(19)日本国特許庁(JP)　(12)公表特許公報(A)　(11)特許出願公表番号
特表2007-501458
(P2007-501458A)
(43)公表日　平成19年1月25日(2007.1.25)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 12/00	(2006.01)	G06F	12/00	535Z	5B065
G06F 3/06	(2006.01)	G06F	3/06	304P	5B082
		G06F	3/06	301Z	

審査請求 未請求　予備審査請求 未請求　（全 25 頁）

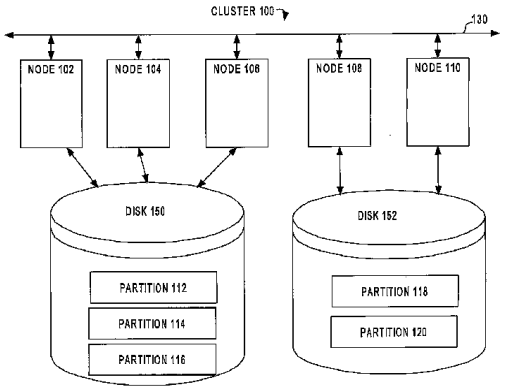
(21)出願番号　特願2006-522078 (P2006-522078)
(86)(22)出願日　平成16年7月28日 (2004.7.28)
(85)翻訳文提出日　平成18年1月30日 (2006.1.30)
(86)国際出願番号　PCT/US2004/024555
(87)国際公開番号　WO2005/013157
(87)国際公開日　平成17年2月10日 (2005.2.10)
(31)優先権主張番号　60/492,019
(32)優先日　平成15年8月1日 (2003.8.1)
(33)優先権主張国　米国 (US)
(31)優先権主張番号　10/831,401
(32)優先日　平成16年4月23日 (2004.4.23)
(33)優先権主張国　米国 (US)

(71)出願人　502303739
オラクル・インターナショナル・コーポレイション
アメリカ合衆国、９４０６５　カリフォルニア州、レッドウッド・ショアーズ、オラクル・パークウェイ、５００
(74)代理人　100064746
弁理士　深見　久郎
(74)代理人　100085132
弁理士　森田　俊雄
(74)代理人　100083703
弁理士　仲村　義平
(74)代理人　100096781
弁理士　堀井　豊

最終頁に続く

(54)【発明の名称】データの所有権の動的な再割当

(57)【要約】
非共有データベースシステムの性能を改善するためのさまざまな技術を記載する。非共有データベースシステムにおいて、この非共有データベースシステムを稼動する少なくとも２つのノードは、ディスクへの共有アクセスを有する。具体的には、特定の動作の性能に対してどのノードが最も効率の良い所有者であるか等の因子に基づき、非共有データベース内のデータの所有権を動的に変更するための技術を提供する。データの所有権は、一旦決定されると、新規の所有者に永続的に変更され得るか、または、特定の動作が持続する間、一時的に変更され得る。



【特許請求の範囲】
【請求項１】
データを管理するための方法であって、
複数のノードにとってアクセス可能な永続的な記憶上に複数の永続的なデータ項目を保存するステップを含み、前記永続的なデータ項目は、前記永続的な記憶上の特定の位置に記憶された特定のデータ項目を含み、前記方法はさらに、
前記ノードの１つに、前記永続的なデータ項目の各々の排他的な所有権を割当てるステップを含み、前記複数のノードの特定のノードには、前記特定のデータ項目の排他的な所有権が割当てられ、
いずれかのノードが前記特定のデータ項目を必要とする演算が実行されることを望むと、前記演算が実行されることを望む前記ノードは、前記特定のデータ項目が前記特定のノードによって排他的に所有されている時に、前記特定のノードが前記特定のデータ項目についての前記演算を実行するように前記特定のノードに前記演算を転送し、前記方法はさらに、
システム性能および作業負荷の少なくとも１つに関する情報を収集することによって統計を収集するステップと、
システム性能およびスループットの少なくとも１つを改善するために、前記統計に基づいて前記永続的なデータ項目の所有権を動的に再割当するステップとを含む、方法。
【請求項２】
情報を収集する前記ステップは、前記永続的なデータ項目を必要とする演算をどのノードが要求しているかを監視するステップを含む、請求項１に記載の方法。
【請求項３】
前記複数のノードは、マルチノードデータベースシステムのノードである、請求項１に記載の方法。
【請求項４】
動的に再割当する前記ステップは、前記永続的なデータ項目が再割当されるべきノードを指定するユーザ入力を受取ることなく実行される、請求項１に記載の方法。
【請求項５】
動的に再割当する前記ステップは、前記マルチノードデータベースシステムがデータベースアプリケーションからのデータベース指令の処理を継続する間に実行される、請求項３に記載の方法。
【請求項６】
動的に再割当する前記ステップは、或る一定の期間の後に実行され、前記或る一定の期間の間に前記永続的なデータ項目を必要とする演算をどのノードが要求したかに基づく、請求項１に記載の方法。
【請求項７】
動的に再割当する前記ステップは、前記特定のデータ項目を必要とする演算を最も頻繁に要求したノードに対し、特定のデータ項目の所有権を再割当するステップを含む、請求項６に記載の方法。
【請求項８】
動的に再割当する前記ステップは、第１のノードに前記特定のデータ項目の所有権を再割当するステップを含み、
前記方法はさらに、前記第１のノードに前記特定のデータ項目の所有権を再割当することに応じて、前記第１のノードから第２のノードに第２のデータ項目の所有権を動的に再割当するステップを含む、請求項１に記載の方法。
【請求項９】
前記第１のノードに前記特定のデータ項目を割当てることにより、前記第１のノードに関連するしきい値が上回ることが生じ、
前記第２のデータ項目の所有権を動的に再割当する前記ステップは、前記しきい値が上回ったことに応じて実行される、請求項８に記載の方法。

【請求項１０】
データを管理するための方法であって、
複数のノードにとってアクセス可能な永続的な記憶上に複数の永続的なデータ項目を保存するステップを含み、前記永続的なデータ項目は、前記永続的な記憶上の特定の位置に記憶された特定のデータ項目を含み、前記方法はさらに、
前記ノードの１つに、前記永続的なデータ項目の各々の排他的な所有権を割当てるステップを含み、前記複数のノードの第１のノードには、前記特定のデータ項目の排他的な所有権が割当てられ、
いずれかのノードが前記特定のデータ項目を必要とする演算が実行されることを望むと、前記演算が実行されることを望む前記ノードは、前記特定のデータ項目が前記第１のノードによって排他的に所有されている時に、前記第１のノードが前記特定のデータ項目についての演算を実行するように第１のノードに前記演算を転送し、前記方法はさらに、
前記特定のデータ項目の排他的な所有権が前記第１のノードによって保持されている間に、前記特定のデータ項目についての演算が実行されることを要求する指令を受信するステップと、
前記第１のノードとは異なる第２のノードによって前記演算が実行されるようにするステップとを含む、方法。
【請求項１１】
第２のノードによって前記演算が実行されるようにする前記ステップは、
前記第２のノードが前記指令についてのサブ演算を実行するのに必要とされる時間と少なくとも同じ長さの時間に、前記特定のデータ項目の排他的な所有権を第２のノードに一時的に再割当するステップを含み、前記サブ演算は前記特定のデータ項目を必要とし、第２のノードによって演算が実行されるようにする前記ステップはさらに、
前記時間の後に、前記特定のデータ項目の前記排他的な所有権を前記第１のノードに再び自動的に再割当するステップを含む、請求項１０に記載の方法。
【請求項１２】
一時的に再割当する前記ステップは、並列化された演算のサブ演算が、複数のノードに存在するスレーブ間に分配されるように実行される、請求項１１に記載の方法。
【請求項１３】
一時的に再割当する前記ステップは、前記指令によって要求される演算が、前記第２のノードを含むものの前記第１のノードを含まない一組の１つ以上のノードにおいて統合されるように実行される、請求項１１に記載の方法。
【請求項１４】
第２のノードによって演算が実行されるようにする前記ステップは、
前記第２のノードが前記特定のデータ項目の排他的な所有権を獲得することなく前記第２のノードに演算を実行させるステップを含み、
前記第２のノードは、前記指令のサブ演算を実行するために前記特定のデータ項目にアクセスすることが許可され、前記サブ演算は前記特定のデータ項目を必要とし、第２のノードによって演算が実行されるようにする前記ステップはさらに、
前記第２のノードが前記サブ演算の実行を完了した後に、前記第２のノードに前記特定のデータ項目にアクセスすることを許可するのを中止するステップを含む、請求項１０に記載の方法。
【請求項１５】
非コミット読出分離レベルを前記指令に適用することを決定するステップと、
前記第１のノードが前記特定のデータ項目のいずれかのダーティなバージョンをディスクにフラッシュすることを要求せずに前記第２のノードに前記サブ演算を実行させるステップとをさらに含む、請求項１４に記載の方法。
【請求項１６】
コミット済読出分離レベルを前記指令に適用することを決定するステップと、
前記第１のノードが前記特定のデータ項目のいずれかのダーティなバージョンをディス

