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

特許第6382958号
(P6382958)

(45)発行日　平成30年8月29日(2018.8.29)　(24)登録日　平成30年8月10日(2018.8.10)

| (51)Int.Cl. | | F I | | |
|---|---|---|---|---|
| G06F 3/06 | (2006.01) | G06F | 3/06 | 304H |
| G06F 13/10 | (2006.01) | G06F | 13/10 | 340A |
| G06F 13/14 | (2006.01) | G06F | 13/14 | 310Z |
| | | G06F | 13/14 | 310B |

請求項の数 19　(全 23 頁)

(21)出願番号　特願2016-516710 (P2016-516710)
(86)(22)出願日　平成26年5月27日 (2014.5.27)
(65)公表番号　特表2016-526229 (P2016-526229A)
(43)公表日　平成28年9月1日 (2016.9.1)
(86)国際出願番号　PCT/US2014/039480
(87)国際公開番号　WO2014/193770
(87)国際公開日　平成26年12月4日 (2014.12.4)
審査請求日　平成29年5月16日 (2017.5.16)
(31)優先権主張番号　13/904,989
(32)優先日　平成25年5月29日 (2013.5.29)
(33)優先権主張国　米国 (US)

(73)特許権者　314015767
マイクロソフト　テクノロジー　ライセンシング，エルエルシー
アメリカ合衆国　ワシントン州　９８０５２　レッドモンド　ワン　マイクロソフト　ウェイ
(74)代理人　100140109
弁理士　小野　新次郎
(74)代理人　100075270
弁理士　小林　泰
(74)代理人　100101373
弁理士　竹内　茂雄
(74)代理人　100118902
弁理士　山本　修

最終頁に続く

(54)【発明の名称】クラスタにおけるストレージ防御の分散

(57)【特許請求の範囲】
【請求項１】
ストレージデバイスへのアクセスを可能にするための方法であって、
特定のノードにより、前記ストレージデバイスへ書込み要求を送信するステップと、
前記特定のノードにより、前記ストレージデバイスから登録テーブルを受信するステップであって、前記登録テーブルはノードのクラスタ内の１以上の他のノードを識別する、ステップと、
前記特定のノードにより、前記特定のノードの第１の登録キーが前記登録テーブル内に存在しないかを決定するステップと、
前記特定のノードにより、第２の登録キーを、前記クラスタへの加入再承認及び前記ストレージデバイスへの書込みアクセスのために前記ストレージデバイスへ送信するステップと、
前記第２の登録キーを送信するステップに基づいて、前記特定のノードにより、第１の登録タイマを設定するステップであって、前記ストレージデバイスにより維持される前記登録テーブルを変更するために、前記第１の登録タイマは、前記ノードのクラスタ内の１以上のノードに、時間期間を提供する、ステップと、
前記第１の登録タイマの期限切れと同時に、前記特定のノードにより前記ストレージデバイスから前記登録テーブルを受信するステップであって、前記登録テーブルは、前記時間期間の間のノードのクラスタ内の前記１以上の他のノードからの変更を含む、ステップと、

前記特定のノードにより、前記第２の登録キーが前記登録テーブル内に記憶されているかを決定するステップと、
前記第２の登録キーが前記登録テーブルに記憶されているときに、前記クラスタ内の前記１つ以上の他のノードを通じて前記ストレージデバイスへのリモート接続を確立し、前記ストレージデバイスへ書込みするステップと
を含む方法。
【請求項２】
前記クラスタに加入するステップ後、前記登録テーブルをスクラブするステップをさらに含む、請求項１に記載の方法。
【請求項３】
前記登録テーブルをスクラブするステップは、
前記登録テーブルを前記ストレージデバイスから受信するステップと、
前記登録テーブルを読み取るステップと、
前記登録テーブルから１つまたは複数の登録キーを抹消するステップであって、前記１つまたは複数の登録キーのそれぞれは、前記クラスタ内の他のノードと関連している、ステップと
を含む、請求項２に記載の方法。
【請求項４】
請求項３に記載の方法であって、さらに第２の登録タイマを設定するステップを備え、前記第１の登録タイマの長さは、前記第２の登録タイマの長さの少なくとも２倍である、方法。
【請求項５】
前記第１の登録キーおよび第２の登録キーのそれぞれは、クラスタ識別子、およびノード識別子を含む、請求項１に記載の方法。
【請求項６】
コンピュータ実行可能な命令を符号化するコンピュータ可読ストレージデバイスであって、前記コンピュータ実行可能な命令は、１つまたは複数のプロセッサによって実行されると、ストレージデバイスへのアクセスを可能にするための方法を行い、前記方法は、
特定のノードにより、前記ストレージデバイスへ書込み要求を送信するステップと、
前記特定のノードにより、前記ストレージデバイスから登録テーブルを受信するステップであって、前記登録テーブルはノードのクラスタ内の１以上の他のノードを識別する、ステップと、
前記特定のノードにより、前記特定のノードの第１の登録キーが前記登録テーブル内に存在しないことを決定するステップと、
前記特定のノードにより、第２の登録キーを、前記ストレージデバイスへ送信するステップと、
前記第２の登録キーを送信するステップに基づいて、前記特定のノードにより、第１の登録タイマを設定するステップであって、前記ストレージデバイスにより維持される前記登録テーブルを変更するために、前記第１の登録タイマは、ノードのクラスタ内の１以上の他のノードに、時間期間を提供する、ステップと、
前記第１の登録タイマの期限切れと同時に、前記特定のノードにより前記ストレージデバイスから前記登録テーブルを受信するステップであって、前記登録テーブルは、前記時間期間の間のノードのクラスタ内の前記１以上のノードからの変更を含む、ステップと、
前記特定のノードにより、前記第２の登録キーが前記登録テーブル内に記憶されているかを決定するステップと、
前記第２の登録キーが前記登録テーブルに記憶されているときに、前記クラスタ内の前記１つ以上のノードを通じて前記ストレージデバイスへのリモート接続を確立し、前記ストレージデバイスへ書込みするステップと
を含む方法。
【請求項７】

前記クラスタに加入するステップ後、前記登録テーブルをスクラブするための命令をさらに含む、請求項６に記載のコンピュータ可読ストレージデバイス。
【請求項８】
前記登録テーブルをスクラブするステップは、
前記ストレージデバイスから前記登録テーブルを受信するステップと、
前記登録テーブルを読み取るステップと、
前記登録テーブルから１つまたは複数の登録キーを抹消するステップであって、前記１つまたは複数の登録キーのそれぞれは、前記クラスタ内の他のノードと関連している、ステップと
を含む、請求項７に記載のコンピュータ可読ストレージデバイス。
【請求項９】
請求項８に記載のコンピュータ可読ストレージデバイスであって、第２の登録タイマを設定するための命令をさらに備え、前記第１の登録タイマの長さは、前記第２の登録タイマの長さの少なくとも２倍である、コンピュータ可読ストレージデバイス。
【請求項１０】
請求項９に記載のコンピュータ可読ストレージデバイスであって、前記ストレージデバイスへの接続を生成するための命令をさらに備える、コンピュータ可読ストレージデバイス。
【請求項１１】
前記接続は物理的接続である、請求項１０に記載のコンピュータ可読ストレージデバイス。
【請求項１２】
前記第１の登録キー及び第２の登録キーのそれぞれは、クラスタ識別子、およびノード識別子を含む、請求項６に記載のコンピュータ可読ストレージデバイス。
【請求項１３】
ストレージデバイスへのアクセスを可能にするための方法であって、
クラスタ内の特定のノードから、前記ストレージデバイスに書込みコマンドを送信するステップであって、前記特定のノードは、関連登録キーを有する、ステップと、
前記書込みコマンドが拒否されたという通知を受信すると同時に、
前記特定のノードにより、前記ストレージデバイスに登録テーブルを要求するステップと、
前記特定のノードにより、前記特定のノードと関連する前記関連登録キーが前記登録テーブルに存在するかどうかを決定するステップと、
前記特定のノードと関連する前記関連登録キーが前記登録テーブルに存在しないとき、
前記特定のノードにより、前記ストレージデバイスに新規の登録キーを送信するステップと、
前記特定のノードにより、第１の登録タイマを設定するステップであって、前記第１の登録タイマは前記ストレージデバイスにより維持された前記登録テーブルを修正するために、ノードのクラスタ内の１以上の他のノードに時間期間を供給する、ステップと、
前記第１の登録タイマの期限切れと同時に、前記特定のノードにより、前記ストレージデバイスから前記登録テーブルを受信するステップであって、前記登録テーブルは、前記時間期間の間にノードのクラスタ内の前記１以上の他のノードからの修正を含む、ステップと、
前記特定のノードにより、前記新規の登録キーが前記登録テーブルに記憶されているかどうかを決定するステップと、
前記新規の登録キーが前記登録テーブルに記憶されているとき、前記クラスタに加入するステップであって、前記クラスタに加入するステップは、前記クラスタ内の１以上のノードを通じて前記ストレージデバイスへのリモート接続を確立することにより、前記特定のノードが前記ストレージデバイスに書き込むことを可能にする、ステップと

を含む方法。
【請求項１４】
前記クラスタに加入するステップ後、前記登録テーブルをスクラブするステップをさらに含む、請求項１３に記載の方法。
【請求項１５】
請求項１４に記載の方法であって、前記登録テーブルをスクラブするステップは、
前記ストレージデバイスから前記登録テーブルを受信するステップと、
前記登録テーブルを読み取るステップと、
前記登録テーブルから１以上の登録キーを抹消するステップであって、１以上の登録キーのそれぞれは、前記クラスタ内の他のノードと関連している、ステップと、
を含む方法。
【請求項１６】
請求項１５に記載の方法であって、第２の登録タイマを設定するステップをさらに含み、前記第１の登録タイマの長さは、前記第２の登録タイマの長さの少なくとも２倍である、方法。
【請求項１７】
請求項１５に記載の方法であって、１以上の登録キーを抹消するステップは、関連する前記ノードが前記ストレージデバイスへのアクセスを拒否されるようにする、方法。
【請求項１８】
請求項１３に記載の方法であって、前記新規の登録キーは、クラスタ識別子、およびノード識別子を含む、方法。
【請求項１９】
前記第１の登録タイマは、前記ノードのクラスタ内の全てのノードが前記登録テーブルをスクラブするのに十分な時間期間で構成されている、請求項１に記載の方法。
【発明の詳細な説明】
【背景技術】
【０００１】
[0001]ノードのクラスタがストレージデバイスにアクセスできるという通常の共有ストレージ状況では、クラスタ内の少なくとも１つのノードが、ストレージデバイスに接続される。結果として、ストレージデバイスに接続されているノードは、ストレージデバイスの防御の対処を担う。しかしながら、クラスタが複数のストレージデバイスにアクセスできる状況では、クラスタ内の単一のノードを各ストレージデバイスに接続することはできない。結果として、ストレージデバイスの一部は、保護されない場合がある。
【発明の概要】
【発明が解決しようとする課題】
【０００２】
[0002]これらおよび他の概括的考慮事項に関して、諸実施形態は作成されている。また、比較的具体的な問題を論じているが、諸実施形態は、この背景技術で特定された具体的な問題を解決することに限定されるべきでないことを理解されたい。
【課題を解決するための手段】
【０００３】
[0003]この概要は、「発明を実施するための形態」の項でさらに後述する概念選択を簡略化された形態で紹介するために提供される。この概要は、特許請求される主題の主要な特徴または本質的な特徴を特定するように意図するものでも、特許請求される主題の範囲を決定する際の助けとして使用すべきように意図するものでもない。
【０００４】
[0004]本開示の実施形態は、ストレージデバイスへのアクセスを可能にし、クラスタ内の様々なノードによってアクセス可能な１つまたは複数のストレージデバイスを保護するための方法およびシステムを提供する。具体的には、１つまたは複数の実施形態は、ノードがいかにクラスタに加入承認され得、それによって、ノードのクラスタ内の少なくとも

