(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第5872731号
(P5872731)

(45)発行日　平成28年3月1日(2016.3.1)　(24)登録日　平成28年1月22日(2016.1.22)

(51)Int.Cl.		F I		
H04L 12/28	(2006.01)	H04L	12/28	200M
H04L 12/713	(2013.01)	H04L	12/713	
H04L 12/70	(2013.01)	H04L	12/70	100Z

請求項の数 15　(全 14 頁)

(21)出願番号	特願2015-501687(P2015-501687)
(86)(22)出願日	平成25年2月28日(2013.2.28)
(65)公表番号	特表2015-511099(P2015-511099A)
(43)公表日	平成27年4月13日(2015.4.13)
(86)国際出願番号	PCT/US2013/028346
(87)国際公開番号	WO2013/142023
(87)国際公開日	平成25年9月26日(2013.9.26)
審査請求日	平成27年7月1日(2015.7.1)
(31)優先権主張番号	13/425,127
(32)優先日	平成24年3月20日(2012.3.20)
(33)優先権主張国	米国(US)

早期審査対象出願

(73)特許権者　501113353
シマンテック　コーポレーション
Ｓｙｍａｎｔｅｃ　Ｃｏｒｐｏｒａｔｉｏｎ
アメリカ合衆国，カリフォルニア州　９４０４３，マウンテン　ビュー，エリス　ストリート　３５０
(74)代理人　100147485
弁理士　杉村　憲司
(74)代理人　100134119
弁理士　奥町　哲行

最終頁に続く

(54)【発明の名称】クラスタの複数のノードのそれぞれに対してリンクの障害の検出を伝えるためのコンピュータ実装方法、非一時的なコンピュータ可読媒体およびコンピュータシステム

(57)【特許請求の範囲】

【請求項１】

クラスタの複数のノードのそれぞれに対してリンクの障害の検出を伝えるためのコンピュータ実装方法であって、該方法が、

該クラスタの特定のノードの特定のリンクの障害を含む、リンクダウン事象を検出する工程と、

該クラスタの各ノードに通信可能に連結される別のリンクであってリンクダウン事象通知メッセージをブロードキャストするための専用のリンクを使用して、該クラスタの該複数のノードに該特定のリンクの該障害の通知を伝播する工程であって、これにより該ノードが該リンクダウン事象を平行して処理する、工程と、

を含み、該特定のリンクの該障害の該通知の該伝播は、該クラスタの該ノードが、対応するハートビートの期限切れにより該リンクダウン事象を知るよりも前に、該通知を受け取るように実行される、コンピュータ実装方法。

【請求項２】

前記クラスタの前記複数のノードに前記特定のリンクの前記障害の通知を伝播する工程が、前記クラスタの前記ノードにメッセージをブロードキャストする工程であって該メッセージは前記特定のリンクの前記障害を前記クラスタの前記ノードに通知する、工程を更に含む、請求項１に記載の方法。

【請求項３】

前記別のリンクを使用して前記クラスタの該ノードに該メッセージをブロードキャスト

する工程を更に含む、請求項２に記載の方法。
【請求項４】
前記クラスタの各ノードに通信可能に連結された別のリンクを確保する工程であって、該リンクがリンクダウン事象通知メッセージのブロードキャスト専用である、工程が、該別のリンクを、前記クラスタのユーザーに公開されていないプライベートリンクとして確保する工程を更に含む、請求項３に記載の方法。
【請求項５】
前記リンクダウン事象を処理する前に前記ブロードキャストされたメッセージが前記クラスタの前記ノードに受信されたことを確実にする工程を更に含む、請求項２に記載の方法。
【請求項６】
前記リンクダウン事象を処理する前に前記ブロードされたキャストメッセージが、前記クラスタの前記ノードにより受信されたことを確実にする工程が、前記リンクダウン事象を処理する前に前記クラスタの前記ノードのそれぞれから前記ブロードキャストされたメッセージを受信したことを確認する確認応答を受け取るのを待つ工程を更に含む、請求項５に記載の方法。
【請求項７】
所与の時間内に前記ブロードキャストされたメッセージの受信を確認する確認応答を受信していないことに応答して、前記ブロードキャストされたメッセージを再送信する工程を更に含む、請求項５に記載の方法。
【請求項８】
前記リンクの障害検出に対応して、前記リンクの該障害の前記通知を伝播する前に特定の猶予期間待つ工程を更に含む、請求項１に記載の方法。
【請求項９】
前記クラスタの前記複数のノードに前記特定のリンクの前記障害の通知を伝播する工程が、前記クラスタの前記ノードのすべてにアクセス可能な中央集中型コンピューティングデバイスに、前記特定のリンクの前記障害の通知を送信する工程を更に含む、請求項１に記載の方法。
【請求項１０】
前記クラスタの各ノードが、リンクの障害の通知のために、特定の周波数で、前記中央集中型コンピューティングデバイスのポーリングを行い、これにより前記クラスタの前記ノードが、前記中央集中型コンピューティングデバイスのポーリングにより前記特定のリンクの前記障害を知る、請求項９に記載の方法。
【請求項１１】
前記中央集中型コンピューティングデバイスが、前記クラスタの前記ノードに、前記特定のリンクの前記障害の前記通知を送信する、請求項９に記載の方法。
【請求項１２】
前記クラスタの前記特定のノードの前記特定のリンクの障害を検出する工程が、前記特定のノードのオペレーティングシステムから、前記特定のリンクの該障害の通知を受け取る工程を更に含む、請求項１に記載の方法。
【請求項１３】
クラスタの複数のノードそれぞれにリンクの障害の検出を伝えるための、コンピュータプログラムを格納する少なくとも１つの非一時的なコンピュータ可読媒体であって、該コンピュータプログラムが、
該クラスタの特定のノードの特定のリンクの障害を含む、リンクダウン事象を検出するためのプログラムコードと、
該クラスタの各ノードに通信可能に連結される別のリンクであってリンクダウン事象通知メッセージをブロードキャストするための専用のリンクを使用して、該クラスタの該複数のノードに該特定のリンクの該障害の通知を伝播し、これにより該ノードが該リンクダウン事象を平行して処理するためのプログラムコードと、

