

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号

特許第7295461号

(P7295461)

(45)発行日 令和5年6月21日(2023.6.21)

(24)登録日 令和5年6月13日(2023.6.13)

(51)国際特許分類

G 0 6 F 16/2458(2019.01)

F I

G 0 6 F 16/2458

請求項の数 5 (全17頁)

(21)出願番号	特願2021-541901(P2021-541901)	(73)特許権者	000004226 日本電信電話株式会社 東京都千代田区大手町一丁目5番1号
(86)(22)出願日	令和1年8月29日(2019.8.29)	(74)代理人	100083806 弁理士 三好 秀和
(86)国際出願番号	PCT/JP2019/033914	(74)代理人	100129230 弁理士 工藤 理恵
(87)国際公開番号	WO2021/038795	(72)発明者	岩越 さやか 東京都千代田区大手町一丁目5番1号 日本電信電話株式会社内
(87)国際公開日	令和3年3月4日(2021.3.4)	(72)発明者	持田 誠一郎 東京都千代田区大手町一丁目5番1号 日本電信電話株式会社内
審査請求日	令和4年2月10日(2022.2.10)	(72)発明者	山本 直人 東京都千代田区大手町一丁目5番1号 最終頁に続く

(54)【発明の名称】 データベースシステム、分散処理装置、データベース装置、分散処理方法、および、分散処理プログラム

(57)【特許請求の範囲】

【請求項1】

分散処理装置と、複数のデータベース装置とを備えるデータベースシステムであって、前記分散処理装置は、

複数の前記データベース装置に関連するクエリの実行計画を列挙し、各実行計画のデータ転送時間に基づいていずれかの実行計画を選択する選択部と、

選択した実行計画に従って前記クエリを分割し、分割された分割クエリと分割クエリの実行結果の転送先とを含む指示を、対応するデータベース装置にそれぞれ送信する送信部と、

データベース装置から前記クエリの実行結果を受信し、出力する出力部と、を備え、

前記選択部は、前記実行計画の中から、他のデータベース装置への転送が禁止されたデータを、他のデータベース装置に転送する実行計画を除いたいずれかの実行計画を選択し、

前記データベース装置は、

前記分散処理装置から受信した指示に含まれる分割クエリを実行し、実行結果を、前記指示に含まれる転送先の他のデータベース装置、または前記分散処理装置に送信する実行部を、備え、前記指示に含まれる転送先の他のデータベース装置は、次に分割クエリを実行するデータベース装置である

データベースシステム。

【請求項2】

10

20

前記データベース装置は、

他のデータベース装置または前記分散処理装置との間のネットワーク性能を測定し、測定した性能情報を前記分散処理装置に送信する測定部を備え、

前記分散処理装置の前記選択部は、前記データベース装置から収集した前記性能情報と、各データベース装置の転送データ量とを用いて、各実行計画のデータ転送時間を算出する請求項 1 に記載のデータベースシステム。

【請求項 3】

分散処理装置であって、

複数のデータベース装置に関連するクエリの実行計画を列挙し、各実行計画のデータ転送時間に基づいていずれかの実行計画を選択する選択部と、

選択した実行計画に従って前記クエリを分割し、分割された分割クエリと分割クエリの実行結果の転送先とを含む指示を、対応するデータベース装置にそれぞれ送信する送信部と、

データベース装置から前記クエリの実行結果を受信し、出力する出力部と、を備え、

前記選択部は、前記実行計画の中から、他のデータベース装置への転送が禁止されたデータを、他のデータベース装置に転送する実行計画を除いたいずれかの実行計画を選択する分散処理装置。

【請求項 4】

分散処理装置と、複数のデータベース装置とを備えるデータベースシステムの分散処理方法であって、

前記分散処理装置は、

複数の前記データベース装置に関連するクエリの実行計画を列挙し、各実行計画のデータ転送時間に基づいていずれかの実行計画を選択する選択ステップと、

選択した実行計画に従って前記クエリを分割し、分割された分割クエリと分割クエリの実行結果の転送先とを含む指示を、対応するデータベース装置にそれぞれ送信する送信ステップと、

データベース装置から前記クエリの実行結果を受信し、出力する出力ステップと、を行い、

前記選択ステップは、前記実行計画の中から、他のデータベース装置への転送が禁止されたデータを、他のデータベース装置に転送する実行計画を除いたいずれかの実行計画を選択し、

前記データベース装置は、

前記分散処理装置から受信した指示に含まれる分割クエリを実行し、実行結果を前記指示に含まれる転送先の他のデータベース装置または前記分散処理装置に送信する実行ステップを行い、前記指示に含まれる転送先の他のデータベース装置は、次に分割クエリを実行するデータベース装置である

分散処理方法。

【請求項 5】

請求項 3 に記載の分散処理装置として、コンピュータを機能させる分散処理プログラム。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、データベースシステム、分散処理装置、データベース装置、分散処理方法、および、分散処理プログラムに関する。

【背景技術】

【0002】

ネットワークを介して分散したデータベース間を横断するクエリを処理する技術として、外部データベースを仮想的に統合する手法が知られている（非特許文献 1 参照）。

【0003】

非特許文献 1 では、ネットワークを介して分散するデータベースのデータを、1 つのサ

10

20

30

40

50

ーバに集約してからクエリを処理する。

【先行技術文献】

【非特許文献】

【0004】

【文献】「DB選定基準編要件から採択するPostgreSQL」、PostgreSQLエンタープライズ・コンソーシアム技術部会 WG # 2、P 28 - 30、[online]、インターネット<URL : https://www.pgecons.org/wp-content/uploads/PGECons/2015/WG2/14_ReferenceForDatabaseSelection.pdf>

【発明の概要】

【発明が解決しようとする課題】

【0005】

ネットワークを介して分散するデータベースのデータを、1つのサーバに集約してクエリ処理する場合、転送データ量が多くなり、データ転送に時間を要する。また、大量のデータを転送する場合、高額な転送コストが発生する。

【0006】

本発明は、上記事情に鑑みてなされたものであり、本発明の目的は、ネットワークを介した複数のデータベースのデータを1つの装置に集約することなく、複数のデータベースに関連するクエリを処理する技術を提供することにある。

