

(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
2021年3月4日(04.03.2021)



(10) 国際公開番号

WO 2021/038795 A1

(51) 国際特許分類:

G06F 16/2458 (2019.01)

(21) 国際出願番号:

PCT/JP2019/033914

(22) 国際出願日:

2019年8月29日(29.08.2019)

(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: 日本電信電話株式会社 (NIPPON TELEGRAPH AND TELEPHONE CORPORATION) [JP/JP]; 〒1008116 東京都千代田区大手町一丁目5番1号 Tokyo (JP).

(72) 発明者: 岩越 さやか (IWAKOSHI Sayaka); 〒1808585 東京都武蔵野市緑町3丁目9-

1 1 N T T 知的財産センタ内 Tokyo (JP). 持田 誠一郎(MOCHIDA Seiichiro); 〒1808585 東京都武蔵野市緑町3丁目9-1 1 N T T 知的財産センタ内 Tokyo (JP). 山本 直人(YAMAMOTO Naoto); 〒1808585 東京都武蔵野市緑町3丁目9-1 1 N T T 知的財産センタ内 Tokyo (JP). 柿本 真司(KAKIMOTO Shinji); 〒1808585 東京都武蔵野市緑町3丁目9-1 1 N T T 知的財産センタ内 Tokyo (JP).

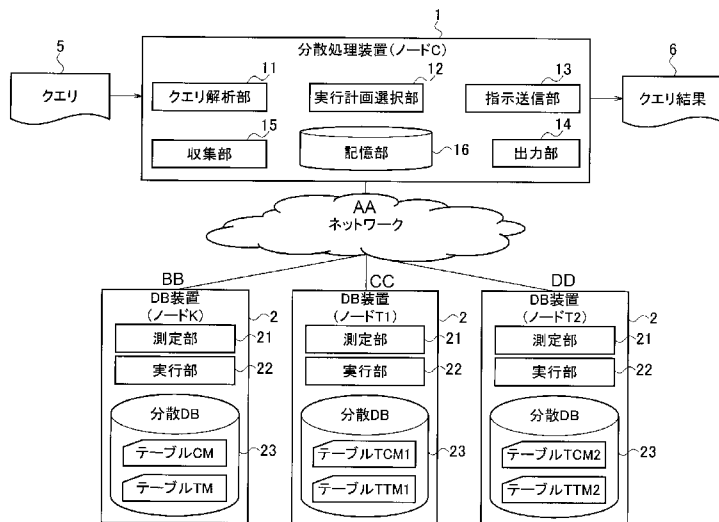
(74) 代理人: 三好 秀和, 外(MIYOSHI Hidekazu et al.); 〒1050001 東京都港区虎ノ門1丁目2番8号 虎ノ門琴平タワー Tokyo (JP).

(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,

(54) Title: DATABASE SYSTEM, DISTRIBUTED PROCESSING DEVICE, DATABASE DEVICE, DISTRIBUTED PROCESSING METHOD, AND DISTRIBUTED PROCESSING PROGRAM

(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

[図1]



1... DISTRIBUTED PROCESSING DEVICE (NODE C)
5... QUERY
6... QUERY RESULT
11... QUERY ANALYSIS UNIT
12... EXECUTION PLAN SELECTION UNIT
13... INSTRUCTION TRANSMISSION UNIT
14... OUTPUT UNIT
15... COLLECTION UNIT
16... STORAGE UNIT
21... MEASUREMENT UNIT
22... EXECUTION UNIT
23... DISTRIBUTED DATABASE
CM, TCM1, TCM2... TABLE
TM, TTM1, TTM2... TABLE
AA... NETWORK
BB... DATABASE DEVICE (NODE K)
CC... DATABASE DEVICE (NODE T1)
DD... DATABASE DEVICE (NODE T2)

(57) Abstract: A distributed processing device 1 is provided with: a selection unit 12 which lists execution plans of a query associated with a plurality of database devices 2, and selects one of the execution plans on the basis of the data transfer time associated with each execution plan; a transmission unit 13 which divides the query into divided queries in accordance with the selected execution plan, and transmits, to each database device 2, an instruction that includes a respective one of the divided queries and a forwarding destination for the execution result of the divided query; and an output unit 14 which receives the query execution result from each database device 2, and outputs the same. Each database

BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, 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, ST, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, 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, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

添付公開書類:

- 一 国際調査報告 (条約第21条(3))

device 2 is provided with an execution unit 22 which executes the divided query included in the instruction received from the distributed processing device 1, and transmits the execution result either to another database device 2 that is the forwarding destination included in the instruction, or to the distributed processing device 1.

(57) 要約: 分散処理装置 1 は、複数のデータベース装置 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選択部 1 2 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、対応するデータベース装置 2 にそれぞれ送信する送信部 1 3 と、データベース装置 2 からクエリの実行結果を受信し、出力する出力部 1 4 とを備え、データベース装置 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データベース装置 2 または分散処理装置 1 に送信する実行部 2 2 を備える。

明 細 書

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

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] 非特許文献1：「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と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な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。

[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。

[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。

[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DB23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DB23には少なくとも1つのデータベースが格納される。

[0025] 図2から図4は、図1に示す各分散DB23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DB23に格納されたテーブルを示す。図2の分散DB23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、6Mである。TMテーブルのレコード数は、60Mである。

[0026] 図3に、ノードT1のDB装置2の分散DB23に格納されたテーブルを示す。図3の分散DB23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。T

T M 2 テーブルは、テナント 2 のユーザ I D (T U i d) と、百貨店の C M テーブルのユーザ I D (U i d) とが関連付けて記憶されている。

[0028] (分散 D B システムの動作)

以下に、本実施形態の分散 D B システムの動作を説明する。ここでは、図 1 に示す 3 つの D B 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 D B システムの動作を示すフローチャートである。各 D B 装置 2 は、自身の D B 装置 2 と、他の D B 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 (S 1 1) 。そして、 D B 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 : b p s) を用いるが、これに限定されない。分散処理装置 1 は、各 D B 装置 2 からネットワーク性能情報を収集し、記憶部 1 6 に記憶する (S 1 2) 。

[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 T 1 との間のネットワーク帯域は 1 0 M b p s である。ノード K とノード T 2 との間のネットワーク帯域は 1 0 M b p s 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 5 M b p s である。

[0031] なお、 S 1 1 および S 1 2 の処理は、 S 1 3 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 1 6 にネットワーク性能情報が既に記憶されている場合は、 S 1 1 および S 1 2 は行われず、分散処理装置 1 は、記憶部 1 6 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 S 1 1 および S 1 2 を実行し、記憶部 1 6 に記憶されたネットワーク性能情報を更新してもよい。

[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 (S 1 3) 。

[0033] 図 7 は、複数の D B 装置 2 に関連するクエリの一例である。図示するクエ

りは、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）。

[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（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）。

[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。

[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。

[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。

[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 S 14 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。

[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。

[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。

[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。

[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。

[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

。

[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、CPU 901 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。

[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。

[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

- [0064] 1 : 分散処理装置（ノードC）
- 1 1 : クエリ解析部
- 1 2 : 実行計画選択部
- 1 3 : 指示送信部
- 1 4 : 出力部
- 1 5 : 収集部
- 1 6 : 記憶部
- 2 : DB装置（ノードK、T 1、T 2）
- 2 1 : 測定部
- 2 2 : 実行部
- 2 3 : 分散DB
- 5 : クエリ
- 6 : クエリ結果

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
- データベースシステム。
- [請求項2] 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] 前記分散処理装置の前記選択部は、前記実行計画の中から、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
- 請求項1または2に記載のデータベースシステム。

[請求項4]

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、を備える

分散処理装置。

[請求項5]

データベース装置であって、

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と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

データベース装置。

[請求項6]

