(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6763580号
(P6763580)
(45)発行日　令和2年9月30日(2020.9.30)　(24)登録日　令和2年9月14日(2020.9.14)

(51)Int.Cl.　F I
G06F 8/656 (2018.01)　G06F 8/656
G06F 16/11 (2019.01)　G06F 16/11
G06F 16/182 (2019.01)　G06F 16/182

請求項の数 32　(全 39 頁)

(21)出願番号　特願2018-529541 (P2018-529541)
(86)(22)出願日　平成28年12月19日 (2016.12.19)
(65)公表番号　特表2019-502202 (P2019-502202A)
(43)公表日　平成31年1月24日 (2019.1.24)
(86)国際出願番号　PCT/CN2016/110722
(87)国際公開番号　WO2017/114213
(87)国際公開日　平成29年7月6日 (2017.7.6)
審査請求日　令和1年9月27日 (2019.9.27)
(31)優先権主張番号　201511034171.7
(32)優先日　平成27年12月31日 (2015.12.31)
(33)優先権主張国・地域又は機関
中国 (CN)

(73)特許権者　510330264
アリババ・グループ・ホールディング・リミテッド
ＡＬＩＢＡＢＡ　ＧＲＯＵＰ　ＨＯＬＤＩＮＧ　ＬＩＭＩＴＥＤ
英国領，ケイマン諸島，グランド・ケイマン，ジョージ・タウン，ワン・キャピタル・プレイス，フォース・フロア，ピー．オー．ボックス　８４７
(74)代理人　110000877
龍華国際特許業務法人

最終頁に続く

(54)【発明の名称】分散記憶システムをアップグレードするための方法および装置

(57)【特許請求の範囲】
【請求項１】
クライアントに適用可能な、分散記憶システムをアップグレードするための方法であって、
複数のデータサーバに対して同じ書き込み対象データに関する書き込み要求を送信するステップと、
前記複数のデータサーバの各々により返される応答を受信するステップと、
前記応答に基づいて、成功した書き込みの数が所定の数より多いかどうかを判定するステップと、
前記成功した書き込みの前記数が前記所定の数より多い場合に、成功した書き込みで応答するそれぞれのデータサーバに対して第１フィードバック情報を送信するステップと、
前記成功した書き込みの前記数が前記所定の数より多くない場合に、成功した書き込みで応答するそれぞれのデータサーバに対して第２フィードバック情報を送信するステップと
を備え、
前記第１フィードバック情報または前記第２フィードバック情報は、前記データサーバが自らの状態を判定するために使用され、前記状態にはアップグレード可能な状態およびアップグレード不可能な状態が含まれ、データサーバの前記状態は、アップグレード制御サーバが前記データサーバに、アップグレード動作を実行するようローリング方式で

通知するために使用される、方法。
【請求項２】
前記成功した書き込みの前記数が前記所定の数より多くない場合に、前記成功した書き込みを有するそれぞれのデータサーバに対して第２フィードバック情報を前記送信するステップであって、前記第２フィードバック情報は、前記データサーバが自らの状態をアップグレード不可能と判定するために使用される、前記送信するステップは、
前記成功した書き込みの前記数が前記所定の数と等しい場合に、前記成功した書き込みを有するそれぞれのデータサーバに対して前記第２フィードバック情報を送信するステップと、
前記成功した書き込みの前記数が前記所定の数より少ない場合に、前記複数のデータサーバ以外の少なくとも１つのデータサーバに対して前記書き込み対象データに関する書き込み要求を送信するステップと、
前記少なくとも１つのデータサーバにより返される応答を受信し、前記応答に従って、成功した書き込みの現在の数が前記成功した書き込みの以前の数と組み合わせて前記所定の数と等しくなるかどうかを判定するステップと、
前記成功した書き込みの前記現在の数が前記所定の数と等しい場合に、前記成功した書き込みを有するそれぞれのデータサーバに対して前記第２フィードバック情報を送信するステップと
を含む、請求項１に記載の方法。
【請求項３】
前記成功した書き込みを有するそれぞれのデータサーバに対して第２フィードバック情報を前記送信するステップの後に、
新たな書き込み対象データが現れた場合に、以前に前記成功した書き込みがあった前記データサーバを含む前記複数のデータサーバを対象として使用するステップと、
前記複数のデータサーバに対して前記同じ書き込み対象データに関する書き込み要求を送信するステップと
を更に備える、請求項１または２に記載の方法。
【請求項４】
前記第１フィードバック情報および前記第２フィードバック情報は、クライアント識別子を含み、次いで、前記第１フィードバック情報または前記第２フィードバック情報は、前記データサーバが自らの状態を判定するために使用され、
前記第２フィードバック情報を受信して前記データサーバの前記状態をアップグレード不可能なものとしてマーキングした後、前記データサーバが前記第２フィードバック情報内の前記クライアント識別子をアップグレード不可能なリストへ書き込むために、前記第２フィードバック情報を使用するステップを備え、
前記データサーバが前記第１フィードバック情報を受信した後、前記第１フィードバック情報は、前記データサーバが、前記第１フィードバック情報内の前記クライアント識別子を前記アップグレード不可能なリストから削除すること、および、前記アップグレード不可能なリストが空であると判定した後に自らの状態をアップグレード可能なものとしてマーキングすることを行うために使用される、請求項３に記載の方法。
【請求項５】
複数のデータサーバに対して前記同じ書き込み対象データに関する書き込み要求を前記送信するステップの前に、
データサーバへアクセスするときに前記データサーバにより送信される第２アップグレード通知を受信するステップと、
前記第２アップグレード通知に従ってアップグレード準備状態へと移行するステップと
を更に備える、請求項１に記載の方法。
【請求項６】
データサーバに適用される、分散記憶システムをアップグレードするための方法であっ

て、
クライアントから送信される第１フィードバック情報または第２フィードバック情報を受信するステップであって、前記第１フィードバック情報または前記第２フィードバック情報は、前記クライアントが複数のデータサーバに対して同じ書き込み対象データに関する書き込み要求を送信した後に、成功した書き込みの数と所定の数とを比較した結果に従って取得される、受信するステップと、
データサーバがアップグレードされなかった場合に、前記第１フィードバック情報または前記第２フィードバック情報に従って前記データサーバの状態を判定するステップであって、前記状態にはアップグレード可能な状態およびアップグレード不可能な状態が含まれ、前記状態は、アップグレード制御サーバがローリング方式で、前記データサーバを選択し、前記データサーバに、アップグレード動作を実行するよう通知するために使用される、判定するステップと
を備える方法。
【請求項７】
前記第１フィードバック情報および前記第２フィードバック情報は、クライアント識別子を含み、次いで、前記データサーバがアップグレードされていない状況において、前記第１フィードバック情報または前記第２フィードバック情報に従って前記データサーバの状態を前記判定するステップは、
前記第２フィードバック情報を受信すると、前記第２フィードバック情報内の前記クライアント識別子をアップグレード不可能なリストに書き込むこと、および、自らの状態をアップグレード不可能なものとしてマーキングすることを行うステップと、
前記第１フィードバック情報を受信すると前記第１フィードバック情報内の前記クライアント識別子を前記アップグレード不可能なリストから削除すること、および、前記アップグレード不可能なリストが空であると判定した後に自らの状態をアップグレード可能なものとしてマーキングすることを行うステップと
を含む、請求項６に記載の方法。
【請求項８】
前記第１フィードバック情報内の前記クライアント識別子を前記アップグレード不可能なリストから削除すること、および、前記アップグレード不可能なリストが空であると判定した後に自らの状態をアップグレード可能なものとしてマーキングすることを前記行うステップは、
前記アップグレード不可能なリストが前記第１フィードバック情報内の前記クライアント識別子を含むかどうかを判定するステップと、
前記アップグレード不可能なリストが前記第１フィードバック情報内の前記クライアント識別子を含む場合に、前記クライアント識別子を前記アップグレード不可能なリストから削除するステップと、
前記アップグレード不可能なリストが空であるかどうかを判定するステップと、
前記アップグレード不可能なリストが空であれば、自らの状態をアップグレード可能なものとしてマーキングするステップと
を含む、請求項７に記載の方法。
【請求項９】
前記第２フィードバック情報を受信すると、前記第２フィードバック情報内の前記クライアント識別子をアップグレード不可能なリストに書き込むこと、および、自らの状態をアップグレード不可能なものとしてマーキングすることを前記行うステップの後に、
前記アップグレード不可能なリスト内のクライアント識別子について、対応するクライアントの第２フィードバックメッセージが所定の数の期間内に受信されなかったかどうかを判定すること、前記対応するクライアントの前記第２フィードバックメッセージが前記所定の数の期間内に受信されなかった場合に、前記クライアント識別子を前記アップグレード不可能なリストから削除することを行うステップ
を更に備える、請求項７に記載の方法。

【請求項１０】
クライアントにより送信される第１フィードバック情報または第２フィードバック情報を前記受信するステップの前に、
前記アップグレード制御サーバにより送信される第１アップグレード通知を受信するステップと、
前記第１アップグレード通知に従ってアップグレード準備状態へと移行すること、および、前記クライアントが前記アップグレード準備状態へと移行するよう、前記クライアントのアクセス要求を受信した後に第２アップグレード通知を前記クライアントへ送信することを行うステップと
を更に備える、請求項６に記載の方法。
【請求項１１】
アップグレード制御サーバに適用される、分散記憶システムをアップグレードするための方法であって、
複数のデータサーバのデータサーバごとに状態を取得するステップであって、前記状態にはアップグレード可能な状態およびアップグレード不可能な状態が含まれ、それぞれのデータサーバは一方の状態を有し、前記データサーバの前記状態は、第１フィードバック情報または第２フィードバック情報に従って判定され、前記第１フィードバック情報または前記第２フィードバック情報は、クライアントが複数のデータサーバに対して同じ書き込み対象データに関する書き込み要求を送信した後に、成功した書き込みの数と所定の数とを比較した結果に従って取得される、取得するステップと、
アップグレード可能な状態にある少なくとも１つのデータサーバに、アップグレード動作を実行するようローリング方式で通知するステップであって、前記データサーバは、前記通知に従って前記アップグレード動作を実行する、通知するステップと
を備える方法。
【請求項１２】
アップグレード可能な状態にある少なくとも１つのデータサーバに、アップグレード動作を実行するようローリング方式で前記通知するステップは、
毎回アップグレード可能な状態にある少なくとも１つのデータサーバを選択すること、および、前記アップグレード可能な状態にある前記少なくとも１つのデータサーバに、前記アップグレード動作を実行するよう通知することを行うステップと、
前記アップグレード可能な状態にある前記少なくとも１つのデータサーバが前記アップグレード動作を完全に終了しているかどうかをモニタリングするステップと、
前記アップグレード可能な状態にある前記少なくとも１つのデータサーバが前記アップグレード動作を完全に終了している場合に、次回に備えてアップグレード可能な状態にある少なくとも１つのデータサーバを選択すること、および、前記アップグレード可能な状態にある前記少なくとも１つのデータサーバに、前記アップグレード動作を実行するよう通知することを行うステップと
を含む、請求項１１に記載の方法。
【請求項１３】
任意のデータサーバの前記アップグレード動作の結果がアップグレードの失敗であることがモニタリングされた場合に、前記データサーバをアップグレードブラックリストに追加すること、および、前記データサーバのアップグレードを中断することを行うステップを更に備える、請求項１２に記載の方法。
【請求項１４】
前記データサーバは、少なくとも２つのラック上に設置され、アップグレード可能な状態にある少なくとも１つのデータサーバに、アップグレード動作を実行するようローリング方式で前記通知するステップは、
毎回アップグレード可能な状態にある前記データサーバを最も多く有するラックを選択すること、および、前記ラック内の前記データサーバに、アップグレード動作を実行するよう通知することを行うステップであって、前記ラック内のそれぞれのデータサーバは

、前記通知に従って自らの状態をチェックする、行うステップと、
　データサーバがアップグレード可能な状態にある場合に、前記データサーバを再起動およびアップグレードするステップと、
　データサーバがアップグレード不可能な状態またはアップグレード終了状態にある場合に、前記データサーバの再起動およびアップグレードを拒否するステップと
を含む、請求項１１から１３の何れか一項に記載の方法。
【請求項１５】
　毎回アップグレード可能な状態にある前記データサーバを最も多く有するラックを選択すること、および、前記ラック内の前記データサーバに、アップグレード動作を実行するよう通知することを前記行うステップの後に、
　前記ラック内の全ての前記データサーバが前記アップグレード動作を終了しているかどうかをモニタリングするステップと、
　前記ラック内の全ての前記データサーバが前記アップグレード動作を終了している場合に、次回に備えてアップグレード可能な状態にある前記データサーバを最も多く有するラックを選択すること、および、前記ラック内の前記データサーバに、アップグレード動作を実行するよう通知することを行うステップと
を更に備える、請求項１４に記載の方法。
【請求項１６】
　データサーバごとに前記状態を取得する前に、
　それぞれのデータサーバがアップグレード準備状態へと移行するよう、かつ、前記クライアントのアクセス要求を受信した後にそれぞれのデータサーバが第２アップグレード通知を前記クライアントへ送信することで、前記クライアントが前記アップグレード準備状態へと移行するよう、第１アップグレード通知をそれぞれのデータサーバに送信するステップ
を更に備える、請求項１１に記載の方法。
【請求項１７】
　クライアントに適用される、分散記憶システムをアップグレードするための装置であって、
　複数のデータサーバに対して同じ書き込み対象データに関する書き込み要求を送信する要求送信モジュールと、
　前記複数のデータサーバの各々により返される応答を受信すること、および、前記応答に基づいて、成功した書き込みの数が所定の数より多いかどうかを判定することを行う判定モジュールと、
　前記成功した書き込みの前記数が前記所定の数より多い場合に、前記成功した書き込みを有するそれぞれのデータサーバに対して第１フィードバック情報を送信する第１フィードバックモジュールと、
　前記成功した書き込みの前記数が前記所定の数より多くない場合に、前記成功した書き込みを有するそれぞれのデータサーバに対して第２フィードバック情報を送信する第２フィードバックモジュールと
を備え、
　前記第１フィードバック情報または前記第２フィードバック情報は、前記データサーバが自らの状態を判定するために使用され、前記状態にはアップグレード可能な状態およびアップグレード不可能な状態が含まれ、前記データサーバの前記状態は、アップグレード制御サーバが前記データサーバに、アップグレード動作を実行するようローリング方式で通知するために使用される、装置。
【請求項１８】
　前記第２フィードバックモジュールは、
　前記成功した書き込みの現在の数が前記所定の数と等しい場合に、前記成功した書き込みを有するそれぞれのデータサーバに対して前記第２フィードバック情報を送信する第２フィードバック情報送信サブモジュールと、

前記成功した書き込みの前記数が前記所定の数より少ない場合に、前記複数のデータサーバ以外の少なくとも１つのデータサーバに対して前記書き込み対象データに関する書き込み要求を送信する書き込み要求送信サブモジュールと、

前記少なくとも１つのデータサーバにより返される応答を受信すること、および、前記応答に従って、前記成功した書き込みの前記現在の数が前記成功した書き込みの以前の数と組み合わせて前記所定の数と等しくなるかどうかを判定すること、および、前記成功した書き込みの前記現在の数が前記所定の数と等しい場合に、前記第２フィードバック情報送信サブモジュールへと移行することを行う判定サブモジュールと