【課題を解決するための手段】

【0007】

上記目的を達成するため、本発明の一態様は、分散処理装置と、複数のデータベース装置とを備えデータベースシステムであって、前記分散処理装置は、複数の前記データベース装置に関連するクエリの実行計画を列挙し、各実行計画のデータ転送時間に基づいていずれかの実行計画を選択する選択部と、選択した実行計画に従って前記クエリを分割し、分割された分割クエリと分割クエリの実行結果の転送先を含む指示を、対応するデータベース装置にそれぞれ送信する送信部と、データベース装置から前記クエリの実行結果を受信し、出力する出力部と、を備え、前記データベース装置は、前記分散処理装置から受信した指示に含まれる分割クエリを実行し、実行結果を前記指示に含まれる転送先の他のデータベース装置または前記分散処理装置に送信する実行部を、備える。

【0008】

本発明の一態様の分散処理装置は、複数のデータベース装置に関連するクエリの実行計画を列挙し、各実行計画のデータ転送時間に基づいていずれかの実行計画を選択する選択部と、選択した実行計画に従って前記クエリを分割し、分割された分割クエリと分割クエリの実行結果の転送先を含む指示を、対応するデータベース装置にそれぞれ送信する送信部と、データベース装置から前記クエリの実行結果を受信し、出力する出力部と、を備える。

【0009】

本発明の一態様のデータベース装置は、自データベース装置および他データベース装置に関連するクエリの実行計画に従って、前記クエリが分割された分割クエリと、分割クエリの実行結果の転送先を含む指示を分散処理装置から受信し、前記分割クエリを実行し、実行結果を前記転送先の他データベース装置または前記分散処理装置に送信する実行部と、他のデータベース装置または前記分散処理装置との間のネットワーク性能を測定し、測定した性能情報を前記分散処理装置に送信する測定部を備え、前記実行計画は、前記性能情報を用いて算出された、自データベース装置および他データベース装置のデータ転送時間の合計が最小となる実行計画である。

【0010】

本発明の一態様は、分散処理装置と、複数のデータベース装置とを備えるデータベースシステムの分散処理方法であって、前記分散処理装置は、複数の前記データベース装置に関連するクエリの実行計画を列挙し、各実行計画のデータ転送時間に基づいていずれかの実行計画を選択する選択ステップと、選択した実行計画に従って前記クエリを分割し、分

10

20

30

40

50

割された分割クエリと分割クエリの実行結果の転送先とを含む指示を、対応するデータベース装置にそれぞれ送信する送信ステップと、データベース装置から前記クエリの実行結果を受信し、出力する出力ステップと、を行い、前記データベース装置は、前記分散処理装置から受信した指示に含まれる分割クエリを実行し、実行結果を前記指示に含まれる転送先の他のデータベース装置または前記分散処理装置に送信する実行ステップを行う。

【0011】

本発明の一態様は、上記分散処理装置として、コンピュータを機能させる分散処理プログラムである。

【0012】

本発明の一態様は、上記データベース装置として、コンピュータを機能させる分散処理プログラムである。

【発明の効果】

【0013】

本発明によれば、ネットワークを介した複数のデータベースのデータを1つの装置に集約することなく、複数のデータベースに関連するクエリを処理する技術を提供することができる。

【図面の簡単な説明】

【0014】

【図1】本発明の実施形態の分散データベースシステムの構成例を示す図である。

【図2】DB装置の分散DBに格納されたテーブルの例である。

【図3】DB装置の分散DBに格納されたテーブルの例である。

【図4】DB装置の分散DBに格納されたテーブルの例である。

【図5】分散データベースシステムの動作を示すフローチャートである。

【図6】装置間（ノード間）のネットワーク性能を模式的に示す図である。

【図7】クエリの一例である。

【図8】クエリ木の一例である。

【図9】直接転送と迂回転送とを説明する説明図である。

【図10】実行計画の実行コストを説明する説明図である。

【図11】転送時間を用いた場合の実行計画の実行コストの例である。

【図12】実行計画1のノードT2の処理を模式的に示す図である。

【図13】実行計画1のノードT1の処理を模式的に示す図である。

【図14】実行計画1のノードKの処理を模式的に示す図である。

【図15】変形例1のポリシーエンフォースメントが存在する場合の実行計画を示す図である。

【図16】変形例2の分散データベースシステムの構成例である。

【図17】本実施形態の比較例1を説明する説明図である。

【図18】本実施形態の比較例2を説明する説明図である。

【図19】本実施形態の方式を説明する説明図である。

【図20】分散処理装置およびDB装置のハードウェア構成例である。

【発明を実施するための形態】

【0015】

以下、本発明の実施の形態について、図面を参照して説明する。図面の記載において、同一部分には同一符号を付して説明を省略する。

【0016】

（分散DBシステムの構成）

図1は、本実施形態の分散DBシステム（データベースシステム）の構成例である。図示する分散DBシステムは、分散処理装置1と、複数のDB装置2とを備える。分散処理装置1および複数のDB装置2は、ネットワークを介して他の装置と通信可能のように接続されている。図示する例では、DB装置は3つであるが、DB装置の数は3に限定されない。DB装置2は少なくとも2つであればよい。なお、分散処理装置1およびDB装置

10

20

30

40

50

2 は、「ノード」ともいう。

【0017】

分散処理装置 1（ノード C）は、ネットワークを介して分散された複数の DB 装置間を横断するクエリを、ネットワーク性能に基づいて処理する。図示する分散処理装置 1 は、クエリ解析部 11 と、実行計画選択部 12 と、指示送信部 13 と、出力部 14 と、収集部 15 と、記憶部 16 とを備える。

【0018】

クエリ解析部 11 は、入力されたクエリ 5 を解析し、当該クエリ 5 を木構造で表現したクエリ木を生成する。本実施形態のクエリ 5 は、複数の DB 装置 2 に関連するクエリ、すなわち、複数の DB 装置 2 を横断的に処理するクエリである。

10

【0019】

実行計画選択部 12（選択部）は、クエリ 5 の実行計画を列挙し、各実行計画のデータ転送時間に基づいていずれかの実行計画を選択する。具体的には、実行計画選択部 12 は、クエリ木造に基づいた複数の実行計画を生成し、複数の実行計画からネットワーク性能に基づいて最適な実行計画を選択する。選択部 12 は、DB 装置 2 から収集したネットワーク性能情報と、各 DB 装置 2 の転送データ量とを用いて、各実行計画のデータ転送時間を算出する。選択される実行計画は、例えば、複数の DB 装置 2 のデータ転送時間の合計が最小となる実行計画である。

【0020】

指示送信部 13（送信部）は、選択した実行計画に従って、複数の DB 装置 2 にクエリ 5 を分散処理させる。具体的には、指示送信部 13 は、選択した実行計画に従ってクエリ 5 を分割し、分割された分割クエリと分割クエリの実行結果の転送先とを含む指示を、対応する DB 装置 2 にそれぞれ送信する。

