(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4838878号
(P4838878)
(45)発行日　平成23年12月14日(2011.12.14)　(24)登録日　平成23年10月7日(2011.10.7)

(51)Int.Cl.　F I
G06F 12/00 (2006.01)　G06F 12/00 501B
G06F 12/00 545A
G06F 12/00 520E

請求項の数 9　(全 54 頁)

(21)出願番号　特願2009-276025 (P2009-276025)
(22)出願日　平成21年12月4日 (2009.12.4)
(65)公開番号　特開2011-118712 (P2011-118712A)
(43)公開日　平成23年6月16日 (2011.6.16)
審査請求日　平成23年2月17日 (2011.2.17)

(73)特許権者　000005223
富士通株式会社
神奈川県川崎市中原区上小田中４丁目１番１号
(74)代理人　100092152
弁理士　服部　毅巖
(72)発明者　田村　雅寿
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内
(72)発明者　野口　泰生
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内
(72)発明者　荻原　一隆
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内
最終頁に続く

(54)【発明の名称】データ管理プログラム、データ管理装置、およびデータ管理方法

(57)【特許請求の範囲】
【請求項１】
複数のディスクノードそれぞれで管理されるストレージ装置にデータを分散格納するマルチノードストレージシステムを構成する１つの前記ディスクノードで実行すべきデータ管理処理をコンピュータに実行させるデータ管理プログラムにおいて、
前記コンピュータに、
仮想的な記憶領域を定義した論理ボリュームを複数に分割して得られるデータユニットを指定した書き込み要求があると、前記コンピュータに接続されたストレージ装置内の単位記憶領域の１つを指定された前記データユニットに割り当て、前記データユニットに割り当てられた前記単位記憶領域に対してデータの書き込みを行い、
書き込み対象となった前記データユニットに対応付けて、現在の時刻を最終書き込み時刻としてデータユニット情報記憶手段に格納し、
前記データユニット情報記憶手段を参照し、前記最終書き込み時刻から重複排除化猶予期間以上経過している前記データユニットを検出し、
仮想的な記憶領域を定義した重複排除ボリュームを複数に分割して得られる重複排除ユニットのうち重複排除化の対象とされた重複排除対象データの格納に使用されている重複排除ユニットの識別情報および前記重複排除ユニットを管理している前記ディスクノードの識別子を含む重複排除アドレスと、前記重複排除対象データに所定の演算を行うことで得られる前記重複排除対象データの固有値とを対応付けて管理するインデックスサーバから、検出された前記データユニットに割り当てられた前記単位記憶領域内のデータの前記

固有値に対応付けられた前記重複排除アドレスを取得し、
検出された前記データユニットに対応付けて、取得した前記重複排除アドレスを前記データユニット情報記憶手段に格納すると共に、検出された前記データユニットへの前記単位記憶領域の割り当てを解除する、
処理を実行させることを特徴とするデータ管理プログラム。
【請求項２】
前記コンピュータに、さらに、
前記インデックスサーバから、前記重複排除ユニットの識別情報を指定した重複排除ユニット割当要求を受信すると、指定された前記重複排除ユニットに前記単位記憶領域の１つを割り当て、指定された前記重複排除ユニットをデータの読み出し先とする前記ディスクノードから前記重複排除対象データを取得し、指定された前記重複排除ユニットに割り当てた前記単位記憶領域に格納し、
前記重複排除ユニットの識別情報を指定したデータの読み出し要求を受信すると、指定された前記重複排除ユニットに割り当てられた前記単位記憶領域内の前記重複排除対象データを応答する、
処理を実行させることを特徴とする請求項１記載のデータ管理プログラム。
【請求項３】
前記コンピュータに、さらに、
前記重複排除ユニットを前記重複排除対象データの読み出し先とする前記ディスクノードが存在しなくなった場合、前記重複排除ユニットへの前記単位記憶領域の割り当てを解除する、
処理を実行させることを特徴とする請求項２記載のデータ管理プログラム。
【請求項４】
前記コンピュータにさらに、
前記重複排除ユニットに対応付けて、前記重複排除ユニットに割り当てられた前記単位記憶領域から最後にデータの読み出しが行われた時刻から保存期間経過後の保存期限満了時刻を重複排除ユニット情報記憶手段に格納し、
前記重複排除ユニット情報記憶手段を参照し、現在の時刻が前記保存期限満了時刻を過ぎている前記重複排除ユニットを検出し、検出した前記重複排除ユニットへの前記単位記憶領域の割り当てを解除する、
処理を実行させることを特徴とする請求項３記載のデータ管理プログラム。
【請求項５】
前記コンピュータに、さらに、
前記データユニットを指定した読み出し要求があると、読み出し対象の前記データユニットに割り当てられた前記単位記憶領域がない場合、前記データユニット情報記憶手段を参照し、読み出し対象の前記データユニットに対応付けられた前記重複排除アドレスに示される前記ディスクノードから、前記重複排除アドレスに示される前記重複排除ユニットの識別情報を指定して前記重複排除対象データを取得する、
処理を実行させることを特徴とする請求項１記載のデータ管理プログラム。
【請求項６】
複数のディスクノードそれぞれで管理されるストレージ装置にデータを分散格納するマルチノードストレージシステムの記憶領域管理処理をコンピュータに実行させるデータ管理プログラムにおいて、
前記コンピュータに、
仮想的な記憶領域を定義した重複排除ボリュームを複数に分割して得られる重複排除ユニットのうち前記ディスクノードで管理している前記重複排除ユニットの使用の有無、使用されている前記重複排除ユニットに割り当てられている単位記憶領域に格納された重複排除対象データに所定の演算を行うことで得られる固有値、および前記重複排除ユニットの識別情報を含む重複排除ユニット情報が、前記重複排除ユニットを管理している前記ディスクノードの識別子に対応付けて重複排除ユニット情報記憶手段に格納されており、

