

(19) 日本国特許庁 (JP)

(12) 特 許 公 報 (B2)

(11) 特許番号

特許第5247882号
(P5247882)

(45) 発行日 平成25年7月24日 (2013. 7. 24)

(24) 登録日 平成25年4月19日 (2013. 4. 19)

(51) Int. Cl.

F I

H04L 12/701 (2013.01)

H04L 12/56 100Z

請求項の数 13 (全 16 頁)

(21) 出願番号	特願2011-515327 (P2011-515327)	(73) 特許権者	391030332
(86) (22) 出願日	平成21年6月19日 (2009. 6. 19)		アルカテルルーセント
(65) 公表番号	特表2011-525766 (P2011-525766A)		フランス国、75007・パリ、 アブニ
(43) 公表日	平成23年9月22日 (2011. 9. 22)		ユ・オクターブ・グレアール、 3
(86) 国際出願番号	PCT/EP2009/057702	(74) 代理人	100094112
(87) 国際公開番号	W02009/156352		弁理士 岡部 譲
(87) 国際公開日	平成21年12月30日 (2009. 12. 30)	(74) 代理人	100085176
審査請求日	平成22年12月24日 (2010. 12. 24)		弁理士 加藤 伸晃
(31) 優先権主張番号	08290625.6	(74) 代理人	100104352
(32) 優先日	平成20年6月27日 (2008. 6. 27)		弁理士 朝日 伸光
(33) 優先権主張国	欧州特許庁 (EP)	(74) 代理人	100128657
			弁理士 三山 勝巳
		(74) 代理人	100160967
			弁理士 ▲濱▼口 岳久

最終頁に続く

(54) 【発明の名称】 後続リストを提供する方法

(57) 【特許請求の範囲】

【請求項 1】

複数のノード（0、20、50、87、112、140、179、200、211、223、240）を備えるピアツーピア・オーバーレイ・ネットワーク（N）においてルーティングのために後続リスト（L）を提供する方法であって、前記後続リスト（L）は、前記ピアツーピア・オーバーレイ・ネットワーク（N）のトポロジ構造に関して、前記複数のノード（0、20、50、87、112、140、179、200、211、223、240）のうちの第1のノード（200）に続いて後続となる少なくとも2つの後続ノード（211、223）の識別子を指定し、前記方法が、

前記後続リスト（L）にリストされた前記少なくとも2つの後続ノード（211、223）が同一の物理ハードウェア上で稼働するかどうかを決定するステップと、

前記少なくとも2つの後続ノード（211、223）が同一の物理ハードウェア上で稼働する場合、前記ピアツーピア・オーバーレイ・ネットワーク（N）の前記トポロジ構造に関して前記少なくとも2つの後続ノード（211、223）の後続となり、前記少なくとも2つの後続ノード（211、223）とは異なる物理ハードウェア上で稼働する、前記複数のノード（0、20、50、87、112、140、179、200、211、223、240）のうちの少なくとも1つの追加のノード（240）の識別子を前記後続リスト（L）に追加するステップと、

結合ノード（220）が前記後続リスト（L）にリストされた前記後続ノード間の前記ピアツーピア・オーバーレイ・ネットワーク（N）を結合するとき、前記結合ノード（2

10

20

20)の前記識別子を前記後続リスト(L)に追加するステップと、
前記結合ノード(220)が稼働する前記物理ハードウェアを決定するステップと、
前記少なくとも1つの追加のノード(240)を除く、異なる物理ハードウェアで稼働する前記後続リスト(L)にリストされた後続ノードの前記合計数が、事前定義された冗長因数と等しい場合、前記少なくとも1つの追加のノード(240)の識別子を前記後続リスト(L)から削除するステップと
を備える方法。

【請求項2】

前記少なくとも2つの後続ノード(211、223)が同一の物理ハードウェア上で稼働する場合、前記異なる物理ハードウェア上で稼働する前記後続リスト(L)にリストされた後続ノードの合計数が事前定義された冗長因数と等しくなるように、前記ピアツーピア・オーバーレイ・ネットワーク(N)の前記トポロジ構造に関して前記少なくとも2つの後続ノード(211、223)の後続となり、前記少なくとも2つの後続ノード(211、223)とは異なる物理ハードウェア上で稼働する、前記複数のノード(0、20、50、87、112、140、179、200、211、223、240)の同数の追加のノード(240)の識別子を前記後続リスト(L)に追加するステップをさらに備える、請求項1に記載の方法。

10

【請求項3】

前記少なくとも2つの後続ノード(211、223)および／または前記少なくとも1つの追加のノード(240)が稼働する物理ハードウェアを、それぞれ前記少なくとも2つの後続ノード(211、223)および／または前記少なくとも1つの追加のノード(240)のプラットフォームから情報を取り出すことによって決定するステップをさらに備える、請求項1に記載の方法。

20

【請求項4】

前記結合ノード(220)が稼働する物理ハードウェアを、前記結合ノード(220)のプラットフォームから情報を取り出すことによって決定するステップをさらに備える、請求項1に記載の方法。

【請求項5】

前記少なくとも2つの後続ノード(211、223)および／または前記少なくとも1つの追加のノード(240)が稼働する物理ハードウェアを、それぞれ前記少なくとも2つの後続ノード(211、223)および／または前記少なくとも1つの追加のノード(240)とリンクされた外部管理ユニットから情報を取り出すことによって決定するステップをさらに備える、請求項1に記載の方法。

30

【請求項6】

前記結合ノード(220)が稼働する物理ハードウェアを、前記結合ノード(220)とリンクされた外部管理ユニットから情報を取り出すことによって決定するステップをさらに備える、請求項1に記載の方法。

【請求項7】

前記少なくとも2つの後続ノード(211、223)および／または前記少なくとも1つの追加のノード(240)が稼働する物理ハードウェアを、それぞれ前記ピアツーピアネットワーク(N)のネットワークオペレータおよび／または前記少なくとも2つの後続ノード(211、223)および／または前記少なくとも1つの追加のノード(240)のユーザから対応する情報を取り出す手段を用いて決定するステップをさらに備える、請求項1に記載の方法。

40

【請求項8】

前記結合ノード(220)が稼働する物理ハードウェアを、前記ピアツーピアネットワーク(N)のネットワークオペレータおよび／または前記結合ノード(220)のユーザから対応する情報を取り出す手段を用いて決定するステップをさらに備える、請求項1に記載の方法。