を含む、請求項１７に記載の装置。

【請求項１９】

前記第２フィードバックモジュールに次いで、

以前に前記成功した書き込みがあった前記データサーバを含む前記複数のデータサーバを対象として使用すること、および、新たな書き込み対象データが現れた場合は前記要求送信モジュールへと移行することを行う、新たな書き込み対象データの送信モジュールを更に備える、請求項１７または１８に記載の装置。

【請求項２０】

前記第１フィードバック情報および前記第２フィードバック情報は、クライアント識別子を含み、次いで、前記第１フィードバック情報または前記第２フィードバック情報は、前記データサーバが自らの状態を判定するために使用され、

前記第２フィードバック情報を受信した後に、前記データサーバが、前記第２フィードバック情報内の前記クライアント識別子をアップグレード不可能なリストに書き込むこと、および、前記データサーバの状態をアップグレード不可能なものとしてマーキングすることを行うために、前記第２フィードバック情報を使用することと、

前記データサーバが前記第１フィードバック情報を受信した後に前記データサーバが、前記第１フィードバック情報内の前記クライアント識別子を前記アップグレード不可能なリストから削除するために、前記第１フィードバック情報を使用すること、および、前記アップグレード不可能なリストが空であると判定した後に前記データサーバの状態をアップグレード可能なものとしてマーキングすることを行うことと

を備える、請求項１９に記載の装置。

【請求項２１】

複数のデータサーバに対して前記同じ書き込み対象データに関する書き込み要求を前記送信するステップの前に、

データサーバへアクセスするときに前記データサーバにより送信される第２アップグレード通知を受信する第２アップグレード通知受信モジュールと、

前記第２アップグレード通知に従ってアップグレード準備状態へと移行する第２アップグレード準備モジュールと

を更に備える、請求項１７に記載の装置。

【請求項２２】

データサーバに適用される、分散記憶システムをアップグレードするための装置であって、

クライアントから送信される第１フィードバック情報または第２フィードバック情報を受信するフィードバック情報受信モジュールであって、前記第１フィードバック情報または前記第２フィードバック情報は、前記クライアントが複数のデータサーバに対して同じ書き込みデータに関する書き込み要求を送信した後に、成功した書き込みの数と所定の数とを比較した結果に従って取得される、フィードバック情報受信モジュールと、

前記データサーバがアップグレードされなかった場合に、前記第１フィードバック情報または前記第２フィードバック情報に従って前記データサーバの状態を判定する状態判定モジュールであって、前記状態にはアップグレード可能な状態およびアップグレード不可能な状態が含まれ、前記状態は、アップグレード制御サーバがローリング方式で、前記データサーバを選択し、前記データサーバに、アップグレード動作を実行するよう通知す

るために使用される、状態判定モジュールと
を備える装置。
【請求項２３】
前記状態判定モジュールは、
前記第２フィードバック情報内のクライアント識別子をアップグレード不可能なリストに書き込むこと、および、前記第２フィードバック情報を受信すると、前記データサーバの状態をアップグレード不可能なものとしてマーキングすることを行うアップグレード不可能な状態判定サブモジュールと、
前記第１フィードバック情報を受信すると、前記第１フィードバック情報内の前記クライアント識別子を前記アップグレード不可能なリストから削除すること、および、前記アップグレード不可能なリストが空であると判定した後に、前記データサーバの状態をアップグレード可能なものとしてマーキングすることを行うアップグレード可能な判定サブモジュールと
を含む、請求項２２に記載の装置。
【請求項２４】
前記アップグレード可能な判定サブモジュールは、
前記アップグレード不可能なリストが前記第１フィードバック情報内の前記クライアント識別子を含むかどうかを判定すること、および、含む場合は第１削除サブモジュールへと移行することを行うクライアント識別子判定サブモジュールであって、前記第１削除サブモジュールは、前記クライアント識別子を前記アップグレード不可能なリストから削除する、クライアント識別子判定サブモジュールと、
前記アップグレード不可能なリストが空であるかどうかを判定すること、および、空であればアップグレード可能なマーキングサブモジュールへと移行することを行うアップグレード不可能なリスト判定サブモジュールであって、前記アップグレード可能なマーキングサブモジュールは、前記データサーバの状態をアップグレード可能なものとしてマーキングする、アップグレード不可能なリスト判定サブモジュールと
を含む、請求項２３に記載の装置。
【請求項２５】
前記アップグレード不可能な状態判定サブモジュールに次いで、
前記アップグレード不可能なリスト内のクライアント識別子について、対応するクライアントの第２フィードバックメッセージが所定の数の期間内に受信されなかったかどうかを判定すること、および、前記対応するクライアントの前記第２フィードバックメッセージが前記所定の数の期間内に受信されなかった場合に、第２削除サブモジュールへと移行することを行う時間判定サブモジュールであって、
前記第２削除サブモジュールは、前記クライアント識別子を前記アップグレード不可能なリストから削除する、時間判定サブモジュール
を更に備える、請求項２３に記載の装置。
【請求項２６】
前記フィードバック情報受信モジュールより前に、
前記アップグレード制御サーバにより送信される第１アップグレード通知を受信する第１アップグレード通知受信モジュールと、
前記第１アップグレード通知に従ってアップグレード準備状態へと移行すること、および、前記クライアントが前記アップグレード準備状態へと移行するよう、前記クライアントのアクセス要求を受信した後に第２アップグレード通知を前記クライアントへ送信することを行う第１アップグレード準備モジュールと
を更に備える、請求項２２に記載の装置。
【請求項２７】
アップグレード制御サーバに適用される、分散記憶システムをアップグレードするための装置であって、
複数のデータサーバのそれぞれのデータサーバの状態を取得する状態取得モジュール

であって、前記状態にはアップグレード可能な状態およびアップグレード不可能な状態が含まれ、それぞれのデータサーバは一方の状態を有し、前記データサーバの前記状態は、第１フィードバック情報または第２フィードバック情報に従って判定され、前記第１フィードバック情報または前記第２フィードバック情報は、クライアントが複数のデータサーバに対して同じ書き込み対象データに関する書き込み要求を送信した後に、成功した書き込みの数と所定の数とを比較した結果に従って取得される、状態取得モジュールと、

アップグレード可能な状態にある少なくとも１つのデータサーバに、アップグレード動作を実行するようローリング方式で通知するアップグレード通知モジュールであって、前記データサーバは、前記通知に従って前記アップグレード動作を実行する、アップグレード通知モジュールと

を備える装置。

【請求項２８】

前記アップグレード通知モジュールは、

毎回アップグレード可能な状態にある少なくとも１つのデータサーバを選択すること、および、前記アップグレード可能な状態にある前記少なくとも１つのデータサーバに、前記アップグレード動作を実行するよう通知することを行う第１選択サブモジュールと、

前記アップグレード可能な状態にある前記少なくとも１つのデータサーバが前記アップグレード動作を完全に終了しているかどうかをモニタリングすること、および、終了している場合は前記第１選択サブモジュールへと移行することを行う第１モニタリングサブモジュールと

を含む、請求項２７に記載の装置。

【請求項２９】

任意のデータサーバの前記アップグレード動作の結果がアップグレードの失敗であることがモニタリングされた場合に、前記データサーバをアップグレードブラックリストに追加すること、および、前記データサーバのアップグレードを中断することを行う中断サブモジュール

を更に備える、請求項２８に記載の装置。

【請求項３０】

前記データサーバは、少なくとも２つのラック上に設置され、前記アップグレード通知モジュールは、

毎回アップグレード可能な状態にある前記データサーバを最も多く有するラックを選択すること、および、前記ラック内の前記データサーバに、アップグレード動作を実行するよう通知することを行うことであって、前記ラック内のそれぞれのデータサーバは、前記通知に従って自らの状態をチェックする、行うこととと、データサーバがアップグレード可能な状態にある場合に、前記データサーバを再起動およびアップグレードすることと、データサーバがアップグレード不可能な状態またはアップグレード終了状態にある場合に、前記データサーバの再起動およびアップデートを拒否することとを行うアップグレード通知サブモジュール

を含む、請求項２７から２９の何れか一項に記載の装置。

【請求項３１】

前記アップグレード通知サブモジュールに次いで、

前記ラック内の前記データサーバが前記アップグレード動作を完全に終了しているかどうかをモニタリングすること、および、終了している場合は前記アップグレード通知サブモジュールへと移行することを行う第２モニタリングサブモジュール

を更に備える、請求項３０に記載の装置。

【請求項３２】

前記状態取得モジュールの前に、

それぞれのデータサーバがアップグレード準備状態へと移行するように、かつ、前記クライアントのアクセス要求を受信した後にそれぞれのデータサーバが第２アップグレード通知を前記クライアントへ送信することで、前記クライアントが前記アップグレード準

備状態へと移行するよう、第１アップグレード通知をそれぞれのデータサーバに送信するアップグレード通知送信モジュール
を更に備える、請求項２７に記載の装置。
【発明の詳細な説明】
【技術分野】
【０００１】
［関連出願の相互参照］
本開示は、２０１５年１２月３１日に出願された「ＭＥＴＨＯＤ　ＡＮＤ　ＡＰＰＡＲＡＴＵＳ　ＦＯＲ　ＵＰＧＲＡＤＩＮＧ　ＤＩＳＴＲＩＢＵＴＥＤ　ＳＴＯＲＡＧＥ　ＳＹＳＴＥＭ」と題する中国特許出願第２０１５１１０３４１７１．７号、および２０１６年１２月１９日に出願された「ＭＥＴＨＯＤ　ＡＮＤ　ＡＰＰＡＲＡＴＵＳ　ＦＯＲ　ＵＰＧＲＡＤＩＮＧ　ＤＩＳＴＲＩＢＵＴＥＤ　ＳＴＯＲＡＧＥ　ＳＹＳＴＥＭ」と題する国際特許出願第ＰＣＴ／ＣＮ１６／１１０７２２号に基づく優先権を主張するものである。これらの出願は両方とも、全体が参照により本明細書に組み込まれる。
【０００２】
本開示は、分散コンピュータ技術の分野、特に分散記憶システムをアップグレードするための方法および装置に関する。
【背景技術】
【０００３】
高可用性分散記憶システムが、高可用性サービスを構築するための基盤となっている。分散記憶システムは、分散データサーバで構成され、高信頼性データおよび高可用性アクセスを提供する。高信頼性は、データ冗長マルチバックアップコードまたは消去コードにより実装され、高可用性は、迅速な例外処理およびフェイルオーバにより実装される。分散記憶システムのバージョンアップグレードを実行するには、当該システムのそれぞれのデータサーバが再起動されてバージョンアップデートを完了する必要がある。
【０００４】
現行のシステムでは、以下の解決策を用いて分散記憶システムのアップグレードが実行され得る。
【０００５】
第１解決策では、高水準サービスが無効となり、分散記憶システム全体が停止する。次いで、分散記憶システムの全てのデータサーバが再起動およびアップグレードされる。全てのデータサーバがアップグレードを完了した後、高水準サービスが再開する。しかしながら、この解決策では、全ての高水準サービスが利用不可能となり、こうした非可用性が長期間にわたって続くため、高可用性を必要とするサービスには適用可能でない。
【０００６】
第２解決策では、高水準サービスが停止するのではなく、むしろそれぞれのデータサーバがローリング方式で再起動される。クライアントが複数のデータサーバ（データサーバの数は、バックアップを必要とするデータサーバのデフォルト数である）に対して書き込み要求を送信して、当該データサーバに書き込み対象データを書き込む。要求が失敗すると、クライアントは、当該データサーバの残りで再試行してクライアントのアクセスを確保する。しかしながら、この解決策では、クライアントがデータサーバのリカバリを待ってから再試行した場合に、この再試行手段がサーバの応答時間に大きな影響を及ぼすことになる。なぜなら、データサーバの再起動リカバリには通常何秒もの時間を必要とし、リアルタイム性の高いデータアクセスに１０～１００ｍｓの遅れが生じるからである。
【０００７】
第３の解決策では、高水準サービスが停止するのではなく、むしろそれぞれのデータサーバがローリング方式で再起動される。クライアントが複数のデータサーバ（当該複数のデータサーバの数は、バックアップを必要とするデータサーバのデフォルト数である）に対して書き込み要求を送信して、当該データサーバに書き込み対象データを書き込む。要求が失敗すると、クライアントは、失敗したデータサーバを無視して再試行することでク

ライアントのアクセスを確保する。しかしながら、この解決策では、たとえダウンしているデータサーバをクライアントが一時的に無視したとしても、分散記憶システムがこれらのデータサーバのローリングアップグレードを継続する。分散記憶システム内のデータサーバがあまりにも急速にローリング方式で再起動された場合、クライアントが１つのデータサーバにだけ成功した書き込みを実行することで１つのデータサーバだけが成功した書き込みを有することになる場合、および、このデータサーバのディスクまたは機械がローリング再起動中に損傷を受けた場合は、クライアントの書き込み対象データが他のデータサーバに書き込まれていないのでユーザデータが失われる。たとえローリング再起動時間が延長されても、長時間のローリング中に分散記憶システムの大きなクラスタで予期せぬディスクおよび機械の例外が生じる。従って、１つのデータサーバだけがクライアントのデータの成功した書き込みを有するといった状況が存続し、ユーザデータの消失という大きな危険をもたらす。故に、この解決策では、待ち時間の問題は解消されるが、データの信頼性は低い。

【発明の概要】

【０００８】

前述の問題を考慮して、分散記憶システムをアップグレードするための方法、および、それに対応する、分散記憶システムをアップグレードするための装置を提供することでこれらの問題を克服するか、または少なくとも部分的に解消するために、本開示の実施形態が紹介される。

【０００９】

前述の問題を解消すべく、本開示は、クライアントに適用される、分散記憶システムをアップグレードするための方法を開示する。当該方法は、複数のデータサーバに対して同じ書き込み対象データに関する書き込み要求を送信するステップと、当該複数のデータサーバの各々により返される応答を受信するステップと、当該応答に基づいて、成功した書き込みの数が所定の数より多いかどうかを判定するステップと、成功した書き込みの数が所定の数より多い場合に、成功した書き込みで応答するそれぞれのデータサーバに対して第１フィードバック情報を送信するステップと、成功した書き込みの数が所定の数より多くない場合に、成功した書き込みで応答するそれぞれのデータサーバに対して第２フィードバック情報を送信するステップとを備える。ここで、第１フィードバック情報または第２フィードバック情報は、データサーバが自らの状態を判定するために使用され、当該状態にはアップグレード可能な状態およびアップグレード不可能な状態が含まれ、データサーバの状態は、アップグレード制御サーバがデータサーバに、アップグレード動作を実行するようローリング方式で通知するために使用される。

【００１０】

本開示は更に、データサーバに適用される、分散記憶システムをアップグレードするための方法を開示する。当該方法は、クライアントから送信される第１フィードバック情報または第２フィードバック情報を受信するステップを備える。ここで、第１フィードバック情報または第２フィードバック情報は、クライアントが複数のデータサーバに対して同じ書き込み対象データに関する書き込み要求を送信した後に、成功した書き込みの数と所定の数とを比較した結果に従って取得される。データサーバがアップグレードされなかった場合は、第１フィードバック情報または第２フィードバック情報に従って当該データサーバの状態を判定する。当該状態にはアップグレード可能な状態およびアップグレード不可能な状態が含まれ、当該状態は、アップグレード制御サーバがローリング方式で、データサーバを選択し、当該データサーバに、アップグレード動作を実行するよう通知するために使用される。

【００１１】

本開示は更に、アップグレード制御サーバに適用される、分散記憶システムをアップグレードするための方法を開示する。当該方法は、複数のデータサーバのデータサーバごとに状態を取得するステップであって、当該状態にはアップグレード可能な状態およびアップグレード不可能な状態が含まれ、それぞれのデータサーバは一方の状態を有し、データ