を含み、該特定のリンクの該障害の該通知の該伝播は、該クラスタの該ノードが、対応するハートビートの期限切れにより該リンクダウン事象を知るよりも前に、該通知を受け取るように実行される、コンピュータ可読媒体。

【請求項１４】

前記クラスタの前記複数のノードに前記特定のリンクの前記障害の通知を伝播するための前記プログラムコードが、前記クラスタの前記ノードにメッセージをブロードキャストするためのプログラムコードを更に含み、該メッセージは、前記特定のリンクの前記障害を前記クラスタの前記ノードに通知する、請求項１３に記載のコンピュータプログラム。

【請求項１５】

クラスタの複数のノードのそれぞれに対してリンクの障害の検出を伝播するためのコンピュータシステムであって、該コンピュータシステムが、

プロセッサと、

コンピュータメモリと、

該クラスタの特定のノードの特定のリンクの障害を含むリンクダウン事象を検出するための手段と、

該クラスタの各ノードに通信可能に連結される別のリンクであってリンクダウン事象通知メッセージをブロードキャストするための専用のリンクを使用して、該クラスタの該複数のノードに該特定のリンクの該障害の通知を伝播し、これにより該ノードが該リンクダウン事象を平行して処理する、手段と、

を含み、該特定のリンクの該障害の該通知の該伝播は、該クラスタの該ノードが、対応するハートビートの期限切れにより該リンクダウン事象を知るより前に、該通知を受け取るように実行される、コンピュータシステム。

【発明の詳細な説明】

【技術分野】

【０００１】

本開示は、概ね、コンピュータクラスタの管理に関し、より具体的には、リンク障害のクラスタ全体の一貫した検出に関する。

【背景技術】

【０００２】

高可用性クラスタ（別名、ＨＡクラスタ、又はフェイルオーバークラスタ）は、ダウンタイムを最小限に抑え、サーバアプリケーションの実行をサポートする、コンピュータ（ノード）群である。高可用性クラスタは、個々のシステムコンポーネントに障害が生じた場合にも連続したサービスを提供するために、冗長なコンピュータリソース群を使用する。より具体的には、高可用性クラスタは、複数のサーバ、複数のネットワーク接続、冗長なデータ記憶装置などを提供することにより、単一の障害ポイントを排除する。クラスタがない場合、特定のアプリケーションを実行しているサーバに障害が生じた場合、このアプリケーションは、このサーバが復帰するまで利用できないことになる。高可用性クラスタリングでは、サーバ（又は、例えばネットワークアダプタ、記憶装置などの、これにより使用されている特定のコンピューティングリソース）の障害が検出される。障害が生じているサーバで実行中であったアプリケーションを使用できる状態に維持できるようにするため、工程が自動的に行われる。これは、異なるネットワークリソース（例えばネットワークアダプタ）を使用するか、又は別のコンピューティングシステム（すなわちクラスタの別のノード）上でアプリケーションを自動的に再起動する、サーバとアプリケーションの再起動の形態である場合がある。このプロセスは「フェイルオーバー」と呼ばれる。高可用性クラスタはしばしば、企業データベース、重要な業務用アプリケーション、電子商取引ウェブサイトなどの重要なサーバアプリケーション用に使用される。そのようなアプリケーションでは、特に、銀行業務及び通信などの特定の業種において、たとえ短時間であってもダウンタイムが起こると非常に高価なものとなる場合がある。

【０００３】

コンピュータ記憶装置においては、論理ボリューム管理が、大容量記憶装置のスペースを割り当てる柔軟な方法である。特に、ボリュームマネージャは、下位の物理的パーティションを連結させ、整合させ、又は組み合わせて、より大きな仮想パーティションにすることができる。よって管理者は、可能性としてはシステム利用を中断することなく、論理ボリュームのリサイズ又は移動を行うことができる。クラスタボリュームマネージャは、クラスタの複数のノードにわたってボリューム論理管理を拡張し、これにより各ノードは同じ論理ボリュームレイアウトを認識し、全ノードの全ボリュームリソースの同じ状態を認識する。クラスタボリューム管理の下では、ディスク又はボリューム構成に対して、クラスタ内のどのノードから変更が行われても、クラスタの全ノードにより認識される。クラスタレベルのボリューム管理をサポートするため、新しいノードがクラスタに加わったとき、及び既存のノードがクラスタから離れるときに、クラスタが再構成される。

【発明の概要】

【発明が解決しようとする課題】

【０００４】

クラスタは、データ用ノードと管理通信との間の相互接続（リンク）を使用している。リンクに障害が生じた場合は、修正処置を行う必要がある。フェイルオーバー及びその他のクラスタ処置からの復帰の効率と適時性は、クラスタがリンク障害を検出及び処理可能なスピードに依存する。個々のノードは、リンクのいずれかに障害が生じているかどうかを検出するのに、オペレーティングシステムのサービスを使用することができる。２つのノードがスイッチなしで（クロスオーバーケーブルを使って）直接接続されているとき、両方のノードがリンクダウン通知を受け取り、これにより両ノードが平行してこの事象に対処することができる。しかしながら、２つ以上のノードがスイッチを使用して連結されている場合は、ローカルノードだけがリンク障害をリアルタイムで知る。よって、クラスタのネットワークトポロジーにより、典型的には単一のノードのみ、又はローカライズされたノードのサブセットのみが、リンク障害を即座に知る。従来、他のノードはその後に、クラスタ内の各ノードの状況を監視するのに使用されるクラスタ全体のハートビート機構を介して、接続障害を知る。ハートビート機構は、個々のノードがローカルリンク障害を知るのにかかる時間に比べ、比較的遅い。クラスタの適切な管理は、修正処置を行う前に、クラスタの各ノードがリンク障害を認識することに依存し得る。ハートビート機構を介して各ノードが障害を知るまでの時間によって、ダウンした接続に必要なクラスタ再構成又はフェイルオーバー処置を遅らせることになり、これにより、クラスタ管理（例えばフェンシングアービトレーション（arbitration）判断の実行）に更なる問題を引き起こす。更に、ノード又はクラスタが異なるタイミングでリンク障害を知ると、不正又は望ましくない修正処置が生じる場合がある。

