(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第3712791号
(P3712791)
(45)発行日　平成17年11月2日(2005.11.2)　(24)登録日　平成17年8月26日(2005.8.26)

(51)Int.Cl.[7]　F I
G06F 15/16　G06F 15/16　620F
G06F 12/00　G06F 12/00　512

請求項の数 2　(全 7 頁)

(21)出願番号　特願平8-154097
(22)出願日　平成8年6月14日(1996.6.14)
(65)公開番号　特開平10-3458
(43)公開日　平成10年1月6日(1998.1.6)
審査請求日　平成11年9月7日(1999.9.7)
審判番号　不服2002-17073(P2002-17073/J1)
審判請求日　平成14年9月5日(2002.9.5)

(73)特許権者　000005108
株式会社日立製作所
東京都千代田区丸の内一丁目６番６号
(74)代理人　100068504
弁理士　小川　勝男
(72)発明者　野元　純一
神奈川県横浜市戸塚区戸塚町５０３０番地　株式会社日立製作所ソフトウェア開発本部内
(72)発明者　木村　耕治
神奈川県横浜市戸塚区戸塚町５０３０番地　株式会社日立製作所ソフトウェア開発本部内

最終頁に続く

(54)【発明の名称】データベース管理方法及びその情報処理装置

(57)【特許請求の範囲】

【請求項１】

第１のデータベースを含む第１の計算機と第２のデータベースを含む第２の計算機とを有するデータベース管理システムにおけるデータベース管理方法において、

前記第１のデータベースと前記第２のデータベースとの結合要求を入力したとき、前記結合要求に基づいて結合キーについて前記第１のデータベースと前記第２のデータベースとを検索し、前記第１のデータベースでヒットした検索行が前記第２のデータベースでヒットした検索行よりも多い場合は前記結合要求の処理を第１の計算機に割り当て、前記第２のデータベースでヒットした検索行が前記第１のデータベースでヒットした検索行よりも多い場合は前記結合要求の処理を第２の計算機に割り当て、前記第１のデータベースでヒットした検索行が前記第２のデータベースでヒットした検索行と同じ場合は未処理件数に応じて第１の計算機と第２の計算機のいずれかを選択することを特徴とするデータベース管理方法。

【請求項２】

第１のデータベースのデータベース管理を行う第１の情報処理装置と第２のデータベースのデータベース管理を行う第２の情報処理装置とを有する並列計算機システムにおいて、

前記第１の情報処理装置および前記第２の情報処理装置の各々は、前記結合要求に基づいて結合キーについて前記第１のデータベースと前記第２のデータベースのうち所有する方のデータベースを検索する手段と、当該情報処理装置における未処理分の処理件数を更

新維持する手段とを有し、前記第１のデータベースと前記第２のデータベースとの結合要求を受け付けた情報処理装置は、前記第１のデータベースでヒットした検索行が前記第２のデータベースでヒットした検索行よりも多い場合は前記結合要求の処理を第１の情報処理装置に割り当てる手段と、前記第２のデータベースでヒットした検索行が前記第１のデータベースでヒットした検索行よりも多い場合は前記結合要求の処理を第２の情報処理装置に割り当てる手段と、前記第１のデータベースでヒットした検索行が前記第２のデータベースでヒットした検索行と同じ場合は未処理件数に応じて第１の情報処理装置と第２の情報処理装置のいずれかを選択する手段とを有することを特徴とする情報処理装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、データベース管理システムの負荷分散方法に係わり、特に複数の情報処理装置がネットワークを介して相互に接続されるシステムにおいて情報処理装置の負荷を分散させる方法に関する。

【０００２】

【従来の技術】

複数の処理装置がネットワークを介して相互に接続されるシステムにおいて、各処理装置のＣＰＵ利用率を計測しＣＰＵ利用率に応じて負荷を分散させる方法が行われている。すなわちこの方法は、新しいジョブや処理をいずれかの処理装置に割り当てる時点でＣＰＵ利用率が最も低い処理装置を選択することによって処理装置の負荷を分散させるものである。しかしこの方法は、一時点のＣＰＵ利用率に基づいて処理装置にかかっている負荷の大きさを推測するものであり、必ずしも各処理装置が実際に保持している負荷の大きさを反映しているとは言えない。

