(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2006-508470
(P2006-508470A)
(43)公表日　平成18年3月9日(2006.3.9)

(51)Int.Cl.		F I		テーマコード（参考）
G06F 11/30	(2006.01)	G06F 11/30	H	5B042

審査請求 未請求　予備審査請求 未請求　（全 23 頁）

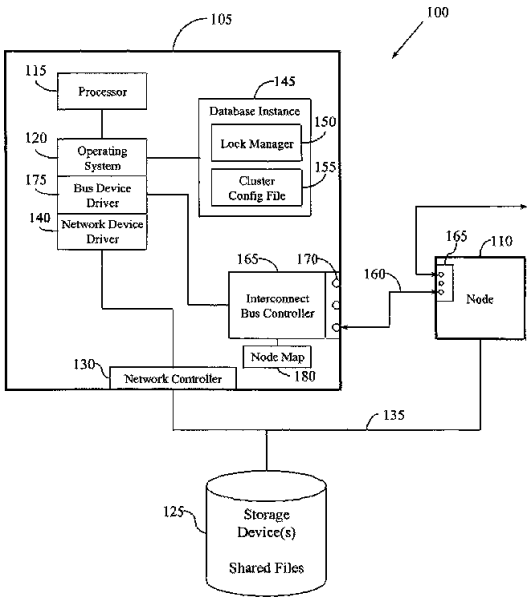
(21)出願番号	特願2004-557239 (P2004-557239)
(86)(22)出願日	平成15年11月19日 (2003.11.19)
(85)翻訳文提出日	平成17年5月26日 (2005.5.26)
(86)国際出願番号	PCT/US2003/037172
(87)国際公開番号	WO2004/051479
(87)国際公開日	平成16年6月17日 (2004.6.17)
(31)優先権主張番号	10/305,483
(32)優先日	平成14年11月27日 (2002.11.27)
(33)優先権主張国	米国 (US)

(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】クラスタシステムのためのハートビート機構

(57)【要約】
クラスタシステムのためのハートビートシステムおよび方法を提供する。クラスタは複数のノードを含む。ノードは、ネットワークを介して１つ以上のデータ記憶装置上のファイルにアクセスする。当該システムは定数ファイルを含み、当該定数ファイルは、複数のノードから、それらがアクティブであることを示すハートビートメッセージを受信しかつ記憶する。ネットワークコントローラは、ＩＥＥＥ１３９４通信プロトコルに従って定数ファイルを複数のノードに接続する。



【特許請求の範囲】
【請求項１】
複数のノードを含むクラスタのためのハートビート機構であって、システムは、
前記複数のノードからハートビートメッセージを受信するための定数ファイルと、
前記定数ファイルを前記複数のノードに接続するためのネットワークコントローラとを含み、前記ネットワークコントローラは、ＩＥＥＥ１３９４プロトコルに従って前記定数ファイルとの通信を提供する、ハートビート機構。
【請求項２】
前記ネットワークコントローラはＩＥＥＥ１３９４カードである、請求項１に記載のハートビート機構。
【請求項３】
前記定数ファイルは記憶装置内に割当てられるメモリを含む、請求項１に記載のハートビート機構。
【請求項４】
前記記憶装置は前記定数ファイルと前記複数のノードによって共有されるファイルとをともに含む、請求項３に記載のハートビート機構。
【請求項５】
前記ネットワークコントローラによって保持され、前記ＩＥＥＥ１３９４プロトコルに基づいてアクティブなノードを識別するためのノードマップをさらに含む、請求項１に記載のハートビート機構。
【請求項６】
前記ノードマップによって書込まれた前記定数ファイルにおけるハートビートメッセージと前記ノードとを比較することにより前記複数のノードからノードの状態を決定するための状態ロジックをさらに含む、請求項５に記載のハートビート機構。
【請求項７】
前記ハートビートメッセージはタイムスタンプを含む、請求項１に記載のハートビート機構。
【請求項８】
前記クラスタはデータベースクラスタである、請求項１に記載のハートビートシステム。
【請求項９】
クラスタ化されたシステムであって、
１つ以上の共有ファイルと、
ともにクラスタ化され、前記共有ファイルへの通信アクセスを有する複数のノードと、
ＩＥＥＥ１３９４プロトコルに基づいて前記複数のノードの各々と通信するよう構成された定数ファイルとを含み、前記定数ファイルは前記複数のノードの各々からの状態メッセージを記憶する、クラスタ化されたシステム。
【請求項１０】
前記定数ファイルは１つ以上の第１の記憶装置上に保持され、前記共有ファイルは前記第１の記憶装置とは別個の１つ以上の第２の記憶装置上で保持される、請求項９に記載のクラスタ化されたシステム。
【請求項１１】
前記定数ファイルおよび前記共有ファイルは１つ以上の共通の記憶装置上で保持される、請求項９に記載のクラスタ化されたシステム。
【請求項１２】
前記共有ファイルと前記複数のノードとの間に通信を提供するための第１のネットワークと、
前記定数ファイルと前記複数のノードとの間に通信を提供するための第２のネットワークとをさらに含む、請求項９に記載のクラスタ化されたシステム。
【請求項１３】

前記第１のネットワークおよび前記第２のネットワークはＩＥＥＥ１３９４プロトコルに基づいてデータを通信する、請求項１２に記載のクラスタ化されたシステム。

【請求項１４】

前記第１のネットワークおよび前記第２のネットワークは同じネットワークである、請求項１２に記載のクラスタ化されたシステム。

【請求項１５】

前記状態メッセージはタイムスタンプを含む、請求項９に記載のクラスタ化されたシステム。

【請求項１６】

前記１つ以上の共有ファイルと前記定数ファイルとの両方に前記複数のノードを接続するための共通ネットワークをさらに含む、請求項９に記載のクラスタ化されたシステム。

【請求項１７】

前記複数のノードの各々に保持され、状態メッセージを定期的に生成し、前記ＩＥＥＥ１３９４プロトコルに従って前記状態メッセージを前記定数ファイルに伝送するハートビートロジックをさらに含む、請求項９に記載のクラスタ化されたシステム。

【請求項１８】

前記クラスタ化されたシステムはデータベースクラスタを含む、請求項９に記載のクラスタ化されたシステム。

【請求項１９】

クラスタにおけるノードを監視する方法であって、

前記クラスタにおけるノードから受取った状態メッセージを記憶するための定数ファイルを割当てるステップと、

ノードがアクティブであることを示す状態メッセージを前記クラスタにおけるノードから定期的に受信するステップとを含み、前記状態メッセージはＩＥＥＥ１３９４通信プロトコルに基づいて受信される、方法。

【請求項２０】

前記ＩＥＥＥ１３９４通信プロトコルに基づいて前記クラスタにおいてアクティブなノードのノードマップを保持するステップをさらに含む、請求項１９に記載の方法。

【請求項２１】

前記定数ファイルにおける状態メッセージと前記ノードマップとを比較することによりノードがアクティブであるかどうかを判断するステップをさらに含む、請求項２０に記載の方法。

【請求項２２】

選択された時間間隔で状態メッセージを生成し、前記ＩＥＥＥ１３９４通信プロトコルに従って前記状態メッセージを前記定数ファイルに伝送するステップをさらに含む、請求項１９に記載の方法。

【請求項２３】

前記割当てるステップは、前記定数ファイルの部分を前記ノードの各々に等しく割当てるステップを含む、請求項１９に記載の方法。

【請求項２４】

前記割当てるステップは、前記ノードによってアクセスされるファイルを含む記憶装置において前記定数ファイルを規定するステップを含む、請求項１９に記載の方法。

【請求項２５】

前記割当てるステップは、ＩＥＥＥ１３９４コントローラを含む定数装置において前記定数ファイルを規定するステップを含む、請求項１９に記載の方法。

【請求項２６】

ハートビートシステムであって、

複数のノードと、

前記複数のノードから送信されたハートビートメッセージ情報を記憶するための定数区域と、

前記定数ファイルを前記複数のノードに伝達するためのネットワークコントローラとを含み、前記ネットワークコントローラはＩＥＥＥ１３９４プロトコルに従って前記定数ファイルとの通信を提供する、ハートビートシステム。

【請求項２７】

前記定数区域は前記複数のノードとは異なるネットワーク上で保持される、請求項２６に記載のハートビートシステム。

【請求項２８】

前記定数区域は定数ファイルを割当てるための手段を含む、請求項２６に記載のハートビートシステム。

【請求項２９】

前記複数のノードの各々に保持され、ハートビートメッセージ情報を定期的に生成し、これを前記ＩＥＥＥ１３９４プロトコルに従って前記定数区域に伝送するハートビートロジックをさらに含む、請求項２６に記載のハートビートシステム。

