

(19) 日本国特許庁 (JP)

(12) 特 許 公 報 (B2)

(11) 特許番号

特許第4489455号
(P4489455)

(45) 発行日 平成22年6月23日 (2010. 6. 23)

(24) 登録日 平成22年4月9日 (2010. 4. 9)

(51) Int. Cl.

F I

G O 6 F 12/00 (2006. 01)

G O 6 F 12/00 5 3 1 D

G O 6 F 3/06 (2006. 01)

G O 6 F 3/06 3 0 4 F

請求項の数 15 (全 34 頁)

(21) 出願番号 特願2004-38169 (P2004-38169)
 (22) 出願日 平成16年2月16日 (2004. 2. 16)
 (65) 公開番号 特開2005-228217 (P2005-228217A)
 (43) 公開日 平成17年8月25日 (2005. 8. 25)
 審査請求日 平成18年12月18日 (2006. 12. 18)

(73) 特許権者 000005108
 株式会社日立製作所
 東京都千代田区丸の内一丁目6番6号
 (74) 代理人 110000176
 一色国際特許業務法人
 (72) 発明者 室谷 暁
 神奈川県小田原市中里322番2号 株式
 会社日立製作所 R A I Dシステム事業部
 内
 (72) 発明者 石川 篤
 神奈川県小田原市中里322番2号 株式
 会社日立製作所 R A I Dシステム事業部
 内

最終頁に続く

(54) 【発明の名称】 ディスク制御装置及びディスク制御装置の制御方法

(57) 【特許請求の範囲】

【請求項 1】

情報処理装置と通信可能に接続され、

一又は複数の論理ボリュームが形成されている第一の記憶デバイスと、

一又は複数の論理ボリュームが形成されている第二の記憶デバイスと、

第三の記憶デバイスと

に対するデータの書き込みまたは読み出しを行うディスク制御装置であって、

前記第一の記憶デバイスの前記論理ボリュームの識別子である主論理ボリュームの識別子と前記第二の記憶デバイスの前記論理ボリュームの識別子である副論理ボリュームの識別子とが対応付けられたペア管理テーブルと、任意の値に設定された検出時間と、を記憶するメモリと、

計時機構と、

前記情報処理装置から前記第一の記憶デバイスの前記論理ボリュームに対するデータの書き込み要求と前記データとを受信する書き込み要求受信部と、

前記書き込み要求を受信すると前記データを前記第一の記憶デバイスの前記論理ボリュームに書き込む第一の書き込み部と、

前記データが書き込まれている前記第一の記憶デバイスの前記論理ボリュームの識別子、当該論理ボリューム内の前記データが格納されている位置情報、前記計時機構より取得される現在時刻である更新時刻、及び前記データで構成されるジャーナルデータを前記第三の記憶デバイスに書き込むジャーナル書き込み部と、

10

20

前記第三の記憶デバイスに記憶されている複数の前記ジャーナルデータのそれぞれの前記更新時刻と前記メモリに記憶されている前記検出時間とを参照し、前記計時機構より取得される現在時刻と前記更新時刻との差が前記検出時間を超えている前記ジャーナルデータを前記複数のジャーナルデータの中から選択し、選択された前記ジャーナルデータの前記更新時刻が早い順に前記ジャーナルデータの前記論理ボリュームの識別子と前記位置情報と前記データとを参照し、当該論理ボリュームの識別子が前記主論理ボリュームの識別子である前記副論理ボリュームの識別子を前記ペア管理テーブルから取得し、当該データを前記第二の記憶デバイスの当該副論理ボリュームの識別子で示される前記論理ボリュームの当該位置情報で示される場所に書き込む第二の書き込み部と、

前記情報処理装置から前記第二の記憶デバイスへの前記データの書き込みを中断するスプリット指示命令を受信するスプリット指示命令受信部と、

前記情報処理装置から前記第二の記憶デバイスへの前記データの書き込みを再開するスプリット解除命令を受信するスプリット解除命令受信部と、

前記スプリット指示命令を受信すると前記スプリット指示命令を受信したことを示すデータと前記計時機構より取得される現在時刻であるスプリット時刻とで構成されるスプリット指示データを前記第三の記憶デバイスに書き込むスプリット指示記憶部と、

前記スプリット解除命令を受信すると前記スプリット解除命令を受信したことを示すデータであるスプリット解除データを前記第三の記憶デバイスに書き込むスプリット解除記憶部と

を有し、

前記第二の書き込み部は前記第三の記憶デバイスに前記スプリット指示データが記憶され、かつ、前記スプリット解除データが記憶されていない場合は、前記スプリット指示データの前記スプリット時刻を参照し、前記更新時刻が当該スプリット時刻より遅い前記ジャーナルデータの前記データを前記第二の記憶デバイスに書き込まない

ことを特徴とするディスク制御装置。

【請求項 2】

請求項 1 に記載のディスク制御装置において、

前記第一の記憶デバイスを構成する一又は複数のハードディスクドライブと、前記第二の記憶デバイス及び前記第三の記憶デバイスを構成する一又は複数のハードディスクドライブとが物理的に別であることを特徴とするディスク制御装置。

【請求項 3】

請求項 1 に記載のディスク制御装置において、

前記第一の記憶デバイスを構成する一又は複数のハードディスクドライブと、前記第二の記憶デバイスを構成する一又は複数のハードディスクドライブとが物理的に別であることを特徴とするディスク制御装置。

【請求項 4】

請求項 1 に記載のディスク制御装置において、

前記第二の記憶デバイスの前記論理ボリュームに対応する仮想的な論理ボリュームである仮想論理ボリュームが設けられ、前記論理ボリュームの識別子と前記仮想論理ボリュームの識別子との対応付けが前記メモリに仮想論理ボリューム管理テーブルとして記憶され、

前記第三の記憶デバイスを参照し、前記スプリット指示データが記憶され、かつ、前記スプリット解除データが記憶されていない場合は、前記更新時刻が前記スプリット指示データの前記スプリット時刻より早い前記ジャーナルデータの前記論理ボリュームの識別子と前記位置情報とで構成される未反映情報を前記メモリに記憶する未反映情報記憶部と、

前記情報処理装置から前記仮想論理ボリュームの識別子と位置情報とが設定されているデータの読み出し要求を受信する読み出し要求受信部と、

前記読み出し要求の前記仮想論理ボリュームの識別子に対応する前記第二の記憶デバイスの前記論理ボリュームの識別子を前記仮想論理ボリューム管理テーブルから取得する識別子取得部と、

10

20

30

40

50

前記メモリに記憶されている前記未反映情報を参照し、前記識別子取得部により取得された前記第二の記憶デバイスの前記論理ボリュームの識別子と前記未反映情報の前記論理ボリュームの識別子とが同じ前記未反映情報において、前記読み出し要求に設定されている前記位置情報が示す範囲と前記未反映情報の前記位置情報が示す範囲とを比較して重複する範囲を取得する重複範囲取得部と、

前記読み出し要求に設定されている前記位置情報が示す範囲のうち前記重複する範囲については前記第三の記憶デバイスに記憶されている前記ジャーナルデータの前記データを読み出し、前記読み出し要求に設定されている前記位置情報が示す範囲のうち前記重複する範囲以外の範囲については前記識別子取得部により取得された前記第二の記憶デバイスの前記論理ボリュームの識別子で示される前記論理ボリュームに記憶されている前記データを読み出す仮想論理ボリューム読み出し部と、

10

前記仮想論理ボリューム読み出し部により読み出された前記データを前記情報処理装置に送信する読み出しデータ送信部と
を有することを特徴とするディスク制御装置。

【請求項 5】

請求項 1 に記載のディスク制御装置において、

前記第二の書き込み部は前記第三の記憶デバイスに前記スプリット指示データが記憶され、かつ、前記スプリット解除データが記憶されていない場合は、前記更新時刻が当該スプリット時刻より早い前記ジャーナルデータの前記データを、前記更新時刻と前記計時機構より取得される現在時刻との差が前記検出時間を超えているかどうかにかかわらず、前記第二の記憶デバイスに書き込むこと
を特徴とするディスク制御装置。

20

【請求項 6】

請求項 1 に記載のディスク制御装置において、

前記第一の記憶デバイスの前記論理ボリュームにグループ ID が付与され、前記論理ボリュームの識別子と前記グループ ID との対応付けが前記メモリにグループ ID 管理テーブルとして記憶され、

前記ジャーナル書き込み部は、前記第一の記憶デバイスの前記データが書き込まれている前記論理ボリュームの識別子に対応する前記グループ ID を前記グループ ID 管理テーブルから取得し、当該グループ ID を前記ジャーナルデータに設定して前記第三の記憶デバイスに書き込み、

30

前記情報処理装置から前記第二の記憶デバイスへの前記データの書き込みを中断するスプリット指示命令を受信するスプリット指示命令受信部と、

前記情報処理装置から前記第二の記憶デバイスへの前記データの書き込みを再開するスプリット解除命令を受信するスプリット解除命令受信部と、

前記スプリット指示命令を受信すると前記スプリット指示命令に設定されている前記グループ ID と前記スプリット指示命令を受信したことを示すデータと前記計時機構より取得される現在時刻であるスプリット時刻とで構成されるスプリット指示データを前記第三の記憶デバイスに書き込むスプリット指示記憶部と、

前記スプリット解除命令を受信すると前記スプリット解除命令に設定されている前記グループ ID と前記スプリット解除命令を受信したことを示すデータであるスプリット解除データを前記第三の記憶デバイスに書き込むスプリット解除記憶部と

40

を有し、

前記第二の書き込み部は、前記第三の記憶デバイスに前記スプリット指示データが記憶され、かつ、当該スプリット指示データの前記グループ ID と同じ前記グループ ID が設定されている前記スプリット解除データが記憶されていない場合は、前記スプリット指示データの前記グループ ID と前記スプリット時刻とを参照し、前記ジャーナルデータの前記グループ ID が前記スプリット指示データの当該グループ ID と同じで、かつ、前記更新時刻が当該スプリット時刻より遅い前記ジャーナルデータの前記データを前記第二の記憶デバイスに書き込まないこと

50

を特徴とするディスク制御装置。

【請求項 7】

請求項 6 に記載のディスク制御装置において、

前記第二の記憶デバイスの前記論理ボリュームに対応する仮想的な論理ボリュームである仮想論理ボリュームが設けられ、前記論理ボリュームの識別子と前記仮想論理ボリュームの識別子との対応付けが前記メモリに仮想論理ボリューム管理テーブルとして記憶され、

前記第三の記憶デバイスを参照し、前記第三の記憶デバイスに前記スプリット指示データが記憶され、かつ、当該スプリット指示データの前記グループ ID と同じ前記グループ ID が設定されている前記スプリット解除データが記憶されていない場合は、当該スプリット指示データの前記グループ ID と同じ前記グループ ID が設定され、かつ、前記更新時刻が前記スプリット指示データの前記スプリット時刻より早い前記ジャーナルデータの前記論理ボリュームの識別子と前記位置情報とで構成される未反映情報を前記メモリに記憶する未反映情報記憶部と、

前記情報処理装置から前記仮想論理ボリュームの識別子と位置情報とが設定されているデータの読み出し要求を受信する読み出し要求受信部と、

前記読み出し要求の前記仮想論理ボリュームの識別子に対応する前記第二の記憶デバイスの前記論理ボリュームの識別子を前記仮想論理ボリューム管理テーブルから取得する識別子取得部と、

前記メモリに記憶されている前記未反映情報を参照し、前記識別子取得部により取得された前記第二の記憶デバイスの前記論理ボリュームの識別子と前記未反映情報の前記論理ボリュームの識別子とが同じ前記未反映情報において、前記読み出し要求に設定されている前記位置情報が示す範囲と前記未反映情報の前記位置情報が示す範囲とを比較して重複する範囲を取得する重複範囲取得部と、

前記読み出し要求に設定されている前記位置情報が示す範囲のうち前記重複する範囲については前記第三の記憶デバイスに記憶されている前記ジャーナルデータに設定されている前記データを読み出し、前記読み出し要求に設定されている前記位置情報が示す範囲のうち前記重複する範囲以外の範囲については前記識別子取得部により取得された前記第二の記憶デバイスの前記論理ボリュームの識別子で示される前記論理ボリュームに記憶されている前記データを読み出す仮想論理ボリューム読み出し部と、

前記仮想論理ボリューム読み出し部により読み出された前記データを前記情報処理装置に送信する読み出しデータ送信部と

を有することを特徴とするディスク制御装置。

【請求項 8】

請求項 6 に記載のディスク制御装置において、

前記第二の書き込み部は、前記第三の記憶デバイスに前記スプリット指示データが記憶され、かつ、当該スプリット指示データの前記グループ ID と同じ前記グループ ID が設定されている前記スプリット解除データが記憶されていない場合は、前記ジャーナルデータの前記グループ ID が前記スプリット指示データの当該グループ ID と同じで、かつ、前記更新時刻が当該スプリット時刻より早い前記ジャーナルデータの前記データを、前記更新時刻と前記計時機構より取得される現在時刻との差が前記検出時間を超えているかどうかにかかわらず、前記第二の記憶デバイスに書き込むことを特徴とするディスク制御装置。

【請求項 9】

情報処理装置と通信可能に接続され、

—又は複数の論理ボリュームが形成されている第一の記憶デバイスと、

—又は複数の論理ボリュームが形成されている第二の記憶デバイスと、

第三の記憶デバイスと、

前記第一の記憶デバイスの前記論理ボリュームの識別子である主論理ボリュームの識別子と前記第二の記憶デバイスの前記論理ボリュームの識別子である副論理ボリュームの識

10

20

30

40

50

別子とが対応付けられたペア管理テーブルと、任意の値に設定された検出時間と、を記憶するメモリと、

計時機構と

を有し、

前記第1～第3の記憶デバイスに対するデータの書き込みまたは読み出しを行うディスク制御装置の制御方法であって、

前記情報処理装置から前記第一の記憶デバイスの前記論理ボリュームに対するデータの書き込み要求と前記データとを受信するステップと、

前記書き込み要求を受信すると前記データを前記第一の記憶デバイスの前記論理ボリュームに書き込むステップと、

前記データが書き込まれている前記第一の記憶デバイスの前記論理ボリュームの識別子、当該論理ボリューム内の前記データが格納されている位置情報、前記計時機構より取得される現在時刻である更新時刻、及び前記データで構成されるジャーナルデータを前記第三の記憶デバイスに書き込むステップと、

