーブサーバの配置関係を決定する。また、配置管理部 160 は、処理結果を表示するための表示情報を生成する。
- [0043] 構成情報管理部 170 は、クラスタ上に構築された分散 KVS の構成を管理するための構成情報 171 を管理する。構成情報管理部 170 は、必要に応じて、構成情報 171 を更新する。構成情報 171 の詳細については、図 4 を用いて後述する。
- [0044] 情報共有部 180 は、サーバ 100 の性能に関する情報を管理し、また、

他のサーバ１００との間でサーバ１００の性能に関する情報を共有する。情報共有部１８０は、例えばハートビート等を用いることによって実現できる。

[0045] また、情報共有部１８０は、性能管理情報１８１を保持する。性能管理情報１８１は、計算機システムに含まれる全サーバ１００の性能を管理するための情報を格納する。性能管理情報１８１の詳細については、図５を用いて後述する。

[0046] 情報共有部１８０は、サーバ１００の性能に関する情報を収集し、収集された情報に基づいて、当該サーバ１００の性能情報を生成する。情報共有部１８０は、各サーバ１００から取得した性能情報に基づいて、性能管理情報１８１を生成し、又は更新する。

[0047] クライアント２００は、プロセッサ（図示省略）、メモリ（図示省略）及びネットワークインタフェース（図示省略）等を備える計算機であって、分散ＫＶＳに対する各種処理の実行を要求する。クライアント２００は、ＵＡＰ２１０及びデータ送受信部２２０を有する。

[0048] ＵＡＰ２１０は、各種機能を提供するアプリケーションであって、そのＵＡＰ２１０によって各種処理の要求がサーバ１００に送信される。データ送受信部２２０は、ＵＡＰ２１０から出力された要求をサーバ１００に送信し、また、サーバ１００からの処理結果を受信する。

[0049] 図２は、本発明の第１の実施例におけるサーバ１００のハードウェア構成の一例を示す説明図である。

[0050] サーバ１００は、プロセッサ１０１、メモリ１０２、ネットワークインタフェース１０３、ストレージインタフェース１０４及び記憶装置１０５を有する。

[0051] プロセッサ１０１は、メモリ１０２に格納されるプログラムを実行する。プロセッサ１０１がプログラムを実行することによって、サーバ１００が備える機能を実現することができる。以下、プログラムを主語に処理を説明する場合には、プロセッサ１０１によって、プログラムが実行されていること

を示すものとする。

[0052] メモリ 102 は、プロセッサ 101 が実行するプログラム及び当該プログラムの実行に必要な情報を格納する。

[0053] 本実施例のメモリ 102 上には、データ管理部 110、複製制御部 120 及びクラスタ構成管理部 150 を実現するためのプログラムが格納される。また、メモリ 102 上には、必要な情報として、構成情報 171 及び性能管理情報 181 が格納される。

[0054] さらに、メモリ 102 上には、分散 KVS を構成するデータベースであるメモリストア 140 が構築される。メモリストア 140 には、キーとバリューとを一組としたデータが格納される。メモリストア 140 に格納されるデータの詳細については、図 3 を用いて後述する。

[0055] 記憶装置 105 は、各種情報を格納する。記憶装置 105 は、例えば、HDD 又は SSD 等が考えられる。記憶装置 105 上に分散 KVS を構築するディスクストア 130 が構築される。また、記憶装置 105 には、データ管理部 110 等を実現するプログラムが格納されてもよい。この場合、プロセッサ 101 が記憶装置 105 からプログラムを読み出し、読み出されたプログラムをメモリ 102 上にロードし、ロードされたプログラムを実行する。

[0056] 図 3 は、本発明の第 1 の実施例におけるメモリストア 140 に格納されるデータの一例を示す説明図である。なお、ディスクストア 130 にも同様の形式のデータが格納される。

[0057] 本実施例では、メモリストア 140 は、データ管理情報 400 を格納する。データ管理情報 400 には、キーとバリューとがペアとなったデータが複数含まれる。以下、キーとバリューとがペアとなったデータをキーバリュー型データとも記載する。

[0058] データ管理情報 400 は、Key 401 及び Value 402 を含む。Key 401 は、データを識別するための識別子（キー）を格納する。Value 402 は、実際のデータ（バリュー）を格納する。

[0059] クライアント 200 を操作するユーザは、Key 401 を指定して分散 K

V Sにデータを保存し、また、K e y 4 0 1を指定して分散K V Sから所望のデータを取得することができる。

[0060] 各サーバ1 0 0は、所定の管理範囲毎にキーバリュ型データを管理する。すなわち、管理範囲毎にキーバリュ型データが各サーバ1 0 0に分散して配置される。サーバ1 0 0は、指定された管理範囲に含まれるデータのマスターサーバとして処理を実行することとなる。これによって、大量のデータを並列的かつ高速に処理できる。

[0061] 図4は、本発明の第1の実施例における構成情報1 7 1の一例を示す説明図である。

[0062] 構成情報1 7 1は、各サーバ1 0 0の管理範囲に関する情報を格納する。具体的には、構成情報1 7 1は、サーバI D 5 0 1及び管理範囲5 0 2を含む。

[0063] サーバI D 5 0 1は、サーバ1 0 0を一意に識別するための識別子を格納する。サーバI D 5 0 1には、サーバ1 0 0の識別子、I Pアドレス及びM A Cアドレス等が格納される。

[0064] 管理範囲5 0 2は、サーバ1 0 0が担当する管理範囲の値を格納する。本実施例では管理範囲の値としてハッシュ値が格納される。

[0065] 管理範囲5 0 2のマスター5 0 5は、サーバI D 5 0 1に対応するサーバ1 0 0がマスターとして管理する管理範囲のハッシュ値を格納する。管理範囲5 0 2のスレーブ5 0 6は、サーバI D 5 0 1に対応するサーバ1 0 0が保持する複製データの管理範囲のハッシュ値を格納する。

[0066] なお、本実施例では、一つのサーバ1 0 0が複製データを保持する構成となっているが本発明はこれに限定されない。すなわち、同一の複製データを保持するスレーブサーバが二つ以上あってもよい。

[0067] 図5は、本発明の第1の実施例における性能管理情報1 8 1の一例を示す説明図である。

[0068] 性能管理情報1 8 1は、各サーバ1 0 0の性能に関する情報を格納する。具体的には、性能管理情報1 8 1は、サーバI D 6 0 1、スペック6 0 2及

びレンジ割合603を含む。

[0069] サーバID601は、サーバID501と同一のものである。

[0070] スペック602は、サーバID601に対応するサーバ100の性能に関する情報を格納する。図5に示す例では、スペック602には、プロセッサ性能、メモリ性能及び通信性能が含まれる。なお、スペック602には、プロセッサ101とメモリ102とを接続するチャンネル数、メモリ102のクロック数、記憶装置105の容量、及びHDDの回転数等その他の情報が含まれてもよい。

[0071] プロセッサ性能は、サーバ100が有するプロセッサ101の性能を示す情報を格納する。本実施例では、プロセッサ性能には、プロセッサ101の周波数が格納される。

[0072] メモリ性能は、サーバ100が有するメモリ102の性能を示す情報を格納する。本実施例では、メモリ性能には、メモリ102の容量が格納される。

[0073] 通信性能は、サーバ100が有するネットワークインタフェース103の通信性能を示す情報を格納する。本実施例では、通信性能には、ネットワークインタフェース103の通信速度が格納される。

[0074] レンジ割合603は、サーバ100に割り当てる管理範囲の幅（割当比率）を示す情報を格納する。また、レンジ割合603は、スレーブサーバが保持する複製データのデータ量にも対応する。

[0075] 本実施例では、右隣の列（エントリ）に対応するサーバ100の複製データを保持するようにスレーブサーバが設定されるものとする。なお、最も左の列に対応するサーバ100は、最も右の列に対応するサーバ100の複製データを保持するスレーブサーバとして設定される。したがって、本実施例では、性能管理情報181のエントリ（サーバ100）の配列が複製データを保持するスレーブサーバの配置関係に対応する。

[0076] 図6は、本発明の第1の実施例におけるサーバ100が実行する処理を説明するフローチャートである。

- [0077] 以下で説明する処理は、一つ以上のサーバ１００が実行するものとする。
- [0078] サーバ１００は、クライアント２００からクラスタの構成要求を受け付けると処理を開始する。クラスタの構成要求には、少なくともスレーブサーバの配置関係を決定するための指標情報が含まれる。
- [0079] なお、クライアント２００は、一つ以上のサーバ１００にクラスタの構成要求を送信できる。例えば、クライアント２００は、マルチキャストで全サーバ１００にクラスタの構成要求を送信する。
- [0080] 前述した指標情報は、例えば、図７に示すような入力画面７００を用いて入力される。
- [0081] 図７は、本発明の第１の実施例における入力画面７００の一例を示す説明図である。図７に示す入力画面７００は、クライアント２００に表示されるものとする。ただし、入力画面７００は、サーバ１００に表示されてもよい。
- [0082] 入力画面７００は、優先順位選択領域７１０及びＥＸＩＴ操作ボタン７２０を含む。
- [0083] 優先順位選択領域７１０は、スレーブサーバの配置関係を決定するために用いられる指標の優先順位を選択するための表示領域である。優先順位選択領域７１０は、Ｓｅｌｅｃｔ ７１１、Ｐｒｉｏｒｉｔｙ ７１２、及びＴｅｒｍ ７１３を含む。
- [0084] Ｓｅｌｅｃｔ ７１１は、使用する指標を選択するための表示部である。例えば、ユーザがＳｅｌｅｃｔ ７１１を操作することによって、対応する指標が選択される。Ｐｒｉｏｒｉｔｙ ７１２は、選択された指標の優先順位を設定する表示部である。Ｔｅｒｍ ７１３は、具体的な指標の内容を示す表示部である。
- [0085] ＥＸＩＴ操作ボタン７２０は、入力画面７００を用いた操作を終了するための操作ボタンである。ユーザがＥＸＩＴ操作ボタン７２０を操作することによって、優先順位選択領域７１０に設定された情報がサーバ１００に送信される。
- [0086] 図７に示す例では、メモリ容量が最も優先順位の高い指標として選択され

、通信速度が次に優先順位の高い指標として選択されている。

[0087] サーバ１００は、指標情報を含むクラスタの構成要求を受信すると、メモリ１０２上に当該指標情報を格納する。

[0088] 図６の説明に戻る。

[0089] サーバ１００は、サーバ１００自身の性能情報を収集する（ステップＳ１０１）。具体的には、情報共有部１８０が、サーバ１００の性能情報を収集する。

[0090] サーバ１００は、他のサーバ１００の性能情報を取得し、取得された性能情報に基づいて性能管理情報１８１を生成する（ステップＳ１０３）。

[0091] 具体的には、情報共有部１８０が、他のサーバ１００から送信された性能情報を受信し、収集された自身の性能情報及び受信した他のサーバ１００の性能情報に基づいて、性能管理情報１８１を生成する。このとき、レンジ割合６０３は空欄のままである。

[0092] また、情報共有部１８０は、収集された自身の性能情報をマルチキャストで他のサーバ１００に送信する。

[0093] サーバ１００は、計算機システムに含まれる全サーバ１００を処理対象に設定し、また、指標の優先順位の値 n を「１」に設定する（ステップＳ１０５）。このとき、サーバ１００は、指標情報を参照し、Priority 712に「１」が設定された指標を特定する。

[0094] サーバ１００は、指標情報及び性能管理情報１８１に基づいて、スレーブサーバの配置関係（サーバ１００のエントリの配列）を決定するための配列決定処理を実行する（ステップＳ１０７）。配列決定処理の詳細については、図８を用いて後述する。

[0095] サーバ１００は、配列決定処理の結果、スレーブサーバの配置関係を一意に決定できるか否かを判定する（ステップＳ１０９）。例えば、メモリ容量に基づいて、スレーブサーバの配置関係を決定する場合、メモリ容量が同一であるサーバ１００が複数存在すると、スレーブサーバの配置関係を一意に決定することができないと判定される。

- [0096] スレーブサーバの配置関係（サーバ１００のエントリの配列）を一意に決定できると判定された場合、サーバ１００は、決定されたスレーブサーバの配置関係（サーバ１００のエントリの配列）に基づいて、クラスタを構成し、処理を終了する（ステップＳ１１１）。
- [0097] 具体的には、サーバ１００は、スレーブサーバの配置関係（決定されたサーバ１００のエントリの配列）に基づいて、構成情報１７１を生成する。生成された構成情報１７１は、情報共有部１８０によって各サーバ１００に送信される。
- [0098] スレーブサーバの配置関係（サーバ１００のエントリの配列）を一意に決定できないと判定された場合、サーバ１００は、指標情報を参照し、次に優先順位が低い指標が存在するか否かを判定する（ステップＳ１１３）。例えば、サーバ１００は、指標情報を参照して、優先順位の値ｎが「２」に設定された指標の情報が存在するか否かを判定する。
- [0099] 次に優先順位が低い指標が存在しないと判定された場合、サーバ１００は、所定の基準にしたがってスレーブサーバの配置関係を決定し、ステップ１１１に進む。例えば、サーバＩＤに基づいてスレーブサーバの配置関係を決定する方法が考えられる。
- [0100] 次に優先順位が低い指標が存在すると判定された場合、サーバ１００は、スレーブサーバの配置関係（サーバ１００のエントリの配列）が一意に決定しないサーバ１００のみを処理対象として設定し、また、優先順位の値ｎに「１」をインクリメントした値を設定し（ステップＳ１１５）、ステップＳ１０７に戻る。
- [0101] これによって、スレーブサーバの配置関係（サーバ１００のエントリの配列）が一意に決定しないサーバ１００に対してのみ配列決定処理を実行することができる。
- [0102] なお、サーバ１００は、同一の性能管理情報１８１を保持するため、通常、処理結果である構成情報１７１は同一となる。ただし、構成情報１７１が異なる場合、サーバ１００は、他のサーバ１００から受信した構成情報１７