【請求項9】

50

前記少なくとも2つの後続ノード（211、223）および／または前記少なくとも1つの追加のノード（240）が稼働する物理ハードウェアを、前記ピアツーピア・オーバーレイ・ネットワーク（N）のアプリケーションレイヤの待ち時間測定を用いて決定するステップをさらに備える、請求項1に記載の方法。

【請求項10】

前記結合ノード（220）が稼働する物理ハードウェアを、前記ピアツーピア・オーバーレイ・ネットワーク（N）のアプリケーションレイヤの待ち時間測定を用いて決定するステップをさらに備える、請求項1に記載の方法。

【請求項11】

複数のノード（0、20、50、87、112、140、179、200、211、223、240）を備えるピアツーピア・オーバーレイ・ネットワーク（N）の第1のノード（200）であって、前記ピアツーピア・オーバーレイ・ネットワーク（N）においてルーティングするために後続リスト（L）を保持し、前記後続リスト（L）は、前記ピアツーピア・オーバーレイ・ネットワーク（N）のトポロジ構造に関して前記第1のノード（200）に続いて後続となる少なくとも2つの後続ノード（211、223）の識別子を指定し、前記第1のノード（200）は、前記後続リスト（L）にリストされた前記少なくとも2つの後続ノード（211、223）が同一の物理ハードウェア上で稼働するかどうかを決定し、前記少なくとも2つの後続ノード（211、223）が同一の物理ハードウェア上で稼働する場合、前記ピアツーピア・オーバーレイ・ネットワーク（N）の前記トポロジ構造に関して前記少なくとも2つの後続ノード（211、223）の後続となり、前記少なくとも2つの後続ノード（211、223）とは異なる物理ハードウェア上で稼働する、前記複数のノード（0、20、50、87、112、140、179、200、211、223、240）のうちの少なくとも1つの追加のノード（240）の識別子を前記後続リスト（L）に追加し、結合ノード（220）が前記後続リスト（L）にリストされた前記後続ノード間の前記ピアツーピア・オーバーレイ・ネットワーク（N）を結合するとき、前記結合ノード（220）の前記識別子を前記後続リスト（L）に追加し、前記結合ノード（220）が稼働する前記物理ハードウェアを決定し、前記少なくとも1つの追加のノード（240）を除く、異なる物理ハードウェアで稼働する前記後続リスト（L）にリストされた後続ノードの前記合計数が、事前定義された冗長因数と等しい場合、前記少なくとも1つの追加のノード（240）の識別子を前記後続リスト（L）から削除するように適合された制御ユニットを備える第1のノード（200）。

【請求項12】

前記制御ユニットは、ルーティング・ポリシーに基づいて前記第1のノード（200）から後続リスト（L）のノードにピアツーピア・メッセージをルーティングするようにさらに適合され、前記後続リスト（L）の前記ノードが前記ピアツーピア・オーバーレイ・ネットワーク（N）において到達可能ではない場合、前記ルーティング・ポリシーに基づいて前記ピアツーピア・メッセージを前記第1のノード（200）から前記少なくとも1つの追加のノード（240）にルーティングする、請求項11に記載の第1のノード（200）。

【請求項13】

複数のノード（0、20、50、87、112、140、179、200、211、223、240）を備えるピアツーピア・オーバーレイ・ネットワーク（N）においてルーティングのために後続リスト（L）を提供するコンピュータプログラム製品であって、前記後続リスト（L）は、前記ピアツーピア・オーバーレイ・ネットワーク（N）のトポロジ構造に関して、前記複数のノード（0、20、50、87、112、140、179、200、211、223、240）のうちの第1のノード（200）に続いて後続となる少なくとも2つの後続ノード（211、223）の識別子を指定し、前記コンピュータプログラム製品は、前記第1のノード（211）によって実行されるとき、

前記後続リスト（L）にリストされた前記少なくとも2つの後続ノード（211、223）が同一の物理ハードウェア上で稼働するかどうかを決定するステップと、

前記少なくとも2つの後続ノード（211、223）が同一の物理ハードウェア上で稼働する場合、前記ピアツーピア・オーバーレイ・ネットワーク（N）の前記トポロジ構造に関して前記少なくとも2つの後続ノード（211、223）の後続となり、前記少なくとも2つの後続ノード（211、223）とは異なる物理ハードウェア上で稼働する、前記複数のノード（0、20、50、87、112、140、179、200、211、223、240）のうちの少なくとも1つの追加のノード（240）の識別子を前記後続リスト（L）に追加するステップと、

結合ノード（220）が前記後続リスト（L）にリストされた前記後続ノード間の前記ピアツーピア・オーバーレイ・ネットワーク（N）を結合するとき、前記結合ノード（220）の前記識別子を前記後続リスト（L）に追加するステップと、

10

前記結合ノード（220）が稼働する前記物理ハードウェアを決定するステップと、
前記少なくとも1つの追加のノード（240）を除く、異なる物理ハードウェアで稼働する前記後続リスト（L）にリストされた後続ノードの前記合計数が、事前定義された冗長因数と等しい場合、前記少なくとも1つの追加のノード（240）の識別子を前記後続リスト（L）から削除するステップとを実行する、コンピュータプログラム。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、ピアツーピア・オーバーレイ・ネットワークにおいてルーティングのために後続リスト（successor list）を提供する方法、ならびに前記方法を実行するためのネットワーク・ノードおよびコンピュータプログラム製品に関する。

20

【背景技術】

【0002】

ピアツーピア（=P2P）・オーバーレイ・ネットワークは、VoIP（たとえば、Skype）またはファイル共有（たとえば、eMule）のような、幅広い多数のアプリケーションに使用される（VoIP=Voice over Internet Protocol）。高スケーラビリティ、自己組織化、および耐障害性のような特徴は、集約的に提供され使用されるリソースで分散されたアーキテクチャおよびデータストレージによって達成される。P2Pネットワークは、ピアとしての、つまり、以下の説明においてネットワークのクライアントおよびサーバの両方としてのノードアクションから成り、「ノード」および「ピア」という用語は同義的に使用される。ネットワークの各ノードは、オーバーレイ・アルゴリズムに関して確立され、メッセージ伝送のために使用される他のノードへの1つまたは複数の論理リンクを保持する。

