

(19) 日本国特許庁 (JP)

(12) 特 許 公 報 (B2)

(11) 特許番号

特許第6833990号
(P6833990)

(45) 発行日 令和3年2月24日 (2021.2.24)

(24) 登録日 令和3年2月5日 (2021.2.5)

(51) Int. Cl.

F I

G 0 6 F 3 / 0 6 (2006.01)
G 0 6 F 1 3 / 1 0 (2006.01)G 0 6 F 3 / 0 6 5 4 0
G 0 6 F 3 / 0 6 3 0 1 X
G 0 6 F 3 / 0 6 3 0 5 C
G 0 6 F 3 / 0 6 3 0 1 K
G 0 6 F 3 / 0 6 3 0 2 A

請求項の数 22 (全 64 頁) 最終頁に続く

(21) 出願番号 特願2019-521779 (P2019-521779)
 (86) (22) 出願日 平成29年3月29日 (2017.3.29)
 (65) 公表番号 特表2019-532440 (P2019-532440A)
 (43) 公表日 令和1年11月7日 (2019.11.7)
 (86) 国際出願番号 PCT/CN2017/078579
 (87) 国際公開番号 WO2018/176265
 (87) 国際公開日 平成30年10月4日 (2018.10.4)
 審査請求日 平成31年4月23日 (2019.4.23)

(73) 特許権者 503433420
 華為技術有限公司
 HUAWEI TECHNOLOGIES
 CO., LTD.
 中華人民共和国 518129 広東省深
 ▲チェン▼市龍崗区坂田 華為總部▲ベン
 ▼公樓
 Huawei Administration Building, Bantian,
 Longgang District, Shenzhen, Guangdong
 518129, P. R. China

(74) 代理人 100107766
 弁理士 伊東 忠重

最終頁に続く

(54) 【発明の名称】 分散型ストレージシステムにアクセスするための方法、関係する装置及び関係するシステム

(57) 【特許請求の範囲】

【請求項1】

データアクセス方法であって、当該方法は分散型ストレージシステムに適用され、前記分散型ストレージシステムはm個のストレージノードを含み、前記m個のストレージノードは第1のストレージノードを含み、各ストレージノードは少なくとも1つのメモリを含み、各メモリは不揮発性記憶媒体を含み、mは1より大きい整数であり、当該方法は、

前記第1のストレージノードにより、端末から第1のデータ読み取り要求を受信するステップであり、前記第1のデータ読み取り要求は第1の位置情報を搬送し、前記第1の位置情報は、第1のデータが属するデータボディ内の前記第1のデータの位置を記述するために使用される、ステップと、

前記第1のストレージノードが、前記第1の位置情報に基づいて、前記第1のデータが位置するメモリが第1のメモリであると決定し、前記第1のメモリが前記第1のストレージノードに属する場合、状態情報セットが、前記第1のメモリが信頼されたアクセス状態にあることを現在記録しているとき、前記第1のストレージノードにより、前記第1のメモリから前記第1のデータを読み取り、前記第1のデータ読み取り要求に応答するために使用される第1のデータ読み取り応答を前記端末に送信するステップであり、前記第1のデータ読み取り応答は前記読み取られた第1のデータを搬送する、ステップと

を含み、

各メモリは、オンライン状態及びオフライン状態を含み、前記オンライン状態は、データ再構築状態及び信頼されたアクセス状態を含み、前記信頼されたアクセス状態は、前記

10

20

メモリに割り当てられたデータが全て成功して前記メモリに書き込まれたことを示す、方法。

【請求項 2】

前記第 1 のメモリが前記 m 個のストレージノード内の第 2 のストレージノードに属し、前記第 1 のメモリが信頼されたアクセス状態にあり、前記第 1 のストレージノードと、状態情報セットをリリースするように構成された管理ノードとが接続状態にある場合、前記第 1 のストレージノードにより、前記第 1 のデータ読み取り要求を前記第 2 のストレージノードに転送するステップと、

前記第 1 のストレージノードから前記第 1 のデータ読み取り要求を受信した後に、前記第 2 のストレージノードにより、前記第 1 のメモリから前記第 1 のデータを読み取り、前記第 1 のデータ読み取り要求に応答するために使用される前記第 1 のデータ読み取り応答を前記第 1 のストレージノード又は前記端末に送信するステップであり、前記第 1 のデータ読み取り応答は前記読み取られた第 1 のデータを搬送する、ステップと

を更に含む、請求項 1 に記載の方法。

【請求項 3】

前記第 1 のメモリが前記 m 個のストレージノード内の第 2 のストレージノードに属し、前記第 1 のメモリが信頼されたアクセス状態にある場合、或いは前記第 1 のメモリが前記 m 個のストレージノード内の第 2 のストレージノードに属し、前記第 1 のメモリが信頼されたアクセス状態にあり、前記第 1 のストレージノードと、状態情報セットをリリースするように構成された管理ノードとが接続状態にある場合、前記第 1 のストレージノードにより、前記第 1 のストレージノードにより現在キャッシュされている前記状態情報セットの識別子を前記第 1 のデータ読み取り要求に追加し、前記状態情報セットの前記識別子が追加された前記第 1 のデータ読み取り要求を前記第 2 のストレージノードに送信するステップと、

前記第 2 のストレージノードが前記第 1 のストレージノードから前記第 1 のデータ読み取り要求を受信した後に、前記第 1 のデータ読み取り要求で搬送された前記状態情報セットの前記識別子が、前記第 2 のストレージノードにより現在キャッシュされている状態情報セットの識別子と同じであると比較により決定された場合、前記第 1 のメモリから前記第 1 のデータを読み取り、前記第 1 のデータ読み取り要求に応答するために使用される前記第 1 のデータ読み取り応答を前記第 1 のストレージノード又は前記端末に送信するステップであり、前記第 1 のデータ読み取り応答は前記読み取られた第 1 のデータを搬送する、ステップと

を更に含む、請求項 1 に記載の方法。

【請求項 4】

前記第 1 のデータは、ストリップにセグメント化され、前記ストリップは、イレジャーコード (EC) ストライプを形成し、

当該方法は、

前記第 1 のメモリが信頼されていないアクセス状態にあると決定され、前記 EC ストライプの前記ストリップが位置する N 個のメモリ内の x 個のメモリが前記第 1 のストレージノードに属し、前記 N 個のメモリ内の N-x 個のメモリが前記第 1 のストレージノードとは異なる y 個のストレージノードに属する場合、前記第 1 のストレージノードにより、前記 EC ストライプのストライプ識別子と、前記状態情報セットの識別子とを搬送するデータ読み取り要求を前記 y 個のストレージノードに送信するステップと、

前記第 1 のストレージノードから前記データ読み取り要求を受信した後に、前記 y 個のストレージノードのそれぞれにより、前記データ読み取り要求で搬送された前記状態情報セットの前記識別子を、前記ストレージノードにより現在キャッシュされている状態情報セットの識別子と比較し、前記データ読み取り要求で搬送された前記状態情報セットの前記識別子が、前記ストレージノードにより現在キャッシュされている前記状態情報セットの前記識別子と同じであると比較により決定された場合、前記ストレージノードに含まれる対応するメモリから前記 EC ストライプの対応するストリップを読み取り、前記 EC ストラ

イブの前記読み取られた対応するストリップを前記第1のストレージノードに送信するステップと、

前記第1のストレージノードにより、前記 y 個のストレージノード及び前記第1のストレージノードから収集された前記ECストライプのストリップに基づいて、前記第1のデータを取得するための検査演算を実行し、前記第1のデータ読み取り要求に応答するために使用される前記第1のデータ読み取り応答を前記端末に送信するステップであり、前記第1のデータ読み取り応答は前記取得された第1のデータを搬送する、ステップと

を更に含む、請求項1に記載の方法。

【請求項5】

前記 m 個のストレージノードのうち1つは管理ノードであり、

当該方法は、

前記第1のメモリの状態が第1の状態から第2の状態に変化したとき、前記第1のストレージノードにより、メモリ状態変化報告を前記管理ノードに送信するステップであり、前記メモリ状態変化報告は、前記第1のメモリが前記第2の状態にあることを示し、前記第1の状態は前記第2の状態とは異なり、前記第1の状態及び前記第2の状態は、以下の状態、すなわち、前記オフライン状態、前記データ再構築状態又は前記信頼されたアクセス状態のうちいずれか1つをそれぞれ含む、ステップと、

前記第1のストレージノードから前記メモリ状態変化報告を受信した後に、前記管理ノードにより、前記管理ノードによりキャッシュされた状態情報セットに記録された前記第1のメモリの状態を前記第2の状態に更新し、前記管理ノードによりキャッシュされた前記状態情報セットのバージョン番号を更新し、前記管理ノードにより、更新された状態情報セットを前記 m 個のストレージノード内の前記管理ノード以外のストレージノードに送信するステップと、

前記第1のストレージノードにより、前記第1のストレージノードにより現在キャッシュされている前記状態情報セットを更新するために、前記管理ノードからの前記状態情報セットを使用するステップと

を更に含む、請求項1乃至4のうちいずれか1項に記載の方法。

【請求項6】

前記第1のストレージノードにより、前記端末から第1のデータ書き込み要求を受信するステップであり、前記第1のデータ書き込み要求は第2のデータ及び第2の位置情報を搬送し、前記第2の位置情報は、前記第2のデータが属するデータボディ内の前記第2のデータの位置を記述するために使用される、ステップと、

前記第1のストレージノードにより、前記第2の位置情報に基づいて、前記第2のデータを書き込むことに関する W 個のメモリを決定し、前記第2のデータを W - T 個のデータストリップにセグメント化し、前記 W - T 個のデータストリップを使用することによる計算を用いて T 個のパリティストリップを取得するステップであり、前記 T 個のパリティストリップ及び前記 W - T 個のデータストリップは、 W 個のストリップを含むECストライプを形成し、前記 W 個のメモリは前記 W 個のストリップと1対1の対応関係にあり、 T 及び W は正の整数であり、 T は W 未満である、ステップと、

前記第1のストレージノードによりキャッシュされた前記状態情報セットが、前記 W 個のメモリ内の $W1$ 個のメモリが非オフライン状態にあることを現在記録しているとき、前記ECストライプのストリップ及び前記状態情報セットの前記識別子を搬送するデータ書き込み要求を、前記 $W1$ 個のメモリが属する $y2$ 個のストレージノードに送信するステップと、

前記第1のストレージノードから前記データ書き込み要求を受信した後に、前記 $y2$ 個のストレージノードのそれぞれにより、前記データ書き込み要求で搬送された前記状態情報セットの前記識別子を、前記ストレージノードにより現在キャッシュされている状態情報セットの識別子と比較し、前記データ書き込み要求で搬送された前記状態情報セットの前記識別子が、前記ストレージノードにより現在キャッシュされている前記状態情報セットの前記識別子と同じであると比較により決定された場合、前記ECストライプの対応するストリップを前記ストレージノードに含まれる対応するメモリに書き込むか、或いは前記デ

10

20

30

40

50

ータ書き込み要求で搬送された前記状態情報セットの前記識別子が、前記ストレージノードにより現在キャッシュされている前記状態情報セットの前記識別子とは異なると比較により決定された場合、前記ECストライプの対応するストリップを前記ストレージノードに含まれる対応するメモリに書き込むことを拒否するステップと

を更に含む、請求項1乃至5のうちいずれか1項に記載の方法。

【請求項7】

前記第1のストレージノードにより、前記端末から第2のデータ書き込み要求を受信するステップであり、前記第2のデータ書き込み要求は第3のデータ及び第3の位置情報を搬送し、前記第3の位置情報は、前記第3のデータが属するデータボディ内の前記第3のデータの位置を記述するために使用される、ステップと、

前記第3の位置情報に基づいて、前記第3のデータを書き込むことに関するW個のメモリを決定した後に、前記第1のストレージノードにより、前記第3のデータをW-T個のデータストリップにセグメント化し、前記W-T個のデータストリップを使用することによる計算を用いてT個のパリティストリップを取得するステップであり、前記T個のパリティストリップ及び前記W-T個のデータストリップは、W個のストリップを含むECストライプを形成し、前記W個のメモリは前記W個のストリップと1対1の対応関係にあり、T及びWは正の整数であり、TはW未満である、ステップと、

前記第1のストレージノードによりキャッシュされた前記状態情報セットが、前記W個のメモリ内のW1個のメモリが非オフライン状態にあることを現在記録しており、前記第1のストレージノードと、状態情報セットをリリースするように構成された前記管理ノードとが接続状態にある場合、前記ECストライプのストリップを搬送するデータ書き込み要求を、前記W1個のメモリが属するy2個のストレージノードに送信するステップと、

前記第1のストレージノードから前記データ書き込み要求を受信した後に、前記y2個のストレージノードのそれぞれにより、前記ECストライプの対応するストリップを前記ストレージノードに含まれる対応するメモリに書き込むステップと

を更に含む、請求項1乃至5のうちいずれか1項に記載の方法。

【請求項8】

W1はW未満であり、前記第1のストレージノードによりキャッシュされた前記状態情報セットは、前記W個のメモリ内のW2個のメモリがオフライン状態にあることを現在記録しており、

当該方法は、

前記第1のストレージノードにより、第1の再構築ログを生成するステップであり、前記第1の再構築ログは、前記W2個のメモリ内の第2のメモリのメモリ識別子を記録し、前記第1の再構築ログは、前記W個のストリップ内の前記第2のメモリに対応する第1のストリップのストリップ識別子を更に記録し、前記第1の再構築ログは、前記ECストライプの前記ストライプ識別子を更に記録し、前記第2のメモリは前記W2個のメモリ内のいずれかのメモリである、ステップと、

前記W2個のメモリ内の前記第2のメモリが再びオンラインになり、前記第2のメモリが属する前記第2のストレージノードが前記第2のメモリのオフライン期間中に生成された前記第1の再構築ログを収集した後に、前記第2のストレージノードにより、前記第1の再構築ログに記録された、前記第2のメモリに書き込まれる必要がある前記第1のストリップの前記ストリップ識別子を取得し、前記ECストライプに含まれる前記第1のストリップ以外のW-T個のストリップが位置するW-T個のメモリを決定し、前記W-T個のメモリから前記W-T個のストリップを読み取り、前記第2のメモリに書き込まれる必要がある前記第1のストリップを再構築するために前記W-T個のストリップを使用することにより、検査演算を実行し、前記再構築された第1のストリップを前記第2のメモリに書き込むステップと

を更に含む、請求項6又は7に記載の方法。

【請求項9】

m個のストレージノードを含む分散型ストレージシステムであって、各ストレージノード

10

20

30

40

50

ドは少なくとも1つのメモリを含み、各メモリは不揮発性記憶媒体を含み、 m は1より大きい整数であり、

前記 m 個のストレージノード内の第1のストレージノードは、

端末から第1のデータ読み取り要求を受信し、前記第1のデータ読み取り要求は第1の位置情報を搬送し、前記第1の位置情報は、第1のデータが属するデータボディ内の前記第1のデータの位置を記述するために使用され、前記第1の位置情報に基づいて、前記第1のデータが位置する第1のメモリを決定し、

前記第1のメモリが前記第1のストレージノードに属する場合、状態情報セットが、前記第1のメモリが信頼されたアクセス状態にあることを現在記録しているとき、前記第1のメモリから前記第1のデータを読み取り、前記第1のデータ読み取り要求に応答するために使用される第1のデータ読み取り応答を前記端末に送信し、前記第1のデータ読み取り応答は前記読み取られた第1のデータを搬送する、ように構成され、

各メモリは、オンライン状態及びオフライン状態を含み、前記オンライン状態は、データ再構築状態及び信頼されたアクセス状態を含み、前記信頼されたアクセス状態は、前記メモリに割り当てられたデータが全て成功して前記メモリに書き込まれたことを示す、分散型ストレージシステム。

【請求項10】

当該分散型ストレージシステムは、状態情報セットをリリースするように構成された管理ノードを更に含み、

前記第1のストレージノードは、前記第1のメモリが前記 m 個のストレージノード内の第2のストレージノードに属し、前記第1のメモリが信頼されたアクセス状態にあり、前記第1のストレージノードと、状態情報セットをリリースするように構成された前記管理ノードとが接続状態にある場合、前記第1のデータ読み取り要求を前記第2のストレージノードに転送するように更に構成され、

前記第2のストレージノードは、前記第1のストレージノードから前記第1のデータ読み取り要求を受信した後に、前記第1のメモリから前記第1のデータを読み取り、前記第1のデータ読み取り要求に応答するために使用される前記第1のデータ読み取り応答を前記第1のストレージノード又は前記端末に送信するように構成され、前記第1のデータ読み取り応答は前記読み取られた第1のデータを搬送する、請求項9に記載の分散型ストレージシステム。

【請求項11】

前記第1のストレージノードは、前記第1のメモリが前記 m 個のストレージノード内の第2のストレージノードに属し、前記第1のメモリが信頼されたアクセス状態にある場合、前記第1のストレージノードにより現在キャッシュされている前記状態情報セットの識別子を前記第1のデータ読み取り要求に追加し、前記状態情報セットの前記識別子が追加された前記第1のデータ読み取り要求を前記第2のストレージノードに送信するように更に構成され、

前記第2のストレージノードは、前記第1のストレージノードから前記第1のデータ読み取り要求を受信した後に、前記第1のデータ読み取り要求で搬送された前記状態情報セットの前記識別子を、前記第2のストレージノードにより現在キャッシュされている状態情報セットの識別子と比較し、前記第1のデータ読み取り要求で搬送された前記状態情報セットの前記識別子が、前記第2のストレージノードにより現在キャッシュされている前記状態情報セットの前記識別子と同じであると比較により決定された場合、前記第1のメモリから前記第1のデータを読み取り、前記第1のデータ読み取り要求に応答するために使用される前記第1のデータ読み取り応答を前記第1のストレージノード又は前記端末に送信するように構成され、前記第1のデータ読み取り応答は前記読み取られた第1のデータを搬送する、請求項9に記載の分散型ストレージシステム。

【請求項12】

前記第1のデータは、ストリップにセグメント化され、前記ストリップは、イレジャーコード（EC）ストライプを形成し、

10

20

30

40

50

前記第1のストレージノードは、前記第1のメモリが信頼されていないアクセス状態にあると決定され、前記ECストライプの前記ストリップが位置するN個のメモリ内のx個のメモリが前記第1のストレージノードに属し、前記N個のメモリ内のN-x個のメモリが前記第1のストレージノードとは異なるy個のストレージノードに属する場合、前記ECストライプのストライプ識別子と、前記状態情報セットの識別子とを搬送するデータ読み取り要求を前記y個のストレージノードに送信するように更に構成され、

前記y個のストレージノードのそれぞれは、前記第1のストレージノードから前記データ読み取り要求を受信した後に、前記データ読み取り要求で搬送された前記状態情報セットの前記識別子を、前記ストレージノードにより現在キャッシュされている状態情報セットの識別子と比較し、前記データ読み取り要求で搬送された前記状態情報セットの前記識別子が、前記ストレージノードにより現在キャッシュされている前記状態情報セットの前記識別子と同じであると比較により決定された場合、前記ストレージノードに含まれる対応するメモリから前記ECストライプの対応するストリップを読み取り、前記ECストライプの前記読み取られた対応するストリップを前記第1のストレージノードに送信するように構成され、

前記第1のストレージノードは、前記y個のストレージノード及び前記第1のストレージノードから収集された前記ECストライプのストリップに基づいて、前記第1のデータを取得するための検査演算を実行し、前記第1のデータ読み取り要求に応答するために使用される前記第1のデータ読み取り応答を前記端末に送信するように更に構成され、前記第1のデータ読み取り応答は前記取得された第1のデータを搬送する、請求項9に記載の分散型ストレージシステム。

【請求項13】

前記m個のストレージノードのうち1つは管理ノードであり、

前記第1のストレージノードは、前記第1のメモリの状態が第1の状態から第2の状態に変化したとき、メモリ状態変化報告を前記管理ノードに送信するように更に構成され、前記メモリ状態変化報告は、前記第1のメモリが前記第2の状態にあることを示し、前記第1の状態は前記第2の状態とは異なり、前記第1の状態及び前記第2の状態は、以下の状態、すなわち、前記オフライン状態、前記データ再構築状態又は前記信頼されたアクセス状態のうちいずれか1つをそれぞれ含み、

前記管理ノードは、前記第1のストレージノードから前記メモリ状態変化報告を受信した後に、前記管理ノードによりキャッシュされた状態情報セットに記録された前記第1のメモリの状態を前記第2の状態に更新し、前記管理ノードによりキャッシュされた前記状態情報セットのバージョン番号を更新するように構成され、前記管理ノードは、更新された状態情報セットを前記m個のストレージノード内の前記管理ノード以外のストレージノードに送信し、

前記第1のストレージノードは、前記第1のストレージノードにより現在キャッシュされている前記状態情報セットを更新するために、前記管理ノードからの前記状態情報セットを使用するように更に構成される、請求項9乃至12のうちいずれか1項に記載の分散型ストレージシステム。

【請求項14】

前記第1のストレージノードは、前記端末から第1のデータ書き込み要求を受信するように更に構成され、前記第1のデータ書き込み要求は第2のデータ及び第2の位置情報を搬送し、前記第2の位置情報は、前記第2のデータが属するデータボディ内の前記第2のデータの位置を記述するために使用され、

前記第1のストレージノードは、前記第2の位置情報に基づいて、前記第2のデータを書き込むことに関するW個のメモリを決定し、前記第2のデータをW-T個のデータストリップにセグメント化し、前記W-T個のデータストリップを使用することによる計算を用いてT個のパリティストリップを取得するように更に構成され、前記T個のパリティストリップ及び前記W-T個のデータストリップは、W個のストリップを含むECストライプを形成し、前記W個のメモリは前記W個のストリップと1対1の対応関係にあり、T及びWは正の整数であ

り、TはW未満であり、

前記第1のストレージノードは、前記第1のストレージノードによりキャッシュされた前記状態情報セットが、前記W個のメモリ内のW1個のメモリが非オフライン状態にあることを現在記録しているとき、前記ECストライプのストリップ及び前記状態情報セットの前記識別子を搬送するデータ書き込み要求を、前記W1個のメモリが属するy2個のストレージノードに送信するように更に構成され、

前記y2個のストレージノードのそれぞれは、前記第1のストレージノードから前記データ書き込み要求を受信した後に、前記データ書き込み要求で搬送された前記状態情報セットの前記識別子を、前記ストレージノードにより現在キャッシュされている状態情報セットの識別子と比較し、前記データ書き込み要求で搬送された前記状態情報セットの前記識別子が、前記ストレージノードにより現在キャッシュされている前記状態情報セットの前記識別子と同じであると比較により決定された場合、前記ECストライプの対応するストリップを前記ストレージノードに含まれる対応するメモリに書き込むか、或いは前記データ書き込み要求で搬送された前記状態情報セットの前記識別子が、前記ストレージノードにより現在キャッシュされている前記状態情報セットの前記識別子とは異なると比較により決定された場合、前記ECストライプの対応するストリップを前記ストレージノードに含まれる対応するメモリに書き込むことを拒否するように構成される、請求項9乃至12のうちいずれか1項に記載の分散型ストレージシステム。

【請求項15】

前記第1のストレージノードは、前記端末から第2のデータ書き込み要求を受信するように更に構成され、前記第2のデータ書き込み要求は第3のデータ及び第3の位置情報を搬送し、前記第3の位置情報は、前記第3のデータが属するデータボディ内の前記第3のデータの位置を記述するために使用され、

前記第1のストレージノードは、前記第3の位置情報に基づいて、前記第3のデータに関するW個のメモリを決定した後に、前記第3のデータをW-T個のデータストリップにセグメント化し、前記W-T個のデータストリップを使用することによる計算を用いてT個のパリティストリップを取得するように更に構成され、前記T個のパリティストリップ及び前記W-T個のデータストリップは、W個のストリップを含むECストライプを形成し、前記W個のメモリは前記W個のストリップと1対1の対応関係にあり、T及びWは正の整数であり、TはW未満であり、

前記第1のストレージノードは、前記第1のストレージノードによりキャッシュされた前記状態情報セットに現在記録されている前記W個のメモリ内の状態を決定し、前記W個のメモリ内のW1個のメモリが非オフライン状態にあると決定され、前記第1のストレージノードと、状態情報セットをリリースするように構成された前記管理ノードとが接続状態にある場合、前記ECストライプのストリップを搬送するデータ書き込み要求を、前記W1個のメモリが属するy2個のストレージノードに送信するように更に構成され、

前記y2個のストレージノードのそれぞれは、前記第1のストレージノードから前記データ書き込み要求を受信した後に、前記ECストライプの対応するストリップを前記ストレージノードに含まれる対応するメモリに書き込むように構成される、請求項9乃至12のうちいずれか1項に記載の分散型ストレージシステム。

【請求項16】

W1はW未満であり、前記第1のストレージノードによりキャッシュされた前記状態情報セットは、前記W個のメモリ内のW2個のメモリがオフライン状態にあることを現在記録しており、

前記第1のストレージノードは、第1の再構築ログを生成するように更に構成され、前記第1の再構築ログは、前記W2個のメモリ内の第2のメモリのメモリ識別子を記録し、前記第1の再構築ログは、前記W個のストリップ内の前記第2のメモリに対応する第1のストリップのストリップ識別子を更に記録し、前記第1の再構築ログは、前記ECストライプの前記ストライプ識別子を更に記録し、前記第2のメモリは前記W2個のメモリ内のいずれかのメモリであり、

前記W2個のメモリ内の前記第2のメモリが再びオンラインになり、前記第2のメモリが属する前記第2のストレージノードが前記第2のメモリのオフライン期間中に生成された前記第1の再構築ログを収集した後に、前記第2のストレージノードは、前記第1の再構築ログに記録された、前記第2のメモリに書き込まれる必要がある前記第1のストリップの前記ストリップ識別子を取得し、

前記ECストライプに含まれる前記第1のストリップ以外のW-T個のストリップが位置するW-T個のメモリを決定し、前記W-T個のメモリから前記W-T個のストリップを読み取り、前記第2のメモリに書き込まれる必要がある前記第1のストリップを再構築するために前記W-T個のストリップを使用することにより、検査演算を実行し、前記再構築された第1のストリップを前記第2のメモリに書き込む、請求項14又は15に記載の分散型ストレージシステム。

10

【請求項17】

ストレージノードであって、当該ストレージノードは、分散型ストレージシステムに含まれるm個のストレージノードのうち1つであり、各ストレージノードは少なくとも1つのメモリを含み、各メモリは不揮発性記憶媒体を含み、mは1より大きい整数であり、当該ストレージノードは、

端末から第1のデータ読み取り要求を受信するように構成された通信ユニットであり、前記第1のデータ読み取り要求は第1の位置情報を搬送し、前記第1の位置情報は、第1のデータが属するデータボディ内の前記第1のデータの位置を記述するために使用される、通信ユニットと、

20

前記第1の位置情報に基づいて、前記第1のデータが位置するメモリが第1のメモリであると決定され、前記第1のメモリが前記ストレージノードに属する場合、状態情報セットが、前記第1のメモリが信頼されたアクセス状態にあることを現在記録しているとき、前記第1のメモリから前記第1のデータを読み取るように構成された処理制御ユニットとを含み、

前記通信ユニットは、前記第1のデータ読み取り要求に応答するために使用される第1のデータ読み取り応答を前記端末に送信するように更に構成され、前記第1のデータ読み取り応答は前記読み取られた第1のデータを搬送し、

