【公報種別】特許法第１７条の２の規定による補正の掲載
【部門区分】第６部門第３区分
【発行日】平成23年5月26日(2011.5.26)

【公表番号】特表2003-515813(P2003-515813A)
【公表日】平成15年5月7日(2003.5.7)
【出願番号】特願2001-540586(P2001-540586)
【国際特許分類】
*Ｇ０６Ｆ　15/177　(2006.01)*
*Ｇ０６Ｆ　 9/46　(2006.01)*
*Ｇ０６Ｆ　12/00　(2006.01)*
【ＦＩ】
Ｇ０６Ｆ　15/177　６８２Ｇ
Ｇ０６Ｆ　 9/46　３６０Ｅ
Ｇ０６Ｆ　12/00　５０１Ａ
Ｇ０６Ｆ　12/00　５１４Ｅ

【誤訳訂正書】
【提出日】平成23年4月7日(2011.4.7)
【誤訳訂正１】
【訂正対象書類名】明細書
【訂正対象項目名】全文
【訂正方法】変更
【訂正の内容】
【書類名】明細書
【発明の名称】ストレージネットワーク内のクォーラムリソースアービタ
【特許請求の範囲】
【請求項１】　複数のコンピュータノードと、１つまたは複数の記憶デバイス内に含まれる物理クォーラムリソースを具備する記憶サブシステムとを備えるシステムにおいて、前記物理クォーラムリソースがストレージネットワークの前記コンピュータノードに利用可能である前記ストレージネットワークを形成する方法であって、
前記システムの前記コンピュータノードのそれぞれにより前記記憶サブシステムの前記物理クォーラムリソースについての何れかの現在の所有権をプロセッサが終了させるステップと、
新しいストレージネットワークの構成情報を更新するために前記システムのそれぞれの他のコンピュータノードを前記プロセッサがスキャンするステップと、
１つまたは複数の前記コンピュータノードにより前記物理クォーラムリソースの所有権を決定するためにアービトレーションを前記プロセッサが呼び出すステップと、
前記物理クォーラムリソースを備えるクォーラムボリュームを前記ストレージネットワークのメンバーに前記プロセッサが組み込むステップとを備え、
前記現在の所有権を終了させるステップと、前記アービトレーションを呼び出すステップは、第１のプログラムモジュールにより実行され、前記それぞれの他のコンピュータノードをスキャンするステップと、前記クォーラムボリュームを組み込むステップは、前記第１のプログラムモジュールと異なる第２のプログラムモジュールにより実行されることを特徴とする方法。
【請求項２】　前記現在の所有権を終了させるステップは、前記記憶サブシステムへのアクセスをブロックすることを含むことを特徴とする請求項１に記載の方法。
【請求項３】　前記現在の所有権を終了させるステップは、それぞれのコンピュータノードのバスをリセットすることを含むことを特徴とする請求項１に記載の方法。
【請求項４】　前記現在の所有権を終了させるステップは、アービトレーションを呼

び出す前に所定の遅延期間待機することを含むことを特徴とする請求項１に記載の方法。

【請求項５】　前記ストレージネットワークの前記コンピュータノードにより所有された物理クォーラムリソースを含む前記記憶サブシステムのそれぞれの記憶デバイスを識別する構成情報を前記プロセッサが生成するステップを更に備えることを特徴とする請求項１に記載の方法。

【請求項６】　内部構成データベースを再構築するために、前記生成された構成情報に含まれるそれぞれの記憶デバイスからの前記クォーラムボリュームの情報を前記プロセッサが処理するステップを更に備えることを特徴とする請求項５に記載の方法。

【請求項７】　前記クォーラムボリュームを組み込むステップは、コンピュータノードがクォーラムボリュームに必要な全ての記憶デバイスの所有権を獲得したとき実行されることを特徴とする請求項１に記載の方法。

【請求項８】　前記クォーラムボリュームを組み込むステップは、コンピュータノードが前記物理クォーラムリソースの過半数の所有権を獲得したとき実行されることを特徴とする請求項１に記載の方法。

【請求項９】　１以上の前記物理クォーラムリソースは、１つまたは複数の連結されかつストライプされたボリュームエクステントを具備し、

前記クォーラムボリュームを組み込むステップは、所有権が１より大きい前記連結されかつストライプされたボリュームエクステントのために獲得されたとき実行されることを特徴とする請求項１に記載の方法。

【請求項１０】　１以上の前記物理クォーラムリソースは、１つまたは複数の連結されかつストライプされたボリュームエクステントを具備し、

前記クォーラムボリュームを組み込むステップは、所有権が前記１つまたは複数の連結されかつストライプされたボリュームエクステントの単純な過半数のために獲得されたとき実行されることを特徴とする請求項１に記載の方法。

【請求項１１】　複数のコンピュータノードと、１つまたは複数の記憶デバイス内に含まれる物理クォーラムリソースを具備する記憶サブシステムとを備えるシステム内で、前記物理クォーラムリソースがストレージネットワークの前記コンピュータノードに利用可能である前記ストレージネットワークを形成するための各ステップを実施するコンピュータ実行可能命令を記憶するコンピュータ可読記録媒体であって、前記各ステップは、

前記システムの前記コンピュータノードのそれぞれにより前記記憶サブシステムの前記物理クォーラムリソースについての何れかの現在の所有権をプロセッサが終了させるステップと、

新しいストレージネットワークの構成情報を更新するために前記システムのそれぞれの他のコンピュータノードを前記プロセッサがスキャンするステップと、

１つまたは複数の前記コンピュータノードにより前記物理クォーラムリソースの所有権を決定するためにアービトレーションを前記プロセッサが呼び出すステップと、

前記物理クォーラムリソースを備えるクォーラムボリュームを前記ストレージネットワークのメンバーに前記プロセッサが組み込むステップとを備え、

前記現在の所有権を終了させるステップと、前記アービトレーションを呼び出すステップは、第１のプログラムモジュールにより実行され、前記それぞれの他のコンピュータノードをスキャンするステップと、前記クォーラムボリュームを組み込むステップは、前記第１のプログラムモジュールと異なる第２のプログラムモジュールにより実行されることを特徴とするコンピュータ可読記録媒体。

