

(19) 日本国特許庁 (JP)

(12) 特 許 公 報 (B2)

(11) 特許番号

特許第6347802号
(P6347802)

(45) 発行日 平成30年6月27日 (2018.6.27)

(24) 登録日 平成30年6月8日 (2018.6.8)

(51) Int. Cl.

F I

G06F 12/00 (2006.01)
G06F 3/06 (2006.01)
G06F 11/14 (2006.01)

G06F 12/00 531R
 G06F 3/06 54O
 G06F 3/06 305C
 G06F 3/06 301M
 G06F 12/00 517

請求項の数 9 外国語出願 (全 38 頁) 最終頁に続く

(21) 出願番号 特願2016-111409 (P2016-111409)
 (22) 出願日 平成28年6月3日 (2016.6.3)
 (65) 公開番号 特開2017-4514 (P2017-4514A)
 (43) 公開日 平成29年1月5日 (2017.1.5)
 審査請求日 平成28年7月26日 (2016.7.26)
 (31) 優先権主張番号 201510304374.7
 (32) 優先日 平成27年6月4日 (2015.6.4)
 (33) 優先権主張国 中国 (CN)

(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 Administrati
 on Building, Bantia
 n, Longgang Distric
 t, Shenzhen, Guangd
 ong 518129, P. R. Ch
 ina
 (74) 代理人 100107766
 弁理士 伊東 忠重

最終頁に続く

(54) 【発明の名称】 データ記憶方法、データ回復方法、関係する装置及びシステム

(57) 【特許請求の範囲】

【請求項 1】

分散型記憶システムに適用されるデータ記憶方法であって、前記分散型記憶システムは、イレージャーコーディング (EC) ストライプを使用することによりデータを記憶し、各 EC ストライプは、データ部及びパリティ部を有し、各 EC ストライプの前記データ部は、 m 個のデータブロックを有し、各 EC ストライプの前記パリティ部は、前記 m 個のデータブロックにおいて実行されるパリティコーディングの後に取得された k 個のパリティブロックを有し、前記 m 個のデータブロック及び前記 k 個のパリティブロックは、各 EC ストライプの EC block を構成し、前記分散型記憶システムは、複数の記憶ノードグループを有し、各記憶ノードグループは、 $m+k$ 個未満ではない記憶ノードを有し、1つのプライマリ記憶ノードが各記憶ノードグループ内で指定され、前記プライマリ記憶ノードを除く各記憶ノードグループ内のノードはセカンダリ記憶ノードであり、 m 及び k は共に正の整数であり、前記方法は、前記記憶ノードグループのいずれか 1 つの中のプライマリ記憶ノードにより実行されるデータ記憶方法であって、

記憶されるべきデータセグメントを受信し、ターゲット EC ストライプを取得するために、前記記憶されるべきデータセグメントにおいて EC コーディングを実行するステップであり、前記記憶されるべきデータセグメントのサイズは Z より小さくなく、 Z は m 個のデータブロックのサイズであり、前記ターゲット EC ストライプは、 $m+k$ 個のターゲット EC block を有し、前記 $m+k$ 個のターゲット EC block は、 m 個のターゲットデータブロック及び k 個のターゲットパリティブロックを有するステップと、

10

20

m+k個のターゲット記憶ノードを決定するステップであり、前記m+k個のターゲット記憶ノードは、前記m+k個のターゲットEC blockを記憶するように構成され、各ターゲット記憶ノードは、前記ターゲットEC blockの1つを記憶するステップと、

準備メッセージを前記m+k個のターゲット記憶ノードに別々に送信するステップであり、各準備メッセージは、各ターゲット記憶ノードにより記憶されるべきターゲットEC blockを有するステップと、

前記m+k個のターゲット記憶ノードのいずれかのターゲット記憶ノードにより送信された応答メッセージを受信するステップであり、前記応答メッセージは、前記応答メッセージを現在送信している前記ターゲット記憶ノードが前記準備メッセージに従って準備ログを生成したことを示すために使用され、前記準備ログは、前記応答メッセージを現在送信している前記ターゲット記憶ノードにより記憶されるべきターゲットEC blockを有するステップと、

10

前記m+k個のターゲット記憶ノード内の全てのターゲット記憶ノードの応答メッセージを受信した後に、実行メッセージを前記m+k個のターゲット記憶ノードに送信するステップであり、前記実行メッセージは、前記m+k個のターゲット記憶ノードに対して、前記準備ログに含まれる前記ターゲットEC blockを書き込むように命令するために使用されるステップと

を有するデータ記憶方法。

【請求項2】

前記準備メッセージは、前記ターゲットECストライプのバージョン番号を更に有し、

20

前記準備ログは、前記ターゲットECストライプの前記バージョン番号を更に有し、

前記実行メッセージは、前記m+k個のターゲット記憶ノードに対して、前記準備ログにある前記ターゲットEC block及び前記ターゲットECストライプの前記バージョン番号を書き込むように命令するために使用される、請求項1に記載のデータ記憶方法。

【請求項3】

準備メッセージを前記ターゲット記憶ノードに送信した後に、

故障した記憶ノードが前記m+k個のターゲット記憶ノードに存在することが検出された場合、前記故障した記憶ノードの故障回復の後に、前記m+k個のターゲット記憶ノードにおける前記ターゲットECストライプのバージョン番号が同じであるか否かを決定するステップであり、前記m+k個のターゲット記憶ノードにおける前記ターゲットECストライプの前記バージョン番号は、準備ログを生成したターゲット記憶ノードにおける準備ログ内の前記ターゲットECストライプの前記バージョン番号及び/又は準備ログを有さないターゲット記憶ノードにより記憶された前記ターゲットECストライプのバージョン番号を有するステップと、

30

前記m+k個のターゲット記憶ノード内の全てのターゲット記憶ノードにおける前記ターゲットECストライプの前記バージョン番号が同じである場合、実行メッセージを、準備ログを生成した前記ターゲット記憶ノードに送信するステップであり、前記実行メッセージは、前記ターゲット記憶ノードに対して、前記準備ログに含まれる前記ターゲットEC blockを書き込むように命令するために使用されるステップと

を更に有する、請求項2に記載のデータ記憶方法。

40

【請求項4】

前記m+k個のターゲット記憶ノード内の前記ターゲット記憶ノードにおける前記ターゲットECストライプの前記バージョン番号が異なる場合、削除メッセージを、準備ログを生成した前記ターゲット記憶ノードに送信するステップであり、前記削除メッセージは、前記ターゲット記憶ノードに対して、前記準備ログを削除するように命令するために使用されるステップを更に有する、請求項3に記載のデータ記憶方法。

【請求項5】

分散型記憶システムに適用されるデータ記憶装置であり、前記分散型記憶システムは、イレージャーコーディング(EC)ストライプを使用することによりデータを記憶し、各ECストライプは、データ部及びパリティ部を有し、各ECストライプの前記データ部は、m個

50

のデータブロックを有し、各ECストライプの前記パリティ部は、前記m個のデータブロックにおいて実行されるパリティコーディングの後に取得されたk個のパリティブロックを有し、前記m個のデータブロック及び前記k個のパリティブロックは、各ECストライプのEC blockを構成し、前記分散型記憶システムは、複数の記憶ノードグループを有し、各記憶ノードグループは、m+k個未満ではない記憶ノードを有し、1つのプライマリ記憶ノードが各記憶ノードグループ内で指定され、前記プライマリ記憶ノードを除く各記憶ノードグループ内のノードはセカンダリ記憶ノードであり、m及びkは共に正の整数であり、前記データ記憶装置は、前記記憶ノードグループのいずれか1つの中のプライマリ記憶ノードに配置されるデータ記憶装置であって、

記憶されるべきデータセグメントを受信し、ターゲットECストライプを取得するために、前記記憶されるべきデータセグメントにおいてECコーディングを実行するように構成されたデータコーディングモジュールであり、前記記憶されるべきデータセグメントのサイズはZより大きくなく、Zはm個のデータブロックのサイズであり、前記ターゲットECストライプは、m+k個のターゲットEC blockを有し、前記m+k個のターゲットEC blockは、m個のターゲットデータブロック及びk個のターゲットパリティブロックを有するデータコーディングモジュールと、

m+k個のターゲット記憶ノードを決定するように構成されたノード決定モジュールであり、前記m+k個のターゲット記憶ノードは、前記m+k個のターゲットEC blockを記憶するように構成され、各ターゲット記憶ノードは、前記ターゲットEC blockの1つを記憶するノード決定モジュールと、

準備メッセージを前記m+k個のターゲット記憶ノードに別々に送信するように構成されたメッセージ送信モジュールであり、各準備メッセージは、各ターゲット記憶ノードにより記憶されるべきターゲットEC blockを有するメッセージ送信モジュールと、

前記m+k個のターゲット記憶ノードのいずれかのターゲット記憶ノードにより送信された応答メッセージを受信するように構成されたメッセージ受信モジュールであり、前記応答メッセージは、前記応答メッセージを現在送信している前記ターゲット記憶ノードが前記準備メッセージに従って準備ログを生成したことを示すために使用され、前記準備ログは、前記応答メッセージを現在送信している前記ターゲット記憶ノードにより記憶されるべきターゲットEC blockを有するメッセージ受信モジュールと

を有し、

前記メッセージ送信モジュールは、前記メッセージ受信モジュールが前記m+k個のターゲット記憶ノード内の全てのターゲット記憶ノードの応答メッセージを受信した後に、実行メッセージを前記m+k個のターゲット記憶ノードに送信するように更に構成され、前記実行メッセージは、前記m+k個のターゲット記憶ノードに対して、前記準備ログに含まれる前記ターゲットEC blockを書き込むように命令するために使用されるデータ記憶装置

【請求項6】

前記準備メッセージは、前記ターゲットECストライプのバージョン番号を更に有し、

前記準備ログは、前記ターゲットECストライプの前記バージョン番号を更に有し、

前記実行メッセージは、前記m+k個のターゲット記憶ノードに対して、前記準備ログにある前記ターゲットEC block及び前記ターゲットECストライプの前記バージョン番号を書き込むように命令するために使用される、請求項5に記載のデータ記憶装置。

【請求項7】

故障した記憶ノードが前記m+k個のターゲット記憶ノードに存在することが検出された場合、前記故障した記憶ノードの故障回復の後に、前記m+k個のターゲット記憶ノードにおける前記ターゲットECストライプのバージョン番号が同じであるか否かを決定するように構成されたバージョン決定モジュールであり、前記m+k個のターゲット記憶ノードにおける前記ターゲットECストライプの前記バージョン番号は、準備ログを生成したターゲット記憶ノードにおける準備ログ内の前記ターゲットECストライプの前記バージョン番号及び/又は準備ログを有さないターゲット記憶ノードにより記憶されたターゲットECストラ

イプのバージョン番号を有するバージョン決定モジュールを更に有し、

前記メッセージ送信モジュールは、前記 $m+k$ 個のターゲット記憶ノード内の全てのターゲット記憶ノードにおける前記ターゲットECストライプの前記バージョン番号が同じである場合、実行メッセージを、準備ログを生成した前記ターゲット記憶ノードに送信するように構成され、前記実行メッセージは、前記ターゲット記憶ノードに対して、前記準備ログに含まれる前記ターゲットEC blockを書き込むように命令するために使用される、請求項6に記載のデータ記憶装置。

【請求項8】

前記メッセージ送信モジュールは、前記 $m+k$ 個のターゲット記憶ノード内の全てのターゲット記憶ノードにおける前記ターゲットECストライプの前記バージョン番号が異なる場合、削除メッセージを、準備ログを生成した前記ターゲット記憶ノードに送信するように更に構成され、前記削除メッセージは、前記ターゲット記憶ノードに対して、前記準備ログを削除するように命令するために使用される、請求項7に記載のデータ記憶装置。

【請求項9】

コンピュータ実行可能命令を有するコンピュータ読み取り可能記憶媒体であって、コンピュータのプロセッサが前記コンピュータ実行可能命令を実行した場合、前記コンピュータは、請求項1ないし4のうちいずれか1項に記載の方法を実行するコンピュータ読み取り可能記憶媒体。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、データ記憶の分野に関し、特にデータ記憶方法、データ回復方法、関係する装置及びシステムに関する。

【背景技術】

【0002】

現在、分散型記憶システムは、データを記憶するためにイレージャーコーディング（EC、Erasure coding）技術をますます使用している。イレージャーコーディング技術の原理は、データを m 個のデータブロックに分割し、 k 個のパリティブロックを生成するために、冗長アルゴリズムを使用することにより m 個のデータブロックにおいてパリティコーディングを実行することである。 m 個のデータブロック及び k 個のパリティブロックは、1つのECストライプと呼ばれる。データブロック又はパリティブロックは、EC blockとも呼ばれ、各ECストライプは、 k 個のEC blockの喪失を許容する。

【0003】

記憶の信頼性を確保するため、分散型記憶システムは、記憶のために、 $m+k$ 個の記憶ノードに各ECストライプの $m+k$ 個のEC blockを分配する。このように、ノードが故障した場合であっても、故障したノードに記憶されたEC blockは、他のノードに記憶されたデータに従って回復することができる。EC技術を使用する分散型記憶システムは、 k 個の記憶ノードの同時故障を許容することが認識され得る。

【0004】

しかし、EC blockを記憶ノードに書き込むことは、或る期間を必要とする。ECストライプを記憶するために使用される $m+k$ 個の記憶ノード内の記憶ノードがEC blockを書き込むときに故障している場合、記憶ノードの一部がECストライプのEC blockを有するが、記憶ノードの他の部分がECストライプのEC blockを有さない場合が生じ、ECストライプの記憶の不一致を引き起こす。従って、分散型記憶システムの記憶の一貫性が悪い。

【発明の概要】

【0005】

本発明の実施例は、分散型記憶システムの記憶の一貫性を改善するために使用されるデータ回復方法を提供する。

【0006】

本発明の実施例の第1の態様は、分散型記憶システムに適用されるデータ記憶方法を提

10

20

30

40

50

供し、分散型記憶システムは、イレージャーコーディング（EC）ストライプを使用することによりデータを記憶し、各ECストライプは、データ部及びパリティ部を含み、各ECストライプのデータ部は、 m 個のデータブロックを含み、各ECストライプのパリティ部は、パリティコーディングが m 個のデータブロックにおいて実行された後に取得された k 個のパリティブロックを含み、 m 個のデータブロック及び k 個のパリティブロックは、各ECストライプのEC blockを構成し、分散型記憶システムは、複数の記憶ノードグループを含み、各記憶ノードグループは、 $m+k$ 個未満ではない記憶ノードを含み、1つのプライマリ記憶ノードが各記憶ノードグループ内で指定され、プライマリ記憶ノードを除く各記憶ノードグループ内のノードはセカンダリ記憶ノードであり、 m 及び k は共に正の整数であり、この方法は、記憶ノードグループのいずれか1つの中のプライマリ記憶ノードにより実行され、記憶されるべきデータセグメントを受信し、ターゲットECストライプを取得するために、記憶されるべきデータセグメントにおいてECコーディングを実行するステップであり、記憶されるべきデータセグメントのサイズは Z より大きくなく、 Z は m 個のデータブロックのサイズであり、ターゲットECストライプは、 $m+k$ 個のターゲットEC blockを含み、 $m+k$ 個のターゲットEC blockは、 m 個のターゲットデータブロック及び k 個のターゲットパリティブロックを含むステップと、 $m+k$ 個のターゲット記憶ノードを決定するステップであり、 $m+k$ 個のターゲット記憶ノードは、 $m+k$ 個のターゲットEC blockを記憶するように構成され、各ターゲット記憶ノードは、ターゲットEC blockの1つを記憶するステップと、準備メッセージを $m+k$ 個のターゲット記憶ノードに別々に送信するステップであり、各準備メッセージは、各ターゲット記憶ノードにより記憶されるべきターゲットEC blockを含むステップと、 $m+k$ 個のターゲット記憶ノードのいずれかのターゲット記憶ノードにより送信された応答メッセージを受信するステップであり、応答メッセージは、応答メッセージを現在送信しているターゲット記憶ノードが準備メッセージに従って準備ログを生成したことを示すために使用され、準備ログは、応答メッセージを現在送信しているターゲット記憶ノードにより記憶されるべきターゲットEC blockを含むステップと、 $m+k$ 個のターゲット記憶ノード内の全てのターゲット記憶ノードの応答メッセージを受信した後に、実行メッセージを $m+k$ 個のターゲット記憶ノードに送信するステップであり、実行メッセージは、 $m+k$ 個のターゲット記憶ノードに対して、準備ログに含まれるターゲットEC blockを書き込むように命令するために使用されるステップとを含む。

【0007】

本発明の実施例の第1の態様を参照して、本発明の実施例の第1の態様の第1の実現方式では、準備メッセージは、ターゲットECストライプのバージョン番号を更に含み、準備ログは、ターゲットECストライプのバージョン番号を更に含み、実行メッセージは、 $m+k$ 個のターゲット記憶ノードに対して、準備ログにあるターゲットEC block及びターゲットECストライプのバージョン番号を書き込むように命令するために具体的に使用される。