１つのノードに接続されているストレージデバイスへの読取り／書込みアクセスを取得することができるかについて記載する。加えて、１つまたは複数の実施形態は、ノードがクラスタと関連する登録テーブルを監視し、認識されていないノードからのエントリを除去することが可能であることを提供する。除去されたそれらのノードに関しては、ノードは、クラスタに対する再承認を求めるために、登録テーブルに再登録しようと試みることができる。

【０００５】

[0005]以下に説明されるように、ストレージデバイスへのアクセスを求めるノードは、クラスタ通信プトロコルを使用して、クラスタ内に入る。ノードは、一旦、クラスタに加入承認されると、クラスタによって利用される１つまたは複数のストレージデバイスへのアクセスを得る資格を有することができる。ストレージデバイスへのアクセスを得るために、ストレージデバイスへのアクセスを求めるノードは、ストレージデバイスに登録キーを送信する。ストレージデバイスに登録後、ノードは、登録タイマを設定する。諸実施形態では、登録タイマは、要求されたアクセスがストレージデバイスへのアクセスを求めるノードに付与されるべきであるかどうかをクラスタ内の各ノードが決定する機会を得るその間の時間期間に相当する。登録タイマの期限切れと同時に、ストレージデバイスへのアクセスを求めるノードは、ストレージデバイスから登録テーブルを受信する。一旦、登録テーブルが受信されると、ノードは、その登録キーが登録テーブルに記憶されているかどうかを決定する。登録キーが登録テーブルに記憶されている場合、ノードは、ストレージデバイスへのアクセスが許可される。より具体的には、ノードには、ストレージデバイスへの書込みアクセスが付与される。

【０００６】

[0006]諸実施形態は、コンピュータ処理、コンピューティングシステムとして、またはコンピュータプログラム製品もしくはコンピュータ可読媒体などの製造品として実装可能である。コンピュータプログラム製品は、コンピュータシステムによって読取り可能で、コンピュータ処理を実行するためのコンピュータプログラム命令を符号化した、コンピュータストレージ媒体であってよい。コンピュータプログラム製品はまた、コンピューティングシステムによって読取り可能で、コンピュータ処理を実行するためのコンピュータプログラム命令を符号化した、搬送波上の伝播信号であってもよい。

【０００７】

[0007]限定的でなく、包括的でない実施形態は、以下の図を参照して説明される。

【図面の簡単な説明】

【０００８】

【図１】[0008]本開示の１つまたは複数の実施形態による、クラスタ内の複数のノードが、それぞれのストレージデバイスに接続されているシステムを示す図である。

【図２】[0009]本開示の１つまたは複数の実施形態による、クラスタにおけるメンバシップを要求するための方法を示す図である。

【図３】[0010]本開示の１つまたは複数の実施形態による、ノードのクラスタと関連するストレージデバイスへのアクセスを決定するための方法を示す図である。

【図４】[0011]本開示の１つまたは複数の実施形態による、クラスタにおける加入再承認を要求するための方法を示す図である。

【図５】[0012]本開示の１つまたは複数の実施形態による、クラスタ内の様々なノードがいかにして物理的ストレージデバイスに接続可能であるかを示すブロック図である。

【図６】[0013]本開示の１つまたは複数の実施形態とともに使用可能なコンピューティングデバイスの例示的な物理的構成要素を示すブロック図である。

【図７Ａ】[0014]本開示の１つまたは複数の実施形態とともに使用可能なモバイルコンピューティングデバイスの簡略化されたブロック図である。

【図７Ｂ】本開示の１つまたは複数の実施形態とともに使用可能なモバイルコンピューティングデバイスの簡略化されたブロック図である。

【図８】[0015]本開示の１つまたは複数の実施形態とともに使用可能な分散されたコンピ

ューティングシステムの簡略化されたブロック図である。
【発明を実施するための形態】
【０００９】
[0016]様々な実施形態が、その一部を成し、具体的な例示的実施形態を示す添付の図面を参照して、より十分に後述される。しかしながら、諸実施形態は、多数の種々の形態で実装可能であり、本明細書に示される実施形態に限定されるものと見なすべきでなく、そうではなくて、これらの実施形態は、本開示が、徹底的で完全になるように、および当業者に対して実施形態の範囲を十分に伝えることになるように提供される。諸実施形態は、方法、システム、またはデバイスとして実践可能である。したがって、実施形態は、ハードウェアの実装形態、全面的にソフトウェアの実装形態、またはソフトウェアの態様とハードウェアの態様とを組み合わせた実装形態の形態を取ることができる。そのため、以下の詳細な説明は、限定する意味で解釈すべきではない。
【００１０】
[0017]図１は、本開示の１つまたは複数の実施形態による、クラスタ１０２内の複数のノードが、それぞれのストレージデバイスに接続されているシステム１００を示している。図１に示すように、クラスタ１０２は、複数のノード１０２Ａ～１０２Ｄを含むことができる。４つのノードが示されているが、クラスタ１０２が、４つより多いノード、または４つより少ないノードを有する場合があることが企図される。特定の実施形態では、ノードが、例えば、パーソナルコンピュータ、タブレット、ラップトップ、スマートフォン、およびパーソナルデジタルアシスタントなど、コンピューティングデバイスであってよい。他の実施形態では、ノードが、サーバコンピューティングデバイスであってもよい。
【００１１】
[0018]図１はまた、クラスタ１０２内の各ノードが、１つまたは複数のストレージデバイスに接続されていることを示している。特定の実施形態では、ストレージデバイスは、ダイレクトアタッチドストレージデバイス（すなわち、ホストシステムまたはホストデバイスに直接、接続されているストレージデバイス）とすることができる。また、ストレージデバイスが、１つまたは複数のパスを使用して、クラスタ内のいくつかのノードによってアクセス可能であることも企図される。例えば、１つまたは複数のノードは、ストレージデバイスに物理的に接続可能である一方、クラスタ内の他のノードは、リモートパスを使用して、ストレージデバイスに接続することができる。加えて、単一のノードが、様々なストレージデバイスに対する複数の物理的接続と、様々なストレージデバイスに対する１つまたは複数のリモート接続とを有することができる。また、クラスタ内の各ノードは、クラスタ内の他のノードのそれぞれの活動および接続をビューでき得ることが企図される。つまり、システム１００は、一部のストレージデバイスが、一部のノードにとって利用可能である一方、他のストレージデバイスは、それらのノードにとって利用できないという点で非対称とすることができる。
【００１２】
[0019]例えば、図１に示すように、ノード１０２Ａとノード１０２Ｂは、ストレージデバイス１０４に接続され、ノード１０２Ｃは、ストレージデバイス１０４とストレージデバイス１０６に接続され、ノード１０２Ｄは、ストレージデバイス１０６とストレージデバイス１０８に接続されている。特定の実施形態では、ストレージデバイス１０４～１０８は、ストレージプールを備える。ストレージプール内の各ストレージデバイスにアクセスできる単一のノードがクラスタ１０２内に存在しないとき、クラスタ１０２内の各ノードは、防御アルゴリズムを動作させて、確実に、クラスタの一部であるノードだけがストレージデバイスに読取り／書込みアクセスできるようにする役割を担う。したがって、クラスタ１０２内の各ノードは、それらが、ストレージプールにおいて接続されているストレージデバイスを同時に保護する。
【００１３】
[0020]図１に戻って参照すると、ノード１０２Ａ、１０２Ｂ、および１０２Ｃのそれぞれは、ストレージデバイス１０４に接続されている。論じたように、ノードのそれぞれは

、ストレージデバイス１０４への物理的接続、またはリモート接続（すなわち、ストレージデバイス１０４への物理的接続を有するノードを通じたストレージデバイス１０４への接続）を有することができる。ノード１０２Ａ、１０２Ｂ、および１０２Ｃは、ストレージデバイス１０４に接続されているので、各ノードは、ストレージデバイス１０４に読取り／書込みアクセスできる。さらには、クラスタ１０２内のノード１０２Ａ、１０２Ｂ、および１０２Ｃのそれぞれは、クラスタ１０２内の他のノードの存在を検出し、クラスタ内の他のノードのそれぞれの活動を決定することができる。

【００１４】

[0021]諸実施形態では、特定のストレージデバイスに対する権利は、永続的な予約（persistent reservation）によって決定される。つまり、例えば、ストレージデバイス１０４などのストレージデバイスは、オフラインであるときでも、または再起動したときでも、特定のノードの予約を維持する。議論を進めるために、特定のノードの予約は、あるノードがある特定のストレージデバイスを予約したときに行われ、別の認可されていないノードがそのストレージデバイスにアクセスするのを防ぐ。

【００１５】

[0022]上記の例に戻って参照すると、ノード１０２Ａ、１０２Ｂ、および１０２Ｃのそれぞれは、クラスタ１０２の一部であるという理由から、ストレージデバイス１０４に読取り／書込みアクセスできる。下記に詳細に説明されることになるが、クラスタ１０２内の各ノードは、時間ｔにおいて防御アルゴリズムを動作させて、クラスタ１０２内の他のいずれのノードも、（ｉ）クラスタ内の他のノードに対する接続性か、または（ｉｉ）ストレージデバイス１０４に対する接続性を失っているかどうかを決定する。