クにフラッシュするまで、前記第２のノードが前記サブ演算を実行することを防止するステップとをさらに含む、請求項１４に記載の方法。

【請求項１７】

１つ以上のプロセッサによって実行されると、前記１つ以上のプロセッサに請求項１に記載の方法を実行させる命令の１つ以上のシーケンスを搬送する、コンピュータ読取可能な媒体。

【請求項１８】

１つ以上のプロセッサによって実行されると、前記１つ以上のプロセッサに請求項２に記載の方法を実行させる命令の１つ以上のシーケンスを搬送する、コンピュータ読取可能な媒体。

【請求項１９】

１つ以上のプロセッサによって実行されると、前記１つ以上のプロセッサに請求項３に記載の方法を実行させる命令の１つ以上のシーケンスを搬送する、コンピュータ読取可能な媒体。

【請求項２０】

１つ以上のプロセッサによって実行されると、前記１つ以上のプロセッサに請求項４に記載の方法を実行させる命令の１つ以上のシーケンスを搬送する、コンピュータ読取可能な媒体。

【請求項２１】

１つ以上のプロセッサによって実行されると、前記１つ以上のプロセッサに請求項５に記載の方法を実行させる命令の１つ以上のシーケンスを搬送する、コンピュータ読取可能な媒体。

【請求項２２】

１つ以上のプロセッサによって実行されると、前記１つ以上のプロセッサに請求項６に記載の方法を実行させる命令の１つ以上のシーケンスを搬送する、コンピュータ読取可能な媒体。

【請求項２３】

１つ以上のプロセッサによって実行されると、前記１つ以上のプロセッサに請求項７に記載の方法を実行させる命令の１つ以上のシーケンスを搬送する、コンピュータ読取可能な媒体。

【請求項２４】

１つ以上のプロセッサによって実行されると、前記１つ以上のプロセッサに請求項８に記載の方法を実行させる命令の１つ以上のシーケンスを搬送する、コンピュータ読取可能な媒体。

【請求項２５】

１つ以上のプロセッサによって実行されると、前記１つ以上のプロセッサに請求項９に記載の方法を実行させる命令の１つ以上のシーケンスを搬送する、コンピュータ読取可能な媒体。

【請求項２６】

１つ以上のプロセッサによって実行されると、前記１つ以上のプロセッサに請求項１０に記載の方法を実行させる命令の１つ以上のシーケンスを搬送する、コンピュータ読取可能な媒体。

【請求項２７】

１つ以上のプロセッサによって実行されると、前記１つ以上のプロセッサに請求項１１に記載の方法を実行させる命令の１つ以上のシーケンスを搬送する、コンピュータ読取可能な媒体。

【請求項２８】

１つ以上のプロセッサによって実行されると、前記１つ以上のプロセッサに請求項１２に記載の方法を実行させる命令の１つ以上のシーケンスを搬送する、コンピュータ読取可能な媒体。

【請求項２９】
１つ以上のプロセッサによって実行されると、前記１つ以上のプロセッサに請求項１３に記載の方法を実行させる命令の１つ以上のシーケンスを搬送する、コンピュータ読取可能な媒体。
【請求項３０】
１つ以上のプロセッサによって実行されると、前記１つ以上のプロセッサに請求項１４に記載の方法を実行させる命令の１つ以上のシーケンスを搬送する、コンピュータ読取可能な媒体。
【請求項３１】
１つ以上のプロセッサによって実行されると、前記１つ以上のプロセッサに請求項１５に記載の方法を実行させる命令の１つ以上のシーケンスを搬送する、コンピュータ読取可能な媒体。
【請求項３２】
１つ以上のプロセッサによって実行されると、前記１つ以上のプロセッサに請求項１６に記載の方法を実行させる命令の１つ以上のシーケンスを搬送する、コンピュータ読取可能な媒体。
【発明の詳細な説明】
【技術分野】
【０００１】
発明の分野
この発明は、共有されたディスクハードウェア上で稼動する非共有データベースシステムにおいてデータを管理するための技術に関する。
【背景技術】
【０００２】
発明の背景
マルチプロセッシングコンピュータシステムは一般に、３つのカテゴリ、すなわち、全共有システム、共有ディスクシステム、および非共有システムに分類される。全共有システムにおいて、すべてのプロセッサ上のプロセスは、システム内のすべての揮発性メモリデバイス（以降、包括的に「メモリ」と称する）と、すべての不揮発性メモリデバイス（以降、包括的に「ディスク」と称する）とに対して直接のアクセスを有する。したがって、全共有の機能を提供するために、コンピュータのさまざまな構成要素間には高度な配線が必要とされる。加えて、全共有アーキテクチャには、スケーラビリティの限界が存在する。
【発明の開示】
【発明が解決しようとする課題】
【０００３】
共有ディスクシステムでは、プロセッサおよびメモリがノードにグループ化される。共有ディスクシステム内の各ノードは、それ自体が、複数のプロセッサおよび複数のメモリを含む全共有システムを構成し得る。すべてのプロセッサ上のプロセスは、システム内のすべてのディスクにアクセス可能であるが、特定のノードに属するプロセッサ上のプロセスのみが、特定のノード内のメモリに直接アクセスできる。共有ディスクシステムは一般に、必要とする配線が、全共有システムよりも少ない。共有ディスクシステムはまた、作業負荷が不均衡な状態にも容易に適合する。なぜなら、すべてのノードがすべてのデータにアクセスできるためである。しかしながら、共有ディスクシステムは、コヒーレンスオーバヘッドの影響を受けやすい。たとえば、第１のノードがデータを変更し、かつ、第２のノードがその同じデータの読出または変更を望む場合、そのデータの正しいバージョンが第２のノードに確実に提供されるように、さまざまなステップを実行しなければならないことが考えられる。
【０００４】
非共有システムでは、すべてのプロセッサ、メモリ、およびディスクがノードにグルー

プ化される。非共有システムは、共有ディスクシステムと同様に、各ノード自体が全共有システムまたは共有ディスクシステムを構成し得る。特定のノード上で稼動するプロセスのみが、特定のノード内のメモリおよびディスクに直接アクセス可能である。マルチプロセッシングシステムの３つの一般的なタイプのうち、さまざまなシステム構成要素間で必要とされる配線の量は、一般に非共有システムが最も少ない。しかしながら、非共有システムは、作業負荷が不均衡な状態の影響を最も受けやすい。たとえば、特定のタスク中にアクセスされるべきすべてのデータが、特定のノードのディスク上に存在し得る。したがって、他のノード上のプロセスがアイドル状態であるにも関わらず、粒度の細かい仕事を実行するために、そのノード内で稼動するプロセスしか使用することができない。

【０００５】

マルチノードシステム上で稼動するデータベースは一般に、２つのカテゴリ、すなわち、共有ディスクデータベースおよび非共有データベースに分類される。

【０００６】

共有ディスクデータベース

共有ディスクデータベースは、データベースシステムによって管理されるすべてのデータが、データベースシステムにとって利用可能なすべての処理ノードの管理下にある（visible）という前提に基づき、仕事を調整する。その結果、共有ディスクデータベースでは、仕事中にアクセスされるであろうデータを含むディスクの位置に関係なく、サーバがいずれかのノード上のプロセスにいずれかの仕事を割当てることができる。

【０００７】

すべてのノードが同じデータへのアクセスを有し、各ノードがそれ自体の専用キャッシュを有しているため、同じデータ項目の多数のバージョンが、多くのノード中のどのような数のノードのキャッシュ内にも存在し得る。残念ながら、このことは、１つのノードが特定のデータ項目の特定のバージョンを要求する際に、そのノードが他のノードと連携して、要求を行なっているノードにそのデータ項目の特定のバージョンが転送されるようにしなければならないことを意味する。したがって、共有ディスクデータベースは、「データ転送」の概念で作動すると言われており、データは、そのデータに仕事を行なうように割当てられたノードに転送されなければならない。

【０００８】

このようなデータ転送要求は、「ピング（ping）」を生じ得る。具体的に、ピングは、１つのノードが必要とするデータ項目の複製が、別のノードのキャッシュ内に存在する際に生じる。ピングは、データ項目がディスクに書込まれた後にディスクから読出されることを必要とし得る。ピングにより必要とされるディスク動作の性能は、データベースシステムの性能を著しく下げる恐れがある。

【０００９】

共有ディスクデータベースは、非共有コンピュータシステムおよび共有ディスクコンピュータシステムのいずれの上でも稼動され得る。非共有コンピュータシステム上で共有ディスクデータベースを稼動させるために、オペレーティングシステムにソフトウェアサポートを追加するか、または、追加のハードウェアを設けて、プロセスが遠隔ディスクへのアクセスを有し得るようにすることが可能である。

【００１０】

非共有データベース

非共有データベースは、プロセスと同じノードに属するディスクにデータが含まれている場合に限り、そのプロセスがそのデータにアクセス可能であるものと想定する。したがって、特定のノードが、別のノードによって所有されるデータ項目についての演算が実行されることを望む場合、その特定のノードは、他のノードがその演算を実行するように、他のノードに要求を送信しなければならない。したがって、非共有データベースは、ノード間でデータを転送する代わりに「機能の転送」を実行すると言われる。