【0008】

本発明の実施例の第1の態様の第1の実現方式を参照して、本発明の実施例の第1の態様の第2の実現方式は、故障した記憶ノードが $m+k$ 個のターゲット記憶ノードに存在することが検出された場合、故障した記憶ノードの故障回復の後に、 $m+k$ 個のターゲット記憶ノードにおけるターゲットECストライプのバージョン番号が同じであるか否かを決定するステップであり、 $m+k$ 個のターゲット記憶ノードにおけるターゲットECストライプのバージョン番号は、準備ログを生成したターゲット記憶ノードにおける準備ログ内のターゲットECストライプのバージョン番号及び/又は準備ログを有さないターゲット記憶ノードにより記憶されたターゲットECストライプのバージョン番号を含むステップと、 $m+k$ 個のターゲット記憶ノード内の全てのターゲット記憶ノードにおけるターゲットECストライプのバージョン番号が同じである場合、実行メッセージを、準備ログを生成したターゲット記憶ノードに送信するステップであり、実行メッセージは、ターゲット記憶ノードに対して、準備ログに含まれるターゲットEC blockを書き込むように命令するために使用されるステップとを更に含む。

【0009】

本発明の実施例の第1の態様の第2の実現方式を参照して、本発明の実施例の第1の態様の第3の実現方式は、 $m+k$ 個のターゲット記憶ノード内のターゲット記憶ノードにおけるターゲットECストライプのバージョン番号が異なる場合、削除メッセージを、準備ログを生成したターゲット記憶ノードに送信するステップであり、削除メッセージは、ターゲット記憶ノードに対して、準備ログを削除するように命令するために使用されるステップを更に含む。

【0010】

本発明の実施例の第2の態様は、分散型記憶システムに適用されるデータ回復方法を提供し、分散型記憶システムは、イレジャーコーディング(EC)ストライプを使用することによりデータを記憶し、各ECストライプは、データ部及びパリティ部を含み、各ECストライプのデータ部は、 m 個のデータブロックを含み、各ECストライプのパリティ部は、パリティコーディングが m 個のデータブロックにおいて実行された後に取得された k 個のパリティブロックを含み、 m 個のデータブロック及び k 個のパリティブロックは、各ECストライプのEC blockを構成し、各ECストライプはkey値に対応し、データ部の同じアドレス範囲を有するECストライプのkey値は同じであり、データ部の異なるアドレス範囲を有するECストライプのkey値は異なり、分散型記憶システムは、複数の記憶ノードグループを含み、各記憶ノードグループは、 $m+k$ 個未満ではない記憶ノードを含み、1つのプライマリ記憶ノードが各記憶ノードグループ内で指定され、プライマリ記憶ノードを除く各記憶ノードグループ内のノードはセカンダリ記憶ノードであり、 m 及び k は共に正の整数であり、故障した記憶ノードが記憶ノードグループのいずれか1つの中のセカンダリ記憶ノードに存在する場合、この方法は、記憶ノードグループのいずれか1つの中のプライマリ記憶ノードにより実行され、故障した記憶ノードの故障回復の後に、故障した記憶ノードにより送信された回復メッセージを受信するステップと、記憶ノードグループ内の故障していない記憶ノードにより記憶されたEC blockに従った計算を通じて、故障した記憶ノードの N 個の喪失したEC blockを取得するステップであり、 N は正の整数であるステップと、記憶を実行するために、 N 個の喪失したEC blockを故障した記憶ノードに順次に送信するステップと、 N 個の喪失したEC blockを故障した記憶ノードに順次に送信する過程において、プライマリ記憶ノードが第1のEC blockを生成し、第1のEC blockが故障した記憶ノードにより記憶されるべきであると決定した場合、第1のEC blockと同じkey値を有する第2のEC blockが、故障した記憶ノードにより現在成功して記憶されていない喪失したEC blockに存在するか否かを決定するステップと、第2のEC blockが存在すると決定された場合、故障した記憶ノードが第2のEC blockを成功して記憶した後に、記憶を実行するために、第1のEC blockを故障した記憶ノードに送信するステップとを含む。

【0011】

本発明の実施例の第2の態様を参照して、本発明の実施例の第2の態様の第1の実現方式では、記憶のために、 N 個の喪失したEC blockを故障した記憶ノードに順次に送信するステップは、第 n の喪失したEC blockを故障した記憶ノードに送信するステップであり、 $1 \leq n \leq N$ であるステップと、第 n の喪失したEC blockに対応する応答メッセージを受信するステップであり、応答メッセージは、故障した記憶ノードにより送信され、応答メッセージは、故障した記憶ノードが第 n の喪失したEC blockを成功して記憶したことを示すために使用されるステップと、第 n の喪失したEC blockに関して故障した記憶ノードの応答メッセージを受信した後に、 $n < N$ である場合、 n の値に1を加算し、第 n の喪失したEC blockを故障した記憶ノードに送信するステップを再実行するステップとを含む。

【0012】

本発明の実施例の第2の態様の第1の実現方式を参照して、本発明の実施例の第2の態様の第2の実現方式では、現在成功して記憶されていない喪失したEC blockは、故障した記憶ノードに現在送信されていない喪失したEC blockと、故障した記憶ノードに現在送信されているが、それに対応する応答メッセージが受信されていない喪失したEC blockとを含む。

【0013】

本発明の実施例の第2の態様又は第2の態様の第1若しくは第2の実現方式を参照して、本発明の実施例の第2の態様の第3の実現方式では、分散型記憶システムの論理ボリュームは、複数の論理パーティションを含み、各論理パーティションは、 k 個のデータブロックのサイズであり、互いに重複せず、各論理パーティションは、1つのkey値に一意に対応し、各EC blockが入る論理パーティションのkey値は、EC blockのkey値である。

【0014】

本発明の実施例の第3の態様は、データ記憶装置を提供し、分散型記憶システムは、イレジャーコーディング(EC)ストライプを使用することによりデータを記憶し、各ECストライプは、データ部及びパリティ部を含み、各ECストライプのデータ部は、 m 個のデータブロックを含み、各ECストライプのパリティ部は、パリティコーディングが m 個のデータブロックにおいて実行された後に取得された k 個のパリティブロックを含み、 m 個のデータブロック及び k 個のパリティブロックは、各ECストライプのEC blockを構成し、分散型記憶システムは、複数の記憶ノードグループを含み、各記憶ノードグループは、 $m+k$ 個未満ではない記憶ノードを含み、1つのプライマリ記憶ノードが各記憶ノードグループ内で指定され、プライマリ記憶ノードを除く各記憶ノードグループ内のノードはセカンダリ記憶ノードであり、 m 及び k は共に正の整数であり、データ記憶装置は、記憶ノードグループのいずれか1つの中のプライマリ記憶ノードに配置され、データ記憶装置は、記憶されるべきデータセグメントを受信し、ターゲットECストライプを取得するために、記憶されるべきデータセグメントにおいてECコーディングを実行するように構成されたデータコーディングモジュールであり、記憶されるべきデータセグメントのサイズは Z より大きくなく、 Z は m 個のデータブロックのサイズであり、ターゲットECストライプは、 $m+k$ 個のターゲットEC blockを含み、 $m+k$ 個のターゲットEC blockは、 m 個のターゲットデータブロック及び k 個のターゲットパリティブロックを含むデータコーディングモジュールと、 $m+k$ 個のターゲット記憶ノードを決定するように構成されたノード決定モジュールであり、 $m+k$ 個のターゲット記憶ノードは、 $m+k$ 個のターゲットEC blockを記憶するように構成され、各ターゲット記憶ノードは、ターゲットEC blockの1つを記憶するノード決定モジュールと、準備メッセージを $m+k$ 個のターゲット記憶ノードに別々に送信するように構成されたメッセージ送信モジュールであり、各準備メッセージは、各ターゲット記憶ノードにより記憶されるべきターゲットEC blockを含むメッセージ送信モジュールと、 $m+k$ 個のターゲット記憶ノードのいずれかのターゲット記憶ノードにより送信された応答メッセージを受信するように構成されたメッセージ受信モジュールであり、応答メッセージは、応答メッセージを現在送信しているターゲット記憶ノードが準備メッセージに従って準備ログを生成したことを示すために使用され、準備ログは、応答メッセージを現在送信しているターゲット記憶ノードにより記憶されるべきターゲットEC blockを含むメッセージ受信モジュールとを含み、メッセージ送信モジュールは、メッセージ受信モジュールが $m+k$ 個のターゲット記憶ノード内の全てのターゲット記憶ノードの応答メッセージを受信した後に、実行メッセージを $m+k$ 個のターゲット記憶ノードに送信するように更に構成され、実行メッセージは、 $m+k$ 個のターゲット記憶ノードに対して、準備ログに含まれるターゲットEC blockを書き込むように命令するために使用される。

【0015】

本発明の実施例の第3の態様を参照して、本発明の実施例の第3の態様の第1の実現方式では、準備メッセージは、ターゲットECストライプのバージョン番号を更に含み、準備ログは、ターゲットECストライプのバージョン番号を更に含み、実行メッセージは、 $m+k$ 個のターゲット記憶ノードに対して、準備ログにあるターゲットEC block及びターゲットECストライプのバージョン番号を書き込むように命令するために具体的に使用される。

【0016】

本発明の実施例の第3の態様の第1の実現方式を参照して、本発明の実施例の第3の態様の第2の実現方式では、この装置は、故障した記憶ノードが $m+k$ 個のターゲット記憶ノードに存在することが検出された場合、故障した記憶ノードの故障回復の後に、 $m+k$ 個のターゲット記憶ノードにおけるターゲットECストライプのバージョン番号が同じであるか

否かを決定するように構成されたバージョン決定モジュールであり、 $m+k$ 個のターゲット記憶ノードにおけるターゲットECストライプのバージョン番号は、準備ログを生成したターゲット記憶ノードにおける準備ログ内のターゲットECストライプのバージョン番号及び/又は準備ログを有さないターゲット記憶ノードにより記憶されたターゲットECストライプのバージョン番号を含むバージョン決定モジュールを更に含み、メッセージ送信モジュールは、 $m+k$ 個のターゲット記憶ノード内の全てのターゲット記憶ノードにおけるターゲットECストライプのバージョン番号が同じである場合、実行メッセージを、準備ログを生成したターゲット記憶ノードに送信するように具体的に構成され、実行メッセージは、ターゲット記憶ノードに対して、準備ログに含まれるターゲットEC blockを書き込むように命令するために使用される。

10

【0017】

本発明の実施例の第3の態様の第2の実現方式を参照して、本発明の実施例の第3の態様の第3の実現方式では、メッセージ送信モジュールは、 $m+k$ 個のターゲット記憶ノード内のターゲット記憶ノードにおけるターゲットECストライプのバージョン番号が異なる場合、削除メッセージを、準備ログを生成したターゲット記憶ノードに送信するように更に構成され、削除メッセージは、ターゲット記憶ノードに対して、準備ログを削除するように命令するために使用される。

【0018】

本発明の実施例の第4の態様は、分散型記憶システムに適用されるデータ回復装置を提供し、分散型記憶システムは、イレージャーコーディング(EC)ストライプを使用することによりデータを記憶し、各ECストライプは、データ部及びパリティ部を含み、各ECストライプのデータ部は、 m 個のデータブロックを含み、各ECストライプのパリティ部は、パリティコーディングが m 個のデータブロックにおいて実行された後に取得された k 個のパリティブロックを含み、 m 個のデータブロック及び k 個のパリティブロックは、各ECストライプのEC blockを構成し、各ECストライプはkey値に対応し、データ部の同じアドレス範囲を有するECストライプのkey値は同じであり、データ部の異なるアドレス範囲を有するECストライプのkey値は異なり、分散型記憶システムは、複数の記憶ノードグループを含み、各記憶ノードグループは、 $m+k$ 個未満ではない記憶ノードを含み、1つのプライマリ記憶ノードが各記憶ノードグループ内で指定され、プライマリ記憶ノードを除く各記憶ノードグループ内のノードはセカンダリ記憶ノードであり、 m 及び k は共に正の整数であり、故障した記憶ノードが分散型記憶システム内のいずれかの記憶ノードグループ内のセカンダリ記憶ノードに存在する場合、データ回復装置は、故障した記憶ノードが属する記憶ノードグループ内のプライマリ記憶ノードに配置され、データ回復装置は、故障した記憶ノードの故障回復の後に、故障した記憶ノードにより送信された回復メッセージを受信するように構成された回復メッセージモジュールと、記憶ノードグループ内の故障していない記憶ノードにより記憶されたEC blockに従った計算を通じて、故障した記憶ノードの N 個の喪失したEC blockを取得するように構成されたデータ計算モジュールであり、 N は正の整数であるデータ計算モジュールと、記憶を実行するために、 N 個の喪失したEC blockを故障した記憶ノードに順次送信するように構成されたデータ送信モジュールと、 N 個の喪失したEC blockを故障した記憶ノードに順次送信する過程において、プライマリ記憶ノードが第1のEC blockを生成し、第1のEC blockが故障した記憶ノードにより記憶されるべきであると決定した場合、第1のEC blockと同じkey値を有する第2のEC blockが、故障した記憶ノードにより現在成功して記憶されていない喪失したEC blockに存在するか否かを決定するように構成されたkey値決定モジュールとを含み、データ送信モジュールは、第2のEC blockが存在するとkey値決定モジュールが決定した場合、故障した記憶ノードが第2のEC blockを成功して記憶した後に、記憶を実行するために、第1のEC blockを故障した記憶ノードに送信するように更に構成される。

20

30

40

【0019】

本発明の実施例の第4の態様を参照して、本発明の実施例の第4の態様の第1の実現方式では、データ送信モジュールは、第 n の喪失したEC blockを故障した記憶ノードに送信

50

するように具体的に構成され、 $1 \leq n \leq N$ であり、第 n の喪失したEC blockに対応する応答メッセージを受信するように具体的に構成され、応答メッセージは、故障した記憶ノードにより送信され、応答メッセージは、故障した記憶ノードが第 n の喪失したEC blockを成功して記憶したことを示すために使用され、第 n の喪失したEC blockに関して故障した記憶ノードの応答メッセージを受信した後に、 $n < N$ である場合、 n の値に1を加算し、第 n の喪失したEC blockを故障した記憶ノードに送信するステップを再実行するように具体的に構成される。

【0020】

本発明の実施例の第4の態様の第1の実現方式を参照して、本発明の実施例の第4の態様の第2の実現方式では、現在成功して記憶されていない喪失したEC blockは、故障した記憶ノードに現在送信されていない喪失したEC blockと、故障した記憶ノードに現在送信されているが、それに対応する応答メッセージが受信されていない喪失したEC blockを含む。

【0021】

本発明の実施例の第4の態様又は第4の態様の第1若しくは第2の実現方式を参照して、本発明の実施例の第4の態様の第3の実現方式では、分散型記憶システムの論理ボリュームは、複数の論理パーティションを含み、各論理パーティションは、 k 個のデータブロックのサイズであり、互いに重複せず、各論理パーティションは、1つのkey値に一意に対応し、各EC blockが入る論理パーティションのkey値は、EC blockのkey値である。

【0022】

本発明の実施例の第5の態様は、データ記憶装置を提供し、データ記憶装置は、入力装置と、出力装置と、プロセッサと、メモリとを含み、メモリに記憶された動作命令を呼び出すことにより、プロセッサは、記憶されるべきデータセグメントを受信し、ターゲットECストライプを取得するために、記憶されるべきデータセグメントにおいてECコーディングを実行するように構成され、記憶されるべきデータセグメントのサイズは Z より大きくなく、 Z は m 個のデータブロックのサイズであり、ターゲットECストライプは、 $m+k$ 個のターゲットEC blockを含み、 $m+k$ 個のターゲットEC blockは、 m 個のターゲットデータブロック及び k 個のターゲットパリティブロックを含み、 $m+k$ 個のターゲット記憶ノードを決定するように構成され、 $m+k$ 個のターゲット記憶ノードは、 $m+k$ 個のターゲットEC blockを記憶するように構成され、各ターゲット記憶ノードは、ターゲットEC blockの1つを記憶し、準備メッセージを $m+k$ 個のターゲット記憶ノードに別々に送信するように構成され、各準備メッセージは、各ターゲット記憶ノードにより記憶されるべきターゲットEC blockを含み、 $m+k$ 個のターゲット記憶ノードのいずれかのターゲット記憶ノードにより送信された応答メッセージを受信するように構成され、応答メッセージは、応答メッセージを現在送信しているターゲット記憶ノードが準備メッセージに従って準備ログを生成したことを示すために使用され、準備ログは、応答メッセージを現在送信しているターゲット記憶ノードにより記憶されるべきターゲットEC blockを含み、 $m+k$ 個のターゲット記憶ノード内の全てのターゲット記憶ノードの応答メッセージを受信した後に、実行メッセージを $m+k$ 個のターゲット記憶ノードに送信するように構成され、実行メッセージは、 $m+k$ 個のターゲット記憶ノードに対して、準備ログに含まれるターゲットEC blockを書き込むように命令するために使用される。