各メモリは、オンライン状態及びオフライン状態を含み、前記オンライン状態は、データ再構築状態及び信頼されたアクセス状態を含み、前記信頼されたアクセス状態は、前記メモリに割り当てられたデータが全て成功して前記メモリに書き込まれたことを示す、ストレージノード。

30

【請求項18】

前記通信ユニットは、前記第1のメモリが前記m個のストレージノード内の第2のストレージノードに属し、前記第1のメモリが信頼されたアクセス状態にあり、前記ストレージノードと、状態情報セットをリリースするように構成された管理ノードとが接続状態にある場合、前記第1のデータ読み取り要求を前記第2のストレージノードに転送するように更に構成され、

前記第1のデータ読み取り要求は、前記第2のストレージノードに対して、前記ストレージノードから前記第1のデータ読み取り要求を受信した後に、前記第1のメモリから前記第1のデータを読み取り、前記第1のデータ読み取り要求に応答するために使用される前記第1のデータ読み取り応答を前記ストレージノード又は前記端末に送信するようにトリガーするために使用され、前記第1のデータ読み取り応答は前記読み取られた第1のデータを搬送する、請求項17に記載のストレージノード。

40

【請求項19】

前記通信ユニットは、前記第1のメモリが前記m個のストレージノード内の第2のストレージノードに属し、前記第1のメモリが信頼されたアクセス状態にある場合、或いは前記第1のメモリが前記m個のストレージノード内の第2のストレージノードに属し、前記第1のメモリが信頼されたアクセス状態にあり、前記ストレージノードと、状態情報セットをリリースするように構成された管理ノードとが接続状態にある場合、前記ストレージ

50

ノードにより現在キャッシュされている前記状態情報セットの識別子を前記第1のデータ読み取り要求に追加し、前記状態情報セットの前記識別子が追加された前記第1のデータ読み取り要求を前記第2のストレージノードに送信するように更に構成され、

前記第1のデータ読み取り要求は、前記第2のストレージノードに対して、前記ストレージノードから前記第1のデータ読み取り要求を受信した後に、前記第1のデータ読み取り要求で搬送された前記状態情報セットの前記識別子を、前記第2のストレージノードにより現在キャッシュされている状態情報セットの識別子と比較し、前記第1のデータ読み取り要求で搬送された前記状態情報セットの前記識別子が、前記第2のストレージノードにより現在キャッシュされている前記状態情報セットの前記識別子と同じであると比較により決定された場合、前記第1のメモリから前記第1のデータを読み取り、前記第1のデータ読み取り要求に応答するために使用される前記第1のデータ読み取り応答を前記ストレージノード又は前記端末に送信するようにトリガーするために使用され、前記第1のデータ読み取り応答は前記読み取られた第1のデータを搬送する、請求項17に記載のストレージノード。

【請求項20】

前記 m 個のストレージノードのうち1つは管理ノードであり、

前記通信ユニットは、前記第1のメモリの状態が第1の状態から第2の状態に変化したとき、メモリ状態変化報告を前記管理ノードに送信するように更に構成され、前記メモリ状態変化報告は、前記第1のメモリが前記第2の状態にあることを示し、前記第1の状態は前記第2の状態とは異なり、前記第1の状態及び前記第2の状態は、以下の状態、すなわち、前記オフライン状態、前記データ再構築状態又は前記信頼されたアクセス状態のうちいずれか1つをそれぞれ含み、前記メモリ状態変化報告は、前記管理ノードに対して、前記ストレージノードから前記メモリ状態変化報告を受信した後に、前記管理ノードによりキャッシュされた状態情報セットに記録された前記第1のメモリの状態を前記第2の状態に更新し、前記管理ノードによりキャッシュされた前記状態情報セットのバージョン番号を更新し、更新された状態情報セットを前記 m 個のストレージノード内の前記管理ノード以外のストレージノードに送信するようにトリガーするために使用され、

前記処理制御ユニットは、前記ストレージノードにより現在キャッシュされている前記状態情報セットを更新するために、前記管理ノードからの前記状態情報セットを使用するように構成される、請求項17乃至19のうちいずれか1項に記載のストレージノード。

【請求項21】

第1のストレージノードに適用されるデータアクセス方法であって、前記第1のストレージノードは m 個のストレージノードを含む分散型ストレージシステムに位置し、各ストレージノードは少なくとも1つのメモリを含み、各メモリは不揮発性記憶媒体を含み、 m は1より大きい整数であり、当該方法は、

前記第1のストレージノードにより、第1のデータ読み取り要求を受信するステップであり、前記第1のデータ読み取り要求は第1の位置情報を搬送し、前記第1の位置情報は、第1のデータが属するデータボディ内の前記第1のデータの位置を記述するために使用される、ステップと、

前記第1のストレージノードが、前記第1の位置情報に基づいて、前記第1のデータが位置するメモリが第1のメモリであると決定し、前記第1のメモリが前記第1のストレージノードに属する場合、状態情報セットが、前記第1のメモリが信頼されたアクセス状態にあることを現在記録しているとき、前記第1のストレージノードにより、前記第1のメモリから前記第1のデータを読み取り、前記第1のデータ読み取り要求に応答するために使用される第1のデータ読み取り応答を送信するステップであり、前記第1のデータ読み取り応答は前記読み取られた第1のデータを搬送する、ステップと

を含み、

各メモリは、オンライン状態及びオフライン状態を含み、前記オンライン状態は、データ再構築状態及び信頼されたアクセス状態を含み、前記信頼されたアクセス状態は、前記メモリに割り当てられたデータが全て成功して前記メモリに書き込まれたことを示す、方

10

20

30

40

50

法。

【請求項 2 2】

ストレージノードであって、当該ストレージノードは、分散型ストレージシステムに含まれる m 個のストレージノードのうち1つであり、各ストレージノードは少なくとも1つのメモリを含み、各メモリは不揮発性記憶媒体を含み、 m は1より大きい整数であり、

当該ストレージノードは、互いに結合されたプロセッサ及び通信インタフェースを含み、

前記プロセッサは、前記通信インタフェースを使用することにより、第1のデータ読み取り要求を受信し、前記第1のデータ読み取り要求は第1の位置情報を搬送し、前記第1の位置情報は、第1のデータが属するデータボディ内の前記第1のデータの位置を記述するために使用され、前記第1の位置情報に基づいて、前記第1のデータが位置するメモリが第1のメモリであると決定され、前記第1のメモリが前記ストレージノードに属する場合、状態情報セットが、前記第1のメモリが信頼されたアクセス状態にあることを現在記録しているとき、前記第1のメモリから前記第1のデータを読み取り、前記通信インタフェースを使用することにより、前記第1のデータ読み取り要求に応答するために使用される第1のデータ読み取り応答を送信し、前記第1のデータ読み取り応答は前記読み取られた第1のデータを搬送する、ように構成され、

各メモリは、オンライン状態及びオフライン状態を含み、前記オンライン状態は、データ再構築状態及び信頼されたアクセス状態を含み、前記信頼されたアクセス状態は、前記メモリに割り当てられたデータが全て成功して前記メモリに書き込まれたことを示す、ストレージノード。

【発明の詳細な説明】

【技術分野】

【0001】

この出願はコンピュータ技術の分野に関し、具体的には、分散型ストレージシステムにアクセスするための方法、関係する装置及び関係するシステムに関する。

【背景技術】

【0002】

従来のネットワークストレージシステムは、全てのデータを記憶するために集中型ストレージサーバを使用する。このようなアーキテクチャでは、ストレージサーバは、システム性能のボトルネックと、信頼性及びセキュリティにおける弱点になり、大規模ストレージアプリケーションの要件を満たすことが困難である。分散型ストレージシステムでは、データは、分散化された方式で複数の別々のデバイス（すなわち、ストレージノード）に記憶される。分散型ストレージシステムは、スケーラブルなシステム構造を使用し、ストレージ負荷を共有するために複数のストレージノードを使用する。分散型ストレージシステムは、システム信頼性、可用性、アクセス効率及びスケーラビリティを改善するのに役立つ。

【0003】

ストレージ容量が絶えず増加するにつれて、分散型ストレージシステムにおけるストレージノード及びハードディスクの数も増加する。ますます多くの分散型ストレージシステムは、データ信頼性を改善するために、異なるハードディスク上の記憶のためにデータを分散化するために、イレージャーコード（英文：erasure code，略称：EC）技術を使用する。システム信頼性を確保しつつ、EC技術はシステム性能を有効に更に改善できる。明らかに、EC技術に加えて、他の同様の冗長検査技術もまた、分散型ストレージシステムにおいて使用されてもよい。

【0004】

ストライプ（stripe）（例えば、ECストライプ）は、複数のストリップ（英文：strip）を含み、具体的には、 N 個のデータストリップ及び M 個のパリティストリップ（例えば、2つのデータstrip及び1つのパリティstrip）を含んでもよい。 M 個のパリティストリッ

プは、N個のデータストリップに基づく計算を用いて取得される。データストリップの長さは、1Mbまででもよい。従来の技術では、読み取り操作において読み取られるべきターゲットデータがストライプの小部分のみを占有しているとしても、データ正確性検査が実行できるように、ターゲットデータが属するストライプ内の全てのデータストリップ及びパリティストリップが各読み取り操作において読み取られる必要があり、それにより、ターゲットデータの正確性を確保する。しかし、ストライプの長さは通常では長く、したがって、ランダム読み取りのようなシナリオでは、ハードディスクの負荷が増加し、システム性能が大きく低減される。

【発明の概要】

【0005】

この出願の実施例は、データアクセス方法並びに関係する装置及びシステムを提供する。

【0006】

第1の態様によれば、この出願の実施例はデータ読み取り方法を提供する。当該方法は、m個のストレージノードを含む分散型ストレージシステムに適用され、各ストレージノードは少なくとも1つのメモリ（例えば、メモリはハードディスク又は他の形式のメモリでもよい）を含み、mは1より大きい整数である。当該方法は以下のステップを含んでもよい。例えば、端末が分散型ストレージシステムから第1のデータを読み取ることを要求する必要があるとき、m個のストレージノード内の第1のストレージノードは、端末から第1のデータ読み取り要求を受信してもよく、第1のデータ読み取り要求は第1の位置情報を搬送し、第1の位置情報は、第1のデータが属するデータボディ内の第1のデータの位置を記述するために使用される。第1のストレージノードが、第1の位置情報に基づいて、第1のデータが位置するメモリが第1のメモリであると決定し、第1のメモリが第1のストレージノードに属する場合、第1のストレージノードは、状態情報セット（第1のストレージノードによりキャッシュされた状態情報セット）に現在記録されている第1のメモリの状態（例えば、第1のメモリは信頼されたアクセス状態又は信頼されていないアクセス状態にあってもよい）を決定する。第1のメモリが信頼されたアクセス状態にあると決定されたとき、第1のストレージノードは、第1のメモリから第1のデータを読み取る。第1のストレージノードは、第1のデータ読み取り要求に応答するために使用される第1のデータ読み取り応答を端末に送信し、第1のデータ読み取り応答は読み取られた第1のデータを搬送し、第1のデータは、第1のデータが属するストリップ内のデータの一部又は全部である。

【0007】

この出願の実施例におけるデータボディは、例えば、ファイル、オブジェクト、ファイルセグメント、データベースレコード、データセグメント等でもよい。

【0008】

この出願の実施例における状態情報セットは、ハードディスクの状態等を記録するために使用され、状態情報セットのフォーマットは、例えば、ファイル又は他のデータフォーマットでもよい。状態情報セットのフォーマットがファイルである場合、状態情報セットはまた、「状態ファイル」とも呼ばれてもよい。

【0009】

前述の解決策では、メモリ状態を記録するために使用される状態情報セットが分散型ストレージシステムに導入され、したがって、状態情報セットが分散型ストレージシステムにおけるメモリのアクセス状態を記録及び管理するために使用されてもよいことが習得できる。状態情報セットにおいて信頼されたアクセス状態になっているとして記録されたメモリについて、要求されたデータは、メモリから直接読み取られて端末にフィードバックできる。信頼されたアクセスのこのような場合、関係するストライプの全てのストリップを読み取り、次に、正確性検査を実行する従来の方式における複雑なステップは必要なく、データ読み取り量が低減され、それにより、メモリ負荷を低減するのに役立ち、さらに、システム性能を改善するのに役立つ。

10

20

30

40

50

【0010】

さらに、この出願のいくつかの可能な実現方式では、第1のストレージノードが状態情報セットに現在記録されている第1のメモリの状態を決定する前に、第1のストレージノードは、まず、状態情報セットがキャッシュされているか否かを決定してもよく、状態情報セットがキャッシュされている場合、第1のストレージノードは、キャッシュされた状態情報セットに現在記録されている第1のメモリの状態を決定してもよく、或いは状態情報セットがキャッシュされておらず、第1のストレージノードと、状態情報セットをリリースするように構成された管理ノードとが接続状態にある場合、第1のストレージノードは、まず、キャッシュするために管理ノードから状態情報セットを取得し、次に、キャッシュされた状態情報セットに現在記録されている第1のメモリの状態を決定する。

10

【0011】

さらに、この出願のいくつかの可能な実現方式では、第1のストレージノードが状態情報セットに現在記録されている第1のメモリの状態を決定する前に、第1のストレージノードは、まず、状態情報セットがキャッシュされているか否かを決定してもよく、状態情報セットがキャッシュされている場合、第1のストレージノードは、キャッシュされた状態情報セットが取得されている持続時間が持続時間閾値を超えるか否かを決定してもよい（持続時間閾値は、例えば、2分、10分又は30分でもよく、一般的に、現在キャッシュされている状態情報セットが取得されている、より長い持続時間は、状態情報セットの無効のより高い可能性を示す）。キャッシュされた状態情報セットが取得されている持続時間が持続時間閾値を超え、第1のストレージノードと、状態情報セットをリリースするように構成された管理ノードとが接続状態にある場合、第1のストレージノードは、まず、管理ノードから状態情報セットを取得し、以前にキャッシュされた状態情報セットを更新するために、最新の取得された状態情報セットを使用し、次に、キャッシュされた状態情報セットに現在記録されている第1のメモリの状態を決定する。キャッシュされた状態情報セットが取得されている持続時間が持続時間閾値を超えないか、或いは第1のストレージノードと、状態情報セットをリリースするように構成された管理ノードとが接続状態にない場合、第1のストレージノードは、キャッシュされた状態情報セットに現在記録されている第1のメモリの状態を決定する。

20

【0012】

さらに、この出願のいくつかの可能な実現方式では、当該方法は以下のステップを更に含んでもよい。第1のメモリがm個のストレージノード内の第2のストレージノードに属し、第1のメモリが信頼されたアクセス状態にあると決定され（具体的には、例えば、第1のストレージノードは、状態情報セットに基づいて、第1のメモリが信頼されたアクセス状態にあると決定し、すなわち、第1のストレージノードによりキャッシュされた状態情報セットが、第1のメモリが信頼されたアクセス状態にあることを記録する）、第1のストレージノードと、状態情報セットをリリースするように構成された管理ノードとが接続状態にある場合（第1のストレージノードと、状態情報セットをリリースするように構成された管理ノードとが接続状態にあるとき、第1のストレージノードによりキャッシュされた状態情報セットが、管理ノードによりリリースされた最新の状態情報セットであるという比較的高い可能性が存在し、したがって、第1のストレージノードにより現在キャッシュされている状態情報セットが現在有効であるという比較的高い可能性が存在する。したがって、第1のストレージノードが、キャッシュされた状態情報セットに基づいて、第1のメモリが信頼されたアクセス状態にあると決定する場合は、第1のメモリの実際状況と一致するという非常に高い可能性が存在すると考えられてもよい）、第1のストレージノードは、第1のデータ読み取り要求を第2のストレージノードに転送する。第1のストレージノードから第1のデータ読み取り要求を受信した後に、第2のストレージノードは、第1のメモリから第1のデータを読み取り、第1のデータ読み取り要求に応答するために使用される第1のデータ読み取り応答を第1のストレージノード又は端末に送信し、第1のデータ読み取り応答は読み取られた第1のデータを搬送する。

30

40

【0013】

50

この出願のいくつかの可能な実現方式では、当該方法は以下のステップを更に含んでもよい。第1のメモリがm個のストレージノード内の第2のストレージノードに属し、第1のメモリが信頼されたアクセス状態にある（具体的には、例えば、第1のストレージノードにより現在キャッシュされている状態情報セットが、第1のメモリが信頼されたアクセス状態にあることを記録する）場合、或いは第1のメモリがm個のストレージノード内の第2のストレージノードに属し、第1のメモリが信頼されたアクセス状態にあり、第1のストレージノードと、状態情報セットをリリースするように構成された管理ノードとが接続状態にある場合（第1のストレージノードと、管理ノードとが接続状態にあるとき、第1のストレージノードが管理ノードによりリリースされた最新の状態情報セットを取得するという非常に高い可能性が存在する）、第1のストレージノードは、第1のストレージノードにより現在キャッシュされている状態情報セットの識別子を第1のデータ読み取り要求に追加し、状態情報セットの識別子が追加された第1のデータ読み取り要求を第2のストレージノードに送信してもよい。

10

【0014】

第1のストレージノードから第1のデータ読み取り要求を受信した後に、第2のストレージノードは、第1のデータ読み取り要求で搬送された状態情報セットの識別子を、第2のストレージノードにより現在キャッシュされている状態情報セットの識別子と比較し、第1のデータ読み取り要求で搬送された状態情報セットの識別子が、第2のストレージノードにより現在キャッシュされている状態情報セットの識別子と同じであると比較により決定された場合（第1のデータ読み取り要求で搬送された状態情報セットの識別子が、第2のストレージノードにより現在キャッシュされている状態情報セットの識別子と同じであるとき、これは、第1のストレージノード及び第2のストレージノードが同じ状態情報セットを現在キャッシュしていることを示し、この場合、第1のストレージノードにより現在キャッシュされている状態情報セットが現在有効であるという非常に高い可能性が存在する。この場合、第1のストレージノードと、管理ノードとが接続状態にある場合、第1のストレージノードにより現在キャッシュされている状態情報セットが現在有効であるという可能性が更に増加する。したがって、第1のストレージノードが、第1のストレージノードにより現在キャッシュされている状態情報セットに基づいて、第1のメモリが信頼されたアクセス状態にあると決定する場合は、第1のメモリの実際の状況と一致するという非常に高い可能性が存在する）、第1のメモリから第1のデータを読み取り、第1のデータ読み取り要求に応答するために使用される第1のデータ読み取り応答を第1のストレージノード又は端末に送信してもよく、第1のデータ読み取り応答は読み取られた第1のデータを搬送する。

20

30

【0015】

この出願の実施例では、状態情報セットの識別子は、例えば、状態情報セットのバージョン番号及び／又は状態情報セットのデータダイジェストを含む。或るバージョンの状態情報セットの識別子は、具体的には、このバージョンの状態情報セットを表すことができる如何なる情報でもよいことが理解され得る。状態情報セットのバージョン番号及び／又はデータダイジェストは、状態情報セットの識別子の単に可能な例である。

【0016】

40

一般的に、管理ノードの役目は、特定のノード（例えば、分散型ストレージシステムにおける特定のストレージノード）により果たされる。しかし、いくつかの特別な場合（例えば、現在の管理ノードが故障になったか、或いは十分な処理能力を有さない）、管理ノードの役目を果たすストレージノードはまた、変化することを許容されてもよい。例えば、ストレージノードAは、或る期間内に管理ノードの役目を果たすが（この場合、管理ノードのノード識別子はまた、ストレージノードAのノード識別子である）、他の期間内では、ストレージノードAは、もはや管理ノードの役目を果たさず、その代わりに、ストレージノードBが管理ノードの役目を果たす（この場合、管理ノードのノード識別子は、ストレージノードBのノード識別子である）。すなわち、管理ノードの役目を果たすストレージノードはいくつかの特別な場合に变化し得るため、管理ノードのノード識別子（すな

50

わち、管理ノードの役目を果たすストレージノードのノード識別子)は変化してもよい。分散型ストレージシステムにおける各ストレージノードは、最新の管理ノードとして現在考えられているストレージノードのノード識別子をキャッシュしてもよく、管理ノードの役目を果たすストレージノードが変化したと見出したとき、ストレージノードは、ストレージノードにより以前にキャッシュされた管理ノードのノード識別子を更新するためにストレージノードにより見出された最新の管理ノードのノード識別子を使用する。

【0017】

この出願のいくつかの可能な実現方式では、当該方法は以下のステップを更に含んでもよい。第1のメモリがm個のストレージノード内の第2のストレージノードに属し、第1のメモリが信頼されたアクセス状態にあり(具体的には、例えば、第1のストレージノードによりキャッシュされた状態情報セットが、第1のメモリが信頼されたアクセス状態にあることを記録する)、第1のストレージノードと、状態情報セットをリリースするように構成された管理ノードとが接続状態にあるとき、第1のストレージノードは、第1のストレージノードにより現在キャッシュされている、状態情報セットをリリースするように構成された管理ノードのノード識別子を第1のデータ読み取り要求に追加し、管理ノードのノード識別子が追加された第1のデータ読み取り要求を第2のストレージノードに送信する。

【0018】

第2のストレージノードが第1のストレージノードから第1のデータ読み取り要求を受信した場合、或いは第2のストレージノードが第1のストレージノードから第1のデータ読み取り要求を受信し、第2のストレージノードと、状態情報セットをリリースするように構成された管理ノードとが接続状態にある場合、第2のストレージノードは、第1のデータ読み取り要求で搬送された管理ノードのノード識別子を、第2のストレージノードにより現在キャッシュされている管理ノードのノード識別子と比較し、第1のデータ読み取り要求で搬送された管理ノードのノード識別子が、第2のストレージノードにより現在キャッシュされている管理ノードのノード識別子と同じであると比較により決定された場合(第1のストレージノード及び第2のストレージノードが管理ノードの同じノード識別子をキャッシュするとき、これは、第1のストレージノード及び第2のストレージノードが管理ノードとして同じストレージノードを考えていることを示し、第1のストレージノードと第2のストレージノードとの双方により管理ノードとして考えられているストレージノードが、分散型ストレージシステムにおける現在の最新の管理ノードであるという非常に高い可能性が存在する。この場合、第1のストレージノードと、管理ノードとが接続状態にある場合、第1のストレージノードによりキャッシュされた状態情報セットが現在有効であるという非常に高い可能性が存在する。したがって、第1のストレージノードが、第1のストレージノードによりキャッシュされた状態情報セットに基づいて、第1のメモリが信頼されたアクセス状態にあると決定する場合は、第1のメモリの実際の状況と一致するという非常に高い可能性が存在すると考えられてもよい)、第1のメモリから第1のデータを読み取り、第1のデータ読み取り要求に応答するために使用される第1のデータ読み取り応答を第1のストレージノード又は端末に送信し、第1のデータ読み取り応答は読み取られた第1のデータを搬送する。

【0019】

この出願のいくつかの可能な実現方式では、当該方法は以下のステップを更に含んでもよい。第1のメモリがm個のストレージノード内の第2のストレージノードに属し、第1のメモリが信頼されたアクセス状態にある(具体的には、例えば、第1のストレージノードによりキャッシュされた状態情報セットが、第1のメモリが信頼されたアクセス状態にあることを記録する)場合、或いは第1のメモリがm個のストレージノード内の第2のストレージノードに属し、第1のメモリが信頼されたアクセス状態にあり(例えば、第1のストレージノードによりキャッシュされた状態情報セットが、第1のメモリが信頼されたアクセス状態にあることを記録する)、第1のストレージノードと、状態情報セットをリリースするように構成された管理ノードとが接続状態にある場合、第1のストレージノード

ドは、第1のストレージノードにより現在キャッシュされている状態情報セットの識別子と、状態情報セットをリリースするように構成された管理ノードのノード識別子とを第1のデータ読み取り要求に追加し、状態情報セットの識別子及び管理ノードのノード識別子が追加された第1のデータ読み取り要求を第2のストレージノードに送信する。

【0020】

第2のストレージノードが第1のストレージノードから第1のデータ読み取り要求を受信した場合、或いは第2のストレージノードが第1のストレージノードから第1のデータ読み取り要求を受信し、第2のストレージノードと、状態情報セットをリリースするように構成された管理ノードとが接続状態にある場合、第2のストレージノードは、第1のデータ読み取り要求で搬送された状態情報セットの識別子を、第2のストレージノードにより現在キャッシュされている状態情報セットの識別子と比較し、第1のデータ読み取り要求で搬送された管理ノードのノード識別子を、第2のストレージノードにより現在キャッシュされている管理ノードのノード識別子と比較し、第1のデータ読み取り要求で搬送された状態情報セットの識別子が、第2のストレージノードにより現在キャッシュされている状態情報セットの識別子と同じであると比較により決定され、第1のデータ読み取り要求で搬送された管理ノードのノード識別子が、第2のストレージノードにより現在キャッシュされている管理ノードのノード識別子と同じであると比較により決定された場合（第1のストレージノード及び第2のストレージノードが管理ノードの同じノード識別子をキャッシュするとき、これは、第1のストレージノード及び第2のストレージノードが管理ノードとして同じストレージノードを考えていることを示し、第1のストレージノードと第2のストレージノードとの双方により管理ノードとして考えられているストレージノードが、分散型ストレージシステムにおける現在の最新の管理ノードであるという非常に高い可能性が存在する。この場合、第1のストレージノード及び第2のストレージノードが同じ状態情報セットを現在キャッシュしている場合、第1のストレージノードにより現在キャッシュされている状態情報セットが現在有効であるという非常に高い可能性が存在する。さらに、第1のストレージノード及び／又は第1のストレージノードと、管理ノードとが接続状態にある場合、第1のストレージノードにより現在キャッシュされている状態情報セットが現在有効であるというより高い可能性が存在する。したがって、第1のストレージノードが、第1のストレージノードにより現在キャッシュされている状態情報セットに基づいて、第1のメモリが信頼されたアクセス状態にあると決定する場合は、第1のメモリの実際の状況と一致するという非常に高い可能性が存在する）、第1のメモリから第1のデータを読み取り、第1のデータ読み取り要求に応答するために使用される第1のデータ読み取り応答を第1のストレージノード又は端末に送信し、第1のデータ読み取り応答は読み取られた第1のデータを搬送する。

【0021】

この出願のいくつかの可能な実現方式では、当該方法は以下のステップを更に含む。第1のストレージノードが、第1のストレージノードが信頼されていないアクセス状態にあると決定した場合、第1のストレージノードは、第1のデータが属するストライプを決定し、第1のストレージノードは、ストライプのストリップが位置するN個のメモリを決定する。

【0022】

N個のメモリが全て第1のストレージノードに属する場合、第1のストレージノードは、N個のメモリからストライプのストリップを読み取り、第1のストレージノードは、ストライプの読み取られたストリップに基づいて、第1のデータを取得するための検査演算を実行し（具体的には、第1のストレージノードは、ストライプの読み取られたストリップに基づいて、第1のデータが属するストリップを取得するための検査演算を実行し、次に、第1のデータが属するストリップから第1のデータを取得する）、第1のデータ読み取り要求に応答するために使用される第1のデータ読み取り応答を端末に送信し、第1のデータ読み取り応答は取得された第1のデータを搬送する。

