(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4939440号
(P4939440)
(45)発行日　平成24年5月23日(2012.5.23)　(24)登録日　平成24年3月2日(2012.3.2)

(51)Int.Cl.　F I
G06F 12/00 (2006.01)　G06F 12/00 518A
G06F 12/00 531D
G06F 12/00 533J

請求項の数 23　(全 14 頁)

(21)出願番号　特願2007-556381 (P2007-556381)
(86)(22)出願日　平成18年2月17日(2006.2.17)
(65)公表番号　特表2008-530716 (P2008-530716A)
(43)公表日　平成20年8月7日(2008.8.7)
(86)国際出願番号　PCT/US2006/005909
(87)国際公開番号　WO2006/089263
(87)国際公開日　平成18年8月24日(2006.8.24)
審査請求日　平成21年1月27日(2009.1.27)
(31)優先権主張番号　11/061,152
(32)優先日　平成17年2月18日(2005.2.18)
(33)優先権主張国　米国(US)

(73)特許権者　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】データベースシステムにおいて報告トランザクションを処理する方法およびメカニズム

(57)【特許請求の範囲】
【請求項１】
データベースシステムにおいて報告トランザクションを処理するコンピュータによって実行される方法であって、プロセッサを用いて以下の処理を実行することを含み、当該以下の処理は、
データベースのスナップショットを取得することを備え、前記データベースは、プライマリノードおよび前記データベースの複製されたデータベースを用いないフェイルオーバノードにリンクされ、前記処理はさらに、
前記プライマリノードで１つ以上の非報告トランザクションを実行することを備え、前記１つ以上の非報告トランザクションは、１つ以上の第１のデータベースクエリを、前記１つ以上の第１のデータベースクエリの実行において前記データベースの最新の更新を用いることなく実行し、前記処理はさらに、
前記プライマリノードで前記１つ以上の非報告トランザクションを実行するのと同時に、前記フェイルオーバノードで報告トランザクションを実行するために前記スナップショットを利用することを備え、
前記報告トランザクションは、１つ以上の第２のデータベースクエリを、前記１つ以上の第２データベースクエリの実行において前記データベースの最新の更新を用いて実行し、
前記プライマリノードのトランザクションは、前記データベースを修正することを許可される一方で、前記フェイルオーバノードのトランザクションは、前記データベースを直

接的に修正することを許可されない、コンピュータによって実行される方法。
【請求項２】
前記フェイルオーバノードで１つ以上の一時的なテーブルを作成することをさらに備え、前記１つ以上の一時的なテーブルは、前記報告トランザクションが前記フェイルオーバノードで実行されるときに使用される、請求項１に記載のコンピュータによって実行される方法。
【請求項３】
前記１つ以上の一時的なテーブルは、前記報告トランザクションにおけるクエリスクリプトを通じて作成される、請求項２に記載のコンピュータによって実行される方法。
【請求項４】
前記１つ以上の一時的なテーブルのうち少なくとも１つは、前記報告トランザクションにおける２つ以上のクエリにアクセス可能である、請求項２に記載のコンピュータによって実行される方法。
【請求項５】
前記データベースにおける１つ以上のスキーマを修正することをさらに備え、前記１つ以上のスキーマは、前記報告トランザクションが前記フェイルオーバノードで実行されるときに使用される、請求項１に記載のコンピュータによって実行される方法。
【請求項６】
前記１つ以上のスキーマは、前記プライマリノードで実行する前記１つ以上の非報告トランザクションにアクセス可能ではない、請求項５に記載のコンピュータによって実行される方法。
【請求項７】
前記１つ以上のスキーマのうち少なくとも１つは１つ以上のテーブルを含む、請求項５に記載のコンピュータによって実行される方法。
【請求項８】
前記プライマリノードで１つ以上のユーザ定義プロシージャにアクセスすることをさらに備え、前記１つ以上のユーザ定義プロシージャは、前記報告トランザクションが前記フェイルオーバノードで実行されるときに使用される、請求項１に記載のコンピュータによって実行される方法。
【請求項９】
前記データベースに一時的なスペースを確保することをさらに備え、前記一時的なスペースは、前記報告トランザクションが前記フェイルオーバノードで実行されるときに使用される、請求項１に記載のコンピュータによって実行される方法。
【請求項１０】
前記プライマリノードおよび前記フェイルオーバノードはクラスタの一部である、請求項１に記載のコンピュータによって実行される方法。
【請求項１１】
前記クラスタは１つ以上のさらなるフェイルオーバノードを含む、請求項１０に記載のコンピュータによって実行される方法。
【請求項１２】
前記１つ以上の非報告トランザクションのうち少なくとも１つはリード・ライトトランザクションである、請求項１に記載のコンピュータによって実行される方法。
【請求項１３】
前記報告トランザクションおよび前記１つ以上の非報告トランザクションはワークロードの一部である、請求項１に記載のコンピュータによって実行される方法。
【請求項１４】
前記報告トランザクションはリアルタイムに近い報告を与える、請求項１に記載のコンピュータによって実行される方法。
【請求項１５】
前記プライマリノードのみが前記データベースを修正できる、請求項１に記載のコンピ