1の集計し、同一内容の構成情報171の数が一番多いものを優先的に選択する方法が考えられる。

[0103] ここで、本実施例において適用されるスレーブサーバの配置関係（サーバ100のエントリの配列）のアルゴリズムについて説明する。以下の説明では、優先順位が「1」の指標としてメモリ容量が選択されているものとする。また、メモリ容量が「3GB」のサーバA、メモリ容量が「2GB」のサーバB、メモリ容量が「3GB」のサーバC、メモリ容量が「1GB」のサーバD、メモリ容量が「4GB」のサーバE及びメモリ容量が「2GB」のサーバFを例に説明する。

[0104] 本実施例では、下り階段型アルゴリズムと山型アルゴリズムの二つがある。

[0105] 図8は、本発明の第1の実施例における下り階段型アルゴリズムの一例を示す説明図である。

[0106] 下り階段型アルゴリズムでは、性能の高い順（メモリ容量の大きい順）にサーバ100のエントリが配列される。本実施例では、片側（右側又は左側）のサーバ100との性能差が小さくなるようにエントリが配置される。

[0107] 図8に示す例では、左からメモリ容量の大きい順にサーバ100のエントリが配列される。具体的には、左から順にサーバE、サーバA、サーバC、サーバF、サーバB及びサーバDとエントリが配列される。自身のエントリの右側のサーバ100が、当該エントリの対応するサーバ100が管理するマスタデータの複製データを保持するスレーブサーバとなる。また、自身のエントリの左側のサーバが、当該エントリに対応するサーバ100が保持する複製データのマスタデータを保持するマスタサーバとなる。

[0108] これによって、右側のサーバ100とのメモリ容量の差が小さくなるため、KVS全体としてアクセス性能が非常に高くなる。

[0109] 下り階段型アルゴリズムは、第1のサーバの性能と、第1のサーバ100が管理するマスタデータの複製データを保持する第2のサーバの性能とを考慮したアルゴリズムである。

[0110] より一般的には、識別子が i 、 j であるサーバ 100 のメモリ容量を $M[i]$ 、 $M[j]$ と定義した場合、下式 (1) が最小となるサーバ 100 の組み合わせが下り階段型アルゴリズムに対応する。なお、マスターデータを保持するサーバ 100 の性能が複製データを保持するサーバ 100 の性能以上であることを条件とする。ただし、全てのサーバ 100 について前述した条件を満たさなくてもよい。

[数1]

$$\sum_{i,j=1}^n |M(i) - M(j)| \quad \text{ただし } i \neq j \quad \dots (1)$$

ここで、識別子が j であるサーバ 100 が、識別子が i であるサーバ 100 の複製データを保持するスレーブサーバとなる。

[0111] 図 9 は、本発明の第 1 の実施例における山型アルゴリズムの一例を示す説明図である。

[0112] 山型アルゴリズムでは、最も性能の高い（メモリ容量が最大の）サーバ 100 のエントリを中心に山型になるようにサーバ 100 のエントリが配列される。本実施例では、両側のサーバ 100 との性能差が小さくなるようにエントリが配置される。自身のエントリの右側のサーバが、当該エントリに対応するサーバ 100 が管理するマスターデータの複製データを保持するスレーブサーバとなる。また、自身のエントリの左側のサーバが、当該エントリに対応するサーバ 100 が保持する複製データのマスターデータを保持するマスターサーバとなる。

[0113] 図 9 に示す例では、メモリ容量の大きいサーバ 100 を中心に山型にサーバ 100 のエントリが配置される。具体的には、左から順にサーバ D、サーバ F、サーバ A、サーバ E、サーバ C 及びサーバ B とエントリが配列される。

[0114] これによって、両側のサーバ 100 とのメモリ容量の差が小さくなり、かつ、メモリ使用効率が高くなる。これは、両隣のメモリ容量の差が小さく、

複製データを効率的に格納できるためである。

[0115] 山型アルゴリズムは、第1のサーバの性能と、第1のサーバが管理するマスタデータの複製データを保持する第2サーバの性能と、第1のサーバ100が保持する複製データのマスタデータを管理する第3のサーバの性能とを考慮したアルゴリズムである。

[0116] より一般的には、識別子が*i*、*j*、*k*であるサーバ100のメモリ容量を*M* [*i*]、*M* [*j*]、*M* [*k*] と定義した場合、下式(2)が最小となるサーバ100の組み合わせが山型アルゴリズムに対応する。

[数2]

$$\sum_{i,j,k=1}^n (|M(i) - M(j)| + |M(j) - M(k)|) \quad \text{ただし } i \neq j \neq k \quad \dots (2)$$

ここで、識別子が*k*であるサーバ100が、識別子が*j*であるサーバ100の複製データを保持するスレーブサーバとなる。また、識別子が*j*であるサーバ100が、識別子が*i*であるサーバの複製データを保持するスレーブサーバとなる。

[0117] 本実施例では、サーバ100の性能比に基づいて、各サーバ100が担当する管理範囲の幅が決定される。これによって、アクセス性能を高め、かつ、メモリ使用効率を高めることができる。

[0118] 下り階段型アルゴリズムでは、右側のサーバ100との性能差(メモリ容量)が小さいため、複製データの容量差も小さくできる。そのため、メモリストア140に全ての複製データを格納できる。したがって、メモリストア140のみを用いてクライアントからのアクセス処理に対応できる。

[0119] ただし、下り階段型アルゴリズムでは、メモリ容量が最大のサーバ100が、メモリ容量が最低のサーバの複製データを格納するため、メモリ使用効率が低くなる場合がある。

[0120] 一方、山型アルゴリズムでは、両側のサーバ100との性能差(メモリ容量)が小さいため、下り階段型アルゴリズムのような問題は発生しない。しかし、下り階段型アルゴリズムと比較して、サーバ100間の性能差が大き

くなるため、複製データがディスクストア 130 に格納される場合がある。
この場合、アクセス性能は下り階段型アルゴリズムより低くなる。

[0121] 本実施例では、サーバ 100 は、ユーザによって指定された性能（指標）に基づいて、前述した二つのアルゴリズムを切り替える。すなわち、高いアクセス性能が求められる場合には、サーバ 100 は下り階段型アルゴリズムを適用し、アクセス性能及びメモリ使用効率が求められる場合には、サーバ 100 は山型アルゴリズムを適用する。

[0122] 以下の説明では、下り階段型アルゴリズムを適用したサーバ 100 のエントリの配列を下り階段配列と記載し、山型アルゴリズムを適用したサーバ 100 のエントリの配列を山型配列と記載する。

[0123] 図 10 は、本発明の第 1 の実施例における配置決定処理の詳細を説明するフローチャートである。配列決定処理は、配置管理部 160 によって実行される。

[0124] 配置管理部 160 は、性能管理情報 181 のエントリを性能の降順に並べ替える（ステップ S201）。すなわち、下り階段配列にエントリが並び替えられる。

[0125] 具体的には、配置管理部 160 は、指標情報の値 n に対応する指標を特定する。配置管理部 160 は、特定された指標に基づいて、各サーバ 100 の性能を比較して、性能の高いサーバ 100 のエントリを左から順に並べる。

[0126] なお、使用する指標の値が同一であるエントリがある場合、配置管理部 160 は、他の指標に対応する性能を比較することによってエントリを並べ替える（ステップ S113、ステップ S115）。使用する指標がない場合には、サーバ ID に基づいて、エントリが並び替えられる。

[0127] 次に、配置管理部 160 は、山型アルゴリズムを適用するか否かを判定する（ステップ S203）。具体的には、以下のような処理が実行される。

[0128] 配置管理部 160 は、指標情報を参照し、使用する指標としてメモリ容量及び通信速度が選択されているか否かを判定する。

[0129] 使用する指標としてメモリ容量及び通信速度が選択されている場合、配置

管理部 160 は、山型アルゴリズムを適用すると判定する。

[0130] なお、メモリ容量又は通信速度のいずれか一方のみが使用する指標として選択されている場合、配置管理部 160 は、山型アルゴリズムを適用しないと判定する。

[0131] 山型アルゴリズムを適用しないと判定された場合、すなわち、下り階段型アルゴリズムを適用すると判定された場合、配置管理部 160 は、ステップ S207 に進む。

[0132] ここでは、ステップ S203 において、性能管理情報 181 のエントリは下り階段配列に並び替えられているため、性能管理情報 181 のエントリの順番は変更されない。

[0133] 山型アルゴリズムを適用すると判定された場合、配置管理部 160 は、項目の性能に基づいて、性能管理情報 181 のエントリを山型配列に並び替える（ステップ S205）。山型配列の並び替え方法としては、例えば、以下のような方法が考えられる。

[0134] 配置管理部 160 は、まず、性能が最大のサーバ 100 に対応する第 1 のエントリの配置を決定する。例えば、性能管理情報 181 の中央に第 1 のエントリを配置する方法が考えられる。

[0135] 配置管理部 160 は、次に性能が高いサーバ 100 に対応する第 2 のエントリを検索し、第 1 のエントリの左側に第 2 のエントリを配置する。配置管理部 160 は、次に性能が高いサーバ 100 に対応する第 3 のエントリを検索し、第 1 のエントリの右側に第 3 のエントリを配置する。以下、配置管理部 160 は、偶数番目のエントリを第 1 のエントリの左側に順に配置し、奇数番目のエントリを第 1 のエントリの右側に順に配置する。

[0136] 配置管理部 160 は、全てのエントリに対して前述した手順を実行することによって、山型配列に性能管理情報 181 のエントリを並び替えることができる。

[0137] なお、前述の配列方法は一例であって、本発明はこれに限定されない。両側のサーバ 100 の性能差が小さくなるように配置できれば、どのような方

法であってもよい。

[0138] 配置管理部 160 は、性能管理情報 181 に基づいて、各サーバ 100 が担当する管理範囲の幅を決定する（ステップ S207）。

[0139] 例えば、メモリ容量の比に基づいて、サーバ 100 の管理範囲の幅を決定する方法が考えられる。

[0140] 配置管理部 160 は、処理結果に基づいてクラスタを構成し、処理を終了する（ステップ S209）。具体的には、以下のような処理が実行される。

[0141] 配置管理部 160 は、性能管理情報 181 のエントリの順番に基づいて、各サーバ 100 に対するスレーブサーバを設定する。本実施例では、所定のエントリの右側のエントリに対応するサーバ 100 がスレーブサーバとして設定される。

[0142] また、配置管理部 160 は、決定された管理範囲の幅に基づいて分散アルゴリズムを適用し、サーバ 100 が担当する管理範囲を決定する。

[0143] なお、配置決定処理では、処理の一貫性が保たれれば処理の順番を変更してもよい。例えば、配置管理部 160 は、ステップ S203 の処理を実行し、山型アルゴリズムを適用しない場合にはステップ S201 の処理を実行し、山型アルゴリズムを適用する場合にはステップ S205 の処理を実行してもよい。

[0144] また、配置管理部 160 は、サーバ 100 が担当する管理範囲の幅、及び決定されたスレーブサーバの配置関係に基づいて、構成情報 171 を生成する。生成された構成情報 171 は構成情報管理部 170 に出力される。

[0145] ここで、具体的な例を用いて本発明の実施例について説明する。

[0146] 以下の説明では、メモリ容量が指標として選択されたものとする。また、計算システムには、サーバ A、サーバ B、サーバ C、サーバ D 及びサーバ E が含まれるものとする。また、サーバ A のメモリ容量を「3 GB」、サーバ B のメモリ容量を「2 GB」、サーバ C のメモリ容量を「3 GB」、サーバ D のメモリ容量を「1 GB」、サーバ E のメモリ容量を「4 GB」とする。

[0147] まず、図 11A、図 11B 及び図 11C を用いて下り階段型アルゴリズム

を適用した場合の配列決定処理の具体例について説明する。なお、説明の簡単のため、スペック 602 にはメモリ容量のみを記載する。

[0148] ステップ S201 において、配置管理部 160 は、メモリ容量に基づいて、下り階段配列に性能管理情報 181 のエントリを並び替える。その結果、図 11A から図 11B に示すように性能管理情報 181 のエントリが並び替えられる。

[0149] ここで、サーバ A 及びサーバ C のメモリ容量は同一であるため、配置管理部 160 は、サーバ D のアルファベット順にサーバ A 及びサーバ C のエントリを並び替える。

[0150] ステップ S207 において、配置管理部 160 は、メモリ容量に基づいて、各サーバ 100 が担当する管理範囲の幅を決定する。

[0151] 具体的には、配置管理部 160 は、全エントリのメモリ容量の合計値を算出する。図 11B に示す例では、合計値は「13GB」と算出される。次に、配置管理部 160 は、合計値を分母に、自己のメモリ容量の値を分子に設定する。当該分数は、データ範囲における管理範囲の比率を示す値である。

[0152] ステップ S207 の処理の結果、図 11C に示すように性能管理情報 181 が更新される。

[0153] ステップ S209 では、配置管理部 160 は、サーバ 100 が担当する管理範囲を決定し、また、複製データを保持するサーバ 100 を設定することによってクラスタを構成する。具体的には、以下のような処理を実行する。

[0154] 配置管理部 160 は、全データ範囲を 13 等分し、サーバ E、サーバ A、サーバ C、サーバ B 及びサーバ D の管理範囲の幅を「4 : 3 : 3 : 2 : 1」の比率に決定する。配置管理部 160 は、決定された管理範囲の幅に基づいて分散アルゴリズムを適用することによって、各サーバの管理範囲のハッシュ値を決定する。

[0155] また、配置管理部 160 は、サーバ E の複製データをサーバ A が、サーバ A の複製データをサーバ C が、サーバ C の複製データをサーバ B が、サーバ B の複製データをサーバ D が、さらに、サーバ D の複製データをサーバ E が