20

【0021】

出力部 14 は、DB 装置 2 からクエリ 5 の最終の実行結果を受信し、クエリ結果 6 として出力する。本実施形態の出力部 14 は、選択された実行計画の最後に設定された 1 つの DB 装置から、クエリ 5 の実行結果を受信する。出力部 14 は、受信した実行結果を Tableau などの可視化ツールを用いて可視化し、可視化したクエリ結果 6 を出力してもよい。

【0022】

収集部 15 は、DB 装置 2（ノード）間のネットワーク性能情報（ネットワーク帯域情報など）を、DB 装置 2 から収集し、記憶部 16 に記憶する。記憶部 16 には、収集部 15 により収集されたネットワーク性能情報が記憶される。

30

【0023】

各 DB 装置 2（ノード K、T1、T2）は、測定部 21 と、実行部 22 と、分散 DB 23 とを備える。測定部 21 は、他の DB 装置 2 または分散処理装置 1 との間のネットワーク性能を測定し、測定した性能情報を分散処理装置 1 に送信する。すなわち、測定部 21 は、ノード間のネットワーク性能を測定する。

【0024】

実行部 22 は、分散処理装置 1 から受信した指示に含まれる分割クエリを実行し、実行結果を前記指示に含まれる転送先の他の DB 装置 2 または分散処理装置 1 に送信する。分散 DB 23 には少なくとも 1 つのデータベースが格納される。

40

【0025】

図 2 から図 4 は、図 1 に示す各分散 DB 23 に格納されたテーブルの例である。図 2 に、ノード K の DB 装置 2 の分散 DB 23 に格納されたテーブルを示す。図 2 の分散 DB 23 は、百貨店のデータベースであって、CM テーブル（顧客管理テーブル）と、TM テーブル（購買履歴テーブル）とを有する。CM テーブルのレコード数は、6 M である。TM テーブルのレコード数は、60 M である。

【0026】

図 3 に、ノード T1 の DB 装置 2 の分散 DB 23 に格納されたテーブルを示す。図 3 の分散 DB 23 は、百貨店に入っているテナント 1 のデータベースであって、TCM1 テー

50

ブル（顧客管理テーブル）と、TTM1テーブル（購買履歴テーブル）とを有する。TCM1テーブルのレコード数は、50000である。TTM1テーブルのレコード数は、500000である。TTM1テーブルは、テナント1のユーザID（Tuid）と、百貨店のCMテーブルのユーザID（Uid）とが関連付けて記憶されている。

【0027】

図4に、ノードT2のDB装置2の分散DB23に格納されたテーブルを示す。図4の分散DB23は、百貨店に入っているテナント2のデータベースであって、TCM2テーブル（顧客管理テーブル）と、TTM2テーブル（購買履歴テーブル）とを有する。TCM2テーブルのレコード数は、20000である。TTM2テーブルのレコード数は、200000である。TTM2テーブルは、テナント2のユーザID（Tuid）と、百貨店のCMテーブルのユーザID（Uid）とが関連付けて記憶されている。

10

【0028】

（分散DBシステムの動作）

以下に、本実施形態の分散DBシステムの動作を説明する。ここでは、図1に示す3つのDB装置2に関連するクエリとして、エステに行った人は、その直後に洋服または靴を購入する確率が高いという仮説を検証するためのクエリを実行する場合について説明する。

【0029】

図5は、本実施形態の分散DBシステムの動作を示すフローチャートである。各DB装置2は、自身のDB装置2と、他のDB装置2または分散処理装置1との間のネットワーク性能を測定する（S11）。そして、DB装置2は、測定したネットワーク性能情報を分散処理装置1に送信する。ここでは、ネットワーク性能としてネットワーク帯域（データ転送速度：bps）を用いるが、これに限定されない。分散処理装置1は、各DB装置2からネットワーク性能情報を収集し、記憶部16に記憶する（S12）。

20

【0030】

図6は、ノード間のネットワーク性能を模式的に示す図である。図示する例では、例えば、ノードKとノードT1との間のネットワーク帯域は10Mbpsである。ノードKとノードT2との間のネットワーク帯域は10Mbpsであり、ノードKとノードCとの間のネットワーク帯域は5Mbpsである。

【0031】

なお、S11およびS12の処理は、S13以降の処理を行う度に、毎回行う必要はない。例えば、記憶部16にネットワーク性能情報が既に記憶されている場合は、S11およびS12は行われず、分散処理装置1は、記憶部16に記憶されたネットワーク性能情報を用いてもよい。また、定期的またはオペレータの指示などの所定のタイミングでS11およびS12を実行し、記憶部16に記憶されたネットワーク性能情報を更新してもよい。

30

【0032】

次に、分散処理装置1は、ユーザにより入力されたクエリを受け付け、当該クエリを解析し、木構造のクエリ木を生成する（S13）。

【0033】

図7は、複数のDB装置2に関連するクエリの一例である。図示するクエリは、エステに行った人は、その直後に洋服または靴を購入する確率が高いという仮説を検証するために、商品またはサービスの分類が「服」、「靴」および「エステ」を購入した百貨店ユーザの購買履歴を抽出する検索条件である。図示するクエリが対象とするデータは、ノードKの分散DB（百貨店）のテーブルCM、ノードT1の分散DB（テナント1）のテーブルTTM1、および、ノードT2の分散DB（テナント2）のテーブルTTM2である（図2－図4参照）。

40

【0034】

図8は、図7のクエリから生成されたクエリ木の一例である。

【0035】

分散処理装置1は、入力されたクエリを実行可能な少なくとも1つの実行計画を列挙（

50

生成)する(S14)。そして、分散処理装置1は、各実行計画の実行コスト(実行時間)を算出し、実行コストに基づいて最適な実行計画を選択する(S15)。具体的には、分散処理装置1は、各実行計画のデータ転送時間に基づいていずれかの実行計画を選択する。図6に示すネットワーク構成の場合、生成される実行計画(実行ルート)は、以下の6通りとなる。

