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

特許第6382454号
(P6382454)

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

(51)Int.Cl.　　F I
G06F 12/00 (2006.01)　　G06F 12/00 545A
G06F 3/06 (2006.01)　　G06F 12/00 531R
　　G06F 3/06 304F

請求項の数 12　(全 58 頁)

(21)出願番号　特願2017-539482(P2017-539482)
(86)(22)出願日　平成26年11月6日(2014.11.6)
(65)公表番号　特表2017-534133(P2017-534133A)
(43)公表日　平成29年11月16日(2017.11.16)
(86)国際出願番号　PCT/CN2014/090445
(87)国際公開番号　WO2016/070375
(87)国際公開日　平成28年5月12日(2016.5.12)
審査請求日　平成29年4月19日(2017.4.19)

(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　華為総部▲ベン▼公楼
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】分散ストレージ及びレプリケーションシステム、並びに方法

(57)【特許請求の範囲】
【請求項１】
分散ストレージ及びレプリケーションシステムであって、前記システムは、少なくとも１つのメタデータ制御（ＭＤＣ）モジュールと、複数のＩＯルーティングモジュールと、複数のオブジェクトベースのストレージデバイス（ＯＳＤ）ノードとを備え、前記ＭＤＣモジュールは、各ＯＳＤノードに関して、各ＯＳＤノードによって管理される物理ストレージリソースに対応する少なくとも１つの論理パーティション（ｐａｒｔｉｔｉｏｎ）を構成し、前記少なくとも１つのｐａｒｔｉｔｉｏｎは、一次ｐａｒｔｉｔｉｏｎ、二次ｐａｒｔｉｔｉｏｎ、又は一次ｐａｒｔｉｔｉｏｎと二次ｐａｒｔｉｔｉｏｎの任意の組合せであり、一次ｐａｒｔｉｔｉｏｎ、及び前記一次ｐａｒｔｉｔｉｏｎに対応する二次ｐａｒｔｉｔｉｏｎは、ｐａｒｔｉｔｉｏｎグループを構成し、同一のｐａｒｔｉｔｉｏｎグループの中の一次ｐａｒｔｉｔｉｏｎと二次ｐａｒｔｉｔｉｏｎは、異なるＯＳＤノード上に位置付けられ、一次ｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤノードは、前記一次ｐａｒｔｉｔｉｏｎを備えるｐａｒｔｉｔｉｏｎグループの一次ＯＳＤノードであり、二次ｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤノードは、前記二次ｐａｒｔｉｔｉｏｎを備えるｐａｒｔｉｔｉｏｎグループの二次ＯＳＤノードであり、並びに前記ｐａｒｔｉｔｉｏｎによりｐａｒｔｉｔｉｏｎビューを生成し、前記ｐａｒｔｉｔｉｏｎビューは、ｐａｒｔｉｔｉｏｎグループの中のｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤについての情報を備える、ように適応され、前記ＩＯルーティングモジュールは、受信されたＩＯ要求をＯＳＤノードへルーティングするように適応され、前記ＯＳＤノードは、前記ＩＯ

要求により、前記ＩＯ要求に対応するデータの記憶を実行するように適応され、
前記ＭＤＣは、当該システムの中のＯＳＤノードが障害を有するＯＳＤノードであると決定し、前記障害を有するＯＳＤノード上のｐａｒｔｉｔｉｏｎを決定し、前記障害を有するＯＳＤノード上の前記ｐａｒｔｉｔｉｏｎを備えるｐａｒｔｉｔｉｏｎグループのｐａｒｔｉｔｉｏｎビューを更新し、前記更新されたｐａｒｔｉｔｉｏｎビューにおける一次ＯＳＤノードに更新通知を送信し、
前記一次ＯＳＤノードは、前記ＭＤＣモジュールによって送信された前記更新通知を受信した後、前記更新されたｐａｒｔｉｔｉｏｎビューにより、前記ＩＯ要求に対応する前記データのレプリケーションを処理するように適応され、
前記ｐａｒｔｉｔｉｏｎビューは、ｐａｒｔｉｔｉｏｎグループの中のｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤのものである一次／二次ＩＤ及び対応するｐａｒｔｉｔｉｏｎステータスを備え、前記一次ＯＳＤノードは、前記更新されたｐａｒｔｉｔｉｏｎビューにより、前記一次ＯＳＤノード上にローカルで記憶されたｐａｒｔｉｔｉｏｎビューを更新するようにさらに適応され、
前記更新されたｐａｒｔｉｔｉｏｎビューにより、前記ＩＯ要求に対応する前記データのレプリケーションを前記処理することは、
前記更新されたローカルで記憶されたｐａｒｔｉｔｉｏｎビューにより、前記ＩＯルーティングモジュールからの前記ＩＯ要求に対応する前記データを、ｐａｒｔｉｔｉｏｎステータスが前記更新されたローカルで記憶されたｐａｒｔｉｔｉｏｎビューにおいて整合性がある二次ＯＳＤノード上に、或いはｐａｒｔｉｔｉｏｎステータスが前記更新されたローカルで記憶されたｐａｒｔｉｔｉｏｎビューにおいて整合性がある二次ＯＳＤノード、及びｐａｒｔｉｔｉｏｎステータスが前記更新されたローカルで記憶されたｐａｒｔｉｔｉｏｎビューにおいて不整合であるが、データを回復している二次ＯＳＤノードの上にレプリケートすることを備える、
システム。
【請求項２】
前記ＭＤＣモジュールは、ＩＯビューを生成し、前記ＩＯビューは、ｐａｒｔｉｔｉｏｎグループの一次ＯＳＤノードの識別子を備え、並びに前記ＩＯルーティングモジュール、及び前記ｐａｒｔｉｔｉｏｎビューにおいて前記ｐａｒｔｉｔｉｏｎが位置付けられる前記ＯＳＤノードに前記ＩＯビューを送信するように適応され、
前記一次ＯＳＤノードは、前記更新されたｐａｒｔｉｔｉｏｎビューにより、前記一次ＯＳＤノード上にローカルで記憶されたＩＯビューを更新し、及び前記更新されたローカルで記憶されたＩＯビューにより、前記ＩＯ要求に対応する前記データのレプリケーションを処理するようにさらに適応される、請求項１に記載のシステム。
【請求項３】
前記ＭＤＣモジュールは、前記障害を有するＯＳＤノード上の前記ｐａｒｔｉｔｉｏｎが一次ｐａｒｔｉｔｉｏｎを備えると決定した場合、前記一次ｐａｒｔｉｔｉｏｎを備えるｐａｒｔｉｔｉｏｎグループのＩＯビューを更新し、前記更新されたＩＯビューについて前記更新されたｐａｒｔｉｔｉｏｎビューにおける二次ＯＳＤノードに通知するようにさらに適応され、
前記更新されたｐａｒｔｉｔｉｏｎビューにおける前記二次ＯＳＤノードは、前記更新されたＩＯビューにより、ローカルで記憶されたＩＯビューを更新し、前記更新されたローカルで記憶されたＩＯビューにより、前記ＩＯ要求に対応する前記データのレプリケーションを処理するように適応される、請求項２に記載のシステム。
【請求項４】
前記ＭＤＣモジュールは、前記更新されたＩＯビューについて前記ＩＯルーティングモジュールに通知するようにさらに適応され、
前記ＩＯルーティングモジュールは、前記更新されたＩＯビューにより、前記ＩＯルーティングモジュール上にローカルで記憶されたＩＯビューを更新し、前記更新されたローカルで記憶されたＩＯビューにより、前記ＩＯ要求の転送を処理する、請求項３に記載の

システム。
【請求項５】
前記障害を有するＯＳＤノード上の前記ｐａｒｔｉｔｉｏｎを備えるｐａｒｔｉｔｉｏｎグループのｐａｒｔｉｔｉｏｎビューを前記更新することは、
前記障害を有するＯＳＤノード上の前記ｐａｒｔｉｔｉｏｎが二次ｐａｒｔｉｔｉｏｎを備える場合、前記二次ｐａｒｔｉｔｉｏｎを備えるｐａｒｔｉｔｉｏｎグループのｐａｒｔｉｔｉｏｎビューにおいて前記障害を有するＯＳＤノードのｐａｒｔｉｔｉｏｎステータスに不整合であるという印を付けること、並びに
前記障害を有するＯＳＤノード上の前記ｐａｒｔｉｔｉｏｎが前記一次ｐａｒｔｉｔｉｏｎを備える場合、前記一次ｐａｒｔｉｔｉｏｎを備える前記ｐａｒｔｉｔｉｏｎグループのｐａｒｔｉｔｉｏｎビューにおいて一次ＯＳＤノードの役割をする前記障害を有するＯＳＤノードを新たな二次ＯＳＤノードとして設定すること、前記新たな二次ＯＳＤノードに対応するｐａｒｔｉｔｉｏｎステータスに不整合であるという印を付けること、ｐａｒｔｉｔｉｏｎステータスが整合性がある二次ＯＳＤノードを、前記一次ｐａｒｔｉｔｉｏｎを備える前記ｐａｒｔｉｔｉｏｎグループの前記ｐａｒｔｉｔｉｏｎビューにおける元の二次ＯＳＤノードから選択すること、及び前記選択された二次ＯＳＤノードを新たな一次ＯＳＤノードとして設定することを備える、請求項３に記載のシステム。
【請求項６】
前記障害を有するＯＳＤノードがフェイルバック及びデータ回復を完了した後、前記ＭＤＣモジュールは、前記更新されたｐａｒｔｉｔｉｏｎビュー及び前記更新されたＩＯビューをさらに更新し、前記さらに更新されたｐａｒｔｉｔｉｏｎビューにおける一次ＯＳＤノードに更新通知を送信し、及び前記さらに更新されたｐａｒｔｉｔｉｏｎビューにおける二次ＯＳＤノードに更新通知を送信するように適応され、
前記さらに更新されたｐａｒｔｉｔｉｏｎビューにおける前記一次ＯＳＤノードは、前記さらに更新されたｐａｒｔｉｔｉｏｎビューにより、前記ＩＯ要求に対応する前記データのレプリケーションを処理するように適応され、
前記さらに更新されたｐａｒｔｉｔｉｏｎビューにおける前記二次ＯＳＤノードは、前記さらに更新されたＩＯビューにより、前記ＩＯ要求に対応する前記データのレプリケーションを処理するように適応される、請求項５に記載のシステム。
【請求項７】
前記システムは複数のホストを備え、前記ＭＤＣモジュール、前記ＩＯルーティングモジュール、及び前記ＯＳＤノードは、前記複数のホストのうちの少なくとも１つに別々に配置され、前記ＯＳＤノードは、前記ホスト上の物理ストレージリソースを管理するように適応される、請求項１から６のいずれか一項に記載のシステム。
【請求項８】
分散ストレージ及びレプリケーションシステムであって、前記システムは、少なくとも１つのメタデータ制御（ＭＤＣ）モジュールと、複数のＩＯルーティングモジュールと、複数のオブジェクトベースのストレージデバイス（ＯＳＤ）ノードとを備え、前記ＭＤＣモジュールは、各ＯＳＤノードに関して、各ＯＳＤノードによって管理される物理ストレージリソースに対応する少なくとも１つの論理パーティション（ｐａｒｔｉｔｉｏｎ）を構成し、前記少なくとも１つのｐａｒｔｉｔｉｏｎは、一次ｐａｒｔｉｔｉｏｎ、二次ｐａｒｔｉｔｉｏｎ、又は一次ｐａｒｔｉｔｉｏｎと二次ｐａｒｔｉｔｉｏｎの任意の組合せであり、一次ｐａｒｔｉｔｉｏｎ、及び一次ｐａｒｔｉｔｉｏｎに対応する二次ｐａｒｔｉｔｉｏｎは、ｐａｒｔｉｔｉｏｎグループを構成し、同一のｐａｒｔｉｔｉｏｎグループの中の一次ｐａｒｔｉｔｉｏｎと二次ｐａｒｔｉｔｉｏｎは、異なるＯＳＤノード上に位置付けられ、一次ｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤノードは、前記一次ｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループの一次ＯＳＤノードであり、二次ｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤノードは、前記二次ｐａｒｔｉｔｉｏｎを備えるｐａｒｔｉｔｉｏｎグループの二次ＯＳＤノードであり、並びに前記ｐａｒｔｉｔｉｏｎによりｐａｒｔｉｔｉｏｎビューを生成し、前記ｐａｒｔｉｔｉｏｎビューは、ｐａｒ

ｔｉｔｉｏｎグループの中のｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤについての情報、並びにｐａｒｔｉｔｉｏｎグループの中のｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤのものである一次／二次ＩＤ及び対応するｐａｒｔｉｔｉｏｎステータスを備える、ように適応され、前記ＩＯルーティングモジュールは、受信されたＩＯ要求をＯＳＤノードへルーティングするように適応され、前記ＯＳＤノードは、前記ＩＯ要求により、前記ＩＯ要求に対応するデータの記憶を実行するように適応され、

前記システムは、メモリと、プロセッサとを備え、

前記メモリは、コンピュータ可読命令を記憶するように適応され、前記命令は、前記ＭＤＣモジュール、前記ＩＯルーティングモジュール、及び前記ＯＳＤノードの機能を実行するのに使用され、

前記プロセッサは、前記メモリに接続され、前記メモリの中の前記命令を読み取り、かつ前記命令により、前記プロセッサが、

前記システムにおけるＯＳＤノードが障害を有するＯＳＤノードであると決定し、前記障害を有するＯＳＤノード上のｐａｒｔｉｔｉｏｎを決定する工程と、

前記障害を有するＯＳＤノード上の前記ｐａｒｔｉｔｉｏｎを備えるｐａｒｔｉｔｉｏｎグループのｐａｒｔｉｔｉｏｎビューを更新する工程と、

前記更新されたｐａｒｔｉｔｉｏｎビューにおける前記ｐａｒｔｉｔｉｏｎグループが位置付けられる一次ＯＳＤノードに更新通知を送信して、前記一次ＯＳＤノードが、前記更新されたｐａｒｔｉｔｉｏｎビューにより、前記一次ＯＳＤノードにローカルで記憶されたｐａｒｔｉｔｉｏｎビューを更新し、前記更新されたローカルで記憶されたｐａｒｔｉｔｉｏｎビューにより、前記ＩＯルーティングモジュールからの前記ＩＯ要求に対応する前記データを、ｐａｒｔｉｔｉｏｎステータスが前記更新されたローカルで記憶されたｐａｒｔｉｔｉｏｎビューにおいて整合性がある二次ＯＳＤノード上に、或いはｐａｒｔｉｔｉｏｎステータスが前記更新されたローカルで記憶されたｐａｒｔｉｔｉｏｎビューにおいて整合性がある二次ＯＳＤノード、及びｐａｒｔｉｔｉｏｎステータスが前記更新されたローカルで記憶されたｐａｒｔｉｔｉｏｎビューにおいて不整合であるが、データを回復している二次ＯＳＤノードの上にレプリケートするようにする工程と、

を実行することを可能にするように適応される、システム。

【請求項９】

分散ストレージシステムにおいてデータ記憶及びデータレプリケーションを管理するための方法であって、前記システムは、少なくとも１つのメタデータ制御（ＭＤＣ）モジュールと、複数のＩＯルーティングモジュールと、複数のオブジェクトベースのストレージデバイス（ＯＳＤ）ノードとを備え、前記ＭＤＣモジュールは、各ＯＳＤノードに関して、各ＯＳＤノードによって管理される物理ストレージリソースに対応する少なくとも１つの論理パーティション（ｐａｒｔｉｔｉｏｎ）を構成し、前記少なくとも１つのｐａｒｔｉｔｉｏｎは、一次ｐａｒｔｉｔｉｏｎ、二次ｐａｒｔｉｔｉｏｎ、又は一次ｐａｒｔｉｔｉｏｎと二次ｐａｒｔｉｔｉｏｎの任意の組合せであり、一次ｐａｒｔｉｔｉｏｎ、及び前記一次ｐａｒｔｉｔｉｏｎに対応する二次ｐａｒｔｉｔｉｏｎは、ｐａｒｔｉｔｉｏｎグループを構成し、同一のｐａｒｔｉｔｉｏｎグループの中の一次ｐａｒｔｉｔｉｏｎと二次ｐａｒｔｉｔｉｏｎは、異なるＯＳＤノード上に位置付けられ、一次ｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤノードは、前記一次ｐａｒｔｉｔｉｏｎを備えるｐａｒｔｉｔｉｏｎグループの一次ＯＳＤノードであり、二次ｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤノードは、前記二次ｐａｒｔｉｔｉｏｎを備えるｐａｒｔｉｔｉｏｎグループの二次ＯＳＤノードであり、並びに前記ｐａｒｔｉｔｉｏｎによりｐａｒｔｉｔｉｏｎビューを生成し、前記ｐａｒｔｉｔｉｏｎビューは、ｐａｒｔｉｔｉｏｎグループの中のｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤについての情報を備える、ように適応され、前記ＩＯルーティングモジュールは、ＯＳＤノードに受信されたＩＯ要求をルーティングするように適応され、前記ＯＳＤノードは、前記ＩＯ要求により、前記ＩＯ要求に対応するデータの記憶を実行するように適応され、前記方法は、

前記システムにおけるＯＳＤノードが障害を有するＯＳＤノードであると決定し、前記

障害を有するＯＳＤノード上のｐａｒｔｉｔｉｏｎを決定し、前記障害を有するＯＳＤノード上の前記ｐａｒｔｉｔｉｏｎを備えるｐａｒｔｉｔｉｏｎグループのｐａｒｔｉｔｉｏｎビューを更新すること、及び前記更新されたｐａｒｔｉｔｉｏｎビューにおける一次ＯＳＤノードに更新通知を送信する工程を備え、

前記一次ＯＳＤノードは、前記ＭＤＣモジュールによって送信された前記更新通知を受信した後、前記更新されたｐａｒｔｉｔｉｏｎビューにより、前記ＩＯ要求に対応する前記データのレプリケーションを処理するように適応され、

前記ｐａｒｔｉｔｉｏｎビューは、ｐａｒｔｉｔｉｏｎグループの中のｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤのものである一次／二次ＩＤ及び対応するｐａｒｔｉｔｉｏｎステータスを特に備え、前記一次ＯＳＤノードは、前記更新されたｐａｒｔｉｔｉｏｎビューにより、前記一次ＯＳＤノード上にローカルで記憶されたｐａｒｔｉｔｉｏｎビューを更新するようにさらに適応され、

前記更新されたｐａｒｔｉｔｉｏｎビューにより、前記ＩＯ要求に対応する前記データのレプリケーションを前記処理することは、

前記更新されたローカルで記憶されたｐａｒｔｉｔｉｏｎビューにより、前記ＩＯルーティングモジュールからの前記ＩＯ要求に対応する前記データを、ｐａｒｔｉｔｉｏｎステータスが前記更新されたローカルで記憶されたｐａｒｔｉｔｉｏｎビューにおいて整合性がある二次ＯＳＤノード上に、或いはｐａｒｔｉｔｉｏｎステータスが前記更新されたローカルで記憶されたｐａｒｔｉｔｉｏｎビューにおいて整合性がある二次ＯＳＤノード、及びｐａｒｔｉｔｉｏｎステータスが前記更新されたローカルで記憶されたｐａｒｔｉｔｉｏｎビューにおいて不整合であるが、データを回復している二次ＯＳＤノードの上にレプリケートすることを備える、方法。

【請求項１０】

前記方法は、

前記障害を有するＯＳＤノード上の前記ｐａｒｔｉｔｉｏｎが一次ｐａｒｔｉｔｉｏｎを備えると決定した場合、前記ＭＤＣモジュールによる、前記一次ｐａｒｔｉｔｉｏｎを備えるｐａｒｔｉｔｉｏｎグループのＩＯビューを更新し、及び前記更新されたＩＯビューについて前記更新されたｐａｒｔｉｔｉｏｎビューにおける二次ＯＳＤノードに通知する工程と、

前記更新されたｐａｒｔｉｔｉｏｎビューにおける前記二次ＯＳＤノードによる、前記更新されたＩＯビューにより、ローカルで記憶されたＩＯビューを更新し、及び前記更新されたローカルで記憶されたＩＯビューにより、前記ＩＯ要求に対応する前記データのレプリケーションを処理する工程と、

をさらに備える請求項９に記載の方法。

【請求項１１】

前記障害を有するＯＳＤノード上の前記ｐａｒｔｉｔｉｏｎを備えるｐａｒｔｉｔｉｏｎグループのｐａｒｔｉｔｉｏｎビューを前記更新することは、

前記障害を有するＯＳＤノード上の前記ｐａｒｔｉｔｉｏｎが二次ｐａｒｔｉｔｉｏｎを備える場合、前記二次ｐａｒｔｉｔｉｏｎを備えるｐａｒｔｉｔｉｏｎグループのｐａｒｔｉｔｉｏｎビューにおいて前記障害を有するＯＳＤノードのｐａｒｔｉｔｉｏｎステータスに不整合であるという印を付けること、並びに

前記障害を有するＯＳＤノード上の前記ｐａｒｔｉｔｉｏｎが前記一次ｐａｒｔｉｔｉｏｎを備える場合、前記一次ｐａｒｔｉｔｉｏｎを備える前記ｐａｒｔｉｔｉｏｎグループのｐａｒｔｉｔｉｏｎビューにおいて一次ＯＳＤノードの役割をする前記障害を有するＯＳＤノードを新たな二次ＯＳＤノードとして設定し、前記新たな二次ＯＳＤノードに対応するｐａｒｔｉｔｉｏｎステータスに不整合であるという印を付け、ｐａｒｔｉｔｉｏｎステータスが整合性がある二次ＯＳＤノードを、前記一次ｐａｒｔｉｔｉｏｎを備える前記ｐａｒｔｉｔｉｏｎグループの前記ｐａｒｔｉｔｉｏｎビューにおける元の二次ＯＳＤノードから選択し、及び前記選択された二次ＯＳＤノードを新たな一次ＯＳＤノードとして設定することを備える、請求項１０に記載の方法。

【請求項１２】
前記方法は、
前記障害を有するＯＳＤノードがフェイルバック及びデータ回復を完了した後、前記ＭＤＣモジュールによる、前記更新されたｐａｒｔｉｔｉｏｎビュー及び前記更新されたＩＯビューをさらに更新し、前記さらに更新されたｐａｒｔｉｔｉｏｎビューにおける一次ＯＳＤノードに更新通知を送信し、及び前記さらに更新されたｐａｒｔｉｔｉｏｎビューにおける二次ＯＳＤノードに更新通知を送信する工程と、
前記さらに更新されたｐａｒｔｉｔｉｏｎビューにおける前記一次ＯＳＤノードによる、前記さらに更新されたｐａｒｔｉｔｉｏｎビューにより、前記ＩＯ要求に対応する前記データのレプリケーションを処理する工程と、
前記さらに更新されたｐａｒｔｉｔｉｏｎビューにおける前記二次ＯＳＤノードによる、前記さらに更新されたＩＯビューにより、前記ＩＯ要求に対応する前記データのレプリケーションを処理する工程と、
をさらに備える請求項１１に記載の方法。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、ＩＴ（Ｉｎｆｏｒｍａｔｉｏｎ　Ｔｅｃｈｎｏｌｏｇｙ）情報技術分野に関し、詳細には、分散ストレージ及びレプリケーションシステム、並びに方法に関する。
【背景技術】
【０００２】
情報技術の急速な発展、及びインターネットの広い応用とともに、人々によって生成されるデータは、爆発的な様態で増加し、このことが、データストレージの拡張性により高い要件を課す。従来のストレージアレイシステムと比べて、分散ストレージシステムは、より良い拡張性、及び一般的なハードウェアデバイス互換性を有し、将来においてデータストレージの要件をより良く満たすことが可能である。
【０００３】
分散ストレージシステムにおいて、一般に、大量のストレージノードが分散システムを形成するように編成され、データ信頼性は、異なるノード間のデータレプリケーション及びバックアップによって確実にされて、データが異なるすべてのストレージノード上にレプリカを有するようになる。複数のデータレプリカのデータ整合性をどのようにして確実にすべきかが、長期にわたって分散ストレージシステムの直面する問題となっている。データ整合性を確実にする事例において、システムパフォーマンス及び利用可能性もまた、重要性の高まる考慮事項となる。
【０００４】
図１は、既存の２フェーズコミットプロトコル（２　Ｐｈａｓｅ　Ｃｏｍｍｉｔ，２ＰＣ）を示し、これは、典型的な集中型の強整合性レプリカ制御プロトコルであり、多くの分散データベースシステムにおいてレプリカ整合性を確実にするのに使用される。
【０００５】
２フェーズコミットプロトコルにおいて、システムは、一般に、２つのタイプのノード、すなわち、コーディネータ（ｃｏｏｒｄｉｎａｔｏｒ）と、参加者（ｐａｒｔｉｃｉｐａｎｔ）とを含む。コーディネータは、データ更新に関して投票を開始すること、及び投票決定を通知することを実行することを担い、参加者は、データ更新に関する投票に参加し、投票決定を実行する。
【０００６】
２フェーズコミットプロトコルは、２つのフェーズを含み、すなわち、フェーズ１は、コミット要求フェーズであり、コーディネータが、データ変更に関して投票するよう参加者に指示し、参加者が、参加者自らの投票結果、すなわち、はい又はいいえについてコーディネータに通知し、フェーズ２は、コミットフェーズであり、コーディネータが、第１のフェーズにおける投票結果により決定、すなわち、コミット又は取消しを行う。

【０００７】
２フェーズコミットプロトコルを１回、実行することに成功することは、４つのメッセージを用いてコーディネータと各参加者の間で少なくとも２ラウンドの対話を要求し、対話の過度の回数がパフォーマンスを低下させる。さらに、２フェーズコミットプロトコルにおいて、ノードが障害を有するようになった場合、又は応答を有さないことが続く場合、別のＩＯ要求は、ブロックされ、最終的に、タイムアウトのために失敗し、データロールバックが実行される必要がある。２フェーズコミットプロトコルは、比較的低い障害耐性及び利用可能性を有する。
【発明の概要】
【０００８】
本発明の実施形態が、分散ストレージ及びレプリケーションシステム、並びに分散ストレージシステムにおいてデータストレージ及びレプリケーションを管理するための方法を提供して、既存の整合性レプリケーションプロトコルの低いパフォーマンス及び低い利用可能性の問題を解決するようにする。
【０００９】
第１の態様によれば、分散ストレージ及びレプリケーションシステムが提供され、システムは、少なくとも１つのメタデータ制御（ＭＤＣ）モジュールと、複数のＩＯルーティングモジュールと、複数のオブジェクトベースのストレージデバイス（ＯＳＤ）ノードとを含み、ＭＤＣモジュールは、各ＯＳＤノードに関して、各ＯＳＤノードによって管理される物理ストレージリソースに対応する少なくとも１つの論理パーティション（ｐａｒｔｉｔｉｏｎ）を構成し、少なくとも１つのｐａｒｔｉｔｉｏｎは、一次ｐａｒｔｉｔｉｏｎ、二次ｐａｒｔｉｔｉｏｎ、又は一次ｐａｒｔｉｔｉｏｎと二次ｐａｒｔｉｔｉｏｎの任意の組合せであり、一次ｐａｒｔｉｔｉｏｎ、及び一次ｐａｒｔｉｔｉｏｎに対応する二次ｐａｒｔｉｔｉｏｎは、ｐａｒｔｉｔｉｏｎグループを構成し、同一のｐａｒｔｉｔｉｏｎグループの中の一次ｐａｒｔｉｔｉｏｎと二次ｐａｒｔｉｔｉｏｎは、異なるＯＳＤノード上に位置付けられ、一次ｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤノードは、一次ｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループの一次ＯＳＤノードであり、二次ｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤノードは、二次ｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループの二次ＯＳＤノードであり、並びにｐａｒｔｉｔｉｏｎによりｐａｒｔｉｔｉｏｎビューを生成し、ｐａｒｔｉｔｉｏｎビューは、ｐａｒｔｉｔｉｏｎグループの中のｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤについての情報を含む、ように適応され、ＩＯルーティングモジュールは、ＯＳＤノードに受信されたＩＯ要求をルーティングするように適応され、ＯＳＤノードは、ＩＯ要求により、ＩＯ要求に対応するデータの記憶を実行するように適応され、ＭＤＣは、システムにおけるＯＳＤノードが障害を有するＯＳＤノードであると決定し、障害を有するＯＳＤノード上のｐａｒｔｉｔｉｏｎを決定し、障害を有するＯＳＤノード上のｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループのｐａｒｔｉｔｉｏｎビューを更新し、更新されたｐａｒｔｉｔｉｏｎビューにおける一次ＯＳＤノードに更新通知を送信するように適応され、一次ＯＳＤノードは、ＭＤＣモジュールによって送信された更新通知を受信した後、更新されたｐａｒｔｉｔｉｏｎビューにより、ＩＯ要求に対応するデータのレプリケーションを処理するように適応される。
【００１０】
第１の態様の第１の可能な実施様態において、ｐａｒｔｉｔｉｏｎビューは、ｐａｒｔｉｔｉｏｎグループの中のｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤのものである一次／二次ＩＤ及び対応するｐａｒｔｉｔｉｏｎステータスを特に含み、一次ＯＳＤノードは、更新されたｐａｒｔｉｔｉｏｎビューにより、一次ＯＳＤノード上にローカルで記憶されたｐａｒｔｉｔｉｏｎビューを更新するようにさらに適応され、更新されたｐａｒｔｉｔｉｏｎビューにより、ＩＯ要求に対応するデータのレプリケーションを処理することは、更新されたローカルで記憶されたｐａｒｔｉｔｉｏｎビューにより、ＩＯルーティングモジュールからのＩＯ要求に対応するデータを、ｐａｒｔｉｔｉｏｎステータスが更新さ

れたローカルで記憶されたｐａｒｔｉｔｉｏｎビューにおいて整合性がある二次ＯＳＤノード上に、或いはｐａｒｔｉｔｉｏｎステータスが更新されたローカルで記憶されたｐａｒｔｉｔｉｏｎビューにおいて整合性がある二次ＯＳＤノード、及びｐａｒｔｉｔｉｏｎステータスが更新されたローカルで記憶されたｐａｒｔｉｔｉｏｎビューにおいて不整合であるが、データを回復している二次ＯＳＤノードの上にレプリケートすることを特に含む。

【００１１】

第１の態様の第１の可能な実施様態を参照して、第２の可能な実施様態において、ＭＤＣモジュールは、ＩＯビューを生成し、ＩＯビューは、ｐａｒｔｉｔｉｏｎグループの一次ＯＳＤノードの識別子を含み、並びにＩＯルーティングモジュール、及びｐａｒｔｉｔｉｏｎビューにおいてｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤノードにＩＯビューを送信するように適応され、一次ＯＳＤノードは、更新されたｐａｒｔｉｔｉｏｎビューにより、一次ＯＳＤノード上にローカルで記憶されたＩＯビューを更新し、及び更新されたローカルで記憶されたＩＯビューにより、ＩＯ要求に対応するデータのレプリケーションを処理するようにさらに適応される。