【００１１】

いずれかの所定のデータ片が１つのノードによってのみ所有されているため、その１つ

のノード（データの「所有者」）のみが、そのキャッシュ内にそのデータの複製を有する。したがって、共有ディスクデータベースシステムで必要とされたタイプのキャッシュのコヒーレンスのメカニズムが必要とされない。さらに、非共有システムは、ピングにまつわる性能の不利益を被らない。なぜなら、別のノードがそのキャッシュにデータ項目をロードすることができるように、そのデータ項目を所有するノードが、そのデータ項目のキャッシュされたバージョンをディスクに保存するように求められないためである。

【００１２】

非共有データベースは、共有ディスクマルチプロセッシングシステムおよび非共有マルチプロセッシングシステムのいずれの上でも稼動され得る。共有されたディスクマシン上で非共有データベースを稼動させるために、データベースをセグメント化して各パーティションの所有権を特定のノードに割当てるためのメカニズムを設けることができる。

【００１３】

所有を行なうノードのみがデータ片に作用し得るということは、非共有データベース内の作業負荷が極めて不均衡になり得ることを意味する。たとえば、１０個のノードのシステムにおいて、仕事の全要求の９０％が、それらのノードのうちの１つによって所有されるデータを必要とするかもしれない。したがって、その１つのノードが酷使され、他のノードの計算リソースが十分に活用されない。作業負荷の「均衡を取り戻す」ために、非共有データベースをオフラインにすることができ、データ（およびその所有権）をノード間で再分配することができる。しかしながら、このプロセスは、潜在的に大量のデータの移動を伴い、作業負荷の偏りを一時的にしか解決しない恐れがある。

【課題を解決するための手段】

【００１４】

この発明は、添付の図面において限定ではなく例示として示される。これらの図面では、同じ参照番号が同じ要素を指す。

【発明を実施するための最良の形態】

【００１５】

発明の詳細な説明

共有ディスク記憶システムを含む非共有データベースシステムの性能を改善するためのさまざまな技術を以下に説明する。以下の記載内容では、説明のために多数の特定の詳細を明示して、この発明の完全な理解を図る。しかしながら、このような特定の詳細を用いなくてもこの発明を実施し得ることが明らかであろう。場合によっては、周知の構造およびデバイスをブロック図の形で示し、この発明をむやみに不明瞭にしないようにする場合もある。

【００１６】

機能上の概観

非共有データベースシステムを稼動する少なくとも２つのノードがディスクへの共有されたアクセスを有する非共有データベースシステムの性能を改善するためのさまざまな技術を、以下に説明する。データベースシステムの非共有アーキテクチャによって規定されるように、各データ片は依然として、いずれかの所定の時点において１つのノードによってのみ所有される。しかしながら、非共有データベースシステムを稼動する少なくともいくつかのノードがディスクへの共有されたアクセスを有するという点を利用して、非共有データベースシステムの均衡をより効率よく取り戻し、非共有データベースシステムをより効率よく回復する。

【００１７】

具体的には、特定の動作の性能に対してどのノードが最も効率の良い所有者であるか等の因子に基づき、非共有データベース内のデータの所有権を動的に変更するための技術を提供する。データの所有権は、一旦決定されると、新規の所有者に永続的に変更され得るか、または、特定の動作が持続する間に一時的に変更され得る。

【００１８】

従来の所有権の再割当の動作にまつわるオーバーヘッドを回避するために、５０２７７

－２２７７に記載された技術を用いて所有権の再割当が実行され得る。これらの技術は、再割当されているデータが、そのデータが永続的な記憶上に存在する位置から移動されることを必要としない。

【００１９】

共有ディスクシステムを含む例示的なクラスタ

図１は、この発明の実施例が実現され得るクラスタ１００を示すブロック図である。クラスタ１００は、相互接続１３０によって結合される５個のノード１０２、１０４、１０６、１０８、および１１０を含み、相互接続１３０は、これらのノードが互いに通信することを可能にする。クラスタ１００は、２つのディスク１５０および１５２を含む。ノード１０２、１０４および１０６は、ディスク１５０へのアクセスを有し、ノード１０８および１１０は、ディスク１５２へのアクセスを有する。したがって、ノード１０２，１０４および１０６とディスク１５０とを含むサブシステムは、第１の共有ディスクシステムを構成し、ノード１０８および１１０とディスク１５２とを含むサブシステムは、第２の共有ディスクシステムを構成する。

【００２０】

クラスタ１００は、２つの共有ディスクサブシステムを含み、かつ、それらの共有ディスクサブシステム間で帰属関係が重複しない、相対的に単純なシステムの一例である。実際のシステムは、クラスタ１００よりもより一層複雑であることが考えられ、数百個のノード、数百個の共有ディスク、および、これらのノードとこれらの共有ディスクとの間に多対多の関係を有する。このようなシステムでは、多くのディスクへのアクセスを有する１つのノードが、たとえば、いくつかの別個の共有ディスクサブシステムのメンバーであることが考えられ、ここでは、共有ディスクサブシステムの各々が、共有されたディスクのうちの１つと、その共有されたディスクへのアクセスを有するすべてのノードとを含む。

【００２１】

共有ディスクシステム上の非共有データベース

例示のため、非共有データベースシステムがクラスタ１１０上で稼動しているものと想定されたい。ここで、非共有データベースシステムによって管理されるデータベースは、ディスク１５０および１５２に記憶される。データは、データベースシステムの非共有特性に基づき、５個のグループまたはパーティション１１２、１１４、１１６、１１８、および１２０に分離され得る。各パーティションは、対応するノードに割当てられる。パーティションに割当てられたノードは、そのパーティション内に存在するすべてのデータの排他的な所有者であると考えられる。この例において、ノード１０２、１０４、１０６、１０８、および１１０はそれぞれ、パーティション１１２、１１４、１１６、１１８、および１２０を所有する。ディスク１５０へのアクセスを有するノード（ノード１０２、１０４および１０６）によって所有されるパーティション１１２、１１４および１１８は、ディスク１５０に記憶される。同様に、ディスク１５２へのアクセスを有するノード（ノード１０８および１１０）によって所有されるパーティション１１８および１２０は、ディスク１５２に記憶される。

【００２２】

クラスタ１００上で稼動するデータベースシステムの非共有特性によって規定されるように、いずれかのデータ片は、いずれかの所定の時点において、多くても１つのノードによって所有される。加えて、共有されたデータへのアクセスは、機能の転送により調整される。たとえば、ＳＱＬ言語をサポートするデータベースシステムの場合、特定のデータ片を所有しないノードは、そのデータ片を所有するノードにＳＱＬステートメントのフラグメントを転送することにより、そのデータに関する演算が実行されるようにすることができる。

【００２３】

所有権のマップ

機能の転送を効率よく実行するために、すべてのノードは、どのノードがどのデータを

所有しているかを認識しなければならない。したがって、所有権のマップが規定され、所有権のマップは、データーノード間の所有権の割当を示す。実行時に、さまざまなノードが所有権のマップに照会し、実行時に、ＳＱＬフラグメントの経路を適正なノードに指定する。

【００２４】

一実施例に従うと、データーノード間のマップピングは、ＳＱＬ（または他のいずれかのデータベースアクセス言語）のステートメントのコンパイル時に決定される必要はない。むしろ、以下により詳細に説明するように、データーノード間のマッピングは、実行時に規定および修正され得る。以下に説明する技術を用いると、データが存在するディスクへのアクセスを有する１つのノードから、データが存在するディスクへのアクセスを有する別のノードに所有権が変更する際に、データをディスク上の永続的な位置から移動させることなく所有権の変更が実行される。

【００２５】

ロッキング

ロックは、リソースへのアクセスを有するいくつかのエンティティ間でリソースへのアクセスを調整するために使用される構造である。非共有データベースシステムの場合、非共有データベース内のユーザデータへのアクセスを調整するためのグローバルロッキングが必要ではなくなる。なぜなら、いずれかの所定のデータ片が１つのノードによってのみ所有されているためである。しかしながら、非共有データベースのすべてのノードが、所有権のマップへのアクセスを要求するため、所有権のマップに対する整合性のない更新を防ぐために、何らかのロッキングが必要とされ得る。

【００２６】

一実施例に従うと、データ片の所有権が１つのノード（「以前の所有者」）から別のノード（「新規の所有者」）に再割当されている際に、２ノードロッキング方式が使用される。さらに、非共有データベースに関連するメタデータへのアクセスを制御するために、グローバルロッキング機構を使用することができる。このようなメタデータは、たとえば所有権のマップを含み得る。

【００２７】

バケットベースのセグメント化

上述のように、非共有データベースにより管理されるデータがセグメント化され、各パーティション内のデータが１つのノードにより排他的に所有される。一実施例に従うと、これらのパーティションは、データを論理バケットに割当てた後に、それらのバケットの各々を１つのパーティションに割当てることによって規定される。したがって、所有権のマップ内のデーターノード間のマッピングは、データーバケット間のマッピングおよびバケットーノード間のマッピングを含む。