【0023】

この出願のいくつかの可能な実現方式では、第1のデータが属するストライプのストリップが位置するN個のメモリ内のx個のメモリ（x個のメモリは第1のメモリを含む）が第1のストレージノードに属し、N個のメモリ内のN-x個のメモリが第1のメモリとは異なるy個のストレージノードに属するとき、当該方法は以下のステップを更に含んでもよい。第1のストレージノードが、第1のメモリが信頼されていないアクセス状態にあると決定し、N個のメモリ内の第1のメモリ以外のメモリが信頼されたアクセス状態にあると決定した場合（或いは第1のストレージノードが、第1のメモリが信頼されていないアクセス状態にあると決定し、N個のメモリ内の第1のストレージノード以外のメモリが信頼されたアクセス状態にあると決定し、第1のストレージノードと、状態情報セットをリリースするように構成された管理ノードとが接続状態にある場合）、第1のストレージノードは、ストライプのストライプ識別子と、第1のストレージノードによりキャッシュされた状態情報セットの識別子とを搬送するデータ読み取り要求をy個のストレージノードに送信する。対応して、第1のストレージノードからデータ読み取り要求を受信した後に、y個のストレージノードのそれぞれは、データ読み取り要求で搬送された状態情報セットの識別子を、ストレージノードにより現在キャッシュされている状態情報セットの識別子と比較し、データ読み取り要求で搬送された状態情報セットの識別子が、ストレージノードにより現在キャッシュされている状態情報セットの識別子と同じであると比較により決定された場合（データ読み取り要求で搬送された状態情報セットの識別子が、データ読み取り要求を受信したストレージノードにより現在キャッシュされている状態情報セットの識別子と同じであるとき、これは、第1のストレージノード及びデータ読み取り要求を受信したストレージノードが同じ状態情報セットを現在キャッシュしていることを示し、この場合、第1のストレージノードにより現在キャッシュされている状態情報セットが現在有効であるという非常に高い可能性が存在する。この場合、第1のストレージノードと、管理ノードとが接続状態にある場合、第1のストレージノードにより現在キャッシュされている状態情報セットが現在有効であるという可能性が更に増加する。したがって、第1のストレージノードが、第1のストレージノードにより現在キャッシュされている状態情報セットに基づいて、第1のメモリが信頼されていないアクセス状態にあると決定する場合は、第1のメモリの実際の状況と一致するという非常に高い可能性が存在する）、ストレージノードに含まれる対応するメモリからストライプの対応するストリップを読み取り、ストライプの読み取られた対応するストリップを第1のストレージノードに送信してもよい。

【0024】

対応して、第1のストレージノードは、y個のストレージノード及び第1のストレージノードから収集されたストライプのストリップ（これは第1のデータが属するストリップを含まない）に基づいて、第1のデータを取得するための検査演算を実行し、第1のデータ読み取り要求に応答するために使用される第1のデータ読み取り応答を端末に送信し、第1のデータ読み取り応答は取得された第1のデータを搬送する。

【0025】

この出願のいくつかの可能な実現方式では、m個のストレージノードのうち1つは管理ノードである。当該方法は以下のステップを更に含んでもよい。第1のメモリの状態が第1の状態から第2の状態に変化したとき、第1のストレージノードは、メモリ状態変化報告を管理ノードに送信し、メモリ状態変化報告は、第1のメモリが第2の状態にあることを示し、第1の状態は第2の状態とは異なり、第1の状態及び第2の状態は、以下の状態、すなわち、オフライン状態、データ再構築状態又は信頼されたアクセス状態のうちいずれか1つをそれぞれ含む。

【0026】

対応して、第1のストレージノードからメモリ状態変化報告を受信した後に、管理ノードは、管理ノードによりキャッシュされた状態情報セットに記録された第1のメモリの状態を第2の状態に更新し、管理ノードによりキャッシュされた状態情報セットのバージョン番号を更新してもよく、管理ノードは、更新された状態情報セットをm個のストレージ

ノード内の管理ノード以外のストレージノードに送信する。

【0027】

対応して、第1のストレージノードは、第1のストレージノードにより現在キャッシュされている状態情報セットを更新するために、管理ノードからの状態情報セットを使用する。

【0028】

この出願のいくつかの可能な実現方式では、当該方法は以下のステップを更に含む。例えば、端末が第2のデータを分散型ストレージシステムに書き込むことを要求する必要があるとき、第1のストレージノードは、端末から第1のデータ書き込み要求を受信してもよく、第1のデータ書き込み要求は第2のデータ及び第2の位置情報を搬送し、第2の位置情報は、第2のデータが属するデータボディ内の第2のデータの位置を記述するために使用される。第1のストレージノードは、第2の位置情報に基づいて、第2のデータを書き込むことに関するW個のメモリを決定する。第1のストレージノードは、第2のデータをW-T個のデータストリップにセグメント化し、W-T個のデータストリップを使用することによる計算を用いてT個のパリティストリップを取得する。T個のパリティストリップ及びW-T個のデータストリップは、W個のストリップを含むストライプを形成し、W個のメモリはW個のストリップと1対1の対応関係にあり、T及びWは正の整数であり、TはW未満である。

10

【0029】

第1のストレージノードは、第1のストレージノードによりキャッシュされた状態情報セットに現在記録されているW個のメモリの状態を決定し、W個のメモリ内のW1個のメモリが非オフライン状態にある（且つ、例えば、W個のメモリ内のW2個のメモリがオフライン状態にある）と決定されたとき、ストライプのストリップ及び状態情報セットの識別子を搬送するデータ書き込み要求を、W1個のメモリが属するy2個のストレージノードに送信する。

20

【0030】

対応して、第1のストレージノードからデータ書き込み要求を受信した後に、y2個のストレージノードのそれぞれは、データ書き込み要求で搬送された状態情報セットの識別子を、ストレージノードにより現在キャッシュされている状態情報セットの識別子と比較し、例えば、データ書き込み要求で搬送された状態情報セットの識別子が、ストレージノードにより現在キャッシュされている状態情報セットの識別子と同じであると比較により決定された場合、ストライプの対応するストリップをストレージノードに含まれる対応するメモリに書き込むか、或いはデータ書き込み要求で搬送された状態情報セットの識別子が、ストレージノードにより現在キャッシュされている状態情報セットの識別子とは異なると比較により決定された場合、ストライプの対応するストリップをストレージノードに含まれる対応するメモリに書き込むことを拒否する。

30

【0031】

他の例では、この出願のいくつかの可能な実現方式では、当該方法は以下のステップを更に含んでもよい。例えば、端末が第3のデータを分散型ストレージシステムに書き込むことを要求する必要があるとき、第1のストレージノードは、端末から第2のデータ書き込み要求を受信し、第2のデータ書き込み要求は第3のデータ及び第3の位置情報を搬送し、第3の位置情報は、第3のデータが属するデータボディ内の第3のデータの位置を記述するために使用される。第1のストレージノードは、第3の位置情報に基づいて、第3のデータを書き込むことに関するW個のメモリを決定する。第1のストレージノードは、第3のデータをW-T個のデータストリップにセグメント化し、W-T個のデータストリップを使用することによる計算を用いてT個のパリティストリップを取得する。T個のパリティストリップ及びW-T個のデータストリップは、W個のストリップを含むストライプを形成し、W個のメモリはW個のストリップと1対1の対応関係にあり、T及びWは正の整数であり、TはW未満である。

40

【0032】

50

第1のストレージノードは、第1のストレージノードによりキャッシュされた状態情報セットに現在記録されているW個のメモリの状態を決定し、W個のメモリ内のW1個のメモリが非オフライン状態にある（且つ、例えば、W個のメモリ内のW2個のメモリがオフライン状態にある）と決定され、第1のストレージノードと、状態情報セットをリリースするように構成された管理ノードとが接続状態にある場合、ストライプのストリップを搬送するデータ書き込み要求を、W1個のメモリが属するy2個のストレージノードに送信する。対応して、第1のストレージノードからデータ書き込み要求を受信した後に、y2個のストレージノードのそれぞれは、ストライプの対応するストリップをストレージノードに含まれる対応するメモリに書き込んでもよい。

【0033】

他の例では、この出願のいくつかの可能な実現方式では、当該方法は以下のステップを更に含んでもよい。第1のストレージノードは、端末から第3のデータ書き込み要求を受信し、第3のデータ書き込み要求は第4のデータ及び第4の位置情報を搬送し、第4の位置情報は、第4のデータが属するデータボディ内の第4のデータの位置を記述するために使用される。第4の位置情報に基づいて、第4のデータを書き込むことに関するW個のメモリを決定した後に、第1のストレージノードは、第4のデータをW-T個のデータストリップにセグメント化し、W-T個のデータストリップを使用することによる計算を用いてT個のパリティストリップを取得する。T個のパリティストリップ及びW-T個のデータストリップは、W個のストリップを含むストライプを形成し、W個のメモリはW個のストリップと1対1の対応関係にあり、T及びWは正の整数であり、TはW未満である。第1のストレージノードによりキャッシュされた状態情報セットが、W個のメモリ内のW1個のメモリが非オフライン状態にある（且つ、例えば、W個のメモリ内のW2個のメモリがオフライン状態にある）ことを現在記録しており、第1のストレージノードと、状態情報セットをリリースするように構成された管理ノードとが接続状態にあるとき、第1のストレージノードは、ストライプのストリップと、状態情報セットをリリースするように構成された管理ノードのノード識別子とを搬送するデータ書き込み要求を、W1個のメモリが属するy2個のストレージノードに送信する。

【0034】

対応して、第1のストレージノードからデータ書き込み要求を受信した後に、y2個のストレージノードのそれぞれは、データ書き込み要求で搬送された管理ノードのノード識別子を、ストレージノードにより現在キャッシュされている管理ノードのノード識別子と比較し、データ書き込み要求で搬送された管理ノードのノード識別子が、ストレージノードにより現在キャッシュされている管理ノードのノード識別子と同じであると比較により決定された場合、ストライプの対応するストリップをストレージノードに含まれる対応するメモリに書き込むか、或いはデータ書き込み要求で搬送された管理ノードのノード識別子が、ストレージノードにより現在キャッシュされている管理ノードのノード識別子とは異なると比較により決定された場合、ストライプの対応するストリップをストレージノードに含まれる対応するメモリに書き込むことを拒否する。

【0035】

この出願のいくつかの可能な実現方式では、メモリがオフライン等のため、メモリ（例えば、第2のメモリ）に割り当てられたストリップがメモリに書き込まれるのに失敗したとき、メモリが再びオンラインになった後に関係するストリップ再構築を実行するために再構築機構が使用されてもよい。例えば、W1はW未満であり、第1のストレージノードによりキャッシュされた状態情報セットは、W個のメモリ内のW2個のメモリがオフライン状態にあることを現在記録している。当該方法は以下のステップを更に含んでもよい。第1のストレージノードは、第1の再構築ログを生成し、第1の再構築ログは、W2個のメモリ内の第2のメモリのメモリ識別子を記録し、第1の再構築ログは、W個のストリップ内の第2のメモリに対応する第1のストリップのストリップ識別子を更に記録し、第1の再構築ログは、ストライプのストライプ識別子を更に記録し、第2のメモリはW2個のメモリ内のいずれかのメモリである。

【0036】

対応して、W2個のメモリ内の第2のメモリが再びオンラインになり、第2のメモリが属する第2のストレージノードが第2のメモリのオフライン期間中に生成された第1の再構築ログを収集した後に、第2のストレージノードは、第1の再構築ログに記録された、第2のメモリに書き込まれる必要がある第1のストリップの識別子を取得し、ストライプに含まれる第1のストリップ以外のW-T個のストリップが位置するW-T個のメモリを決定し、W-T個のメモリからW-T個のストリップを読み取り、第2のメモリに書き込まれる必要がある第1のストリップを再構築するためにW-T個のストリップを使用することにより、検査演算を実行し、再構築された第1のストリップを第2のメモリに書き込む。

【0037】

10

前述の例示の解決策は、状態情報セットの有効性を或る程度検証するために使用できるいくつかの可能な実現方式を提供することが習得できる。これらの可能な実現方式は、複数の信頼性レベルの検証要件を満たすのに役立ち、さらに、複数の読み取り／書き込みアクセス効率要件を満たすのに役立つ。

【0038】

第2の態様によれば、この出願の実施例は分散型ストレージシステムを更に提供する。分散型ストレージシステムはm個のストレージノードを含み、各ストレージノードは少なくとも1つのメモリを含み、各メモリは不揮発性記憶媒体を含み、mは1より大きい整数である。

【0039】

20

m個のストレージノード内の第1のストレージノードは、端末から第1のデータ読み取り要求を受信し、第1のデータ読み取り要求は第1の位置情報を搬送し、第1の位置情報は、第1のデータが属するデータボディ内の第1のデータの位置を記述するために使用され、第1の位置情報に基づいて、第1のデータが位置する第1のメモリを決定し、

第1のメモリが第1のストレージノードに属する場合、状態情報セットが、第1のメモリが信頼されたアクセス状態にあることを現在記録しているとき、第1のメモリから第1のデータを読み取り、第1のデータ読み取り要求に応答するために使用される第1のデータ読み取り応答を端末に送信し、第1のデータ読み取り応答は読み取られた第1のデータを搬送する、ように構成される。第1のデータは、例えば、第1のデータが属するストリップ内のデータの一部又は全部である。

30

【0040】

この出願のいくつかの可能な実現方式では、分散型ストレージシステムは、状態情報セットをリリースするように構成された管理ノードを更に含む。

【0041】

第1のストレージノードは、第1のメモリがm個のストレージノード内の第2のストレージノードに属し、第1のメモリが信頼されたアクセス状態にあり、第1のストレージノードと、状態情報セットをリリースするように構成された管理ノードとが接続状態にある場合、第1のデータ読み取り要求を第2のストレージノードに転送するように更に構成される。

【0042】

40

第2のストレージノードは、第1のストレージノードから第1のデータ読み取り要求を受信した後に、第1のメモリから第1のデータを読み取り、第1のデータ読み取り要求に応答するために使用される第1のデータ読み取り応答を第1のストレージノード又は端末に送信するように更に構成され、第1のデータ読み取り応答は読み取られた第1のデータを搬送する。

【0043】

この出願のいくつかの可能な実現方式では、第1のストレージノードは、第1のメモリがm個のストレージノード内の第2のストレージノードに属し、第1のメモリが信頼されたアクセス状態にある場合、第1のストレージノードにより現在キャッシュされている状態情報セットの識別子を第1のデータ読み取り要求に追加し、状態情報セットの識別子が

50

追加された第1のデータ読み取り要求を第2のストレージノードに送信するように更に構成される。

【0044】

第2のストレージノードは、第1のストレージノードから第1のデータ読み取り要求を受信した後に、第1のデータ読み取り要求で搬送された状態情報セットの識別子を、第2のストレージノードにより現在キャッシュされている状態情報セットの識別子と比較し、第1のデータ読み取り要求で搬送された状態情報セットの識別子が、第2のストレージノードにより現在キャッシュされている状態情報セットの識別子と同じであると比較により決定された場合、第1のメモリから第1のデータを読み取り、第1のデータ読み取り要求に応答するために使用される第1のデータ読み取り応答を第1のストレージノード又は端

10

【0045】

この出願のいくつかの可能な実現方式では、第1のストレージノードは、第1のメモリが m 個のストレージノード内の第2のストレージノードに属し、第1のメモリが信頼されたアクセス状態にある場合、第1のストレージノードにより現在キャッシュされている状態情報セットの識別子と、状態情報セットをリリースするように構成された管理ノードのノード識別子とを第1のデータ読み取り要求に追加し、状態情報セットの識別子及び管理ノードのノード識別子が追加された第1のデータ読み取り要求を第2のストレージノードに送信するように更に構成される。

20

【0046】

対応して、第2のストレージノードは、第1のストレージノードから第1のデータ読み取り要求を受信した後に、第1のデータ読み取り要求で搬送された状態情報セットの識別子を、第2のストレージノードにより現在キャッシュされている状態情報セットの識別子と比較し、第1のデータ読み取り要求で搬送された管理ノードのノード識別子を、第2のストレージノードにより現在キャッシュされている管理ノードのノード識別子と比較し、第1のデータ読み取り要求で搬送された状態情報セットの識別子が、第2のストレージノードにより現在キャッシュされている状態情報セットの識別子と同じであると比較により決定され、第1のデータ読み取り要求で搬送された管理ノードのノード識別子が、第2のノードにより現在キャッシュされている管理ノードのノード識別子と同じであると比較により決定された場合、第1のメモリから第1のデータを読み取り、第1のデータ読み取り要求に応答するために使用される第1のデータ読み取り応答を第1のストレージノード又は端末に送信するように構成され、第1のデータ読み取り応答は読み取られた第1のデータを搬送する。

30

【0047】

この出願のいくつかの可能な実現方式では、第1のストレージノードは、第1のメモリが信頼されていないアクセス状態にあると決定され、第1のデータが属するストライプのストリップが位置する N 個のメモリ内の x 個のメモリが第1のストレージノードに属し、 N 個のメモリ内の $N-x$ 個のメモリが第1のストレージノードとは異なる y 個のストレージノードに属する場合、ストライプのストライプ識別子と、状態情報セットの識別子とを搬送するデータ読み取り要求を y 個のストレージノードに送信するように更に構成される。

40

【0048】

y 個のストレージノードのそれぞれは、第1のストレージノードからデータ読み取り要求を受信した後に、データ読み取り要求で搬送された状態情報セットの識別子を、ストレージノードにより現在キャッシュされている状態情報セットの識別子と比較し、データ読み取り要求で搬送された状態情報セットの識別子が、ストレージノードにより現在キャッシュされている状態情報セットの識別子と同じであると比較により決定された場合、ストレージノードに含まれる対応するメモリからストライプの対応するストリップを読み取り、ストライプの読み取られた対応するストリップを第1のストレージノードに送信するように構成される。

50

【0049】

第1のストレージノードは、 y 個のストレージノード及び第1のストレージノードから収集されたストライプのストリップに基づいて、第1のデータを取得するための検査演算を実行し、第1のデータ読み取り要求に応答するために使用される第1のデータ読み取り応答を端末に送信するように更に構成され、第1のデータ読み取り応答は取得された第1のデータを搬送する。

【0050】

この出願のいくつかの可能な実現方式では、例えば、 m 個のストレージノードのうち1つは管理ノードである。第1のストレージノードは、第1のメモリの状態が第1の状態から第2の状態に変化したとき、メモリ状態変化報告を管理ノードに送信するように更に構成され、メモリ状態変化報告は、第1のメモリが第2の状態にあることを示し、第1の状態は第2の状態とは異なり、第1の状態及び第2の状態は、以下の状態、すなわち、オフライン状態、データ再構築状態又は信頼されたアクセス状態のうちいずれか1つをそれぞれ含む。

10

【0051】

管理ノードは、第1のストレージノードからメモリ状態変化報告を受信した後に、管理ノードによりキャッシュされた状態情報セットに記録された第1のメモリの状態を第2の状態に更新し、管理ノードによりキャッシュされた状態情報セットのバージョン番号を更新するように構成され、管理ノードは、更新された状態情報セットを m 個のストレージノード内の管理ノード以外のストレージノードに送信する。

20

【0052】

対応して、第1のストレージノードは、第1のストレージノードにより現在キャッシュされている状態情報セットを更新するために、管理ノードからの状態情報セットを使用するように更に構成される。対応して、他のストレージノードもまた、他のストレージノードにより現在キャッシュされている状態情報セットを更新するために、管理ノードからの状態情報セットを使用する。

【0053】

この出願のいくつかの可能な実現方式では、第1のストレージノードは、端末から第1のデータ書き込み要求を受信するように更に構成され、第1のデータ書き込み要求は第2のデータ及び第2の位置情報を搬送し、第2の位置情報は、第2のデータが属するデータボディ内の第2のデータの位置を記述するために使用される。

30

【0054】

第1のストレージノードは、第2の位置情報に基づいて、第2のデータを書き込むことに関する W 個のメモリを決定し、第2のデータを W - T 個のデータストリップにセグメント化し、 W - T 個のデータストリップを使用することによる計算を用いて T 個のパリティストリップを取得するように更に構成され、 T 個のパリティストリップ及び W - T 個のデータストリップは、 W 個のストリップを含むストライプを形成し、 W 個のメモリは W 個のストリップと1対1の対応関係にあり、 T 及び W は正の整数であり、 T は W 未満である。

【0055】

第1のストレージノードは、第1のストレージノードによりキャッシュされた状態情報セットが、 W 個のメモリ内の $W1$ 個のメモリが非オフライン状態にあり、 W 個のメモリ内の $W2$ 個のメモリがオフライン状態にあることを現在記録しているとき、ストライプのストリップ及び状態情報セットの識別子を搬送するデータ書き込み要求を、 $W1$ 個のメモリが属する $y2$ 個のストレージノードに送信するように更に構成される。

40

【0056】

対応して、 $y2$ 個のストレージノードのそれぞれは、第1のストレージノードからデータ書き込み要求を受信した後に、データ書き込み要求で搬送された状態情報セットの識別子を、ストレージノードにより現在キャッシュされている状態情報セットの識別子と比較し、データ書き込み要求で搬送された状態情報セットの識別子が、ストレージノードにより現在キャッシュされている状態情報セットの識別子と同じであると比較により決定された

50

場合、ストライプの対応するストリップをストレージノードに含まれる対応するメモリに書き込むか、或いはデータ書き込み要求で搬送された状態情報セットの識別子が、ストレージノードにより現在キャッシュされている状態情報セットの識別子とは異なると比較により決定された場合、ストライプの対応するストリップをストレージノードに含まれる対応するメモリに書き込むことを拒否するように構成される。

【0057】

この出願のいくつかの可能な実現方式では、第1のストレージノードは、端末から第2のデータ書き込み要求を受信するように更に構成され、第2のデータ書き込み要求は第3のデータ及び第3の位置情報を搬送し、第3の位置情報は、第3のデータが属するデータボディ内の第3のデータの位置を記述するために使用される。

10

【0058】

第1のストレージノードは、第3の位置情報に基づいて、第3のデータを書き込むことに関するW個のメモリを決定した後に、第3のデータをW-T個のデータストリップにセグメント化し、W-T個のデータストリップを使用することによる計算を用いてT個のパリティストリップを取得するように更に構成され、T個のパリティストリップ及びW-T個のデータストリップは、W個のストリップを含むストライプを形成し、W個のメモリはW個のストリップと1対1の対応関係にあり、T及びWは正の整数であり、TはW未満である。

【0059】

第1のストレージノードは、第1のストレージノードによりキャッシュされた状態情報セットに現在記録されているW個のメモリの状態を決定し、W個のメモリ内のW1個のメモリが非オフライン状態にある（且つ、例えば、W個のメモリ内のW2個のメモリがオフライン状態にある）と決定され、第1のストレージノードと、状態情報セットをリリースするように構成された管理ノードとが接続状態にある場合、ストライプのストリップを搬送するデータ書き込み要求を、W1個のメモリが属するy2個のストレージノードに送信するように更に構成される。

20

【0060】

対応して、y2個のストレージノードのそれぞれは、第1のストレージノードからデータ書き込み要求を受信した後に、ストライプの対応するストリップをストレージノードに含まれる対応するメモリに書き込むように構成される。

【0061】

この出願のいくつかの可能な実現方式では、第1のストレージノードは、端末から第3のデータ書き込み要求を受信するように更に構成され、第3のデータ書き込み要求は第4のデータ及び第4の位置情報を搬送し、第4の位置情報は、第4のデータが属するデータボディ内の第4のデータの位置を記述するために使用される。

30

【0062】

第1のストレージノードは、第4の位置情報に基づいて、第4のデータを書き込むことに関するW個のメモリを決定した後に、第4のデータをW-T個のデータストリップにセグメント化し、W-T個のデータストリップを使用することによる計算を用いてT個のパリティストリップを取得するように更に構成され、T個のパリティストリップ及びW-T個のデータストリップは、W個のストリップを含むストライプを形成し、W個のメモリはW個のストリップと1対1の対応関係にあり、T及びWは正の整数であり、TはW未満である。

40

【0063】

第1のストレージノードは、第1のストレージノードによりキャッシュされた状態情報セットが、W個のメモリ内のW1個のメモリが非オフライン状態にある（且つ、例えば、W個のメモリ内のW2個のメモリがオフライン状態にある）ことを現在記録しており、第1のストレージノードと、状態情報セットをリリースするように構成された管理ノードとが接続状態にあるとき、ストライプのストリップと、状態情報セットをリリースするように構成された管理ノードのノード識別子とを搬送するデータ書き込み要求を、W1個のメモリが属するy2個のストレージノードに送信するように更に構成される。

【0064】

50

対応して、y2個のストレージノードのそれぞれは、第1のストレージノードからデータ書き込み要求を受信した後に、データ書き込み要求で搬送された管理ノードのノード識別子を、ストレージノードにより現在キャッシュされている管理ノードのノード識別子と比較し、データ書き込み要求で搬送された管理ノードのノード識別子が、ストレージノードにより現在キャッシュされている管理ノードのノード識別子と同じであると比較により決定された場合、ストライプの対応するストリップをストレージノードに含まれる対応するメモリに書き込むか、或いはデータ書き込み要求で搬送された管理ノードのノード識別子が、ストレージノードにより現在キャッシュされている管理ノードのノード識別子とは異なると比較により決定された場合、ストライプの対応するストリップをストレージノードに含まれる対応するメモリに書き込むことを拒否するように構成される。

10

【0065】

この出願のいくつかの可能な実現方式では、第1のストレージノードは、第1の再構築ログを生成するように更に構成され、第1の再構築ログは、W2個のメモリ内の第2のメモリのメモリ識別子を記録し、第1の再構築ログは、W個のストリップ内の第2のメモリに対応する第1のストリップのストリップ識別子を更に記録し、第1の再構築ログは、ストライプのストライプ識別子を更に記録し、第2のメモリはW2個のメモリ内のいずれかのメモリである。

【0066】

対応して、例えば、W2個のメモリ内の第2のメモリが再びオンラインになり、第2のメモリが属する第2のストレージノードが第2のメモリのオフライン期間中に生成された第1の再構築ログを収集した後に、第2のストレージノードは、第1の再構築ログに記録された、第2のメモリに書き込まれる必要がある第1のストリップの識別子を取得し、ストライプに含まれる第1のストリップ以外のW-T個のストリップが位置するW-T個のメモリを決定し、W-T個のメモリからW-T個のストリップを読み取り、第2のメモリに書き込まれる必要がある第1のストリップを再構築するためにW-T個のストリップを使用することにより、検査演算を実行し、再構築された第1のストリップを第2のメモリに書き込む。

20

【0067】

第3の態様によれば、この出願の実施例はストレージノードを更に提供する。ストレージノードは、分散型ストレージシステムに含まれるm個のストレージノードのうち1つであり、各ストレージノードは少なくとも1つのメモリを含み、各メモリは不揮発性記憶媒体を含み、mは1より大きい整数である。ストレージノードは、通信ユニットと処理制御ユニットとを含む。

30

【0068】

通信ユニットは、端末から第1のデータ読み取り要求を受信するように構成され、第1のデータ読み取り要求は第1の位置情報を搬送し、第1の位置情報は、第1のデータが属するデータボディ内の第1のデータの位置を記述するために使用される。

【0069】

処理制御ユニットは、第1の位置情報に基づいて、第1のデータが位置するメモリが第1のメモリであると決定され、第1のメモリがストレージノードに属する場合、状態情報セットが、第1のメモリが信頼されたアクセス状態にあることを現在記録しているとき、第1のメモリから第1のデータを読み取るように構成される。

40

【0070】

通信ユニットは、第1のデータ読み取り要求に応答するために使用される第1のデータ読み取り応答を端末に送信するように更に構成され、第1のデータ読み取り応答は読み取られた第1のデータを搬送する。