前記第三の記憶デバイスに記憶されている複数の前記ジャーナルデータのそれぞれの前記更新時刻と前記メモリに記憶されている前記検出時間とを参照し、前記計時機構より取得される現在時刻と前記更新時刻との差が前記検出時間を超えている所定の時間以上である前記ジャーナルデータを前記複数のジャーナルデータの中から選択し、選択された前記ジャーナルデータの前記更新時刻が早い順に前記ジャーナルデータの前記論理ボリュームの識別子と前記位置情報と前記データとを参照し、当該論理ボリュームの識別子が前記主論理ボリュームの識別子である前記副論理ボリュームの識別子を前記ペア管理テーブルから取得し、当該データを前記第二の記憶デバイスの当該副論理ボリュームの識別子で示される前記論理ボリュームの当該位置情報で示される場所に書き込むステップと、

前記情報処理装置から前記第二の記憶デバイスへの前記データの書き込みを中断するスプリット指示命令を受信するステップと、

前記情報処理装置から前記第二の記憶デバイスへの前記データの書き込みを再開するスプリット解除命令を受信するステップと、

前記スプリット指示命令を受信すると前記スプリット指示命令を受信したことを示すデータと前記計時機構より取得される現在時刻であるスプリット時刻とで構成されるスプリット指示データを前記第三の記憶デバイスに書き込むステップと、

前記スプリット解除命令を受信すると前記スプリット解除命令を受信したことを示すデータであるスプリット解除データを前記第三の記憶デバイスに書き込むステップと

を有し、

前記第三の記憶デバイスに記憶されている更新データを前記第二の記憶デバイスの前記論理ボリュームに書き込む前記ステップは、前記第三の記憶デバイスに前記スプリット指示データが記憶され、かつ、前記スプリット解除データが記憶されていない場合は、前記スプリット指示データの前記スプリット時刻を参照し、前記更新時刻が当該スプリット時刻より遅い前記ジャーナルデータの前記データを前記第二の記憶デバイスに書き込まないステップである

ことを特徴とするディスク制御装置の制御方法。

【請求項10】

請求項9に記載のディスク制御装置の制御方法において、

前記第二の記憶デバイスの前記論理ボリュームに対応する仮想的な論理ボリュームである仮想論理ボリュームが設けられ、前記論理ボリュームの識別子と前記仮想論理ボリュームの識別子との対応付けが前記メモリに仮想論理ボリューム管理テーブルとして記憶され、

前記第三の記憶デバイスを参照し、前記スプリット指示データが記憶され、かつ、前記スプリット解除データが記憶されていない場合は、前記更新時刻が前記スプリット指示データの前記スプリット時刻より早い前記ジャーナルデータの前記論理ボリュームの識別子と前記位置情報とで構成される未反映情報を前記メモリに記憶するステップと、

前記情報処理装置から前記仮想論理ボリュームの識別子と位置情報とが設定されているデータの読み出し要求を受信するステップと、

前記読み出し要求の前記仮想論理ボリュームの識別子に対応する前記第二の記憶デバイスの前記論理ボリュームの識別子を前記仮想論理ボリューム管理テーブルから取得するステップと、

前記メモリに記憶されている前記未反映情報を参照し、前記識別子取得部により取得された前記第二の記憶デバイスの前記論理ボリュームの識別子と前記未反映情報の前記論理ボリュームの識別子とが同じ前記未反映情報において、前記読み出し要求に設定されている前記位置情報が示す範囲と前記未反映情報の前記位置情報が示す範囲とを比較して重複する範囲を取得するステップと、

10

前記読み出し要求に設定されている前記位置情報が示す範囲のうち前記重複する範囲については前記第三の記憶デバイスに記憶されている前記ジャーナルデータの前記データを読み出し、前記読み出し要求に設定されている前記位置情報が示す範囲のうち前記重複する範囲以外の範囲については前記識別子取得部により取得された前記第二の記憶デバイスの前記論理ボリュームの識別子で示される前記論理ボリュームに記憶されている前記データを読み出すステップと、

前記仮想論理ボリューム読み出し部により読み出された前記データを前記情報処理装置に送信するステップと

を有することを特徴とするディスク制御装置の制御方法。

【請求項 11】

20

請求項9に記載のディスク制御装置の制御方法において、

前記第三の記憶デバイスに記憶されている更新データを前記第二の記憶デバイスの前記論理ボリュームに書き込む前記ステップは、前記第三の記憶デバイスに前記スプリット指示データが記憶され、かつ、前記スプリット解除データが記憶されていない場合は、前記更新時刻が当該スプリット時刻より早い前記ジャーナルデータの前記データを、前記更新時刻と前記計時機構より取得される現在時刻との差が前記検出時間を超えているかどうかにかかわらず、前記第二の記憶デバイスに書き込むステップであること
を特徴とするディスク制御装置の制御方法。

【請求項 12】

請求項9に記載のディスク制御装置の制御方法において、

30

前記第一の記憶デバイスの前記論理ボリュームにグループIDが付与され、前記論理ボリュームの識別子と前記グループIDとの対応付けが前記メモリにグループID管理テーブルとして記憶され、

前記ジャーナルデータを前記第三の記憶デバイスに書き込む前記ステップは、前記第一の記憶デバイスの前記データが書き込まれている前記論理ボリュームの識別子に対応する前記グループIDを前記グループID管理テーブルから取得し、当該グループIDを前記ジャーナルデータに設定して前記第三の記憶デバイスに書き込むステップであり、

前記情報処理装置から前記第二の記憶デバイスへの前記データの書き込みを中断するスプリット指示命令を受信するステップと、

前記情報処理装置から前記第二の記憶デバイスへの前記データの書き込みを再開するスプリット解除命令を受信するステップと、

40

前記スプリット指示命令を受信すると前記スプリット指示命令に設定されている前記グループIDと前記スプリット指示命令を受信したことを示すデータと前記計時機構より取得される現在時刻であるスプリット時刻とで構成されるスプリット指示データを前記第三の記憶デバイスに書き込むステップと、

前記スプリット解除命令を受信すると前記スプリット解除命令に設定されている前記グループIDと前記スプリット解除命令を受信したことを示すデータであるスプリット解除データを前記第三の記憶デバイスに書き込むステップと

を有し、

前記第三の記憶デバイスに記憶されている更新データを前記第二の記憶デバイスの前記

50

論理ボリュームに書き込む前記ステップは、前記第三の記憶デバイスに前記スプリット指示データが記憶され、かつ、当該スプリット指示データの前記グループIDと同じ前記グループIDが設定されている前記スプリット解除データが記憶されていない場合は、前記スプリット指示データの前記グループIDと前記スプリット時刻とを参照し、前記ジャーナルデータの前記グループIDが前記スプリット指示データの当該グループIDと同じで、かつ、前記更新時刻が当該スプリット時刻より遅い前記ジャーナルデータの前記データを前記第二の記憶デバイスに書き込まないステップであること
を特徴とするディスク制御装置の制御方法。

【請求項13】

請求項12に記載のディスク制御装置の制御方法において、

10

前記第二の記憶デバイスの前記論理ボリュームに対応する仮想的な論理ボリュームである仮想論理ボリュームが設けられ、前記論理ボリュームの識別子と前記仮想論理ボリュームの識別子との対応付けが前記メモリに仮想論理ボリューム管理テーブルとして記憶され、

前記第三の記憶デバイスを参照し、前記第三の記憶デバイスに前記スプリット指示データが記憶され、かつ、当該スプリット指示データの前記グループIDと同じ前記グループIDが設定されている前記スプリット解除データが記憶されていない場合は、当該スプリット指示データの前記グループIDと同じ前記グループIDが設定され、かつ、前記更新時刻が前記スプリット指示データの前記スプリット時刻より早い前記ジャーナルデータの前記論理ボリュームの識別子と前記位置情報とで構成される未反映情報を前記メモリに記憶するステップと、

20

前記情報処理装置から前記仮想論理ボリュームの識別子と位置情報とが設定されているデータの読み出し要求を受信するステップと、

前記読み出し要求の前記仮想論理ボリュームの識別子に対応する前記第二の記憶デバイスの前記論理ボリュームの識別子を前記仮想論理ボリューム管理テーブルから取得するステップと、

前記メモリに記憶されている前記未反映情報を参照し、前記識別子取得部により取得された前記第二の記憶デバイスの前記論理ボリュームの識別子と前記未反映情報の前記論理ボリュームの識別子とが同じ前記未反映情報において、前記読み出し要求に設定されている前記位置情報が示す範囲と前記未反映情報の前記位置情報が示す範囲とを比較して重複する範囲を取得するステップと、

30

前記読み出し要求に設定されている前記位置情報が示す範囲のうち前記重複する範囲については前記第三の記憶デバイスに記憶されている前記ジャーナルデータに設定されている前記データを読み出し、前記読み出し要求に設定されている前記位置情報が示す範囲のうち前記重複する範囲以外の範囲については前記識別子取得部により取得された前記第二の記憶デバイスの前記論理ボリュームの識別子で示される前記論理ボリュームに記憶されている前記データを読み出すステップと、

前記仮想論理ボリューム読み出し部により読み出された前記データを前記情報処理装置に送信するステップと

を有することを特徴とするディスク制御装置の制御方法。

40

【請求項14】

請求項12に記載のディスク制御装置の制御方法において、

前記第三の記憶デバイスに記憶されている更新データを前記第二の記憶デバイスの前記論理ボリュームに書き込む前記ステップは、前記第三の記憶デバイスに前記スプリット指示データが記憶され、かつ、当該スプリット指示データの前記グループIDと同じ前記グループIDが設定されている前記スプリット解除データが記憶されていない場合は、前記ジャーナルデータの前記グループIDが前記スプリット指示データの当該グループIDと同じで、かつ、前記更新時刻が当該スプリット時刻より早い前記ジャーナルデータの前記データを、前記更新時刻と前記計時機構より取得される現在時刻との差が前記検出時間を超えているかどうかにかかわらず、前記第二の記憶デバイスに書き込むステップであるこ

50

と

を特徴とするディスク制御装置の制御方法。

【請求項 15】

情報処理装置と通信可能に接続され、

一又は複数のハードディスクドライブにより構成される第一の記憶デバイスと、

一又は複数のハードディスクドライブにより構成される第二の記憶デバイスと、

一又は複数のハードディスクドライブにより構成される第三の記憶デバイスと

に対するデータの書き込みまたは読み出しを行うディスク制御装置であって、

前記第一の記憶デバイスに一又は複数の論理ボリュームが形成され、

前記第二の記憶デバイスに一又は複数の論理ボリュームが形成され、

10

前記情報処理装置から前記論理ボリュームに対するデータの書き込み要求を受信するチャンネル制御部と、

前記チャンネル制御部により受信された前記データを前記論理ボリュームに書き込むディスク制御部と、

前記第一の記憶デバイスの前記論理ボリュームの識別子である主論理ボリュームの識別子と前記第二の記憶デバイスの前記論理ボリュームの識別子である副論理ボリュームの識別子とが対応付けられたペア管理テーブルと、任意の値に設定された検出時間と、を記憶する共有メモリと、

計時機構と

前記チャンネル制御部と前記ディスク制御部と前記共有メモリとを通信可能に接続するスイッチング制御部と

20

を有し、

前記ディスク制御部は、

前記チャンネル制御部により受信された前記書き込み要求に伴い前記第一の記憶デバイスの前記論理ボリュームに前記データを書き込むと、前記データが書き込まれている前記第一の記憶デバイスの前記論理ボリュームの識別子、当該論理ボリューム内の前記データが格納されている位置情報、前記計時機構より取得される現在時刻である更新時刻、及び前記データで構成されるジャーナルデータを前記第三の記憶デバイスに書き込むジャーナル書き込み部と、

前記第三の記憶デバイスに記憶されている複数の前記ジャーナルデータのそれぞれの前記更新時刻と前記メモリに記憶されている前記検出時間とを参照し、前記計時機構より取得される現在時刻と前記更新時刻との差が前記共有メモリに記憶されている前記検出時間を超えている前記ジャーナルデータを前記複数のジャーナルデータの中から選択し、選択された前記ジャーナルデータの前記更新時刻が早い順に前記ジャーナルデータの前記論理ボリュームの識別子と前記位置情報と前記データとを参照し、当該論理ボリュームの識別子が前記主論理ボリュームの識別子である前記副論理ボリュームの識別子を前記ペア管理テーブルから取得し、当該データを前記第二の記憶デバイスの当該副論理ボリュームの識別子で示される前記論理ボリュームの当該位置情報で示される場所に書き込む第二の書き込み部と

30

前記情報処理装置から前記第二の記憶デバイスへの前記データの書き込みを中断するスプリット指示命令を受信するスプリット指示命令受信部と、

40

前記情報処理装置から前記第二の記憶デバイスへの前記データの書き込みを再開するスプリット解除命令を受信するスプリット解除命令受信部と、

前記スプリット指示命令を受信すると前記スプリット指示命令を受信したことを示すデータと前記計時機構より取得される現在時刻であるスプリット時刻とで構成されるスプリット指示データを前記第三の記憶デバイスに書き込むスプリット指示記憶部と、

前記スプリット解除命令を受信すると前記スプリット解除命令を受信したことを示すデータであるスプリット解除データを前記第三の記憶デバイスに書き込むスプリット解除記憶部と

を有し、

50

前記第二の書き込み部は前記第三の記憶デバイスに前記スプリット指示データが記憶され、かつ、前記スプリット解除データが記憶されていない場合は、前記スプリット指示データの前記スプリット時刻を参照し、前記更新時刻が当該スプリット時刻より遅い前記ジャーナルデータの前記データを前記第二の記憶デバイスに書き込まない

ことを特徴とするディスク制御装置。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、ディスク制御装置及びディスク制御装置の制御方法に関する。

【背景技術】

10

【0002】

近年、インターネットを利用した商品販売システムなど、24時間無停止で運用される情報処理システムが増加している。このような情報処理システムで用いられるデータは、データ量の増加や耐障害性の向上の要求に伴い、ディスクアレイ装置上のデータベースに格納されることが多い。無停止で運用される情報処理システムにおいては、データベースの更新処理を停止することなく、データベースに格納されているデータをバックアップする必要がある。そこで、ディスクアレイ装置内にバックアップ用のレプリカボリュームを設け、データベースに格納されているデータをバックアップ用のレプリカボリュームにも記憶しておく方法が用いられている。レプリカボリュームへのデータの書き込みを停止し、レプリカボリュームに記憶されているデータを磁気テープ等の外部記憶媒体にコピーすることにより、データベースのある時点のバックアップを取得することができる。

20

【0003】

特許文献1においては、レプリカボリュームへのデータの書き込みを停止している間にデータベースに書き込まれたデータを、バックアップの完了後にレプリカボリュームに書き込むリシンク処理が開示されている。

リシンク処理を実行している最中は、レプリカボリュームの内容は保証されない。そこで、1つのデータベースに対して2つのレプリカボリュームを設け、リシンク処理を2つのレプリカボリュームで交互に行っている場合もある。この場合、データベースを格納するデータ領域にハードウェア障害が発生すると、いずれか一方のレプリカボリュームとデータベースのREDOログとを用いてデータベースを復旧することが可能である。