【0036】

実行計画1: ノードT2→ノードT1→ノードK →ノードC

実行計画2: ノードT2→ノードK →ノードT1→ノードC

実行計画3: ノードT1→ノードT2→ノードK →ノードC

実行計画4: ノードT1→ノードK →ノードT2→ノードC

実行計画5: ノードK →ノードT1→ノードT2→ノードC

実行計画6: ノードK →ノードT2→ノードT1→ノードC

なお、1つの経路(例えば、ノードK→ノードT2)において、複数の転送方法(直接転送、迂回転送)があるが、ここでは直接転送のみを用いた実行計画を生成する。

【0037】

図9は、直接転送と、迂回転送とを説明する説明図である。迂回転送は、直接転送以外の転送方法である。図9は、図6に示すネットワーク構成において、ノードKからノードT2への複数の転送方法を示す。直接転送91では、データは、KからT2へ直接転送される。迂回転送92では、Cを経由したノードK→ノードC→ノードT2と、ノードT1を経由したノードK→ノードT1→ノードT2と、ノードC、T1を経由したノードK→ノードC→ノードT1→ノードT2の3つの迂回転送を示している。

【0038】

図10は、実行計画1の実行コストの算出を説明するための説明図である。実行計画1では、最初にノードT2が、ノードT2用の指示51に含まれる分割クエリをテーブルTTM2に対して実行し、実行結果であるTEMPをノードT1に送信する。ノードT1は、ノードT1用の指示52に含まれる分割クエリをテーブルTTM1およびTEMPに対して実行し、実行結果であるTをノードKに送信する。ノードKは、ノードK用の指示53に含まれる分割クエリをテーブルCMおよびTに対して実行し、最終の実行結果であるResultをノードC(分散処理装置1)に送信する。

【0039】

図10に示す実行コストテーブル90では、分散処理装置1により算出された各ノード(T1、T2、K)のクエリ処理時間と、他のノードへの実行結果のデータ転送時間との合計を実行コスト(実行時間)として示す。クエリ処理時間は、対象とするテーブルのデータサイズを用いて算出される。データ転送時間は、転送レコード数および転送速度を用いて算出される。分散処理装置1は、クエリ最適化の機能を用いて、各ノードの転送レコード数を推定する。図10では、データ転送時間とクエリ処理時間との合計時間を、実行コストとしている。ただし、クエリ処理時間を加味することなく、大きな割合を占めるデータ転送時間のみを実行コストとしてもよい。

【0040】

図11は、データ転送時間のみを用いた場合の各実行計画に実行コスト(実行時間)を示す図である。この場合、分散処理装置1は、実行コストが最小(61.33)の実行計画1を選択する(S15)。

【0041】

分散処理装置1は、選択した実行計画1に従って、入力されたクエリを分割して各ノード用の分割クエリを生成する。そして、分散処理装置1は、ノードT1、T2、K(DB装置2)毎に、分割クエリと、当該分割クエリの実行結果の転送先とを含む指示を生成し、各ノードに対応する指示を送信する(S16)。図10に、前記指示の一例51-53を示す。分割クエリは、入力されたクエリを各ノードの実行内容に分割したものである。各ノードは、指示に従って分割クエリを実行し、実行結果を指示された転送先に転送する(S17)。

10

20

30

40

50

【0042】

図12は、ノードT2の処理を示す図である。実行計画1において、最初にノードT2は、自身のテーブルTTM2に対して指示51の分割クエリを実行し、実行結果をTEMPとして指示されたノードT1に転送する。ここでは、ノードT2は、テーブルTTM2から分類が「服」、「靴」および「エステ」のレコードを抽出し、抽出したレコードをTEMPとしてノードT1に転送する。

【0043】

図13は、ノードT1の処理を示す図である。ノードT1は、ノードT2から受信したTEMP（ノードT2の実行結果）と、自身のテーブルTTM1に対して指示52の分割クエリを実行し、実行結果をTとして指示されたノードKに転送する。ここでは、ノードT1は、テーブルTTM1から分類が「服」、「靴」および「エステ」のレコードを抽出し、抽出したレコードとTEMPとを統合し、TとしてノードKに転送する。

10

【0044】

図14は、ノードKの処理を示す図である。ノードKは、ノードT1から受信したT（ノードT1の実行結果）と、自身のテーブルCMに対して指示53の分割クエリを実行し、最終の実行結果をResultとして指示されたノードCに転送する。ここでは、ノードKは、百貨店のユーザごとの「服」、「靴」および「エステ」の購入履歴のレコードをResultとしてノードCに転送する。

【0045】

分散処理装置1（ノードC）は、ノードKからクエリ5の最終の実行結果を受信し、当該実行結果を出力する（S18）。分散処理装置1は、受信した実行結果をTableauなどの可視化ツールを用いて可視化し、可視化したクエリ結果を出力してもよい。

20

【0046】

（変形例1）

次に、本実施形態の変形例1について説明する。変形例1では、ポリシーエンフォースメントが存在する場合、分散処理装置1は、当該ポリシーエンフォースメントに反する実行計画を除外する。ポリシーエンフォースメントとしては、例えば、クエリ処理前のデータを、他のノードに転送できないなどがある。このような動かせないデータが存在する場合、分散処理装置1は、当該データについては、分割クエリの実行結果のみを他のノードに転送する。

30

【0047】

図15は、ノードKのテーブルCMの転送を禁止するポリシーエンフォースメントが存在する場合の実行計画を示す図である。分散処理装置1（実行計画選択部12）は、図15のA行の転送元にノードKが設定された実行計画、および、B行の転送元にノードKが設定された実行計画は、ポリシーエンフォースメントに反する実行計画であると判定し、除外する。したがって、分散処理装置1は、図5のS14において、実行計画2、4-6を除外して、実行計画1、3のみを列挙し、最小コストの実行計画1を選択する。

【0048】

変形例1では、所定のノードのクエリ実行前のデータを、他のノードに転送することなく、クエリを実行できる。すなわち、本変形例を含む本実施形態では、分散DBのデータを一か所に集約する必要がないため、個人情報などの機密データの転送を禁止するポリシーエンフォースメントが存在する場合であっても、転送が禁止されたデータを含む複数の分散DBに関連するクエリに適用することができる。したがって、本実施形態では、外部への転送が禁止されたデータを分析することができる。

40

【0049】

（変形例2）

図16は、本実施形態の変形例2の分散DBシステムの構成例である。変形例2の分散DBシステムは、分散処理装置1が収集部15を備えず、DB装置2が測定部21を備えない点において、図1に示す分散DBシステムと異なる。この場合、分散処理装置1の記憶部16には、あらかじめ測定または設計されたノード間のネットワーク性能情報が記憶

50

されている。このように、DB装置2は、測定部21を備えていても、測定部21を備えていなくてもよい。

【0050】

(本実施形態の効果)