【請求項３０】

前記複数のノードは共有ディスクアーキテクチャまたは非共有アーキテクチャにおいてクラスタ化される、請求項２６に記載のハートビートシステム。

【請求項３１】

クラスタにおけるノードを監視する方法であって、

前記クラスタにおけるノードから受取った状態メッセージを記憶するための定数ファイルを割当てるステップと、

前記ノードがアクティブであることを示す状態メッセージを前記クラスタにおけるノードから定期的に受信するステップとを含み、前記状態メッセージはＩＥＥＥ１３９４通信プロトコルに基づいて受信される、方法。

【請求項３２】

アクティブなノードのノードマップを保持するステップをさらに含む、請求項３１に記載の方法。

【請求項３３】

前記定数ファイルにおける状態メッセージと前記ノードマップとを比較することによりノードがアクティブであるかどうかを判断するステップをさらに含む、請求項３２に記載の方法。

【請求項３４】

前記割当てるステップは、前記ノードによってアクセスされるファイルを含む記憶装置において前記定数ファイルを規定するステップを含む、請求項３１に記載の方法。

【請求項３５】

前記割当てるステップは、前記クラスタにおける前記ノードとは別個のネットワーク上で保持される記憶装置において前記定数ファイルを規定するステップを含む、請求項３１に記載の方法。

【発明の詳細な説明】

【技術分野】

【０００１】

発明の分野

この発明はクラスタリング技術に関する。この発明は、ハートビートシステムおよび方法を有するクラスタリングシステムへの特定の用途を見出す。

【背景技術】

【０００２】

発明の背景

クラスタは単一のシステムとして協働する独立したサーバのグループである。主要なクラスタ構成要素はプロセッサノード、クラスタ相互接続（専用ネットワーク）およびディスクサブシステムである。クラスタはディスクアクセスや、データを管理するリソースを共有するが、別個の各ハードウェアクラスタノードはメモリを共有しない。各ノードはそ

れ自体の専用のシステムメモリ、ならびにそれ自体のオペレーティングシステム、データベースインスタンスおよびアプリケーションソフトウェアを有する。クラスタは、単一の対称的なマルチプロセッサシステムにわたって高度な障害許容力およびモジュールの増分システム成長を提供し得る。サブシステムが故障した場合には、クラスタリングにより高い可用性が確実にされる。冗長なハードウェア構成要素、たとえば付加的なノード、相互接続および共有ディスクはより高い可用性を提供する。このような冗長なハードウェアアーキテクチャにより単一障害点が避けられ、障害許容力が与えられる。

【０００３】

データベースクラスタにおいては、各ノードに対するＣＰＵおよびメモリ要件がデータベースアプリケーションごとに異なり得る。性能および費用の要件もデータベースアプリケーションごとに異なる。性能に寄与する１つの要因は、クラスタにおける各ノードがその健康状態および構成を当該クラスタにおける他のノードに知らせ続ける必要があることである。これは、ハートビートと称されるネットワークメッセージをネットワークにわたって定期的に同報通信することによってなされてきた。ハートビート信号は通常、専用のネットワーク、すなわちノード間通信に用いられるクラスタ相互接続を介して送信される。しかしながら、ハートビートメッセージが損失または遅延したりすると、ノードが機能していないという誤った報告がなされるおそれがある。

【０００４】

先行技術のシステムにおいては、クラスタ相互接続は、各ノードにネットワークカードを取付け、適切なネットワークケーブルでこれらを接続し、ワイヤ全体にわたらせるようソフトウェアプロトコルを構成することによって構築されてきた。相互接続は、典型的には、ＴＣＰ／ＩＰもしくはＵＤＰを実行する低コスト／低速のイーサネット（登録商標）カード、または、ＲＤＧ（Reliable DataGram）を実行するコンパック（Compaq）のメモリチャネル（Memory Channel）もしくはＨＭＰ（Hyper Messaging Protocol）を用いるヒューレット・パッカード（Hewlett-Packard）のHyperfabric／２のような高コスト／高速のプロプラエタリの相互接続であった。低コスト／高速の相互接続はユーザのためにクラスタリングのコストを下げ、実行中の待ち時間を短くするだろう。

【０００５】

この発明は、上述の問題に対処するクラスタリングの新しく有用な方法およびシステムを提供する。

【発明の開示】

【課題を解決するための手段】

【０００６】

発明の概要

一実施例においては、複数のノードを含むクラスタシステムのためのハートビート機構が提供される。一局面においては、当該システムは、複数のノードからハートビートメッセージを受信する定数ファイルを含む。ネットワークコントローラは定数ファイルを複数のノードに接続し、ここで、当該ネットワークコントローラがＩＥＥＥ１３９４プロトコルに従って定数ファイルと通信を行なう。

【０００７】

この発明の別の実施例に従うと、クラスタにおけるノードを監視する方法が提供される。定数ファイルは、クラスタにおけるノードから受取った状態メッセージを記憶するために割当てられる。ノードがアクティブであることを示す状態メッセージがクラスタにおけるノードから定期的に受信される。この場合、状態メッセージはＩＥＥＥ１３９４通信プロトコルに基づいて受信される。

【０００８】

明細書に援用され、明細書の一部を構成する添付の図面においてシステムおよび方法の実施例が示されるが、当該実施例は、下記の詳細な説明とともに当該システムおよび方法の具体的な実施例を説明するのに役立つ。図面に示される要素の境界（たとえば箱または箱のグループ）が境界の一例を表わすことが理解されるだろう。１つの要素が複数の要素

として設計され得るかまたは複数の要素が１つの要素として設計され得ることを当業者は理解するだろう。別の要素の内部の構成要素として示される要素が外部の構成要素として実現され得、逆の場合も同様に実現され得る。

【発明を実施するための最良の形態】

【０００９】

図示される実施例の詳細な説明

以下は、開示全体を通じて用いられる選択された用語の定義を含む。すべての用語の単数形および複数形はともに各々の意味の範囲内である。

【００１０】

この明細書中で用いられる「コンピュータ読取可能媒体」は、信号、命令および／またはデータを実行のために直接的または間接的にプロセッサに供給することに関わるいかなる媒体をも指す。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、これらに限定されない多くの形を取り得る。不揮発性媒体はたとえば光ディスクまたは磁気ディスクを含み得る。揮発性媒体は動的メモリを含み得る。伝送媒体は同軸ケーブル、銅ワイヤおよび光ファイバケーブルを含み得る。伝送媒体はまた、電波および赤外線データ通信中に生成されるような音波または光波の形を取り得る。コンピュータ読取可能媒体の一般的な形は、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、磁気テープもしくは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、孔のパターンを備えた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、搬送波／パルス、またはコンピュータが読取ることのできる他のいずれかの媒体を含む。

【００１１】

この明細書中で用いられる「ロジック」は、機能もしくは動作を実行するために、および／または別の構成要素から機能もしくは動作を引起すためにハードウェア、ファームウェア、ソフトウェアおよび／または各々の組合せを含むが、これらには限定されない。たとえば、所望の用途または必要性に基づき、ロジックはソフトウェア制御のマイクロプロセッサ、特定用途向け集積回路（ＡＳＩＣ）などのディスクリートロジックまたはプログラミングされた他の論理素子を含み得る。ロジックはまたソフトウェアとして十分に具体化され得る。

【００１２】

この明細書中で用いられる「信号」は、１つ以上の電気信号、アナログもしくはデジタル信号、信号状態の変化（たとえば電圧上昇／降下）、１つ以上のコンピュータ命令、メッセージ、ビットもしくはビットストリーム、または受信、伝送および／もしくは検出が可能な他の手段を含むが、これらに限定されない。

【００１３】

この明細書中で用いられる「ソフトウェア」は、コンピュータまたは他の電子素子に所望の態様で機能を実行させたり、動作を実行させたり、および／または作動させたりする１つ以上のコンピュータ読取可能および／または実行可能な命令を含むが、これらに限定されない。命令は、動的にリンクされたライブラリからの別個のアプリケーションまたはコードを含むルーチン、アルゴリズム、モジュールまたはプログラムなどのさまざまな形で実現され得る。ソフトウェアはまた、独立プログラム、関数呼出、サーブレット、アプレット、メモリに記憶された命令、オペレーティングシステムの一部、または他の種類の実行可能な命令などのさまざまな形で実現され得る。ソフトウェアの形がたとえば所望のアプリケーションの要件、それが実行される環境、および／または設計者／プログラマの要望などに依存することを当業者は理解するだろう。

【００１４】

図１には、この発明の一実施例に従った単純なクラスタ化されたデータベースシステム１００の一実施例が示される。２つのノード、すなわちノード１０５およびノード１１０がこの例において示されるが、異なる数のノードが用いられてもよく、異なる構成でクラ