【0023】

本発明の実施例の第5の態様を参照して、本発明の実施例の第5の態様の第1の実現方式では、準備メッセージは、ターゲットECストライプのバージョン番号を更に含み、準備ログは、ターゲットECストライプのバージョン番号を更に含み、実行メッセージは、 $m+k$ 個のターゲット記憶ノードに対して、準備ログにあるターゲットEC block及びターゲットECストライプのバージョン番号を書き込むように命令するために具体的に使用される。

【0024】

本発明の実施例の第5の態様の第1の実現方式を参照して、本発明の実施例の第5の態様の第2の実現方式では、プロセッサは、故障した記憶ノードが $m+k$ 個のターゲット記憶

10

20

30

40

50

ノードに存在することが検出された場合、故障した記憶ノードの故障回復の後に、 $m+k$ 個のターゲット記憶ノードにおけるターゲットECストライプのバージョン番号が同じであるか否かを決定するように更に構成され、 $m+k$ 個のターゲット記憶ノードにおけるターゲットECストライプのバージョン番号は、準備ログを生成したターゲット記憶ノードにおける準備ログ内のターゲットECストライプのバージョン番号及び/又は準備ログを有さないターゲット記憶ノードにより記憶されたターゲットECストライプのバージョン番号を含み、 $m+k$ 個のターゲット記憶ノード内の全てのターゲット記憶ノードにおけるターゲットECストライプのバージョン番号が同じである場合、実行メッセージを、準備ログを生成したターゲット記憶ノードに送信するように更に構成され、実行メッセージは、ターゲット記憶ノードに対して、準備ログに含まれるターゲットEC blockを書き込むように命令するために使用される。

10

【0025】

本発明の実施例の第5の態様の第2の実現方式を参照して、本発明の実施例の第5の態様の第3の実現方式では、プロセッサは、 $m+k$ 個のターゲット記憶ノード内のターゲット記憶ノードにおけるターゲットECストライプのバージョン番号が異なる場合、削除メッセージを、準備ログを生成したターゲット記憶ノードに送信するように更に構成され、削除メッセージは、ターゲット記憶ノードに対して、準備ログを削除するように命令するために使用される。

【0026】

本発明の実施例の第6の態様は、データ回復装置を提供し、データ回復装置は、入力装置と、出力装置と、プロセッサと、メモリとを含み、メモリに記憶された動作命令を呼び出すことにより、プロセッサは、故障した記憶ノードの故障回復の後に、故障した記憶ノードにより送信された回復メッセージを受信し、故障した記憶ノードが属する記憶ノードグループ内の故障していない記憶ノードにより記憶されたEC blockに従った計算を通じて、故障した記憶ノードの N 個の喪失したEC blockを取得するように構成され、 N は正の整数であり、記憶を実行するために、 N 個の喪失したEC blockを故障した記憶ノードに順次に送信し、 N 個の喪失したEC blockを故障した記憶ノードに順次に送信する過程において、プライマリ記憶ノードが第1のEC blockを生成し、第1のEC blockが故障した記憶ノードにより記憶されるべきであると決定した場合、第1のEC blockと同じkey値を有する第2のEC blockが、故障した記憶ノードにより現在成功して記憶されていない喪失したEC blockに存在するか否かを決定し、第2のEC blockが存在すると決定された場合、故障した記憶ノードが第2のEC blockを成功して記憶した後に、記憶を実行するために、第1のEC blockを故障した記憶ノードに送信するように構成される。

20

30

【0027】

本発明の実施例の第6の態様を参照して、本発明の実施例の第6の態様の第1の実現方式では、プロセッサは、第 n の喪失したEC blockを故障した記憶ノードに送信するように更に構成され、 $1 \leq n \leq N$ であり、第 n の喪失したEC blockに対応する応答メッセージを受信するように更に構成され、応答メッセージは、故障した記憶ノードにより送信され、応答メッセージは、故障した記憶ノードが第 n の喪失したEC blockを成功して記憶したことを示すために使用され、第 n の喪失したEC blockに関して故障した記憶ノードの応答メッセージを受信した後に、 $n < N$ である場合、 n の値に1を加算し、第 n の喪失したEC blockを故障した記憶ノードに送信するステップを再実行するように更に構成される。

40

【0028】

本発明の実施例の第6の態様の第1の実現方式を参照して、本発明の実施例の第6の態様の第2の実現方式では、現在成功して記憶されていない喪失したEC blockは、故障した記憶ノードに現在送信されていない喪失したEC blockと、故障した記憶ノードに現在送信されているが、それに対応する応答メッセージが受信されていない喪失したEC blockを含む。

【0029】

本発明の実施例の第6の態様又は第6の態様の第1若しくは第2の実現方式を参照して

50

、本発明の実施例の第6の態様の第3の実現方式では、分散型記憶システムの論理ボリュームは、複数の論理パーティションを含み、各論理パーティションは、 k 個のデータブロックのサイズであり、互いに重複せず、各論理パーティションは、1つのkey値に一意に対応し、各EC blockが入る論理パーティションのkey値は、EC blockのkey値である。

【0030】

本発明の実施例の第7の態様は、コンピュータ実行可能命令を含むコンピュータ読み取り可能媒体を提供し、コンピュータのプロセッサがコンピュータ実行可能命令を実行した場合、コンピュータは、以下の方法、すなわち、記憶されるべきデータセグメントを受信し、ターゲットECストライプを取得するために、記憶されるべきデータセグメントにおいてECコーディングを実行するステップであり、記憶されるべきデータセグメントのサイズは Z より大きくなく、 Z は m 個のデータブロックのサイズであり、ターゲットECストライプは、 $m+k$ 個のターゲットEC blockを含み、 $m+k$ 個のターゲットEC blockは、 m 個のターゲットデータブロック及び k 個のターゲットパリティブロックを含むステップと、 $m+k$ 個のターゲット記憶ノードを決定するステップであり、 $m+k$ 個のターゲット記憶ノードは、 $m+k$ 個のターゲットEC blockを記憶するように構成され、各ターゲット記憶ノードは、ターゲットEC blockの1つを記憶するステップと、準備メッセージを $m+k$ 個のターゲット記憶ノードに別々に送信するステップであり、各準備メッセージは、各ターゲット記憶ノードにより記憶されるべきターゲットEC blockを含むステップと、 $m+k$ 個のターゲット記憶ノードのいずれかのターゲット記憶ノードにより送信された応答メッセージを受信するステップであり、応答メッセージは、応答メッセージを現在送信しているターゲット記憶ノードが準備メッセージに従って準備ログを生成したことを示すために使用され、準備ログは、応答メッセージを現在送信しているターゲット記憶ノードにより記憶されるべきターゲットEC blockを含むステップと、 $m+k$ 個のターゲット記憶ノード内の全てのターゲット記憶ノードの応答メッセージを受信した後に、実行メッセージを $m+k$ 個のターゲット記憶ノードに送信するステップであり、実行メッセージは、 $m+k$ 個のターゲット記憶ノードに対して、準備ログに含まれるターゲットEC blockを書き込むように命令するために使用されるステップとを実行する。

【0031】

本発明の実施例の第7の態様を参照して、本発明の実施例の第7の態様の第1の実現方式では、準備メッセージは、ターゲットECストライプのバージョン番号を更に含み、準備ログは、ターゲットECストライプのバージョン番号を更に含み、実行メッセージは、 $m+k$ 個のターゲット記憶ノードに対して、準備ログにあるターゲットEC block及びターゲットECストライプのバージョン番号を書き込むように命令するために具体的に使用される。

【0032】

本発明の実施例の第7の態様の第1の実現方式を参照して、本発明の実施例の第7の態様の第2の実現方式では、コンピュータは、以下の方法、すなわち、故障した記憶ノードが $m+k$ 個のターゲット記憶ノードに存在することが検出された場合、故障した記憶ノードの故障回復の後に、 $m+k$ 個のターゲット記憶ノードにおけるターゲットECストライプのバージョン番号が同じであるか否かを決定するステップであり、 $m+k$ 個のターゲット記憶ノードにおけるターゲットECストライプのバージョン番号は、準備ログを生成したターゲット記憶ノードにおける準備ログ内のターゲットECストライプのバージョン番号及び/又は準備ログを有さないターゲット記憶ノードにより記憶されたターゲットECストライプのバージョン番号を含むステップと、 $m+k$ 個のターゲット記憶ノード内の全てのターゲット記憶ノードにおけるターゲットECストライプのバージョン番号が同じである場合、実行メッセージを、準備ログを生成したターゲット記憶ノードに送信するステップであり、実行メッセージは、ターゲット記憶ノードに対して、準備ログに含まれるターゲットEC blockを書き込むように命令するために使用されるステップとを更に実行する。

【0033】

本発明の実施例の第7の態様の第2の実現方式を参照して、本発明の実施例の第7の態様の第3の実現方式では、コンピュータは、以下の方法、すなわち、 $m+k$ 個のターゲット

記憶ノード内のターゲット記憶ノードにおけるターゲットECストライプのバージョン番号が異なる場合、削除メッセージを、準備ログを生成したターゲット記憶ノードに送信するステップであり、削除メッセージは、ターゲット記憶ノードに対して、準備ログを削除するように命令するために使用されるステップを更に実行する。

【0034】

本発明の実施例の第8の態様は、コンピュータ実行可能命令を含むコンピュータ読み取り可能媒体を提供し、コンピュータのプロセッサがコンピュータ実行可能命令を実行した場合、コンピュータは、以下の方法、すなわち、故障した記憶ノードの故障回復の後に、故障した記憶ノードにより送信された回復メッセージを受信するステップと、記憶ノードグループ内の故障していない記憶ノードにより記憶されたEC blockに従った計算を通じて、故障した記憶ノードのN個の喪失したEC blockを取得するステップであり、Nは正の整数であるステップと、記憶を実行するために、N個の喪失したEC blockを故障した記憶ノードに順次に送信するステップと、N個の喪失したEC blockを故障した記憶ノードに順次に送信する過程において、プライマリ記憶ノードが第1のEC blockを生成し、第1のEC blockが故障した記憶ノードにより記憶されるべきであると決定した場合、第1のEC blockと同じkey値を有する第2のEC blockが、故障した記憶ノードにより現在成功して記憶されていない喪失したEC blockに存在するか否かを決定するステップと、第2のEC blockが存在すると決定された場合、故障した記憶ノードが第2のEC blockを成功して記憶した後に、記憶を実行するために、第1のEC blockを故障した記憶ノードに送信するステップとを実行する。

【0035】

本発明の実施例の第8の態様を参照して、本発明の実施例の第8の態様の第1の実現方式では、コンピュータは、以下の方法、すなわち、第nの喪失したEC blockを故障した記憶ノードに送信するステップであり、 $1 \leq n \leq N$ であるステップと、第nの喪失したEC blockに対応する応答メッセージを受信するステップであり、応答メッセージは、故障した記憶ノードにより送信され、応答メッセージは、故障した記憶ノードが第nの喪失したEC blockを成功して記憶したことを示すために使用されるステップと、第nの喪失したEC blockに関して故障した記憶ノードの応答メッセージを受信した後に、 $n < N$ である場合、nの値に1を加算し、第nの喪失したEC blockを故障した記憶ノードに送信するステップを再実行するステップとを更に実行する。

【0036】

本発明の実施例の第8の態様の第1の実現方式を参照して、本発明の実施例の第8の態様の第2の実現方式では、現在成功して記憶されていない喪失したEC blockは、故障した記憶ノードに現在送信されていない喪失したEC blockと、故障した記憶ノードに現在送信されているが、それに対応する応答メッセージが受信されていない喪失したEC blockとを含む。

【0037】

本発明の実施例の第8の態様又は第8の態様の第1若しくは第2の実現方式を参照して、本発明の実施例の第8の態様の第3の実現方式では、分散型記憶システムの論理ボリュームは、複数の論理パーティションを含み、各論理パーティションは、k個のデータブロックのサイズであり、互いに重複せず、各論理パーティションは、1つのkey値に一意に対応し、各EC blockが入る論理パーティションのkey値は、EC blockのkey値である。

【0038】

本発明の実施例の第9の態様は、記憶ノードを提供し、本発明の実施例の第3の態様若しくは第3の態様の第1～第3の実現方式のいずれか1つによるデータ記憶装置及び/又は本発明の実施例の第4の態様若しくは第4の態様の第1～第3の実現方式のいずれか1つによるデータ回復装置を含み、及び/又は、本発明の実施例の第5の態様若しくは第5の態様の第1～第3の実現方式のいずれか1つによるデータ記憶装置及び/又は本発明の実施例の第6の態様若しくは第6の態様の第1～第3の実現方式のいずれか1つによるデータ回復装置を含み、及び/又は、本発明の実施例の第7の態様若しくは第7の態様の第

1 ~ 第3の実現方式のいずれか1つによるコンピュータ読み取り可能媒体及び/又は本発明の実施例の第8の態様若しくは第8の態様の第1 ~ 第3の実現方式のいずれか1つによるコンピュータ読み取り可能媒体を含む。

【0039】

本発明の実施例により提供されるデータ記憶方法は、プライマリノードにより、ターゲットECストライプを取得するために、記憶されるべきデータセグメントにおいてECコーディングを実行するステップであり、ターゲットECストライプは、 $m+k$ 個のターゲットEC blockを含むステップと、記憶ノードグループ内のセカンダリ記憶ノードの中で、 $m+k$ 個のターゲットEC blockを記憶するために使用される $m+k$ 個のターゲット記憶ノードを決定するステップと、準備メッセージをターゲット記憶ノードに送信するステップであり、各準備メッセージは、ターゲット記憶ノードにより記憶されるべきターゲットEC blockを搬送するステップと、ターゲット記憶ノードにより送信された応答メッセージを受信するステップと、全てのターゲット記憶ノードの応答メッセージを受信した後に、ターゲット記憶ノードに対して、準備ログにあるターゲットEC blockを書き込むように命令するために、実行メッセージを全てのターゲット記憶ノードに送信するステップとを含む。この実施例により提供される方法では、ターゲットECストライプを記憶するために使用される $m+k$ 個のターゲット記憶ノードは、記憶されるべきターゲットEC blockを直接書き込まないが、まず、記憶されるべきターゲットEC blockをローカルに生成された準備ログに書き込み、全てのターゲット記憶ノードが準備ログを生成した後に、ターゲットEC blockを書き込む。全てのターゲット記憶ノードが準備ログを生成した場合にのみ、ターゲット記憶ノードは、ターゲットEC blockを正確な位置に書き込む動作を実行する。ターゲット記憶ノードによりターゲットEC blockを書き込む過程において故障が発生した場合であっても、ターゲットEC blockは、ターゲット記憶ノードの準備ログに記憶されている。このように、ターゲットEC blockを書き込む動作が成功したか否かに拘わらず、ターゲットEC blockが喪失されないことが確保され得る。従って、ターゲット記憶ノードの一部がターゲットECストライプのターゲットEC blockを有するが、ターゲット記憶ノードの他の部分がターゲットECストライプのターゲットEC blockを有さない場合が回避され、ターゲットECストライプが一貫して記憶されることができ、分散型記憶システムの記憶の一貫性が改善される。

【図面の簡単な説明】

【0040】

【図1(a)】ECストライプの構成図

【図1(b)】本発明の実施例による分散型記憶システムの概略構成図

【図2】本発明の実施例によるデータ記憶方法の実施例のフローチャート

【図3】本発明の実施例によるデータ記憶方法の他の実施例のフローチャート

【図4】本発明の実施例によるデータ記憶方法の他の実施例のフローチャート

【図5】本発明の実施例によるデータ回復方法の実施例のフローチャート

【図6】本発明の実施例によるデータ記憶装置の実施例の構成図

【図7】本発明の実施例によるデータ記憶装置の他の実施例の構成図

【図8】本発明の実施例によるデータ回復装置の他の実施例の構成図

【図9】本発明の実施例によるデータ記憶装置の他の実施例の構成図

【図10】本発明の実施例によるデータ記憶装置の他の実施例の構成図

【図11】本発明の実施例によるデータ回復装置の他の実施例の構成図

【発明を実施するための形態】

【0041】

本発明の実施例は、分散型記憶システムの記憶の一貫性を改善するために使用されるデータ記憶方法を提供する。本発明の実施例は、以下に別々に説明するデータ回復方法、関係する装置及びシステムを更に提供する。

【0042】

分散型記憶システムは、複数の記憶ノードを含む。上位レイヤデータ配送装置は、記憶

されるべきデータを複数のデータブロックに分割し、次に、記憶のために、分散型記憶システムの複数の記憶ノードにデータブロックを配送する。データの分散型記憶は、データ喪失リスクを低減し、分散型記憶システムの信頼性を改善することができる。データ配送装置は、クライアント、サーバ、又は他のデバイスでもよい。