【００２８】

一実施例に従うと、データーバケット間のマッピングは、各データ項目の名前にハッシュ関数を適用することによって規定される。同様に、バケットーノード間のマッピングは、バケットに関連する識別子に別のハッシュ関数を適用することによって規定され得る。代替的に、マッピングの一方または両方は、範囲ベースのセグメント化もしくはリスト区分を用いることにより、または、個々の関係の１つ１つを単に列挙することにより、規定され得る。たとえば、１００万個のデータ項目の名前空間を５０個の範囲に分割することにより、それらのデータ項目は、５０個のバケットにマッピングされ得る。５０個のバケットは次に、各バケットに関する、（１）そのバケットを識別し、（２）現時点でバケットが割当てられたノードを識別する記録を記憶することにより、５個のノードにマッピングされ得る。

【００２９】

バケットを使用することにより、各データ項目に関して別個のマッピング記録が記憶されていたマッピングに比べ、所有権のマッピングのサイズが大幅に縮小される。さらに、バケットの数がノードの数を上回る実施例では、バケットの使用により、所定のノードが

所有するデータのサブセットへの所有権の再割当が相対的に容易になる。たとえば、１つの新規のノードには、現時点で１０個のバケットが割当てられている１つのノードから、１つのバケットが割当てられ得る。このような再割当は、そのバケットについてのバケット－ノード間のマッピングを示す記録の修正を伴うに過ぎない。再割当されたデータのデーター－バケット間のマッピングは、変更されなくてよい。

【００３０】

上述のように、データ－バケット間のマッピングは、以下のものに限定されないが、ハッシュパーティション、範囲パーティション、またはリスト値を含むさまざまな技術のうちのいずれか１つを用いることによって規定され得る。範囲ベースのセグメント化が用いられ、かつ、範囲の数がノードの数よりもそれほど著しく大きくない場合は、データ項目をセグメント化するのに使用される範囲キーが変化しない値（日付等）である限り、データベースサーバは、粒度のより細かい（より狭い）範囲を用いて所望の数のバケットを得ることができる。範囲キーが変化し得る値である場合、データ項目は、特定のデータ項目についての範囲キーの値に対する変化に応じて古いバケットから除去されて、そのデータ項目の範囲キーの新規の値に対応するバケットに追加される。

【００３１】

所有権の最初の割当の規定

上述のマッピング技術を用いると、１つのテーブルまたは索引の所有権が複数のノード間で共有され得る。最初に、所有権の割当は無作為であることが考えられる。たとえば、ユーザは、データ－バケット間のマッピングに対してキーおよびセグメント化の技術（ハッシュ、範囲、リスト等）と、バケット－ノード間のマッピングに対してセグメント化の技術を選択することができるが、バケットの、ノードへの最初の割当は指定する必要がない。データベースサーバが次に、データ－バケット間のマッピングに対するキーに基づいて、バケット－ノード間のマッピングに対するキーを決定することができ、バケットによって表される特定のデータおよびデータベースオブジェクトに関係なく、バケット－ノード間の最初の割当を行うことができる。

【００３２】

たとえば、ユーザがキーＡに基づいてオブジェクトをセグメント化することを選択する場合、データベースサーバはキーＡを用いて、バケット－ノード間のマッピングを決定する。場合によっては、データベースサーバは、データ－バケット間のマッピングに使用されるキーに対し、さらに別のキーを追加するか、または、異なる機能（その機能がデーター－バケット間のマッピングを保持する限り）を適用することができる。たとえば、オブジェクトが、キーＡの使用により４個のデータバケットにハッシュセグメント化される場合、データベースサーバは、ハッシュ関数をキーＢに適用してバケット－ノード間のマッピングを決定するか、または、ハッシュ値の数を１２へと単に増大することにより、それらの４個のバケットの各々を３個のバケットに細分することができる（そして、バケットの、ノードへの柔軟性の高い割当を可能にする）。ハッシュがモジュロ関数である場合、０番目、４番目、および８番目のバケット－ノード間のバケットは、０番目のデーター－バケット間のバケットに対応し、１番目、５番目、および９番目のバケット－ノード間のバケットは、１番目のデーター－バケット間のバケットに対応する、等である。

【００３３】

別の例として、オブジェクトがDATE型を有するキーＡ上に範囲区分される場合、データ－バケット間のマッピングは、年を返す関数year(date)を用いることによって指定され得る。しかしながら、バケット－ノード間のマッピングは、month_and_year(date)を用いることにより、データベースサーバによって内部で計算され得る。各年の区分は、１２個のバケット－ノード間のバケットに分割される。すなわち、データベースサーバは、特定の年（一般には現時点での年）の日付に対してアクセスが頻繁に行なわれていると判断した場合に、これらの１２個のバケットを他のノード間で再分配することができる。

【００３４】

上で提示したいずれの例においても、バケット－ノード間のbucket＃を考慮すると、デ

ータベースサーバは、データ－バケット間のbucket#を一意に決定することができる。また、これらの例において、ユーザは、データ－バケット間のマッピングに対してキーおよびセグメント化の技術を選択する。しかしながら、代替的な実施例において、ユーザは、データ－バケット間のマッピングに対してキーおよびセグメント化の技術を選択しないことが考えられる。むしろ、データ－バケット間のマッピングに対するキーおよびセグメント化の技術は、データベースサーバにより自動的に決定されることも考えられる。

【００３５】

一実施例に従うと、データベースサーバは、各ノードにいくつのバケットが割当てられるべきであるかに基づき、バケット－ノード間の最初の割当を行なう。たとえば、より大きな容量を有するノードには、より多くのバケットが割当てられ得る。しかしながら、最初の割当において、どの特定のバケットがどのノードに割当てられるべきかについての決定は無作為である。

【００３６】

代替的な実施例によると、データベースサーバは、バケット－ノード間の割当を行なう際に、どのデータが１つのバケットによって表わされているかを考慮する。たとえば、特定のテーブルに対するデータがいくつかのバケット間で分割されているものと想定されたい。データベースサーバは、それらのバケットのすべてを同じノードに作為的に割当てることができ、または、それらのバケットの所有権を多くのノード間に作為的に分配することができる。同様に、データベースサーバは、最初の割当において、テーブルに関連するバケットを、それらのテーブルに対する索引に関連するバケットと同じノードに割当てようと試みることが考えられる。反対に、データベースサーバは、テーブルに関連するバケットを、それらのテーブルに対する索引に関連するバケットが割当てられたノードとは異なるノードに割当てようと試みることが考えられる。

【００３７】

所有権の、自動的かつ作業負荷ベースの再割当

最初の割当が行なわれる態様に関わらず、最初の割当により、データベースサーバが実行するように求められるすべての動作の最適性能を確実に生じるようにすることは事実上不可能である。したがって、この発明の一実施例に従うと、所有権は、データベースシステムの実際の実行時の動作を監視することによって収集された統計に基づき、自動的に再割当される。

【００３８】

たとえば、一実施例において、非共有データベースシステムは監視機構を含む。この監視機構は、ノードによって行なわれた要求を監視し、各バケットに関し、非所有者ノードが、そのバケットからのデータを必要とする演算をどれだけ頻繁に要求するかに関する統計を保存する。一実施例によると、所有権のマップは永続的な記憶に保存されるが、監視機構によって保存される統計は、揮発性メモリに保存される。

【００３９】

データベースサーバは、統計に基づき、特定の非所有者ノードが他の任意のノードに比べ、より一層高い頻度で特定のバケットからのデータを必要とする演算を要求していると判断することができる。データベースサーバは、この情報に基づき、バケットの所有権をその特定のノードに自動的に再割当することができる。

【００４０】

非所有者のノードがバケットを必要とする演算を要求する頻度は、所有権を再割当する決定に関わることが考えられる多くの性能因子の単なる一例である。たとえば、監視機構は、所有者ノードがそれ自体のために、特定のバケットからのデータについての演算（「自己に有利な演算」）を実行する頻度を追跡することもできる。特定のバケットの所有者ノードが特定のバケットからのデータに対して自己に有利な演算を実行する頻度が、いずれかの非所有ノードが特定のバケットに対する演算を要求する頻度よりも高い場合、データベースサーバは、バケットの所有権の再割当を行なわないことを選択することができる。非所有者ノードが特定のバケットに対する演算を要求する頻度が、所有者ノードがその

バケットに対して自己に有利な演算を実行する頻度よりも高い場合でも、データベースサーバは、この使用量の差が或る一定のしきい値を超えたときにのみ、所有権を転送するように構成され得る。

【００４１】

一実施例に従うと、テーブル等の特定のリソースが、ほぼ同じ頻度でいくつかのノードによってアクセスされている場合、データベースサーバは、そのテーブルに関連するバケットの所有権を、それらのノード間に均等に分散させることができる。その結果、テーブルへのアクセスが非常に多い場合でも、そのテーブルにアクセスするという作業は、利用可能なノード間でより均一に分散される。

【００４２】

時間の経過とともに、どのノードがバケット内のデータに最も頻繁にアクセスしているかに基づいてバケットの所有権が調節されるのに伴い、論理的に関連するデータを表わすバケットは、同じノードによって所有される傾向を有する。たとえば、特定のテーブルからのデータに対応するバケットは、そのテーブルに構築された索引に対応するバケットと同じノードによって所有される傾向を有する。