【０００３】

なおこの種の技術として関連するものには、例えば特開平４－１２９３９９号公報や特開平５－１５１１７７号公報などがある。

【０００４】

【発明が解決しようとする課題】

ＣＰＵ利用率は必ずしも処理装置が将来消化すべき負荷の大きさを反映するものではないので、ＣＰＵ利用率に従って負荷分散制御を行うと、正しく負荷分散ができず、１つの処理装置に処理が集中してしまう場合があった。

【０００５】

本発明の目的は、データベース管理システムにおいて未処理分の負荷の大きさに応じて負荷の分散制御を行い、上記問題を解決することにある。

【０００６】

【課題を解決するための手段】

本発明は、データベースと接続された複数の情報処理装置で構成される並列計算機システムにおけるデータベース管理方法であって、処理を割り当てる情報処理装置を特定するために、要求によって情報処理装置間で発生するデータ転送量を比較し、当該データ転送量が同等である場合は、未処理件数に応じて処理を割り当てる情報処理装置を決定するデータベース管理方法及びその情報処理装置を特徴とする。

【０００７】

【発明の実施の形態】

以下、本発明の一実施形態について図面を用いて説明する。

【０００８】

図１は、本実施形態の並列計算機システムの構成図である。並列計算機システムは、複数の処理装置１，２，３とこれらの処理装置を相互に接続するネットワーク１０から構成される。各処理装置１，２，３は、それぞれデータベース１４，２４，３４を接続する。これらのデータベースは、リレーショナル・データベースであり、データベース１４は表Ａと表Ｂ、データベース２４は表Ｃと表Ｄ、データベース３４は表Ｅから構成される。デー

タベース管理システム１１，２１，３１は、それぞれデータベース１４，２４，３４を管理するプログラムであり、それぞれ検索結果を結合するときの未処理分の累積処理行数を保持する処理行数カウンタ１２，２２，３２を有する。ＵＡＰ（ユーザ・アプリケーション・プログラム）１３，２３，３３は、それぞれデータベース管理システム１１，２１，３１へデータベースの検索要求を発行するプログラムである。データベース管理テーブル１５は、データベースを構成する表の名称とその表を所有する処理装置との対応を格納するテーブルであり、処理装置１に接続される記憶装置上に格納される。図示していないが、データベース管理テーブル１５は処理装置２及び３にもそれぞれ接続される。

【０００９】

ＵＡＰ１３から表ｘと表ｙを検索した後に表の結合を行うＳＱＬが発行されたとき、データベース管理システム１１はデータベース管理テーブル１５を参照して表ｘ及び表ｙを所有する処理装置の番号を求め、各表を有する処理装置のデータベース管理システムへ検索要求を発行する。各データベース管理システムはそれぞれ表ｘ及び表ｙを検索し、検索処理が終了したときヒットした検索行数x_n及びy_nをデータベース管理システム１１に報告する。データベース管理システム１１は、検索結果及び結合結果のネットワーク１０を介するデータ転送量を削減するように表ｘ又は表ｙを所有する処理装置のいずれか一方に表の結合処理要求を発行する。結合処理要求を受けた処理装置のデータベース管理システムは、他方の表を所有する処理装置のデータベース管理システムへ検索結果を要求し、検索結果を受けたとき処理行数カウンタ１２，２２又は３２に結合処理の処理行数を加算してから結合処理を行う。結合処理が終了したとき、処理行数カウンタから結合処理行数を減算した後、データベース管理システム１１へ結合結果を送信し、データベース管理システム１１がＵＡＰ１３に結合結果を渡す。表ｘ又は表ｙを所有する処理装置のいずれを選択しても処理装置間転送が同等になるとき、データベース管理システム１１は表ｘ及び表ｙを有する各処理装置の処理行数カウンタの処理行数を比較し、負荷の少ない処理装置のデータベース管理システムへ表の結合要求を発行する。