【０００５】

これらの問題に対処することが望ましいと考えられる。

【課題を解決するための手段】

【０００６】

リンク障害の通知は、クラスタ内の複数のノードのそれぞれに伝えられる。例えば特定のノードのオペレーティングシステムを介して、クラスタの特定のノードの特定のリンクの障害が検出される。特定のリンクの障害の通知は、クラスタの複数のノードに伝播され、これによりノードがリンクダウン事象を平行して処理する。リンク障害の通知の伝播は、クラスタのノードが、対応するハートビートの期限切れによりリンクダウン事象を知るよりも前に、通知を受け取るように実行される。いくつかの実施形態において、リンクの障害の通知は、リンクが即座に復帰した場合、特定の猶予期間待った後に限り、伝播される。

【０００７】

一実施形態において、リンク障害の通知は、特定のリンクの障害をクラスタのノードに通知するメッセージをブロードキャストすることにより、伝播される。別のリンクが確保され、これはクラスタの各ノードに接続され、リンクダウン事象通知メッセージをブロー

ドキャストするための専用となる。この別のリンクは、クラスタのノードにメッセージをブロードキャストするのに使用することができる。この別のリンクは、クラスタのユーザーに公開されないプライベートリンクの形態であり得るが、必ずしもプライベートリンクである必要はない。一実施形態において、ブロードキャストメッセージがクラスタのノードに受信されたことを確実にしてから、リンクダウン事象が処理される。これは、例えば、クラスタのノードそれぞれからブロードキャストメッセージを受信したことを確認する確認応答を受け取るのを待つことによりなされてもよい。一実施形態において、ブロードキャストメッセージの受信を確認する確認応答が、所与の時間以内に各ノードから受信されない場合、そのブロードキャストメッセージは再送信される。

【０００８】

別の一実施形態において、リンク障害の通知は、ノードのすべてにアクセス可能な中央集中型コンピューティングデバイスに通知を送信することにより、クラスタのノードに伝播される。クラスタのノードはリンク障害の通知のため特定の周波数で中央集中型コンピューティングデバイスのポーリングを行い、そのポーリングから特定のリンクの障害を知ることができる。一実施形態において、ポーリングの代わりに（又はこれに加えて）、中央集中型コンピューティングデバイスは、特定のリンクの障害の通知を、クラスタのノードに送信する。

【０００９】

この「発明の概要」及び後述の「発明を実施するための形態」に記述される特徴及び利点は、すべてを包含したものではなく、特に、本特許の図面、明細書、及び請求項の見地から、関連分野の当業者には、数多くの付加的な特徴及び利点が明らかとなろう。更に、本明細書に使用されている表現は、読みやすさと説明の目的のために主に選択されており、発明の主題を限定又は制限するために選択されたものではなく、そのような発明の主題を決定するには、請求項に依ることが必要である。

【図面の簡単な説明】

【００１０】

【図１】いくつかの実施形態による、リンク障害管理システムが実装され得る代表的なネットワークアーキテクチャのブロック図である。

【図２】いくつかの実施形態による、リンク障害管理システムを実装するのに好適なコンピュータシステムのブロック図である。

【図３】いくつかの実施形態による、例示的なクラスタトポグラフィーにおいてリンクの障害を検出する、リンク障害管理システムのブロック図である。

【図４】いくつかの実施形態による、ブロードキャストメッセージを介してリンク障害の通知をクラスタのノードに伝播する、リンク障害管理システムのブロック図である。

【図５】いくつかの実施形態による、中央集中型コンピューティングデバイスを介してリンク障害の通知をクラスタのノードに伝播する、リンク障害管理システムのブロック図である。

【図６】一実施形態によるリンク障害管理システムのオペレーションのフローチャートである。

【図７】別の一実施形態によるリンク障害管理システムのオペレーションのフローチャートである。

【００１１】

図には、単に例示目的のため、様々な実施形態が示されている。当業者には、本明細書に記述されている原理から逸脱することなく、本明細書に示されている構造及び方法の別の実施形態を採用し得ることが、下記の議論から容易に理解されよう。

【発明を実施するための形態】

【００１２】

図１は、リンク障害管理システム１０１が実装され得る代表的なネットワークアーキテクチャ１００を示すブロック図である。図示のネットワークアーキテクチャ１００は、複数のクライアント１０３Ａ、１０３Ｂ及び１０３Ｎ、並びに複数のサーバ１０５Ａ及び１

０５Ｎを含む。図１において、リンク障害管理システム１０１は、サーバ１０５Ａ上にあるものとして図示されている。これは単に一例であり、様々な実施形態においてこのシステム１０１の様々な機能性が、サーバ１０５、クライアント１０３で実例を示すことができ、あるいは、複数のクライアント１０３及び／又はサーバ１０５間に分散できることが理解されよう。

【００１３】

クライアント１０３及びサーバ１０５は、例えば図２及び下記に示すもののような、コンピュータシステム２１０を使用して実装することができる。クライアント１０３及びサーバ１０５は、図２に関連して下記で述べられるように、例えばネットワークインタフェース２４８又はモデム２４７を介して、ネットワーク１０７に通信可能に連結されている。クライアント１０３は、例えばウェブブラウザ又はその他のクライアントソフトウェア（図示なし）を用いて、サーバ１０５上のアプリケーション及び／又はデータにアクセス可能である。

【００１４】

図１には３つのクライアント及び２つのサーバが一例として示されているが、実際にはもっと多くの（又はもっと少ない）数のクライアント１０３及び／又はサーバ１０５を配備することができる。一実施形態において、ネットワーク１０７はインターネットの形態である。他の実施形態において、他のネットワーク１０７又はネットワークベースの環境を使用することができる。