【００４３】

実際のバケットの使用量に関する統計に基づいて所有権を再割当することにより、より複雑な他の割当の決定を行なう必要がなくなる。たとえば、所有権は、ＳＱＬのWHERE句、JOIN条件、またはAGGREGATIONを考慮するクエリのプロファイルに基づく必要がない。同様に、ユーザは、トランザクションの、データへの相性の良さを明示的に示す必要がない。さらに、割当者ノードおよび被割当者ノードの両方によって共有されるディスクにデータが存在する時に、非共有データベースサーバは、データを物理的に再分配することなく、作業負荷の偏りまたは変化に迅速に適応することができる。共有されたディスク上に存在するデータによって再割当のコストが下がることにより、非共有データベースサーバは、どのような外部ツールも使用せずに、データの所有権の変更および性能の測定を効率よく行なうことができる。

【００４４】

再割当の均衡化

特定のノードがその最初のバケットのすべてを継続して所有する場合、および、特定のノードがそれらのバケット内のデータに頻繁にアクセスしているために、その特定のノードにいくつかの新規のバケットが動的に割当てられている場合、ノードが酷使された状態になることが考えられ、それにより、データベースシステムの作業負荷は偏ってしまう。同様に、バケットが特定のノードに再割当されることなく、その特定のノードの最初のバケットの多くが他のノードに再割当される場合にも、作業負荷に偏りが生じる。

【００４５】

したがって、バケットの所有権が特定のノードに割当てられる際に、そのノードから別のノードに異なるバケットの所有権が再割当されることが望ましいと考えられる。同様に、バケットの所有権が特定のノードから割当てられる際に、その特定のノードに異なるバケットの所有権を再割当することが望ましいと考えられる。このような再割当は、他の再割当のひずみ効果に逆らうように行なわれ、この明細書では「再割当の均衡化」と称される。

【００４６】

再割当動作が何らかのオーバーヘッドを伴うため、或る一定の偏りしきい値が満たされた後にのみ再割当の均衡化を実行することが望ましいと考えられる。たとえば、データベースサーバは、各ノードに対して「目標バケット数」を保存することができる。特定のノードによって所有されるバケットの数が、予め定められた量だけ、特定のノードの目標バケット数を下回ると、再割当の均衡化が実行されて、他のノードからその特定のノードに１つ以上のバケットが割当てられ得る。そこからバケットが再割当されるノードは、たとえば、それらのノードが所有するいくつのバケットが、それぞれの目標バケット数を上回っているかに基づいて選択され得る。

【００４７】
同様に、特定のノードにより所有されるバケットの数が、予め定められた量だけ、特定のノードの目標バケット数を上回った場合、再割当の均衡化が実行されて、その特定のノードから他のノードに１つ以上のバケットが割当てられ得る。それに対してバケットが再割当されるノードは、たとえば、それらのノードが所有するいくつの数のバケットが、それぞれの目標バケット数を下回るかに基づいて選択され得る。
【００４８】
再割当の均衡化の間にどのバケットを再割当すべきかを決定する際に、データベースサーバによってさまざまな因子が使用され得る。たとえば、データベースサーバは、割当者ノードの自己に有利な演算に関して頻度が最も少ない、割当者ノードのバケットを選択することができる。代替的に、データベースサーバは、被割当者ノードによって要求される演算に関する頻度が最も高い、割当者ノードのバケットを選択することができる。一実施例に従うと、データベースサーバは、バケットが割当者ノードの自己に有利な演算に関する頻度、および、バケットが被割当者ノードによって要求される演算に関する頻度を含む多くの因子を考慮する。
【００４９】
所有権の一時的な割当
これまでのセクションでは、監視期間中にどのノードがデータのどのバケットに対する演算を要求しているか等の因子に基づいて所有権を再割当するための技術を説明している。しかしながら、この態様で行なわれる割当は長期的に見て最適であるかもしれないが、いずれかの所定の演算に対してはこれらの割当が最適ではないことがあり得る。したがって、一実施例に従うと、データベースサーバは、一組の１つ以上の演算が持続する間、バケットの所有権の割当を一時的に変更するための論理を含む。一組の演算が完了した後に、バケットは、それらの以前の所有者に再割当される。
【００５０】
たとえば、データベースサーバは、膨大な演算が持続する間にのみ、データの所有権を変更することができる。結果として、或るテーブルに対応するバケットのすべては、その日一日中、１つのノードにより所有され得る。したがって、その一日中、そのテーブルに対するすべての要求は、そのノードに経路指定される。一日の終わりの報告用にその同じテーブルを「再セグメント化」して、一日の終わりの報告用のデータを検索するクエリを並列化することができる。
【００５１】
別の例として、データベースサーバは、回復動作が持続する間、データの所有権を変更することができる。具体的に、データベースサーバは、すべてのノードにバケットを一時的に均等に再割当することができ、各ノードは、各ノードが他のノードに対して並列に所有するバケット内で、順方向の再実行－回復および逆方向のトランザクションロールバックを実行することができる。このような所有権の再分配を、データベース回復動作および媒体回復動作のいずれにも用いることができる。たとえば、媒体回復動作中に、最も新しいバックアップが復元された後に、各ノードは、並列なアーカイブから再実行を適用することができる。このような状況下において、並列な媒体回復に携わっている各ノードは、適切な再実行ログおよびアーカイブを読出すことができなくてはならない。
【００５２】
現時点でデータを所有していないノードに対する、並列クエリのスレーブの柔軟的な配置
上述のように、データの所有権は、演算が持続する間、一時的に変更され得る。このような一時的な割当を用いて、たとえば並列化され得る演算を指定するクエリの性能を改善することができる。演算が並列化されると、その演算は、互いに並列に実行され得るいくつかのサブタスクに分割される。このようなサブタスクを実行するのに使用されるプロセスを、並列クエリのスレーブと称する。
【００５３】

非共有データベースシステムが共有ディスク環境で実現される際に、並列クエリのスレーブの配置は、データの物理的な位置によって規定されない。たとえば、テーブルＴ内のデータが、（１）２つのノード１および２によって共有されるディスク上に存在し、かつ、（２）テーブルＴの区分に対応する２つのバケットＢ１およびＢ２に属すると想定されたい。テーブルＴの走査を要求するクエリは、２つのサブタスク、すなわち、バケットＢ１内のデータの走査およびバケットＢ２内のデータの走査に分割され得る。データベースサーバがクエリを実行するように要求された時点で、ノード１がバケットＢ１およびＢ２の両方の所有者であると想定されたい。このような状況下において、Ｂ２の所有権はノード２に一時的に再割当され得、それにより、ノード２上のスレーブは、ノード１上のスレーブがＢ１の走査を実行する間にＢ２の走査を実行することができる。

【００５４】

上述の例では、並列化され得るクエリの処理により多くのノードを関与させるために、１つのバケットの所有権が別のノードに一時的に割当てられた。反対に、クエリの処理に関与するノードの数を減らす態様で、所有権を一時的に再割当することが望ましいことがあり得る。たとえば、データが第１の組のスレーブにより走査された後に第２の組のスレーブに再分配され、それにより、第２の組のスレーブが、そのデータに対し、以降の何らかの演算を実行し得るようにすることをクエリが要求すると想定されたい。多くのノードにわたって分散されたスレーブによって走査が実行される場合、クエリによって要求される再分配は、大量のノード間通信を生じる。ノード間通信の量を減らすために、テーブルＴの区分の所有権を統合することができる。

【００５５】

たとえば、クエリが開始された時点で、Ｂ１がノード１によって所有され、Ｂ２がノード２によって所有されている場合、テーブルＴの所有権は、クエリが持続する間、ノード２からノード１にＢ２を一時的に割当てることによって統合され得る。このような状況下では、第１および第２の組のスレーブがいずれもノード１上に存在する。したがって、走査の後に行なわれた再分配は、ノード間通信を生じない。

【００５６】

所有権を必要としない仕事

上述のように、非共有データベースシステムにおいて、データ片の所有者のみが、そのデータを必要とするタスクを実行することが許可される。しかしながら、この発明の一局面によると、データ片の非所有者がそのデータに対して演算を行なうことが許可されるかどうかは、その演算が要求する分離レベルに依存する。

【００５７】

分離レベルの概念は、たとえば、「データベースシステムに分離レベルを設けるための方法および装置（Method and Apparatus For Providing Isolation Levels In A Database System）」と題された、米国特許第５，８７０，７５８号に詳細に論じられる。データベースシステムの場合、いくつかの分離レベルが規定される。規定された分離レベルは、「非コミット読出」分離レベルおよび「コミット済読出」分離レベルを含む。非コミット読出分離レベルは、ダーティリード、反復不能リード、およびファントムリードから保護されておらず、或る一定の動作、たとえばデータマイニングおよび統計的クエリ等には十分である。このような動作に関しては、データ片に作用する大きなクエリのフラグメントが、それらのデータ片を所有していないノード上で稼動し得る。

【００５８】

非コミット読出分離レベルとは異なり、コミット済読出分離レベルは、ダーティリードを防止する。コミット済読出分離レベルを要求する演算に関し、データ項目の組を所有するノードは、それらのデータ項目を必要とする演算の実行前に、それらのデータ項目を必要とするコミット済みのページをディスクにフラッシュすることができる。ページがディスクにフラッシュされた後に、それらのデータ項目を所有しないノードは、それらのデータ項目を必要とする演算を実行することができる。コミット済みのページがディスクにフラッシュされたため、コミット済みのページは、ディスクへの共有アクセスを有する非所