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、

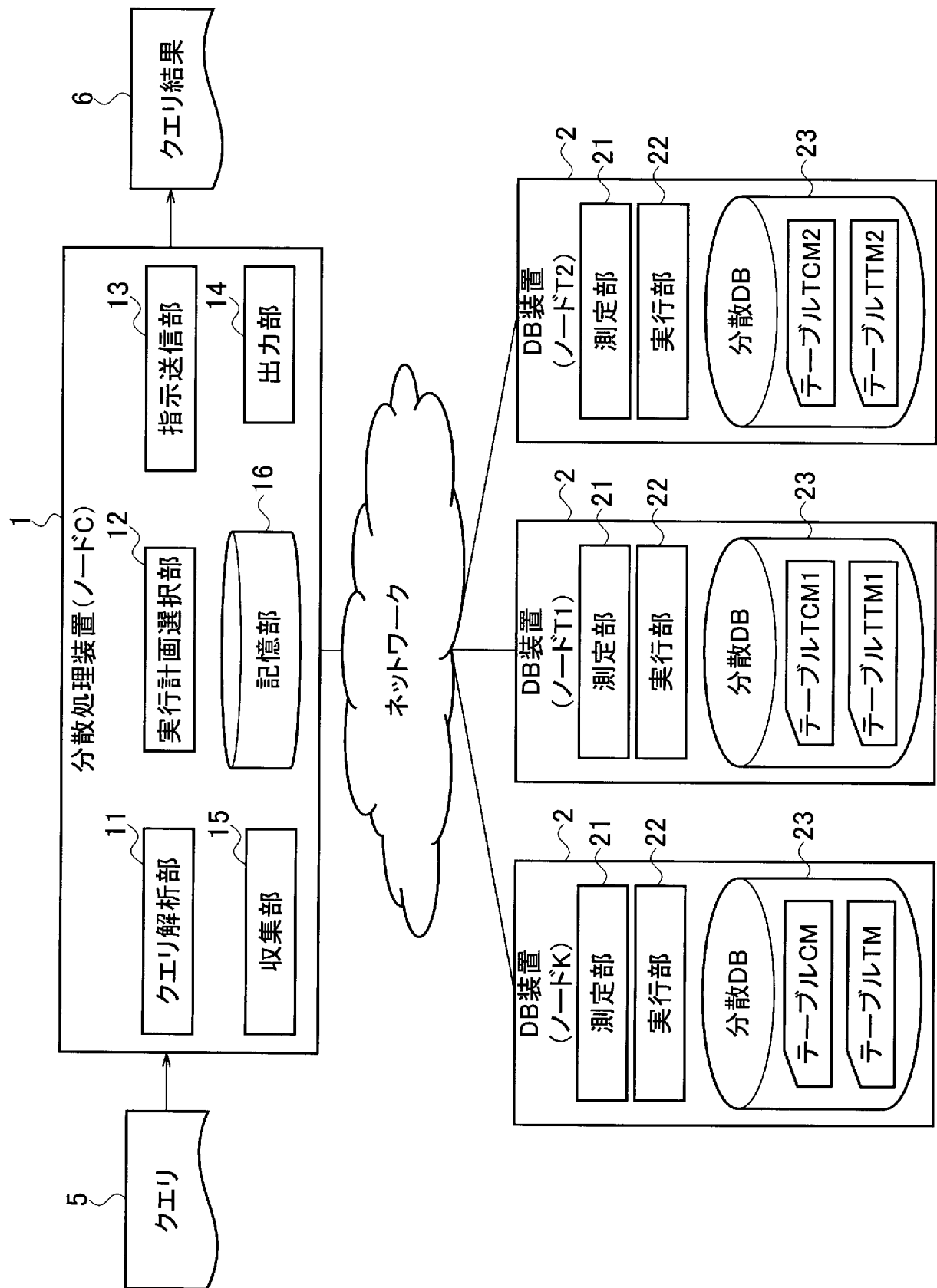
前記分散処理装置から受信した指示に含まれる分割クエリを実行し、実行結果を前記指示に含まれる転送先の他のデータベース装置または前記分散処理装置に送信する実行ステップを行う

分散処理方法。

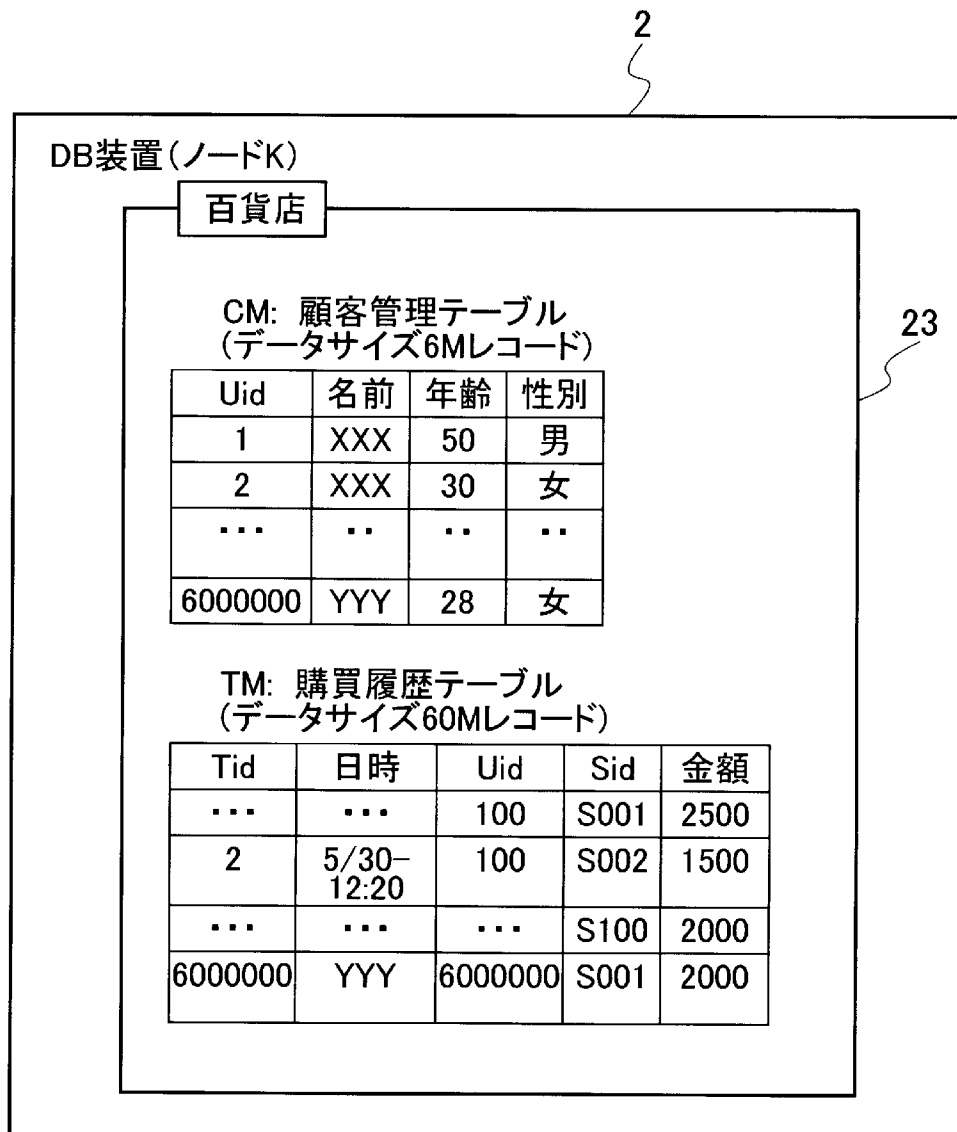
[請求項7] 請求項4項に記載の分散処理装置として、コンピュータを機能させる分散処理プログラム。

[請求項8] 請求項5項に記載の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。

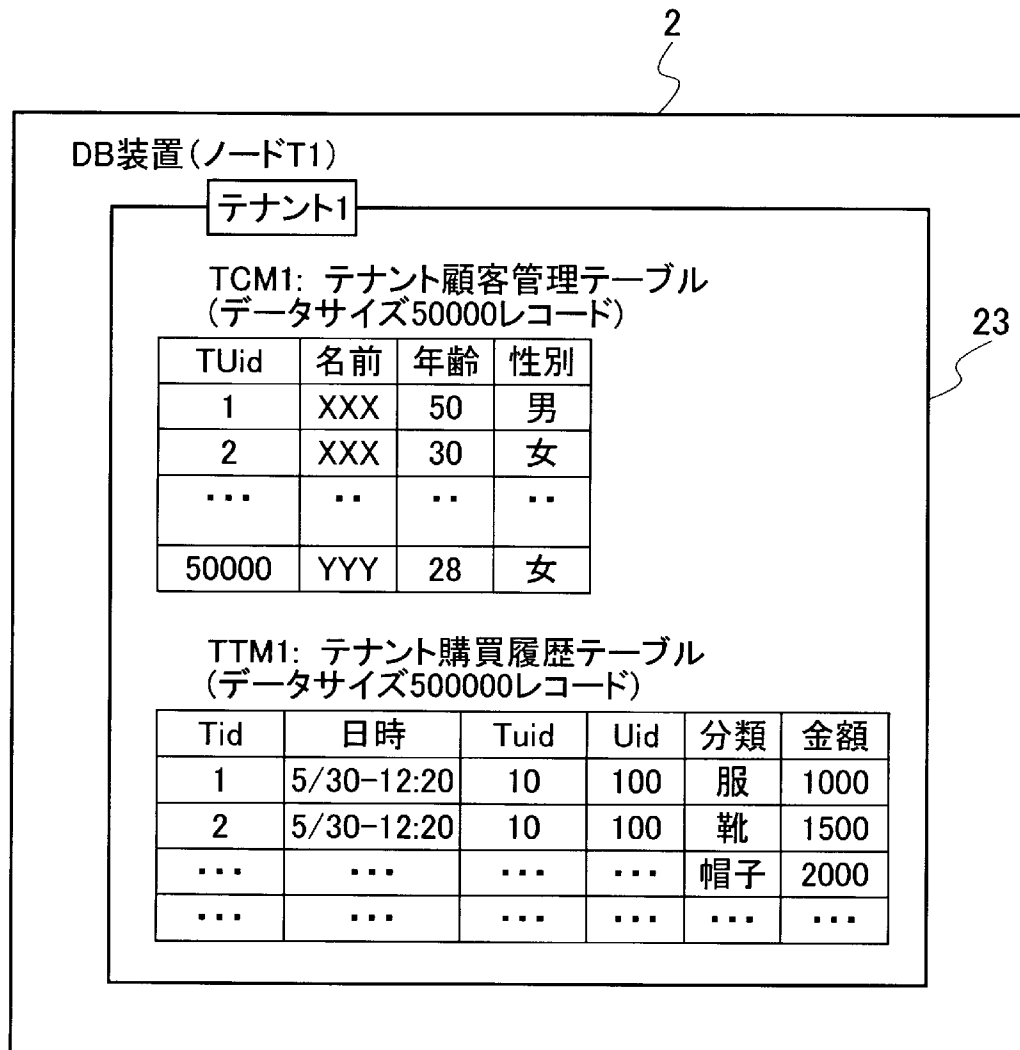
[図1]



