(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5998206号
(P5998206)
(45)発行日　平成28年9月28日(2016.9.28)　(24)登録日　平成28年9月2日(2016.9.2)

(51)Int.Cl.　F I
G06F 12/00 (2006.01)　G06F 12/00 545Z

請求項の数 9　(全 17 頁)

(21)出願番号	特願2014-511463 (P2014-511463)
(86)(22)出願日	平成24年5月15日 (2012.5.15)
(65)公表番号	特表2014-513852 (P2014-513852A)
(43)公表日	平成26年6月5日 (2014.6.5)
(86)国際出願番号	PCT/US2012/037997
(87)国際公開番号	WO2012/158718
(87)国際公開日	平成24年11月22日 (2012.11.22)
審査請求日	平成27年4月24日 (2015.4.24)
(31)優先権主張番号	61/486,701
(32)優先日	平成23年5月16日 (2011.5.16)
(33)優先権主張国	米国 (US)
(31)優先権主張番号	13/239,253
(32)優先日	平成23年9月21日 (2011.9.21)
(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
アメリカ合衆国カリフォルニア州９４０６５レッドウッド・シティー，オラクル・パークウェイ５００
(74)代理人　110001195
特許業務法人深見特許事務所
(72)発明者　リー，ロバート・エイチ
アメリカ合衆国、９４０７０　カリフォルニア州、サン・カルロス、サンセット・ドライブ、９２８

最終頁に続く

(54)【発明の名称】クラスタデータグリッドにおける拡張可能な中央集中型動的リソース分散

(57)【特許請求の範囲】
【請求項１】
拡張可能な集中型の動的なリソース分散を提供するために１つ以上のコンピュータノードで実現される方法であって、前記方法は、
コンピュータノードのクラスタに亘って分散しているデータセットの複数のパーティションを記憶することと、
前記クラスタのグローバル状態を示す情報を収集し、前記グローバル状態へのアクセスを提供する分散コーディネータとなる前記コンピュータノードの１つを指定することとを含み、前記グローバル状態は、以下のうち少なくとも１つを含み、
（ａ）クラスタ内の各コンピュータノードに対して前記パーティションのうちどれが割り当てられるかを示す情報、
（ｂ）クラスタ内の各コンピュータノードに対する現在の処理負荷を示す情報（前記現在の処理負荷は、前記分散コーディネータにランタイムフィードバックの統計を定期的に送信する各コンピュータノードによって決定される）、
（ｃ）クラスタ内の各コンピュータノードのメモリ容量とプロセッサ容量とを示す情報、
定期的に、分散コーディネータによるクラスタのグローバル状態を分析して、前記コンピュータノードの間での前記パーティションの分散に対して変更を加えるべきであるかどうかを判断することと、
前記パーティションの分散に対して変更を加えるべきとの判断に応答して、前記分散コ

ーディネータによって、新たな分散プランを生成し、コンピュータノードのクラスタ中のすべてのコンピュータノードに、前記新たな分散プランへのアクセスを提供することを含み、前記新たな分散プランは、特定のパーティションが前記クラスタ内の指定されたコンピュータノード上に位置することを特定し、

前記新たな分散プランを実行するために前記コンピュータノードに関連付けられた個々のパーティション転送をどのように実行するかをクラスタ内の各コンピュータノードによって独立して判断するための分散アルゴリズムを使用することとを含む、方法。

【請求項２】

前記パーティションの分散に対して変更を加えるべきであるかどうかを判断するために、定期的に、分散コーディネータによるクラスタのグローバル状態を分析することは、

前記分散コーディネータによってプラガブルな分散ロジックモジュールを呼び出すことを含み、前記プラガブルな分散ロジックモジュールは、コンピュータノードのクラスタの間で前記パーティションを分散するために利用されるアルゴリズムを調節するためにランタイムで切り替えられ得る、請求項１に記載の方法。

【請求項３】

分散アルゴリズムを使用することは、さらに、

送信ノードと受信ノードとの間の直接的な非同期通信の結果として、コンピュータノードのクラスタにおける２つのコンピュータノード間で二地点間での個々のパーティションの転送を実行することを含み、前記分散コーディネータは前記パーティションの転送の命令に関与しない、請求項１または２に記載の方法。

【請求項４】

前記複数のパーティションは、さらに、基本パーティションの組とバックアップパーティションの組とを有し、分散コーディネータは、各基本パーティションが当該基本パーティションに関連付けられたバックアップパーティションとは異なる物理的ノード上に位置されることを確実にする、請求項１～３のいずれか１項に記載の方法。

【請求項５】

分散コーディネータによって生成された新たな分散プランは、所与の各コンピュータノード毎に限られた数のコンピュータノード、前記所与のコンピュータノード上に位置する基本パーティションに関連付けられたバックアップパーティションを含むことを許される制限を課す、請求項４に記載の方法。

【請求項６】

前記分散コーディネータは、前記クラスタ内のコンピュータノードの間での前記パーティションの分散のための単一の調節点を提供する、請求項１～５のいずれか１項に記載の方法。

【請求項７】

請求項１～６のいずれか１項に記載の方法のすべてのステップを実行するように構成された１つ以上のコンピュータノードを含む装置。

【請求項８】

拡張可能な集中型の動的なリソース分散を提供するためのシステムであって、前記システムは、

データセットの複数のパーティションを記憶しているコンピュータノードのクラスタを含み、前記パーティションは、コンピュータノードの前記クラスタに亘って分散しており、さらに、

コンピュータノードから選択された分散コーディネータを含み、前記分散コーディネータは、前記クラスタのグローバル状態を示す情報を収集し、定期的にグローバル状態を分析して、前記コンピュータノードの間での前記パーティションの分散に対して変更を加えるべきであるかどうかを判断し、前記パーティションの分散に対して変更を加えるべきとの判断に応答して、新たな分散プランを生成し、コンピュータノードのクラスタ内のすべてのコンピュータノードに前記新たな分散プランへのアクセスを提供し、