保持するように設定する。

[0156] 以上の処理によって、構成情報 171 を生成することができる。

[0157] 次に、図 12A、図 12B 及び図 12C を用いて山型アルゴリズムを適用した場合の配列決定処理の具体例について説明する。

[0158] ステップ S201 において、配置管理部 160 は、メモリ容量に基づいて、下り階段配列に性能管理情報 181 のエントリを並び替える。その結果、図 12A から図 11B に示すように性能管理情報 181 のエントリが並び替えられる。

[0159] ステップ S205 において、配置管理部 160 は、山型配列に性能管理情報 181 のエントリを並び替える。その結果、図 12B に示すような性能管理情報 181 のエントリが並び替えられる。

[0160] ステップ S207 において、配置管理部 160 は、メモリ容量に基づいて、各サーバ 100 が担当する管理範囲の幅を決定する。管理範囲の幅の決定方法は、下り階段型アルゴリズムの場合と同一のものをを用いる。

[0161] ステップ S207 の処理の結果、図 12C に示すように性能管理情報 181 が更新される。

[0162] ステップ S209 では、配置管理部 160 は、サーバ 100 が担当する管理範囲を決定し、また、複製データを保持するサーバ 100 を設定することによってクラスタを構成する。具体的には、以下のような処理を実行する。

[0163] 配置管理部 160 は、全データ範囲を 13 等分し、サーバ D、サーバ A、サーバ E、サーバ C 及びサーバ B の管理範囲の幅を「1 : 3 : 4 : 3 : 2」の比率に決定する。配置管理部 160 は、決定された管理範囲の幅に基づいて分散アルゴリズムを適用することによって、各サーバの管理範囲のハッシュ値を決定する。

[0164] また、配置管理部 160 は、サーバ D の複製データをサーバ A が、サーバ A の複製データをサーバ E が、サーバ E の複製データをサーバ C が、サーバ C の複製データをサーバ B が、さらに、サーバ B の複製データをサーバ D が保持するように設定する。

- [0165] 以上の処理によって、構成情報 171 を生成することができる。
- [0166] 本実施例の配置管理部 160 は、さらに、配列決定処理が終了すると、配列決定処理の結果を表示するための表示情報を生成し、サーバ 100 又はクライアント 200 に確認画面を表示する。
- [0167] 図 13A 及び図 13B は、本発明の第 1 の実施例における確認画面 900 の一例を示す説明図である。
- [0168] 図 13A は、下り階段型アルゴリズムを適用した配列決定処理の結果を表示する確認画面 900 を示す。図 13B は、山型アルゴリズムを適用した配列決定処理の結果を表示する確認画面 900 を示す。
- [0169] 確認画面 900 は、サーバ ID 911、Position 912、メモリ容量 913、プロセッサ 914 及び通信速度 915 を表示する。
- [0170] サーバ ID 911 は、サーバ 100 を一意に識別するための識別子を表示する。Position 912 は、サーバ 100 が担当する管理範囲の幅を表す情報を表示する。本実施例では、データ範囲の先頭から、決定された管理範囲の幅だけデータが割り当てられるものとする。なお、管理範囲を示す値を表示してもよい。
- [0171] メモリ容量 913 は、サーバ ID 911 に対応するサーバ 100 が有するメモリ 102 の容量に関する情報を表示する。本実施例では、メモリ容量を指標とした下り階段型アルゴリズムを適用しているため、メモリ容量 913 には、メモリ容量が下り階段状になるようにエントリが配列されていることを示すアイコンが表示される。
- [0172] プロセッサ 914 は、サーバ ID 911 に対応するサーバ 100 が有するプロセッサ 101 の性能に関する情報を表示し、通信速度 915 は、サーバ ID 911 に対応するサーバ 100 が有するネットワークインタフェースの性能に関する情報を表示する。
- [0173] なお、確認画面 900 に表示される情報は一例であって、記憶装置 105 の容量、プロセッサ 101 の使用率及びプロセッサ 101 の負荷等を表示してもよい。

[0174] なお、本実施例では、一つのマスタデータに対して一つのスレーブサーバを選択する場合を説明したが、一つのマスタデータに対して複数のスレーブサーバを選択する場合であっても同様のアルゴリズムを適用できる。

[0175] 例えば、一つのマスタデータに対して二つのスレーブサーバを選択する場合、下り階段型アルゴリズム又は山型アルゴリズムの何れの場合であっても、一つのスレーブサーバを決定した後、当該決定されたスレーブサーバをマスタサーバとみなして同一の配置決定処理をすればよい。

[0176] (変形例)

[0177] 山型アルゴリズムでは、さらに、アクセス速度及びメモリ使用効率の優先順位に基づいて、異なる形状の山型配列に並び替える方法がある。以下、ステップS205の処理の変形例について説明する。

[0178] 図14は、本発明の第1の実施例の変形例における山型配列処理を説明するフローチャートである。

[0179] 配置管理部160は、指標情報を参照し、アクセス速度の優先順位とメモリ容量の優先順位とを比較する（ステップS401）。

[0180] 配置管理部160は、比較の結果、アクセス速度の優先順位がメモリ容量の優先順位より高いか否かを判定する（ステップS403）。すなわち、メモリ容量より通信速度を重要とするか否かが判定される。

[0181] アクセス速度の優先順位がメモリ容量の優先順位より高いと判定された場合、配置管理部160は、性能管理情報181における第1のエントリの配置位置を中心から所定範囲左の位置に決定する（ステップS405）。例えば、エントリ三つ分左側にずらして第1のエントリを配置する方法が考えられる。なお、第1のエントリは、性能が最大のサーバ100に対応する。

[0182] アクセス速度の優先順位がメモリ容量の優先順位より低いと判定された場合、配置管理部160は、性能管理情報181における第1のエントリの配置位置を中心から所定範囲右の位置に決定し（ステップS409）、ステップS407に進む。例えば、エントリ三つ分右側にずらして第1のエントリを配置する方法が考えられる。

[0183] 配置管理部 160 は、第 2 のエントリを含むその他のエントリの配置を決定し（ステップ S407）、処理を終了する。第 2 のエントリを含むその他のエントリの配置の決定方法は、第 1 の実施例と同一である。

第 2 の実施例

[0184] 第 2 の実施例では、クラスタの運用中に、サーバ 100 が追加又は削除された場合の処理について説明する。以下、第 1 の実施例との差異を中心に説明する。

[0185] 計算機システムの構成、サーバ 100 及びクライアント 200 の構成は第 1 の実施例と同一であるため説明を省略する。

[0186] 図 15 は、本発明の第 2 の実施例におけるクラスタ構成変更処理を説明するフローチャートである。

[0187] サーバ 100 は、計算機システムの構成の変更を検出したか否かを判定する（ステップ S501）。例えば、クライアント 200 からサーバ 100 の追加又は削除指示を受け付けた場合、サーバ 100 は、計算機システムの構成変更を検出したと判定する。

[0188] 計算機システムの構成の変更を検出していないと判定された場合、サーバ 100 は、ステップ S501 に戻り、計算機システムの構成が変更されるまで待ち続ける。

[0189] 計算機システムの構成の変更を検出したと判定された場合、サーバ 100 は、新規サーバ 100 が追加されたか否かを判定する（ステップ S503）。

[0190] 新規サーバ 100 が追加されたと判定された場合、サーバ 100 は、新規サーバ 100 の性能情報を取得し、新規サーバ 100 に対する配置決定処理を実行する（ステップ S505、ステップ S507）。

[0191] このとき、サーバ 100 は、取得された新規サーバ 100 の性能情報に基づいて、配置決定処理を実行することによって性能管理情報 181 を更新する。具体的には、性能管理情報 181 に、新規サーバ 100 に対応するエントリを追加する。具体的な、配置決定処理の詳細は、図 16A 及び図 16B

を用いて後述する。

[0192] サーバ100は、クラスタの構成を更新し、処理を終了する（ステップS509）。具体的には、サーバ100は、更新された性能管理情報181に基づいて、構成情報171を更新する。

[0193] ステップS503において、新規サーバ100が追加されていない、すなわち、サーバ100の削除であると判定された場合、サーバ100は、削除対象のサーバ100の停止を確認する（ステップS511）。具体的には、情報共有部180が、削除対象のサーバ100の停止を検出する。

[0194] サーバ100は、性能管理情報181から削除対象のサーバ100に対応するエントリを削除する（ステップS513）。

[0195] サーバ100は、クラスタの構成を更新し、処理を終了する（ステップS509）。

[0196] 図16A及び図16Bは、本発明の第2の実施例における新規サーバ100の配置決定処理を説明するフローチャートである。

[0197] 以下の説明では、性能管理情報181に左から「1」から「n」までの識別番号*i*が付与されているものとする。また、*i*番目のエントリのメモリ容量をM[*i*]と記載し、また、新規サーバ100のメモリ容量をaと記載する。

[0198] 配置管理部160は、性能管理情報181を参照し、山型アルゴリズムが適用された配列であるか否かを判定する（ステップS601）。

[0199] 山型アルゴリズムが適用された配列でない、すなわち、下り階段型アルゴリズムが適用された配列であると判定された場合、配置管理部160は、識別番号*i*を「1」に設定する（ステップS603）。

[0200] 次に、配置管理部160は、新規サーバ100のメモリ容量が、*i*番目のエントリに対応するサーバ100のメモリ容量以上であるか否かを判定する（ステップS605）。

[0201] ステップS605の条件を満たさない場合、配置管理部160は、識別番号*i*を「1」加算してステップS605に戻り、同様の処理を実行する（ス

テップS 6 0 9)。

[0202] ステップS 6 0 5の条件を満たすと判定された場合、配置管理部1 6 0は、新規サーバ1 0 0のエントリの配列を決定し、性能管理情報を更新する(ステップS 6 1 1)。

[0203] 具体的には、配置管理部1 6 0は、メモリ容量M [i] であるサーバ1 0 0の左側に新規サーバ1 0 0のエントリを追加することによって、性能管理情報1 8 1を更新する。ここで、性能管理情報1 8 1の更新方法の一例について説明する。

[0204] 図1 7は、本発明の第2の実施例における性能管理情報1 8 1の更新方法の一例を示す説明図である。

[0205] 図1 7の(a)は、更新前の性能管理情報1 8 1を示す図である。図1 7の(b)は、更新後の性能管理情報1 8 1である。

[0206] ここで、新規サーバ1 0 0のサーバIDを「サーバF」とし、また、メモリ容量を「2 GB」とする。このとき、配置管理部1 6 0は、図1 7の(a)に示すような形式の性能情報1 0 0 0に基づいて、性能管理情報1 8 1の各エントリについて比較処理を実行する。

[0207] ステップ6 0 5及びステップS 6 0 7において、新規サーバFとサーバEとの比較処理の結果、ステップS 6 0 5の条件を満たさないと判定される。サーバA及びサーバCについても同様の結果となる。サーバBとの比較処理の結果、ステップS 6 0 5の条件を満たすため、配置管理部1 6 0は、新規サーバFのエントリの配置をサーバBの左側に決定する。その結果、図1 7の(b)に示すような性能管理情報1 8 1に更新される。

[0208] なお、図1 6 Aに示す、下り階段型アルゴリズムにおける新規サーバ1 0 0の追加方法は一例であって、例えば、メモリ容量の小さいサーバ1 0 0から順に比較する方法などでもよい。

[0209] 図1 6の説明に戻る。

[0210] ステップS 6 0 1において、山型アルゴリズムが適用された配列であると判定された場合、配置管理部1 6 0は、新規サーバ1 0 0のメモリ容量が、

計算機システムにおけるメモリ容量の最大値以上であるか否かを判定する（ステップS 6 1 1）。

[0211] ステップS 6 1 1の条件を満たさないと判定された場合、配置管理部160は、新規サーバ100のメモリ容量が、計算機システムにおけるメモリ容量の最小値以下であるか否かを判定する（ステップS 6 1 3）。

[0212] ステップS 6 1 3の条件を満たさないと判定された場合、配置管理部160は、山型配列の頂上を起点に、メモリ容量が新規サーバ100のメモリ以下であるサーバ100を検索する（ステップS 6 1 5）。検索されたサーバ100を処理対象サーバ100と記載する。

[0213] なお、山型アルゴリズムでは、ステップS 6 1 5の条件を満たすエントリが山型配列の頂点の右側及び左側に一つずつ存在する。そのため、右側の処理対象サーバ100を第1の処理対象サーバ100と記載し、左側の処理対象サーバ100を第2の処理対象サーバ100と記載する。

[0214] 配置管理部160は、第1の処理対象サーバ100と新規サーバ100とのメモリ差（第1のメモリ差）、及び、新規サーバ100と第1の処理対象サーバ100の一つ左隣のサーバ100とのメモリ差（第2のメモリ差）を算出する（ステップS 6 1 7）。ここでは、各サーバ100のメモリ差の絶対値を算出するものとする。

[0215] また、配置管理部160は、第2の処理対象サーバ100と新規サーバ100とのメモリ差（第3のメモリ差）、及び、新規サーバ100と第2の処理対象サーバ100の一つ右隣のサーバ100とのメモリ差（第4のメモリ差）を算出する（ステップS 6 1 9）。ここでは、各サーバ100のメモリ差の絶対値を算出するものとする。

[0216] 配置管理部160は、第1のメモリ差と第2のメモリ差との合計値が、第3のメモリ差と第4のメモリ差との合計値以上であるか否かを判定する（ステップS 6 2 1）。

[0217] ステップS 6 2 1の条件を満たすと判定された場合、配置管理部160は、第1の処理対象サーバ100の左に新規サーバ100を配置し、処理を終

了する（ステップS 6 2 3）。

[0218] ステップS 6 2 1の条件を満たさないと判定された場合、配置管理部1 6 0は、第2の処理対象サーバ1 0 0の右に新規サーバ1 0 0を配置し、処理を終了する（ステップS 6 2 5）。