ュータによって実行される方法。
【請求項１６】
前記スナップショットはユーザコマンドに応答して取得される、請求項１に記載のコンピュータによって実行される方法。
【請求項１７】
前記スナップショットはリード・オンリである、請求項１に記載のコンピュータによって実行される方法。
【請求項１８】
前記スナップショットは前記プライマリノードによって修正されることができない、請求項１に記載のコンピュータによって実行される方法。
【請求項１９】
前記スナップショットおよび前記データベースはディスクスペースを共有する、請求項１に記載のコンピュータによって実行される方法。
【請求項２０】
前記スナップショットは最新のものである、請求項１に記載のコンピュータによって実行される方法。
【請求項２１】
前記スナップショットは、前記フェイルオーバノードで前記報告トランザクションを実行するために直接に使用される、請求項１に記載のコンピュータによって実行される方法。
【請求項２２】
コンピュータによって実行されることで、請求項１～２１のいずれかに記載の方法を実行する、プログラム。
【請求項２３】
請求項１～２１のいずれかに記載の方法を実行する、システム。
【発明の詳細な説明】
【技術分野】
【０００１】
背景および概要
この発明はデータベースシステムに関する。より詳細には、この発明は、データベースシステムにおいて報告トランザクション（reporting transaction）を処理する方法およびメカニズムに向けられる。
【背景技術】
【０００２】
多くのデータベースシステムは、ペースの速い今日の市場において極めて重要である高可用性を保証するためにフェイルオーバクラスタを利用する。フェイルオーバクラスタでは、データベースはプライマリノードおよび少なくとも１つのフェイルオーバノード（スペアノードとしても公知である）にリンクされる。データベースおよびウェブサーバなどのアプリケーションは、誤動作するまでプライマリノードで動作する。誤動作が発生すると、アプリケーションはフェイルオーバノードで再開される。フェイルオーバノードおよびプライマリノードが単一のクラスタに属しているので、プライマリノードの故障を検出するために標準的なハートビートメカニズムを使用できる。
【発明の開示】
【発明が解決しようとする課題】
【０００３】
フェイルオーバクラスタに関する１つの問題は、フェイルオーバノードをプライマリノードと同時に使用できないことである。したがって、プライマリハードウェアが故障したときにのみ使用される追加のハードウェアを購入するコストを正当化することは困難であり得る。ある特定の並列データベースシステムは、２つ以上のノードがクラスタにおけるデータベースに同時にアクセスできるアクティブ／アクティブクラスタを利用することに

よってこの問題を解決する。しかしながら、アクティブ／アクティブクラスタは、クラスタにおけるすべてのノードからの同時の読取および修正が存在する状態でデータベースが確実に一貫性があるようにするために複雑な並行処理制御メカニズムを必要とする。

【０００４】

ユーザが直面する別の問題は、報告トランザクションが他のトランザクションと同時に実行される混合ワークロードを動作させる必要があることである。理想的には、リアルタイムの報告は各報告トランザクションによって与えられる。すなわち、最新の更新からの結果はトランザクションにおけるクエリによって使用される。さらに、ユーザは、非報告トランザクション（non-reporting transaction）と報告トランザクションとの間の（たとえば、ＣＰＵまたはメモリについての）ハードウェアリソースの競合を回避するために、別個に報告トランザクションを動作させることを好む。

【０００５】

アクティブ／アクティブクラスタリングをサポートしないデータベースシステムでは、報告のために複製データベースが作成され、使用され得る。しかしながら、複製データベースがプライマリデータベースの完全なコピーであるので、この解決法は格納コストを２倍にする。さらに、複製データベースはしばしばプライマリデータベースに遅れをとる。なぜなら、プライマリデータベースにおける変更を瞬時に複製することが実現可能でない可能性があるためである。たとえ瞬時の複製が実現可能であったとしても、プライマリデータベースでのすべてのコミットが同期して報告データベースに複製される必要があるだろうという理由で、プライマリデータベースのスループットは大幅に影響を受けるであろう。

【０００６】

したがって、フェイルオーバクラスタを利用するデータベースシステムにおいて報告トランザクションを実行することに関するこれらのおよび他の問題に方法およびメカニズムが対処する必要がある。