【0071】

この出願のいくつかの可能な実現方式では、通信ユニットは、第1のメモリがm個のストレージノード内の第2のストレージノードに属し、第1のメモリが信頼されたアクセス状態にあり、ストレージノードと、状態情報セットをリリースするように構成された管理ノードとが接続状態にある場合、第1のデータ読み取り要求を第2のストレージノードに

50

転送するように更に構成される。

【0072】

第1のデータ読み取り要求は、第2のストレージノードに対して、第1のストレージノードから第1のデータ読み取り要求を受信した後に、第1のメモリから第1のデータを読み取り、第1のデータ読み取り要求に応答するために使用される第1のデータ読み取り応答をストレージノード又は端末に送信するようにトリガーするために使用され、第1のデータ読み取り応答は読み取られた第1のデータを搬送する。

【0073】

この出願のいくつかの可能な実現方式では、通信ユニットは、第1のメモリがm個のストレージノード内の第2のストレージノードに属し、第1のメモリが信頼されたアクセス状態にある場合、第1のストレージノードにより現在キャッシュされている状態情報セットの識別子を第1のデータ読み取り要求に追加し、状態情報セットの識別子が追加された第1のデータ読み取り要求を第2のストレージノードに送信するように更に構成される。

10

【0074】

第1のデータ読み取り要求は、第2のストレージノードに対して、ストレージノードから第1のデータ読み取り要求を受信した後に、第1のデータ読み取り要求で搬送された状態情報セットの識別子を、第2のストレージノードにより現在キャッシュされている状態情報セットの識別子と比較し、第1のデータ読み取り要求で搬送された状態情報セットの識別子が、第2のストレージノードにより現在キャッシュされている状態情報セットの識別子と同じであると比較により決定された場合、第1のメモリから第1のデータを読み取り、第1のデータ読み取り要求に応答するために使用される第1のデータ読み取り応答をストレージノード又は端末に送信するようにトリガーするために使用され、第1のデータ読み取り応答は読み取られた第1のデータを搬送する。

20

【0075】

この出願のいくつかの可能な実現方式では、m個のストレージノードのうち1つは管理ノードである。

【0076】

通信ユニットは、第1のメモリの状態が第1の状態から第2の状態に変化したとき、メモリ状態変化報告を管理ノードに送信するように更に構成され、メモリ状態変化報告は、第1のメモリが第2の状態にあることを示し、第1の状態は第2の状態とは異なり、第1の状態及び第2の状態は、以下の状態、すなわち、オフライン状態、データ再構築状態又は信頼されたアクセス状態のうちいずれか1つをそれぞれ含む。メモリ状態変化報告は、管理ノードに対して、第1のストレージノードからメモリ状態変化報告を受信した後に、管理ノードによりキャッシュされた状態情報セットに記録された第1のメモリの状態を第2の状態に更新し、管理ノードによりキャッシュされた状態情報セットのバージョン番号を更新し、更新された状態情報セットをm個のストレージノード内の管理ノード以外のストレージノードに送信するようにトリガーするために使用される。

30

【0077】

処理制御ユニットは、ストレージノードにより現在キャッシュされている状態情報セットを更新するために、管理ノードからの状態情報セットを使用するように構成される。

40

【0078】

第4の態様によれば、この出願の実施例は、第1のストレージノードに適用されるデータアクセス方法を提供する。第1のストレージはm個のストレージノードを含む分散型ストレージシステムに位置し、各ストレージノードは少なくとも1つのメモリを含み、各メモリは不揮発性記憶媒体を含み、mは1より大きい整数である。

【0079】

当該方法は、第1のストレージノードにより、（端末から）第1のデータ読み取り要求を受信するステップであり、第1のデータ読み取り要求は第1の位置情報を搬送し、第1の位置情報は、第1のデータが属するデータボディ内の第1のデータの位置を記述するために使用される、ステップと、第1のストレージノードが、第1の位置情報に基づいて、

50

第1のデータが位置するメモリが第1のメモリであると決定し、第1のメモリが第1のストレージノードに属する場合、状態情報セットが、第1のメモリが信頼されたアクセス状態にあることを現在記録しているとき、第1のストレージノードにより、第1のメモリから第1のデータを読み取り、第1のデータ読み取り要求に応答するために使用される第1のデータ読み取り応答を送信するステップであり、第1のデータ読み取り応答は読み取られた第1のデータを搬送する、ステップとを含む。

【0080】

第5の態様によれば、この出願の実施例はストレージノードを提供する。ストレージノードは、分散型ストレージシステムに含まれるm個のストレージノードのうち1つであり、各ストレージノードは少なくとも1つのメモリを含み、各メモリは不揮発性記憶媒体を含み、mは1より大きい整数である。

10

【0081】

ストレージノードは、互いに結合されたプロセッサ及び通信インタフェースを含む。プロセッサは、前述の態様において第1のストレージノード（又は他のストレージノード）により実行される方法のステップの一部又は全部を実行するように構成される。

【0082】

例えば、プロセッサは、通信インタフェースを使用することにより、第1のデータ読み取り要求を受信し、第1のデータ読み取り要求は第1の位置情報を搬送し、第1の位置情報は、第1のデータが属するデータボディ内の第1のデータの位置を記述するために使用され、第1の位置情報に基づいて、第1のデータが位置するメモリが第1のメモリであると決定され、第1のメモリが第1のストレージノードに属する場合、状態情報セットが、第1のメモリが信頼されたアクセス状態にあることを現在記録しているとき、第1のメモリから第1のデータを読み取り、通信インタフェースを使用することにより、第1のデータ読み取り要求に応答するために使用される第1のデータ読み取り応答を送信し、第1のデータ読み取り応答は読み取られた第1のデータを搬送する、ように構成される。

20

【0083】

第6の態様によれば、この出願の実施例はコンピュータ読み取り可能記憶媒体を提供する。コンピュータ読み取り可能記憶媒体はプログラムコードを記憶する。プログラムコードは、第1の態様におけるいずれかのストレージノード（例えば、第1のストレージノード又は第2のストレージノード）により実行される方法のステップの一部又は全部を実行するために使用される命令を含む。

30

【0084】

第7の態様によれば、この出願の実施例は命令を含むコンピュータプログラムプロダクトを提供する。コンピュータプログラムプロダクトがコンピュータ（コンピュータは、例えばストレージノードである）上で実行したとき、コンピュータは、前述の態様におけるいずれかのストレージノード（例えば、第1のストレージノード又は第2のストレージノード）により実行される方法のステップの一部又は全部を実行する。

【0085】

第8の態様によれば、この出願の実施例は、互いに結合されたプロセッサ及び通信インタフェースと、メモリとを含むストレージノードを提供する。プロセッサは、前述の態様におけるいずれかのストレージノードにより実行される方法のステップの一部又は全部を実行するように構成される。

40

【0086】

第9の態様によれば、この出願の実施例は、分散型ストレージシステムと端末とを含んでもよいサービスシステムを更に提供する。分散型ストレージシステム及び端末は通信可能に接続される。分散型ストレージシステムは、この出願の実施例において提供されるいずれかの分散型ストレージシステムである。

【図面の簡単な説明】

【0087】

【図1-A】 この出願の実施例に従って例示されるいくつかの分散型ストレージシステム

50

の概略アーキテクチャ図である。

【図１－Ｂ】この出願の実施例に従って例示されるいくつかの分散型ストレージシステムの概略アーキテクチャ図である。

【図１－Ｃ】この出願の実施例に従って例示されるメモリ状態遷移の概略図である。

【図１－Ｄ】この出願の実施例に従って例示される状態情報セットの編成形式の概略図である。

【図１－Ｅ】この出願の実施例に従って例示されるいくつかのストライプ編成形式の概略図である。

【図１－Ｆ】この出願の実施例に従って例示されるいくつかのストライプ編成形式の概略図である。

10

【図２－Ａ】この出願の実施例に従って例示されるいくつかのファイル編成形式の概略図である。

【図２－Ｂ】この出願の実施例に従って例示されるいくつかのファイル編成形式の概略図である。

【図３－Ａ】この出願の実施例に従って例示される状態更新方法の概略フローチャートである。

【図３－Ｂ】この出願の実施例に従って例示される状態情報セットのいくつかの編成形式の概略図である。

【図３－Ｃ】この出願の実施例に従って例示される状態情報セットのいくつかの編成形式の概略図である。

20

【図４－Ａ】この出願の実施例に従って例示される他の状態更新方法の概略フローチャートである。

【図４－Ｂ】この出願の実施例に従って例示される状態情報セットのいくつかの編成形式の概略図である。

【図４－Ｃ】この出願の実施例に従って例示される状態情報セットのいくつかの編成形式の概略図である。

【図５－Ａ】この出願の実施例に従って例示される他のデータアクセス方法の概略フローチャートである。

【図５－Ｂ】この出願の実施例に従って例示される再構築ログの編成形式の概略図である。

30

【図６】この出願の実施例に従って例示される他のデータアクセス方法の概略フローチャートである。

【図７－Ａ】この出願の実施例に従って例示されるデータ再構築方法の概略フローチャートである。

【図７－Ｂ】この出願の実施例に従って例示される状態情報セットのいくつかの編成形式の概略図である。

【図７－Ｃ】この出願の実施例に従って例示される状態情報セットのいくつかの編成形式の概略図である。

【図８】この出願の実施例に従って例示される他のデータアクセス方法の概略フローチャートである。

40

【図９】この出願の実施例に従って例示される他のデータアクセス方法の概略フローチャートである。

【図１０】この出願の実施例に従って例示される他のデータアクセス方法の概略フローチャートである。

【図１１】この出願の実施例に従って例示される分散型ストレージシステムの概略図である。

【図１２】この出願の実施例に従って例示されるストレージノードの概略図である。

【図１３】この出願の実施例に従って例示される他のストレージノードの概略図である。

【発明を実施するための形態】

【００８８】

50

この出願の明細書、特許請求の範囲及び添付図面において言及される「含む」、「持つ」という用語、又はこれらの他の変形は、非排他的な包含をカバーすることを意図する。例えば、一連のステップ又はユニットを含むプロセス、方法、システム、プロダクト又はデバイスは、記載のステップ又はユニットに限定されず、任意選択で、記載されていないステップ又はユニットを更に含んでもよく、或いは任意選択で、プロセス、方法、プロダクト又はデバイスの他の固有のステップ又はユニットを更に含む。さらに、「第1」、「第2」、「第3」、「第4」等の用語は、特定の順序を記述するのではなく、対象を区別するために使用される。

【0089】

添付図面を参照して、以下に、まず、分散型ストレージシステムのいくつかの可能なアーキテクチャについて説明する。

【0090】

図1-A及び図1-Bを参照すると、この出願の実施例におけるいくつかの技術的解決策は、図1-A若しくは図1-Bに例示されるアーキテクチャの分散型ストレージシステム、又は分散型ストレージシステムの変形アーキテクチャに基づいて具体的に実現されてもよい。分散型ストレージシステムは、インターネットを使用することにより接続された複数のストレージノードを含む（いくつかのシナリオでは、ストレージノードはまた、「ストレージサービングノード」と呼ばれてもよい）。端末は、ネットワークを使用することにより、分散型ストレージシステムへの読み取り／書き込みのようなアクセスを実行してもよい。

【0091】

この出願の実施例における端末（terminal）のプロダクト形式は、例えば、モバイルインターネットデバイス、ノートブックコンピュータ、サーバ、タブレットコンピュータ、パームトップコンピュータ、デスクトップコンピュータ、携帯電話、又はデータ読み取り要求若しくはデータ書き込み要求のようなデータアクセス要求を送信可能な他のプロダクト形式の端末デバイスでもよいことが理解され得る。この出願の実施例では、端末はまた、ホスト（host）により置換されてもよい。

【0092】

ストレージノードは1つ以上のメモリを含む。この出願の実施例におけるメモリは不揮発性記憶媒体を含み、不揮発性記憶媒体を含むメモリはまた、不揮発性メモリとも呼ばれてもよい。実施例では、メモリは、例えば、ハードディスク、消去可能コンパクトディスク又はソリッドステートディスクのような記憶媒体でもよい。図1-A及び図1-Bでは、メモリがハードディスクである例が説明のために使用される。

【0093】

ストレージノードは、ネットワークインタフェースカード、内部メモリ、プロセッサ、拡張カード等を更に含んでもよい。プロセッサは、例えば、拡張カードを使用することにより、メモリに相互接続されてもよい。ストレージノードは、ストレージノードのネットワークインタフェースカードを使用することにより、他のストレージノード又は他の周辺デバイスに相互接続されてもよい。ネットワークインタフェースカードは、ストレージノードの通信インタフェースとして考えられてもよい。

【0094】

プロセッサは機能ユニットを含んでもよく、これらの機能ユニットは、ハードウェア（例えば、プロセッサ／フィールドプログラマブルゲートアレイ（英文：field programmable gate array, 略称：FPGA）回路、又はプロセッサ／FPGA回路と他の補助ハードウェアとの組み合わせ）でもよく、或いはソフトウェアでもよい。例えば、図1-Bに例示されるアーキテクチャでは、プロセッサは、クライアント（英文：client）、データサービス（英文：data service, 略称：DS）ユニット及び状態モニタ（monitor）のような機能を実行してもよい。実行されているとき、これらの機能は、プロセッサ内のモジュールとして考えられてもよい。実行されていないとき、これらの機能を実現するためのプログラムは、内部メモリ内のプログラムでもよく、この場合、これらのモジュールは内部メ

メモリに位置するとも考えられてもよい。一般的に、これらの機能は各ストレージノードにあり、具体的には、各ストレージノード内のプロセッサと内部メモリとの組み合わせにある。説明の便宜上のため、これらのモジュールは、図1-Bではプロセッサ内に描かれている。いくつかの特別な場合、例えば、システム・オン・チップ(system on chip, SoC)では、プロセッサ自体がストレージ機能を有し、プログラムコードはプロセッサに直接焼かれてもよい。この場合、内部メモリはもはや必須のコンポーネントではない。

【0095】

clientは、対応するストレージノードのシステム入出力(英文:input output, 略称:IO)インタフェースであり、主に、他のデバイス(例えば、端末又は他のストレージノード)により送信されたアクセス要求を受信する役目をする。アクセス要求はまた、IO要求とも呼ばれる。アクセス要求は、例えば、データ読み取り要求又はデータ書き込み要求である。DSユニットは、主に、clientにより送信されたアクセス要求を受信し、アクセス要求に基づいてローカルメモリ上で読み取り／書き込み操作を実行し、読み取り／書き込み操作の結果をclientに返信する役目をする。monitorは、対応するストレージノードに含まれるメモリの状態を監視する役目をしてよく、明らかに、monitorは、対応するストレージノードの状態等を監視する役目を更にしてもよい。

【0096】

例えば、monitorは、分散型ストレージシステムにおいて各ストレージノードに配置されてもよく、各ストレージノード内のmonitorは、このストレージノードに含まれるメモリの状態のような情報を監視するように構成されてもよい。さらに、全てのストレージノードのmonitorは、monitorクラスタを形成するように一緒に接続されてもよい。さらに、1つの「プライマリmonitor」は、管理者により指定された方式で或いは選択アルゴリズムを使用することにより、monitorクラスタから選択され、残りのmonitorは「セカンダリmonitor」である。プライマリmonitorが位置するストレージノードは「管理ノード」と呼ばれてもよい。例えば、各セカンダリmonitorは、セカンダリmonitorが位置するストレージノードに含まれる各メモリの現在状態をプライマリmonitorに送信してもよい。プライマリmonitorは、各ストレージノードに含まれるメモリの状態のような収集された情報に基づいて、状態情報セットを生成してもよい。状態情報セットは、例えば、簡単には「ノードマップ(英文:node map)」、「状態ファイル」等と呼ばれてもよい。状態情報セットは、各メモリの状態のような情報を記録する。さらに、例えば、セカンダリmonitorは、セカンダリmonitorが位置するストレージノードの現在状態をプライマリmonitorに送信してもよい。状態情報セットは、各ストレージノードの状態等を更に記録してもよい。前述のように、monitorは、内部メモリ内のプログラムを実行することにより、プロセッサにより実現される機能である。一般的に、monitorはプロセッサ及び内部メモリの機能であるか、或いはmonitorはストレージノードの機能であると考えられてもよい。したがって、以下の説明では、プライマリmonitorの機能は、プライマリプロセッサにより、或いはプライマリプロセッサとプロセッサと内部メモリとにより、或いはプライマリストレージノードにより実行されてもよく、セカンダリmonitorの機能は、セカンダリプロセッサにより、或いはセカンダリプロセッサとプロセッサと内部メモリとにより、或いはセカンダリストレージノードにより実行されてもよい。

【0097】

プライマリmonitorは、最新の状態情報セットを他のストレージノードにアクティブにプッシュしてもよく、或いは他のストレージノードは、プライマリmonitorから最新の状態情報セットをアクティブに要求してもよい。他のストレージノード内のmonitor(すなわち、セカンダリmonitor)は、受信した状態情報セットを他のストレージノードの内部メモリにキャッシュしてもよく、内部メモリにキャッシュされた状態情報セットは、対応するストレージノード(対応するmonitor)によりローカルで維持される状態情報セットである。セカンダリmonitorがプライマリmonitorから状態情報セットを受信する毎に、セカンダリmonitorは、セカンダリmonitorにより現在キャッシュされている状態情報セットを更新するために、プライマリmonitorから最後に受信した状態情報セットを使用しても

10

20

30

40

50

よい。すなわち、セカンダリmonitorは、現在キャッシュされている状態情報セットが、セカンダリmonitorにより受信された最新の状態情報セットであることを確保する。

【0098】

セカンダリmonitorとプライマリmonitorとが接続状態にあるとき、セカンダリmonitorにより現在キャッシュされている状態情報セットは、通常ではプライマリmonitorによりリリースされた最新の状態情報セットであることが理解され得る。しかし、セカンダリmonitorとプライマリmonitorとが接続状態にないとき、セカンダリmonitorはプライマリmonitorによりリリースされた最新の状態情報セットを見逃す可能性があり、したがって、セカンダリmonitorにより現在キャッシュされている状態情報セットは、プライマリmonitorによりリリースされた最新の状態情報セットではない可能性がある。明らかに、セカンダリmonitorとプライマリmonitorとが接続状態にない期間内に、プライマリmonitorが新たな状態情報セットをリリースしていない場合、セカンダリmonitorにより現在キャッシュされている状態情報セットは、依然としてプライマリmonitorによりリリースされた最新の状態情報セットである可能性がある。接続状態のデバイスは、互いに直接的に或いは間接的に通信できる。したがって、管理ノードに接続されている状態にあるストレージノードは、管理ノードから最新バージョンの状態情報セットを取得できる。

10

【0099】

メモリの状態（アクセス状態）は、オンライン（英文：online）状態と、オフライン（英文：offline）状態とを含む。オンライン状態は、データ再構築状態と、信頼されたアクセス状態とを含む。データ再構築状態はまた、キャッチアップ（英文：catchup）状態とも呼ばれてもよい。信頼されたアクセス状態はまた、正常（英文：normal）状態とも呼ばれてもよい。

20

【0100】

データ再構築が実行される必要がある場合、メモリは、データ再構築が完了した後に、信頼されたアクセス状態に入ってもよく、或いはデータ再構築が実行される必要がない場合、メモリは、例えば、オフライン状態から信頼されたアクセス状態に直接変化してもよいことが理解され得る。一般的に、オフライン状態とデータ再構築状態との双方は、信頼されていないアクセス状態と考えられることができる。

【0101】

メモリの状態は、メモリ内のデータの有効性（すなわち、信頼できるか否か）を反映する。メモリが信頼されたアクセス状態にある場合、これは、メモリに割り当てられたデータが全て成功してメモリに書き込まれたことを示す。すなわち、割り当てられたデータが全て書き込まれたメモリは、信頼されたアクセス状態にある。例えば、メモリが信頼されたアクセス状態にあるとき、このような状態のメモリ上のデータは有効であると考えられてもよい。例えば、メモリがデータ再構築状態にあるとき、これは、この時点ではメモリがオフライン期間中に失われた関係するデータを再構築していることを示す。この場合、メモリ上のデータはおそらく無効であり、すなわち、メモリは信頼されていないアクセス状態にあると考えられてもよい。メモリがオフライン状態であるとき、この場合、メモリ上のデータは無効であり、すなわち、メモリは信頼されていないアクセス状態にあると考えられてもよい。オフライン状態のメモリはアクセスできず、データがメモリから読み取られることも、メモリに書き込まれることもできないことが理解され得る。

30

40

【0102】

図1-Cを参照すると、図1-Cは、メモリの異なる状態の間の遷移の可能な方式を例示する。メモリは、オンライン状態からオフライン状態に変化してもよく、或いはオフライン状態からオンライン状態に変化してもよい。具体的には、例えば、メモリは、オフライン状態からデータ再構築状態又は信頼されたアクセス状態に変化してもよく、或いはデータ再構築状態から信頼されたアクセス状態に変化してもよく、或いはデータ再構築状態又は信頼されたアクセス状態からオフライン状態に変化してもよい。

【0103】

メモリは、複数の要因のため、offlineになってもよい。例えば、ストレージノード内

50

のメモリがストレージノードにより検出できないとき、或いはストレージノードがメモリから故障状態報告（故障状態報告は、メモリがofflineであることを示す）を受信したとき、ストレージノードは、対応するメモリの状態をoffline状態に設定してもよい。他の例では、管理ノード（プライマリmonitor）がストレージノードからハートビートメッセージを受信するのに複数の連続する回数失敗したとき、これは、ストレージノードがoffline状態であることを示し、管理ノードは、ストレージノードに含まれる全てのメモリをoffline状態になるように設定してもよい。

【0104】

状態情報セットは、メモリIDと各メモリの状態とを記録してもよく、各メモリが属するストレージノードのノードIDと状態とを更に記録してもよい。図1-Dを参照すると、図1-Dは、状態情報セットの可能な表現形式を例示する。状態情報セットは、状態情報セットのバージョン番号と、各メモリ（メモリは、例えば、図面におけるハードディスクである）の識別子及び状態（状態は、例えば、オフライン状態、データ再構築状態又は信頼されたアクセス状態でもよい）を含む。

【0105】

メモリの状態は時間と共に変化するため、状態情報セットは、メモリの状態の変化時に適応的に更新される。例えば、管理ノード（プライマリmonitor）がメモリ状態変化報告（メモリの状態が変化したこと、例えば、メモリがプラグを抜かれて電源オフになったこと、又は再び電源オンになったこと）を受信したか、或いはメモリがオフラインである（例えば、管理ノードは特定の持続時間内に対応するストレージノードからハートビートメッセージを受信していない）と決定した後に、管理ノードは、状態情報セットを更新する条件が満たされたと考え、管理ノードは、状態情報セットを更新する。更新された状態情報セットは新しいバージョン番号に対応し、更新された状態情報セットは、変化したメモリ状態を記録する。管理ノードは、更新された状態情報セットを他のストレージノード（セカンダリmonitor）にプッシュする。状態情報セットのバージョン番号は、グローバル変数に基づいてプライマリmonitorにより生成されてもよい。具体的には、例えば、状態情報セットが更新される毎に、バージョン番号は現状に基づいて+1又は+2だけインクリメントする。したがって、異なるバージョン番号は異なる状態情報セットを示してもよく、2つの状態情報セットが同じであるか否かは、2つの状態情報セットのバージョン番号を比較することにより決定できる。

【0106】

分散型ストレージシステムにおけるメモリは、様々なデータボディを記憶してもよい（データボディは、メモリにより記憶できるデータを示し、データボディは、例えば、ファイル、オブジェクト、レコード、又は他のフォーマットのデータでもよい）。一般的に、システム信頼性のような要因を考えると、同じデータボディの異なるストリップは、それぞれ異なるメモリに記憶される。パリティストリップは、データストリップに対して冗長検査演算（例えば、EC演算）を実行することにより取得できる。検査冗長構成（例えば、EC構成）は、ストライプ内のデータストリップとパリティストリップとの間の比例関係を記述するために使用される。例えば、ファイルの検査冗長構成はA:Tとして記されてもよく、この場合、Aはストライプに含まれるデータストリップの数であり、T（Tは冗長を示す）はストライプに含まれるパリティストリップの数である。例えば、冗長検査アルゴリズムがECアルゴリズムであると仮定すると、検査冗長構成（すなわち、EC構成）が2:1であるとき、2はストライプ（ECストライプ）に含まれるデータストリップの数であり、1はストライプに含まれるパリティストリップの数である。すなわち、この場合、ストライプは2個のデータストリップと1個のパリティストリップとを含み、ストライプに含まれる2個のデータストリップ及び1個のパリティストリップは、それぞれ異なるメモリに記憶される。すなわち、ストライプに含まれる2個のデータストリップ及び1個のパリティストリップを記憶するために、3個のメモリが必要である。他の例では、検査冗長構成（すなわち、EC構成）が16:4であると仮定すると、16はストライプ（ECストライプ）に含まれるデータストリップの数であり、4はストライプに含まれるパリティストリップの数である。

すなわち、この場合、ストライプは16個のデータストリップと4個のパリティストリップとを含み、この場合、ストライプに含まれる20個のストリップをそれぞれ記憶するために、20個のメモリが必要である。他の場合も類推して理解され得る。図1-Eは、2個のデータストリップと1個のパリティストリップとを含むストライプを例示する。図1-Fは、複数のデータストリップと複数のパリティストリップとを含むストライプを示す。

【0107】

さらに、1つの分散型ストレージシステムは1つ以上の検査冗長構成をサポートでき、各検査冗長構成は1つ以上の検査メンバグループに対応してもよく、各検査メンバグループは複数のメモリを含む。明らかに、異なる検査メンバグループに含まれるメモリは、完全に異なってもよく、或いは部分的に異なってもよく、すなわち、各メモリは、1つ以上の検査メンバグループに仕えてもよい。例えば、検査冗長構成が2:1である場合、検査冗長構成に対応する各検査メンバグループは、3個(2+1=3)のメモリを含む。例えば、検査冗長構成が4:2である場合、検査冗長構成に対応する各検査メンバグループは、6個(4+2=6)のメモリを含む。例えば、検査冗長構成が16:4である場合、検査冗長構成に対応する各検査メンバグループは、20個(16+4=20)のメモリを含み、以下同様である。

【0108】

さらに、ファイルの長さが大きすぎる(例えば、ファイルの長さが1Gbより大きい)とき、場合によって、管理の便宜上のため、ファイルはファイルセグメント、例えば、500Mb又は1Gbより大きくないファイルセグメントに分割されてもよい。図2-Aを参照すると、図2-Aは、ファイルがファイルセグメントに分割される概略図である。ファイルの各ファイルセグメントは同じ或いは異なる検査冗長構成に対応してもよく、異なるファイルセグメントの検査メンバグループは同じでもよく、或いは異なってもよい。この場合、ファイルセグメントはファイルとして考えられてもよい。例えば、図2-Bを参照すると、図2-Bは、ファイルがファイルセグメントに分割されない概略図である。

【0109】

以下に、いくつかの実施例を使用することにより、この出願におけるいくつかの関係する方法について説明する。

【0110】

実施例1

添付図面を参照して、以下に、まず、状態更新方法を説明するための例を使用する。状態更新方法は分散型ストレージシステムに適用される。例えば、分散型ストレージシステムはm個のストレージノードを含む。m個のストレージノードは、例えば、ストレージノードNd2、Nd1、Nd3及びNd4を含んでもよい。例えば、ストレージノードNd1は管理ノードである。分散型ストレージシステムがより多くの或いはより少ないストレージノードを含む場合における実現方式も類推により理解され得る。以下の例示の解決策では、非管理ノードは、主に、管理ノードに対して、状態情報セットを更新及びリリースするようにアクティブにトリガーする。