以上説明した本実施形態の分散DBシステムでは、分散処理装置1は、複数のDB装置2に関連するクエリの実行計画を列挙し、各実行計画のデータ転送時間に基づいていずれかの実行計画を選択する実行計画選択部12と、選択した実行計画に従ってクエリを分割し、分割された分割クエリと分割クエリの実行結果の転送先とを含む指示を、対応するDB装置2にそれぞれ送信する指示送信部13と、DB装置2からクエリの実行結果を受信し、出力する出力部14とを備え、DB装置2は、分散処理装置1から受信した指示に含まれる分割クエリを実行し、実行結果を指示に含まれる転送先の他のDB装置2または分散処理装置1に送信する実行部22を、備える。

10

【0051】

これにより、本実施形態では、ネットワークを介した複数のDB装置2のデータを1つの装置に集約することなく、複数のDB装置2に関連するクエリを処理することができる。したがって、本実施形態では、特定のネットワークへの負荷の集中を回避し、効率的にクエリを実行することが可能となる。また、データ転送時間およびデータ転送コストを削減することができる。

【0052】

また、本実施形態では、列挙した実行計画のデータ転送時間に基づいていずれかの実行計画を選択する。これにより、本実施形態では、ネットワーク性能に応じて、クエリの最適な実行計画を選択することができ、クエリの実行コストを削減することができる。

20

【0053】

また、本実施形態の変形例1では、分散処理装置1は、実行計画の中から、ポリシーエンフォースメントに反する実行計画を除外する。本実施形態では、複数のDB装置2のデータを1つの装置に集約することなく、各ノード間で実行結果を送受信してクエリを実行するため、外部へのデータ転送を禁止するポリシーエンフォースメントが存在する場合であっても、本実施形態の分散クエリ処理方式を適用することができる。

【0054】

(比較例)

図17および図18は、本実施形態の比較例1、2を説明する説明図である。ノードAの分散DB-AとノードBの分散DB-Bは、それぞれオンプレミスに存在し、分散DB-Aは100万レコード、分散DB-Bは100レコードを有する。ここでは、レコード数が50程度になるような分散DB-Aおよび分散DB-Bをクロスするクエリを実行する場合の動作を説明する。

30

【0055】

図17に示す比較例1は、BI (Business Intelligence) ツールを用いた比較例である。比較例1では、各ノードは、自身が備える分散DBの全レコードを集約ノードに転送し、集約ノードは、転送されたレコードに対してクエリを実行する。比較例1では、1箇所の集約ノードにデータを集約するため、転送データ量が多くなり、データ転送時間および転送コストが増大する。

40

【0056】

図18に示す比較例2は、PostgreSQLのForeign Data Wrapper(FDW)を用いて、ノードAの分散DB-AとノードBの分散DB-Bに、集約演算をPush Down (委譲) する方式である。すなわち、比較例2では、一部のクエリはPush Down可能であり、各ノードが自身の分散DBに対して処理した結果のデータを集約ノードに転送する。集約ノードは、各ノードから転送されたデータの結合などを行う。

【0057】

具体的には、ノードAは、Push Downのクエリ処理により分散DB-Aのレコードを50万レコードに削減して、集約ノードに転送する。ノードBも同様に、分散DB-Bのレコ

50

ードを70レコードに削減して集約ノードに転送する。集約ノードは、各ノードから送信されたレコードを結合し50レコードとする。

【0058】

比較例2では、クエリの一部を下位ノードにPush Downすることで、集約ノードへ転送するデータをフィルタリング（削減）することができ、データ転送時間を圧縮することは可能である。しかしながら、分散DB-A単体で可能なフィルタリング処理は分散DB-AにPush Downできるが、分散DB-Bのデータが必要なフィルタリング処理は分散DB-AにPush Downできない。このため、Push Downによるフィルタリング処理の効果は限定的である。すなわち、比較例2では、複数のDB間のデータ結合には処理対象のデータを集約ノードに集める必要があり、特定のネットワークに負荷が集中してしまう。

10

【0059】

これに対し、図18に示す本実施形態の方式では、ノード間のネットワーク性能および転送データ量を考慮して最適な実行計画を選択し、ノード間でデータを相互に転送し、最終の実行結果のみを分散処理装置1に送信する。これにより本実施形態では、データ転送時間を圧縮することができる。

【0060】

（分散処理装置およびDB装置のハードウェア構成）

上記説明した分散処理装置1およびDB装置2は、例えば、図20に示すような汎用的なコンピュータシステムを用いることができる。図示するコンピュータシステムは、CPU（Central Processing Unit、プロセッサ）901と、メモリ902と、ストレージ903（HDD：Hard Disk Drive、SSD：Solid State Drive）と、通信装置904と、入力装置905と、出力装置906とを備える。メモリ902およびストレージ903は、記憶装置である。このコンピュータシステムにおいて、CPU901がメモリ902上にロードされた所定のプログラムを実行することにより、各装置の各機能が実現される。例えば、分散処理装置1およびDB装置2の各機能は、分散処理装置1用のプログラムの場合は分散処理装置1のCPUが、DB装置2用のプログラムの場合はDB装置2のCPUが、それぞれ実行することにより実現される。

20

【0061】

また、分散処理装置1およびDB装置2は、1つのコンピュータで実装されてもよく、あるいは複数のコンピュータで実装されても良い。また、分散処理装置1およびDB装置2は、コンピュータに実装される仮想マシンであっても良い。

30

【0062】

分散処理装置1用のプログラムおよびDB装置2用のプログラムは、HDD、SSD、USB（Universal Serial Bus）メモリ、CD（Compact Disc）、DVD（Digital Versatile Disc）などのコンピュータ読取り可能な記録媒体に記憶することも、ネットワークを介して配信することもできる。