【課題を解決するための手段】

【０００７】

この発明の実施例は、データベースシステムにおいて報告トランザクションを処理するための改良された方法、システムおよび媒体を提供する。実施例によれば、データベースのスナップショットが取得される。データベースはプライマリノードおよびフェイルオーバノードにリンクされる。次いで１つ以上の非報告トランザクションがプライマリノードで実行され、プライマリノードで１つ以上の非報告トランザクションを実行するのと同時にフェイルオーバノードで報告トランザクションを実行するためにスナップショットが利用される。

【発明を実施するための最良の形態】

【０００８】

この発明の局面、目的および利点のさらなる詳細について、詳細な説明、図面および特許請求の範囲において以下で説明する。先の一般的な説明および以下の詳細な説明の両方は例示的および説明的なものであり、この発明の範囲に関して限定的であるように意図されるものではない。

【０００９】

添付の図面は、この発明をさらに理解できるようにするために含まれ、詳細な説明とともにこの発明の原理を説明するのに役立つ。

【００１０】

詳細な説明

データベースシステムにおける報告トランザクションの処理を開示する。複雑な整合性およびルーティングメカニズムを必要とするアクティブ／アクティブクラスタを利用するか、または追加のハードウェアの購入を必然的に伴い、潜在的にデータが古い別個の複製データベースを有するのではなく、報告トランザクションは、プライマリノードで動作する非報告トランザクションと同時に、データベーススナップショットを使用して、フェイ

ルオーバノードで実行される。これは、そうでなければアイドルのままであろうフェイルオーバノードを利用し、最新のスナップショットが使用されるときにリアルタイムに近い報告を与える。

【００１１】

データベースシステムにおいて報告トランザクションを処理する方法を図１に示す。１０２において、データベースのスナップショットが取得される。データベースはプライマリノードおよびフェイルオーバノードにリンクされる。いくつかの実施例では、プライマリノードのみがデータベースを修正することを許可される。クライアント接続部は、すべての報告トランザクションをフェイルオーバノードに向け、すべての他のトランザクションをプライマリノードに向けるよう構成され得るであろう。フェイルオーバノードが、場合によってはデータベースを修正し得るであろうトランザクションを自動的にプライマリノードにルーティングすることも可能であり得る。このルーティングは、セッションがデータベースを修正することになるかどうかを識別するトランザクションにリード・ライト（READ-WRITE）またはリード・オンリ（READ-ONLY）という印をつけることによってなされ得る。

【００１２】

次いで１つ以上の非報告トランザクションがプライマリノードで実行され（１０４）、プライマリノードで１つ以上の非報告トランザクションを実行するのと同時にフェイルオーバノードで報告トランザクションを実行するためにスナップショットが利用される（１０６）。報告トランザクションおよび非報告トランザクションの各々は、１つ以上のクエリを備える。そして、非報告トランザクションはリード・ライトまたはリード・オンリトランザクションであってもよいが、報告トランザクションは通常リード・オンリトランザクションである。

【００１３】

スナップショットは、データベースのある時点のコピーであり、スナップショットが取得された後に修正されるデータベースブロックを除いて、データベースと同一のディスクスペースを共有する。これは、スナップショットが修正されないままであるように、変更されたブロックが新しい場所に書込まれる標準的なコピー・オン・ライトメカニズムによって達成されることができる。スナップショットがリード・オンリであり、プライマリノードによって修正されることができないので、フェイルオーバノードで動作するクエリは、プライマリノードとの調整を必要とすることなく、使用されるスナップショットと一貫性のある結果を返すことになる。そして、スナップショットが一貫性があり、データベース全体のためのものである（すなわち、クエリにおいて参照されるスナップショットおよびテーブルの中の索引がすべて一貫性がある）ので、既存のクエリ実行エンジンは修正される必要がない。さまざまなスナップショット方法論が利用可能であり、ファイル、アプリケーション、システムまたはデータベースレベルで実現され得る。たとえば、ファイルレベルのスナップショットの作成についての説明は、http://www.netapp.com/tech library/3002.htmlにおいて見ることができる。

【００１４】

スナップショットは、すべての変更されないデータについてデータベースと同一のディスク記憶装置を使用するので、ディスクスペースおよびＣＰＵ使用率の両方の点で比較的安価に作成される。したがって、データベースシステムはかなり頻繁に、たとえば１０秒ごとにスナップショットを取得するよう構成され得る。しかしながら、ユーザコマンドに応答して、たとえば報告セッションまたは他のこのようなメトリクスによって所望されるサービスの質に基づいてデータベースシステムがスナップショットを生成することも可能である。最新のスナップショットを使用してフェイルオーバノードで報告トランザクションを実行することにより、リアルタイムに近い報告が与えられることになる。なぜなら、最新の更新は報告トランザクションにおけるクエリによって使用されることになるためである。しかしながら、ユーザは取得された最新のスナップショットよりも古いスナップショットの使用を指定することも許可され得る。