サーバの状態は、第１フィードバック情報または第２フィードバック情報に従って判定され、第１フィードバック情報または第２フィードバック情報は、クライアントが複数のデータサーバに対して同じ書き込み対象データに関する書き込み要求を送信した後に、成功した書き込みの数と所定の数とを比較した結果に従って取得される、取得するステップと、アップグレード可能な状態にある少なくとも１つのデータサーバに、アップグレード動作を実行するようローリング方式で通知するステップであって、この通知に従って当該データサーバはアップグレード動作を実行する、通知するステップとを備える。

【００１２】

本開示は更に、クライアントに適用される、分散記憶システムをアップグレードするための装置を開示する。当該装置は、複数のデータサーバに対して同じ書き込み対象データに関する書き込み要求を送信するように構成される要求送信モジュールと、当該データサーバの各々により返される応答を受信すること、および、当該応答に基づいて、成功した書き込みの数が所定の数より多いかどうかを判定することを行うように構成される判定モジュールと、成功した書き込みの数が所定の数より多い場合に、成功した書き込みを有するそれぞれのデータサーバに対して第１フィードバック情報を送信するように構成される第１フィードバックモジュールと、成功した書き込みの数が所定の数より多くない場合に、成功した書き込みを有するそれぞれのデータサーバに対して第２フィードバック情報を送信するように構成される第２フィードバックモジュールとを備える。ここで、第１フィードバック情報または第２フィードバック情報は、データサーバが自らの状態を判定するために使用され、当該状態にはアップグレード可能な状態およびアップグレード不可能な状態が含まれ、データサーバの状態は、アップグレード制御サーバがデータサーバに、アップグレード動作を実行するようローリング方式で通知するために使用される。

【００１３】

本開示は更に、データサーバに適用される、分散記憶システムをアップグレードするための装置を開示する。当該装置は、クライアントから送信される第１フィードバック情報または第２フィードバック情報を受信するように構成されるフィードバック情報受信モジュールであって、第１フィードバック情報または第２フィードバック情報は、クライアントが複数のデータサーバに対して同じ書き込み対象データに関する書き込み要求を送信した後に、成功した書き込みの数と所定の数とを比較した結果に従って取得される、フィードバック情報受信モジュールと、データサーバがアップグレードされなかった場合に、第１フィードバック情報または第２フィードバック情報に従ってデータサーバの状態を判定するように構成される状態判定モジュールであって、当該状態にはアップグレード可能な状態およびアップグレード不可能な状態が含まれ、当該状態は、アップグレード制御サーバがローリング方式で、データサーバを選択し、当該データサーバに、アップグレード動作を実行するよう通知するために使用される、状態判定モジュールとを備える。

【００１４】

本開示は更に、アップグレード制御サーバに適用される、分散記憶システムをアップグレードするための装置を開示する。当該装置は、複数のデータサーバのデータサーバごとに状態を取得するように構成される状態取得モジュールであって、当該状態にはアップグレード可能な状態およびアップグレード不可能な状態が含まれ、それぞれのデータサーバは一方の状態を有し、データサーバの状態は、第１フィードバック情報または第２フィードバック情報に従って判定され、第１フィードバック情報または第２フィードバック情報は、クライアントが複数のデータサーバに対して同じ書き込み対象データに関する書き込み要求を送信した後に、成功した書き込みの数と所定の数とを比較した結果に従って取得される、状態取得モジュールと、アップグレード可能な状態にある少なくとも１つのデータサーバに、アップグレード動作を実行するようローリング方式で通知するように構成されるアップグレード通知モジュールであって、この通知に従って当該データサーバはアップグレード動作を実行する、アップグレード通知モジュールとを備える。本開示の実施形態には以下の利点がある。

【００１５】

本開示の実施形態において、分散記憶システムがアップグレード手続きを開始した後、分散記憶システムへアクセスしているそれぞれのクライアントは、同じ書き込み対象データに関する書き込み要求を複数のデータサーバへ同時に送信する。それぞれのクライアントは次いで、書き込み対象データが幾つのデータサーバへ成功裏に書き込まれているかを解析し、成功した書き込みの数が所定の数より多いかどうかを判定し、その判定結果に従って、成功した書き込みを有するそれぞれのデータサーバに対して第１フィードバック情報または第２フィードバック情報を送信する。受信される第１フィードバック情報または第２フィードバック情報に従って、データサーバは、自らがアップグレード可能な状態にあるのか、アップグレード不可能な状態にあるのかを判定する。次いで、データサーバの状態に従って、アップグレード制御サーバはローリング方式で、データサーバを選択し、当該データサーバに、アップグレード動作を実行するよう通知してよい。前述のプロセスにより、本開示の実施形態において、アップグレード制御サーバが、ローリング方式でアップグレードするようデータサーバを制御する場合は、高水準サービスが停止する必要はない。なぜなら、クライアントがデータサーバの状態を制御し、任意のクライアントの書き込み対象データがバックアップのために少なくとも所定の数のデータサーバへ書き込まれることが保証されるからである。更には、分散記憶システムからクライアントへのより短い応答時間が保証され得ることで、データの信頼性が高まり、ユーザデータの損失という危険が著しく低下する。

【図面の簡単な説明】

【００１６】

【図１】本開示の幾つかの実施形態に係る、クライアント側で分散記憶システムをアップグレードするための方法を示すフロー図である。

【００１７】

【図２】本開示の幾つかの実施形態に係る、クライアント側で分散記憶システムをアップグレードするための方法を示すフロー図である。

【００１８】

【図３】本開示の幾つかの実施形態に係る、アップグレード制御サーバ側で分散記憶システムをアップグレードするための方法を示すフロー図である。

【００１９】

【図４】本開示の幾つかの実施形態に係る、分散記憶システムをアップグレードするための方法を示すフロー図である。

【００２０】

【図５】本開示の幾つかの実施形態に係る、クライアント側で分散記憶システムをアップグレードするための装置のブロック図である。

【００２１】

【図６】本開示の幾つかの実施形態に係る、クライアント側で分散記憶システムをアップグレードするための装置のブロック図である。

【００２２】

【図７】本開示の幾つかの実施形態に係る、データサーバ側で分散記憶システムをアップグレードするための装置のブロック図である。

【００２３】

【図８】本開示の幾つかの実施形態に係る、アップグレード制御サーバ側で分散記憶システムをアップグレードするための装置のブロック図である。

【００２４】

【図８Ａ】本開示の幾つかの実施形態に係る、分散記憶システムのアーキテクチャを示している。

【発明を実施するための形態】

【００２５】

上述した本開示の目的、特徴および利点をより明瞭で理解し易くするために、本開示は、添付図および具体的な実装と併せて更に詳しく後述される。

【００２６】
開示される実施形態の核となる着想のうち１つは、分散記憶システムのデータサーバおよびアップグレード制御サーバに対する創造的変更が為されるというものであり、データサーバへアクセスしているクライアントに対して新たな実行論理が提供されるというものでもある。分散記憶システムがアップグレード手続きを開始した後、分散記憶システムへアクセスしているそれぞれのクライアントは、同じ書き込み対象データに関する書き込み要求を複数のデータサーバへ同時に送信する。次いで、書き込み対象データが成功裏に書き込まれたデータサーバの数が解析される。成功した書き込みの数が所定の数より多いかどうかが判定される。この判定に従って、成功した書き込みを有するそれぞれのデータサーバに対して第１フィードバック情報または第２フィードバック情報が送信される。受信される第１フィードバック情報または第２フィードバック情報に従って、データサーバは、自らがアップグレード可能な状態にあるのか、アップグレード不可能な状態にあるのかを判定する。次いで、データサーバの状態に従って、アップグレード制御サーバはローリング方式で、データサーバを選択し、当該データサーバに、アップグレード動作を実行するよう通知してよい。前述のプロセスにより、アップグレード制御サーバが、ローリング方式でアップグレードするようアップグレード可能な状態にあるデータサーバを制御する場合は、高水準サービスが停止する必要はない。なぜなら、クライアントがデータサーバの状態を制御し、任意のクライアントの書き込み対象データがバックアップのために少なくとも所定の数のデータサーバへ書き込まれることが保証されるからである。更には、分散記憶システムからクライアントへのより短い応答時間が保証され得ることで、データの信頼性が高まり、ユーザデータの損失という危険が著しく低下する。更に、このプロセスは、サービスが影響を受けないように保証しつつ、アップグレード中の機械の予期せぬ例外を許容できる。
【００２７】
後述される本開示の実施形態は、クライアントに適用されてよい。
【００２８】
図１は、本開示の幾つかの実施形態に係る、分散記憶システムをアップグレードするための方法を示すフロー図である。具体的に言うと、当該方法は以下のステップを含んでよい。
【００２９】
ステップ１１０：複数のデータサーバに対して書き込み対象データに関する書き込み要求を伝送する。一実施形態では、書き込み要求ごとに書き込み対象データが同じである。
【００３０】
本開示のこの実施形態において、クライアント（Ａ）が分散記憶システムのアップグレード中にデータＡ１をデータサーバへ書き込もうとする場合は、クライアントＡが複数のデータサーバに対してデータＡ１に関する書き込み要求を送信する。
【００３１】
本開示のこの実施形態では、複数のデータサーバの数Ｒが、例えば１０に予め設定されてよい。本開示で具体的な数Ｒが限定されることはない。
【００３２】
次いで、クライアントＡは、Ｒ個のデータサーバに対して書き込み対象データに関する書き込み要求を送信し、次いで、ステップ１２０へと進んでよい。
【００３３】
実際の適用においては、まず、クライアントＡが分散記憶システムのスケジューリングサーバに対してＲ個の書き込み要求を送信してよいことが理解され得る。次いで、スケジューリングサーバは、どのＲ個のデータサーバに当該Ｒ個の書き込み要求が割り当てられるかを制御する。
【００３４】
異なるクライアントがＲ個のデータサーバに対して自らの書き込み対象データに関する書き込み要求を送信する場合は、Ｒ個のデータサーバがそれぞれ同じであってもよいし、

異なっていてもよいことに留意すべきである。
【００３５】
本開示の別の実施形態において、当該方法は、ステップ１１０の前にステップ１０１および１０２を更に含む。
【００３６】
ステップ１０１：データサーバへアクセスするときに当該データサーバにより送信される第２アップグレード通知を受信する。
【００３７】
ステップ１０２：第２アップグレード通知に従って、アップグレード準備状態へと移行する。
【００３８】
本開示のこの実施形態では、まず、分散記憶システムのアップグレード制御サーバが、それぞれのデータサーバにアップグレード準備状態への移行を通知する。次いで、それぞれのデータサーバは、クライアントがアップグレード準備状態へと移行するよう、データサーバへアクセスしているクライアントに対して第２アップグレード通知を送信してよい。
【００３９】
それに応じて、アップグレード準備状態へと移行しているデータサーバにクライアントＡがアクセスした後、当該データサーバは、第２アップグレード通知をクライアントに返す。第２アップグレード通知を受信した後、クライアントは、第２アップグレード通知に従ってアップグレード準備状態へと移行する。
【００４０】
本開示のこの実施形態において、クライアントがブラウザのウェブページを介してデータサーバにアクセスする場合は、ブラウザにより開かれたウェブページを介して、アップグレードスクリプトがブラウザに送信されてよい。当該スクリプトを受信した後、ブラウザは、クライアントがアップグレード準備状態へと移行するよう当該スクリプトを実行してよい。クライアントがモバイルアプリケーション（例えばＡＬＩＰＡＹアプリケーション）を介してデータサーバにアクセスする場合は、当該アプリケーションにアップグレード処理論理が予め追加されてよい。アップグレード処理論理は、クライアントがアップグレード準備状態へと移行するよう、アプリケーションが第２アップグレード通知を受信した後に有効となる。もちろん、本開示で具体的なやり方が限定されることはない。
【００４１】
クライアントがアップグレード準備状態へと移行するとき、クライアントにより実行される、本開示の一実施形態におけるステップ１１０、１２０、１３０および１４０等のステップへとフローが移行してよい。
【００４２】
ステップ１２０：データサーバの各々により返される書き込み応答を受信し、当該応答に従って、成功した書き込みの数が所定の数より多いかどうかを判定する。成功した書き込みの数が所定の数より多い場合は、ステップ１３０へと移行する。成功した書き込みの数が所定の数より多くない場合は、ステップ１４０へと移行する。
【００４３】
本開示のこの実施形態において、正常な状況下で、それぞれのデータサーバは、クライアントにより送信される、同じ書き込み対象データに関する書き込み要求を受信した後、相応してクライアントへ応答を返す。もちろん、アップグレード中の特定のデータサーバまたはダウンしている特定のデータサーバがクライアントに応答を返すことはない。
【００４４】
本開示のこの実施形態において、クライアントは、Ｒ個のデータサーバに対して書き込み対象データＡ１に関する書き込み要求を送信した後、受信される応答を定期的にチェックしてよい。例えば、特定の１つまたは複数のデータサーバの１つまたは複数の応答が受信されたかどうかが定期的にチェックされる。指定された期間内に特定の１つまたは複数

のデータサーバの１つまたは複数の応答が受信されなかった場合、このことは、書き込み対象データＡ１が特定の１つまたは複数のデータサーバへ成功裏に書き込まれていないことを示す。指定された期間内に特定の１つまたは複数のデータサーバの１つまたは複数の応答が受信された場合は、当該応答が成功した書き込み応答であるのか、失敗した書き込み応答であるのかが判定されてよい。

【００４５】

応答が成功した書き込み応答であれば、このことは、当該成功した書き込み応答に対応するデータサーバが書き込み対象データＡ１を成功裏にバックアップしたこと、すなわち成功裏に書き込んだことを示す。応答が失敗した書き込み応答であれば、当該失敗した書き込み応答に対応するデータサーバが書き込み対象データＡ１をバックアップするのに失敗したこと、すなわち書き込むのに失敗したことを示す。

【００４６】

Ｒ個のデータサーバの応答については、ＭがＮより大きいかどうかをクライアントＡが定期的にチェックすることに留意すべきである（パラメータＭおよびＮについては、本明細書で更に説明される）。期間は例えば１ｍｓ（マイクロ秒）である。もちろん、期間は必要に従って決定されてよく、本開示により限定されることはない。次いで、前述の決定に基づいて、データサーバに対する書き込み対象データＡ１の成功した書き込みの数が算出されてよい。

【００４７】

本開示のこの実施形態では、成功した書き込みの所定の数が予め設定されてよい。所定の数は、成功した書き込みの必要最低限の数であり、クライアントの書き込み対象データのバックアップに成功したデータサーバの数と理解されてもよい。所定の数は例えばＮである。ここで、Ｎ＜Ｒであり、ＮおよびＲは両方とも正の整数である。実際には、デフォルト設定によりＮ＝３およびＲ＝１０であってよい。もちろん、ＮおよびＲの値は実際の必要に従って設定されてよく、本開示で限定されることはない。

【００４８】

次いで、クライアントＡが現在のＲ個のデータサーバに対して書き込み対象データＡ１に関する書き込み要求を送信した後、Ｍ個のデータサーバが成功した書き込み応答を返す場合は、Ｍ＞Ｎかどうかが判定される。Ｍ＞Ｎならば、フローがステップ１３０へと移行する。Ｍ≦Ｎならば、フローがステップ１４０へと移行する。Ｍ≧０であり、Ｍは整数である。

【００４９】

実際の適用においては、データサーバがラック上に設置される。次いで、ディザスタリカバリ・バックアップの効果を更に高めるために、Ｒ個のデータサーバが異なるラック上に設置されてよい。Ｒ個のデータサーバは、全て異なるラック上に設置されるのが最適である。