30

【特許文献1】米国特許第6141773号明細書

【発明の開示】

【発明が解決しようとする課題】

【0004】

データベースの運用においては、上記のハードウェア障害の他に、ソフトウェアの不具合やユーザの操作ミス等によりデータベースに不正なデータが書き込まれる場合がある。このような障害をソフトウェア障害と呼んでいる。ソフトウェア障害が発生すると、磁気テープ等に保存されているある時点のデータを復元し、復元されたデータとREDOログとを用いてデータベースを復旧する必要がある。磁気テープ等からのデータ復元には相当の時間を要するため、システムの停止時間が長くなってしまう。

40

【0005】

そこで、ソフトウェア障害が発生した際に、磁気テープ等の外部記憶媒体からデータを復元することなく、ディスクアレイ装置内のデータを用いて迅速にデータベースを復旧することが求められている。ここで、前述した2つのレプリカボリュームを用いる場合、データベースを格納するボリュームの3倍の記憶容量が必要となり、ディスクアレイ装置を導入する際の費用が増大する。

【0006】

そのため、ハードウェア障害及びソフトウェア障害に対して、データベースをバックアップするために必要となる記憶容量を低減させることが求められている。

また、リシンク処理においては、レプリカボリュームに書き込むためのデータをデータ

50

ベースから読み出す必要がある。そのため、リシンク処理中は外部からデータベースへのアクセス性能が劣化するという問題がある。そこで、外部からデータベースへのアクセス性能を低下させずに、データベースを迅速に復旧することが求められている。

また、データベースの迅速な復旧とは別に、データベースの更新処理を停止することなく、万が一に備えてデータベースのある時点のバックアップを取得することが求められている。

【0007】

本発明は上記課題を鑑みてなされたものであり、ディスク制御装置及びディスク制御装置の制御方法を提供することを主たる目的とする。

【課題を解決するための手段】

【0008】

上記課題を解決するために、本発明のうち主たる発明では、情報処理装置と通信可能に接続され、一又は複数の論理ボリュームが形成されている第一の記憶デバイスと、一又は複数の論理ボリュームが形成されている第二の記憶デバイスと、第三の記憶デバイスとに対するデータの書き込みまたは読み出しを行うディスク制御装置であって、前記第一の記憶デバイスの前記論理ボリュームの識別子である主論理ボリュームの識別子と前記第二の記憶デバイスの前記論理ボリュームの識別子である副論理ボリュームの識別子とが対応付けられたペア管理テーブルと、任意の値に設定された検出時間と、を記憶するメモリと、計時機構と、前記情報処理装置から前記第一の記憶デバイスの前記論理ボリュームに対するデータの書き込み要求と前記データとを受信する書き込み要求受信部と、前記書き込み要求を受信すると前記データを前記第一の記憶デバイスの前記論理ボリュームに書き込む第一の書き込み部と、前記データが書き込まれている前記第一の記憶デバイスの前記論理ボリュームの識別子、当該論理ボリューム内の前記データが格納されている位置情報、前記計時機構より取得される現在時刻である更新時刻、及び前記データで構成されるジャーナルデータを前記第三の記憶デバイスに書き込むジャーナル書き込み部と、前記第三の記憶デバイスに記憶されている複数の前記ジャーナルデータのそれぞれの前記更新時刻と前記メモリに記憶されている前記検出時間とを参照し、前記計時機構より取得される現在時刻と前記更新時刻との差が前記検出時間を超えている前記ジャーナルデータを前記複数のジャーナルデータの中から選択し、選択された前記ジャーナルデータの前記更新時刻が早い順に前記ジャーナルデータの前記論理ボリュームの識別子と前記位置情報と前記データとを参照し、当該論理ボリュームの識別子が前記主論理ボリュームの識別子である前記副論理ボリュームの識別子を前記ペア管理テーブルから取得し、当該データを前記第二の記憶デバイスの当該副論理ボリュームの識別子で示される前記論理ボリュームの当該位置情報で示される場所に書き込む第二の書き込み部と、前記情報処理装置から前記第二の記憶デバイスへの前記データの書き込みを中断するスプリット指示命令を受信するスプリット指示命令受信部と、前記情報処理装置から前記第二の記憶デバイスへの前記データの書き込みを再開するスプリット解除命令を受信するスプリット解除命令受信部と、前記スプリット指示命令を受信すると前記スプリット指示命令を受信したことを示すデータと前記計時機構より取得される現在時刻であるスプリット時刻とで構成されるスプリット指示データを前記第三の記憶デバイスに書き込むスプリット指示記憶部と、前記スプリット解除命令を受信すると前記スプリット解除命令を受信したことを示すデータであるスプリット解除データを前記第三の記憶デバイスに書き込むスプリット解除記憶部とを有し、前記第二の書き込み部は前記第三の記憶デバイスに前記スプリット指示データが記憶され、かつ、前記スプリット解除データが記憶されていない場合は、前記スプリット指示データの前記スプリット時刻を参照し、前記更新時刻が当該スプリット時刻より遅い前記ジャーナルデータの前記データを前記第二の記憶デバイスに書き込まないこととする。

【発明の効果】

【0009】

ディスク制御装置及びディスク制御装置の制御方法を提供することができる。

【発明を実施するための最良の形態】

【 0 0 1 0 】

＝ ディスクアレイ装置 ＝

本実施の形態におけるディスクアレイ装置の第一の形態を図 1 に示す。ディスクアレイ装置 1 0 は、ディスク制御装置 1 1 0 と複数のハードディスクドライブ 1 2 0 とを備えている。ディスクアレイ装置 1 0 は情報処理装置 2 0 と通信手段により接続されている。通信手段は、例えば、L A N (Local Area Network) や S A N (Storage Area Network)、i S C S I (Internet Small Computer System Interface)、E S C O N (Enterprise Systems Connection) (登録商標)、F I C O N (Fibre Connection) (登録商標) などである。

【 0 0 1 1 】

10

情報処理装置 2 0 は、C P U (Central Processing Unit) やメモリを備えるコンピュータであり、パーソナルコンピュータやワークステーション、メインフレームなどのコンピュータである。情報処理装置 2 0 は、結合された複数台のコンピュータで構成されることもある。情報処理装置 2 0 ではオペレーティングシステムが動作している。オペレーティングシステム上ではアプリケーションソフトウェアが動作している。アプリケーションソフトウェアは、例えば、銀行の自動預金預け払いシステムや航空機の座席予約システムの機能を提供する。

【 0 0 1 2 】

ディスク制御装置 1 1 0 はディスクアレイ装置 1 0 全体の制御を司る。ディスク制御装置 1 1 0 は、情報処理装置 2 0 から受信したコマンドに従ってハードディスクドライブ 1 2 0 に対する制御を行う。例えば情報処理装置 2 0 からデータの入出力要求を受信して、ハードディスクドライブ 1 2 0 に記憶されているデータの入出力のための処理を行う。

20

【 0 0 1 3 】

ディスク制御装置 1 1 0 は、チャンネル制御部 1 3 1、ディスク制御部 1 3 2、共有メモリ 1 3 3、キャッシュメモリ 1 3 4、これらの間を通信可能に接続するクロスバスイッチなどで構成されるスイッチング制御部 1 3 5、及び管理端末 1 3 6などを備えて構成される。また、ディスク制御装置 1 1 0 を構成する各部 1 3 1 ~ 1 3 6 は耐障害性を高めるために冗長化されてもよい。

【 0 0 1 4 】

キャッシュメモリ 1 3 4 は、主としてチャンネル制御部 1 3 1 とディスク制御部 1 3 2 との間で授受されるデータを一時的に記憶するために用いられる。例えばチャンネル制御部 1 3 1 が情報処理装置 2 0 から受信したデータ入出力コマンドが書き込みコマンドである場合には、チャンネル制御部 1 3 1 は情報処理装置 2 0 から受信した書き込みデータをキャッシュメモリ 1 3 4 に書き込む。またディスク制御部 1 3 2 はキャッシュメモリ 1 3 4 から書き込みデータを読み出してハードディスクドライブ 1 2 0 に書き込む。なお、キャッシュメモリ 1 3 4 は不揮発とすることもできる。この場合、チャンネル制御部 1 3 1 が情報処理装置 2 0 から受信したデータがキャッシュメモリ 1 3 4 に書き込まれた段階で、情報処理装置 2 0 に書込完了通知を送信することとしてもよい。

30

【 0 0 1 5 】

ディスク制御部 1 3 2 は、チャンネル制御部 1 3 1 により共有メモリ 1 3 3 に書き込まれたデータ入出力要求を読み出してそのデータ入出力要求に設定されているコマンド (例えば、S C S I (Small Computer System Interface) 規格のコマンド) に従ってハードディスクドライブ 1 2 0 にデータの書き込みや読み出しなどの処理を実行する。ディスク制御部 1 3 2 はハードディスクドライブ 1 2 0 から読み出したデータをキャッシュメモリ 1 3 4 に書き込む。またデータの書き込み完了通知や読み出し完了通知などをチャンネル制御部 1 3 1 に送信する。ディスク制御部 1 3 2 は、ハードディスクドライブ 1 2 0 をいわゆる R A I D (Redundant Array of Inexpensive Disks) 方式に規定される R A I D レベル (例えば、0, 1, 5) で制御する機能を備えることもある。

40

【 0 0 1 6 】

ハードディスクドライブ 1 2 0 により提供される記憶領域は、この記憶領域上に論理的

50

に設定されるボリュームである論理ボリューム 121 を単位として管理されている。ハードディスクドライブ 120 へのデータの書き込みや読み出しは、論理ボリュームに付与される識別子を指定して行なうことができる。

【0017】

管理端末 136 はディスクアレイ装置 10 を保守・管理するためのコンピュータである。チャンネル制御部 131 やディスク制御部 132 において実行されるソフトウェアやパラメータの変更は、管理端末 136 からの指示により行われる。管理端末 136 はディスクアレイ装置 10 に内蔵される形態とすることもできるし、別体とすることもできる。

【0018】

共有メモリ 133 はチャンネル制御部 131、ディスク制御部 132、及び管理端末 136 からアクセスが可能である。チャンネル制御部 131 とディスク制御部 132 との間におけるデータ入出力要求コマンドの受け渡しに利用される他、ディスクアレイ装置 10 の管理情報等が記憶される。

【0019】

図 2 はチャンネル制御部 131 の構成を示すブロック図である。チャンネル制御部 131 は、インタフェース部 201、メモリ 202、CPU 203、NVRAM (Nonvolatile Random-Access Memory) 204、コネクタ 205 を備え、これらが一枚もしくは複数枚の回路基板に一体的なユニットとして形成されている。

【0020】

インタフェース部 201 は、情報処理装置 20 との間で通信を行うためのインタフェースを備えている。通信を行うためのインタフェースとは、例えば、ファイバチャネルに対応したコネクタやイーサネット（登録商標）に対応したコネクタなどである。

【0021】

コネクタ 205 は、チャンネル制御部 131 をディスクアレイ装置 10 に接続するためのコネクタである。コネクタ 205 がディスクアレイ装置 10 側のコネクタと嵌合することにより、チャンネル制御部 131 が形成される基板はディスクアレイ装置 10 と電氣的に接続される。チャンネル制御部 131 はコネクタ 205 を介してスイッチング制御部 135 に接続され、共有メモリ 133、キャッシュメモリ 134、ディスク制御部 132 などにアクセスすることができる。

【0022】

CPU 203 はチャンネル制御部 131 の全体の制御を司る。CPU 203 はメモリ 202 や NVRAM 203 に格納されている各種プログラムを実行することによりチャンネル制御部 131 の機能を実現する。NVRAM 203 は各種プログラムや設定データなどを格納する不揮発性のメモリである。NVRAM 203 に記憶される各種プログラムや設定データなどの内容は管理端末 136 からの指示により書き換えを行うことができる。

【0023】

図 3 はディスク制御部 132 の構成を示すブロック図である。ディスク制御部 132 は、インタフェース部 301、メモリ 302、CPU 303、NVRAM 304、コネクタ 305 を備え、これらが一枚もしくは複数枚の回路基板に一体的なユニットとして形成されている。

【0024】

インタフェース部 301 は、ハードディスクドライブ 120 との間で通信を行うためのインタフェースを備えている。通信を行うためのインタフェースとは、例えば、SCSI 規約に従うコネクタやファイバチャネル規約に従うコネクタなどである。

【0025】

コネクタ 305 は、ディスク制御部 132 をディスクアレイ装置 10 に接続するためのコネクタである。コネクタ 305 がディスクアレイ装置 10 側のコネクタと嵌合することにより、ディスク制御部 132 が形成される基板はディスクアレイ装置 10 と電氣的に接続される。ディスク制御部 132 はコネクタ 305 を介してスイッチング制御部 135 に接続され、共有メモリ 133、キャッシュメモリ 134、チャンネル制御部 131 などにア

10

20

30

40

50

クセスすることができる。

【0026】

CPU303はディスク制御部132の全体の制御を司る。CPU303はメモリ302やNVRAM303に格納されている各種プログラムを実行することによりディスク制御部132の機能を実現する。NVRAM303は各種プログラムや設定データなどを格納する不揮発性のメモリである。NVRAM303に記憶される各種プログラムや設定データなどの内容は管理端末136からの指示により書き換えを行うことができる。

【0027】

次に、本実施の形態におけるディスクアレイ装置の第二の形態を図4に示す。第二の形態においては、ディスク制御装置110の構成が図1に示した第一の形態と異なる。ディスク制御装置110は、CPU141、メモリ142、ホストインタフェース143、ディスクインタフェース144、キャッシュメモリ145、データコントローラ146を備えている。

10

【0028】

CPU141はディスクアレイ装置10の全体の制御を司る。CPU141はメモリ142に格納されているプログラムを実行することによりハードディスクドライブ120の管理やブロックアクセス要求の解釈など様々な機能を実現することができる。

【0029】

ホストインタフェース143は情報処理装置20との間で通信を行うインタフェースである。ホストインタフェース143はファイバチャネルプロトコルに従ってブロックアクセス要求を受け付ける機能を備える。

20

ディスクインタフェース144はハードディスクドライブ120との間でデータのやり取りを行うインタフェースである。ディスクインタフェース144はハードディスクドライブ120を制御するコマンドなどを規定するプロトコルに従ってハードディスクドライブ120に対するデータ入出力要求を送信する機能を備える。ディスクインタフェース144は、SCSIやファイバチャネルなどのプロトコルに従ってハードディスクドライブ120に対してデータの書き込みや読み出しのコマンドを送信することができる。

【0030】

キャッシュメモリ145は、ホストインタフェース143とディスクインタフェース144との間で授受されるデータが記憶されるメモリである。

30

データコントローラ146は、CPU141の制御によりホストインタフェース143とキャッシュメモリ145との間、あるいはキャッシュメモリ145とディスクインタフェース144との間でのデータ転送を行うものである。データコントローラ146は、例えば、特定用途向けICにロジックを形成する回路とすることができる。