[0219] ステップS 6 1 1において、ステップS 6 1 1の条件を満たすと判定された場合、配置管理部1 6 0は、ステップS 6 1 7に進む。

[0220] 計算機システム内においてメモリ容量が最大であるサーバ1 0 0が複数ある場合、識別番号iが一番小さいエントリに対応するサーバを第2の処理対象サーバ1 0 0とし、識別番号iが一番大きいエントリに対応するサーバを第1の処理対象サーバ1 0 0とする。

[0221] なお、計算機システム内においてメモリ容量が最大であるサーバ1 0 0が一つだけである場合、第1の処理対象サーバ1 0 0と第2の処理対象サーバ1 0 0とは同一のものになる。

[0222] ステップS 6 1 7以下の処理は同一であるため説明を省略する。

[0223] ステップS 6 1 3において、ステップS 6 1 3の条件を満たすと判定された場合、配置管理部1 6 0は、ステップS 6 1 7に進む。

[0224] 計算機システム内においてメモリ容量が最小であるサーバ1 0 0が複数ある場合、識別番号iが一番小さいものを第2の処理対象サーバ1 0 0とし、識別番号iが一番大きいものを第1の処理対象サーバ1 0 0とする。

[0225] ステップS 6 1 7以下の処理は同一であるため説明を省略する。

[0226] 図1 8は、本発明の第2の実施例における性能管理情報1 8 1の更新方法の一例を示す説明図である。

[0227] 図1 8の（a）は、更新前の性能管理情報1 8 1を示す図である。図1 8の（b）は、更新後の性能管理情報1 8 1である。

[0228] ここで、新規サーバ1 0 0のサーバIDを「サーバF」とし、また、メモリ容量を「2 GB」とする。このとき、配置管理部1 6 0は、図1 8の（a）に示すような形式の性能情報1 1 0 0に基づいて、ステップS 6 1 1～ステップS 6 2 5の処理を実行する。

- [0229] ステップS 6 1 1 及びステップS 6 1 3 の条件を満たさないため、配置管理部 1 6 0 は、ステップS 6 1 5 に進む。
- [0230] ステップS 6 1 5 において、新規サーバFのメモリ以下のサーバ1 0 0 として、サーバB及びサーバDが検索される。ここでは、サーバDが第2の処理対象サーバ1 0 0 となり、サーバBが第1の処理対象サーバ1 0 0 となる。
- [0231] ステップS 6 2 1 では、第1のメモリ差と第2のメモリ差との合計値が「1 G B」となり、第3のメモリ差と第4のメモリ差との合計値が「2 G B」となるため、ステップS 6 2 1 の条件を満たさないと判定される。
- [0232] したがって、配置管理部 1 6 0 は、サーバFの配置をサーバDの右側に決定する。その結果図 1 8 の（b）に示すような性能管理情報 1 8 1 に更新される。
- [0233] 本発明の一形態によれば、サーバ1 0 0 間の性能差を考慮して、分散K V Sを最適に構築することができる。具体的には、マスタサーバとスレーブサーバとの性能差が小さくなるようにスレーブサーバを設定でき、また、マスタサーバが担当する管理範囲の幅（複製データのデータ量）も最適に設定することができる。また、サーバ1 0 0 が追加又は削除された場合も、動的に、スレーブサーバの配置関係及び管理範囲の幅を設定することができる。
- [0234] なお、本実施例で例示した種々のソフトウェアは、電磁的、電子的及び光学式等の種々の記録媒体に格納可能であり、インターネット等の通信網を通じて、コンピュータにダウンロード可能である。
- [0235] さらに、本実施例では、ソフトウェアによる制御を用いた例について説明したが、その一部をハードウェアによって実現することも可能である。
- [0236] 以上、本発明を添付の図面を参照して詳細に説明したが、本発明はこのような具体的構成に限定されるものではなく、添付した請求の範囲の趣旨内における様々な変更及び同等の構成を含むものである。

請求の範囲

[請求項1]

ネットワークを介して複数の計算機が接続され、前記複数の計算機が有する記憶領域から構成されたデータベースを用いて業務を実行する計算機システムであって、

前記複数の計算機の各々は、

プロセッサと、前記プロセッサに接続される記憶装置と、前記ネットワークを介して他の前記計算機と通信するためのネットワークインタフェースとを有し、

前記複数の計算機の各々の性能を管理する性能管理情報を保持し、

前記データベースには、キー及びデータ値から構成されるデータが格納され、

前記複数の計算機の各々には、

前記複数の計算機の各々に割り当てる前記キーの範囲を示す管理範囲を決定するための分散アルゴリズムに基づいて、前記複数の計算機の各々がマスタとして管理するマスタデータが配置され、

他の前記複数の計算機のうちの一つの計算機が管理する前記マスタデータの複製データが配置され、

前記計算機システムは、

前記複数の計算機の各々から性能に関する情報を取得し、取得された情報に基づいて性能管理情報を更新する性能情報管理部と、

前記性能管理情報に基づいて、前記複数の計算機の各々が管理する前記マスタデータの前記管理範囲、及び、前記複数の計算機の各々が管理するマスタデータの複製データを保持する副計算機を決定するクラスタ構成管理部と、を備えることを特徴とする計算機システム。

[請求項2]

請求項1に記載の計算機システムであって、

前記性能管理情報は、前記複数の計算機の各々が有する前記記憶装置の記憶容量に関する情報を含み、

前記クラスタ構成管理部は、

前記性能管理情報を参照して、前記複数の計算機の各々が有する前記記憶装置の記憶容量を比較し、

前記複数の計算機のうちの一つの計算機が有する前記記憶装置の記憶容量と、当該計算機に対する前記副計算機が有する前記記憶装置の記憶容量との差が小さくなるように、前記複数の計算機の各々の管理範囲及び前記副計算機を決定することを特徴とする計算機システム。

[請求項3]

請求項2に記載の計算機システムであって、

前記複数の計算機の各々が有する前記記憶装置は、メモリ及びディスク装置を含み、

前記データベースは、前記複数の計算機の各々が有する前記メモリの記憶領域から構成されるメモリストアと、前記複数の計算機の各々が有する前記ディスク装置から構成されるディスクストアとを含み、

前記クラスタ構成管理部は、

前記メモリの記憶容量の差が小さくなる計算機と副計算機との組み合わせを算出し、

前記算出された計算機と副計算機との組み合わせに基づいて、前記複数の計算機の各々に対する前記副計算機を設定し、

前記算出された計算機と副計算機との組み合わせに基づいて、前記複数の計算機の各々の前記管理範囲の割当幅を決定し、

前記決定された管理範囲の割当幅に基づいて、前記分散アルゴリズムを適用することによって、前記複数の計算機の各々の前記管理範囲を決定することを特徴とする計算機システム。

[請求項4]

請求項3に記載の計算機システムであって、

前記複数の計算機の各々は、前記データベースの処理性能を指定する指標情報を保持し、

前記クラスタ構成管理部は、

前記指標情報を参照して、前記データベースのアクセス性能が要求されているか否かを判定し、

前記データベースのアクセス性能が要求されていると判定された場合、前記複数の計算機の各々が有する前記メモリの記憶容量が、前記複数の計算機の各々に対する前記副計算機が有する前記メモリの記憶容量以上であり、かつ、前記メモリの記憶容量の差が小さくなる前記計算機と副計算機との組合わせを算出し、

前記複数の計算機の各々が有するメモリの記憶容量の大きさに基づいて、前記複数の計算機の各々の管理範囲の割当幅を決定することを特徴とする計算機システム。

[請求項5]

請求項3に記載の計算機システムであって、

前記複数の計算機の各々は、前記データベースの処理性能を指定する指標情報を保持し、

前記クラスタ構成管理部は、

前記指標情報を参照して、前記データベースのアクセス性能と前記データベースの記憶容量の使用効率とが要求されているか否かを判定し、

前記データベースのアクセス性能と前記データベースの記憶容量の使用効率とが要求されていると判定された場合、前記複数の計算機の各々が保持する前記複製データに対応する前記マスタデータを管理する計算機が有するメモリの記憶容量と、前記複数の計算機の各々が有するメモリの記憶容量と、前記複数の計算機の各々に対する前記副計算機が有するメモリの記憶容量との差が小さくなる前記計算機と副計算機との組合わせを算出し、

前記複数の計算機の各々が有するメモリの記憶容量の大きさに基づいて、前記複数の計算機の各々の管理範囲の割当幅を決定することを特徴とする計算機システム。

[請求項6]

請求項4又は請求項5に記載の計算機システムであって、

前記クラスタ構成管理部は、前記指標情報を入力するための入力画面を表示し、

前記入力画面を介した操作に基づいて、前記指標情報を取得することを特徴とする計算機システム。

[請求項7] 請求項1から請求項6のいずれか一つに記載の計算機システムであって、

前記クラスタ構成管理部は、前記複数の計算機と前記複数の計算機の各々に対する副計算機との性能差、及び、前記複数の計算機の各々の管理範囲の割当幅を表示するための表示情報を生成することを特徴とする計算機システム。

[請求項8] ネットワークを介して複数の計算機が接続され、前記複数の計算機が有する記憶領域から構成されたデータベースを用いて業務を実行する計算機システムにおけるデータ配置管理方法であって、

前記複数の計算機の各々は、

プロセッサと、前記プロセッサに接続される記憶装置と、前記ネットワークを介して他の前記計算機と通信するためのネットワークインタフェースとを有し、

前記複数の計算機の各々の性能を管理する性能管理情報を保持し、

前記データベースには、キー及びデータ値から構成されるデータが格納され、

前記複数の計算機の各々には、

前記複数の計算機の各々に割り当てる前記キーの範囲を示す管理範囲を決定するための分散アルゴリズムに基づいて、前記複数の計算機の各々がマスタとして管理するマスタデータが配置され、

他の前記複数の計算機のうちの一つの計算機が管理する前記マスタデータの複製データが配置され、

前記方法は、

前記複数の計算機の各々が、前記複数の計算機の各々から性能に関する情報を取得し、取得された情報に基づいて性能管理情報を更新する第1のステップと、

前記複数の計算機の各々が、前記性能管理情報に基づいて、前記複数の計算機各々が管理する前記マスタデータの前記管理範囲、及び、前記複数の計算機の各々が管理するマスタデータの複製データを保持する副計算機を決定する第2のステップと、を含むことを特徴とするデータ配置管理方法。

[請求項9]

請求項8に記載のデータ配置管理方法であって、

前記性能管理情報は、前記複数の計算機の各々が有する前記記憶装置の記憶容量に関する情報を含み、

前記第2のステップは、

前記性能管理情報を参照して、前記複数の計算機の各々が有する前記記憶装置の記憶容量を比較する第3のステップと、

前記複数の計算機のうちの一つの計算機が有する前記記憶装置の記憶容量と、当該計算機に対する前記副計算機が有する前記記憶装置の記憶容量との差が小さくなるように、前記複数の計算機の各々の管理範囲及び前記副計算機を決定する第4のステップと、を含むことを特徴とするデータ配置管理方法。

[請求項10]

請求項9に記載のデータ配置管理方法であって、

前記複数の計算機の各々が有する前記記憶装置は、メモリ及びディスク装置を含み、

前記データベースは、前記複数の計算機の各々が有する前記メモリの記憶領域から構成されるメモリストアと、前記複数の計算機の各々が有する前記ディスク装置から構成されるディスクストアとを含み、

前記第4のステップは、

前記メモリの記憶容量と差が小さくなる計算機と副計算機との組み合わせを算出する第5のステップと、

前記算出された計算機と副計算機との組み合わせに基づいて、前記複数の計算機の各々に対する前記副計算機を設定する第6のステップと、

、

前記算出された計算機と副計算機との組合わせに基づいて、前記複数の計算機の各々の管理範囲の割当幅を決定する第7のステップと、

前記決定された管理範囲の割当幅に基づいて、前記分散アルゴリズムを適用することによって前記複数の計算機の各々の管理範囲を決定する第8のステップと、を含むことを特徴とするデータ配置管理方法。

[請求項11]

請求項10に記載のデータ配置管理方法であって、

前記複数の計算機の各々は、前記データベースの処理性能を指定する指標情報を保持し、

前記第5のステップは、

前記指標情報を参照して、前記データベースのアクセス性能が要求されているか否かを判定するステップと、

前記データベースのアクセス性能が要求されていると判定された場合、前記複数の計算機の各々が有する前記メモリの記憶容量が、前記複数の計算機の各々に対する前記副計算機が有する前記メモリの記憶容量より大きく、かつ、前記メモリの記憶容量の差が小さくなる前記計算機と副計算機との組合わせを算出するステップと、を含み、

前記第7のステップでは、前記複数の計算機の各々が有するメモリの記憶容量の大きさに基づいて、前記複数の計算機の各々の管理範囲の割当幅を決定することを特徴とするデータ配置管理方法。

[請求項12]

請求項10に記載のデータ配置管理方法であって、

前記複数の計算機各々は、前記データベースの処理性能を指定する指標情報を保持し、

前記第5のステップは、

前記指標情報を参照して、前記データベースのアクセス性能と前記データベースの記憶容量の使用効率とが要求されているか否かを判定するステップと、

前記データベースのアクセス性能と前記データベースの記憶容量の

使用効率とが要求されていると判定された場合、前記複数の計算機の各々が保持する前記複製データに対応する前記マスタデータを管理する計算機が有するメモリの記憶容量と、前記複数の計算機の各々が有するメモリの記憶容量と、前記複数の計算機の各々に対する前記副計算機が有するメモリの記憶容量との差が小さくなる前記計算機と副計算機との組み合わせを算出するステップと、を含み、

前記第 7 のステップでは、前記複数の計算機の各々が有するメモリの記憶容量の大きさに基づいて、前記複数の計算機の各々の管理範囲の割当幅を決定することを特徴とするデータ配置管理方法。

[請求項13]