【００１６】

[0023]例えば、ノード１０２Ａが、ノード１０２Ｂと１０２Ｃに対する接続性、またはストレージデバイス１０４に対する接続性を失う場合、ノード１０２Ａはもはや、ストレージデバイス１０４に（少なくとも）書込みアクセスできない、したがって、ストレージデバイスへのアクセスが禁止されるべきであることを、ノード１０２Ｂまたはノード１０２Ｃは独立に決定する。接続が失われている場合、ノード１０２Ｂおよび１０２Ｃは、ノード１０２Ａのワークロードを取り、また、ストレージデバイスへの書込みをノード１０２Ａに許可することにより、ストレージデバイス１０４上のデータを破損させる可能性があると、ノード１０２Ａがもはや、ストレージデバイス１０４への書込みを確実にできないようにするステップを取る。ノード１０２Ａは、ノード１０２Ｂと１０２Ｃに対する接続性を失っている場合はあるものの、ストレージデバイス１０４に対する接続性をなおも有することは可能であることが企図される。同様に、ノード１０２Ａは、ストレージデバイス１０４に対する接続性を失った場合であっても、なおもノード１０２Ｂおよび／またはノード１０２Ｃに接続可能であることが企図される。

【００１７】

[0024]上記の例に戻ると、ノード１０２Ａがストレージデバイス１０４に書き込むことを禁止するためには、ノード１０２Ｂまたは１０２Ｃは、ノード１０２Ａをノード登録テーブルから抹消する要求をストレージデバイス１０４に送信する。より具体的には、要求は、ノード１０２Ａと関連する登録キーをノード登録テーブルから抹消するように、ストレージデバイスに送信される。結果として、ストレージデバイス１０４はもはや、ノード１０２Ａと関連する物理パス、またはリモートパスから書込みコマンドを受け付けなくなる。特定の実施形態では、ノード１０２Ａは、ストレージデバイス１０４に書込みアクセスできなくなるものの、なおもストレージデバイス１０４への読取りアクセスはできることになる。

【００１８】

[0025]特定の実施形態では、ノード１０２Ｂまたは１０２Ｃはいずれも、互いに独立に抹消要求を送信することができる。別の実施形態では、クラスタのノードは、特定のノードが抹消されるべきであるかどうかに関して合意に達するように求められる場合がある。さらなる別の実施形態では、ノードがクラスタから除去されるべきであることをノード自

体が決定することができる。例えば、ノード１０２Ａは、ノード１０２Ａが、他のノードのうちの１つに対する接続、またはストレージデバイス１０４に対する接続を失っていると決定する場合、それ自体からストレージデバイスへの１つまたは複数のパスを除去し、あるいはノード登録テーブルからその登録キーを除去するようにストレージデバイス１０４に命令することができる。

【００１９】

[0026]ノード登録テーブルを参照すると、特定の実施形態では、ノード登録テーブルは、ストレージデバイス１０４によって維持され、ストレージデバイス１０４に対して書込みアクセスできるノードをリストする。諸実施形態では、ノード登録テーブルは、ストレージデバイスに対して書込みアクセスできるノードごとに登録キーを含む。特定の実施形態では、登録キーは、以下の形式、すなわち、（ｉ）（クラスタ内のすべてのノードについて同じである）クラスタグローバル一意識別子の３２ビットハッシュ、（ｉｉ）８ビットキーリビジョン(key revision)、（ｉｉｉ）８ビットノード番号、および（ｉｖ）１６ビットシグネチャを有する６４ビット整数を含む。登録キーの具体的なサイズおよび構成が示されているが、登録キーは、登録キーが各ノードに固有である限り、任意の数のビットを有しても、また様々な構成を有してもよいことが企図される。

【００２０】

[0027]以下により詳細に説明されることになるように、ノードは、一旦、抹消されると、クラスタへの加入再承認を要求することができる。加入再承認を要求するために、抹消されたノードは、更新済みの登録キーをストレージデバイスに送信することができる。一旦、ノードがストレージデバイスに再登録されると、クラスタ内の他のノードのそれぞれは、ノードがクラスタに加入再承認されるべきであるかどうかに関しての決定を行う。クラスタ内のノードの決定に基づいて、加入再承認を求めるノードには、加入再承認が付与され得る、または加入再承認が拒否され得る。諸実施形態では、クラスタ内のノードは、加入再承認を求めるノードの接続速度、加入再承認を求めるノードの信頼度、および加入再承認を求めるノードが、ストレージプール内の他のストレージデバイスに対してできるアクセスなどを含む任意の数の要因に基づいて、その決定を行うことができる。

【００２１】

[0028]図２は、本開示の１つまたは複数の実施形態による、クラスタにおけるメンバシップを要求するための方法２００を示している。特定の実施形態では、方法２００は、図１のクラスタ１０２などのクラスタにおけるメンバシップを要求するノードによって使用され得る。上記に論じたように、ノードは、一旦、クラスタに加入承認されると、その特定のストレージデバイスに対して読取り／書込みアクセスでき得る。

【００２２】

[0029]具体的には、加入するノードは、アクティブなクラスタへの加入承認を得るために、クラスタ通信プロトコルを使用して他のノードと通信しようと試みることができる。この場合、加入するノードは、一旦、アクティブ状態に入ると、方法２００を実行して、ストレージへのアクセスを得ることになる。加入するノードが、クラスタプロトコルを介して、他の加入するノード、またはアクティブなノードと通信するのに失敗し、アクティブなクラスタは存在しない場合もあり得ると考える場合、加入するノードは、方法２００を実行して、ストレージへのアクセスを得て、したがって、第１のアクティブなノードになることができる。議論を進めるために、ノードは、クラスタへのアクセスを要求し、加入承認を得るとき、アクティブなノードとして、またはアクティブ状態に入っていると見られる。例えば、クラスタ通信プロトコルを動作中であり、現在、クラスタメンバシップに参加中であるすべてのノードは、アクティブなノードと見なされる。加えて、クラスタと関連する１つまたは複数のストレージデバイスにアクセスできるノードは、ストレージノードと見なされる。諸実施形態では、ストレージノードは、アクティブなノードセットのサブセットである。

【００２３】

[0030]図２に示すように、方法２００は、ノードが、クラスタ通信プロトコルを使用し

て、クラスタに加入承認された後に開始する。一旦、クラスタに加入承認されると、１つまたは複数の実施形態は、登録キーを使用してストレージデバイスに登録すること（２１０）によって、ノードが、例えば、ストレージデバイス１０４（図１）など、クラスタと関連する１つまたは複数のストレージデバイスにアクセスするように求めることを実現する。論じたように、登録キーは、上記に論じた様々な構成要素を有する６４ビット整数を含むことができる。

【００２４】

[0031]一旦、登録キーがストレージデバイスに送信されると、フローは動作２２０に進み、ここで、登録タイマが設定される。特定の実施形態では、この登録タイマは、ストレージデバイスへのアクセス、またはストレージプールへのアクセスを要求しているノードによって維持され得る。しかしながら、ストレージデバイスまたはクラスタ内の別のノードもまた、登録タイマを維持することが可能であり得ることが企図される。諸実施形態では、登録タイマの長さは、時間期間ｔに相当する。特定の実施形態では、時間期間ｔは、３秒である。より具体的には、時間期間ｔは、クラスタ内の他のいずれのノードも、（ＣＰＵロード、およびＩ／Ｏ待ち時間などに起因して生じ得るいずれの遅延も考慮して）３秒毎に行われるべきスクラブを行うための時間に相当する。

【００２５】

[0032]タイマの期限切れと同時に、フローは動作２３０に進み、ここで、ノード登録テーブルは、ストレージデバイスから読み取られる。上記に論じたように、ノード登録テーブルは、ストレージデバイスによって（またはストレージプール内の少なくとも１つのストレージデバイスによって）維持され、クラスタ内のあらゆるノードと関連する各登録キーのリストを含んでいる。

【００２６】

[0033]一旦、登録テーブルが受信され、読み取られると、フローは動作２４０に進み、ここで、ノードの登録キーがノード登録テーブルに含まれているかどうかが決定される。ノードの登録キーがノード登録テーブルに含まれている場合、クラスタ内のノードのそれぞれは、防御アルゴリズムを動作させており、ストレージデバイスは、ストレージデバイスへのアクセスを求めるノードの要求を受け付けている。結果として、フローは動作２５０に進み、ここで、ノードは、ストレージデバイスにアクセスするように、より具体的には、ストレージデバイスに書込みアクセスできるように許可される。

【００２７】

[0034]しかしながら、要求するノードの登録キーが、ノード登録テーブル内に存在しないことが動作２４０で決定される場合、フローは動作２１０に戻るように進み、ノードは、ストレージデバイスに登録しようと２回目の試みを行う。方法は、繰り返し、ストレージデバイスへのアクセスを要求するノードは、再度、登録テーブルを要求し、読み取って、その登録キーがノード登録テーブルに記憶されているかどうかを決定する。

【００２８】

[0035]図３は、本開示の１つまたは複数の実施形態による、ノードのクラスタと関連するストレージデバイスへのアクセスを決定するための方法３００を示している。特定の実施形態では、方法３００は、アクティブなノードとして見られるクラスタ内の各ノード（すなわち、クラスタ通信プロトコルを動作中であり、現在、クラスタメンバシップに加入中であるすべてのノード）によって行われる。方法３００はまた、ストレージノードと見なされるノード（すなわち、クラスタと関連する１つまたは複数のストレージデバイスにアクセスできるいずれのノード）によっても行われ得る。

【００２９】

[0036]方法３００は、ノードが、ノード登録テーブルからの登録キーを「スクラブする（scrub）」（３１０）ときに開始する。具体的には、ノードは、アクティブなクラスタの一部ではない他のノードを探す。登録キーがディスク登録テーブルからスクラブされるとき、スクラブされた登録キーと関連するノードはもはや、クラスタと関連する特定のストレージデバイスまたはストレージプールに、（少なくとも）書込みアクセスできない。

特定の実施形態では、クラスタ内の１つのノードが、クラスタ内の別のノードはもはや、特定のストレージデバイスまたはストレージプールに書込みアクセスできないはずであると考えるとき、登録キーがスクラブされる。これは、ノードがクラスタ内の別のノードへの接続を失うこと、ノードがクラスタからそれ自体を除去すること、またはノードからストレージデバイスへの接続を失うことのうちの１つの結果とすることができる。ノードが、それ自体をクラスタから除去している状況では、そのノードは、その登録キーがノード登録テーブルから除去されるべきであるということを示す要求をストレージデバイスに送信することができる。別の実施形態では、クラスタ内の他のノードのうちの１つは、スクラビング処理中、ノードがクラスタから抹消される（すなわち、除去される）ことを要求することができる。特定の実施形態では、ノードがクラスタから抹消される場合、他のノードは、除去されたノードからのコマンドがストレージデバイスに達するのを防ぐように構成可能である。