スタ化されてもよい。データベースクラスタが一例として用いられるが、当該システムは他の種類のクラスタ化されたシステムにも適用可能である。各ノードは、ソフトウェアを実行しかつ情報を処理するコンピュータシステムである。コンピュータシステムはパーソナルコンピュータ、サーバまたは他の計算装置であってもよい。各ノードはさまざまな構成要素および装置、たとえば、１つ以上のプロセッサ１１５、オペレーティングシステム１２０、メモリ、データ記憶装置、データ通信バスおよびネットワーク通信装置を含み得る。各ノードは他のノードとは異なる構成を有し得る。一種のクラスタリングシステムの一例が、「１つのノードのキャッシュから別のノードのキャッシュにデータを転送するための方法および装置（“Method and Apparatus for Transferring Data from the Cache of One Node to the Cache of Another Node”）」と題され、この発明の譲渡人に譲渡され、その全体がすべての目的のために引用によりこの明細書中に援用される米国特許第６，３５３，８３６号に記載される。

【００１５】

図１をさらに参照すると、ノード１０５が、クラスタ化されたデータベースシステム１００におけるノードの構成例を説明するために用いられる。この実施例においては、ノードは、各ノードが１つ以上のデータ記憶装置１２５にアクセスできるデータ共有構成でネットワーク接続される。データ記憶装置１２５は、クラスタにおいて接続されたノードによって共有され得るデータベースファイルなどのさまざまなファイルを保持し得る。ネットワークコントローラ１３０はノード１０５をネットワーク１３５に接続する。オペレーティングシステム１２０は、ノード１０５上で実行するソフトウェアアプリケーションとネットワークコントローラ１３０との間の通信インターフェイスを含む。たとえば、当該インターフェイスは、ネットワーク１３５の選択された通信プロトコルに従ってプログラミングされたネットワークデバイスドライバ１４０であってもよい。

【００１６】

ネットワークコントローラ１３０およびネットワーク１３５のために用いられ得る通信プロトコルの例には、ファイバチャネル（Fibre Channel）ＡＮＳＩ規格Ｘ３．２３０および／またはＳＣＳＩ－３　ＡＮＳＩ規格Ｘ３．２７０が含まれる。ファイバチャネルアーキテクチャは、シリアル通信およびストレージＩ／Ｏの両方に高速のインターフェイスリンクをもたらす。ネットワークコントローラ１３０の他の実施例は、とりわけ、Fast－４０（Ultra-SCSI）、シリアル・ストレージ・アーキテクチャ（ＳＳＡ）、ＩＥＥＥ規格１３９４、非同期転送モード（ＡＴＭ）、スケーラブル・コヒーレント・インターフェイス（ＳＣＩ）ＩＥＥＥ規格１５９６－１９９２または上述のいくつかの組合せを利用する実施例などの記憶装置１２５とノード１０５、１１０とを接続する他の方法をサポートし得る。

【００１７】

ノード１０５はさらに、１つ以上の記憶装置１２５において保持されるデータへのアクセスを管理および制御するデータベースインスタンス１４５を含む。クラスタ化されたデータベースシステム１００における各ノードがデータベースインスタンスを実行することにより、その特定のノードが記憶装置１２５において共有データベース上のデータにアクセスしかつ当該データを処理することが可能となるので、ロックマネージャ１５０が設けられる。当該ロックマネージャ１５０は、記憶装置１２５に格納される共有データベースなどの１つ以上のリソース上のロックを認可したり、待ち行列に入れたり、追跡したりすることを担うエンティティである。プロセスが共有データベース上で動作を実行し得る前に、当該プロセスは、データベース上で所望の動作を実行する権利を当該プロセスに与えるロックを得る必要がある。ロックを得るために、プロセスはロックの要求をロックマネージャに伝送する。ネットワークシステムにおけるリソースの使用を管理するために、ロックマネージャがネットワークにおける１つ以上のノード上で実行される。

【００１８】

ロックは、特定のプロセスがリソースに関する或る権利を与えられたことを示すデータ構造である。多くの種類のロックがある。多くのプロセスで共有され得る種類のロックも

あれば、同じリソース上で他のいずれかのロックが認可されるのを妨げる種類のロックもある。ロック管理システムの一例のより詳細な説明が、「ロック管理システムにおける予期ロックモード変換（“Anticipatory Lock Mode Conversions in a Lock Management System”）」と題され、この発明の譲渡人に譲渡され、その全体がすべての目的のために引用によりこの明細書中に援用される米国特許第６，４０５，２７４　Ｂ１号に見出される。

【００１９】

記憶装置１２５にアクセスし得るノードをネットワーク上で追跡しかつ管理するために、クラスタ構成ファイル１５５が保持される。クラスタ構成ファイル１５５はクラスタにおけるアクティブなノードの現在のリストを含み、これにはノードアドレス、ノードＩＤおよび接続構造（たとえば隣接ノード、親子ノード）などの識別情報が含まれる。当然、他の種類の情報がこのような構成ファイルに含まれてもよく、その種類のネットワークシステムに基づいて異なっていてもよい。トポロジ変化がクラスタにおいて発生すると、ノードが識別され、クラスタ構成ファイル１５５がクラスタノードの現在の状態を反映するよう更新される。トポロジ変化の例は、ノードがいつ追加されるか、いつ除去されるかまたはいつ動作を停止させるかを含む。

【００２０】

図１をさらに参照すると、データベースクラスタシステム１００はさらに、ノード１０５とノード１１０との間にノード間通信をもたらす相互接続ネットワーク１６０を含む。相互接続ネットワーク１６０は、ネットワーク上のすべてのノードが互いに双方向通信することを可能にするバスを備える。相互接続１６０は、同じバスを介して各ノードとメッセージおよびデータのやりとりを行なうためのアクティブな通信プロトコルを提供する。相互接続ネットワーク１６０に接続されるように、各ノードは、ノードのＰＣＩスロットに差込まれる周辺カードであり得る相互接続バスコントローラ１６５を含む。コントローラ１６５はノード間でケーブルを接続するための１つ以上の接続ポート１７０を含む。３つの接続ポートがポート１７０に図示されているが、異なる数のポートが用いられてもよい。

【００２１】

一実施例においては、相互接続バスコントローラ１６５は、ファイアワイヤまたはｉ．ＬＩＮＫとしても公知であるＩＥＥＥ１３９４プロトコルに従って動作する。データベースインスタンス１４５またはノード１０５上で実行する他のアプリケーションを相互接続バス１６０と通信させるために、バスデバイスドライバ１７５が設けられる。バスデバイスドライバ１７５はオペレーティングシステム１２０と作動して、相互接続バスコントローラ１６５とアプリケーションとのインターフェイスを取る。たとえば、データベースインスタンス１４５からのデータベースコマンドが、バスデバイスドライバ１６５によってＩＥＥＥ１３９４コマンドまたはオープン・ホスト・コントローラ・インターフェイス（ＯＨＣＩ）コマンドに翻訳される。ＩＥＥＥ１３９４ＯＨＣＩ規格は、ＩＥＥＥ１３９４バスに接続するための標準的なハードウェアおよびソフトウェアを規定する。ＯＨＣＩは、標準的なレジスタアドレスおよび機能、データ構造、ならびにダイレクトメモリアクセス（ＤＭＡ）モデルを規定する。

【００２２】

ＩＥＥＥ１３９４は、使いやすく低コストで高速の通信を提供するバスプロトコルである。当該プロトコルは非常に拡張可能であり、非同期アプリケーションおよび等時性アプリケーション（isochronous application）の両方を備え、大量のメモリマップドアドレス空間へのアクセスを可能にし、ピアツーピア通信を可能にする。相互接続バスコントローラ１６５がＩＥＥＥ１３９４ａ、１３９４ｂなどの他のバージョンのＩＥＥＥ１３９４プロトコルや、他の将来の変更および増強に対応するよう変更可能であることを当業者は理解するだろう。

【００２３】

ＩＥＥＥ１３９４プロトコルは、ポイント・ツー・ポイントシグナリング環境を備えた

ピアツーピアネットワークである。バス１６０上のノードは、それらの上にいくつかのポート、たとえばポート１７０を有し得る。これらのポートの各々は中継器として機能し、ノード内の他のポートが受信するいずれのデータパケットをも再伝送する。各ノードは、ネットワークトポロジ／構成の現在の状態を追跡するノードマップ１８０を保持する。ＩＥＥＥ１３９４プロトコルは、その現在の形では、単一のバス上で６３個までの装置をサポートし、装置への接続は電話機のプラグ差込口に差込むのと同じくらい容易である。ノードおよび他の装置は、最初にノードの電源を切ったりネットワークを再起動したりしなくても直ちに接続され得る。データベースクラスタトポロジの管理が以下により詳細に記載される。