請求項 1 1 又は請求項 1 2 に記載のデータ配置管理方法であって、前記方法は、前記指標情報を入力するための入力画面を表示するステップを含み、

前記入力画面を介した操作に基づいて、前記指標情報を取得することを特徴とするデータ配置管理方法。

[請求項14]

請求項 8 から請求項 1 3 のいずれか一つに記載のデータ配置管理方法であって、

前記方法は、前記複数の計算機と前記複数の計算機の各々に対する前記副計算機との性能差、及び、前記複数の計算機の各々の管理範囲の割当幅を表示するための表示情報を生成するステップを含むことを特徴とするデータ配置管理方法。

[請求項15]

ネットワークを介して複数の計算機が接続され、前記複数の計算機が有する記憶領域から構成されたデータベースを用いて業務を実行する計算機システムに含まれる前記複数の計算機の各々が実行するプログラムであって、

前記複数の計算機の各々は、

プロセッサと、前記プロセッサに接続される記憶装置と、前記ネットワークを介して他の前記計算機と通信するためのネットワークインタフェースとを有し、

前記複数の計算機の各々の性能を管理する性能管理情報を保持し、
前記データベースには、キー及びデータ値から構成されるデータが
格納され、

前記複数の計算機の各々には、

前記複数の計算機の各々に割り当てる前記キーの範囲を示す管理範
囲を決定するための分散アルゴリズムに基づいて、前記複数の計算機
の各々がマスタとして管理するマスタデータが配置され、

他の前記複数の計算機のうちに一つの計算機が管理する前記マスタ
データの複製データが配置され、

前記プログラムは、

前記複数の計算機の各々から性能に関する情報を取得し、取得され
た情報に基づいて性能管理情報を更新する第1の手順と、

前記性能管理情報に基づいて、複数の計算機の各々が管理する前記
マスタデータの前記管理範囲、及び、前記複数の計算機の各々が管理
するマスタデータの複製データを保持する副計算機を決定する第2の
手順と、を前記複数の計算機の各々に実行させることを特徴とするプ
ログラム。

[請求項16]

請求項15に記載のプログラムであって、

前記性能管理情報は、前記複数の計算機の各々が有する前記記憶装
置の記憶容量に関する情報を含み、

前記第2の手順は、

前記性能管理情報を参照して、前記複数の計算機の各々が有する前
記記憶装置の記憶容量を比較する第3の手順と、

前記複数の計算機のうちの一つの計算機が有する前記記憶装置の記
憶容量と、当該計算機に対する前記副計算機が有する前記記憶装置の
記憶容量との差が小さくなるように、前記複数の計算機の各々の管理
範囲及び前記副計算機を決定する第4の手順と、を含むことを特徴と
するプログラム。

[請求項17]

請求項 16 に記載のプログラムであって、

前記複数の計算機の各々が有する前記記憶装置は、メモリ及びディスク装置を含み、

前記データベースは、前記複数の計算機の各々が有する前記メモリの記憶領域から構成されるメモリストアと、前記複数の計算機の各々が有する前記ディスク装置から構成されるディスクストアとを含み、

前記第 4 の手順は、

前記メモリの記憶容量の差が小さくなる計算機と副計算機との組合わせを算出する第 5 の手順と、

前記算出された計算機と副計算機との組合わせに基づいて、前記複数の計算機の各々に対する前記副計算機を設定する第 6 の手順と、

前記算出された計算機と副計算機との組合わせに基づいて、前記複数の計算機の各々の管理範囲の割当幅を決定する第 7 の手順と、

前記決定された管理範囲の割当幅に基づいて、前記分散アルゴリズムを適用することによって前記複数の計算機の各々の管理範囲を決定する第 8 の手順と、を含むことを特徴とするプログラム。

[請求項18]

請求項 17 に記載のプログラムであって、

前記複数の計算機の各々は、前記データベースの処理性能を指定する指標情報を保持し、

前記第 5 の手順は、

前記指標情報を参照して、前記データベースのアクセス性能が要求されているか否かを判定する手順と、

前記データベースのアクセス性能が要求されていると判定された場合、前記複数の計算機の各々が有する前記メモリの記憶容量が、前記複数の計算機の各々に対する前記副計算機が有する前記メモリの記憶容量より大きく、かつ、前記メモリの記憶容量の差が小さくなる前記計算機と副計算機との組合わせを算出する手順と、を含み、

前記第 7 の手順では、前記複数の計算機の各々が有するメモリの記

憶容量の大きさに基づいて、前記複数の計算機の各々の管理範囲の割当幅を決定させることを特徴とするプログラム。

[請求項19]

請求項17に記載のプログラムであって、

前記複数の計算機の各々は、前記データベースの処理性能を指定する指標情報を保持し、

前記第5の手順は、

前記指標情報を参照して、前記データベースのアクセス性能と前記データベースの記憶容量の使用効率とが要求されているか否かを判定する手順と、

前記データベースのアクセス性能と前記データベースの記憶容量の使用効率とが要求されていると判定された場合、前記複数の計算機の各々が保持する前記複製データに対応する前記マスタデータを管理する計算機が有するメモリの記憶容量と、前記複数の計算機の各々が有するメモリの記憶容量と、前記複数の計算機の各々に対する前記副計算機が有するメモリの記憶容量との差が小さくなる前記計算機と副計算機との組み合わせを算出する手順と、を含み、

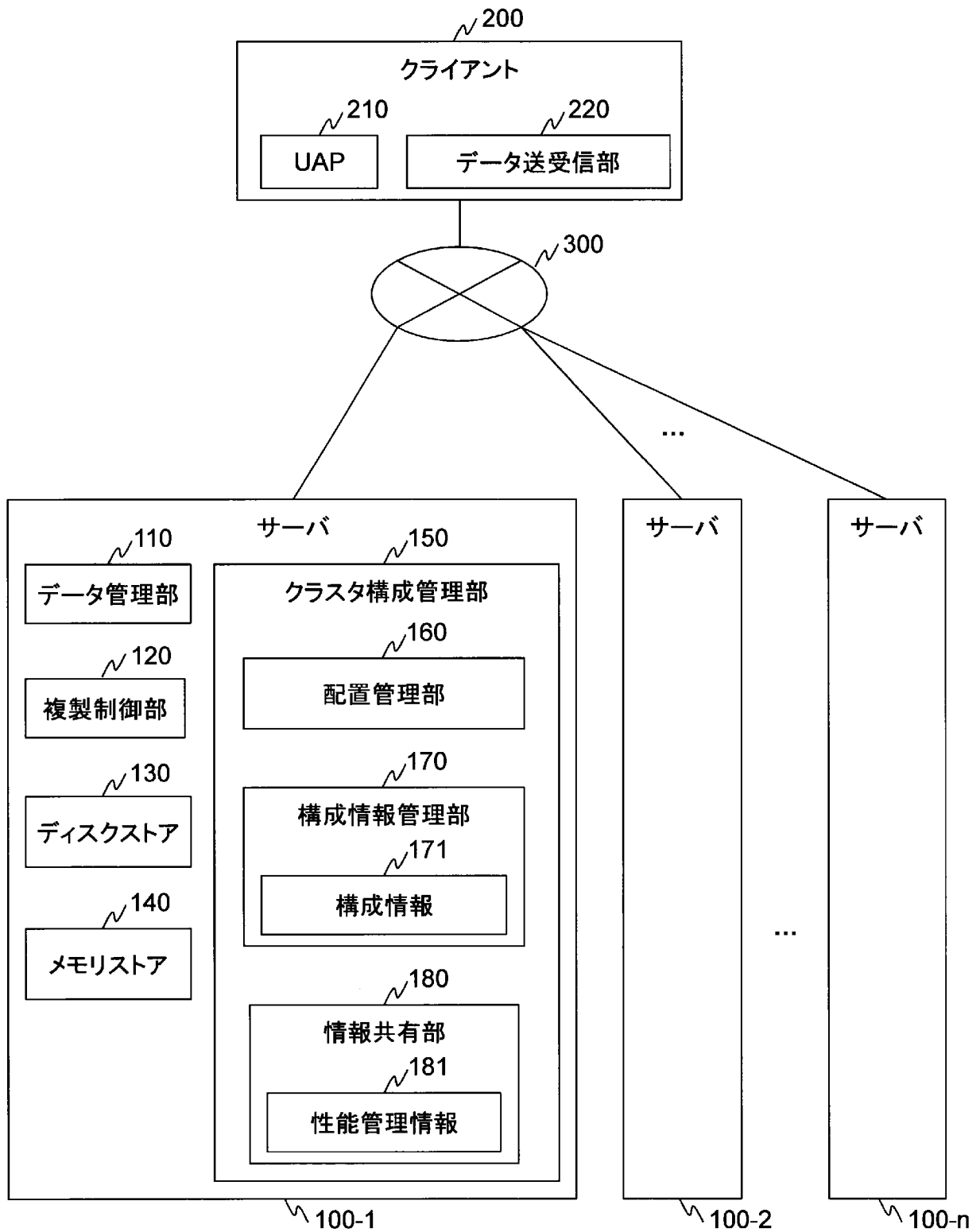
前記第7の手順では、前記複数の計算機の各々が有するメモリの記憶容量の大きさに基づいて、前記複数の計算機の各々の管理範囲の割当幅を決定させることを特徴とするプログラム。

[請求項20]

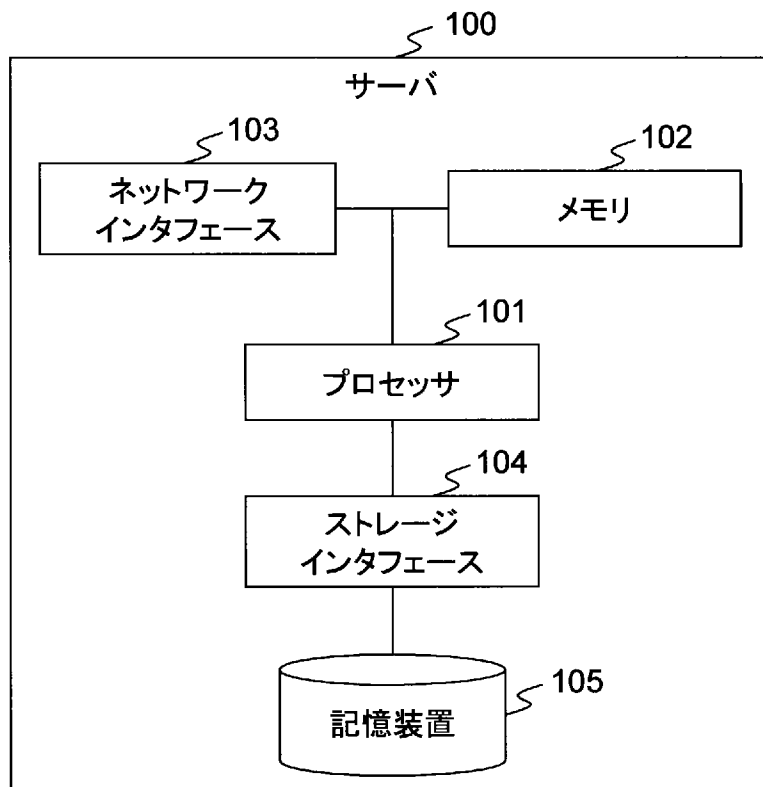
請求項16から請求項19のいずれか一つに記載のプログラムであって、

前記プログラムは、前記複数の計算機と前記複数の計算機の各々に対する前記副計算機との性能差、及び、前記複数の計算機の各々の管理範囲の割当幅を表示するための表示情報を生成する手順を前記複数の計算機の各々に実行させることを特徴とするプログラム。

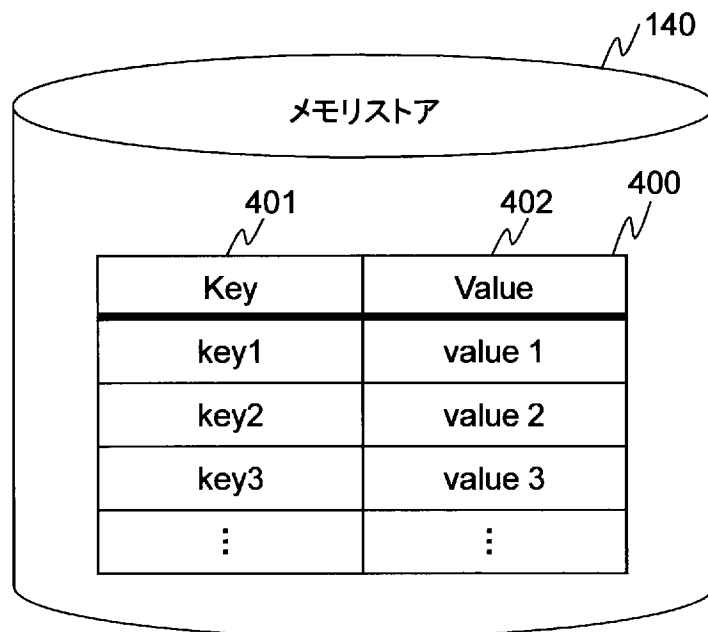
[図1]



[図2]



[図3]



[図4]

構成情報

サーバ ID	管理範囲	
	マスタ	スレーブ
サーバ E	1 ~ 400	1201 ~ 1300
サーバ A	401 ~ 700	1 ~ 400
サーバ C	701 ~ 1000	401 ~ 700
サーバ B	1001 ~ 1200	701 ~ 1000
サーバ D	1201 ~ 1300	1001 ~ 1200

[図5]