【0063】

なお、本発明は上記実施形態および変形例に限定されるものではなく、その要旨の範囲内で数々の変形が可能である。

【符号の説明】

40

【0064】

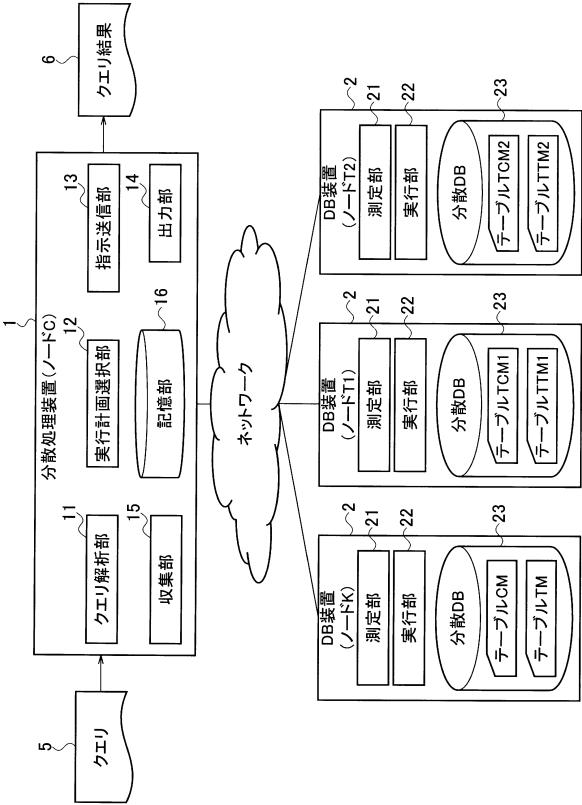
- 1：分散処理装置（ノードC）
- 11：クエリ解析部
- 12：実行計画選択部
- 13：指示送信部
- 14：出力部
- 15：収集部
- 16：記憶部
- 2：DB装置（ノードK、T1、T2）
- 21：測定部

50

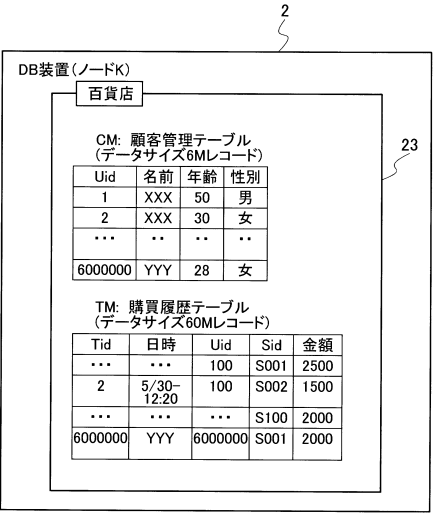
- 2 2 : 実行部
- 2 3 : 分散 D B
- 5 : クエリ
- 6 : クエリ結果

【図面】

【図 1】



【図 2】



10

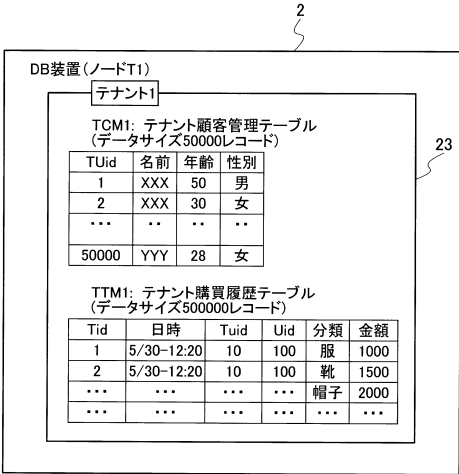
20

30

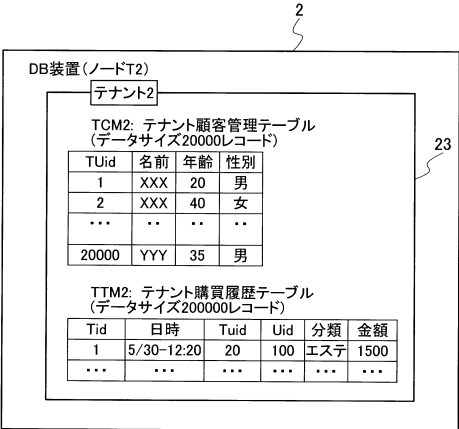
40

50

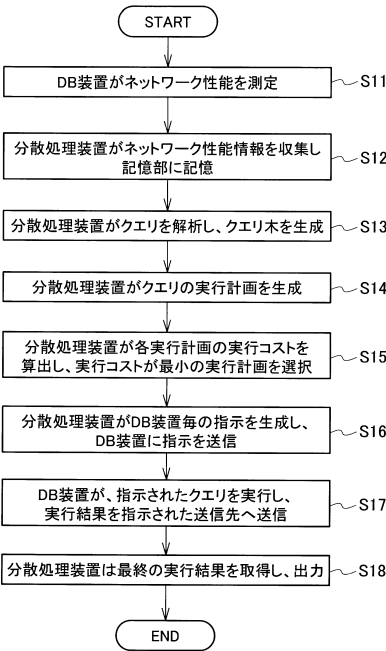
【図 3】



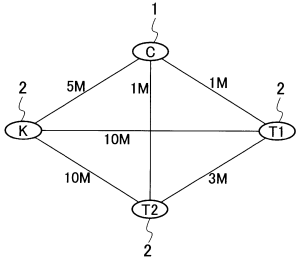
【図 4】



【図 5】



【図 6】



10

20

30

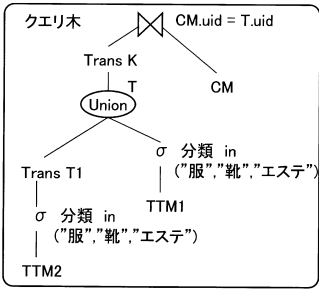
40

50

【図 7】

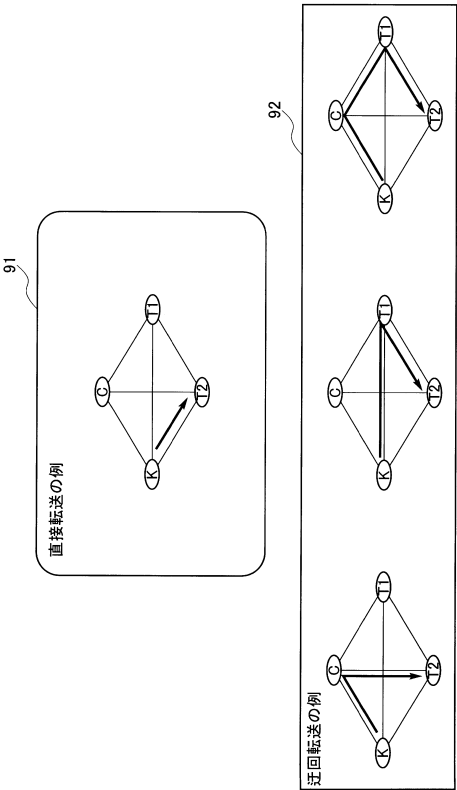
```
> select CM.uid, T.分類, T.金額 from CM,
      (select * from TTM1) union (select * from TTM2) as T
where
  CM.uid = T.uid and
  T.分類 in ("服","靴","エステ")
```