【００２４】

相互接続ネットワーク１６０を用いれば、ノード１０５におけるデータベース１４５は、ノード１１０またはクラスタにおける他のノード上の実行中のデータベースアプリケーションに対し直接データを要求するかデータを送受信するかまたはメッセージを送信し得る。これにより、１つ以上の中間ステップや付加的なディスクＩ／Ｏが必要になったり、待ち時間が増えたりするような、メッセージまたはデータパケットを記憶装置１２５に送信せざるを得ない状態が回避される。

【００２５】

図２には、ＩＥＥＥ１３９４規格に基づいた相互接続バスコントローラ１６５の例が示される。これは3つのＩＳＯプロトコル層、すなわちトランザクション層２００、リンク層２０５および物理層２１０を含む。当該層は、上述において規定されハードウェア、ソフトウェアまたはこれらの両方を含むロジックで実現され得る。トランザクション層２００は、３つの基本的な動作、すなわち読出、書込およびロックを用いてバストランザクションを実行するための完全な要求－応答プロトコルを規定する。リンク層２０５は中間レベル層であり、トランザクション層２００および物理層２１０の両方と相互作用して、データパケットのために非同期および等時性転送サービスを提供する。データ転送を制御する構成要素はデータパケット送信機、データパケット受信機およびクロックサイクルコントローラを含む。

【００２６】

物理層２１０は、コントローラ１６５と相互接続バス１６０の一部を形成するケーブルとの間に電気的および機械的なインターフェイスを提供する。これは物理ポート１７０を含む。当該物理層２１０はまた、すべてのノードがアービトレーション機構を用いてバスに公平にアクセスできることを確実にする。たとえば、ノードは、バスにアクセスする必要がある場合、その親ノードに要求を送信し、当該親ノードが当該要求をルートノードに転送する。当該ルートが受信した第１の要求が受入れられると、他のすべての要求が拒否され取消される。ノードがルートに近ければ近いほど受入れられる可能性が高くなる。結果として生ずるアービトレーションの不公平性を解決するために、バスアクティビティの期間がいくつかの間隔に分割される。ある間隔中に各ノードが一度伝送され、次の間隔まで待機する。当然、アービトレーションのために他の方法が用いられてもよい。

【００２７】

物理層２１０の他の機能は、データ再同期、符号化および復号化、バス初期設定ならびに信号レベルの制御を含む。上述のとおり、各ノードの物理層はまた中継器として機能し、ポイント・ツー・ポイント接続を仮想の同報通信バスに翻訳する。標準的なＩＥＥＥ１３９４ケーブルは１．５アンペアまでのＤＣ電力を供給して、遠隔装置を、それらの電源が切られているときでも「認識している（aware）」状態に維持する。物理層はまた、ＩＥＥＥ１３９４に基づいて、ノードが単一の媒体上でさまざまな速度でデータを伝送することを可能にする。データレート能力が異なるノードまたは他の装置はより遅い装置速度で通信する。

【００２８】

ＩＥＥＥ１３９４プロトコルに基づいて動作する相互接続バスコントローラ１６５はアクティブなポートであり、自己監視／自己構成シリアルバスを備える。これは、たとえバ

スがアクティブであってもユーザが装置を追加したり取除いたりすることを可能にするホットプラグアンドプレイとして公知である。こうして、ノードおよび他の装置が、ネットワーク動作を遮らずに接続および切断され得る。自己監視／自己構成ロジック２１５は、相互接続バス信号における変化に基づいてクラスタシステムにおけるトポロジ変化を自動的に検出する。ノードのバスコントローラ１６５は、当該ノードがバスに接続されると、相互接続バス１６０上にバイアス信号を配置する。自己監視ロジック２１５を介する隣接ノードが、電圧の変化として現われ得るバイアス信号を自動的に検出する。こうして、検出されたバイアス信号は、ノードが追加されたことおよび／またはノードが依然としてアクティブであることを示す。逆に、バイアス信号がないということは、ノードが除去されたかまたは機能を停止したことを示す。この態様では、トポロジ変化は、ノード間で伝送されるポーリングメッセージを用いなくても検出することができる。ロジック２１５の自己構成局面が図６および図７に関連してより詳細に説明される。

【００２９】

アプリケーションプログラムインターフェイス（ＡＰＩ）層２２０は、バスデバイスドライバ１７５へのインターフェイスとしてバスコントローラ１６５に含まれ得る。これは概して、データ、エンドシステム設計およびアプリケーションをまとめるより高いレベルのシステムガイドライン／インターフェイスを含む。ＡＰＩ層２２０は、データベースインスタンス１４５（および他のアプリケーション）と相互接続バスコントローラ１６５との間の通信をカスタマイズするよう所望の特徴でプログラミングされ得る。随意には、ＡＰＩ層２２０の機能は、トランザクション層２００またはバスデバイスドライバ１７５内で全体または一部が実現され得る。

【００３０】

図３を参照すると、この発明のシステムおよび方法が実現され得るデータベースクラスタアーキテクチャ３００の一実施例が示される。当該アーキテクチャ３００は共有ディスクアーキテクチャとして一般に公知であり、付加的なノードが示されている以外は図１に類似している。概して共有ディスクデータベースアーキテクチャにおいては、ファイルおよび／またはデータはノード間で論理的に共有され、各々のデータベースインスタンスはすべてのデータにアクセスできる。共有ディスクアクセスが、たとえばファイルを保持する１つ以上の記憶装置３０５への直接的なハードウェア接続性によって達成される。随意には、接続は、すべてのノード上におけるすべての記憶装置３０５の単一のビューを提供するオペレーティングシステム抽象層を用いることによって実行され得る。ノードＡ～Ｄはまた、ノード相互接続１６０を介して接続されてノード間通信を提供する。共有ディスクアーキテクチャにおいては、ノード内のいずれかのデータベースインスタンス上で実行されるトランザクションは、記憶装置３０５上のデータベースのいずれかの部分を直接読出すかまたは変更することができる。アクセスは、上述のように１つ以上のロックマネージャによって制御される。

【００３１】

図４を参照すると、この発明のシステムおよび方法を組込み得るクラスタアーキテクチャの別の実施例が示される。クラスタアーキテクチャ４００は典型的には非共有アーキテクチャと称される。非共有アーキテクチャの一例が、「ハイブリッド非共有／共有ディスクデータベースシステム（“Hybrid Shared Nothing/Shared Disk Database System”）」と題され、この発明の譲渡人に譲渡され、その全体がすべての目的のために引用によりこの明細書中に援用される米国特許第６，３２１，２１８号に記載される。純粋な非共有アーキテクチャにおいては、データベースファイルは、たとえば、ノードＡ～Ｄ上で実行するデータベースインスタンス間で分割される。各データベースインスタンスまたはノードはデータの別個のサブセットの所有権を有し、このデータへの全アクセスがこの「所有」インスタンスによって排他的に実行される。ノードはまた相互接続１６０と接続される。

【００３２】

たとえば、記憶装置Ａ～Ｄに記憶されるデータファイルが従業員ファイルを含む場合、

当該データファイルは、ノードＡが文字Ａ～Ｇで始まる従業員名に対する従業員ファイルを制御し、ノードＢが従業員名Ｈ～Ｎに対する従業員ファイルを記憶装置Ｂ上で制御し、ノードＣが名前「Ｏ～Ｕ」に対する従業員ファイルを記憶装置Ｃ上で制御し、ノードＤが記憶装置Ｄ上で従業員ファイル名「Ｖ～Ｚ」を制御するように分割され得る。他のノードからデータにアクセスするために、このようなデータを要求するメッセージが送られるだろう。たとえば、ノードＤがノードＡによって制御される従業員ファイルを所望する場合、データファイルを要求するメッセージがノードＡに送られるだろう。次いで、ノードＡが記憶装置Ａからデータファイルを検索し、データをノードＤに伝送するだろう。この発明のシステムおよび方法が、ツリー構造などの他のクラスタアーキテクチャおよび構成上で、かつ特定の用途に対し所望のとおりに他のデータアクセス権および／または制限で実現され得ることが理解されるだろう。