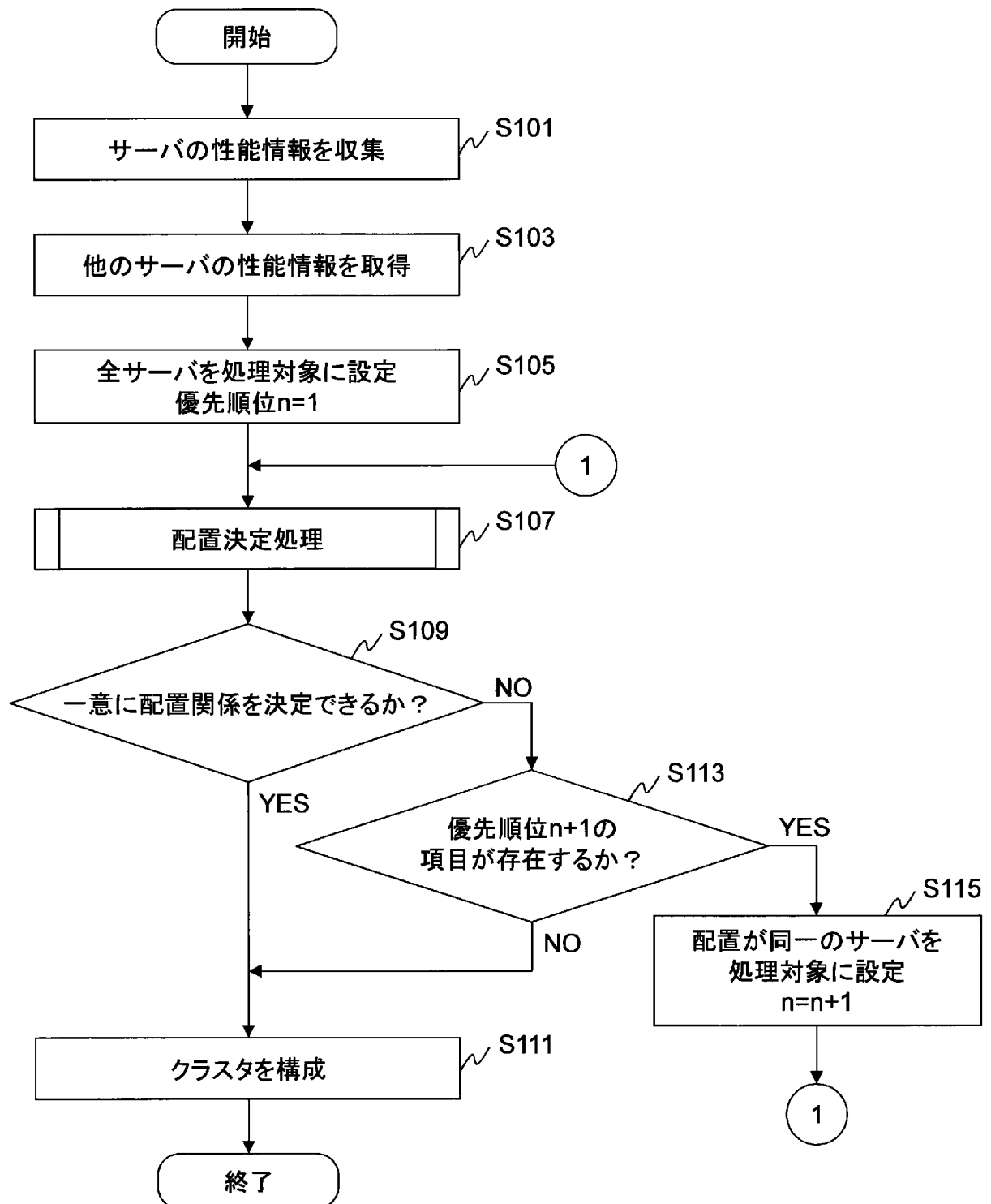
性能管理情報

601	サーバID		サーバA	サーバB	サーバC	サーバD	サーバE
	602	プロセッサ性能	3GHz	2GHz	2GHz	2GHz	2GHz
		メモリ性能	3GB	2GB	3GB	1GB	4GB
		通信性能	100Mbps	10Mbps	100Mbps	10Mbps	100Mbps
603	レンジ割当		-	-	-	-	-

181

[図6]

クラスタ構成処理



[図7]

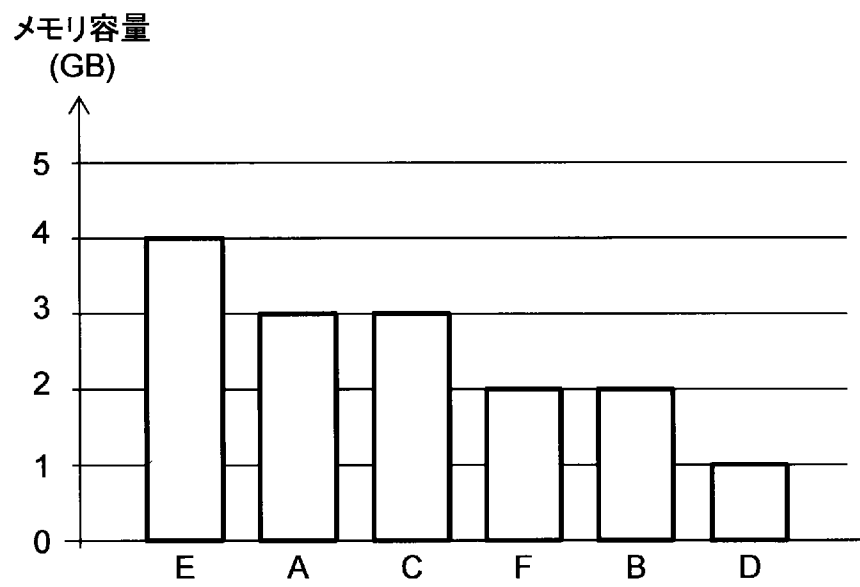
700

— ✕

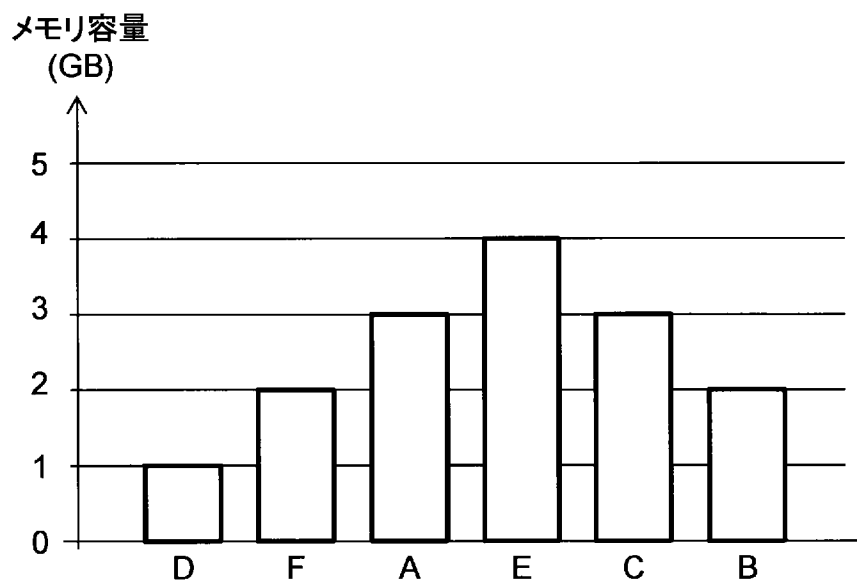
711 Select	712 Priority	713 Term
☒	1 ▲ ▼	メモリ容量
☐	▲ ▼	CPU性能
☒	2 ▲ ▼	通信速度
☐	▲ ▼	ディスク容量
⋮	⋮	⋮

720
EXIT

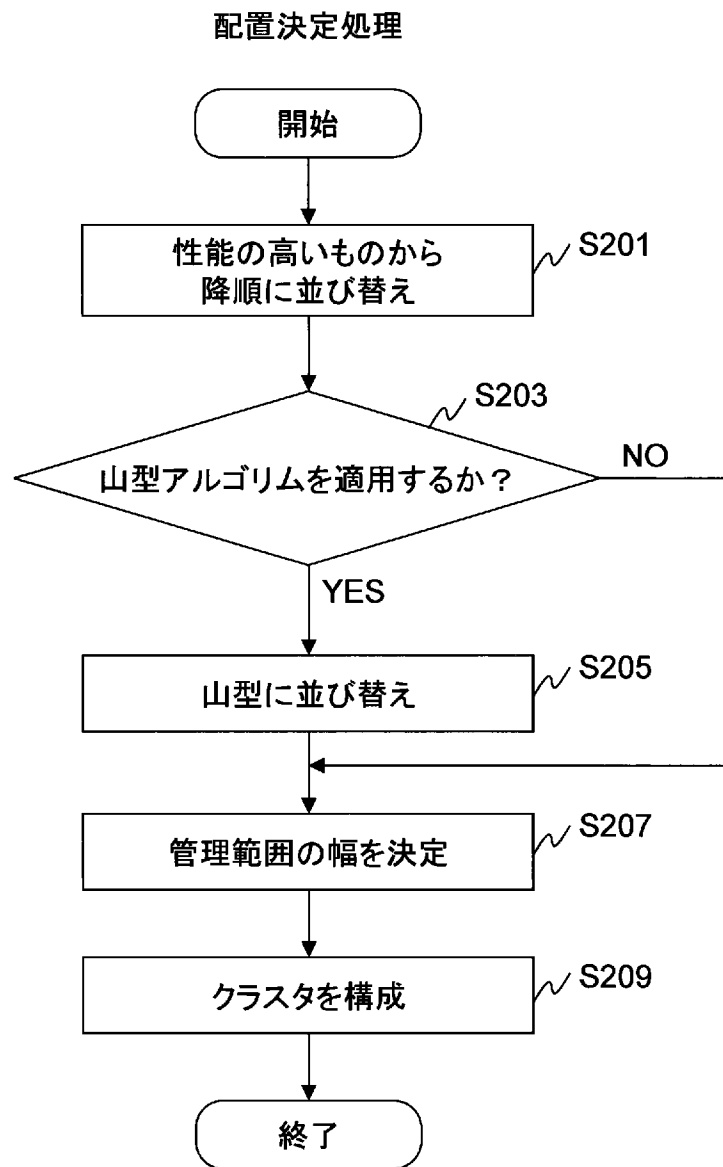
[図8]



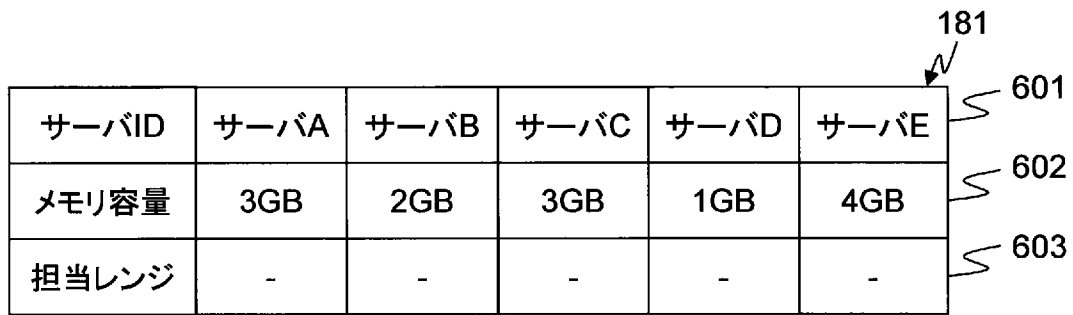
[図9]



[図10]

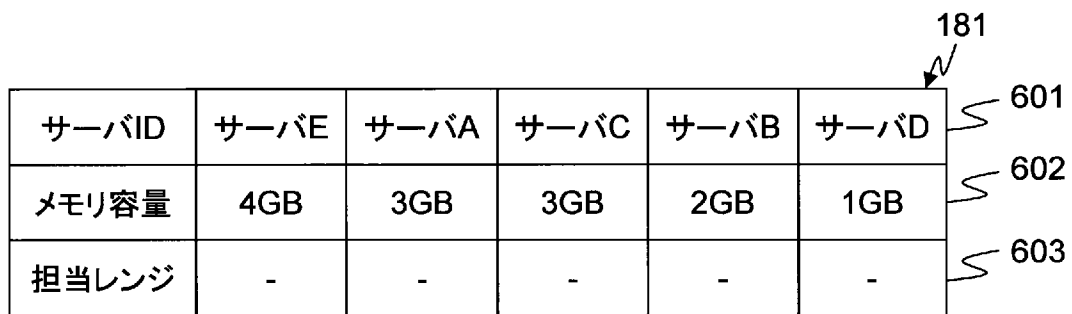


[図11A]



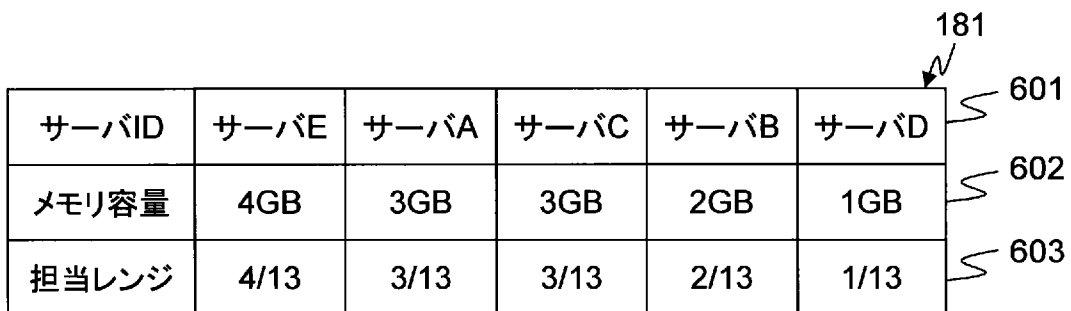
サーバID	サーバA	サーバB	サーバC	サーバD	サーバE
メモリ容量	3GB	2GB	3GB	1GB	4GB
担当レンジ	-	-	-	-	-

[図11B]



サーバID	サーバE	サーバA	サーバC	サーバB	サーバD
メモリ容量	4GB	3GB	3GB	2GB	1GB
担当レンジ	-	-	-	-	-

[図11C]