コンピュータノードの前記クラスタは、前記分散コーディネータによって生成された前

記新たな分散プランを実行するために前記コンピュータノードに関連付けられた個々のパーティション転送をどのように実行するかをクラスタ内の各コンピュータノードによって独立して判断するための分散アルゴリズムを使用し、
前記グローバル状態は、以下のうち少なくとも１つを含み、
（ａ）クラスタ内の各コンピュータノードに対して前記パーティションのうちどれが割り当てられるかを示す情報、
（ｂ）クラスタ内の各コンピュータノードに対する現在の処理負荷を示す情報（前記現在の処理負荷は、前記分散コーディネータにランタイムフィードバックの統計を定期的に送信する各コンピュータノードによって決定される）、
（ｃ）クラスタ内の各コンピュータノードのメモリ容量とプロセッサ容量とを示す情報、
前記新たな分散プランは、特定のパーティションが前記クラスタ内の指定されたコンピュータノード上に位置することを特定する、システム。
【請求項９】
拡張可能な集中型の動的なリソース分散を提供するための方法であって、前記方法は、
コンピュータノードのクラスタに亘って分散しているデータセットの複数のパーティションを記憶することと、
定期的に、分散コーディネータによるクラスタのグローバル状態を分析して、前記コンピュータノードの間での前記パーティションの分散に対して変更を加えるべきであるかどうかを判断することとを含み、前記分散コーディネータは前記コンピュータノードの中の１つであり、前記グローバル状態は、以下のうち少なくとも１つを含み、
（ａ）クラスタ内の各コンピュータノードに対して前記パーティションのうちどれが割り当てられるかを示す情報、
（ｂ）クラスタ内の各コンピュータノードに対する現在の処理負荷を示す情報（前記現在の処理負荷は、前記分散コーディネータにランタイムフィードバックの統計を定期的に送信する各コンピュータノードによって決定される）、
（ｃ）クラスタ内の各コンピュータノードのメモリ容量とプロセッサ容量とを示す情報、
前記パーティションの分散に対して変更を加えるべきとの判断に応答して、前記分散コーディネータによって、新たな分散プランを生成し、コンピュータノードのクラスタの全体に、前記新たな分散プランへのアクセスを提供することを含み、前記新たな分散プランは、特定のパーティションが前記クラスタ内の指定されたコンピュータノード上に位置することを特定し、
前記新たな分散プランを実行するために前記コンピュータノードに関連付けられた個々のパーティション転送をどのように実行するかをクラスタ内の各コンピュータノードによって独立して判断するための分散アルゴリズムを使用することとを含む、方法。
【発明の詳細な説明】
【技術分野】
【０００１】
著作権表示
この特許文献の開示の一部分は、著作権保護の対象である資料を含む。著作権者は、特許商標庁の特許ファイルまたは記録に見られるような特許文献または特許の開示の任意の者による完全な複製に対しては異議がないが、それ以外の件についてはそれが何であっても全ての著作権を留保する。
【０００２】
発明の分野
本発明は、分散コンピューティング環境におけるデータキャッシング技術に関し、特に、クラスタネットワークのメンバ間でのデータのパーティショニングに関する。
【背景技術】
【０００３】

背景

分散コンピューティングおよび分散アルゴリズムは、性能と負荷容量の増加、高い利用可能性および障害迂回、ならびに、データへのより迅速なアクセスという理由から、多様な背景において、普及してきている。分散コンピューティングは、典型的には、アプリケーションの実行のようなタスクを解決し、複雑な計算問題を解き、または、ユーザに様々なサービスへのアクセスを提供するための、互いに通信する、多くの自律型コンピュータ（ノードとも呼ばれる）を含む。コンピュータノードの各々は、典型的にはその独自のプロセッサと、メモリと、他のノードへの通信リンクとを含む。コンピュータは、特定の場所（たとえば、クラスタネットワーク）内に位置してもよく、または、インターネット等の広域ネットワーク（ＬＡＮ）を介して接続してもよい。たいていの場合、分散コンピュータは、互いに通信するために、かつ、タスク処理およびデータマネジメントを連係させるために、メッセージを利用する。

【０００４】

データマネジメントは、分散コンピューティングにおける重要な課題である。クラスタネットワークの背景において、大きなデータセットが、クラスタの様々なノード間でパーティショニングされ得る。各ノードは、通常、多くのそのようなパーティション（データセット全体の部分）を格納し、かつ、当該パーティションに対してトランザクションを実行する。多くの場合、パーティションは、障害迂回の目的で、メンバ間に分散された、基本データおよびデータのバックアップコピーを含む。このように区切られた態様でのデータの分散は、管理容易性、性能、および、情報の利用可能性を向上することができる。

【発明の概要】

【発明が解決しようとする課題】

【０００５】

この背景には、理想的なデータの分散に影響する、多くの制約や懸念事項が存在する。たとえば、あるサーバから別のサーバへのデータの移動は、時間および／またはプロセッサ容量の消費を要する。高い利用可能性の理由から、基本データおよびデータのバックアップコピーを、しばしば、物理的に異なる装置に位置させておくことが好ましい。さらに、性能、拡張性、および、容量の理由から、しばしば、利用可能なストレージサーバ間で多少は均等にデータの分散のバランスをとり、また、ノードがクラスタに追加されるまたはクラスタから削除されたときには分散を調節することが好ましい。

【０００６】

実際の使用事例において、一層のアプリケーションに特有の選択が望まれる場合もあり得る。たとえば、ある状況下では、特定のデータセットを特定のサーバ上に位置させるように特定することが有用である場合がある。さらには、分散が、ノードメンバ間でデータを配列するためにランタイムフィードバックおよび入力を利用するように特定することが望ましいことがある。上記に照らして、必要とされていることは、データ分散における懸念事項のすべてまたは多く、かつ、分散されたクラスタのメンバ間でのパーティションのバランス調整を最適化するための、単純かつ効率的なやり方である。

【０００７】

簡単な概要

発明の様々な実施の形態に従うと、パーティショニングを実行する実際のメカニクスは分散アルゴリズムのままでありながら、クラスタノード間におけるデータのパーティショニングの決定部は集中化される。中央の分散コーディネータは、分散プランを生成する、中央集中型のロジックまたはアルゴリズム（ストラテジ）を実行し得る。分散プランは、データがクラスタの全てに亘ってどのようにパーティショニングされるかを制御する。この分散プランを実行するための作業は、すべてのメンバによって、個別にかつ非同期的に、分散アルゴリズムに従って、実行される。分散プランは、クラスタのすべてのメンバに通信することが可能で、そして、各メンバは、当該各メンバがそれ自体のみに関連付けるように、パーティショニングを実行する。たとえば、分散アルゴリズムに従って、あるノードは分散プランに照らして特定のパーティションを取得して、そのデータを取得するの

に必要なステップを実行する必要があると判断してもよい一方で、他のノードはそれらの特定のノードに関連する他の個々のパーティションの転送を非同期で実行していてもよい。このように、データ分散のための中央集中型の管理および制御を同時に実現可能にしながら、単一の連係点によって引き起こされるボトルネックを回避できる。