【００３０】

[0037]図３に示すように、スクラビング処理は、様々なサブ動作を有する。スクラビング処理は、サブ動作３１１で開始し、ここで、ノードは、ストレージデバイスによって維持されるノード登録テーブルを読み取る。上記に論じたように、ノード登録テーブルは、クラスタ内の各ノードと関連する登録キーすべてのリストを含んでいる。

【００３１】

[0038]次いで、フローはサブ動作３１２に進み、ここで、クラスタ内のアクティブなメンバシップを有していない１つまたは複数のノードが、クラスタから抹消される。諸実施形態では、クラスタ内の各ノードは、クラスタ内の他のあらゆるノードによってビューでき、リモート接続、または物理接続のいずれかによって、１つまたは複数のストレージデバイスに接続され得る。クラスタ内の各ノードが、クラスタ内の他のあらゆるノードのビューを有すると、ノード登録テーブルを読取り中であるノードは、クラスタ内のどのノードが、ノード登録テーブル内の関連登録キー有しているのかを決定することができる。したがって、ノードは、ストレージデバイスから受信した登録テーブルをスクラブする。登録キーはテーブル内にあるが、ノードがアクティブでない場合、ノードは抹消される。

【００３２】

[0039]特定の実施形態では、ノードは、クラスタ内の複数のノードが類似の決定（すなわち、抹消されようとしているノードはノード登録テーブルに登録キーを有していない）に達するまで、抹消され得ない。他の実施形態では、抹消されるべきノードはノード登録テーブルに関連登録キーを有していないという決定に単一のノードが達するとき、ノードが、抹消され得る。

【００３３】

[0040]一旦、ノードが抹消されるように、要求が送信されると、アルゴリズムを動作させるノードは、ノードが成功裏に抹消されたかどうかを決定する（３１３）。ノードが成功裏に抹消されなかった場合、アルゴリズムを動作させるノードは、自己チェックを行って、それ自体の登録キーがノード登録テーブル内に存在するかどうかを決定する。それ自体の登録キーがノード登録テーブル内に存在しない場合、フローはサブ動作３１４に進み、ノードは、「ゲスト状態（ｇｕｅｓｔ　ｓｔａｔｅ）」に入り、１つまたは複数のクラスタプロトコルを使用して、クラスタへの加入再承認を求める。ノードが成功裏に抹消された場合、フローは動作３１５に進み、ストレージデバイスがまだ予約されていない場合、ノードは、ストレージデバイスを予約する。つまり、ノードは、ストレージデバイスの所有権を取り、次いで、このストレージデバイスへのアクセスを、同じクラスタ内にある他のノードと共有する。

【００３４】

[0041]ノード登録テーブル内の登録キーがスクラブ済みである場合、フローは動作３２に進み、ここで、第２のタイマが設定される。諸実施形態では、第２のタイマは、図２で論じた登録タイマに比例する。例えば、スクラビングアルゴリズムは、すべてのノードにおいて並行して動作する（例えば、各ノードは、すべての他のノードと独立に３秒毎にキ

ーをスクラブする。特定の実施形態では、スクラビングアルゴリズムは、別のノードもやはり、登録テーブルをスクラブしているかどうか、またはノードのうちの１つがストレージへのその接続を失っており、そのため、登録テーブルをスクラブすることができないかどうかを１つのノードが信頼可能に見分けることができない可能性があるという理由から、並行して動作する。

【００３５】

[0042]ノードは、クラスタに加入するとき、クラスタに登録し、加入承認されるようになるのを待つ。ノードが加入承認される場合、１つの実施形態は、できる限りすぐにノードが、ディスクをスクラブし、明らかにする（surface）ことを実現する。３秒スクラビングタイマが設定され、ノードは、ストレージ上のそれ自体の登録を認証することができる。

【００３６】

[0043]タイマの期限切れと同時に、フローは動作３３０に進み、ここで、ノードは、ストレージデバイスへの１つまたは複数のパスを明らかにする。つまり、ノードは、物理パス、またはリモートパスのいずれのパスが、クラスタ内の他のノードおよび／またはストレージデバイスに接続される必要があるかを決定することができる。例えば、図１を参照すると、ノード１０２Ｂが、クラスタ１０２に加入し、ノード１０２Ａが、ストレージデバイス１０４への物理接続を確立した場合、または有する場合、物理パスは、クラスタ内の他のノードにアドバタイズされ、クラスタ内の他のノード、例えば、１０２Ｄは、ノード１０２Ａに対するリモートパスを確立し、ノード１０２Ａとストレージデバイス１０４との間の物理接続を利用することができる。

【００３７】

[0044]図４は、本開示の１つまたは複数の実施形態による、クラスタへの加入再承認を要求するための方法４００を示している。特定の実施形態では、クラスタへの加入再承認を求めるノードは、それ自体、クラスタから除去されている可能性があり、または図３に関して上記に論じたように、クラスタ内の別のノードによって除去されている可能性もある。諸実施形態では、方法４００は、クラスタ内のノードが、ストレージデバイスへの書込みコマンドを送信し、書込みが成功裏でなかったことを通知されるときに開始する。書込みが成功裏でなかった場合、ノードは、コマンドが、異なるパスの下方に送信されることを要求することができる。加えて、または代替として、ノードは、進行中である他のすべての保留になっているコマンドがキャンセルされることも要求することができる。通知を受信すると同時に、ノードは、ストレージデバイスにノード登録テーブルを要求する（４１０）。上記に論じたように、ノード登録テーブルは、ストレージデバイスによって維持され、クラスタ内のそれぞれのノードと関連している様々な登録キーを含む。

【００３８】

[0045]ノード登録テーブルが、要求しているノードによって受信されたとき、ノードは、登録テーブルを読み取って（４２０）、それ自体の登録キーが登録テーブルに含まれているかどうかを決定する。ノードの登録キーが、ノード登録テーブルに含まれていないことが決定されるとき、ノードは、別の登録キーを使用して、ストレージデバイスに登録する（４３０）。上記に論じたように、登録キーは、クラスタグローバル一意識別子の３２ビットハッシュ、８ビットキーリビジョン、８ビットノード番号、および１６ビットシグネチャを有する６４ビット整数とすることができる。特定の実施形態では、ノードは、再登録しなくてはならないとき、新規の登録キーを生成することができ、その場合、登録キーの少なくとも一部分がインクリメントされ、または変更される。したがって、ノード、ストレージデバイス、またはクラスタ内の他のノードは、新規のパスがノードのために設定されなくてはならない回数を追跡することができる。いくつかの実施形態では、ノードが、加入再承認を要求した回数は、クラスタに対する加入再承認を取得するノードに影響をもたらし得る。加えて、登録キーに対する変更は、（ノードがクラスタに加入再承認されるときに）ノードからおよび／またはノードと関連するパスから生じる書込みコマンドが、ストレージデバイスに書き込まれるのをなおも待っている可能性がある、ノードから

の古くなった書込みコマンドと確実に区別できるようにすることの助けにもなる。

【００３９】

[0046]一旦、ノードが、その新規に生成された登録キーをストレージデバイスに登録すると、登録タイマが設定される（４４０）。先に論じたように、登録タイマの長さは、時間期間ｔの約２．５倍に相当する。つまり、時間期間ｔは、クラスタ内の各ノードが、本明細書に開示される防御アルゴリズムを動作させるのにかかる時間に相当する。

【００４０】

[0047]タイマの期限切れと同時に、フローは動作４５０に進み、ストレージデバイスへの書込みアクセスを求めるノードは、ストレージデバイスによって維持される登録テーブルを読み取る。新規に生成された登録キーが、ノード登録テーブルに存在することが決定される場合（４６０）、ノードには、ストレージデバイスへの書込みアクセスが付与される。

【００４１】

[0048]しかしながら、動作４６０で、ノードの新規に生成された登録キーが、ノード登録テーブルに存在しないことが決定される場合、フローは動作４３０に戻って進み、ノードは再登録し、登録タイマはリセットされる。特定の実施形態では、ノードが、決められた回数、拒否された場合、ノードはもはや、ストレージデバイスへの書込みアクセスを求めなくなる。特定の実施形態では、ノードは、決められた時間期間が経過した後に、再度、クラスタに対する加入承認、またはストレージデバイスへのアクセスを求めることができる。

【００４２】

[0049]図５は、本開示の１つまたは複数の実施形態による、クラスタ内の２つのノードが、リモートパスおよび／または物理パスを使用して、物理ディスクにそれによってアクセスすることが可能なシステム５００を示すブロック図である。図１から図４に関して上記に論じたように、クラスタ内の様々なノードは、ストレージプール内の１つまたは複数のストレージデバイスに接続可能である。それらの接続（またはスピンドル）は、物理接続またはリモート接続とすることができる。以下に説明されるように、ノードは、様々なパスを利用して、１つまたは複数のストレージデバイスに接続することができる。

【００４３】

[0050]図５に示される例示的な実施形態では、クラスタは、２つのノード、Ｎｏｄｅ（ノード）Ａの５１０およびＮｏｄｅＢの５２０を有することができる。各ノードは、物理Ｄｉｓｋ５３０への物理接続を有することができる。２つのノードのみが示されているが、クラスタは、３つ以上のノードで構成されてよいことが企図される。加えて、各ノードは、１つまたは複数のストレージデバイスへの物理接続を有することができることも企図される。図５に示すように、ノードは、いくつかの異なるパスを通じて物理ディスクにアクセスでき得る。例えば、ＮｏｄｅＡの５１０は、物理ディスク５３０への物理パスを有し、また、ＮｏｄｅＢの５２０のターゲット５２４を経由して物理ディスク５３０へのリモートパスも有する。特定の実施形態では、単一のノードが、同じディスクへの複数の物理パスを有することができる。そのような実施形態では、ノードは、クラスタ内の他のノードのすべてに対してすべてのこれらの物理パスを曝すことになる。

【００４４】

[0051]また、図５に示すように、ＮｏｄｅＡは、仮想ディスク５１１と、物理ディスク５３０への複数の物理パスおよびリモートパスを統合するマルチパスオブジェクト５１２と、例えば、ＮｏｄｅＢの５２０など、別のノードを通じて、物理ディスク５３０への１つまたは複数のリモートパスをインスタンス化するリモートパスオブジェクト５１３と、ＮｏｄｅＡの５１０と物理ディスク５３０との間の１つまたは複数の物理接続をアドバタイズするように働き、クラスタ内の他のノードがノードＡの５１０を介して物理ディスク５３０に対する接続性を（例えば、リモートパスを通じて）得ることを可能にするターゲットオブジェクト５１４にアクセスでき、物理パスオブジェクト５１５が、物理ディスク５３０への１つまたは複数の物理接続もしくはパスをインスタンス化する。