【請求項１２】　前記現在の所有権を終了させるステップは、前記記憶サブシステムへのアクセスをブロックすることを含むことを特徴とする請求項１１に記載のコンピュータ可読記録媒体。

【請求項１３】　前記現在の所有権を終了させるステップは、それぞれのコンピュータノードのバスをリセットすることを含むことを特徴とする請求項１１に記載のコンピュータ可読記録媒体。

【請求項１４】　前記現在の所有権を終了させるステップは、アービトレーションを

呼び出す前に所定の遅延期間待機することを含むことを特徴とする請求項１１に記載のコンピュータ可読記録媒体。

【請求項１５】　前記ストレージネットワークの前記コンピュータノードにより所有された物理クォーラムリソースを含む前記記憶サブシステムのそれぞれの記憶デバイスを識別する構成情報を前記プロセッサが生成するステップを更に備えることを特徴とする請求項１１に記載のコンピュータ可読記録媒体。

【請求項１６】　内部構成データベースを再構築するために、前記生成された構成情報に含まれるそれぞれの記憶デバイスからの前記クォーラムボリュームの情報を前記プロセッサが処理するステップを更に備えることを特徴とする請求項１５に記載のコンピュータ可読記録媒体。

【請求項１７】　前記クォーラムボリュームを組み込むステップは、コンピュータノードがクォーラムボリュームに必要な全ての記憶デバイスの所有権を獲得したとき実行されることを特徴とする請求項１１に記載のコンピュータ可読記録媒体。

【請求項１８】　前記クォーラムボリュームを組み込むステップは、コンピュータノードが前記物理クォーラムリソースの過半数の所有権を獲得したとき実行されることを特徴とする請求項１１に記載のコンピュータ可読記録媒体。

【請求項１９】　１以上の前記物理クォーラムリソースは、１つまたは複数の連結されかつストライプされたボリュームエクステントを具備し、

前記クォーラムボリュームを組み込むステップは、所有権が１より大きい前記連結されかつストライプされたボリュームエクステントのために獲得されたとき実行されることを特徴とする請求項１１に記載のコンピュータ可読記録媒体。

【請求項２０】　１以上の前記物理クォーラムリソースは、１つまたは複数の連結されかつストライプされたボリュームエクステントを具備し、

前記クォーラムボリュームを組み込むステップは、所有権が前記１つまたは複数の連結されかつストライプされたボリュームエクステントの単純な過半数のために獲得されたとき実行されることを特徴とする請求項１１に記載のコンピュータ可読記録媒体。

【発明の詳細な説明】

【０００１】

（関連出願）

本出願は、本出願と同日に出願され同じ譲受人に譲渡された、以下のすべての出願に関連するものである。

第０９／４４９５７７号（整理番号７７７２４５ＵＳ１）「Ｓｔｏｒａｇｅ Ｍａｎａｇｅｍｅｎｔ Ｓｙｓｔｅｍ Ｈａｖｉｎｇ Ｃｏｍｍｏｎ Ｖｏｌｕｍｅ Ｍａｎａｇｅｒ」

第０９／４５０３６４号（整理番号７７７２４６ＵＳ１）「Ｓｔｏｒａｇｅ Ｍａｎａｇｅｍｅｎｔ Ｓｙｓｔｅｍ Ｈａｖｉｎｇ Ａｂｓｔｒａｃｔｅｄ Ｖｏｌｕｍｅ Ｐｒｏｖｉｄｅｒｓ」

第０９／４５１２１９号（整理番号７７７２４７ＵＳ１）「Ｖｏｌｕｍｅ Ｓｔａｃｋｉｎｇ Ｍｏｄｅｌ」

第０９／４５０３００号（整理番号７７７２４８ＵＳ１）「Ｖｏｌｕｍｅ Ｃｏｎｆｉｇｕｒａｔｉｏｎ Ｄａｔａ Ａｄｍｉｎｉｓｔｒａｔｉｏｎ」

第０／４５１２２０号（整理番号７７７２４９ＵＳ１）「Ｖｏｌｕｍｅ Ｍｉｇｒａｔｉｏｎ Ｂｅｔｗｅｅｎ Ｖｏｌｕｍｅ Ｇｒｏｕｐｓ」

【０００２】

（発明の分野）

本発明は、一般にデータ記憶デバイスに関し、より詳細には、ストレージネットワーク内での論理クォーラムリソースのアービトレーション機構に関する。

【０００３】

（版権の通知／許可）

本特許文書の開示の一部には、版権保護を対象とする資料が含まれている。版権所有者

は、特許商標事務所の特許ファイルまたは記録に表示されているものと同じ特許文書または特許開示を複製することに異議はないが、そうでない場合はあらゆる版権を保有する。以下の通知は、下記および図面に記載するようなソフトウェアおよびデータに適用される。



【０００４】

（発明の背景）

コンピュータシステムを発展させる場合、磁気ディスクや光ディスクなどのデータ記憶デバイスを利用および構成することができる。たとえば、これらの記憶デバイスは、バスを介してコンピュータシステムに接続するか、または無線または有線のネットワークを介してコンピュータシステムに接続することができる。さらにこれらの記憶デバイスは、分離するかまたは単一キャビネット内に配置することができる。

【０００５】

ストレージネットワークとは、ノードと呼ばれ、単一の記憶資源として動作する、相互接続されたコンピューティングシステムの集まりのことである。ストレージネットワークでは、ハードウェアまたはソフトウェアに障害が発生した場合でもシステムの動作を続行させることが可能であり、ノードを簡単に追加できるようにすることでスケーラビリティを高くし、管理者がノードを単一のシステムとして管理できるようにすることで管理を簡略化する。