【００１０】

図２は、データベース管理テーブル１５のデータ形式を示す図である。データベース管理テーブル１５は、表の識別子とその表を有する処理装置の番号との対応を保持する。

【００１１】

図３は、ＵＡＰ１３から表ｘと表ｙを検索した後に表の結合を行うＳＱＬが発行された後のデータベース管理システム１１の処理の流れを示すフローチャートである。ただしステップ４１，４３及び４５は、ネットワーク１０を介する処理装置のデータ転送及び他のデータベース管理システムの処理を伴う場合がある。データベース管理システム１１は、データベース管理テーブル１５を参照して表ｘと表ｙを所有する処理装置の番号を求め、検索処理を実行する（ステップ４１）。すなわち表ｘと表ｙを所有する処理装置のデータベース管理システムへそれぞれトランザクションＩＤと表ｘ又は表ｙを指定して検索要求を送信する。処理装置１が表ｘと表ｙのいずれか一方又は両方を所有するのであれば、データベース管理システム１１中の検索処理プログラムへ検索要求を発行するだけである。検索処理が終了したとき、データベース管理システム１１は同一のトランザクションＩＤ、検索の結果ヒットした行数x_n，y_n及び処理行数カウンタの値を受け取る。次にデータベース管理システム１１は、表ｘと表ｙが同一処理装置であれば（ステップ４２ＹＥＳ）、表ｘのある処理装置による結合処理を行う（ステップ４３）。表ｘと表ｙを所有する処理装置が異なるとき（ステップ４２ＮＯ）、表ｘと表ｙのヒットした検索行数を比較する（ステップ４４）。表ｘの検索行数が大きければ（$x_n > y_n$）、表ｘのある処理装置による結合処理を行う（ステップ４３）。表ｙの検索行数が大きければ（$x_n < y_n$）、表ｙのある処理装置による結合処理を行う（ステップ４５）。表ｘの検索行数と表ｙの検索行数が等しいとき（$x_n = y_n$）、ＵＡＰ１３と表ｘが同一処理装置にあれば（ステップ４６ＹＥＳ）、表ｘのある処理装置による結合処理を行う（ステップ４３）。ＵＡＰ１３と表ｘのある処理装置が異なり（ステップ４６ＮＯ）、ＵＡＰ１３と表ｙが同一処理装置にあれば（ステップ４７ＹＥＳ）、表ｙのある処理装置による結合処理を行う（ステップ４５）。

ＵＡＰ１３，表ｘ及び表ｙが別々の処理装置にあれば（ステップ４７ＮＯ）、データベース管理システム１１は、表ｘを有する処理装置の処理行数ｐ（ｘ）$_m$と表ｙを有する処理装置の処理行数ｐ（ｙ）$_m$とを比較する（ステップ４８）。処理行数ｐ（ｘ）$_m$の方が小さいか等しければ（ｐ（ｘ）$_m$≦ｐ（ｙ）$_m$）、表ｘのある処理装置による結合処理を行う（ステップ４３）。処理行数ｐ（ｙ）$_m$の方が小さければ（ｐ（ｘ）$_m$＞ｐ（ｙ）$_m$）、表ｙのある処理装置による結合処理を行う（ステップ４５）。

【００１２】

図４は、ステップ４３及びステップ４５の結合処理をさらに展開して示すフローチャートである。ＵＡＰ１３のある処理装置１のデータベース管理システム１１は、一方の表を有し結合処理を行う処理装置へ検索時のトランザクションＩＤと表ｘ，表ｙを指定して結合処理要求を送信する（ステップ５１）。結合処理を行う処理装置は、この結合処理要求を受信したとき（ステップ５２）、他方の表を有する処理装置へ検索時のトランザクションＩＤを指定して検索結果要求を送信する（ステップ５３）。他方の表を有する処理装置は、この検索結果要求を受信すると（ステップ５４）、結合処理を行う処理装置へトランザクションＩＤを指定して検索結果とヒットした検索行数を送信する（ステップ５５）。結合処理を行う処理装置は、この検索結果を受信すると（ステップ５６）、処理行数カウンタに結合処理の処理行数を加算することによって処理行数カウンタを更新し（ステップ５７）、表ｘと表ｙの検索結果について結合処理を行う（ステップ５８）。結合処理が終了すると、処理行数カウンタから結合処理の処理行数を減算することによって処理行数カウンタを更新する（ステップ５９）。次に結合処理を行う処理装置は、処理装置１へ結合処理要求時のトランザクションＩＤを指定して結合結果を送信する（ステップ６０）。データベース管理システム１１はこの結合結果を受信し（ステップ６１）、ＵＡＰ１３へ渡す（ステップ６２）。ＵＡＰ１３のある処理装置１と結合処理を行う処理装置とが同一処理装置であれば、同一処理装置の処理はデータベース管理システム１１内の処理となり、ネットワーク１０を介する結合処理要求と結合結果の転送はない。また結合処理を行う処理装置と他方の表を有する処理装置とが同一処理装置であれば、同一処理装置の処理は同一データベース管理システム内の処理となり、ネットワーク１０を介する検索結果要求と検索結果の転送はない。なお図４の例では結合処理の処理行数をヒットした検索行数x_nとy_nの和としたが、他にx_nとy_nの積のようにx_nとy_nの関数で表現できるものであればよい。