【0031】

情報処理装置20がハードディスクドライブ120に対するデータの書き込み要求をディスクアレイ装置10に送信すると、ディスクアレイ装置10では、ホストインタフェース143が書き込み要求を受け付け、データコントローラ146が書き込み要求に付随する書き込みデータをキャッシュメモリ145に転送する。書き込みデータがキャッシュメモリ145に転送されると、データコントローラ146がキャッシュメモリ145からディスクインタフェース144に書き込みデータを読み出し、ディスクインタフェース144がハードディスクドライブ120に対して書き込みを指示するコマンドを送信する。

40

【0032】

なお、キャッシュメモリ145は不揮発とすることもできる。この場合、ホストインタフェース143が情報処理装置20から受信したデータがキャッシュメモリ145に書き込まれた段階で、CPU141が情報処理装置20に書込完了通知を送信することとしてもよい。

【0033】

以上、ディスクアレイ装置10の構成について説明した。なお、ディスクアレイ装置10は、以上に説明した構成のもの以外にも、例えば、NFS(Network File System)な

50

どのプロトコルにより情報処理装置 20 からファイル名指定によるデータ入出力要求を受け付けるように構成された N A S (Network Attached Storage) として機能するものなどであってもよい。

【0034】

== データベースの運用形態 ==

次に、本実施の形態における、ディスクアレイ装置 10 に構築されたデータベースの運用形態について説明する。

図 5 は、クライアント端末 30、データベースサーバ 40、ディスクアレイ装置 10 で構成されるデータベースシステムを示す図である。データベースサーバ 40 は、図 1 および図 4 における情報処理装置 20 に相当する。

図 6 は、クライアント端末 30 の構成を示すブロック図である。クライアント端末 30 は、CPU 601、メモリ 602、記憶装置 603、ポート 604、記録媒体読取装置 605、入力装置 606、出力装置 607 を備える。

【0035】

CPU 601 はクライアント端末 30 の全体の制御を司るもので、メモリ 602 や記憶装置 603 に格納されたプログラムを実行することにより各種機能を実現する。記憶装置 603 は、例えばハードディスクドライブなどの記憶装置である。記録媒体読取装置 605 は、記録媒体 608 に記録されているプログラムやデータを読み取るための装置である。読み取られたプログラムやデータはメモリ 602 や記憶装置 603 に格納される。記録媒体 608 としてはフレキシブルディスクや CD-ROM、半導体メモリ等を用いることができる。記録媒体読取装置 605 はクライアント端末 30 に内蔵されている形態とすることもできるし、外付されている形態とすることもできる。入力装置 606 はオペレータ等によるクライアント装置 30 へのデータ入力等のために用いられる。入力装置 606 としては例えばキーボードやマウス等が用いられる。出力装置 607 は情報を外部に出力するための装置である。出力装置 607 としては例えばディスプレイやプリンタ等が用いられる。ポート 604 はデータベースサーバ 40 と通信を行うための装置である。

【0036】

図 7 は、データベースサーバ 40 の構成を示すブロック図である。データベースサーバ 40 は、CPU 701、メモリ 702、記憶装置 703、LAN インタフェース 704、ストレージインタフェース 705、記録媒体読取装置 706 を備える。

【0037】

CPU 701 はデータベースサーバ 40 の全体の制御を司るもので、メモリ 702 や記憶装置 703 に格納されたプログラムを実行することにより各種機能を実現する。記憶装置 703 は、例えばハードディスクドライブなどの記憶装置である。記録媒体読取装置 706 は、記録媒体 707 に記録されているプログラムやデータを読み取るための装置である。読み取られたプログラムやデータはメモリ 702 や記憶装置 703 に格納される。記録媒体 707 としてはフレキシブルディスクや CD-ROM、半導体メモリ等を用いることができる。記録媒体読取装置 706 はデータベースサーバ 40 に内蔵されている形態とすることもできるし、外付されている形態とすることもできる。

【0038】

LAN インタフェース 704 はクライアント端末 30 と LAN 等の通信手段により通信を行うためのインタフェースである。また、ストレージインタフェース 705 は、ストレージ装置 10 と SAN や LAN 等の通信手段により通信を行うためのインタフェースである。

【0039】

図 5 に示すように、ディスクアレイ装置 10 には、データ領域 501 と REDO ログ領域 502 とが設けられている。データ領域 501 と REDO ログ 502 領域は、一つ又は複数のハードディスクドライブ 120 により構成される記憶領域である。クライアント端末 30 は、データ領域 501 に格納されているテーブル 503、504 の更新要求をデータベースサーバ 40 に送信する。データベースサーバ 40 には、データベースバッファ 5

10

20

30

40

50

05とREDOログバッファ506とが設けられている。データベースバッファ505とREDOログバッファ506は後述する記憶装置703に記憶されている。データベースバッファ505に格納されているデータは、データ領域501に格納されているデータのキャッシュデータとして使用される。これにより、データベースサーバ40はディスクアレイ装置10へのアクセスを行うことなく、クライアント端末30からのテーブル更新要求に応答することができ、クライアント端末30への応答時間を短縮することが可能となる。同様に、REDOログバッファ506に格納されているデータは、REDOログ領域502に格納されているデータのキャッシュデータとして使用される。

【0040】

データベースの更新処理においては、複数のテーブルの更新を保証することが要求される。例えば、銀行の預金口座間における振込の場合、振込元の口座を管理するテーブルに記憶されている預金残高を減額し、振込先の口座を管理するテーブルに記憶されている預金残高を増額する必要がある。つまり、振込元のテーブルの更新と振込先のテーブルの更新とが完了しないと、更新処理が正しく行われたことにならない。データベースサーバ40は、このように複数のテーブルの更新を保証する機能を備えている。クライアント端末30は、複数のテーブルの更新処理を実行した後に、これら複数のテーブルの更新処理の保証を要求するためのコミット要求をデータベースサーバ40に送信する。データベースサーバ40は、これら複数のテーブルの更新処理の途中で何らかの障害が発生した場合は、これら複数のテーブルを更新前の状態に戻すことができる。

【0041】

図8は、データ領域501のテーブル503, 504を更新する処理を示すフローチャートである。クライアント端末30は、データベースサーバ40にテーブル503の更新要求を送信する(S801)。更新要求にはテーブル503における更新位置と更新データとが含まれている。更新要求を受信すると、データベースサーバ40は更新位置情報をもとに更新データをデータベースバッファ505に格納する(S802)。そして、データベースサーバ40は、更新位置情報と更新データとをREDOログバッファ506に格納し(S803)、クライアント端末30に更新完了通知を送信する(S804)。続いて、クライアント端末30はテーブル504の更新要求をデータベースサーバ40に送信する。データベースサーバ40は、同様にデータベースバッファ505及びREDOログバッファ506の更新を行い(S806, S807)、更新完了通知をクライアント端末30に送信する(S808)。テーブル504の更新完了通知を受信すると、クライアント端末30は、当該更新処理のコミット要求をデータベースサーバ40に送信する(S809)。

コミット要求を受信すると、データベースサーバ40は、REDOログバッファ506に格納されているテーブル503及びテーブル504の更新における更新位置情報と更新データとをREDOログ領域502に書き込む更新要求をディスクアレイ装置10に送信する(S810)。ディスクアレイ装置10はREDOログ領域502に更新位置情報と更新データとを格納し(S811)、更新完了通知をデータベースサーバ40に送信する(S812)。データベースサーバ40は、ディスクアレイ装置10からREDOログ領域502の更新完了通知を受信すると、コミット完了通知をクライアント端末30に送信する(S813)。クライアント端末30は、データベースサーバ40からコミット完了通知を受信し(S814)、テーブル503, 504の更新が保証されたことを認識する。

【0042】

データベースバッファ505に格納された更新データは、クライアント端末30との間で行われる一連の処理(S801~S814)とは非同期に、ディスクアレイ装置10のデータ領域501に格納される。まず、データベースサーバ40は、データ領域501から更新前のデータの読み出す要求をディスクアレイ装置10に送信する(S851)。ディスクアレイ装置10は、更新前のデータをデータ領域501から読み出し、データベースサーバ40に送信する(S852)。データベースサーバ40は、更新前のデータと更

新位置情報とを R E D O ログ領域 5 0 2 に書き込む要求をディスクアレイ装置 1 0 に送信する (S 8 5 3)。ディスクアレイ装置 1 0 は、更新前のデータと更新位置情報とを R E D O ログ領域 5 0 2 に格納し (S 8 5 4)、更新完了通知をデータベースサーバ 4 0 に送信する (S 8 5 5)。データベースサーバ 4 0 は、R E D O ログ領域 5 0 2 の更新完了通知を受信すると、データベースバッファ 5 0 5 に格納されている更新データをデータ領域 5 0 1 に書き込む要求をディスクアレイ装置 1 0 に送信する (S 8 5 6)。そして、ディスクアレイ装置 1 0 は更新データをデータ領域 5 0 1 に格納する (S 8 5 7)。

このように、データベースサーバ 4 0 は、時間のかかるデータ領域 5 0 1 への更新データの格納を、データベースバッファ 5 0 5 への更新データの格納とは別のタイミングで行っている。これにより、データベースサーバ 4 0 は、クライアント端末 3 0 からの更新要求に対する応答時間を短縮している。

10

【 0 0 4 3 】

また、データベースサーバ 4 0 に障害が発生した場合は、R E D O ログ領域 5 0 2 に格納されているデータ (以後、「R E D O ログ」と称する)を確認することにより、データ領域 5 0 1 のデータをクライアント端末 3 0 からのコミット要求を反映した内容に復旧することが可能である。つまり、R E D O ログに、S 8 1 1 の処理で格納されるコミット要求に伴う更新データが存在せず、かつ、S 8 5 4 の処理で格納される更新前データが存在する場合、更新前データをデータ領域 5 0 1 に書き込む。これを、ロールバック処理という。また、R E D O ログに、S 8 1 1 の処理で格納されるコミット要求に伴う更新データが存在し、かつ、S 8 5 4 の処理で格納される更新前データが存在しない場合、更新データをデータ領域 5 0 1 に書き込む。これを、ロールフォワード処理という。

20

このように、データ領域 5 0 1 のデータをクライアント端末 3 0 からのコミット要求を反映した内容に復旧する処理は、データベースのリカバリ処理と呼ばれている。リカバリ処理は、前述した手順に限られない。例えば、更新前データの全てをロールバックし、コミット要求に伴う更新データを全てロールフォワードすることにより実現されるものとしてもよい。

【 0 0 4 4 】

次に、図 9 にて、複数台のハードディスクドライブ 1 2 0 に障害が発生し、R A I D の冗長性等を用いてデータ領域 5 0 1 のデータを復旧することができない場合 (以後、「データ領域の障害」と称する)における、データ領域 5 0 1 の一般的な復旧手順について説明する。

30

【 0 0 4 5 】

図 9 は、R E D O ログの記録を 0 時に開始し、1 2 時にデータ領域 5 0 1 に障害が発生した場合を示している。また、データ領域 5 0 1 の 6 時の状態がデータ領域 5 0 1 を構成するハードディスクドライブ 1 2 0 とは別のハードディスクドライブ 1 2 0 又は磁気テープ等の記憶媒体にバックアップされている (S 9 0 1)。なお、データ領域 5 0 1 のある時点のデータ (以後、「静止化データ」と称する)をバックアップする一般的な手順については、後述する。

【 0 0 4 6 】

1 2 時にデータ領域 5 0 1 の障害が発生した場合、障害が発生している一つ又は複数のハードディスクドライブ 1 2 0 を交換した後、バックアップされている 6 時のデータをデータ領域 5 0 1 に復元する (S 9 0 2)。そして、前述したリカバリ処理の手順に従い、0 時以降の更新前データの全てをロールバックし (S 9 0 3)、0 時から障害発生前までのコミット要求に伴う更新データをロールフォワードする (S 9 0 4)。これにより、障害直前の状態までデータ領域 5 0 1 を復旧することができる。

40

【 0 0 4 7 】

また、ユーザの操作ミスによる障害やソフトウェア障害によりデータ領域 5 0 1 が不正な状態となった場合も、静止化データと R E D O ログとを用いてデータ領域 5 0 1 を復旧することができる。ユーザの操作ミスによる障害とは、例えば、ファイルシステム上でデータ領域 5 0 1 をフォーマットしてしまうことによりデータ領域 5 0 1 を失ってしまう場

50

合などである。また、ソフトウェア障害とは、アプリケーションプログラムの不具合等により、データ領域 501 に格納されているデータに不整合が生じてしまう場合などである。このような障害が発生した場合は、障害時刻をある程度特定し、REDO ログを用いたロールフォワード処理を障害時刻の前までとする。これにより、障害前の状態のデータ領域 501 を復旧することができる。

【0048】

以上のように、データ領域 501 に障害が発生した場合は、静止化データと REDO ログとを用いて、データ領域 501 を復旧することが可能である。また、静止化データをディスクアレイ装置 10 内に保持しておくことにより、磁気テープ等からのデータ復元を行わずに、迅速にデータ領域 501 の復旧を行うことも可能である。

10

【0049】

== データバックアップ方式 ==

次に、データのバックアップ方式について説明する。図 10 はレプリカ方式、図 11 はスナップショット方式を示す図である。

【0050】

まず、図 10 を用いてレプリカ方式について説明する。レプリカ方式では、データ領域 501 とは別の記憶領域としてレプリカボリューム 1001 が設けられている。レプリカボリューム 1001 は一つ又は複数のハードディスクドライブ 120 により構成される。ディスク制御装置 110 は、データベースサーバ 40 からデータ領域 501 の更新要求を受信すると、データ領域 501 とレプリカボリューム 1001 の双方に更新データを書き込む。このように、双方に更新データを書き込んでいる状態を同期状態と称する。ディスク制御部 110 は、データベースサーバ 40 等からレプリカボリューム 1001 への更新データの書き込みを停止するスプリット指示命令を受信する。スプリット指示命令を受信すると、ディスク制御部 110 は、レプリカボリューム 1001 への更新データの書き込みを停止する。つまり、レプリカボリューム 1001 はスプリット指示命令を受信した時点における、データ領域 501 の静止化データである。

20

【0051】

ディスク制御装置 110 は、レプリカボリューム 1001 の更新を停止している間にデータ領域 501 に書き込まれた更新データの位置情報を、メモリ 1003 の差分セグメントビットマップ 1004 に記憶する。メモリ 1003 は、図 1 の共有メモリ 133、キャッシュメモリ 134、図 4 のキャッシュメモリ 145 などである。ディスク制御装置 110 は、データベースサーバ 40 等からレプリカボリューム 1001 を同期状態にするリシンク指示命令を受信すると、差分セグメントビットマップ 1004 に記憶されている位置情報で示されるデータをデータ領域 501 から読み出してレプリカボリューム 1001 に書き込む。このリシンク指示命令に伴う処理をリシンク処理と称する。