【０００６】

クラスタソフトウェアはそれぞれのノード上にあり、ストレージネットワークのクラスタに特有のすべての活動を管理する。クラスタソフトウェアは、しばしば、ノードのスタートアップを自動的に実行する。この際、クラスタソフトウェアはローカルの非共有デバイスを構成して装着する。また、クラスタソフトウェアは、ストレージネットワークの他のメンバが動作可能であるかどうかを判定するための「発見」プロセスも使用する。クラスタソフトウェアは既存のクラスタを発見すると、認証シーケンスを実行することによってそのクラスタの接合を試みる。既存クラスタのクラスタマスタは新規クラスタを認証し、接合ノードの認証が成功すれば、その状況を戻す。ノードがメンバとして認識されない場合は、接合要求が拒否される。

【０００７】

発見プロセス中にクラスタが見つからない場合、ノードは独自のクラスタの形成を試みる。このプロセスは、ノードが自分の属するクラスタと通信できないときには何回でも繰り返される。従来のコンピューティングシステムでは、ノードは、ストレージネットワークを形成するために、ディスクなどの物理「クォーラムリソース」（ｐｈｙｓｉｃａｌ“ｑｕｏｒｕｍ　ｒｅｓｏｕｒｃｅ”）に対してアービトレーションを実行する。最近のシステムでは、クォーラムリソースは、１つまたは複数の物理クォーラムリソースを含むボリュームなどの論理リソースが可能である。たとえばボリュームは、ディスクの一部分、ディスク全体、複数ディスクの一部分、または複数ディスクである可能性のある論理記憶ユニットである。

【０００８】

従来のシステムでは、クラスタの所有権を決定する、すなわちアービトレーションプロセスを実施するための役割および情報が、いくつかの構成要素および／またはソフトウェアモジュール間で分配されることが多い。基礎となる記憶デバイスを構成および管理するための役割も、同様に分配されることが多い。このように役割を明確に分割しないと、所与の構成要素またはソフトウェアモジュールを変更することが困難になる。したがって、当分野では、クラスタ管理からのクラスタアービトレーションの役割と、ボリューム管理および基礎となる記憶デバイスの役割とを、より明確に分離することが求められている。

【０００９】

（発明の概要）

本発明は、上記の短所、欠点、および問題点に対処するものである。本発明のクラスタ管理ソフトウェアおよびボリューム管理ソフトウェアは、ストレージネットワークのノード上で実行され、クォーラムボリュームなどの論理クォーラムリソースをアービトレーションするために、基礎となるオペレーティングシステムと協働して動作する。本発明によれば、クラスタ管理ソフトウェアは論理クォーラムリソースをアービトレーションし、基礎となる物理クォーラムリソースの知識を持たずにストレージネットワークを形成する。この方法では、クラスタ管理ソフトウェアはハードウェア特有のものではない。さらに、クラスタ管理ソフトウェアは、基礎となる物理クォーラムリソースから論理クォーラムリソースがどのように形成されるかを知っておく必要がない。たとえば、ボリューム管理ソフトウェアが管理する役割は、論理クォーラムボリュームの形成および装着のみである。ボリューム管理ソフトウェアは、アービトレーションプロセスおよび所有権決定に関する詳細な知識を持たずに、ボリューム管理を実行する。

【００１０】

（発明の詳細な説明）

本発明の例示的な実施形態を詳細に説明する以下の記述では、本明細書の一部であり、本発明が実施可能な特有の例示的な実施形態が例として示されている、添付の図面を参照する。これらの実施形態は、当分野の技術者が本発明を実施できるように詳細に記述されており、他の実施形態が利用可能であること、および本発明の範囲を逸脱することなく変更が可能であることが理解されよう。したがって、以下の詳細な説明は限定的なものではなく、本発明の範囲は特許請求の範囲によってのみ定義される。

【００１１】

詳細な説明は、４つのセクションに分けられる。第１のセクションは用語集である。第２のセクションでは、本発明の実施形態がそれと共に実施可能なハードウェアおよび動作環境について説明する。第３のセクションでは、本発明のシステムレベルの概要を示す。最後に第４のセクションでは、詳細な説明の結果を示す。

【００１２】

[用語の定義]

損失：フォールトトレラントボリュームに、１つまたは複数のディスクまたはボリュームエクステントがないことを示す状況。たとえば、現在１つのミラーしか使用できないミラーセット。

【００１３】

構成データ：物理リソースを論理ボリュームにマッピングすることを示す。

【００１４】

指示された構成：プロバイダに、論理ブロックの再マッピングを選択するための規則が明示的に与えられること。

【００１５】

ディスクプラッタ：ディスクパックからボリュームをエクスポートまたはインポートするのに使用される、ディスクパックのサブセット。

【００１６】

ディスクパック：論理ボリュームと基礎となるディスクの集まり。ディスクパックとは、ボリュームの推移閉包の単位である。

【００１７】

エクスポート：ディスクプラッタおよびそのプラッタに含まれるすべてのボリュームを、１つのディスクパックから移動すること。

【００１８】

露出された：ボリュームに、関連するボリューム名（ドライブ文字）または装着位置がある場合、ボリュームはオペレーティングシステムに対して露出されている。ボリュームは、ファイルシステムまたは他のデータ格納に使用可能にする
ことができる。