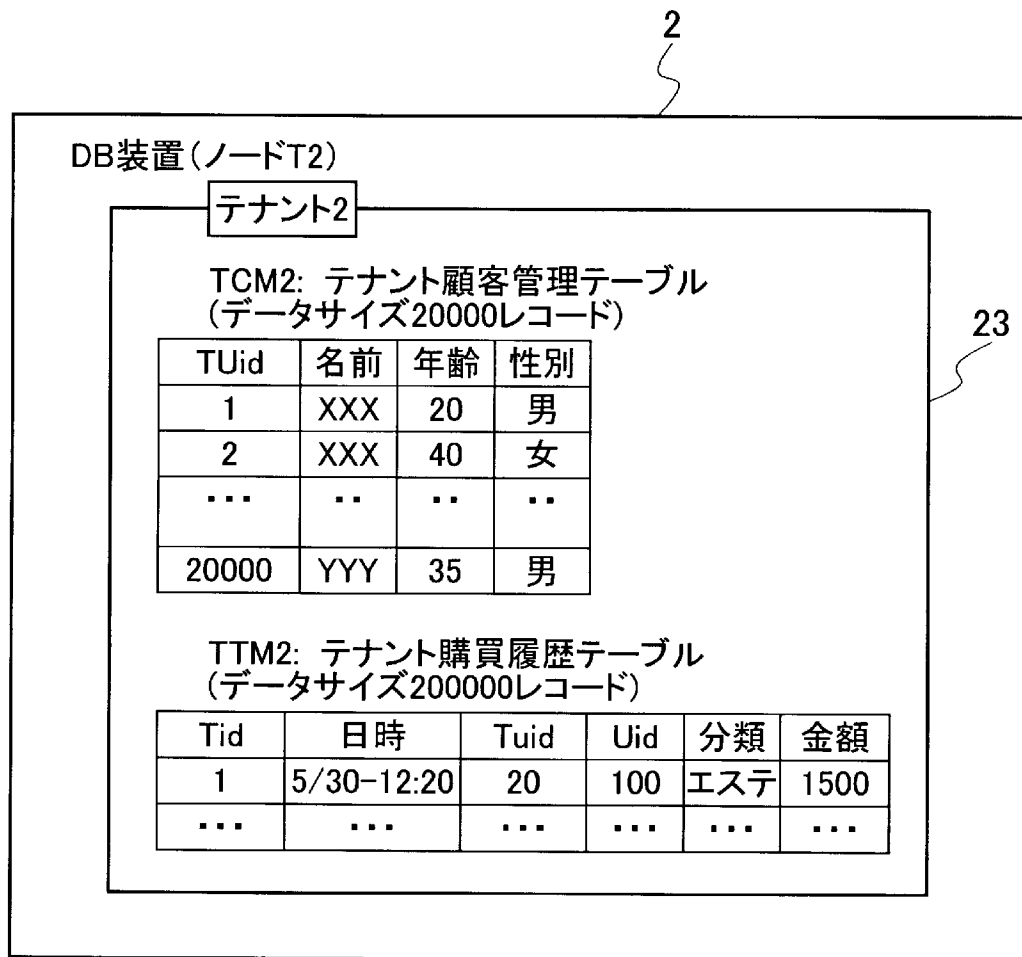
[図2]



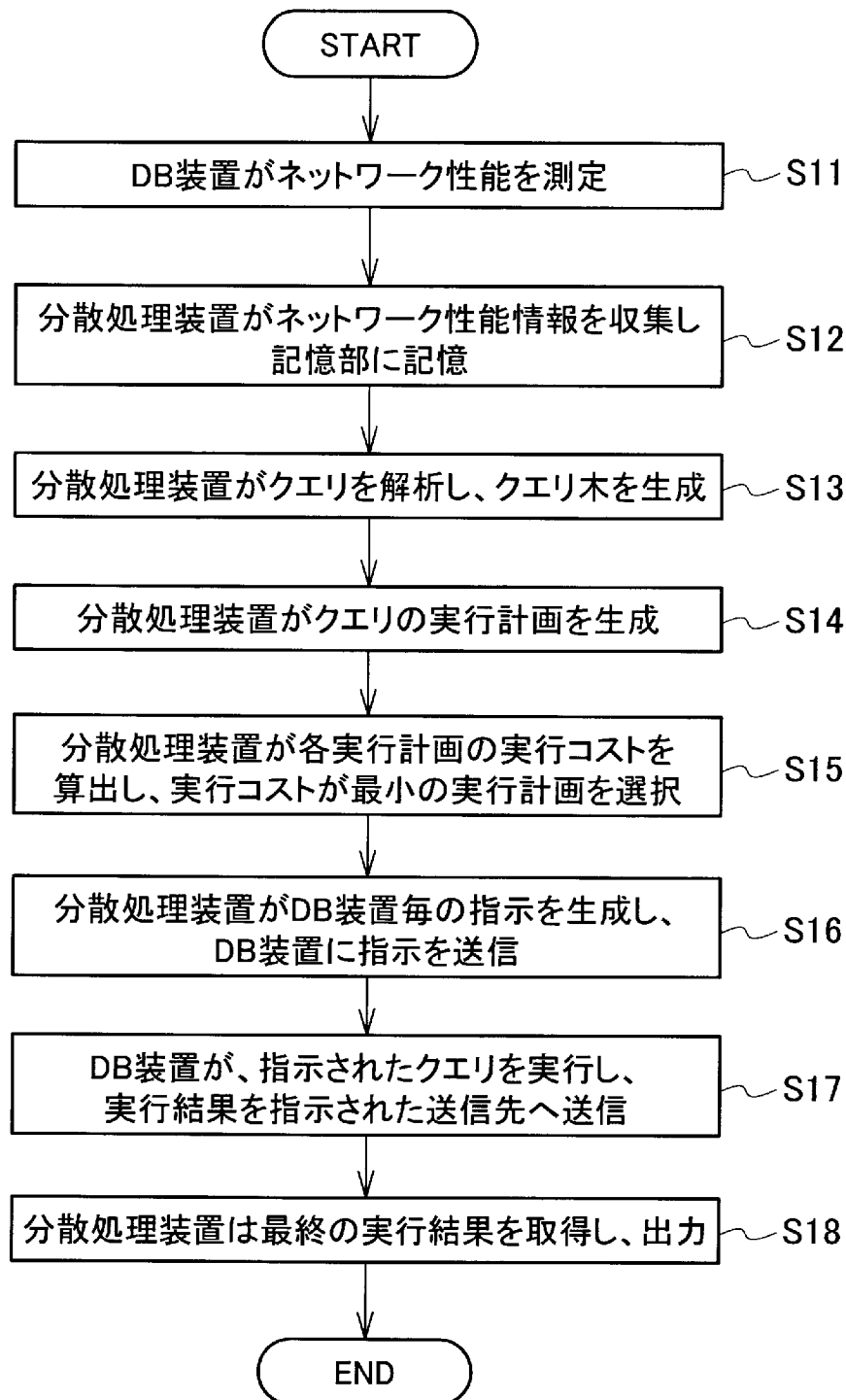
[図3]