【0043】

EC技術の原理は、記憶されるべきデータを m 個のデータブロックに分割し、 k 個のパリティブロックを生成するために、冗長アルゴリズムを使用することにより m 個のデータブロックにおいてパリティコーディングを実行することである。 m 個のデータブロック及び k 個のパリティブロックは、1つのECストライプを構成する。ECストライプの基本的な構成について、図1(a)を参照する。各データブロック又はパリティブロックは、EC block 10と呼ばれてもよく、 m 個のデータブロック及び k 個のパリティブロックは、各ECストライプの全てのEC blockを構成する。各ECストライプは、 k 個のEC blockの喪失を許容可能である。分散型記憶システムがデータ記憶を実行するためにEC技術を使用する場合、データ配送装置は、記憶されるべきデータに従って複数のECストライプを生成し、記憶のために、分散型記憶システムの $m+k$ 個の記憶ノードに各ECストライプの $m+k$ 個のEC blockを配送する。このように、分散型記憶システム内のノードが故障した場合、故障したノードに記憶されたEC blockは、故障していないノードにおけるEC blockに従って回復することができる。ECストライプを記憶する $m+k$ 個の記憶ノード内の k 個より多くの記憶ノードが故障しない限り、ECストライプのデータは、データ配送装置により完全に読み取り又は書き込み可能である。従って、データを記憶するためにEC技術を使用する分散型記憶システムは 20、非常に高い信頼性を有する。

【0044】

分散型記憶システムに記憶されたECストライプは、データ配送装置により読み取り又は上書き可能である。ECストライプを読み取る場合、データ配送装置は、ECストライプからデータを記憶するために使用されるデータブロックを読み取る必要がある点に留意すべきである。しかし、ECストライプが上書きされる場合、例えば、ECストライプの m 個のデータブロック内の第 i のデータブロックが上書きされる場合、ECストライプの他のデータブロックが変更される必要はないが、ECストライプのパリティブロックを再生成することが必要である。

【0045】

特に、本発明の実施例では、分散型記憶システム内の記憶ノードがグループ化される。具体的に、記憶ノードは、複数の記憶ノードグループにグループ化され、各記憶ノードグループに含まれる記憶ノードの量は、 $m+k$ 未満ではなく、1つのプライマリ記憶ノード（以下では略してプライマリノード）が各記憶ノードグループ内で指定される。他の記憶ノードは、セカンダリ記憶ノード（以下では略してセカンダリノード）である。 m 及び k は共に正の整数である。異なる記憶ノードグループの記憶ノードは、繰り返し可能でもよい。例えば、記憶ノードグループ1内のプライマリノードは、記憶ノードグループ2内のセカンダリノードでもよい。図1(b)は、本発明の実施例による分散型記憶システムの概略構成図である。図1(b)は、分散型記憶システムの構成のみを示すために使用され、分散型記憶システム内の記憶ノードはまた、他の接続構成でもよい。 40

【0046】

前述の分散型記憶システムに基づいて、本発明の実施例は、分散型記憶システムのいずれかの記憶ノードグループ内のプライマリノードに適用されるデータ記憶方法を提供する。図2を参照すると、この方法の基本的な手順は以下を含む。

【0047】

201. ターゲットECストライプを取得するために、記憶されるべきデータセグメントにおいてECコーディングを実行する。

【0048】

プライマリノードは、データ配送装置により配送された記憶されるべきデータセグメントを受信し、記憶されるべきデータセグメントのサイズは Z より大きくなく、 Z は m 個のデ 50

ータブロックのサイズである。コーディング後のターゲットECストライプは、 $m+k$ 個のターゲットEC blockを含み、これらは、 m 個のターゲットデータブロック及び k 個のターゲットパリティブロックを具体的に含む。

【0049】

202. $m+k$ 個のターゲット記憶ノードを決定する。

【0050】

$m+k$ 個のターゲットEC blockを生成した後に、プライマリノードは、プライマリノードが属する記憶ノードグループにおいて、 $m+k$ 個のターゲットEC blockを記憶するために使用される $m+k$ 個のターゲット記憶ノードを決定する。 $m+k$ 個のターゲット記憶ノードは、プライマリノード及び $m+k-1$ 個のセカンダリノードを含んでもよく、 $m+k$ 個のセカンダリノードを含んでもよい。各ターゲット記憶ノードは、最大で1つのEC blockのみを記憶する。

10

【0051】

特に、記憶ノードグループが $m+k$ 個の記憶ノード（プライマリノード及び $m+k-1$ 個のセカンダリノード）のみを含む場合、プライマリノードは、 $m+k$ 個の記憶ノードがターゲット記憶ノードであると直接決定する。

【0052】

203. 準備メッセージをターゲット記憶ノードに送信する。

【0053】

ターゲット記憶ノードを決定した後に、プライマリノードは、準備メッセージを $m+k$ 個のターゲット記憶ノードに別々に送信し、各準備メッセージは、各ターゲット記憶ノードにより記憶されるべきターゲットEC blockを搬送する。

20

【0054】

セカンダリノードであるターゲット記憶ノードでは、プライマリノードによりターゲット記憶ノードに送信された準備メッセージは、ネットワークを通じて送信されるノード間メッセージである。しかし、プライマリノードもまたターゲット記憶ノードの1つでもよいことが、ステップ202において言及された。プライマリノードがターゲット記憶ノードである場合、このステップにおいて、プライマリノードもまた、準備メッセージをプライマリノード自体に送信するが、プライマリノード自体に送信される準備メッセージは、ネットワークを通過しないノード内メッセージである。

30

【0055】

204. ターゲット記憶ノードにより送信された応答メッセージを受信する。

【0056】

プライマリノードにより送信された準備メッセージを受信した後に、ターゲット記憶ノードは、準備メッセージに従って準備ログを生成し、準備ログは、ターゲット記憶ノードにローカルに記憶される。準備ログの目的は、主に、ターゲット記憶ノードにより記憶されるべきEC blockを一時的に記憶するためのものである。従って、準備ログの記憶位置は、ターゲットEC blockの最終的な記憶位置ではない。実際の適用では、準備ログは、ログ又はデータブロックでもよく、他の形式でもよい。これは、この実施例では限定されないが、準備ログは、ターゲット記憶ノードにより受信された準備メッセージ内のターゲットEC blockを含む必要がある。

40

【0057】

準備ログを生成した後に、ターゲット記憶ノードは、応答メッセージを現在送信しているターゲットセカンダリ記憶ノードが準備メッセージに従って準備ログを生成したことをプライマリノードに通知するために使用される応答メッセージをプライマリノードに送信する。

【0058】

セカンダリノードとしてのターゲット記憶ノードにより送信された応答メッセージは、ネットワークを通じて送信されるノード間メッセージである。しかし、プライマリノードもまたターゲット記憶ノードの1つである場合、このステップにおいて、プライマリノード

50

ドもまた、応答メッセージをプライマリノード自体に送信するが、プライマリノード自体に送信される応答メッセージは、ネットワークを通過しないノード内メッセージである。

【 0 0 5 9 】

プライマリノードは、 $m+k$ 個のターゲット記憶ノード内のいずれか 1 つ以上のターゲット記憶ノードにより送信された応答メッセージを受信し、 $m+k$ 個のターゲット記憶ノード内の全てのターゲット記憶ノードにより送信された応答メッセージを受信した後に、ステップ205を実行する。

【 0 0 6 0 】

205. 実行メッセージをターゲット記憶ノードに送信する。

【 0 0 6 1 】

全てのターゲット記憶ノードにより送信された応答メッセージを受信した後に、プライマリノードは、 $m+k$ 個のターゲット記憶ノードに対して、準備ログに含まれるターゲットEC blockを書き込むように命令するために、実行メッセージを $m+k$ 個のターゲット記憶ノードに送信する。実行メッセージを受信した後に、 $m+k$ 個のターゲット記憶ノードは、準備ログの記憶位置から準備ログを読み取り、準備ログに含まれるターゲットEC blockを、ターゲットEC blockを記憶するために使用されるターゲット記憶ノードの位置に書き込む。

【 0 0 6 2 】

セカンダリノードであるターゲット記憶ノードでは、プライマリノードによりターゲット記憶ノードに送信された実行メッセージは、ネットワークを通じて送信されるノード間メッセージである。しかし、プライマリノードもまたターゲット記憶ノードの 1 つでもよいことが、ステップ202において言及された。プライマリノードがターゲット記憶ノードである場合、このステップにおいて、プライマリノードもまた、実行メッセージをプライマリノード自体に送信するが、プライマリノード自体に送信される実行メッセージは、ネットワークを通過しないノード内メッセージである。

【 0 0 6 3 】

この実施例は、データ記憶方法を提供し、ターゲットECストライプを取得するために、記憶されるべきデータセグメントにおいてECコーディングを実行するステップであり、ターゲットECストライプは、 $m+k$ 個のターゲットEC blockを含むステップと、記憶ノードグループ内のセカンダリ記憶ノードの中で、 $m+k$ 個のターゲットEC blockを記憶するために使用される $m+k$ 個のターゲット記憶ノードを決定するステップと、準備メッセージをターゲット記憶ノードに送信するステップであり、準備メッセージは、ターゲット記憶ノードにより記憶されるべきターゲットEC blockを搬送するステップと、ターゲット記憶ノードにより送信された応答メッセージを受信するステップと、全てのターゲット記憶ノードの応答メッセージを受信した後に、ターゲット記憶ノードに対して、準備ログにあるターゲットEC blockを書き込むように命令するために、実行メッセージを全てのターゲット記憶ノードに送信するステップとを含む。この実施例により提供される方法では、ターゲットECストライプを記憶するために使用される $m+k$ 個のターゲット記憶ノードは、記憶されるべきターゲットEC blockを直接書き込まないが、まず、記憶されるべきターゲットEC blockをローカルに生成された準備ログに書き込み、全てのターゲット記憶ノードが準備ログを生成した後に、ターゲットEC blockを書き込む。全てのターゲット記憶ノードが準備ログを生成した場合にのみ、ターゲット記憶ノードは、ターゲットEC blockを正確な位置に書き込む動作を実行する。ターゲット記憶ノードによりターゲットEC blockを書き込む過程において故障が発生した場合であっても、ターゲットEC blockは、ターゲット記憶ノードの準備ログに記憶されている。このように、ターゲットEC blockを書き込む動作が成功したか否かに拘わらず、ターゲットEC blockが喪失されないことが確保され得る。従って、ターゲット記憶ノードの一部がターゲットECストライプのターゲットEC blockを有するが、ターゲット記憶ノードの他の部分がターゲットECストライプのターゲットEC blockを有さない場合が回避され、ターゲットECストライプが一貫して記憶されることができ、分散型記憶システムの記憶の一貫性が改善される。

10

20

30

40

50

【 0 0 6 4 】

好ましくは、本発明の更に他の実施例では、ターゲット記憶ノードが故障したときに、ターゲット記憶ノードにより生成された準備ログが喪失されないことを確保するために、ターゲット記憶ノードは、準備ログを停電保護媒体に書き込んでよい。

【 0 0 6 5 】

ECストライプのkey値の概念が、本発明のこの実施例に導入される。具体的に、ECストライプのkey値は、ECストライプのデータ部のアドレス範囲に対応し、データ部の同じアドレス範囲を有するECストライプのkey値は同じであるが、データ部の異なるアドレス範囲を有するECストライプのkey値は異なる。key値は、ECストライプの各EC blockで搬送されてもよい。

10

【 0 0 6 6 】

ECストライプのkey値は、データ配送装置によりプライマリノードに配送されたデータセグメントから生じてよい。例えば、分散型記憶システムの論理ボリュームは、複数の論理パーティションに分割されてもよく、各論理パーティションは、k個のデータブロックのサイズである。データ配送装置は、記憶されるべきデータを複数のデータセグメントに分割し、具体的に、1つの論理パーティションに入るデータを1つのデータセグメントとして使用する。分散型記憶システムの各論理パーティションに一意的key値が設定され、各データセグメントが入る論理パーティションに対応するkey値は、データセグメントに対応するkey値である。データ配送装置は、key値をデータセグメントに追加し、プライマリノードにデータセグメントを配送し、データセグメントに従ってプライマリノードにより生成されたECストライプは、データセグメントと同じkey値を有する。

20

【 0 0 6 7 】

key値はデータのアドレス範囲を識別するために使用されてもよいことが、前述の説明から分かる。更に、ターゲット記憶ノードが実行メッセージを受信した後に、ターゲット記憶ノードは、準備ログの記憶位置から準備ログを読み取り、準備ログに含まれるターゲットEC blockを、ターゲットEC blockを記憶するために使用されるターゲット記憶ノードの位置に書き込むことが、ステップ205において言及された。具体的に、ターゲット記憶ノードが準備ログにターゲットEC blockで搬送されたkey値と同じkey値を有するEC blockを記憶していない場合、ターゲット記憶ノードは、新たな記憶位置をターゲットEC blockに割り当てる。ターゲット記憶ノードが準備ログにターゲットEC blockで搬送されたkey値と同じkey値を有するEC blockを記憶している場合、これは、古いデータがターゲットEC blockのアドレス範囲に書き込まれていることを示し、ターゲット記憶ノードは、古いデータを新たなデータで上書きするために、ターゲットEC blockを、ターゲットEC blockのkey値と同じkey値を有するEC blockを記憶するために元々使用された位置に書き込む。新たなデータが古いデータを上書きするため、各記憶ノードにおいて記憶された複数のEC blockのkey値が互いに異なることが認識され得る。

30

【 0 0 6 8 】

ECストライプのバージョン番号の概念が、本発明のこの実施例に更に導入される。具体的に、バージョン番号は、同じkey値を有する2つのECストライプの新しい程度又は古い程度を識別するために使用される。書き込まれた新たなECストライプは古いECストライプを上書きする。例えば、ある時点において、プライマリノードは、ECストライプを生成し、データ部のアドレス範囲は4M~8Mであり、key値は2であり、ECストライプのバージョン番号は1であり、プライマリノードは、記憶のためにバージョン番号1を有するECストライプを6個の記憶ノードに割り当てる。1時間後に、プライマリノードは、他のECストライプを再び生成し、データ部のアドレス範囲は4M~8Mであり、key値は2であり、ECストライプのバージョン番号は2である。プライマリノードは、バージョン番号2を有するECストライプがバージョン番号1を有するECストライプを上書きするように、バージョン番号2を有するECストライプを、バージョン番号1を有するECストライプを記憶するために元々使用された位置に書き込む。ターゲットECストライプのバージョン番号は、ターゲットECストライプの各ターゲットEC blockで搬送されてもよい。

40

50

【 0 0 6 9 】

本発明のこの実施例では、プライマリノードによりターゲット記憶ノードに送信された準備メッセージは、ターゲットECストライプのバージョン番号を更に含み、更に、準備メッセージに従ってターゲット記憶ノードにより生成された準備ログもまた、ターゲットECストライプのバージョン番号を含む。実行メッセージは、 $m+k$ 個のターゲット記憶ノードに対して、準備ログにあるターゲットEC block及びターゲットECストライプのバージョン番号を書き込むように具体的に命令する。バージョン番号の概念が導入された後に、ECストライプの記憶が一貫しているか否かは、バージョン番号に従って決定されてもよい。ターゲットECストライプを記憶するターゲット記憶ノードに書き込まれたターゲットECストライプのバージョン番号が同じである場合、これは、ターゲットECストライプの記憶が一貫していることを示す。ターゲットECストライプを記憶するターゲット記憶ノードに書き込まれたターゲットECストライプのバージョン番号が異なる場合、これは、ターゲットECストライプの記憶が一貫していないことを示す。

10

【 0 0 7 0 】

分散型記憶システムの記憶ノードグループ内の記憶ノードはいずれかの時点で故障する可能性があるため、記憶ノードグループは維持管理される必要がある。記憶ノードグループは、複数の方式で維持管理されてもよく、例えば、更なるノードグループ管理モジュールが配置されてもよい。ノードグループ管理モジュールは、データ配送装置に配置されてもよく、記憶ノードに配置されてもよい。これは、本発明の実施例では限定されない。ノードグループ管理モジュールは、各記憶ノードグループを維持管理することを担う。具体的に、ノードグループ管理モジュールは、各記憶ノードグループに含まれる記憶ノードについての情報を収集し、プライマリノードを決定し、ハートビート接続を使用することにより各記憶ノードの状態をリアルタイムで監視することを担ってもよい。記憶ノードの状態が変化した場合、又はノードグループ内のプライマリノードが変化した場合、ノードグループ管理モジュールは、収集された情報を更新する必要がある。ノードグループ管理モジュールは、各データ配送装置及び記憶ノードが各ノードグループについての情報を習得することができるように、記録するために、収集された情報を各記憶ノードに通知する。ノードグループ管理モジュールにより収集された情報は、簡単に表1の形式で表されてもよい。

20

【 表 1 】

30

表 1

ノードグループ ID	ノード	ノードアイデンティティ	ノード状態
ノードグループ 1	A	プライマリノード	正常
	B	セカンダリノード	正常
	...	...	...
	C	セカンダリノード	故障