30

【0003】

最新技術のP2Pネットワークは、基礎となるネットワーク・インフラストラクチャが不可知のオーバーレイ固有のアルゴリズムに基づいて論理トポロジ構造を構築する。各ノードは、ブロードキャストの方法で（非構造化オーバーレイ、たとえばGnutella）、またはIDベースの方法で（分散ハッシュテーブル（=DHT: Distributed Hash Table）を使用する構造化オーバーレイ、たとえばChord）、メッセージルーティングに使用される他のノードへの1つまたは複数のリンクを保持する（ID=identification（識別）/identifier（識別子））。一部のシステムは、ピアの非構造化グループとの混成アーキテクチャを使用し、それらのグループはより大規模なトポロジで構造化される（たとえば、Skype）。

40

【0004】

P2Pネットワークは、全体として非集中的である。オーバーレイに参加しているピアは、不適切に、しかも不定期にネットワークを去ることもある。ノードに障害が発生した場合にP2Pネットワークのトポロジ構造を保持するために、各ノードが直接の隣接ノードへの複数のリンクを保持することが必須である。Chordにおいて、たとえば、各ピアは、後続の隣接ピアのうちの1つまたは複数が正常終了せずにP2Pオーバーレイを去る場合にP2Pリング・トポロジを安定させる、いわゆる「冗長後続リスト」を保持する

50

。

【0005】

図1aは、複数のノード0、20、50、87、112、140、179、200、211、223、240を備える従来技術の構造化P2Pオーバーレイ・ネットワークNを示す。特に、図1aはChordリングNを表す。ノードの参照符号0、20、50、87、112、140、179、200、211、223、240は、ノード0、20、50、87、112、140、179、200、211、223、240のノードIDを表すことも意図されている。Chordにおいて、ノードは、それらの先行ノードおよび後続ノードへの直接接続を保持し、その結果リング・トポロジが形成される。

【0006】

P2PのChordリングNにおいて、リング・トポロジ構造は、冗長後続リスト(=RSL:redundant successor list)と呼ばれることもある後続リストを使用することにより安定状態が保持される。図1aに示されるように、複数のノード0、20、50、87、112、140、179、200、211、223、240の第1のノード200は、第1のノード200に続いて後続となる2つの後続ノード211、223の識別子を指定する後続リストを保持する。図1bに示されるように、第1の後続ノード211がクラッシュしてリングを去る場合、第1のノード200は引き続き、後続リストにリストされたその他の後続ノード223へのリンクを有している。後続リスト内の後続ノードの冗長性を用いて、ピアツーピア・オーバーレイ・ネットワークのトポロジ構造は保持することができる。第1のノード200に隣接する複数の次の後続ノードがクラッシュして、リングが閉じられうるように十分な後続ノードが後続リストにリストされない場合、トポロジ構造は保持することができなくなり、リングが破損する。したがって、リング・トポロジの安定性は、後続リストにリストされた後続ノードの数によって決まる、つまり後続ノードが多ければ、それに応じてリングもより安定する。

【発明の概要】

【発明が解決しようとする課題】

【0007】

本発明の目的は、ピアツーピア・オーバーレイ・ネットワークにおいてルーティングのために後続リストを提供する改良された方法を提供することである。さらに、本発明の目的は、この改良された方法を実行するために、対応するネットワーク・ノードおよび対応するコンピュータプログラム製品を提供することである。

【0008】

本発明の目的は、複数のノードを備えるピアツーピア・オーバーレイ・ネットワークにおいてルーティングのために後続リストを提供する方法であって、前記後続リストは、ピアツーピア・オーバーレイ・ネットワークのトポロジ構造に関して、複数のノードのうちの第1のノードに続いて後続となる少なくとも2つの後続ノードの識別子を指定し、後続リストにリストされた少なくとも2つの後続ノードが同一の物理ハードウェア上で稼働するかどうかを決定するステップと、少なくとも2つの後続ノードが同一の物理ハードウェア上で稼働する場合、ピアツーピア・オーバーレイ・ネットワークのトポロジ構造に関して少なくとも2つの後続ノードの後続となり、少なくとも2つの後続ノードとは異なる物理ハードウェア上で稼働する、複数のノードのうちの少なくとも1つの追加のノードの識別子を後続リストに追加するステップとを備える方法によって達成される。さらに、本発明の目的は、複数のノードを備えるピアツーピア・オーバーレイ・ネットワークの第1のノードであって、ピアツーピア・オーバーレイ・ネットワークにおいてルーティングするために後続リストを保持し、後続リストはピアツーピア・オーバーレイ・ネットワークのトポロジ構造に関して第1のノードに続いて後続となる少なくとも2つの後続ノードの識別子を指定し、第1のノードは後続リストにリストされた少なくとも2つの後続ノードが同一の物理ハードウェア上で稼働するかどうかを決定するように適合された制御ユニットを備え、少なくとも2つの後続ノードが同一の物理ハードウェア上で稼働する場合、ピアツーピア・オーバーレイ・ネットワークのトポロジ構造に関して少なくとも2つの後続ノ

ードの後続となり、少なくとも2つの後続ノードとは異なる物理ハードウェア上で稼働する、複数のノードのうちの少なくとも1つの追加のノードの識別子を後続リストに追加する第1のノードによって達成される。また、本発明の目的は、複数のノードを備えるピアツーピア・オーバーレイ・ネットワークにおいてルーティングのために後続リストを提供するコンピュータプログラム製品であって、後続リストは、ピアツーピア・オーバーレイ・ネットワークのトポロジ構造に関して、複数のノードのうちの第1のノードに続いて後続となる少なくとも2つの後続ノードの識別子を指定し、コンピュータプログラム製品は、第1のノードによって実行されるとき、後続リストにリストされた少なくとも2つの後続ノードが同一の物理ハードウェア上で稼働するかどうかを決定するステップと、少なくとも2つの後続ノードが同一の物理ハードウェア上で稼働する場合、ピアツーピア・オーバーレイ・ネットワークのトポロジ構造に関して少なくとも2つの後続ノードの後続となり、少なくとも2つの後続ノードとは異なる物理ハードウェア上で稼働する、複数のノードのうちの少なくとも1つの追加のノードの識別子を後続リストに追加するステップとを実行するコンピュータプログラム製品によって達成される。