有者ノードによって視認され得る。したがって、コミット済読出分離レベルが保存される。演算の実行中に、それらの演算が完了するまで、所有者は、それらのデータ項目を読出専用と標示することができる。

【００５９】

コーディネーター－スレーブ間の通信のための共有ディスクの使用

クエリが並列化されると、１つのノードがしばしば、クエリの実行に参加するさまざまなスレーブを調整する役割を果たす。そのノードは、しばしば「コーディネータ」と呼ばれ、参加しているスレーブによって生成される一時結果を受取る。コーディネータがクエリスレーブから受取るデータの量は、かなり多いことが考えられる。したがって、この発明の局面に従うと、クエリスレーブからコーディネータに通信されるべき一時結果は、通信されるべきデータの量がノード間の相互接続を混乱させる場合、共有ディスクを介して通信される。具体的には、スレーブによって生成される一時結果の量が或る一定のしきい値を超えた場合、それらの結果は、スレーブが存在するノードの揮発性メモリからコーディネータが存在するノードの揮発性メモリに直接送られる代わりに、スレーブとコーディネータとの間で共有されるディスクに書込まれる。この態様で共有ディスクを媒介として使用することは、子の演算子が完了するまで演算子の消費者が待たなければならない（すなわち、演算子間のパイプラインが存在しない）場合に、演算子を阻止するのに特に有用である。

【００６０】

ハードウェアの概観

図２は、この発明の一実施例が実現され得るコンピュータシステム２００を示すブロック図である。コンピュータシステム２００は、バス２０２または情報を通信するための他の通信機構と、バス２０２に結合されて情報を処理するためのプロセッサ２０４とを含む。コンピュータシステム２００は、バス２０２に結合されてプロセッサ２０４が実行する命令および情報を記憶するためのメインメモリ２０６、たとえばランダムアクセスメモリ（ＲＡＭ）または他の動的記憶装置も含む。メインメモリ２０６は、プロセッサ２０４が実行する命令の実行中に、一時的数値変数または他の中間情報を記憶するためにも使用可能である。コンピュータシステム２００は、バス２０２に結合されてプロセッサ２０４に対する静的情報および命令を記憶するための読出専用メモリ（ＲＯＭ）２０８または他の静的記憶装置をさらに含む。磁気ディスクまたは光学ディスク等の記憶装置２１０が設けられてバス２０２に結合され、情報および命令を記憶する。

【００６１】

コンピュータシステム２００は、コンピュータユーザに情報を表示するためのディスプレイ２１２、たとえば陰極線管（ＣＲＴ）に、バス２０２を介して結合され得る。英数字キーおよび他のキーを含む入力装置２１４がバス２０２に結合されて、情報および指令選択をプロセッサ２０４に通信する。別の種類のユーザ入力装置が、方向情報および指令選択をプロセッサ２０４に通信してディスプレイ２１２上のカーソルの動作を制御するためのカーソル制御機器２１６、たとえばマウス、トラックボール、またはカーソル方向キーである。この入力装置は一般に、２つの軸、すなわち第１の軸（ｘ等）および第２の軸（ｙ等）において２自由度を有し、これによって入力装置は平面上で位置を特定することができる。

【００６２】

この発明は、この明細書に記載された技術を実現するためにコンピュータシステム２００を用いることに関する。この発明の一実施例によると、これらの技術は、メインメモリ２０６に含まれる１つ以上の命令の１つ以上のシーケンスをプロセッサ２０４が実行することに応じて、コンピュータシステム２００により実行される。このような命令は、別のコンピュータ読取可能な媒体、たとえば記憶装置２１０からメインメモリ２０６内に読出すことができる。メインメモリ２０６に含まれる命令のシーケンスを実行することにより、プロセッサ２０４はこの明細書に記載されたプロセスのステップを実行する。代替的な実施例では、ソフトウェア命令の代わりに、またはソフトウェア命令と組合せて結線回路

を用いて、この発明を実施することができる。したがって、この発明の実施例は、ハードウェア回路およびソフトウェアのいずれかの特定の組合せに限定されない。

【００６３】

この明細書で用いられる「コンピュータ読取可能な媒体」という用語は、プロセッサ２０４に対して実行のために命令を提供することに携わる、いずれかの媒体を指す。このような媒体は、不揮発性媒体、揮発性媒体、および伝送媒体を含む多くの形態を取り得るが、これらに限定されない。不揮発性媒体には、たとえば記憶装置２１０等の光学または磁気ディスクが含まれる。揮発性媒体には、メインメモリ２０６等の動的メモリが含まれる。伝送媒体には、同軸ケーブル、銅線、および光ファイバが含まれ、バス２０２を有するワイヤが含まれる。伝送媒体は、電波データ通信および赤外線データ通信の際に生成されるもの等の音波または光波の形を取り得る。

【００６４】

コンピュータ読取可能な媒体の一般的な形態には、たとえばフロッピー（登録商標）ディスク、フレキシブルディスク、ハードディスク、磁気テープ、他のいずれかの磁気媒体、ＣＤ－ＲＯＭ、他のいずれかの光学媒体、パンチカード、紙テープ、孔のパターンを有する他のいずれかの物理的媒体、ＲＡＭ、ＰＲＯＭ、ＥＰＲＯＭ、ＦＬＡＳＨ－ＥＰＲＯＭ、他のいずれかのメモリチップもしくはカートリッジ、以下に述べる搬送波、またはコンピュータが読出すことのできる他のいずれかの媒体が含まれる。

【００６５】

プロセッサ２０４に対して実行のために１つ以上の命令の１つ以上のシーケンスを搬送することに対し、コンピュータ読取可能な媒体のさまざまな形態が関与し得る。たとえば、命令は、最初に遠隔コンピュータの磁気ディスクで搬送され得る。遠隔コンピュータはそれらの命令をそれ自体の動的メモリにロードして、それらの命令を、モデムを用いて電話回線経由で送信することができる。コンピュータシステム２００に対してローカルなモデムが電話回線上のデータを受信して、赤外線送信機を用いてそのデータを赤外線信号に変換することができる。赤外線信号によって搬送されたデータは赤外線検出器によって受信され得、適切な回路がそのデータをバス２０２上に出力することができる。バス２０２はそのデータをメインメモリ２０６に搬送し、そこからプロセッサ２０４が命令を取り出して実行する。メインメモリ２０６が受信した命令は、プロセッサ２０４による実行前または実行後のいずれかに、記憶装置２１０に任意に記憶され得る。

【００６６】

コンピュータシステム２００は、バス２０２に結合された通信インターフェイス２１８も含む。通信インターフェイス２１８は、ローカルネットワーク２２２に接続されたネットワークリンク２２０に対する双方向のデータ通信結合を提供する。たとえば通信インターフェイス２１８は、対応する種類の電話回線に対するデータ通信接続を設けるためのサービス統合デジタル網（ＩＳＤＮ）カードまたはモデムであり得る。別の例として、通信インターフェイス２１８は、互換性を有するローカルエリアネットワーク（ＬＡＮ）にデータ通信接続を設けるためのＬＡＮカードであり得る。無線リンクもまた実現することができる。このようなどの実現例においても、通信インターフェイス２１８は、さまざまな種類の情報を表わすデジタルデータストリームを搬送する電気信号、電磁信号、または光信号を送受信する。

【００６７】

ネットワークリンク２２０は一般に、１つ以上のネットワーク経由で他のデータ装置に対してデータ通信を提供する。たとえば、ネットワークリンク２２０は、ローカルネットワーク２２２経由で、ホストコンピュータ２２４か、またはインターネットサービスプロバイダ（Internet Service Provider）（ＩＳＰ）２２６により運営されるデータ装置に接続を提供することができる。ＩＳＰ２２６は次いで、現在一般に「インターネット」２２８と呼ばれるワールドワイドパケットデータ通信網を介してデータ通信サービスを提供する。ローカルネットワーク２２２およびインターネット２２８はともに、デジタルデータストリームを搬送する電気信号、電磁信号、または光信号を用いる。さまざまなネット

ワークを経由する信号と、ネットワークリンク２２０上の、または、通信インターフェイス２１８経由の信号とは、コンピュータシステム２００との間でデジタルデータを搬送し、情報を運ぶ搬送波の例示的形態である。

【００６８】

コンピュータシステム２００は、ネットワーク、ネットワークリンク２２０、および通信インターフェイス２１８を介してメッセージを送信して、プログラムコードを含むデータを受信することができる。インターネットの例では、サーバ２３０は、インターネット２２８、ＩＳＰ２２６、ローカルネットワーク２２２、および通信インターフェイス２１８経由で、アプリケーションプログラムに対して要求されたコードを送信することができる。

【００６９】

受信されたコードは、受信されたときにプロセッサ２０４によって実行され得、および／または後の実行のために記憶装置２１０もしくは他の不揮発性記憶装置に記憶され得る。このようにして、コンピュータシステム２００は搬送波の形でアプリケーションコードを得ることができる。