【 0 0 7 1 】

表1は、ノードグループ管理モジュールにより収集された情報の単なる表現形式である。ノードグループ管理モジュールは、他の形式で収集された情報を表してもよい。これは、本発明の実施例では限定されない。

40

【 0 0 7 2 】

ステップ202においてターゲット記憶ノードを決定した後に、故障した記憶ノードが $m+k$ 個のターゲット記憶ノードに存在することを、プライマリノードがノードグループ管理モジュールを使用することにより検出した場合、プライマリノードは、データ記憶手順を中断し、故障した記憶ノードを回復させる必要がある。故障した記憶ノードが回復した後に、プライマリノードは、データ記憶手順を続ける必要がある。詳細については、ステップ1~3を参照する。

【 0 0 7 3 】

ステップ1: $m+k$ 個のターゲット記憶ノードにおけるターゲットECストライプのバージョ

50

ン番号が同じであるか否かを決定する。ターゲット記憶ノードが準備ログを生成している場合、ターゲット記憶ノードにおけるターゲットECストライプのバージョン番号は、準備ログ内のターゲットECストライプのバージョン番号でもよい。ターゲット記憶ノードが準備ログを有さない場合、ターゲット記憶ノードにおけるターゲットECストライプのバージョン番号は、ターゲット記憶ノードにより記憶されているターゲットECストライプのバージョン番号でもよい。

【0074】

m+k個のターゲット記憶ノード内の全てのターゲット記憶ノードにおけるターゲットECストライプのバージョン番号が同じである場合、これは、準備ログを有さないターゲット記憶ノードが記憶されるべきターゲットEC blockを書き込む動作を完了したことを示し、ステップ2が実行される。

10

【0075】

m+k個のターゲット記憶ノードにおけるターゲットECストライプのバージョン番号が異なる場合、これは、準備ログを有さないターゲット記憶ノードが準備ログを生成しておらず、更に、記憶されるべきターゲットEC blockを書き込んでないことを示し、ステップ3が実行される。

【0076】

ステップ2：実行メッセージを、準備ログを生成したターゲット記憶ノードに送信し、実行メッセージは、ターゲット記憶ノードに対して、準備ログに含まれるターゲットEC blockを書き込むように命令するために使用される。実行メッセージを受信した後に、準備ログを生成したターゲット記憶ノードは、ターゲットECストライプがターゲット記憶ノードに一貫して記憶されるように、準備ログに含まれるターゲットEC blockを書き込む。

20

【0077】

ステップ3：削除メッセージを、準備ログを生成したターゲット記憶ノードに送信し、削除メッセージは、ターゲット記憶ノードに対して、準備ログを削除するように命令するために使用される。削除メッセージを受信した後に、準備ログを生成したターゲット記憶ノードは、ターゲットECストライプのターゲットEC blockがいずれのターゲット記憶ノードにも書き込まれないように、ローカルに記憶された準備ログを削除する。

【0078】

30

特に、故障した記憶ノードがプライマリノードである場合、ノードグループ管理モジュールは、セカンダリノードを新たなプライマリノードとして指定し、次に、ステップ1~3を再実行してもよい。

【0079】

前述の実施例では、本発明のこの実施例により提供されるデータ記憶方法は、プライマリノードの観点から記載された。以下の実施例では、本発明のこの実施例により提供されるデータ記憶方法は、セカンダリノードの観点から説明される。図3を参照すると、この方法の基本的な手順は以下を含む。

【0080】

301. プライマリノードにより送信された準備メッセージを受信する。

40

【0081】

プライマリノードは、ターゲットECストライプを生成し、ターゲットECストライプを記憶するために使用されるm+k個のターゲット記憶ノードを決定する。本発明のこの実施例におけるセカンダリノードは、ターゲット記憶ノード内のノードである。

【0082】

セカンダリノードは、プライマリノードにより送信された準備メッセージを受信し、準備メッセージは、セカンダリノードがターゲットEC blockを記憶するために使用されるターゲット記憶ノードであるとプライマリノードが決定したときに送信され、準備メッセージは、セカンダリノードにより記憶されるべきターゲットEC blockを含む。

【0083】

50

302. 準備メッセージに従って準備ログを生成する。

【0084】

プライマリノードにより送信された準備メッセージを受信した後に、セカンダリノードは、準備メッセージに従って準備ログを生成し、準備ログは、セカンダリノードにローカルに記憶される。準備ログの目的は、主に、セカンダリノードにより記憶されるべきEC blockを一時的に記憶するためのものである。従って、準備ログの記憶位置は、ターゲットEC blockの最終的な記憶位置ではない。実際の適用では、準備ログは、ログ又はデータブロックでもよく、他の形式でもよい。これは、この実施例では限定されないが、準備ログは、準備メッセージで搬送されたターゲットEC blockを含む必要がある。準備ログは、セカンダリ記憶ノードにより記憶されるべきターゲットEC blockを含む。

10

【0085】

303. 応答メッセージをプライマリノードに送信する。

【0086】

準備ログを生成した後に、セカンダリノードは、応答メッセージをプライマリノードに送信し、応答メッセージは、セカンダリノードが準備メッセージに従って準備ログを生成したことを示すために使用される。

【0087】

記憶ノードグループ内の他のターゲット記憶ノードもまた、応答メッセージをプライマリノードに送信することが認識され得る。

【0088】

20

304. プライマリノードにより送信された実行メッセージを受信する。

【0089】

セカンダリノードは、プライマリノードにより送信された実行メッセージを受信し、実行メッセージは、プライマリノードが全てのターゲット記憶ノードの応答メッセージを受信した後に送信され、実行メッセージは、準備ログに含まれるターゲットEC blockを書き込むように命令するために使用される。

【0090】

305. 準備ログに含まれるターゲットEC blockを書き込む。

【0091】

実行メッセージを受信した後に、セカンダリノードは、準備ログで搬送されたターゲットEC blockを書き込む。

30

【0092】

この実施例では、本発明のこの実施例により提供されるデータ記憶方法は、セカンダリノードの観点から記載された。セカンダリノードは、プライマリノードにより送信された準備メッセージを受信し、準備メッセージに従って準備ログを生成し、応答メッセージをプライマリノードに送信し、プライマリノードにより送信された実行メッセージを受信し、実行メッセージに従って準備ログに含まれるターゲットEC blockを書き込む。この実施例により提供される方法では、セカンダリノードは、記憶されるべきターゲットEC blockを直接書き込まないが、まず、記憶されるべきターゲットEC blockをローカルに生成された準備ログに書き込み、プライマリノードにより送信された実行メッセージを受信した後に、ターゲットEC blockを書き込む。従って、セカンダリノードがターゲットEC blockを書き込む過程において故障した場合であっても、ターゲットEC blockは、セカンダリノードの準備ログに記憶されている。このように、ターゲットEC blockを書き込む動作が成功したか否かに拘わらず、ターゲットEC blockが喪失されないことが確保され得る。プライマリノードにより決定された全てのターゲット記憶ノードにおけるターゲットEC blockが全く喪失されない場合、ターゲット記憶ノードの一部がターゲットECストライプのターゲットEC blockを有するが、ターゲット記憶ノードの他の部分がターゲットECストライプのターゲットEC blockを有さない場合が回避され、ターゲットECストライプが一貫して記憶されることができ、分散型記憶システムの記憶の一貫性が改善される。

40

50

【 0 0 9 3 】

好ましくは、本発明の更に他の実施例では、プライマリノードにより送信されてセカンダリノードにより受信された準備メッセージは、プライマリノードにより生成されたターゲットECストライプのバージョン番号を更に含み、更に、準備メッセージに従ってセカンダリノードにより生成された準備ログもまた、ターゲットECストライプのバージョン番号を含む。セカンダリノードはまた、実行メッセージに従ってターゲットEC blockを書き込むときにターゲットECストライプのバージョン番号を書き込む。

【 0 0 9 4 】

好ましくは、本発明の更に他の実施例では、セカンダリノードが故障したときに、生成された準備ログが喪失されないことを確保するために、セカンダリノードは、準備ログを停電保護媒体に書き込んでよい。

10

【 0 0 9 5 】

図1に示す実施例では、ターゲット記憶ノードにおけるターゲットECストライプのバージョン番号が異なるとプライマリノードが決定した場合、プライマリノードは、削除メッセージを、準備ログを生成したターゲット記憶ノードに送信することが、ステップ3において言及された。従って、好ましくは、本発明の更に他の実施例では、ステップ304の前に、セカンダリノードは、プライマリノードにより送信された削除メッセージを更に受信してもよい。削除メッセージを受信した後に、セカンダリノードは、ローカルに記憶された準備ログを削除する。

20

【 0 0 9 6 】

前述の実施例を理解するのを容易にするために、以下では、説明のための例として前述の実施例の具体的な適用シナリオを使用する。具体的な手順について、図4を参照する。

【 0 0 9 7 】

分散型記憶システムは、ECストライプの形式でデータを記憶し、各ECストライプは、4つのデータブロック及び2つのパリティブロックを含む。分散型記憶システムは、100個の記憶ノードグループを含む。記憶ノードグループ1は6個の記憶ノードを含み、プライマリノードはノードAであり、セカンダリノードはノードB、ノードC、ノードD、ノードE及びノードFである。

【 0 0 9 8 】

401. ターゲットECストライプを生成する。ノードAは、データ配送装置により配送されたデータセグメントを受信し、ターゲットECストライプを取得するために、データセグメントにおいてECコーディングを実行する。ターゲットECストライプのバージョン番号は3であり、ターゲットECストライプのデータ部のアドレス範囲は4M~8Mである。

30

【 0 0 9 9 】

402. ターゲット記憶ノードを決定する。ノードAは、ターゲットECストライプの6個のEC blockを記憶するために使用されるターゲット記憶ノードがノードA~Fであると決定する。

【 0 1 0 0 】

403. 準備メッセージを送信する。ノードAは、準備メッセージをノードA~Fに送信し、各ノードに送信される準備メッセージは、ターゲットEC blockと、ターゲットECストライプのバージョン番号とを搬送し、ノードA~Fは、準備メッセージを受信する。

40

【 0 1 0 1 】

404. 準備ログを生成する。ノードA~Fは、準備ログを生成し、準備ログは、記憶されるべきターゲットEC blockと、ターゲットECストライプのバージョン番号とを含む。

【 0 1 0 2 】

405. 応答メッセージを送信する。ノードA~Fは、準備ログを生成した後に、応答メッセージをノードAに送信する。ノードAは、ノードA~Fにより送信された応答メッセージを受信する。

【 0 1 0 3 】

406. 実行メッセージを送信する。ノードA~Fにより送信された応答メッセージを受信し

50

た後に、ノードAは、実行メッセージをノードA~Fに送信し、ノードA~Fは、実行メッセージを受信する。

【0104】

407. ノードA~Eは、準備ログにあるターゲットEC block及びバージョン番号を書き込み、ノードFは故障している。ノードAが実行メッセージを送信した後に、ノードA~Eは、準備ログにあるターゲットEC block及びバージョン番号を書き込み、ノードFは、実行メッセージを受信する過程において故障している。

【0105】

408. ターゲットECストライプのバージョン番号が同じであると決定する。ノードAは、ノードFが故障していると決定し、従って、ノードFが故障から回復した後に、ノードA~FにおけるターゲットECストライプのバージョン番号が同じであるか否かを決定する。ノードA~Eは準備ログを有さず、記憶されたターゲットEC blockで搬送されたターゲットECストライプのバージョン番号は3である。ノードFは準備ログを有し、準備ログ内のターゲットEC blockで搬送されたターゲットECストライプのバージョン番号は3である。

【0106】

409. ノードAは、実行メッセージをノードFに送信する。ノードFは、実行メッセージを受信する。

【0107】

410. ノードFは、準備ログにあるターゲットEC block及びバージョン番号を書き込む。

【0108】

ターゲット記憶ノードが故障している場合、データ記憶手順は、故障した記憶ノードの故障回復の後に実行される必要があることが、前述の実施例において言及された。しかし、故障は、故障した記憶ノードにおいてデータの喪失を引き起こす可能性がある。この場合、プライマリノードは、記憶ノードグループ内の他の記憶ノードにより記憶されたEC blockに従ったEC技術を使用することによる計算を通じて、故障した記憶ノードの喪失したEC blockを取得し、次に、計算を通じて取得されたEC blockを回復した故障した記憶ノードに対して同期させる必要がある。従って、本発明の実施例は、以下を具体的に含むデータ回復方法を提供する。

【0109】

501. 故障した記憶ノードにより送信された回復メッセージを受信する。

【0110】

故障した記憶ノードは、分散型記憶システム内のいずれかの記憶ノードグループに存在する。故障した記憶ノードが回復した後に、故障した記憶ノードは、回復メッセージをプライマリノードに送信し、回復メッセージは、故障した記憶ノードが回復したことを示すために使用される。プライマリノードは、回復メッセージを受信する。

【0111】

502. 計算を通じて、故障した記憶ノードのN個の喪失したEC blockを取得する。

【0112】

前述のように、分散型記憶システムは、EC技術を使用することにより、ECストライプの形式でデータを記憶し、1つのECストライプは、記憶ノードグループ内のm+k個の記憶ノードにそれぞれ記憶されたm+k個のEC blockを含み、1つのECストライプは、k個のEC blockの喪失を許容可能である。ECストライプAを一例として使用すると、ECストライプAのEC blockが故障した記憶ノードにおいて喪失した場合、プライマリノードは、故障していないノードにより記憶されたECストライプAの他のEC blockに従ったEC技術を使用することによる計算を通じて、故障した記憶ノードの喪失したEC blockを取得してもよい。

【0113】

従って、この実施例では、回復メッセージを受信した後に、プライマリノードは、プライマリノードが属する記憶ノードグループ内の故障していない記憶ノード（すなわち、故障した記憶ノードを除く記憶ノード）により記憶されたEC blockに従った計算を通じて、故障した記憶ノードのN個の喪失したEC blockを取得する。

【 0 1 1 4 】

前述のように、各記憶ノードに記憶された複数のEC blockのkey値は互いに異なり、従って、故障した記憶ノードのN個の喪失したEC blockは、異なるkey値を搬送するべきである。

【 0 1 1 5 】

503. 記憶を実行するために、N個の喪失したEC blockを故障した記憶ノードに順次に送信する。N個の喪失したEC blockを故障した記憶ノードに順次に送信する過程において、プライマリノードが第1のEC blockを生成し、第1のEC blockが故障した記憶ノードにより記憶されるべきであると決定した場合、第1のEC blockと同じkey値を有する第2のEC blockが、故障した記憶ノードにより現在成功して記憶されていない喪失したEC blockに存在するか否かを決定する。決定結果が肯定的である場合、故障した記憶ノードが第2のEC blockを成功して記憶した後に、記憶を実行するために、第1のEC blockを故障した記憶ノードに送信する。

10

【 0 1 1 6 】

計算を通じて故障した記憶ノードのN個の喪失したEC blockを取得した後に、プライマリノードは、N個の喪失したEC blockを送信キューに配置し、N個の喪失したEC blockを故障した記憶ノードに順次に送信する。

【 0 1 1 7 】

N個の喪失したEC blockを故障した記憶ノードに順次に送信する過程において、プライマリノードが新たなECストライプを再び生成し、新たなECストライプ内の第1のEC blockが故障した記憶ノードにより記憶されるべきであると決定した場合、プライマリノードは、第2のEC blockが故障した記憶ノードにより現在成功して記憶されていない喪失したEC blockに存在するか否かを決定し、第2のEC blockは第1のEC blockと同じkey値を有する。

20

【 0 1 1 8 】

プライマリノードが第1のEC blockを生成した後に、記憶を実行するために、第1のEC blockを故障した記憶ノードに直接送信したが、第2のEC blockが故障した記憶ノードにより現在成功して記憶されていない喪失したEC blockに存在する場合、プライマリノードがN個の喪失したEC blockを故障した記憶ノードに順次に送信する動作を実行し続けるときに、第2のEC blockが第1のEC blockのデータを上書きする場合は引き起こされる可能性があることが認識され得る。新たな第1のEC blockを古い第2のEC blockで置換することを回避するために、この実施例では、第2のEC blockが故障した記憶ノードにより現在成功して記憶されていない喪失したEC blockに存在するとプライマリノードが決定した場合、プライマリノードは、送信キューにおいて第2のEC blockの後に第1のEC blockを挿入する、すなわち、プライマリノードは、第1のEC blockを故障した記憶ノードに最初に送信しないが、故障した記憶ノードが第2のEC blockを成功して記憶した後に、記憶を実行するために、第1のEC blockを故障した記憶ノードに送信する。