【００１５】

図２は、リンク障害管理システム１０１を実装するのに好適なコンピュータシステム２１０のブロック図である。クライアント１０３とサーバ１０５の両方が、そのようなコンピュータシステム２１０の形態で実装され得る。図示のように、コンピュータシステム２１０の１つの構成要素がバス２１２である。バス２１２は、コンピュータシステム２１０の他の構成要素、例えば、少なくとも１つのプロセッサ２１４、システムメモリ２１７（例えばランダムアクセスメモリ（ＲＡＭ）、リードオンリーメモリ（ＲＯＭ）、フラッシュメモリ）、入出力（Ｉ／Ｏ）コントローラ２１８、スピーカシステム２２０などの外部音声装置に通信可能に連結された音声出力インタフェース２２２、ディスプレイ画面２２４などの外部ビデオ出力装置に通信可能に連結されたディスプレイアダプタ２２６、１つ以上のインタフェース（例えばシリアルポート２３０、ユニバーサルシリアルバス（ＵＳＢ）のレセプタクル２３０、パラレルポート（図示なし）など）、キーボード２３２に通信可能に連結されたキーボードコントローラ２３３、少なくとも１つのハードディスク２４４（又はその他の形態の磁気媒体）に通信可能に連結された記憶装置インタフェース２３４、フロッピーディスク２３８（「フロッピー」は登録商標、以下同じ）を受容するよう構成されたフロッピーディスクドライブ２３７、ファイバーチャネル（ＦＣ）ネットワーク２９０と接続するよう構成されたホストバスアダプタ（ＨＢＡ）インタフェースカード２３５Ａ、ＳＣＳＩバス２３９に接続するよう構成されたＨＢＡインタフェースカード２３５Ｂ、光ディスク２４２を受容するよう構成された光ディスクドライブ２４０、バス２１２に（例えばＵＳＢレセプタクル２２８を介して）連結されたマウス２４６（又はその他のポインティングデバイス）、バス２１２に（例えばシリアルポート２３０を介して）接続されたモデム２４７、及び、例えばバス２１２に直接連結されたネットワークインタフェース２４８と、通信可能に連結される。

【００１６】

他の構成要素（図示なし）は、同様の方法で接続され得る（例えば文書スキャナ、デジタルカメラ、プリンタなど）。逆に、図２に示されているすべての構成要素は必ずしも存在する必要はない。構成要素は、図２に示されるものとは異なる手法で相互接続することができる。

【００１７】

バス２１２は、プロセッサ２１４と、上述のように、ＲＯＭ及び／又はフラッシュメモリ並びにＲＡＭを含み得るシステムメモリ２１７との間のデータ通信を可能にする。ＲＡ

Ｍは典型的に、オペレーティングシステム及びアプリケーションプログラムがロードされるメインメモリである。ＲＯＭ及び／又はフラッシュメモリは、他のコードと共に、特定の基本的なハードウェアオペレーションを制御するベーシックインプット・アウトプットシステム（ＢＩＯＳ）を含み得る。アプリケーションプログラムはローカルコンピュータ可読媒体（例えばハードディスク２４４、光ディスク２４２）に格納され、システムメモリ２１７にロードされ、そしてプロセッサ２１４により実行され得る。アプリケーションプログラムは、例えばネットワークインタフェース２４８又はモデム２４７を介して、離れた場所（すなわち、離れたところに配置されたコンピュータシステム２１０）からシステムメモリ２１７にロードすることもできる。図２において、リンク障害管理システム１０１は、システムメモリ２１７内にあるものとして図示されている。リンク障害管理システム１０１のはたらきは、図３に関連して下記で更に詳しく説明される。

【００１８】

記憶装置インタフェース２３４は、１つ以上のハードディスク２４４（及び／又はその他の標準記憶媒体）に連結されている。（複数の）ハードディスク２４４は、コンピュータシステム２１０の一部であってもよく、又は、物理的に別であってもよく、他のインタフェースシステムを介してアクセスされてもよい。

【００１９】

ネットワークインタフェース２４８及び／又はモデム２４７は、例えばインターネットなどのネットワーク１０７に、直接的又は間接的に、通信可能に連結され得る。そのような連結は、有線又は無線であり得る。

【００２０】

図３は、いくつかの実施形態による、一例のクラスタ３０１トポグラフィーにおいてリンク３０５の障害を検出する、リンク障害管理システム１０１を示す。上述のように、リンク障害管理システム１０１の機能性はクライアント１０３又はサーバ１０５上にあるか、又は、複数のコンピュータシステム２１０間で分散され得る。これには、リンク障害管理システム１０１の機能性がネットワーク１０７にわたるサービスとして提供されるクラウドベースのコンピューティング環境内が含まれる。リンク障害管理システム１０１は図３において単独のエンティティとして示されているが、図示されているリンク障害管理システム１０１は、機能性の集まりを示し、これは所望に応じて単一又は複数のモジュールとして実例を示すことができる（リンク障害管理システム１０１の具体的な複数モジュールの実例は図３及び４に示されている）。例示目的のため、リンク障害管理システム１０１は、図示のクラスタ３０１の各ノード３０３上にあり、かつノード３０３上で起こる機能性を管理しているものとして図示されている。実際には、リンク障害管理システム１０１は所望に応じて、中央集中化されてよく、又はクラスタ３０１の複数ノード３０３にわたって分散されていてもよい。