【００３３】

図５には、図３または図４のクラスタシステムに関連付けられる方法論の一実施例が示される。当該実施例は、相互接続バス１６０を用いたノード間におけるデータの直接的な送受信を説明する。例示された要素は「処理ブロック」を示し、コンピュータに動作を実行させかつ／または決定を下させるコンピュータソフトウェア命令または命令のグループを表わす。代替的には、処理ブロックは、デジタル信号プロセッサ回路または特定用途向け集積回路（ＡＳＩＣ）などの機能的に同等の回路によって実行される機能および／または動作を表わし得る。図ならびに他の例示された図はいずれかの特定のプログラミング言語の構文を示すものではない。むしろ、当該図は、回路を製作し、コンピュータソフトウェアまたはハードウェアおよびソフトウェアの組合せを作成して例示された処理を実行するために当業者が使用し得る関数情報を例示する。電子的アプリケーションおよびソフトウェアアプリケーションが動的かつフレキシブルなプロセスを含み得るので、例示されたブロックが図示されるものとは異なる他のシーケンスで実行され得、および／またはブロックが組合され得るかもしくは付加的な構成要素に分けられ得ることが理解されるだろう。これらはまた、機械語、手続き型、オブジェクト指向および／または人工知能技術などのさまざまなプログラミング手法を用いて実現され得る。上述のことは、この明細書中に記載されるすべての方法論に適用される。

【００３４】

図５を参照すると、図５００は、ノード間相互接続ネットワーク１６０を用いてノード間でデータを伝達する一例である。ノード（要求ノード）が別のノードからのデータを所望する場合、データ要求メッセージが相互接続バス１６０を介して宛先ノードに伝送される（ブロック５０５）。当該データ要求は、ノード名および／またはアドレスを当該要求に添付することによって１つ以上の選択された宛先ノードに直接送信され得る。要求されたデータの位置が未知である場合、データ要求が相互接続ネットワークにおける各ノードに同報通信され得る。

【００３５】

データ要求が適切なノードによって受信されると、データベースインスタンスは、データがそのノード上で利用可能であるかどうかを判断する（ブロック５１０）。データが利用可能でない場合、データが利用可能でないというメッセージが要求ノードに伝送される（ブロック５１５）。データが利用可能である場合、データはダイレクトメモリアクセスによってローカルメモリから検索され（ブロック５２０）、相互接続バスを介して要求ノードに伝送される（ブロック５２５）。リモートダイレクトメモリアクセスがまた、直接的なメモリ間転送を実行するよう実現され得る。この態様では、メッセージおよびデータは、メッセージまたはデータを共有記憶装置に伝送する必要なしにノード間で直接伝送され得る。ノード間通信は待ち時間を減らし、ディスク入出力の数を減らす。

【００３６】

図６には、ＩＥＥＥ１３９４バスプロトコルに基づいたクラスタアーキテクチャを再構成する例示的な方法論が示される。データベースクラスタにおけるノードが追加されるか、除去されるかまたは機能を停止する場合、データベースクラスタは変更を検出し、ノー

ドを識別する必要があり、当該クラスタは適切に再構成される必要がある。上述のように、ＩＥＥＥ１３９４プロトコルに基づいて動作する相互接続バスコントローラ１６５（図１）はアクティブなポートであり、自己構成シリアルバスを備える。こうして、ノードおよび他の装置がネットワーク動作を遮らずに接続および切断され得る。

【００３７】

たとえば、ノードがバスに追加されると、当該バスがリセットされる（ブロック６０５）。追加されたノードの相互接続コントローラ１６５はバス上でバイアス信号を自動的に送信し、隣接ノードがそのバイアス信号を検出し得る（ブロック６１０）。同様に、ノードが除去されるときにノードのバイアス信号がないことが検出され得る。すなわち、隣接ノードの相互接続コントローラ１６５は、ノードを追加するかまたは除去することによって引起されるバス信号強度の変化などの相互接続バス１６０上の信号の変化を検出し得る。次いで、トポロジ変化が、データベースクラスタにおける他のすべてのノードに伝送される。バスノードマップが当該変化に応じて再構築される（ブロック６１５）。一実施例においては、ノードマップは当該変化に応じて更新され得る。データベースインスタンスが通知され、これが、ロックマネージャに対するアクティブなノードを追跡するようクラスタ構成ファイルを更新する（ブロック６２０）。当然、図示されたシーケンスの順は他のやり方で実現されてもよい。

【００３８】

ＩＥＥＥ１３９４プロトコルを用いれば、相互接続コントローラ１６５は上述の自己監視／自己構成機構を含むアクティブなポートとなる。この機構を用いる場合、データベースクラスタシステムは、ポーリング機構に伴うさらなる待ち時間なしに再構成され得る。というのも、ノードがトポロジにおける変化を実質的に直ちに検出し得るからである。アクティブなポートはまた、ネットワークの電源を切る必要なしにクラスタの再構成を可能にする。

【００３９】

図７には、クラスタを検出および再構成する別の実施例が示される。各ノードは、バイアス信号の有無などのバス信号の変化を検出するために相互接続バスを監視する（ブロック７０５）。ノードがトポロジ変化を検出すると（ブロック７１０）、当該ノードはバスを介してバスリセット信号を送信し、自己構成機構を開始する。物理層２１０によって管理されるこの機構は３つの段階、すなわち、バス初期設定、ツリー識別および自己識別を含み得る。バス初期設定中にアクティブなノードが識別され、樹状の論理トポロジが構築される（ブロック７１５）。各々のアクティブなノードはアドレスが割当てられ、ルートノードが動的に割当てられ、ノードマップが新しいトポロジで再構築または更新される（ブロック７２０）。バス自体が構成されると、ノードはバスにアクセスできる。各ノード上のデータベースインスタンスがトポロジ変化について通知される（ブロック７２５）と、データベースロックマネージャは、共有データベースがクラスタ全体にわたって適切に管理され得るように当該変化に応じて再構成される（ブロック７３０）。

【００４０】

ネットワーク１３５などのネットワーク接続が他の方法で実現され得ることが理解されるだろう。たとえば、これは、ノベル（Novell）、マイクロソフト（Microsoft）、アーティソフト（Artisoft）および他の販売業者から入手できるソフトウェアなどの通信またはネットワーキングソフトウェアを含み得、ＴＣＰ／ＩＰ、ＳＰＸ、ＩＰＸ、ならびにツイストペア、同軸もしくは光ファイバケーブル、電話線、衛星、マイクロ波中継装置、無線周波数信号、変調されたＡＣ電力線、および／または当業者に公知の他のデータ伝送線を介する他のプロトコルを用いて動作し得る。ネットワーク１３５は、ゲートウェイまたは類似の機構を介して他のネットワークに接続可能であり得る。相互接続バス１６０のプロトコルが無線バージョンを含み得ることも理解されるだろう。

【００４１】

図８を参照すると、データベースクラスタ８００のためのハートビートシステムの一実施例が示される。ハートビートシステムは、ノードが、それらがアクティブであり機能し

ていることを示す信号またはメッセージを周期的に生成する機構である。当該機構はまた、ノードが、生成された信号に基づいてクラスタにおける他のノードの健康状態または状態を判断することを可能にする。図示のとおり、クラスタ８００はノード８０５および８１０を含むが、ノードがいくつクラスタに接続されてもよい。例示されたノードは図１に示されるノードと類似の構成を有し得る。しかしながら、例示の目的で簡略化された構成が示される。

【００４２】

ノード８０５、８１０は、データベースファイルなどのファイルを保持する記憶装置８１５へのアクセスを共有する。当該ノードは、共有ストレージネットワーク８２０によって記憶装置８１５に接続される。一実施例においては、ネットワーク８２０はＩＥＥＥ１３９４通信プロトコルに基づいている。互いに通信するために、ノード８０５、８１０および記憶装置８１５はＩＥＥＥ１３９４ネットワークコントローラ８２５を含む。ネットワークコントローラ８２５は相互接続バスコントローラ１６５に類似しており、一実施例においては、各々の装置に差込まれるネットワークカードである。代替的には、コントローラはノード内に固定され得る。ネットワークコントローラ８２５は、ケーブルが各装置間に接続され得るように１つ以上のポートを含む。加えて、他の種類のネットワーク接続、たとえばＩＥＥＥ１３９４プロトコルまたは他の類似のプロトコル規格に基づいた無線接続が用いられてもよい。

【００４３】

図８をさらに参照すると、各ノードは、記憶装置８１５上のファイルへのアクセスを制御するデータベースインスタンス８３０を含む。リソースがデータベースクラスタ８００におけるノード間で共有されるので、各ノードはそれらの健康状態を他のノードに知らせるためのロジックを含み、ネットワーク上の他のノードの健康状態を判断するためのロジックを含む。たとえば、ハートビートロジック８３５は、予め定められた時間間隔内でハートビートメッセージを生成しかつ伝送するようプログラミングされる。ハートビートメッセージは状態信号とも称される。予め定められた時間間隔は選択されたいかなる間隔であってもよいが、典型的には、数ミリ秒から数秒のオーダ、たとえば３００ミリ秒から５秒のオーダである。このため、当該間隔が１秒である場合、各ノードは１秒ごとにハートビートメッセージを伝送するだろう。