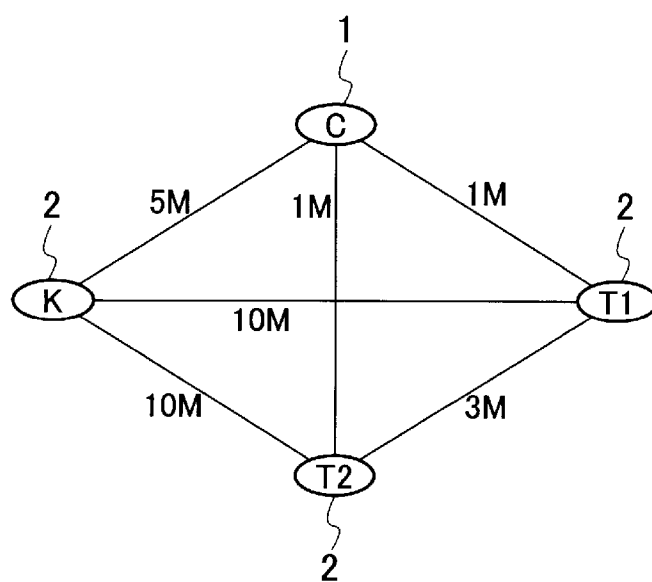
[図4]



[図5]



[図6]



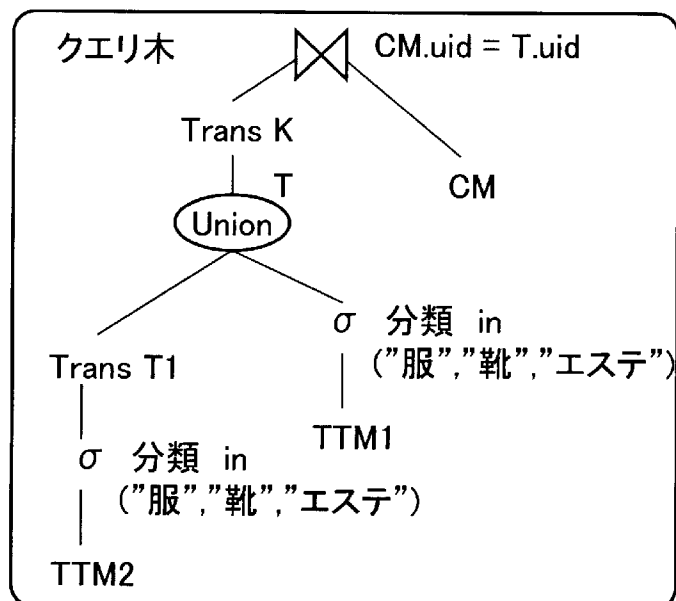
[図7]

```

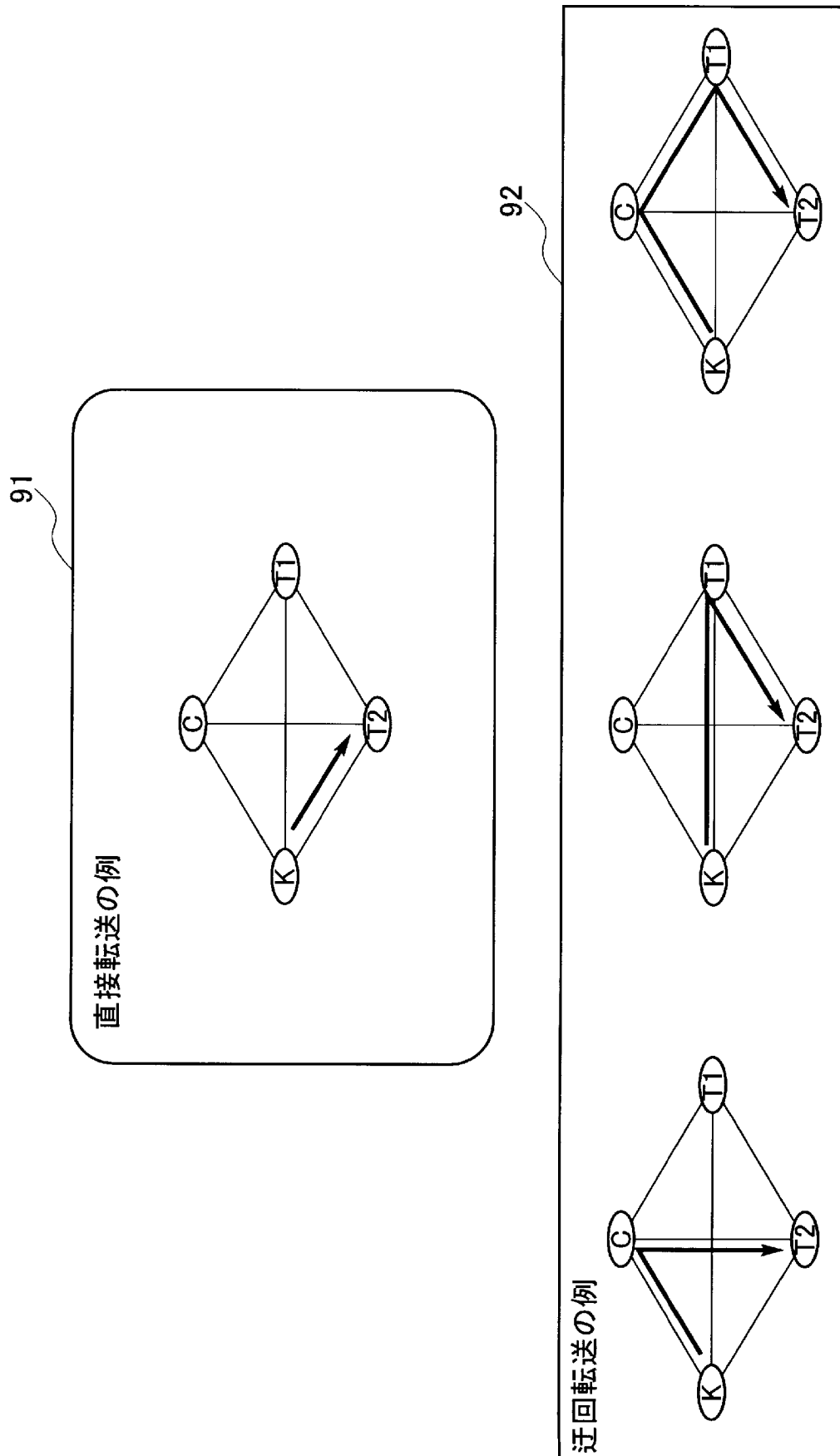
> select CM.uid, T.分類, T.金額 from CM,
      (select * from TTM1) union (select * from TTM2) as T
where
  CM.uid = T.uid and
  T.分類 in ("服", "靴", "エステ")