【００２１】

リンク障害管理システム１０１のモジュールは、任意のコンピュータシステム２１０のシステムメモリ２１７（例えばＲＡＭ、ＲＯＭ、フラッシュメモリ）内で実例を示すことができ（例えばオブジェクトコード又は実行可能画像として）、これによって、コンピュータシステム２１０のプロセッサ２１４がモジュールを処理する際、コンピュータシステム２１０がその関連する機能性を実行することが理解されよう。本明細書において使用される用語「コンピュータシステム」、「コンピュータ」、「クライアント」、「クライアントコンピュータ」、「サーバ」、「サーバコンピュータ」及び「コンピューティングデバイス」は、記述される機能性を実行するよう設定及び／又はプログラムされた、１つ以上のコンピュータを意味する。加えて、リンク障害管理システム１０１の機能性を実装するためのプログラムコードは、コンピュータ可読記憶媒体に格納することができる。この文脈において、例えば磁気又は光学記憶媒体など、任意の形態の有形のコンピュータ可読記憶媒体を使用することができる。本明細書において使用される用語「コンピュータ可読記憶媒体」は、下位の物理的媒体とは別体の電気信号は意味しない。

【００２２】

図３に示すように、リンク障害管理システム１０１は、障害のあるリンク３０５の検出を、クラスタ３０１全体に迅速に伝えることを可能にする。明確にするため、図３は３つのノード３０３Ａ、３０３Ｂ、及び３０３Ｃを含むクラスタ３０１を示す。この分野において、クラスタ３０１は典型的に、この数桁以上の数のノード３０３を含み得ることが理解されよう。ノード３０３は、ネットワークリンク３０５及びハブ３０７を使用して接続されている。図３に示されている例において、各ノード３０３は、別のリンク３０５で、３つの異なるハブ３０７Ａ、３０７Ｂ、及び３０７Ｃに接続されている。別の実施形態において、クラスタ３０１のノード３０３を接続するネットワークトポロジーを形成するのに、配備するリンク３０５及びハブ３０７の数は、これより多くても少なくてもよい。

【００２３】

図３に示すように、ノード３０３Ａにあるリンク障害管理システム１０１のリンク障害検出モジュール３０９は、ノード３０３Ａの任意のリンク３０５（すなわちリンク３０５Ａ～Ｃ）が障害を生じた時点を検出する。（他のノード３０３それぞれの類似のモジュールは、それぞれのローカルリンク３０５の障害を検出する）。リンク障害検出モジュール３０９は、ローカルリンク３０５がダウンしたときに即座に通知を受け取るため、ローカルノード３０３のオペレーティングシステムサービスを利用することができる。例えば、リンク３０５Ａが壊れた場合、ノード３０３Ａのリンク障害検出モジュール３０９は、ノード３０３Ａ上のオペレーティングシステム（図示なし）から即座に通知を受け取ることができる。しかしながら、リンク障害検出モジュール３０９はそのローカルノード３０３のリンク３０５の障害を検出するだけであるため、ノード３０３Ｂ、３０３Ｃはこのリンクダウン事象に気づいていないことになる。所与のノード（例えば３０３Ａ）のオペレーティングシステムは、そのノード３０３にとってローカルであるリンク３０５の障害を検出するだけであるため、このようになる。

【００２４】

従来、クラスタ３０１の他のノード３０３は、ハートビートを監視することにより、ノード３０３Ａのリンクの障害を知り得る。しかしながら、上述のように、ハートビートに依存すると、離れたノード３０３が、離れたリンク３０５に障害が生じたと結論づけられるまでに、過度に長い遅れが必要となる。図示の例において、ノード３０３Ａはリンク３０５Ａの障害を即座に知り得るが、ノード３０３Ｂ及び３０３Ｃが、リンク３０５Ａに障害が生じたと結論づけられるようになるまでには、対応するハートビートが期限切れになるのを待たなければならない。上記で説明したように、このような状況は問題である。それは、クラスタ３０１の管理にとって、各ノード３０３がリンクダウン事象を平行して処理する（すなわち、ダウンしているリンク３０５に対応する）ことが重要であり、また同時に、ハートビート機構に依存して促進され得るよりも迅速にこれを成し遂げるのが望ましいためである。

【００２５】

図４～５に示すように、ローカルノード３０３のリンク障害管理システム１０１は、クラスタ３０１の他のノード３０３に、リンク３０５の障害の通知を伝播し、これによりクラスタ３０１のノード３０３がリンクダウン事象を平行して処理できるようにする。下記で詳しく記述されるように、異なる実施形態において、リンク障害管理システム１０１は、別の方法を使用して、クラスタ３０１のノード３０３に、リンク３０５の障害の通知伝播を実行する。これらの異なる実施形態において、リンク３０５の障害の通知伝播は、対応するハートビートが期限切れになる前に、クラスタ３０１のノード３０３がその通知を受け取るように実行されることが理解される。よって、リンク障害管理システム１０１によって、クラスタ３０１のノード３０３が、リンクダウン事象を平行に、ハートビート機構に依存して達成され得るよりも迅速に処理することができるようになる。

【００２６】

図４は、リンク障害管理システム１０１が、リンク３０５の障害に関する情報を、ブロードキャストメッセージ４０３を介して、クラスタ３０１のノード３０３に伝播する一実施形態を示す。一実施形態において、ローカルリンク（例えば図３に示される例示的なト

ポロジーにおけるリンク３０５Ａ）の障害を知ったことに対応して、そのローカルノード（例えばノード３０３Ａ）上にあるリンク障害管理システム１０１のブロードキャストモジュール４０１が、検出されたリンク３０５の障害のことをクラスタ３０１のノード３０３に通知する、ブロードキャストメッセージ４０３を生成かつ送信する。よって、クラスタ３０１の他のノード３０３はすべて、リンク３０５の障害の通知を同時に受け取り、よってこの事象の処理を平行して行うことができる。別の実施形態において、この目的にどのリンク３０５が利用できるかに応じて、別のリンク３０５を使用して、ブロードキャストメッセージ４０３を送信することができる。