30

【0052】

更新を停止している間にデータ領域 501 に書き込まれたデータの量が多い場合、リシンク処理には相当の時間を要する。リシンク処理では、データ領域 501 におけるデータの更新順序とは関係なく、セグメント単位でレプリカボリューム 1001 のデータが更新される。つまり、リシンク処理の実行中は、レプリカボリューム 1001 はデータ領域 501 のある時点の状態とはなっていない。したがって、リシンク処理の実行中にデータ領域 501 に障害が発生した場合、レプリカボリューム 1001 をデータ領域 501 を復旧するために用いることができない。

40

【0053】

そこで、ディスクアレイ装置 10 内に、もう一つレプリカボリューム 1002 が設けられている場合もある。この場合、2つのレプリカボリューム 1001, 1002 で交互に更新の停止とリシンク処理を行うことで、ディスクアレイ装置 10 内に必ずデータ領域 501 の静止化データを確保することができる。

【0054】

次に、図 11 を用いてスナップショット方式について説明する。スナップショット方式

50

では、データ領域 5 0 1 とは別の記憶領域としてプール 1 1 0 1 が設けられている。プール 1 0 0 1 は一つ又は複数のハードディスクドライブ 1 2 0 により構成される。また、仮想的なレプリカボリュームである仮想レプリカボリューム 1 1 0 2 が提供される。ディスク制御装置 1 1 0 は、データベースサーバ 4 0 等から仮想レプリカボリューム 1 1 0 2 の作成指示を受信する。ディスク制御装置 1 1 0 は、データベースサーバ 4 0 からデータ領域 5 0 1 の更新要求を受信すると、データ領域 5 0 1 の更新予定位置に格納されている更新前データを読み出してプール 1 1 0 1 に書き込む。そして、ディスク制御装置 1 1 0 は、データ領域 5 0 1 に更新データを書き込み、メモリ 1 0 0 3 の差分セグメントビットマップ 1 0 0 4 に更新データの位置情報を記憶する。ディスク制御装置 1 1 0 は、仮想レプリカボリューム 1 1 0 2 に対する読み出し要求を受信すると、差分セグメントビットマップ 1 0 0 4 を参照し、読み出しデータの位置情報が記憶されている場合はプール 1 1 0 1 からデータを読み出し、そうでない場合はデータ領域 5 0 1 からデータを読み出す。つまり、仮想レプリカボリューム 1 1 0 2 は、仮想レプリカボリューム 1 1 0 2 の作成指示を受信した時点における、データ領域 5 0 1 の静止化データである。

【 0 0 5 5 】

以上に説明したレプリカ方式とスナップショット方式には、次のような欠点がある。レプリカ方式では、リシンク処理の際にデータ領域 5 0 1 に格納されているデータを読み出す必要がある。そのため、データベースサーバ 4 0 から受信するデータ領域 5 0 1 の更新要求に伴うデータ領域 5 0 1 の更新処理と競合し、データベースサーバ 4 0 からデータ領域 5 0 1 へのアクセス性能が低下してしまう。また、2つのレプリカボリューム 1 0 0 1 , 1 0 0 2 を設ける場合、データ領域 5 0 1 の3倍の記憶容量が必要となり、ディスクアレイ装置 1 0 の導入コストが高くなってしまう。

【 0 0 5 6 】

スナップショット方式では、静止化データである仮想レプリカボリューム 1 1 0 2 を読み出す際に、データ領域 5 0 1 に格納されているデータを読み出す必要がある。そのため、レプリカ方式の場合と同様、データベースサーバ 4 0 から受信するデータ領域 5 0 1 の更新要求に伴うデータ領域 5 0 1 の更新処理と競合し、データベースサーバ 4 0 からデータ領域 5 0 1 へのアクセス性能が低下してしまう。また、データ領域 5 0 1 において障害が発生した場合、ディスクアレイ装置 1 0 内にデータ領域 5 0 1 の静止化データが存在しないこととなる。そのため、データ領域 5 0 1 を迅速に復旧することができない。

【 0 0 5 7 】

次に、前述したレプリカ方式とスナップショット方式を改良した、本実施の形態におけるデータバックアップ方式を、図 1 2 を用いて説明する。

【 0 0 5 8 】

ディスクアレイ装置 1 0 には、データ領域 5 0 1 が格納される第一の記憶デバイス 1 2 0 1、第二の記憶デバイス 1 2 0 2、及び第三の記憶デバイス 1 2 0 3 が設けられている。第一乃至第三の記憶デバイス 1 2 0 1 ~ 1 2 0 3 は、それぞれ一つ又は複数のハードディスクドライブ 1 2 0 により構成されている。また、第一の記憶デバイス 1 2 0 1 及び第二の記憶デバイス 1 2 0 2 には、一つ又は複数の論理ボリューム 1 2 1 が形成されている。また、第一の記憶デバイス 1 2 0 1 を構成するハードディスクドライブ 1 2 0 と第二の記憶デバイス 1 2 0 2 及び第三の記憶デバイス 1 2 0 3 を構成するハードディスクドライブ 1 2 0 は物理的に別のものである。なお、REDOログは、ディスクアレイ装置 1 0 内の第一の記憶デバイス 1 2 0 1 を構成するハードディスクドライブ 1 2 0 とは別のハードディスクドライブ 1 2 0 に格納されている。また、REDOログは、ディスクアレイ装置 1 0 の外部にある記憶装置に格納されているものとしてもよい。ディスク制御装置 1 1 0 には、計時機構 1 2 0 4 が備えられている。計時機構 1 2 0 4 は、外部からの要求に応じて日付及び時刻を提供する。計時機構 1 2 0 4 は、クロック信号により時刻をカウントする計時回路などである。メモリ 1 0 0 3 には、所定の時間である検出時間 1 2 0 5 が記憶されている。検出時間 1 2 0 5 は、管理端末 1 3 6 や情報処理装置 2 0 などから登録される。

【 0 0 5 9 】

本方式における、データの更新処理の概略を説明する。ディスク制御装置 1 1 0 は、データベースサーバ 4 0 から第一の記憶デバイス 1 2 0 1 の論理ボリューム 1 2 1 の更新要求を受信する。ディスク制御装置 1 1 0 は、当該更新要求を受信すると、第一の記憶デバイス 1 2 0 1 の論理ボリュームに更新データを書き込む。そして、ディスク制御部 1 1 0 は第一の記憶デバイスに書き込んだ更新データと更新データが書き込まれた位置情報、更新時刻等で構成されるジャーナルデータを第三の記憶デバイスに書き込む。ディスク制御部 1 1 0 は、第三の記憶デバイスに書き込まれているジャーナルデータを所定の間隔で参照し、更新時刻と現在時刻との差が検出時間 1 2 0 5 を超えているジャーナルデータについて、更新時刻順に更新データを第二の記憶デバイス 1 2 0 2 の論理ボリューム 1 2 1 に書き込む。第二の記憶デバイス 1 2 0 2 の論理ボリューム 1 2 0 2 は、第一の記憶デバイス 1 2 0 1 の論理ボリューム 1 2 1 の検出時間 1 2 0 5 だけ前の状態となっている。つまり、第一の記憶デバイス 1 2 0 1 に障害が発生した場合、第二の記憶デバイス 1 2 0 2 と第三の記憶デバイス 1 2 0 3 又は R E D O ログとを用いて、第一の記憶デバイス 1 2 0 1 を復旧することが可能である。

10

【 0 0 6 0 】

図 1 3 は、本実施の形態に係るディスク制御装置 1 1 0 における前述したデータの更新処理を実現する機能を示すブロック図である。ディスク制御装置 1 1 0 は、書き込み要求受信部 1 3 0 1、第一の書き込み部 1 3 0 2、ジャーナル書き込み部 1 3 0 3、第二の書き込み部 1 3 0 4、スプリット指示命令受信部 1 3 0 5、スプリット解除命令受信部 1 3 0 6、スプリット指示記憶部 1 3 0 7、スプリット解除記憶部 1 3 0 8、未反映情報記憶部 1 3 0 9、読み出し要求受信部 1 3 1 0、識別子取得部 1 3 1 1、重複範囲取得部 1 3 1 2、仮想論理ボリューム読み出し部 1 3 1 3、読み出しデータ送信部 1 3 1 4 を備えている。各部 1 3 0 1 ~ 1 3 1 4 は、図 1 ~ 図 3 に示すディスクアレイ装置 1 0 における C P U 2 0 3 及び C P U 3 0 3、図 4 に示すディスクアレイ装置 1 0 における C P U 1 4 1 がメモリ 2 0 2、3 0 2、1 4 2 又は N V R A M 2 0 4、3 0 4 に格納されているプログラムを実行することにより実現される。

20

【 0 0 6 1 】

== ペア及びグループ ==

図 1 4 は、第一の記憶デバイス 1 2 0 1 の論理ボリューム 1 2 1 と第二の記憶デバイス 1 2 0 2 の論理ボリューム 1 2 1 との関係を示す図である。第一の記憶デバイス 1 2 0 1 の論理ボリューム 1 2 1 は、第二の記憶デバイス 1 2 0 2 の論理ボリューム 1 2 1 と対応付けられている。この対応付けにおける第一の記憶デバイス 1 2 0 1 の論理ボリューム 1 2 1 を主論理ボリューム、第二の記憶デバイス 1 2 0 2 の論理ボリューム 1 2 1 を副論理ボリュームと称する。この対応付けはペアと呼ばれ、図 1 5 に示すペア管理テーブル 1 5 0 1 に主論理ボリュームの識別子と副論理ボリュームの識別子とが対応付けられて記憶されている。また、第一の記憶デバイス 1 2 0 1 の論理ボリューム 1 2 1 は一つ又は複数の論理ボリューム 1 2 1 をまとまりとするグループを形成している。グループは、例えば、データベースサーバ 4 0 ごとや、データベースサーバ 4 0 が提供するデータベース・インスタンスごとに形成される。つまり、データベースサーバ 4 0 は第一の記憶デバイス 1 2 0 1 の論理ボリューム 1 2 1 をグループ単位で使用していると言うことができる。従って、第一の記憶デバイス 1 2 0 1 の論理ボリューム 1 2 1 をグループ単位でバックアップすることが必要となる。グループにはグループを示すグループ ID が付与され、図 1 6 に示すグループ管理テーブル 1 6 0 1 に論理ボリューム 1 2 1 の識別子とグループ ID との対応付けが記憶されている。ペア管理テーブル 1 5 0 1 及びグループ管理テーブル 1 6 0 1 はメモリ 1 0 0 3 に記憶されており、管理端末 1 3 6 や情報処理装置 2 0 などから登録される。

30

40

【 0 0 6 2 】

== ジャーナルデータ ==

図 1 7 は、第三の記憶デバイス 1 2 0 3 に書き込まれるジャーナルデータを示す図であ

50

る。ジャーナルデータ 1701 は、ジャーナル部 1702 及び更新ストリーム部 1703 からなる。

【0063】

ジャーナル部 1702 は、グループ ID、シーケンス ID、更新時刻、更新位置情報、及び更新ストリームオフセットで構成されている。データベースサーバ 40 から第一の記憶デバイス 1201 の論理ボリューム 121 に更新データを書き込む要求を受信すると、ディスク制御装置 110 は第三の記憶デバイス 1203 にジャーナルデータ 1701 を記憶する。グループ ID は、当該論理ボリューム 121 が属するグループのグループ ID である。シーケンス ID は、第一の記憶デバイス 1201 における更新データの更新順序を示すものであり、グループごとに管理されている連続番号である。ジャーナルデータ 1701 に格納されている更新データは、グループごとにシーケンス ID 順に第二の記憶デバイス 1202 の論理ボリューム 121 に書き込まれる。更新時刻は、ディスク制御装置 110 がジャーナルデータ 1701 を記憶する際に計時機構 1201 から取得した日付と時刻である。したがって、シーケンス ID の順序と更新時刻の順序は同じである。更新位置情報には、当該論理ボリューム 121 の識別子と当該論理ボリューム 121 において更新データが格納された位置情報とが記憶されている。なお、データが格納された位置情報とは、例えば、開始アドレス及び開始アドレスからのデータ長などで示される情報である。更新ストリームオフセットには、更新データの更新ストリーム部 1703 における位置情報が記憶されている。更新ストリーム部 1703 には、更新データが更新ストリームオフセットで示される位置に記憶されている。

【0064】

また、ディスク制御部 110 が後述するスプリット指示命令を受信した場合は、更新位置情報にスプリット指示命令を受信したことを示すデータが記憶される。グループ ID には、スプリット指示命令で指定されたグループ ID が記憶される。シーケンス ID には、スプリット指示命令で指定されたグループ ID のシーケンス ID が記憶される。更新時刻には、計時機構 1201 から取得した日付と時刻が記憶される。このように、スプリット指示命令に伴いジャーナル部 1701 に書き込まれたデータをスプリット指示データと称する。そして、ディスク制御部 110 が後述するスプリット解除命令を受信した場合は、更新位置情報にスプリット解除命令を受信したことを示すデータが記憶される。このように、スプリット解除命令に伴いジャーナル部 1701 に書き込まれたデータをスプリット解除データと称する。

【0065】

なお、第二の記憶デバイス 1202 の論理ボリューム 121 への反映が完了した更新データを示すジャーナルデータ 1701 の記憶領域は空き領域として管理され、別の更新データを示すジャーナルデータ 1701 の記憶領域として再利用される。

【0066】

== スプリット運用 ==

次に、本実施の形態において静止化データを作成してデータのバックアップを行う概要について説明する。図 18 は、静止化データを作成するためのスプリット指示命令を受信した場合の、ディスクアレイ装置 10 の状態遷移を示す図である。

【0067】

まず、通常運用の状態 (S1801) においては、第三の記憶デバイス 1203 のジャーナルデータ 1701 から、更新時刻と現在時刻との差が検出時間 1205 を超えている更新データが抽出され、第二の記憶デバイス 1202 に書き込まれている。

【0068】

ディスクアレイ装置 10 は、データベースサーバ 40 またはバックアップサーバ 1803 等からスプリット指示命令を受信すると、仮想デバイス 1801 を作成する状態 (S1802) へと遷移する。バックアップサーバ 1803 とは、図 1 及び図 4 における情報処理装置 10 に相当する装置である。バックアップサーバ 1803 は、ディスクアレイ装置 10 に格納されているデータを読み出し、外部の記憶媒体にバックアップする機能を備え