【００１９】

フリーエージェントドライブ：ディスクパックのメンバでないディスクドライブ。フリーエージェントドライブは、露出された論理ボリュームを含むことはできない。
【００２０】
健全：ボリューム障害管理状況。ボリュームの状況には、初期、健全、損失、不健全、または再構築がある。
【００２１】
健全な：有効データを含んでいるかまたは含むことができる。
【００２２】
ホットスポッティング：ボリュームまたはボリュームエクステントの集まりを一時的にプレクシング（ｐｌｅｘｉｎｇ）すること。
【００２３】
インポート：ディスクプラッタおよびそのプラッタに含まれるすべてのボリュームを、１つのディスクパックに移動すること。
【００２４】
初期化：ボリュームがボリューム構成を再発見していることを示す状況。
【００２５】
ＬＢＮ：論理ブロック番号
【００２６】
論理ブロックマッピング：論理ボリュームプロバイダに対して露出された論理ブロックと、同じプロバイダによって露出された論理ブロックとの関係。
【００２７】
論理クォーラムリソース：ストレージネットワークの形成に必要な論理リソース。論理ボリュームなどの論理クォーラムリソースには、ディスクなどの１つまたは複数の物理クォーラムリソースが含まれる。
【００２８】
論理ボリューム：ディスクの一部分、ディスク全体、複数ディスクの一部分、または複数ディスクである可能性のある、論理記憶ユニット。
【００２９】
論理ボリュームプロバイダ：論理ボリュームを露出するソフトウェア。プロバイダには、ランタイムサービス、構成データ、および管理サービスが含まれる。
【００３０】
管理サービス：ボリュームの構成、モニタリング、または障害処理を実行するために、まれにしか実行されないソフトウェア。
【００３１】
マッピング済みボリューム：単一のより大きなボリュームを露出するためにボリュームを連結する、単純な線形の論理ブロックマッピング。
【００３２】
ミラーリング済みボリューム：２つまたはそれ以上の同一のデータコピーを維持する論理ボリューム。ＲＡＩＤ １とも呼ばれる。
【００３３】
パリティストライピング済みボリューム：パリティチェック情報ならびにデータを維持する論理ボリューム。厳密なマッピングおよび保護方式は、ベンダ特有のものである。ＲＡＩＤ３、４、５、６が含まれる。
【００３４】
プレクシング済みボリューム：動的なミラーボリューム。プレクシングは、フォールトトレランスを提供するのではなく、ボリュームのコピーを作成するのに使用される。ミラーは、コンテンツが同期化された後に除去する意図で、ボリュームに追加される。
【００３５】
ＲＡＩＤ：Ｒｅｄｕｎｄａｎｔ Ａｒｒａｙ ｏｆ Ｉｎｄｅｐｅｎｄｅｎｔ Ｄｉｓｋｓ

【００３６】
再構築：以前に損失したフォールトトレラントボリュームがすべてのボリュームエクステントデータを再同期化することを示す状況。
【００３７】
ランタイムサービス：ＩＯごとの要求ベースで実行されるソフトウェア。
【００３８】
ＳＣＳＩ：Ｓｍａｌｌ－Ｃｏｍｐｕｔｅｒ　Ｓｙｓｔｅｍｓ　Ｉｎｔｅｒｆａｃｅ
【００３９】
スタック済みボリューム：複数の論理ブロックマッピング動作によって構築されたボリューム。一例として、ミラーボリュームのストライプセットが挙げられる。スタッキングには、ストリッピング、マッピング、およびプレクシングが含まれる。
【００４０】
ストライピング済みボリューム：連続する論理ボリュームエクステントを複数のボリュームにまたがって分配する、論理ブロックマッピング。ＲＡＩＤ０とも呼ばれる。
【００４１】
不健全：非フォールトトレラントボリュームに、１つまたは複数のディスクまたはボリュームエクステントがないことを示す状況。不健全ボリュームに含まれるデータにアクセスしてはいけない。
【００４２】
ボリューム構成安定性：ボリュームの論理対物理マッピングが変化しているかどうか。ボリュームには、安定、拡張、縮小、プレクシング、または再マッピングがある。
【００４３】
ボリュームエクステント：ボリューム上に含まれる論理ブロックの連続する領域。ボリュームエクステントは、最小管理対象論理ボリューム単位である。
【００４４】
ボリューム状況：システムによる現在のボリューム使用状況。ボリュームには、未使用、ホットスペア、マップ済み、使用、または未知がある。
【００４５】
[ハードウェアおよび動作環境]
図１は、本発明の実施形態がそれと共に実施可能なハードウェアおよび動作環境を示す図である。図１の説明は、本発明がそれと共に実施可能な、適切なコンピュータハードウェアおよび適切なコンピューティング環境を、簡単かつ概略的に説明することを意図している。必須ではないが、本発明は、パーソナルコンピュータなどのコンピュータによって実行される、プログラムモジュールなどのコンピュータ実行可能命令の一般的な情況で説明される。一般に、プログラムモジュールには、特定のタスクを実行するか、または、特定の抽象データ型を実施する、ルーチン、プログラム、オブジェクト、構成要素、データ構造などが含まれる。
【００４６】
さらに当分野の技術者であれば、本発明が、ハンドヘルドデバイス、マルチプロセッサシステム、マイクロプロセッサベースまたはプログラム可能な大衆消費電子製品、ネットワークＰＣ、ミニコンピュータ、メインフレームコンピュータなどを含む、他のコンピュータシステム構成でも実施可能であることを理解されよう。本発明は、通信ネットワークを介してリンクされたリモート処理デバイスによってタスクが実行される、分散型コンピューティング環境でも実施可能である。分散型コンピューティング環境では、プログラムモジュールは、ローカルおよびリモートのどちらのメモリ記憶デバイスにも配置可能である。
【００４７】
本発明を実施するための図１の例示的なハードウェアおよび動作環境には、コンピュータ２０の形式の汎用コンピューティングデバイスが含まれ、これには、処理ユニット２１、システムメモリ２２、ならびに、システムメモリ２２を含む様々なシステム構成要素を

処理ユニット２１に動作可能に結合するシステムバス２３が含まれる。処理ユニット２１は１つだけでも複数であってもよく、その結果、コンピュータ２０のプロセッサには、単一の中央処理ユニット（ＣＰＵ）が含まれるか、または一般に並列処理環境と呼ばれる複数の処理ユニットが含まれる。コンピュータ２０は、従来のコンピュータ、分散型コンピュータ、または任意の他のタイプのコンピュータが可能であり、本発明はそのように限定されていない。