【００２７】

例えば、一実施形態において、クラスタ３０１内のいくつかのリンク３０５が、標準クラスタリンク３０５として構成され（例えば図３に示す例示的なトポロジーにおいて、ハブ３０７Ａを介して相互接続しているリンク３０５Ａ、３０５Ｄ、及び３０５Ｇ、並びに、ハブ３０７Ｂを介して相互接続しているリンク３０５Ｂ、３０５Ｅ、及び３０５Ｈ）、ここにおいて一組のリンク３０５は、リンクダウン事象ブロードキャストメッセージ４０３のために確保される（例えば、ハブ３０７Ｃを介して相互接続しているリンク３０５Ｃ、３０５Ｆ及び３０５Ｉ）。このシナリオにおいて、ノード３０３Ａのリンク障害検出モジュール３０９が、リンク３０５Ａ及び／又はリンク３０５Ｂがダウンしたことを検出すると、専用のリンク３０５Ｃを使用して、ブロードキャストメッセージ４０３をノード３０３Ｂ及び３０３Ｃへと送信することができる。一実施形態において、障害検出ブロードキャストメッセージ４０３専用の別のリンク３０５は、リンク障害管理システム１０１のプライベートリンクであり、クラスタ３０１のユーザーに対して公開されない。このシナリオにおいて、専用リンク３０５は、クラスタ３０１内のすべてのノード３０３に接続され、リンクダウン事象情報を送信する際にのみ使用される。これにより、リンク３０５の障害検出を示すブロードキャストメッセージ４０３の通信のための専用チャネルが確保される。

【００２８】

別の一実施形態において、これらのリンク３０５は、リンク障害管理システム１０１のプライベートリンクであるが、リンクダウン事象情報以外のトラフィックに使用することができる。更に別の一実施形態において、リンクダウン事象情報の送信に使用されるリンク３０５は、リンク障害管理システム１０１のためのプライベートリンクではなく、例えば利用可能な帯域幅に基づいて、他の当事者により他のトラフィックのために使用することができる。いくつかの実施形態において、リンクダウン事象情報を通信するための専用リンク３０５は利用できず、この場合、他のリンク３０５をこの目的のために使用することができる（例えば、高優先度リンク３０５又は低優先度リンクを含む既存の公共リンク３０５で、例えば他のものがダウンした場合など）。別の実施形態において、利用できるリンク３０５が何であれ、所望に応じてリンクダウン事象情報を送信するのに使用できることが理解されよう。

【００２９】

リンクダウン事象メッセージ４０３がブロードキャストされると、ローカルノード（例えば図３の３０３Ａ）のリンク障害管理システム１０１が工程を実施し、ブロードキャストメッセージ４０３をクラスタ３０１の他のノード３０３が確実に受け取ってから、リンクダウン事象の処理が行われる。これにより、クラスタ３０１の複数のノード３０３が、リンクダウン事象を同時に確実に処理することができる。一実施形態において、リンク障害管理システム１０１の確認応答受信モジュール４０５が、クラスタ３０１の他のノード３０３のそれぞれから、ブロードキャストメッセージ４０３の受信を確認するための、確認応答（ＡＣＫ）４０７の受信を待つ。他のノード３０３のそれぞれからＡＣＫ　４０７を受信するということは、クラスタ内のすべてのノード３０３がリンク３０５の障害を知っていることを示す。ブロードキャストメッセージ４０３自体、又は１つ以上のノード３０３からのＡＣＫ　４０７が、ネットワークにより中断されている場合、確認応答受信モジュール４０５は、クラスタ３０１の他のノード３０３それぞれからのＡＣＫ　４０７を

受信しない。一実施形態において、所与の一定時間内に、クラスタ３０１の各ノード３０３からＡＣＫ　４０７を受信していないことに対応して、ブロードキャストモジュール４０１はブロードキャストメッセージ４０３を再送信する。いくつかの例において、ブロードキャストメッセージ４０３は、１つ以上の予期されるＡＣＫ　４０７を受信していないことに対応して、複数回再送信することができる。再送信までの待ち時間の長さ、並びに再送信の回数は、実施形態間で変えることができる設計パラメータであることが理解されよう。いかなる場合でも、これらのパラメータは典型的に、確認応答受信モジュール４０５がＡＣＫ　４０７を待つ合計時間が、ノード３０３のステータス情報のクラスタ３０１の全体への伝播のために、ハートビート機構によって使用される時間スケールよりも更に短くなるように、設定される。やがてクラスタの他のノード３０３がハートビート機構を介してリンク３０５の障害を知ることが理解されよう。よって、ハートビート機構を介してリンク３０５の障害をノード３０３が知るのにかかる時間内に、すべてのＡＣＫ　４０７が受信されなかった場合のバックアップとして、クラスタ３０１のノード３０３は、リンクダウン事象を従来の方法で知ることができる。リンクダウン事象に関する情報をブロードキャストするため、及びその確認応答受信に使用されるハンドシェーキングに使用されるプロトコル及びフォーマットの具体的な実装は、所望に応じて実施形態間で変えることができることが理解されよう（例えばアトミックブロードキャスト、２相コミットなど）。

【００３０】

いくつかの実施形態において、リンク障害検出モジュール３０９がリンク３０５の障害を検出したとき、ブロードキャストモジュール４０１は猶予期間待ってから、ブロードキャストメッセージ４０３を送信する。この猶予期間の目的は、リンクがダウンしたけれどもほぼ即座に元に戻った場合に（例えば不安定なリンク）、リンクダウン事象のブロードキャストを避けることである。猶予期間の長さは可変の設計パラメータであるが、典型的には、ハートビート機構の時間スケールに比べ、かなり短い。例えば、０．５秒、１秒、又は２秒の猶予期間が使用され得る。そのような実施形態において、障害が生じたリンク３０５が猶予期間内に元に戻った場合、ブロードキャストメッセージ４０３は送信されない。一方、猶予期間が経過し、リンク３０５が依然としてダウンしている場合は、ブロードキャストモジュール４０１がメッセージ４０３を他のノード３０３に送信する手続きを行う。