【００５０】

ステップ１３０：成功した書き込みを有するそれぞれのデータサーバに対して第１フィードバック情報を送信する。

【００５１】

クライアントの書き込み要求が成功裏に受信され得ること、および、内部の書き込み対象データがバックアップされ得ることから、Ｍ個のデータサーバが利用可能であると言える。次いで、クライアントは、Ｍ個のデータサーバに対して第１フィードバックメッセージを送信してよい。第１フィードバックメッセージは、例えばＯＫメッセージであってよく、ＯＫメッセージは、データサーバに対してそれが自らの状態をアップグレード可能なものに設定できることを知らせる。

【００５２】

実際のところ、本開示の図２では、データサーバが第１フィードバックメッセージに従って自らの状態をアップグレード可能なものに設定する方法が詳しく説明されている。

【００５３】

ステップ１４０：成功した書き込みを有するそれぞれのデータサーバに対して第２フィードバック情報を送信する。

【００５４】

第１フィードバック情報または第２フィードバック情報は、データサーバが自らの状態を判定するために使用される。当該状態にはアップグレード可能な状態およびアップグレード不可能な状態が含まれる。データサーバの状態は、アップグレード制御サーバがデータサーバに、アップグレード動作を実行するようローリング方式で通知するために使用される。

【００５５】

Ｍ≦Ｎならば、Ｍ個のデータサーバが利用可能となるので、クライアントは、Ｍ個のデータサーバに対して第２フィードバックメッセージを送信してよい。第２フィードバックメッセージは、例えばＨＯＬＤハートビートメッセージであってよく、ＨＯＬＤメッセージは、データサーバに対してそれが自らの状態をアップグレード不可能なものに設定できることを知らせる。　実際のところ、本開示の図２では、データサーバが第２フィードバックメッセージに従って自らの状態をアップグレード不可能なものに設定する方法が詳しく説明されている。

【００５６】

Ｍ＜Ｎという状況については、Ｍが少なくともＮと等しいことが保証されるまで、クライアントが書き込み要求を新たなデータサーバへ再び送信することに留意すべきである。

【００５７】

分散記憶システム内のアップグレードされなかったそれぞれのデータサーバについては、データサーバは、データサーバが第１フィードバック情報または第２フィードバック情報を受信するときに、第１フィードバック情報または第２フィードバック情報に従って自らがアップグレード可能な状態にあるのか、アップグレード不可能な状態にあるのかを判定してよい。アップグレード可能な状態は例えばＯＫであり、アップグレード不可能な状態は例えばＨＯＬＤである。例えば、特定のデータサーバが第２フィードバック情報を受信する場合は、データサーバが自らの状態をアップグレード不可能なものに設定してよい。特定のデータサーバが第１フィードバック情報を受信する場合は、データサーバが自らの状態をアップグレード可能なものに設定してよい。

【００５８】

もちろん、実際の適用においては、高水準サービスが停止するのではなく、データサーバが継続的にアップグレードされるので、前述のＲ個のデータサーバが、アップグレードを完了するデータサーバを含んでもよい。次いで、アップグレードされたデータサーバについては、データサーバにより受信される任意の前述のフィードバック情報が更に処理されることはなく、データサーバはアップグレードされた状態のままである。

【００５９】

本開示の図２では、受信される第１フィードバック情報または第２フィードバック情報に従ってデータサーバが自らの状態を判定する方法の具体的なプロセスが詳しく説明されている。

【００６０】

更に、それぞれのデータサーバの状態は、アップグレード制御サーバがデータサーバに、アップグレード動作を実行するようローリング方式で通知するために使用されてよい。本開示の図３では、アップグレード制御サーバがデータサーバに、アップグレード動作を実行するようローリング方式で通知する方法の具体的なプロセスが説明されている。

【００６１】

本開示の別の実施形態において、ステップ１４０は、サブステップ１４１～１４４を含む。

【００６２】

サブステップ１４１：成功した書き込みの数が所定の数と等しい場合は、成功した書き込みを有するそれぞれのデータサーバに対して第２フィードバック情報を送信する。

【００６３】
例えば、前述のクライアントＡは、Ｒ個のデータサーバに対して書き込み対象データＡ１を初めて送信する。次いで、モニタリングされた成功した書き込み応答の数がＭ＝Ｎであれば、成功した書き込みを有するＭ個のデータサーバに対して第２フィードバック情報が直接送信され得る。
【００６４】
サブステップ１４２：成功した書き込みの数が所定の数より少なければ、当該複数のデータサーバ以外の少なくとも１つのデータサーバに対して書き込み対象データに関する書き込み要求を送信する。
【００６５】
サブステップ１４３：当該少なくとも１つのデータサーバにより返される応答を受信し、当該応答に従って、成功した書き込みの現在の数が成功した書き込みの以前の数と組み合わせて所定の数と等しくなるかどうかを判定する。成功した書き込みの現在の数が所定の数と等しい場合は、フローがサブステップ１４４へと移行する。成功した書き込みの現在の数が所定の数より少なければ、フローがサブステップ１４２へと移行する。
【００６６】
サブステップ１４４：成功した書き込みの数が所定の数と等しい場合は、成功した書き込みを有するそれぞれのデータサーバに対して第２フィードバック情報を送信する。
【００６７】
クライアントＡがＲ個のデータサーバに対して書き込み対象データＡ１を初めて送信するサブステップ１４２～１４４については、モニタリングされた成功した応答の数がＭ＜Ｎであれば、少なくとも（Ｎ－Ｍ）個のデータサーバに対して書き込み要求が改めて送信される。次いで（Ｎ－Ｍ）個のデータサーバの応答が受信され、（Ｎ－Ｍ）個のデータサーバにより返される成功した書き込み応答の数が、Ｒ個のデータサーバの成功した書き込み応答の以前の数に追加されてＭとなる。この時点でＭ＝Ｎならば、成功した書き込みを有するそれぞれのデータサーバに対して第２フィードバック情報が送信される。この時点でＭの値は変化するが、Ｍがこの時点で依然としてＮより小さければ、少なくとも（Ｎ－Ｍ）個のデータサーバに対して書き込み要求が再び送信される。このプロセスはＭ＝Ｎとなるまで継続し、次いで、成功した書き込みを有するそれぞれのデータサーバに対して第２フィードバック情報が送信される。
【００６８】
更に、Ｒ＝１０であり、Ｎ＝５であり、かつ、クライアントＡが１０個のデータサーバに対して書き込み対象データＡ１を初めて送信する場合は、モニタリングされた成功した書き込み応答の数がＭ＝３であれば、少なくとも（５－３＝２）個のデータサーバに対して書き込み要求が改めて送信される。２つのデータサーバの成功した書き込み応答の数が更にモニタリングされ、一方の成功した書き込み応答だけが受信されたことが分かり、次いで、予め記録されている３に１が追加されてＭ＝４となる。次いで、少なくとも（５－４＝１）個のデータサーバに対して書き込み要求が改めて送信される。前述のやり方で、モニタリングおよび判定が更に実行される。Ｍ＝Ｎ＝５ならば、成功した書き込みを有するそれぞれのデータサーバに対して第２フィードバック情報が送信される。実際の適用において、Ｍ＜Ｎならば、前述のプロセスにおいて第２フィードバックメッセージは、書き込み要求が送信されるたびに、成功した書き込みを有するデータサーバに対して直接送信されてよく、Ｍ＝Ｎになるまで待ってからまとめて送信しなくてもよい。
【００６９】
もちろん、実際の適用においては、データサーバの応答が定期的にチェックされるので、指定された期間Ｔ、例えば３*ＴにおいてＭ＜Ｎであると更に判定された場合は、少なくとも（Ｎ－Ｍ）個のデータサーバに対して書き込み要求が改めて送信される。ＭおよびＮに関するそのような判定は、周期ごとに為される。一般的には、複数の周期の後にＭ＝Ｎであると判定され得る。応答時間は一般的に短く、クライアントの応答遅延に大きな影響を及ぼすことはない。ユーザの観点からすれば、前述の状況における応答遅延は、再起

動を待っている際の応答遅延よりもはるかに短いことが分かる。

【００７０】

このように、複数のデータベースにおけるクライアントのバックアップ対象データの成功したバックアップが、アップロードサービスの通常使用に影響を及ぼすことなく保証される。

【００７１】

本開示の別の実施形態において、当該方法は、ステップ１４０の後にステップ１５０を更に含む。

【００７２】

ステップ１５０：新たな書き込み対象データが現れた場合は、以前に成功した書き込みがあったデータサーバを含む複数のデータサーバが対象として使用され、ステップ１２０へと移行する。

【００７３】

Ｍ＝Ｎという状況については、成功した書き込みを有するそれぞれのデータサーバに対してクライアントがＨＯＬＤハートビート情報を返すので、これらのデータサーバが自らをアップグレード不可能な状態、例えばＨＯＬＤ状態に設定することが理解され得る。これらのデータサーバにおけるクライアントの制限をより簡便に取り除くべく、本開示のこの実施形態において、クライアントはその後、以前のＲ個のデータサーバに対して新たな書き込み対象データに関する書き込み要求を送信する。

【００７４】

例えば、クライアントＡの書き込み対象データＡ１については、１０個のデータサーバＵ１、Ｕ２・・・Ｕ１０に対して書き込み要求が予め送信され、Ｍ＝Ｎと判定される。次いで、クライアントＡの書き込み対象データＡ２については、１０個のデータサーバＵ１、Ｕ２・・・Ｕ１０に対して書き込み要求が更に送信されてよく、ステップ１２０での判定する段階が継続する。

【００７５】

サブステップ１４２については、Ｒ個のデータサーバが、成功した書き込みを有するＭ個のデータサーバを含むこと、次いで、失敗した書き込みを有するデータサーバから残り（Ｒ－Ｍ個）のデータサーバが選択されることが理解され得る。例えば、Ｒ＝１０かつＮ＝３であり、クライアントＡの書き込み対象データＡ１については、１０個のデータサーバＵ１、Ｕ２・・・Ｕ１０に対して書き込み要求が初めて送信され、Ｕ１およびＵ２だけが成功した書き込みを有し、次いで、Ｕ１１およびＵ１２に対して２度目の書き込み要求が送信され、Ｕ１１が成功した書き込みを有する。次いで、クライアントの書き込み対象データＡ２については、１０個のデータサーバＵ１、Ｕ２、Ｕ１１、Ｕ４、Ｕ５・・・Ｕ１０が選択されてよく、当該１０個のデータサーバに対して書き込み要求が送信されるが、もちろん、Ｕ１、Ｕ２、Ｕ１１および他のデータサーバから成る１０個のデータサーバが選択されてもよい。

【００７６】

次いで、クライアントＡの書き込み対象データＡ２の成功した書き込みの数がＭ＞Ｎであれば、成功した書き込みを有するそれぞれのデータサーバに対してＯＫメッセージが送信されてよく、データサーバは、ＯＫメッセージに従って自らの状態をアップグレード可能に変更するかどうかを判定してよい。

【００７７】

このように、Ｍ＜Ｎという状況では、クライアントが書き込み対象データを同じＲ個のデータサーバへ再び書き込んで、当該データサーバの状態をアクティブにアップデートしてよい。その結果、当該データサーバがアップグレード不可能な状態にある時間が短縮され得る。

【００７８】

本開示の別の実施形態において、第１フィードバック情報および第２フィードバック情報は、クライアント識別子を含む。

【００７９】
例えば、クライアントＡについては、第１フィードバック情報または第２フィードバック情報をサーバへ送信するとき、第１フィードバック情報および第２フィードバック情報の両方がクライアント識別子「クライアントＡ」を含む。
【００８０】
次いで、データサーバが自らの状態を判定するために第１フィードバック情報または第２フィードバック情報が使用されることは、第２フィードバック情報を受信した後に、データサーバが、第２フィードバック情報内のクライアント識別子をアップグレード不可能なリストに書き込むこと、および、自らの状態をアップグレード不可能なものとしてマーキングすることを行うために、第２フィードバック情報が使用されることを含む。
【００８１】
本開示のこの実施形態では、アップグレード不可能なリストがデータサーバ側で設定される。クライアントＡを例として用いると、クライアントＡの第２フィードバック情報を受信した後、データサーバは、自らの状態をアップグレード不可能な状態に設定して、クライアントＡをアップグレード不可能なリストに書き込む。
【００８２】
データサーバが第１フィードバック情報を受信した後、第１フィードバック情報は、データサーバが、第１フィードバック情報内のクライアント識別子をアップグレード不可能なリストから削除すること、および、アップグレード不可能なリストが空であると判定した後に自らの状態をアップグレード可能なものとしてマーキングすることを行うために使用される。
【００８３】
本開示のこの実施形態では、このステップが実行された後、成功した書き込みを有するそれぞれのデータサーバに対してクライアントＡが第２フィードバック情報を送信し、新たな書き込み対象データが出現し、成功した書き込みを有する以前のデータサーバを含む複数のデータサーバが対象として使用される。次いで、クライアントＡにより送信される第２フィードバックメッセージを受信するデータサーバはその後、クライアントＡにより送信されるメッセージを受信する。次いで、データサーバにより更に受信されるクライアントＡのメッセージが第１フィードバックメッセージであれば、クライアントＡの記録がアップグレード不可能なリストから削除されてよい。
【００８４】
もちろん、実際の適用においては、データサーバが特定のクライアントの第１フィードバックメッセージを受信し、対応するクライアント識別子は、アップグレード不可能なリストに記録されなくてもよい。その場合は、削除プロセスが実行される必要はない。
【００８５】
次いで、アップグレード不可能なリストが空であるとデータサーバが判定すれば、データサーバは、自らの状態をアップグレード可能なものに設定する。
【００８６】
もちろん、本開示のこの実施形態において、クライアントは、データサーバにより送信される、アップグレード準備状態を終了するようにとの終了通知を更に受信し、当該終了通知に従ってアップグレード準備状態を終了してよい。次いで、クライアントは、通常の要求送信論理に従って書き込み要求をデータサーバへ送信する。
【００８７】
本開示のこの実施形態では、本開示における分散記憶システムをアップグレードするための方法がクライアント側から導入され、それぞれのクライアントが、同じ書き込み対象データに関する書き込み要求を複数のデータサーバへ同時に送信し、次いで、幾つのデータサーバが成功した書き込みを有しているかが解析され、成功した書き込みの数が所定の数より多いかどうかが判定され、その判定結果に従って、成功した書き込みを有するそれぞれのデータサーバに対して第１フィードバック情報または第２フィードバック情報が送信される。　データサーバが第１フィードバック情報を受信した後、第１フィードバック

情報は、データサーバが、第１フィードバック情報内のクライアント識別子をアップグレード不可能なリストから削除すること、および、アップグレード不可能なリストが空であると判定した後に自らの状態をアップグレード可能なものとしてマーキングすることを行うために使用される。従って、本開示の実施形態において、アップグレード制御サーバが、ローリング方式でアップグレードするようデータサーバを制御する場合は、高水準サービスが停止する必要はない。なぜなら、クライアントがデータサーバの状態を制御し、任意のクライアントの書き込み対象データがバックアップのために少なくとも所定の数のデータサーバへ書き込まれることが保証されるからである。更には、分散記憶システムからクライアントへのより短い応答時間が保証され得ることで、データの信頼性が高まり、ユーザデータの損失という危険が著しく低下する。

【００８８】

本開示のこの実施形態は、分散記憶システムのデータサーバ側に適用される。

【００８９】

図２は、本開示の幾つかの実施形態に係る、分散記憶システムをアップグレードするための方法を示すフロー図である。具体的に言うと、当該方法は以下のステップを含んでよい。