【００１５】
図２は、プライマリノード２０２、フェイルオーバノード２０４およびデータベース２０６を有するクラスタ２００を示す。データベース２０６のスナップショット２０８が取得されている。複数の非報告トランザクション２１０ａおよび２１０ｂがプライマリノード２０２で動作している間、スナップショット２０８はフェイルオーバノード２０４で報告トランザクション２１２を実行するために使用される。いくつかの実施例では、非報告トランザクション２１０ａおよび２１０ｂならびに報告トランザクション２１２はワークロードの一部である。
【００１６】
データベースシステムにおいて報告トランザクションを処理するための方法のプロセスフローを図３に示す。この実施例によれば、プライマリノードおよびフェイルオーバノードにリンクされるデータベースのスナップショットが取得される（３０２）。３０４において、１つ以上の非報告トランザクションがプライマリノードで実行される。プライマリノードで１つ以上の非報告トランザクションを実行するのと同時にフェイルオーバノードで報告トランザクションを実行するためにスナップショットが利用される（３０６）。次いで、報告トランザクションがフェイルオーバノードで実行されるときに１つ以上の一時的なテーブルが作成され、使用される（３０８）。
【００１７】
クラスタ４００を図４に示す。クラスタ４００は、プライマリノード４０２、フェイルオーバノード４０４およびデータベース４０６を含む。この例では、スナップショット４０８ａが取得され、非報告トランザクション４１０がプライマリノード４０２で動作している間にフェイルオーバノード４０４で報告トランザクション４１２を実行するために使用される。報告トランザクション４１２の実行中、一時的な結果を格納するためにトランザクション４１２におけるクエリスクリプトを通じて一時的なテーブル４１４ａおよび４１４ｂが作成される。これらの一時的なテーブル４１４ａおよび４１４ｂはプライマリノード４０２に透過的に送られ、プライマリノード４０２は次いで一時的なテーブル４１４ａおよび４１４ｂのためにデータベース４０６においてスペースを割当てる。フェイルオーバノード４０４において一時的なテーブル４１４ａおよび４１４ｂに後に保存される変更はプライマリノード４０２に送られる必要はない。
【００１８】
図４では、データベース４０６の新しいスナップショット４０８ｂが取得されて、報告トランザクション４１２における後続のクエリが一時的なテーブル４１４ａおよび４１４ｂにアクセスできるようにする。しかしながら、他の実施例では、作成されるすべての一時的なテーブルよりも少ない一時的なテーブルが後続のクエリによるアクセスのために保有されることになる。したがって、クエリの完了後、フェイルオーバノードは、テーブルのために割当てられたデータベーススペースを解放するために、一時的なテーブルを削除でき、削除部分をプライマリノードに送ることができる。
【００１９】
一貫性のある結果を保証するために、単一のクエリは通常同一のスナップショットを使用することになる。しかしながら、図４の例に見られるように、同一のセッションまたはトランザクション内の後続のクエリは、以前のクエリによって使用されたスナップショットと同一のスナップショットまたはそれよりも最近のスナップショットを使用してもよい。
【００２０】
データベースシステムにおいて報告トランザクションを処理する別の方法を図５に示す。５０２において、データベースのスナップショットが取得される。この実施例では、データベースはプライマリノードおよびフェイルオーバノードにリンクされる。次いで１つ以上の非報告トランザクションがプライマリノードで実行され（５０４）、プライマリノードで１つ以上の非報告トランザクションを実行するのと同時にフェイルオーバノードで報告トランザクションを実行するためにスナップショットが利用される（５０６）。５０

８において、報告トランザクションがフェイルオーバノードで実行されるときにデータベースにおける１つ以上のスキーマが修正され、使用される。１つ以上のスキーマは、プライマリノードで作成されていた可能性があり、フェイルオーバノードで報告トランザクションが使用するために「印をつけられていた」または「確保されていた」可能性がある。さらに、１つ以上のスキーマへの変更はプライマリノードと調整することなくなされ得る。

【００２１】

データベーススキーマはオブジェクトの集まりである。スキーマオブジェクトは、たとえばテーブル、ビュー、シーケンスおよびストアドプロシージャを含むが、それらに限定されない。テーブルは、概してデータベースにおける編成の基本単位であり、それぞれの行および列に格納されたデータを備える。ビューは、１つ以上のテーブルでのデータの特別仕立ての提示である。ビューは、データが基づいているテーブル、すなわちベーステーブルからデータを導き出す。さらには、ベーステーブルがテーブルである可能性もあれば、ベーステーブル自体がビューである可能性もある。ビューの一例は、テーブルからテーブルのデータの列のうち２列をマイナスしたものである。

【００２２】

シーケンスは、１つ以上のデータベーステーブルの数値列を識別する固有の数字の連続的なリストである。シーケンスは概して、単一のテーブルまたは複数のテーブルの行について固有の数値を自動的に生成することによってアプリケーションプログラミングを単純化する。シーケンスを使用することによって、２人以上のユーザが概して同時にテーブルにデータを入力できる。ストアドプロシージャは概して、特定のタスクを行なうために実行可能な単位としてともにグループ分けされる１組のコンピュータ命令文である。