【００４５】
[0052]同様に、ＮｏｄｅＢの５２０は、仮想ディスク５２１と、ＮｏｄｅＢの５２０から物理ディスク５３０への複数の物理パスおよびリモートパスを統合するマルチパスオブジェクト５２２と、例えば、ＮｏｄｅＡの５１０など、別のノードを通じて、物理ディスク５３０への１つまたは複数のリモートパスをインスタンス化するリモートパスオブジェクト５２３その５１３と、クラスタ内の他のノードに物理ディスク５３０への物理パスをアドバタイズするターゲットオブジェクト５２４と、物理ディスク５３０への１つまたは複数の物理接続もしくはパスをインスタンス化する物理パスオブジェクト５２５にアクセスできる。ＮｏｄｅＡの５１０とＮｏｄｅＢの５２０の両方に１つのリモートパスが示されているが、単一のノードが複数のリモートパスを有し得ることが企図される。また、各ノードが複数の物理パスを有し得ることも企図される。
【００４６】
[0053]諸実施形態では、様々なコマンドが物理ディスク５３０に送信されるのに経由する好ましいパスは、物理パスである。例えば、新規のディスクが検出されるとき、クラスタの１つまたは複数のノードは、ディスクを登録し、または予約することになる。上記に論じたように、この処理は、図２に関して上記に論じた防御アルゴリズムを動作させること、続いて、ノードから物理ディスクへの物理パスを創出することを含む。特定の実施形態では、各物理パス、または物理パスの新規の各インスタンスは、クラスタ識別子、ノード識別子、および再生識別子（物理パスがインスタンス化されるたびにインクリメントされる物理パスの固有の番号）を含む登録キーを有する。諸実施形態では、パスの登録キーは、関連ノードの登録キーに相当してよい。一旦、物理接続が確立され、ノードが登録キーを使用してディスクに登録されると、ノードのマルチパスオブジェクトおよびターゲットオブジェクトは、新規に確立された物理パスについて通知される。次いで、その情報が、クラスタ内の他のノードに伝送され、したがって、他のノードは、物理ディスクへの物理接続を有するノードのターゲットを介してリモートパスを確立することができる。
【００４７】
[0054]上記に論じたように、１つまたは複数のノードが、クラスタ内の１つまたは複数の他のノードへの、あるいは物理ディスクへの接続を失う場合があることが企図される。そのような事象においては、クラスタ内の接続されたノードのうちの１つは、接続解除されたノードからの１つまたは複数のパスが、除去されることを要求し、また、ストレージデバイスが、接続解除されたノードと関連する１つまたは複数のパス（例えば、物理パスもしくはリモートパス）からの書込み要求を取るのを中止することも要求することになる。同様に、接続解除されたノードへのリモート接続を有する各ノードと関連するターゲットもまた、接続解除されたノードからのコマンドを受け取るのを中止することができる。そのようなアクションにより、接続解除されたノードは、通信線上にはあるが、まだ完了していない可能性があるストレージデバイスに対する追加のおよび／または二重の書込みを送信しないようになる。つまり、登録キーをストレージデバイスから除去すること、およびターゲットを通じて書込みコマンドをブロックすることは、接続解除されたノードが、物理パスまたはリモートパスを使用して、ディスクに書き込むことが確実にできないようにすることの助けになる。
【００４８】
[0055]例えば、図５を参照すると、ＮｏｄｅＡの５１０は、その物理パスオブジェクト５１５を介して物理ディスク５３０へのその物理接続を失っている場合がある。しかしながら、図示するように、ＮｏｄｅＡの５１０はまた、ＮｏｄｅＢの５２０のターゲットオブジェクト５２４を通じた物理ディスク５３０へのリモートパス５１３も有する。加えて、物理ディスク５３０への接続を失う前に、ＮｏｄｅＡの５１０は、まだ完了していない物理ディスク５３０への書込みコマンドを送信していた場合もある。ＮｏｄｅＡの５１０は、物理ディスク５３０への接続性を失ったとき、その書込みコマンドが実行されたのか、または拒否されたのかについての知識が全くない場合がある。
【００４９】

[0056]しかしながら、ＮｏｄｅＡの５１０が、物理ディスク５３０とすぐに再接続することが許容され、実行された可能性がある、もしくは実行されていない可能性があるコマンドをいずれも再提出することが許容された場合、またはＮｏｄｅＡの５１０が、物理ディスク５３０に対する追加のコマンド（ＮｏｄｅＡの５１０がその接続を失ったことにより、順序から外れてもよい）を送信することが許可された場合、そのようなアクションは、物理ディスク５３０内のデータを破損させることになる可能性がある。そのような破損を防ぐために、ＮｏｄｅＢの５２０は、ＮｏｄｅＡの５１０と関連する物理パスおよび／またはすべてのリモートパスをプリエンプト（preempt）する。

【００５０】

[0057]一旦、ＮｏｄｅＡの５１０と関連する物理パスおよび／またはリモートパスがプリエンプトされると、物理ディスク５３０は、ＮｏｄｅＡの５１０と関連するパスからのコマンドを受け付けなくなる。各ノードについての各パスが関連識別子を有すると、物理ディスク５３０は、どのコマンドが、パスのそれぞれの識別子に基づいてノードと関連付けられているかを決定することができる。特定の実施形態では、物理ディスク５３０は、物理パス間を区別する。したがって、物理ディスクの視点から見れば、Ｉ／Ｏがリモートパスを経由して来た場合、リモートパスが接続されているターゲットをホストするノードからＩ／Ｏが来たかのように見えることになる。要するに、リモートパスＩ／Ｏフェンシングが、ターゲットにおいて実行される一方、物理パスＩ／Ｏフェンシングは、物理ディスク５３０レベルにおいて実行される。

【００５１】

[0058]例を進めるために、クラスタ内の各ノードは、他のあらゆるノードの各スピンドルまたはパスを見ることができる。したがって、ＮｏｄｅＢの５２０は、ＮｏｄｅＡの５１０が、物理ディスク５３０への接続を失ったことが分かり得る。結果として、ＮｏｄｅＢの５２０は、リモートパス５２３を破棄することになる。しかしながら、ＮｏｄｅＡの５１０がクラスタ内の他のノードと通信することができない場合、ＮｏｄｅＢの５２０は、物理ディスク５３０に、ＮｏｄｅＡの５１０からの書込みコマンドを拒否するように命令することができる。

【００５２】

[0059]特定の実施形態では、一旦、物理ディスク５３０が、ＮｏｄｅＡの５１０の物理パスからのコマンドの拒否を開始すると、ＮｏｄｅＡの５１０のマルチパスオブジェクト５１２は、コマンドの拒否を検出する。結果として、マルチパスオブジェクト５１２は、すべての他の既存の物理パスに、いずれかが有効であるかどうかを決定するようにクエリすることができる。１つの物理パスがなおも、有効である場合、有効物理パスは、マルチパスオブジェクト５１２に追加される。しかしながら、有効な物理パスオブジェクトが全くない場合、新規のマルチパスオブジェクトが創出され、物理パスオブジェクト５１５は、新規の登録キーを含む新規の物理パスをインスタンス化する。新規の物理パスおよびその関連登録キーは、生成されると、古い物理パスと関連した今や機能していない（defunct）識別子とは別に設定する新規の再生識別子を有することになる。

【００５３】

[0060]加えて、例えば、ＮｏｄｅＡの５１０などのノードが、新規の識別子を使用して、クラスタに加入再承認を要求するとき、新規の識別子は、クラスタ内の他のノードにアドバタイズされる。したがって、他のノードのリモートパスオブジェクトは、ＮｏｄｅＡの５１０の物理パスの新規の識別子を使用して、物理ディスク５３０に接続することができる。上記に論じたように、物理ディスク５３０は、古い物理パスからのコマンドを受け付けないことを知ると、新規の物理パスからのコマンドおよびその関連識別子を受け付け、そのとき、ＮｏｄｅＡの５１０が、図２～図４に関して上述した方法によりクラスタに加入再承認を求める。

【００５４】

[0061]図５に戻って参照すると、ＮｏｄｅＡの５１０およびＮｏｄｅＢの５２０は、互いへの接続性を失う場合、仮想ディスクに書き込まれていないアプリケーションのキャッ

シュにデータが存在することも、または物理ディスクに書き込まれていない仮想ディスクにデータが存在することもある。したがって、諸実施形態は、接続解除されたノードからのパス上の残存するすべてのコマンドを消失させる（drain）こと、および接続解除されたノードと関連するパスからのさらなるコマンドが受け付けられないことを実現する。

【００５５】

[0062]本明細書に記載の実施形態および機能性は、デスクトップコンピュータシステム、ワイヤードおよびワイヤレスのコンピューティングシステム、モバイルコンピューティングシステム（例えば、モバイル電話、ネットブック、タブレットタイプまたはスレートタイプのコンピュータ、ノートブックコンピュータ、およびラップトップコンピュータ）、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のまたはプログラマブルの家庭用電化製品、ミニコンピュータ、ならびにメインフレームコンピュータを限定することなく含めた数多くのコンピューティングシステムによって動作することができる。

【００５６】

[0063]加えて、本明細書に記載の実施形態および機能性は、分散されたシステム（例えば、クラウドベースのコンピューティングシステム）を通じて動作することができ、その場合、アプリケーション機能性、メモリ、データの記憶および検索ならびに様々な処理機能は、Ｉｎｔｅｒｎｅｔ、またはイントラネットなど、分散されたコンピューティングネットワークを通じて互いとリモートに動作可能である。様々なタイプのユーザインターフェイスおよび情報は、オンボードコンピューティングデバイス表示部を介して、あるいは１つまたは複数のコンピューティングデバイスと関連するリモート表示装置を介して表示可能である。例えば、様々なタイプのユーザインターフェイスおよび情報は、様々なタイプのユーザインターフェイスおよび情報が投影される壁面に表示することができ、その壁面と相互作用することができる。本発明の実施形態がそれにより実践可能な数多くのコンピューティングシステムとの相互作用は、キーストローク入力、タッチ画面入力、音声または他の音響入力、およびジェスチャ入力などを含み、このジェスチャ入力では、関連コンピューティングデバイスに、コンピューティングデバイスの機能性を制御するようにユーザジェスチャを取り込み、解釈するための検出（例えば、カメラ）機能性が備わっている。

【００５７】

[0064]図６～図８および関連の記載では、本発明の実施形態がその中で実践可能である種々の動作環境についての議論が行われる。しかしながら、図６～図８に関して図示され、論じられるデバイスおよびシステムは、例示および図示の目的のためであり、本明細書に記載される本発明の実施形態を実践するのに使用可能な膨大な数のコンピューティングデバイス構成を限定するものではない。