【００４４】

一実施例においては、ネットワークロードはハートビート時間間隔を決定する際に要因として用いられる。たとえば、ハートビートメッセージが同じネットワーク上でデータとして伝送される場合、ネットワーク上のハートビートメッセージの周波数が高いことにより、データ伝送プロセスに遅延がもたらされる可能性がある。図８は、この状況によって影響を受ける可能性のあるネットワークを示し、図９は、異なるネットワーク上でハートビートシステムを実現することによりネットワークトラフィックの量を減ずるネットワークを示す。図８および図９のネットワークも非共有アーキテクチャとして構成され得ることがさらに理解されるだろう。

【００４５】

図８をさらに参照すると、各ノードからのハートビートメッセージが集められ、定数ファイル８４０に記憶される。この実施例においては、定数ファイル８４０は、記憶装置８１５内に規定される１つ以上のファイルまたは区域であり、当該記憶装置８１５は共有ファイルも保持する。クラスタ８００における各ノードは定数ファイル８４０内にアドレス空間が割当てられて、そこにそのハートビートメッセージが記憶される。定数ファイル８４０の空間は典型的には等しく分割され、各ノードに割当てられるが、他の構成も可能であり得る。こうして、定数ファイル８４０は、ファイルが１つのデータ構造として論理的に規定され得るにもかかわらず、クラスタ全体に対する１つのファイルとしてではなく各ノードに対する別個のファイルとして実現され得る。定数ファイルは、１つ以上の記憶場所、レジスタまたは他の種類の記憶区域に記憶されるスタック、アレイ、表、リンクされたリスト、テキストファイルまたは他の種類のデータ構造として実現され得る。ノードの

定数空間が一杯になれば、新しいメッセージが受信されるとその空間における最も古いメッセージが押出されるかまたは上書きされる。

【００４６】

図９には、データベースクラスタ９００およびハートビートシステムの別の実施例が示される。この実施例においては、ノード９０５および９１０は定数ネットワーク９２０を介して定数装置９１５と通信する。定数ネットワーク９２０は共有ストレージネットワーク９２５とは別個のネットワークである。こうして、ノードは、定数ネットワークとは異なるネットワークバスを用いて記憶装置９３０上の共有ファイルにアクセスする。定数ネットワーク９２０は、上述のようにノード間相互接続ネットワークの一部であり得る。定数装置９１５は、クラスタにおけるノードから受取ったハートビートメッセージを記憶するために定数ファイルを保持するよう構成されたデータストレージを含む。

【００４７】

図９をさらに参照すると、ノード９０５、９１０は定数装置９１５に接続され、ＩＥＥＥ１３９４通信プロトコルに従って互いに通信する。各ノードおよび定数装置９１５は、先述のコントローラに類似のＩＥＥＥ１３９４コントローラ９３５を含む。別個のネットワークがファイルとのデータ通信のために構成されるので、各ノードは、記憶装置９３０と通信を行なう別個の共有ネットワークコントローラ９４０を含む。共有ネットワークコントローラ９４０はＩＥＥＥ１３９４コントローラであり得るかまたはファイバチャネルプロトコルなどの他のネットワークプロトコルであり得る。各ノード内のデータベースインスタンス９４５は共有ネットワークコントローラ９４０を介してデータ要求を処理する。

【００４８】

ハートビートロジック９５０はハートビート機構を制御し、ＩＥＥＥ１３９４コントローラ９３５を用いて定数装置９１５と通信する。このアーキテクチャを用いると、既存のデータベースクラスタ９００内における定数装置９１５の追加または交換は、既存のネットワークへの影響を最小限に抑えて容易に実行可能である。また、ハートビート機構が別個のネットワークを介して処理されるので、共有ストレージネットワーク９２５上のトラフィックを減らすことにより、データ処理要求に対するより迅速な応答が可能となる。図８および図９のクラスタがノード間相互接続ネットワークを含み得ることも理解されるだろう。

【００４９】

図１０には、以下において共に定数ファイルと称される定数ファイル８４０または定数装置９１５で実行されるハートビートシステムの例示的な方法論１０００が示される。定数ファイルがデータベースクラスタ内で構成されかつ起動されると、定数ファイル内のメモリがクラスタにおけるノードの各々に割当てられる（ブロック１００５）。定数ファイルは等しく分割され得、さらに、各ノードに割当てられ得るかまたは他の割当が規定され得る。定数ファイルがアクティブになると、当該定数ファイルはＩＥＥＥ１３９４プロトコルに従って各ノードからハートビートメッセージを受信する（ブロック１０１０）。各ハートビートメッセージはノード識別子を含み、当該ノード識別子が、メッセージと当該メッセージの時間を示すタイムスタンプとを送信するノードを識別する。次いで、定数ファイルが受信した各メッセージがそのノードの割当てられた位置に記憶され（ブロック１０１５）、受信された各々のハートビートメッセージのために当該プロセスが繰返される。

【００５０】

各ノードのために、ハートビートメッセージが、受信される順序で定数ファイルに記憶される。こうして、最後に受信されたタイムスタンプと現在の時間とを比較することにより、当該システムは、どのノードがそれらのハートビートメッセージをアクティブに送信しているかを決定することができる。この情報は、ノードがアクティブであるか否かを示し得る。たとえば、ノードが予め定められた数の連続したタイムスタンプを見落とした場合、問題が起り得ると想定され得る。１つのメッセージを含む各ノードのために任意の数

のメッセージが記憶され得る。上述のように、各ノードのハートビートロジックは、予め定められた間隔でハートビートメッセージを生成および伝送するようプログラミングされる。こうして、定数ファイルからデータを読出すことにより、当該ロジックは、いくつかの間隔が見落とされたかどうかを判断することができる。この種類の状態チェックロジックはハートビートロジック８３５または９５０の一部であってもよく、図１１に関連してより詳細に説明される。

【００５１】

図１１は、ノードの健康状態または状態を判断するための方法論の例を示す。上述のとおり、ハートビートロジックは、予め定められた時間間隔で各ハートビートメッセージを生成し、かつ当該メッセージを定数ファイルに伝送するためのロジックを含む。いかなる所望のときにも、ノードのハートビートロジックはそのクラスタ構成ファイルを更新して、アクティブなノードの現在の組を決定し、いずれかのノードが機能を停止したかまたはさもなければネットワークから除去されたかどうかを決定し得る。また、クラスタ全体に亘ってこの決定の同期を取ることもできる。状態チェックロジック（図示せず）が、以下のとおりにこのタスクを実行するようハートビートロジックの一部としてプログラミングされ得る。

【００５２】

状態チェックを開始するために、定数ファイルが、ノードの各々に対するタイムスタンプされた情報を検査するために読出される（ブロック１１０５）。各ノードのために記憶されるタイムスタンプされたデータに基づき、ロジックは、特定のノードが定数ファイルに書込まれた最後のメッセージの時間に基づいて依然として機能しているかどうかを判断し得る（ブロック１１１０）。しきい値が設定されることにより、問題が存在し得ることが当該決定によって示される前に予め定められた数のタイムスタンプを見落とすことが可能となる。たとえば、ノードは２つの連続したタイムスタンプを見落としてもよいが、第３のスタンプが見落とされた場合、当該ノードは適切に機能し得ない。しきい値は、他の値、たとえば１の値に設定されてもよい。

【００５３】

ノードが指定された量のタイムスタンプメッセージを見落とした場合（ブロック１１２０）、それは必ずしもノードが機能を停止したことを意味するものではないかもしれない。ノードがＩＥＥＥ１３９４規格に従って定数ファイルに接続されるので、付加的な状態チェックを実行することができる。先に説明したとおり、ＩＥＥＥ１３９４バスはアクティブであり、当該バスに接続された各装置は、隣接するノードが機能を停止しているかどうかまたはネットワークから取除かれているかどうかを検出し得る。この付加的な情報は、ノードの健康状態をよりよく判断するのに役立ち得る。状態ロジックは、定数ファイルからのタイムスタンプ情報とＩＥＥＥ１３９４コントローラによって保持されるノードマップデータとを比較し得る。

【００５４】

たとえば、ノードがそのタイムスタンプを見落とし（ブロック１１２０）、当該ノードがノードマップにおいてアクティブなノードでない（ブロック１１２５）場合、当該ノードがダウンしていると推定されるかまたはネットワークから除去されたと判断される（ブロック１１３０）。しかしながら、ノードがそのタイムスタンプを見落とすが当該ノードがノードマップにおいて依然としてアクティブである場合、ことによると当該ノードがハングアップするか、または他の何らかの遅延がクラスタに存在する可能性がある（ブロック１１３５）。この場合には、プロセスは、そのノードに対する定数ファイルを随意に再チェックして、新しいタイムスタンプが受信されたかどうか、起こり得る遅延を示すメッセージが生成され得るかどうか、および／またはノードがアクティブなノードのリストから除去され得るかどうかを判断し得る。