10

【0009】

冗長後続リストが、従来技術によるシステムの場合のように、オーバーレイ固有のノードIDのみに基づく場合、すなわち、基礎となる物理ハードウェアを考慮しない場合、後続リストに従ってルーティングするノードの次に後続となる2つの後続ノードが、同一の物理マシン（たとえば、パーソナルコンピュータ）上に位置する可能性もある。この物理ハードウェアに障害が発生する場合、2つの後続ノードは同時にDHTを終了し、ノードによって保持される後続リストが障害の発生したハードウェア上で稼働しているピアへのリンクしか有していない場合、P2Pオーバーレイ・ネットワークのトポロジ構造は破損してしまう。本発明は、従来技術において遭遇するこの問題を解決する。

20

【課題を解決するための手段】**【0010】**

本発明は、オーバーレイ・トポロジの特性に基づくが、同時にP2Pオーバーレイ・ネットワークの基礎となる物理ハードウェアを尊重する後続リストを提供するためのアルゴリズムを導入する。本発明は、ネットワークのノードが稼働している物理ハードウェアに関する情報を、後続リストの提示および後続リストに基づくルーティングに加えるネットワーク・ノード（＝論理ピア）を説明する。

30

【0011】

（たとえば、XEN、VMwareのような、オペレーティングシステムの仮想化を通じて）同一の物理マシン上に位置するノードがこのことを認識していない従来技術の場合とは異なり、本発明は、ネットワーク・ノードの1つまたは複数稼働している物理ハードウェアを識別して、適切な冗長後続リストのために物理ハードウェアに関するこの情報を考慮する後続リストを提供するためのシステムを提示する。本発明によれば、後続リストが同一のハードウェアA上で稼働している後続ノードのみをリストする場合、後続リストもまた異なるハードウェアB上で稼働している1つまたは複数の後続ノードを備えるように、後続リストは修正される。

【0012】

40

本発明は、後続リストを提供する確実な方法をもたらす。後続リストは、ノードに障害が発生した場合にトポロジ構造の破損を防ぐための戦略であるため、本発明による後続リストは、潜在的または実際の後続ノードの物理ハードウェアを尊重して、後続ノードが同一の物理ノード上で稼働しないようにする。これは、従来技術のDHTアルゴリズムによっては考慮されない。特に、小規模な配置（最大100ノード）においては、ノードは（たとえば、VMware、Xenなどのような、仮想化環境において）同一の物理ノード上でホスティングされる。そのため、ノードの物理ハードウェアが無視される従来技術においては、複数のピアに障害が発生した場合にトポロジ構造が破損する可能性が高い。本発明は、ノードの物理ハードウェアを考慮する後続リストを提示することにより、このリスクを回避する。

50

【0013】

したがって、本発明は特に、隣接ノードが同一の物理マシン上でホスティングされる可能性が高まる小規模な配置において有用である。本発明は、物理ロケーションおよびオーバーレイ・ノードIDのマッピングを用いてトポロジ構造の破損を回避する。

【0014】

本発明は、仮想環境においてP2Pトポロジの冗長性を保持する。さらに、本発明は、後続リストが正しい長さを有するように自動的に最適化されるので、より高い効率をもたらす。

【0015】

後続リストがオーバーレイの特性に基づく現在のP2Pオーバーレイとは対照的に、本発明は、P2Pネットワークにおいてオーバーレイ・トポロジ作成および冗長後続リストの提示のための分散された物理ハードウェア・インフラストラクチャの上部で仮想化ソリューションを考慮するソリューションを提供する。

10

【0016】

さらに、本発明は、サーバ・クラスタまたは純正P2Pテレフォニーシステムのような、分散環境において有利に使用されうる基盤技術の改良について説明する。特にこの種の技術は、IETF P2P SIP (IETF=Internet Engineering Task Force、SIP=Session Initiation Protocol) に含まれる場合もある。

【0017】

さらなる利点は、従属請求の範囲により示される本発明の実施形態により達成される。

20

【0018】

本発明の好ましい実施形態によれば、少なくとも2つの後続ノードが同一の物理ハードウェア上で稼働する場合、異なる物理ハードウェア上で稼働する後続リストにリストされた後続ノードの合計数が事前定義された冗長因数と等しくなるように、ピアツーピア・オーバーレイ・ネットワークのトポロジ構造に関して少なくとも2つの後続ノードの後続となり、少なくとも2つの後続ノードとは異なる物理ハードウェア上で稼働する、複数のノードの同数の追加のノードの識別子が後続リストに追加される。

【0019】

好ましい実施形態によれば、結合ノードが後続リストにリストされた後続ノード間のピアツーピア・オーバーレイ・ネットワークを結合するとき、結合ノードの識別子は後続リストに追加される。結合ノードが稼働する物理ハードウェアが決定される。好ましくは、前記決定は、第1のノードによって実行される。異なる物理ハードウェアで稼働する後続リストにリストされた後続ノードの合計数が、事前定義された冗長因数と等しい場合、少なくとも1つの追加のノードの識別子が後続リストから削除されるが、合計数は少なくとも1つの追加のノードを除く、つまり前記合計数は少なくとも1つの追加のノードをカウントせずに決定される。

30

【0020】

第1のノードが、それ自身のハードウェアに関する情報、およびネットワークのその他のノードのハードウェアに関する情報を検出することが可能である。この検出の目的は、複数のノードのうちのいずれのノードが同一の物理ハードウェア上で稼働しているかを見出すことである。物理ハードウェアを検出する可能な方法はいくつかある。

40

【0021】

本発明の好ましい実施形態によれば、少なくとも2つの後続ノードが稼働する前記物理ハードウェアは、少なくとも2つの後続ノードのプラットフォームから情報を取り出すことによって決定される。好ましくは、前記取り出しは、第1のノードによって実行される。少なくとも2つの後続ノードが稼働する前記物理ハードウェアの決定に加えて、またはその代替として、少なくとも1つの追加のノードが稼働する物理ハードウェアは、少なくとも1つの追加のノードのプラットフォームから情報を取り出すことによって決定される。好ましくは、前記取り出しは、第1のノードによって実行される。少なくとも2つの後