【００５８】

[0065]図６は、本発明の実施形態がそれにより実践可能なコンピューティングデバイス１０５の物理的構成要素（すなわち、ハードウェア）を示すブロック図である。後述のコンピューティングデバイスの構成要素は、上記に記載のノードまたはコンピューティングデバイスに適し得る。基本構成では、コンピューティングデバイス１０５は、少なくとも１つの処理装置６０２、およびシステムメモリ６０４を含むことができる。コンピューティングデバイスの構成およびタイプに応じて、システムメモリ６０４は、これらに限定されないが、揮発性ストレージ（例えば、ランダムアクセスメモリ）、不揮発性ストレージ（例えば、読取り専用メモリ）、フラッシュメモリ、またはそのようなメモリの任意の組合せを含むことができる。システムメモリ６０４は、オペレーティングシステム６０５と、ソフトウェアの様々なアプリケーション６２０を動作させるのに適している１つまたは複数のプログラムモジュール６０６とを含むことができる。オペレーティングシステム６０５は、例えば、コンピューティングデバイス１０５の動作を制御するのに適し得る。さらには、本発明の実施形態は、グラフィックスライブラリ、他のオペレーティングシステム、または任意の他のアプリケーションプログラムと併せて実践可能であり、いずれかの

特定のアプリケーションまたはシステムに限定されない。この基本構成は、図６に、破線６０８内のそれらの構成要素によって示されている。コンピューティングデバイス１０５は、追加の特徴または機能性を有することができる。例えば、コンピューティングデバイス１０５はまた、例えば、磁気ディスク、光ディスク、またはテープなどの追加のデータストレージデバイス（リムーバブルおよび／または非リムーバブル）を含むことができる。そのような追加のストレージは、図６に、リムーバブルストレージデバイス６０９および非リムーバブルストレージデバイス６１０によって示されている。

【００５９】

[0066]上記のように、いくつかのプログラムモジュールおよびデータファイルが、システムメモリ６０４に記憶され得る。処理装置６０２において実行する一方、プログラムモジュール６０６は、これらに限定されないが、図１～図４に示す方法の段階のうちの１つまたは複数を含む処理を行うことができる。本発明の実施形態により使用可能な他のプログラムモジュールには、電子メール、および連絡先アプリケーション、ワードプロセシングアプリケーション、スプレッドシートアプリケーション、データベースアプリケーション、スライドプレゼンテーションアプリケーション、描画またはコンピュータ支援アプリケーションのプログラムなどを含めることができる。

【００６０】

[0067]さらには、本発明の実施形態は、個別の電子要素、論理ゲートを含んだパッケージ化もしくは集積化された電子チップ、マイクロプロセッサを利用する回路、または電子要素もしくはマイクロプロセッサを含んだ単一のチップを含む電気回路で実践可能である。例えば、本発明の実施形態は、システムオンチップ（ｓｙｓｔｅｍ－ｏｎ－ａ－ｃｈｉｐ：ＳＯＣ）により実践可能であり、その場合、図６に示す構成要素のそれぞれ、または多くが、単一の集積回路上に集積化可能である。そのようなＳＯＣデバイスは、１つまたは複数の処理装置、グラフィックス装置、通信装置、システム仮想化装置、および様々なアプリケーション機能性を含むことができ、それらのすべては、単一の集積回路としてチップ基板上に集積化される（または焼き込まれる）。本明細書に記載される機能性は、ＳＯＣにより動作するとき、単一の集積回路（チップ）上のコンピューティングデバイス１０５の他の構成要素とともに集積化されるアプリケーションに固有の論理機構により動作可能である。本発明の実施形態はまた、これらに限定されないが、機械技術、光技術、流体技術、および量子技術を含んだ、例えば、ＡＮＤ、ＯＲ、およびＮＯＴなどの論理演算を行うことができる他の技術を使用しても実践可能である。加えて、本発明の実施形態は、汎用のコンピュータ内で、または任意の他の回路もしくはシステムにおいて実践可能である。

【００６１】

[0068]コンピューティングデバイス１０５はまた、キーバード、マウス、ペン、音入力デバイス、タッチ入力デバイスなどの１つまたは複数の入力デバイス６１２を有することができる。表示部、スピーカ、プリンタなどの出力デバイス（複数可）６１４もまた、含められ得る。上述のデバイスは、例であり、他のデバイスが使用されてもよい。コンピューティングデバイス１０４は、他のコンピューティングデバイス６１８との通信を可能にする１つまたは複数の通信接続部６１６を含むことができる。適切な通信接続部６１６の例は、これらに限定されないが、ＲＦ伝送器、受信器、および／または送受信器回路、ユニバーサルシリアルバス（ｕｎｉｖｅｒｓａｌ　ｓｅｒｉａｌ　ｂｕｓ：ＵＳＢ）、パラレルポート、および／またはシリアルポートを含む。

【００６２】

[0069]本明細書に使用される用語コンピュータ可読媒体は、コンピュータストレージ媒体を含むことができる。コンピュータストレージ媒体には、コンピュータ可読命令、データ構造、またはプログラムモジュールなど、情報を記憶するための任意の方法または技術で実装される揮発性と不揮発性の、リムーバブルと非リムーバブルの媒体を含めることができる。システムメモリ６０４、リムーバブルストレージデバイス６０９、および非リムーバブルストレージデバイス６１０はすべて、コンピュータストレージ媒体の例（すなわ

ち、メモリストレージ）である。コンピュータストレージ媒体には、ＲＡＭ、ＲＯＭ、電気的に消去可能な読取り専用メモリ（ＥＥＰＲＯＭ）、フラッシュメモリ、もしくは他のメモリ技術、ＣＤ－ＲＯＭ、デジタル多用途ディスク（ＤＶＤ）、もしくは他の光ストレージ、磁気カセット、磁気テープ、磁気ディスクストレージ、もしくは他の磁気ストレージデバイス、または情報を記憶するのに使用可能であり、コンピューティングデバイス１０５によってアクセス可能な任意の他の製造品を含めることができる。任意のそのようなコンピュータストレージ媒体は、コンピューティングデバイス１０５の一部とすることができる。コンピュータストレージ媒体は、搬送波、または他の伝播もしくは変調データ信号を含まない。

【００６３】

[0070]通信媒体は、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、または搬送波もしくは他の輸送機構などの変調データ信号における他のデータによって実現可能であり、任意の情報配信媒体を含む。用語「変調データ信号」は、信号における情報を符号化するような形で、１つまたは複数の特性を設定または変更した信号について説明することができる。限定ではなく、例として、通信媒体は、ワイヤードネットワークまたは直接ワイヤード接続などのワイヤード媒体、ならびに音響、無線周波数（ＲＦ）、赤外線、および他のワイヤレス媒体など、ワイヤレス媒体を含むことができる。

【００６４】

[0071]図７Ａおよび図７Ｂは、本発明の実施形態がそれにより実践可能なモバイルコンピューティングデバイス７００、例えば、モバイル電話、スマートフォン、タブレットパーソナルコンピュータ、およびラップトップコンピュータなどを示している。図７Ａに関しては、諸実施形態を実装するためのモバイルコンピューティングデバイス７００の１つの実施形態が示されている。基本構成では、モバイルコンピューティングデバイス７００は、入力要素と出力要素との両方を有するハンドヘルドコンピュータである。モバイルコンピューティングデバイス７００は、典型的には、表示部７０５と、ユーザがモバイルコンピューティングデバイス７００に情報を入力することを可能にする１つまたは複数の入力ボタン７１０とを含む。モバイルコンピューティングデバイス７００の表示部７０５はまた、入力デバイス（例えば、タッチ画面表示部）として機能することができる。オプションのサイド入力要素７１５が含まれている場合、このサイド入力要素７１５は、さらなるユーザ入力を可能にする。サイド入力要素７１５は、ロータリスイッチ、ボタンまたは任意の他のタイプの手動入力要素とすることができる。代替の実施形態では、モバイルコンピューティングデバイス７００は、より多くの入力要素、またはより少ない入力要素を組み込むことができる。例えば、表示部７０５は、いくつかの実施形態では、タッチ画面でない場合もある。さらなる別の代替の実施形態では、モバイルコンピューティングデバイス７００は、セルラ電話など、携帯電話システムである。モバイルコンピューティングデバイス７００はまた、オプションのキーパッド７３５を含むことができる。オプションのキーパッド７３５は、物理キーパッドであっても、またはタッチ画面表示部において作成される「ソフト」キーパッドであってもよい。様々な実施形態では、出力要素は、グラフィカルユーザインターフェイス（ｇｒａｐｈｉｃａｌ　ｕｓｅｒ　ｉｎｔｅｒｆａｃｅ：ＧＵＩ）を示すために表示部７０５、視覚指示器７２０（例えば、発光ダイオード）、および／または音響トランスデューサ（例えば、スピーカ）を含む。いくつかの実施形態では、モバイルコンピューティングデバイス７００は、ユーザに触覚フィードバックを提供するための振動トランスデューサを組み込んでいる。さらなる別の実施形態では、モバイルコンピューティングデバイス７００は、音響入力（例えば、マイクロフォンジャック）、音響出力（例えば、ヘッドフォンジャック）など、入力ポートおよび／または出力ポート、ならびに外部デバイスに信号を送信し、もしくは外部デバイスから信号を受信するための映像出力（例えば、ＨＤＭＩ（登録商標）ポート）を組み込んでいる。

【００６５】

[0072]図７Ｂは、モバイルコンピューティングデバイスの１つの実施形態のアーキテクチャを示すブロック図である。つまり、モバイルコンピューティングデバイス７００は、

いくつかの実施形態を実装するシステム（すなわち、アーキテクチャ）７０２を組み込んでいることができる。１つの実施形態では、システム７０２は、１つまたは複数のアプリケーション（例えば、ブラウザ、ｅメール、スケジュール管理（calendaring）、連絡先マネージャ、メッセージングクライアント、ゲーム、およびメディアクライアント／プレーヤ）を動作させることができる「スマートフォン」として実装される。いくつかの実施形態では、システム７０２は、集積化されたパーソナルデジタルアシスタント（ｐｅｒｓｏｎａｌ ｄｉｇｉｔａｌ ａｓｓｉｓｔａｎｔ：ＰＤＡ）とワイヤレス電話など、コンピューティングデバイスとして集積化される。