【００１２】

第１の態様の第２の可能な実施様態を参照して、第３の可能な実施様態において、ＭＤＣモジュールは、障害を有するＯＳＤノード上のｐａｒｔｉｔｉｏｎが一次ｐａｒｔｉｔｉｏｎを含むと決定した場合、一次ｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループのＩＯビューを更新し、更新されたＩＯビューについて更新されたｐａｒｔｉｔｉｏｎビューにおける二次ＯＳＤノードに通知するようにさらに適応され、更新されたｐａｒｔｉｔｉｏｎビューにおける二次ＯＳＤノードは、更新されたＩＯビューにより、ローカルで記憶されたＩＯビューを更新し、更新されたローカルで記憶されたＩＯビューにより、ＩＯ要求に対応するデータのレプリケーションを処理するように適応される。

【００１３】

第１の態様の第３の可能な実施様態を参照して、第４の可能な実施様態において、障害を有するＯＳＤノード上のｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループのｐａｒｔｉｔｉｏｎビューを更新することは、障害を有するＯＳＤノード上のｐａｒｔｉｔｉｏｎが二次ｐａｒｔｉｔｉｏｎを含む場合、二次ｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループのｐａｒｔｉｔｉｏｎビューにおいて障害を有するＯＳＤノードのｐａｒｔｉｔｉｏｎステータスに不整合であるという印を付けること、並びに障害を有するＯＳＤノード上のｐａｒｔｉｔｉｏｎが一次ｐａｒｔｉｔｉｏｎを含む場合、一次ｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループのｐａｒｔｉｔｉｏｎビューにおいて一次ＯＳＤノードの役割をする障害を有するＯＳＤノードを新たな二次ＯＳＤノードとして設定すること、新たな二次ＯＳＤノードに対応するｐａｒｔｉｔｉｏｎステータスに不整合であるという印を付けること、ｐａｒｔｉｔｉｏｎステータスが整合性がある二次ＯＳＤノードを、一次ｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループのｐａｒｔｉｔｉｏｎビューにおける元の二次ＯＳＤノードから選択すること、及び選択された二次ＯＳＤノードを新たな一次ＯＳＤノードとして設定することを特に含む。

【００１４】

第２の態様によれば、分散ストレージ及びレプリケーションシステムが提供され、システムは、少なくとも１つのメタデータ制御（ＭＤＣ）モジュールと、複数のＩＯルーティングモジュールと、複数のオブジェクトベースのストレージデバイス（ＯＳＤ）ノードとを含み、ＭＤＣモジュールは、各ＯＳＤノードに関して、各ＯＳＤノードによって管理される物理ストレージリソースに対応する少なくとも１つの論理パーティション（ｐａｒｔｉｔｉｏｎ）を構成し、少なくとも１つのパーティションは、一次ｐａｒｔｉｔｉｏｎ、二次ｐａｒｔｉｔｉｏｎ、又は一次ｐａｒｔｉｔｉｏｎと二次ｐａｒｔｉｔｉｏｎの任意の組合せであり、一次ｐａｒｔｉｔｉｏｎ、及び一次ｐａｒｔｉｔｉｏｎに対応する二次ｐａｒｔｉｔｉｏｎは、ｐａｒｔｉｔｉｏｎグループを構成し、同一のｐａｒｔｉｔｉｏｎグループの中の一次ｐａｒｔｉｔｉｏｎと二次ｐａｒｔｉｔｉｏｎは、異なるＯＳＤノ

ード上に位置付けられ、一次ｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤノードは、一次ｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループの一次ＯＳＤノードであり、二次ｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤノードは、二次ｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループの二次ＯＳＤノードであり、並びにｐａｒｔｉｔｉｏｎによりｐａｒｔｉｔｉｏｎビュー及びＩＯビューを生成し、ｐａｒｔｉｔｉｏｎビューは、ｐａｒｔｉｔｉｏｎグループの中のｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤについての情報を含み、ＩＯビューは、ｐａｒｔｉｔｉｏｎグループの一次ＯＳＤの識別子を含む、ように適応され、ＩＯルーティングモジュールは、ＯＳＤノードに受信されたＩＯ要求をルーティングするように適応され、ＯＳＤノードは、ＩＯ要求によりＩＯデータ記憶を実行するように適応され、ＩＯルーティングモジュールは、ＩＯ要求を受信し、ＩＯ要求はｋｅｙを含み、ｋｅｙにより、ＩＯ要求に対応するデータが属するｐａｒｔｉｔｉｏｎグループを決定し、及びデータが属するｐａｒｔｉｔｉｏｎグループの一次ＯＳＤノードを決定し、データが属するｐａｒｔｉｔｉｏｎグループのＩＯビューのＩＯビューバージョン情報をＩＯ要求に追加し、及び決定された一次ＯＳＤノードに、ＩＯビューバージョン情報を搬送するＩＯ要求を送信するように適応され、一次ＯＳＤノードは、ＩＯ要求を受信し、ＩＯビューバージョン情報により、ＩＯ要求におけるＩＯビューバージョンが、一次ＯＳＤノード上にローカルで記憶されたＩＯビューバージョンと整合性があると決定した後、ＩＯ要求を実行し、ＩＯビューバージョン情報を搬送するレプリケーション要求を生成し、及びデータが属するｐａｒｔｉｔｉｏｎグループの二次ＯＳＤノードにレプリケーション要求を送信するように適応され、二次ＯＳＤノードは、レプリケーション要求を受信し、及びＩＯビューバージョン情報により、レプリケーション要求におけるＩＯビューバージョンが、二次ＯＳＤノード上にローカルで記憶されたＩＯビューバージョンと整合性があると決定した後、レプリケーション要求を実行して、二次ＯＳＤノード上のＩＯ要求に対応するデータが、一次ＯＳＤノード上のＩＯ要求に対応するデータと整合性を保つように適応される。

【００１５】

第２の態様の第１の可能な実施様態において、一次ＯＳＤノードは、ＩＯビューバージョン情報により、ＩＯ要求におけるＩＯビューバージョンが、一次ＯＳＤ上にローカルで記憶されたＩＯビューバージョンより早期であると決定した後、ＩＯルーティングモジュールにエラーを返し、及びＩＯ要求におけるＩＯビューバージョンが、一次ＯＳＤ上にローカルで記憶されたＩＯビューバージョンより後期であると決定した後、キャッシュキューにＩＯ要求を追加し、及びデータが属するｐａｒｔｉｔｉｏｎグループのＩＯビューのＩＯビューバージョン情報に関してＭＤＣモジュールにクエリを行って、一次ＯＳＤ上にローカルで記憶されたＩＯビューバージョンがＩＯ要求におけるＩＯビューバージョンと整合性があると決定した後、ＩＯ要求を実行するようにさらに適応され、ＩＯルーティングモジュールは、一次ＯＳＤノードによって返されたエラーを受信した後、データが属するｐａｒｔｉｔｉｏｎグループのＩＯビューに関してＭＤＣモジュールにクエリを行い、及び更新されたＩＯビューバージョン情報を獲得した後、更新されたＩＯビューバージョン情報を搬送するＩＯ要求を送信するように適応される。

【００１６】

第２の態様、又は第２の態様の第１の可能な実施様態を参照して、第２の可能な実施様態において、ＩＯビューバージョン情報は、ＩＯビューバージョン番号を含み、一次ＯＳＤノードは、ＩＯ要求に関するシーケンス識別子をさらに生成し、シーケンス識別子を、二次ＯＳＤノードに送信されるレプリケーション要求に追加し、シーケンス識別子は、ＩＯビューバージョン番号と、シーケンス番号とを含み、シーケンス番号は、ＩＯビューバージョン内のＩＯビューにおけるｐａｒｔｉｔｉｏｎグループに対応するデータに対する変更操作の通し番号を示し、二次ＯＳＤノードは、シーケンス識別子によりレプリケーション要求を実行するようにさらに適応される。

【００１７】

第２の態様の第２の可能な実施様態を参照して、第３の可能な実施様態において、レプ

リケーション要求は、ｐａｒｔｉｔｉｏｎグループに関する一次ＯＳＤノードによって送信された前のレプリケーション要求におけるシーケンス識別子をさらに搬送し、二次ＯＳＤノードは、レプリケーション要求を受信した後、前のレプリケーション要求におけるシーケンス識別子が、二次ＯＳＤノード上にローカルで記憶された最大のシーケンス識別子と整合性がある場合、レプリケーション要求を実行するようにさらに適応される。

【００１８】

第２の態様、又は第２の態様の第１、第２、及び第３の可能な実施様態を参照して、第４の可能な実施様態において、ｐａｒｔｉｔｉｏｎビューは、ｐａｒｔｉｔｉｏｎグループの中のｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤのものである一次／二次ＩＤ及び対応するｐａｒｔｉｔｉｏｎステータスを特に含み、ＭＤＣモジュールは、ＩＯ要求を処理するプロセスにおいて、一次ＯＳＤノードが障害を有するようになったことを検出した場合、データが属するｐａｒｔｉｔｉｏｎグループのｐａｒｔｉｔｉｏｎビューにおける一次ＯＳＤノードを新たな二次ＯＳＤノードとして設定し、及び新たな二次ＯＳＤのｐａｒｔｉｔｉｏｎステータスに不整合であるという印を付け、データが属するｐａｒｔｉｔｉｏｎグループのｐａｒｔｉｔｉｏｎビューにおける二次ＯＳＤノードの任意の二次ＯＳＤノードを新たな一次ＯＳＤノードとして設定し、データが属するｐａｒｔｉｔｉｏｎグループの更新されたｐａｒｔｉｔｉｏｎビューについて新たな一次ＯＳＤノードに通知し、新たな一次ＯＳＤノードを使用することによって、データが属するｐａｒｔｉｔｉｏｎグループのＩＯビューを更新し、及びデータが属するｐａｒｔｉｔｉｏｎの更新されたＩＯビューについてＩＯルーティングモジュールに通知するようにさらに適応され、ＩＯルーティングモジュールは、ＭＤＣモジュールによって送信されたｐａｒｔｉｔｉｏｎグループのものである更新されたＩＯビューを受信し、及びｐａｒｔｉｔｉｏｎグループの更新されたＩＯビューにより新たな一次ＯＳＤノードにＩＯ要求を送信するようにさらに適応され、新たな一次ＯＳＤノードは、ＩＯ要求を受信し、ＩＯ要求を実行した後、第２のレプリケーション要求を生成し、及び第２のレプリケーション要求を、ｐａｒｔｉｔｉｏｎステータスが、データが属するｐａｒｔｉｔｉｏｎグループの更新されたｐａｒｔｉｔｉｏｎビューにおいて整合性のある二次ＯＳＤノードに送信するように適応される。

【００１９】

第２の態様、又は第２の態様の第１、第２、及び第３の可能な実施様態を参照して、第５の可能な実施様態において、ｐａｒｔｉｔｉｏｎビューは、ｐａｒｔｉｔｉｏｎグループの中のｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤのものである一次／二次ＩＤ及び対応するｐａｒｔｉｔｉｏｎステータスを特に含み、ＭＤＣモジュールは、ＩＯ要求を処理するプロセスにおいて、二次ＯＳＤノードのうちのいずれかの二次ＯＳＤノードが障害を有するようになったことを検出した場合、データが属するｐａｒｔｉｔｉｏｎグループのｐａｒｔｉｔｉｏｎビューにおけるそのいずれかの二次ＯＳＤのｐａｒｔｉｔｉｏｎステータスに不整合であるという印を付け、及びデータが属するｐａｒｔｉｔｉｏｎグループの更新されたｐａｒｔｉｔｉｏｎビューについて一次ＯＳＤノードに通知するようにさらに適応され、一次ＯＳＤノードは、データが属するｐａｒｔｉｔｉｏｎグループの更新されたｐａｒｔｉｔｉｏｎビューを受信した後、ｐａｒｔｉｔｉｏｎステータスが、更新されたｐａｒｔｉｔｉｏｎビューにおいて整合性がある二次ＯＳＤノードにレプリケーション要求を送信し、及びｐａｒｔｉｔｉｏｎステータスが不整合である二次ＯＳＤノードにレプリケーション要求を送信することを省くように適応される。

【００２０】

第３の態様によれば、分散ストレージ及びレプリケーションシステムが提供され、システムは、少なくとも１つのメタデータ制御（ＭＤＣ）モジュールと、複数のＩＯルーティングモジュールと、複数のオブジェクトベースのストレージデバイス（ＯＳＤ）ノードとを含み、ＭＤＣモジュールは、各ＯＳＤノードに関して、各ＯＳＤノードによって管理される物理ストレージリソースに対応する少なくとも１つの論理パーティション（ｐａｒｔｉｔｉｏｎ）を構成し、少なくとも１つのｐａｒｔｉｔｉｏｎは、一次ｐａｒｔｉｔｉｏｎ、二次ｐａｒｔｉｔｉｏｎ、又は一次ｐａｒｔｉｔｉｏｎと二次ｐａｒｔｉｔｉｏｎの

任意の組合せであり、一次ｐａｒｔｉｔｉｏｎ、及び一次ｐａｒｔｉｔｉｏｎに対応する二次ｐａｒｔｉｔｉｏｎは、ｐａｒｔｉｔｉｏｎグループを構成し、同一のｐａｒｔｉｔｉｏｎグループの中の一次ｐａｒｔｉｔｉｏｎと二次ｐａｒｔｉｔｉｏｎは、異なるＯＳＤノード上に位置付けられ、一次ｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤノードは、一次ｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループの一次ＯＳＤノードであり、二次ｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤノードは、二次ｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループの二次ＯＳＤノードであり、並びにｐａｒｔｉｔｉｏｎによりｐａｒｔｉｔｉｏｎビュー及びＩＯビューを生成し、ｐａｒｔｉｔｉｏｎビューは、ｐａｒｔｉｔｉｏｎグループの中のｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤについての情報を含み、ＩＯビューは、ｐａｒｔｉｔｉｏｎグループの一次ＯＳＤの識別子を含む、ように適応され、ＩＯルーティングモジュールは、ＯＳＤノードに受信されたＩＯ要求をルーティングするように適応され、ＯＳＤノードは、ＩＯ要求により、ＩＯ要求に対応するデータの記憶を実行するように適応され、ＯＳＤノードは、フェイルバックの後、ＭＤＣモジュールにＯＳＤノード上のｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループのＩＯビューを要求するクエリ要求を送信し、ＯＳＤノードは、フェイルバックＯＳＤノードと呼ばれ、クエリ要求は、フェイルバックＯＳＤノードのＯＳＤ識別子を搬送し、ＭＤＣによって返されるＩＯビューを受信し、ＩＯビューにおける一次ＯＳＤに対して、障害中にフェイルバックＯＳＤノードによって更新されたデータを回復することを要求するデータ回復要求を開始し、障害中に更新され、かつ一次ＯＳＤによって送信されるデータを受信し、及びｐａｒｔｉｔｉｏｎグループのものであり、かつフェイルバックＯＳＤノードがデータ回復を完了した後にＭＤＣモジュールによって更新されたｐａｒｔｉｔｉｏｎビューにより、ＩＯ要求のレプリケーションを処理するように適応され、ＭＤＣモジュールは、フェイルバックＯＳＤノードのクエリ要求を受信し、クエリ要求におけるＯＳＤ識別子によりフェイルバックＯＳＤノードにＩＯビューを返し、及びフェイルバックＯＳＤノードがデータ回復を完了した後、ｐａｒｔｉｔｉｏｎグループのｐａｒｔｉｔｉｏｎビューを更新するように適応され、一次ＯＳＤノードは、フェイルバックＯＳＤノードのデータ回復要求を受信し、フェイルバックＯＳＤノードに、障害中に更新されたデータを送信し、及びｐａｒｔｉｔｉｏｎグループのものであり、かつフェイルバックＯＳＤノードがデータ回復を完了した後にＭＤＣモジュールによって更新されたｐａｒｔｉｔｉｏｎビューにより、ＩＯ要求に対応するデータのレプリケーションを処理するように適応される。

【００２１】

第３の態様の第１の可能な実施様態において、一次ＯＳＤノードは、データ回復要求を受信した後、フェイルバックＯＳＤノード上のｐａｒｔｉｔｉｏｎに関するＩＯルーティングモジュールによって送信されたＩＯ要求を受信し、ＩＯ要求を実行し、及びフェイルバックＯＳＤノードに、ＩＯキー情報、及びＩＯ要求に対応するデータを搬送するレプリケーション要求を送信するようにさらに適応され、フェイルバックＯＳＤノードは、ログに、ＩＯキー情報、及びＩＯ要求に対応するデータを搬送するレプリケーション要求を書き込み、データ回復が完了された後、ログのレコードにより、ＩＯ要求に対応するデータを、フェイルバックＯＳＤノードによって管理される物理ストレージリソースに対して更新する。

【００２２】

第３の態様の第１の可能な実施様態を参照して、第２の可能な実施様態において、データ回復要求は、フェイルバックＯＳＤノード上のｐａｒｔｉｔｉｏｎに関するＩＯ操作のものであり、かつフェイルバックＯＳＤノード上にローカルで記録された最大のシーケンス識別子を搬送し、最大のシーケンス識別子は、フェイルバックＯＳＤノード上のｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループのＩＯビューの最新のＩＯビューバージョン番号であり、データに対する変更操作の最大の通し番号が、最新のＩＯビューバージョン番号に対応するＩＯビューにおけるｐａｒｔｉｔｉｏｎに対応し、障害中に更新されたデータをフェイルバックＯＳＤノードに送信することは、データ回復要求における最大

のシーケンス識別子が、一次ＯＳＤノード上にローカルで記憶された現在の最小のシーケンス識別子以上であると決定すること、障害中にフェイルバックＯＳＤノードに欠落しているｅｎｔｒｙをフェイルバックＯＳＤノードに送信すること、ｅｎｔｒｙによりフェイルバックＯＳＤノードによって開始されたデータ回復要求を受信すること、及びｅｎｔｒｙに対応するデータをフェイルバックＯＳＤノードに送信することを含み、最小のシーケンス識別子は、ｐａｒｔｉｔｉｏｎグループのＩＯビューのものであり、かつ一次ＯＳＤノード上に記憶された最小のＩＯビューバージョン番号であり、データに対する変更操作の最小の通し番号は、最小のＩＯビューバージョン番号に対応するＩＯビューにおけるｐａｒｔｉｔｉｏｎに対応する。

【００２３】

第３の態様の第１の可能な実施様態を参照して、第３の可能な実施様態において、データ回復要求は、フェイルバックＯＳＤノード上のｐａｒｔｉｔｉｏｎに関するＩＯ操作のものであり、かつフェイルバックＯＳＤノード上にローカルで記録された最大のシーケンス識別子を搬送し、最大のシーケンス識別子は、フェイルバックＯＳＤノード上のｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループのＩＯビューの最新のＩＯビューバージョン番号を含み、データに対する変更操作の最大の通し番号が、最新のＩＯビューバージョン番号に対応するＩＯビュー内のＩＯビューにおけるｐａｒｔｉｔｉｏｎに対応し、障害中に更新されたデータをフェイルバックＯＳＤノードに送信することは、データ回復要求における最大のシーケンス識別子が、一次ＯＳＤノード上にローカルで記憶された現在の最小のシーケンス識別子より小さいと決定すること、一次ＯＳＤ上にローカルで記憶された現在の最小のシーケンス識別子をフェイルバックＯＳＤノードに送信すること、ｐａｒｔｉｔｉｏｎグループに属し、かつ一次ＯＳＤノード上にある一次ｐａｒｔｉｔｉｏｎに対応するすべてのデータを同期させるために、フェイルバックＯＳＤノードによって開始されたデータ回復要求を受信すること、及び一次ｐａｒｔｉｔｉｏｎに対応するすべてのデータをフェイルバックＯＳＤノードに送信することを含み、最小のシーケンス識別子は、ｐａｒｔｉｔｉｏｎグループのＩＯビューのものであり、かつ一次ＯＳＤノード上に記憶された最小のＩＯビューバージョン番号であり、データに対する変更操作の最小の通し番号が、最小のＩＯビューバージョン番号に対応するＩＯビューにおけるｐａｒｔｉｔｉｏｎに対応する。

【００２４】

第４の態様によれば、分散ストレージ及びレプリケーションシステムが提供され、システムは、少なくとも１つのメタデータ制御（ＭＤＣ）モジュールと、複数のＩＯルーティングモジュールと、複数のオブジェクトベースのストレージデバイス（ＯＳＤ）ノードとを含み、ＭＤＣモジュールは、各ＯＳＤノードに関して、各ＯＳＤノードによって管理される物理ストレージリソースに対応する少なくとも１つの論理パーティション（ｐａｒｔｉｔｉｏｎ）を構成し、少なくとも１つのｐａｒｔｉｔｉｏｎは、一次ｐａｒｔｉｔｉｏｎ、二次ｐａｒｔｉｔｉｏｎ、又は一次ｐａｒｔｉｔｉｏｎと二次ｐａｒｔｉｔｉｏｎの任意の組合せであり、一次ｐａｒｔｉｔｉｏｎ、及び一次ｐａｒｔｉｔｉｏｎに対応する二次ｐａｒｔｉｔｉｏｎは、ｐａｒｔｉｔｉｏｎグループを構成し、同一のｐａｒｔｉｔｉｏｎグループの中の一次ｐａｒｔｉｔｉｏｎと二次ｐａｒｔｉｔｉｏｎは、異なるＯＳＤノード上に位置付けられ、一次ｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤノードは、一次ｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループの一次ＯＳＤノードであり、二次ｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤノードは、二次ｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループの二次ＯＳＤノードであり、並びにｐａｒｔｉｔｉｏｎによりｐａｒｔｉｔｉｏｎビューを生成し、ｐａｒｔｉｔｉｏｎビューは、ｐａｒｔｉｔｉｏｎグループの中のｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤについての情報を含む、ように適応され、ＩＯルーティングモジュールは、ＯＳＤノードに受信されたＩＯ要求をルーティングするように適応され、ＯＳＤノードは、ＩＯ要求により、ＩＯ要求に対応するデータの記憶を実行するように適応され、システムは、メモリと、プロセッサとを含み、メモリは、コンピュータ可読命令を記憶するように適応され、命令は、ＭＤＣモジュー

ル、ＩＯルーティングモジュール、及びＯＳＤノードの機能を実行するのに使用され、プロセッサは、メモリに接続され、メモリの中の命令を読み取り、かつ命令により、プロセッサが、システムにおけるＯＳＤノードが障害を有するＯＳＤノードであると決定し、障害を有するＯＳＤノード上のｐａｒｔｉｔｉｏｎを決定し、障害を有するＯＳＤノード上のｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループのｐａｒｔｉｔｉｏｎビューを更新し、及び更新されたｐａｒｔｉｔｉｏｎビューにおけるｐａｒｔｉｔｉｏｎグループが位置付けられる一次ＯＳＤノードに更新通知を送信して、一次ＯＳＤノードが、更新されたｐａｒｔｉｔｉｏｎビューにより、ＩＯ要求に対応するデータのレプリケーションを処理する、ことを可能にするように適応される。

【００２５】

第５の態様によれば、分散ストレージ及びレプリケーションシステムが提供され、システムは、少なくとも１つのメタデータ制御（ＭＤＣ）モジュールと、複数のＩＯルーティングモジュールと、複数のオブジェクトベースのストレージデバイス（ＯＳＤ）ノードとを含み、ＭＤＣモジュールは、各ＯＳＤノードに関して、各ＯＳＤノードによって管理される物理ストレージリソースに対応する少なくとも１つの論理パーティション（ｐａｒｔｉｔｉｏｎ）を構成し、少なくとも１つのｐａｒｔｉｔｉｏｎは、一次ｐａｒｔｉｔｉｏｎ、二次ｐａｒｔｉｔｉｏｎ、又は一次ｐａｒｔｉｔｉｏｎと二次ｐａｒｔｉｔｉｏｎの任意の組合せであり、一次ｐａｒｔｉｔｉｏｎ、及び一次ｐａｒｔｉｔｉｏｎに対応する二次ｐａｒｔｉｔｉｏｎは、ｐａｒｔｉｔｉｏｎグループを構成し、同一のｐａｒｔｉｔｉｏｎグループの中の一次ｐａｒｔｉｔｉｏｎと二次ｐａｒｔｉｔｉｏｎは、異なるＯＳＤノード上に位置付けられ、一次ｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤノードは、一次ｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループの一次ＯＳＤノードであり、二次ｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤノードは、二次ｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループの二次ＯＳＤノードであり、並びにｐａｒｔｉｔｉｏｎによりｐａｒｔｉｔｉｏｎビュー及びＩＯビューを生成し、ｐａｒｔｉｔｉｏｎビューは、ｐａｒｔｉｔｉｏｎグループの中のｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤについての情報を含み、ＩＯビューは、ｐａｒｔｉｔｉｏｎグループの一次ＯＳＤの識別子を含む、ように適応され、ＩＯルーティングモジュールは、ＯＳＤノードに受信されたＩＯ要求をルーティングするように適応され、ＯＳＤノードは、ＩＯ要求により、ＩＯ要求に対応するデータの記憶を実行するように適応され、システムは、メモリと、プロセッサとを含み、メモリは、コンピュータ可読命令を記憶するように適応され、命令は、ＭＤＣモジュール、ＩＯルーティングモジュール、及びＯＳＤノードの機能を実行するのに使用され、プロセッサは、メモリに接続され、メモリの中の命令を読み取り、かつ命令により、プロセッサが、ＩＯルーティングモジュールが、ＩＯ要求を受信し、ＩＯ要求はｋｅｙを含み、ｋｅｙにより、ＩＯ要求に対応するデータが属するｐａｒｔｉｔｉｏｎグループを決定し、及びデータが属するｐａｒｔｉｔｉｏｎグループの一次ＯＳＤノードを決定し、データが属するｐａｒｔｉｔｉｏｎグループのＩＯビューのＩＯビューバージョン情報をＩＯ要求に追加し、及び決定された一次ＯＳＤノードに、ＩＯビューバージョン情報を搬送するＩＯ要求を送信することを可能にすること、一次ＯＳＤノードが、ＩＯ要求を受信し、ＩＯビューバージョン情報により、ＩＯ要求におけるＩＯビューバージョンが、ローカルで記憶されたＩＯビューバージョンと整合性があると決定した後、ＩＯ要求を実行し、ＩＯビューバージョン情報を搬送するレプリケーション要求を生成し、及びデータが属するｐａｒｔｉｔｉｏｎの二次ＯＳＤノードにレプリケーション要求を送信することを可能にすること、並びに二次ＯＳＤノードが、レプリケーション要求を受信し、及びＩＯビューバージョン情報により、レプリケーション要求におけるＩＯビューバージョンが、二次ＯＳＤノード上にローカルで記憶されたＩＯビューバージョンと整合性があると決定した後、レプリケーション要求を実行して、二次ＯＳＤノード上のＩＯ要求に対応するデータが、一次ＯＳＤノード上のＩＯ要求に対応するデータと整合性を保つことを可能にする、ことを可能にするように適応される。

【００２６】

第６の態様によれば、分散ストレージ及びレプリケーションシステムが提供され、システムは、少なくとも１つのメタデータ制御（ＭＤＣ）モジュールと、複数のＩＯルーティングモジュールと、複数のオブジェクトベースのストレージデバイス（ＯＳＤ）ノードとを含み、ＭＤＣモジュールは、各ＯＳＤノードに関して、各ＯＳＤノードによって管理される物理ストレージリソースに対応する少なくとも１つの論理パーティション（ｐａｒｔｉｔｉｏｎ）を構成し、少なくとも１つのｐａｒｔｉｔｉｏｎは、一次ｐａｒｔｉｔｉｏｎ、二次ｐａｒｔｉｔｉｏｎ、又は一次ｐａｒｔｉｔｉｏｎと二次ｐａｒｔｉｔｉｏｎの任意の組合せであり、一次ｐａｒｔｉｔｉｏｎ、及び一次ｐａｒｔｉｔｉｏｎに対応する二次ｐａｒｔｉｔｉｏｎは、ｐａｒｔｉｔｉｏｎグループを構成し、同一のｐａｒｔｉｔｉｏｎグループの中の一次ｐａｒｔｉｔｉｏｎと二次ｐａｒｔｉｔｉｏｎは、異なるＯＳＤノード上に位置付けられ、一次ｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤノードは、一次ｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループの一次ＯＳＤノードであり、二次ｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤノードは、二次ｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループの二次ＯＳＤノードであり、並びにｐａｒｔｉｔｉｏｎによりｐａｒｔｉｔｉｏｎビュー及びＩＯビューを生成し、ｐａｒｔｉｔｉｏｎビューは、ｐａｒｔｉｔｉｏｎグループの中のｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤについての情報を含み、ＩＯビューは、ｐａｒｔｉｔｉｏｎグループの一次ＯＳＤの識別子を含む、ように適応され、ＩＯルーティングモジュールは、ＯＳＤノードに受信されたＩＯ要求をルーティングするように適応され、ＯＳＤノードは、ＩＯ要求により、ＩＯ要求に対応するデータの記憶を実行するように適応され、システムは、メモリと、プロセッサとを含み、メモリは、コンピュータ可読命令を記憶するように適応され、命令は、ＭＤＣモジュール、ＩＯルーティングモジュール、及びＯＳＤノードの機能を実行するのに使用され、プロセッサは、メモリに接続され、メモリの中の命令を読み取り、かつ命令により、プロセッサが、ＯＳＤノードが、フェイルバックの後、ＭＤＣモジュールにＯＳＤノード上のｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループのＩＯビューを要求するクエリ要求を送信し、ＯＳＤノードは、フェイルバックＯＳＤノードと呼ばれ、クエリ要求は、フェイルバックＯＳＤノードのＯＳＤ識別子を搬送し、ＭＤＣによって返されるＩＯビューを受信し、ＩＯビューにおける一次ＯＳＤに対して、障害中にフェイルバックＯＳＤノードによって更新されたデータを回復することを要求するデータ回復要求を開始し、障害中に更新され、かつ一次ＯＳＤによって送信されるデータを受信し、及びｐａｒｔｉｔｉｏｎグループのものであり、かつフェイルバックＯＳＤノードがデータ回復を完了した後にＭＤＣモジュールによって更新されたｐａｒｔｉｔｉｏｎビューにより、ＩＯ要求のレプリケーションを処理することを可能にし、ＭＤＣモジュールが、フェイルバックＯＳＤノードのクエリ要求を受信し、クエリ要求におけるＯＳＤ識別子によりフェイルバックＯＳＤノードにＩＯビューを返し、及びフェイルバックＯＳＤノードがデータ回復を完了した後、ｐａｒｔｉｔｉｏｎグループのｐａｒｔｉｔｉｏｎビューを更新することを可能にすること、並びに一次ＯＳＤノードが、フェイルバックＯＳＤノードのデータ回復要求を受信し、フェイルバックＯＳＤノードに、障害中に更新されたデータを送信し、及びｐａｒｔｉｔｉｏｎグループのものであり、かつフェイルバックＯＳＤノードがデータ回復を完了した後にＭＤＣモジュールによって更新されたｐａｒｔｉｔｉｏｎビューにより、ＩＯ要求に対応するデータのレプリケーションを処理することを可能にすること、を可能にするように適応される。