【図面の簡単な説明】

【０００８】

【図１】発明の様々な実施形態に従った、クラスタ内に実装されている拡張可能な集中型のリソース分散の説明図である。

【図２】発明の様々な実施形態に従ってクラスタ内のノードによって実施される分散プランの説明図である。

【図３】発明の様々な実施形態に従って集中型のリソース分散を提供するためのプロセスのフローチャートである。

【図４】発明の様々な実施形態に従って分散コーディネータによって実行されるプロセスのフローチャートである。

【図５】発明の様々な実施形態に従ってクラスタ内のノードによって実行されるプロセスのフローチャートである。

【発明を実施するための形態】

【０００９】

詳細な説明

クラスタ化されたデータ・グリッド内の機能の一つは、様々なクラスタ要素（ノード）間のデータストレージパーティションの分散を維持し、達成することである。この背景において、分散は、データストレージを提供する所与のクラスタ要素サーバ上の基本データおよびバックアップコピーのデータの両者の配列と考えることができる。

【００１０】

クラスタ全体に亘るデータパーティショニングを実装するための一つのやり方は、あらゆるストラテジサーバが所有するパーティションを別のサーバへ移動させするか否か、または、別のサーバからパーティションを取得するか否かを、当該ストレージサーバが独室して判断する、分散アルゴリズム（distributed algorithm）を実施することによってである。一例として、クラスタのデータセットは、まず、均等に２つのクラスタメンバ（ノード）に分散している、（対応する２５６のバックアップパーティションを有する）２５６の基本パーティションを含むことができる。なお、各メンバが１２８の基本パーティションと１２８のバックアップパーティションとを格納する。第３のエレメントがクラスタに参加する場合、メンバは、各ノードが８５のパーティションを格納するノードの１つが８６格納するように、互いの間でパーティションの再分散を行なうことができる。たとえば、新たに参加したメンバノードは、クラスタ内の他の２つのノードの各々が４３のパーティションを当該新たなノードに転送するように、要求し得る。これに加えて、当該他の２つのノードは、新しいメンバがクラスタに参加してきたことを判断して、独立して当該新しいノードにパーティションを転送してもよい。

【００１１】

パーティションを割り当てるための分散アルゴリズムを使用する、この自律かつ非連結のアプローチは、より良いスケーラビリティを含む、多くの利点を提供し、単一障害点等を排除する。しかし、パーティションの転送について、メンバノードは独立した判断をしているため、複雑な分散ストラテジを実行することも非常に困難になる。たとえば、どのような頻度で特定のパーティションがアクセスされそうであるかや、特定のメンバノードに対する現在の処理負荷、および各ノードのＣＰＵ／メモリの容量などの、より動的かつ複雑な要因に従ってパーティションを分散することが望ましいかもしれない。さらには、高い利用可能性および障害迂回の理由から、より複雑かつ最適化された態様で基本パーティションおよびバックアップパーティションの配列を調節することが望ましいかもしれない。

【００１２】

種々の実施の形態に従うと、個々のパーティションの転送は二地点間（送信者ノードと受信者ノードとの間の、直接的であるが非同期の通信の結果として）で行なわれるが、単一の調整点（分散コーディネータ）を許容する。この単一の調整点は、現在の分散（データ所有者）をも含む、クラスタのグローバルな状態と、クラスタにおける所与の時点の各ノードに対する処理負荷のような動的なランタイムフィードバックとを与える。ある単一の点でシステムのグローバル・ビュー（global-view of the system）を持つことができることにより、分散アルゴリズムの代替物よりもはるかにより表現力の豊かな分散ロジックが許容される。

【００１３】

ある実施の形態に従うと、クラスタ内の各サーバノードは、分散状態を、稀に、分散コーディネータに通信する。分散コーディネータは、この情報を収集し、定期的に（または加入数の変更の結果として）構成された分散ストラテジを呼び出して、分散を分析する。さらに、現在の分散状態に対するストラテジアクセスを提供し、利用可能なサーバノードの間でのパーティションの新たな分散（配列）を示唆するためにストラテジによって利用される、プラグインが可能なロジック（ストラテジ）へのインタフェースを供給することができる。

【００１４】

ある実施の形態に従うと、分散コーディネータは、分散ストラテジによって行われた提案を、総合的な分散プランまたはゴールに集める。プランは、クラスタ内のすべてのノードに通信される。その後、各サーバは、新たな分散プランにアプローチするために個々のパーティションの転送（データ移動）を開始する。

【００１５】

様々な実施の形態に従えば、拡張可能な中央集中型の分散は、様々な追加特徴のための基礎を形成することもできる。たとえば、適応分散は、クラスタを、アプリケーションの負荷またはデータ分散の変動に動的に適応させることができる。さらに、集中型分散は、ＷＡＮに安全なクラスタリングに有用であり得る。

【００１６】

図１は、発明の様々な実施の形態に従った、クラスタ内に実装されている拡張可能な集中型のリソース分散の説明図である。この図は、構成エレメントを論理的に別々に示しているが、そのような描写は例示の目的のために過ぎない。なお、この図に描かれた構成エレメントを組み合わせる、または別々のソフトウェア、ファームウェアおよび／もしくはハードウェアに分けることができることは当業者には明らかであろう。さらに、そのような構成エレメントは、どのように組み合わせられるかまたは分割されるかに拘わらず、同一のコンピューティングデバイス上で実行することができる、または、１つ以上のネットワークもしくは他の好適な通信手段によって接続された異なるコンピューティングデバイスの間で分散できることも、当業者には明らかであろう。

【００１７】

図示されたように、クラスタは、あらゆる場所に分散されるデータパーティション（１１０、１１１、１１２、１１３、１１４、１１５、１１６）を格納する、多くのノード（１００、１０１、１０２、１０３）から構成されている。クラスタ内の１つのノード１００は、中央の分散コーディネータ１０４であることが特定されている。実施の形態によれば、分散コーディネータは、定期的に多くの要因に基づいて、クラスタ全体のパーティションの分散（配列）を再評価し、それに対して変更がなされるべきか否かを判定する。たとえば、分散コーディネータは、各ノードの現在の要求負荷、特定のパーティションがどの程度の頻度でアクセスされるか、および／または、各ノードのメモリ容量およびCPU容量に基づいて、パーティションの配列を評価することができる。同様に、中央の分散コーディネータ１０４は、バックアップパーティションに対するより複雑な基本分散を実行するために使用されることができる。たとえば、特定のノード上のすべての基本パーティションについては、分散コーディネータは、当該基本パーティションに関連付けられたバックアップパーティションが多数のノードに分散されていないようにすることができる。通

常、バックアップパーティションは、障害迂回の目的から、一般的に、基本パーティションが関連付けられているのとは異なる物理的ノードに配列されていることが好ましい。しかしながら、基本パーティションが更新されたときには、そのバックアップパーティションもアップデートされる必要がある。これは、クラスタ全体のネットワーク・トラフィックに加えて、かなりの数のノードジャンプを引き起こす可能性がある。このため、バックアップパーティション上に位置するノードの数を制限することは有利であり得る。中央の分散コーディネータは、このストラテジだけでなく、いかなる他のパーティション配列ストラテジをも実施できる。