```

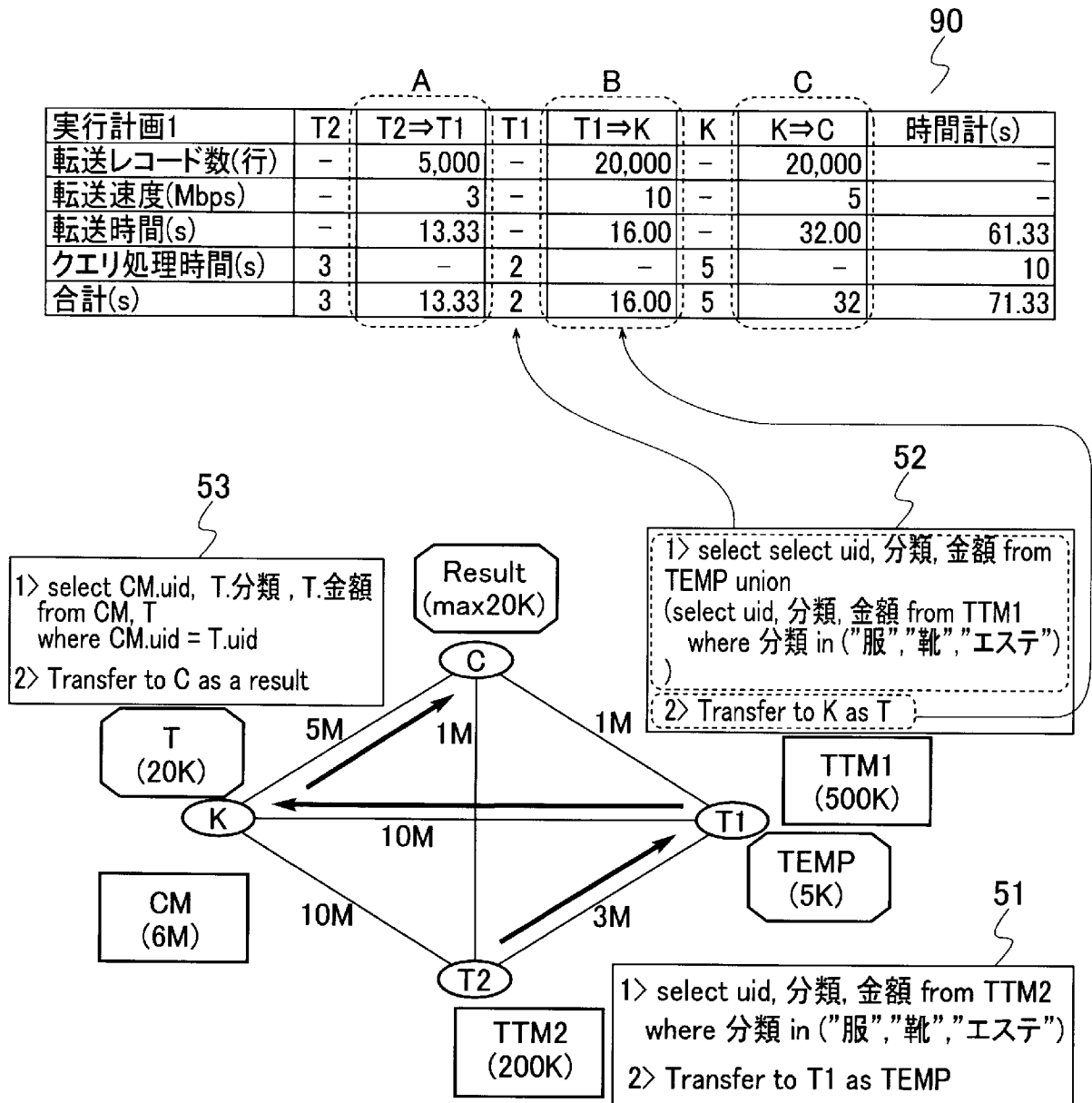
[図8]



[図9]



[図10]



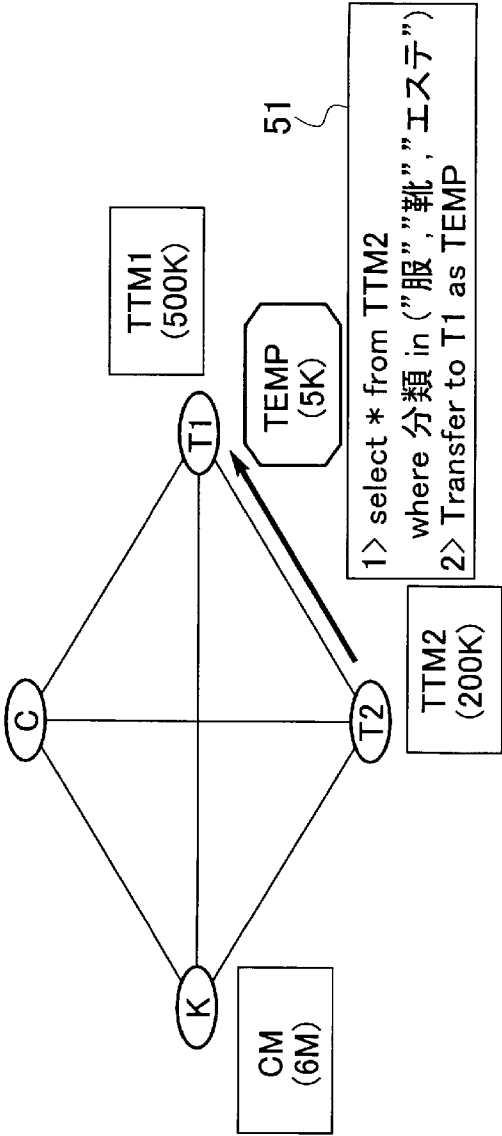
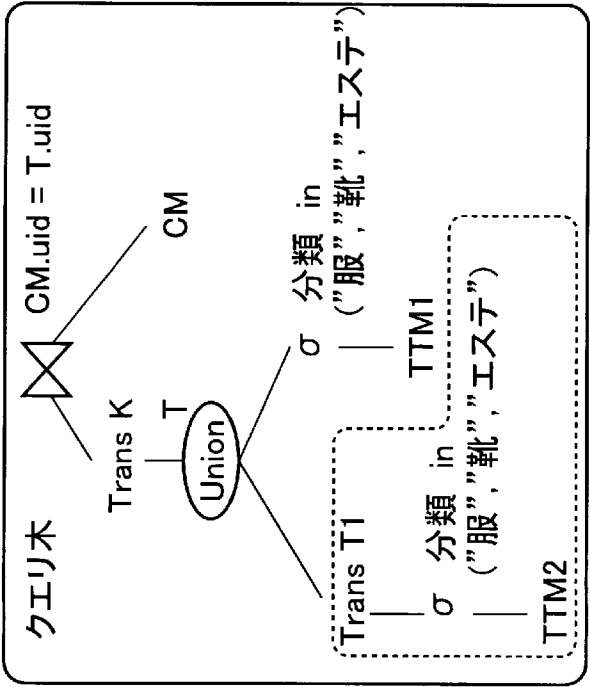
[図11]

	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	16.00	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

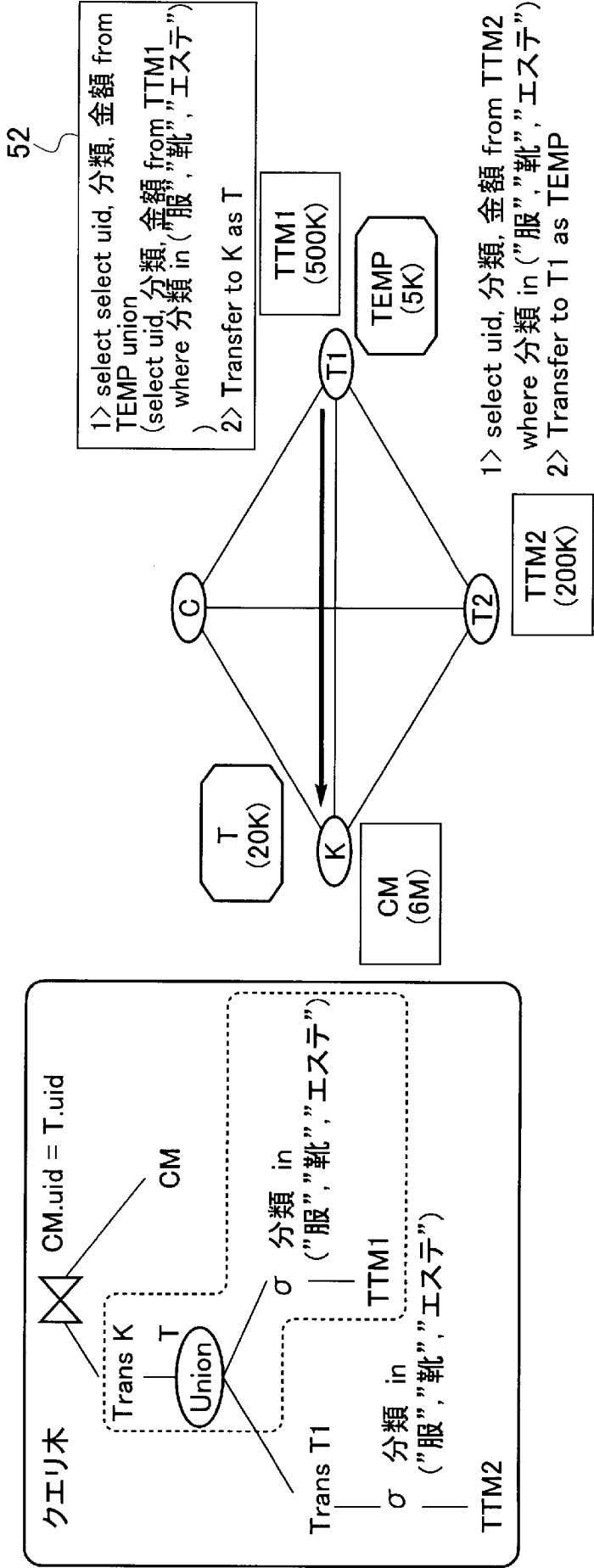
[図12]