【００２７】

第７の態様によれば、分散ストレージシステムにおいてデータ記憶及びデータレプリケーションを管理するための方法が提供され、システムは、少なくとも１つのメタデータ制御（ＭＤＣ）モジュールと、複数のＩＯルーティングモジュールと、複数のオブジェクトベースのストレージデバイス（ＯＳＤ）ノードとを含み、ＭＤＣモジュールは、各ＯＳＤノードに関して、各ＯＳＤノードによって管理される物理ストレージリソースに対応する少なくとも１つの論理パーティション（ｐａｒｔｉｔｉｏｎ）を構成し、少なくとも１つのｐａｒｔｉｔｉｏｎは、一次ｐａｒｔｉｔｉｏｎ、二次ｐａｒｔｉｔｉｏｎ、又は一次

ｐａｒｔｉｔｉｏｎと二次ｐａｒｔｉｔｉｏｎの任意の組合せであり、一次ｐａｒｔｉｔｉｏｎ、及び一次ｐａｒｔｉｔｉｏｎに対応する二次ｐａｒｔｉｔｉｏｎは、ｐａｒｔｉｔｉｏｎグループを構成し、同一のｐａｒｔｉｔｉｏｎグループの中の一次ｐａｒｔｉｔｉｏｎと二次ｐａｒｔｉｔｉｏｎは、異なるＯＳＤノード上に位置付けられ、一次ｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤノードは、一次ｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループの一次ＯＳＤノードであり、二次ｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤノードは、二次ｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループの二次ＯＳＤノードであり、並びにｐａｒｔｉｔｉｏｎによりｐａｒｔｉｔｉｏｎビューを生成し、ｐａｒｔｉｔｉｏｎビューは、ｐａｒｔｉｔｉｏｎグループの中のｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤについての情報を含む、ように適応され、ＩＯルーティングモジュールは、ＯＳＤノードに受信されたＩＯ要求をルーティングするように適応され、ＯＳＤノードは、ＩＯ要求により、ＩＯ要求に対応するデータの記憶を実行するように適応され、方法は、システムにおけるＯＳＤノードが障害を有するＯＳＤノードであると決定し、障害を有するＯＳＤノード上のｐａｒｔｉｔｉｏｎを決定し、障害を有するＯＳＤノード上のｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループのｐａｒｔｉｔｉｏｎビューを更新し、及び更新されたｐａｒｔｉｔｉｏｎビューにおける一次ＯＳＤノードに更新通知を送信する工程を含み、一次ＯＳＤノードは、ＭＤＣモジュールによって送信された更新通知を受信した後、更新されたｐａｒｔｉｔｉｏｎビューにより、ＩＯ要求に対応するデータのレプリケーションを処理するように適応される。

【００２８】

第７の態様の第１の可能な実施様態において、ｐａｒｔｉｔｉｏｎビューは、ｐａｒｔｉｔｉｏｎグループの中のｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤのものである一次／二次ＩＤ及び対応するｐａｒｔｉｔｉｏｎステータスを特に含み、一次ＯＳＤノードは、更新されたｐａｒｔｉｔｉｏｎビューにより、一次ＯＳＤノード上にローカルで記憶されたｐａｒｔｉｔｉｏｎビューを更新するようにさらに適応され、更新されたｐａｒｔｉｔｉｏｎビューにより、ＩＯ要求に対応するデータのレプリケーションを処理することは、更新されたローカルで記憶されたｐａｒｔｉｔｉｏｎビューにより、ＩＯルーティングモジュールからのＩＯ要求に対応するデータを、ｐａｒｔｉｔｉｏｎステータスが更新されたローカルで記憶されたｐａｒｔｉｔｉｏｎビューにおいて整合性がある二次ＯＳＤノード上に、或いはｐａｒｔｉｔｉｏｎステータスが更新されたローカルで記憶されたｐａｒｔｉｔｉｏｎビューにおいて整合性がある二次ＯＳＤノード、及びｐａｒｔｉｔｉｏｎステータスが更新されたローカルで記憶されたｐａｒｔｉｔｉｏｎビューにおいて不整合であるが、データを回復している二次ＯＳＤノードの上にレプリケートすることを特に含む。

【００２９】

第７の態様の第１の可能な実施様態を参照して、第２の可能な実施様態において、方法は、障害を有するＯＳＤノード上のｐａｒｔｉｔｉｏｎが一次ｐａｒｔｉｔｉｏｎを含むと決定した場合、ＭＤＣモジュールによる、一次ｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループのＩＯビューを更新し、及び更新されたＩＯビューについて更新されたｐａｒｔｉｔｉｏｎビューにおける二次ＯＳＤノードに通知する工程と、並びに更新されたｐａｒｔｉｔｉｏｎビューにおける二次ＯＳＤノードによる、更新されたＩＯビューにより、ローカルで記憶されたＩＯビューを更新し、及び更新されたローカルで記憶されたＩＯビューにより、ＩＯ要求に対応するデータのレプリケーションを処理する工程と、をさらに含む。

【００３０】

第７の態様の第２の可能な実施様態を参照して、第３の可能な実施様態において、障害を有するＯＳＤノード上のｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループのｐａｒｔｉｔｉｏｎビューを更新することは、障害を有するＯＳＤノード上のｐａｒｔｉｔｉｏｎが二次ｐａｒｔｉｔｉｏｎを含む場合、二次ｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループのｐａｒｔｉｔｉｏｎビューにおいて障害を有するＯＳＤノードのｐａｒ

ｔｉｔｉｏｎステータスに不整合であるという印を付ける工程と、並びに障害を有するＯＳＤノード上のｐａｒｔｉｔｉｏｎが一次ｐａｒｔｉｔｉｏｎを含む場合、一次ｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループのｐａｒｔｉｔｉｏｎビューにおいて一次ＯＳＤノードの役割をする障害を有するＯＳＤノードを新たな二次ＯＳＤノードとして設定し、新たな二次ＯＳＤノードに対応するｐａｒｔｉｔｉｏｎステータスに不整合であるという印を付けること、ｐａｒｔｉｔｉｏｎステータスが整合性がある二次ＯＳＤノードを、一次ｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループのｐａｒｔｉｔｉｏｎビューにおける元の二次ＯＳＤノードから選択し、及び選択された二次ＯＳＤノードを新たな一次ＯＳＤノードとして設定する工程と、を特に含む。

【００３１】

第８の態様によれば、分散ストレージシステムにおいてデータ記憶及びデータレプリケーションを管理するための方法が提供され、システムは、少なくとも１つのメタデータ制御（ＭＤＣ）モジュールと、複数のＩＯルーティングモジュールと、複数のオブジェクトベースのストレージデバイス（ＯＳＤ）ノードとを含み、ＭＤＣモジュールは、各ＯＳＤノードに関して、各ＯＳＤノードによって管理される物理ストレージリソースに対応する少なくとも１つの論理パーティション（ｐａｒｔｉｔｉｏｎ）を構成し、少なくとも１つのｐａｒｔｉｔｉｏｎは、一次ｐａｒｔｉｔｉｏｎ、二次ｐａｒｔｉｔｉｏｎ、又は一次ｐａｒｔｉｔｉｏｎと二次ｐａｒｔｉｔｉｏｎの任意の組合せであり、一次ｐａｒｔｉｔｉｏｎ、及び一次ｐａｒｔｉｔｉｏｎに対応する二次ｐａｒｔｉｔｉｏｎは、ｐａｒｔｉｔｉｏｎグループを構成し、同一のｐａｒｔｉｔｉｏｎグループの中の一次ｐａｒｔｉｔｉｏｎと二次ｐａｒｔｉｔｉｏｎは、異なるＯＳＤノード上に位置付けられ、一次ｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤノードは、一次ｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループの一次ＯＳＤノードであり、二次ｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤノードは、二次ｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループの二次ＯＳＤノードであり、並びにｐａｒｔｉｔｉｏｎによりｐａｒｔｉｔｉｏｎビュー及びＩＯビューを生成し、ｐａｒｔｉｔｉｏｎビューは、ｐａｒｔｉｔｉｏｎグループの中のｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤについての情報を含み、ＩＯビューは、ｐａｒｔｉｔｉｏｎグループの一次ＯＳＤの識別子を含む、ように適応され、ＩＯルーティングモジュールは、ＯＳＤノードに受信されたＩＯ要求をルーティングするように適応され、ＯＳＤノードは、ＩＯ要求によりＩＯデータ記憶を実行するように適応され、方法は、ＩＯルーティングモジュールが、ＩＯ要求を受信し、ＩＯ要求はｋｅｙを含み、ｋｅｙにより、ＩＯ要求に対応するデータが属するｐａｒｔｉｔｉｏｎグループを決定し、及びデータが属するｐａｒｔｉｔｉｏｎグループの一次ＯＳＤノードを決定し、データが属するｐａｒｔｉｔｉｏｎグループのＩＯビューのＩＯビューバージョン情報をＩＯ要求に追加し、及び決定された一次ＯＳＤノードに、ＩＯビューバージョン情報を搬送するＩＯ要求を送信するように適応されること、一次ＯＳＤノードが、ＩＯ要求を受信し、ＩＯビューバージョン情報により、ＩＯ要求におけるＩＯビューバージョンが、一次ＯＳＤノード上にローカルで記憶されたＩＯビューバージョンと整合性があると決定した後、ＩＯ要求を実行し、ＩＯビューバージョン情報を搬送するレプリケーション要求を生成し、及びデータが属するｐａｒｔｉｔｉｏｎグループの二次ＯＳＤノードにレプリケーション要求を送信するように適応されること、二次ＯＳＤノードが、レプリケーション要求を受信し、及びＩＯビューバージョン情報により、レプリケーション要求におけるＩＯビューバージョンが、二次ＯＳＤノード上にローカルで記憶されたＩＯビューバージョンと整合性があると決定した後、レプリケーション要求を実行して、二次ＯＳＤノード上のＩＯ要求に対応するデータが、一次ＯＳＤノード上のＩＯ要求に対応するデータと整合性を保つように適応されること、を含む。

【００３２】

第８の態様の第１の可能な実施様態を参照して、ｐａｒｔｉｔｉｏｎビューは、ｐａｒｔｉｔｉｏｎグループの中のｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤのものである一次／二次ＩＤ及び対応するｐａｒｔｉｔｉｏｎステータスを特に含み、方法は、ＩＯ要求

を処理するプロセスにおいて、一次ＯＳＤノードが障害を有するようになったことを検出した場合、ＭＤＣモジュールによる、データが属するｐａｒｔｉｔｉｏｎグループのｐａｒｔｉｔｉｏｎビューにおける一次ＯＳＤノードを新たな二次ＯＳＤノードとして設定し、及び新たな二次ＯＳＤのｐａｒｔｉｔｉｏｎステータスに不整合であるという印を付ける工程と、データが属するｐａｒｔｉｔｉｏｎグループのｐａｒｔｉｔｉｏｎビューにおける二次ＯＳＤノードの任意の二次ＯＳＤノードを新たな一次ＯＳＤノードとして設定し、データが属するｐａｒｔｉｔｉｏｎグループの更新されたｐａｒｔｉｔｉｏｎビューについて新たな一次ＯＳＤノードに通知する工程と、新たな一次ＯＳＤノードを使用することによって、データが属するｐａｒｔｉｔｉｏｎグループのＩＯビューを更新し、及びデータが属するｐａｒｔｉｔｉｏｎの更新されたＩＯビューについてＩＯルーティングモジュールに通知する工程と、をさらに含み、ＩＯルーティングモジュールは、ＭＤＣモジュールによって送信されたｐａｒｔｉｔｉｏｎグループのものである更新されたＩＯビューを受信し、及びｐａｒｔｉｔｉｏｎグループの更新されたＩＯビューにより新たな一次ＯＳＤノードにＩＯ要求を送信するようにさらに適応され、新たな一次ＯＳＤノードは、ＩＯ要求を受信し、ＩＯ要求を実行した後、第２のレプリケーション要求を生成し、及び第２のレプリケーション要求を、ｐａｒｔｉｔｉｏｎステータスが、データが属するｐａｒｔｉｔｉｏｎグループの更新されたｐａｒｔｉｔｉｏｎビューにおいて整合性のある二次ＯＳＤノードに送信するように適応される。

【００３３】

第２の態様の第２の可能な実施様態を参照して、ｐａｒｔｉｔｉｏｎビューは、ｐａｒｔｉｔｉｏｎグループの中のｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤのものである一次／二次ＩＤ及び対応するｐａｒｔｉｔｉｏｎステータスを特に含み、方法は、ＩＯ要求を処理するプロセスにおいて、二次ＯＳＤノードのうちのいずれかの二次ＯＳＤノードが障害を有するようになったことを検出した場合、ＭＤＣモジュールによる、データが属するｐａｒｔｉｔｉｏｎグループのｐａｒｔｉｔｉｏｎビューにおけるそのいずれかの二次ＯＳＤノードのｐａｒｔｉｔｉｏｎステータスに不整合であるという印を付け、及びデータが属するｐａｒｔｉｔｉｏｎグループの更新されたｐａｒｔｉｔｉｏｎビューについて一次ＯＳＤノードに通知する工程をさらに含み、一次ＯＳＤノードは、データが属するｐａｒｔｉｔｉｏｎグループの更新されたｐａｒｔｉｔｉｏｎビューを受信した後、ｐａｒｔｉｔｉｏｎステータスが、更新されたｐａｒｔｉｔｉｏｎビューにおいて整合性がある二次ＯＳＤノードにレプリケーション要求を送信し、及びｐａｒｔｉｔｉｏｎステータスが不整合である二次ＯＳＤノードにレプリケーション要求を送信することを省くように適応される。

【００３４】

第９の態様によれば、分散ストレージシステムにおいてデータ記憶及びデータレプリケーションを管理するための方法が提供され、システムは、少なくとも１つのメタデータ制御（ＭＤＣ）モジュールと、複数のＩＯルーティングモジュールと、複数のオブジェクトベースのストレージデバイス（ＯＳＤ）ノードとを含み、ＭＤＣモジュールは、各ＯＳＤノードに関して、各ＯＳＤノードによって管理される物理ストレージリソースに対応する少なくとも１つの論理パーティション（ｐａｒｔｉｔｉｏｎ）を構成し、少なくとも１つのｐａｒｔｉｔｉｏｎは、一次ｐａｒｔｉｔｉｏｎ、二次ｐａｒｔｉｔｉｏｎ、又は一次ｐａｒｔｉｔｉｏｎと二次ｐａｒｔｉｔｉｏｎの任意の組合せであり、一次ｐａｒｔｉｔｉｏｎ、及び一次ｐａｒｔｉｔｉｏｎに対応する二次ｐａｒｔｉｔｉｏｎは、ｐａｒｔｉｔｉｏｎグループを構成し、同一のｐａｒｔｉｔｉｏｎグループの中の一次ｐａｒｔｉｔｉｏｎと二次ｐａｒｔｉｔｉｏｎは、異なるＯＳＤノード上に位置付けられ、一次ｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤノードは、一次ｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループの一次ＯＳＤノードであり、二次ｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤノードは、二次ｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループの二次ＯＳＤノードであり、並びにｐａｒｔｉｔｉｏｎによりｐａｒｔｉｔｉｏｎビュー及びＩＯビューを生成し、ｐａｒｔｉｔｉｏｎビューは、ｐａｒｔｉｔｉｏｎグループの中のｐａ

ｒｔｉｔｉｏｎが位置付けられるＯＳＤについての情報を含み、ＩＯビューは、ｐａｒｔｉｔｉｏｎグループの一次ＯＳＤの識別子を含む、ように適応され、ＩＯルーティングモジュールは、ＯＳＤノードに受信されたＩＯ要求をルーティングするように適応され、ＯＳＤノードは、ＩＯ要求により、ＩＯ要求に対応するデータの記憶を実行するように適応され、方法は、ＯＳＤノードが、フェイルバックの後、ＭＤＣモジュールにＯＳＤノード上のｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループのＩＯビューを要求するクエリ要求を送信し、ＯＳＤノードは、フェイルバックＯＳＤノードと呼ばれ、クエリ要求は、フェイルバックＯＳＤノードのＯＳＤ識別子を搬送し、ＭＤＣによって返されるＩＯビューを受信し、ＩＯビューにおける一次ＯＳＤに対して、障害中にフェイルバックＯＳＤノードによって更新されたデータを回復することを要求するデータ回復要求を開始し、障害中に更新され、かつ一次ＯＳＤによって送信されるデータを受信し、及びｐａｒｔｉｔｉｏｎグループのものであり、かつフェイルバックＯＳＤノードがデータ回復を完了した後にＭＤＣモジュールによって更新されたｐａｒｔｉｔｉｏｎビューにより、ＩＯ要求のレプリケーションを処理するように適応されること、ＭＤＣモジュールが、フェイルバックＯＳＤノードのクエリ要求を受信し、クエリ要求におけるＯＳＤ識別子によりフェイルバックＯＳＤノードにＩＯビューを返し、及びフェイルバックＯＳＤノードがデータ回復を完了した後、ｐａｒｔｉｔｉｏｎグループのｐａｒｔｉｔｉｏｎビューを更新するように適応されること、及び一次ＯＳＤノードが、フェイルバックＯＳＤノードのデータ回復要求を受信し、フェイルバックＯＳＤノードに、障害中に更新されたデータを送信し、及びｐａｒｔｉｔｉｏｎグループのものであり、かつフェイルバックＯＳＤノードがデータ回復を完了した後にＭＤＣモジュールによって更新されたｐａｒｔｉｔｉｏｎビューにより、ＩＯ要求に対応するデータのレプリケーションを処理するように適応されること、を含む。

【００３５】

本発明の実施形態における技術的ソリューションをより明確に説明するのに、以下に、本発明を説明するのに要求される添付の図面について簡単に説明する。明らかに、以下の説明における添付の図面は、本発明のいくつかの実施形態を示すに過ぎず、当業者は、それでも、創造的な努力なしにこれらの添付の図面から他の図面を導き出すことが可能である。

【図面の簡単な説明】

【００３６】

【図１】従来技術における２フェーズコミットプロトコルを示す流れ図である。

【図２Ａ】本発明の実施形態による分散ストレージ及びレプリケーションシステムを示す概略アーキテクチャ図である。

【図２Ｂ】本発明の別の実施形態による分散ストレージ及びレプリケーションシステムを示す概略アーキテクチャ図である。

【図２Ｃ】本発明の別の実施形態による分散ストレージ及びレプリケーションシステムを示す概略アーキテクチャ図である。

【図３】本発明の実施形態によるクラスタビューを示す概略図である。

【図４】本発明の実施形態によるＯＳＤビューステータス遷移を示す概略図である。

【図５】本発明の別の実施形態による分散ストレージ及びレプリケーションシステムを示す概略構造図である。

【図６】本発明の実施形態によるビュー初期化を示す流れ図である。

【図７Ａ】本発明の実施形態によるＩＯ要求処理を示す流れ図である。

【図７Ｂ】本発明の実施形態によるＩＯ要求処理を示す流れ図である。

【図８Ａ】本発明の実施形態によるＯＳＤ障害処理を示す流れ図である。

【図８Ｂ】本発明の実施形態によるＯＳＤ障害処理を示す流れ図である。

【図９Ａ】本発明の実施形態によるＯＳＤフェイルバック処理を示す流れ図である。

【図９Ｂ】本発明の実施形態によるＯＳＤフェイルバック処理を示す流れ図である。

【図９Ｃ】本発明の実施形態によるＯＳＤフェイルバック処理を示す流れ図である。

【図１０Ａ】本発明の実施形態によるＯＳＤフェイルバック処理プロセスにおけるデータ回復を示す流れ図である。
【図１０Ｂ】本発明の実施形態によるＯＳＤフェイルバック処理プロセスにおけるデータ回復を示す流れ図である。
【図１１Ａ】本発明の実施形態によるＯＳＤがクラスタから出た後に実行される処理を示す流れ図である。
【図１１Ｂ】本発明の実施形態によるＯＳＤがクラスタから出た後に実行される処理を示す流れ図である。
【図１２Ａ】本発明の実施形態による新たなＯＳＤがクラスタに参加した後に実行される処理を示す流れ図である。
【図１２Ｂ】本発明の実施形態による新たなＯＳＤがクラスタに参加した後に実行される処理を示す流れ図である。
【発明を実施するための形態】
【００３７】
以下に、本発明の実施形態における添付の図面を参照して、本発明の実施形態における技術的ソリューションを明確に、かつ完全に説明する。明らかに、説明される実施形態は、本発明の実施形態のすべてではなく、いくつかに過ぎない。創造的な努力なしに本発明の実施形態に基づいて当業者によって得られる他のすべての実施形態は、本発明の保護範囲に入るべきものとする。
【００３８】
図２Ａに示されるとおり、本発明の特定の実施形態は、本発明のこの実施形態において言及されるデータ記憶及びレプリケーションを管理及び制御するために、分散ストレージ及びレプリケーション制御システムを提供する。分散ストレージ及びレプリケーション制御システムは、３つの副層、すなわち、ステータス層、インターフェース層、及びデータ層を主に含む。ステータス層は、メタデータ制御Ｍｅｔａ　Ｄａｔａ　Ｃｏｎｔｒｏｌｌｅｒ（ＭＤＣ）モジュール２０２を含み、実際の応用において、ＭＤＣの二次ＭＤＣが構成される必要があるかどうか、及び二次ＭＤＣの数量は、要件に応じて決定されることが可能であり、二次ＭＤＣは、一次ＭＤＣモジュールが障害を有するようになった場合、一次ＭＤＣの役割をするように適応される。インターフェース層は、複数のＩＯルーティングモジュール（Ｉｎｐｕｔ／Ｏｕｔｐｕｔ入出力ルーティングモジュール）２０４（クライアントとも呼ばれ、この２つの概念は、本発明の実施において互換的であり得る）を含む。データ層は、複数のオブジェクトベースのストレージデバイスＯｂｊｅｃｔ　Ｓｔｏｒａｇｅ　Ｄｅｖｉｃｅ（ＯＳＤ）ノード２０６を含む。ステータス層は、ステータスビューメッセージを使用することによってインターフェース層及びデータ層と通信する。例えば、ＭＤＣモジュール２０２が、ステータスビューメッセージを使用することによってＩＯルーティングモジュール２０４及びＯＳＤノード２０６に更新通知を送信して、ＩＯルーティングモジュール２０４及びＯＳＤノード２０６がローカルクラスタビュー（ｖｉｅｗと呼ばれることも可能であり、この２つの概念は、本発明の実施において互換的であり得る）を更新するようにする、又はＭＤＣモジュール２０２によって生成された、若しくは更新されたクラスタビューをＩＯルーティングモジュール２０４及びＯＳＤノード２０６に直接に送信するように指示する。インターフェース層とデータ層は、サービスメッセージを使用することによって互いに通信する。例えば、ＩＯルーティングモジュール２０４が、ＩＯデータ記憶及びレプリケーションを要求するＩＯ要求メッセージをＯＳＤノード２０６に送信する。
【００３９】
ＭＤＣモジュール２０２は、クラスタ構成情報の供給のための入口として、各ＯＳＤノードにアプリケーションストレージスペースにおける論理ストレージリソースの論理パーティション（ｐａｒｔｉｔｉｏｎ）を割り当て、ｐａｒｔｉｔｉｏｎによりクラスタビューを生成し、クラスタビューを維持し、かつ更新し、及びクラスタビュー更新について、対応するＩＯルーティングモジュール２０４及びＯＳＤノード２０６に通知するように適

応される。
【００４０】
ＩＯルーティングモジュール２０４は、クラスタビューにより対応するＯＳＤノードに上位層アプリケーションのＩＯ要求をルーティングするように、かつ転送するように適応される。
【００４１】
ＯＳＤノード２０６は、データバックアップ整合性を実現するために、クラスタビューによりＩＯ要求に対して関連するＩＯ操作を実行し、関連するＩＯ操作は、データを記憶しかつレプリケートすることを含み、及びＯＳＤノード２０６によって管理される物理ストレージリソース（例えば、ローカルディスク又は外部ストレージリソース）上でデータ操作を編成する、ように適応される。
【００４２】
前述のＭＤＣモジュール、ＩＯルーティングモジュール、及びＯＳＤノードは、ハードウェア、ファームウェア、ソフトウェア、又は以上の組合せによって実施され得ることが理解され得る。実際のアプリケーションにおいて、特定の実施様態は、製品設計要件又は製造費用を考慮して決定され、本発明は、特定の実施様態に限定されるべきものではない。
【００４３】
本発明の特定の実施形態において、分散ストレージ及びレプリケーションシステム全体が、独立したプラットフォーム又はサーバ（例えば、前述の図２Ｂにおける分散ストレージ及びレプリケーション制御プラットフォーム）に配置されて、そのプラットフォーム又はサーバに接続された分散ストレージシステムにおけるデータレプリケーション及びデータ記憶を管理するようにする。
【００４４】
本発明の別の特定の実施形態において、分散ストレージ及びレプリケーション制御システムは、図２Ｃに示される分散ストレージシステムにおいて分散された様態で配置され得る。分散ストレージシステムは、複数のサーバ又はホストを含み、この実施形態におけるホスト又はサーバは、物理ホスト又は物理サーバであり、すなわち、プロセッサ及びメモリなどのハードウェアを含む。前述のＭＤＣモジュール２０２は、分散ストレージシステムにおいて１つだけのサーバ若しくはホストの上に（二次ＭＤＣなし）、又は２つのサーバ若しくはホストの上に（１つの一次ＭＤＣモジュール、及び１つの二次ＭＤＣモジュール）、又は３つのサーバ若しくはホスト上に（１つの一次ＭＤＣモジュール、及び２つの二次ＭＤＣモジュール）配置されることが可能である。ＩＯルーティングモジュール２０４は、分散ストレージシステムにおいて各サーバ又は各ホストの上に配置される。ＯＳＤノード２０６は、分散ストレージシステムにおいてストレージリソースを有する各サーバ又は各ホストの上に配置されて、ローカルストレージリソース又は外部ストレージリソースを管理し、かつ制御するようにする。実際の応用形態において、ＩＯルーティングモジュール又はＯＳＤノード、或いはＩＯルーティングモジュールとＯＳＤノードの両方が、１つのホスト上に配置されてもよく、特定の配置様態は、本発明において限定されない、実際の特定の状況により決定されることが可能である。図２ＣにおけるＭＤＣモジュール２０２、ＩＯルーティングモジュール２０４、及びＯＳＤノード２０６が、図２Ｂに示される分散ストレージシステムにおいて分散レプリケーションプロトコル層と呼ばれる分散ストレージ制御システムを構成する。分散ストレージシステムは、分散レプリケーションプロトコル層を使用することによってストレージ層におけるストレージリソースに対するＩＯデータ記憶及びレプリケーションを制御する。ストレージ層は、複数のサーバ又はホストの上のローカルストレージリソースを含み、分散レプリケーションプロトコル層にあり、かつサーバ又はホストの上に分散されたモジュールは、ネットワーク層における交換データネットワークを使用することによって互いに対話する。特定の実施様態において、イーサネット又はｉｎｆｉｎｉｂａｎｄが使用されてもよい。前述のイーサネット又はｉｎｆｉｎｉｂａｎｄは、本発明のこの実施形態において限定されない、本発明のこの実施

形態において使用される高速交換データネットワークの例示的な実施様態に過ぎないことを理解されたい。

【００４５】

以下に、特定の実施形態及び実施様態を使用して、前述の分散ストレージ及びレプリケーション制御システムにおけるＭＤＣモジュール２０２、ＩＯルーティングモジュール２０４、及びＯＳＤノード２０６の接続及び対話、特定の機能などについて詳細に説明する。

【００４６】

本発明の特定の実施形態において、ＭＤＣモジュールのパーティション機能は、ＭＤＣモジュールが、各ＯＳＤノードによって管理される物理ストレージリソースのステータスにより、各ＯＳＤノードに関して、各ＯＳＤノードによって管理される物理ストレージリソースに対応する論理パーティション（ｐａｒｔｉｔｉｏｎ）を構成することを特に含み得る。ｐａｒｔｉｔｉｏｎは、アプリケーションストレージスペースにおける特定の数量のデータブロックを含む。ストレージ層における物理ストレージスペースと比べて、アプリケーション層におけるアプリケーションストレージスペースは、アプリケーション層によってユーザに割り当てられた特定の量の論理ストレージスペースであり、ストレージ層における物理ストレージスペースの論理マッピングである。すなわち、本明細書におけるｐａｒｔｉｔｉｏｎの概念は、物理ストレージスペースパーティションの概念とは異なる。データが記憶される際、アプリケーションストレージスペースにおける１つのｐａｒｔｉｔｉｏｎのスペースが、物理ストレージスペースにおける１つ又は複数のパーティションにマップされることが可能である。ｐａｒｔｉｔｉｏｎの特定の粒度は、クラスタ構成情報から獲得されることが可能であり、又は特定の規則によりＭＤＣモジュールによって決定されることが可能であり、又は別の様態によって決定されることが可能であり、このことは、本発明において限定されない。

【００４７】

特定の実施様態において、ＭＤＣモジュールは、ｐａｒｔｉｔｉｏｎサイズ構成情報、ローカルストレージリソースステータス、及び外部ストレージリソースステータス（例えば、通常どおりにアクセスされるＳＡＮ（Ｓｔｏｒａｇｅ Ａｒｅａ Ｎｅｔｗｏｒｋ，ストレージエリアネットワーク）のＬＵＮ（Ｌｏｇｉｃａｌ Ｕｎｉｔ Ｎｕｍｂｅｒ，論理単位番号）情報）などの情報によりｐａｒｔｉｔｉｏｎのクラスタビューを生成することが可能である。

【００４８】

一般に、データ信頼性及びデータ利用可能性を確実にするのに、ｐａｒｔｉｔｉｏｎは、異なるＯＳＤノード上に記憶されたレプリカを有し、ｐａｒｔｉｔｉｏｎレプリカ数量は、構成ファイルを使用することによって構成されることが可能であり、又は特定のアルゴリズムによりＭＤＣによって決定されることが可能である。分類により一次ｐａｒｔｉｔｉｏｎ及び二次ｐａｒｔｉｔｉｏｎが存在する。すなわち、ｐａｒｔｉｔｉｏｎの複数のレプリカから１つのレプリカが一次レプリカとして選択され、その一次レプリカが一次ｐａｒｔｉｔｉｏｎと呼ばれ、ｐａｒｔｉｔｉｏｎの一次レプリカ以外のレプリカは、二次ｐａｒｔｉｔｉｏｎと呼ばれる。一次ｐａｒｔｉｔｉｏｎ、及び一次ｐａｒｔｉｔｉｏｎに対応する二次ｐａｒｔｉｔｉｏｎは、ｐａｒｔｉｔｉｏｎグループを構成し、一次ｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤノードが、一次ｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループの一次ＯＳＤノードと呼ばれ、この実施形態において説明される一次ＯＳＤは、ｐａｒｔｉｔｉｏｎグループの一次ＯＳＤを指し、二次ｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤノードが、二次ｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループの二次ＯＳＤノードと呼ばれ、この実施形態において説明される二次ＯＳＤは、ｐａｒｔｉｔｉｏｎグループの二次ＯＳＤを指す。