【００１８】

ある実施の形態によれば、分散コーディネータ１０４は、グローバルなクラスタの状態105の概念を維持する。当該グローバル状態（global state）は、ノード間でのパーティションの位置、各ノードの処理負荷、各パーティションに格納されたデータに対する可能性の高い要求、各ノードのCPUおよび／または記憶容量など（ただし、これらに限定されない）を含む情報を含み得る。ある実施の形態によれば、分散コーディネータは、定期的に（またはノードの構成エレメントの変化に応答して）、クラスタ全体のパーティション分散を再評価するために、グローバル状態を採用する。クラスタのグローバル状態は、分散コーディネータ上に格納される必要はなく、選択的に、リモートで他の装置上に格納されるのと同様に、クラスタの他のエレメントにも格納され得る。

【００１９】

ある実施の形態によれば、分散コーディネータ１０４は、パーティションの分散を評価するためにプラガブル（pluggable）なロジックコンポーネント１０６を呼び出すことができる。プラガブルなロジック１０６は、クラスタに使用されるべき特定のカスタム分散ストラテジを特定することができる。分散コーディネータは、その方針に従ってパーティション分散を評価し、それに対して変更がなされるべきかどうかを判断することができる。

【００２０】

分散コーディネータ１０４は、パーティション分散に対して変更が行われるべきであると判断した場合には、分散プラン１０７を生成することができる。そして、この分散プランは、クラスタ内の各ノードに利用可能にされ得る。ある実施の形態によれば、分散プラン１０７は、どのパーティションがどのノードに位置されるべきかを特定することができる。新しい分散プランが一旦利用可能にされると、種々のノードは、後述するように、分散された態様で適切なパーティション転送を行なうことができるようになる。

【００２１】

図２は、本発明の様々な実施の形態に従って、クラスタ内のノードによって実施される分散プランの説明図である。この図は、構成エレメントを論理的に分離されたものとして示しているが、そのような描写は例示の目的のために過ぎない。なお、この図に描かれた構成エレメントは、組み合わされる、または別々のソフトウェア、ファームウェアおよび／もしくはハードウェアに分けられ得ることが、当業者には明らかであろう。さらに、そのような構成エレメントは、それらがどのように組み合わされるまたは分割されるかに関係なく、同一のコンピューティングデバイス上で実行することができ、または１つ以上のネットワークもしくは他の好適な通信手段によって接続された異なるコンピューティングデバイスの間で分散できることも当業者には明らかであろう。

【００２２】

図示された実施の形態によれば、分散コーディネータが分散プランを一旦生成すれば、当該分散プランはクラスタ内の各ノードが利用できるようになり得る。ノードは、独立して、分散プランに従って、正しいノードにパーティションを転送するために必要な決定とステップとを実行することができる。このように、パーティションの配列の決定とストラテジとは中央集中化される一方で、パーティションを分散するメカニクスは分散アルゴリズムのままである。

【００２３】

図示されたように、新たな分散プラン１０７は、ノード１００がパーティション１１０を格納すること、ノード１０１がパーティション１１１を格納すること、ノード１０２がパーティション１１３および１１４を格納すること、ならびに、ノード１０３がパーティション１１５および１１６を格納することを、特定する。現在、パーティション１１２がノード１０１上に位置しているため、ノード１００は、新たな分散プランを調査し、ノード１０１のパーティション１１２を要求し得る。同様に、ノード１０３は、ノード１０２のパーティション１１５を要求し得る。代替的な実施の形態によれば、ノード１０１および１０２は、ノード１００および１０３からの要求を待つことなく、分散プランを受信した後、ノード１００および１０３に必要なパーティションを転送することができる。

【００２４】

図３は、本発明の様々な実施の形態に従って、集中型のリソース分散を提供するためのプロセスのフローチャートである。この図は、例示の目的のために特定の順序で機能的なステップを示しているが、プロセスは、必ずしもこの特定の順序またはステップに限定されるものではない。当業者は、この図に描かれた様々なステップが、変更され、再配列され、並行して実行され、または、様々な方法で適合されることが可能であることを理解するであろう。さらに、あるステップまたはステップの配列が本発明の精神および範囲から逸脱することなくこのプロセスに追加またはこのプロセスから省略することができることが、理解されるべきである。

【００２５】

ステップ３００に示されるように、コンピュータノードのクラスタは、データを多くのパーティションとして格納する。これらのパーティションは、クラスタ内のノードに亘って分散される。ステップ301において、いずれかのコンピュータノードが中央の分散コーディネータになるように特定される。ある実施の形態では、分散コーディネータは、クラスタのグローバル状態を指示するデータを収集し、グローバル状態へのアクセスを提供する。さらにステップ302に示されているように、分散コーディネータは、定期的に、クラスタのグローバル状態を分析し、クラスタにおけるノード間でパーティションの分散に変更がなされるべきであるかどうかを判断する。代替的に、分散コーディネータは、定期的にではなく、クラスタ内のエレメントの変更に応じて、パーティションの分散を再評価することができる。

【００２６】

分散コーディネータは、変更がなされるべきであると判断した場合には、ステップ303に示されるように、それらの変更に基づいて新しい分散プランを生成し、クラスタ内のすべてのノードへの分散プランへのアクセスを提供する。ステップ304に示されるように、その後、分散アルゴリズムを用いて、個々のパーティションの転送が実行され得る。換言すれば、各ノードは、新しい分散プランを最適に実現するために、独立して、当該ノードに関連する個々のパーティションの転送をどのように実行するかを決定することができる。

【００２７】

図４は、本発明の様々な実施の形態に従って、分散コーディネータによって実行されるプロセスのフローチャートである。この図は、例示の目的のために特定の順序で機能的なステップを示しているが、プロセスは、必ずしもこの特定の順序またはステップに限定されるものではない。当業者は、この図に描かれた様々なステップが、変更され、再配列され、並行して行なわれ、または様々な方法で適合されることが可能であることを理解するであろう。さらに、工程の特定のステップまたはステップの配列は、本発明の精神および範囲から逸脱することなく、このプロセスから追加またはこのプロセスから省略され得ることが理解されるべきである。

【００２８】

図示された実施の形態によれば、プロセスはステップ400で開始される。一旦開始されると、分散コーディネータは、継続的に、クラスタのグローバル状態をコンパイルするために、クラスタ内の各ノードからランタイムフィードバックおよび他のデータを受信する