実行計画1: [T2] → T1 → K → C



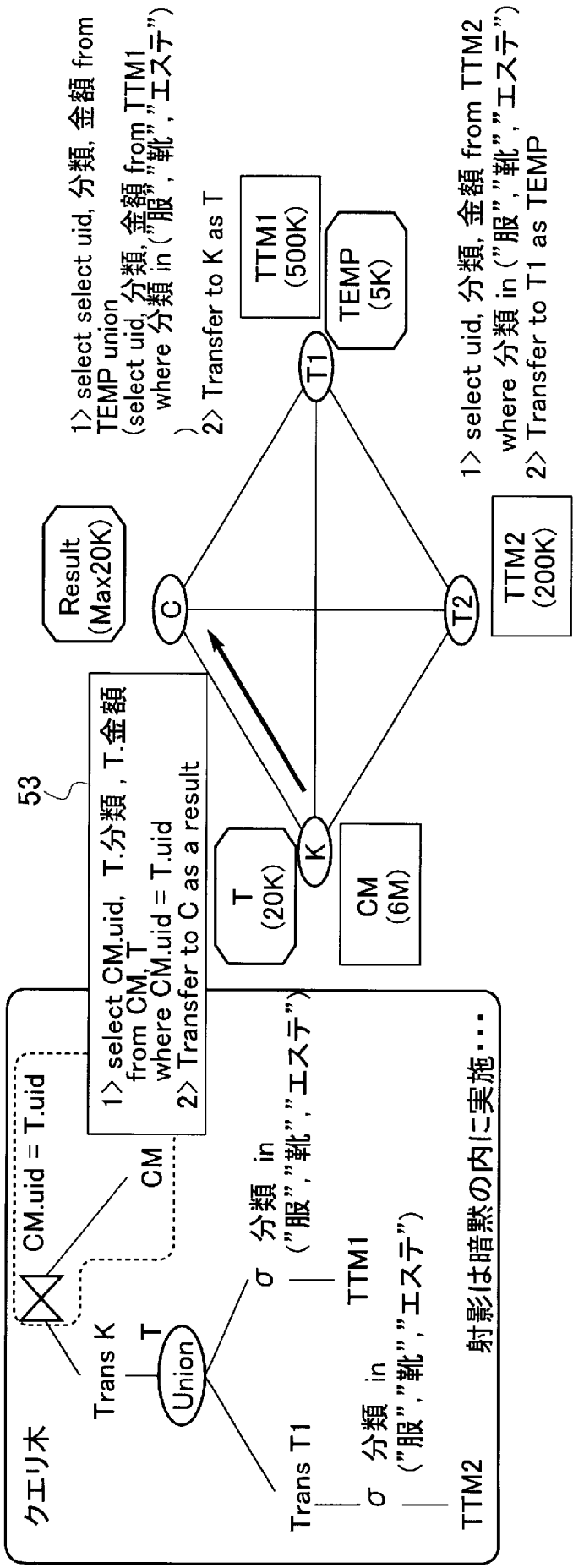
[図13]

実行計画1: T2→T1→K→C

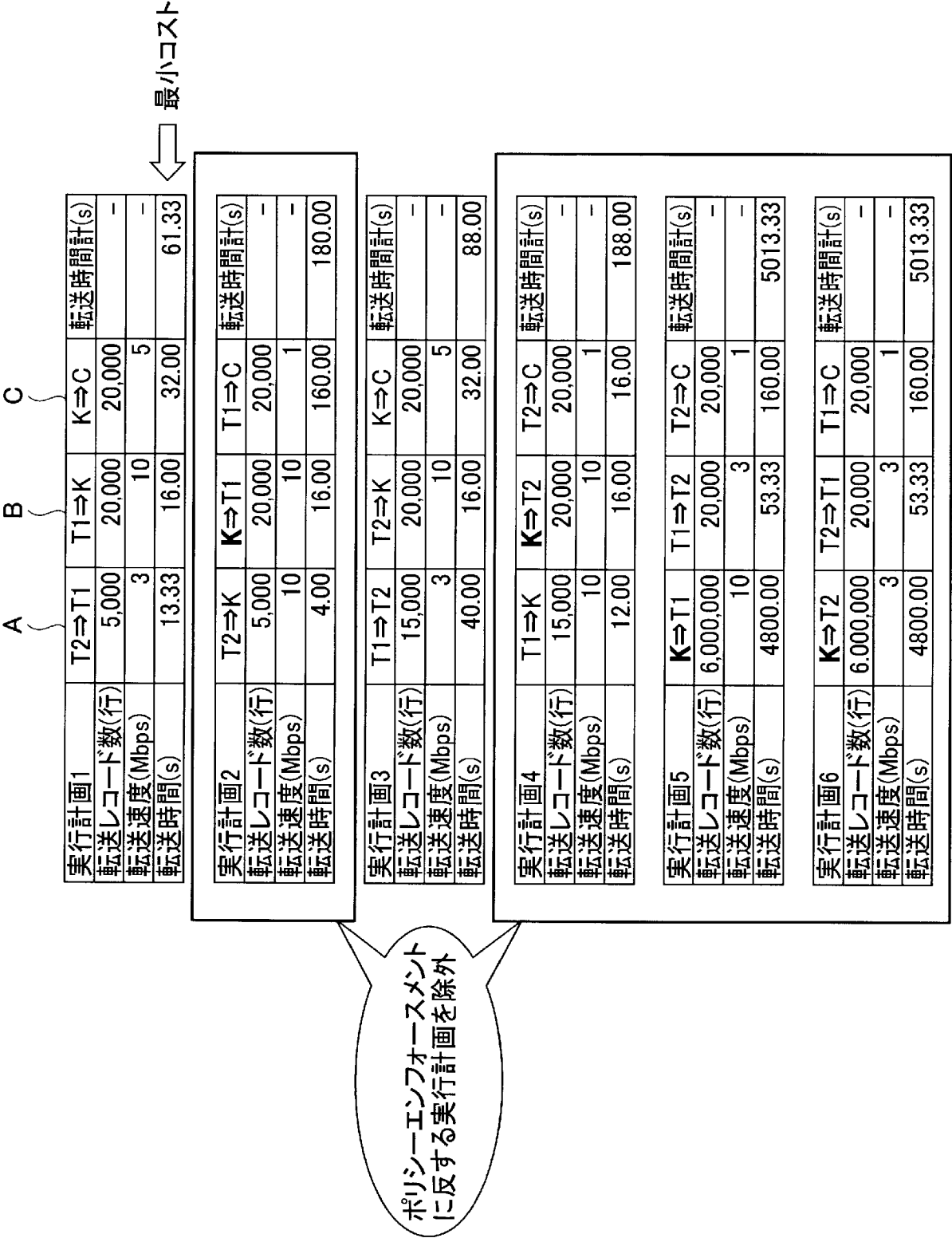


[図14]