50

続ノードおよび少なくとも1つの追加のノードがそれぞれ稼働する前記物理ハードウェアの決定に加えて、またはその代替として、結合ノードが稼働するハードウェアは、結合ノードのプラットフォームから情報を取り出すことによって決定される。好ましくは、前記取り出しは、第1のノードによって実行される。プラットフォームは、少なくとも2つの後続ノードおよび／または少なくとも1つの追加のノードおよび／または結合ノードが稼働する、たとえば、CPU(=Central Processing Unit(中央演算処理装置))、コンピュータ、ネットワークユニットなどの、物理ハードウェアであってもよい。情報は、少なくとも2つの後続ノードおよび／または少なくとも1つの追加のノードおよび／または結合ノードのプラットフォームに関連付けられている、たとえばCPU ID、MACアドレス、ロケーション、IPアドレス、ロケーションなどであってもよい(MAC=Media Access Control(メディアアクセス制御))。これは、ソフトウェア・アプリケーションが、ネットワーク・ノードのオペレーティングシステムの助けを借りて見出すことができる任意の方法を含む。たとえば、MACアドレスは通常、物理ホストを決定する比較的確実な方法であるが、これはハードウェア仮想化技術が使用される場合には当てはまらない場合もある。

10

【0022】

本発明は、Chordピアによって実行されるレプリケーション・メソッドを変更するので、物理ノードに障害が発生した場合にデータの可用性を向上させることができる。本発明は、各ピアノードを、たとえばCPU ID、MACアドレスなど、ピアノードをホスティングするハードウェア・プラットフォームを識別するハードウェア識別子に関連付ける。ピアノードは、ハードウェア識別子を使用して、たとえばピアIDに関してなど、トポロジ構造に関して好ましくは最も近接する後続である、異なるハードウェア・プラットフォーム上で稼働している少なくとも1つのピアノードに責任を負うリソースのセットを複製する。

20

【0023】

本発明のもう1つの好ましい実施形態によれば、少なくとも2つの後続ノードが稼働する前記物理ハードウェアは、少なくとも2つの後続ノードにリンクされるか、またはこれらに関連付けられている外部管理ユニットから情報を取り出すことによって決定される。好ましくは、前記取り出しは、第1のノードによって実行される。少なくとも2つの後続ノードが稼働する前記物理ハードウェアの決定に加えて、またはその代替として、少なくとも1つの追加のノードが稼働する物理ハードウェアは、少なくとも1つの追加のノードにリンクされるか、またはこれに関連付けられている外部管理ユニットから情報を取り出すことによって決定される。好ましくは、前記取り出しは、第1のノードによって実行される。少なくとも2つの後続ノードおよび少なくとも1つの追加のノードがそれぞれ稼働する前記物理ハードウェアの決定に加えて、またはその代替として、結合ノードが稼働するハードウェアは、結合ノードにリンクされるか、またはこれに関連付けられている外部管理ユニットから情報を取り出すことによって決定される。好ましくは、前記取り出しは、第1のノードによって実行される。外部管理ユニットはまた、外部クラスタ管理エンティティであってもよく、「クラスタ」という用語は、ネットワーク・ノードのクラスタとしてのP2Pオーバーレイ・ネットワークを指す。

30

40

【0024】

たとえばノードIDを構成するために、外部クラスタ管理エンティティが使用可能な場合、外部クラスタ管理エンティティはまた、クラスタの物理トポロジに関する情報を提供することもできる。外部クラスタ管理エンティティは、中央登録サーバ、ブート・ストラップ・サーバ、またはネットワーク管理および監視サーバであってもよい。

【0025】

ソフトウェア・アプリケーションが必要なハードウェア情報を検出することができない場合、ノードのマネージャによるノードの手動(事前)構成が解決策となる。好ましくは、少なくとも2つの後続ノードが稼働する前記物理ハードウェアは、ピアツーピアネットワークのネットワークオペレータから、および／または少なくとも2つの後続ノードのユ

50

ーザから対応する情報を取り出すことによって決定される。好ましくは、前記取り出しは、第1のノードによって実行される。少なくとも2つの後続ノードが稼働する前記物理ハードウェアの決定に加えて、またはその代替として、少なくとも1つの追加のノードが稼働する物理ハードウェアは、ピアツーピアネットワークのネットワークオペレータから、および／または少なくとも1つの追加のノードのユーザから対応する情報を取り出すことによって決定される。好ましくは、前記取り出しは、第1のノードによって実行される。少なくとも2つの後続ノードおよび少なくとも1つの追加のノードがそれぞれ稼働する前記物理ハードウェアの決定に加えて、またはその代替として、結合ノードが稼働するハードウェアは、ピアツーピアネットワークのネットワークオペレータから、および／または結合ノードのユーザから情報を取り出すことによって決定される。

10

【0026】

好ましい実施形態において、少なくとも2つの後続ノードが稼働する前記物理ハードウェアは、少なくとも2つの後続ノードに関連付けられているアプリケーションレイヤの測定を用いて決定される。好ましくは、前記決定は、第1のノードによって実行される。少なくとも2つの後続ノードが稼働する前記物理ハードウェアの決定に加えて、またはその代替として、少なくとも1つの追加のノードが稼働する物理ハードウェアは、少なくとも1つの追加のノードに関連付けられているアプリケーションレイヤの測定を用いて決定される。好ましくは、前記決定は、第1のノードによって実行される。少なくとも2つの後続ノードおよび少なくとも1つの追加のノードがそれぞれ稼働する前記物理ハードウェアの決定に加えて、またはその代替として、結合ノードが稼働するハードウェアは、結合ノードに関連付けられているアプリケーションレイヤの測定を用いて決定される。好ましくは、前記決定は、第1のノードによって実行される。

20

【0027】

アプリケーションレイヤの前記測定は、「ping」メッセージの交換のような、アプリケーションレベルのプロブ機構であってもよい。前記pingメッセージの交換に基づいて、ノードは、その他のノードが同一のデバイス上でホスティングされる場合、接続待ち時間から推定することができる。また、ノードが同一のハードウェア上にある場合、その他の既知の方法を使用して推定することもできる。たとえば、ネットワーク調整システムは、すべてのリンクの明示的な測定を必要とすることなく、一部のノード間待ち時間を測定することによって、基礎となるネットワークに関する知識を得る。これにより、ノードが物理トポロジにおいて近接する場合、ノードを論理オーバーレイ内で互いに近づけることができる。もう1つの例は、その他のノードの待ち時間を測定することにより、位置についてそのルーティング・テーブルを継続的に調節するPASTRYである。