（ステップ４０１）。ステップ４０２に示されるように、分散コーディネータは、定期的に、クラスタ全体のパーティションの分散を分析するために、このグローバル状態を使用することができる。変更が必要とされない場合は（ステップ４０３）、分散コーディネータは、パーティション分散を評価する必要がある次の時間まで、何もアクションを実行し得ない。一方、分散コーディネータは、分散に変更を加えるべきであると判断した場合は、ステップ４０４に示されるように、当該変更を含む新しい分散プランを生成することができる。ステップ４０５で、分散コーディネータは、クラスタ内のすべてのノードに分散プランを提供することができる。

【００２９】

図５は、本発明の様々な実施の形態に従う、クラスタ内のノードによって実行されるプロセスのフローチャートである。この図は、例示の目的のために特定の順序で機能的なステップを示しているが、プロセスは、必ずしもこの特定の順序またはステップに限定されるものではない。当業者は、この図に描かれた様々なステップが、変更され、再配列され、並行して実行され、または、様々な方法で適合されることが可能であることを理解するであろう。さらに、特定のステップまたはステップの配列は、本発明の精神および範囲から逸脱することなく、このプロセスに追加されまたはこのプロセスから省略され得ることを理解すべきである。

【００３０】

図示された実施の形態に従えば、プロセスはステップ５００で開始される。ステップ５０１に示されるように、一旦開始されると、各ノードは、定期的に、分散コーディネータにランタイムフィードバック、負荷統計、および他のデータを送信することができる。ステップ５０１は、さらに、それ自体に戻る矢印を含み、これは、新たなプランが生成されないという可能性を表し、単に、メンバは実行し続けて、定期的に統計を集めて送信する。

【００３１】

ステップ５０２では、ノードは、分散コーディネータから新しい分散プランを受信することができる。この時点で、ノードは、分散プランを調査し、当該プランがこの特定のノードが関連する変更を特定しているかどうかを判断することができる（ステップ５０３）。新たなプランが当該ノードに係らない場合には、当該ノードは転送を実行しないことが可能であり、分散コーディネータに定期的にランタイムフィードバックを送信し続けることができる。一方、分散プランが当該ノードに係るパーティションの変更を含む場合には、当該ノードは、分散プランに従って、他のノードから必要なパーティションを取得していく、および／または、他のノードに必要なパーティションを提供することができる（ステップ５０４）。

【００３２】

本開示で説明される様々なコンテキストを通して、本発明の実施の形態は、さらに、コンピュータ装置、コンピューティングシステムと、前述のシステムおよび方法を実行するように構成された機械可読媒体を包含する。コンピュータ技術における当業者に明らかであるように、特別に設計された集積回路または他の電子機器からなる実施の形態に加えて、本発明は、従来の汎用又は本開示の教示に従ってプログラムされた専用デジタルコンピュータ又はマイクロプロセッサを用いて、簡便に、実施されることができる。

【００３３】

一般に、本発明は、拡張可能な集中型の動的なリソース分散を提供するためのシステムであって、上記システムは、コンピュータノードのクラスタに亘って分散したデータセットの複数のパーティションを格納するための手段と、上記クラスタのグローバル状態を示す情報を収集し、上記グローバル状態へのアクセスを提供する中央の分散コーディネータになる上記コンピュータノードの１つを特定するための手段と、上記コンピュータノードの間の上記パーティションの分散に対して変更がなされるべきであるかどうかを判断するために上記分散コーディネータによってクラスタのグローバル状態を定期的に分析するための手段と、上記パーティションの分散に対する変更に基づいた新たな分散プランを上記

分散コーディネータによって生成し、コンピュータノードのクラスタの全てに分散プランへのアクセスを提供するための手段と、新たな分散プランを実施するために上記ノードに関連付けられた個々のパーティションの転送をどのように実行するかをクラスタ内の各ノードによって独室して判断するための分散アルゴリズムを使用するための手段とを含む。

【００３４】

上記システムにおいて、上記コンピュータノードの間の上記パーティションの分散に対して変更がなされるべきであるかどうかを判断するために上記分散コーディネータによってクラスタのグローバル状態を定期的に分析するための上記手段は、さらに、上記分散コーディネータによってプラガブルな分散ロジックモジュールを呼び出すための手段を有し、上記プラガブルな分散ロジックモジュールはコンピュータノードのクラスタに上記パーティションを分散させるために利用されるアルゴリズムを調節するためにランタイムで切り替えられ得る。

【００３５】

上記システムにおいて、分散アルゴリズムを使用するための手段は、さらに、送信者ノードと受信者ノードとの間の直接的な非同期通信の結果として、２つのコンピュータノード間の二地点間で個々のパーティションの転送を実行するための手段を含み、分散コーディネータは上記パーティションの転送の命令に関与しない。

【００３６】

上記システムにおいて、上記クラスタのグローバル状態は、クラスタ内の各コンピュータノードに対して上記パーティションのうちどれが割り当てられるかを示す情報を含む。

【００３７】

上記システムにおいて、上記クラスタのグローバル状態は、クラスタ内の各コンピュータノードの現在の処理負荷を示す情報を含み、現在の処理負荷は、分散コーディネータにランタイムフィードバックの統計を定期的に送信する各コンピュータノードによって決定される。

【００３８】

上記システムにおいて、クラスタのグローバル状態は、クラスタ内の各コンピュータノードのメモリ容量とプロセッサ容量を示す情報とを含む。

【００３９】

上記システムにおいて、上記複数のパーティションは、基本パーティションの組とバックアップパーティションの組とを有し、分散コーディネータは、各基本パーティションが当該基本パーティションに関連付けられたバックアップパーティションとは異なる物理的ノードに位置されることを確実にする。

【００４０】

上記システムにおいて、分散コーディネータによって生成された新たな分散プランは、以下の制限、つまり、所与の各コンピュータノードに、上記所与のコンピュータノード上に位置する基本パーティションに関連付けられたバックアップパーティションを含むことを許可される多数のコンピュータノードを限定すること、を課す。

【００４１】

上記システムにおいて、分散コーディネータは、クラスタ内のコンピュータノードの間の上記パーティションの分散のための単一の調整点を提供する。

【００４２】

上記システムにおいて、新たな分散プランは、特定のパーティションがクラスタ内の特定されたコンピュータノード上に位置することを特定する。

【００４３】

好適なソフトウェアコーディングは、ソフトウェア技術の当業者には明らかなように、本開示の教示に基づき、熟練したプログラマによって、容易に準備され得る。当業者には容易に明らかであろうように、本発明は、また、特定用途向け集積回路を準備することにより又は従来のコンポーネント回路の好適なネットワークを相互接続することにより実現されてもよい。