30

【 0 1 1 9 】

この実施例は、データ回復方法を提供し、故障した記憶ノードが回復した後に、プライマリノードは、計算を通じて故障した記憶ノードのN個の喪失したEC blockを取得し、記憶を実行するために、N個の喪失したEC blockを故障した記憶ノードに順次に送信する。N個の喪失したEC blockを故障した記憶ノードに順次に送信する過程において、プライマリノードが新たなECストライプを再び生成し、新たなECストライプ内の第1のEC blockが故障した記憶ノードにより記憶されるべきであると決定した場合、プライマリノードは、第2のEC blockが故障した記憶ノードにより現在成功して記憶されていない喪失したEC blockに存在するか否かを決定し、第2のEC blockは第1のEC blockと同じkey値を有する。第2のEC blockが存在すると決定された場合、故障した記憶ノードが第2のEC blockを成功して記憶した後に、プライマリノードは、記憶を実行するために、第1のEC blockを故障した記憶ノードに送信する。この方法は、新たなデータを古いデータで置

40

50

換することを回避し、回復の後に故障した記憶ノードに記憶されたEC blockが最新のバージョンであることを確保することができる。

【0120】

ステップ503において、プライマリノードは、記憶を実行するために、N個の喪失したEC blockを故障した記憶ノードに順次に送信する。具体的に、プライマリノードは、まず、N個の喪失したEC block内の第1の喪失したEC blockを故障した記憶ノードに送信し、第nの喪失したEC blockを故障した記憶ノードに送信した後に、プライマリノードは、第nの喪失したEC blockに対応する応答メッセージを受信し、応答メッセージは、故障した記憶ノードにより送信され、故障した記憶ノードが第nの喪失したEC blockを成功して記憶したことを示すために使用される。第nの喪失したEC blockに対応する応答メッセージを受信した後に、プライマリノードは、次の喪失したEC blockを故障した記憶ノードに再び送信し、 $1 \leq n \leq N$ である。プライマリノードがN個の喪失したEC blockを故障した記憶ノードに順次に送信することは、以下のステップにまとめられてもよい。

10

【0121】

ステップ1: プライマリノードは、第nの喪失したEC blockを故障した記憶ノードに送信する。この手順が開始するときのnの初期値は1であり、nはNより大きくない。

【0122】

ステップ2: 第nの喪失したEC blockを故障した記憶ノードに送信した後に、プライマリノードは、第nの喪失したEC blockに対応する応答メッセージを受信する。

【0123】

20

ステップ3: 第nの喪失したEC blockに対応する応答メッセージを受信した後に、 $n < N$ である場合、nの値に1を加算し、ステップ1を再実行する。

【0124】

プライマリノードは、ステップ1～ステップ3を循環的に実行することにより、記憶を実行するために、N個の喪失したEC blockを故障した記憶ノードに送信してもよい。

【0125】

故障した記憶ノードが喪失したEC blockを成功して記憶したことを示すために応答メッセージが使用される場合、ステップ503において、故障した記憶ノードが第2のEC blockを成功して記憶した後に、プライマリノードが、記憶のために、第1のEC blockを故障した記憶ノードに送信することは、具体的に、第2のEC blockを故障した記憶ノードに送信し、第2のEC blockに関して故障した記憶ノードにより送信された応答メッセージを受信した後に、プライマリノードが、記憶を実行するために、第1のEC blockを故障した記憶ノードに送信することであるべきである。

30

【0126】

ステップ503において、プライマリノードは、第2のEC blockがN個の喪失したEC blockの中の故障した記憶ノードにより現在成功して記憶されていない喪失したEC blockに存在するか否かを決定する必要がある。故障した記憶ノードにより現在成功して記憶されていない喪失したEC blockは、プライマリノードにより故障した記憶ノードに現在送信されていない喪失したEC blockと、故障した記憶ノードに現在送信されているが、それに対応する応答メッセージが受信されていない喪失したEC blockとを含むべきであることが認識され得る。

40

【0127】

前述のように、分散型記憶システムの論理ボリュームは、複数の論理パーティションに分割されてもよく、各論理パーティションは、k個のデータブロックのサイズであり、互いに重複しない。分散型記憶システムの各論理パーティションに一意的key値が設定され、各データセグメントが入る論理パーティションに対応するkey値は、データセグメントに対応するkey値である。データ配送装置は、key値をデータセグメントに追加し、プライマリノードにデータセグメントを配送し、ECストライプのkey値は、ECストライプのEC blockで搬送されるが、データセグメントに従ってプライマリノードにより生成されたECストライプは、データセグメントと同じkey値を有する。従って、各EC blockが入る論理パ

50

ーティションのkey値は、EC blockのkey値であることが認識され得る。

【0128】

前述の実施例を理解するのを容易にするために、以下では、説明のために前述の実施例の具体的な適用シナリオを使用する。

【0129】

分散型記憶システムは、ECストライプの形式でデータを記憶し、各ECストライプは、4つのデータブロック及び2つのパリティブロックを含む。分散型記憶システムは、100個の記憶ノードグループを含む。記憶ノードグループ1は6個の記憶ノードを含み、プライマリノードはノードAであり、セカンダリノードはノードB、ノードC、ノードD、ノードE及びノードFである。ある時点において、ノードFが故障する。3時間後に、ノードFは回復するが、データが喪失する。

10

【0130】

ノードFは、ノードFが故障から回復したことをノードAに通知するために、回復メッセージをノードAに送信する。ノードAは、回復メッセージを受信する。

【0131】

回復メッセージを受信した後に、ノードAは、ノードA、ノードB、ノードC、ノードD及びノードEにより記憶された全てのEC blockに従った計算を通じて、ノードFの200個の喪失したEC blockを取得する。

【0132】

プライマリノードは、喪失したEC blockをノードFに毎回送信し、喪失したEC blockに関してノードFから応答メッセージを受信した後に、次の喪失したEC blockを送信する。このように、200個の喪失したEC blockは、ノードFに順次送信される。

20

【0133】

第151の喪失したEC blockがノードFに送信された後に、ノードAは、データ配送装置により配送されたデータセグメントに従って新たなECストライプを生成し、新たなECストライプ内の第1のEC blockがノードFにより記憶されるべきであると決定し、第1のEC blockのkey値は25である。

【0134】

ノードAは、第151の喪失したEC blockに関してノードFにより送信された応答メッセージを受信していないため、ノードAは、key値25を有する第2のEC blockが第151～第200の喪失したEC blockに存在するか否かを決定する。ノードAは、第190の喪失したEC blockのkey値が25であると決定する。

30

【0135】

ノードAは、第1のEC blockをノードFに一時的に送信せず、第190の喪失したEC blockをノードFに送信して、第190の喪失したEC blockに関してノードFにより送信された応答メッセージを受信した後に、第1のEC blockをノードFに送信する。

【0136】

ノードAは、全ての200個の喪失したEC blockをノードFに送信し、ノードFは、データ回復を完了する。

40

【0137】

前述の実施例は、本発明の実施例により提供されるデータ記憶又はデータ回復方法を提供する。前述の方法を実現するための対応する装置は、以下に提供される。

【0138】

本発明の実施例は、データ記憶装置を提供し、データ記憶装置は、いずれかの記憶ノードグループ内のプライマリ記憶ノードに配置され、記憶されるべきデータセグメントを受信し、ターゲットECストライプを取得するために、記憶されるべきデータセグメントにおいてECコーディングを実行するように構成されたデータコーディングモジュール601であり、記憶されるべきデータセグメントのサイズはZより大きくなく、Zはm個のデータブロックのサイズであり、ターゲットECストライプは、m個のターゲットデータブロック及びk個のターゲットパリティブロックを具体的に含むm+k個のターゲットEC blockを含むデー

50

タコーディングモジュール601と、記憶ノードグループ内で、 $m+k$ 個のターゲットEC blockを記憶するために使用される $m+k$ 個のターゲット記憶ノードを決定するように構成されたノード決定モジュール602であり、各ターゲット記憶ノードは、1つのターゲットEC blockを記憶するように構成されるノード決定モジュール602と、準備メッセージを $m+k$ 個のターゲット記憶ノードに別々に送信するように構成されたメッセージ送信モジュール603であり、各準備メッセージは、各ターゲット記憶ノードにより記憶されるべきターゲットEC blockを含むメッセージ送信モジュール603と、 $m+k$ 個のターゲット記憶ノードのいずれかのターゲット記憶ノードにより送信された応答メッセージを受信するように構成されたメッセージ受信モジュール604であり、応答メッセージは、応答メッセージを現在送信しているターゲットセカンダリ記憶ノードが準備メッセージに従って準備ログを生成したことを示すために使用され、準備ログは、応答メッセージを現在送信しているターゲット記憶ノードにより記憶されるべきターゲットEC blockを含むメッセージ受信モジュール604とを具体的に含み、メッセージ送信モジュール603は、メッセージ受信モジュール604が $m+k$ 個のターゲット記憶ノード内の全てのターゲット記憶ノードの応答メッセージを受信した後に、実行メッセージを $m+k$ 個のターゲット記憶ノードに送信するように更に構成され、実行メッセージは、 $m+k$ 個のターゲット記憶ノードに対して、準備ログに含まれるターゲットEC blockを書き込むように命令するために使用される。

【0139】

この実施例により提供されるデータ記憶装置では、データコーディングモジュール601は、ターゲットECストライプを取得するために、記憶されるべきデータセグメントにおいてECコーディングを実行し、ノード決定モジュール602は、記憶ノードグループ内のセカンダリ記憶ノードの中で、 $m+k$ 個のターゲットEC blockを記憶するために使用される $m+k$ 個のターゲット記憶ノードを決定し、メッセージ送信モジュール603は、準備メッセージをターゲット記憶ノードに送信し、メッセージ受信モジュール604は、ターゲット記憶ノードにより送信された応答メッセージを受信し、メッセージ送信モジュール603は、全てのターゲット記憶ノードの応答メッセージを受信した後に、ターゲット記憶ノードに対して、準備ログにあるターゲットEC blockを書き込むように命令するために、実行メッセージを全てのターゲット記憶ノードに送信する。この実施例により提供されるデータ記憶装置を使用することにより、ターゲットECストライプを記憶するために使用される $m+k$ 個のターゲット記憶ノードは、記憶されるべきターゲットEC blockを直接書き込まないが、まず、記憶されるべきターゲットEC blockをローカルに生成された準備ログに書き込み、全てのターゲット記憶ノードが準備ログを生成した後に、ターゲットEC blockを書き込む。全てのターゲット記憶ノードが準備ログを生成した場合にのみ、ターゲット記憶ノードは、ターゲットEC blockを正確な位置に書き込む動作を実行する。ターゲット記憶ノードによりターゲットEC blockを書き込む過程において故障が発生した場合であっても、ターゲットEC blockは、ターゲット記憶ノードの準備ログに記憶されている。このように、ターゲットEC blockを書き込む動作が成功したか否かに拘わらず、ターゲットEC blockが喪失されないことが確保され得る。従って、ターゲット記憶ノードの一部がターゲットECストライプのターゲットEC blockを有するが、ターゲット記憶ノードの他の部分がターゲットECストライプのターゲットEC blockを有さない場合が回避され、ターゲットECストライプが一貫して記憶されることができ、分散型記憶システムの記憶の一貫性が改善される。

【0140】

好ましくは、本発明の更に他の実施例では、ターゲット記憶ノードが故障したときに、ターゲット記憶ノードにより生成された準備ログが喪失されないことを確保するために、ターゲット記憶ノードは、準備ログを停電保護媒体に書き込んでよい。

【0141】

好ましくは、本発明の更に他の実施例では、プライマリノードによりターゲット記憶ノードに送信された準備メッセージは、ターゲットECストライプのバージョン番号を更に含み、更に、準備メッセージに従ってターゲット記憶ノードにより生成された準備ログもま

10

20

30

40

50

た、ターゲットECストライプのバージョン番号を含む。実行メッセージは、 $m+k$ 個のターゲット記憶ノードに対して、準備ログにあるターゲットEC block及びターゲットECストライプのバージョン番号を書き込むように具体的に命令する。

【0142】

好ましくは、本発明の更に他の実施例では、データ記憶装置は、故障した記憶ノードが $m+k$ 個のターゲット記憶ノードに存在することが検出された場合、故障した記憶ノードの故障回復の後に、 $m+k$ 個のターゲット記憶ノードにおけるターゲットECストライプのバージョン番号が同じであるか否かを決定するように構成されたバージョン決定モジュール605であり、 $m+k$ 個のターゲット記憶ノードにおけるターゲットECストライプのバージョン番号は、準備ログを生成したターゲット記憶ノードにおける準備ログ内のターゲットECストライプのバージョン番号及び/又は準備ログを有さないターゲット記憶ノードにより記憶されたターゲットECストライプのバージョン番号を含むバージョン決定モジュール605を更に含んでもよい。メッセージ送信モジュール603は、 $m+k$ 個のターゲット記憶ノード内の全てのターゲット記憶ノードにおけるターゲットECストライプのバージョン番号が同じである場合、実行メッセージを、準備ログを生成したターゲット記憶ノードに送信し、 $m+k$ 個のターゲット記憶ノード内のターゲット記憶ノードにおけるターゲットECストライプのバージョン番号が異なる場合、削除メッセージを、準備ログを生成したターゲット記憶ノードに送信するように具体的に構成され、削除メッセージは、ターゲット記憶ノードに対して、準備ログを削除するように命令するために使用される。

【0143】

本発明の実施例は、セカンダリノードに配置されるデータ記憶装置を更に提供する。図7を参照すると、データ記憶装置の構成は、プライマリノードにより送信された準備メッセージを受信するように構成されたメッセージ受信モジュール701であり、準備メッセージは、データ記憶装置が位置するセカンダリ記憶ノードにより記憶されるべきターゲットEC blockを含むメッセージ受信モジュール701と、準備メッセージに従って準備ログを生成するように構成されたログ生成モジュール702であり、準備ログは、セカンダリ記憶ノードにより記憶されるべきターゲットEC blockを含むログ生成モジュール702と、応答メッセージをプライマリ記憶ノードに送信するように構成されたメッセージ送信モジュール703であり、応答メッセージは、セカンダリ記憶ノードが準備メッセージに従って準備ログを生成したことを示すために使用されるメッセージ送信モジュール703とを含み、メッセージ受信モジュール701は、プライマリ記憶ノードにより送信された実行メッセージを受信するように更に構成され、実行メッセージに従って準備ログに含まれるターゲットEC blockを書き込むように構成されたデータ書き込みモジュール704を含む。

【0144】

この実施例により提供されるデータ記憶装置では、メッセージ受信モジュール701は、プライマリノードにより送信された準備メッセージを受信し、ログ生成モジュール702は、準備メッセージに従って準備ログを生成し、メッセージ送信モジュール703は、応答メッセージをプライマリノードに送信し、メッセージ受信モジュール701は、プライマリノードにより送信された実行メッセージを受信し、データ書き込みモジュール704は、実行メッセージに従って準備ログに含まれるターゲットEC blockを書き込む。この実施例により提供されるデータ記憶装置は、記憶されるべきターゲットEC blockを直接書き込まないが、まず、記憶されるべきターゲットEC blockをローカルに生成された準備ログに書き込み、プライマリノードにより送信された実行メッセージを受信した後に、ターゲットEC blockを書き込む。従って、セカンダリノードがターゲットEC blockを書き込む過程において故障した場合であっても、ターゲットEC blockは、セカンダリノードの準備ログに記憶されている。このように、ターゲットEC blockを書き込む動作が成功したか否かに拘わらず、ターゲットEC blockが喪失されないことが確保され得る。プライマリノードにより決定された全てのターゲット記憶ノードにおけるターゲットEC blockが全く喪失されない場合、ターゲット記憶ノードの一部がターゲットECストライプのターゲットEC blockを有するが、ターゲット記憶ノードの他の部分がターゲットECストライプの

ターゲットEC blockを有さない場合が回避され、ターゲットECストライプが一貫して記憶されることができ、分散型記憶システムの記憶の一貫性が改善される。

【0145】

好ましくは、本発明の更に他の実施例では、セカンダリ記憶ノードにより記憶されるべきターゲットEC blockは、プライマリ記憶ノードにより生成されたターゲットECストライプ内のデータブロック又はパリティブロックであり、準備メッセージは、プライマリノードにより生成されたターゲットECストライプのバージョン番号を更に含み、更に、準備メッセージに従ってセカンダリノードにより生成された準備ログもまた、ターゲットECストライプのバージョン番号を含む。セカンダリノードはまた、実行メッセージに従ってターゲットEC blockを書き込むときにターゲットECストライプのバージョン番号を書き込む。

10

【0146】

好ましくは、本発明の更に他の実施例では、メッセージ受信モジュール701は、プライマリ記憶ノードにより送信された削除メッセージを受信するように更に構成され、データ記憶装置は、削除メッセージに従って準備ログを削除するように構成されたデータ削除モジュール705を更に含んでもよい。

【0147】

記憶ノードのアイデンティティがプライマリノードであるかセカンダリノードであるかは、自由に設定されてもよく、1つの記憶ノードは、記憶ノードグループ内でプライマリノードあり、また、他の記憶ノードグループ内でセカンダリノードでもよい。従って、図6及び図7に示すデータ記憶装置は、同じ記憶ノードに位置してもよい。