【００１３】

上記実施形態によれば、検索処理は検索対象の表を管理するデータベース管理システムのみが実行できるから、データベース管理システム１１はそれぞれの表を管理する処理装置へ検索要求を発行する。表ｘと表ｙが同一処理装置にあれば（ステップ４２ＹＥＳ）、検索結果の送信のためのネットワーク１０の負荷がかからないから、表ｘ及び表ｙのある同一処理装置で結合処理を行えばよい。表ｘと表ｙを所有する処理装置が異なるとき（ステップ４２ＮＯ）、ヒットした検索行数x_nとy_nのうち検索行数の多い方の処理装置で結合処理を行うので、検索結果の転送のためにネットワーク１０にかかる負荷が少ない。検索行数x_n，y_nが同じ場合（x_n＝y_n）には、ＵＡＰと一方の表が同一処理装置にあるような処理装置で結合処理を行うので、結合結果の転送のためにネットワーク１０の負荷がかからない。なお一般に検索結果のデータ量に比べて結合結果のデータ量が小さいとみなしている。検索行数x_n，y_nが等しく、ＵＡＰ，表ｘ，表ｙがそれぞれ別の処理装置にある場合（ステップ４７ＮＯ）には、表ｘのある処理装置で結合処理を行っても表ｙのある処理装置で結合処理を行ってもネットワーク１０の負荷は同じであるので、データベース管理システム１１は、表ｘを有する処理装置の処理行数カウンタの内容と表ｙを有する処理装置の処理行数カウンタの内容を比較して、いずれか累積負荷の小さい方の処理装置へ結合処理要求を発行する。結合処理を行う処理装置の負荷は、結合処理の処理行数に応じて増加するものとみなし得る。

【００１４】

なお上記実施形態では、ＵＡＰはいずれかの処理装置内に所在したが、ＵＡＰが処理装置

１，２，３とは別にネットワーク１０に接続されたクライアント装置内にあってもよい。このとき処理装置１，２及び３は、データベースを管理するサーバとして機能する。クライアントからネットワーク１０を介して処理装置１，２又は３へＳＱＬが発行されると、これを受けた処理装置のデータベース管理システムは、関連する処理装置と共同して図３及び図４に示す上記処理を行った後、ネットワーク１０を介して結合結果をクライアントへ送信する。

上記実施形態によれば、結合処理の累積処理行数を常時更新することによって各処理装置が保持する未処理分の負荷の大きさを予測できるため、この処理行数に基づいて結合処理の処理装置への割当をすれば、従来問題であった負荷の集中を防ぐことができる。

【００１５】

【発明の効果】

本発明によれば、処理件数を常時更新することによって各情報処理装置が保持する未処理分の負荷の大きさを予測できるため、この処理件数に基づいて新しい処理の情報処理装置への割当をすれば、従来問題であった負荷の集中を防ぐことができる。