【図 8】

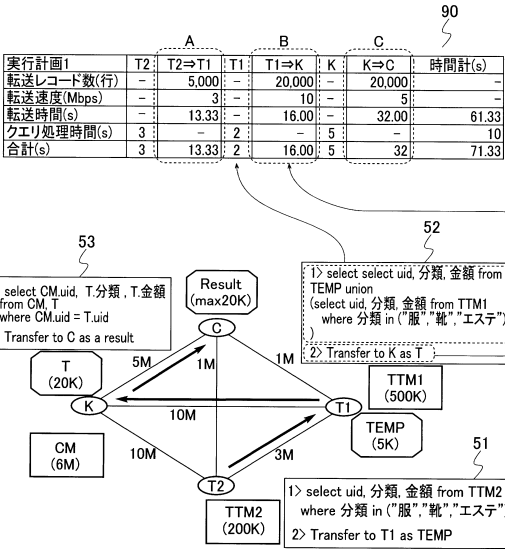


10

【図 9】



【図 10】



20

30

40

50

【図 1 1】

	A	B	C	
実行計画1	T2⇒T1	T1⇒K	K⇒C	転送時間計(s)
転送レコード数(行)	5,000	20,000	20,000	-
転送速度(Mbps)	3	10	5	-
転送時間(s)	13.33	16.00	32.00	61.33

← 最小コスト

実行計画2	T2⇒K	K⇒T1	T1⇒C	転送時間計(s)
転送レコード数(行)	5,000	20,000	20,000	-
転送速度(Mbps)	10	10	1	-
転送時間(s)	4.00	16.00	160.00	180.00

実行計画3	T1⇒T2	T2⇒K	K⇒C	転送時間計(s)
転送レコード数(行)	15,000	20,000	20,000	-
転送速度(Mbps)	3	10	5	-
転送時間(s)	40.00	16.00	32.00	88.00

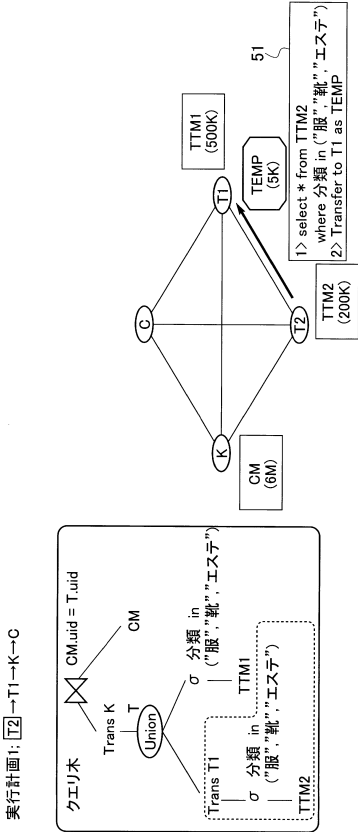
実行計画4	T1⇒K	K⇒T2	T2⇒C	転送時間計(s)
転送レコード数(行)	15,000	20,000	20,000	-
転送速度(Mbps)	10	10	1	-
転送時間(s)	12.00	16.00	160.00	188.00

実行計画5	K⇒T1	T1⇒T2	T2⇒C	転送時間計(s)
転送レコード数(行)	6,000,000	20,000	20,000	-
転送速度(Mbps)	10	3	1	-
転送時間(s)	4800.00	53.33	160.00	5013.33

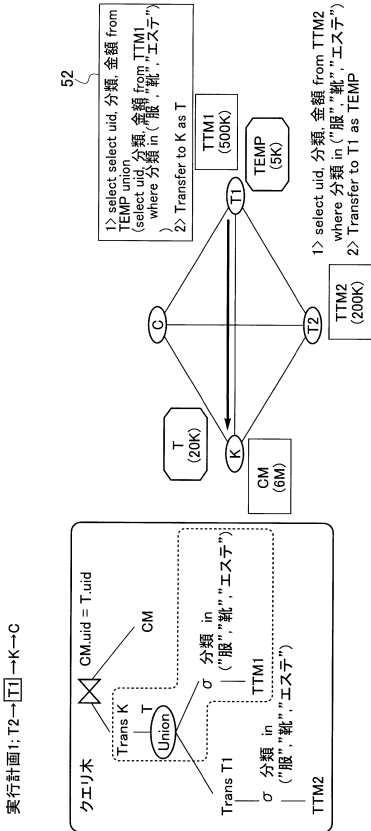
実行計画6	K⇒T2	T2⇒T1	T1⇒C	転送時間計(s)
転送レコード数(行)	6,000,000	20,000	20,000	-
転送速度(Mbps)	3	3	1	-
転送時間(s)	4800.00	53.33	160.00	5013.33