【0111】

図3-Aを参照すると、図3-Aは、この出願の実施例による状態更新方法の概略フローチャートである。図3-Aを参照すると、この出願の実施例において提供される状態更新方法は以下のステップを含んでもよい。

【0112】

301. ストレージノードNd2内のメモリNd2-D1の状態が変化したとき、ストレージノードNd2は、メモリ状態変化報告Report1を管理ノードNd1に送信し、メモリ状態変化報告Report1は、メモリNd2-D1の最新の変化した状態を示すために使用される。

【0113】

メモリ状態変化報告Report1は、メモリNd2-D1のメモリ識別子、状態識別子等を搬送してもよい。例えば、メモリNd2-D1の最新の変化した状態がoffline状態であるとき、メモリ状態変化報告Report1で搬送されるメモリNd2-D1の状態識別子は、offline状態を示す。他の例では、メモリNd2-D1の最新の変化した状態がデータ再構築状態であるとき、メモリ

状態変化報告Report1で搬送されるメモリNd2-D1の状態識別子は、データ再構築状態を示す。更に他の例では、メモリNd2-D1の最新の変化した状態が信頼されたアクセス状態であるとき、メモリ状態変化報告Report1で搬送されるメモリNd2-D1の状態識別子は、信頼されたアクセス状態を示し、以下同様である。

【0114】

302.管理ノードNd1は、ストレージノードNd2からメモリ状態変化報告Report1を受信し、管理ノードNd1は、メモリ状態変化報告Report1に基づいて、管理ノードNd1により現在キャッシュされている状態情報セットを更新する。

【0115】

具体的には、例えば、管理ノードNd1は、管理ノードNd1により現在キャッシュされている状態情報セットに記録されたメモリNd2-D1の状態を、メモリ状態変化報告Report1により示されるメモリNd2-D1の状態に更新し、状態情報セット内のバージョン番号（すなわち、状態情報セット自体のバージョン番号）を更新する。

10

【0116】

例えば、monitorが各ストレージノードに配置されるとき、具体的には、ストレージノードNd2内のセカンダリmonitorは、メモリ状態変化報告Report1をストレージノードNd1に送信してもよい。対応して、管理ノードNd1内のプライマリmonitorは、ストレージノードNd2からメモリ状態変化報告Report1を受信してもよい。

【0117】

例えば、管理ノードNd1により更新されていない状態情報セットが図3-Bに例示される。例えば、管理ノードNd1により更新されている状態情報セットが図3-Cに例示される。明らかに、実際の用途では、状態情報セットはこのような例示の形式に限定されない。

20

【0118】

303.管理ノードNd1は、最新の更新された状態情報セットを他のストレージノードに送信する。

【0119】

対応して、ストレージノードNd2、Nd3及びNd4は、管理ノードNd1から更新された状態情報セットを別々に受信する。ストレージノードNd2、Nd3及びNd4は、これらのそれぞれ現在キャッシュされている状態情報セットを更新するために、管理ノードNd1からの状態情報セットを別々に使用する。このように、ストレージノード（Nd1、Nd2、Nd3及びNd4）によりキャッシュされた状態情報セットは、相対的に同期され続けることができる。

30

【0120】

前述の解決策は、非管理ノードが、主に、管理ノードに対して、状態情報セットを更新及びリリースするようにアクティブにトリガーする機構を提供する。具体的には、ストレージノード内のメモリの状態が変化したとき、ストレージノードは、メモリ状態変化報告を管理ノードにアクティブに送信し、それにより、管理ノードに対して、メモリ状態変化報告に基づいてタイムリーに状態情報セットを更新及びリリースするようにトリガーする。この機構は、分散型ストレージシステムにおける各ストレージノードが状態情報セットを最大限まで同期させることを可能にするのに役立ち、各ストレージノードが分散型ストレージシステムにおける各メモリの状態を比較的正確に習得することを可能にするのに役立ち、メモリの状態に基づいて実行される以降の読み取り／書き込み操作の基礎としての役割をする。

40

【0121】

実施例2

添付図面を参照して、以下に、他の状態更新方法を説明するための例を使用する。状態更新方法は分散型ストレージシステムに適用される。例えば、分散型ストレージシステムはm個のストレージノードを含んでもよい。m個のストレージノードは、例えば、ストレージノードNd2、Nd1、Nd3、…、及びNd4を含んでもよい。例えば、ストレージノードNd1は管理ノードである。分散型ストレージシステムがより多くの或いはより少ないストレージ

50

ノードを含む場合における実現方式も類推により理解され得る。以下の例示の解決策では、管理ノードは、主に、他のストレージノードに対するハートビート監視の結果に基づいて状態情報セットを更新及びリリースする。

【0122】

図4-Aを参照すると、図4-Aは、この出願の他の実施例による他の状態更新方法の概略フローチャートである。図4-Aを参照すると、この出願の他の実施例において提供される他の状態更新方法は以下のステップを含んでもよい。

【0123】

401. 管理ノードNd1は、各ストレージノードにより周期的に送信されるハートビートメッセージを受信する。

10

【0124】

402. 管理ノードNd1が指定の持続時間（指定の持続時間は、例えば、5分、10分、20分、1分又は他の持続時間でもよい）内に、ストレージノードNd2により送信されるハートビートメッセージを受信していないとき、管理ノードNd1は、ストレージノードNd2がオフライン状態にあると考えてもよく、ストレージノードNd2がオフライン状態にあると考えたとき、管理ノードNd1は、管理ノードNd1により現在キャッシュされている状態情報セットを更新する。

【0125】

具体的には、管理ノードNd1は、管理ノードNd1に現在キャッシュされている状態情報セットに記録された、ストレージノードNd2に含まれる全てのメモリの状態をオフライン状態に更新し、状態情報セットのバージョン番号を更新する。

20

【0126】

例えば、monitorが各ストレージノードに配置されるとき、具体的には、管理ノードNd1内のプライマリmonitorは、各ストレージノード内のセカンダリmonitorにより周期的に送信されるハートビートメッセージを受信してもよい。

【0127】

例えば、ストレージノードNd2はメモリNd2-D1、Nd2-D2及びNd2-D3を含む。この場合、例えば、管理ノードNd1により更新されていない状態情報セットが図4-Bに例示され、例えば、管理ノードNd1により更新されている状態情報セットが図4-Cに例示される。明らかに、実際の用途では、状態情報セットはこのような例示の形式に限定されない。

30

【0128】

403. 管理ノードNd1は、最新の更新された状態情報セットを他のストレージノードに送信する。

【0129】

対応して、ストレージノードNd3及びNd4は、管理ノードNd1から更新された状態情報セットを別々に受信する。ストレージノードNd3及びNd4は、これらのそれぞれ現在キャッシュされている状態情報セットを更新するために、管理ノードNd1からの状態情報セットを別々に使用する。このように、管理ノードNd1によりキャッシュされた状態情報セットは、他のストレージノード（例えば、Nd3及びNd4）によりキャッシュされた状態情報セットと相対的に同期され続けることができる。しかし、ストレージノードNd2はオフライン状態にあり、したがって、管理ノードNd1によりリリースされた最新の状態情報セットを受信できない。したがって、オフライン期間中に、ストレージノードNd2によりキャッシュされた状態情報セットは、管理ノードによりキャッシュされた状態情報セットと同期され続けることはできない。ストレージノードNd2が再びオンラインになったとき、ストレージノードNd2は、ストレージノードNd2にキャッシュされた状態情報セットを更新するために、管理ノードから最新の状態情報セットをアクティブに要求してもよい。明らかに、ストレージノードNd2は、代替として、管理ノードが最新の状態情報セットを再びリリースするまで、パッシブに待機し、ストレージノードNd2によりキャッシュされた状態情報セットを更新するために、管理ノードにより再びリリースされた最新の状態情報セットを使用してもよい。

40

50

【0130】

前述の解決策は、管理ノードが、主に、他のストレージノードに対するハートビート監視の結果に基づいて状態情報セットを更新及びリリースする機構を提供する。この機構は、分散型ストレージシステムにおける各ストレージノードが状態情報セットを最大限まで同期させることを可能にするのに役立ち、各ストレージノードが分散型ストレージシステムにおける各メモリの状態を比較的正確に習得することを可能にするのに役立ち、メモリの状態に基づいて実行される以降の読み取り／書き込み操作の基礎としての役割をする。

【0131】

いくつかの可能な実現方式では、図3-A及び図4-Aにおいて説明した状態更新方法は共に、同じ分散型ストレージシステムに適用されてもよいことが理解され得る。すなわち、管理ノードは、他のストレージノードに対するハートビート監視の結果に基づいて状態情報セットを更新及びリリースし、代替として、非管理ノードは、管理ノードに対して、状態情報セットを更新及びリリースするようにアクティブにトリガーしてもよい。

【0132】

実施例3

添付図面を参照して、以下に、データアクセス方法を説明するための例を更に使用し、これは、端末が分散型ストレージシステムからのデータを書き込む可能なシナリオに主に使用される。データアクセス方法は分散型ストレージシステムに適用される。例えば、分散型ストレージシステムはm個のストレージノードを含み、m個のストレージノードは、ストレージノードNd1及びNd2を含む。メモリNd2-D1、メモリNd2-D2及びメモリNd2-D3は全てストレージノードNd2に位置する。この実施例は、主に、対応する検査冗長構成が具体的に2:1であり、すなわち、ストライプが2個のデータストリップと1個のパリティストリップとを含む例を使用することにより説明する。他の検査冗長構成の場合も類推により理解され得る。

【0133】

図5-Aを参照すると、図5-Aは、この出願の実施例によるデータアクセス方法の概略フローチャートである。図5-Aに例示するように、この出願の実施例において提供されるデータアクセス方法は以下のステップを含んでもよい。

【0134】

501. 端末がデータを分散型ストレージシステムに書き込む必要があるとき、端末は、書き込み対象のデータを搬送するデータ書き込み要求を分散型ストレージシステムに送信する。以下の説明の便宜上のため、この実施例ではデータ書き込み要求はWq3という名で呼ばれる。このステップでは、ストレージノードNd1は端末からデータ書き込み要求Wq3を受信する。

【0135】

データ書き込み要求Wq3は、データData1（データData1は書き込み対象のデータである）と、データData1が属するファイルのファイル識別子と、データData1の位置情報とを搬送してもよい。データData1の位置情報は、データData1が属するファイル内のデータData1の位置を記述するために使用される。データData1の位置情報は、例えば、データData1のファイルオフセットアドレス（データData1のファイルオフセットアドレスはファイル内のデータData1の開始位置を示す）、データData1の長さ等を含む。

【0136】

502. ストレージノードNd1は、データData1を2個のデータストリップ（以下の説明の便宜上のため、この実施例では2個のデータストリップはデータストリップPd1及びデータストリップPd2という名で呼ばれる）にセグメント化し、ストレージノードNd1は、2個のデータストリップを使用することによる計算を用いて、1個のパリティストリップ（以下の説明の便宜上のため、この実施例ではパリティストリップはパリティストリップPj1という名で呼ばれる）を取得し、1個のパリティストリップ及び2個のデータストリップは、3個のストリップを含む1個のストライプを形成する。例えば、データストリップPd1及びデータストリップPd2は長さが同じでもよく、或いは異なってもよい。

【0137】

503. ストレージノードNd1は、データData1の位置情報に基づいて、データData1を書き込むことに関するメモリを決定する。

【0138】

例えば、ストレージノードNd1により決定された、データData1を書き込むことに関するメモリは、メモリNd2-D1、メモリNd2-D2及びメモリNd2-D3である。

【0139】

一般的に、メモリへのファイルデータの割り当て中に、ファイルデータのセグメントが記憶のために具体的に割り当てられるメモリは、通常では、ファイル内のファイルデータの位置に基づいて決定される。したがって、データData1を書き込むことに関するメモリは、データData1の位置情報に基づいて決定されてもよい。

10

【0140】

前述の3個のメモリ（Nd2-D1、Nd2-D2及びNd2-D3）は3個のストリップと1対1の対応関係にある。具体的には、例えば、パリティストリップPj1はメモリNd2-D3に対応し、データストリップPd1はメモリNd2-D1に対応し、データストリップPd2はメモリNd2-D2に対応する。

【0141】

504. ストレージノードNd1は、ストレージノードNd1によりローカルにキャッシュされた状態情報セットに現在記録されているメモリNd2-D1、メモリNd2-D2及びメモリNd2-D3の状態を決定する。

20

【0142】

ストレージノードNd1によりローカルにキャッシュされた状態情報セットが、メモリNd2-D2及びNd2-D3がオンライン状態（信頼されたアクセス状態又はデータ再構築状態）にあり、メモリNd2-D1がオフライン状態にあることを現在記録している場合が、以下の説明のための例として使用される。

【0143】

505. メモリNd2-D2及びメモリNd2-D3がオンライン状態にあり、メモリNd2-D1がオフライン状態にあると決定されたとき、ストレージノードNd2はデータストリップPd2をメモリNd2-D2に書き込み、ストレージノードNd2はパリティストリップPj1をメモリNd2-D3に書き込む。

30

【0144】

EC検査アルゴリズムによれば、特定の数のストリップが書き込まれるのに失敗したとしても、書き込まれるのに失敗したストリップは、成功して書き込まれたストリップを使用することによる再構築により回復でき、データロスが引き起こされないため、このステップにおいて、データは、オフラインストレージノードが存在するとしても、通常では残りの非オフラインストレージノードに書き込まれる。書き込まれるのに失敗したストリップの数がパリティストリップの数を超えないという条件で、EC検査アルゴリズムは通常通り機能できる。明らかに、他の検査アルゴリズムも同様である。

【0145】

506. ストレージノードNd2内のメモリNd2-D1はオフライン状態にあり、したがって、データストリップPd1はメモリNd2-D1に書き込まれるのに失敗するため、ストレージノードNd1又はストレージノードNd2は再構築ログlog1を生成する。

40

【0146】

再構築ログlog1は、書き込まれるのに失敗したストリップ（データストリップPd1）のストリップ識別子と、ストリップが書き込まれるのに失敗したメモリ（メモリNd2-D1）のメモリ識別子とを記録する。再構築ログlog1は、書き込まれるのに失敗したストリップ（データストリップPd1）が属するストライプのストライプ識別子等を更に記録してもよい。再構築ログlog1は、メモリNd2-D1が再びオンラインになった後に、データストリップPd1の以降の再構築に使用される。図5-Bは、再構築ログlog1の可能な表現形式を例示する。明らかに、他の表現形式も使用されてもよい。

50

【0147】

メモリNd2-D1もまた信頼されたアクセス状態にある場合、ストレージノードNd2は、データストリップPd1をメモリNd2-D1に直接書き込んでもよいことが理解され得る。この場合、ストレージノードNd1又はストレージノードNd2は、再構築ログlog1を生成する関係するステップを実行する必要はない。

【0148】

前述の解決策が主に使用されるデータ書き込みシナリオは、ストライプ内のストリップに対応するメモリが全て同じストレージノードに位置する可能なシナリオである。この可能なシナリオでは、メモリ状態を記録するために使用される状態情報セットが分散型ストレージシステムに導入され、したがって、状態情報セットは、分散型ストレージシステムにおけるメモリのアクセス状態を記録及び管理するために使用されてもよい。データは、オンライン状態のメモリに直接書き込まれることができる。オフライン状態のメモリについて、関係する再構築ログは、データ書き込みの試みが失敗した後に、関係する再構築ログを生成する必要なく、直接生成できる。この実施例で例示されるシナリオにおいて、従来技術では、まず、データ書き込みの試みが行われ、関係する再構築ログは、データ書き込みの試みが失敗した後に生成される。この実施例の解決策では、無駄な試みが低減でき、それにより、データ書き込み効率を改善し、さらに、システム性能を改善するのに役立つ。

【0149】

実施例4

添付図面を参照して、以下に、他のデータアクセス方法を説明するための例を更に使用し、これは、端末が分散型ストレージシステムからのデータを書き込む他の可能なシナリオに主に使用される。前の実施例では、ストリップに対応するメモリは同じストレージノードに位置する。この実施例のシナリオでは、これらのメモリは異なるストレージノードに位置する。

【0150】

例えば、分散型ストレージシステムはm個のストレージノードを含み、m個のストレージノードは、ストレージノードNd2、Nd1、Nd3及びNd4を含む。メモリNd2-D1はストレージノードNd2に属し、メモリNd3-D1はストレージノードNd3に属し、メモリNd4-D1はストレージノードNd4に属する。この実施例は、主に、検査冗長構成が具体的に2:1であり、すなわち、ストライプが2個のデータストリップと1個のパリティストリップとを含む例を使用することにより説明する。他の検査冗長構成の場合も類推により理解され得る。

【0151】

図6を参照すると、図6は、この出願の他の実施例による他のデータアクセス方法の概略フローチャートである。図6に例示するように、この出願の他の実施例において提供される他のデータアクセス方法は以下のステップを含んでもよい。

【0152】

601及び602.ステップ601及び602はステップ501及び502と同じである。したがって、関係する説明について、対応するステップ501及び502を参照し、詳細はここでは再び説明しない。

【0153】

603.ストレージノードNd1は、データData1の位置情報に基づいて、データData1を書き込むことに関するメモリを決定する。

【0154】

例えば、ストレージノードNd1により決定された、データData1を書き込むことに関するメモリは、メモリNd2-D1、メモリNd3-D1及びメモリNd4-D1である。前述の3個のメモリ(Nd2-D1、Nd3-D1及びNd4-D1)は3個のストリップと1対1の対応関係にある。具体的には、例えば、パリティストリップPj1はメモリNd4-D1に対応し、データストリップPd1はメモリNd2-D1に対応し、データストリップPd2はメモリNd3-D1に対応する。

【0155】

604. ストレージノードNd1は、ストレージノードNd1によりローカルにキャッシュされた状態情報セットに現在記録されているメモリNd2-D1、メモリNd3-D1及びメモリNd4-D1の状態を決定する。ストレージノードNd1によりローカルにキャッシュされた状態情報セットが、メモリNd3-D1及びNd4-D1がオンライン状態にあり、メモリNd2-D1がオフライン状態にあることを現在記録している場合が、説明のための例として使用される。

【0156】

具体的には、ストレージノードNd1と、管理ノードとが接続状態にある場合、ストレージノードNd1は、ストレージノードNd1によりローカルにキャッシュされた状態情報セットに現在記録されているメモリNd2-D1、メモリNd3-D1及びメモリNd4-D1の状態を決定できる。

10

【0157】

605. メモリNd3-D1及びメモリNd4-D1がオンライン状態にあり、メモリNd2-D1がオフライン状態にあると決定されたとき、ストレージノードNd1は、データ書き込み要求Wq1を、メモリNd3-D1が属するストレージノードNd3に送信し、ストレージノードNd1は、データ書き込み要求Wq2を、メモリNd4-D1が属するストレージノードNd4に送信する。

【0158】

例えば、データ書き込み要求Wq1は、データストリップPd2と、現在ストレージノードNd1によりローカルにキャッシュされている状態情報セットのバージョン番号とを搬送する。さらに、データ書き込み要求Wq1は、データストリップPd2の長さ（長さは、例えば、110Kbでもよい）、データストリップPd2のファイルオフセットアドレス、データストリップPd2が属するストライプのストライプ識別子、データストリップPd2のストリップ識別子、データストリップPd2が属するファイルのファイル識別子等を更に搬送してもよい。

20

【0159】

他の例では、データ書き込み要求Wq2は、パリティストリップPj1と、現在ストレージノードNd1によりローカルにキャッシュされている状態情報セットのバージョン番号とを搬送してもよい。さらに、データ書き込み要求Wq2は、パリティストリップPj1の長さ、パリティストリップPj1が属するストライプのストライプ識別子、パリティストリップPj1のストリップ識別子等を更に搬送してもよい。

【0160】

ストレージノードNd1は、ストレージノードNd1によりキャッシュされた状態情報セットのバージョン番号を、他のストレージノードに送信されるデータ書き込み要求に追加し、それにより、他のストレージノードは、バージョン番号を比較することにより、ストレージノードNd1によりキャッシュされた状態情報セットが有効であるか否かを検証できることが理解され得る。

30

【0161】

606. ストレージノードNd1からデータ書き込み要求Wq1を受信した後に、ストレージノードNd3は、データ書き込み要求Wq1で搬送された状態情報セットのバージョン番号を、ストレージノードNd3により現在キャッシュされている状態情報セットのバージョン番号と比較する。

【0162】

データ書き込み要求Wq1で搬送された状態情報セットのバージョン番号が、ストレージノードNd3により現在キャッシュされている状態情報セットのバージョン番号と同じであると比較により決定され、ストレージノードNd3と、管理ノードとが接続状態にある（すなわち、ストレージノードNd3内のセカンダリmonitorが通常通り管理ノード内のプライマリmonitorに接続される）場合、ストレージノードNd3内のセカンダリmonitorは、管理ノード内のプライマリmonitorによりリリースされた最新バージョンの状態情報セットを通常通りタイムリーに受信できると考えられてもよい。したがって、この場合、ストレージノードNd3により現在キャッシュされている状態情報セットは、プライマリmonitorにより維持される最新バージョンの状態情報セットであると考えられてもよい。データ書き込み要求Wq1で搬送された状態情報セットのバージョン番号は、ストレージノードNd3により現

40

50

在キャッシュされている状態情報セットのバージョン番号と同じであり、すなわち、ストレージノードNd1により現在キャッシュされている状態情報セットもまた、プライマリmonitorにより維持される最新バージョンの状態情報セットである。この場合、ストレージノードNd1により現在キャッシュされている状態情報セットに記録されたメモリNd3-D1の状態（信頼されたアクセス状態）もまた、実際の状況と一致すると想定され、ストレージノードNd3は、データストリップPd2をメモリNd3-D1に書き込む。

【0163】

さらに、データ書き込み要求Wq1で搬送された状態情報セットのバージョン番号が、ストレージノードNd3により現在キャッシュされている状態情報セットのバージョン番号と異なると比較により決定された場合、Nd1に記録されたNd3-D1の状態が正確でないと考えられてもよい。言い換えると、Nd1に記録されたNd3-D1の状態は信頼されたアクセス状態であるが、Nd3-D1の実際の状態が信頼されたアクセス状態であることが確保できない。Nd3-D1は信頼されていないアクセス状態にあってもよい。そのため、ストレージノードNd3は、データストリップPd2をメモリNd3-D1に書き込むことを拒否してもよい。

10

【0164】

ストレージノードと、管理ノードとが接続状態にあるとき、ストレージノードによりキャッシュされた状態情報セットが、管理ノードによりリリースされた最新の状態情報セットであるという比較的高い可能性が存在する点に特に留意すべきである。しかし、いくつかの比較的特別な場合、ストレージノードと、管理ノードとが接続状態にあるとしても、ちょうど管理ノードが状態情報セットをリリースした時点でストレージノードが一時的に故障になった場合、或いはちょうど管理ノードが状態情報セットをリリースした時点でストレージノードと管理ノードとの間の通信の品質が極めて悪い場合、ストレージノードが管理ノードによりリリースされた最新の状態情報セットを取得できないことがあり得る。すなわち、ストレージノードと、管理ノードとが接続状態にあるとしても、ストレージノードによりキャッシュされた状態情報セットが、管理ノードによりリリースされた最新の状態情報セットでないという非常に低い可能性が依然として存在する。したがって、ストレージノードと、管理ノードとが接続状態にあるとき、これらのそれぞれ現在キャッシュされている状態情報セットのバージョン番号が更に比較され、バージョン番号が同じである場合、これらのそれぞれ現在キャッシュされている状態情報セットが有効であると考えられてもよい。このやり方は信頼性を更に改善できる。

20

30

【0165】

さらに、ストレージノードNd3は、データ書き込み要求Wq1に応答するために使用されるデータ書き込み応答Wr1をストレージノードNd1に送信する。データ書き込み応答Wr1は、データストリップPd2に対する書き込み操作の結果を搬送する。具体的には、ストレージノードNd3がデータストリップPd2をメモリNd3-D1に成功して書き込んだとき、書き込み操作の結果は書き込み成功であり、或いはストレージノードNd3がデータストリップPd2をメモリNd3-D1に書き込むのに失敗した（例えば、ストレージノードNd2がデータストリップをメモリNd3-D1に書き込むことを拒否する）とき、書き込み操作の対応する結果は書き込み失敗である。

【0166】

607. ストレージノードNd1からデータ書き込み要求Wq2を受信した後に、ストレージノードNd4は、パリティストリップPj1をメモリNd4-D1に書き込むか否かを決定するために、ステップ606と同様に、状態情報セットのバージョン番号の間の比較を行う（違いは、ストレージノードNd3がストレージノードNd4に変わり、データ書き込み要求Wq1がデータ書き込み要求Wq2に変わることにある）。ステップ607の具体的な実行プロセスはここでは再び説明しない。

40

【0167】

ステップ606及びステップ607は、同時に実行されてもよく、或いは1つずつ実行されてもよい。

【0168】

50

608. ストレージノードNd1は、ストレージノードNd3からデータ書き込み応答Wr1を受信し、ストレージノードNd1は、ストレージノードNd4からデータ書き込み応答Wr2を受信する。

【0169】

さらに、データ書き込み応答が対応する書き込み失敗を示す場合、ストレージノードNd1は、再書き込みを要求するために、対応するデータ書き込み要求を再送信してもよい。例えば、データ書き込み応答Wr1が対応する書き込み失敗を示す場合、ストレージノードNd1は、データ書き込み要求Wq1をストレージノードNd3に再送信してもよい。

【0170】

609. ストレージノードNd2内のメモリNd2-D1はオフライン状態にあり、データストリップPd1はメモリNd2-D1に書き込まれるのに失敗するため、ストレージノードNd1は再構築ログlog1を生成し、ストレージノードNd1は、再構築ログlog1をストリップが成功して書き込まれたストレージノード（例えば、ストレージノードNd3及び／又はストレージノードNd4）に送信する。対応して、ストレージノードNd3及び／又はストレージノードNd4は、再構築ログlog1を受信及び記憶してもよく、再構築ログlog1は、メモリNd2-D1が再びオンラインになった後に、データストリップPd1の以降の再構築に使用される。

【0171】

再構築ログlog1は、書き込まれるのに失敗したストリップ（データストリップPd1）のストリップ識別子と、ストリップが書き込まれるのに失敗したメモリ（メモリNd2-D1）のメモリ識別子とを記録する。再構築ログlog1は、書き込まれるのに失敗したストリップ（データストリップPd1）が属するストライプのストライプ識別子等を更に記録してもよい。

【0172】

前述の解決策が主に使用されるデータ書き込みシナリオは、ストライプ内のストリップに対応するメモリが同じストレージノードに位置しない可能なシナリオである。この可能なシナリオでは、メモリ状態を記録するために使用される状態情報セットが分散型ストレージシステムに導入され、したがって、状態情報セットは、分散型ストレージシステムにおけるメモリのアクセス状態を記録及び管理するために使用されてもよい。