【図面の簡単な説明】

【図１】実施形態の並列計算機システムの構成図である。

【図２】実施形態のデータベース管理テーブル１５のデータ形式を示す図である。

【図３】実施形態の結合処理の負荷を分散させる処理の流れを示すフローチャートである。

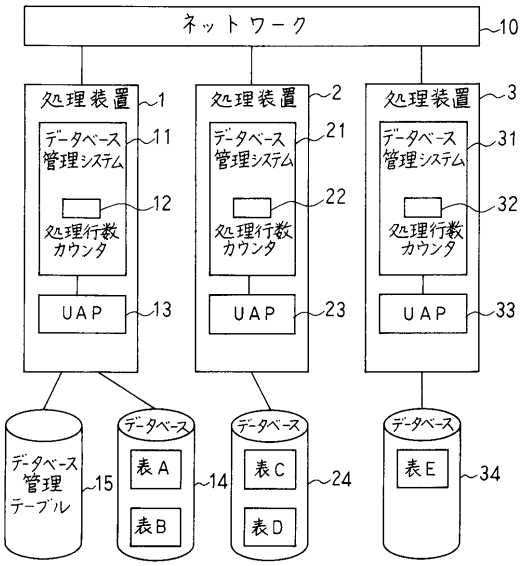
【図４】図３の結合処理を展開して示すフローチャートである。

【符号の説明】

１，２，３：処理装置、１１，２１，３１：データベース管理システム、１２，２２，３２：処理行数カウンタ、１４，２４，３４：データベース

【図１】

図1



【図２】

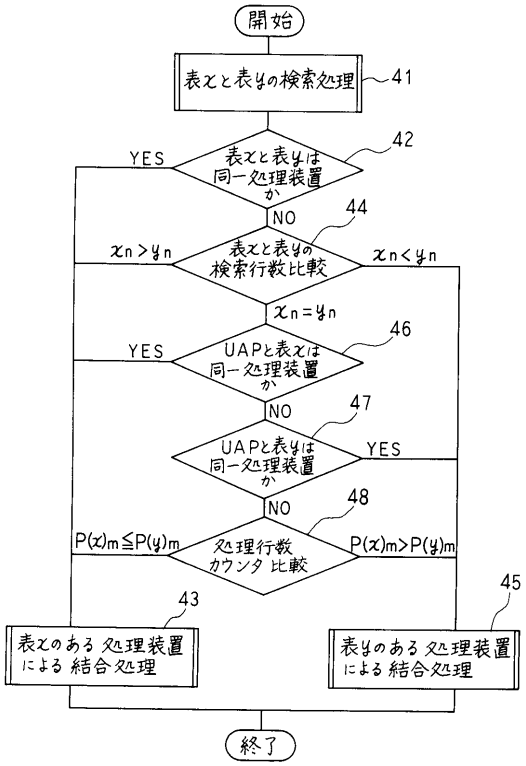
図２

15

表名	処理装置番号
表A	1
表B	1
表C	2
表D	2
表E	3
・・	・

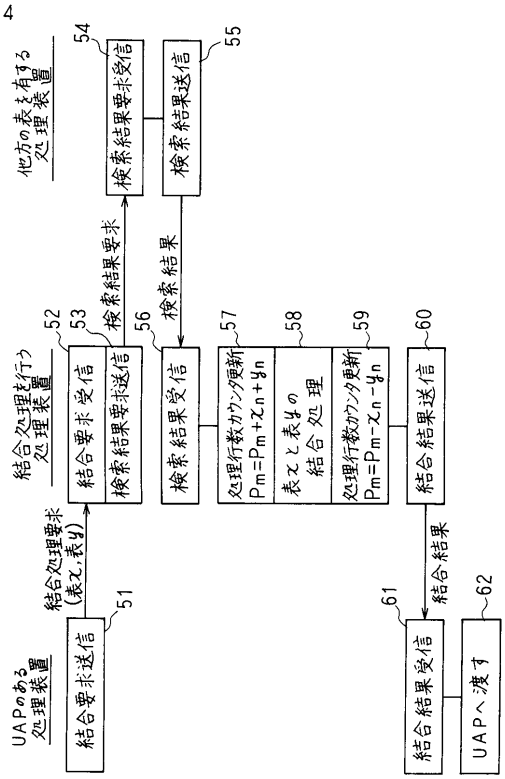
【図３】

図3


開始
表xと表yの検索処理
41
表xと表yは同一処理装置か
42
YES
NO
表xと表yの検索行数比較
44
xn>yn
xn<yn
xn=yn
UAPと表xは同一処理装置か
46
YES
NO
UAPと表yは同一処理装置か
47
YES
NO
処理行数カウンタ比較
48
P(x)m≦P(y)m
P(x)m>P(y)m
表xのある処理装置による結合処理
43
表yのある処理装置による結合処理
45
終了

【図４】

図4


他方の表を有する処理装置
54
検索結果要求受信
55
検索結果送信
検索結果要求
検索結果
結合処理を行う処理装置
52
53
結合要求受信
検索結果要求送信
56
検索結果受信
57
処理行数カウンタ更新 Pm=Pm+Zn+yn
58
表xと表yの結合処理
59
処理行数カウンタ更新 Pm=Pm-Zn-yn
60
結合結果送信
結合処理要求 (表x,表y)
UAPのある処理装置
51
結合要求送信
61
結合結果受信
結合結果
62
UAPへ渡す

フロントページの続き

合議体
審判長　吉岡　浩
審判官　冨吉　伸弥
審判官　林　毅

(56)参考文献　特開平０５－２６５９７５号公報
特開平０７－１０５１４９号公報
特開昭６３－０７６０５０号公報
特開平０４－０３３０５０号公報
富田眞治、末吉敏則共著，「並列処理マシン」，株式会社オーム社，１９８９．０５．２５，Ｐ．１５３－１５５

(58)調査した分野(Int.Cl.[7]，ＤＢ名)
G06F15/16,G06F12/00