【００４４】
様々な実施の形態は、ここに表されたすべての特徴を実行するための汎用のまたは専用のコンピューティングプロセッサ（単数または複数）／デバイス（単数または複数）をプログラムするために使用することができる命令を記憶された記憶媒体であるコンピュータプログラム製品を含む。記憶媒体は、以下のものの一つ以上のものを含み得るが、これらには限られない。フロッピー（登録商標）ディスク、光ディスク、ＤＶＤ、ＣＤ－ＲＯＭ、マイクロドライブ、光磁気ディスク、ホログラム記憶装置、ＲＯＭ、ＲＡＭ、ＰＲＡＭ、ＥＰＲＯＭ、ＥＥＰＲＯＭ、ＤＲＡＭ、ＶＲＡＭ、フラッシュメモリ装置、磁気もしくは光カード、ナノシステム（分子メモリＩＣを含む）を含む物理的媒体、紙もしくは紙ベースの媒体、ならびに、命令および／もしくは情報を記憶するための好適な任意の種類の媒体または装置。コンピュータプログラム製品は、全体的にまたは部分的に送信され得、また、１つ以上の公衆のおよび／または専用のネットワークを介して送信され得、そして、当該送信は、本明細書中に提示された特徴の任意のものを実行するために１つ以上のプロセッサによって用いられ得る命令を含む。送信は、複数の別個の送信を含むことができる。しかしながら、ある実施の形態に従えば、命令を含むコンピュータ記憶媒体は、非一時的（すなわち、伝送される過程になり）であるが、むしろ物理的デバイス上に保持されている。

【００４５】
１．拡張可能な集中型の動的なリソース分散を提供するための方法であって、上記方法は、
コンピュータノードのクラスタにわたって分散しているデータセットの複数のパーティションを記憶することと、
上記クラスタのグローバル状態を示す情報を収集し、上記グローバル状態へのアクセスを提供する中央の分散コーディネータとなる上記コンピュータノードの１つを特定することと、
上記コンピュータノードの間の上記パーティションの分散に対して変更を加えるべきであるかどうかを判断するために、定期的に、分散コーディネータによるクラスタのグローバル状態を分析することと、
上記パーティションの分散に対する上記変更に基づいて、上記分散コーディネータによって、新たな分散プランを生成し、コンピュータノードのクラスタの全てに、上記分散プランへのアクセスを提供することと、
上記新たな分散プランを実施するために、上記ノードに関連付けられた個々のパーティション転送をどのように実行するかをクラスタ内の各ノードによって独立して判断するために分散アルゴリズムを使用することとを含む。

【００４６】
２．請求項１に記載の方法において、上記コンピュータノードの間で上記パーティションの分散に対して変更を加えるべきであるかどうかを判断するために、定期的に、分散コーディネータによるクラスタのグローバル状態を分析することは、さらに、
上記分散コーディネータによってプラガブルな分散ロジックモジュールを呼び出すことを含み、上記プラガブルな分散ロジックモジュールは、コンピュータノードのクラスタの間で上記パーティションを分散するために利用されるアルゴリズムを調節するためにランタイムで切り替えられ得る。

【００４７】
３．請求項１に記載の方法において、分散アルゴリズムを使用することは、さらに、
送信者ノードと受信者ノードとの間の直接的な非同期通信の結果として、２つのコンピュータノード間を二地点間で個々のパーティションの転送を実行することを含み、上記分散コーディネータは上記パーティションの転送の命令に関与しない。

【００４８】
４．請求項１に記載の方法において、上記クラスタのグローバル状態は、クラスタ内の各コンピュータノードに対して上記パーティションのうちどれが割り当てられるかを示す

情報を含む。
【００４９】
５．請求項１に記載の方法において、上記クラスタのグローバル状態は、クラスタ内の各コンピュータノードの現在の処理負荷を示す情報を含み、現在の処理負荷は、上記分散コーディネータにランタイムフィードバックの統計を定期的に送信する各コンピュータノードによって判断される。
【００５０】
６．請求項１に記載の方法において、上記クラスタのグローバル状態は、クラスタ内の各コンピュータノードのメモリ容量とプロセッサ容量とを示す情報を含む。
【００５１】
７．請求項１に記載の方法において、上記複数のパーティションは、さらに、基本パーティションの組とバックアップパーティションの組とを含み、分散コーディネータは、各基本パーティションが当該基本パーティションに関連付けられたバックアップパーティションとは異なる物理的ノード上に位置されるようにする。
【００５２】
８．請求項７に記載の方法において、分散コーディネータによって生成された新たな分散プランは、以下の制限を課す、
所与の各コンピュータノード毎に、上記所与のコンピュータノード上に位置する基本パーティションに関連付けられたバックアップパーティションを含むことを許される多数のコンピュータノードを限定する。
【００５３】
９．請求項１に記載の方法において、上記分散コーディネータは、上記クラスタ内のコンピュータノードの間での上記パーティションの分散のための単一の調整点を提供する。
【００５４】
１０．請求項１に記載の方法において、上記新しい分散プランは、特定のパーティションが上記クラスタ内の特定されたコンピュータノード上に位置することを特定する。
【００５５】
１１．拡張可能な集中型の動的なリソース分散を提供するためのシステムであって、上記システムは、
データセットの複数のパーティションを記憶しているコンピュータノードのクラスタを含み、上記パーティションは、コンピュータノードの上記クラスタにわたって分散しており、さらに、
コンピュータノードから選択された分散コーディネータを含み、上記分散コーディネータは、上記クラスタのグローバル状態を示す情報を収集し、上記コンピュータノードの間の上記パーティションの分散に対して変更を加えるべきであるかどうかを判断するために、定期的にグローバル状態を分析し、上記変更に基づいて新たな分散プランを生成し、コンピュータノードのクラスタの全てに上記分散プランへのアクセスを提供し、
コンピュータノードの上記クラスタは、上記分散コーディネータによって生成された上記新たな分散プランを実行するために上記ノードに関連付けられた個々のパーティション転送をどのように実行するかをクラスタ内の各ノードによって独室して判断するための分散アルゴリズムを使用する。
【００５６】
１２．請求項１１に記載のシステムにおいて、上記分散コーディネータは、上記変更がなされるべきであるかどうかを判断するためにプラガブルな分散ロジックモジュールを呼び出し、上記プラガブルな分散ロジックモジュールは、コンピュータノードのクラスタの間で上記パーティションを分散するために利用されるアルゴリズムを調節するためにランタイムで切り替えられ得る。
【００５７】
１３．請求項１１に記載のシステムにおいて、分散アルゴリズムをコンピュータノードによって使用することは、さらに、

送信者ノードと受信者ノードとの間の直接的な非同期通信の結果として、２つのコンピュータノード間を二地点間で個々のパーティションを送信することを有し、上記分散コーディネータは上記パーティションの転送の命令に関与しない。

【００５８】

１４．請求項１１に記載のシステムにおいて、上記クラスタのグローバル状態は、クラスタ内の各コンピュータノードに対して上記パーティションのうちどれが割り当てられるかを示す情報を含む。