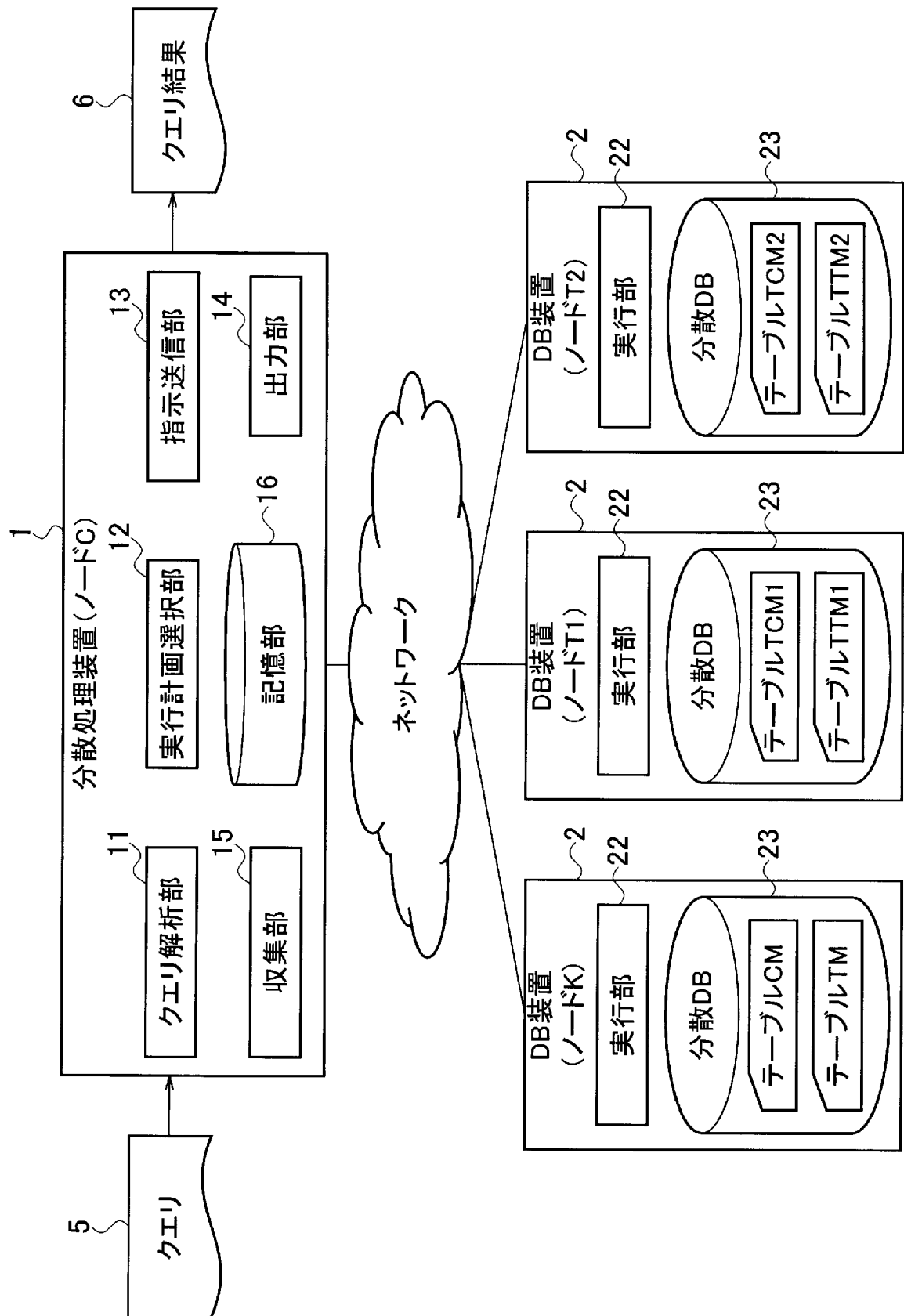
T2→T1→[K]→C



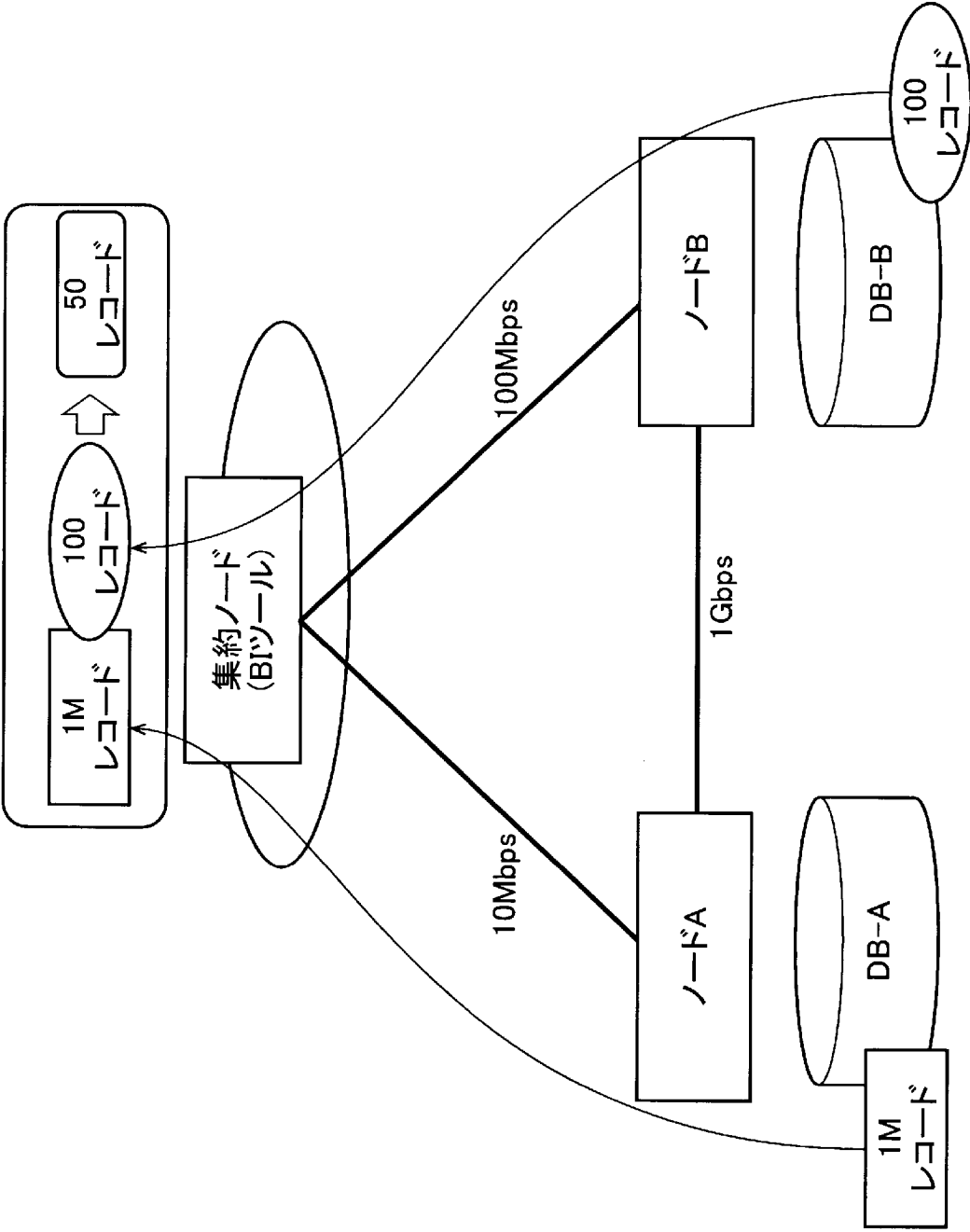
[図15]



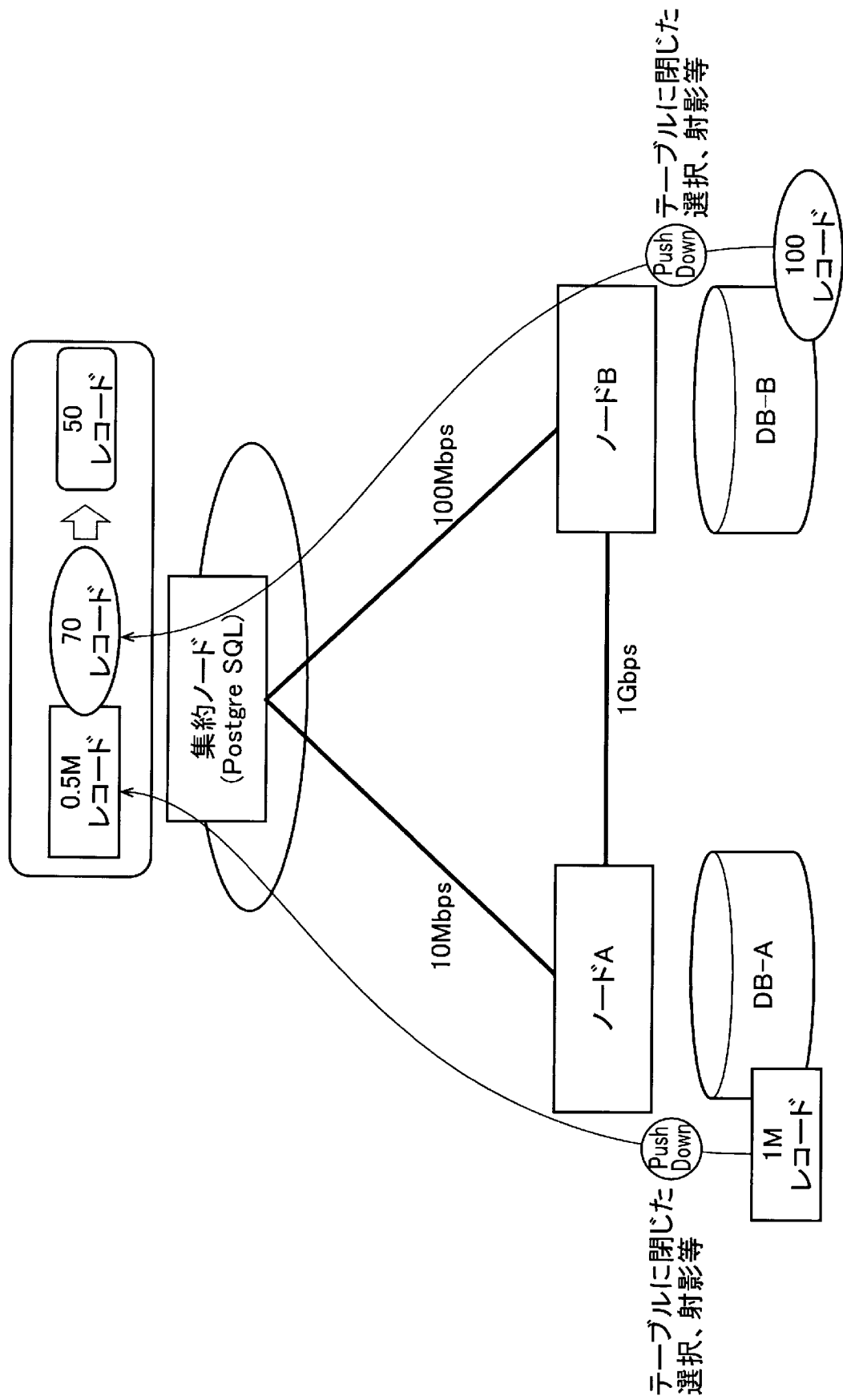
[図16]