【００３１】

図５は別の一実施形態を示し、ここにおいてリンク障害管理システム１０１は、中央集中型コンピューティングデバイス５０３を介してクラスタ３０１の他のノード３０３に、リンク３０５の障害の通知５０１を伝播する。この実施形態において、リンク障害検出モジュール３０９がリンク３０５の障害を検出した場合、他のノード３０３にメッセージ４０３をブロードキャストする代わりに（又はそれに加えて）、リンク障害管理システム１０１の送信モジュール５０５が、中央集中型コンピューティングデバイス５０３（例えば、サーバ１０５、ディスク２４４など）に通知５０１を送信する。この実施形態において、クラスタ３０１の各ノード３０３上のリンク障害管理システム１０１のポーリングモジュール５０７は、リンク３０５の障害の通知５０１のために、特定の周波数で中央集中型コンピューティングデバイス５０３のポーリングを行い、その過程でリンク３０５の障害を知る。使用する具体的なポーリング周波数は、可変の設計パラメータである。ポーリングの代わりに又はこれに加えて、中央集中型コンピューティングデバイス５０３は、リンクダウン事象に関する通知５０１を受信すると、その通知５０１をクラスタ３０１のノード３０３すべてに送信することができる。これらの実施形態は、例えば、リンクダウン事象のブロードキャストメッセージ４０３を送信するための、所与のノード３０３上のブロードキャストモジュール４０１が利用できるクラスタリンク３０５がない場合に、使用することができる。

【００３２】

図６は、一実施形態によるリンク障害管理システム１０１のオペレーションの工程を示

す。リンク障害検出モジュール３０９が、クラスタ３０１の特定のノード３０３の特定のリンク３０５の障害を検出する６０１。ブロードキャストモジュール４０１が、検出されたリンク３０５の障害を、クラスタ３０１のノード３０３を通知するメッセージ４０３をブロードキャストし６０３、これによりノード３０３は、対応するハートビートの期限切れによりリンクダウン事象を知るより前に、メッセージ４０３を受信する。確認応答受信モジュール４０５が、クラスタ３０１のノード３０３からのブロードキャストメッセージ４０３の受信を確認する確認応答４０７を受信する６０５。クラスタ３０１の各ノード３０３が、リンクダウン事象を平行に処理する６０７。

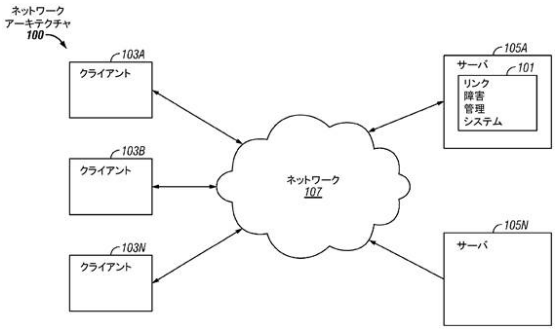
【００３３】

図７は、別の一実施形態によるリンク障害管理システム１０１のオペレーションの工程を示す。リンク障害検出モジュール３０９が、クラスタ３０１の特定のノード３０３の特定のリンク３０５の障害を検出する７０１。送信モジュール５０５が、リンクダウン事象に関する通知５０１を、中央集中型コンピューティングデバイス５０３に送信する７０３。クラスタ３０１のノード３０３のポーリングモジュール５０７が、リンク３０５の障害の通知５０１のため、特定の周波数で中央集中型コンピューティングデバイス５０３のポーリングを行い７０５、これによって、クラスタ３０１のノード３０３が、対応するハートビートの期限切れによって知るよりも前に、ポーリングからリンクダウン事象のことを知る。クラスタ３０１の各ノード３０３が、リンクダウン事象を平行に処理する７０７。

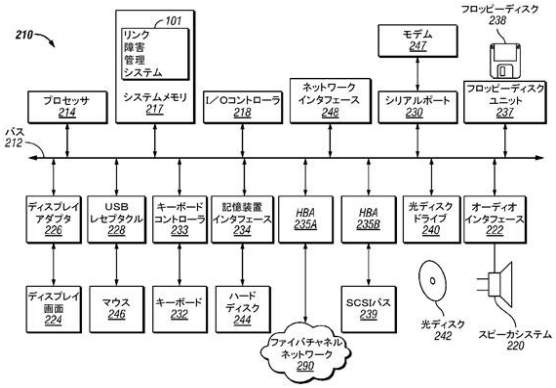
【００３４】

当該技術分野を知る業者には理解されるように、本発明は、その趣旨又は本質的特徴から逸脱することなく、他の具体的な形態で実施することができる。同様に、部分、モジュール、エージェント、マネージャー、構成要素、機能、手順、処置、階層、特徴、属性、方法論、データ構造、及びその他の態様の具体的な名称及び区分は、必須のものでも重要なものでもなく、本発明又はその特徴を実施するメカニズムは、別の名称、区分、及び／又は形式を有し得る。説明目的のための上記の記述は、特定の実施形態を参照して記述されている。しかしながら、上記の例示的議論は網羅的であることを意図したものではなく、開示されている正確な形態に限定することを意図したものでもない。上記の教示の見地から、数多くの改変及び変化が可能である。実施形態は、関連する原理及びその実際的な適用を最適に説明するよう選択かつ記述されており、これにより、他の当業者が、想到される特定の使用に好適であり得るような様々な改変を行うか行わないかにかかわらず、様々な実施形態を最適に利用できる。

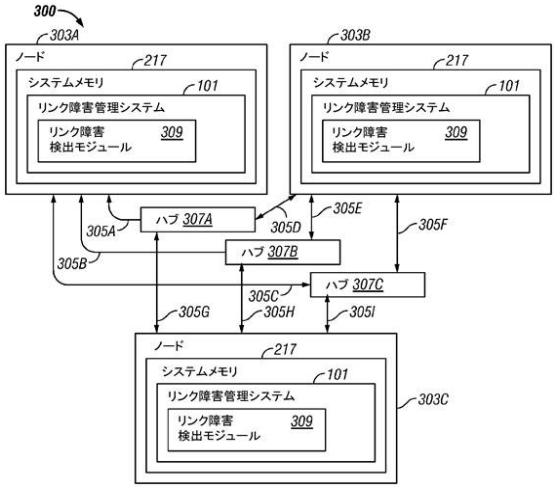
【図１】


ネットワーク
アーキテクチャ
100
103A
クライアント
103B
クライアント
103N
クライアント
ネットワーク
107
105A
サーバ
101
リンク
障害
管理
システム
105N
サーバ