【００９０】

ステップ２１０：クライアントから送信される第１フィードバック情報または第２フィードバック情報を受信する。ここで、第１フィードバック情報または第２フィードバック情報は、クライアントが複数のデータサーバに対して同じ書き込み対象データに関する書き込み要求を送信した後に、成功した書き込みの数と所定の数とを比較した結果に従って取得される。

【００９１】

図１におけるクライアント側の説明を参照すると、クライアントがＲ個のデータサーバに対して書き込み対象データＡ１に関する書き込み要求を送信した後、成功した書き込みの数Ｍが所定の数Ｎより大きければ、Ｍ個のデータサーバに対して第１フィードバックメッセージが送信される。ＭがＮと等しい場合は、Ｍ個のデータサーバに対して第２フィードバックメッセージが送信される。Ｍ＜Ｎならば、当該複数のデータサーバ以外の少なくとも１つのデータサーバに対して書き込み対象データＡ１に関する書き込み要求が送信される。次いで、当該少なくとも１つのデータサーバにより返される応答が受信され、当該応答に従って、全ての成功した書き込みの現在の数Ｍが成功した書き込みの以前の数と組み合わせて所定の数Ｎと等しくなるかどうかが判定される。Ｍ＝Ｎならば、Ｍ個のデータサーバに対して第２フィードバック情報が送信される。クライアントが第１フィードバック情報および第２フィードバック情報を返す具体的なプロセスについては、図１の説明が参照されてよく、本明細書で再び説明されることはない。

【００９２】

関連して言うと、分散記憶システム内のそれぞれのデータサーバは、それぞれのクライアントにより返されるＯＫメッセージ等の第１フィードバック情報を受信してよい。または、それぞれのクライアントにより返されるＨＯＬＤメッセージ等の第２フィードバックメッセージが受信されてもよい。

【００９３】

一実施形態において、当該方法は、ステップ２１０の前に以下のステップを更に含む。

【００９４】

ステップ２０１：アップグレード制御サーバにより送信される第１アップグレード通知を受信する。

【００９５】

ステップ２０２：第１アップグレード通知に従ってアップグレード準備状態へと移行すること、および、クライアントがアップグレード準備状態へと移行するよう、クライアントのアクセス要求を受信した後に第２アップグレード通知をクライアントへ送信することを行う。

【００９６】
本開示のこの実施形態では、まず、分散記憶システムのアップグレード制御サーバが第１アップグレード通知をそれぞれのデータサーバへ送信する。それに応じて、それぞれのデータサーバは、第１アップグレード通知を受信し、次いで、第１アップグレード通知に従ってアップグレード準備状態へと移行する。
【００９７】
次いで、アップグレード準備状態に移行するデータサーバへのアクセスについては、特定のクライアントのアクセス要求が受信された後に、第２アップグレード通知がクライアントに返される。クライアントは次いで、第２アップグレード通知に従ってアップグレード準備状態へと移行する。
【００９８】
データサーバがアップグレード準備状態へと移行するときには、データサーバにより実行される、本開示のこの実施形態におけるステップ２１０および２２０等のステップへとフローが移行してよい。
【００９９】
ステップ２２０：データサーバがアップグレードされなかった場合は、第１フィードバック情報または第２フィードバック情報に従ってデータサーバの状態を判定する。当該状態にはアップグレード可能な状態およびアップグレード不可能な状態が含まれる。
【０１００】
当該状態は、アップグレード制御サーバがローリング方式で、データサーバを選択し、当該データサーバに、アップグレード動作を実行するよう通知するために使用される。
【０１０１】
実際の適用においては、全てのデータサーバがアップグレード準備状態へと移行したばかりであれば、それぞれのデータサーバがアップグレードされることはない。しかしながら、幾つかのデータサーバが継続的にアップグレードされているときには、これらのデータサーバがアップグレードされた状態にある。例えば、成功裏にアップグレードされたデータサーバが自らの状態をＤＯＮＥに設定する。この時点で、成功裏にアップグレードされたデータサーバは、第１フィードバック情報または第２フィードバック情報も受信する。しかしながら、データサーバが第１フィードバック情報または第２フィードバック情報を更に処理することはなく、アップグレードされた状態のままである。
【０１０２】
アップグレードされていない状態にあるデータサーバだけが、第１フィードバック情報または第２フィードバック情報に従ってデータサーバ自らの状態を判定する。
【０１０３】
分散クラスタ内のアップグレードされなかったそれぞれのデータサーバについては、データサーバは、データサーバが第１フィードバック情報または第２フィードバック情報を受信するときに、第１フィードバック情報または第２フィードバック情報に従って自らがアップグレード可能な状態にあるのか、アップグレード不可能な状態にあるのかを判定してよいことが理解され得る。アップグレード可能な状態は例えばＯＫであり、アップグレード不可能な状態は例えばＨＯＬＤである。例えば、特定のデータサーバが第２フィードバック情報を受信する場合は、データサーバが自らの状態をアップグレード不可能なものに設定してよい。特定のデータサーバが第１フィードバック情報を受信する場合は、データサーバが自らの状態をアップグレード可能なものに設定してよい。
【０１０４】
本開示の別の実施形態において、第１フィードバック情報および第２フィードバック情報は、クライアント識別子を含む。次いで、ステップ２２０は、図１のステップ１５０に基づいてサブステップ２２１および２２２を含んでよい。
【０１０５】
サブステップ２２１：第２フィードバック情報を受信すると、第２フィードバック情報内のクライアント識別子をアップグレード不可能なリストに書き込んで、自らの状態をア

ップグレード不可能なものとしてマーキングする。

【０１０６】

本開示のこの実施形態では、アップグレード不可能なリストがそれぞれのデータサーバに予め設定され、ＨＯＬＤメッセージを送信するクライアントの識別子等の情報を記録するのに使用される。

【０１０７】

例えば、クライアントＡが複数のデータサーバに対して書き込み対象データＡ１に関する書き込み要求を送信する。ここで、データサーバＵ１、Ｕ２、Ｕ３は成功した書き込みを有する。Ｍ＝Ｎならば、クライアントがＭ個の対応するデータサーバに対してＨＯＬＤメッセージを送信する。ＨＯＬＤメッセージは、ＨＯＬＤ命令およびクライアント識別子を含む。次いで、Ｍ個のデータサーバは、ＨＯＬＤメッセージを受信するとクライアントＡをアップグレード不可能なリストに記録する。もちろん、相応して現在時刻が記録されてよい。

【０１０８】

例えば、クライアントＢが複数のデータサーバに対して書き込み対象データＢ１に関する書き込み要求を送信する。ここで、複数のデータサーバにはデータサーバＵ１も含まれ、Ｍ＝Ｎである。次いで、データサーバＵ１のアップグレード不可能なリストがクライアントＢを更に記録する。もちろん、現在時刻も記録されてよい。

【０１０９】

表１には、データサーバＵ１のアップグレード不可能なリストの記録の例が示されている。

【表１】

クライアント識別子	時刻
クライアントＡ	２０１５．１０．０１、１０：００：００
クライアントＢ	２０１５．１０．０１、１２：０１：００

【０１１０】

同じクライアントにより再び送信されるＨＯＬＤメッセージについては、アップグレード不可能なリストは、クライアントに対応するクライアント識別子を１つだけ記録してもよいし、クライアントに対応するクライアント識別子を複数記録してもよい。実際の適用において、同じクライアントにより異なる時刻に送信されるＨＯＬＤメッセージについては、クライアント識別子が１つだけ記録されてよく、次いで、それぞれの時刻が時刻フィールドに記録される。

【０１１１】

ＨＯＬＤメッセージが受信された後、データサーバ自らの状態は、ＨＯＬＤ状態へと更に変更され、データサーバがアップグレード不可能であることを示す。

【０１１２】

一実施形態において、当該方法は、サブステップ２２１の後にサブステップＢ２１を更に含む。

【０１１３】

サブステップＢ２１：アップグレード不可能なリスト内のクライアント識別子については、対応するクライアントの第２フィードバックメッセージが所定の数の期間内に受信さ

れなかったかどうかを判定し、対応するクライアントの第２フィードバックメッセージが所定の数の期間内に受信されなかった場合に、クライアント識別子をアップグレード不可能なリストから削除する。

【０１１４】

本開示のこの実施形態では、期間Ｔ１が予め設定されてよく、アップグレード不可能なリスト内のクライアント識別子に対応するクライアントについては、クライアントにより送信されるＨＯＬＤメッセージをデータサーバが所定の数の期間内に再び受信していない場合は、アップグレード不可能なリスト内のクライアント識別子の記録が削除される。

【０１１５】

例えば、前述のデータサーバＵ１の表１については、クライアントＡのＨＯＬＤメッセージが３＊Ｔ１という時間内に受信されなかった場合に、表１内のクライアントＡの記録が削除される。

【０１１６】

このステップにより、データサーバがアップグレード不可能な状態へ移行した後、アップグレード不可能な状態にずっと留まるのを防ぐことができる。

【０１１７】

サブステップ２２２：第１フィードバック情報を受信すると、第１フィードバック情報内のクライアント識別子をアップグレード不可能なリストから削除し、アップグレード不可能なリストが空であると判定した後に自らの状態をアップグレード可能なものとしてマーキングする。

【０１１８】

例えば、クライアントＡにより送信されるＯＫメッセージを前述のデータサーバＵ１が再び受信し、次いで、クライアントＡの記録が表１から削除される。データサーバが表１内の全ての記録を削除した後に表１が空であれば、データサーバは、自らの状態をアップグレード可能なものに設定する。

【０１１９】

本開示の別の実施形態におけるサブステップ２２２は、サブステップＢ１１～Ｂ１４を含む。

【０１２０】

サブステップＢ１１：アップグレード不可能なリストが第１フィードバック情報内のクライアント識別子を含むかどうかを判定し、含む場合はサブステップＢ１２へと移行する。

【０１２１】

サブステップＢ１２：クライアント識別子をアップグレード不可能なリストから削除する。

【０１２２】

１つのデータサーバが使用を目的として異なるクライアントへ割り当てられてよいので、異なるクライアントの第１フィードバック情報が受信されてよい。次いで、受信される特定のクライアントの第１フィードバック情報に基づいて、アップグレード不可能なリストが第１フィードバック情報のクライアント識別子を含むかどうかが判定される。アップグレード不可能なリストが第１フィードバック情報のクライアント識別子を含む場合は、アップグレード不可能なリストからクライアント識別子が削除される。アップグレード不可能なリストが第１フィードバック情報のクライアント識別子を含まない場合は、その後の動作が実行されなくてもよい。

【０１２３】

例えば、前述のデータサーバＵ１については、クライアントＡにより改めて送信される、クライアント識別子を含むＯＫメッセージが受信された場合は、表１内のクライアント識別子とのマッチングに当該クライアント識別子が使用されてよい。クライアントＡが存在することが分かった場合は、次いで、クライアントＡの記録が消去される。クライアントＣにより送信されるＯＫメッセージが受信されて、クライアントＣが表１に記録されて

いないことが分かった場合は、何の動作も実行されない。

【０１２４】

前述のステップにより、データサーバが簡単な論理で自らのアップグレード状態を管理するのが容易になる。

【０１２５】

サブステップＢ１３：アップグレード不可能なリストが空であるかどうかを判定し、空であればサブステップＢ１４へと移行する。アップグレード不可能なリストが空でない場合は、データサーバが自らの状態をアップグレード不可能なものに維持する。

【０１２６】

サブステップＢ１４：自らの状態をアップグレード可能なものとしてマーキングする。

【０１２７】

アップグレード不可能なリストが空でない場合は、データサーバが自らの状態をアップグレード不可能なものとして維持する。

【０１２８】

例えば、前述の例において、データサーバＵ１は、まずクライアントＡにより送信されるＯＫメッセージを受信し、クライアントＡの記録を消去し、表１がまだ空ではないと判定する。従って、データサーバＵ１はＨＯＬＤ状態のままである。次いで、クライアントＢのＯＫメッセージが受信された場合は、表１内のクライアントＢの記録がサブステップ２２１で削除され、この時点で表１が空であると判定される。次いで、データサーバＵ１は、自らのＨＯＬＤ状態をＯＫ状態に変更し、データサーバＵ１がアップグレード可能であることを示す。

【０１２９】

次いで、それぞれのデータサーバの状態に従って、アップグレード制御サーバはローリング方式で、データサーバを選択し、当該データサーバに、アップグレード動作を実行するよう通知してよい。アップグレード制御サーバがそれぞれのデータサーバのアップグレードを具体的に制御するプロセスについては、図３の説明が参照されてよい。

【０１３０】

もちろん、本開示のこの実施形態において、データサーバは、アップグレード制御サーバにより送信される、アップグレード準備状態を終了するようにとの終了通知を更に受信し、当該終了通知に従ってアップグレード準備状態を終了してよい。一方で、データサーバは、クライアントがアップグレード準備状態を終了するよう、当該終了通知に従って終了通知をクライアントへ送信してよい。次いで、データサーバは、通常の処理論理に従ってクライアントの書き込み要求を処理する。

【０１３１】

本開示のこの実施形態では、本開示における分散記憶システムをアップグレードするための方法がデータサーバ側で導入される。クライアントにより送信される、受信される第１フィードバック情報または第２フィードバック情報に従って、データサーバは、自らがアップグレード可能な状態にあるのか、アップグレード不可能な状態にあるのかを判定する。第１フィードバック情報または第２フィードバック情報は、クライアントが複数のデータサーバに対して同じ書き込み対象データに関する書き込み要求を送信した後、成功した書き込みの数と所定の数とを比較した結果に従って取得され、当該状態にはアップグレード可能な状態およびアップグレード不可能な状態が含まれ、当該状態は、アップグレード制御サーバがローリング方式で、データサーバを選択し、当該データサーバに、アップグレード動作を実行するよう通知するために使用される。従って、クライアントがデータサーバの状態を制御することにより、任意のクライアントの書き込み対象データがバックアップのために少なくとも所定の数のデータサーバへ書き込まれることが保証される。アップグレード制御サーバが、ローリング方式でアップグレードするようそれぞれのデータサーバを制御する場合は、高水準サービスが停止する必要はない。更には、分散記憶システムからクライアントへのより短い応答時間が保証され得ることで、データの信頼性が高まり、ユーザデータの損失という危険が著しく低下する。