【００４８】

システムバス２３は、任意の様々なバスアーキテクチャを使用する、メモリバスまたはメモリ制御装置、周辺バス、およびローカルバスを含む、いくつかのタイプのバス構造のうちいずれかでよい。システムメモリは単にメモリと呼ぶことも可能であり、読取り専用メモリ（ＲＯＭ）２４およびランダムアクセスメモリ（ＲＡＭ）２５を含む。起動中などにコンピュータ２０内の要素間で情報を転送するのを助ける基本ルーチンを含む基本入出力システム（ＢＩＯＳ）２６が、ＲＯＭ２４に格納される。さらにコンピュータ２０は、図示されていないハードディスクからの読取りおよびハードディスクへの書込みのためのハードディスクドライブ２７、取外し可能磁気ディスク２９からの読取りまたはこのディスクへの書込みのための磁気ディスクドライブ２８、ならびに、ＣＤ－ＲＯＭまたは他の光学式媒体などの取外し可能光ディスク３１からの読取りまたはこのディスクへの書込みのための光ディスクドライブ３０を含む。

【００４９】

ハードディスクドライブ２７、磁気ディスクドライブ２８、および光ディスクドライブ３０は、それぞれハードディスクドライブインターフェース３２、磁気ディスクドライブインターフェース３３、および光ディスクドライブインターフェース３４によって、システムバス２３に接続される。ドライブおよびこれらに関連付けられたコンピュータ可読媒体は、コンピュータ可読命令、データ構造、プログラムモジュール、およびコンピュータ２０に関する他のデータの、不揮発性記憶域を提供する。当分野の技術者であれば、磁気カセット、フラッシュメモリカード、デジタルビデオディスク、ベルヌイカートリッジ、ランダムアクセスメモリ（ＲＡＭｓ）、読取り専用メモリ（ＲＯＭｓ）などの、コンピュータによってアクセス可能なデータを格納できる任意のタイプのコンピュータ可読媒体が、例示的動作環境で使用可能であることを理解されたい。

【００５０】

いくつかのプログラムモジュールが、ハードディスク２７、磁気ディスク２９、光ディスク３１、ＲＯＭ２４、あるいは、オペレーティングシステム３５、１つまたは複数のアプリケーションプログラム３６、他のプログラムモジュール３７、およびプログラムデータ３８を含むＲＡＭ２５上に格納可能である。ユーザは、コマンドおよび情報を、キーボード４０およびポインティングデバイス４２などの入力デバイスを介して、パーソナルコンピュータ２０に入力することができる。他の入力デバイス（図示せず）には、マイクロフォン、ジョイスティック、ゲームパッド、衛星放送用アンテナ、スキャナなどが含まれる。これらおよび他の入力デバイスは、システムバスに結合されたシリアルポートインターフェース４６を介して、処理ユニット２１に接続されることが多いが、パラレルポート、ゲームポート、またはユニバーサルシリアルバス（ＵＳＢ）などの他のインターフェースによって接続されることも可能である。モニタ４７または他のタイプの表示デバイスも、ビデオアダプタ４８などのインターフェースを介して、システムバス２３に接続される。コンピュータには、モニタに加えて、典型的にはスピーカおよびプリンタなどの他の周辺出力デバイス（図示せず）が含まれる。

【００５１】

コンピュータ２０は、リモートコンピュータ４９などの１つまたは複数のリモートコンピュータへの論理接続を使用して、ネットワーク化された環境で動作することができる。これらの論理接続は、ローカルコンピュータであるコンピュータ２０に結合されたかまたはこの一部である通信デバイスによって達成され、本発明は特定タイプの通信デバイスに限定されるものではない。リモートコンピュータ４９は、他のコンピュータ、サーバ、ル

ータ、ネットワークＰＣ、クライアント、ピア（ｐｅｅｒ）デバイス、または他の共通ネットワークノードであってよく、典型的には、コンピュータ２０に関連した前述の要素の多くまたはすべてを含むが、図１にはメモリ記憶デバイス５０のみが図示されている。図１に示されている論理接続には、ローカルエリアネットワーク（ＬＡＮ）５１およびワイドエリアネットワーク（ＷＡＮ）５２が含まれる。こうしたネットワーキング環境は、事務所、全社的なコンピュータネットワーク、イントラネット、およびインターネットで、一般的に見られるものである。

【００５２】

コンピュータ２０は、ＬＡＮネットワーキング環境で使用される場合、通信デバイスの一種であるネットワークインターフェースまたはアダプタ５３を介して、ローカルネットワーク５１に接続される。コンピュータ２０は、ＷＡＮネットワーキング環境で使用される場合、典型的には、モデム５４、通信デバイスの一種、あるいはインターネットなどのワイドエリアネットワーク５２を介して通信を確立するための他の任意のタイプの通信デバイスを含む。モデム５４は内部または外部であってよく、シリアルポートインターフェース４６を介してシステムバス２３に接続される。ネットワーク化された環境では、パーソナルコンピュータ２０に関連して、またはその一部として示されたプログラムモジュールは、リモートメモリ記憶デバイスに格納することができる。図示されたネットワーク接続は例示的なものであり、コンピュータ間に通信リンクを確立するための他の手段および通信デバイスが使用可能であることが理解されよう。

【００５３】

以上、本発明の実施形態がそれと共に実施可能なハードウェアおよび動作環境について説明してきた。本発明の実施形態がそれと共に実施可能なコンピュータは従来のコンピュータ、分散型コンピュータ、または任意の他のタイプのコンピュータであってよく、本発明はそのように限定されていない。こうしたコンピュータには、典型的にはその処理装置として１つまたは複数の処理ユニット、およびメモリなどのコンピュータ可読媒体が含まれる。コンピュータは、ネットワークアダプタまたはモデムなどの通信デバイスも含むことができるため、他のコンピュータと通信上で結合することができる。

【００５４】

［システムレベルの概要］

図２は、ネットワーク１２０を介してノード１１０に通信上で結合されたノード１０５を含む、ストレージネットワーク１００のシステムレベルの概要を示す構成図である。ノード１０５および１１０は、図１に示されたローカルコンピュータ２０またはリモートコンピュータ４９などの、任意の適切なコンピューティングシステムを表す。