30

【0028】

好ましくは、制御ユニットは、ルーティング・ポリシーに基づいて、第1のノードから後続リストのノードにピアツーピア・メッセージをルーティングするように適合される。後続リストの前記ノードが、たとえば後続リストの前記ノードが稼働している／稼働していた物理ハードウェアの障害により、ピアツーピア・オーバーレイ・ネットワークで到達可能ではない場合、制御ユニットは、ルーティング・ポリシーに基づいて第1のノードから少なくとも1つの追加のノードにピアツーピア・メッセージをルーティングするように適合される。

40

【0029】

好ましくは、ピアツーピア・メッセージは、ルーティング・ポリシーに基づいて、第1のノードから少なくとも2つの後続ノードのうちの1つまたは複数のノードにルーティングされる。少なくとも2つの後続ノードのうちの前記1つまたは複数のノードが、たとえば少なくとも2つの後続ノードが稼働している／稼働していた物理ハードウェアの障害により、ピアツーピア・オーバーレイ・ネットワークで到達可能ではない場合、ピアツーピア・メッセージは、ルーティング・ポリシーに基づいて第1のノードから少なくとも1つの追加のノードにルーティングされる。

【0030】

50

ルーティング・ポリシーは、後続リストのエントリを使用する順序を指定する。後続リストがノードの識別子のシーケンスを備えることは可能である。次いで、ルーティング・ポリシーは、第1のノードが最初に、識別子が後続リストの第1の識別子である後続リストの後続ノードにP2Pメッセージをルーティングしようと試みるよう規定することができる。これが成功しない場合、つまりP2Pメッセージがその宛先に到達しない場合、第1のノードは、識別子が後続リストの第2の識別子である少なくとも1つの追加のノードにP2Pメッセージをルーティングしようと試みる、というように続行する。

【0031】

本発明によれば、2つの異なる物理ハードウェア・ユニットに同時に障害が発生する可能性は低いので、少なくとも1つの追加のノードおよび少なくとも2つの後続ノードが異なる物理ハードウェア上で稼働すれば、少なくとも2つの後続ノードが到達可能ではないときに、少なくとも1つの追加のノードは到達可能になる。しかし、さらに高度な冗長性が必要とされる場合、異なる物理ハードウェア上で稼働する後続リストにリストされた後続ノードの合計数が事前定義された冗長因数と等しくなるように、ピアツーピア・オーバーレイ・ネットワークのトポロジ構造に関して少なくとも2つの後続ノードの後続となり、少なくとも2つの後続ノードとは異なる物理ハードウェア上で稼働する、複数のノードの同数の追加のノードの識別子が、後続リストに追加される。

【0032】

好ましくは、本発明は、同一の物理ハードウェア上で稼働している2つまたはそれ以上の次の後続ノードがオーバーレイを終了する場合に、P2Pオーバーレイを保持するために動的冗長後続リスト・アルゴリズムを導入する。動的冗長後続リスト・アルゴリズムは、RSL内の1つまたは複数の後続ノードが、たとえば仮想化ソフトウェアによって、同一ハードウェア上で稼働している場合、後続ノードの識別子が冗長後続リストに追加されるよう規定する。各ピアが各自のハードウェア上で稼働する純正P2P環境の場合と同様にChordの冗長因数を仮想環境において保持するため、ピアは、純正P2P環境の場合と同数の冗長後続ノードを異なるハードウェアに対して保持する必要がある。したがって、同一の仮想プラットフォーム上で稼働しているピアの数は重要ではない。それに加えて、動的冗長後続リスト・アルゴリズムは、RSLにリストされたノードとは異なるハードウェア上で稼働している新しいピアが、P2Pネットワークを結合して、冗長後続リストに追加されている場合、RSLから削除されるよう規定する。

【0033】

本発明の前述およびさらなる特徴および利点は、添付の図面と併せて現在好ましい例示の実施形態の以下の詳細な説明を読むことにより深く理解されるであろう。

【図面の簡単な説明】

【0034】

【図1a】複数のノードを備える従来技術の構造化P2Pオーバーレイ・ネットワークを示す図である。

【図1b】従来技術の、第1の後続ノードがクラッシュしてリングを去る場合を示す図である。

【図2】本発明の1つの実施形態によるP2Pオーバーレイ・ネットワークを示す図である。

【図3a】図2に示されるP2Pオーバーレイ・ネットワークにおける物理ハードウェアの障害を示す図である。

【図3b】図2に示されるP2Pオーバーレイ・ネットワークにおける物理ハードウェアの障害を示す図である。

【図4a】本発明の1つの実施形態による図2に示されるP2Pオーバーレイ・ネットワークにおける追加アルゴリズムを示す図である。

【図4b】本発明の1つの実施形態による図2に示されるP2Pオーバーレイ・ネットワークにおける追加アルゴリズムを示す図である。

【図5a】本発明の1つの実施形態による図2に示されるP2Pオーバーレイ・ネットワ

ークにおける削除アルゴリズムを示す図である。

【図5b】本発明の1つの実施形態による図2に示されるP2Pオーバーレイ・ネットワークにおける削除アルゴリズムを示す図である。

【発明を実施するための形態】

【0035】

図2は、本発明の1つの実施形態によるP2Pオーバーレイ・ネットワークを示す。特に、図2は、本発明の1つの実施形態によるChordリングNを表す。ChordリングNは、参照符号0、20、50、87、112、140、179、200、211、223、240が、ノード0、20、50、87、112、140、179、200、211、223、240のノードIDを表すことも意図される複数のノード0、20、50、87、112、140、179、200、211、223、240を備える。ノードのノードIDは、オーバーレイ固有のDHTアルゴリズムによって割り当てられ、たとえば登録サーバに基づくか、またはノード自身によってランダムに選択されてもよい。各ノードは、ChordリングNのピアを表す。

10

【0036】

方法の第1のステップにおいて、第1のノード200は、第1のノード200によって保持される後続リストに含まれる後続ノードが稼働している物理ハードウェアを決定する。第1のノード200が、2つのエントリを持つ後続リストLを保持し、1つのエントリは、ノード211の、たとえばノードIDのような識別子であり、もう一方のエントリは、ノード223の、たとえばノードIDのような識別子であると仮定する。