※1レコード = 1kB = 8kbit

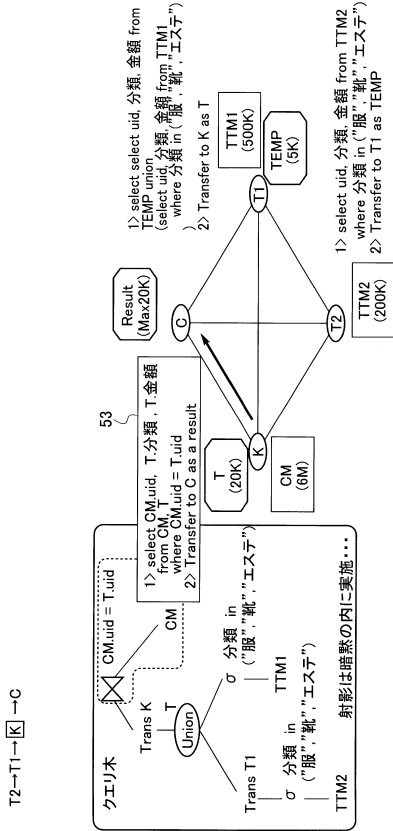
【図 1 2】



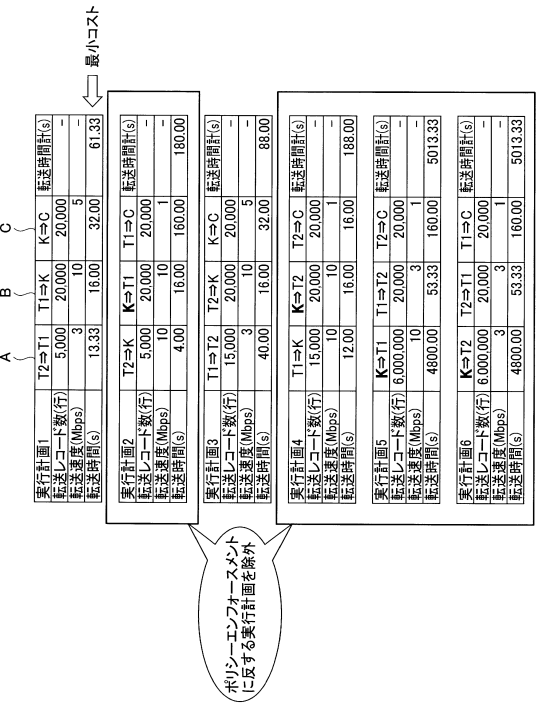
【図 1 3】



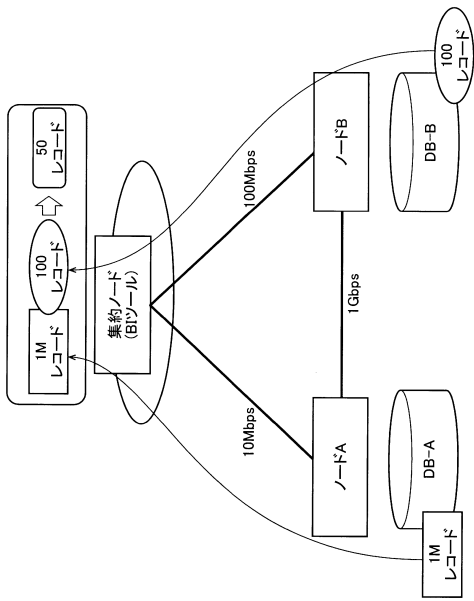
【図 1 4】



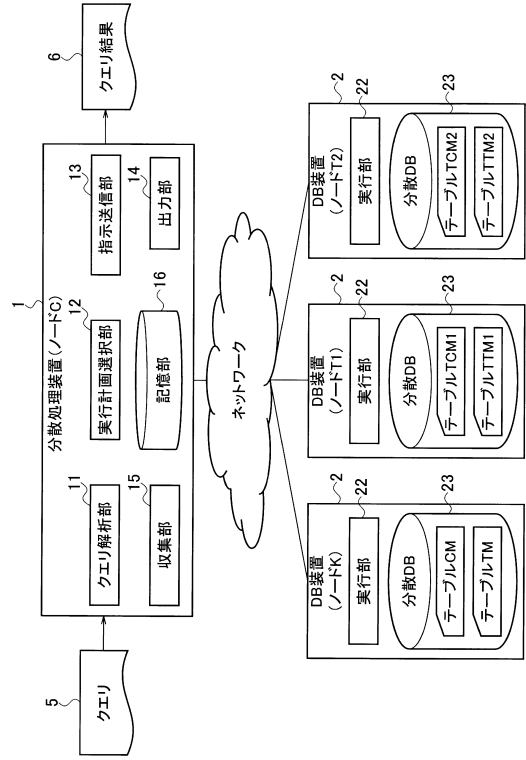
【図 15】



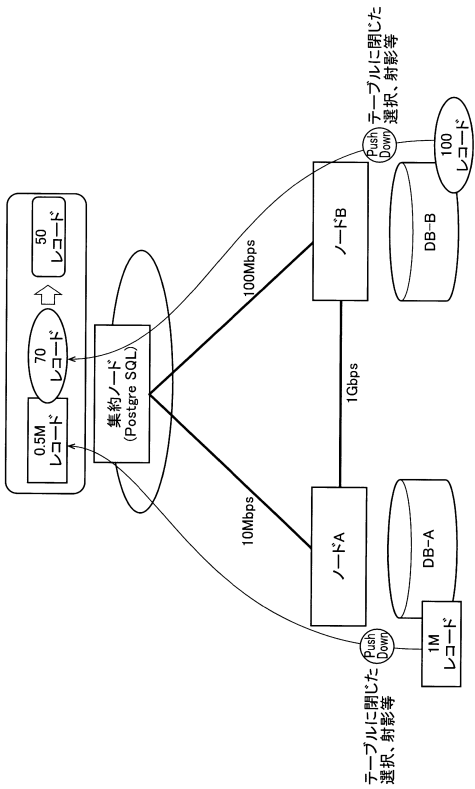
【図 17】



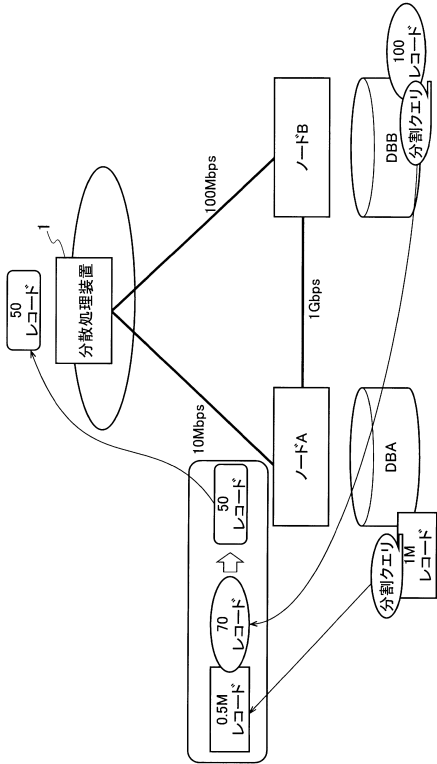
【図 16】



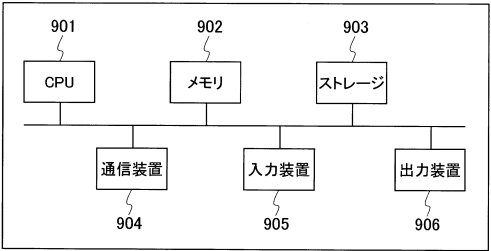
【図 18】



【図 19】



【図 20】



10

20

30

40

50

フロントページの続き

日本電信電話株式会社内

(72)発明者 柿本 真司

東京都千代田区大手町一丁目5番1号 日本電信電話株式会社内

審査官 原 秀人

(56)参考文献 国際公開第2017/138062 (WO, A1)

特開2016-095561 (JP, A)

米国特許出願公開第2015/0234895 (US, A1)

油井 誠 外, Remote Proxyを利用した並列分散XML問合せ処理手法の提案,
電子情報通信学会技術研究報告, 日本, 社団法人電子情報通信学会, 2007年06月25日, V
ol. 107 No. 131, pp. 217-222

(58)調査した分野 (Int.Cl., DB名)

G06F 16/00-16/958