【００５５】

ストレージネットワーク１００は、記憶デバイス１０７、記憶デバイス１０８、および記憶デバイス１０９を含む、記憶サブシステム１０６をさらに含む。これらのデバイスは、単一の内部ディスク、複数の外部ディスク、またはＲＡＩＤキャビネットなどの、任意の適切な記憶媒体であってよい。記憶サブシステム１０６は、デュアル接続ＳＣＳＩ（Ｓｍａｌｌ－Ｃｏｍｐｕｔｅｒ　Ｓｙｓｔｅｍｓ　Ｉｎｔｅｒｆａｃｅ）、ファイバチャネルなどの任意の適切な相互接続メカニズムである、バス１１２を介して結合される。

【００５６】

ストレージネットワーク１００を形成するために、ノード１０５および１１０は、クォーラムボリュームなどの論理クォーラムリソースをアービトレーションする。図２では、論理クォーラムリソースが、物理クォーラムリソース１１１によって集合的に形成されたクォーラムボリュームとして示されており、本実施形態では、データ記憶デバイス１０８およびデータ記憶デバイス１０９内のデータ記憶エクステントである。ノード１０５または１１０のいずれかがすべての物理クォーラムリソース１１１の所有権を得ることに成功すると、成功したノードはストレージネットワーク１００を形成することができる。以下で説明するように、本発明のクラスタ管理ソフトウェアおよびボリューム管理ソフトウェアは各ノード上で実行し、物理クォーラムリソース１１１の所有権がノード１０５と１１

０とで分けられている状況を解決する。各ノード上で、クラスタ管理ソフトウェアおよびボリューム管理ソフトウェアが、ストレージネットワーク１００を形成するために基礎となるオペレーティングシステムと協働する。以下で説明するように、アービトレーションおよび管理の役割は、クラスタ管理ソフトウェアとボリューム管理ソフトウェアで分担され、その結果、クラスタ管理ソフトウェアは、ボリューム管理の詳細および記憶サブシステム１０６について知らずに、アービトレーションプロセスを処理する。ボリューム管理ソフトウェアは、ストレージネットワーク１００がどのように形成されるかを知らずに、記憶サブシステム１０６の構成および管理を処理する。

【００５７】

図３は、様々な協働するソフトウェア構成要素が本発明のアービトレーション技法を実行する、図２のノード１０５またはノード１１０などのノード２００の一実施形態を示す構成図である。ノード２００では、クラスタマネージャ２０２がすべてのクラスタに特有の活動を監視し、ディスク制御装置２０６を介して記憶サブシステム１０６のバス１１２（図２）と通信する。クラスタマスタとして、クラスタマネージャ２０２、ボリュームマネージャ２０４、およびオペレーティングシステム３５が協働して、ストレージネットワーク１００および対応する物理クォーラムリソース１１１のクォーラムボリュームを管理する。より具体的に言えば、クラスタマネージャ２０２は、ボリューム管理および記憶サブシステム１０６の詳細を知らずに、アービトレーションプロセスを処理する。ボリュームマネージャは、すべてのボリュームマッピングおよびストレージネットワーク１００の記憶サブシステム１０６の構成を処理する。ディスク制御装置２０６は、記憶サブシステム１０６とのすべての通信を処理し、ＳＣＳＩ、ＩＰなどの様々なデータ通信プロトコルのうち１つを実施することができる。アプリケーション２１０は、ストレージネットワーク１００と対話する、任意のユーザモードソフトウェアモジュールを表す。以上、本セクションでは、本発明の例示的な実施形態のシステムレベルの動作概要について、詳細に説明した。

【００５８】

（本発明の例示的な実施形態の方法）

前のセクションでは、本発明の例示的な実施形態のシステムレベルの動作概要について説明した。このセクションでは、例示的な実施形態を実行中のコンピュータによって実行される特定の方法について、一連の流れ図を参照しながら説明する。コンピュータによって実行される方法は、コンピュータ実行可能命令で構成されるコンピュータプログラムとなる。流れ図を参照しながらこの方法について説明することで、当分野の技術者は、適切なコンピュータ（コンピュータ読み取り可能媒体からの命令を実行中のコンピュータのプロセッサ）上で方法を実行するためのこのような命令を含む、このようなプログラムを展開することができる。

【００５９】

図４は、本発明が、クラスタ管理の役割とボリューム管理の役割を明確に分離する方法を示す図である。より具体的に言えば、アービトレーションサイクル３００は、ストレージネットワーク１００の各ノード上でクラスタマネージャ２０２およびボリュームマネージャによって実行されるときの、本発明の変形アービトレーション方法の一実施形態を示す。アービトレーションサイクル３００は、ストレージネットワーク１００がまだ確立されていないとき、たとえば、ノード１０５または１１０のいずれかが最初にブートされるときか、またはストレージネットワーク１００があらかじめ形成されているがノード１０５と１１０の間の通信が故障しているときにいつでも、などに呼び出される。

【００６０】

アービトレーションサイクル３００は、ノード１０５またはノード１１０のいずれかが、ブロック３０２からブロック３０４に進むことによって開始される。ブロック３０４では、クラスタマネージャ２０２（図３）が、記憶サブシステム１０６の現在の所有権をすべて終了させる。一実施形態では、これはバス１１２をリセットすることによって達成される。次にこの動作によって、ストレージネットワーク１００のすべての他のノードにア

ービトレーションサイクル３００を実行させ、すべてのボリュームをオフラインモードにする。このモードでは、ボリュームマネージャ２０４が記憶サブシステム１０６のすべてのアクセスをブロックする。一実施形態では、アービトレーション実行ノードは、ストレージネットワーク１００のすべてのノードが確実にアービトレーションに入るようにするために、アービトレーションサイクル３００が進行するまでの所定の遅延期間、待機する。

【００６１】

ブロック３０６では、それぞれの新しいかまたは除去された記憶デバイス１０６についての構成情報を更新するために、クラスタマネージャ２０２が、ストレージネットワーク１００内のすべての他のノードをスキャンするように、ボリュームマネージャ２０４に命令する。ブロック３０６の終わりには、他のノードによって所有されていた情報が変更されている場合があるために、ボリュームマネージャ２０４によって維持される構成情報は、部分的にのみ完了することになる。したがって、ブロック３０８では、クラスタマネージャ２０２が、以前にストレージネットワーク１００のノードによって所有されていた、それらの記憶サブシステム１０６を識別するリストを生成するように、ボリュームマネージャ２０４に命令する。