20

【0037】

第1のノード200は、1つまたは複数の連結されたコンピュータ、つまり、ハードウェア・プラットフォーム、ハードウェア・プラットフォームに基づくソフトウェアプラットフォーム、ならびにソフトウェアおよびハードウェア・プラットフォームによって形成されるシステム・プラットフォームによって実行される複数のアプリケーション・プログラムから成る。第1のノード200の機能は、それらのアプリケーション・プログラムの実行により提供される。アプリケーション・プログラムまたはそれらのアプリケーション・プログラムの選択された部分は、システム・プラットフォーム上で実行されるとき、以下に説明するようなルーティングサービスを提供するコンピュータ・ソフトウェア製品を構成する。さらに、そのようなコンピュータ・ソフトウェア製品は、それらのアプリケーション・プログラムまたはアプリケーション・プログラムの前記選択された部分を格納する記憶媒体によって構成される。

30

【0038】

機能の観点からは、第1のノード200は、第1のノードの機能の制御のための制御ユニット、データを格納するための記憶媒体、および他のノードとの情報の交換のためのインターフェイスを備える。

【0039】

物理ハードウェアを決定するために、第1のノード200は、後続ノード211、223が稼働している物理ハードウェアの識別に役立つ任意の使用可能な情報を使用することができる。第1のノード200が、たとえばCPU ID、MACアドレス、ロケーション、IPアドレスなど、後続ノード211、223のプラットフォームから情報を収集することにより、後続ノード211、223の物理ハードウェアに関する情報を取り出すことが可能である。次いで、第1のノード200は、ChordリングNのその他のノード0、20、50、87、112、140、179、240から対応する情報を収集して、収集された情報の断片を比較することができる。

40

【0040】

後続ノード211、223のプラットフォームから使用可能な情報（たとえば、CPU ID、MACアドレス、ロケーション、IP、・・・）に加えて、第1のノード211は、アプリケーションレイヤの測定を用いて収集される、後続ノード211、223が稼働しているハードウェアのユーザによって実行された手動構成を用いて提供される、外部

50

管理エンティティから取り出される、などした情報を使用することができる。同様に、その他の各ノード0、20、50、87、112、140、179、211、223、240は、それぞれが稼働している物理ハードウェアの識別に役立つ任意の使用可能な情報を使用することができる。その他の各ノード0、20、50、87、112、140、179、211、223、240が、それぞれの物理ハードウェアに関して収集された情報を第1のノード200に伝達することが可能である。また、第1のノード200が、その他のノード0、20、50、87、112、140、179、211、223、240が稼働している物理ハードウェアにアクセスして、その他のノード0、20、50、87、112、140、179、211、223、240からそれぞれの物理ハードウェアに関する情報を収集することも可能である。

10

【0041】

物理ハードウェアを分類するこのステップの結果は、ハードウェアの分類を示し、それをノード0、20、50、87、112、140、179、200、211、223、240の異なるアイコン形状によって表す図2に示される。各形状は、別々の物理マシンを表す。それはつまり、ノード0、50、140、211、および223が第1のマシン上で稼働するということである。ノード20、112、179、および240は、第2のマシン上で稼働する。ノード87および200は、第3のマシン上で稼働する。

【0042】

従来技術において遭遇し、本発明により解決される問題は、図3aおよび図3bを用いて示される。(たとえば、XEN、VMwareのような、オペレーティングシステムの仮想化を通じて)図3aにおいて円で指示されるように、ノード211およびノード223はいずれも同一の物理ハードウェア上で稼働する。図3bは、後続リストLに含まれる2つの後続ノードが稼働する物理ハードウェアの障害を示す。ノード211および223は、同一のマシン上で稼働している。図3bにおいて、このマシンはクラッシュして、ノード211および223はリングNを去る。その結果、ノード200は、リング・トポロジを保持するためのアクティブなピアへのリンクから外れる。この場合、第1のノード200に関連付けられているRSLで保持された後続ノード211、223のいずれもハードウェア障害のために到達可能ではないため、第1のノード200は後続ノードへの接続を失う。

20

【0043】

図4aは、本発明による動的冗長後続リスト適合のための方法を示す。ノード0、20、50、87、112、140、179、200、211、223、240は、それぞれが稼働している物理ハードウェアを識別する。第1のノード200は、同一の物理ハードウェア上でホスティングされる後続の隣接ピア211、223がクラッシュした場合、P2Pオーバーレイの安定性を維持するために動的冗長後続リスト・アルゴリズムを導入する。それにより、このアルゴリズムは、追加アルゴリズムおよび削除アルゴリズムという2つのサブアルゴリズムを備える。

30

【0044】

追加アルゴリズムは、RSLにリストされた複数の後続隣接ピアが同一の物理マシン上で稼働している場合に、ピアをRSLに追加する。

40

【0045】

図4aに示されるように、第1のノード200は、第1のノード200の後続リストLに追加のノード240を追加して、後続リストLを拡張する。これを追加する理由は、その識別子だけがこれまで第1のノード200の後続リストLのエントリであった2つの後続ノード211、223が、同一の物理マシン上で稼働しているからである。したがって、2つの後続ノード211、223が稼働するマシンがクラッシュする場合、第1のノード200には引き続き第1のノード200の後続リストLにノード240が残っており、第1のノード200がネットワークNのリング・トポロジを閉じることができる。

【0046】

各ピアが各自のハードウェア上で稼働する通常のChord環境の場合と同じ冗長因数

50

を得るため、ピアは、冗長因数が指定するピアの数と同数の、異なるハードウェア上で稼働するピアを後続リストに追加する必要がある。

【0047】

図4bは、ChordリングNが、複数のノード0、20、50、87、112、140、179、200、211、219、223、240、および249を備える状況を示す。本発明の方法により、3の冗長因数が実現される事例を仮定する。3のChord冗長因数では、ノード211および223が同一の物理マシン上で稼働しており、ノード219および240も同一の物理マシン上で稼働しているので、5つのノード211、219、223、240、および249が後続リストLに含まれることが必要となる。

【0048】

以前追加された追加のノードが必要ではなくなる場合、削除アルゴリズムがRSLからピアを削除する。これは、ノードが、RSLによってカバーされるIDスペースに加わるか、または去る場合である。

【0049】

図5aは、第1のノード200が、2つの元の後続ノード211および223、ならびに後に追加された追加のノード240を含む後続リストLを保持する状況を示す。2つの後続ノード211および223が稼働する物理ハードウェアは、同一である。追加のノード240が稼働する物理ハードウェアは、2つの元の後続ノード211および223が稼働する物理ハードウェア上とは異なる。