【００５９】

１５．請求項１１に記載のシステムにおいて、上記クラスタのグローバル状態は、クラスタ内の各コンピュータノードの現在の処理負荷を示す情報を含み、現在の処理負荷は、分散コーディネータにランタイムフィードバックの統計を定期的に送信する各コンピュータノードによって決定される。

【００６０】

１６．請求項１１に記載のシステムにおいて、上記クラスタのグローバル状態は、クラスタ内の各コンピュータノードのメモリ容量とプロセッサ容量とを示す情報を含む。

【００６１】

１７．請求項１１に記載のシステムにおいて、上記複数のパーティションは、基本パーティションの組とバックアップパーティションの組とを有し、分散コーディネータは、各基本パーティションが当該基本パーティションに関連付けられたバックアップパーティションとは異なる物理的ノードに位置されることを確実にする。

【００６２】

１８．請求項１７に記載のシステムにおいて、分散コーディネータによって生成された新たな分散プランは、以下の制限を課す、

所与の各コンピュータノードに対して、上記所与のコンピュータノード上に位置する基本パーティションに関連付けられたバックアップパーティションを含むことを許可されるコンピュータノードの数を限定する。

【００６３】

１９．請求項１１に記載の方法において、上記新たな分散プランは、特定のパーティションが、上記クラスタ内の特定されたコンピュータノード上に位置することを特定する。

【００６４】

２０．以下のステップの組を実行するために１つ以上のプロセッサによって実行可能な命令の１つ以上のシーケンスを格納する、非一時的コンピュータ可読媒体であって、上記ステップの組は、

コンピュータノードのクラスタに亘って分散しているデータセットの複数のパーティションを記憶することと、

上記クラスタのグローバル状態を示す情報を収集し、上記グローバル状態へのアクセスを提供する中央の分散コーディネータとなるように上記コンピュータノードの１つを特定することと、

上記コンピュータノードの間での上記パーティションの分散に対して変更を加えるべきであるかどうかを判断するために、定期的に、分散コーディネータによりクラスタのグローバル状態を分析することと、

上記パーティションの分散に対する上記変更に基づいて、上記分散コーディネータによって、新たな分散プランを生成し、コンピュータノードのクラスタの全てに、上記分散プランへのアクセスを提供することと、

上記新たな分散プランを実施するために上記ノードに関連付けられた個々のパーティション転送をどのように実行するかをクラスタ内の各ノードによって独立して判断するために分散アルゴリズムを使用することとを含む。

【００６５】

２１．拡張可能な集中型の動的なリソース分散を提供するためのシステムであって、上記システムは、

コンピュータノードのクラスタに亘って分散しているデータセットの複数のパーティションを記憶するための手段と、
上記クラスタのグローバル状態を示す情報を収集し、上記グローバル状態へのアクセスを提供する中央の分散コーディネータとなるように上記コンピュータノードの１つを指定するための手段と、
上記コンピュータノードの間の上記パーティションの分散に対して変更を加えるべきであるかどうかを判断するために、定期的に、分散コーディネータによるクラスタのグローバル状態を分析するための手段と、
上記パーティションの分散に対する上記変更に基づいて、上記分散コーディネータによって、新たな分散プランを生成し、コンピュータノードのクラスタの全てに、上記分散プランへのアクセスを提供するための手段と、
上記新たな分散プランを実施するために上記ノードに関連付けられた個々のパーティション転送をどのように実行するかをクラスタ内の各ノードによって独立して決定するために分散アルゴリズムを使用するための手段とを含む。
【００６６】
２２．請求項２１に記載のシステムにおいて、上記コンピュータノードの間の上記パーティションの分散に対して変更を加えるべきであるかどうかを判断するために、定期的に、分散コーディネータによるクラスタのグローバル状態を分析するための手段は、
上記分散コーディネータによってプラガブルな分散ロジックモジュールを呼び出すための手段を有し、上記プラガブルな分散ロジックモジュールは、コンピュータノードのクラスタの間で上記パーティションを分散するために利用されるアルゴリズムを調節するためにランタイムで切り替えられ得る。
【００６７】
２３．請求項２１に記載のシステムにおいて、分散アルゴリズムを使用するための手段は、さらに、
送信者ノードと受信者ノードとの間の直接的な非同期通信の結果として、２つのコンピュータノード間を二地点間で個々のパーティションを送信するための手段を有し、上記分散コーディネータは上記パーティションの転送の命令に関与しない。
【００６８】
２４．請求項２１に記載のシステムにおいて、上記クラスタのグローバル状態は、クラスタ内の各コンピュータノードに対して上記パーティションのうちどれが割り当てられるかを示す情報を含む。
【００６９】
２５．請求項２１に記載のシステムにおいて、上記クラスタのグローバル状態は、クラスタ内の各コンピュータノードの現在の処理負荷を示す情報を含み、現在の処理負荷は、上記分散コーディネータにランタイムフィードバックの統計を定期的に送信する各コンピュータノードによって決定される。
【００７０】
２６．請求項２１に記載のシステムにおいて、上記クラスタのグローバル状態は、クラスタ内の各コンピュータノードのメモリ容量とプロセッサ容量とを示す情報を含む。
【００７１】
２７．請求項２１に記載のシステムにおいて、上記複数のパーティションは、基本パーティションの組とバックアップパーティションの組とを有し、分散コーディネータは、各基本パーティションが当該基本パーティションに関連付けられたバックアップパーティションとは異なる物理的ノードに位置されることを確実にする。
【００７２】
２８．請求項２７に記載のシステムにおいて、分散コーディネータによって生成された新たな分散プランは、以下の制限を課す、
所与の各コンピュータノードに対して、上記所与のコンピュータノード上に位置する基本パーティションに関連付けられたバックアップパーティションを含むことを許される多

数のコンピュータノードを限定する。

【００７３】

２９．請求項２１に記載のシステムにおいて、上記分散コーディネータは、上記クラスタ内のコンピュータノードの間での上記パーティションの分散のための単一の調整点を提供する。

【００７４】

３０．請求項２１に記載のシステムにおいて、上記新しい分散プランは、特定のパーティションが上記クラスタ内の特定されたコンピュータノード上に位置することを特定する。

【００７５】

本発明の好ましい実施の形態の上記の説明は、例示および説明の目的で提供されている。網羅的であることまたは開示された厳格な形態に本発明を限定することを意図するものではない。多くの修正および変更が当業者には明らかであり得る。実施の形態は、本発明の原理およびその実際の応用を最も適切に説明することによって、関連技術に精通する他者が本発明を理解できるように、選択され説明された。なお、本発明の範囲は以下の請求の範囲およびその均等物によって規定されることが意図される。