【００６２】

ブロック３０９では、ボリュームマネージャ２０４が、生成されたリストの各記憶デバイス１０６から情報を読み取って処理する。ボリュームマネージャ２０４は、内部構成データベースを再構築する。この動作によって、アービトレーションサイクル３００以前に、クォーラムリソース全体が異なるノードによって所有されていた場合であっても、アービトレーション実行ノードがストレージネットワーク１００に関するクォーラムリソースを確実に発見する。ブロック３０９が終わると、ボリュームマネージャ２０４は、すべての記憶サブシステム１０６およびそのすべてのボリュームに関する情報を得る。

【００６３】

次に、ブロック３１０で、クラスタマネージャ２０２は、クォーラムボリュームに関連付けられたすべての物理クォーラムリソース１１１を識別するように、ボリュームマネージャ２０４に要求する。ボリュームマネージャ２０４は、すべての記憶サブシステム１０６が物理クォーラムリソース１１１を有することを判定し、ストレージネットワーク１００に関するクォーラムボリューム情報を再構築する。たとえば図１を参照すると、確実にボリュームをオンライン上に持ち出すことができるようにするために、ボリュームマネージャ２０４が記憶デバイス１０８および１０９の所有権を必要に応じて識別する。ブロック３１０が完了すると、クォーラムボリューム情報は、ストレージネットワーク１００のすべてのノードに対して一貫したものとなる。この時点で、クラスタマネージャ２０２は記憶デバイス１０８および１０９の所有権を得ようと試みる。

【００６４】

ブロック３１２では、クラスタマネージャ２０２が、物理クォーラムリソース、すなわち記憶デバイス１０８および１０９をアービトレーションするために、ＳＣＳＩプロトコルによって指定された技法などの、バス１１２によって提供された従来のアービトレーション技法を呼び出す。これらの従来のメカニズムが終わると、ノード１０５または１１０のいずれかが、記憶デバイス１０８および１０９の両方を所有することができるか、あるいは従来のアービトレーション技法に提示された競合条件により、物理クォーラムリソース１１１の所有権を分けることができる。

【００６５】

物理クォーラムリソース１１１のアービトレーションが完了した後、ボリュームマネージャ２０４は、ローカルノード、すなわちクラスタマネージャ２０２を実行中のノードが、クォーラムボリュームに必要な記憶デバイス１０８および１０９の両方の所有権を首尾よく取得したかどうかを判定する。首尾よく取得した場合、ボリュームマネージャ２０４はクォーラムボリュームを装着し、ブロック３１６でクラスタマネージャ２０２は、ローカルノードがクラスタマスタになることを宣言し、ストレージネットワーク１００が形成

されたことを他のノードに通知する。この時点で、他のノードはアービトレーションを終了し、ストレージネットワーク１００を接合する。

【００６６】

ローカルノードが記憶デバイス１０８および１０９の両方の所有権を有していない場合、ボリュームマネージャ２０４はブロック３１４から３１８に進み、ローカルノードが任意のクォーラムボリュームリソース、すなわち記憶デバイス１０８または１０９のいずれかの所有権を取得したかどうかを判定する。ローカルノードがいずれかの所有権を有していない場合、制御はクラスタマネージャ２０２に渡り、これがブロック３２０でアービトレーションを終了して、最終的にクラスタマスタとなる他のノードからの通信を待つ。

【００６７】

アービトレーション実行ノードが、記憶デバイス１０８および１０９の両方ではなく一方の所有権を有する場合、ボリュームマネージャ２０４はブロック３１８からブロック３２２に進み、ボリュームリストがクォーラムを形成するのに十分であるかどうかを判定する。ボリュームマネージャは、ボリュームリストが適切であるかどうかを判定する際に、単純な多数決または加重投票方式などの、いくつかの異なるアルゴリズムを使用する。ボリュームリストが不十分な場合、ボリュームマネージャ２０４は何らかのクォーラムリソースを解放する。クラスタマネージャ２０２はブロック３２０に進み、最終的にクラスタマスタになる他のノードからの通信を待つ。

【００６８】

ただし、ボリュームマネージャ２０４が、ボリュームリストが十分であると判定した場合、ボリュームマネージャ２０４はブロック３２２からブロック３２３に進み、クォーラムボリュームの装着が安全かどうかを判定する。この判定は、ボリューム特有の情報に基づいて行われる。たとえば、クォーラムボリュームが連結されたかまたはストライピングされたエクステントを使用する場合、ボリュームマネージャ２０４は、常に、１つのエクステントのみが所有されているときにはクォーラムボリュームの装着は安全でないと判定する。他の例として、クォーラムボリュームがＲＡＩＤＶである場合、ボリュームマネージャ２０４は、１つを除くすべてのエクステントが必要であるというような、「マイナス１」アルゴリズムを適用することができる。さらに、ボリュームマネージャ２０４はユーザが選択可能な基準を適用することもできる。たとえば、クォーラムボリュームがミラーである場合、ユーザはすべてのエクステントを要求するか、または単純な多数決を要求するように、ボリュームマネージャ２０４を構成することができる。ボリュームマネージャ２０４がクォーラムボリュームを安全に装着できる場合、ボリュームマネージャ２０４はクォーラムボリュームを装着し、クラスタマネージャ２０２はブロック３１６に進んで、ローカルノードをクラスタマスタとして宣言する。

【００６９】

ただし、ボリュームマネージャ２０４がクォーラムボリュームの装着が安全に実行できないと判定した場合、クラスタマネージャ２０２は所定の期間待機する。ブロック３２６で、所定の時間内にクラスタマスタからの通信が受け取られなかった場合、クラスタマネージャ２０２はブロック３０４に戻り、本発明のアービトレーション方法を繰り返す。一実施形態では、アービトレーションサイクル３００が反復されるごとに、遅延期間が増加する。