【００７０】

上述の明細書では、この発明の実施例を実現例ごとに異なり得る多数の特定の詳細を参照して説明してきた。したがって、この発明が何であるか、およびこの発明を目指して出願人が何を意図しているかを排他的に示す唯一のものが、この出願から発生して特有の形態をとった一組の請求項である。特有の形態においてこのような請求項は、今後のどのような補正をも含んで発生する。このような請求項に含まれる用語に対してここで明示されたどのような定義も、請求項で用いられる用語の意味を決定するものとする。したがって、請求項に明示的に記載されていない限定、要素、特性、特徴、利点または属性は、このような請求項の範囲を決して限定しない。したがって、明細書および図面は限定的な意味ではなく例示的な意味で捉えられるべきである。

【図面の簡単な説明】

【００７１】

【図１】この発明の一実施例に従った、２つの共有ディスクサブシステムを含むクラスタを示すブロック図である。

【図２】この発明の実施例が実現され得るコンピュータシステムのブロック図である。

【図１】

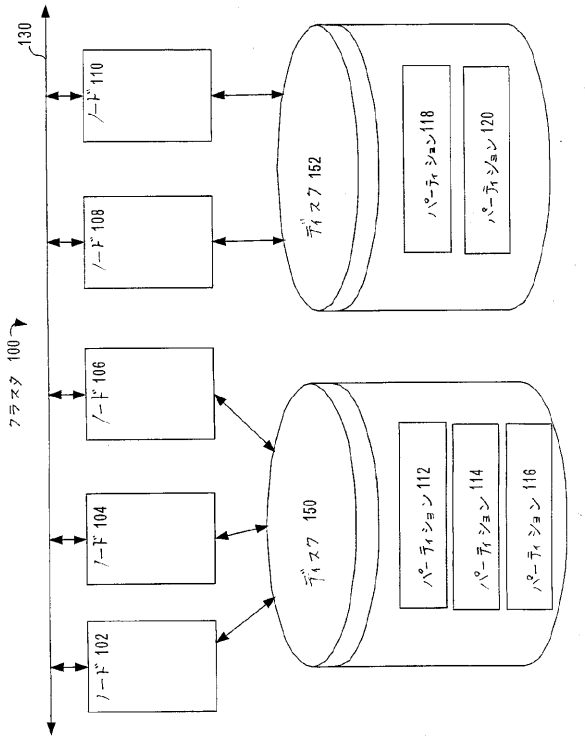


FIG. 1

【図２】

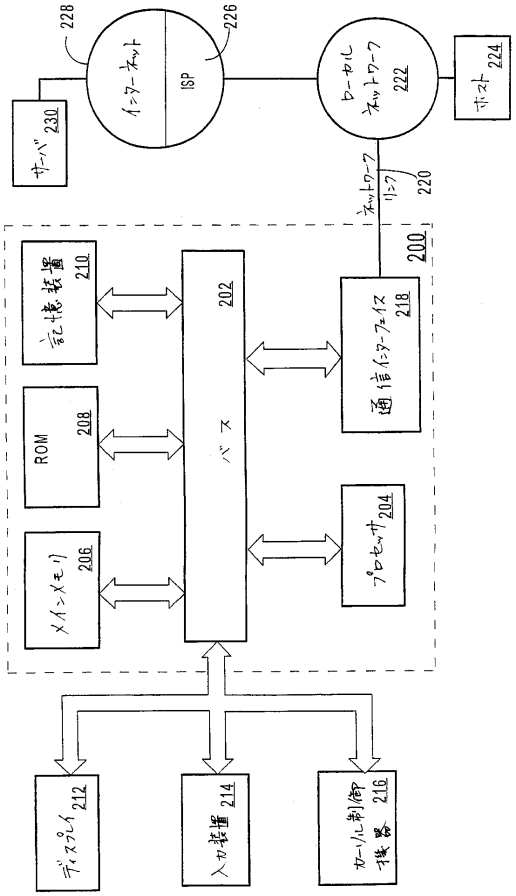


FIG. 2

【手続補正書】
【提出日】平成17年10月31日(2005.10.31)
【手続補正１】
【補正対象書類名】特許請求の範囲
【補正対象項目名】全文
【補正方法】変更
【補正の内容】
【特許請求の範囲】
【請求項１】
データを管理するための方法であって、
複数のノードにとってアクセス可能な永続的な記憶上に複数の永続的なデータ項目を保存するステップを含み、前記永続的な記憶は、複数のノードにとってアクセス可能であり、前記永続的なデータ項目は、前記永続的な記憶上の特定の位置に記憶された特定のデータ項目を含み、前記方法はさらに、
前記ノードの１つに、前記永続的なデータ項目の各々の排他的な所有権を割当てるステップを含み、前記複数のノードの特定のノードには、前記特定のデータ項目の排他的な所有権が割当てられ、
いずれかのノードが前記特定のデータ項目を必要とする演算が実行されることを望むと、前記演算が実行されることを望む前記ノードは、前記特定のデータ項目が前記特定のノードによって排他的に所有されている時に、前記特定のノードが前記特定のデータ項目についての前記演算を実行するように前記特定のノードに前記演算を転送し、前記方法はさらに、
システム性能および作業負荷の少なくとも１つに関する情報を収集することによって統計を収集するステップと、
システム性能およびスループットの少なくとも１つを改善するために、前記統計に基づ

いて前記永続的なデータ項目の所有権を動的に再割当するステップとを含む、方法。
【請求項２】
情報を収集する前記ステップは、前記永続的なデータ項目を必要とする演算をどのノードが要求しているかを監視するステップを含む、請求項１に記載の方法。
【請求項３】
前記複数のノードは、マルチノードデータベースシステムのノードである、請求項１に記載の方法。
【請求項４】
動的に再割当する前記ステップは、前記永続的なデータ項目が再割当されるべきノードを指定するユーザ入力を受取ることなく実行される、請求項１に記載の方法。
【請求項５】
動的に再割当する前記ステップは、前記マルチノードデータベースシステムがデータベースアプリケーションからのデータベース指令の処理を継続する間に実行される、請求項３に記載の方法。
【請求項６】
動的に再割当する前記ステップは、或る一定の期間の後に実行され、前記或る一定の期間の間に前記永続的なデータ項目を必要とする演算をどのノードが要求したかに基づく、請求項１に記載の方法。
【請求項７】
動的に再割当する前記ステップは、前記特定のデータ項目を必要とする演算を最も頻繁に要求したノードに対し、特定のデータ項目の所有権を再割当するステップを含む、請求項６に記載の方法。
【請求項８】
動的に再割当する前記ステップは、第１のノードに前記特定のデータ項目の所有権を再割当するステップを含み、
前記方法はさらに、前記第１のノードに前記特定のデータ項目の所有権を再割当することに応じて、前記第１のノードから第２のノードに第２のデータ項目の所有権を動的に再割当するステップを含む、請求項１に記載の方法。
【請求項９】
前記第１のノードに前記特定のデータ項目を割当てることにより、前記第１のノードに関連するしきい値が上回ることが生じ、
前記第２のデータ項目の所有権を動的に再割当する前記ステップは、前記しきい値が上回ったことに応じて実行される、請求項８に記載の方法。
【請求項１０】
データを管理するための方法であって、
複数のノードにとってアクセス可能な永続的な記憶上に複数の永続的なデータ項目を保存するステップを含み、前記永続的なデータ項目は、前記永続的な記憶上の特定の位置に記憶された特定のデータ項目を含み、前記方法はさらに、
前記ノードの１つに、前記永続的なデータ項目の各々の排他的な所有権を割当てるステップを含み、前記複数のノードの第１のノードには、前記特定のデータ項目の排他的な所有権が割当てられ、
いずれかのノードが前記特定のデータ項目を必要とする演算が実行されることを望むと、前記演算が実行されることを望む前記ノードは、前記特定のデータ項目が前記第１のノードによって排他的に所有されている時に、前記第１のノードが前記特定のデータ項目についての演算を実行するように第１のノードに前記演算を転送し、前記方法はさらに、
前記特定のデータ項目の排他的な所有権が前記第１のノードによって保持されている間に、前記特定のデータ項目についての演算が実行されることを要求する指令を受信するステップと、
前記第１のノードとは異なる第２のノードによって前記演算が実行されるようにするステップとを含む、方法。