【００２３】

図６は、プライマリノード６０２、２つのフェイルオーバノード６０４ａおよび６０４ｂ、ならびにデータベース６０６を有するクラスタ６００を示す。データベース６０６のスナップショット６０８が取得されている。この実施例では、スナップショット６０８を介してしかフェイルオーバノード６０４ａおよび６０４ｂに開いていない残りのデータベース６０６とは異なって、リード・ライトモードでデータベース６０６内のスキーマ６１４ａおよび６１４ｂをフェイルオーバノード６０４ａおよび６０４ｂが利用できる。この状況下で、スキーマ６１４ａおよび６１４ｂは、フェイルオーバノード６０４ａおよび６０４ｂで動作する報告トランザクション６１２ａおよび６１２ｂによってそれぞれに修正され得る。スキーマ６１４ａおよび６１４ｂに含まれるデータがフェイルオーバノード６０４ａ、６０４ｂとプライマリノード６０２との間で共有されないので、プライマリノード６０２で実行する非報告トランザクション６１０はデータベース６０６におけるスキーマ６１４ａおよび６１４ｂにアクセスできない。

【００２４】

データベースシステムにおいて報告トランザクションを処理するための方法のフローチャートを図７に示す。７０２において、プライマリノードおよびフェイルオーバノードにリンクされるデータベースのスナップショットが取得される。７０４において、１つ以上の非報告トランザクションがプライマリノードで実行される。次いで、プライマリノードで１つ以上の非報告トランザクションを実行するのと同時にフェイルオーバノードで報告トランザクションを実行するためにスナップショットが利用される（７０６）。

【００２５】

この実施例では、報告トランザクションがフェイルオーバノードで実行されるときに、プライマリノードの１つ以上のユーザ定義プロシージャがアクセスされ、使用される（７０８）。ユーザ定義プロシージャは、一般に複雑な報告の準備をより容易にするために使用され、通常はプライマリノードで作成され、コンパイルされる。これらのプロシージャには、ちょうど他のデータベースオブジェクトと同様に、フェイルオーバノードからアクセス可能である。

【００２６】

データベースシステム８００を図８に示す。この図はユーザ８０２、クライアント８０４、プライマリノード８０６、フェイルオーバノード８０８およびデータベース８１０しか示していないが、システム８００は他のクラスタ、ノード、ユーザ、データベースおよびクライアントを含んでもよい。この例では、ユーザ８０２はクライアント８０４を介してプライマリノード８０６でプロシージャ８１８ａおよび８１８ｂを定義している。データベース８１０のスナップショット８１２が取得された後、スナップショット８１２ならびにユーザ定義プロシージャ８１８ａおよび８１８ｂを使用して、プライマリノード８０６で非報告トランザクション８１４を動作させるのと同時に報告トランザクション８１６がフェイルオーバノード８０８で実行される。図８に示すように、スナップショット８１２の使用は、ユーザ定義プロシージャ８１８ａおよび８１８ｂとは異なって、直接的である。すなわち、スナップショット８１２はプライマリノード８０６を経ることなく使用される。

【００２７】

データベースシステムにおいて報告トランザクションを処理する別の方法を図９に示す。この方法によれば、９０２において、データベースのスナップショットが取得される。データベースはプライマリノードおよびセカンダリノードにリンクされる。次いで、９０４において、１つ以上の非報告トランザクションがプライマリノードで実行され、９０６において、プライマリノードで１つ以上の非報告トランザクションを実行するのと同時にフェイルオーバノードで報告トランザクションを実行するためにスナップショットが利用される。報告トランザクションがフェイルオーバノードで実行されるときにデータベースにおける一時的なスペースが確保され、使用される（９０８）。

【００２８】

データベースに一時的なスペースを確保するために、フェイルオーバノードはメッセージをプライマリノードに送信することができる。なぜなら、確保することは通常、整合性の問題を回避するためにプライマリノードによって行なわれるカタログの変更を必要とするためである。一旦フェイルオーバノードのためにスクラッチディスクスペースが確保されると、プライマリノードからの介入なしに一時的なスペース自体への書込を行なうことができる。スクラッチスペースによって一時的なファイルを作成することができる。これらの一時的なファイルは時には、メインメモリに適合しない一時的な動作の結果、たとえばもろもろの中間結果、ＪＯＩＮ法において使用されるハッシュテーブルなどを格納するために必要である。