【００５５】

判断ブロック１１２０を再び参照すると、ノードがそのタイムスタンプを見落とさなければ、当該ノードはおそらく適切に機能している。しかしながら、ノードがノードマップ

においてアクティブであるかどうかをチェックすることにより追加の決定が下され得る（ブロック１１４０）。ノードがアクティブであれば（ブロック１１４５）、当該ノードは適切に機能している。ノードがアクティブでなければ（ブロック１１５０）、ネットワークバスのエラーが存在する可能性がある。こうして、定数ファイルおよびＩＥＥＥ１３９４バスのノードマップの両方からの情報を用いて、ノードの健康状態のより詳細な分析が決定され得る。さらに、共有ストレージネットワーク９２５がＩＥＥＥ１３９４バスでもある実施例における図９に示されるクラスタ構成においては、２つの別個のネットワークノードマップが保持される。付加的なノードマップがまた、上述の比較プロセスおよび状態チェックに含まれてもよい。

【００５６】

図１１を再び参照すると、簡略化された実施例が実現され得る。判断ブロック１１２０では、ノードがそのタイムスタンプを書込むのに失敗した場合、ロジックはそのノードが機能していないことを宣言し、データベースインスタンスのクラスタ構成ファイルからそれを除去し得る。このプロセスにおいてはノードマップは検査されない。

【００５７】

この明細書中に記載されるさまざまな記憶装置が定数ファイルを割当てるための定数装置を含み、多数の方法で実現され得ることが理解されるだろう。たとえば、記憶装置は、磁気ディスクドライブまたは光ディスクドライブ、テープドライブ、電子メモリなどの１つ以上の専用の記憶装置を含み得る。記憶装置はまた、コンピュータ、サーバ、携帯用処理装置、または、データを保持するためのストレージ、メモリもしくはこれらの組合せを含む類似の装置を含み得る。記憶装置はまた、いかなるコンピュータ読取可能媒体であってもよい。

【００５８】

この発明のシステムおよび方法のさまざまな構成要素を実現するための好適なソフトウェアは、ここに呈示される教示やプログラミング言語およびツール、たとえばJava（登録商標）、Pascal、Ｃ＋＋、Ｃ、ＣＧＩ、Perl、ＳＱＬ、ＡＰＩ、ＳＤＫ、アセンブリ、ファームウェア、マイクロコードならびに／または他の言語およびツールなどを用いて当業者によって容易に提供される。ソフトウェアとして具体化される構成要素は、コンピュータを所定の態様で動作させるコンピュータ読取可能／実行可能な命令を含む。当該ソフトウェアは製品としてであってもよく、および／または、先に規定したようにコンピュータ読取可能媒体に記憶されてもよい。

【００５９】

この発明がその実施例を説明することにより例示され、その実施例がかなり詳細に説明されてきたが、出願人の意図は、添付の特許請求の範囲をこのような詳細に制限するかまたは何らかの方法で限定することではない。付加的な利点および変形例が当業者には容易に明らかとなるだろう。したがって、この発明は、そのより広範な局面においては、特定の詳細、代表的な装置ならびに図示および説明される具体例には限定されない。したがって、出願人の一般的な発明の概念の精神または範囲から逸脱せずにこのような詳細からの逸脱が可能である。

【図面の簡単な説明】

【００６０】

【図１】この発明に従ったクラスタノードの一実施例を示すシステム図である。
【図２】図１の相互接続バスコントローラを示す例図である。
【図３】共有ディスククラスタアーキテクチャの一例を示す図である。
【図４】非共有クラスタアーキテクチャの一例を示す図である。
【図５】相互接続バスを用いてデータを通信する方法論の一例を示す図である。
【図６】トポロジ変化を検出する方法論の一例を示す図である。
【図７】トポロジ変化を検出する方法論の別の例を示す図である。
【図８】ハートビートシステムを含むクラスタの別の実施例を示す図である。
【図９】ハートビートシステムの別の実施例を示す図である。