【請求項１１】
第２のノードによって前記演算が実行されるようにする前記ステップは、
前記第２のノードが前記指令についてのサブ演算を実行するのに必要とされる時間と少なくとも同じ長さの時間に、前記特定のデータ項目の排他的な所有権を第２のノードに一時的に再割当するステップを含み、前記サブ演算は前記特定のデータ項目を必要とし、第２のノードによって演算が実行されるようにする前記ステップはさらに、
前記時間の後に、前記特定のデータ項目の前記排他的な所有権を前記第１のノードに再び自動的に再割当するステップを含む、請求項１０に記載の方法。
【請求項１２】
一時的に再割当する前記ステップは、並列化された演算のサブ演算が、複数のノードに存在するスレーブ間に分配されるように実行される、請求項１１に記載の方法。
【請求項１３】
一時的に再割当する前記ステップは、前記指令によって要求される演算が、前記第２のノードを含むものの前記第１のノードを含まない一組の１つ以上のノードにおいて統合されるように実行される、請求項１１に記載の方法。
【請求項１４】
第２のノードによって演算が実行されるようにする前記ステップは、
前記第２のノードが前記特定のデータ項目の排他的な所有権を獲得することなく前記第２のノードに演算を実行させるステップを含み、
前記第２のノードは、前記指令のサブ演算を実行するために前記特定のデータ項目にアクセスすることが許可され、前記サブ演算は前記特定のデータ項目を必要とし、第２のノードによって演算が実行されるようにする前記ステップはさらに、
前記第２のノードが前記サブ演算の実行を完了した後に、前記第２のノードに前記特定のデータ項目にアクセスすることを許可するのを中止するステップを含む、請求項１０に記載の方法。
【請求項１５】
非コミット読出分離レベルを前記指令に適用することを決定するステップと、
前記第１のノードが前記特定のデータ項目のいずれかのダーティなバージョンをディスクにフラッシュすることを要求せずに前記第２のノードに前記サブ演算を実行させるステップとをさらに含む、請求項１４に記載の方法。
【請求項１６】
コミット済読出分離レベルを前記指令に適用することを決定するステップと、
前記第１のノードが前記特定のデータ項目のいずれかのダーティなバージョンをディスクにフラッシュするまで、前記第２のノードが前記サブ演算を実行することを防止するステップとをさらに含む、請求項１４に記載の方法。
【請求項１７】
１つ以上のプロセッサによって実行されると、前記１つ以上のプロセッサに請求項１～１６のいずれかに記載の方法を実行させる命令の１つ以上のシーケンスを搬送する、コンピュータ読取可能な媒体。

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International Application No
PCT/US2004/024555

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G06F9/46

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)
IPC 7 G06F

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

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal, INSPEC

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 5 675 791 A (BHIDE ET AL) 7 October 1997 (1997-10-07) column 2, line 62 - column 3, line 16 column 5, line 23 - column 7, line 22	1-32
X	WO 97/04384 A (EMC CORPORATION) 6 February 1997 (1997-02-06) abstract page 8, line 28 - page 9, line 7 page 11, line 9 - line 13 page 13, line 9 - page 15, line 5	1-10, 17-26
A	US 5 539 883 A (ALLON ET AL) 23 July 1996 (1996-07-23) column 3, line 32 - line 38	10,26
	-/--	

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in 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 document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
22 July 2005	03/08/2005

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Michel, T

Form PCT/ISA/210 (second sheet) (January 2004)

INTERNATIONAL SEARCH REPORT

International Application No
PCT/US2004/024555

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CA 2 435 388 A1 (ORACLE INTERNATIONAL CORPORATION) 9 January 2003 (2003-01-09) abstract page 11, line 17 - page 13, line 12; figure 1 page 3, line 8 - line 15	1,10,17, 26
A	WO 02/073416 A (ORACLE INTERNATIONAL CORPORATION) 19 September 2002 (2002-09-19) the whole document	15,16, 31,32
A	WO 99/44130 A (SUN MICROSYSTEMS, INC) 2 September 1999 (1999-09-02) claim 1	11,27

Form PCT/ISA/210 (continuation of second sheet) (January 2004)

INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No
PCT/US2004/024555

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 5675791 A	07-10-1997	US 5625811 A	29-04-1997
WO 9704384 A	06-02-1997	US 6173306 B1	09-01-2001
		US 5860137 A	12-01-1999
		DE 69619531 D1	04-04-2002
		DE 69619531 T2	31-10-2002
		EP 0842464 A1	20-05-1998
		JP 11509659 T	24-08-1999
		JP 3459263 B2	20-10-2003
		WO 9704384 A1	06-02-1997
US 5539883 A	23-07-1996	DE 69227665 D1	07-01-1999
		DE 69227665 T2	22-07-1999
		EP 0540151 A2	05-05-1993
		JP 2685067 B2	03-12-1997
		JP 5216845 A	27-08-1993
CA 2435388 A1	09-01-2003	WO 03003252 A1	09-01-2003
		EP 1399847 A1	24-03-2004
		EP 1521184 A2	06-04-2005
		JP 2004531005 T	07-10-2004
		US 2001039550 A1	08-11-2001
WO 02073416 A	19-09-2002	US 2002099729 A1	25-07-2002
		CA 2438262 A1	12-09-2002
		CA 2440277 A1	19-09-2002
		CN 1524226 A	25-08-2004
		CN 1496510 A	12-05-2004
		EP 1412858 A2	28-04-2004
		EP 1366420 A2	03-12-2003
		JP 2005505808 T	24-02-2005
		JP 2005506598 T	03-03-2005
		WO 02073416 A2	19-09-2002
		WO 02071229 A2	12-09-2002
		US 2002095403 A1	18-07-2002
		US 2005065907 A1	24-03-2005
WO 9944130 A	02-09-1999	US 6247026 B1	12-06-2001
		AU 2680299 A	15-09-1999
		AU 2680399 A	15-09-1999
		AU 2680499 A	15-09-1999
		AU 2686699 A	15-09-1999
		AU 2686799 A	15-09-1999
		AU 2766199 A	15-09-1999
		AU 2769899 A	15-09-1999
		AU 2770199 A	15-09-1999
		AU 2770299 A	15-09-1999
		AU 2770399 A	15-09-1999
		AU 2770499 A	15-09-1999
		AU 2770599 A	15-09-1999
		AU 2787699 A	15-09-1999
		AU 2787799 A	15-09-1999
		AU 2787899 A	15-09-1999
		AU 2876899 A	15-09-1999
		AU 2876999 A	15-09-1999
		AU 2878399 A	15-09-1999
		AU 2878499 A	15-09-1999
		AU 3297199 A	15-09-1999

INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No
PCT/US2004/024555

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 9944130 A		AU 3297299 A	15-09-1999
		AU 3300499 A	15-09-1999
		AU 3300599 A	15-09-1999
		AU 3309199 A	15-09-1999
		CN 1292115 A	18-04-2001
		CN 1292116 A	18-04-2001
		CN 1292113 A	18-04-2001
		CN 1292117 A	18-04-2001
		CN 1292192 A	18-04-2001
		CN 1292118 A	18-04-2001
		CN 1298502 A	06-06-2001
		CN 1298503 A	06-06-2001
		CN 1298504 A	06-06-2001
		CN 1298505 A	06-06-2001
		CN 1298523 A	06-06-2001
		CN 1298524 A	06-06-2001
		CN 1298506 A	06-06-2001
		CN 1298507 A	06-06-2001
		CN 1298508 A	06-06-2001
		CN 1298509 A	06-06-2001
		CN 1298510 A	06-06-2001
		CN 1298511 A	06-06-2001
		CN 1298512 A	06-06-2001
		CN 1298513 A	06-06-2001
		CN 1298514 A	06-06-2001
		CN 1298515 A	06-06-2001
		CN 1298516 A	06-06-2001
		CN 1298525 A	06-06-2001
		DE 69903711 D1	05-12-2002

フロントページの続き

(81)指定国 AP(BW,GH,GM,KE,LS,MW,MZ,NA,SD,SL,SZ,TZ,UG,ZM,ZW),EA(AM,AZ,BY,KG,KZ,MD,RU,TJ,TM),EP(AT,BE,BG,CH,CY,CZ,DE,DK,EE,ES,FI,FR,GB,GR,HU,IE,IT,LU,MC,NL,PL,PT,RO,SE,SI,SK,TR),OA(BF,BJ,CF,CG,CI,CM,GA,GN,GQ,GW,ML,MR,NE,SN,TD,TG),AE,AG,AL,AM,AT,AU,AZ,BA,BB,BG,BR,BW,BY,BZ,CA,CH,CN,CO,CR,CU,CZ,DE,DK,DM,DZ,EC,EE,EG,ES,FI,GB,GD,GE,GH,GM,HR,HU,ID,IL,IN,IS,JP,KE,KG,KP,KR,KZ,LC,LK,LR,LS,LT,LU,LV,MA,MD,MG,MK,MN,MW,MX,MZ,NA,NI,NO,NZ,OM,PG,PH,PL,PT,RO,RU,SC,SD,SE,SG,SK,SL,SY,TJ,TM,TN,TR,TT,TZ,UA,UG,US,UZ,VC,VN,YU,ZA,ZM,ZW

(74)代理人 100098316
弁理士 野田 久登
(74)代理人 100109162
弁理士 酒井 將行
(72)発明者 バンフォード, ロジャー・ジェイ
アメリカ合衆国、９４０６２ カリフォルニア州、ウッドサイド、マンザニータ・ウェイ、５５５
(72)発明者 チャンドラセカラン, サシカンス
アメリカ合衆国、９４００２ カリフォルニア州、ベルモント、カールモント・ドライブ、２５４５、ナンバー・２４
(72)発明者 プルスチーノ, アンジェロ
アメリカ合衆国、９４０２２ カリフォルニア州、ロス・アルトス、ディステル・ドライブ、４３６

Ｆターム(参考) 5B065 BA01 EK07 ZA15
5B082 FA17 GA03