ている。

【 0 0 6 9 】

図 1 8 の S 1 8 0 2 は、ディスクアレイ装置 1 0 がスプリット指示命令を 1 8 時（スプリット時刻）に受信してから 5 分経過した状態を示している。なお、図 1 8 の例では検出時間は 2 時間である。S 1 8 0 2 において、第二の記憶デバイス 1 2 0 2 には、現在時刻の 2 時間前である 1 6 : 0 5 までの更新データが反映されている。つまり、1 6 : 0 5 から 1 8 : 0 0 までの更新データは、まだ第二の記憶デバイス 1 2 0 2 に反映されず、第三の記憶デバイス 1 2 0 3 に格納されている。

【 0 0 7 0 】

そこで、1 8 : 0 0 における第一の記憶デバイスの静止化データとして仮想デバイス 1 8 0 1 が提供される。仮想デバイス 1 8 0 1 には、仮想的な論理ボリュームである仮想論理ボリュームが設けられている。仮想論理ボリュームの識別子と第二の記憶デバイス 1 2 0 2 の論理ボリューム 1 2 1 の識別子との対応は、図 1 9 に示す仮想論理ボリューム管理テーブル 1 9 0 1 に記憶されている。仮想論理ボリューム管理テーブル 1 9 0 1 はメモリ 1 0 0 3 に記憶され、管理端末 1 3 6 や情報処理装置 2 0 などから登録される。

【 0 0 7 1 】

ディスク制御装置 1 1 0 は、更新時刻がスプリット時刻より早いジャーナルデータ 1 7 0 1 の更新位置情報を抽出し、未反映情報としてメモリ 1 0 0 3 に記憶している。未反映情報の記憶方式は、例えば、第二の記憶デバイス 1 2 0 2 の記憶領域を複数のセグメントに分割し、このセグメント単位に未反映の更新データがあるかどうかを示すビットマップ方式などである。ディスク制御装置 1 1 0 は、バックアップサーバ 1 8 0 2 から仮想論理ボリュームに対するデータの読み出し要求を受信すると、未反映情報で示される位置に格納されているデータについては、第三の記憶デバイス 1 2 0 3 からデータを読み出す。また、ディスク制御装置 1 1 0 は、未反映情報で示される位置以外に格納されているデータについては、第二の記憶デバイス 1 2 0 2 から読み出す。これにより、バックアップサーバ 1 8 0 2 に対して第一の記憶デバイス 1 2 0 1 の 1 8 時の静止化データを提供することができる。

【 0 0 7 2 】

仮想デバイス 1 8 0 1 が提供されている間も、第三の記憶デバイス 1 2 0 3 から第二の記憶デバイス 1 2 0 2 への更新データの反映は行われている。そのため、ディスク制御装置 1 1 0 は、仮想デバイス 1 8 0 1 を静止化データとするために、更新時刻がスプリット時刻以降の更新データを第二の記憶デバイス 1 2 0 2 に反映しない。この状態を示しているのが、S 1 8 0 3 である。S 1 8 0 3 においては、第二の記憶デバイス 1 2 0 2 にはスプリット時刻である 1 8 時までの更新データが反映されている。したがって、第二の記憶デバイス 1 2 0 2 と仮想デバイス 1 8 0 1 とは一致している。

【 0 0 7 3 】

ディスクアレイ装置 1 0 は、データベースサーバ 4 0 またはバックアップサーバ 1 8 0 1 等からスプリット解除命令を受信すると、第三の記憶デバイス 1 2 0 3 に格納されている、更新時刻と現在時刻との差が検出時間 1 2 0 5 を超えている更新データの第二の記憶デバイス 1 2 0 2 への反映を再開する。S 1 8 0 4 は、この状態を示している。そして、この反映が完了すると、通常運用の状態（S 1 8 0 1）に戻る。

【 0 0 7 4 】

なお、仮想デバイス 1 8 0 1 を用いずに、スプリット指示命令を受信するとディスク制御装置 1 1 0 は更新時刻がスプリット時刻より早い更新データを、更新時刻と現在時刻との差が検出時間 1 2 0 5 を超えているかどうかにかかわらず、第二の記憶デバイス 1 2 0 2 に反映するものとしてもよい。この場合、バックアップサーバ 1 8 0 2 は、仮想デバイス 1 8 0 1 ではなく、第二の記憶デバイス 1 2 0 2 からデータを読み出してバックアップを行う。

【 0 0 7 5 】

= = 動作の説明 = =

10

20

30

40

50

以上に説明した、本実施の形態におけるバックアップ方式を実現するための、ディスク制御装置 110 の各部 1301 ~ 1314 の動作を説明する。

【0076】

== 第三の記憶デバイスへの書き込み ==

図 20 は、ディスク制御装置 110 がデータベースサーバ 40 から第一の記憶デバイス 1201 の論理ボリューム 121 に対するデータの書き込み要求を受信した際の処理を示すフローチャートである。

【0077】

書き込み要求受信部 1301 は、データベースサーバ 40 から第一の記憶デバイス 1201 の論理ボリューム 121 に対するデータの書き込み要求と書き込み用のデータとを受信する (S2001)。第一の書き込み部 1302 は、当該データを第一の記憶デバイス 1201 の当該書き込み要求で指定される論理ボリューム 121 に書き込む (S2002)。ジャーナル書き込み部 1303 は、当該データに対するジャーナルデータ 1701 を第三の記憶デバイス 1203 に書き込む (S2003)。そして、ジャーナル書き込み部 1303 は、データベースサーバ 40 に対して当該データの書き込み完了通知を送信する (S2004)。

【0078】

なお、当該ジャーナルデータ 1701 の生成は次のように行われる。ジャーナル書き込み部 1303 は、当該論理ボリュームの識別子に対応するグループ ID をグループ管理テーブル 1601 から取得し、当該ジャーナルデータ 1701 のグループ ID に設定する。ジャーナル書き込み部 1303 は、取得したグループ ID において前回付与されたシーケンス ID に 1 加算し、当該ジャーナルデータ 1701 のシーケンス ID に設定する。ジャーナル書き込み部 1303 は、計時機構 1204 から現在の日時を取得し、当該ジャーナルデータ 1701 の更新時刻に設定する。ジャーナル書き込み部 1303 は、当該論理ボリュームの識別子と当該データが書き込まれた位置情報とを当該ジャーナルデータ 1701 の更新位置情報に設定する。ジャーナル書き込み部 1303 は、更新ストリーム部 1703 において次にデータが格納される位置を当該ジャーナルデータ 1701 の更新ストリームオフセットに設定する。そして、ジャーナル書き込み部 1303 は、当該データを更新ストリーム部 1703 に格納する。

【0079】

また、ジャーナルデータ 1701 は、第三の記憶デバイス 1203 の論理ボリューム 121 に書き込まれる前に、メモリ 1003 に書き込まれる。メモリ 1003 が冗長構成で、かつ不揮発性である場合、メモリ 1003 にジャーナルデータ 1701 が書き込まれた時点でジャーナル書き込み部 1303 がデータベースサーバ 40 に対して当該データの書き込み完了通知を送信することとしてもよい。

【0080】

== スプリット指示 ==

図 21 は、ディスク制御装置 110 がスプリット指示命令を受信した際の処理を示すフローチャートである。

【0081】

スプリット指示命令受信部 1305 は、データベースサーバ 40 やバックアップサーバ 1801 等からスプリット指示命令を受信する (S2101)。スプリット指示記憶部 1307 は、スプリット指示データを第三の記憶デバイス 1203 に書き込む (S2102)。なお、スプリット指示データの生成は次のように行われる。スプリット指示記憶部 1307 は、スプリット指示命令で指定されるグループ ID を当該スプリット指示データのグループ ID に設定する。スプリット指示記憶部 1307 は、当該グループ ID において前回付与されたシーケンス ID に 1 加算し、当該スプリット指示データのシーケンス ID に設定する。スプリット指示記憶部 1307 は、計時機構 1204 から現在の日時を取得し、当該スプリット指示データの更新時刻に設定する。スプリット指示記憶部 1307 は、スプリット指示命令を受信したことを示すデータを当該スプリット指示データの更新位

10

20

30

40

50

置情報に設定する。

そして、未反映情報記憶部 1309 は、更新時刻がスプリット時刻より早いジャーナルデータ 1701 の更新位置情報である未反映情報をメモリ 1003 に記憶する。なお、更新時刻がスプリット時刻より早いかどうかの判断は、更新時刻を直接比較することにより行われてもよいし、シーケンス ID を比較することにより行われてもよい。

【0082】

== スプリット解除 ==

図 22 は、ディスク制御装置 110 がスプリット解除命令を受信した際の処理を示すフローチャートである。

スプリット解除命令受信部 1306 は、データベースサーバ 40 やバックアップサーバ 1801 等からスプリット解除命令を受信する (S2201)。スプリット解除記憶部 1308 は、スプリット解除データを第三の記憶デバイス 1203 に書き込む (S2202)。つまり、スプリット解除記憶部 1308 は、スプリット指示命令で指定されるグループ ID が設定されているスプリット指示データの更新位置情報にスプリット解除命令を受信したことを示すデータを設定する。

【0083】

== 第二の記憶デバイスへの更新データの反映 ==

図 23 は、ディスク制御装置 110 が第三の記憶デバイス 1203 に格納されている更新データを第二の記憶デバイス 1202 に反映する処理を示すフローチャートである。

第二の書き込み部 1304 は、メモリ 1003 に格納されている検出時間 1205 を取得する (S2301)。第二の書き込み部 1304 は、計時機構 1204 から日付と時刻 (現在時刻) を取得する (S2302)。第二の書き込み部 1304 は、第三の記憶デバイス 1203 に格納されているジャーナルデータ 1701 を参照し (S2303)、スプリット指示データがあり、かつ、スプリット解除データがないかどうかを確認する (S2304)。

【0084】

スプリット指示データがあり、かつ、スプリット解除データがない場合、第二の書き込み部 1304 は、当該スプリット指示データとグループ ID が同一で、更新時刻と現在時刻との差が検出時間 1205 を超えており、更新時刻がスプリット時刻より早いジャーナルデータを選択する (S2305)。条件に該当するジャーナルデータ 1701 がない場合は、再び S2302 からの処理を実行する。なお、更新時刻がスプリット時刻より早いジャーナルの選択は、更新時刻を比較することにより行われてもよいし、シーケンス ID を比較することにより行われてもよい。

スプリット指示データがないか、またはスプリット解除データがある場合、第二の書き込み部 1304 は、更新時刻と現在時刻との差が検出時間 1205 を超えているジャーナルデータ 1701 を選択する (S2306)。条件に該当するジャーナルデータ 1701 がない場合は、再び S2302 からの処理を実行する。

【0085】

第二の書き込み部 1304 は、選択したジャーナルデータ 1701 の更新位置情報に設定されている論理ボリューム 121 の識別子に対応する副論理ボリュームの識別子をペア管理テーブル 1501 から取得する (S2307)。第二の書き込み部 1304 は、選択したジャーナルデータ 1701 の更新データを、取得した副論理ボリュームの識別子で示される第二の記憶デバイス 1202 の論理ボリューム 121 のジャーナルデータ 1701 の位置情報で示される場所書き込む (S2308)。なお、ジャーナルデータ 1701 が複数ある場合、第二の書き込み部 1304 は、シーケンス ID 順、つまり更新時刻の早い順に更新データを第二の記憶デバイス 1202 の論理ボリューム 121 に書き込む。第二の記憶デバイス 1202 の論理ボリューム 121 への書き込みが完了すると、第二の書き込み部 1304 は、メモリ 1003 に記憶されている当該更新データの未反映情報を削除する (S2309)。

【0086】

10

20

30

40

50

なお、前述のS 2 3 0 5のステップを、第二の書き込み部1 3 0 4が更新時刻と現在時刻との比較を行わず、更新時刻がスプリット時刻より早いジャーナルデータを選択すると変更することにより、仮想デバイス1 8 0 1を設けずに第二の記憶デバイス1 2 0 2を第一の記憶デバイス1 2 0 1の静止化データとすることができる。

【0 0 8 7】

= = 仮想論理ボリュームの読み出し = =

図2 4は、ディスク制御装置1 1 0が仮想デバイス1 8 0 1の仮想論理ボリュームに対する読み出し要求を受信した際の処理を示すフローチャートである。

【0 0 8 8】

読み出し要求受信部1 3 1 0は、バックアップサーバ1 8 0 2から仮想デバイス1 8 0 1の仮想論理ボリュームの識別子が指定された読み出し要求を受信する(S 2 4 0 1)。識別子取得部1 3 1 1は、読み出し要求で指定された仮想論理ボリュームの識別子に対応する論理ボリューム1 2 1の識別子を仮想論理ボリューム管理テーブル1 9 0 1から取得する(S 2 4 0 2)。重複範囲取得部1 3 1 2は、読み出し要求で指定される位置情報とメモリ1 0 0 3に記憶されている未反映情報の位置情報とで重複している範囲を取得する(S 2 4 0 3)。未反映情報が開始アドレスとブロック長とで示されている場合、重複している範囲とは、読み出し要求で指定される位置情報で示される範囲と未反映情報の位置情報で示される範囲とで重なりあう部分である。また、未反映情報がセグメント単位でのビットマップで表現されている場合、重複している範囲とは、読み出し要求で指定される位置情報で示される範囲のうちで未反映情報のビットマップで未反映のビットが設定されているセグメントである。

【0 0 8 9】

仮想論理ボリューム読み出し部1 3 1 3は、重複している範囲については第三の記憶デバイス1 2 0 3に格納されているジャーナルデータ1 7 0 1から読み出し(S 2 4 0 4)、重複していない範囲については第二の記憶デバイス1 2 0 2の論理ボリューム1 2 1から読み出す(S 2 4 0 5)。そして、読み出しデータ送信部1 3 1 4は、読み出されたデータをバックアップサーバ1 8 0 2に送信する(S 2 4 0 6)。

【0 0 9 0】

以上、本実施の形態におけるバックアップ方式を実現するための、ディスク制御装置1 1 0の各部1 3 0 1 ~ 1 3 1 4の動作を説明した。

【0 0 9 1】

これにより、第二の記憶デバイス1 2 0 2の論理ボリューム1 2 1を第一の記憶デバイス1 2 0 1の論理ボリューム1 2 1の検出時間1 2 0 5前の状態とすることができる。ソフトウェアの障害等により不正なデータが第一の記憶デバイス1 2 0 1の論理ボリューム1 2 1に書き込まれた場合、検出時間1 2 0 5の間は、第二の記憶デバイス1 2 0 2の論理ボリューム1 2 1には不正なデータが書き込まれない。