【００４９】

理解を容易にするのに、以下に、図２Ｃにおいて提供される実施形態を参照してさらなる説明を提供する。図２Ｃに示されるとおり、ホスト又はサーバｓｅｒｖｅｒ＿１（本発

明のこの実施形態におけるホスト及びサーバの概念は、互換的であり得る）の上でＯＳＤによって管理されるストレージリソースが、ｐａｒｔｉｔｉｏｎ１、ｐａｒｔｉｔｉｏｎ２、及びｐａｒｔｉｔｉｏｎ３（略してＰ１、Ｐ２、Ｐ３）、並びにｐａｒｔｉｔｉｏｎ４’、ｐａｒｔｉｔｉｏｎ５’、及びｐａｒｔｉｔｉｏｎ６’（略してＰ４’、Ｐ５’、Ｐ６’）に分割され、Ｐ４’、Ｐ５’、Ｐ６’は、それぞれ、サーバｓｅｒｖｅｒ＿２上のＯＳＤノード上のｐａｒｔｉｔｉｏｎ４、ｐａｒｔｉｔｉｏｎ５、及びｐａｒｔｉｔｉｏｎ６（略してＰ４、Ｐ５、及びＰ６）のレプリカである。ｓｅｒｖｅｒ＿１上のＯＳＤ上のｐａｒｔｉｔｉｏｎとストレージ層における物理ストレージリソースの間に対応するマッピング関係が存在する。例えば、ＯＳＤ上の１つのｐａｒｔｉｔｉｏｎのスペースが、物理ストレージスペースにおける１つ又は複数のＢｌｏｃｋにマップされる。

【００５０】

ホスト又はサーバｓｅｒｖｅｒ＿１の上のＯＳＤが、一次ｐａｒｔｉｔｉｏｎ（Ｐ１、Ｐ２、Ｐ３）及び二次ｐａｒｔｉｔｉｏｎ（Ｐ４’、Ｐ５’、Ｐ６’）を管理し、ＯＳＤは、別々に、Ｐ１及びＰ１’を含むｐａｒｔｉｔｉｏｎグループの一次ＯＳＤノード、Ｐ２及びＰ２’を含むｐａｒｔｉｔｉｏｎグループの一次ＯＳＤノード、並びにＰ３及びＰ３’を含むｐａｒｔｉｔｉｏｎグループの一次ＯＳＤノードであり、さらに、ＯＳＤは、別々に、Ｐ４及びＰ４’を含むｐａｒｔｉｔｉｏｎの二次ＯＳＤノード、Ｐ５及びＰ５’を含むｐａｒｔｉｔｉｏｎの二次ＯＳＤノード、並びにＰ６及びＰ６’を含むｐａｒｔｉｔｉｏｎの二次ＯＳＤノードである。異なるｐａｒｔｉｔｉｏｎグループに関して、同一のＯＳＤノードが、同時に一次ＯＳＤノードと二次ＯＳＤノードの役割をすることが可能であることが知られ得る。

【００５１】

前述のパーティション、及び対応するレプリカは、以下の要因により設定されることが可能であり、特定の応用形態において、別の要因がディスクパーティションを設定するのに、かつ計画するのに考慮に入れられることが可能である。

【００５２】

第１に、データセキュリティであり、すなわち、各パーティションのレプリカが可能な限り異なるホスト又はサーバに分散されるべきである。データセキュリティの要点は、ｐａｒｔｉｔｉｏｎの複数のレプリカが同一のホスト又はサーバの上に配置されることを許されないことである。第２に、データ平衡であり、各ＯＳＤ上のｐａｒｔｉｔｉｏｎの数量が可能な限り同一に保たれる。各ＯＳＤ上の一次ｐａｒｔｉｔｉｏｎの数量、二次ｐａｒｔｉｔｉｏｎ１の数量、及び二次ｐａｒｔｉｔｉｏｎ２の数量は、可能な限り同一に保たれて、すべてのＯＳＤ上で処理されるサービスが平衡されて、ホットスポットが出現しないようにされる。第３に、データ拡散であり、すなわち、各ＯＳＤ上のｐａｒｔｉｔｉｏｎのレプリカが、可能な限り均等に異なる他のＯＳＤに分散されるべきであり、同一の要件が、より高レベルの物理構成要素にも当てはまる。

【００５３】

図３に示されるとおり、本発明の特定の実施形態において、ＭＤＣモジュールがクラスタビュー情報を生成することは、ＭＤＣが、管理者によって供給されるクラスタ構成情報、及びパーティション状況によりクラスタビュー情報を生成することを特に含むことが可能である。特に、クラスタビュー情報は、３つの次元におけるクラスタビュー、すなわち、ＯＳＤビュー（ＯＳＤ　ｖｉｅｗ）、ＩＯビュー（ＩＯ　ｖｉｅｗ）、及びＰａｒｔｉｔｉｏｎビュー（ｐａｒｔｉｔｉｏｎ　ｖｉｅｗ）を含む。

【００５４】

ＯＳＤ　ｖｉｅｗは、クラスタにおけるＯＳＤノードのステータス情報を含む。特定の実施様態において、ＯＳＤ　ｖｉｅｗは、ＯＳＤノードのＩＤと、ＯＳＤノードのステータス情報とを含むことが可能であり、ＯＳＤ　ＩＤは、ＯＳＤマーカ又はＯＳＤ番号である。図４に示される本発明の実施形態において、ＯＳＤステータスは、ＯＳＤが障害を有するかどうかにより、「アップ（ＵＰ）」状態と「ダウン（ＤＯＷＮ）」状態に特に分類されること、及びＯＳＤがクラスタから出るかどうかにより、「アウト（ＯＵＴ）」状態

と「イン（ＩＮ）」状態に分類されることが可能である。図４に示されるとおり、特定のステータス遷移は、フェイルバックの後、ＯＳＤノードが初期化され、又は再起動され、次に、「イン（ＩＮ）」かつ「ダウン（ＤＯＷＮ）」状態から「イン（ＩＮ）」かつ「アップ（ＵＰ）」状態に遷移することを含む。ＯＳＤの障害が特定の閾値を超えて（例えば、５分を超えて）続く場合、ＯＳＤノードは、クラスタから追放され、相応するように、ＯＳＤノードは、「イン（ＩＮ）」かつ「ダウン（ＤＯＷＮ）」状態から「アウト（ＯＵＴ）」かつ「ダウン（ＤＯＷＮ）」状態に遷移する。本発明の特定の実施形態において、ＯＳＤ ｖｉｅｗは、ＯＳＤ ｖｉｅｗバージョン番号、ＯＳＤ ｖｉｅｗ ＩＤ、又はビューバージョンに印を付ける他の任意の情報などのＯＳＤ ｖｉｅｗバージョン情報をさらに含むことが可能である。

【００５５】

ＩＯ ｖｉｅｗは、ｐａｒｔｉｔｉｏｎグループの一次ＯＳＤノードを識別する識別子を含む。特定の実施様態において、ＩＯ ｖｉｅｗは、ｐａｒｔｉｔｉｏｎグループＩＤと、ｐａｒｔｉｔｉｏｎグループＩＤに対応するｐａｒｔｉｔｉｏｎグループの一次ＯＳＤノードの識別子とを含むことが可能である。各ＩＯ ｖｉｅｗは、ＩＯ ｖｉｅｗを識別するＩＯ ｖｉｅｗバージョン情報を有し、ＩＯ ｖｉｅｗバージョン情報は、ＩＯ ｖｉｅｗのバージョンを識別して、異なるモジュールがＩＯ ｖｉｅｗバージョンを比較するのを助けるようにするのに使用されるＩＯ ｖｉｅｗ ＩＤ（ＩＯ ｖｉｅｗバージョン番号とも呼ばれ得る）であることが可能である。特定の実施形態において、ＩＯ ｖｉｅｗバージョン情報は、ＩＯ ｖｉｅｗに含められても、ＩＯ ｖｉｅｗから除外されてもよい。

【００５６】

ｐａｒｔｉｔｉｏｎ ｖｉｅｗは、ｐａｒｔｉｔｉｏｎグループにおけるｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤについての情報を含む。特定の実施様態において、ｐａｒｔｉｔｉｏｎ ｖｉｅｗは、ｐａｒｔｉｔｉｏｎグループＩＤ、ｐａｒｔｉｔｉｏｎグループＩＤに対応するｐａｒｔｉｔｉｏｎグループにおける各パーティションが位置付けられるＯＳＤ、及びそのＯＳＤの一次／二次ＩＤ、並びに各パーティションのＯＳＤに対応するｐａｒｔｉｔｉｏｎステータスを含むことが可能である。ｐａｒｔｉｔｉｏｎ ｖｉｅｗは、一次ｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤノードについての情報（ＯＳＤノードＩＤ、ＯＳＤノードの一次／二次ＩＤ、及び一次ｐａｒｔｉｔｉｏｎのＯＳＤノードに対応するｐａｒｔｉｔｉｏｎステータスなどの）と、一次ｐａｒｔｉｔｉｏｎに対応する二次ｐａｒｔｉｔｉｏｎ（１つ又は複数の二次ｐａｒｔｉｔｉｏｎが存在することが可能である）が位置付けられるＯＳＤノードについての情報（ＯＳＤノードＩＤ、ＯＳＤノードの一次／二次ＩＤ、及びそのＯＳＤに対応する二次ｐａｒｔｉｔｉｏｎのｐａｒｔｉｔｉｏｎステータスなどの）とを含む。特定の実施形態において、ｐａｒｔｉｔｉｏｎステータスは、２つのタイプ、すなわち、「整合性がある」と「不整合である」に分類されることが可能であり、「整合性がある」は、二次ｐａｒｔｉｔｉｏｎにおけるデータが一次ｐａｒｔｉｔｉｏｎにおけるものと整合性があることを示し、「不整合である」は、二次ｐａｒｔｉｔｉｏｎにおけるデータが一次ｐａｒｔｉｔｉｏｎにおけるものと不整合であり得ることを示す。各ｐａｒｔｉｔｉｏｎ ｖｉｅｗは、ｐａｒｔｉｔｉｏｎ ｖｉｅｗを識別するｐａｒｔｉｔｉｏｎ ｖｉｅｗバージョン情報を有し、ｐａｒｔｉｔｉｏｎ ｖｉｅｗバージョン情報は、モジュールが ｖｉｅｗを比較するようにｐａｒｔｉｔｉｏｎ ｖｉｅｗ ＩＤ（ｐａｒｔｉｔｉｏｎ ｖｉｅｗバージョン番号とも呼ばれ得る）であり得る。特定の実施形態において、ｐａｒｔｉｔｉｏｎ ｖｉｅｗバージョン情報は、ｐａｒｔｉｔｉｏｎ ｖｉｅｗに含められても、ｐａｒｔｉｔｉｏｎ ｖｉｅｗから除外されてもよい。特定の実施形態において、ＩＯ ｖｉｅｗは、ｐａｒｔｉｔｉｏｎ ｖｉｅｗのサブセットである、すなわち、ｐａｒｔｉｔｉｏｎ ｖｉｅｗは、ＩＯ ｖｉｅｗ情報を含むので、ｐａｒｔｉｔｉｏｎ ｖｉｅｗは、ＩＯ ｖｉｅｗバージョン情報をさらに含むことが可能である。

【００５７】

ＭＤＣは、クラスタビューを維持し、管理し、かつ更新し、障害、フェイルバック、障害の後にクラスタから出て、フェイルバックの後にクラスタに再び参加し、及びクラスタに新たに参加することなどのＯＳＤノードステータスにより、クラスタビューを更新し、並びにクラスタビュー更新について関連するモジュールに通知して、関連するモジュールが、更新されたクラスタビューにより、対応するＩＯ要求に対応するデータのレプリケーションを処理する、ようにさらに適応される。

【００５８】

特定の実施様態において、対話を減らし、かつ管理リソース及びストレージリソースを節約するのに、ＯＳＤ　ｖｉｅｗが、ＭＤＣ上にだけ存在することが可能であり、ｐａｒｔｉｔｉｏｎ　ｖｉｅｗが、ＭＤＣモジュール及び一次ＯＳＤノードの上にだけ存在することが可能であり、ＩＯ　ｖｉｅｗが、ＭＤＣモジュール、ＩＯルーティングモジュール、一次ＯＳＤノード、及び二次ＯＳＤノードの上に存在する。ＭＤＣモジュールは、ｐａｒｔｉｔｉｏｎ　ｖｉｅｗにおけるｐａｒｔｉｔｉｏｎが位置付けられる一次ＯＳＤノードにだけｐａｒｔｉｔｉｏｎ　ｖｉｅｗを送信し、又はｐａｒｔｉｔｉｏｎ　ｖｉｅｗにおけるｐａｒｔｉｔｉｏｎが位置付けられる一次ＯＳＤノードにだけローカルｐａｒｔｉｔｉｏｎ　ｖｉｅｗを更新するよう指示し、ｐａｒｔｉｔｉｏｎ　ｖｉｅｗの一部分を構成するＩＯ　ｖｉｅｗ（すなわち、ＩＯ　ｖｉｅｗは、ｐａｒｔｉｔｉｏｎ　ｖｉｅｗのサブｖｉｅｗと見なされ得る）をＩＯルーティングモジュール、一次ＯＳＤノード、及び二次ＯＳＤノードに送信し、又は対応するモジュールにローカルで記憶されたＩＯ　ｖｉｅｗを更新するよう指示する。特定の実施プロセスに関しては、後段の特定の手順、及びＯＳＤがクラスタに参加する手順を参照されたい。実際の応用形態において、ＭＤＣモジュールは、構成情報若しくは特定のポリシーにより、かつクラスタビューの基本的機能により、クラスタビューを異なる形態で設定してもよく、このことは、本発明のこの実施形態において限定されない。

【００５９】

本発明の特定の実施形態において、ＩＯルーティングモジュールは、ＩＯ要求ルーティング機能を実施するように主に適応される。ＩＯルーティングモジュールは、ＭＤＣモジュールからクラスタにおけるすべてのｐａｒｔｉｔｉｏｎのＩＯ　ｖｉｅｗを獲得し、ＩＯ　ｖｉｅｗをキャッシュする。サービスＩＯ要求がＩＯルーティングモジュールに着信すると、ＩＯルーティングモジュールは、ＩＯ要求の中のｋｅｙを使用することによって計算を用いて（ハッシュアルゴリズム又は別のアルゴリズムが計算方法において使用され得る）、ＩＯが属するｐａｒｔｉｔｉｏｎグループを獲得し、次に、ローカルで記憶されたＩＯ　ｖｉｅｗを探索して、ｐａｒｔｉｔｉｏｎグループに対応する一次ＯＳＤノードを見出し、一次ＯＳＤノードにＩＯ要求を送信する。ＩＯルーティングモジュールは、ＭＤＣモジュールから受信されたＩＯ　ｖｉｅｗ更新通知を処理し、更新通知は、例えば、更新されるべきコンテンツを示す更新されたＩＯ　ｖｉｅｗ又は対応する更新インジケーション情報を含み得、更新通知によりローカルで記憶されたＩＯ　ｖｉｅｗを更新し、及び更新されたローカルで記憶されたＩＯ　ｖｉｅｗによりＩＯ要求をルーティングする。特定の実施プロセスに関しては、以下の特定の手順を参照されたい。

【００６０】

本発明の特定の実施形態において、ＯＳＤノードが、クラスタビューによりＩＯ要求を処理してＩＯ操作を実行することは、以下を特に含む。すなわち、

【００６１】

ＯＳＤノードが一次ＯＳＤノードの役割をする場合、一次ＯＳＤノードは、ＩＯルーティングモジュールによって送信されたＩＯ要求を受信すること、ＩＯ要求を実行すること、及び対応する二次ＯＳＤノードにレプリケーション要求を送信して、ＩＯデータ記憶及びレプリケーションを実行するように主に適応される。一次ＯＳＤノードは、ＭＤＣモジュールから一次ＯＳＤノード上のｐａｒｔｉｔｉｏｎのｐａｒｔｉｔｉｏｎ　ｖｉｅｗを受信し、ｐａｒｔｉｔｉｏｎ　ｖｉｅｗを記憶する。一次ＯＳＤノードは、ｐａｒｔｉｔｉｏｎ　ｖｉｅｗによりＩＯ要求のレプリケーションを処理する。一次ＯＳＤノードはさ

らに、ＭＤＣモジュールから、ｐａｒｔｉｔｉｏｎ　ｖｉｅｗについての更新通知を受信し、更新通知によりローカルで記憶されたｐａｒｔｉｔｉｏｎ　ｖｉｅｗを更新し、及び更新されたｐａｒｔｉｔｉｏｎ　ｖｉｅｗにより、ＩＯ要求に対応するデータのレプリケーションを処理し、更新通知は、更新されたｐａｒｔｉｔｉｏｎ　ｖｉｅｗ又は対応する更新情報を含むことが可能であり、それにより、ＯＳＤノードが、更新されたｐａｒｔｉｔｉｏｎ　ｖｉｅｗ又は更新情報によりローカルで記憶されたｐａｒｔｉｔｉｏｎ　ｖｉｅｗ及びＩＯ　ｖｉｅｗを更新するようになっている。ＯＳＤノードが二次ＯＳＤノードの役割を果たす場合、二次ＯＳＤノードが、一次ＯＳＤノードのレプリケーション要求を受信し、レプリケーション要求によりデータレプリケーション及びデータバックアップを実行し、ＭＤＣモジュールから、二次ＯＳＤノード上のデータが属するｐａｒｔｉｔｉｏｎのＩＯ　ｖｉｅｗを受信し、ＩＯ　ｖｉｅｗを記憶し、ＩＯ　ｖｉｅｗにより、ＩＯ要求に対応するデータのレプリケーションを処理し、ＭＤＣモジュールから、ＩＯ　ｖｉｅｗについての更新通知をさらに受信し、更新通知によりローカルで記憶されたＩＯ　ｖｉｅｗを更新し、更新されたＩＯ　ｖｉｅｗにより、ＩＯ要求に対応するデータのレプリケーションを処理するように適応される。特定の実施プロセスに関しては、後段の特定の手順を参照されたい。

【００６２】

本発明の特定の実施形態において、前述の実施形態における分散ストレージ及びレプリケーションシステム（図２Ａ、図２Ｂ、及び図２Ｃに示される）が、図５に示されるシステムに基づいて実施され得る。図５に示されるとおり、システムは、１つ又は複数のメモリ５０２、１つ又は複数の通信インターフェース５０４、及び１つ又は複数のプロセッサ５０６、或いは別のデータ対話ネットワーク（複数のプロセッサとメモリの間の対話のために使用され、図示されない）を含むことが可能である。

【００６３】

メモリ５０２は、読取り専用メモリ（Ｒｅａｄ　Ｏｎｌｙ　Ｍｅｍｏｒｙ，ＲＯＭ）、スタティックストレージデバイス、ダイナミックストレージデバイス、又はランダムアクセスメモリ（Ｒａｎｄｏｍ　Ａｃｃｅｓｓ　Ｍｅｍｏｒｙ，ＲＡＭ）などの様々なタイプのメモリであることが可能である。メモリ５０２は、オペレーティングシステム、並びに別のアプリケーションプログラムの命令及びアプリケーションデータを記憶することが可能であり、命令は、本発明の様々な実施形態におけるＭＤＣモジュール、ＩＯルーティングモジュール、及びＯＳＤノードの機能を実行するために使用される命令を含む。

【００６４】

メモリ５０２に記憶された命令は、プロセッサ５０６によって実施され、実行される。

【００６５】

通信インターフェース５０４は、メモリ５０２とプロセッサ５０６の間の通信、プロセッサ間の通信、メモリ間の通信、及びシステムと別のデバイス若しくは通信ネットワークの間の通信を実施するように適応される。

【００６６】

プロセッサ５０６は、汎用中央処理装置（Ｃｅｎｔｒａｌ　Ｐｒｏｃｅｓｓｉｎｇ　Ｕｎｉｔ，ＣＰＵ）、マイクロプロセッサ、特定用途向け集積回路（Ａｐｐｌｉｃａｔｉｏｎ　Ｓｐｅｃｉｆｉｃ　Ｉｎｔｅｇｒａｔｅｄ　Ｃｉｒｃｕｉｔ，ＡＳＩＣ）、或いは１つ又は複数の集積回路であることが可能であり、関連するプログラムを実行して、本発明の様々な実施形態において説明されるＭＤＣモジュール、ＩＯルーティングモジュール、及びＯＳＤノードの間の対話の手順、並びに実施される機能を実行するように適応される。

【００６７】

理解を容易にし、不必要な繰り返しの説明を回避するのに、後段の特定の実施形態が、本発明の実施形態において説明されるＭＤＣモジュール、ＩＯルーティングモジュール、及びＯＳＤノードの間の対話の手順、並びに実施される機能がどのように、図５に示されるシステムにおいて実行されるかを説明するのに使用される。本発明のすべての実施形態

に基づいて、当業者は、図５に示されるシステムが、他の様々な実施形態において説明される事例を実施するのに適応されることが可能であり、それらの事例のすべては、本発明の記録され開示される範囲に入ることを理解されよう。

【００６８】

実施形態１

プロセッサは、メモリに接続されて、メモリの中の命令を読み取ることであって、命令は、ＭＤＣモジュール、ＩＯルーティングモジュール、及びＯＳＤノードの機能を実行するために使用される命令を含むこと、及び命令により、プロセッサが以下の操作、すなわち、

ＭＤＣモジュールが、システムにおけるＯＳＤノードが障害を有するＯＳＤノードであると決定すること、障害を有するＯＳＤノード上のｐａｒｔｉｔｉｏｎを決定すること、障害を有するＯＳＤノード上のｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループのｐａｒｔｉｔｉｏｎ　ｖｉｅｗを更新すること、及び更新されたｐａｒｔｉｔｉｏｎ　ｖｉｅｗにおけるｐａｒｔｉｔｉｏｎグループが位置付けられている一次ＯＳＤノードに更新通知を送信して、一次ＯＳＤノードが、更新されたｐａｒｔｉｔｉｏｎ　ｖｉｅｗにより、ＩＯ要求に対応するデータのレプリケーションを処理することを可能にすること、

を実行することを可能にするように適応される。

【００６９】

前述の実施形態において、前述のＭＤＣモジュール、ＩＯルーティングモジュール、及びＯＳＤノードの機能は、１つのホストによって実施され得ることに留意されたい。この事例において、前述のＭＤＣモジュール、ＩＯルーティングモジュール、及びＯＳＤノードの機能を実施するための命令は、ホストのメモリの中に存在することが可能であり、ホストのプロセッサが、前述のＭＤＣモジュール、ＩＯルーティングモジュール、及びＯＳＤノードの機能を実施するための命令をメモリから読み取る。別の実施形態において、前述のＭＤＣモジュール、ＩＯルーティングモジュール、及びＯＳＤノードの機能は、複数のホストによって対話様態で実施され得る。この事例において、前述のＭＤＣモジュール、ＩＯルーティングモジュール、及びＯＳＤノードは、異なるホストのメモリに分散様態で記憶される。例えば、ホスト１のプロセッサが、前述のＭＤＣモジュールの機能を実行し、ホスト２のプロセッサが、一次ＯＳＤノードの機能を実行し、ホスト３が、ＩＯルーティングモジュールの機能を実行する。

【００７０】

実施形態２

プロセッサは、メモリに接続されて、メモリの中の命令を読み取ることであって、命令は、ＭＤＣモジュール、ＩＯルーティングモジュール、及びＯＳＤノードの機能を実行するために使用される命令を含むこと、及び命令により、プロセッサが以下の操作、すなわち、

ＩＯルーティングモジュールが、ＩＯ要求を受信することであって、ＩＯ要求は、ｋｅｙを含む、受信すること、ｋｅｙにより、ＩＯ要求に対応するデータが属するｐａｒｔｉｔｉｏｎグループを決定すること、データが属するｐａｒｔｉｔｉｏｎグループの一次ＯＳＤノードを決定すること、データが属するｐａｒｔｉｔｉｏｎグループのＩＯ　ｖｉｅｗのＩＯ　ｖｉｅｗバージョン情報をＩＯ要求に追加すること、及び決定された一次ＯＳＤノードに、ＩＯ　ｖｉｅｗバージョン情報を搬送するＩＯ要求を送信すること、を可能にすること、

一次ＯＳＤノードが、ＩＯ要求を受信すること、ＩＯ　ｖｉｅｗバージョン情報により、ＩＯ要求におけるＩＯ　ｖｉｅｗバージョンが、ローカルで記憶されたＩＯ　ｖｉｅｗバージョンと整合性があると決定した後、ＩＯ要求を実行すること、ＩＯ　ｖｉｅｗバージョン情報を搬送するレプリケーション要求を生成すること、及びレプリケーション要求を、データが属するｐａｒｔｉｔｉｏｎの二次ＯＳＤノードに送信すること、を可能にすること、並びに

二次ＯＳＤノードが、レプリケーション要求を受信し、ＩＯ　ｖｉｅｗバージョン情報

により、レプリケーション要求におけるＩＯ　ｖｉｅｗバージョンが、二次ＯＳＤノード上にローカルで記憶されたＩＯ　ｖｉｅｗバージョンと整合性があると決定した後、レプリケーション要求を実行して、二次ＯＳＤノード上のＩＯ要求に対応するデータが、一次ＯＳＤノード上のＩＯ要求に対応するデータと整合性を保つことを可能にすること、
を実行することを可能にするように適応される。

【００７１】

同様に、前述の実施形態において、前述のＭＤＣモジュール、ＩＯルーティングモジュール、及びＯＳＤノードの機能は、１つのホストによって実施されることが可能である。この事例において、前述のＭＤＣモジュール、ＩＯルーティングモジュール、及びＯＳＤノードの機能を実施するための命令は、ホストのメモリの中に存在することが可能であり、ホストのプロセッサが、前述のＭＤＣモジュール、ＩＯルーティングモジュール、及びＯＳＤノードの機能を実施するための命令をメモリから読み取る。別の実施形態において、前述のＭＤＣモジュール、ＩＯルーティングモジュール、及びＯＳＤノードの機能は、複数のホストによって対話様態で実施され得る。この事例において、前述のＭＤＣモジュール、ＩＯルーティングモジュール、及びＯＳＤノードは、異なるホストのメモリに分散様態で記憶される。例えば、ホスト１のプロセッサが、前述のＩＯルーティングモジュールの機能を実行し、ホスト２のプロセッサが、一次ＯＳＤノードの機能を実行し、ホスト３のプロセッサが、二次ＯＳＤノードの機能を実行し、ホスト４のプロセッサが、ＭＤＣモジュールの機能を実行する。

【００７２】

実施形態３

プロセッサは、メモリに接続されて、メモリの中の命令を読み取ることであって、命令は、ＭＤＣモジュール、ＩＯルーティングモジュール、及びＯＳＤノードの機能を実行するために使用される命令を含むこと、及び命令により、プロセッサが以下の操作、すなわち、

ＯＳＤノードが、フェイルバックの後、ＭＤＣモジュールに、ＯＳＤノード上のｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループのＩＯ　ｖｉｅｗを要求するクエリ要求を送信することであって、ＯＳＤノードは、フェイルバックＯＳＤノードと呼ばれ、クエリ要求は、フェイルバックＯＳＤノードのＯＳＤ識別子を搬送する、送信すること、ＭＤＣによって返されるＩＯ　ｖｉｅｗを受信すること、ＩＯ　ｖｉｅｗにおける一次ＯＳＤに対してデータ回復要求を開始して、障害中にフェイルバックＯＳＤノードによって更新されたデータを回復することを要求すること、障害中に更新され、かつ一次ＯＳＤによって送信されるデータを受信すること、及びｐａｒｔｉｔｉｏｎグループのものであり、かつフェイルバックＯＳＤノードがデータ回復を完了した後にＭＤＣモジュールによって更新されたｐａｒｔｉｔｉｏｎ　ｖｉｅｗにより、ＩＯ要求のレプリケーションを処理すること、を可能にすること、

ＭＤＣモジュールが、フェイルバックＯＳＤノードのクエリ要求を受信すること、クエリ要求におけるＯＳＤ識別子によりフェイルバックＯＳＤノードにＩＯ　ｖｉｅｗを返すこと、及びフェイルバックＯＳＤノードがデータ回復を完了した後、ｐａｒｔｉｔｉｏｎグループのｐａｒｔｉｔｉｏｎ　ｖｉｅｗを更新すること、を可能にすること、並びに

一次ＯＳＤノードが、フェイルバックＯＳＤノードのデータ回復要求を受信すること、フェイルバックＯＳＤノードに、障害中に更新されたデータを送信すること、及びｐａｒｔｉｔｉｏｎグループのものであり、かつフェイルバックＯＳＤノードがデータ回復を完了した後にＭＤＣモジュールによって更新されたｐａｒｔｉｔｉｏｎ　ｖｉｅｗにより、ＩＯ要求に対応するデータのレプリケーションを処理すること、を可能にすること、
を実行することを可能にするように適応される。

【００７３】

同様に、前述の実施形態において、前述のＭＤＣモジュール、ＩＯルーティングモジュール、及びＯＳＤノードの機能は、１つのホストによって実施されることが可能である。この事例において、前述のＭＤＣモジュール、ＩＯルーティングモジュール、及びＯＳＤ

ノードの機能を実施するための命令は、ホストのメモリの中に存在することが可能であり、ホストのプロセッサが、前述のＭＤＣモジュール、ＩＯルーティングモジュール、及びＯＳＤノードの機能を実施するための命令をメモリから読み取る。別の実施形態において、前述のＭＤＣモジュール、ＩＯルーティングモジュール、及びＯＳＤノードの機能は、複数のホストによって対話様態で実施され得る。この事例において、前述のＭＤＣモジュール、ＩＯルーティングモジュール、及びＯＳＤノードは、異なるホストのメモリに分散様態で記憶される。例えば、ホスト１のプロセッサが、前述のフェイルバックＯＳＤノードの機能を実行し、ホスト２のプロセッサが、一次ＯＳＤノードの機能を実行し、ホスト３のプロセッサが、ＭＤＣモジュールの機能を実行し、ホスト４のプロセッサが、ＩＯルーティングモジュールの機能を実行する。

【００７４】

以下に、複数の特定の手順実施形態を使用して、前述の分散ストレージ及びレプリケーション制御システムにおけるＭＤＣモジュール２０２、ＩＯルーティングモジュール２０４、及びＯＳＤノード２０６の接続及び対話、特定の機能などを詳細にさらに説明する。これらの特定の手順実施形態は、クラスタビューの初期化された生成及び獲得の手順、ＩＯ要求処理手順、ＯＳＤ障害処理手順、ＯＳＤノードフェイルバック手順、データ回復手順、ＯＳＤノードが障害の後にクラスタから出る手順、新たなＯＳＤノードがクラスタに参加する手順を含み、これらの手順が、以下に１つずつ詳細に説明される。