【００２９】

図１０は、プライマリノード１００２ならびに３つのフェイルオーバノード１００４ａ、１００４ｂおよび１００４ｃを有するクラスタ１０００を示し、それらはすべてデータベース１００６にリンクされる。この図では、ユーザ定義プロシージャ１０１２は、リード・ライトトランザクション１０１０ａおよびリード・オンリトランザクション１０１０ｂとともに、プライマリノード１００２で見られることができる。報告トランザクション１０１４ａおよび１０１４ｂはフェイルオーバノード１００４ａで動作している。さらに、報告トランザクション１０１４ｄ、１０１４ｅおよび１０１４ｆがフェイルオーバノード１００４ｃで動作している間、報告トランザクション１０１４ｃはファイルオーバノード１００４ｂで動作している。データベース１００６の３つのスナップショット１００８ａ、１００８ｂおよび１００８ｃは異なるときに取得された。報告トランザクションの各々はスナップショットのうちの１つを使用して実行され得る。しかしながら、同一のフェイルオーバノードでの報告トランザクションは同一のスナップショットを利用する必要はない。たとえば、フェイルオーバノード１００４ｃでの報告トランザクション１０１４ｄ、１０１４ｅおよび１０１４ｆは各々が異なるスナップショット１００８を使用できる。

【００３０】

図１０に示すように、３つの一時的なスペース１０１６ａ、１０１６ｂおよび１０１６ｃは、フェイルオーバノード１００４ａ、１００４ｂおよび１００４ｃのためにそれぞれにデータベース１００６に確保されている。フェイルオーバノード１００４ａ、１００４

ｂおよび１００４ｃの各々は要求をプライマリノード１００２に送信して、それぞれのスクラッチスペースを確保する。他の実施例では、フェイルオーバノード１００４ａ、１００４ｂおよび１００４ｃは１つ以上の一時的なスペースを共有してもよい。

【００３１】

システムアーキテクチャの概要

図１１は、この発明の実施例を実現するのに好適なコンピュータシステム１１００のブロック図である。コンピュータシステム１１００は、プロセッサ１１０４、システムメモリ１１０６（たとえばＲＡＭ）、静的記憶装置１１０８（たとえばＲＯＭ）、ディスクドライブ１１１０（たとえば磁気もしくは光学）、通信インターフェイス１１１２（たとえばモデムもしくはイーサネット（登録商標）カード）、ディスプレイ１１１４（たとえばＣＲＴもしくはＬＣＤ）、入力装置１１１６（たとえばキーボード）およびカーソル制御装置１１１８（たとえばマウスもしくはトラックボール）などのサブシステムおよび装置を相互接続する、情報を通信するためのバス１１０２または他の通信メカニズムを含む。

【００３２】

この発明の一実施例によれば、コンピュータシステム１１００は、システムメモリ１１０６に含まれる１つ以上の命令の１つ以上のシーケンスを実行するプロセッサ１１０４によって特定の動作を行なう。このような命令は、静的記憶装置１１０８またはディスクドライブ１１１０などの別のコンピュータ可読媒体からシステムメモリ１１０６に読取られることができる。代替的な実施例では、この発明を実現するために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られてハードワイヤード回路が使用されてもよい。

【００３３】

本明細書において使用される「コンピュータ可読媒体」という用語は、実行のためにプロセッサ１１０４に命令を与えることに関与する任意の媒体を指す。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がそれらに限定されない多くの形態を取り得る。不揮発性媒体はたとえばディスクドライブ１１１０などの光学または磁気ディスクを含む。揮発性媒体はシステムメモリ１１０６などのダイナミックメモリを含む。伝送媒体はバス１１０２を備えるワイヤを含む同軸ケーブル、銅線および光ファイバを含む。伝送媒体は、電波および赤外線データ通信中に発生するものなどの音波または光波の形態も取り得る。

【００３４】

コンピュータ可読媒体の一般的な形態は、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、他の磁気媒体、ＣＤ－ＲＯＭ、他の光学媒体、パンチカード、紙テープ、穴のパターンを有する他の物理的な媒体、ＲＡＭ、ＰＲＯＭ、ＥＰＲＯＭ、ＦＬＡＳＨ－ＥＰＲＯＭ、他のメモリチップもしくはカートリッジ、搬送波、またはコンピュータが読取ることができる他の媒体を含む。

【００３５】

この発明の実施例では、この発明を実施するための命令のシーケンスの実行は単一のコンピュータシステム１１００によって行なわれる。この発明の他の実施例によれば、通信リンク１１２０（たとえばＬＡＮ、ＰＴＳＮまたはワイヤレスネットワーク）によって結合される２つ以上のコンピュータシステム１１００が互いに連携してこの発明を実施するのに必要な命令のシーケンスを実行してもよい。

【００３６】

コンピュータシステム１１００は、通信リンク１１２０および通信インターフェイス１１１２を介して、プログラムすなわちアプリケーションコードを含むメッセージ、データおよび命令を伝送および受信できる。受信されたプログラムコードは、受信したままでプロセッサ１１０４によって実行されてもよく、および／または後に実行するためにディスクドライブ１１１０もしくは他の不揮発性記憶装置に格納されてもよい。