【0 0 9 2】

例えば、情報処理装置2 0がデータベースサーバ4 0であり、第一の記憶デバイス1 2 0 1の論理ボリューム1 2 1がデータベースのデータ格納領域であるとする。このような場合、データベースサーバ4 0は第一の記憶デバイス1 2 0 1の論理ボリューム1 2 1に書き込んだデータに対するある時点からのR E D Oログを別の記憶領域に記憶している。したがって、第一の記憶デバイス1 2 0 1の論理ボリューム1 2 1に不正なデータが書き込まれた場合、第二の記憶デバイス1 2 0 2の論理ボリューム1 2 1に記憶されている検出時間1 2 0 5前のデータとR E D Oログとを用いて、第一の記憶デバイス1 2 0 1の論理ボリューム1 2 1を不正なデータが書き込まれる直前の状態にまで復元することができる。つまり、磁気テープ等の媒体に保存されているデータを用いることなく、ディスクアレイ装置1 0内のデータを用いて第一の記憶デバイス1 2 0 1の論理ボリューム1 2 1のデータを復元することができるため、復元作業に要する時間を短縮することが可能である。

【0 0 9 3】

10

20

30

40

50

また、第一の記憶デバイス１２０１の論理ボリューム１２１を主論理ボリュームとし、主論理ボリュームに対する副論理ボリューム（レプリカボリューム）を２つ設ける従来のデータバックアップ方式と比較して、記憶容量を削減することが可能である。副論理ボリュームを２つ設ける従来のデータバックアップ方式では、主論理ボリュームと副論理ボリュームとで、第一の記憶デバイス１２０１の３倍の記憶容量が必要であった。本発明の方式で必要となる記憶容量は、第三の記憶デバイス１２０３に格納するジャーナルデータの量に依存している。通常、１日のデータベース更新量は主論理ボリュームの２０％までであることが多い。つまり、検出時間１２０５を１日とする場合、第三の記憶デバイス１２０３の記憶容量は第一の記憶デバイス１２０１の２５％を確保すれば十分である。したがって、第一の記憶デバイス１２０１と第二の記憶デバイス１２０２と第三の記憶デバイス１２０３とで、第一の記憶デバイス１２０１の２．２５倍の記憶容量でよいこととなる。検出時間１２０５を数時間とすることができる場合は、必要な記憶容量を更に削減することが可能である。つまり、障害が発生した場合にディスクアレイ装置１０内に存在するデータで第一の記憶デバイス１２０１の論理ボリューム１２１を復旧するバックアップ方式において、必要な記憶容量を削減でき、ディスクアレイ装置１０の導入コストを削減することができる。

10

【００９４】

また、検出時間１２０５がメモリ１００３に記憶されていることにより、検出時間１２０５を管理端末１３６や情報処理装置２０等から登録することが可能となる。つまり、ソフトウェアの不具合や人為的な操作ミス等により発生する第一の記憶デバイス１２０１の論理ボリューム１２１に不正なデータが書き込まれていることを検出可能な時間を、業務やソフトウェアの特性に応じて変更することが可能である。したがって、第三の記憶デバイス１２０３にジャーナルデータ１７０１を格納するために必要な記憶容量を、業務やソフトウェアの特性に応じて変更することができる。

20

【００９５】

また、第一の記憶デバイス１２０１を構成するハードディスクドライブ１２０と第二の記憶デバイス１２０２を構成するハードディスクドライブ１２０とを物理的に別のものとすることができる。第一の記憶デバイス１２０１を構成するハードディスクドライブ１２０にハードウェア障害が発生した場合でも、第二の記憶デバイス１２０２には、検出時間１２０５前に第一の記憶デバイス１２０１の論理ボリューム１２１に記憶されていたデータが記憶されている。情報処理装置２０がデータベースサーバ４０で、ディスクアレイ装置１０内に第一の記憶デバイス１２０１の論理ボリューム１２１に書き込まれたＲＥＤＯログが存在する場合、第二の記憶デバイス１２０２の論理ボリューム１２１とＲＥＤＯログとで、第一の記憶デバイス１２０１の論理ボリューム１２１のデータを復元することができる。データの復元には、磁気テープ等の媒体に保存されているデータを用いないため、復元作業に要する時間を短縮することが可能である。

30

【００９６】

つまり、第一の記憶デバイス１２０１の論理ボリューム１２１のある時点の静止化データを用いることなく、継続的に更新が行われている第二の記憶デバイス１２０１を用いて第一の記憶デバイス１２０１の論理ボリューム１２１のデータを迅速に復旧することができる。

40

【００９７】

また、第一の記憶デバイス１２０１を構成するハードディスクドライブ１２０と第二の記憶デバイス１２０２及び第三の記憶デバイス１２０３を構成するハードディスクドライブ１２０とを物理的に別のものとすることができる。第三の記憶デバイス１２０３に格納されているジャーナルデータ１７０１を用いて第二の記憶デバイス１２０２の論理ボリューム１２１を更新する際に、第一の記憶デバイス１２０１からデータを読み出す必要がない。つまり、情報処理装置２０から第一の記憶デバイス１２０１に対するデータ入出力処理の性能に与える影響が少ない。

【００９８】

50

また、第一の記憶デバイス1201を構成するハードディスクドライブ120にハードウェア障害が発生した場合、第二の記憶デバイス1202の論理ボリューム121と第三の記憶デバイス1203に格納されているジャーナルデータ1701とを用いて第一の記憶デバイス1201の論理ボリューム121のデータを復元することができる。この場合においても、データの復元には磁気テープ等の媒体に保存されているデータを用いないため、復元作業に要する時間を短縮することが可能である。

【0099】

また、スプリット時刻から検出時間を経過後、第一の記憶デバイス1201のスプリット時刻の状態を第二の記憶デバイス1202上に作成することが可能となる。この状態において、第二の記憶デバイス1202の論理ボリューム121を磁気テープ等の記憶媒体にバックアップすることが可能となる。したがって、ディスクアレイ装置10全体の障害や広域災害等に備えたデータのバックアップを行うことが可能となる。

【0100】

また、スプリット指示後直ちに、第二の記憶デバイス1202の論理ボリューム121の仮想論理ボリュームを用いて、第一の記憶デバイス1201の論理ボリューム121のスプリット時刻におけるデータを磁気テープ等にバックアップすることが可能である。つまり、第三の記憶デバイス1203に格納されている更新時刻がスプリット時刻より早いジャーナルデータが第二の記憶デバイス1202の論理ボリューム121に反映されるのを待つ必要がない。そのため、データのバックアップを容易に行うことが可能となる。

【0101】

また、スプリット指示命令を受信すると、検出時間にかかわらず、更新時刻がスプリット時刻より早いジャーナルデータの更新データを第二の記憶デバイス1202の論理ボリューム121に反映することもできる。これにより、第二の記憶デバイス1202の論理ボリューム121を用いて、第一の記憶デバイス1201の論理ボリューム121のスプリット時刻におけるデータを磁気テープ等にバックアップすることが可能である。つまり、バックアップ処理中に仮想論理ボリュームを形成する必要がないため、ディスクアレイ装置10の負荷が軽減され、バックアップ処理時間が短縮される。

【0102】

また、スプリット指示命令を第一の記憶デバイス1201の論理ボリューム121のグループIDを指定して行うことができる。第一の記憶デバイス1201を、複数の情報処理装置20が使用している場合や、情報処理装置10上で複数の業務アプリケーションが稼働している場合等に、情報処理装置20や業務アプリケーションごとに論理ボリューム121にグループIDを付与することで、グループID単位でデータのバックアップを行うことが可能となる。

【0103】

例えば、情報処理装置20がデータベースサーバであり、データベースサーバが使用しているデータベース・インスタンスが2つあるとする。各データベース・インスタンスがデータ格納領域として用いている第一の記憶デバイス1201の論理ボリューム121に、異なるグループIDを付与する。スプリット指示命令及びスプリット解除命令は、このグループID単位で行われる。つまり、データベース・インスタンス単位でデータのバックアップを行うことが可能となる。また、例えば、業務アプリケーションごとに使用する論理ボリューム121が分かれている場合、業務アプリケーション単位でのデータのバックアップを行うことが可能となる。

【0104】

また、グループIDを考慮する場合においても、当該グループIDに対する仮想論理ボリュームを用いて、第一の記憶デバイス1201の論理ボリューム121のスプリット時刻におけるデータを磁気テープ等にバックアップすることが可能である。つまり、バックアップを開始する前に、第三の記憶デバイスに格納されている更新時刻がスプリット時刻より早いジャーナルデータが第二の記憶デバイスの論理ボリュームに反映されるのを待つ必要がない。そのため、データのバックアップを容易に行うことが可能となる。

【 0 1 0 5 】

また、グループIDを考慮する場合においても、スプリット指示命令を受信すると、当該グループIDについては検出時間にかかわらず更新時刻がスプリット時刻より早いジャーナルデータの更新データを第二の記憶デバイス1202の論理ボリューム121に反映することもできる。これにより、第二の記憶デバイス1202の論理ボリューム121を用いて、第一の記憶デバイス1201の論理ボリューム121のスプリット時刻におけるデータを磁気テープ等にバックアップすることが可能である。つまり、バックアップ処理中に仮想論理ボリュームを形成する必要がないため、ディスクアレイ装置10の負荷が軽減され、バックアップ処理時間が短縮される。

【 0 1 0 6 】

以上、本実施の形態について説明したが、上記実施例は本発明の理解を容易にするためのものであり、本発明を限定して解釈するためのものではない。本発明は、その趣旨を逸脱することなく、変更、改良され得ると共に、本発明にはその等価物も含まれる。

【図面の簡単な説明】

【 0 1 0 7 】

【図1】本実施の形態に係るディスクアレイ装置の第一の形態の構成を示すブロック図である。

【図2】本実施の形態に係るチャネル制御部の構成を示すブロック図である。

【図3】本実施の形態に係るディスク制御部の構成を示すブロック図である。

【図4】本実施の形態に係るディスクアレイ装置の第二の形態の構成を示すブロック図である。

【図5】本実施の形態に係るクライアント端末、データベースサーバ、ディスクアレイ装置で構成されるデータベースシステムを示す図である。

【図6】本実施の形態に係るクライアント端末の構成を示すブロック図である。

【図7】本実施の形態に係るデータベースサーバの構成を示すブロック図である。

【図8】本実施の形態に係るデータベースシステムにおけるテーブル更新処理を示すフローチャートである。

【図9】REDOログを用いてデータを復旧する一般的な手順を示す図である。

【図10】レプリカボリュームを用いてデータのバックアップを行う方式を示す図である。

【図11】スナップショットを用いてデータのバックアップを行う方式を示す図である。

【図12】本実施の形態に係るバックアップ方式を示す図である。

【図13】本実施の形態に係るバックアップ方式を実現する機能を示すブロック図である。

【図14】本実施の形態に係る第一の記憶デバイスの論理ボリュームと第二の記憶デバイスの論理ボリュームとの関係を示す図である。

【図15】本実施の形態に係るペア管理テーブルを示す図である。

【図16】本実施の形態に係るグループ管理テーブルを示す図である。

【図17】本実施の形態に係るジャーナルデータを示す図である。

【図18】本実施の形態に係るスプリット指示命令を受信した場合のディスクアレイ装置の状態遷移を示す図である。

【図19】本実施の形態に係る仮想論理ボリューム管理テーブルを示す図である。

【図20】本実施の形態に係るディスク制御装置が第一の記憶デバイスの論理ボリュームに対する書き込み要求を受信した際の処理を示すフローチャートである。

【図21】本実施の形態に係るディスク制御装置がスプリット指示命令を受信した際の処理を示すフローチャートである。

【図22】本実施の形態に係るディスク制御装置がスプリット解除命令を受信した際の処理を示すフローチャートである。

【図23】本実施の形態に係るディスク制御装置がジャーナルデータを第二の記憶デバイスに反映する処理を示すフローチャートである。

10

20

30

40

50

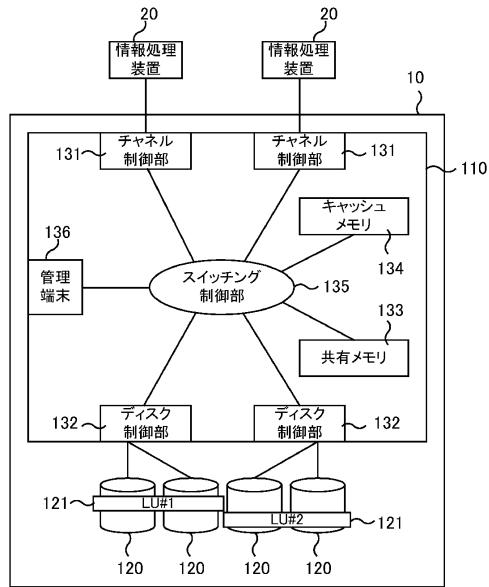
【図 2 4】本実施の形態に係るディスク制御装置が仮想論理ボリュームに対する読み出し要求を受信した際の処理を示すフローチャートである。

【符号の説明】

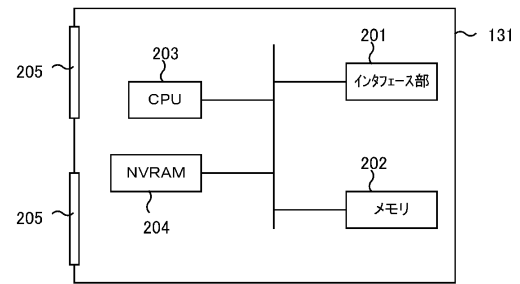
【 0 1 0 8 】

1 0	ディスクアレイ装置	2 0	情報処理装置	
3 0	クライアント端末	4 0	データベースサーバ	
1 1 0	ディスク制御装置	1 2 0	ハードディスクドライブ	
1 2 1	論理ボリューム	1 3 1	チャネル制御部	
1 3 2	ディスク制御部	1 3 3	共有メモリ	
1 3 4	キャッシュメモリ	1 3 5	スイッチング制御部	10
1 3 6	管理端末	1 4 1	C P U	
1 4 2	メモリ	1 4 3	ホストインタフェース	
1 4 4	ディスクインタフェース	1 4 5	キャッシュメモリ	
1 4 6	データコントローラ	2 0 1	インタフェース部	
2 0 2	メモリ	2 0 3	C P U	
2 0 4	N V R A M	2 0 5	コネクタ	
3 0 1	インタフェース部	3 0 2	メモリ	
3 0 3	C P U	3 0 4	N V R A M	
3 0 5	コネクタ	5 0 1	データ領域	
5 0 2	R E D O ログ領域	5 0 3	テーブル	20
5 0 4	テーブル	5 0 5	データベースバッファ	
5 0 6	R E D O ログバッファ	6 0 1	C P U	
6 0 2	メモリ	6 0 3	記憶装置	
6 0 4	ポート	6 0 5	記録媒体読取装置	
6 0 6	入力装置	6 0 7	出力装置	
6 0 8	記録媒体	7 0 1	C P U	
7 0 2	メモリ	7 0 3	記憶装置	
7 0 4	L A N インタフェース	7 0 5	ストレージインタフェース	
7 0 6	記録媒体読取装置	7 0 7	記録媒体	
1 0 0 1	レプリカボリューム	1 0 0 2	レプリカボリューム	30
1 0 0 3	メモリ	1 0 0 4	差分セグメントビットマップ	
1 1 0 1	プール	1 1 0 2	仮想レプリカボリューム	
1 2 0 1	第一の記憶デバイス	1 2 0 2	第二の記憶デバイス	
1 2 0 3	第三の記憶デバイス	1 2 0 4	計時機構	
1 2 0 5	検出時間	1 3 0 1	書き込み要求受信部	
1 3 0 2	第一の書き込み部	1 3 0 3	ジャーナル書き込み部	
1 3 0 4	第二の書き込み部	1 3 0 5	スプリット指示命令受信部	
1 3 0 6	スプリット解除命令受信部	1 3 0 7	スプリット指示記憶部	
1 3 0 8	スプリット解除記憶部	1 3 0 9	未反映情報記憶部	
1 3 1 0	読み出し要求受信部	1 3 1 1	識別子取得部	40
1 3 1 2	重複範囲取得部	1 3 1 3	仮想論理ボリューム読み出し部	
1 3 1 4	読み出しデータ送信部	1 5 0 1	ペア管理テーブル	
1 6 0 1	グループ管理テーブル	1 7 0 1	ジャーナルデータ	
1 7 0 2	ジャーナル部	1 7 0 3	更新ストリーム部	
1 9 0 1	仮想論理ボリューム管理テーブル			