【0050】

結合ノード220は、後続リストLにリストされた後続ノード211と223の間のピアツーピア・オーバーレイ・ネットワークNを結合する。たとえばノードID220のような、結合ノード220の識別子は、第1のノード200の後続リストに追加される。好ましくは、第1のノード220は、その後続リストLへの修正を実行する。

【0051】

第1のノード220は、結合ノード220が稼働する物理ハードウェアを決定する。第1のノード220は、結合ノード220が稼働する物理ハードウェアが、2つの元の後続ノード211および223が稼働する物理ハードウェアとは異なり、追加のノード240が稼働する物理ハードウェアとは異なることを検出する。

【0052】

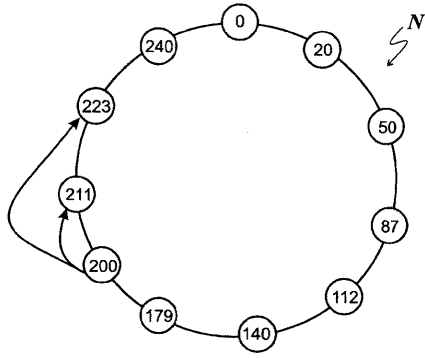
ルーティング・ポリシーが、2の冗長因数が十分であることを指定するので、第1のノード200は、追加のノード240の識別子を後続リストLから削除する。次いで、識別子が後続リストLから削除されている少なくとも1つの追加のノード240をカウントすることなく、異なる物理ハードウェア上で稼働する後続リストLにリストされた後続ノード211、220の合計数は、事前定義された2の冗長因数と等しくなる。

10

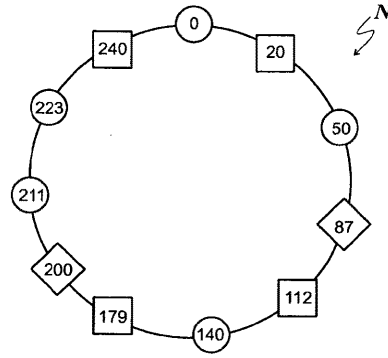
20

30

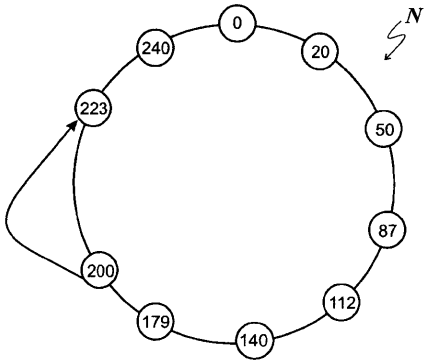
【図 1 a】



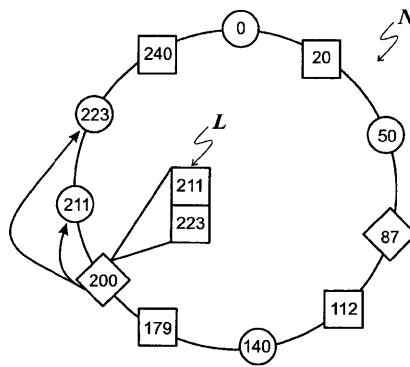
【図 2】



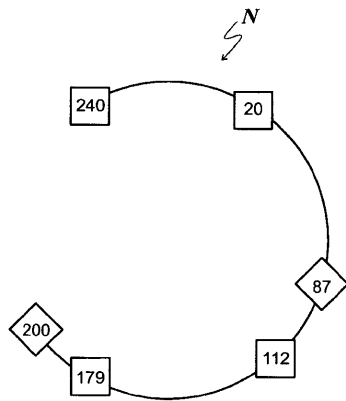
【図 1 b】



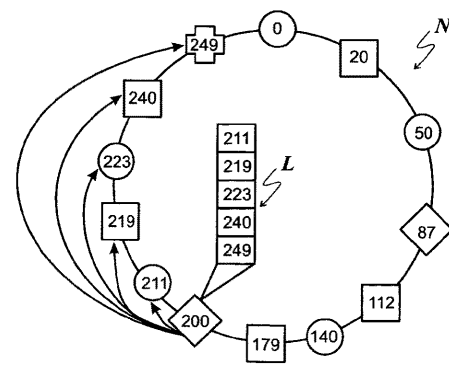
【図 3 a】



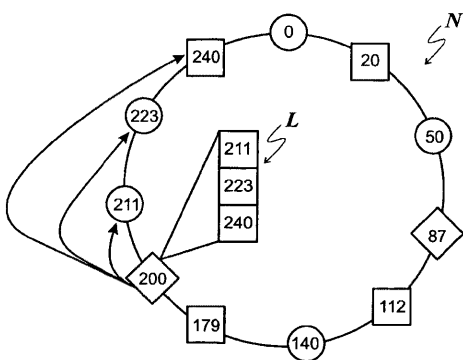
【図 3 b】



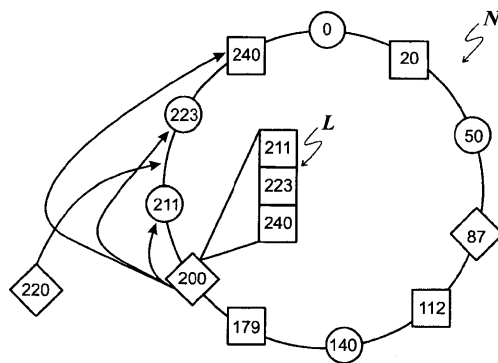
【図 4 b】



【図 4 a】



【図 5 a】



フロントページの続き

(72)発明者 シュワン, ニコ

ドイツ 70190 シュトゥットガルト, シューバルトシュトラッセ 9

(72)発明者 ストラウス, トーマス

ドイツ 73734 エスリンゲン, ジュラシュトラッセ 61

(72)発明者 トムス, マルコ

ドイツ 71254 ディッツィンゲン, アーダルベルトーシュティフターーシュトラッセ 12

審査官 衣嶋 文彦

(56)参考文献 Waagan, Kristian, Building a Replicated Data Store using Berkeley DB and the Chord DHT, Student thesis, 2005年, URL, <http://ntnu.diva-portal.org/smash/record.jsf?pid=diva2:348050&rvn=2>

吉永 浩和 他, Peer-to-Peerネットワークにおける分散協調ストレージ技術の提案, 電子情報通信学会技術研究報告, 2005年 6月16日, 第105巻, 第127号, p.81~84

(58)調査した分野(Int.Cl., DB名)

H04L 12/70