【００７５】

前述の実施形態におけるすべてのステップ、又はすべての機能が必要であるわけではなく、ステップのシーケンスは、説明を簡単にすることを目的としているに過ぎず、本発明の問題解決法の原理により必然的に要求されるのでない限り、特に限定されず、さらに、ステップにおける特定の実施様態は、説明的な例として使用されるに過ぎず、本発明の保護範囲に対する特定の限定を構成しないことに留意されたい。本発明の明細書全体に基づいて、当業者は、実際の状況により前述のステップに対して、対応する追加又は削除、或いは非独創的な変更又は置換を行うこと、並びにそれらのステップにおける様々な実施様態を、他のステップにおける実施様態と組み合わせて、異なる実施形態を形成することが可能であり、以上のすべては、本発明の記録され開示される範囲に入る。

【００７６】

クラスタビューの初期化された生成及び獲得の手順

【００７７】

図６が、本発明によるクラスタビューの初期化された生成及び獲得の実施形態を示す。この実施形態において、ＭＤＣモジュールは、管理者によって供給されるクラスタ構成情報により初期クラスタビューを生成する。ＩＯルーティングモジュール及びＯＳＤノードが、初期化中にｖｉｅｗに関してＭＤＣモジュールにクエリを行う。この実施形態は、図２Ａないし図２Ｃ、及び図５において説明される実施形態において言及されるＭＤＣモジュール、ＩＯルーティングモジュール、及びＯＳＤノードによって実行される。

【００７８】

システム初期化中、ＭＤＣモジュールがまず開始され、次に、ＩＯルーティングモジュール及びＯＳＤノードによる、ＭＤＣからｖｉｅｗを獲得することが開始される。特定のプロセスは、以下のステップを含む。すなわち、

【００７９】

６０２　ユーザ又は管理者が、ＭＤＣモジュールにクラスタ構成情報を供給し、クラスタ構成情報は、クラスタトポロジ情報、ｐａｒｔｉｔｉｏｎ数量、及びレプリカ数量などのシステム構成パラメータを含むことが可能であり、クラスタトポロジ情報は、サーバの数量及びサーバのＩＰアドレス、ラック情報、各サーバ上のＯＳＤノードの数量、並びにＯＳＤノードによって管理される物理ストレージリソースについての情報（例えば、ＯＳＤノードに対応するローカルディスク情報）などを主に含む。

【００８０】

６０４　ＭＤＣが、供給されたクラスタ構成により初期クラスタビューを生成し、３つ

のタイプのクラスタビュー（ＯＳＤ　ｖｉｅｗ、ｐａｒｔｉｔｉｏｎ　ｖｉｅｗ、及びＩＯ　ｖｉｅｗ）が既に前段で説明されており、ＭＤＣは、構成されたＯＳＤ情報によりＯＳＤ　ｖｉｅｗを生成し、ｐａｒｔｉｔｉｏｎ割当てアルゴリズム、構成されたｐａｒｔｉｔｉｏｎ数量、レプリカ数量、及びＯＳＤノード数量を使用することによってｐａｒｔｉｔｉｏｎ　ｖｉｅｗを生成し、ｐａｒｔｉｔｉｏｎ　ｖｉｅｗのサブセットであるＩＯ　ｖｉｅｗをさらに生成する必要はなく、ｐａｒｔｉｔｉｏｎ　ｖｉｅｗが生成される場合、ｐａｒｔｉｔｉｏｎ割当て平衡（各ＯＳＤノード上のｐａｒｔｉｔｉｏｎの数量が可能な限り同一に保たれる）及びセキュリティ（ｐａｒｔｉｔｉｏｎレプリカが存在するＯＳＤノードが異なるサーバ、又は異なるラックにある）が一般に考慮される必要がある。

【００８１】

６０６　ＩＯルーティングモジュールの初期化を開始し、ＩＯ　ｖｉｅｗに関してＭＤＣモジュールにクエリを行い、開始される場合、ＩＯルーティングモジュールは、ＭＤＣから関連するｖｉｅｗを獲得して、正常に機能するようにする必要がある。

【００８２】

６０８　ＯＳＤノードの初期化を開始し、ｐａｒｔｉｔｉｏｎ　ｖｉｅｗ及びＩＯ　ｖｉｅｗに関してＭＤＣモジュールにクエリを行い、ＯＳＤノードは、ＯＳＤノード上に分散された一次ｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループのｐａｒｔｉｔｉｏｎ　ｖｉｅｗ、及びＯＳＤノード上に分散された二次ｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループのＩＯ　ｖｉｅｗを獲得する必要がある。

【００８３】

６１０　ＭＤＣモジュールが、すべてのｐａｒｔｉｔｉｏｎグループのＩＯ　ｖｉｅｗをＩＯルーティングモジュールに返す。

【００８４】

６１２　ＭＤＣモジュールが、ＯＳＤノードに、ＯＳＤノード上に分散された一次ｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループのｐａｒｔｉｔｉｏｎ　ｖｉｅｗ、及びＯＳＤノード上に分散された二次ｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループのＩＯ　ｖｉｅｗを返す。

【００８５】

ＩＯ要求処理手順

【００８６】

図７Ａ及び図７Ｂが、本発明によるＩＯ要求手順の実施形態を示す。この実施形態は、図２Ａないし図２Ｃ、及び図５において説明される実施形態において言及されるＩＯルーティングモジュール及びＯＳＤノードによって実行される。理解を容易にするのに、ｐａｒｔｉｔｉｏｎグループ（Ｐａｒｔｉｔｉｏｎ　Ｘ）が、この実施形態における説明のための例として使用され、Ｐａｒｔｉｔｉｏｎ　Ｘは、本発明のこの実施形態における分散レプリケーションプロトコル層によって管理され、保持される任意のパーティションであることが可能であり、Ｐａｒｔｉｔｉｏｎ　Ｘが位置付けられる一次ＯＳＤノードは、Ｐａｒｔｉｔｉｏｎ　Ｘ一次ＯＳＤノードと略記され、Ｐａｒｔｉｔｉｏｎ　Ｘにおける二次ｐａｒｔｉｔｉｏｎが位置付けられる二次ＯＳＤ１ノードは、Ｐａｒｔｉｔｉｏｎ　Ｘ二次ＯＳＤ１ノードと略記され、Ｐａｒｔｉｔｉｏｎ　Ｘが位置付けられる二次ｐａｒｔｉｔｉｏｎ上の二次ＯＳＤ２ノードは、Ｐａｒｔｉｔｉｏｎ　Ｘ二次ＯＳＤ２ノードと略記される。ｐａｒｔｉｔｉｏｎＸは、以下の特定の説明における例として使用される。この実施形態におけるＩＯ操作手順（例えば、書込み操作又は変更操作）は、以下のステップを特に含む。すなわち、

【００８７】

７０２　ＩＯルーティングモジュールが、ホスト（例えば、ＩＯルーティングモジュールが位置付けられる図２Ｃに示されるサーバ）によって送信されるＩＯ要求を受信する。

【００８８】

７０４　ＩＯルーティングモジュールが、受信されたＩＯ要求により、ＩＯ要求に対応するデータ（ＩＯデータとも呼ばれ得る）のｐａｒｔｉｔｉｏｎを獲得し、そのｐａｒｔ

ｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループの一次ＯＳＤノードを獲得する。
【００８９】
特定の実施様態において、ＩＯルーティングモジュールは、ハッシュアルゴリズムを使用することによって計算を用いて、かつＩＯ要求において搬送されるＫｅｙにより、ＩＯ要求に対応するデータを含むｐａｒｔｉｔｉｏｎグループのｐａｒｔｉｔｉｏｎ　ＩＤを獲得することが可能であり、次に、ｐａｒｔｉｔｉｏｎ　ＩＤを使用することによってＩＯ　ｖｉｅｗを探索して、ｐａｒｔｉｔｉｏｎグループの一次ＯＳＤノードを獲得するようにする。前述したとおり、本明細書におけるｐａｒｔｉｔｉｏｎ　ＩＤに対応するｐａｒｔｉｔｉｏｎグループは、この実施形態におけるｐａｒｔｉｔｉｏｎ　Ｘである。Ｋｅｙは、上位層サービスにおいて定義される数字又は文字ストリングであり、データブロックを識別するのに使用される。
【００９０】
７０６　ＩＯルーティングモジュールが、ｐａｒｔｉｔｉｏｎグループの一次ＯＳＤノードにＩＯ要求を送信し、要求は、ＩＯ　ｖｉｅｗバージョン情報（例えば、ＩＯ　ｖｉｅｗ　ＩＤ）、ＩＯキー情報、及びＩＯデータを搬送する。
【００９１】
前述の説明によれば、この実施形態におけるｐａｒｔｉｔｉｏｎグループは、ｐａｒｔｉｔｉｏｎ　Ｘであり、したがって、ＩＯ要求は、ｐａｒｔｉｔｉｏｎ　Ｘの一次ＯＳＤノード、すなわち、ｐａｒｔｉｔｉｏｎ　Ｘ一次ＯＳＤノードに送信される。特定の実施様態において、ＩＯ　ｖｉｅｗ　ＩＤは、ＩＯ　ｖｉｅｗバージョン番号と呼ばれることも可能であり、ＩＯ　ｖｉｅｗバージョン番号は、ビューバージョンを識別するのに主に使用され、単調増加し、小さいＩＯ　ｖｉｅｗ　ＩＤは、一次ＯＳＤノードによって保持されるビューバージョンが古くなっていることを示し、整合性を確実にするための要件は、ＩＯ処理手順においてすべてのモジュールによって見られるｖｉｅｗに整合性があることである。ＩＯキー情報は、ｋｅｙと、ｏｆｆｓｅｔと、ｌｅｎｇｔｈとを含むことが可能であり、ｏｆｆｓｅｔは、ｋｅｙによって識別されるデータブロックにおける開始位置に対するＩＯデータのオフセットを示し、ｌｅｎｇｔｈは、ＩＯデータの長さを示す。
【００９２】
７０８　ＩＯ　ｖｉｅｗバージョン情報により、ＩＯ要求において搬送されるビューバージョンがローカルで記憶されたビューバージョンと整合性があると決定された後、ＩＯ要求に関してシーケンス（Ｓｅｑ）ＩＤ（シーケンス識別子とも呼ばれ得る）を生成し、特に、ＩＯ要求において搬送されるＩＯ　ｖｉｅｗ　ＩＤがローカルで記憶されたＩＯ　ｖｉｅｗ　ＩＤと整合性があると判定した後、一次ＯＳＤノードが、ＩＯ要求に関するシーケンス識別子を生成する。
【００９３】
特定の実施様態において、ＳｅｑＩＤは、ビューバージョン番号と、シーケンス番号、ｓｅｑｕｅｎｃｅ　ｎｕｍｂｅｒ（Ｓｅｑ　ＮＯ）とを含む。ビューバージョン番号は、ＩＯ　ｖｉｅｗが変化するにつれて単調増加し、Ｓｅｑ　ＮＯは、１つのＩＯ　ｖｉｅｗバージョン内のＩＯ　ｖｉｅｗにおけるｐａｒｔｉｔｉｏｎに対応するデータに対する変更操作（例えば、書込み及び削除）の通し番号を示す。ＩＯ　ｖｉｅｗが変化した後、ＳｅｑＩＤにおけるＳｅｑ　ＮＯは、０から再び増加することを開始する。Ｐａｒｔｉｔｉｏｎ　Ｘ一次ＯＳＤノードが、ＩＯ要求において搬送されるＩＯ　ｖｉｅｗ　ＩＤがローカルで記憶されたＩＯ　ｖｉｅｗ　ＩＤと整合性があるかどうかを比較することは、特に、Ｐａｒｔｉｔｉｏｎ　Ｘ一次ＯＳＤノードがまず、ＩＯ　ｖｉｅｗ　ＩＤを比較し、より大きいＩＯ　ｖｉｅｗ　ＩＤは、より大きいＳｅｑＩＤを示し、ＩＯ　ｖｉｅｗ　ＩＤが等しい場合、次に、Ｓｅｑ　ＮＯを比較し、より大きいＳｅｑ　ＮＯは、より大きいＳｅｑＩＤを示し、ＩＯ　ｖｉｅｗ　ＩＤとＳｅｑ　ＮＯがともに同一である場合に限り、そのことは、ＳｅｑＩＤに整合性があることを示すことであり得る。
【００９４】
７１０　ローカルＩＯ　ｖｉｅｗがより大きい場合、ＩＯルーティングモジュールにエ

ラーを返す、又はローカルＩＯ　ｖｉｅｗがより小さい場合、キャッシュキューにＩＯ要求を追加し、ｐａｒｔｉｔｉｏｎ　ｖｉｅｗに関してＭＤＣモジュールにクエリを行う。

【００９５】

特定の実施様態において、ＩＯ要求におけるＩＯ　ｖｉｅｗ　ＩＤがローカルで記憶されたＩＯ　ｖｉｅｗ　ＩＤ未満であると決定した後、Ｐａｒｔｉｔｉｏｎ　Ｘ一次ＯＳＤノードが、ＩＯルーティングモジュールにエラーを返し、ＩＯルーティングモジュールが、ｐａｒｔｉｔｉｏｎグループのＩＯ　ｖｉｅｗに関してＭＤＣモジュールにクエリを行い、更新されたＩＯ　ｖｉｅｗ　ＩＤを獲得した後、ＩＯ要求を再送信する、或いはＩＯ要求におけるＩＯ　ｖｉｅｗ　ＩＤがローカルで記憶されたＩＯ　ｖｉｅｗ　ＩＤより大きいと決定した後、Ｐａｒｔｉｔｉｏｎ　Ｘ一次ＯＳＤノードが、キャッシュキューにＩＯ要求を追加し、ｐａｒｔｉｔｉｏｎグループのＩＯ　ｖｉｅｗのＩＯ　ｖｉｅｗ　ＩＤに関してＭＤＣモジュールにクエリを行って、ローカルで記憶されたＩＯ　ｖｉｅｗ　ＩＤが、ＩＯ要求におけるＩＯ　ｖｉｅｗ　ＩＤと整合性があると決定した後、ＩＯ要求を実行するようにする。

【００９６】

７１２　１つのｅｎｔｒｙを記録し、そのｅｎｔｒｙは、操作タイプと、ＰａｒｔｉｔｉｏｎグループＩＤと、ＳｅｑＩＤと、ｋｅｙとを含む。

【００９７】

特定の実施様態において、操作タイプは、書込み、削除などを含むことが可能である。書込み操作に関して、ｅｎｔｒｙは、前述のｏｆｆｓｅｔ及びｌｅｎｇｔｈをさらに含むことが可能である。さらに、様々なタイプの操作に関して、ｅｎｔｒｙは、操作が成功したかどうかを記述するのに使用されるステータス情報をさらに含むことが可能である。一般に、同一のｐａｒｔｉｔｉｏｎグループに対するすべての変更操作（書込み操作及び削除操作など）に連続的に番号が付けられる。

【００９８】

７１４　ＩＯ要求を実行し、ｐａｒｔｉｔｉｏｎグループの二次ＯＳＤノードについての情報を獲得するようにｐａｒｔｉｔｉｏｎグループのｐａｒｔｉｔｉｏｎ　ｖｉｅｗにクエリを行う。

【００９９】

特定の実施様態において、ＩＯ要求を実行することは、ＩＯ要求を書き込むことであり、ＩＯデータは、Ｐａｒｔｉｔｉｏｎ　Ｘ一次ＯＳＤノードによって管理されるローカル物理ストレージリソース（例えば、磁気ディスクなどの図２Ｃに示されるキャッシュ層若しくは永続層、又は前述した外部物理ストレージリソースＳＡＮ）に書き込まれ、ＩＯ要求が削除要求である場合、Ｐａｒｔｉｔｉｏｎ　Ｘ一次ＯＳＤノードによって管理されるローカル物理ストレージリソース上の対応するデータが、削除される。特定の実施形態において、Ｐａｒｔｉｔｉｏｎ　Ｘ一次ＯＳＤノードは、レプリケーション要求をさらに生成する。特定の実施において、レプリケーション要求は、ＩＯ要求の制御部分を別個に組み立てることによって生成されたレプリケーション要求であることが可能であり、Ｐａｒｔｉｔｉｏｎ　Ｘの二次ＯＳＤノード上のＩＯ要求に対応するデータは、Ｐａｒｔｉｔｉｏｎ　Ｘ一次ＯＳＤノード上のＩＯ要求に対応するデータと整合性がある。

【０１００】

７１６（１）／（２）　Ｐａｒｔｉｔｉｏｎ　Ｘ二次ＯＳＤ１ノードとＰａｒｔｉｔｉｏｎ　Ｘ二次ＯＳＤ２ノードにレプリケーション要求を別々に送信し、レプリケーション要求は、ＳｅｑＩＤを搬送する。

【０１０１】

特定の実施様態において、レプリケーション要求は、元の要求におけるｋｅｙ、ｏｆｆｓｅｔ、ｌｅｎｇｔｈ及びＩＯデータなどの情報をさらに含むことが可能である。レプリケーション要求が書込みレプリケーション要求である場合、レプリケーション要求は、ｋｅｙ、ｏｆｆｓｅｔ、ｌｅｎｇｔｈ、及びＩＯデータを搬送し、レプリケーション要求が削除レプリケーション要求である場合、レプリケーション要求は、ｋｅｙだけを搬送する

。
【０１０２】
７１８　レプリケーション要求におけるＩＯ　ｖｉｅｗ　ＩＤがローカルで記憶されたＩＯ　ｖｉｅｗ　ＩＤと整合性があると決定された後、レプリケーション要求を処理する。
【０１０３】
特定の実施様態において、前述のステップ７１８に加えて、手順は、以下のステップをさらに含むことが可能である。すなわち、
【０１０４】
レプリケーション要求におけるＩＯ　ｖｉｅｗ　ＩＤがローカルで記憶されたＩＯ　ｖｉｅｗ　ＩＤ未満であると決定した後、Ｐａｒｔｉｔｉｏｎ　Ｘ二次ＯＳＤ１ノード又はＰａｒｔｉｔｉｏｎ　Ｘ二次ＯＳＤ２が、Ｐａｒｔｉｔｉｏｎ　Ｘ一次ＯＳＤノードにエラーを返し、Ｐａｒｔｉｔｉｏｎ　Ｘ一次ＯＳＤノードが、ｐａｒｔｉｔｉｏｎグループのＩＯ　ｖｉｅｗに関してＭＤＣモジュールにクエリを行い、更新されたＩＯ　ｖｉｅｗ　ＩＤを獲得した後、レプリケーション要求を再送信し、又はレプリケーション要求におけるＩＯ　ｖｉｅｗ　ＩＤがローカルで記憶されたＩＯ　ｖｉｅｗ　ＩＤより大きいと決定した後、Ｐａｒｔｉｔｉｏｎ　Ｘ二次ＯＳＤ１ノード又はＰａｒｔｉｔｉｏｎ　Ｘ二次ＯＳＤ２ノードが、キャッシュキューにレプリケーション要求を追加し、ｐａｒｔｉｔｉｏｎグループのＩＯ　ｖｉｅｗバージョン番号に関してＭＤＣモジュールにクエリを行って、ローカルで記憶されたＩＯ　ｖｉｅｗ　ＩＤが、ＩＯ要求におけるＩＯ　ｖｉｅｗ　ＩＤと整合性があると決定した後、レプリケーション要求を実行するようにする。
【０１０５】
既存の２フェーズコミットプロトコルにおいて、参加者が提案を拒否した場合、ＩＯ手順全体がロールバックされる必要があり、このことは、相当に大きいオーバヘッドをもたらす。しかし、この実施形態において、二次ＯＳＤノードが要求を拒否した場合、最新のｖｉｅｗのクエリがまず行われ、次に、処理が続き、ロールバックは、実行される必要がなく、したがって、このことが、システム全体のフォールトトレランス及び利用可能性を向上させる。
【０１０６】
既存の２フェーズコミットプロトコルにおいて、１つの通常のＩＯプロセスにおいて、コーディネータと参加者の間でメッセージ対話が２回、実行される必要がある。しかし、この実施形態において、ＩＯプロセスは、１回だけしか実行される必要がなく、一次ノードと二次ノードの間のメッセージ対話は、１回だけ実行され、このことが、メッセージ対話によってもたらされるＩＯ遅延を短縮し、システム全体の効率及びパフォーマンスを向上させる。
【０１０７】
特定の実施様態において、レプリケーション要求は、ｐａｒｔｉｔｉｏｎ　ＸのＰａｒｔｉｔｉｏｎ　Ｘ一次ＯＳＤノードによって送信された前のレプリケーション要求におけるＳｅｑＩＤをさらに搬送することが可能であり、ＩＯ手順は、ステップ７２０をさらに含むことが可能である。
【０１０８】
７２０　前のレプリケーション要求に入っており、かつレプリケーション要求において搬送されるＳｅｑＩＤを最大のローカルＳｅｑＩＤと比較し、ＳｅｑＩＤが最大のローカルＳｅｑＩＤより大きい場合、欠落している要求を送信するようＰａｒｔｉｔｉｏｎ　Ｘ一次ＯＳＤノードに要求する、又はＳｅｑＩＤが最大のローカルＳｅｑＩＤと整合性がある場合、後続のステップにおいて処理することを続ける。
【０１０９】
レプリケーション要求は、前のレプリケーション要求に入っているＳｅｑＩＤを搬送し、このことが、ＩＯ　ｖｉｅｗバージョン番号の変更によってもたらされる欠落したデータを防止することが可能である。さらに、欠落したデータが送信され、このことが、一次

ＯＳＤノードと二次ＯＳＤノードがすべてのＩＯを実行するシーケンスに整合性があることを確実にし、かつデータバックアップ整合性をさらに向上させることが可能である。

【０１１０】

特定の実施様態において、前のレプリケーション要求に入っており、かつレプリケーション要求において搬送されるＳｅｑＩＤが最大のローカルＳｅｑＩＤ未満の場合、エラーがｐａｒｔｉｔｉｏｎ　Ｘ一次ＯＳＤノードに返され、ｐａｒｔｉｔｉｏｎ　Ｘ一次ＯＳＤノードが、レプリケーション要求を再送信する、又はＳｅｑＩＤが、さらにクエリすることを用いて決定され、ＳｅｑＩＤが決定された後、更新されたＳｅｑＩＤが獲得され、次に、エラーを直接に返す代わりに、処理が継続する。この２つの事例において、処理は、ロールバックのために終了される必要がなく、このことが、システムのフォールトトレランス及び利用可能性、並びにシステム全体のパフォーマンスをさらに向上させる。

【０１１１】

ＩＯ操作が一次ＯＳＤノードと二次ＯＳＤノードの両方によって実行される必要がある場合、一次ＯＳＤノードと二次ＯＳＤノードがすべてのＩＯ操作を実行するシーケンスに整合性があることを確実にすることが要求され、そうであることは、複数のｐａｒｔｉｔｉｏｎレプリカの整合性を確実にする鍵である。

【０１１２】

７２２　１つのｅｎｔｒｙを記録し、レプリケーション操作を実行する。

【０１１３】

特定の実施様態において、ｅｎｔｒｙは、操作タイプ、Ｐａｒｔｉｔｉｏｎ　ＩＤ、ＳｅｑＩＤ、及びｋｅｙを含み、操作タイプは、書込み、削除などを含むことが可能である。書込み操作に関して、ｅｎｔｒｙは、前述のｏｆｆｓｅｔ及びｌｅｎｇｔｈをさらに含むことが可能である。さらに、様々なタイプの操作に関して、ｅｎｔｒｙは、操作が成功したかどうかを記述するのに使用されるステータス情報をさらに含むことが可能である。レプリケーション要求を実行することは、レプリケーション要求が書込みレプリケーション要求である場合、レプリケーション要求がｋｅｙ、ｏｆｆｓｅｔ、ｌｅｎｇｔｈ、及びＩＯデータを搬送すること、及びレプリケーション要求が削除レプリケーション要求である場合、レプリケーション要求がｋｅｙだけを搬送することを特に含む。

【０１１４】

７２４／７２６　ｐａｒｔｉｔｉｏｎ　Ｘ二次ＯＳＤ１ノードとｐａｒｔｉｔｉｏｎ　Ｘ二次ＯＳＤ２ノードが、ｐａｒｔｉｔｉｏｎ　Ｘ一次ＯＳＤノードに応答要求成功メッセージを別々に送信する。

【０１１５】

７２８　ｐａｒｔｉｔｉｏｎ　Ｘ一次ＯＳＤノードが、ＩＯルーティングモジュールに応答要求成功メッセージを送信する。

【０１１６】

前述のＩＯ要求実施形態に基づいて、ＩＯ要求を処理するプロセスにおいて、ｐａｒｔｉｔｉｏｎ　Ｘ一次ＯＳＤノード又はｐａｒｔｉｔｉｏｎ　Ｘ二次ＯＳＤノードが、障害を有するようになる（例えば、ＩＯ要求がｐａｒｔｉｔｉｏｎ　Ｘ一次ＯＳＤノードに着信した後、ｐａｒｔｉｔｉｏｎ　Ｘ一次ＯＳＤノードが障害を有するようになる）こと、或いはｐａｒｔｉｔｉｏｎ　Ｘ二次ＯＳＤノードが、障害を有するようになり、新たなＯＳＤノードが、ｐａｒｔｉｔｉｏｎ　Ｘの二次ＯＳＤノードとしてシステムに参加することなどを想定すると、これらの事例において、前述の実施形態におけるＩＯ要求処理手順は、以下の実施形態において説明される処理プロセスをさらに含むことが可能である。

【０１１７】

ＩＯ要求がｐａｒｔｉｔｉｏｎ　Ｘ一次ＯＳＤノードに着信した後、ｐａｒｔｉｔｉｏｎ　Ｘ一次ＯＳＤノードが障害を有するようになった場合、ＩＯ処理手順は、以下を含む。すなわち、

ｐａｒｔｉｔｉｏｎ　Ｘ一次ＯＳＤノードが障害を有するようになった後、システムにおけるＭＤＣモジュールが、ＩＯ要求を処理するプロセスにおいて、ｐａｒｔｉｔｉｏｎ

Ｘ一次ＯＳＤノードが障害を有するようになったことを検出した場合、ｐａｒｔｉｔｉｏｎグループのｐａｒｔｉｔｉｏｎ ｖｉｅｗにおけるｐａｒｔｉｔｉｏｎ Ｘ一次ＯＳＤノードを新たなｐａｒｔｉｔｉｏｎ Ｘ二次ＯＳＤノードとして設定し、その新たなｐａｒｔｉｔｉｏｎ Ｘ二次ＯＳＤノードのｐａｒｔｉｔｉｏｎステータスに不整合であるという印を付け、ｐａｒｔｉｔｉｏｎグループのｐａｒｔｉｔｉｏｎ ｖｉｅｗにおけるｐａｒｔｉｔｉｏｎ Ｘ二次ＯＳＤ１ノードを新たなｐａｒｔｉｔｉｏｎ Ｘ一次ＯＳＤノードとして設定し、その新たなｐａｒｔｉｔｉｏｎ Ｘ一次ＯＳＤノードにｐａｒｔｉｔｉｏｎグループの更新されたｐａｒｔｉｔｉｏｎ ｖｉｅｗを送信し、ｐａｒｔｉｔｉｏｎグループのＩＯ ｖｉｅｗにおけるｐａｒｔｉｔｉｏｎ Ｘ二次ＯＳＤ１ノードを新たなｐａｒｔｉｔｉｏｎ Ｘ一次ＯＳＤノードとして設定し、ＩＯルーティングモジュールにｐａｒｔｉｔｉｏｎグループの更新されたＩＯ ｖｉｅｗを送信し、

ＩＯルーティングモジュールが、ｐａｒｔｉｔｉｏｎグループのものであり、かつＭＤＣモジュールによって送信される更新されたＩＯ ｖｉｅｗを受信し、ｐａｒｔｉｔｉｏｎグループの更新されたＩＯ ｖｉｅｗにより、その新たなｐａｒｔｉｔｉｏｎ Ｘ一次ＯＳＤノードにＩＯ要求を送信し、かつ

その新たなｐａｒｔｉｔｉｏｎ Ｘ一次ＯＳＤノードが、ＩＯ要求を受信すること、ＩＯ要求を実行した後、レプリケーション要求を生成すること、及びその新たなレプリケーション要求を、ｐａｒｔｉｔｉｏｎの更新されたｐａｒｔｉｔｉｏｎ ｖｉｅｗにおける別の二次ＯＳＤノードに送信するように適応され、レプリケーション要求を生成するステップ、及びレプリケーション要求を送信するステップは、前述のステップ７１４及び７１６と同一である。

【０１１８】

ＩＯ要求がｐａｒｔｉｔｉｏｎ Ｘ一次ＯＳＤノードに着信した後、ｐａｒｔｉｔｉｏｎ Ｘ二次ＯＳＤノードが障害を有するようになった場合、ＩＯ処理手順は、以下を含む。すなわち、

ｐａｒｔｉｔｉｏｎ Ｘ二次ＯＳＤノードが障害を有するようになった後、ＭＤＣモジュールが、ＩＯ要求を処理するプロセスにおいて、ｐａｒｔｉｔｉｏｎ Ｘ二次ＯＳＤノードが障害を有するようになったことを検出した場合、ｐａｒｔｉｔｉｏｎ ｖｉｅｗにおけるｐａｒｔｉｔｉｏｎ Ｘ二次ＯＳＤノードのｐａｒｔｉｔｉｏｎステータスに不整合であるという印を付け、ｐａｒｔｉｔｉｏｎ Ｘ一次ＯＳＤノードにｐａｒｔｉｔｉｏｎグループの更新されたｐａｒｔｉｔｉｏｎ ｖｉｅｗを送信するようにさらに適応され、かつ

一次ＯＳＤノードが、ｐａｒｔｉｔｉｏｎグループの更新されたｐａｒｔｉｔｉｏｎ ｖｉｅｗを受信した後、更新されたｐａｒｔｉｔｉｏｎ ｖｉｅｗにおいてｐａｒｔｉｔｉｏｎステータスに整合性がある別の二次ＯＳＤノードにレプリケーション要求を送信し、ｐａｒｔｉｔｉｏｎステータスが不整合であるｐａｒｔｉｔｉｏｎ Ｘ二次ＯＳＤノードにレプリケーション要求を送信することを省くように適応される。

【０１１９】

２フェーズコミットプロトコルにおいて、コーディネータが障害を有するようになった場合、ＩＯ処理は、中断され、コーディネータが正常に戻った場合に限り、継続する。この実施形態において、一次ＯＳＤノードが障害を有するようになった後、ＭＤＣノードは、投票を用いて新たな一次ＯＳＤノードを迅速に決定することができ、かつＩＯ処理を迅速に再開することができ、したがって、高い利用可能性、及び強いフォールトトレランスが実現される。