20

【0148】

前述の実施例を理解するのを容易にするために、以下では、説明のための例として前述の実施例の具体的な適用シナリオを使用する。具体的な手順について、図4を参照する。

【0149】

分散型記憶システムは、ECストライプの形式でデータを記憶し、各ECストライプは、4つのデータブロック及び2つのパリティブロックを含む。分散型記憶システムは、100個の記憶ノードグループを含む。記憶ノードグループ1は6個の記憶ノードを含み、プライマリノードはノードAであり、セカンダリノードはノードB、ノードC、ノードD、ノードE及びノードFである。

30

【0150】

401. ターゲットECストライプを生成する。ノードAは、データ配送装置により配送されたデータセグメントを受信し、ノードAのデータコーディングモジュールは、ターゲットECストライプを取得するために、データセグメントにおいてECコーディングを実行する。ターゲットECストライプのバージョン番号は3であり、ターゲットECストライプのデータ部のアドレス範囲は4M~8Mである。

【0151】

402. ターゲット記憶ノードを決定する。ノードAのノード決定モジュールは、ターゲットECストライプの6個のEC blockを記憶するために使用されるターゲット記憶ノードがノードA~Fであると決定する。

40

【0152】

403. 準備メッセージを送信する。ノードAのメッセージ送信モジュールは、準備メッセージをノードA~Fに送信し、各ノードに送信される準備メッセージは、ターゲットEC blockと、ターゲットECストライプのバージョン番号とを搬送し、ノードA~Fのメッセージ受信モジュールは、準備メッセージを受信する。

【0153】

404. 準備ログを生成する。ノードA~Fのログ生成モジュールは、準備ログを生成し、準備ログは、記憶されるべきターゲットEC blockと、ターゲットECストライプのバージョン番号とを含む。

【0154】

50

405. 応答メッセージを送信する。ノードA~Fのメッセージ送信モジュールは、準備ログを生成した後に、応答メッセージをノードAに送信する。ノードAのメッセージ受信モジュールは、ノードA~Fにより送信された応答メッセージを受信する。

【0155】

406. 実行メッセージをノードA~Fに送信する。ノードA~Fにより送信された応答メッセージを受信した後に、ノードAのメッセージ送信モジュールは、実行メッセージをノードA~Fに送信し、ノードA~Fのメッセージ受信モジュールは、実行メッセージを受信する。

【0156】

407. ノードA~Eは、準備ログにあるターゲットEC block及びバージョン番号を書き込み、ノードFは故障している。ノードAが実行メッセージを送信した後に、ノードA~Eのデータ書き込みモジュールは、準備ログにあるターゲットEC block及びバージョン番号を書き込み、ノードFは、実行メッセージを受信する過程において故障している。

10

【0157】

408. ターゲットECストライプのバージョン番号が同じであると決定する。ノードAは、ノードFが故障していると決定し、従って、ノードFが故障から回復した後に、ノードAのバージョン決定モジュールは、ノードA~FにおけるターゲットECストライプのバージョン番号が同じであるか否かを決定する。ノードA~Eは準備ログを有さず、記憶されたターゲットEC blockで搬送されたターゲットECストライプのバージョン番号は3である。ノードFは準備ログを有し、準備ログ内のターゲットEC blockで搬送されたターゲットECストライプのバージョン番号は3である。

20

【0158】

409. ノードAのメッセージ送信モジュールは、実行メッセージをノードFに送信する。ノードFのメッセージ受信モジュールは、実行メッセージを受信する。

【0159】

410. ノードFのデータ書き込みモジュールは、準備ログにあるターゲットEC block及びバージョン番号を書き込む。

【0160】

本発明の実施例は、故障した記憶ノードの喪失したデータを回復するように構成されたデータ回復装置を更に提供し、データ回復装置は、故障した記憶ノードが属する記憶ノードグループのプライマリ記憶ノードに配置される。図8を参照すると、データ回復装置の具体的な構成は、故障した記憶ノードの故障回復の後に、故障した記憶ノードにより送信された回復メッセージを受信するように構成された回復メッセージモジュール801であり、回復メッセージは、故障した記憶ノードが故障から回復したことを示すために使用される回復メッセージモジュール801と、故障した記憶ノードが属する記憶ノードグループ内の故障していない記憶ノードにより記憶されたEC blockに従った計算を通じて、故障した記憶ノードのN個の喪失したEC blockを取得するように構成されたデータ計算モジュール802であり、Nは正の整数であるデータ計算モジュール802と、記憶を実行するために、N個の喪失したEC blockを故障した記憶ノードに順次に送信するように構成されたデータ送信モジュール803と、N個の喪失したEC blockを故障した記憶ノードに順次に送信する過程において、プライマリ記憶ノードが第1のEC blockを生成し、第1のEC blockが故障した記憶ノードにより記憶されるべきであると決定した場合、第1のEC blockと同じkey値を有する第2のEC blockが、故障した記憶ノードにより現在成功して記憶されていない喪失したEC blockに存在するか否かを決定するように構成されたkey値決定モジュール804とを含み、データ送信モジュール803は、第2のEC blockが、故障した記憶ノードにより現在成功して記憶されていない喪失したEC blockに存在する場合、故障した記憶ノードが第2のEC blockを成功して記憶した後に、記憶を実行するために、第1のEC blockを故障した記憶ノードに送信するように更に構成される。

30

40

【0161】

この実施例は、データ回復装置を提供し、データ計算モジュール802は、故障した記憶ノードが回復した後に、計算を通じて故障した記憶ノードのN個の喪失したEC blockを取

50

得し、データ送信モジュール803は、記憶を実行するために、N個の喪失したEC blockを故障した記憶ノードに順次に送信する。N個の喪失したEC blockを故障した記憶ノードに順次に送信する過程において、プライマリノードが新たなECストライプを生成し、新たなECストライプ内の第1のEC blockが故障した記憶ノードにより記憶されるべきであると決定した場合、key値決定モジュール804は、第2のEC blockが故障した記憶ノードにより現在成功して記憶されていない喪失したEC blockに存在するか否かを決定し、第2のEC blockは第1のEC blockと同じkey値を有する。第2のEC blockが存在すると決定された場合、故障した記憶ノードが第2のEC blockを成功して記憶した後に、データ送信モジュール803は、記憶を実行するために、第1のEC blockを故障した記憶ノードに送信する。この方法は、新たなデータを古いデータで置換することを回避し、回復の後に故障した記憶ノードに記憶されたEC blockが最新のバージョンであることを確保することができる。

10

【0162】

好ましくは、本発明の更に他の実施例では、データ送信モジュール803は、第nの喪失したEC blockを故障した記憶ノードに送信するように具体的に構成され、 $1 \leq n \leq N$ であり、第nの喪失したEC blockに対応する応答メッセージを受信するように具体的に構成され、応答メッセージは、故障した記憶ノードにより送信され、応答メッセージは、故障した記憶ノードが第nの喪失したEC blockを成功して記憶したことを示すために使用され、第nの喪失したEC blockに関して故障した記憶ノードの応答メッセージを受信した後に、 $n < N$ である場合、nの値に1を加算し、第nの喪失したEC blockを故障した記憶ノードに送信するステップを再実行するように具体的に構成される。

20

【0163】

好ましくは、本発明の更に他の実施例では、現在成功して記憶されていない喪失したEC blockは、故障した記憶ノードに現在送信されていない喪失したEC blockと、故障した記憶ノードに現在送信されているが、それに対応する応答メッセージが受信されていない喪失したEC blockとを含む。

【0164】

好ましくは、本発明の更に他の実施例では、分散型記憶システムの論理ボリュームは、複数の論理パーティションを含み、各論理パーティションは、k個のデータブロックのサイズであり、互いに重複せず、各論理パーティションは、1つのkey値に一意に対応し、各EC blockが入る論理パーティションのkey値は、EC blockのkey値である。

30

【0165】

前述の実施例を理解するのを容易にするために、以下では、説明のために前述の実施例の具体的な適用シナリオを使用する。

【0166】

分散型記憶システムは、ECストライプの形式でデータを記憶し、各ECストライプは、4つのデータブロック及び2つのパリティブロックを含む。分散型記憶システムは、100個の記憶ノードグループを含む。記憶ノードグループ1は6個の記憶ノードを含み、プライマリノードはノードAであり、セカンダリノードはノードB、ノードC、ノードD、ノードE及びノードFである。ある時点において、ノードFが故障する。3時間後に、ノードFは回復するが、データが喪失する。

40

【0167】

ノードFは、ノードFが故障から回復したことをノードAに通知するために、回復メッセージをノードAに送信する。ノードAの回復メッセージモジュール801は、回復メッセージを受信する。

【0168】

ノードAが回復メッセージを受信した後に、データ計算モジュール802は、ノードA、ノードB、ノードC、ノードD及びノードEにより記憶された全てのEC blockに従った計算を通じて、ノードFの200個の喪失したEC blockを取得する。

【0169】

50

ノードAのデータ送信モジュール803は、喪失したEC blockをノードFに毎回送信し、喪失したEC blockに関してノードFから応答メッセージを受信した後に、次の喪失したEC blockを送信する。このように、200個の喪失したEC blockは、ノードFに順次に送信される。

【0170】

データ送信モジュール803が第151の喪失したEC blockをノードFに送信した後に、ノードAは、データ配送装置により配送されたデータセグメントに従って新たなECストライプを生成し、新たなECストライプ内の第1のEC blockがノードFにより記憶されるべきであると決定し、第1のEC blockのkey値は25である。

【0171】

ノードAは、第151の喪失したEC blockに関してノードFにより送信された応答メッセージを受信していないため、ノードAのkey値決定モジュール804は、key値25を有する第2のEC blockが第151～第200の喪失したEC blockに存在するか否かを決定する。ノードAは、第190の喪失したEC blockのkey値が25であると決定する。

【0172】

ノードAのデータ送信モジュール803は、第1のEC blockをノードFに一時的に送信せず、第190の喪失したEC blockをノードFに送信して、第190の喪失したEC blockに関してノードFにより送信された応答メッセージを受信した後に、第1のEC blockをノードFに送信する。

【0173】

ノードAのデータ送信モジュール803は、全ての200個の喪失したEC blockをノードFに送信し、ノードFは、データ回復を完了する。

【0174】

本発明の実施例における各装置は、モジュール機能エンティティの観点から前述された。以下では、ハードウェア処理の観点から本発明の実施例における装置を説明する。

【0175】

図9を参照すると、本発明の実施例により提供されるデータ記憶装置の他の実施例は、入力装置901と、出力装置902と、プロセッサ903と、メモリ904とを含む（データ記憶装置900は、1つ以上のプロセッサ903を含んでもよく、図9では、1つのプロセッサ903が一例として使用される）。本発明の或る実施例では、入力装置901、出力装置902、プロセッサ903及びメモリ904は、バスにより或いは他の方式で接続されてもよく、バスによる接続が図9において一例として使用される。

【0176】

メモリ904に記憶された動作命令を呼び出すことにより、プロセッサ903は、以下のステップ、すなわち、記憶されるべきデータセグメントを受信し、ターゲットECストライプを取得するために、記憶されるべきデータセグメントにおいてECコーディングを実行するステップであり、記憶されるべきデータセグメントのサイズはZより大きくなく、Zはm個のデータブロックのサイズであり、ターゲットECストライプは、m+k個のターゲットEC blockを含み、m+k個のターゲットEC blockは、m個のターゲットデータブロック及びk個のターゲットパリティブロックを含むステップと、m+k個のターゲット記憶ノードを決定するステップであり、m+k個のターゲット記憶ノードは、m+k個のターゲットEC blockを記憶するように構成され、各ターゲット記憶ノードは、ターゲットEC blockの1つを記憶するステップと、準備メッセージをm+k個のターゲット記憶ノードに別々に送信するステップであり、各準備メッセージは、各ターゲット記憶ノードにより記憶されるべきターゲットEC blockを含むステップと、m+k個のターゲット記憶ノードのいずれかのターゲット記憶ノードにより送信された応答メッセージを受信するステップであり、応答メッセージは、応答メッセージを現在送信しているターゲット記憶ノードが準備メッセージに従って準備ログを生成したことを示すために使用され、準備ログは、応答メッセージを現在送信しているターゲット記憶ノードにより記憶されるべきターゲットEC blockを含むステップと、m+k個のターゲット記憶ノード内の全てのターゲット記憶ノードの応答メッセージを

10

20

30

40

50

受信した後に、実行メッセージを $m+k$ 個のターゲット記憶ノードに送信するステップであり、実行メッセージは、 $m+k$ 個のターゲット記憶ノードに対して、準備ログに含まれるターゲットEC blockを書き込むように命令するために使用されるステップとを実行するように構成される。

【0177】

本発明の或る実施例では、準備メッセージは、ターゲットECストライプのバージョン番号を更に含み、準備メッセージに従ってターゲットセカンダリ記憶ノードにより生成された準備ログは、ターゲットECストライプのバージョン番号を更に含み、実行メッセージは、ターゲット記憶ノードに対して、準備メッセージにあるターゲットEC block及びターゲットECストライプのバージョン番号を書き込むように命令するために具体的に使用される。

10

【0178】

本発明の或る実施例では、プロセッサ903は、以下のステップ、すなわち、故障した記憶ノードが $m+k$ 個のターゲット記憶ノードに存在することが検出された場合、故障した記憶ノードの故障回復の後に、 $m+k$ 個のターゲット記憶ノードにおけるターゲットECストライプのバージョン番号が同じであるか否かを決定するステップであり、 $m+k$ 個のターゲット記憶ノードにおけるターゲットECストライプのバージョン番号は、準備ログを生成したターゲット記憶ノードにおける準備ログ内のターゲットECストライプのバージョン番号及び/又は準備ログを有さないターゲット記憶ノードにより記憶されたターゲットECストライプのバージョン番号を含むステップと、 $m+k$ 個のターゲット記憶ノード内の全てのターゲット記憶ノードにおけるターゲットECストライプのバージョン番号が同じである場合、実行メッセージを、準備ログを生成したターゲット記憶ノードに送信するステップであり、実行メッセージは、ターゲット記憶ノードに対して、準備ログに含まれるターゲットEC blockを書き込むように命令するために使用されるステップとを更に実行する。

20

【0179】

本発明の或る実施例では、プロセッサ903は、以下のステップ、すなわち、 $m+k$ 個のターゲット記憶ノード内のターゲット記憶ノードにおけるターゲットECストライプのバージョン番号が異なる場合、削除メッセージを、準備ログを生成したターゲット記憶ノードに送信するステップであり、削除メッセージは、ターゲット記憶ノードに対して、準備ログを削除するように命令するために使用されるステップとを更に実行する。

30

【0180】

図10を参照すると、本発明の実施例により提供されるデータ記憶装置の他の実施例は、入力装置1001と、出力装置1002と、プロセッサ1003と、メモリ1004とを含む(データ記憶装置1000は、1つ以上のプロセッサ1003を含んでもよく、図10では、1つのプロセッサ1003が一例として使用される)。本発明の或る実施例では、入力装置1001、出力装置1002、プロセッサ1003及びメモリ1004は、バスにより或いは他の方式で接続されてもよく、バスによる接続が図10において一例として使用される。

【0181】

メモリ1004に記憶された動作命令を呼び出すことにより、プロセッサ1003は、以下のステップ、すなわち、プライマリ記憶ノードにより送信された準備メッセージを受信するステップであり、準備メッセージは、セカンダリ記憶ノードにより記憶されるべきターゲットEC blockを含むステップと、準備メッセージに従って準備ログを生成するステップであり、準備ログは、セカンダリ記憶ノードにより記憶されるべきターゲットEC blockを含むステップと、応答メッセージをプライマリ記憶ノードに送信するステップであり、応答メッセージは、セカンダリ記憶ノードが準備メッセージに従って準備ログを生成したことを示すために使用されるステップと、プライマリ記憶ノードにより送信された実行メッセージを受信するステップと、実行メッセージに従って準備ログに含まれるターゲットEC blockを書き込むステップとを実行するように構成される。

40

【0182】

本発明の或る実施例では、セカンダリ記憶ノードにより記憶されるべきターゲットEC

50

blockは、プライマリ記憶ノードにより生成されたターゲットECストライプ内のデータブロック又はパリティブロックであり、準備メッセージは、ターゲットECストライプのバージョン番号を更に含み、準備メッセージに従ってセカンダリ記憶ノードにより生成された準備ログは、ターゲットECストライプのバージョン番号を更に含み、実行メッセージは、ターゲット記憶ノードに対して、準備ログにあるターゲットEC block及びターゲットECストライプのバージョン番号を書き込むように命令するために具体的に使用される。

【0183】

本発明の或る実施例では、プロセッサ1003は、以下のステップ、すなわち、プライマリ記憶ノードにより送信された削除メッセージを受信するステップと、削除メッセージに従って準備ログを削除するステップとを更に実行する。