【０１３２】
図３は、本開示の幾つかの実施形態に係る、分散記憶システムをアップグレードするための方法を示すフロー図である。具体的に言うと、当該方法は以下のステップを含んでよい。
【０１３３】
ステップ３１０：それぞれのデータサーバの状態を取得する。当該状態にはアップグレード可能な状態およびアップグレード不可能な状態が含まれ、それぞれのデータサーバは一方の状態を有する。データサーバの状態は、第１フィードバック情報または第２フィードバック情報に従って判定され、第１フィードバック情報または第２フィードバック情報は、クライアントが複数のデータサーバに対して同じ書き込み対象データに関する書き込み要求を送信した後に、成功した書き込みの数と所定の数とを比較した結果に従って取得される。
【０１３４】
図１および図２の説明を参照すると、分散記憶システム内のそれぞれのデータサーバは、クライアントによりフィードバックされる第１フィードバックメッセージおよび／または第２フィードバックメッセージに従って自らの状態を判定してよい。当該状態にはアップグレード可能な状態およびアップグレード不可能な状態が含まれる。
【０１３５】
それぞれのデータサーバは、同時に一方の状態だけを有し得る。例えば、特定のデータサーバがアップグレード可能な状態にある場合は、当該データサーバが他の状態を有することはできない。他の状況についても同じことが当てはまり、本明細書で再び説明されることはない。
【０１３６】
次いで、本開示のこの実施形態では、アップグレード制御サーバがそれぞれのデータサーバの状態を取得できる。
【０１３７】
アップグレード制御サーバがデータサーバの状態を取得するための具体的な取得方法は多数あり得るが、本開示で限定されることはない。
【０１３８】
ステップ３２０：アップグレード可能な状態にある少なくとも１つのデータサーバに、アップグレード動作を実行するようローリング方式で通知する。この通知に従って、当該データサーバはアップグレード動作を実行する。
【０１３９】
本開示のこの実施形態では、それぞれのデータサーバがクライアントのフィードバック情報に従って自らの状態を設定するので、アップグレード制御サーバが、アップグレード動作を実行するようアップグレード可能な状態にあるデータサーバをローリング方式で制御してよい。
【０１４０】
例えば、アップグレード可能なデータサーバの群が毎回選択され、データサーバのこの群がアップグレード動作を実行するよう通知される。データサーバのこの群は、アップグレード通知を受信した後に再起動およびアップグレードされてよい。
【０１４１】
本開示の別の実施形態において、ステップ３２０は、サブステップ３２１および３２２を含む。
【０１４２】
サブステップ３２１：毎回アップグレード可能な状態にある少なくとも１つのデータサーバを選択し、アップグレード可能な状態にある当該少なくとも１つのデータサーバに、アップグレード動作を実行するよう通知する。
【０１４３】
例えば、アップグレード可能な状態にあるデータサーバには、Ｕ１、Ｕ２、Ｕ３、Ｕ１

０、Ｕ１１・・・Ｕ２０等が含まれる。次いで、アップグレード制御サーバが、Ｋ個のデータサーバ、例えば３つのデータサーバを毎回そこから選択し、当該データサーバに、アップグレード動作を実行するよう通知してよい。Ｋは０より大きい整数である。もちろん、Ｋは実際の必要に従って設定されてよく、本開示により限定されることはない。

【０１４４】

サブステップ３２２：アップグレード可能な状態にある少なくとも１つのデータサーバがアップグレード動作を完全に終了しているかどうかをモニタリングし、終了している場合はサブステップ３２１へと移行する。

【０１４５】

例えば、アップグレード制御サーバは、前述のデータサーバＵ１、Ｕ２およびＵ３に、アップグレード動作を実行するよう通知し、次いで、データサーバＵ１、Ｕ２およびＵ３は、再起動およびアップグレードされる。データサーバが成功裏にアップグレードされた後、当該データサーバは自らの状態をアップグレードされた状態、例えばＤＯＮＥに変更し得る。

【０１４６】

次いで、アップグレード制御サーバは、これらのデータサーバの状態がＤＯＮＥであるかどうかをモニタリングしてよく、ＤＯＮＥであればアップグレードは成功である。

【０１４７】

もちろん、実際の適用においては、１つまたは複数のデータサーバが、データサーバのこの群のアップグレード中にアップグレードに失敗することもある。アップグレード制御サーバは次いで、データサーバがアップグレードに成功したのか、またはアップグレードに失敗したのかをモニタリングする。実際の適用において、機械の再起動に失敗した、または再起動後のシステムバージョンが変わらないといった状況が起きた場合は、アップグレードに失敗したと判定されてよい。

【０１４８】

アップグレード動作を終了する前述のステップがアップグレードの成功およびアップグレードの失敗を両方とも含んでよいことに留意すべきである。アップグレード動作を完全に終了するステップは、全てのデータサーバが成功裏にアップグレードされること、および、幾つかのデータサーバだけが成功裏にアップグレードされた場合に残りのデータサーバがアップグレードに失敗することを伴う。

【０１４９】

Ｕ１、Ｕ２およびＵ３の状態が全てＤＯＮＥであれば、次回に備えてアップグレード可能な状態にある少なくとも１つのデータサーバを選択すること、および、アップグレード可能な状態にある当該少なくとも１つのデータサーバに、アップグレード動作を実行するよう通知することを行うステップへとフローが移行する。

【０１５０】

アップグレード制御サーバにより通知される全てのデータサーバがアップグレード動作を完全に終了していることをアップグレード制御サーバがモニタリングした場合は、アップグレード制御サーバは、データサーバの次の群をアップグレード可能な状態にあるデータサーバからアップグレードのためにローリング方式で選択してよい。

【０１５１】

このようにして、極度に頻繁なローリングが回避され、データの消失という危険が低下する。

【０１５２】

本開示の別の実施形態において、当該方法は、データサーバがアップグレード動作を実行した後に以下のステップを更に含む。

【０１５３】

サブステップ３２３：任意のデータサーバのアップグレード動作の結果がアップグレードの失敗であることがモニタリングされた場合は、当該データサーバをアップグレードブラックリストに追加し、当該データサーバのアップグレードを中断する。

【０１５４】
特定のデータサーバがアップグレードに失敗した場合は、アップグレード制御サーバが当該データサーバをアップグレードブラックリストに追加し、当該データサーバのアップグレードを中断する。これらのデータサーバは次いで、オフラインになるのを待つか、または手動での修理を待つ。
【０１５５】
一実施形態において、本開示のこの実施形態における複数のデータサーバは、少なくとも２つのラック上に設置されてよい。実際の適用においては、もちろん、分散記憶システム内の様々なデータサーバが複数のラック上に設置されてよく、１つのラックは、１つのデータサーバ・サブクラスタ用である。
【０１５６】
更に、ステップ３２０はサブステップＣ１１を含む。
【０１５７】
サブステップＣ１１：毎回アップグレード可能な状態にあるデータサーバを最も多く有するラックを選択し、当該ラック内のデータサーバに、アップグレード動作を実行するよう通知する。この通知に従って、当該ラック内のそれぞれのデータサーバは、自らの状態をチェックする。データサーバがアップグレード可能な状態にある場合は、当該データサーバが再起動およびアップグレードされる。データサーバがアップグレード不可能な状態またはアップグレード終了状態にある場合は、当該データサーバが再起動およびアップグレードを拒否する。
【０１５８】
本開示のこの実施形態では、データサーバがラック上に設置され、データサーバの群が１つのラック上に設置される。しかしながら、多数のクライアントについては、異なるラックの異なるデータサーバにアクセスしてよい。それぞれのラックがアップグレード可能な状態にあるデータサーバを有してもよいし、アップグレード不可能な状態にあるデータサーバを有してもよいし、アップグレードを終了するデータサーバを有してもよい。
【０１５９】
アップグレード制御サーバがアップグレード通知を送信しやすくすべく、本開示のこの実施形態では、ラック単位でアップグレード通知が送信される。例えば、ラックのＩＰセグメントは、２００．２００．２００．＊＊＊であり、本開示のこの実施形態におけるアップグレード制御サーバは、２００．２００．２００．＊＊＊に関する通知を１つ生成すること、および、当該ラックのそれぞれのデータサーバが当該通知を受信できるよう、当該通知を当該ラックにブロードキャストすることを行う必要がある。
【０１６０】
次いで、ラック内のデータサーバが前述のアップグレード通知を受信した後、データサーバはまず、自らの状態がＯＫであるかどうかを判定する。状態がＯＫであれば、データサーバは、再起動およびアップグレードされる。代わりに状態がＨＯＬＤまたはＤＯＮＥであれば、データサーバは、再起動およびアップグレードを拒否する。
【０１６１】
本開示のこの実施形態では、アップグレードを迅速に終了すべく、ＯＫ状態にあるデータサーバを最も多く有するラックが選択され、当該ラックに対してアップグレード通知が送信される。
【０１６２】
例えば、アップグレード可能な状態にあるデータサーバには、Ｕ１、Ｕ２、Ｕ３、Ｕ１０、Ｕ１１・・・Ｕ２０等が含まれる。次いで、Ｋ個のデータサーバ、例えば３つのデータサーバが毎回そこから選択されてよく、これらは、アップグレード動作を実行するよう通知される。
【０１６３】
一実施形態において、当該方法は、サブステップＣ１１の後にサブステップＣ１２を更に含む。

【０１６４】
サブステップＣ１２：ラック内のデータサーバがアップグレード動作を完全に終了しているかどうかをモニタリングし、終了している場合はアップグレード通知サブモジュールへと移行する。
【０１６５】
次いで、ラック内のアップグレード可能な状態にあるデータサーバのアップグレード動作が全て終了していることがモニタリングされた場合は、次のラックが選択されてよく、次のラックのデータサーバに対してアップグレード通知が送信される。この周期は、全てのデータサーバがアップグレードを終了するまで繰り返す。
【０１６６】
前述のプロセスにより、アップグレードに失敗したデータサーバを除く全てのデータサーバがアップグレードを終了していることをアップグレード制御サーバがモニタリングした場合、アップグレード制御サーバは、それぞれのデータサーバにアップグレード準備状態を終了するよう通知し、通常の処理論理に戻ってよい。次いで、それぞれのデータサーバは、当該データサーバへアクセスしているクライアントにアップグレード準備状態を終了するよう通知する。次いで、クライアントは、通常の処理論理に戻る。ステップ３１０および３２０が実行される必要はない。
【０１６７】
本開示のこの実施形態では、本開示における分散記憶システムをアップグレードするための方法がアップグレード制御サーバ側で導入される。アップグレード制御サーバが、ローリング方式でアップグレードするようデータサーバを制御する場合は、高水準サービスが停止する必要はない。なぜなら、クライアントがデータサーバの状態を制御し、任意のクライアントの書き込み対象データがバックアップのために少なくとも所定の数のデータサーバへ書き込まれることが保証されるからである。更には、分散記憶システムからクライアントへのより短い応答時間が保証され得ることで、データの信頼性が高まり、ユーザデータの損失という危険が著しく低下する。更に、当該方法は、サービスが影響を受けないように保証しつつ、アップグレード中の機械の予期せぬ例外を許容できる。加えて、大量のデータを移行する必要なしに迅速なアップグレードが実現される。
【０１６８】
図４は、本開示の幾つかの実施形態に係る、分散記憶システムをアップグレードするための方法を示すフロー図である。具体的に言うと、当該方法は以下のステップを含んでよい。
【０１６９】
ステップ４１０：クライアントが、同じ書き込み対象データに関する複数のデータサーバに対して書き込み要求を送信する。
【０１７０】
ステップ４１２：クライアントは、データサーバの各々により返される応答を受信し、当該応答に従って、成功した書き込みの数が所定の数より多いかどうかを判定する。成功した書き込みの数が所定の数より多い場合は、ステップ４１４へと移行する。成功した書き込みの数が所定の数より多くない場合は、ステップ４１６へと移行する。
【０１７１】
ステップ４１４：クライアントは、成功した書き込みを有するそれぞれのデータサーバに対して第１フィードバック情報を送信する。
【０１７２】
ステップ４１６：クライアントは、成功した書き込みを有するそれぞれのデータサーバに対して第２フィードバック情報を送信する。
【０１７３】
本開示の別の実施形態において、第１フィードバック情報および第２フィードバック情報は、クライアント識別子を含む。成功した書き込みを有するそれぞれのデータサーバに対して第２フィードバック情報を送信するステップの後、当該方法は更に以下のステップ

を含む。
【０１７４】
ステップ４１７（図示せず）：新たな書き込み対象データが現れた場合は、以前に成功した書き込みがあったデータサーバを含む当該複数のデータサーバが対象として使用され、当該複数のデータサーバに対して同じ書き込み対象データに関する書き込み要求が送信される。
【０１７５】
ステップ４１８：データサーバは、クライアントにより送信される第１フィードバック情報または第２フィードバック情報を受信する。
【０１７６】
ステップ４２０：データサーバがアップグレードされていない状況では、第１フィードバック情報または第２フィードバック情報に従ってデータサーバの状態を判定する。当該状態にはアップグレード可能な状態およびアップグレード不可能な状態が含まれる。
【０１７７】
一実施形態において、ステップ４２０は、ステップ４１７に基づいてサブステップＤ１１～Ｄ１６を含む。
【０１７８】
サブステップＤ１１：第２フィードバック情報を受信すると、第２フィードバック情報内のクライアント識別子をアップグレード不可能なリストに書き込んで、自らの状態をアップグレード不可能なものとしてマーキングする。
【０１７９】
サブステップＤ１２：アップグレード不可能なリスト内のクライアント識別子については、対応するクライアントの第２フィードバックメッセージが所定の数の期間内に受信されなかったかどうかを判定する。対応するクライアントの第２フィードバックメッセージが所定の数の期間内に受信されなかった場合は、サブステップＤ１３へと移行する。対応するクライアントの第２フィードバックメッセージが所定の数の期間内に受信された場合は、アップグレード不可能な状態が維持される。
【０１８０】
サブステップＤ１３：クライアント識別子をアップグレード不可能なリストから削除する。サブステップＤ１５へと移行する。
【０１８１】
サブステップＤ１４：アップグレード不可能なリストが第１フィードバック情報内のクライアント識別子を含むかどうかを判定し、含む場合はサブステップＤ１３へと移行する。アップグレード不可能なリストが第１フィードバック情報内のクライアント識別子を含まない場合は、フローがサブステップＤ１５へと移行する。
【０１８２】
サブステップＤ１５：アップグレード不可能なリストが空であるかどうかを判定し、空であればサブステップＤ１６へと移行する。
【０１８３】
サブステップＤ１６：自らの状態をアップグレード可能なものとしてマーキングする。
【０１８４】
ステップ４２２：アップグレード制御サーバがそれぞれのデータサーバの状態を取得する。
【０１８５】
ステップ４２４：アップグレード制御サーバは、アップグレード可能な状態にある少なくとも１つのデータサーバに、アップグレード動作を実行するようローリング方式で通知する。
【０１８６】
ステップ４２６：この通知に従って、データサーバは、アップグレード動作を実行する。

【０１８７】
もちろん、この実施形態において、クライアント側のステップの原理については、図１の説明が参照されてよく、データサーバ側のステップの原理については、図２の説明が参照されてよく、アップグレード制御サーバ側のステップの原理については、図３の説明が参照されてよい。本明細書でこの説明が再び提供されることはない。
【０１８８】
本開示のこの実施形態では、本開示における分散記憶システムをアップグレードするための方法が、クライアント、データサーバおよびアップグレード制御サーバを含む３つの側面で導入される。　分散記憶システムへアクセスしているクライアントについては、それぞれのクライアントが、同じ書き込み対象データに関する書き込み要求を複数のデータサーバへ同時に送信し、次いで、幾つのデータサーバが成功した書き込みを有しているかが解析され、成功した書き込みの数が所定の数より多いかどうかが判定され、その判定結果に従って、成功した書き込みを有するそれぞれのデータサーバに対して第１フィードバック情報または第２フィードバック情報が送信される。受信される第１フィードバック情報または第２フィードバック情報に従って、データサーバは、自らがアップグレード可能な状態にあるのか、アップグレード不可能な状態にあるのかを判定する。アップグレード制御サーバは、アップグレード動作を実行するようアップグレード可能な状態にあるデータサーバをローリング方式で制御してよい。前述のプロセスにより、アップグレード制御サーバが、ローリング方式でアップグレードするようアップグレード可能な状態にあるデータサーバを制御する場合は、高水準サービスが停止する必要はない。なぜなら、クライアントがデータサーバの状態を制御し、任意のクライアントの書き込み対象データがバックアップのために少なくとも所定の数のデータサーバへ書き込まれることが保証されるからである。更には、分散記憶システムからクライアントへのより短い応答時間が保証され得ることで、データの信頼性が高まり、ユーザデータの損失という危険が著しく低下する。更に、当該方法は、サービスが影響を受けないように保証しつつ、アップグレード中の機械の予期せぬ例外を許容できる。
【０１８９】
図５は、本開示の幾つかの実施形態に係る、分散記憶システムをアップグレードするための装置のブロック図である。具体的に言うと、当該装置は、要求送信モジュール５１０、判定モジュール５２０、第１フィードバックモジュール５３０および第２フィードバックモジュール５４０を含んでよい。
【０１９０】
要求送信モジュール５１０は、複数のデータサーバに対して同じ書き込み対象データに関する書き込み要求を送信するように構成される。
【０１９１】
要求送信モジュール５１０に加えて、本開示の別の実施形態における当該装置は、データサーバへアクセスするときに当該データサーバにより送信される第２アップグレード通知を受信するように構成される第２アップグレード通知受信モジュールと、第２アップグレード通知に従ってアップグレード準備状態へと移行するように構成される第２アップグレード準備モジュールとを更に備える。
【０１９２】
判定モジュール５２０は、当該データサーバの各々により返される応答を受信すること、および、当該応答に基づいて、成功した書き込みの数が所定の数より多いかどうかを判定することを行うように構成される。
【０１９３】
第１フィードバックモジュール５３０は、成功した書き込みの数が所定の数より多い場合に、成功した書き込みを有するそれぞれのデータサーバに対して第１フィードバック情報を送信するように構成される。
【０１９４】
第２フィードバックモジュール５４０は、成功した書き込みの数が所定の数より多くな