【００３７】

先の明細書では、具体的な実施例を参照してこの発明について説明してきた。しかしな

がら、この発明のより広い精神および範囲から逸脱することなくさまざまな修正および変更がなされ得ることは明白であろう。たとえば、プロセスアクションの特定の順序付けを参照して上述のプロセスフローを説明する。しかしながら、説明するプロセスアクションの多くの順序付けはこの発明の範囲または動作に影響を及ぼすことなく変更され得る。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考えられるべきである。

【図面の簡単な説明】

【００３８】

【図１】この発明の実施例に従ってデータベースシステムにおいて報告トランザクションを処理する方法のフローチャートである。

【図２】この発明の一実施例に従うフェイルオーバクラスタにおける報告トランザクションの実行を示す。

【図３】この発明の別の実施例に従ってデータベースシステムにおいて報告トランザクションを処理するための方法のプロセスフローを示す。

【図４】この発明の別の実施例に従ってクラスタにおいて報告トランザクションがいかに処理されるかの一例である。

【図５】データベースシステムにおいて報告トランザクションを処理する方法の一実施例を示す。

【図６】複数のフェイルオーバノードを有するクラスタを示す。

【図７】データベースシステムにおいて報告トランザクションを処理するための方法の別の実施例を示す。

【図８】サンプルのデータベースシステムを示す。

【図９】この発明のさらなる実施例に従ってデータベースシステムにおいて報告トランザクションを処理するための方法のプロセスフローである。

【図１０】この発明のさらなる実施例に従うフェイルオーバクラスタにおける複数の報告および非報告トランザクションの実行を示す。

【図１１】この発明の実施例が実現され得るシステムアーキテクチャの図である。

【図１】

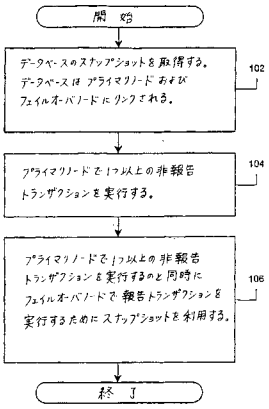

FIG. 1

【図２】

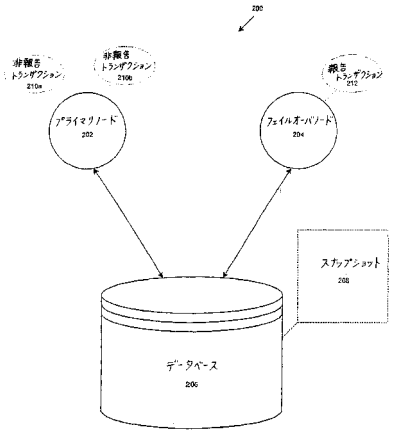