10

【0184】

図11を参照すると、本発明の実施例により提供されるデータ回復装置の他の実施例は、入力装置1101と、出力装置1102と、プロセッサ1103と、メモリ1104とを含む（データ回復装置1100は、1つ以上のプロセッサ1103を含んでもよく、図11では、1つのプロセッサ1103が一例として使用される）。本発明の或る実施例では、入力装置1101、出力装置1102、プロセッサ1103及びメモリ1104は、バスにより或いは他の方式で接続されてもよく、バスによる接続が図11において一例として使用される。

【0185】

メモリ1104に記憶された動作命令を呼び出すことにより、プロセッサ1103は、以下のステップ、すなわち、故障した記憶ノードの故障回復の後に、故障した記憶ノードにより送信された回復メッセージを受信するステップと、故障した記憶ノードが属する記憶ノードグループ内の故障していない記憶ノードにより記憶されたEC blockに従った計算を通じて、故障した記憶ノードのN個の喪失したEC blockを取得するステップであり、Nは正の整数であるステップと、記憶を実行するために、N個の喪失したEC blockを故障した記憶ノードに順次に送信するステップと、N個の喪失したEC blockを故障した記憶ノードに順次に送信する過程において、プライマリ記憶ノードが第1のEC blockを生成し、第1のEC blockが故障した記憶ノードにより記憶されるべきであると決定した場合、第1のEC blockと同じkey値を有する第2のEC blockが、故障した記憶ノードにより現在成功して記憶されていない喪失したEC blockに存在するか否かを決定するステップと、第2のEC blockが存在すると決定された場合、故障した記憶ノードが第2のEC blockを成功して記憶した後に、記憶を実行するために、第1のEC blockを故障した記憶ノードに送信するステップとを実行するように構成される。

20

30

【0186】

本発明の或る実施例では、プロセッサ1103は、以下のステップ、すなわち、第nの喪失したEC blockを故障した記憶ノードに送信するステップであり、 $1 \leq n \leq N$ であるステップと、第nの喪失したEC blockに対応する応答メッセージを受信するステップであり、応答メッセージは、故障した記憶ノードにより送信され、応答メッセージは、故障した記憶ノードが第nの喪失したEC blockを成功して記憶したことを示すために使用されるステップと、第nの喪失したEC blockに関して故障した記憶ノードの応答メッセージを受信した後に、 $n < N$ である場合、nの値に1を加算し、第nの喪失したEC blockを故障した記憶ノードに送信するステップを再実行するステップとを更に実行する。

40

【0187】

本発明の或る実施例では、現在成功して記憶されていない喪失したEC blockは、故障した記憶ノードに現在送信されていない喪失したEC blockと、故障した記憶ノードに現在送信されているが、それに対応する応答メッセージが受信されていない喪失したEC blockとを含む。

【0188】

本発明の或る実施例では、分散型記憶システムの論理ボリュームは、複数の論理パーティションを含み、各論理パーティションは、k個のデータブロックのサイズであり、互いに重複せず、各論理パーティションは、1つのkey値に一意に対応し、各EC blockが入る

50

論理パーティションのkey値は、EC blockのkey値である。

【0189】

本発明の実施例は、図6～図11のいずれか1つ以上に示す装置を含む記憶ノードを更に提供する。

【0190】

便利且つ簡潔な説明の目的で、前述のシステム、装置及びユニットの詳細な動作処理について、前述の方法の実施例における対応する処理に参照が行われてもよく、詳細はここでは再び説明しないことが当業者により明確に認識され得る。

【0191】

本出願で提供される複数の実施例では、開示のシステム、装置及び方法は、他の方式で実現されてもよいことが認識されるべきである。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、他のシステムに統合されてもよく、いくつかの機能が無視されてもよく実行されなくてもよい。更に、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通じて実現されてもよい。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、機械的又は他の形式で実現されてもよい。

【0192】

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別々でなくてもよく、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物理的なユニットでなくてもよく、1つの位置に存在してもよく、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、実施例の対策の目的を達成するために、実際のニーズに従って選択されてもよい。

【0193】

更に、本発明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、ユニットのそれぞれが物理的に単独で存在してもよく、2つ以上のユニットが1つのユニットに統合されてもよい。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、ソフトウェアの機能ユニットの形式で実現されてもよい。

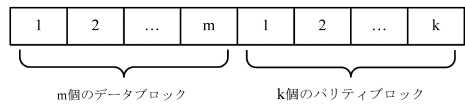
【0194】

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、独立したプロダクトとして販売又は使用される場合、統合されたユニットは、コンピュータ読み取り可能記憶媒体に記憶されてもよい。このような理解に基づいて、基本的に本発明の技術的対策若しくは従来技術に寄与する部分、又は技術的対策の一部若しくは全部は、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、記憶媒体に記憶され、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等でもよい）に対して本発明の実施例に記載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複数の命令を含む。前述の記憶媒体は、USBフラッシュドライブ、取り外し可能ハードディスク、読み取り専用メモリ（ROM、Read-Only Memory）、ランダムアクセスメモリ（RAM、Random Access Memory）、磁気ディスク又は光ディスクのようなプログラムコードを記憶することができるいずれかの媒体を含む。

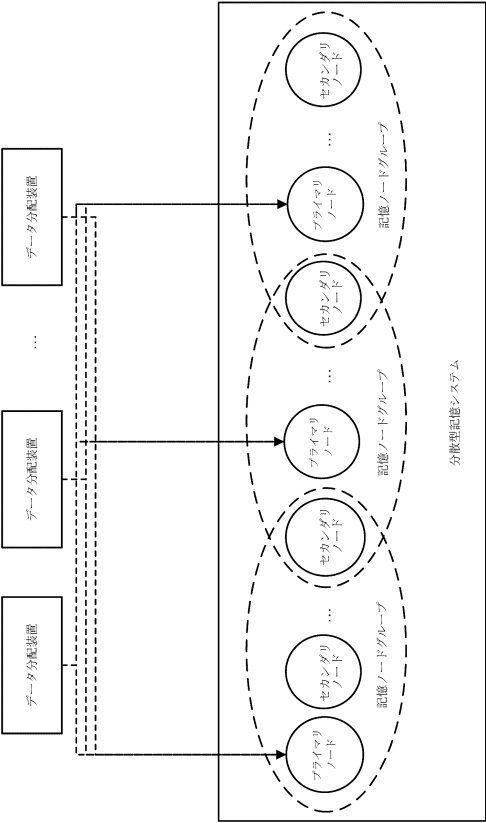
【0195】

前述の実施例は、単に本発明の技術的対策を説明することを意図するものであり、本発明を限定することを意図するものではない。前述の実施例を参照して本発明について詳細に説明したが、当業者は、本発明の実施例の技術的対策の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対策に変更を行ってもよく、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認識するべきである。

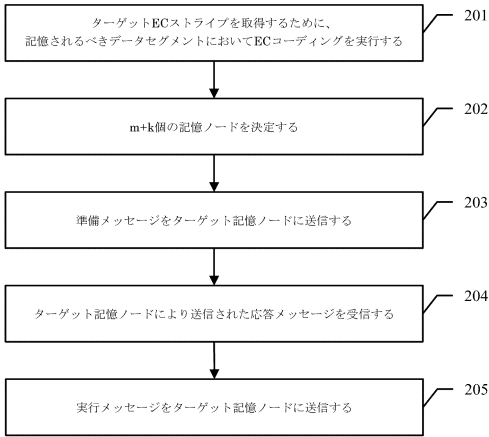
【図 1 (a) 】



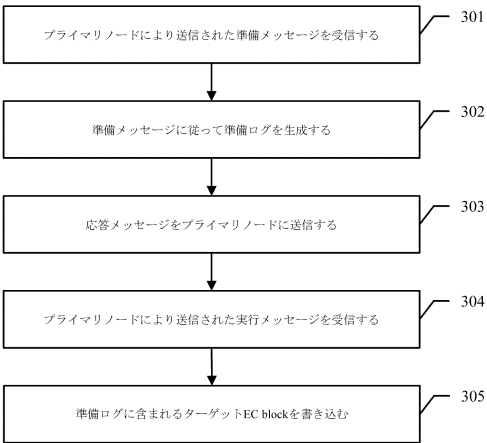
【図 1 (b) 】



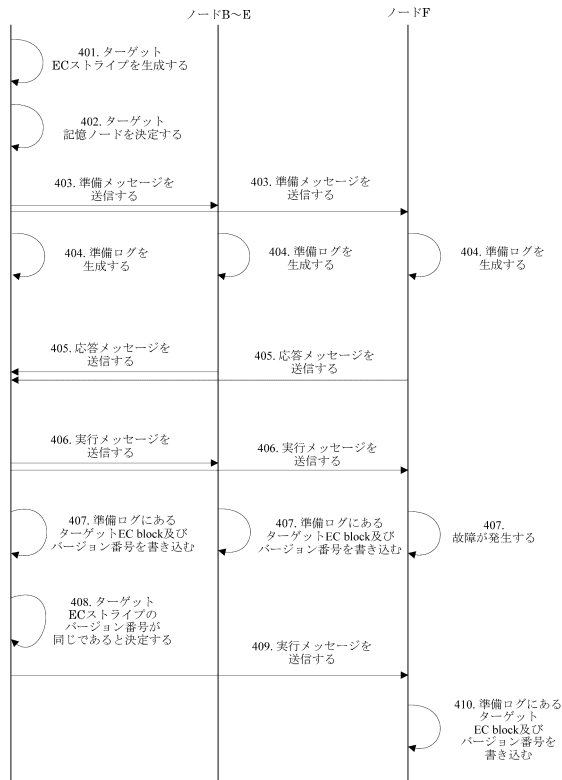
【図 2 】



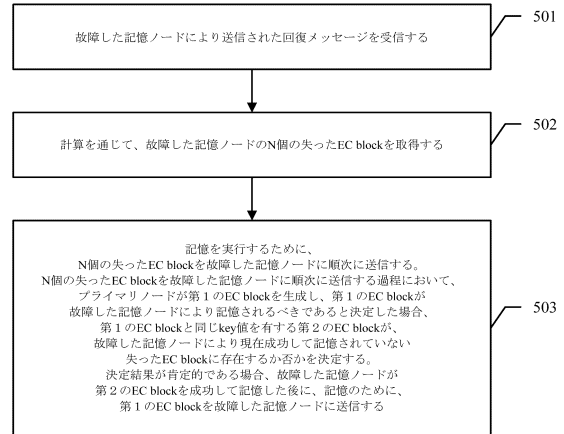
【図 3 】



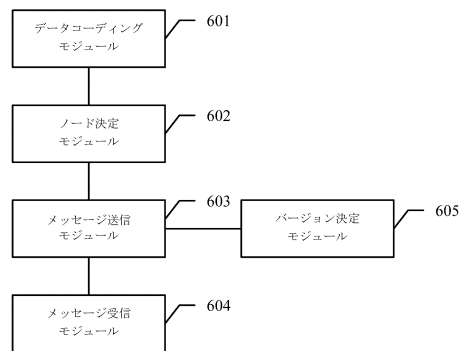
【図 4】



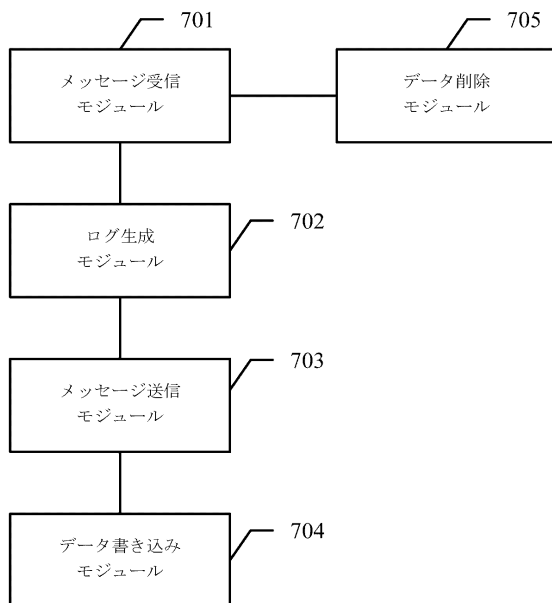
【図 5】



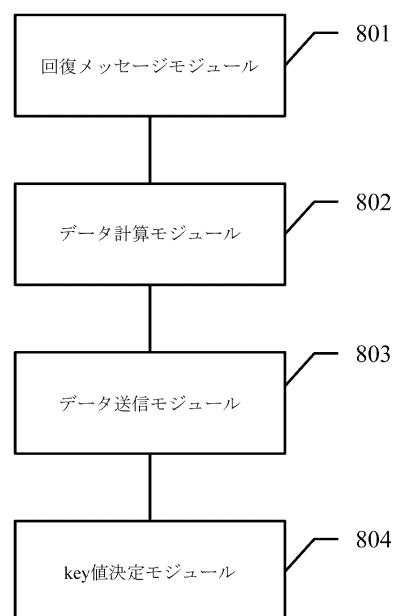
【図 6】



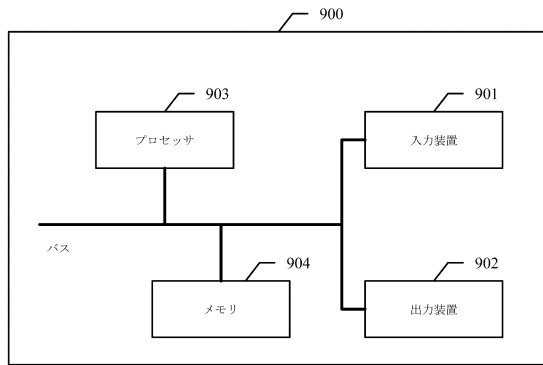
【図 7】



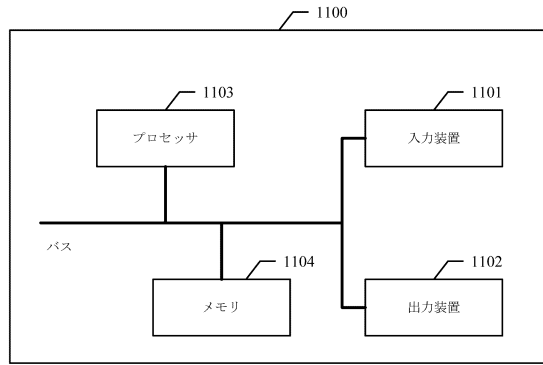
【図 8】



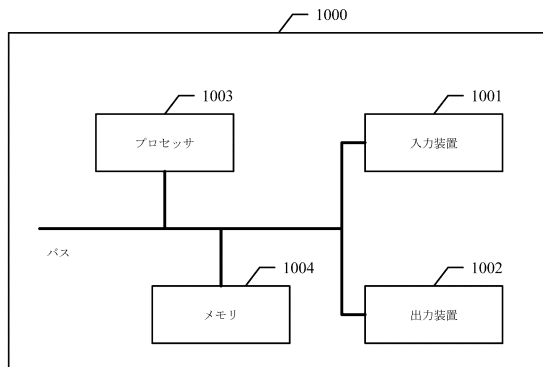
【図 9】



【図 11】



【図 10】



フロントページの続き

(51)Int.Cl.

F I

G 0 6 F	12/00	5 4 5 A
G 0 6 F	11/14	6 6 4
G 0 6 F	11/14	6 6 9

(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 王 道▲輝▼

中国 5 1 8 1 2 9 ▲広▼▲東▼省深▲チェン▼市▲龍▼▲崗▼区坂田▲華▼▲為▼▲総▼部▲辦▼
公楼

(72)発明者 ▲張▼ 烽

中国 5 1 8 1 2 9 ▲広▼▲東▼省深▲チェン▼市▲龍▼▲崗▼区坂田▲華▼▲為▼▲総▼部▲辦▼
公楼

(72)発明者 樊 ▲偉▼

中国 5 1 8 1 2 9 ▲広▼▲東▼省深▲チェン▼市▲龍▼▲崗▼区坂田▲華▼▲為▼▲総▼部▲辦▼
公楼

(72)発明者 ▲張▼ 志▲榮▼

中国 5 1 8 1 2 9 ▲広▼▲東▼省深▲チェン▼市▲龍▼▲崗▼区坂田▲華▼▲為▼▲総▼部▲辦▼
公楼

(72)発明者 ▲曾▼ 永▲強▼

中国 5 1 8 1 2 9 ▲広▼▲東▼省深▲チェン▼市▲龍▼▲崗▼区坂田▲華▼▲為▼▲総▼部▲辦▼
公楼

審査官 福田 正悟

(56)参考文献 国際公開第 2 0 1 5 / 0 5 7 2 4 0 (W O , A 1)

特開昭 6 2 - 2 9 3 3 5 5 (J P , A)

特表 2 0 1 4 - 5 2 9 1 2 6 (J P , A)

米国特許出願公開第 2 0 1 1 / 0 2 0 8 9 9 5 (U S , A 1)

米国特許出願公開第 2 0 1 5 / 0 1 4 9 8 1 9 (U S , A 1)

(58)調査した分野(Int.Cl. , D B 名)

G 0 6 F 1 2 / 0 0

G 0 6 F 3 / 0 6

G 0 6 F 1 1 / 1 4