い場合に、成功した書き込みを有するそれぞれのデータサーバに対して第２フィードバック情報を送信するように構成される。ここで、第１フィードバック情報または第２フィードバック情報は、データサーバが自らの状態を判定するために使用され、当該状態にはアップグレード可能な状態およびアップグレード不可能な状態が含まれ、データサーバの状態は、アップグレード制御サーバがデータサーバに、アップグレード動作を実行するようローリング方式で通知するために使用される。

【０１９５】

本開示の別の実施形態において、第２フィードバックモジュール５４０は、成功した書き込みの現在の数が所定の数と等しい場合に、成功した書き込みを有するそれぞれのデータサーバに対して第２フィードバック情報を送信するように構成される第２フィードバック情報送信サブモジュールと、成功した書き込みの数が所定の数より少ない場合に、当該複数のデータサーバ以外の少なくとも１つのデータサーバに対して書き込み対象データに関する書き込み要求を送信するように構成される書き込み要求送信サブモジュールと、当該少なくとも１つのデータサーバにより返される応答を受信し、当該応答に従って、成功した書き込みの現在の数が成功した書き込みの以前の数と組み合わせて所定の数と等しくなるかどうかを判定すること、成功した書き込みの現在の数が所定の数と等しい場合に、第２フィードバック情報送信サブモジュールへと移行すること、第２フィードバック情報内のクライアント識別子をアップグレード不可能なリストに書き込んで、自らの状態をアップグレード不可能なものとしてマーキングすることを行うように構成される判定サブモジュールとを備える。

【０１９６】

データサーバが第１フィードバック情報を受信した後、第１フィードバック情報は、データサーバが、第１フィードバック情報内のクライアント識別子をアップグレード不可能なリストから削除すること、および、アップグレード不可能なリストが空であると判定した後に自らの状態をアップグレード可能なものとしてマーキングすることを行うために使用される。

【０１９７】

図６は、本開示の幾つかの実施形態に係る、分散記憶システムをアップグレードするための装置のブロック図である。具体的に言うと、当該装置は、フィードバック情報受信モジュール６１０および状態判定モジュール６２０を含んでよい。

【０１９８】

フィードバック情報受信モジュール６１０は、クライアントから送信される第１フィードバック情報または第２フィードバック情報を受信するように構成される。ここで、第１フィードバック情報または第２フィードバック情報は、クライアントが複数のデータサーバに対して同じ書き込み対象データに関する書き込み要求を送信した後に、成功した書き込みの数と所定の数とを比較した結果に従って取得される。

【０１９９】

フィードバック情報受信モジュール６１０に加えて、本開示の別の実施形態における当該装置は、アップグレード制御サーバにより送信される第１アップグレード通知を受信するように構成される第１アップグレード通知受信モジュールと、第１アップグレード通知に従ってアップグレード準備状態へと移行すること、および、クライアントがアップグレード準備状態へと移行するよう、クライアントのアクセス要求を受信した後に第２アップグレード通知をクライアントへ送信することを行うように構成される第１アップグレード準備モジュールとを更に備える。

【０２００】

状態判定モジュール６２０は、データサーバがアップグレードされていない状況において、第１フィードバック情報または第２フィードバック情報に従ってデータサーバの状態を判定するように構成される。当該状態にはアップグレード可能な状態およびアップグレード不可能な状態が含まれ、当該状態は、アップグレード制御サーバがローリング方式で、データサーバを選択し、当該データサーバに、アップグレード動作を実行するよう通知

するために使用される。
【０２０１】
本開示の別の実施形態において、状態判定モジュール６２０は、アップグレード可能な判定サブモジュールおよびアップグレード不可能な状態判定サブモジュールを備える。
【０２０２】
アップグレード可能な判定サブモジュールは、第１フィードバック情報を受信すると、第１フィードバック情報内のクライアント識別子をアップグレード不可能なリストから削除すること、および、アップグレード不可能なリストが空であると判定した後に自らの状態をアップグレード可能なものとしてマーキングすることを行うように構成される。
【０２０３】
一実施形態において、アップグレード可能な判定サブモジュールは、アップグレード不可能なリストが第１フィードバック情報内のクライアント識別子を含むかどうかを判定すること、および、含む場合は第１削除サブモジュールへと移行することを行うように構成されるクライアント識別子判定サブモジュールであって、当該第１削除サブモジュールは、クライアント識別子をアップグレード不可能なリストから削除するように構成される、クライアント識別子判定サブモジュールと、アップグレード不可能なリストが空であるかどうかを判定すること、および、空であればアップグレード可能なマーキングサブモジュールへと移行することを行うように構成されるアップグレード不可能なリスト判定サブモジュールであって、当該アップグレード可能なマーキングサブモジュールは、自らの状態をアップグレード可能なものとしてマーキングするように更に構成される、アップグレード不可能なリスト判定サブモジュールとを備える。
【０２０４】
アップグレード不可能な状態判定サブモジュールは、第２フィードバック情報を受信すると第２フィードバック情報内のクライアント識別子をアップグレード不可能なリストに書き込むこと、および、自らの状態をアップグレード不可能なものとしてマーキングすることを行うように構成される。
【０２０５】
アップグレード不可能な状態判定サブモジュールに加えて、本開示の別の実施形態における当該装置は、アップグレード不可能なリスト内のクライアント識別子について、対応するクライアントの第２フィードバックメッセージが所定の数の期間内に受信されなかったかどうかを判定すること、および、対応するクライアントの第２フィードバックメッセージが所定の数の期間内に受信されなかった場合に、第２削除サブモジュールへと移行することを行うように構成される時間判定サブモジュールであって、当該第２削除サブモジュールは、クライアント識別子をアップグレード不可能なリストから削除するように構成される、時間判定サブモジュールを更に備える。
【０２０６】
図７は、本開示の幾つかの実施形態に係る、分散記憶システムをアップグレードするための装置のブロック図である。具体的に言うと、当該装置は状態取得モジュール７１０およびアップグレード通知モジュール７２０を含んでよい。
【０２０７】
状態取得モジュール７１０は、データサーバごとに状態を取得するように構成される。当該状態にはアップグレード可能な状態およびアップグレード不可能な状態が含まれ、それぞれのデータサーバは一方の状態を有する。データサーバの状態は、第１フィードバック情報または第２フィードバック情報に従って判定され、第１フィードバック情報または第２フィードバック情報は、クライアントが複数のデータサーバに対して同じ書き込み対象データに関する書き込み要求を送信した後に、成功した書き込みの数と所定の数とを比較した結果に従って取得される。
【０２０８】
状態取得モジュール７１０に加えて、本開示の別の実施形態における当該装置は、それぞれのデータサーバがアップグレード準備状態へと移行するよう、第１アップグレード通

知をそれぞれのデータサーバへ送信するように構成されるアップグレード通知送信モジュールを更に備える。それぞれのデータサーバは、クライアントがアップグレード準備状態へと移行するよう、クライアントのアクセス要求を受信した後に第２アップグレード通知をクライアントへ送信する。

【０２０９】

アップグレード通知モジュール７２０は、アップグレード可能な状態にある少なくとも１つのデータサーバに、アップグレード動作を実行するようローリング方式で通知するように構成される。この通知に従って、当該データサーバはアップグレード動作を実行する。

【０２１０】

本開示の別の実施形態において、アップグレード通知モジュールは、毎回アップグレード可能な状態にある少なくとも１つのデータサーバを選択すること、および、アップグレード可能な状態にある当該少なくとも１つのデータサーバに、アップグレード動作を実行するよう通知することを行うように構成される第１選択サブモジュールと、アップグレード可能な状態にある当該少なくとも１つのデータサーバがアップグレード動作を完全に終了しているかどうかをモニタリングすること、および、終了している場合は第１選択サブモジュールへと移行することを行うように構成される第１モニタリングサブモジュールとを備える。

【０２１１】

第１モニタリングサブモジュールに次いで、本開示の別の実施形態における当該装置は、任意のデータサーバのアップグレード動作の結果がアップグレードの失敗であることがモニタリングされた場合に、データサーバをアップグレードブラックリストに追加すること、および、データサーバのアップグレードを中断することを行うように構成される中断サブモジュールを更に備える。

【０２１２】

本開示の別の実施形態において、データサーバは、少なくとも２つのラック上に設置される。次いで、アップグレード通知モジュール７２０は、毎回アップグレード可能な状態にあるデータサーバを最も多く有するラックを選択すること、および、当該ラック内のデータサーバに、アップグレード動作を実行するよう通知することを行うように構成されるアップグレード通知サブモジュールを備える。この通知に従って、当該ラック内のそれぞれのデータサーバは、自らの状態をチェックする。データサーバがアップグレード可能な状態にある場合は、当該データサーバが再起動およびアップグレードされる。データサーバがアップグレード不可能な状態またはアップグレード終了状態にある場合は、当該データサーバが再起動およびアップグレードを拒否する。

【０２１３】

アップグレード通知サブモジュールに加えて、本開示の別の実施形態における当該装置は、ラック内のデータサーバがアップグレード動作を完全に終了しているかどうかをモニタリングすること、および、終了している場合はアップグレード通知サブモジュールへと移行することを行うように構成されるモニタリングサブモジュールを更に備える。

【０２１４】

図８および図８Ａを参照されたい。これらの図は、本開示の幾つかの実施形態に係る、分散記憶システムをアップグレードするためのシステムのブロック図である。具体的に言うと、当該システムは、以下のモジュール、すなわちクライアント８１０、データサーバ８２０およびアップグレード制御サーバ８３０を含んでよい。

【０２１５】

図８Ａは、一実施形態における分散記憶システムのアーキテクチャを示す概略図である。本開示のこの実施形態において、それぞれのクライアント８１０は、分散記憶システム内のＲ個のデータサーバ８２０に対して書き込み要求を送信してよく、アップグレード制御サーバ８３０は、全てのデータサーバのアップグレードプロセスを制御する。

【０２１６】

図８は、クライアント８１０、データサーバ８２０およびアップグレード制御サーバ８３０間の接続関係を示している。

【０２１７】

クライアント８１０は、複数のデータサーバに対して同じ書き込み対象データに関する書き込み要求を送信するように構成される要求送信モジュール８１１と、当該データサーバの各々により返される応答を受信すること、および、当該応答に基づいて、成功した書き込みの数が所定の数より多いかどうかを判定することを行うように構成される判定モジュール８１２と、成功した書き込みの数が所定の数より多い場合に、成功した書き込みを有するそれぞれのデータサーバに対して第１フィードバック情報を送信するように構成される第１フィードバックモジュール８１３と、成功した書き込みの数が所定の数より多くない場合に、成功した書き込みを有するそれぞれのデータサーバに対して第２フィードバック情報を送信するように構成される第２フィードバックモジュール８１４とを備える。

【０２１８】

データサーバ８２０は、クライアントの書き込み要求を受信すること、および、当該クライアントに応答を返すことを行うように構成されるデータ記憶モジュール８２１と、当該クライアントにより送信される第１フィードバック情報または第２フィードバック情報を受信するように構成されるフィードバック情報受信モジュール８２２と、データサーバ自らがアップグレードされなかった場合に、第１フィードバック情報または第２フィードバック情報に従ってデータサーバ自らの状態を判定するように構成される状態判定モジュール８２３と、アップグレード制御サーバの通知に従ってアップグレード動作を実行するように構成されるアップグレードモジュール８２４とを備える。

【０２１９】

アップグレード制御サーバ８３０は、それぞれのデータサーバの状態を取得するように構成される状態取得モジュール８３１と、アップグレード可能な状態にある少なくとも１つのデータサーバに、アップグレード動作を実行するようローリング方式で通知するように構成されるアップグレード通知モジュール８３２とを備える。

【０２２０】

本開示のこの実施形態において、クライアントのモジュールについては、図５の説明が参照されてよい。データサーバのモジュールについては、図６の説明が参照されてよい。アップグレード制御サーバのモジュールについては、図７の説明が参照されてよい。これらの構成要素により実行される方法は、前述の図で説明される方法と同様であり、本明細書で再び説明されることはない。

【０２２１】

本開示のこの実施形態では、本開示における分散記憶システムをアップグレードするための方法が、クライアント、データサーバおよびアップグレード制御サーバを含む３つの側面で導入される。　分散記憶システムへアクセスしているクライアントについては、それぞれのクライアントが、同じ書き込み対象データに関する書き込み要求を複数のデータサーバへ同時に送信し、次いで、幾つのデータサーバが成功した書き込みを有しているかが解析され、成功した書き込みの数が所定の数より多いかどうかが判定され、その判定結果に従って、成功した書き込みを有するそれぞれのデータサーバに対して第１フィードバック情報または第２フィードバック情報が送信される。受信される第１フィードバック情報または第２フィードバック情報に従って、データサーバは、自らがアップグレード可能な状態にあるのか、アップグレード不可能な状態にあるのかを判定する。アップグレード制御サーバは、アップグレード動作を実行するようアップグレード可能な状態にあるデータサーバをローリング方式で制御してよい。前述のプロセスにより、アップグレード制御サーバが、ローリング方式でアップグレードするようアップグレード可能な状態にあるデータサーバを制御する場合は、高水準サービスが停止する必要はない。なぜなら、クライアントがデータサーバの状態を制御し、任意のクライアントの書き込み対象データがバックアップのために少なくとも所定の数のデータサーバへ書き込まれることが保証されるからである。更には、分散記憶システムからクライアントへのより短い応答時間が保証され

得ることで、データの信頼性が高まり、ユーザデータの損失という危険が著しく低下する。更に、当該方法は、サービスが影響を受けないように保証しつつ、アップグレード中の機械の予期せぬ例外を許容できる。

【０２２２】

装置の実施形態は、方法の実施形態と同様なので、比較的簡潔に説明される。関連する部分については、方法の実施形態の説明の部分が参照されてよい。

【０２２３】

本明細書内の実施形態は、それぞれの実施形態が他の実施形態とは異なる部分を強調する形で漸次説明される。相互参照により、当該実施形態の同一の部分または同様の部分が取得されてよい。