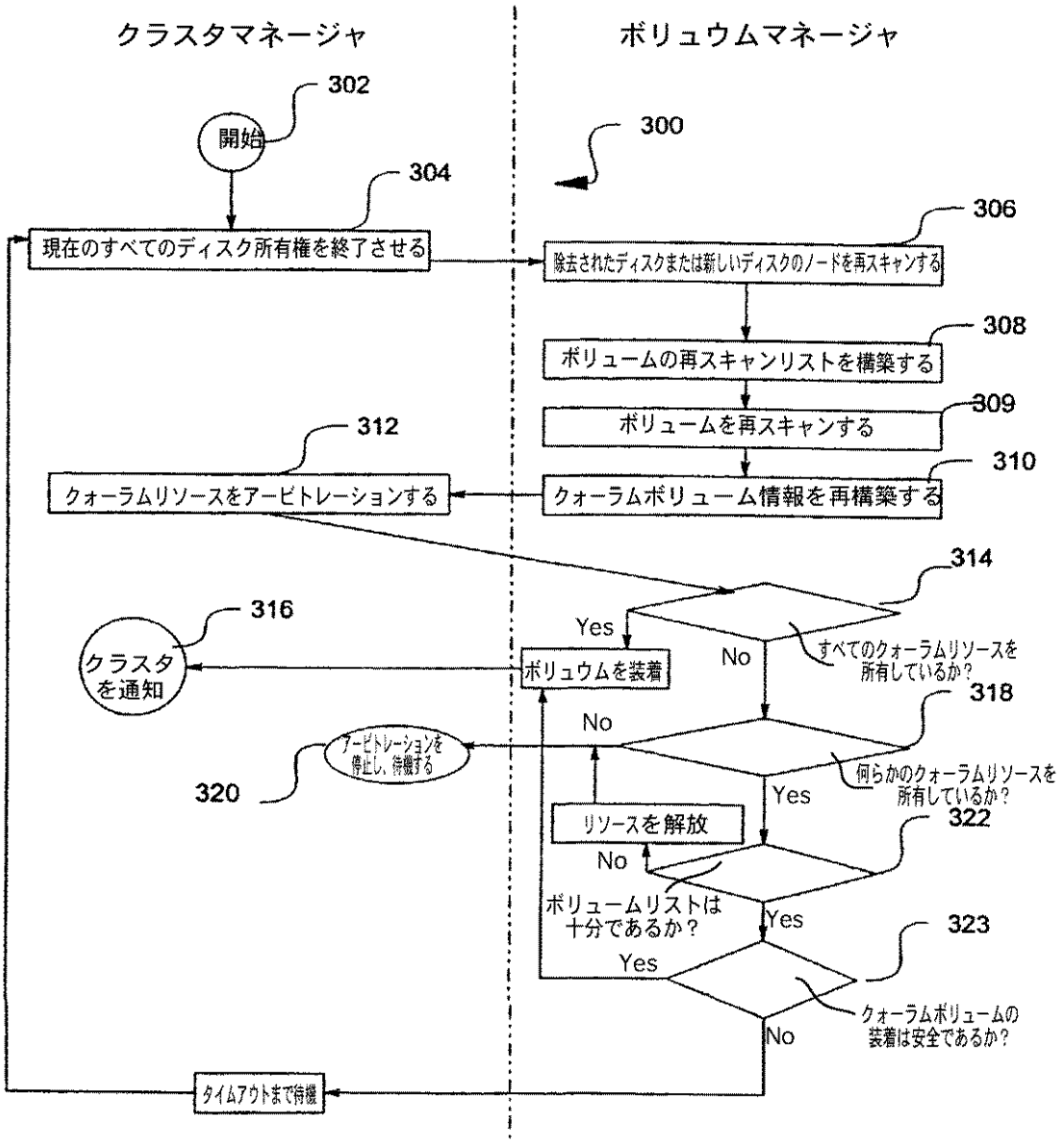
【００７０】

［結論］

以上、ボリューム管理ならびに論理リソースの構成の基礎となる物理的特性についての知識を必要とせずに、クラスタソフトウェアが論理クォーラムリソースのアービトレーションを実行できるようにする、本発明のアービトレーション方式の様々な実施形態について説明してきた。ボリューム管理ソフトウェアは、アービトレーションプロセスを介してクラスタの所有権がどのように確立されるかを知らずに、基礎となる記憶デバイスを管理する。この方法で、本発明は、クラスタ管理の役割とボリューム管理の役割を明確に分離する。本発明は、特許請求の範囲およびその等価物によってのみ限定されるものであるこ

とが意図されている。
【図面の簡単な説明】
【図１】
本発明の実施形態がそれと共に実施可能なハードウェアおよび動作環境を示す図である。
【図２】
２つのコンピューティングシステムおよび様々な記憶デバイスを備えたストレージネットワークのシステムレベルの概要を示すブロック図である。
【図３】
クラスタアービトレーションの役割とボリュームおよび基礎となる記憶デバイスの管理とを明確に分離した、協働するソフトウェア構成要素を有するソフトウェアシステムの一実施形態を示すブロック図である。
【図４】
本発明に従ってシステムが論理クォーラムリソースをアービトレーションする、図３のソフトウェアシステムの一動作モードを示すフローチャートである。
【誤訳訂正２】
【訂正対象書類名】図面
【訂正対象項目名】図４
【訂正方法】変更
【訂正の内容】

【図４】


クラスタマネージャ
ボリュウムマネージャ
302
開始
300
304
現在のすべてのディスク所有権を終了させる
306
除去されたディスクまたは新しいディスクのノードを再スキャンする
308
ボリュームの再スキャンリストを構築する
309
ボリュームを再スキャンする
310
クォーラムボリューム情報を再構築する
312
クォーラムリソースをアービトレーションする
314
すべてのクォーラムリソースを所有しているか？
Yes
No
316
クラスタを通知
ボリュウムを装着
318
何らかのクォーラムリソースを所有しているか？
No
Yes
320
アービトレーションを停止し、待機する
リソースを解放
322
ボリュームリストは十分であるか？
No
Yes
323
クォーラムボリュームの装着は安全であるか？
Yes
No
タイムアウトまで待機