【０１２０】

さらに、２フェーズコミットプロトコルにおいて、参加者が障害を有するようになった場合、又は応答を有さないことが続く場合、別のＩＯ要求は、継続的に阻止され、タイムアウトのために最終的に失敗し、ロールバックが実行される必要がある。この事例において、二次ノードが障害を有するようになった場合、ＭＤＣが、ｖｉｅｗ変更を行うよう一次ノードに指示して、障害を有するＯＳＤノードを隔離する、又は無視するようにし、別

のＩＯ要求に対する処理を阻止することなしにＩＯ要求処理を継続するようにし、このことが、より良好なフォールトトレランスを有し、ノード障害を扱うこと、及びノードフェイルバックを実行することを迅速に行うことができる。例えば、Ｎ＋１のレプリカのうちのＮのレプリカの障害が許容されることが可能であり、このことが、ストレージシステムのパフォーマンス及び利用可能性をさらに向上させる。低い利用可能性を有するシステムは、劣悪な拡張性を不可避に有し、ストレージノード障害は、大規模な分散ストレージシステムにおいてよくあるため、複雑で、大量のプロトコル対話が、システム拡張性をさらに低減する可能性がある。

【０１２１】

さらに、ストレージノード障害の影響範囲が、ｐａｒｔｉｔｉｏｎ粒度におけるクラスタビューの制御を用いて大幅に狭められることが可能であり、したがって、ストレージシステムは、大規模に拡張されることが可能であり、システム拡張性が向上される。

【０１２２】

ＩＯ要求がｐａｒｔｉｔｉｏｎ　Ｘ一次ＯＳＤノードに着信した後、ＭＤＣが、新たなＯＳＤノードがクラスタに参加して、ｐａｒｔｉｔｉｏｎ　Ｘの二次ＯＳＤノードの役割をすることを発見し、ＩＯ処理手順は、以下を含む。すなわち、

ＭＤＣモジュールが、ＩＯ要求を処理するプロセスにおいて、新たなＯＳＤノードがクラスタに参加することを決定した場合、ｐａｒｔｉｔｉｏｎ　Ｘ一次ＯＳＤノードに、その新たなＯＳＤノードが、ｐａｒｔｉｔｉｏｎ　Ｘが位置付けられる新たな二次ＯＳＤノードの役割をすることを通知し、ｐａｒｔｉｔｉｏｎデータ同期が完了された後、ｐａｒｔｉｔｉｏｎグループのｐａｒｔｉｔｉｏｎ　ｖｉｅｗ及びＩＯ　ｖｉｅｗを更新し、ｐａｒｔｉｔｉｏｎ　Ｘ一次ＯＳＤノード上にローカルで記憶されたｐａｒｔｉｔｉｏｎ　ｖｉｅｗを更新するようｐａｒｔｉｔｉｏｎ　Ｘ一次ＯＳＤノードに指示し、かつ

ｐａｒｔｉｔｉｏｎ　Ｘ一次ＯＳＤノードが、ｐａｒｔｉｔｉｏｎ　Ｘ一次ＯＳＤノード上の一次ｐａｒｔｉｔｉｏｎのデータを新たな二次ＯＳＤノードに同期させ、更新されたローカルで記憶されたｐａｒｔｉｔｉｏｎにより、その新たな二次ＯＳＤノードにレプリカ要求を送信する。

【０１２３】

ＯＳＤ障害処理手順

【０１２４】

ストレージノード障害は、大規模な分散ストレージシステムにおいてよくある。いくつかのＯＳＤノードが障害を有する場合、システムは、ＩＯサービスを正常に提供することができる必要がある。本発明のこの実施形態において、すべてのＩＯ要求に対する処理は、ＭＤＣモジュールによって保持されるクラスタビューに依存し、クラスタにおけるＯＳＤノードが障害を有するようになった場合、クラスタビューも相応するように更新されて、ＩＯ要求が適切に、効率的に処理され得るようにする必要がある。

【０１２５】

ＯＳＤノード障害がＩＯ要求のレプリケーションに対する適切な処理に影響を及ぼすのを防止するのに、以下の処理が一般に実行される必要がある。第１に、ＭＤＣモジュールが、ＯＳＤノードのステータスを検出し、ＯＳＤノードが障害を有するようになった場合、ＭＤＣモジュールが、その障害を時間内に発見することが可能であり、第２に、ＯＳＤ障害を発見した後、ＭＤＣモジュールが、ビューに正しい変更を行う処理を時間内に実行し、その変更について関連するＩＯルーティングモジュール及びＯＳＤノードに通知する必要があり、第３に、関連するＩＯルーティングモジュール及びＯＳＤノードが、ＭＤＣの更新通知を受信した後、更新されたビューにより対応するＩＯ要求を処理して、モジュール及びノードが、更新されたビューを時間内に獲得することができるようにして、その結果、ＩＯ要求が平滑に、効果的に処理されることを確実にする。

【０１２６】

ＭＤＣモジュールは、以下の２つのモードにおいてＯＳＤノード障害を検出することが可能である。すなわち、（１）ＭＤＣモジュールが、すべてのＯＳＤノード上の障害検出

を担い、各ＯＳＤノードが、ＭＤＣモジュールにハートビートメッセージを規則的に送信し、ＯＳＤノードが、指定された期間内にＭＤＣモジュールにハートビートメッセージを送信しない場合、ＭＤＣモジュールは、ＯＳＤノードが障害を有するようになったと決定し、（２）ＯＳＤノードが、互いにハートビートメッセージを規則的に送信して障害を検出し、検出する側が、指定された期間内に検出される側のハートビートメッセージを受信しない場合、検出する側が、対応するＯＳＤノードが障害を有するようになったとＭＤＣモジュールに報告する。

【０１２７】

前述したとおり、ｐａｒｔｉｔｉｏｎ ｖｉｅｗとＩＯ ｖｉｅｗはともにｐａｒｔｉｔｉｏｎグループの点で説明され、一般に、ＯＳＤノード上に複数のｐａｒｔｉｔｉｏｎが存在する。１つのＯＳＤノードの障害は、複数のｐａｒｔｉｔｉｏｎを含む複数のｐａｒｔｉｔｉｏｎグループに関するビュー更新と関係し、すべてのｐａｒｔｉｔｉｏｎグループに関するビュー更新は、互いに独立であり、すなわち、

（１）障害を有するＯＳＤノード上のｐａｒｔｉｔｉｏｎが二次ｐａｒｔｉｔｉｏｎを含む場合、二次ｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループのｐａｒｔｉｔｉｏｎ ｖｉｅｗにおける障害を有するＯＳＤノードのｐａｒｔｉｔｉｏｎステータスに不整合であるという印が付けられ、同時に、ｐａｒｔｉｔｉｏｎグループの一次ＯＳＤノードに更新されたｐａｒｔｉｔｉｏｎ ｖｉｅｗについての通知が行われ、ｐａｒｔｉｔｉｏｎプロセスの一次ＯＳＤノードが、更新されたｐａｒｔｉｔｉｏｎ ｖｉｅｗにより、ＩＯ要求に対応するデータのレプリケーションを処理し、

（２）障害を有するＯＳＤノード上のｐａｒｔｉｔｉｏｎが一次ｐａｒｔｉｔｉｏｎを含む場合、一次ｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループのｐａｒｔｉｔｉｏｎ ｖｉｅｗにおいて一次ＯＳＤノードの役割をする障害を有するＯＳＤノードが、新たな二次ＯＳＤノードとして設定され、新たな二次ＯＳＤノードに対応するｐａｒｔｉｔｉｏｎステータスに不整合であるという印が付けられ、ｐａｒｔｉｔｉｏｎステータスに整合性がある二次ＯＳＤノードが、一次ｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループのｐａｒｔｉｔｉｏｎ ｖｉｅｗにおける元の二次ＯＳＤノードから選択され、選択された二次ＯＳＤノードが新たな一次ＯＳＤノードとして設定され、次に、その新たな一次ＯＳＤノードにｐａｒｔｉｔｉｏｎ ｖｉｅｗ更新についての通知が行われ、別の二次ＯＳＤノードにＩＯ ｖｉｅｗ更新についての通知が行われ、ｐａｒｔｉｔｉｏｎグループの一次ＯＳＤが障害を有するようになり、かつすべての二次ｐａｒｔｉｔｉｏｎが位置付けられるＯＳＤノードのｐａｒｔｉｔｉｏｎステータスが不整合である場合、ｐａｒｔｉｔｉｏｎ ｖｉｅｗ及びＩＯ ｖｉｅｗに対して変更はまったく行われず、一次ｐａｒｔｉｔｉｏｎレプリカが、最新の完全なデータを有することを確実にして、その結果、データレプリケーション整合性を確実にすることが要求される。

【０１２８】

関連するＩＯルーティングモジュール及びＯＳＤノードが、ＭＤＣの更新通知を受信した後、更新されたビューにより対応するＩＯ要求を処理することは、

新たな一次ＯＳＤノードが、更新されたローカルで記憶されたｐａｒｔｉｔｉｏｎ ｖｉｅｗにより、ＩＯルーティングモジュールからのＩＯ要求に対応するデータを、更新されたローカルで記憶されたｐａｒｔｉｔｉｏｎ ｖｉｅｗにおける、ｐａｒｔｉｔｉｏｎステータスに整合性があるｐａｒｔｉｔｉｏｎが位置付けられるｐａｒｔｉｔｉｏｎ上の二次ＯＳＤノード上に、又は更新されたローカルで記憶されたｐａｒｔｉｔｉｏｎ ｖｉｅｗに位置付けられる、ｐａｒｔｉｔｉｏｎステータスが不整合であるが、データを回復しているｐａｒｔｉｔｉｏｎ上の二次ＯＳＤノードに障害を隔離するように、かつ適切な、中断のないＩＯ要求処理を確実にするようにして、その結果、システムのフォールトトレランスを向上させ、かつシステムのパフォーマンス及び利用可能性を相応するように向上させることを特に含み得る。さらに、ＯＳＤノード障害の影響範囲が、ｐａｒｔｉｔｉｏｎ粒度におけるクラスタビューの制御を用いて狭められることが可能であり、したがって、システムは、大規模に拡張されることが可能であり、システム拡張性が向上される。

【０１２９】
最後に、障害を有するＯＳＤノードがフェイルバック及びデータ回復を完了した後、ＭＤＣモジュールが、ｐａｒｔｉｔｉｏｎ　ｖｉｅｗ及びＩＯ　ｖｉｅｗをさらに更新し、さらに更新されたｐａｒｔｉｔｉｏｎ　ｖｉｅｗにおけるｐａｒｔｉｔｉｏｎが位置付けられる一次ＯＳＤノードに、そのさらに更新されたｐａｒｔｉｔｉｏｎ　ｖｉｅｗについて通知し、さらに更新されたｐａｒｔｉｔｉｏｎ　ｖｉｅｗにおけるｐａｒｔｉｔｉｏｎが位置付けられる二次ＯＳＤノードに、そのさらに更新されたＩＯ　ｖｉｅｗを送信して、そのさらに更新されたｐａｒｔｉｔｉｏｎ　ｖｉｅｗ又はＩＯ　ｖｉｅｗを受信するモジュール又はＯＳＤノードが、ローカルで記憶されたｐａｒｔｉｔｉｏｎ　ｖｉｅｗ又はＩＯ　ｖｉｅｗを更新し、かつそのさらに更新されたｐａｒｔｉｔｉｏｎ　ｖｉｅｗ又はＩＯ　ｖｉｅｗにより、ＩＯ要求に対応するデータのレプリケーションを処理するようにする。

【０１３０】
ビューは、時間内に更新されて、フェイルバックノードが、クラスタに迅速に参加してＩＯ要求を処理することができるようにされ、このことが、システムのパフォーマンス及び効率を向上させる。

【０１３１】
理解を容易にするのに、以下は、説明のために特定の実施形態を使用する。図８Ａ及び図８Ｂが、本発明によるＯＳＤノード障害処理手順の実施形態を示す。この実施形態は、図２Ａないし図２Ｃ、及び図５において説明される実施形態において言及されるＭＤＣモジュール、ＩＯルーティングモジュール、及びＯＳＤノードによって実行される。理解を容易にするのに、この実施形態におけるＯＳＤノードの中のＯＳＤｘノード、ＯＳＤｙノード、及びＯＳＤｚノードが、説明のための例として使用され、ＯＳＤｘノード、ＯＳＤｙノード、又はＯＳＤｚノードは、本発明のこの実施形態における分散レプリケーションプロトコル層における複数のＯＳＤノードのうちのいずれか１つのＯＳＤノードであることが可能である。さらに、理解を容易にするのに、この実施形態において、ＯＳＤｘノードは、ｐａｒｔｉｔｉｏｎグループ１（略してＰ１）の一次ＯＳＤノードであり、かつｐａｒｔｉｔｉｏｎグループｎ（略してＰｎ）の二次ＯＳＤノードであり、ＯＳＤｙノードは、Ｐｎの一次ＯＳＤノードであり、かつＰ１の二次ＯＳＤノードであり、ＯＳＤｚノードは、Ｐｎの二次ＯＳＤノードであり、かつＰ１の二次ＯＳＤノードであるものと想定される。この実施形態におけるＯＳＤノード障害処理手順は、以下のステップを特に含む。すなわち、

【０１３２】
８０２／８０４／８０６　ＯＳＤｘノード、ＯＳＤｙノード、及びＯＳＤｚノードが、一次ＭＤＣモジュールにハートビートメッセージを別々に、規則的に送信する。

【０１３３】
８０８　ＯＳＤｘノードが、障害を有するようになる。

【０１３４】
実際のアプリケーションにおいて、様々なソフトウェア障害、ハードウェア障害、又はネットワーク障害が存在することが可能である。例えば、ソフトウェアＢＵＧ、一時的ネットワーク切断、サーバ再起動などによってもたらされるプログラムプロセス再起動のため、ＯＳＤノードが、ＩＯ要求を処理することができず、データ記憶機能及びデータレプリケーション機能を実施することができない。

【０１３５】
８１０　ＭＤＣモジュールが、ＯＳＤｘノードが所定の時間内にハートビートメッセージを送信しないことを検出した場合、ＭＤＣモジュールは、ＯＳＤｘノードが障害を有するようになると決定する。

【０１３６】
実際の応用形態において、前述したとおり、ＭＤＣは、別の様態で、ＯＳＤｘノードが障害を有するようになると決定することも可能である。例えば、機能するＯＳＤノードが

、ＯＳＤノードが障害を有するようになることをＭＤＣモジュールに通知する。
【０１３７】
８１２　ＭＤＣモジュールが、障害状況によりビュー更新を実行する。
【０１３８】
特定の実施様態において、ＭＤＣは、決定された障害を有するＯＳＤノード上のｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループにより、対応するｐａｒｔｉｔｉｏｎグループのクラスタビューを更新する。この実施形態において、障害を有するＯＳＤｘノード上のｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎは、Ｐ１及びＰｎを含み、したがって、ＭＤＣは、Ｐ１及びＰｎのクラスタビューを更新する必要があり、このことは、
（１）ＯＳＤ　ｖｉｅｗを更新すること、すなわち、ＯＳＤｘノードのステータスが、「イン（ＩＮ）」かつ「アップ（ＵＰ）」状態から「イン（ＩＮ）」かつ「ダウン（ＤＯＷＮ）」状態に更新されること、
（２）ｐａｒｔｉｔｉｏｎ　ｖｉｅｗを変更すること、すなわち、Ｐ１に関して、ＯＳＤｙノードがＰ１の一次ＯＳＤノードとして設定され（ｐａｒｔｉｔｉｏｎステータスに整合性がある第１の二次ＯＳＤノードが、Ｐ１の二次ＯＳＤノードリストから選択され、ｐａｒｔｉｔｉｏｎグループの一次ＯＳＤノードとして設定される）、ＯＳＤｘがＰ１の二次ＯＳＤノードとして設定され、ＯＳＤｘの対応するｐａｒｔｉｔｉｏｎステータスが「不整合である」に更新され、Ｐｎに関して、Ｐｎの二次ＯＳＤｘのｐａｒｔｉｔｉｏｎステータスが不整合に変更されること、
（３）ＩＯ　ｖｉｅｗを更新すること、すなわち、Ｐ１に関して、元のＯＳＤｘノードが、Ｐ１のＩＯ　ｖｉｅｗにおける一次ＯＳＤノードとしてＯＳＤｙノードで置き換えられ、Ｐｎに関して、障害を有するＯＳＤｘノードは、Ｐｎの二次ＯＳＤノードの役割をするに過ぎず、Ｐｎの一次ＯＳＤｙノードは、障害を有するようにならないため、ＰｎのＩＯ　ｖｉｅｗは更新されないことを特に含み得る。
【０１３９】
８１４　ＯＳＤｙノード（更新の後、Ｐ１の一次ＯＳＤノードの役割をし、依然として、Ｐｎの一次ＯＳＤノードの役割をする）に、Ｐ１及びＰｎのｐａｒｔｉｔｉｏｎ　ｖｉｅｗの更新について通知し、更新することは、ＯＳＤｙノードをＰ１の一次ＯＳＤノードとして設定すること、ＯＳＤｘをＰ１の二次ＯＳＤノードとして設定し、かつＯＳＤｘの対応するｐａｒｔｉｔｉｏｎステータスを「不整合である」に設定すること、及びＰｎの二次ＯＳＤｘノードのｐａｒｔｉｔｉｏｎステータスを「不整合である」に変更することを含む。
【０１４０】
８１６　ＯＳＤｚノードにＰ１のＩＯ　ｖｉｅｗの更新について通知する、すなわち、ＯＳＤｚノードに、Ｐ１の一次ＯＳＤノードがＯＳＤｙノードで置き換えられることを通知する。
【０１４１】
８１８　ＩＯルーティングモジュールにＰ１のＩＯ　ｖｉｅｗの更新について通知する、すなわち、ＩＯルーティングモジュールに、Ｐ１の一次ＯＳＤノードがＯＳＤｙノードで置き換えられることを通知する。
【０１４２】
８２０　ＯＳＤｙノードが、ＭＤＣモジュールの通知を処理し、ローカルで記憶されたビュー情報（ｐａｒｔｉｔｉｏｎ　ｖｉｅｗ及びＩＯ　ｖｉｅｗ）を更新し、ＭＤＣモジュールによって通知される最新のビューにより、ＩＯ要求に対応するデータのレプリケーションを処理する。
【０１４３】
特定の実施様態において、更新の後のＰ１の一次ＯＳＤノードとして、ＯＳＤｙノードが、Ｐ１のｐａｒｔｉｔｉｏｎ　ｖｉｅｗ及びＩＯ　ｖｉｅｗを更新し、Ｐｎの元のＯＳＤノードとして、ＯＳＤｙノードが、Ｐｎのｐａｒｔｉｔｉｏｎ　ｖｉｅｗを更新する。
【０１４４】

Ｐ１上のＩＯ操作に関して、ＩＯルーティングモジュールによって転送されたＩＯ要求を受信した後、ＯＳＤｙノードは、ＩＯ要求を実行し、レプリケーション要求を生成し、更新されたｐａｒｔｉｔｉｏｎ　ｖｉｅｗにおけるＰ１の二次ＯＳＤノード、すなわち、ＯＳＤｚノードにレプリケーション要求を送信し、ＯＳＤｚノードの対応するｐａｒｔｉｔｉｏｎステータスは、「整合性がある」。ＯＳＤｘノードは、Ｐ１の新たな二次ＯＳＤノードの役割をし、ＯＳＤｘノードのｐａｒｔｉｔｉｏｎステータスは、「不整合である」ため、ＯＳＤｙノードは、ＯＳＤｘノードにレプリケーション要求をもはや送信せず、このことが、障害隔離を実施し、Ｐ１上の連続的なＩＯ要求処理に影響を及ぼさない。

【０１４５】

Ｐｎ上のＩＯ操作に関して、ＩＯルーティングモジュールによって転送されたＩＯ要求を受信した後、ＯＳＤｙノードは、ＩＯ要求を実行し、レプリケーション要求を生成し、更新されたｐａｒｔｉｔｉｏｎ　ｖｉｅｗにおけるＰｎの二次ＯＳＤノード、すなわち、ＯＳＤｚノードにレプリケーション要求を送信し、ＯＳＤｚノードの対応するｐａｒｔｉｔｉｏｎステータスは、「整合性がある」。ＯＳＤｘノードは、Ｐｎの二次ＯＳＤノードの役割をし、ＯＳＤｘノードのｐａｒｔｉｔｉｏｎステータスは、「不整合である」ため、ＯＳＤｙノードは、ＯＳＤｘノードにレプリケーション要求をもはや送信せず、このことが、障害隔離を実施し、Ｐ１上の連続的なＩＯ要求処理に影響を及ぼさない。

【０１４６】

２フェーズコミットプロトコルにおいて、参加者が障害を有するようになった場合、又は応答を有さないことが続く場合、別のＩＯ要求は、継続的に阻止され、タイムアウトのために最終的に失敗し、ロールバックが実行される必要がある。この実施形態において、二次ＯＳＤノードが障害を有するようになった場合、ＭＤＣが、ビュー変更を行うよう一次ノードに指示して、障害を有するノードを無視するようにし、障害を有するノードを隔離するようにし、別のＩＯ要求に対する処理を阻止することなしにＩＯ要求処理を継続するようにし、このことが、より良好なフォールトトレランス及び利用可能性を有する。

【０１４７】

８２２　ＭＤＣモジュールのビュー通知を処理し、ローカルで記憶されたＩＯ　ｖｉｅｗ情報を更新する。

【０１４８】

８２４　ＭＤＣモジュールのビュー通知を処理し、ローカルで記憶されたＩＯ　ｖｉｅｗ情報を更新し、ＭＤＣモジュールによって通知される最新のＩＯビューにより、ＩＯ要求ルーティング及び転送を実行する。

【０１４９】

特定の実施様態において、Ｐ１上のＩＯ処理に関して、Ｐ１が位置付けられる元の一次ＯＳＤｘノードが、Ｐ１のＩＯ処理プロセスにおいて障害を有するようになった場合、ＭＤＣが、Ｐ１のｐａｒｔｉｔｉｏｎ　ｖｉｅｗを時間内に更新し、ＩＯルーティングモジュールが、更新されたｐａｒｔｉｔｉｏｎ　ｖｉｅｗにおけるＰ１の新たに選択された一次ＯＳＤｙノードにより、新たに選択されたＯＳＤｙノードにＩＯ要求を再び転送する。

【０１５０】

本発明のこの実施形態において、一次ＯＳＤノードが障害を有するようになった後、ＭＤＣノードは、投票を用いて新たな一次ノードを迅速に決定すること、及びＩＯ処理を迅速に再開することができ、二次ノードが障害を有するようになった場合、ＭＤＣは、ビュー変更を行うよう一次ノードに指示して、障害を有するＯＳＤノードを隔離する、又は無視するようにし、別のＩＯ要求に対する処理を阻止することなしにＩＯ要求処理を継続するようにし、このことが、より良好なフォールトトレランスを有し、ノード障害を迅速に扱うことができる。例えば、Ｎ＋１のレプリカのうちのＮのレプリカの障害が許容されることが可能であり、このことが、ストレージシステムのパフォーマンス及び利用可能性をさらに向上させる。低い利用可能性を有するシステムは、劣悪な拡張性を不可避に有し、ストレージノード障害は、大規模な分散ストレージシステムに一般的であるため、複雑で、大量のプロトコル対話が、システム拡張性をさらに低減する可能性がある。さらに、ス

トレージノード障害の影響範囲が、ｐａｒｔｉｔｉｏｎ粒度におけるクラスタビューの制御を用いて大幅に狭められることが可能であり、したがって、ストレージシステムは、大規模に拡張されることが可能であり、システム拡張性が向上される。

【０１５１】

ＯＳＤノードフェイルバック手順

【０１５２】

新たなデータ変更操作が、ＯＳＤノードの障害中に生じる可能性があり、したがって、障害を有するＯＳＤノードが正常に戻り、クラスタに再び参加してサービスを提供する前に、データ回復及びデータ同期がまず実行されて、障害を有するＯＳＤノードを、一次レプリカのものと整合性がある状態に戻すようにする必要がある。

【０１５３】

本発明の実施形態において、ＯＳＤノードフェイルバック手順は、３つのフェーズに分割され得る。すなわち、

（１）二次ＯＳＤノードが、一次ＯＳＤノードと、障害中に一次ＯＳＤノードによって実行されたデータ変更を同期し、これは、インクリメンタルな同期プロセスであり、無論、実際の応用形態において、ｐａｒｔｉｔｉｏｎのすべてのデータは、実際の状況により同期され得る。

（２）二次ＯＳＤノードが一次ＯＳＤノードのものと整合性がある状態に戻った後、ＭＤＣモジュールが、クラスタビューを変更する。

（３）ＭＤＣモジュールが、各モジュール及び各ノードに更新されたクラスタビューについて通知して、各モジュール及び各ノードが、通知された更新されたクラスタビューによりＩＯ要求レプリケーション又はＩＯ要求転送を処理するようにする。

【０１５４】

実際の応用形態において、一次ＯＳＤノードが、データ回復プロセスにおいてＩＯ要求を受信した後、フェイルバックノードにレプリケーション要求を送信した場合、ＯＳＤノードフェイルバック手順は、以下のフェーズをさらに含むことが可能である。すなわち、

【０１５５】

二次ＯＳＤノードが、一次ＯＳＤノードのレプリケーション要求が受信された後に、一次ＯＳＤとデータを同期するプロセスにおいて記録されたログを再生し、ログに入っているデータを書き込む。このようにして、フェイルバックプロセスにおいて、フェイルバックＯＳＤノードのすべてのデータが一次ＯＳＤノードのものと整合性があることが確実にされて、その結果、一次ＯＳＤノードと二次ＯＳＤノードの間のデータ整合性をさらに向上させることが可能である。

【０１５６】

データ同期プロセスの特定の手順に関しては、後段の図１０Ａ及び図１０Ｂにおいて与えられる特定の実施形態を参照されたい。

【０１５７】

クラスタビュー更新及び更新通知プロセスは、以下の事例を含むことが可能である。すなわち、

（１）ＯＳＤノードが、障害前にｐａｒｔｉｔｉｏｎグループの二次ＯＳＤノードである場合、ｐａｒｔｉｔｉｏｎ　ｖｉｅｗだけが変更されることが可能であり、ｐａｒｔｉｔｉｏｎ　ｖｉｅｗに対応する二次ＯＳＤノードのｐａｒｔｉｔｉｏｎステータスは、「整合性がある」状態に変更され、変更されたｐａｒｔｉｔｉｏｎ　ｖｉｅｗが一次ＯＳＤノードに通知される。

（２）ＯＳＤノードが、障害前にｐａｒｔｉｔｉｏｎグループの一次ＯＳＤノードである場合、ｐａｒｔｉｔｉｏｎ　ｖｉｅｗとＩＯ　ｖｉｅｗの両方が変更され、前のＯＳＤノードがｐａｒｔｉｔｉｏｎグループの新たな一次ＯＳＤノードに設定され、現在の一次ＯＳＤがｐａｒｔｉｔｉｏｎグループの二次ＯＳＤノードとして設定され、新たな一次ＯＳＤノードがｐａｒｔｉｔｉｏｎ　ｖｉｅｗを変更するよう指示され、ＩＯルーティングモジュール及び二次ＯＳＤノードがＩＯ　ｖｉｅｗを変更するよう指示される。

【０１５８】
実際の応用形態において、障害を有するＯＳＤノードが、指定された期間内に正常に戻らない場合、ＭＤＣモジュールが、そのＯＳＤノードをクラスタから追放し、そのＯＳＤノード上に配分されたｐａｒｔｉｔｉｏｎを別のＯＳＤノードに移行させる。特定の手順に関しては、後段の図１１Ａ及び図１１Ｂにおいて与えられる、ＯＳＤノードがクラスタから出る特定の実施形態を参照されたい。
【０１５９】
理解を容易にするのに、以下に、説明のために特定の実施形態を使用する。図９Ａ、図９Ｂ、及び図９Ｃが、本発明によるＯＳＤノードフェイルバック処理手順の実施形態を示す。この実施形態は、図２Ａないし図２Ｃ、及び図５において説明される実施形態において言及されるＭＤＣモジュール、ＩＯルーティングモジュール、及びＯＳＤノードによって実行される。理解を容易にするのに、この実施形態において、フェイルバックＯＳＤノード上のｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループが、ｐａｒｔｉｔｉｏｎグループ１（略してＰ１）と、ｐａｒｔｉｔｉｏｎグループｎ（略してＰｎ）とを含む例が、説明のために使用され、この実施形態におけるＯＳＤノード障害処理手順は、以下のステップを特に含むものと想定される。すなわち、
【０１６０】
９０２　フェイルバックＯＳＤノードが、フェイルバックＯＳＤノードに関するクラスタビュー情報をＭＤＣモジュールに要求し、要求は、そのＯＳＤノードのＯＳＤ　ＩＤを搬送する。
【０１６１】
特定の実施様態に関しては、図６のステップ６０８において説明される前述のプロセスを参照されたい。
【０１６２】
９０４　ＭＤＣモジュールが、そのＯＳＤ　ＩＤによりｐａｒｔｉｔｉｏｎ　ｖｉｅｗのクエリを行って、そのＯＳＤに関するｐａｒｔｉｔｉｏｎ情報を獲得する。
【０１６３】
特定の実施様態において、ＭＤＣは、フェイルバックＯＳＤのＯＳＤ　ＩＤにより、フェイルバックＯＳＤ上のＰ１及びＰｎに対応するｐａｒｔｉｔｉｏｎ　ｖｉｅｗのクエリを別々に行って、Ｐ１のｐａｒｔｉｔｉｏｎ情報とＰｎのｐａｒｔｉｔｉｏｎ情報を別々に獲得する。ｐａｒｔｉｔｉｏｎ情報は、ＩＯ　ｖｉｅｗを含むことが可能であり、ｐａｒｔｉｔｉｏｎステータスをさらに含むことが可能である。
【０１６４】
９０６　ＯＳＤノードにｐａｒｔｉｔｉｏｎ情報を返す。
【０１６５】
９０８／９１０　ＭＤＣによって返されるｐａｒｔｉｔｉｏｎ情報によりｐａｒｔｉｔｉｏｎグループの一次ＯＳＤノードに対してデータ回復プロセスを開始し、障害中に失われたｅｎｔｒｙ情報をＰ１とＰｎの一次ＯＳＤノードから別々に獲得し、ｅｎｔｒｙ情報は、ＳｅｑＩＤを搬送する。
【０１６６】
本発明の特定の実施形態において、一次ＯＳＤノード（例えば、Ｐｎの一次ＯＳＤノード）が、フェイルバックプロセスにおいてＩＯ書込み要求を受信した場合、一次ＯＳＤノードが、Ｐｎのすべての二次ＯＳＤノードにレプリケーション要求を送信する必要があり、フェイルバック手順は、以下のステップ９１２～９１６、及びステップ９１８をさらに含むことが可能である。
【０１６７】
９１２　Ｐｎの一次ＯＳＤノードが、フェイルバック中にホストからＩＯ書込み要求を受信する。
【０１６８】
９１４　Ｐａｒｔｉｔｉｏｎが位置付けられる二次ＯＳＤノードにレプリケーションＩ