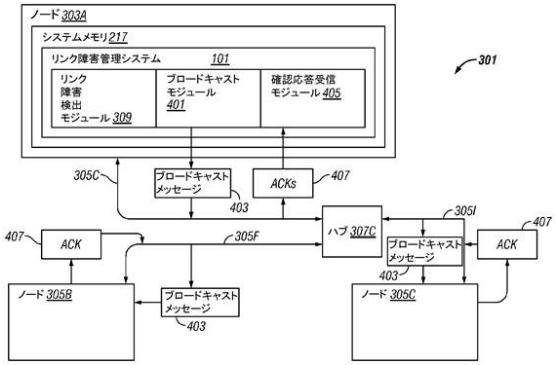
【図２】


210
101
リンク
障害
管理
システム
プロセッサ
214
システムメモリ
217
I／Oコントローラ
218
ネットワーク
インタフェース
248
モデム
247
シリアルポート
230
フロッピーディスク
238
フロッピーディスク
ユニット
237
バス
212
ディスプレイ
アダプタ
226
USB
レセプタクル
228
キーボード
コントローラ
233
記憶装置
インタフェース
234
HBA
235A
HBA
235B
光ディスク
ドライブ
240
オーディオ
インタフェース
222
ディスプレイ
画面
224
マウス
246
キーボード
232
ハード
ディスク
244
SCSIバス
239
光ディスク
242
スピーカシステム
220
ファイバチャネル
ネットワーク
290

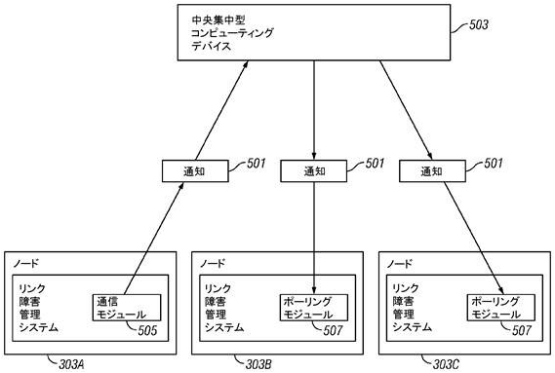
【図３】


300
303A
ノード
217
システムメモリ
101
リンク障害管理システム
リンク障害
検出モジュール
309
303B
ノード
217
システムメモリ
101
リンク障害管理システム
リンク障害
検出モジュール
309
ハブ 307A
ハブ 307B
ハブ 307C
305A
305B
305C
305D
305E
305F
305G
305H
305I
ノード
217
システムメモリ
101
リンク障害管理システム
リンク障害
検出モジュール
309
303C

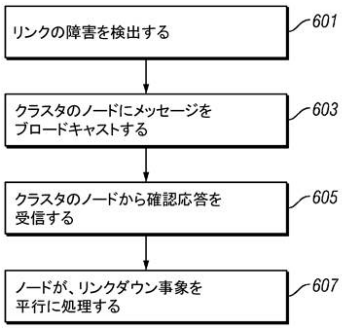
【図４】


ノード 303A
システムメモリ 217
リンク障害管理システム 101
リンク
障害
検出
モジュール 309
ブロードキャスト
モジュール
401
確認応答受信
モジュール 405
301
305C
ブロードキャスト
メッセージ
403
ACKs
407
ハブ 307C
305I
407
ACK
305F
ブロードキャスト
メッセージ
403
ACK
407
ノード 305B
ブロードキャスト
メッセージ
403
ノード 305C

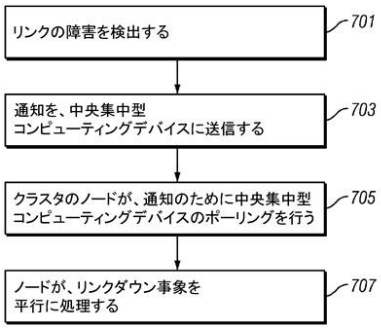
【図５】


中央集中型
コンピューティング
デバイス
503
通知
501
通知
501
通知
501
ノード
リンク
障害
管理
システム
通信
モジュール
505
303A
ノード
リンク
障害
管理
システム
ポーリング
モジュール
507
303B
ノード
リンク
障害
管理
システム
ポーリング
モジュール
507
303C

【図６】


リンクの障害を検出する
601
クラスタのノードにメッセージを
ブロードキャストする
603
クラスタのノードから確認応答を
受信する
605
ノードが、リンクダウン事象を
平行に処理する
607

【図７】


リンクの障害を検出する
701
通知を、中央集中型
コンピューティングデバイスに送信する
703
クラスタのノードが、通知のために中央集中型
コンピューティングデバイスのポーリングを行う
705
ノードが、リンクダウン事象を
平行に処理する
707

フロントページの続き

(72)発明者　カトカー・アモール
インド　マハラシュトラ州　４１１０２０　プネー　ボポディー　バウパティルロード　ラビラヘリテージ　アウィング　３０３
(72)発明者　アガーワル・オーム・プラカシュ
インド　マハラシュトラ州　４１１０３２　プネー　ティンジャーナガー　ロードナンバー１０シー　アサーヴァレジデンシー　フラットナンバー８
(72)発明者　セイカー・バービン
アメリカ合衆国　カリフォルニア州　９４０８７　サニーヴェール　インヴァネスウェイ　７５６

審査官　衣鳩　文彦

(56)参考文献　特開２００５－１２４１７１（ＪＰ，Ａ）
特開２００３－１７９６２９（ＪＰ，Ａ）
特表２００４－５１９０２４（ＪＰ，Ａ）
堀田　勇樹　他，耐故障並列計算を支援する自律的な故障検知機構，情報処理学会論文誌，２００５年　8月１５日，第46巻,第SIG12(ACS 11)号，p.236～244

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｈ０４Ｌ　１２／００～１２／９５５