【０２２４】

当業者であれは、本開示の実施形態が方法、装置またはコンピュータプログラム製品として提供され得ることを理解するはずである。従って、本開示の実施形態は、完全なハードウェアの実施形態、完全なソフトウェアの実施形態、またはソフトウェアとハードウェアとを組み合わせた実施形態として実装されてよい。更に、本開示の実施形態は、内部にコンピュータ使用可能プログラムコードを含む１つまたは複数のコンピュータ使用可能記憶媒体（これらに限定されるわけではないが、磁気ディスクメモリ、ＣＤ‐ＲＯＭ、光学メモリ等が含まれる）上で実装されるコンピュータプログラム製品の形を取ってよい。

【０２２５】

典型的な構成において、コンピュータデバイスは、１つまたは複数のプロセッサ（ＣＰＵ）、入出力インタフェース、ネットワークインタフェースおよびメモリを含む。メモリには、リードオンリメモリ（ＲＯＭ）またはフラッシュメモリ（ＦＬＡＳＨ　ＲＡＭ）といった、非永続的メモリ、ランダムアクセスメモリ（ＲＡＭ）および／または不揮発性メモリ等の形を取った非一時的コンピュータ可読媒体が含まれてよい。メモリは、コンピュータ可読媒体の例である。コンピュータ可読媒体には、任意の方法または技術により情報の記憶を実現できる永続的および非永続的な移動可能媒体並びに移動不可能媒体が含まれる。情報は、コンピュータ可読命令、データ構造、プログラムのモジュール、または他のデータであってよい。コンピュータの記憶媒体の例としては、これらに限定されるわけではないが、相変化メモリ（ＰＲＡＭ）、スタティックランダムアクセスメモリ（ＳＲＡＭ）、ダイナミックランダムアクセスメモリ（ＤＲＡＭ）もしくは他の種類のランダムアクセスメモリ（ＲＡＭ）、リードオンリメモリ（ＲＯＭ）、電気的消去可能プログラマブルリードオンリメモリ（ＥＥＰＲＯＭ）、フラッシュメモリもしくは他のメモリ技術、リードオンリ・コンパクトディスク・リードオンリメモリ（ＣＤ‐ＲＯＭ）、デジタル多用途ディスク（ＤＶＤ）もしくは他の光記憶装置、磁気カセット、磁気テープ、磁気ディスク記憶装置もしくは他の磁気記憶デバイス、または、コンピューティングデバイスによりアクセス可能な情報を記憶するために使用され得る任意の他の非伝送媒体が挙げられる。本明細書における定義を考慮すると、コンピュータ可読媒体に、変調されたデータ信号および搬送波等の非持続的コンピュータ可読媒体（一時的媒体）は含まれない。

【０２２６】

本開示の実施形態は、本開示の実施形態の方法、端末デバイス（システム）およびコンピュータプログラム製品に係るフロー図および／またはブロック図を参照しながら説明される。フロー図および／またはブロック図におけるそれぞれのフローおよび／またはそれぞれのブロック、並びに、フロー図および／またはブロック図におけるフロー同士および／またはブロック同士の組み合わせがコンピュータプログラム命令により実装され得ることを理解すべきである。これらのコンピュータプログラム命令は、機械を生成すべく、汎用コンピュータ、専用コンピュータ、組込みプロセッサ、または任意の他のプログラマブルデータ処理端末デバイスのプロセッサに対して提供されてよい。その結果、任意の他のプログラマブルデータ処理端末デバイスのコンピュータまたはプロセッサにより実行される命令は、フロー図における１つもしくは複数のプロセス、および／または、ブロック図における１つもしくは複数のブロックで指定された機能を実装するための装置を生成する

。

【０２２７】

これらのコンピュータプログラム命令は、コンピュータまたは別のプログラマブルデータ処理端末デバイスに特定のやり方で動作するよう指示できるコンピュータ可読メモリに記憶されてもよい。その結果、コンピュータ可読メモリに記憶される命令は、フロー図の１つもしくは複数のフロー、および／または、ブロック図の１つもしくは複数のブロックで指定された機能を実装する命令手段を含む製造品を生産する。

【０２２８】

これらのコンピュータプログラム命令は、コンピュータまたは任意の他のプログラマブルデータ処理端末デバイス上にロードされてもよい。その結果、一連の動作ステップがコンピュータまたは任意の他のプログラマブル端末デバイス上で実行されることにより、コンピュータ実装プロセスが生成される。従って、コンピュータまたは任意の他のプログラマブル端末デバイス上で実行される命令は、フロー図における１つもしくは複数のプロセス、および／または、ブロック図における１つもしくは複数のブロックで指定された機能を実装するためのステップを提供する。

【０２２９】

本開示の様々な実施形態が説明されてきたが、ひとたび基本的な創造的概念を説明しておけば、当業者はこれらの実施形態の他の変形例および改良例を作ることができる。従って、添付の請求項は、当該実施形態、並びに、本開示の範囲内に入る全ての変形例および改良例を含むものとして解釈されるように意図されている。

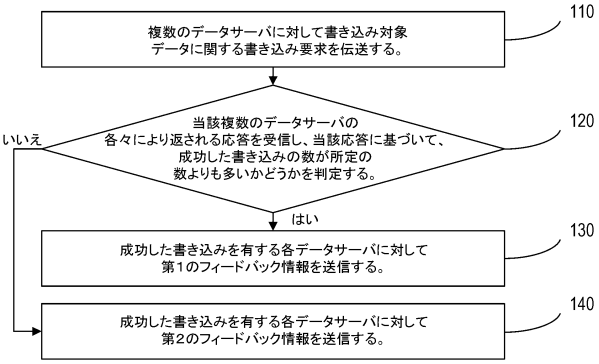
【０２３０】

最後に、本願における第１および第２といった関係用語は、単に１つのエンティティまたは動作を別のエンティティまたは動作と区別するために使用されているに過ぎず、これらのエンティティまたは動作がこうした実際の関係または順序を持つ必要があるわけでもなければ、それを持つと示唆しているわけでもないことを更に留意すべきである。更に「含む（ｉｎｃｌｕｄｅ）」、「含む（ｃｏｍｐｒｉｓｅ）」という用語、またはこれらの他の変形語は、非排他的な包含物をカバーするように意図されている。その結果、一連の要素を含むプロセス、方法、物品または端末デバイスは、当該要素を含むだけではなく、明記されていない他の要素も含むか、または、当該プロセス、当該方法、当該物品もしくは当該端末デバイスに本来備わっている要素を更に含む。「１つの・・・を含む（ｉｎｃｌｕｄｉｎｇ　ｏｎｅ・・・）」という記述により更なる制限なく定義される要素は、当該要素を含むプロセス、方法、物品または端末デバイスにおける追加的な同一要素の存在を除外するものではない。

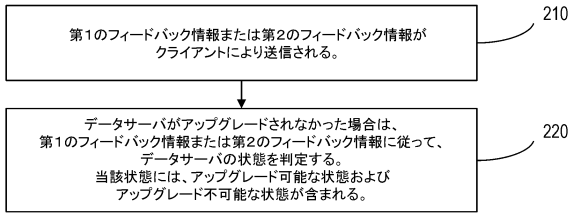
【０２３１】

本開示で提供される、分散記憶システムをアップグレードするための方法、分散記憶装置をアップグレードするための装置、および分散記憶装置をアップグレードするためのシステムが以上に詳しく紹介された。本明細書には、具体的な例を参照しながら本開示の原理および実装が記載されている。上記の実施形態の説明は、単に本開示の方法および本質的な着想を理解する手助けをするために提供されているに過ぎない。当業者であれば、本開示の着想に従って具体的な実装および実施形態に対する変更を加えることができる。上記を考慮すると、本明細書の内容は本開示を限定するものと解釈されるべきではない。

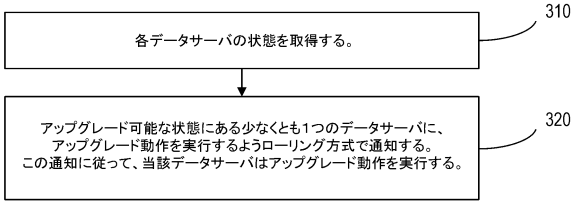
【図１】



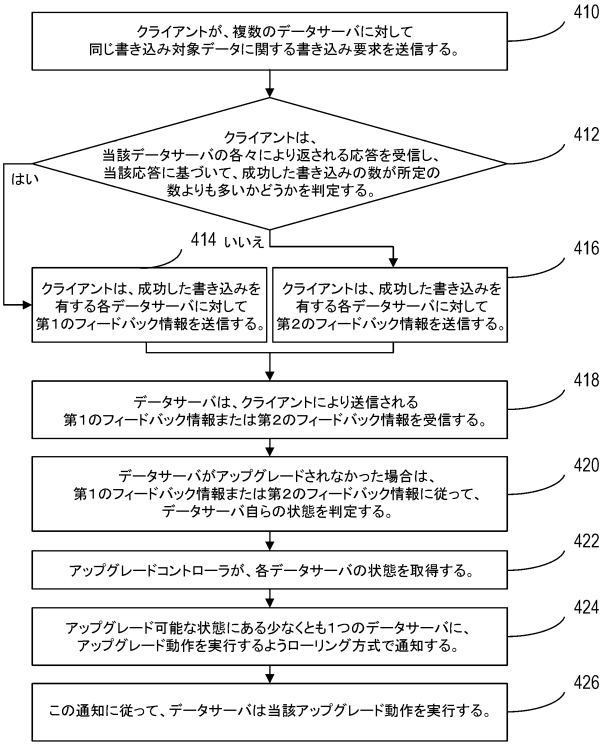
【図２】



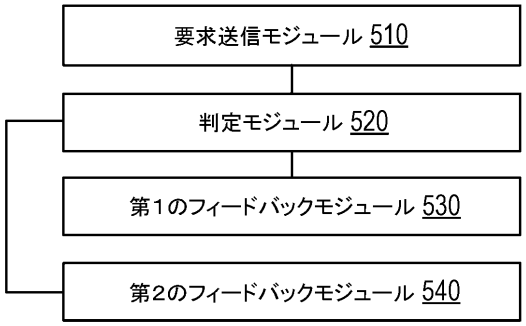
【図３】



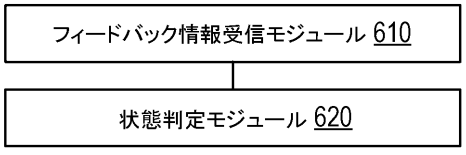
【図４】



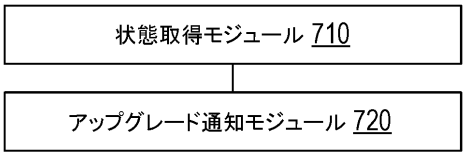
【図５】



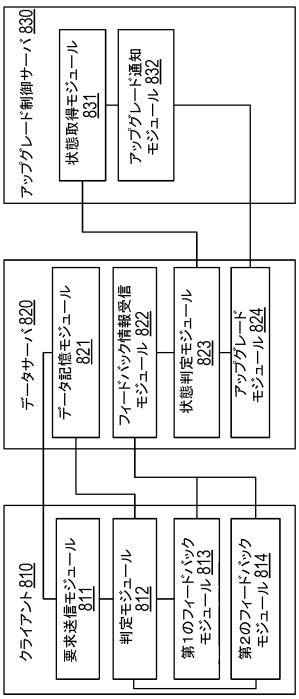
【図６】



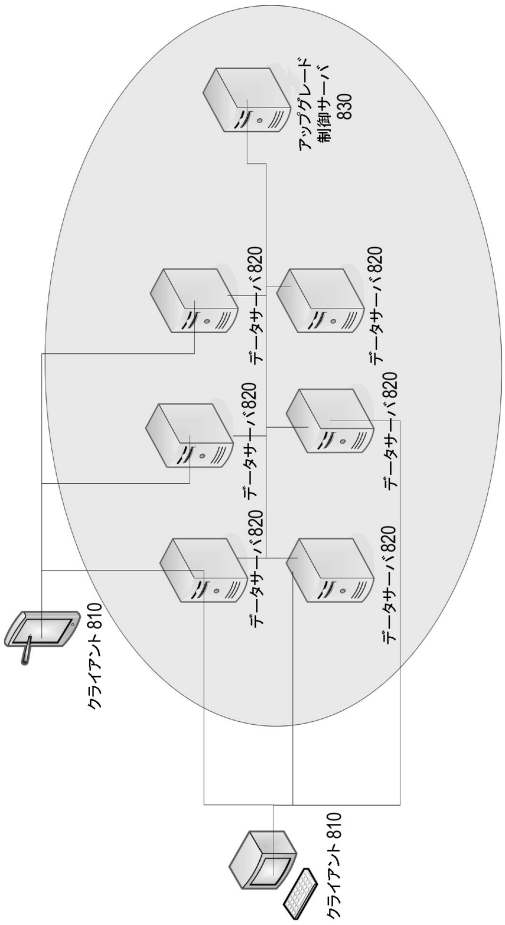
【図７】



【図８】



【図８Ａ】


アップグレード
制御サーバ
830
データサーバ820
データサーバ820
データサーバ820
データサーバ820
データサーバ820
データサーバ820
クライアント810
クライアント810

フロントページの続き

(72)発明者 チュー、ジャジ
英国領，ケイマン諸島，グランド・ケイマン，ジョージ・タウン，ワン・キャピタル・プレイス，フォース・フロア，ピー．オー．ボックス　８４７　アリババ・グループ・ホールディング・リミテッド内
(72)発明者 ジャオ、シューチ
英国領，ケイマン諸島，グランド・ケイマン，ジョージ・タウン，ワン・キャピタル・プレイス，フォース・フロア，ピー．オー．ボックス　８４７　アリババ・グループ・ホールディング・リミテッド内
(72)発明者 リン、ジャンビン
英国領，ケイマン諸島，グランド・ケイマン，ジョージ・タウン，ワン・キャピタル・プレイス，フォース・フロア，ピー．オー．ボックス　８４７　アリババ・グループ・ホールディング・リミテッド内
(72)発明者 グ、ユエシェン
英国領，ケイマン諸島，グランド・ケイマン，ジョージ・タウン，ワン・キャピタル・プレイス，フォース・フロア，ピー．オー．ボックス　８４７　アリババ・グループ・ホールディング・リミテッド内

審査官 久保　光宏

(56)参考文献 奥野幹也，「My SQL Cluster構築・運用バイブル～仕組みからわかる基礎と実践のノウハウ」，日本，株式会社技術評論社，２０１２年　４月２５日，初版，第４０４～４１１頁，ISBN: 978-4-7741-5053-6
西村豪生（外６名），「リソースプールを活用して無停止ソフトウェアアップデートを実現するセッション制御サーバミドルウェア」，電子情報通信学会論文誌B，日本，[online]，電子情報通信学会，２０１４年　５月　１日，Vol.J97-B，No.5，第３９３～４０６頁，ISSN: 1881-0209，[2014年5月1日検索]，インターネット，<URL: http://search.ieice.org/bin/pdf.php?lang=J&year=2014&fname=j97-b_5_393&abst=>
飯沢篤志（外１名），「データベースおもしろ講座 -10- 分散データベースシステム」，ｂｉｔ，日本，共立出版株式会社，１９９２年　２月　１日，1992年2月号（Vol.24，No.2），第１８１～１９２頁，ISSN: 0385-6984

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ８／６５－８／６５８
Ｇ０６Ｆ１６／００－１６／９５８
ＣＳＤＢ（日本国特許庁）
ＩＥＥＥＸｐｌｏｒｅ（ＩＥＥＥ）