Ｏ情報を送信する。
【０１６９】
特定の実施様態において、前述のステップ９１２～９１４に関しては、図７Ａ及び図７Ｂにおける前述のＩＯ操作手順を参照されたい。
【０１７０】
９１６　ログ（ｌｏｇ）にレプリケーションＩＯキー情報を書き込む。
【０１７１】
特定の実施様態において、ＩＯ要求に対応するデータは、ログに書き込まれることも可能であり、このステップにおけるＩＯキー情報に関しては、図７Ａ及び図７Ｂにおける前述のＩＯ操作手順の説明を参照されたい。
【０１７２】
９１８　一次ＯＳＤノードによって返される異なるｅｎｔｒｙ数量情報により、かつｅｎｔｒｙによりデータ要求メッセージを繰り返し送信して、データ回復プロセスが完了されることを保証する。
【０１７３】
特定の実施様態に関しては、後段の図１１Ａ及び図１１Ｂにおけるデータ回復手順を参照されたい。
【０１７４】
９２０　ログに記録されたＩＯ情報により、かつログの記録によりＩＯを書き込む。
【０１７５】
特定の実施様態において、データ回復、及び一次ＯＳＤノードとの同期を完了した後、フェイルバックＯＳＤノードは、フェイルバックＯＳＤノードによって管理される物理ストレージリソースに、ログに記録されたＩＯ情報により、Ｐｎの一次ＯＳＤからのレプリケーション要求に対応するデータを書き込む。
【０１７６】
ＩＯ要求は、フェイルバックプロセスにおいてログに書き込まれ、次に、障害中に生成されたＩＯ要求が、障害中に失われたデータが回復された後に書き込まれ、このことが、一次ＯＳＤノード及び二次ＯＳＤノードがすべてのＩＯ操作を実行するシーケンスに整合性があることを確実にすること、及びデータバックアップ整合性をさらに向上させることが可能である。
【０１７７】
９２２／９２４　データ回復を完了し、Ｐ１とＰｎの一次ＯＳＤノードに別々に通知する。
【０１７８】
特定の実施様態において、ＭＤＣモジュールは、２つの様態で、クラスタビューの更新を実行するよう指示されることが可能であり、様態１において、データ回復が完了された後、フェイルバックＯＳＤノードが一次ＯＳＤノードにＰ１及びＰｎについて通知して、一次ＯＳＤノードが、クラスタビューを更新するようＭＤＣノードに指示するようにする。様態２に関しては、以下のステップ９３０を参照されたい。
【０１７９】
実際の応用形態において、クラスタビューを更新するよう指示する前に、フェイルバックＯＳＤノードは、フェイルバックＯＳＤノード上のｐａｒｔｉｔｉｏｎのｐａｒｔｉｔｉｏｎステータスをさらに決定することが可能であり、ｐａｒｔｉｔｉｏｎステータスが不整合であると決定した後、クラスタビュー更新手順をトリガする。フェイルバックＯＳＤノードは、前述のステップ９０６において返されるｐａｒｔｉｔｉｏｎ情報を収集することによってｐａｒｔｉｔｉｏｎステータス情報をさらに獲得することが可能である。
【０１８０】
９２６／９２８　Ｐ１とＰｎの一次ＯＳＤノードは、ｐａｒｔｉｔｉｏｎ　ｖｉｅｗを変更するようＭＤＣに別々に指示し、通知が、ｐａｒｔｉｔｉｏｎグループＩＤ、二次ＯＳＤのＩＤ、及びビューバージョンを搬送する。

【０１８１】
特定の実施様態において、フェイルバックＯＳＤノード上のｐａｒｔｉｔｉｏｎのｐａｒｔｉｔｉｏｎステータスを整合性があるに更新するようＭＤＣモジュールに要求する通知が、ＭＤＣモジュールに送信され、通知は、ＯＳＤノード上のｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループのｐａｒｔｉｔｉｏｎグループＩＤ、二次ＯＳＤノード（すなわち、フェイルバックＯＳＤノード）のＩＤ、及びビューバージョンを搬送する。本明細書におけるｐａｒｔｉｔｉｏｎグループＩＤは、ｐａｒｔｉｔｉｏｎグループの更新されるべきビューに印を付けるのに使用され、ＯＳＤ　ＩＤは、障害を有するＯＳＤノードに印を付けるのに使用され、本明細書におけるビューバージョンは、Ｐ１及びＰｎが位置付けられる一次ＯＳＤノード上にローカルで記憶された最新のｐａｒｔｉｔｉｏｎ　ｖｉｅｗのビューバージョンであり、ビューバージョンの機能は、通知を受信して、通知におけるｐａｒｔｉｔｉｏｎ　ｖｉｅｗのビューバージョンが、ＭＤＣによってローカルで保持される最新のｐａｒｔｉｔｉｏｎ　ｖｉｅｗのビューバージョンと整合性があると決定した後、ＭＤＣが、クラスタビュー更新処理を実行して、その結果、ＩＯ処理手順においてすべてのモジュール又はすべてのノードによって見られるクラスタビューに整合性があることを確実にすること、及びデータバックアップ整合性を向上させることである。

【０１８２】
実際の応用形態において、通知におけるｐａｒｔｉｔｉｏｎ　ｖｉｅｗのビューバージョンが、ＭＤＣによってローカルで保持される最新のｐａｒｔｉｔｉｏｎ　ｖｉｅｗのビューバージョンと不整合であると決定された場合、最新のｐａｒｔｉｔｉｏｎ　ｖｉｅｗは、Ｐ１及びＰｎが位置付けられる一次ＯＳＤノードにさらに送信され、Ｐ１及びＰｎの一次データと二次データに整合性があると決定された後、クラスタビューが更新される。

【０１８３】
９２８　Ｐａｒｔｉｔｉｏｎ　Ｖｉｅｗを変更するよう、かつフェイルバックＯＳＤノードのステータスを整合性があるに更新するようＭＤＣに指示し、指示が、ｐａｒｔｉｔｉｏｎグループＩＤ、二次ＯＳＤのＩＤ、及びビューバージョンを搬送する。

【０１８４】
９３０　データ回復を完了し、Ｐａｒｔｉｔｉｏｎ　Ｖｉｅｗを変更するよう一次ＭＤＣに指示し、通知が、ＰａｒｔｉｔｉｏｎグループＩＤ、二次ＯＳＤのＩＤ、及びビューバージョンを搬送する。

【０１８５】
特定の実施様態において、データ回復を完了した後、フェイルバックＯＳＤノードは、ＭＤＣモジュールに通知を送信して、フェイルバックＯＳＤノード上のｐａｒｔｉｔｉｏｎのｐａｒｔｉｔｉｏｎステータスを整合性があるに更新するようＭＤＣモジュールに指示する。前述の様態１と様態２の間の違いは、この場合のビューバージョンが、フェイルバックＯＳＤノード上にローカルで記憶された最新のｐａｒｔｉｔｉｏｎ　ｖｉｅｗのｐａｒｔｉｔｉｏｎ　ｖｉｅｗバージョン又は最新のＩＯ　ｖｉｅｗのＩＯ　ｖｉｅｗバージョンであることであり（ビューバージョンがｐａｒｔｉｔｉｏｎ　ｖｉｅｗであるか、又はＩＯ　ｖｉｅｗであるかは、フェイルバックＯＳＤノードが、障害前に一次ＯＳＤノードであるか、二次ノードであるかに特に依存する）、通知におけるビューバージョンにより、ＭＤＣモジュールによってローカルで保持される対応するビューバージョンに整合性があると決定した後、ＭＤＣモジュールが、クラスタビュー更新処理を実行することである。

【０１８６】
９３２　Ｐａｒｔｉｔｉｏｎ　ＶｉｅｗにおけるフェイルバックＯＳＤノードのｐａｒｔｉｔｉｏｎステータスを整合性があるに更新する。

【０１８７】
特定の実施様態において、フェイルバックＯＳＤノード上のＰ１及びＰｎのｐａｒｔｉｔｉｏｎ　ｖｉｅｗにおいてフェイルバックＯＳＤノードに対応するｐａｒｔｉｔｉｏｎステータスが、「整合性がある」に更新される。

【０１８８】
９３４　ビュー比較を用いて、フェイルバックＯＳＤノードが、障害前にいくつかのｐａｒｔｉｔｉｏｎの一次ＯＳＤノードであるかどうかを決定する。
【０１８９】
特定の実施様態において、フェイルバックＯＳＤノード上のＰ１及びＰｎのものであり、かつＭＤＣモジュールによってローカルで保持される最新のｐａｒｔｉｔｉｏｎ　ｖｉｅｗを、Ｐ１及びＰｎの初期化されたｐａｒｔｉｔｉｏｎ　ｖｉｅｗと比較することによって、フェイルバックＯＳＤノードがＰ１の一次ＯＳＤノードであるかどうかが決定される、又はフェイルバックＯＳＤノード上のＰ１及びＰｎのものであり、かつＭＤＣモジュールによってローカルで保持される最新のＩＯ　ｖｉｅｗを、Ｐ１及びＰｎの初期化されたＩＯ　ｖｉｅｗと比較することによって、フェイルバックＯＳＤノードがＰ１の一次ＯＳＤノードであるかどうかが決定される。
【０１９０】
９３６　フェイルバックノードを一次ＯＳＤノードとして再設定し、Ｐａｒｔｉｔｉｏｎ　Ｖｉｅｗを更新する。
【０１９１】
特定の実施様態において、Ｐ１に関して、ＭＤＣが、フェイルバックＯＳＤノードをＰ１の新たな一次ＯＳＤノードとして再設定し、Ｐ１が位置付けられる現在の一次ＯＳＤノードをＰ１の新たな二次ＯＳＤノードとして設定し、Ｐ１のｐａｒｔｉｔｉｏｎ　ｖｉｅｗを更新し、Ｐｎに関して、フェイルバックＯＳＤノードが最初にＰｎの二次ＯＳＤノードであるため、Ｐｎに関して、フェイルバックＯＳＤノードの一次／二次識別変更問題は関与しない。
【０１９２】
９３８　フェイルバックＯＳＤノードが最初に一次ノードである場合、最新のＰａｒｔｉｔｉｏｎ　Ｖｉｅｗをフェイルバックノードに送信する。
【０１９３】
特定の実施様態において、フェイルバックＯＳＤノードが最初にＰ１の一次ノードであり、かつフェイルバックＯＳＤノードが、ステップ９３７によりフェイルバックＯＳＤノードの新たな一次ＯＳＤノードとして設定されるため、Ｐ１の更新された最新のＰａｒｔｉｔｉｏｎ　Ｖｉｅｗが、フェイルバックＯＳＤノードの新たな一次ＯＳＤノード、すなわち、フェイルバックＯＳＤノードに送信される必要がある。フェイルバックＯＳＤノードは最初にＰｎの一次ノードではないため、Ｐｎの最新のｐａｒｔｉｔｉｏｎ　ｖｉｅｗが、フェイルバックＯＳＤノードに送信されないことが可能である。
【０１９４】
９４０／９４２　ローカルＩＯビューを更新するよう二次ＯＳＤに指示する。
【０１９５】
特定の実施様態において、Ｐ１及びＰｎの最新のｐａｒｔｉｔｉｏｎ　Ｖｉｅｗ又は最新のＩＯ　ｖｉｅｗにより、Ｐ１とＰｎの新たな一次ＯＳＤノードが別々に獲得され、Ｐ１とＰｎの最新のＩＯ　Ｖｉｅｗが、それぞれ、Ｐ１の新たな一次ＯＳＤノード、及びＰｎの一次ＯＳＤノードに送信される（Ｐｎに関して、一次ＯＳＤノードの変化は関与しないため、ＩＯ　ｖｉｅｗは、それでも、Ｐｎの元の一次ＯＳＤノードに送信される）。
【０１９６】
９４４　フェイルバックＯＳＤノードが、Ｐａｒｔｉｔｉｏｎ　Ｖｉｅｗとローカルで記憶されたＩＯ　ｖｉｅｗにおける一次ＯＳＤに整合性があるかどうかを決定し、フェイルバックＯＳＤノードが一次ノードとして設定されているかどうかを決定し、ＩＯ　ｖｉｅｗを更新する。
【０１９７】
最新のｐａｒｔｉｔｉｏｎ　ｖｉｅｗを受信した後、フェイルバックＯＳＤノードは、最新のｐａｒｔｉｔｉｏｎ　Ｖｉｅｗとローカルで記憶されたＩＯ　ｖｉｅｗにおける一次ＯＳＤに整合性があるかどうかを決定し、フェイルバックＯＳＤノードが一次ノードと

して設定されているかどうかを決定し、最新のｐａｒｔｉｔｉｏｎ Ｖｉｅｗとローカルで記憶されたＩＯ ｖｉｅｗにおける一次ＯＳＤに整合性があり、かつフェイルバックＯＳＤノードが一次ノードとして設定されている場合、ＩＯ ｖｉｅｗ及びローカルで記憶されたｐａｒｔｉｔｉｏｎ ｖｉｅｗを更新し、更新されたｐａｒｔｉｔｉｏｎ ｖｉｅｗ、及び更新されたＩＯ ｖｉｅｗにより、ＩＯ要求と関係するデータのレプリケーションを処理する。

【０１９８】

９４６ ＩＯビューを更新し、一次ＯＳＤが二次ＯＳＤとして設定されている場合、Ｐａｒｔｉｔｉｏｎ Ｖｉｅｗを削除する。

【０１９９】

特定の実施様態において、Ｐ１及びＰｎの現在の一次ＯＳＤノード（すなわち、データ回復が完了された後、更新が開始する前に存在するＯＳＤノード）が、更新されたＩＯ ｖｉｅｗを受信し、それぞれのローカルで記憶されたＩＯ ｖｉｅｗを更新する。Ｐ１の現在の一次ＯＳＤノードに関して、現在の一次ＯＳＤノードは、ＭＤＣモジュールによってＰ１の新たな二次ＯＳＤノードとして設定されているため、Ｐ１の現在の一次ＯＳＤノードが、ローカルで記憶されたｐａｒｔｉｔｉｏｎ ｖｉｅｗを削除し、更新されたローカルで記憶されたＩＯ ｖｉｅｗによりＩＯ要求のレプリケーションを処理する。

【０２００】

本発明のこの実施形態において、ＯＳＤノードが障害を有するようになった後、ＭＤＣが、ビュー更新を実行するよう別の関連するノードに指示して、障害を有するＯＳＤノードを隔離する、又は無視するようにし、別のＩＯ要求に対する処理を阻止することなしにＩＯ要求処理を継続するようにし、ノードフェイルバックの後、ビューを更新し、関連する各ノードに通知を行って、フェイルバックノードが、作業のためにクラスタに迅速に再び参加することができるようにし、このことが、ノード障害及びフェイルバックを迅速に処理することが可能であり、より良好なフォールトトレランスを有し、ストレージシステムのパフォーマンス及び利用可能性を向上させる。低い利用可能性を有するシステムは、劣悪な拡張性を不可避に有し、ストレージノード障害は、大規模な分散ストレージシステムにおいてよくあるため、複雑で、大量のプロトコル対話が、システム拡張性をさらに低減する可能性がある。さらに、ストレージノード障害の影響範囲が、ｐａｒｔｉｔｉｏｎ粒度におけるクラスタビューの制御を用いて大幅に狭められることが可能であり、したがって、ストレージシステムは、大規模に拡張されることが可能であり、システム拡張性が向上される。

【０２０１】

データ回復手順

【０２０２】

以下に、特定の実施形態を使用することによって図９Ａ、図９Ｂ、及び図９Ｃにおける前述のＯＳＤフェイルバック処理プロセスにおけるデータ回復処理手順を説明する。理解を容易にするのに、この実施形態において、フェイルバックＯＳＤノード上のｐａｒｔｉｔｉｏｎを含むｐａｒｔｉｔｉｏｎグループが、ｐａｒｔｉｔｉｏｎ１（略してＰ１）と、ｐａｒｔｉｔｉｏｎｎ（略してＰｎ）とを含む例が、説明のために使用されるものと想定される。図１０Ａ及び図１０Ｂに示されるとおり、この実施形態におけるＯＳＤノード障害処理手順は、以下のステップを特に含む。すなわち、

【０２０３】

１００２ フェイルバックＯＳＤノードが、各ｐａｒｔｉｔｉｏｎの記録されたｅｎｔｒｙにおける最大ＳｅｑＩＤをローカルで獲得する。

【０２０４】

図７Ａ及び図７Ｂの前述の実施形態におけるステップ７１２及びステップ７２２に示されるとおり、システムにおけるＯＳＤノードが、ＩＯ処理プロセスにおいてｐａｒｔｉｔｉｏｎ上の各ＩＯ操作に関して１つのｅｎｔｒｙを記録する。前述したとおり、ｅｎｔｒｙは、ＩＯ操作タイプと、Ｐａｒｔｉｔｉｏｎ ＩＤと、ＳｅｑＩＤと、ｋｅｙとを含み

、ｅｎｔｒｙは、操作が成功したかどうかを記述するのに使用されるステータス情報をさらに含むことが可能である。さらに、ＩＯ書込み操作に関して、ｅｎｔｒｙは、前述のｏｆｆｓｅｔ及びｌｅｎｇｔｈをさらに含むことが可能である。例えば、この実施形態において、Ｐ１及びＰｎの上のＩＯ書込み操作に関する最大のＳｅｑＩＤが別々に獲得される。

【０２０５】

１００４／１００６　ＯＳＤに欠落しているｅｎｔｒｙを獲得するようＰ１とＰｎの一次ＯＳＤノードに別々に要求し、その要求は、ｐａｒｔｉｔｉｏｎグループＩＤ及びそれぞれの最大のＳｅｑＩＤを搬送する。

【０２０６】

シナリオ１　二次ＯＳＤの最大のＳｅｑＩＤが、一次ＯＳＤによって記録されたｅｎｔｒｙの範囲に入る。

【０２０７】

１００８／１０１０　Ｐ１とＰｎの一次ＯＳＤノードが、フェイルバックＯＳＤノードに欠落しているｅｎｔｒｙをフェイルバックＯＳＤノードに別々に送信する。

【０２０８】

特定の実施様態において、フェイルバックＯＳＤノードの最大のＳｅｑＩＤが１．６であり、一次ＯＳＤの現在の最大のＳｅｑＩＤが１．１６である場合、１．７から１．１６までのＳｅｑＩＤを有する、対応する１０のｅｎｔｒｙが二次ＯＳＤに送信され、このことは、理解を容易にするよう与えられる例に過ぎない。実際の応用形態において、ＳｅｑＩＤ番号付け規則又は番号付け様態は、異なることが可能であり、Ｐ１とＰｎの一次ＯＳＤノードに異なるｅｎｔｒｙが欠落していることが可能である。

【０２０９】

１０１２／１０１４　フェイルバックＯＳＤノードが、前のステップにおいて獲得されたｅｎｔｒｙによりデータ同期を繰り返し実行し、要求は、ｋｅｙ、ｏｆｆｓｅｔ、及びｌｅｎｇｔｈなどのＩＯキー情報を搬送する。

【０２１０】

特定の実施様態において、フェイルバックＯＳＤノードが、獲得されたｅｎｔｒｙにより、データ同期要求をバッチにおいて１つずつ、Ｐ１及びＰｎが位置付けられた一次ＯＳＤノードに送信し、要求は、ｋｅｙ、ｏｆｆｓｅｔ、及びｌｅｎｇｔｈなどのＩＯキー情報を搬送する。

【０２１１】

１０１６　対応するデータを送信する。

【０２１２】

特定の実施様態において、Ｐ１及びＰｎの一次ＯＳＤノードが、獲得されたデータ同期要求における情報により、各ｅｎｔｒｙに対応するデータをフェイルバックＯＳＤノードに送信する。

【０２１３】

シナリオ２　二次ＯＳＤの最大のＳｅｑＩＤが、一次ＯＳＤによって記録されたｅｎｔｒｙの範囲に入らず、一次ＯＳＤの最小のＳｅｑＩＤ未満である。

【０２１４】

１０１８　一次ＯＳＤの最小のＳｅｑＩＤを送信し、ｅｎｔｒｙをまったく送信しない。

【０２１５】

特定の実施様態において、Ｐｎの一次ＯＳＤノードが、フェイルバックＯＳＤノードの最大のＳｅｑＩＤが、一次ＯＳＤによって記録されたｅｎｔｒｙの範囲に入らない、すなわち、最大のＳｅｑＩＤが、Ｐｎの一次ＯＳＤの最小のＳｅｑＩＤ未満であると決定する事例において、Ｐｎの一次ＯＳＤの最小のＳｅｑＩＤがフェイルバックＯＳＤノードに送信され、ｅｎｔｒｙは、フェイルバックＯＳＤノードにまったく送信されず、このことが、フェイルバックＯＳＤノードが、Ｐｎの一次ＯＳＤノードが、インクリメンタルの同期

を用いてデータ回復が完了され得ないほど、データを書き込まないか、又は過度のデータを書き込むかを決定するのを助ける。

【０２１６】

１０２０　ｐａｒｔｉｔｉｏｎデータ同期を繰り返し要求し、要求が、ｐａｒｔｉｔｉｏｎグループＩＤを搬送する。

【０２１７】

フェイルバックＯＳＤノードが、Ｐｎの一次ＯＳＤノードにｐａｒｔｉｔｉｏｎ全体のデータの同期を要求し、例えば、この実施形態におけるＰｎの一次ＯＳＤ上の一次ｐａｒｔｉｔｉｏｎのデータが、フェイルバックＯＳＤノード上のＰｎの二次ｐａｒｔｉｔｉｏｎのデータと同期され、要求は、ｐａｒｔｉｔｉｏｎグループＩＤを搬送する。特定の実施様態において、ｐａｒｔｉｔｉｏｎのデータの量は、一般に、極めて大きく、したがって、データは、１つの要求を使用することによって完全に伝送されることは可能でなく、さらに、一次ノードは、フェイルバックノードのＩＯ能力を知らず、フェイルバックＯＳＤノードにデータを継続的に送信し、フェイルバックノードは、データを処理することに失敗する可能性があり、したがって、一次ノードは、フェイルバックノードがデータ同期を要求する場合に限り、フェイルバックノードにデータを送信する。フェイルバックＯＳＤノードは、状況によりＰｎの一次ＯＳＤノードに同期要求を繰り返し送信して、ｐａｒｔｉｔｉｏｎにおけるすべてのデータが同期されるまで、ｐａｒｔｉｔｉｏｎ全体におけるデータを同期する。実際の応用形態において、ｐａｒｔｉｔｉｏｎ全体が、本発明において限定されない、別の様態で同期されてもよい。

【０２１８】

１０２２　１つ又は複数のｋｅｙに対応するデータを送信する。

【０２１９】

Ｐｎの一次ＯＳＤノードが、各回にフェイルバックＯＳＤノードによって送信された同期要求により、その１つ又は複数のｋｅｙに対応するデータを送信する。

【０２２０】

ＯＳＤノードが障害の後にクラスタから出る手順

【０２２１】

ＯＳＤノードが、障害が事前設定された時間閾値（例えば、５分）を超えて続いた後、又はＯＳＤノード上でハードウェア障害が生じた後、正常に戻り、クラスタに再び参加することが依然としてできない場合、障害を有するＯＳＤノードは、クラスタから追放されて、データ信頼性を保証する必要がある。

【０２２２】

ＯＳＤノードによるクラスタから出ることは、ｐａｒｔｉｔｉｏｎ再分散及びデータ移行のプロセスであり、ｐａｒｔｉｔｉｏｎ再分散において、各ノードの平衡及びレプリカセキュリティが考慮される必要がある。データ移行プロセスにおけるＩＯ処理は、フェイルバック手順及びデータ回復手順における処理と整合性があり、データ移行が完了された後、一次レプリカと二次レプリカは整合性がある状態に到達し、ＭＤＣがビュー更新及び通知を実行するプロセスは、フェイルバックが完了された後に実行されるビュー更新処理と整合性がある。関連する各ＯＳＤノードが、更新されたビューによりＩＯ要求に対するレプリケーション処理又は転送処理を実行するプロセスは、フェイルバックが完了された後に最新のビューにより各ＯＳＤノードによって実行されるＩＯ処理とも整合性がある。

【０２２３】

理解を容易にするのに、以下に、説明のために特定の実施形態を使用する。図１１Ａ及び図１１Ｂは、本発明によるＯＳＤノードフェイルバック処理手順の実施形態を示す。この実施形態は、図２Ａないし図２Ｃ、及び図５において説明される実施形態において言及されるＭＤＣモジュール、ＩＯルーティングモジュール、及びＯＳＤノードによって実行される。理解を容易にするのに、この実施形態において、関連するＯＳＤノードは、ＯＳＤ１ノード、ＯＳＤ２ノード、及びＯＳＤｎノードであり、ＯＳＤ１ノードは、ｐａｒｔｉｔｉｏｎグループ１（略してＰ１）の一次ＯＳＤノード、及びｐａｒｔｉｔｉｏｎグル

ープ２（略してＰ２）の二次ＯＳＤノードであり、ＯＳＤ２ノードは、Ｐ１の二次ＯＳＤノードであり、ＯＳＤｎノードは、Ｐ２の一次ＯＳＤノードであるものと想定される。この実施形態における、ＯＳＤノードが障害を有するようになった後に出る処理手順は、以下のステップを特に含む。すなわち、

【０２２４】

１１００　ＯＳＤ１ノードが障害を有するようになる。

【０２２５】

１１０２　ＭＤＣモジュールが、ＯＳＤ１の障害が所定の閾値を超えて続くこと、又はＯＳＤ１上でハードウェア障害が生じたことを見出した場合、ＭＤＣが、ＯＳＤ１をクラスタから追放し、ビューを変更し、ＯＳＤ１ノード上のｐａｒｔｉｔｉｏｎ（この場合、ｐａｒｔｉｔｉｏｎは、Ｐ１の一次ｐａｒｔｉｔｉｏｎ、及びＰ２の二次ｐａｒｔｉｔｉｏｎである）を他のＯＳＤノード、例えば、この実施形態におけるＯＳＤ２ノード及びＯＳＤｎノードに移行させる。

【０２２６】

１１０４　ＭＤＣモジュールが、ＯＳＤ２ノードにビュー更新について通知し、ＯＳＤ２ノードがＰ１の一次ノードとして設定され、Ｐ２の二次ノードとなる。

【０２２７】

１１０６　ＭＤＣモジュールが、ＯＳＤｎノードにビュー更新について通知し、ＯＳＤｎノードは、Ｐ２の一次ノードであり、Ｐ１の二次ノードとなる。

【０２２８】

１１０８　ＯＳＤ２ノードが、Ｐ２のデータを同期するようＯＳＤｎノードに要求する。

【０２２９】

ＯＳＤｎノードは、Ｐ２の一次ＯＳＤノードであるため、ＯＳＤ２ノードが、ＯＳＤｎ上のＰ２の一次ｐａｒｔｉｔｉｏｎのデータを同期して、ＯＳＤ２上のＰ２の二次ｐａｒｔｉｔｉｏｎのデータがＯＳＤｎ上のＰ２の一次ｐａｒｔｉｔｉｏｎのデータと整合性があるようにするようＯＳＤｎノードに要求する。特定の同期手順は、図１０Ａ及び図１０Ｂに示される前述のデータ回復手順におけるｐａｒｔｉｔｉｏｎ全体のデータの同期手順と同様であり、詳細がここで再び説明されることはない。

【０２３０】

１１１０　ＯＳＤｎノードが、Ｐ１のデータを同期するようＯＳＤ２ノードに要求する。

【０２３１】

ＯＳＤ２ノードは、Ｐ１の元の二次ＯＳＤノードであるため、ＯＳＤ１ノードが、Ｐ１の新たな二次ＯＳＤノードとして障害を有するようになった後、ＯＳＤｎノードは、Ｐ１のデータをＯＳＤ２とだけしか同期することができず、ＯＳＤｎノードは、ＯＳＤ２上のＰ１の一次ｐａｒｔｉｔｉｏｎのデータを同期して、ＯＳＤｎ上のＰ１の二次ｐａｒｔｉｔｉｏｎのデータがＯＳＤ２上のＰ１の一次ｐａｒｔｉｔｉｏｎのデータと整合性があるようにするようＯＳＤ２ノードに要求する。

【０２３２】

１１１２　ｐａｒｔｉｔｉｏｎデータ同期を完了する。

【０２３３】

１１１４　ＯＳＤ２ノードが、ＭＤＣモジュールにＰ２のデータ移行が完了されたことを通知する。

【０２３４】

１１１６　ＯＳＤｎノードが、ＭＤＣモジュールにＰ１のデータ移行が完了されたことを通知する。

【０２３５】

１１１８　ＭＤＣモジュールが、対応する通知によりビュー更新を実行する。

【０２３６】

特定のビュー更新原理は、前述の手順において説明されるものと同一であり、更新プロセスは、前述の手順において説明されるものと同一であり、詳細がここで再び説明されることはない。

【０２３７】

１１２０　ビュー更新を通知し、Ｐ１の二次ＯＳＤｎノードのｐａｒｔｉｔｉｏｎステータスは、整合性がある。

【０２３８】

１１２２　ビュー更新を通知し、Ｐ２の二次ＯＳＤ２ノードのｐａｒｔｉｔｉｏｎステータスは、整合性がある。

【０２３９】

１１２４　ＯＳＤ２ノード及びＯＳＤｎノードが、最新のビューにより、ＩＯ要求に対応するデータのレプリケーションを処理する。

【０２４０】

新たなＯＳＤノードがクラスタに参加する手順

【０２４１】

新たなノードがクラスタに参加し、別のＯＳＤノード上に最初に分散されているｐａｒｔｉｔｉｏｎが、クラスタに参加する新たなＯＳＤノードに移行されて、データ分散平衡を保証する必要がある。新たなノードがクラスタに参加する手順は、データ移行、データ移行後のビュー更新、ビュー更新通知、及び更新されたビューによりＩＯ要求のレプリケーションに対して関連するＯＳＤノードによって実行される処理に主に関する。

【０２４２】

データ移行プロセスにおけるＩＯ処理は、フェイルバック手順及びデータ回復手順における処理と合致しており、データ移行が完了された後、ＭＤＣがビュー更新及びビュー通知を実行するプロセスは、フェイルバックが完了された後に実行されるビュー更新処理と整合性がある。関連する各ＯＳＤノードが、更新されたビューにより、ＩＯ要求に対応するデータのレプリケーションを処理するプロセスもまた、フェイルバックが完了された後に最新のビューにより各ＯＳＤノードによって実行されるＩＯ処理と整合性がある。