FIG. 2

【図３】

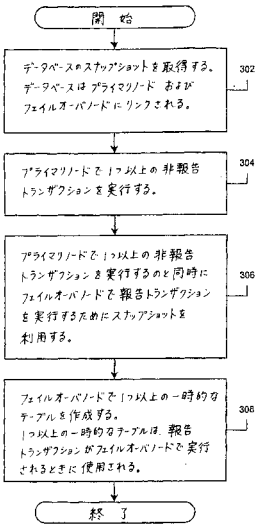

FIG. 3

【図４】

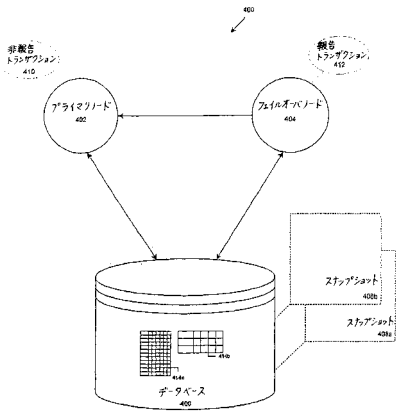

FIG. 4

【図５】

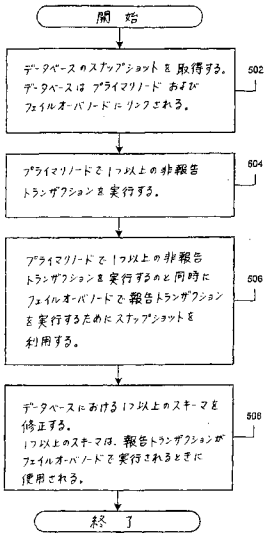

FIG. 5

【図６】

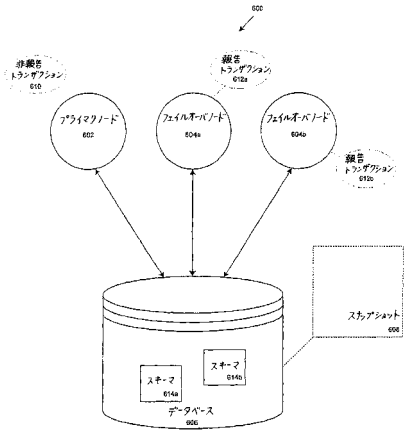

FIG. 6

【図７】

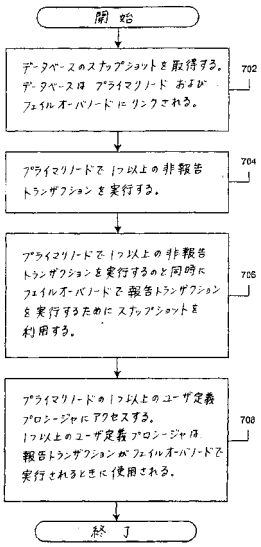

FIG. 7

【図８】

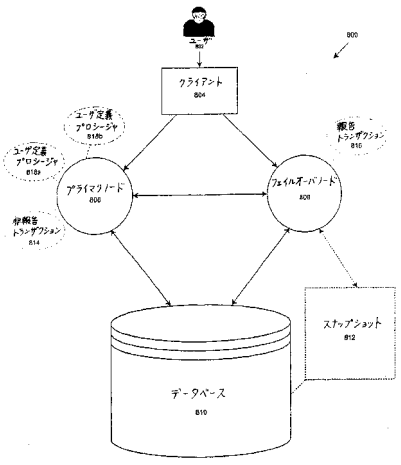

FIG. 8

【図９】

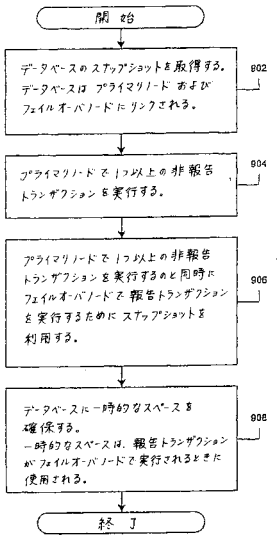

FIG. 9

【図１０】

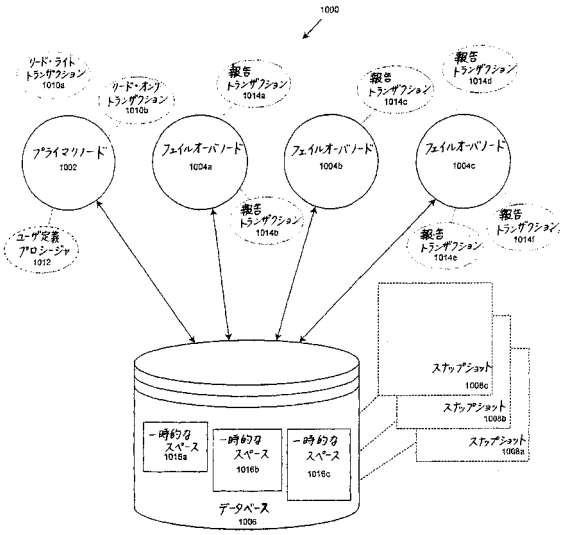

FIG. 10

【図１１】

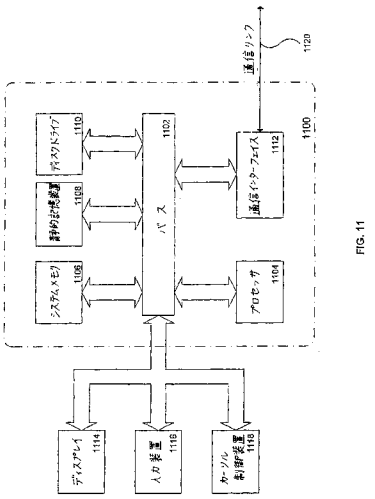

FIG. 11

フロントページの続き

(74)代理人 100098316
弁理士 野田 久登
(74)代理人 100109162
弁理士 酒井 將行
(74)代理人 100111246
弁理士 荒川 伸夫
(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、９５１３４ カリフォルニア州、サン・ノゼ、ルネサンス・ドライブ、４３２５、ナンバー・２１３
(72)発明者 プルシーノ, 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

審査官 桜井 茂行

(56)参考文献 特開２００５－２０２９１５（ＪＰ，Ａ）
特開２００１－１５９９８５（ＪＰ，Ａ）
米国特許出願公開第２００５／０１３８３１２（ＵＳ，Ａ１）
米国特許第６５２９９１７（ＵＳ，Ｂ１）

(58)調査した分野(Int.Cl., ＤＢ名)
G06F 12/00
G06F 11/14
G06F 17/30