サーバID	サーバE	サーバA	サーバC	サーバB	サーバD
メモリ容量	4GB	3GB	3GB	2GB	1GB
担当レンジ	4/13	3/13	3/13	2/13	1/13

[図12A]

サーバID	サーバA	サーバB	サーバC	サーバD	サーバE	181
メモリ容量	3GB	2GB	3GB	1GB	4GB	601
担当レンジ	-	-	-	-	-	602
						603

[図12B]

サーバID	サーバD	サーバA	サーバE	サーバC	サーバB	181
メモリ容量	1GB	3GB	4GB	3GB	2GB	601
担当レンジ	-	-	-	-	-	602
						603

[図12C]

サーバID	サーバD	サーバA	サーバE	サーバC	サーバB	181
メモリ容量	1GB	3GB	4GB	3GB	2GB	601
担当レンジ	1/13	3/13	4/13	3/13	2/13	602
						603

[図13A]

900
~

サーバID	Position	メモリ容量	プロセッサ	通信速度
サーバE	1/13 - 4/13	4GB :	3GHz	100Mbps
サーバA	5/13 - 7/13	3GB :	3GHz	100Mbps
サーバC	8/13 - 10/13	3GB :	2GHz	100Mbps
サーバB	11/13 - 12/13	2GB :	2GHz	10Mbps
サーバD	13/13 - 13/13	1GB :	2GHz	10Mbps

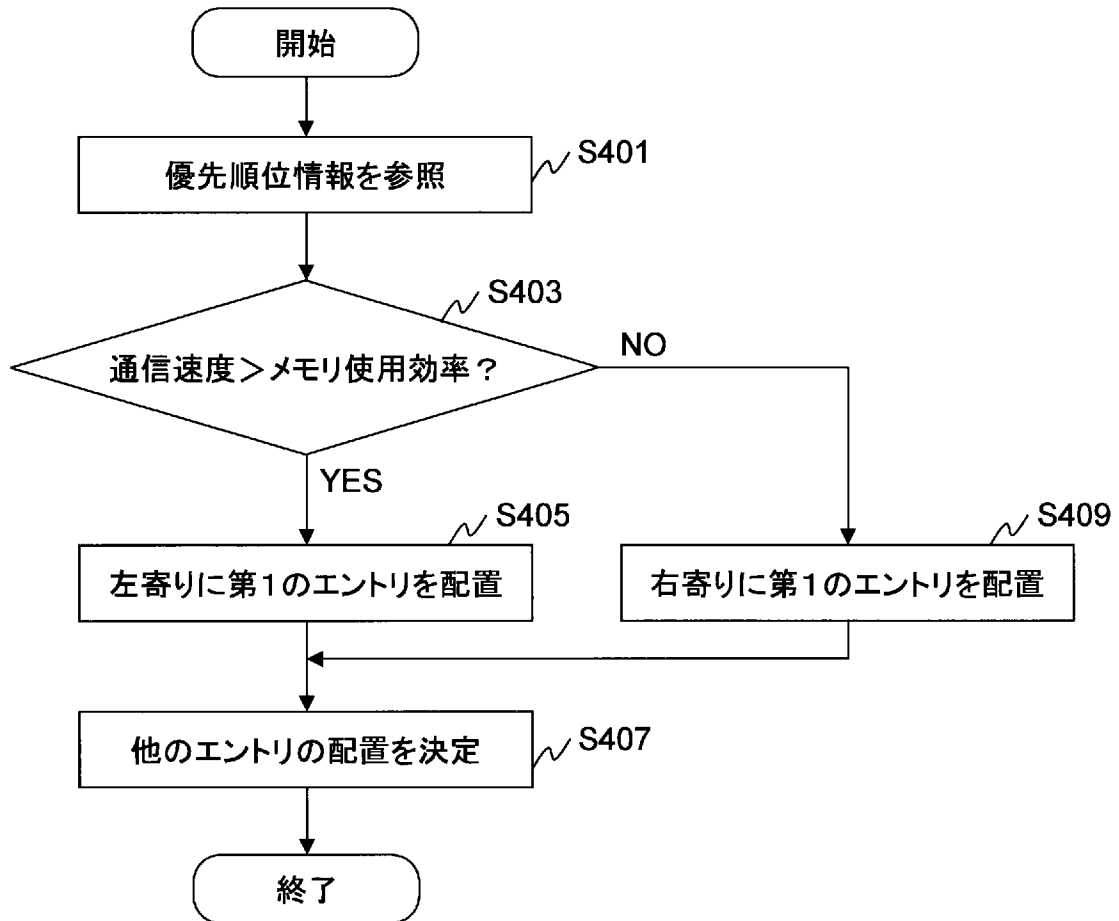
[図13B]

900
~

サーバID	Position	メモリ容量	プロセッサ	通信速度
サーバD	1/13 - 4/13	1GB :	2GHz	10Mbps
サーバA	5/13 - 7/13	3GB :	3GHz	100Mbps
サーバE	8/13 - 10/13	4GB :	3GHz	100Mbps
サーバC	11/13 - 12/13	3GB :	2GHz	100Mbps
サーバB	13/13 - 13/13	2GB :	2GHz	10Mbps

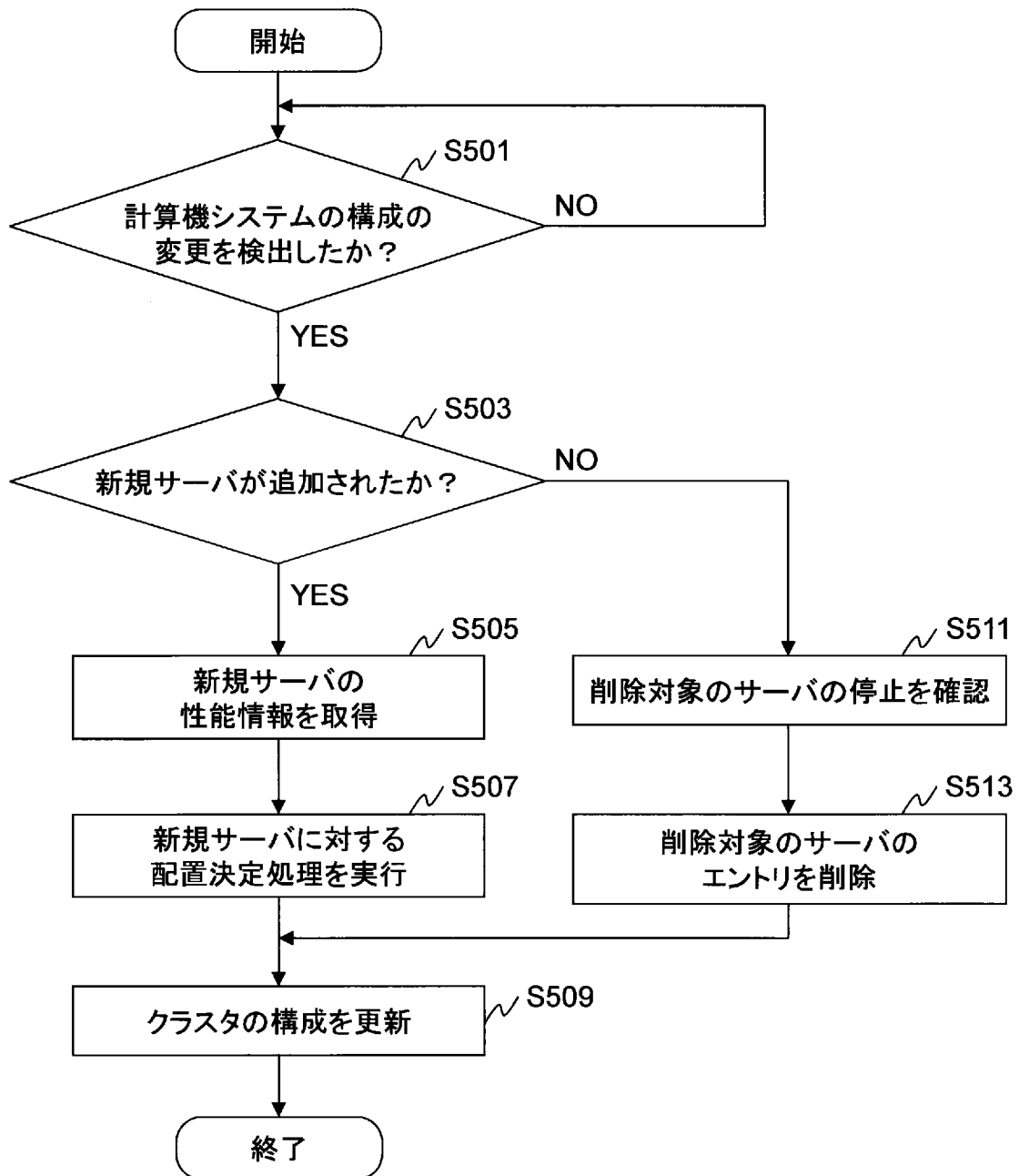
[図14]

山型配置処理

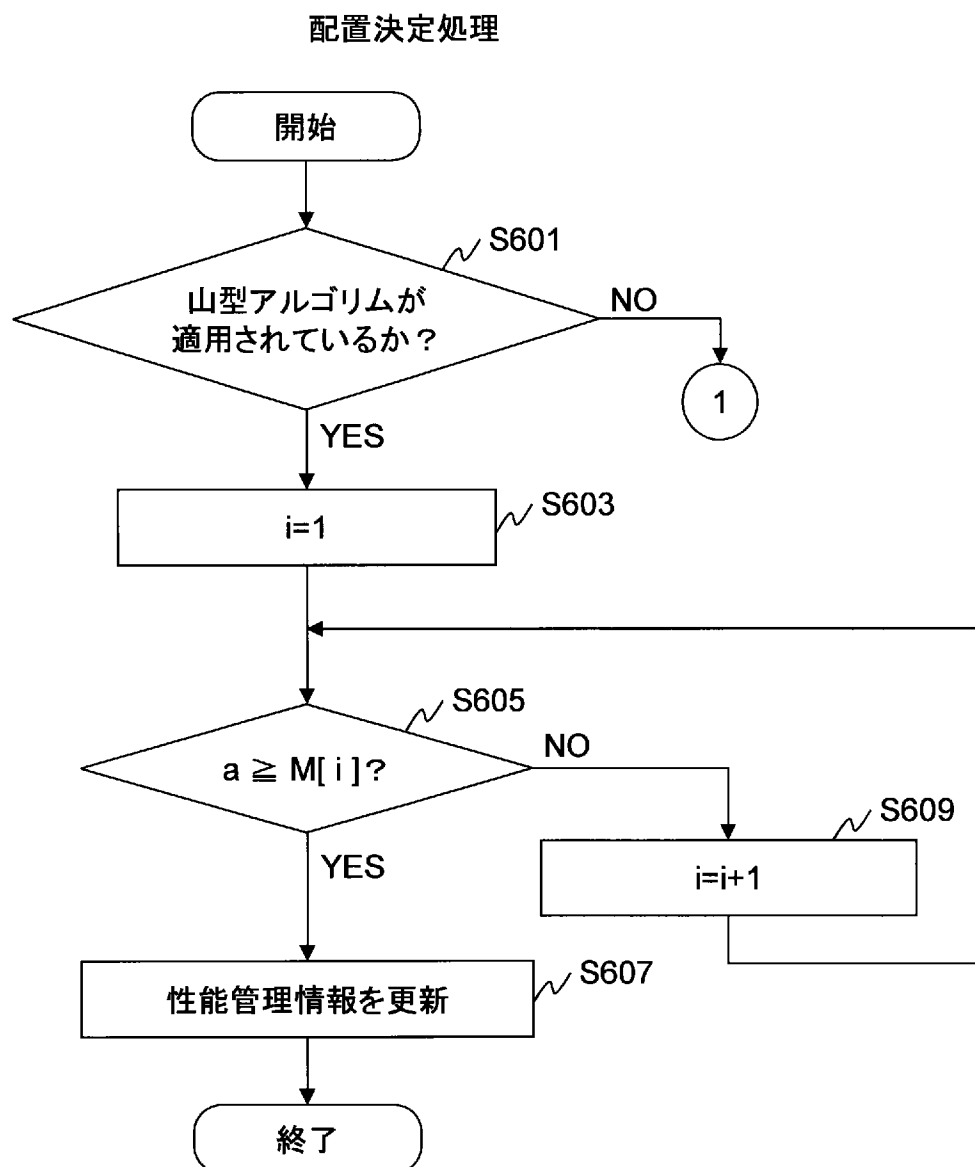


[図15]