【0173】

具体的には、この実施例は、書き込み対象のデータを書き込むための方法を記載しており、書き込み対象のデータに対応するストライプ内のストリップに対応するメモリは、同じストレージノードに位置しない。ストレージノードNd1と、管理ノードとが接続状態にあり、ストレージノードNd1によりキャッシュされた状態情報セットが、メモリNd3-D1及びNd4-D1がオンラインであることを記録するとき、ストレージノードNd1は、データがメモリNd3-D1及びNd4-D1に成功して書き込まれることができると前もって考える。ストレージノードNd1によりキャッシュされた状態情報セットの有効性を検証するために、ストレージノードNd1は、ストレージノードNd1によりキャッシュされた状態情報セットのバージョン番号を、他のストレージノード（ストレージノードNd3及びストレージノードNd4）に送信されるデータ書き込み要求に追加する。他のストレージノードは、データ書き込み要求で搬送された状態情報セットのバージョン番号を、他のストレージノードによりキャッシュされた状態情報セットのバージョン番号と比較する。データ書き込み要求で搬送された状態情報セットのバージョン番号が、他のストレージノードにより現在キャッシュされている状態情報セットのバージョン番号と同じであるとき、ストレージノードNd1によりキャッシュされた状態情報セットが有効であると考えられてもよい。したがって、状態情報セットに記録された関係するメモリの状態は正確であると想定される。この場合、対応するデータ書き込みは通常では成功であり、したがって、データは、関係するメモリに直接書き込まれる。そうでない場合、ストレージノードNd1によりキャッシュされた状態情報セットが無効である可能性があると考えられてもよく、したがって、状態情報セットに記録された関係するメモリの状態は不正確である可能性がある。この場合、対応するデータ書き込みは失敗する可能性が非常に高く、従って、書き込みを拒否することが比較的適

10

20

30

40

50

切なやり方である。

【0174】

この実施例の効果は以下の点で前述の実施例のものと同様である。オフライン状態のメモリについて、関係する再構築ログは、データ書き込みの試みが失敗した後に、従来の方式で関係する再構築ログを生成する必要なく、直接生成でき、それにより、データ書き込み効率を改善し、システム性能を改善するのに役立つ。

【0175】

実施例5

添付図面を参照して、以下に、データ再構築方法を説明するための例を更に使用する。データ再構築方法は分散型ストレージシステムに適用される。例えば、分散型ストレージシステムはm個のストレージノードを含んでもよく、m個のストレージノードは、ストレージノードNd1、Nd2、Nd3及びNd4を含む。ストレージノードNd1は、状態情報セットをリリースする管理ノード（すなわち、プライマリmonitorが位置するストレージノード）である。例えば、メモリNd2-D1はストレージノードNd2に属し、メモリNd3-D1はストレージノードNd3に属し、メモリNd4-D1はストレージノードNd4に属する。データ再構築方法は、例えば、図6に例示される解決策の後に実行されてもよい。

10

【0176】

図7-Aを参照すると、図7-Aは、この出願の実施例によるデータ再構築方法の概略フローチャートである。図7-Aに例示するように、この出願の実施例において提供されるデータ再構築方法は以下のステップを含んでもよい。

20

【0177】

701. ストレージノードNd2が、メモリNd2-D1がオンラインになったと決定したとき、ストレージノードNd2は、メモリ状態変化報告P2を管理ノードNd1に送信し、メモリ状態変化報告P2は、メモリNd2-D1の状態がオフライン状態からデータ再構築状態に変化したことを示してもよい。具体的には、例えば、メモリ状態変化報告P2は、メモリ識別子と、メモリNd2-D1の状態識別子とを搬送してもよい（状態識別子はデータ再構築状態の状態識別子であり、すなわち、状態識別子により示される状態はデータ再構築状態である）。

【0178】

702. 管理ノードNd1は、ストレージノードNd2からメモリ状態変化報告P2を受信し、管理ノードNd1は、キャッシュされた状態情報セットを更新する。具体的には、管理ノードNd1は、管理ノードNd1によりキャッシュされた状態情報セットに記録されたメモリNd2-D1の状態を、メモリ状態変化報告P2により示されるメモリNd2-D1の状態に更新し、状態情報セットに記録されたバージョン番号を更新する。

30

【0179】

例えば、monitorが各ストレージノードに配置されるとき、具体的には、ストレージノードNd2内のセカンダリmonitorは、メモリ状態変化報告P2を管理ノードNd1に送信してもよい。対応して、管理ノードNd1内のプライマリmonitorは、ストレージノードNd2からメモリ状態変化報告P2を受信してもよい。

【0180】

例えば、管理ノードNd1により更新されていない状態情報セットが図3-Cに例示される。例えば、管理ノードNd1により更新されている状態情報セットが図7-Bに例示される。明らかに、実際の用途では、状態情報セットはこのような例示の形式に限定されない。

40

【0181】

703. ストレージノードNd1は、更新された状態情報セットを他のストレージノードに送信する。

【0182】

対応して、ストレージノードNd2、Nd3及びNd4は、管理ノードNd1から更新された状態情報セットを別々に受信する。ストレージノードNd2、Nd3及びNd4は、これらの現在キャッシュされている状態情報セットを更新するために、管理ノードNd1からの状態情報セット

50

を使用する。このように、ストレージノード（Nd1、Nd2、Nd3及びNd4）により維持される状態情報セットは、相対的に同期され続けることができる。

【0183】

704. ストレージノードNd1は、ストレージノードNd2に対して、メモリNd2-D1のためのデータ再構築を実行するように命令する。

【0184】

管理ノードNd1は、更新された状態情報セットを送信することにより、ストレージノードNd2に対して、メモリNd2-D1のためのデータ再構築を実行するように命令してもよく、代替として、管理ノードNd1は、他のメッセージ（例えば、データ再構築通知メッセージ）を使用することにより、ストレージノードNd2に対して、メモリNd2-D1のためのデータ再構築を実行するように命令してもよい。

10

【0185】

705. ストレージノードNd2が管理ノードNd1からメモリNd2-D1のためのデータ再構築の通知を受信した後に、ストレージノードNd2は、ストレージノードNd3及び／又はストレージノードNd4から、メモリNd2-D1のオフライン期間中に生成された関係する再構築ログを収集する。

【0186】

706. ストレージノードNd2は、メモリNd2-D1のオフライン期間中に生成された、収集された関係する再構築ログに基づいて、データを再構築する。

【0187】

20

具体的には、例えば、ストレージノードNd2は、メモリNd2-D1のオフライン期間中に生成された再構築ログlog1を収集する。再構築ログlog1は、メモリNd2-D1のオフライン期間中に書き込まれるのに失敗したストリップ（データストリップPd1）のストリップ識別子と、ストリップがメモリNd2-D1のオフライン期間中に書き込まれるのに失敗したメモリ（メモリNd2-D1）のメモリ識別子とを記録する。さらに、再構築ログlog1は、書き込まれるのに失敗したストリップ（データストリップPd1）が属するストライプのストライプ識別子等を記録してもよい。

【0188】

ストレージノードNd2は、データストリップPd1が属するストライプのストライプ識別子に基づいて、データストリップPd1が属するストライプを決定し、データストリップPd1が属するストライプの他のストリップ（データストリップPd2及びパリティストリップPj1）を取得してもよい。ストレージノードNd2は、データストリップPd1を再構築するために、データストリップPd2及びパリティストリップPj1に基づいて冗長検査演算を実行する。ストレージノードNd2は、再構築されたデータストリップPd1をメモリNd2-D1に書き込む。

30

【0189】

ストレージノードNd2がメモリNd2-D1のオフライン期間中に生成された複数の関係する再構築ログを収集したとき、ストレージノードNd2は、まず、複数の収集された再構築ログに対して重複排除を実行し、次に、重複排除の後に取得された各再構築ログに基づいてデータ再構築を別々に実行してもよいことが理解され得る。再構築ログに基づいてデータ再構築を実行する方式は、前述の例に記載のものと同じでもよい。

40

【0190】

707. メモリNd2-D1のためのデータ再構築が完了した後に、ストレージノードNd2は、メモリNd2-D1の状態を信頼されたアクセス状態に設定する。さらに、ストレージノードNd2は、メモリ状態変化報告P3を管理ノードNd1に送信してもよく、メモリ状態変化報告P3は、メモリNd2-D1の状態がデータ再構築状態から信頼されたアクセス状態に変化したことを示してもよい。

【0191】

708. 管理ノードNd1は、ストレージノードNd2からメモリ状態変化報告P3を受信し、管理ノードNd1は、キャッシュされた状態情報セットを更新する。具体的には、ストレージノードNd1は、ストレージノードNd1によりキャッシュされた状態情報セットに記録されたメ

50

メモリNd2-D1の状態を、メモリ状態変化報告P3により示されるメモリNd2-D1の状態に更新し、状態情報セットに記録されたバージョン番号を更新する。

【0192】

例えば、monitorが各ストレージノードに配置されるとき、具体的には、ストレージノードNd2内のセカンダリmonitorは、メモリ状態変化報告P3を管理ノードNd1に送信してもよい。対応して、管理ノードNd1内のプライマリmonitorは、ストレージノードNd2からメモリ状態変化報告P3を受信してもよい。

【0193】

例えば、管理ノードNd1により更新されていない状態情報セットが図7-Bに例示される。例えば、管理ノードNd1により更新されている状態情報セットが図7-Cに例示される。明らかに、実際の用途では、状態情報セットはこのような例示の形式に限定されない。

10

【0194】

709.管理ノードNd1は、更新された状態情報セットを他のストレージノード（例えば、Nd2、Nd3及びNd4）に送信する。

【0195】

対応して、ストレージノードNd2、Nd3及びNd4は、ストレージノードNd1から更新された状態情報セットを別々に受信してもよい。ストレージノードNd2、Nd3及びNd4は、これらの現在キャッシュされている状態情報セットを更新するために、管理ノードNd1からの状態情報セットを別々に使用してもよい。このように、ストレージノード（Nd1、Nd2、Nd3及びNd4）によりキャッシュされた状態情報セットは、できるだけ相対的に同期され続けることができる。

20

【0196】

前述の解決策は、オフラインのメモリが再びオンラインになった後のデータ再構築シナリオに主に使用され、前述の解決策は、このような可能なシナリオにおいて可能なデータ再構築機構を提供する。元々オフラインであるメモリが再びオンラインになった後に、メモリのストレージノードは管理ノードに通知する。管理ノードは、管理ノードに記録されたメモリの状態を「データ再構築状態」に更新し、次に、管理ノードは、メモリの状態が「データ再構築状態」に更新されていることを他のストレージノードに通知するために、最新の状態情報セットを他のストレージノードにリリースする。以降、関係するストレージノードは、以前にメモリに書き込まれるのに失敗したストリップを再構築する。再構築が完了した後に、再びオンラインになったメモリが位置するストレージノードは、再びオンラインになったメモリの状態を「信頼されたアクセス状態」に更新し、管理ノードに対して、再びオンラインになったメモリの状態を「信頼されたアクセス状態」に更新するように命令する。以降、管理ノードは、再びオンラインになったメモリの状態が「信頼されたアクセス状態」に更新されていることを残りのストレージノードに通知するために、最新の状態情報セットを他のストレージノードに再びリリースする。

30

【0197】

この実施例では、メモリ状態を記録するために使用される状態情報セットが導入され、したがって、状態情報セットは、分散型ストレージシステムにおけるメモリのアクセス状態を記録及び管理するために使用されてもよい。これは、分散型ストレージシステムにおけるストレージノードがメモリの状態切り替え（「オフライン状態」から「データ再構築状態」へ、且つ、その後の「データ再構築状態」から「信頼されたアクセス状態」への切り替え）を比較的正確に習得するのに役立ち、読み取り／書き込みの試みの失敗の可能性を低減するのに役立ち、システム性能を改善するのに役立つ。

40

【0198】

実施例6

関係する添付図面を参照して、以下に、他のデータアクセス方法を説明するための例を更に使用する。これは、端末が分散型ストレージシステムからデータを読み取る可能なシナリオに主に使用される。データアクセス方法は分散型ストレージシステムに適用される

50

。例えば、分散型ストレージシステムはm個のストレージノードを含み、m個のストレージノードは、ストレージノードNd2、Nd1、Nd3及びNd4を含む。メモリNd2-D1はストレージノードNd2に属し、メモリNd3-D1はストレージノードNd3に属し、メモリNd4-D1はストレージノードNd4に属する。この実施例は、データボディのフォーマットがファイルである例を使用することにより説明する。

【0199】

図8を参照すると、図8は、この出願の実施例によるデータアクセス方法の概略フローチャートである。図8に示すように、この出願の実施例において提供されるデータアクセス方法は以下のステップを含んでもよい。

【0200】

10

801. 端末が分散型ストレージシステムからデータを読み取る必要があるとき、端末は、データ読み取り要求を分散型ストレージシステムに送信する。以下の説明の便宜上のため、この実施例ではデータ読み取り要求はRq1という名で呼ばれる。このステップでは、例えば、分散型ストレージシステムにおけるストレージノードNd1は端末からデータ読み取り要求Rq1を受信する。

【0201】

例えば、データ読み取り要求Rq1は、データData2（データData2は読み取り対象のデータである）が属するファイルのファイル識別子と、データData2の位置情報とを搬送する。データData2の位置情報は、データData2が属するファイル内のデータData2の位置を記述するために使用される。データData2の位置情報は、例えば、データData2のファイルオフセットアドレス（データData2のファイルオフセットアドレスはファイル内のデータData2の開始位置又は終了位置を示す）、データData2の長さ等を含んでもよい。

20

【0202】

802. ストレージノードNd1は、データData2の位置情報に基づいて、データData2が位置するメモリを決定する。

【0203】

例えば、ストレージノードNd1は、データData2が位置するメモリがメモリNd2-D1であると決定する。例えば、Data2が属するファイルは、記憶中に複数のデータストリップ（及びパリティストリップ）に分割され、データストリップのうち1つがメモリNd2-D1に記憶され、データData2はこのストリップの一部分である。

30

【0204】

803. ストレージノードNd1は、ストレージノードNd1によりローカルにキャッシュされた状態情報セットに現在記録されているメモリNd2-D1の状態を決定する。この実施例では、ストレージノードNd1と、管理ノードとは接続状態にあってもよく、或いは接続状態になくてもよい。

【0205】

例えば、ストレージノードNd1により現在キャッシュされている状態情報セットが、メモリNd2-D1が信頼されたアクセス状態にあることを記録するとき、ステップ804が実行され、或いはストレージノードNd1により現在維持されている状態情報セットが、メモリNd2-D1がデータ再構築状態又はオフライン状態にあることを記録するとき、ステップ807が実行される。

40

【0206】

804. メモリNd2-D1が信頼されたアクセス状態にあると決定された場合、ストレージノードNd1は、更新されたデータ読み取り要求Rq1を取得するために、現在ストレージノードNd1にローカルにキャッシュされている状態情報セットのバージョン番号をデータ読み取り要求Rq1に追加してもよく、ストレージノードNd1は、更新されたデータ読み取り要求Rq1を、メモリNd2-D1が属するストレージノードNd2に送信する。

【0207】

すなわち、更新されたデータ読み取り要求Rq1は、現在ストレージノードNd1によりローカルにキャッシュされている状態情報セットのバージョン番号を搬送してもよい。さらに

50

、データ読み取り要求Rq1は、データData2の位置情報、データストリップPd2が属するファイルのファイル識別子等を更に搬送してもよい。

【0208】

ストレージノードNd1は、主に、他のストレージノードがバージョン番号を比較することにより、ストレージノードNd1によりキャッシュされた状態情報セットが有効であるか否かを検証するのに役立てるために、ストレージノードNd1によりキャッシュされた状態情報セットのバージョン番号を、他のストレージノードに送信されるデータ読み取り要求に追加することが理解され得る。

【0209】

805. ストレージノードNd1からデータ読み取り要求Rq1を受信した後に、ストレージノードNd2は、データ読み取り要求Rq1で搬送された状態情報セットのバージョン番号を、ストレージノードNd2により現在維持されている状態情報セットのバージョン番号と比較してもよい。

10

【0210】

データ読み取り要求Rq1で搬送された状態情報セットのバージョン番号が、ストレージノードNd2により現在キャッシュされている状態情報セットのバージョン番号と同じであると比較により決定され、ストレージノードNd2と、管理ノードとが接続状態にある場合（すなわち、ストレージノードNd2内のセカンダリmonitorが通常通り管理ノード内のプライマリmonitorに接続され、この場合、ストレージノードNd2により現在キャッシュされている状態情報セットがプライマリmonitorにより維持される最新バージョンの状態情報セットであるという極めて高い可能性が存在すると考えられてもよい）、ストレージノードNd2は、メモリNd2-D1からデータData2を読み取ってもよい。代替として、データ読み取り要求Rq1で搬送された状態情報セットのバージョン番号が、ストレージノードNd2により現在キャッシュされている状態情報セットのバージョン番号とは異なると比較により決定された場合、ストレージノードNd2は、メモリNd2-D1からデータを読み取ることを拒否してもよい。代替として、ストレージノードNd2と、管理ノードとが接続状態にない場合（すなわち、セカンダリノードNd2内のセカンダリmonitorが通常通り管理ノード内のプライマリmonitorに接続されず、この場合、ストレージノードNd2により現在キャッシュされている状態情報セットがプライマリmonitorにより維持される最新バージョンの状態情報セットである可能性があり、或いはプライマリmonitorにより維持される最新バージョンの状態情報セットでない可能性があると考えられてもよい）、ストレージノードNd2は、メモリNd2-D1からデータを読み取ることを拒否してもよい。明らかに、比較的低い検証レベルのいくつかのシナリオでは、ストレージノードNd2と、管理ノードとが接続状態にあるか否かに拘わらず、データ要求Rq1で搬送された状態情報セットのバージョン番号が、ストレージノードNd2により現在キャッシュされている状態情報セットのバージョン番号と同じであるという条件で、ストレージノードNd2がメモリNd2-D1からデータData2を読み取ることができることも可能である。

20

30

【0211】

ストレージノードNd2がメモリNd2-D1からデータData2を成功して読み取ったとき、ストレージノードNd2は、データ読み取り要求Rq1に回答するために使用されるデータ読み取り応答Rr1をストレージノードNd1に送信し、データ読み取り応答Rr1は読み取られたデータData2を搬送する。ストレージノードNd2がメモリNd2-D1からデータData2を読み取るのに失敗した（例えば、ストレージノードNd2がメモリNd2-D1からデータを読み取ることを拒否する）とき、データ読み取り応答Rr1は読み取り操作の結果を搬送してもよい（読み取り操作の結果は読み取り失敗である）。

40

【0212】

806. ストレージノードNd1は、ストレージノードNd2からデータ読み取り応答Rr1を受信し、ストレージノードNd2がデータData2を成功して読み取った場合、ストレージノードNd1は、データ読み取り応答Rr1からデータData2を取得してもよい。

【0213】

50

ストレージノードNd1は、データ読み取り要求Rq1に応答するために使用されるデータ読み取り応答Rr1を端末に送信し、データ読み取り応答Rr1はデータData2を搬送する。次に、この手順が終了する。

【0214】

807. メモリNd2-D1が信頼されていないアクセス状態にあると決定された場合、ストレージノードNd1は、データData2が属するデータストリップPd1を決定し、ストレージノードNd1は、データストリップPd1が属するストライプを決定し、ストライプは、データストリップPd1とデータストリップPd2とパリティストリップPj1とを含む。ストレージノードNd1は、データストリップPd2及びパリティストリップPj1が位置するメモリを決定する。

【0215】

例えば、データストリップPd2が位置するメモリはメモリNd3-D1として記され、メモリNd3-D1が属するストレージノードはストレージノードNd3として記される。例えば、パリティストリップPj1が位置するメモリはメモリNd4-D1として記され、メモリNd4-D1が属するストレージノードはストレージノードNd4として記される。

【0216】

このステップでは、読み取り対象のデータData2が属するストリップが位置するメモリは信頼されていないアクセス状態にある。したがって、このステップでは、読み取り要求を受信したストレージノードNd1は、ストライプの残りのストリップが位置するメモリを決定し、それにより、以降、検査アルゴリズムを使用することにより、Data2が位置するストリップを回復するために残りのストリップを使用し、Data2を取得する。

【0217】

808. ストレージノードNd1は、データ読み取り要求Rq3を、メモリNd3-D1が属するストレージノードNd3に送信し、ストレージノードNd1は、データ読み取り要求Rq4を、メモリNd4-D1が属するストレージノードNd4に送信する。

【0218】

例えば、データ読み取り要求Rq3は、データストリップPd2のストリップ識別子と、ストレージノードNd1により現在キャッシュされている状態情報セットのバージョン番号とを搬送する。さらに、データ読み取り要求Rq3は、データストリップPd2の長さ（長さは、例えば、110Kb又は他の長さである）と、データストリップPd2のファイルオフセットアドレスと、データストリップPd2が属するストライプのストライプ識別子と、データストリップPd2が属するファイルのファイル識別子とを更に搬送してもよく、データストリップPd2が属するファイルセグメントのセグメント識別子ですら更に搬送してもよい。

【0219】

例えば、データ読み取り要求Rq4は、パリティストリップPj1のストリップ識別子と、現在ストレージノードNd1によりローカルにキャッシュされている状態情報セットのバージョン番号とを搬送する。さらに、データ読み取り要求Rq4は、パリティストリップPj1の長さ、パリティストリップPj1が属するストライプのストライプ識別子等を更に搬送してもよい。

【0220】

809. ストレージノードNd1からデータ読み取り要求Rq3を受信した後に、ストレージノードNd3は、データ読み取り要求Rq3で搬送された状態情報セットのバージョン番号を、ストレージノードNd3により現在キャッシュされている状態情報セットのバージョン番号と比較する。データ読み取り要求Rq3で搬送された状態情報セットのバージョン番号が、ストレージノードNd3により現在キャッシュされている状態情報セットのバージョン番号と同じであると比較により決定され、ストレージノードNd3と、管理ノードとが接続状態にある場合（すなわち、ストレージノードNd3内のセカンダリmonitorが通常通り管理ノード内のプライマリmonitorに接続され、この場合、ストレージノードNd3により現在キャッシュされている状態情報セットが、プライマリmonitorにより維持される最新バージョンの状態情報セットであると考えられてもよく、データ読み取り要求Rq3で搬送された状態情報セットのバージョン番号は、ストレージノードNd3により現在キャッシュされている状態

10

20

30

40

50

情報セットのバージョン番号と同じであり、すなわち、ストレージノードNd1により現在キャッシュされている状態情報セットもまた、プライマリmonitorにより維持される最新バージョンの状態情報セットであり、この場合、ストレージノードNd1により現在キャッシュされている状態情報セットに記録されたメモリNd3-D1の状態（信頼されたアクセス状態）も正確である）、ストレージノードNd3は、メモリNd3-D1からデータストリップPd2を読み取る。代替として、データ読み取り要求Rq3で搬送された状態情報セットのバージョン番号が、ストレージノードNd3により現在維持されている状態情報セットのバージョン番号とは異なると比較により決定された場合、ストレージノードNd3は、メモリNd3-D1からデータストリップPd2を読み取ることを拒否してもよい。

【0221】

10

さらに、ストレージノードNd3は、データ読み取り要求Rq3に応答するために使用されるデータ読み取り応答Rr3をストレージノードNd1に送信する。データ読み取り応答Rr3は、データストリップPd2に対する読み取り操作の結果を搬送する。具体的には、ストレージノードNd3がメモリNd3-D1からデータストリップPd2を成功して読み取ったとき、データ読み取り応答Rr3で搬送される読み取り操作の結果は読み取り成功であり、データ読み取り応答Rr3はデータストリップPd2を搬送する。ストレージノードNd3がメモリNd3-D1からデータストリップPd2を読み取るのに失敗した（例えば、ストレージノードNd3がメモリNd3-D1からデータストリップPd2を読み取ることを拒否する）とき、データ読み取り応答Rr3で搬送される読み取り操作の結果は読み取り失敗である。

【0222】

20

810.ストレージノードNd1からデータ読み取り要求Rq4を受信した後に、ストレージノードNd4は、ストレージノードNd3により実行されるものと同様の操作（例えば、状態情報セットのバージョンを比較し、データ読み取り応答Rr4をストレージノードNd1に送信する）を実行する。このステップについてステップ809に対して参照が行われてもよく、したがって、詳細はここでは説明しない。

【0223】

811.ストレージノードNd1は、ストレージノードNd3からデータ読み取り応答Rr3を受信し、ストレージノードNd1は、ストレージノードNd4からデータ読み取り応答Rr4を受信する。

【0224】

30

さらに、データ読み取り応答が対応する読み取り失敗を示す場合、ストレージノードNd1は、再読み取りを要求するために、対応するデータ読み取り要求を再送信してもよい。例えば、データ読み取り応答Rr3が対応する読み取り失敗を示す場合、ストレージノードNd1は、データストリップPd2を読み取ることを再び要求するために、データ読み取り要求をストレージノードNd3に再送信してもよい。他の例では、データ読み取り応答Rr4が対応する読み取り操作失敗を示す場合、ストレージノードNd1は、パリティストリップPj1を読み取ることを再び要求するために、データ読み取り要求Rq4を再送信してもよく、以下同様である。

【0225】

40

812.ストレージノードNd1がデータストリップPd2及びパリティストリップPj1を取得したとき、ストレージノードNd1は、データストリップPd1を取得してデータストリップPd1からデータData2を取得するために、データストリップPd2及びパリティストリップPj1を使用することにより、検査演算を実行する。ストレージノードNd1は、データ読み取り要求Rq1に応答するために使用されるデータ読み取り応答Rr1を端末に送信し、データ読み取り応答Rr1はデータData2を搬送する。

【0226】

ストレージノードNd1がストレージノードNd2からのデータ読み取り応答Rr1からデータData2を取得したとき、ストレージノードNd1は、データ読み取り要求Rq1に応答するために使用されるデータ読み取り応答Rr1を端末に送信し、データ読み取り応答Rr1はデータData2を搬送する。

50

【0227】

前述の解決策では、メモリ状態を記録するために使用される状態情報セットが分散型ストレージシステムに導入され、したがって、状態情報セットは、分散型ストレージシステムにおけるメモリのアクセス状態を記録及び管理するために使用されてもよいことが習得できる。具体的には、機構は以下のように導入される。データ読み取り要求を受信したストレージノードは、データ読み取り要求で搬送された状態情報セットのバージョン番号と、ストレージノードにより現在キャッシュされている状態情報セットのバージョン番号との間の比較の結果に基づいて、対応するデータ読み取り判断を行う。バージョン番号比較は、使用される状態情報セットの有効性をより良く決定するのに役立つ。検証を用いて信頼されたアクセス状態にあると見出されたメモリについて、要求されたデータはメモリから直接読み取られ、端末にフィードバックできる。このような信頼されたアクセスの場合、関係するストライプの全てのストリップを読み取り、次に、正確性検査を実行する従来の方式における複雑なステップは必要なく、データ読み取り量が低減され、それにより、メモリ負荷を低減するのに役立ち、さらに、システム性能を改善するのに役立つ。

10

【0228】

この実施例は、読み取り対象のデータを読み取るための方法を記載しており、読み取り対象のデータは、ストリップの一部分又は全体のストリップである。ストリップが位置するストレージノードが信頼されている場合、読み取り対象のデータは、要求を送信したデバイス（端末又はホスト）に直接返信される。ストリップが位置するストレージノードが信頼されていない（再構築状態又はオフライン状態にある）場合、検査演算は、読み取り対象のデータが属するストライプの他のストリップ（読み取り対象のデータが属するストリップ以外のストリップ）を使用することにより実行され、それにより、読み取り対象のデータが属するストリップを取得し、ストリップから読み取り対象のデータを取得する。