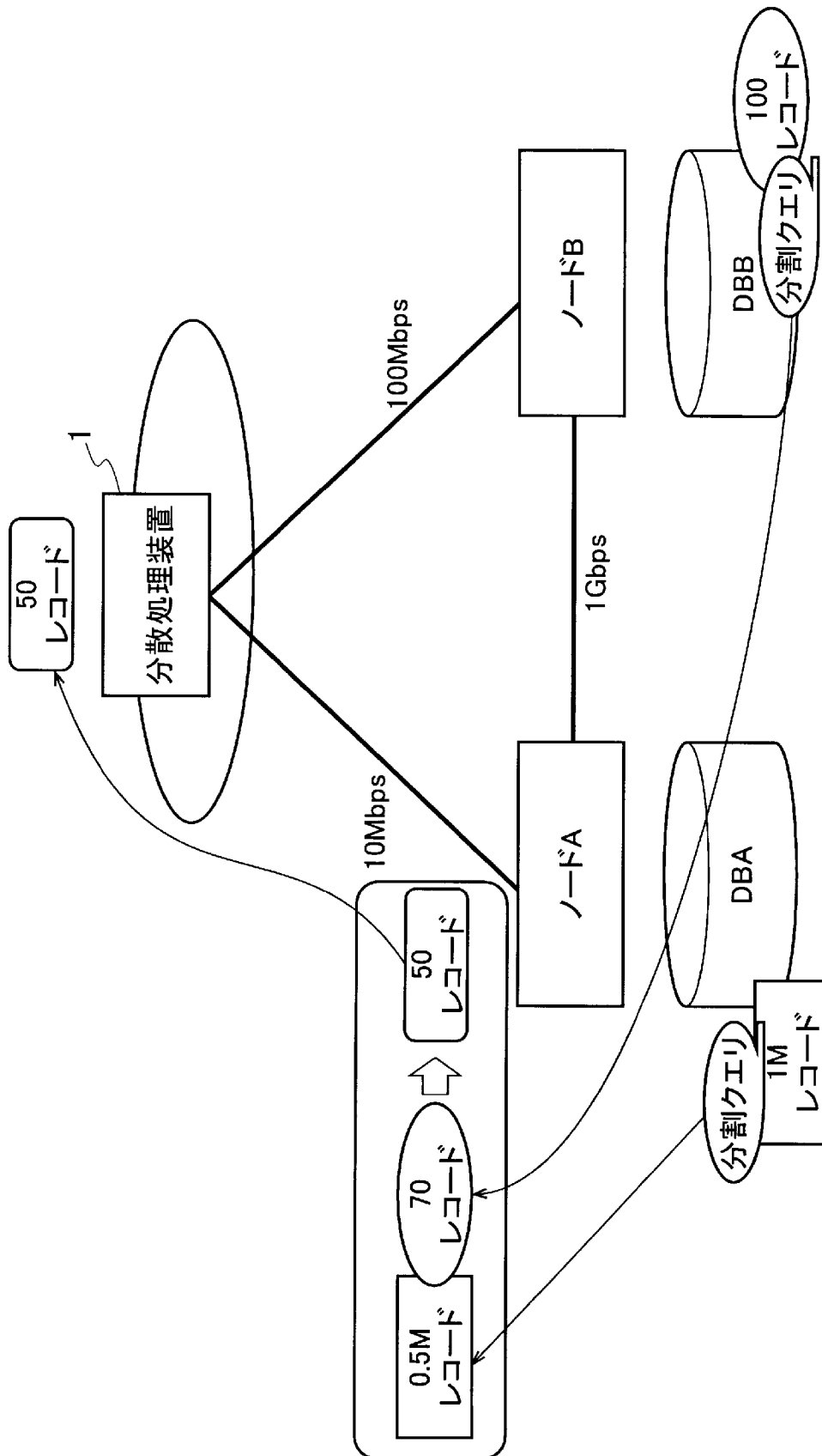
[図17]



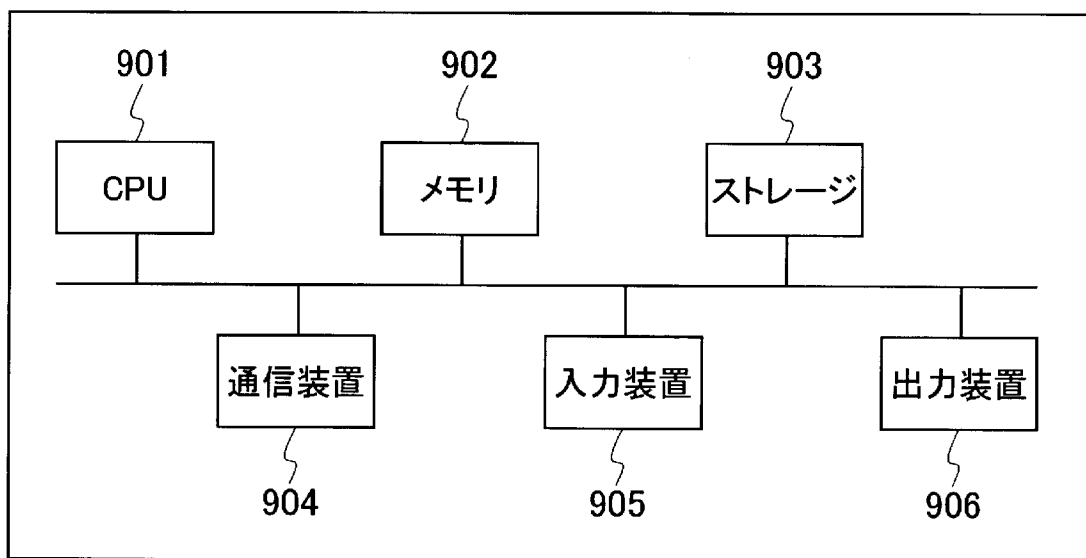
[図18]



[図19]



[図20]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2019/033914

A. CLASSIFICATION OF SUBJECT MATTER

Int. Cl. G06F16/2458 (2019.01) 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)

Int. Cl. G06F16/2458

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Published examined utility model applications of Japan 1922-1996

Published unexamined utility model applications of Japan 1971-2019

Registered utility model specifications of Japan 1996-2019

Published registered utility model applications of Japan 1994-2019

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 2017/138062 A1 (HITACHI, LTD.) 17 August 2017, paragraphs [0003], [0004], [0008], [0056], [0115], fig. 10 (Family: none)	1-8
A	JP 2016-95561 A (NEC CORP.) 26 May 2016, entire text, all drawings (Family: none)	1-8
A	US 2015/0234895 A1 (CITUS DATA BILGI ISLEMLERI TICARET A.S.) 20 August 2015, entire text, all drawings & WO 2015/126343 A1	1-8

☐ 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
22.11.2019

Date of mailing of the international search report
03.12.2019

Name and mailing address of the ISA/
Japan Patent Office
3-4-3, Kasumigaseki, Chiyoda-ku,
Tokyo 100-8915, Japan

Authorized officer

Telephone No.

A. 発明の属する分野の分類（国際特許分類（IPC））

Int.Cl. G06F16/2458(2019.01)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. G06F16/2458

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	1971-2019年
日本国実用新案登録公報	1996-2019年
日本国登録実用新案公報	1994-2019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）

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	WO 2017/138062 A1（株式会社日立製作所）2017.08.17, [0003]-[0004], [0008], [0056], [0115], 図10（ファミリーなし）	1-8
A	JP 2016-95561 A（日本電気株式会社）2016.05.26, 全文全図（ファミリーなし）	1-8
A	US 2015/0234895 A1（CITUS DATA BILGI ISLEMLERI TICARET A.S.） 2015.08.20, 全文全図 & WO 2015/126343 A1	1-8

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

22.11.2019

国際調査報告の発送日

03.12.2019

国際調査機関の名称及びあて先

日本国特許庁（ISA/J P）

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）

原 秀人

5 N

9644

電話番号 03-3581-1101 内線 3586