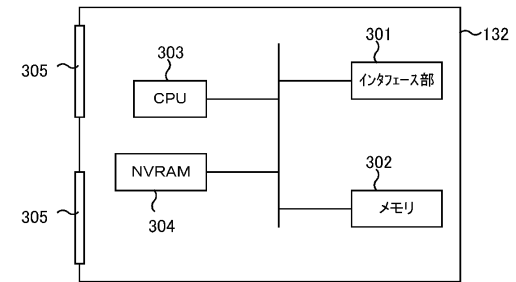
【図 1】



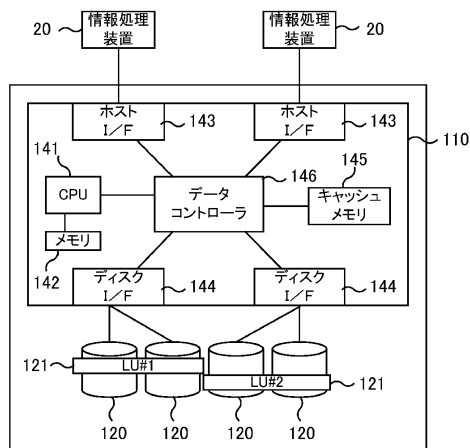
【図 2】



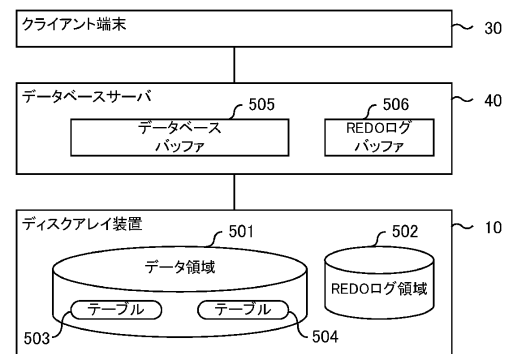
【図 3】



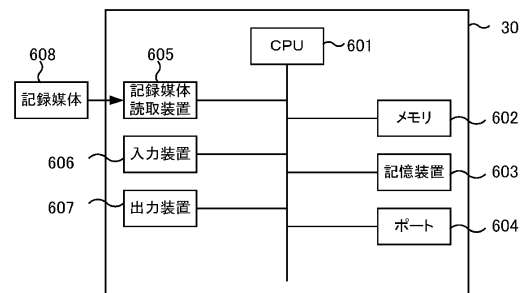
【図 4】



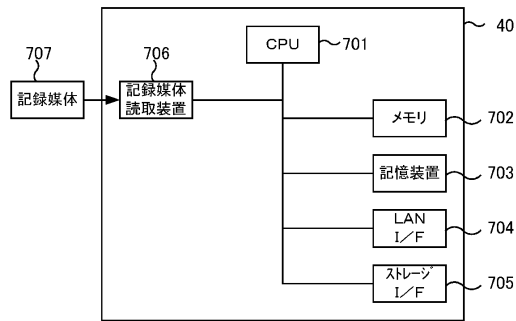
【図 5】



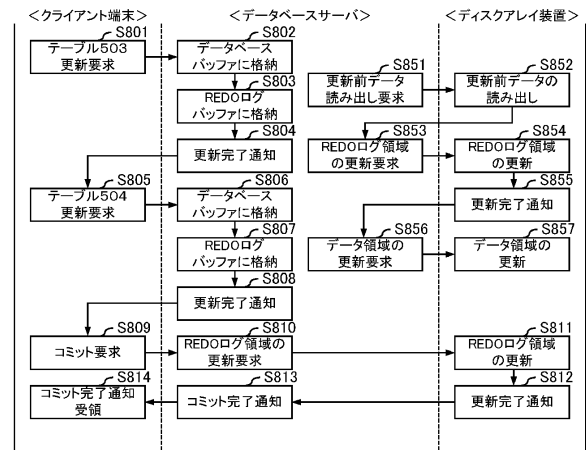
【図 6】



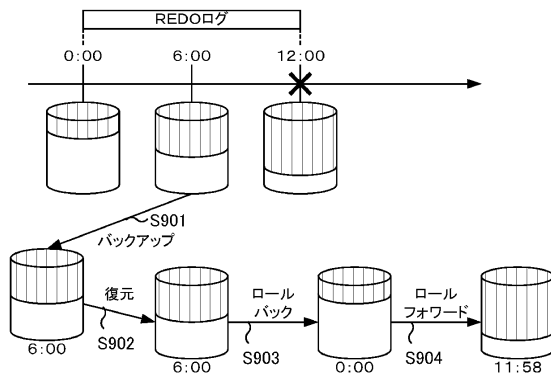
【図 7】



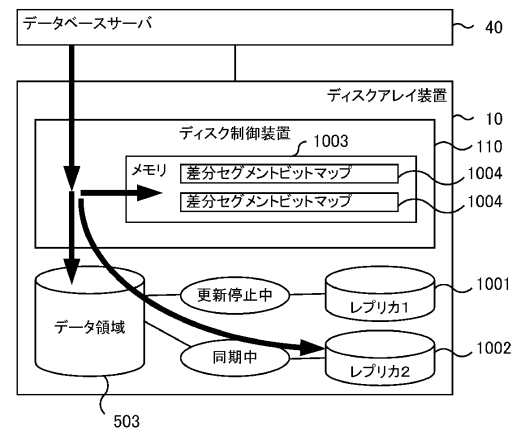
【図 8】



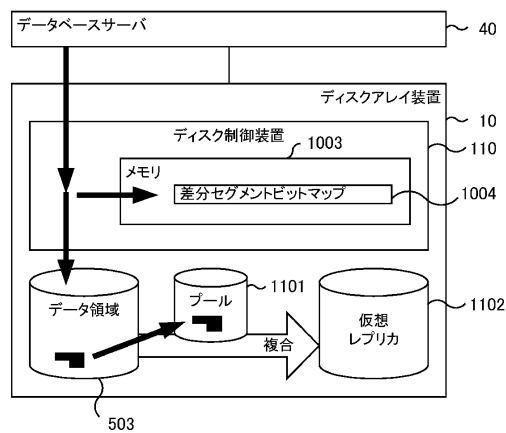
【図 9】



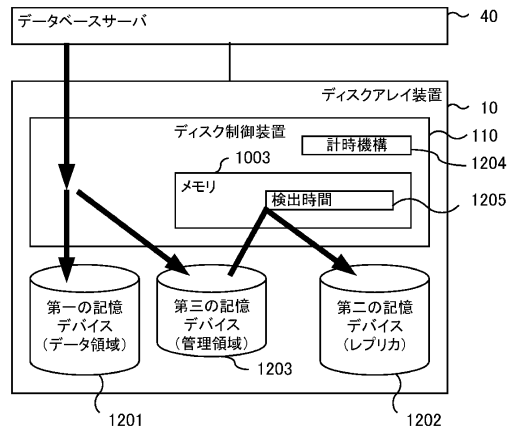
【図 10】



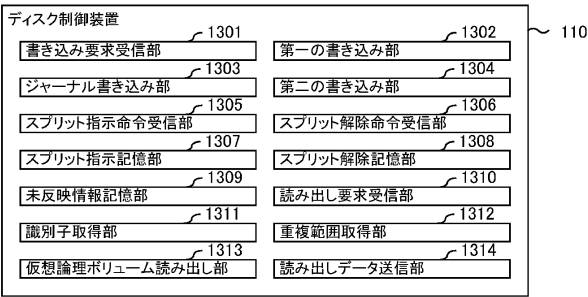
【図 1 1】



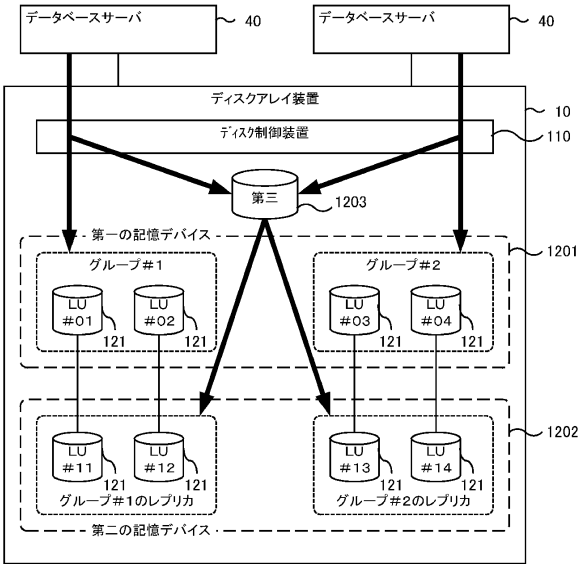
【図 1 2】



【図 1 3】



【図 1 4】



【図 1 5】

主論理ボリューム	副論理ボリューム
LU #01	LU #11
LU #02	LU #12
LU #03	LU #13
LU #04	LU #14
⋮	⋮

1501

【図 1 6】

論理ボリューム	グループID
LU #01	1
LU #02	1
LU #03	2
LU #04	2
⋮	⋮

1601

【図 17】

1701

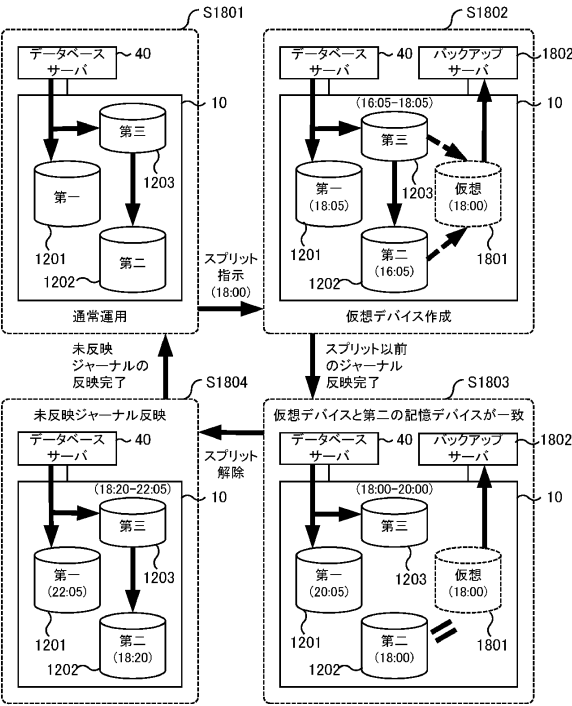
ジャーナル部				
グループID	シーケンスID	更新時刻	更新位置情報	更新ストリームオフセット
1	10011	030115 17:59:55	LU#1/開始LBA10000/Block長8	2048
1	10012	030115 17:59:58	LU#2/開始LBA30000/Block長8	2056
2	00058	030115 17:59:59	LU#4/開始LBA10000/Block長8	2064
1	10013	030115 18:00:01	スプリット指示/スプリット解除	—
2	00059	030115 18:00:02	LU#3/開始LBA10000/Block長8	2072
1	10014	030115 18:00:03	LU#1/開始LBA10000/Block長8	2080

1702

更新ストリーム部		
0	(蓄積更新ブロック:1024Block分)	1023
	(蓄積更新ブロック:1024Block分)	2047
グループID#1 シーケンスID#10011 の更新データ	グループID#1 シーケンスID#10012 の更新データ	グループID#2 シーケンスID#00058 の更新データ
		3071

1703

【図 18】

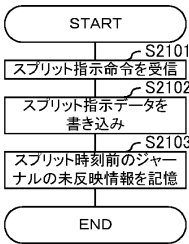


【図 19】

論理ボリューム	仮想論理ボリューム
LU#11	LU#21
LU#12	LU#22
LU#13	LU#23
LU#14	LU#24
⋮	⋮

1901

【図 21】



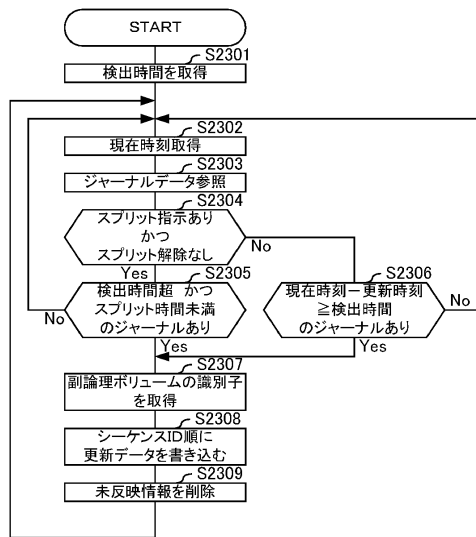
【図 20】



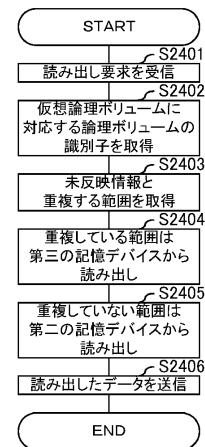
【図 22】



【図 23】



【図 24】



フロントページの続き

(72)発明者 岸本 哲哉

神奈川県小田原市中里322番2号 株式会社日立製作所 RAIDシステム事業部内

審査官 上嶋 裕樹

(56)参考文献 特開平04-124743(JP,A)
特開2003-167684(JP,A)
特開2002-278819(JP,A)
特開平07-262070(JP,A)
特開2001-209565(JP,A)
特開2000-137638(JP,A)
特開2001-312421(JP,A)

(58)調査した分野(Int.Cl., DB名)

G06F 12/00

G06F 3/06