20

【0229】

ストレージノードNd1は、主に、他のストレージノードがバージョン番号を比較することにより、ストレージノードNd1によりキャッシュされた状態情報セットが有効であるか否かを検証するのに役立てるために、ストレージノードNd1によりキャッシュされた状態情報セットのバージョン番号を、他のストレージノードに送信されるデータ読み取り要求に追加する。ストレージノードNd2が通常通り管理ノードに接続されるとき、この場合、データ要求Rq1で搬送された状態情報セットのバージョン番号が、ストレージノードNd2により現在キャッシュされている状態情報セットのバージョン番号と同じである場合、ストレージノードNd1及びNd2が同じ最新の状態情報セットをキャッシュすると考えられてもよく、ストレージノードNd2はメモリNd2-D1からデータData2を読み取ることができる。代替として、ストレージノードNd2が通常通り管理ノードに接続されていないとき、この場合、ストレージノードNd2により現在キャッシュされている状態情報セットは、管理ノードによりリリースされた最新バージョンの状態情報セットである可能性があるか、或いは管理ノードによりリリースされた最新バージョンの状態情報セットでない可能性があると考えられてもよい。ストレージノードNd2により現在キャッシュされている状態情報セットの有効性を決定することは困難であり、したがって、ストレージノードNd2は、メモリNd2-D1からデータData2を読み取ることが拒否してもよい。明らかに、比較的低い検証レベルのいくつかのシナリオでは、ストレージノードNd2と、管理ノードとが接続状態にあるか否かに拘わらず、データ要求Rq1で搬送された状態情報セットのバージョン番号が、ストレージノードNd2により現在キャッシュされている状態情報セットのバージョン番号と同じであるという条件で（2つのバージョン番号が同じであるとき、ストレージノードNd1によりキャッシュされた状態情報セットが有効であるという比較的高い可能性が存在する）、ストレージノードNd2がメモリNd2-D1からデータData2を読み取ることができることも可能である。状態情報セットの有効性を検証するためのいくつかの機構が導入され、したがって、これは、使用される状態情報セットの有効性をより良く決定するのに役立つ。

30

40

【0230】

実施例7

50

添付図面を参照して、以下に、他のデータアクセス方法を更に説明する。データアクセス方法は分散型ストレージシステムに適用される。例えば、分散型ストレージシステムはm個のストレージノードを含み、m個のストレージノードは、ストレージノードNd1、Nd2、Nd3及びNd4を含む。メモリNd2-D1はストレージノードNd2に属する。この実施例は、データボディがファイルである例を使用することにより説明する。

【0231】

図9を参照すると、図9は、この出願の他の実施例による他のデータアクセス方法の概略フローチャートである。図9に示すように、この出願の他の実施例において提供される他のデータアクセス方法は以下のステップを含んでもよい。

【0232】

10

901. 端末が分散型ストレージシステムからデータを読み取る必要があるとき、端末は、データ読み取り要求を分散型ストレージシステムに送信する。以下の説明の便宜上のため、この実施例ではデータ読み取り要求はRq1という名で呼ばれる。このステップでは、例えば、分散型ストレージシステムにおけるストレージノードNd2は端末からデータ読み取り要求Rq1を受信する。

【0233】

例えば、ステップ901においてデータ読み取り要求Rq1で搬送される関係する情報は、ステップ801においてデータ読み取り要求Rq1で搬送される関係する情報と同じである。

【0234】

902. ストレージノードNd2は、データData2の位置情報に基づいて、データData2が位置するメモリを決定する。位置情報に従って、データが位置するメモリを決定する解決策は、他の実施例において既に説明しており、したがって、ここでは再び説明しない。

20

【0235】

例えば、ストレージノードNd2は、データData2が位置するメモリがメモリNd2-D1であると決定する。

【0236】

903. ストレージノードNd2は、ストレージノードNd2によりローカルにキャッシュされた状態情報セットに記録されたメモリNd2-D1の状態を決定する。

【0237】

例えば、ストレージノードNd2により現在キャッシュされている状態情報セットは、メモリNd2-D1が信頼されたアクセス状態にあることを記録し、ストレージノードNd2と、管理ノードとが接続状態にある（すなわち、ストレージノードNd2内のセカンダリmonitorが通常通り管理ノード内のプライマリmonitorに接続され、この場合、ストレージノードNd2により現在キャッシュされている状態情報セットがプライマリmonitorによりリリースされた最新バージョンの状態情報セットであると考えられてもよい）。

30

【0238】

904. メモリNd2-D1が信頼されたアクセス状態にあると決定された場合、ストレージノードNd2は、メモリNd2-D1からデータData2を読み取ってもよく、ストレージノードNd2は、データ読み取り要求Rq1に応答するために使用されるデータ読み取り応答Rr1を端末に送信し、データ読み取り応答Rr1はデータData2を搬送する。

40

【0239】

代替として、メモリNd2-D1が信頼されていないアクセス状態にあると決定されたとき、図8に例示されるシナリオにおける実現方式に参照が行われてもよく、データData2が属するストライプの他のストリップを収集し、他のストリップを使用することによりData2を取得する。具体的な実現方式についての詳細はここでは再び説明しない。

【0240】

前述の解決策では、メモリ状態を記録するために使用される状態情報セットが分散型ストレージシステムに導入され、したがって、状態情報セットは、分散型ストレージシステムにおけるメモリのアクセス状態を記録及び管理するために使用されてもよいことが習得できる。具体的には、機構は以下のように導入される。データ読み取り要求を受信したス

50

ストレージノードは、ストレージノードによりキャッシュされた状態情報セットに記録された関係するメモリの状態と、ストレージノードと状態情報セットをリリースする管理ノードとの間の接続の状態とに基づいて、対応するデータ読み取り判断を行う。これは、使用される状態情報セットの有効性をより良く決定するのに役立つ。検証を用いて信頼されたアクセス状態にあると見出されたメモリについて、要求されたデータはメモリから直接読み取られ、端末にフィードバックできる。このような信頼されたアクセスの場合、関係するストライプの全てのストリップを読み取り、次に、正確性検査を実行する従来の方式における複雑なステップは必要なく、データ読み取り量が低減され、それにより、メモリ負荷を低減するのに役立ち、さらに、システム性能を改善するのに役立つ。

【0241】

10

この解決策では、読み取り対象のデータ（データData2）が、端末／ホストからデータ読み取り要求を受信したストレージノード（Nd2）のメモリ（ND2-D1）に位置し、ストレージノード（Nd2）と、管理ノードとが接続状態にある場合、ストレージノード（Nd2）によりローカルにキャッシュされた状態情報セットに記録されたメモリ状態が比較的信頼性が高いと考えられてもよい。状態情報セットが、メモリ（Nd2-D1）が信頼されたアクセス状態にあることを記録するとき、読み取り対象のデータ（データData2）はメモリから直接読み取られ、次に、端末／ホストに返信できる。

【0242】

実施例8

添付図面を参照して、以下に、他のデータアクセス方法を更に説明する。データアクセス方法は分散型ストレージシステムに適用される。例えば、分散型ストレージシステムはm個のストレージノードを含み、m個のストレージノードは、ストレージノードNd2、Nd1、Nd3及びNd4を含む。メモリNd2-D1はストレージノードNd2に属し、メモリNd3-D1はストレージノードNd3に属し、メモリNd4-D1はストレージノードNd4に属する。この実施例は、データボディがファイルである例を使用することにより説明する。

20

【0243】

図10を参照すると、図10は、この出願の他の実施例による他のデータアクセス方法の概略フローチャートである。図10に示すように、この出願の他の実施例において提供される他のデータアクセス方法は以下のステップを含んでもよい。

【0244】

30

1001及び1002.ステップ1001及び1002はステップ801及び802と同じである。したがって、関係する説明について、対応するステップ801及び802を参照し、詳細はここでは再び説明しない。

【0245】

1003.ストレージノードNd1と、管理ノードとが接続状態にあるとき、ストレージノードNd1は、ストレージノードNd1によりローカルにキャッシュされた状態情報セットに記録されたメモリNd2-D1の状態を決定する。

【0246】

例えば、ストレージノードNd1により現在キャッシュされている状態情報セットが、メモリNd2-D1が信頼されたアクセス状態にあることを記録するとき、ステップ1004が実行され、或いはストレージノードNd1により現在維持されている状態情報セットが、メモリNd2-D1がデータ再構築状態又はオフライン状態にあることを記録するとき、ステップ1007が実行される。

40

【0247】

ストレージノードNd1と、管理ノードとが接続状態にないとき、ストレージノードNd1は、データ読み取り要求Rq1を管理ノードに転送してもよく、或いはデータ読み取り要求Rq1を、管理ノードに接続されている状態にある他のストレージノードに転送してもよく、Rq1を受信した他のストレージノードは、他のストレージノードによりローカルにキャッシュされた状態情報セットに記録されたメモリNd2-D1の状態を決定し、メモリNd2-D1の状態に従って以降の関係する操作を実行する。以降の操作についての詳細について、この実施

50

例におけるステップ1003及び1011を参照する。この実施例では、ストレージノードNd1と、管理ノードとが接続状態にある場合が、説明のための例として使用される。

【0248】

1004.メモリNd2-D1が信頼されたアクセス状態にあると決定された場合、ストレージノードNd1は、データ読み取り要求Rq1を、メモリNd2-D1が属するストレージノードNd2に送信する。

【0249】

1005.ストレージノードNd2がストレージノードNd1からデータ読み取り要求Rq1を受信した後に、ストレージノードNd2は、メモリNd2-D1からデータData2を読み取ってもよい。ストレージノードNd2は、データ読み取り要求Rq1にตอบสนองするために使用されるデータ読み取り応答Rr1をストレージノードNd1に更に送信し、データ読み取り応答Rr1は読み取られたデータData2を搬送する。

10

【0250】

1006.ストレージノードNd1は、ストレージノードNd2からデータ読み取り応答Rr1を受信し、ストレージノードNd2がデータData2を成功して読み取った場合、ストレージノードNd1は、データ読み取り応答Rr1からデータData2を取得してもよい。このステップが実行された後に、手順が終了し、以下のステップ1007及び他のステップはもはや実行されない。

【0251】

1007.メモリNd2-D1が信頼されていないアクセス状態にあると決定された場合、ストレージノードNd1は、データData1が属するデータストリップPd1を決定し、ストレージノードNd1は、データストリップPd1が属するストライプを決定し、ストライプは、データストリップPd1とデータストリップPd2とパリティストリップPj1とを含む。ストレージノードNd1は、データストリップPd2及びパリティストリップPj1が位置するメモリを決定する。

20

【0252】

例えば、データストリップPd2が位置するメモリはメモリNd3-D1として記され、メモリNd3-D1が属するストレージノードはストレージノードNd3として記される。例えば、パリティストリップPj1が位置するメモリはメモリNd4-D1として記され、メモリNd4-D1が属するストレージノードはストレージノードNd4として記される。ストレージノードNd1は、ストレージノードNd1によりローカルにキャッシュされた状態情報セットに記録されたメモリNd3-D1及びNd4-D1の状態を決定する。

30

【0253】

ストレージノードNd1により現在キャッシュされている状態情報セットが、メモリNd3-D1及びNd4-D1が信頼されたアクセス状態にあることを記録することが、以下の説明のための例として使用される。ストレージノードNd1により現在維持されている状態情報セットが、メモリNd3-D1及びNd4-D1が信頼されていないアクセス状態にあることを記録するとき、データ読み取り失敗が直接フィードバックされてもよく、或いは現在キャッシュされている状態情報セットに記録されたメモリNd1-D1、Nd3-D1及びNd4-D1の状態が指定の持続時間の後に再び検査されることを示す。

【0254】

1008.ストレージノードNd1は、データ読み取り要求Rq3を、メモリNd3-D1が属するストレージノードNd3に送信し、ストレージノードNd1は、データ読み取り要求Rq4を、メモリNd4-D1が属するストレージノードNd4に送信する。

40

【0255】

例えば、データ読み取り要求Rq3は、データストリップPd2のストリップ識別子を搬送する。データ読み取り要求Rq3は、データストリップPd2の長さ（例えば、110Kb）と、データストリップPd2のファイルオフセットアドレスと、データストリップPd2が属するストライプのストライプ識別子と、データストリップPd2が属するファイルのファイル識別子とを更に搬送してもよく、データストリップPd2が属するファイルセグメントのセグメント識別子ですら更に搬送してもよい。

【0256】

50

例えば、データ読み取り要求Rq4は、パリティストリップPj1のストリップ識別子を搬送する。さらに、データ読み取り要求Rq4は、パリティストリップPj1の長さ、パリティストリップPj1が属するストライプのストライプ識別子等を更に搬送してもよい。

【0257】

1009. ストレージノードNd3がストレージノードNd1からデータ読み取り要求Rq3を受信した後に、ストレージノードNd3は、メモリNd3-D1からデータストリップPd2を読み取る。

【0258】

さらに、ストレージノードNd3は、データ読み取り要求Rq3に応答するために使用されるデータ読み取り応答Rr3をストレージノードNd1に送信する。データ読み取り応答Rr3は、データストリップPd2に対する読み取り操作の結果を搬送する。具体的には、ストレージノードNd3がメモリNd3-D1からデータストリップPd2を成功して読み取ったとき、データ読み取り応答Rr3で搬送される読み取り操作の結果は読み取り成功であり、データ読み取り応答Rr3はデータストリップPd2を搬送する。ストレージノードNd3がメモリNd3-D1からデータストリップPd2を読み取るのに失敗したとき、データ読み取り応答Rr3で搬送される読み取り操作の結果は読み取り失敗である。

【0259】

1010. ストレージノードNd4がストレージノードNd1からデータ読み取り要求Rq4を受信した後に、ストレージノードNd4は、メモリNd4-D1からパリティストリップPj1を読み取る。

【0260】

さらに、ストレージノードNd4は、データ読み取り要求Rq4に応答するために使用されるデータ読み取り応答Rr4をストレージノードNd1に送信する。データ読み取り応答Rr4は、パリティストリップPj1に対する読み取り操作の結果を搬送する。具体的には、ストレージノードNd4がメモリNd4-D1からパリティストリップPj1を成功して読み取ったとき、データ読み取り応答Rr4で搬送される読み取り操作の結果は読み取り成功であり、データ読み取り応答Rr4はパリティストリップPj1を搬送する。ストレージノードNd4がメモリNd4-D1からパリティストリップPj1を読み取るのに失敗したとき、データ読み取り応答Rr4で搬送される読み取り操作の結果は読み取り失敗である。

【0261】

1011及び1012. ステップ1011及び1012はステップ811及び812と同じである。

【0262】

この実施例では、ストレージノード(Nd1)は読み取り要求を受信するが、読み取られる必要があるデータ(データData1)は、他のストレージノード(Nd2)のメモリ(Nd2-D1)に位置する。Nd2と、管理ノードとが接続状態にあり、Nd2によりローカルにキャッシュされた状態情報セットに記録されたメモリNd2-D1の状態が、信頼されたアクセス状態である場合、読み取られる必要があるデータは、Nd2-D1から直接読み取られることができる。Nd2と、管理ノードとが接続状態にない場合、他のメモリに位置するストリップ(データストリップPd2及びパリティストリップPj1)が位置するメモリ(Nd3-D1及びNd4-D1)が決定され、ストライプは、読み取られる必要があるデータが位置するストリップにある。他のメモリに位置するストリップ(データストリップPd2及びパリティストリップPj1)が位置するメモリ(Nd3-D1及びNd4-D1)が、信頼されたアクセス状態にあり、メモリNd3-D1及びNd4-D1が属するストレージノードと、管理ノードとが接続状態にある場合、他のメモリに位置するストライプが読み取られ、読み取られる必要があるデータ(データData1)が位置するストリップ(Pd2)は、検査アルゴリズムを使用することにより取得され、読み取られる必要があるデータ(データData1)はPd1から取得され、端末/ホストに返信される。

【0263】

前述の解決策では、メモリ状態を記録するために使用される状態情報セットが分散型ストレージシステムに導入され、したがって、状態情報セットは、分散型ストレージシステ

10

20

30

40

50

ムにおけるメモリのアクセス状態を記録及び管理するために使用されてもよいことが習得できる。具体的には、機構は以下のように導入される。データ読み取り要求を受信して転送するストレージノードは、ストレージノードによりキャッシュされた、リリースされた状態情報セットに記録された関係するメモリの状態と、ストレージノードと状態情報セットをリリースする管理ノードとの間の接続の状態とに基づいて、対応するデータ読み取り要求転送判断を行う。これは、使用される状態情報セットの有効性をより良く決定するのに役立つ。信頼されたアクセス状態のメモリについて、関係するストレージノードは、メモリから要求されたデータを読み取り、読み取られたデータを端末にフィードバックするように直接トリガーできる。信頼されたアクセスのこのような場合、関係するストライプの全てのストリップを読み取り、次に、正確性検査を実行する従来の方式における複雑なステップは必要なく、データ読み取り量が低減され、それにより、メモリ負荷を低減するのに役立ち、さらに、システム性能を改善するのに役立つ。

10

【0264】

以下に、前述の解決策を実現するための関係する装置を更に提供する。

【0265】

図11を参照すると、この出願の実施例は、分散型ストレージシステム1100を提供する。分散型ストレージシステムは m 個のストレージノードを含み、各ストレージノードは少なくとも1つのメモリを含み、各メモリは不揮発性記憶媒体を含み、 m は1より大きい整数である。 m 個のストレージノード内の第1のストレージノード1110は、端末から第1のデータ読み取り要求を受信し、第1のデータ読み取り要求は第1の位置情報を搬送し、第1の位置情報は、第1のデータが属するデータボディ内の第1のデータの位置を記述するために使用され、第1の位置情報に基づいて、第1のデータが位置する第1のメモリを決定し、

20

第1のメモリが第1のストレージノードに属する場合、状態情報セットが、第1のメモリが信頼されたアクセス状態にあることを現在記録しているとき、第1のメモリから第1のデータを読み取り、第1のデータ読み取り要求に応答するために使用される第1のデータ読み取り応答を端末に送信し、第1のデータ読み取り応答は読み取られた第1のデータを搬送する、ように構成される。第1のデータは、例えば、第1のデータが属するストリップ内のデータの一部又は全部である。

【0266】

30

この出願のいくつかの可能な実現方式では、分散型ストレージシステムは、状態情報セットをリリースするように構成された管理ノードを更に含む。

【0267】

第1のストレージノードは、第1のメモリが m 個のストレージノード内の第2のストレージノードに属し、第1のメモリが信頼されたアクセス状態にあり、第1のストレージノードと、状態情報セットをリリースするように構成された管理ノードとが接続状態にある場合、第1のデータ読み取り要求を第2のストレージノードに転送するように更に構成される。

【0268】

第2のストレージノード1120は、第1のストレージノードから第1のデータ読み取り要求を受信した後に、第1のメモリから第1のデータを読み取り、第1のデータ読み取り要求に応答するために使用される第1のデータ読み取り応答を第1のストレージノード又は端末に送信するように構成され、第1のデータ読み取り応答は読み取られた第1のデータを搬送する。

40

【0269】

この出願のいくつかの可能な実現方式では、第1のストレージノード1110は、第1のメモリが m 個のストレージノード内の第2のストレージノードに属し、第1のメモリが信頼されたアクセス状態にある場合、第1のストレージノードにより現在キャッシュされている状態情報セットの識別子を第1のデータ読み取り要求に追加し、状態情報セットの識別子が追加された第1のデータ読み取り要求を第2のストレージノードに送信するように更

50

に構成される。

【0270】

第2のストレージノード1120は、第1のストレージノードから第1のデータ読み取り要求を受信した後に、第1のデータ読み取り要求で搬送された状態情報セットの識別子を、第2のストレージノードにより現在キャッシュされている状態情報セットの識別子と比較し、第1のデータ読み取り要求で搬送された状態情報セットの識別子が、第2のストレージノードにより現在キャッシュされている状態情報セットの識別子と同じであると比較により決定された場合、第1のメモリから第1のデータを読み取り、第1のデータ読み取り要求に応答するために使用される第1のデータ読み取り応答を第1のストレージノード又は端末に送信するように構成され、第1のデータ読み取り応答は読み取られた第1のデータを搬送する。

10

【0271】

この出願のいくつかの可能な実現方式では、第1のストレージノード1110は、第1のメモリが m 個のストレージノード内の第2のストレージノードに属し、第1のメモリが信頼されたアクセス状態にある場合、第1のストレージノードにより現在キャッシュされている状態情報セットの識別子と、状態情報セットをリリースするように構成された管理ノードのノード識別子とを第1のデータ読み取り要求に追加し、状態情報セットの識別子及び管理ノードのノード識別子が追加された第1のデータ読み取り要求を第2のストレージノードに送信するように更に構成される。

20

【0272】

対応して、第2のストレージノード1120は、第1のストレージノードから第1のデータ読み取り要求を受信した後に、第1のデータ読み取り要求で搬送された状態情報セットの識別子を、第2のストレージノードにより現在キャッシュされている状態情報セットの識別子と比較し、第1のデータ読み取り要求で搬送された管理ノードのノード識別子を、第2のストレージノードにより現在キャッシュされている管理ノードのノード識別子と比較し、第1のデータ読み取り要求で搬送された状態情報セットの識別子が、第2のストレージノードにより現在キャッシュされている状態情報セットの識別子と同じであると比較により決定され、第1のデータ読み取り要求で搬送された管理ノードのノード識別子が、第2のノードにより現在キャッシュされている管理ノードのノード識別子と同じであると比較により決定された場合、第1のメモリから第1のデータを読み取り、第1のデータ読み取り要求に応答するために使用される第1のデータ読み取り応答を第1のストレージノード又は端末に送信するように構成され、第1のデータ読み取り応答は読み取られた第1のデータを搬送する。

30

【0273】

この出願のいくつかの可能な実現方式では、第1のストレージノード1110は、第1のメモリが信頼されていないアクセス状態にあると決定され、第1のデータが属するストライプのストリップが位置する N 個のメモリ内の x 個のメモリが第1のストレージノードに属し、 N 個のメモリ内の $N-x$ 個のメモリが第1のストレージノードとは異なる y 個のストレージノードに属する場合、ストライプのストライプ識別子と、状態情報セットの識別子とを搬送するデータ読み取り要求を y 個のストレージノードに送信するように更に構成される。

40

【0274】

y 個のストレージノードのそれぞれは、第1のストレージノードからデータ読み取り要求を受信した後に、データ読み取り要求で搬送された状態情報セットの識別子を、ストレージノードにより現在キャッシュされている状態情報セットの識別子と比較し、データ読み取り要求で搬送された状態情報セットの識別子が、ストレージノードにより現在キャッシュされている状態情報セットの識別子と同じであると比較により決定された場合、ストレージノードに含まれる対応するメモリからストライプの対応するストリップを読み取り、ストライプの読み取られた対応するストリップを第1のストレージノードに送信するように構成される。

【0275】

50

第1のストレージノードは、 y 個のストレージノード及び第1のストレージノードから収集されたストライプのストリップに基づいて、第1のデータを取得するための検査演算を実行し、第1のデータ読み取り要求に応答するために使用される第1のデータ読み取り応答を端末に送信するように更に構成され、第1のデータ読み取り応答は取得された第1のデータを搬送する。

【0276】

この出願のいくつかの可能な実現方式では、例えば、 m 個のストレージノードのうち1つは管理ノードである。第1のストレージノードは、第1のメモリの状態が第1の状態から第2の状態に変化したとき、メモリ状態変化報告を管理ノードに送信するように更に構成され、メモリ状態変化報告は、第1のメモリが第2の状態にあることを示し、第1の状態は第2の状態とは異なり、第1の状態及び第2の状態は、以下の状態、すなわち、オフライン状態、データ再構築状態又は信頼されたアクセス状態のうちいずれか1つをそれぞれ含む。

10

【0277】

管理ノードは、第1のストレージノードからメモリ状態変化報告を受信した後に、管理ノードによりキャッシュされた状態情報セットに記録された第1のメモリの状態を第2の状態に更新し、管理ノードによりキャッシュされた状態情報セットのバージョン番号を更新するように構成され、管理ノードは、更新された状態情報セットを m 個のストレージノード内の管理ノード以外のストレージノードに送信する。

20

【0278】

対応して、第1のストレージノードは、第1のストレージノードにより現在キャッシュされている状態情報セットを更新するために、管理ノードからの状態情報セットを使用するように更に構成される。対応して、他のストレージノードもまた、他のストレージノードにより現在キャッシュされている状態情報セットを更新するために、管理ノードからの状態情報セットを使用する。

【0279】

この出願のいくつかの可能な実現方式では、第1のストレージノードは、端末から第1のデータ書き込み要求を受信するように更に構成され、第1のデータ書き込み要求は第2のデータ及び第2の位置情報を搬送し、第2の位置情報は、第2のデータが属するデータボディ内の第2のデータの位置を記述するために使用される。

30

【0280】

第1のストレージノードは、第2の位置情報に基づいて、第2のデータを書き込むことに関する W 個のメモリを決定し、第2のデータを W - T 個のデータストリップにセグメント化し、 W - T 個のデータストリップを使用することによる計算を用いて T 個のパリティストリップを取得するように更に構成され、 T 個のパリティストリップ及び W - T 個のデータストリップは、 W 個のストリップを含むストライプを形成し、 W 個のメモリは W 個のストリップと1対1の対応関係にあり、 T 及び W は正の整数であり、 T は W 未満である。

【0281】

第1のストレージノードは、第1のストレージノードによりキャッシュされた状態情報セットが、 W 個のメモリ内の $W1$ 個のメモリが非オフライン状態にあり、 W 個のメモリ内の $W2$ 個のメモリがオフライン状態にあることを現在記録しているとき、ストライプのストリップ及び状態情報セットの識別子を搬送するデータ書き込み要求を、 $W1$ 個のメモリが属する $y2$ 個のストレージノードに送信するように更に構成される。

40

【0282】

対応して、 $y2$ 個のストレージノードのそれぞれは、第1のストレージノードからデータ書き込み要求を受信した後に、データ書き込み要求で搬送された状態情報セットの識別子を、ストレージノードにより現在キャッシュされている状態情報セットの識別子と比較し、データ書き込み要求で搬送された状態情報セットの識別子が、ストレージノードにより現在キャッシュされている状態情報セットの識別子と同じであると比較により決定された場合、ストライプの対応するストリップをストレージノードに含まれる対応するメモリに

50

書き込むか、或いはデータ書き込み要求で搬送された状態情報セットの識別子が、ストレージノードにより現在キャッシュされている状態情報セットの識別子とは異なると比較により決定された場合、ストライプの対応するストリップをストレージノードに含まれる対応するメモリに書き込むことを拒否するように構成される。

【0283】

この出願のいくつかの可能な実現方式では、第1のストレージノードは、端末から第2のデータ書き込み要求を受信するように更に構成され、第2のデータ書き込み要求は第3のデータ及び第3の位置情報を搬送し、第3の位置情報は、第3のデータが属するデータボディ内の第3のデータの位置を記述するために使用される。

【0284】

第1のストレージノードは、第3の位置情報に基づいて、第3のデータを書き込むことに関するW個のメモリを決定した後に、第3のデータをW-T個のデータストリップにセグメント化し、W-T個のデータストリップを使用することによる計算を用いてT個のパリティストリップを取得するように更に構成され、T個のパリティストリップ及びW-T個のデータストリップは、W個のストリップを含むストライプを形成し、W個のメモリはW個のストリップと1対1の対応関係にあり、T及びWは正の整数であり、TはW未満である。