【００６６】

[0073]１つまたは複数のアプリケーションプログラム７６６は、メモリ７６２にロード可能であり、オペレーティングシステム７６４において、またはオペレーティングシステム７６４と関連して動作することができる。アプリケーションプログラムの例には、電話ダイヤラプログラム、ｅメールプログラム、パーソナル情報管理（ｐｅｒｓｏｎａｌ ｉｎｆｏｒｍａｔｉｏｎ ｍａｎａｇｅｍｅｎｔ：ＰＩＭ）プログラム、ワードプロセシングプログラム、スプレッドシートプログラム、Ｉｎｔｅｒｎｅｔブラウザプログラム、およびメッセージングプログラムなどが含まれる。システム７０２はまた、メモリ７６２内に不揮発性ストレージ領域７６８を含む。不揮発性ストレージ領域７６８は、システム７０２が電源を落とされた場合、失うべきでない永続的情報を記憶するために使用可能である。アプリケーションプログラム７６６は、ｅメールアプリケーションなどによって使用されるｅメールまたは他のメッセージなど、不揮発性ストレージ領域７６８の情報を使用し、記憶することができる。同期化アプリケーション（図示せず）もまた、システム７０２において常駐し、ホストコンピュータに常駐する対応する同期化アプリケーションと相互作用して、不揮発性ストレージ領域７６８に記憶された情報がホストコンピュータにおいて記憶された対応する情報と継続して同期化されるようにプログラミングされる。理解すべきであるように、他のアプリケーションは、メモリ７６２にロード可能であり、モバイルコンピューティングデバイス７００において動作することができる。

【００６７】

[0074]システム７０２は、電源７７０を有し、この電源７７０は、１つまたは複数のバッテリとして実装可能である。電源７７０は、バッテリを補填する、もしくは再充電するＡＣアダプタまたは電力供給されるドッキングクレードルなど、外部電源をさらに含むこともあり得る。

【００６８】

[0075]システム７０２はまた、無線周波数通信を送受信する機能を行う無線機７７２も含むことができる。無線機７７２は、通信キャリアまたはサービスプロバイダを介して、システム７０２と「外部世界（ｏｕｔｓｉｄｅ ｗｏｒｌｄ）」との間のワイヤレス接続性を容易にする。無線機７７２との間の伝送は、オペレーティングシステム７６４の制御の下、実施される。言い換えれば、無線機７７２によって受信される通信は、オペレーティングシステム７６４を介してアプリケーションプログラム７６６に伝えられていくことができ、逆も同様である。

【００６９】

[0076]視覚指示器７２０は、視覚的な通知を提供するために使用可能であり、および／または音響インターフェイス７７４は、音響トランスデューサ７２５を介して可聴通知をもたらすために使用可能である。図示の実施形態では、視覚指示器７２０は、発光ダイオード（ＬＥＤ）であり、音響トランスデューサ７２５は、スピーカである。これらのデバイスは、電源７７０に直接、連結可能であり、それにより、プロセッサ７６０および他の構成要素がバッテリ電力を保存するためにシャットダウンすることがあり得ても、それらのデバイスは、アクティブ状態であるとき、通知機構によって指図された継続時間の間は、オンのままである。ＬＥＤは、ユーザがデバイスの電源オン状態を示すアクションを取るまでは無期限に、オンのままであるようにプログラミング可能である。音響インターフェイス７７４は、ユーザに可聴信号を提供し、ユーザから可聴信号を受信するために使用

される。例えば、音響インターフェイス７７４は、音響トランスデューサ７２５に連結されていることに加えて、電話会話を容易にするなどのために、可聴入力を受信するマイクロフォンにも連結可能である。本発明の実施形態により、マイクロフォンはまた、後述されるように、通知の制御を容易にする音響センサとしても働くことができる。システム７０２は、静止画像、および映像ストリームなどを記録するオンボードカメラ７３０の動作を可能にする映像インターフェイス７７６をさらに含むことができる。

【００７０】

[0077]システム７０２を実装するモバイルコンピューティングデバイス７００は、追加の特徴または機能性を有することができる。例えば、モバイルコンピューティングデバイス７００はまた、磁気ディスク、光ディスク、またはテープなどの追加のデータストレージデバイス（リムーバブルおよび／または非リムーバブル）も含むことができる。そのような追加のストレージは、不揮発性ストレージ領域７６８によって図７Ｂに示されている。

【００７１】

[0078]モバイルコンピューティングデバイス７００によって生成され、または取り込まれ、システム７０２を介して記憶されるデータ／情報は、上記に記載したように、モバイルコンピューティングデバイス７００においてローカルに記憶可能であり、すなわち、データは、無線機７７２を介して、またはモバイルコンピューティングデバイス７００と、モバイルコンピューティングデバイス７００に関連する別個のコンピューティングデバイス、例えば、Ｉｎｔｅｒｎｅｔなど、分散されたコンピューティングネットワークにおけるサーバコンピュータとの間のワイヤード接続を介して、デバイスによってアクセス可能な任意の数のストレージ媒体において記憶可能である。理解すべきであるように、そのようなデータ／情報は、無線機７７２を介して、または分散されたコンピューティングネットワークを介して、モバイルコンピューティングデバイス７００によりアクセス可能である。同様に、そのようなデータ／情報は、電子メールおよびコラボレーション型のデータ／情報共有システムを含む、よく知られているデータ／情報の転送およびストレージ手段に従って、記憶し、使用するためにコンピューティングデバイス間で容易に転送可能である。

【００７２】

[0079]図８は、上述したクラスタにおけるメンバシップを提供し、維持するためのシステムのアーキテクチャの１つの実施形態を示している。例えば、ノード登録テーブル、識別子、ならびにノード間およびノードと物理ディスクとの間の様々なパスは、種々の通信チャネルまたは他のストレージタイプで記憶され得る。例えば、様々な識別子は、ディレクトリサービス８２２、ウェブポータル８２４、メールボックスサービス８２６、インスタントメッセージングストア８２８、またはソーシャルネットワーキングサイト８３０を使用して記憶され得る。サーバ８２０は、データおよび／または接続タイプを、クラスタ内の１つまたは複数の他のサーバもしくはノードに提供することができる。１つの例として、サーバ８２０は、ネットワーク８１５を通じてデータをクライアントへとウェブを介して提供するウェブサーバとすることができる。例として、クライアントコンピューティングデバイスは、コンピューティングデバイス１０５として実装可能であり、パーソナルコンピュータ、タブレットコンピューティングデバイス６１０、および／またはモバイルコンピューティングデバイス７００（例えば、スマートフォン）において実現可能である。クライアントコンピューティングデバイス１０５、６１０、７００のこれらの実施形態のいずれもが、ストア８１６からコンテンツを取得することができる。

【００７３】

[0080]本発明の実施形態は、例えば、本発明の実施形態による方法、システム、およびコンピュータプログラム製品のブロック図および／または動作図を参照して上記に記載されている。ブロックに注記された機能／行為は、任意の流れ図に示される順序から外れて行われてもよい。実際、例えば、連続して示された２つのブロックが、実質的に同時に実行されても、またはブロックが、時として、関与している機能性／行為に応じて、逆の順

序で実行されてもよい。

【００７４】

[0081]本出願において提供された１つまたは複数の実施形態の説明および例示は、特許請求される本発明の範囲をいかなる形でも限定または制限するように意図するものではない。本出願において提供される実施形態、例、および詳細は、所有権を譲渡し、当業者が、特許請求される発明の最良な様態を作成し、使用することを可能にするのに十分であると見なされる。特許請求される本発明は、本出願において提供されるいかなる実施形態、例、または詳細にも限定するものとは見なすべきではない。組み合わせて、または別個に示され、説明されているかどうかにかかわらず、（構造的にも、方法論的にも）様々な特徴は、特定の特徴の組とともに実施形態を生み出すために選択的に含まれ、または省略されるように意図される。本出願の説明および例示が提供されているので、当業者は、特許請求される本発明のより広範な範囲から逸脱しない、本出願において実現される概括的な発明的概念のより広い態様の趣旨の範囲に入る変形形態、修正形態、および代替の実施形態を想定することができる。

【図１】

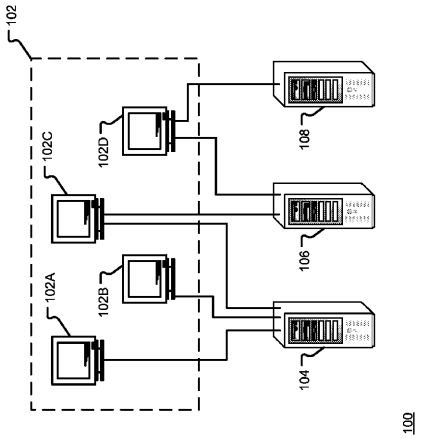
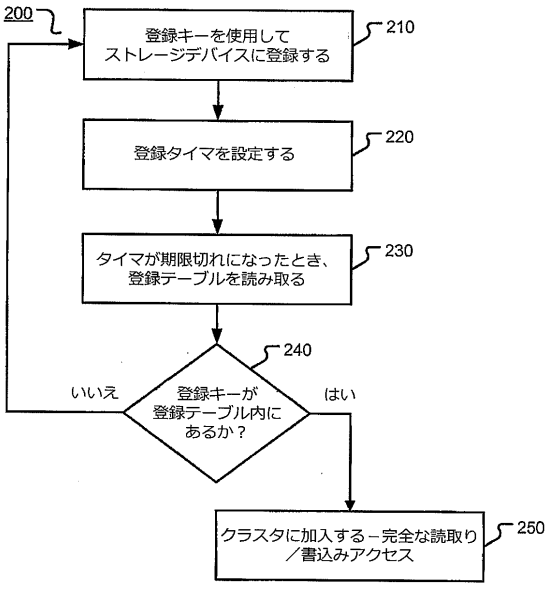
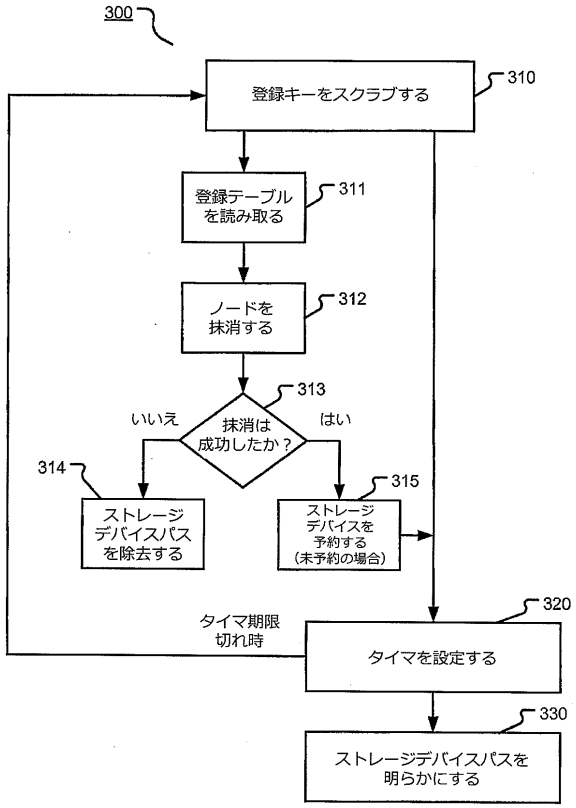