【図１】

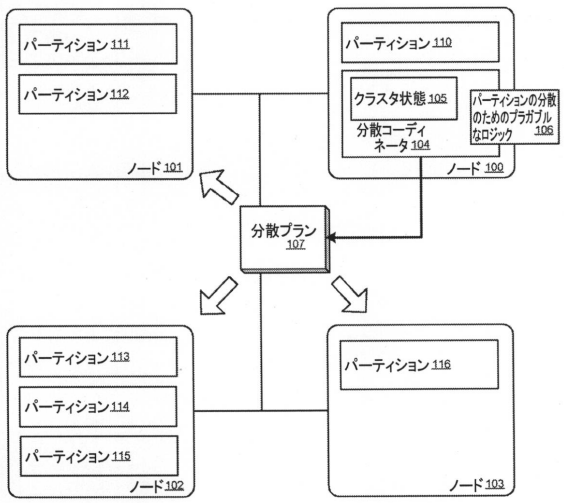

FIGURE 1

【図２】

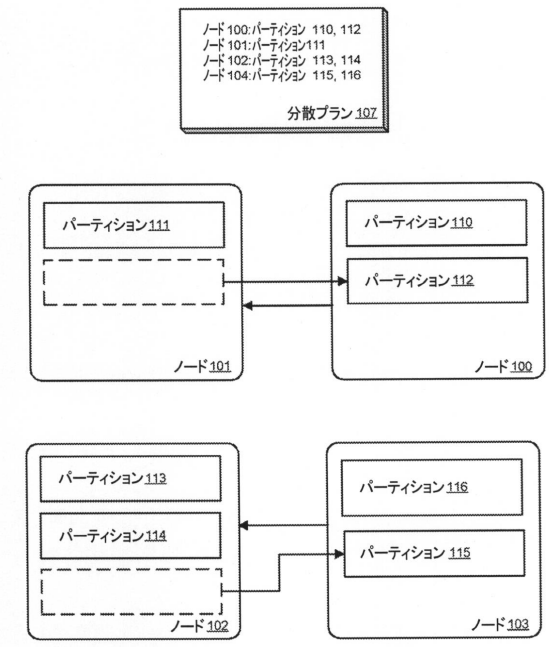

FIGURE 2

【図３】

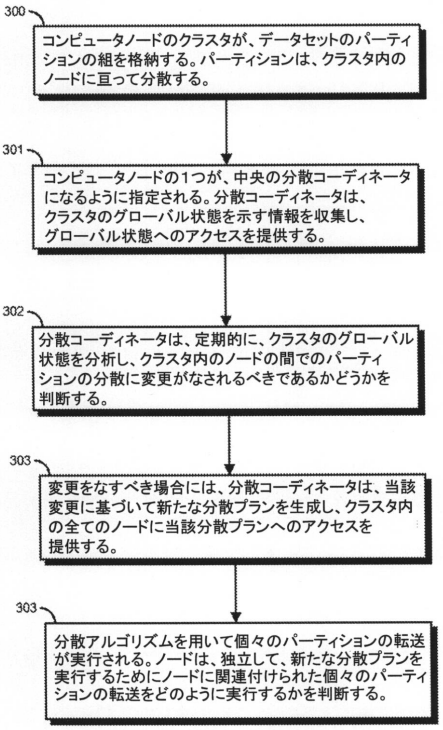

FIGURE 3

【図４】

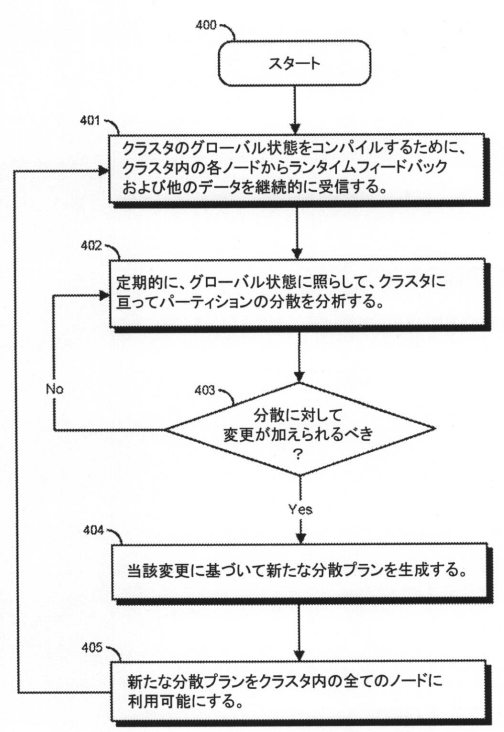

FIGURE 4

【図５】

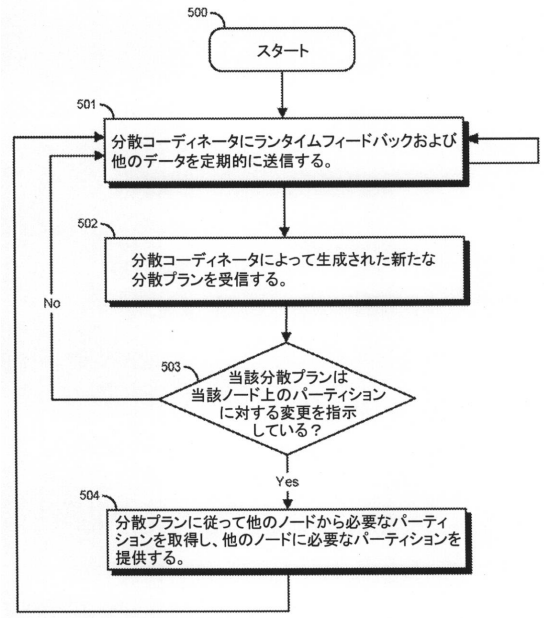

FIGURE 5

フロントページの続き

(72)発明者　グレイザー，ジーン
アメリカ合衆国、０２４２１　マサチューセッツ州、レキシントン、ジョゼフ・カミー・ロード、１
(72)発明者　ファルコ，マーク
アメリカ合衆国、０１８０３　マサチューセッツ州、バーリントン、ウィルミントン・ロード、７４
(72)発明者　パーディ，キャメロン
アメリカ合衆国、０１８０３　マサチューセッツ州、バーリントン、ネットワーク・ドライブ、３５

審査官　田中　幸雄

(56)参考文献　国際公開第２００８／０５６５０７（WO，Ａ１）
特開２０１０－２７７５１７（ＪＰ，Ａ）
特開２００６－１９５６９４（ＪＰ，Ａ）
特表２００８－５１９３１９（ＪＰ，Ａ）
特開２００９－１１６８８４（ＪＰ，Ａ）
米国特許出願公開第２００５／０１４４２８３（ＵＳ，Ａ１）
米国特許第７０８０２２１（ＵＳ，Ｂ１）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１２／００