【０２４３】

特定の実施形態において、データ移行が完了された後、ビュー更新を実行することは、以下、すなわち、（１）クラスタに参加する新たなノードが、依然として、いくつかのｐａｒｔｉｔｉｏｎグループの二次ＯＳＤノードであること、ｐａｒｔｉｔｉｏｎステータスに整合性があること、及びｐａｒｔｉｔｉｏｎグループの元の二次ＯＳＤがもはや、ｐａｒｔｉｔｉｏｎの二次ノードではないこと、（２）クラスタに参加する新たなノードがいくつかのｐａｒｔｉｔｉｏｎグループの一次ＯＳＤノードとして設定されること、及びｐａｒｔｉｔｉｏｎグループがもはや、元の一次ＯＳＤノードに属さないこと（ｐａｒｔｉｔｉｏｎグループがもはや、元の一次ＯＳＤ上に分散されないこと）を含むことが可能である。

【０２４４】

理解を容易にするのに、以下に、説明のために特定の実施形態を使用する。この実施形態は、図２Ａないし図２Ｃ、及び図５において説明される実施形態において言及されるＭＤＣモジュール、ＩＯルーティングモジュール、及びＯＳＤノードによって実行される。理解を容易にするのに、この実施形態において、関連するＯＳＤノードは、ＯＳＤ１ノード、ＯＳＤｎノード、及びクラスタに参加する新たなＯＳＤノードであり、ＯＳＤ１は、ｐａｒｔｉｔｉｏｎグループＰ１（略してＰ１）の一次ＯＳＤノード、及びｐａｒｔｉｔｉｏｎグループＰｎ（略してＰｎ）の二次ＯＳＤノードであり、ＯＳＤｎは、Ｐｎの一次ＯＳＤノード、及びＰ１の二次ＯＳＤノードであり、クラスタに参加する新たなＯＳＤノードは、Ｐ１の二次ＯＳＤノード、及びＰｎの二次ＯＳＤノードであるものと想定される。図１２Ａ及び図１２Ｂに示されるとおり、この実施形態における、新たなＯＳＤノードがクラスタに参加する手順は、以下のステップを特に含む。すなわち、

【０２４５】

１２０２　新たなＯＳＤノードがクラスタに参加する構成コマンドをＭＤＣモジュールに供給する。

【０２４６】

特定の実施様態において、システム管理者が、構成コマンドを使用することによって、新たなＯＳＤノードがクラスタに参加することをＭＤＣモジュールに通知することが可能である。

【０２４７】

１２０４　ＭＤＣモジュールは、ビュー更新を実行し、いくつかのＯＳＤノード上のｐａｒｔｉｔｉｏｎを、クラスタに参加する新たなＯＳＤノードに移行させる。

【０２４８】

この実施形態において、ＭＤＣモジュールは、ＯＳＤ１ノード上のＰ１の二次ｐａｒｔｉｔｉｏｎ、及びＯＳＤｎノード上のＰｎの二次ｐａｒｔｉｔｉｏｎを、クラスタに参加する新たなＯＳＤノードに移行させて、クラスタに参加する新たなＯＳＤノードが、Ｐ１の新たな二次ＯＳＤノード、及びＰｎの新たな二次ＯＳＤノードの役割をするようにする。

【０２４９】

１２０６　ＯＳＤ１ノードにビュー更新について通知し、新たな二次ＯＳＤを追加する。

【０２５０】

特定の実施様態において、ＭＤＣモジュールは、ビュー更新についてＯＳＤ１ノードに通知し、クラスタに参加する新たなＯＳＤノードが、Ｐ１の二次ＯＳＤノードとしてＰ１の新たなｐａｒｔｉｔｉｏｎ　ｖｉｅｗに追加され、対応するｐａｒｔｉｔｉｏｎステータスは「不整合である」（クラスタに参加する新たなＯＳＤノードとＯＳＤ１ノードが、Ｐ１の同期されたデータをまだ有しないため）。

【０２５１】

１２０８　ＯＳＤｎノードにビュー更新について通知し、新たな二次ＯＳＤを追加する。

【０２５２】

特定の実施様態において、クラスタに参加する新たなＯＳＤノードは、Ｐｎの二次ＯＳＤノードとしてＰｎの新たなｐａｒｔｉｔｉｏｎ　ｖｉｅｗに追加され、対応するｐａｒｔｉｔｉｏｎステータスは、「不整合である」（クラスタに参加する新たなＯＳＤノードとＯＳＤｎノードが、Ｐｎの同期されたデータをまだ有しないため）。

【０２５３】

１２１０　クラスタに参加する新たなＯＳＤノードを開始する。

【０２５４】

特定の実施様態において、クラスタに参加する新たなＯＳＤノードが新たなＯＳＤノードとしてクラスタに参加した後、初期化プロセスが実行される。特定のプロセスは、図６におけるクラスタビューの初期化された生成及び獲得の前述の手順と同一であり、詳細がここで再び説明されることはない。

【０２５５】

１２１２　ＯＳＤノードのｐａｒｔｉｔｉｏｎ情報を返す。

【０２５６】

特定の実施様態において、ＭＤＣモジュールが、クラスタに参加する新たなＯＳＤノード上のｐａｒｔｉｔｉｏｎのビュー、すなわち、この実施形態においてＰ１のＩＯ　ｖｉｅｗ及びＰｎのＩＯ　ｖｉｅｗを、クラスタに参加する新たなＯＳＤノードに返す。

【０２５７】

１２１４　クラスタに参加する新たなＯＳＤノードが、ｐａｒｔｉｔｉｏｎデータを同期するよう一次ＯＳＤ１ノードに要求する。

【０２５８】

特定の実施様態において、クラスタに参加する新たなＯＳＤノードが、ＭＤＣによって返されるＰ１のＩＯ　ｖｉｅｗにより、Ｐ１のデータを同期して、すなわち、ＯＳＤ１ノード上のＰ１の一次ｐａｒｔｉｔｉｏｎのデータを同期して、クラスタに参加する新たなＯＳＤノード上のＰ１の二次ｐａｒｔｉｔｉｏｎのデータが、ＯＳＤ１ノード上のＰ１の一次ｐａｒｔｉｔｉｏｎのデータと整合性があるようにするようＰ１の一次ＯＳＤノード、すなわち、ＯＳＤ１に要求する。

【０２５９】

特定の同期手順は、図１０Ａ及び図１０Ｂに示される前述のデータ回復手順におけるｐａｒｔｉｔｉｏｎ全体のデータの同期手順と同様であり、詳細がここで再び説明されることはない。

【０２６０】

１２１６　クラスタに参加する新たなＯＳＤノードが、ｐａｒｔｉｔｉｏｎデータを同期するよう一次ＯＳＤｎノードに要求する。

【０２６１】

特定の実施様態において、クラスタに参加する新たなＯＳＤノードは、ＭＤＣによって返されるＰｎのＩＯビューにより、Ｐｎのデータを同期して、すなわち、ＯＳＤｎノード上のＰｎの一次ｐａｒｔｉｔｉｏｎのデータを同期して、クラスタに参加する新たなＯＳＤノード上のＰｎの二次ｐａｒｔｉｔｉｏｎのデータが、ＯＳＤｎノード上のＰｎの一次ｐａｒｔｉｔｉｏｎのデータと整合性があるようにするようＰｎの一次ＯＳＤノード、すなわち、ＯＳＤｎに要求する。

【０２６２】

特定の同期手順は、図１０Ａ及び図１０Ｂに示される前述のデータ回復手順におけるｐａｒｔｉｔｉｏｎ全体のデータの同期手順と同様であり、詳細がここで再び説明されることはない。

【０２６３】

１２１８　ｐａｒｔｉｔｉｏｎデータ同期を完了する。

【０２６４】

１２２０　二次ノードがｐａｒｔｉｔｉｏｎデータ移行を完了したことをＭＤＣに通知する。

【０２６５】

特定の実施様態において、ＯＳＤ１ノードが、クラスタに参加する新たなＯＳＤノードがＰ１のデータ同期を完了したことをＭＤＣモジュールに通知する。

【０２６６】

１２２２　二次ノードがｐａｒｔｉｔｉｏｎデータ移行を完了したことをＭＤＣに通知する。

【０２６７】

特定の実施様態において、ＯＳＤｎノードが、クラスタに参加する新たなＯＳＤノードがＰｎのデータ同期を完了したことをＭＤＣモジュールに通知する。

【０２６８】

１２２４　ＭＤＣモジュールが、ビュー更新を実行する。

【０２６９】

特定のビュー更新原理は、前述の手順において説明されるものと同一であり、更新プロセスは、前述の手順において説明されるものと同一であり、詳細がここで再び説明されることはない。

【０２７０】

１２２６～１２３０　ビュー更新についてＯＳＤ１ノード、ＯＳＤｎノード、クラスタに参加する新たなＯＳＤノードに別々に通知する。

【０２７１】

１２３２　ＯＳＤ１ノード、ＯＳＤｎノード、クラスタに参加する新たなＯＳＤノードが、更新されたビューにより、ＩＯ要求に対応するデータのレプリケーションを処理する

。

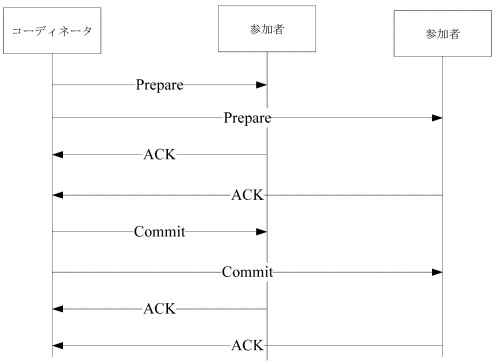
【０２７２】

実施形態の前述の説明に基づいて、当業者は、関係のあるハードウェアに指示するプログラムによって実施され得る実施形態における方法のステップのすべて、又はいくつかを理解することが可能である。そのプログラムは、コンピュータ可読記憶媒体に記憶されることが可能である。そのプログラムが実行されると、方法実施形態のステップが実行される。その記憶媒体は、例えば、ＲＯＭ／ＲＡＭ、磁気ディスク、又は光ディスクである。

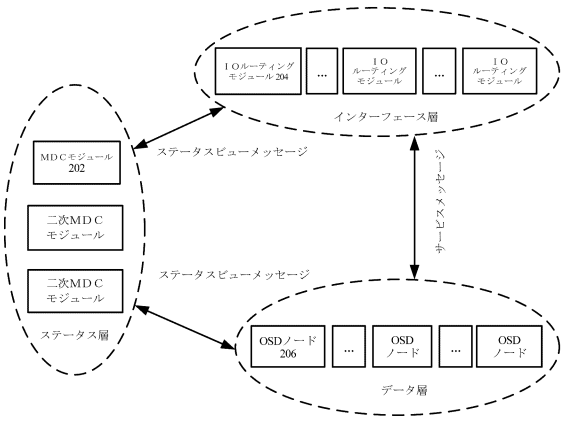
【０２７３】

前述の説明は、本発明の特定の実施形態に過ぎず、本発明の保護範囲を限定することは意図していない。本発明において開示される技術範囲内で当業者によって容易に考案される変形及び置換は、本発明の保護範囲に入るものとする。したがって、本発明の保護範囲は、特許請求の範囲の保護範囲の対象とされるべきものとする。

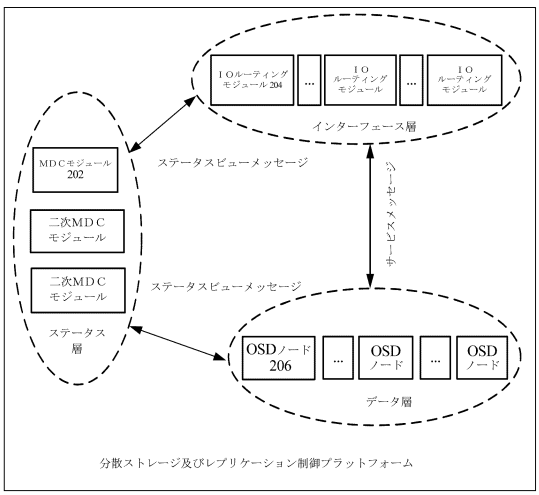
【図１】



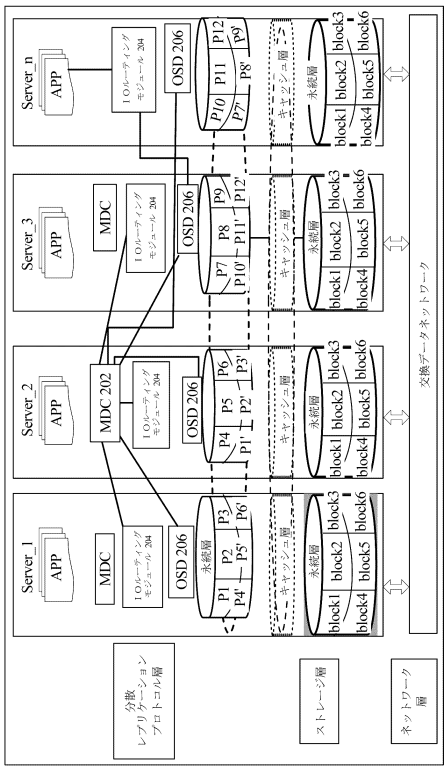
【図２Ａ】



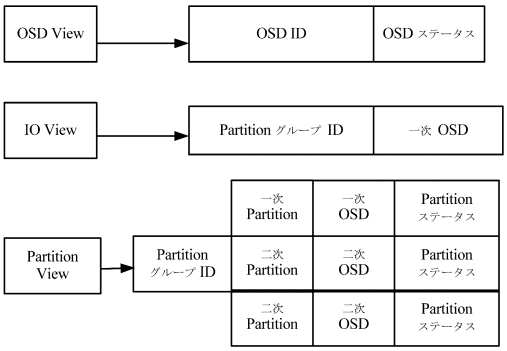
【図２Ｂ】



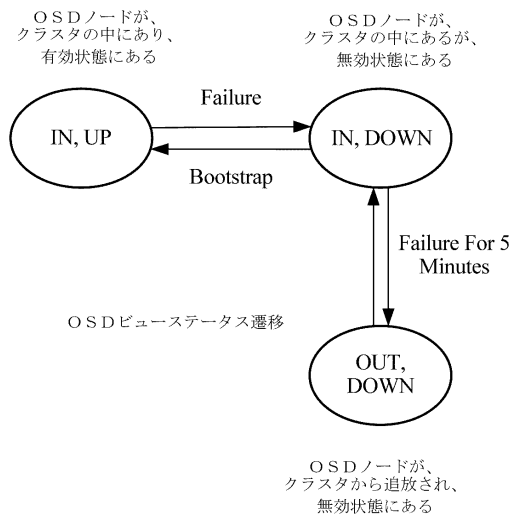
【図２Ｃ】



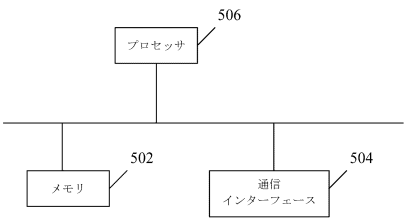
【図３】



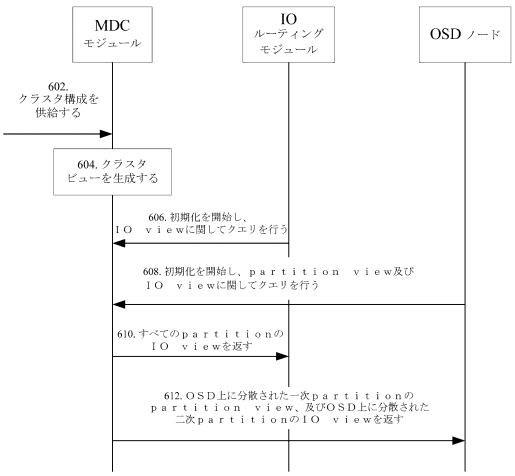
【図４】



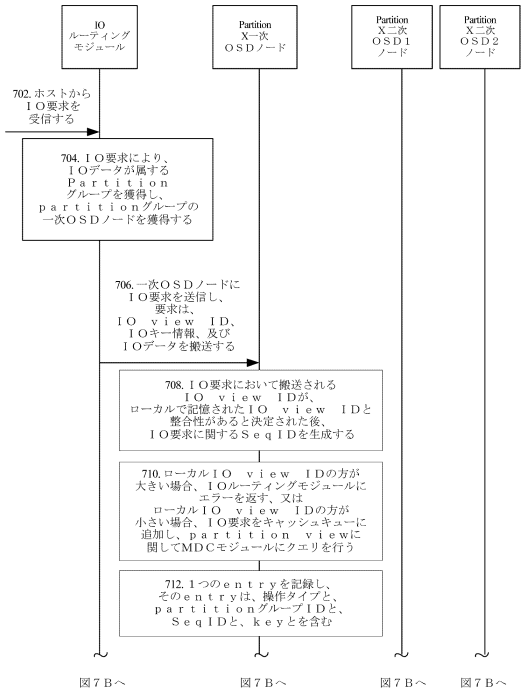
【図５】



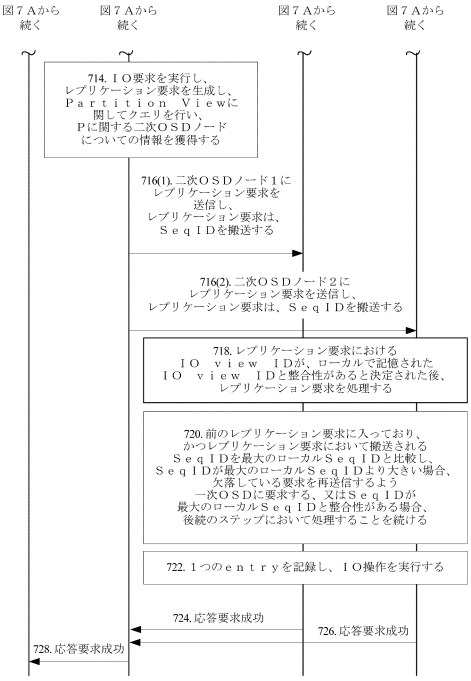
【図６】



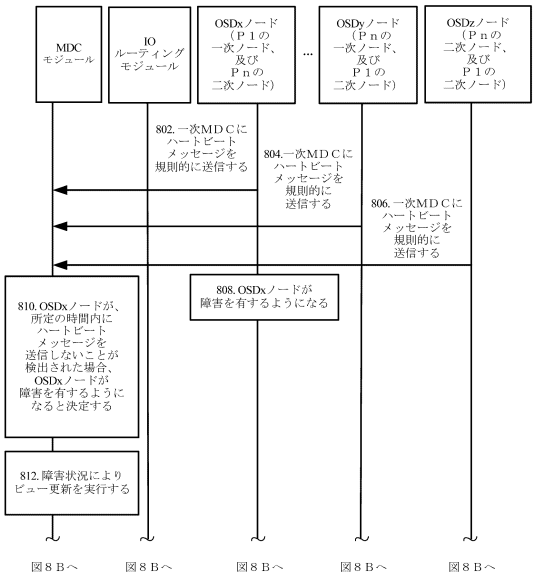
【図７Ａ】



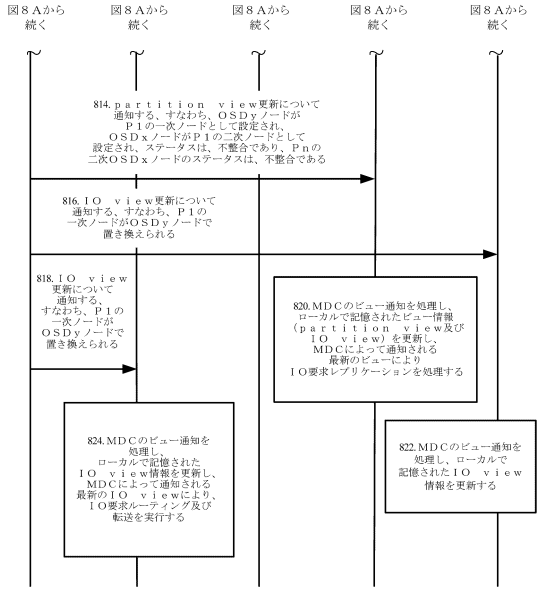
【図７Ｂ】



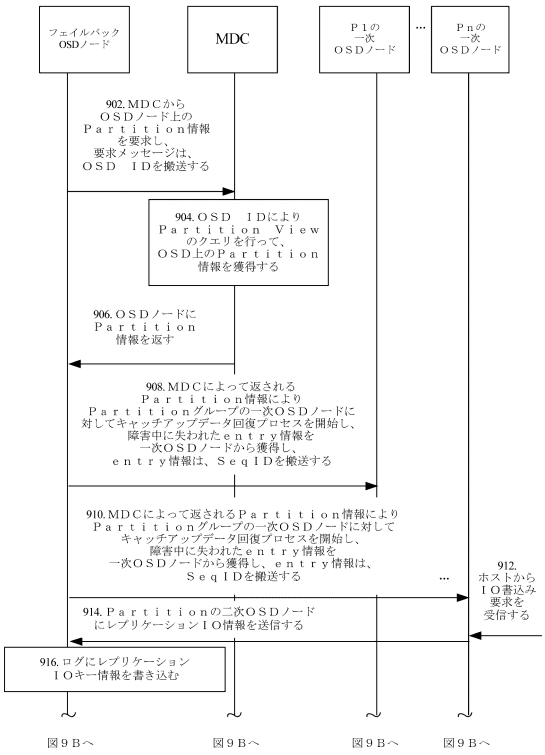
【図８Ａ】



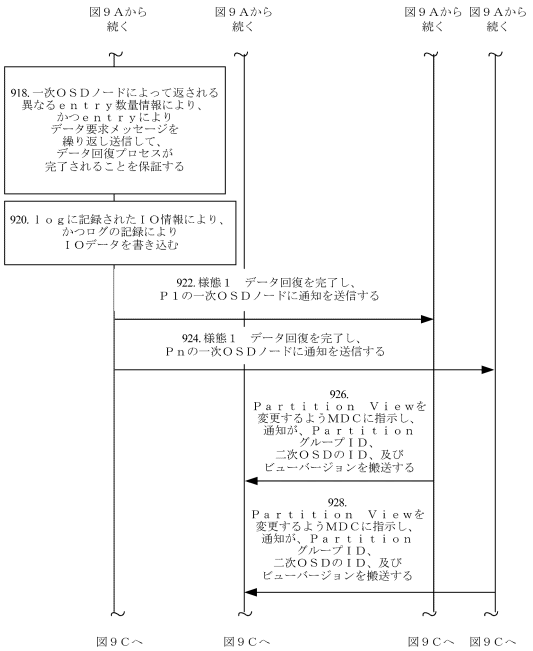
【図８Ｂ】



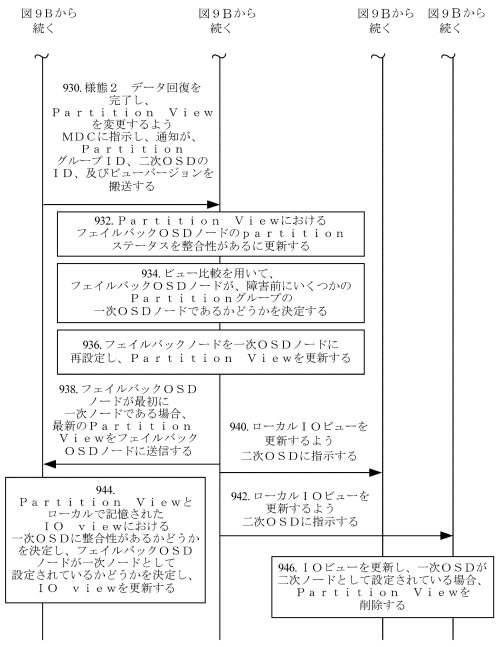
【図９Ａ】



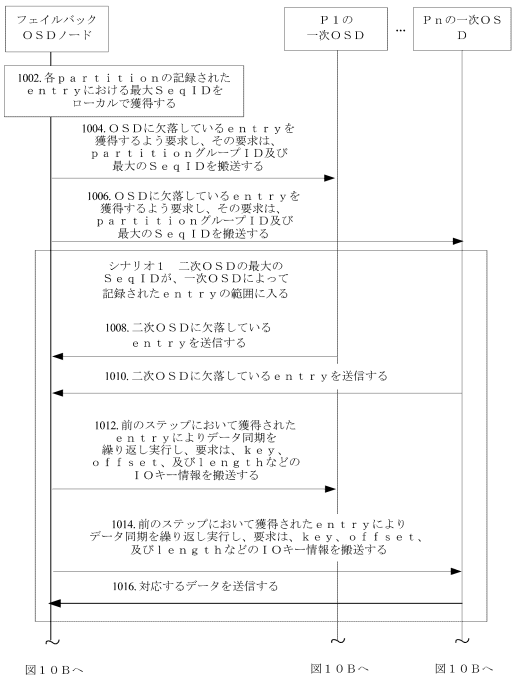
【図９Ｂ】



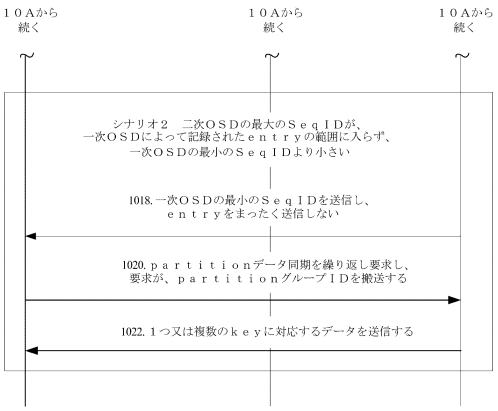
【図９Ｃ】



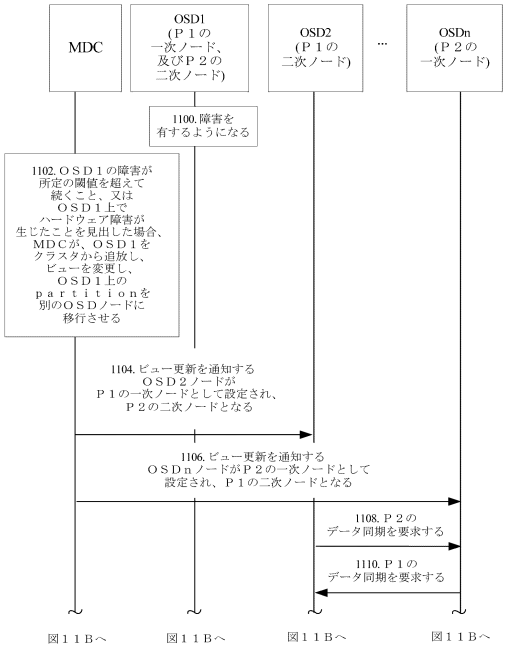
【図１０Ａ】



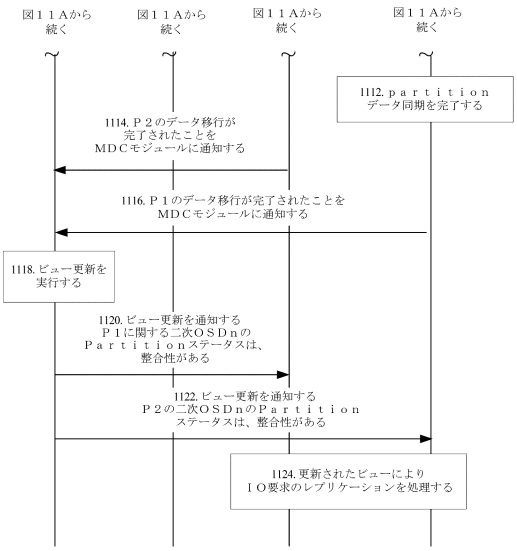
【図１０Ｂ】



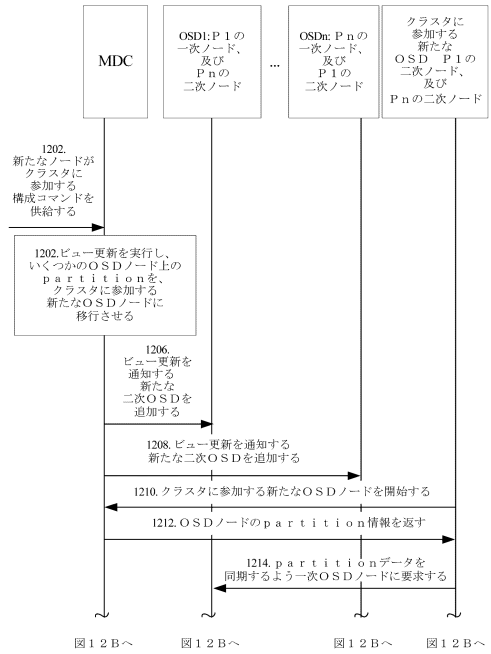
【図１１Ａ】



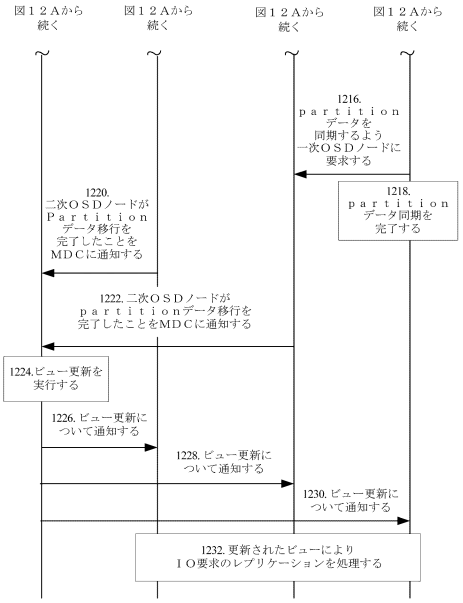
【図１１Ｂ】



【図１２Ａ】



【図１２Ｂ】


図１２Ａから
続く
図１２Ａから
続く
図１２Ａから
続く
図１２Ａから
続く
1216.
ｐａｒｔｉｔｉｏｎ
データを
同期するよう
一次ＯＳＤノードに
要求する
1218.
ｐａｒｔｉｔｉｏｎ
データ同期を
完了する
1220.
二次ＯＳＤノードが
Ｐａｒｔｉｔｉｏｎ
データ移行を
完了したことを
ＭＤＣに通知する
1222. 二次ＯＳＤノードが
ｐａｒｔｉｔｉｏｎデータ移行を
完了したことをＭＤＣに通知する
1224.ビュー更新を
実行する
1226. ビュー更新に
ついて通知する
1228. ビュー更新に
ついて通知する
1230. ビュー更新に
ついて通知する
1232. 更新されたビューにより
ＩＯ要求のレプリケーションを処理する

フロントページの続き

(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５１８１２９▲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５１８１２９▲広▼▲東▼省深▲チェン▼市▲龍▼▲崗▼区坂田▲華▼▲為▼▲総▼部▲辦▼公楼
(72)発明者　▲劉▼　叙友
中国５１８１２９▲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　特開２０１２－２２１４１９（ＪＰ，Ａ）
米国特許出願公開第２０１３／０２９０２４９（ＵＳ，Ａ１）
特開２００６－２０９７７５（ＪＰ，Ａ）
特開２００６－０６５６２９（ＪＰ，Ａ）
特表２０１０－５０９６８６（ＪＰ，Ａ）
米国特許出願公開第２０１１／００９９４２０（ＵＳ，Ａ１）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１２／００
Ｇ０６Ｆ　３／０６