FIG. 1

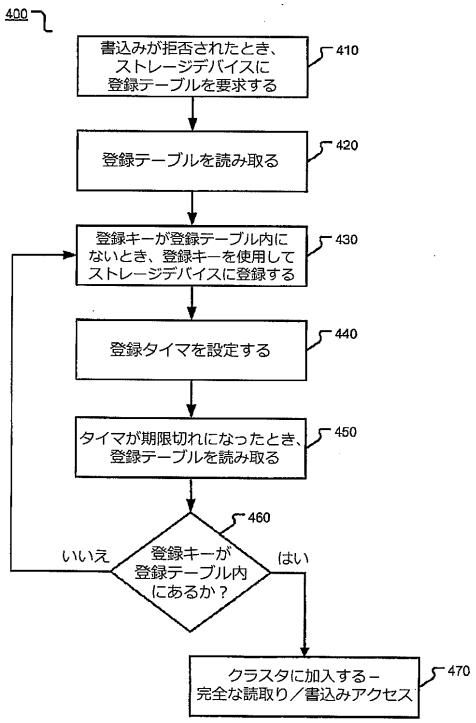
【図２】



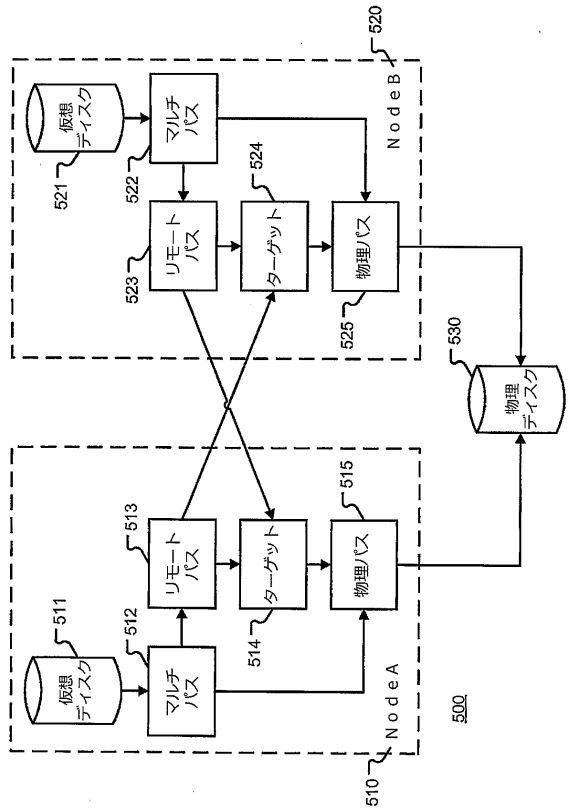
【図３】


300
登録キーをスクラブする
310
登録テーブル
を読み取る
311
ノードを
抹消する
312
抹消は
成功したか？
313
いいえ
はい
314
ストレージ
デバイスパス
を除去する
315
ストレージ
デバイスを
予約する
(未予約の場合)
タイマ期限
切れ時
320
タイマを設定する
330
ストレージデバイスパスを
明らかにする

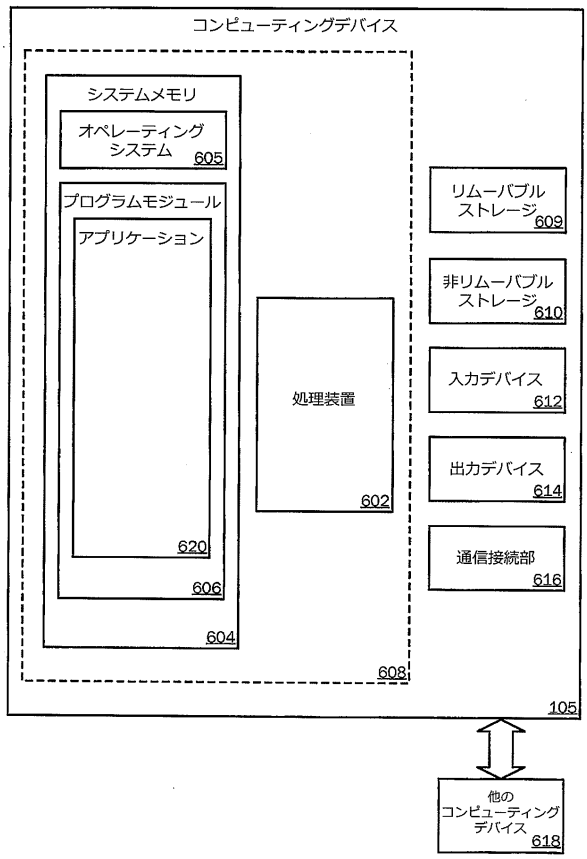
【図４】


400
書込みが拒否されたとき、
ストレージデバイスに
登録テーブルを要求する
410
登録テーブルを読み取る
420
登録キーが登録テーブル内に
ないとき、登録キーを使用して
ストレージデバイスに登録する
430
登録タイマを設定する
440
タイマが期限切れになったとき、
登録テーブルを読み取る
450
登録キーが
登録テーブル内
にあるか？
460
いいえ
はい
クラスタに加入する－
完全な読取り／書込みアクセス
470

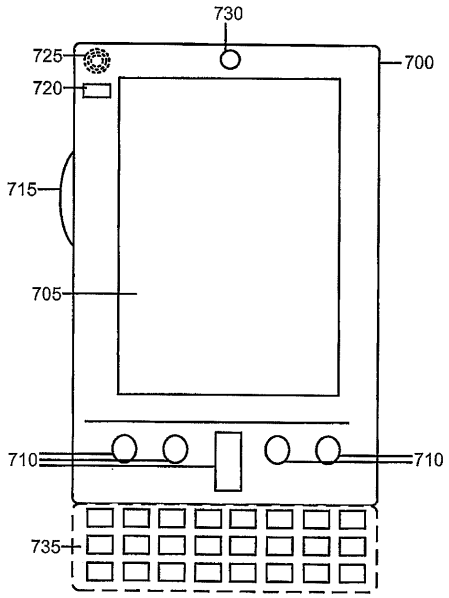
【図５】


520
NodeB
仮想
ディスク
521
マルチ
パス
522
リモート
パス
523
ターゲット
524
物理パス
525
530
物理
ディスク
515
513
リモート
パス
ターゲット
物理パス
511
仮想
ディスク
512
マルチ
パス
514
NodeA
500
510

【図６】

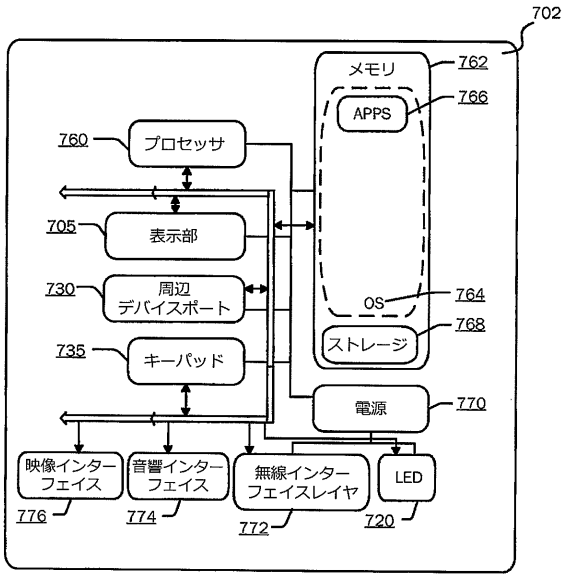

コンピューティングデバイス
システムメモリ
オペレーティング
システム
605
プログラムモジュール
アプリケーション
620
606
604
608
処理装置
602
リムーバブル
ストレージ
609
非リムーバブル
ストレージ
610
入力デバイス
612
出力デバイス
614
通信接続部
616
105
他の
コンピューティング
デバイス
618

【図７Ａ】

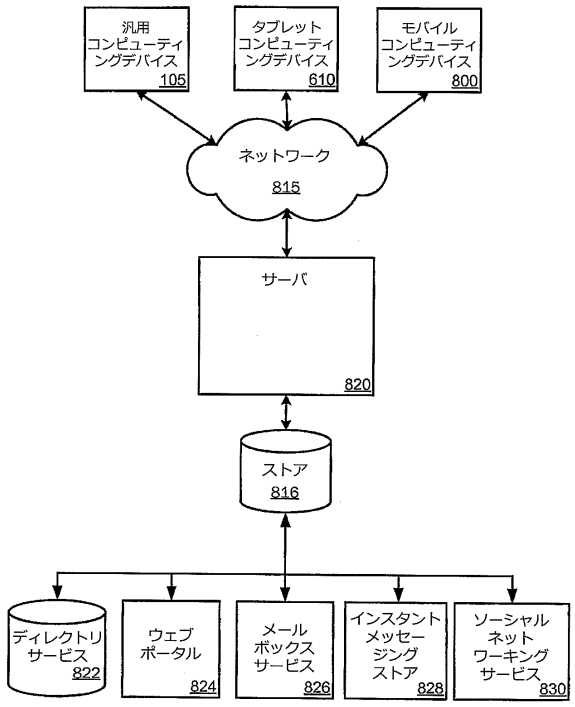



モバイルコンピューティングデバイス

【図７Ｂ】




【図８】



フロントページの続き

(74)代理人 100162846
弁理士 大牧 綾子
(72)発明者 クズネトソフ，ヴャチェスラフ
アメリカ合衆国ワシントン州９８０５２－６３９９，レッドモンド，マイクロソフト コーポレーション，エルシーエイ－インターナショナル・パテンツ（８／１１７２）
(72)発明者 シャンカー，ヴィノッド・アール
アメリカ合衆国ワシントン州９８０５２－６３９９，レッドモンド，マイクロソフト コーポレーション，エルシーエイ－インターナショナル・パテンツ（８／１１７２）
(72)発明者 ダマト，アンドレア
アメリカ合衆国ワシントン州９８０５２－６３９９，レッドモンド，マイクロソフト コーポレーション，エルシーエイ－インターナショナル・パテンツ（８／１１７２）
(72)発明者 ディオン，デーヴィッド・アレン
アメリカ合衆国ワシントン州９８０５２－６３９９，レッドモンド，マイクロソフト コーポレーション，エルシーエイ－インターナショナル・パテンツ（８／１１７２）

審査官 田上 隆一

(56)参考文献 特表２０１２－５０３２４９（ＪＰ，Ａ）
米国特許出願公開第２００３／００６５７８２（ＵＳ，Ａ１）
特表２０１１－５２６０３８（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ ３／０６
Ｇ０６Ｆ １３／１０
Ｇ０６Ｆ １３／１４