【0285】

第1のストレージノードは、第1のストレージノードによりキャッシュされた状態情報セットに現在記録されているW個のメモリの状態を決定し、W個のメモリ内のW1個のメモリが非オフライン状態にある（且つ、例えば、W個のメモリ内のW2個のメモリがオフライン状態にある）と決定され、第1のストレージノードと、状態情報セットをリリースするように構成された管理ノードとが接続状態にある場合、ストライプのストリップを搬送するデータ書き込み要求を、W1個のメモリが属するy2個のストレージノードに送信するように更に構成される。

【0286】

対応して、y2個のストレージノードのそれぞれは、第1のストレージノードからデータ書き込み要求を受信した後に、ストライプの対応するストリップをストレージノードに含まれる対応するメモリに書き込むように構成される。

【0287】

この出願のいくつかの可能な実現方式では、第1のストレージノードは、端末から第3のデータ書き込み要求を受信するように更に構成され、第3のデータ書き込み要求は第4のデータ及び第4の位置情報を搬送し、第4の位置情報は、第4のデータが属するデータボディ内の第4のデータの位置を記述するために使用される。

【0288】

第1のストレージノードは、第4の位置情報に基づいて、第4のデータを書き込むことに関するW個のメモリを決定し、第4のデータをW-T個のデータストリップにセグメント化し、W-T個のデータストリップを使用することによる計算を用いてT個のパリティストリップを取得するように更に構成され、T個のパリティストリップ及びW-T個のデータストリップは、W個のストリップを含むストライプを形成し、W個のメモリはW個のストリップと1対1の対応関係にあり、T及びWは正の整数であり、TはW未満である。

【0289】

第1のストレージノードは、第1のストレージノードによりキャッシュされた状態情報セットが、W個のメモリ内のW1個のメモリが非オフライン状態にある（且つ、例えば、W個のメモリ内のW2個のメモリがオフライン状態にある）ことを現在記録しており、第1のストレージノードと、状態情報セットをリリースするように構成された管理ノードとが接続状態にあるとき、ストライプのストリップと、状態情報セットをリリースするように構成された管理ノードのノード識別子とを搬送するデータ書き込み要求を、W1個のメモリが属するy2個のストレージノードに送信するように更に構成される。

【0290】

対応して、y2個のストレージノードのそれぞれは、第1のストレージノードからデータ

10

20

30

40

50

書き込み要求を受信した後に、データ書き込み要求で搬送された管理ノードのノード識別子を、ストレージノードにより現在キャッシュされている管理ノードのノード識別子と比較し、データ書き込み要求で搬送された管理ノードのノード識別子が、ストレージノードにより現在キャッシュされている管理ノードのノード識別子と同じであると比較により決定された場合、ストライプの対応するストリップをストレージノードに含まれる対応するメモリに書き込むか、或いはデータ書き込み要求で搬送された管理ノードのノード識別子が、ストレージノードにより現在キャッシュされている管理ノードのノード識別子とは異なると比較により決定された場合、ストライプの対応するストリップをストレージノードに含まれる対応するメモリに書き込むことを拒否するように構成される。

【0291】

10

この出願のいくつかの可能な実現方式では、第1のストレージノードは、第1の再構築ログを生成するように更に構成され、第1の再構築ログは、W2個のメモリ内の第2のメモリのメモリ識別子を記録し、第1の再構築ログは、W個のストリップ内の第2のメモリに対応する第1のストリップのストリップ識別子を更に記録し、第1の再構築ログは、ストライプのストライプ識別子を更に記録し、第2のメモリはW2個のメモリ内のいずれかのメモリである。

【0292】

対応して、例えば、W2個のメモリ内の第2のメモリが再びオンラインになり、第2のメモリが属する第2のストレージノードが第2のメモリのオフライン期間中に生成された第1の再構築ログを収集した後に、第2のストレージノードは、第1の再構築ログに記録された、第2のメモリに書き込まれる必要がある第1のストリップの識別子を取得し、ストライプに含まれる第1のストリップ以外のW-T個のストリップが位置するW-T個のメモリを決定し、W-T個のメモリからW-T個のストリップを読み取り、第2のメモリに書き込まれる必要がある第1のストリップを再構築するためにW-T個のストリップを使用することにより、検査演算を実行し、再構築された第1のストリップを第2のメモリに書き込む。

20

【0293】

この実施例における分散型ストレージシステム1100の機能は、前述の方法の実施例における解決策に基づいて実現されてもよいことが理解され得る。記載されていないいくつかの部分について、前述の実施例を参照する。

【0294】

30

図12を参照すると、この出願の実施例は、ストレージノード1200を更に提供する。ストレージノードは、分散型ストレージシステムに含まれるm個のストレージノードのうち1つであり、各ストレージノードは少なくとも1つのメモリを含み、各メモリは不揮発性記憶媒体を含み、mは1より大きい整数である。ストレージノードは、通信ユニットと処理制御ユニットとを含む。

【0295】

通信ユニット1210は、端末から第1のデータ読み取り要求を受信するように構成され、第1のデータ読み取り要求は第1の位置情報を搬送し、第1の位置情報は、第1のデータが属するデータボディ内の第1のデータの位置を記述するために使用される。

【0296】

40

処理制御ユニット1220は、第1の位置情報に基づいて、第1のデータが位置するメモリが第1のメモリであると決定され、第1のメモリがストレージノードに属する場合、状態情報セットが、第1のメモリが信頼されたアクセス状態にあることを現在記録しているとき、第1のメモリから第1のデータを読み取るように構成される。

【0297】

通信ユニット1210は、第1のデータ読み取り要求に応答するために使用される第1のデータ読み取り応答を端末に送信するように更に構成され、第1のデータ読み取り応答は読み取られた第1のデータを搬送する。

【0298】

この出願のいくつかの可能な実現方式では、通信ユニット1210は、第1のメモリがm個

50

のストレージノード内の第2のストレージノードに属し、第1のメモリが信頼されたアクセス状態にあり、ストレージノードと、状態情報セットをリリースするように構成された管理ノードとが接続状態にある場合、第1のデータ読み取り要求を第2のストレージノードに転送するように更に構成される。

【0299】

第1のデータ読み取り要求は、第2のストレージノードに対して、第1のストレージノードから第1のデータ読み取り要求を受信した後に、第1のメモリから第1のデータを読み取り、第1のデータ読み取り要求に応答するために使用される第1のデータ読み取り応答をストレージノード又は端末に送信するようにトリガーするために使用され、第1のデータ読み取り応答は読み取られた第1のデータを搬送する。

10

【0300】

この出願のいくつかの可能な実現方式では、通信ユニットは、第1のメモリがm個のストレージノード内の第2のストレージノードに属し、第1のメモリが信頼されたアクセス状態にある場合、第1のストレージノードにより現在キャッシュされている状態情報セットの識別子を第1のデータ読み取り要求に追加し、状態情報セットの識別子が追加された第1のデータ読み取り要求を第2のストレージノードに送信するように更に構成される。

【0301】

第1のデータ読み取り要求は、第2のストレージノードに対して、ストレージノードから第1のデータ読み取り要求を受信した後に、第1のデータ読み取り要求で搬送された状態情報セットの識別子を、第2のストレージノードにより現在キャッシュされている状態情報セットの識別子と比較し、第1のデータ読み取り要求で搬送された状態情報セットの識別子が、第2のストレージノードにより現在キャッシュされている状態情報セットの識別子と同じであると比較により決定された場合、第1のメモリから第1のデータを読み取り、第1のデータ読み取り要求に応答するために使用される第1のデータ読み取り応答をストレージノード又は端末に送信するようにトリガーするために使用され、第1のデータ読み取り応答は読み取られた第1のデータを搬送する。

20

【0302】

この出願のいくつかの可能な実現方式では、m個のストレージノードのうち1つは管理ノードである。

【0303】

通信ユニットは、第1のメモリの状態が第1の状態から第2の状態に変化したとき、メモリ状態変化報告を管理ノードに送信するように更に構成され、メモリ状態変化報告は、第1のメモリが第2の状態にあることを示し、第1の状態は第2の状態とは異なり、第1の状態及び第2の状態は、以下の状態、すなわち、オフライン状態、データ再構築状態又信頼されたアクセス状態のうちいずれか1つをそれぞれ含む。

30

【0304】

メモリ状態変化報告は、管理ノードに対して、第1のストレージノードからメモリ状態変化報告を受信した後に、管理ノードによりキャッシュされた状態情報セットに記録された第1のメモリの状態を第2の状態に更新し、管理ノードによりキャッシュされた状態情報セットのバージョン番号を更新し、更新された状態情報セットをm個のストレージノード内の管理ノード以外のストレージノードに送信するようにトリガーするために使用される。

40

【0305】

処理制御ユニットは、ストレージノードにより現在キャッシュされている状態情報セットを更新するために、管理ノードからの状態情報セットを使用するように構成される。

【0306】

この実施例におけるストレージノード1200の機能は、前述の方法の実施例における解決策に基づいて具体的に実現されてもよいことが理解され得る。記載されていないいくつかの部分について、前述の実施例を参照する。

【0307】

50

図13を参照すると、この出願の実施例はストレージノード1300を提供する。ストレージノードは、分散型ストレージシステムに含まれる m 個のストレージノードのうち1つであり、各ストレージノードは少なくとも1つのメモリを含み、各メモリは不揮発性記憶媒体を含み、 m は1より大きい整数である。

【0308】

ストレージノードは、互いに結合されたプロセッサ1310及び通信インタフェース1320を含む。プロセッサは、前述の方法の実施例のそれぞれにおいて第1のストレージノード又は他のストレージノードにより実行される方法のステップの一部又は全部を実行するように構成される。

【0309】

メモリ1330は、命令及びデータを記憶するように構成され、プロセッサ1310は、命令を実行するように構成され、通信インタフェース1320は、プロセッサ1310の制御で他のデバイスと通信するように構成される。命令を実行するとき、プロセッサ1310は、命令に従って、この出願の前述の実施例におけるいずれかの方法においていずれかのストレージノードにより実行されるステップの一部又は全部を実行できる。

【0310】

プロセッサ1310はまた、中央処理装置（英文：Central Processing Unit，略称：CPU）とも呼ばれる。具体的な用途では、ストレージノードのコンポーネントは、例えば、バスシステムを使用することにより一緒に結合される。データバスに加えて、バスシステムは、電力バス、制御バス、状態信号バス等を含んでもよい。しかし、説明の明確性のため、様々なバスは図面においてバスシステム1340として記される。この出願の前述の実施例に開示された方法は、プロセッサ1310に適用されてもよく、或いはプロセッサ1310により実現されてもよい。プロセッサ1310は集積回路チップでもよく、信号処理能力を有する。実現プロセスにおいて、前述の方法のステップは、プロセッサ1310内のハードウェアの集積論理回路、又はソフトウェアの形式の命令を使用することにより完成されてもよい。プロセッサ1310は、汎用プロセッサ、デジタルシグナルプロセッサ、特定用途向け集積回路、フィールドプログラマブルゲートアレイ若しくは他のプログラム可能論理コンポーネント、ディスクリートゲート若しくはトランジスタ論理コンポーネント、又はディスクリートハードウェアコンポーネントでもよい。プロセッサ1310は、この出願の実施例に開示された方法、ステップ及び論理ブロック図を実現又は実行してもよい。汎用プロセッサ1310はマイクロプロセッサでもよく、或いはプロセッサはいずれかの従来のプロセッサ等でもよい。この出願の実施例を参照して開示された方法のステップは、ハードウェアデコーディングプロセッサを使用することにより直接実行及び完了されてもよく、或いはデコーディングプロセッサ内のハードウェア及びソフトウェアモジュールの組み合わせを使用することにより実行及び完了されてもよい。ソフトウェアモジュールは、ランダムアクセスメモリ、フラッシュメモリ、フラッシュメモリ、読み取り専用メモリ、プログラム可能読み取り専用メモリ、電氣的消去可能プログラム可能メモリ又はレジスタのような当該分野における成熟した記憶媒体に位置してもよい。記憶媒体はメモリ1330に位置する。例えば、プロセッサ1310は、メモリ1320からの情報を読み取り、ハードウェアと組み合わせて前述の方法のステップを完了できる。

【0311】

例えば、プロセッサ1310は、通信インタフェース1320を使用することにより、第1のデータ読み取り要求を受信し、第1のデータ読み取り要求は第1の位置情報を搬送し、第1の位置情報は、第1のデータが属するデータボディ内の第1のデータの位置を記述するために使用され、第1の位置情報に基づいて、第1のデータが位置するメモリが第1のメモリであると決定され、第1のメモリが第1のストレージノードに属する場合、状態情報セットが、第1のメモリが信頼されたアクセス状態にあることを現在記録しているとき、第1のメモリから第1のデータを読み取り、通信インタフェースを使用することにより、第1のデータ読み取り要求に応答するために使用される第1のデータ読み取り応答を送信し、第1のデータ読み取り応答は読み取られた第1のデータを搬送する、ように構成される

10

20

30

40

50

。

【0312】

この出願の実施例はコンピュータ読み取り可能記憶媒体を提供する。コンピュータ読み取り可能記憶媒体はプログラムコードを記憶する。プログラムコードは、第1の態様におけるいずれかのストレージノード（例えば、第1のストレージノード又は第2のストレージノード）により実行される方法のステップの一部又は全部を実行するために使用される命令を含む。

【0313】

さらに、この出願の実施例は命令を含むコンピュータプログラムプロダクトを提供する。コンピュータプログラムプロダクトがコンピュータ（コンピュータは、例えばストレージノードである）上で実行したとき、コンピュータは、前述の実施例におけるいずれかのストレージノード（例えば、第1のストレージノード又は第2のストレージノード）により実行される方法のステップの一部又は全部を実行する。

【0314】

さらに、この出願の実施例は、分散型ストレージサービスシステムと端末とを含むサービスシステムを更に提供する。分散型ストレージサービスシステム及び端末は通信可能に接続される。分散型ストレージサービスシステムは、この出願の実施例において提供されるいずれかの分散型ストレージシステムである。

【0315】

前述の実施例の全部又は一部は、ソフトウェア、ハードウェア、ファームウェア又はこれらのいずれかの組み合わせを用いて実現されてもよい。ソフトウェアが実施例を実現するために使用されるとき、実施例は、全て或いは部分的にコンピュータプログラムプロダクトの形式で実現されてもよい。コンピュータプログラムプロダクトは1つ以上のコンピュータ命令を含む。コンピュータプログラム命令がコンピュータにロードされて実行されたとき、この出願の実施例による手順又は機能が全て或いは部分的に生成される。コンピュータは、汎用コンピュータ、専用コンピュータ、コンピュータネットワーク又は他のプログラム可能装置でもよい。コンピュータ命令は、コンピュータ読み取り可能記憶媒体に記憶されてもよく、或いはコンピュータ読み取り可能記憶媒体から他のコンピュータ読み取り可能記憶媒体に送信されてもよい。例えば、コンピュータ命令は、有線（例えば、同軸ケーブル、光ファイバ又はデジタル加入者線）又は無線（例えば、赤外線、無線又はマイクロ波）の方式で、ウェブサイト、コンピュータ、サーバ又はデータセンタから他のウェブサイト、コンピュータ、サーバ又はデータセンタに送信されてもよい。コンピュータ読み取り可能記憶媒体は、コンピュータによりアクセス可能ないずれかの使用可能な媒体、又は1つ以上の使用可能な媒体を統合するサーバ若しくはデータセンタのようなデータ記憶デバイスでもよい。使用可能な媒体は、磁気媒体（例えば、フロッピーディスク、ハードディスク又は磁気テープ）、光媒体（例えば、光ディスク）、半導体媒体（例えば、ソリッドステートディスク）等でもよい。前述の実施例では、実施例の説明はそれぞれの焦点を有する。実施例において詳細に記載されていない部分について、他の実施例における関係する説明に参照が行われてもよい。

【0316】

簡単な説明のため、前述の方法の実施例は一連の動作として表される点に留意すべきである。しかし、この出願に従って、いくつかのステップが他の順序で或いは同時に実行されてもよい。当業者は、この出願に記載の動作の順序に限定されないことを認識すべきである。この明細書に記載の実施例は、全て例示の実施例であり、関係する動作及びモジュールは必ずしもこの出願により必要とされるとは限らないことが、当業者により更に認識されるべきである。

【0317】

この出願において提供されるいくつかの実施例において、開示の装置は他の方式で実現されてもよいことが理解されるべきである。例えば、記載の装置の実施例は単なる例である。例えば、ユニットの分割は、単に論理的な機能分割であり、実際の実現方式では他の

分割でもよい。例えば、複数のユニット又はコンポーネントは結合されてもよく、或いは他のシステムに統合されてもよく、或いはいくつかの特徴が無視されてもよく或いは実行されなくてもよい。さらに、表示又は説明した相互結合若しくは直接結合又は通信接続は、いくつかのインタフェースを使用することにより実現されてもよい。装置又はユニットの間の間接結合又は通信接続は電氣的又は他の形式で実現されてもよい。

【0318】

別々の部分として記載したユニットは、物理的に別々でもよく或いは別々でなくともよく、ユニットとして表示された部分は、物理的なユニットでもよく或いは物理的なユニットでなくともよく、1つの場所に位置してもよく、或いは複数のネットワークユニットに分散されてもよい。ユニットの一部又は全部は、実施例の解決策の目的を達成するために、実際の要件に依存して選択されてもよい。

10

【0319】

さらに、この出願の実施例における機能ユニットは1つの処理ユニットに統合されてもよく、或いはユニットのそれぞれが物理的に単独で存在してもよく、或いは2つ以上のユニットが1つのユニットに統合される。統合されたユニットは、ハードウェアの形式で実現されてもよく、或いはソフトウェア機能ユニットの形式で実現されてもよい。

【0320】

統合されたユニットがソフトウェア機能ユニットの形式で実現されて独立したプロダクトとして販売又は使用されるとき、統合されたユニットはコンピュータアクセス可能なメモリに記憶されてもよい。このような理解に基づいて、この出願の技術的解決策は本質的に、或いは従来技術に寄与する部分、又は技術的解決策の全部若しくは一部は、ソフトウェアプロダクトの形式で実現されてもよい。コンピュータソフトウェアプロダクトはメモリに記憶され、コンピュータデバイス（パーソナルコンピュータ、サーバ又はネットワークデバイスでもよく、具体的にはコンピュータデバイス内のプロセッサでもよい）に対してこの出願の実施例に記載の前述の方法のステップの全部又は一部を実行するように命令するための要求を含む。

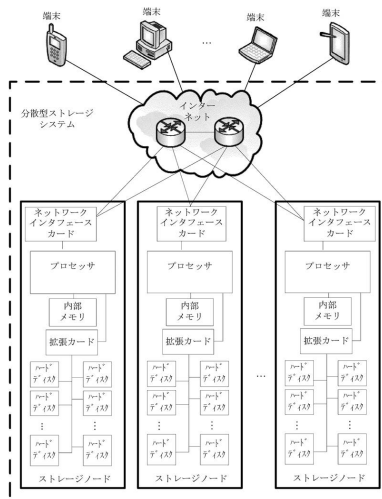
20

【0321】

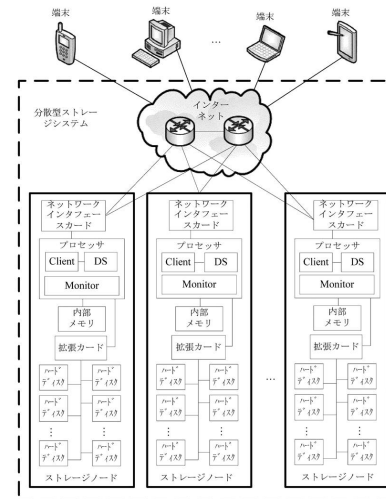
前述の実施例は、単にこの出願の技術的解決策を説明するものであり、この出願を限定することを意図するものではない。この出願について、前述の実施例を参照して詳細に説明したが、当業者は、この出願の実施例の技術的解決策の真意及び範囲を逸脱することなく、依然として前述の実施例に記載の技術的解決策に変更を行ってもよく、或いはそのいくつかの技術的特徴に等価置換を行ってもよいことを理解するべきである。

30

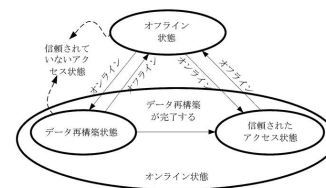
【図 1 - A】



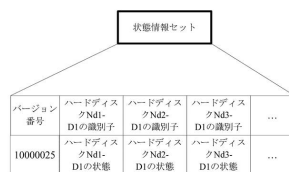
【図 1 - B】



【図 1 - C】



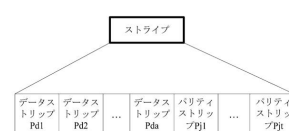
【図 1 - D】



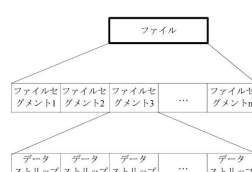
【図 1 - E】



【図 1 - F】



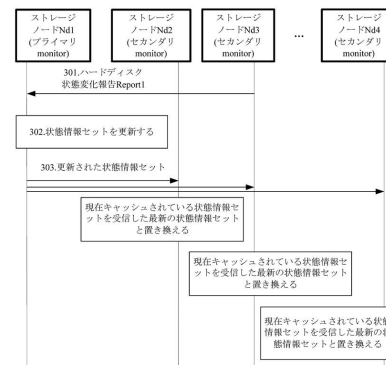
【図 2 - A】



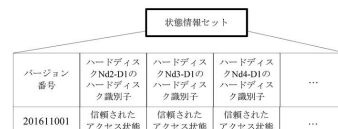
【図 2 - B】



【図 3 - A】



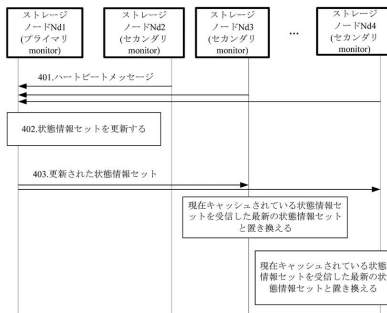
【図 3 - B】



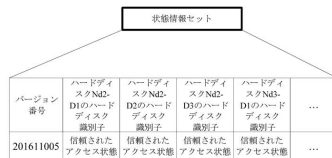
【図 3 - C】



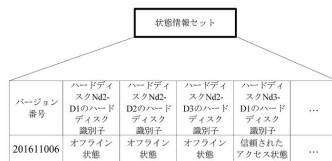
【図 4 - A】



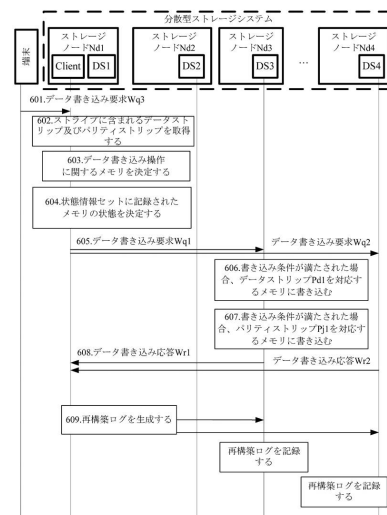
【図 4 - B】



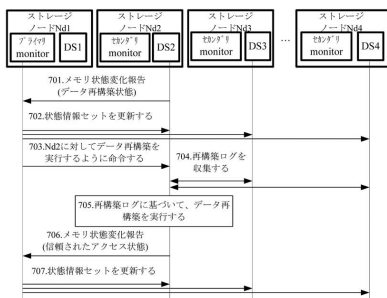
【図 4 - C】



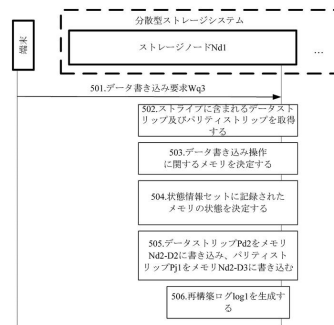
【図 6】



【図 7 - A】



【図 5 - A】



【図 5 - B】



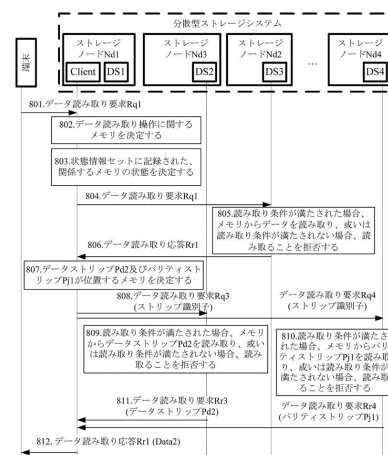
【図 7 - B】



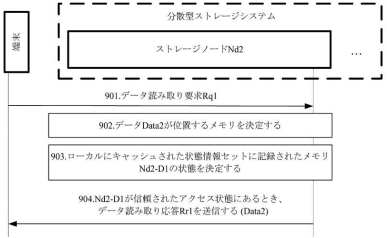
【図 7 - C】



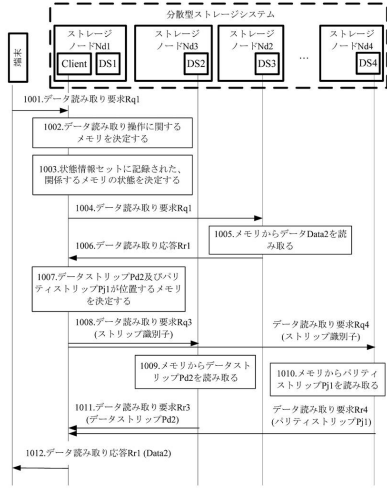
【図 8】



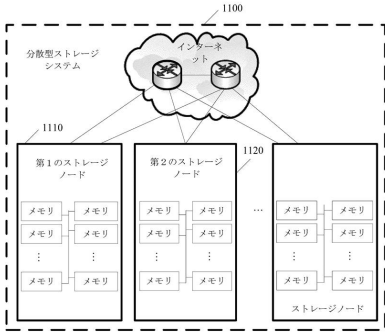
【図 9】



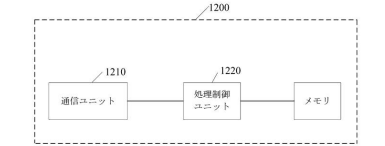
【図 10】



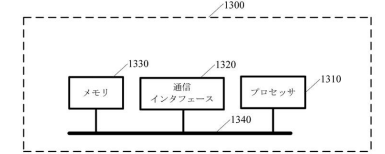
【図 11】



【図 12】



【図 13】



フロントページの続き

(51)Int.Cl. F I
G 0 6 F 13/10 3 4 0 B

(74)代理人 100070150
弁理士 伊東 忠彦

(74)代理人 100091214
弁理士 大貫 進介

(72)発明者 ヤーン, ディーングオ
中国 5 1 8 1 2 9 グァンドン シェンヂェン ロンガン・ディストリクト バンティエン ホ
ァウエイ・アドミニストレーション・ビルディング

審査官 吉田 歩

(56)参考文献 中国特許出願公開第1 0 6 4 0 6 7 5 8 (CN, A)
特開2 0 0 4 - 1 9 2 3 0 5 (JP, A)
特表2 0 1 0 - 5 0 3 9 4 1 (JP, A)

(58)調査した分野(Int.Cl., DB名)
G 0 6 F 1 3 / 1 0
G 0 6 F 3 / 0 6