【図１０】定数ファイルを保持する方法論の一例を示す図である。
【図１１】定数ファイルを用いてノードの状態を決定する方法論の一例を示す図である。

【図１】

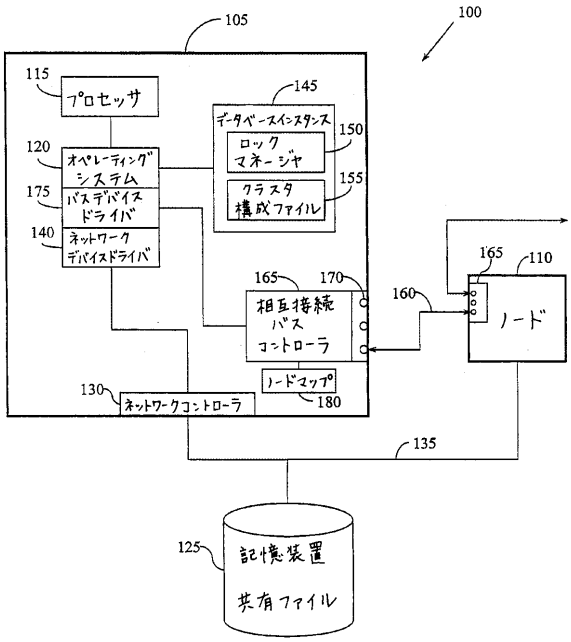


Figure 1

【図２】

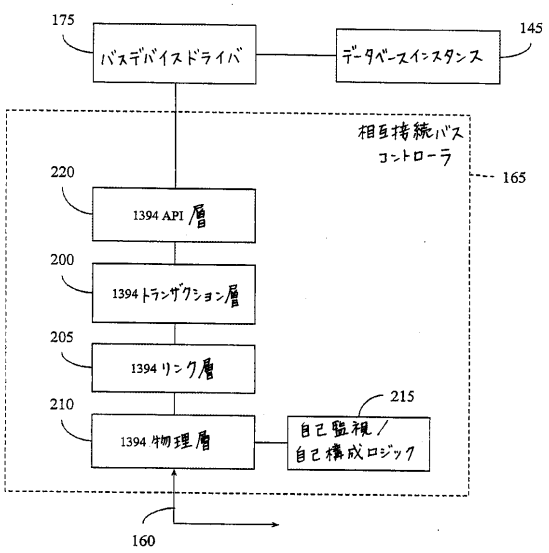


Figure 2

【図３】

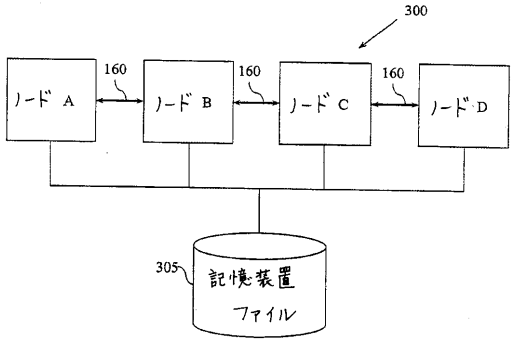


Figure 3

【図４】

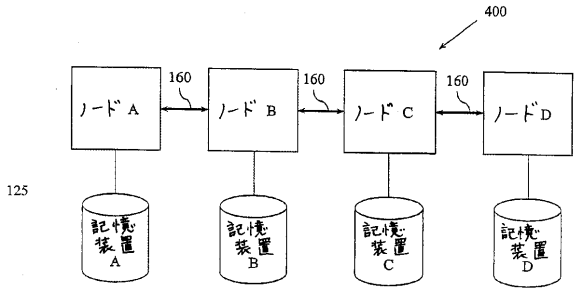


Figure 4

【図５】

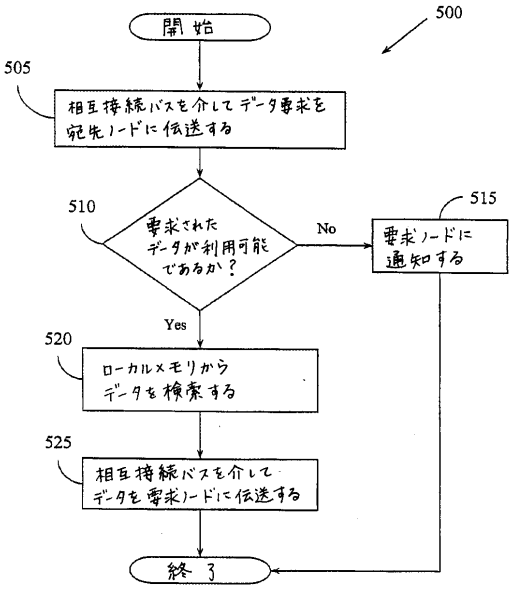


Figure 5

【図６】

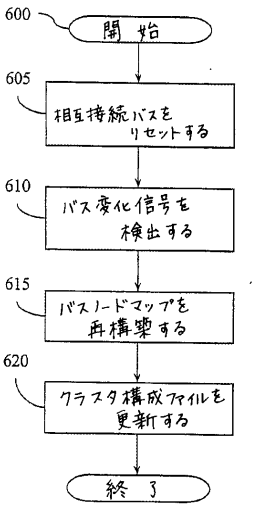


Figure 6

【図７】

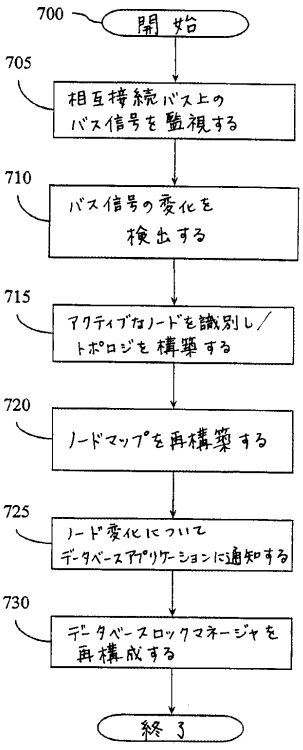


Figure 7

【図８】

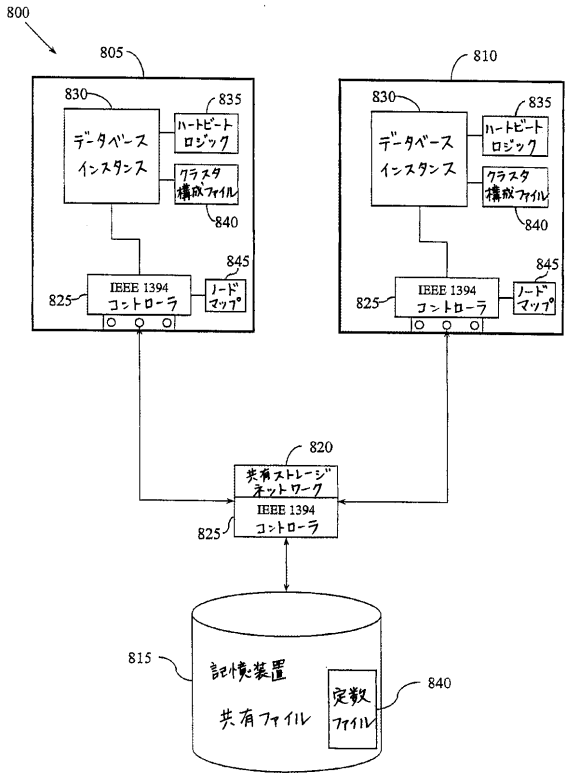


Figure 8

【図９】

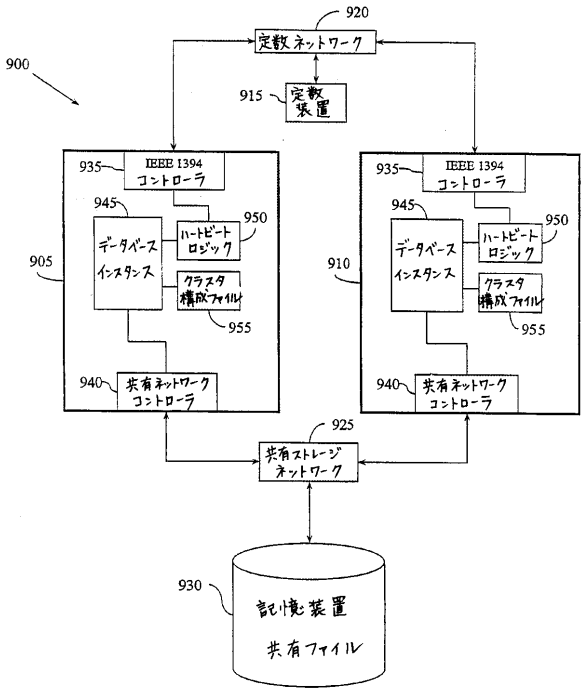


Figure 9

【図１０】

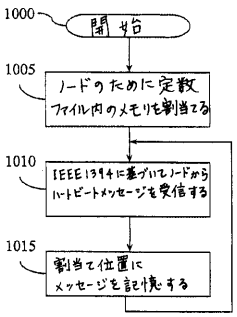


Figure 10

【図１１】

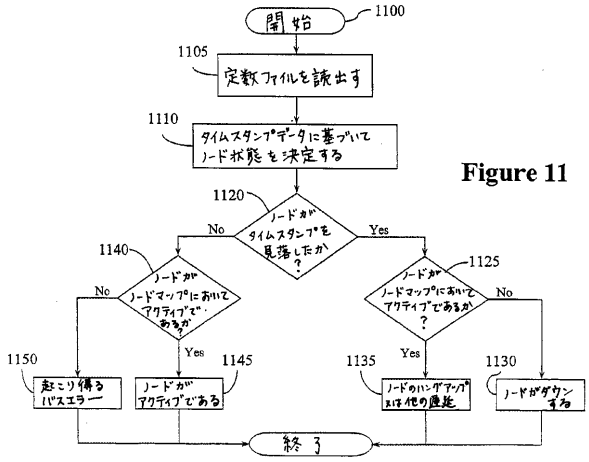


Figure 11

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International Application No
PCT/US 03/37172

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G06F11/14

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, WPI Data, PAJ, INSPEC

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	GREGORY F. PFISTER: "In Search of Clusters" 1998, PRENTICE HALL , NEW JERSEY 013899 , XP002294683 page 401, paragraph 4 page 402, paragraph 1-3 page 419, paragraph 3-5 page 420, paragraphs 4,8-10 page 421, paragraph 1 page 431, paragraph 3 page 432, paragraphs 1,2 figure 109 ----- -/—	1-35

[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
8 September 2004	01/10/2004

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	Noll, J

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

INTERNATIONAL SEARCH REPORT

International Application No
PCT/US 03/37172

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	YAMAGIWA S ET AL: "MEASTRO-LINK: A HIGH PERFORMANCE INTERCONNECT FOR PC CLUSTER" JOURNAL ARTICLE, 1998, pages 421-425, XP000905433 page 422, lines 1-12 figures 1,2	1-35
Y	"Oracle 8i Parallel Server" ORACLE, 'Online! December 1999 (1999-12), XP002275984 Retrieved from the Internet: URL:http://www.lle.rochester.edu/pub/support/Oracle_817/paraserv.817/a76968.pdf> 'retrieved on 2004-03-30! pages 1-2, paragraphs 1,2 pages 2-2, paragraph 1-7 pages 9-14, paragraph 1-3 figures 2-2,2-3, figures 2-6	8,18,30
Y	"VERITAS Cluster Server 3.5" VERITAS, 'Online! July 2002 (2002-07), XP002275985 Retrieved from the Internet: URL:http://ftp.support.veritas.com/pub/support/products/ClusterServer_UNIX/249745.pdf> 'retrieved on 2004-04-02! page 7; figure 1 page 9, paragraph 3 - page 10, paragraph 1	12,14, 16,27,35
A	US 6 393 485 B1 (MCCARTY RICHARD JAMES ET AL) 21 May 2002 (2002-05-21) column 1 - column 5 figures 3-9	1-35
A	"SunCluster3.0 12/01 Concepts Guide" SUN MICROSYSTEMS, 'Online! December 2001 (2001-12), XP002275987 Retrieved from the Internet: URL:http://docs-pdf.sun.com/816-2027/816-2027.pdf> 'retrieved on 2004-03-30! page 10, line 6 - line 10 page 28, line 1 - line 6 page 11; figures 2-1 page 25; figures 3-1	1-35

INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No
PCT/US 03/37172

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 6393485 B1	21-05-2002	CA 2284376 A1	29-04-2000
		KR 2000028685 A	25-05-2000
		TW 497071 B	01-08-2002

フロントページの続き

(81)指定国 AP(BW,GH,GM,KE,LS,MW,MZ,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,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,BY,BZ,CA,CH,CN,CO,CR,CU,CZ,DE,DK,DM,DZ,EC,EE,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,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,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 コーカーツ, ウィム・エイ
アメリカ合衆国、９４０７０ カリフォルニア州、サン・カルロス、ウォールナット・ストリート、８０１、ナンバー・６

Ｆターム(参考) 5B042 GA12 GC08 GC10 JJ08 JJ15 JJ29 KK09