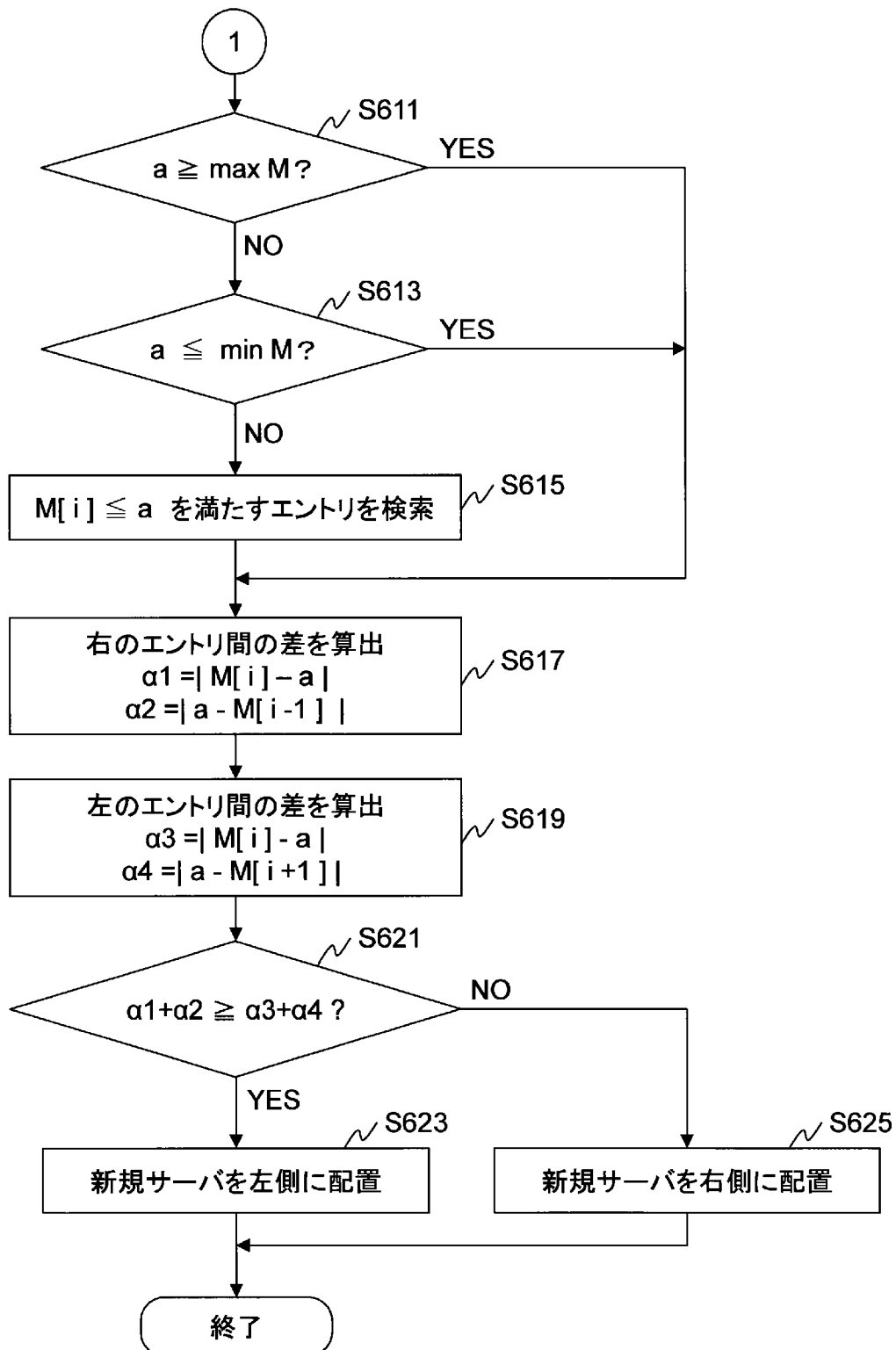
クラスタ構成変更処理



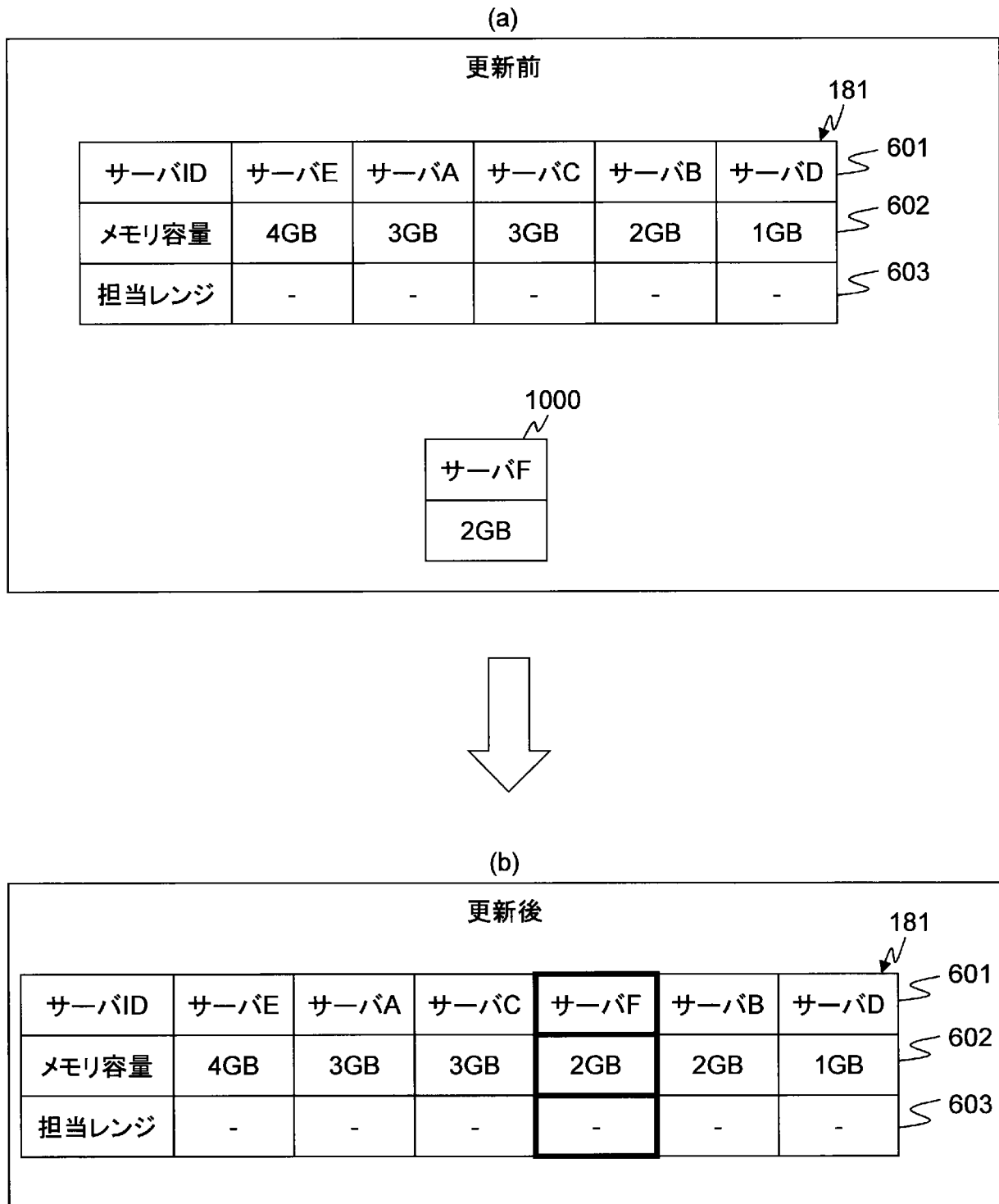
[図16A]



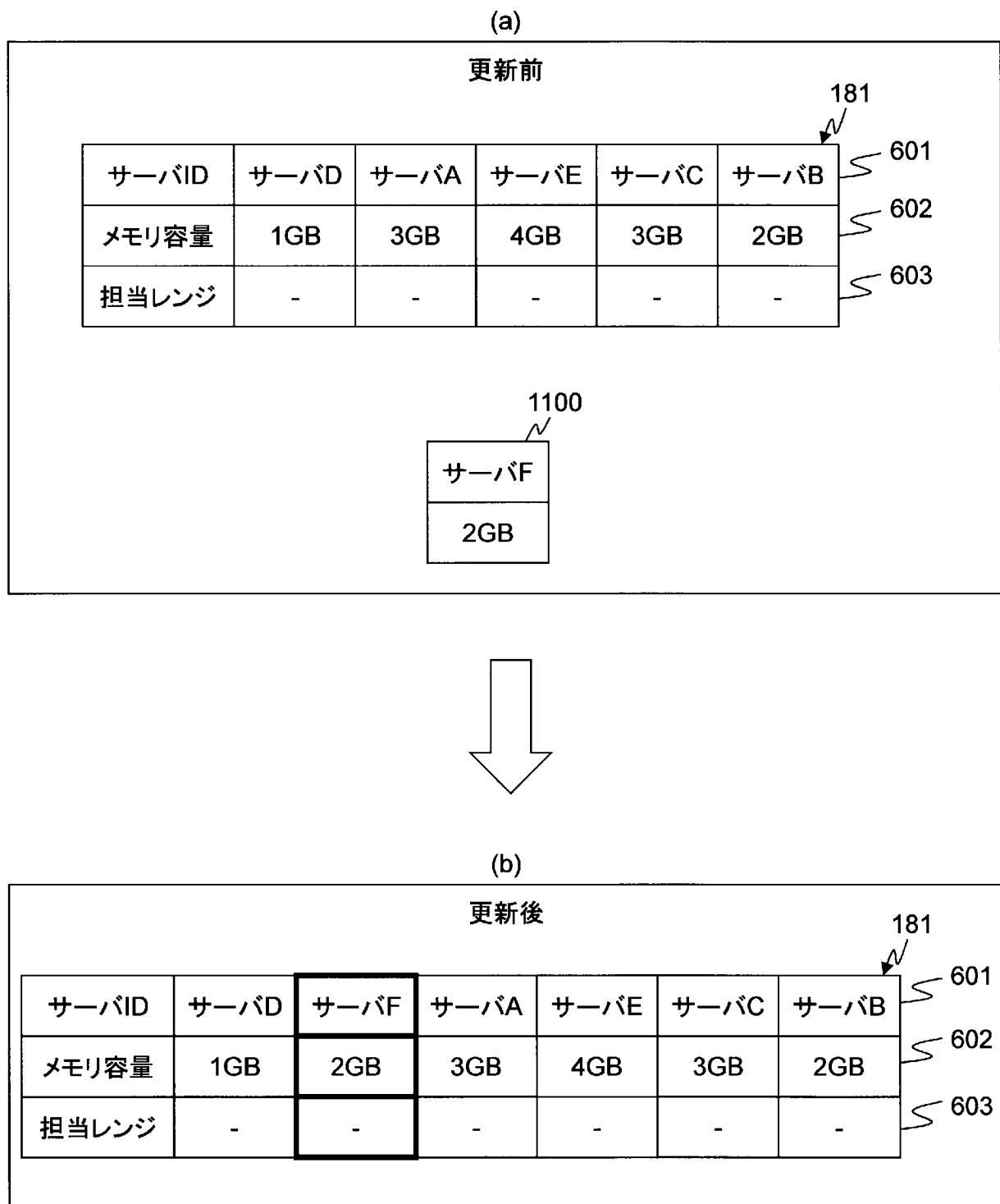
[図16B]



[図17]



[図18]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2012/059972

A. CLASSIFICATION OF SUBJECT MATTER

G06F12/00 (2006.01) i, G06F17/30 (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, G06F17/30

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	1996-2012
Kokai Jitsuyo Shinan Koho	1971-2012	Toroku Jitsuyo Shinan Koho	1994-2012

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

JSTPlus (JDreamII)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y A	Yasutoshi MIYAGI et al., "A study of the method to distribute data replication in distributed database for session control server", 2011 Nen IEICE Communications Society Conference Koen Ronbunshu 2, 30 August 2011 (30.08.2011), page 418	1, 7, 8, 14, 15 2-6, 9-13, 16-20
Y A	JP 5-233390 A (Toshiba Corp.), 10 September 1993 (10.09.1993), entire text; all drawings; particularly, paragraphs [0022] to [0025] (Family: none)	1, 7, 8, 14, 15 2-6, 9-13, 16-20

☒ 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
04 July, 2012 (04.07.12)

Date of mailing of the international search report
17 July, 2012 (17.07.12)

Name and mailing address of the ISA/
Japanese Patent Office

Authorized officer

Facsimile No.

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2012/059972

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	Hikotoshi NAKASATO et al., "A method of virtual server deployment for Consistent Hashing considering server performance and server load change", 2011 Nen IEICE Communications Society Conference Koen Ronbunshu 2, 30 August 2011 (30.08.2011), page 102	1-20
A	Shunsuke NAKAMURA et al., "A Cloud Storage Supporting both Read-Intensive and Write-Intensive Workloads", IPSJ SIG Notes 2011 (Heisei 23) Nendo 1 [CD-ROM], 15 June 2011 (15.06.2011), no.117, pages 1 to 9	1-20

A. 発明の属する分野の分類（国際特許分類（I P C））
Int.Cl. G06F12/00(2006.01)i, G06F17/30(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（I P C））

Int.Cl. G06F12/00, G06F17/30

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	1 9 2 2 - 1 9 9 6 年
日本国公開実用新案公報	1 9 7 1 - 2 0 1 2 年
日本国実用新案登録公報	1 9 9 6 - 2 0 1 2 年
日本国登録実用新案公報	1 9 9 4 - 2 0 1 2 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)

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
Y	宮城安敏ほか2名, セッション制御向け分散データベースのデータ複製に関する一検討, 電子情報通信学会 2 0 1 1 年通信ソサイエティ大会講演論文集 2, 2011.08.30, p. 418	1, 7, 8, 14, 15
A		2-6, 9-13, 16-20
Y	JP 5-233390 A (株式会社東芝) 1993.09.10, 全文・全図、特に段落【0022】～【0025】（ファミリーなし）	1, 7, 8, 14, 15
A		2-6, 9-13, 16-20

☒ 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

0 4 . 0 7 . 2 0 1 2

国際調査報告の発送日

1 7 . 0 7 . 2 0 1 2

国際調査機関の名称及びあて先

日本国特許庁（I S A / J P）

郵便番号 1 0 0 - 8 9 1 5

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）

池田 聡史

電話番号 0 3 - 3 5 8 1 - 1 1 0 1 内線 3 5 6 5

5 U

9 4 7 5

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	中里彦俊ほか3名, サーバ性能・サーバ負荷変動を考慮したコンシ ステントハッシュ法における仮想サーバ配置手法, 電子情報通信学 会 2 0 1 1 年通信ソサイエティ大会講演論文集 2, 2011.08.30, p. 102	1-20
A	中村俊介ほか2名, 読み出し性能と書き込み性能を両立させるクラ ウドストレージ, 情報処理学会研究報告 2 0 1 1 (平成23) 年 度1 [CD-ROM], 2011.06.15, 第117号, pp. 1~9	1-20