前記ディスクノードから、重複排除化データユニットに割り当てられた単位記憶領域内のデータに前記所定の演算を実行して得られた固有値を含む重複排除アドレス照会要求を受け取ると、前記重複排除ユニット情報記憶手段から、前記重複排除アドレス照会要求に示される固有値を含む前記重複排除ユニット情報を検索し、

検索により該当する前記重複排除ユニット情報が検出された場合、検出された前記重複排除ユニット情報に含まれる前記重複排除ユニットの識別情報と、検出された前記重複排除ユニット情報に対応する前記重複排除ユニットを管理している前記ディスクノードの識別子とを含む重複排除アドレスを、前記重複排除アドレス照会要求の送信元である前記ディスクノードに応答し、

検索により該当する前記重複排除ユニット情報が検出されなかった場合、前記重複排除ユニット情報記憶手段を参照し、未使用の前記重複排除ユニットを選択し、選択した前記重複排除ユニットを管理する前記ディスクノードに対して選択した重複排除ユニットへの前記単位記憶領域の割り当て要求を送信し、選択した前記重複排除ユニットの前記単位記憶領域割り当て後の前記重複排除ユニット情報を前記重複排除ユニット情報記憶手段に格納し、選択した前記重複排除ユニットの識別情報と、選択した前記重複排除ユニットを管理する前記ディスクノードの識別子とを含む重複排除アドレスを、前記重複排除アドレス照会要求の送信元である前記ディスクノードに応答する、

処理を実行させることを特徴とするデータ管理プログラム。

【請求項７】

前記コンピュータに、さらに、

前記重複排除ユニット情報には、前記重複排除ユニットに割り当てられた前記単位記憶領域から最後にデータの読み出しが行われた時刻から保存期間経過後の保存期限満了時刻が含められており、前記重複排除ユニット情報記憶手段を参照し、現在の時刻が前記保存期限満了時刻を過ぎている前記重複排除ユニットを検出し、

検出された前記重複排除ユニット情報の取得元の前記ディスクノードから最新の前記重複排除ユニット情報を取得し、前記重複排除ユニット情報記憶手段に反映する、

処理を実行させることを特徴とする請求項６記載のデータ管理プログラム。

【請求項８】

複数のディスクノードそれぞれで管理されるストレージ装置にデータを分散格納するマルチノードストレージシステムを構成する１つの前記ディスクノードで実行すべきデータ管理処理をコンピュータによって実現するデータ管理装置において、

仮想的な記憶領域を定義した論理ボリュームを複数に分割して得られるデータユニットを指定した書き込み要求があると、前記コンピュータに接続されたストレージ装置内の単位記憶領域の１つを指定された前記データユニットに割り当て、前記データユニットに割り当てられた前記単位記憶領域に対してデータの書き込みを行う書き込みアクセス手段と、

書き込み対象となった前記データユニットに対応付けて、現在の時刻を最終書き込み時刻としてデータユニット情報記憶手段に格納する最終書き込み時刻更新手段と、

前記データユニット情報記憶手段を参照し、最終書き込み時刻から重複排除化猶予期間以上経過している前記データユニットを検出する重複排除化データユニット検出手段と、

仮想的な記憶領域を定義した重複排除ボリュームを複数に分割して得られる重複排除ユニットのうち重複排除化の対象とされた重複排除対象データの格納に使用されている重複排除ユニットの識別情報および前記重複排除ユニットを管理している前記ディスクノードの識別子を含む重複排除アドレスと、前記重複排除対象データに所定の演算を行うことで得られる前記重複排除対象データの固有値とを対応付けて管理するインデックスサーバから、検出された前記データユニットに割り当てられた前記単位記憶領域内のデータの前記固有値に対応付けられた前記重複排除アドレスを取得する重複排除アドレス取得手段と、

検出された前記データユニットに対応付けて、取得した前記重複排除アドレスを前記データユニット情報記憶手段に格納すると共に、検出された前記データユニットへの前記単位記憶領域の割り当てを解除する単位記憶領域割当解除手段と、

前記データユニットを指定した読み出し要求があると、読み出し対象の前記データユニットに割り当てられた前記単位記憶領域がある場合、割り当てられた前記単位記憶領域からデータを読み出し、読み出し対象の前記データユニットに割り当てられた前記単位記憶領域がない場合、前記データユニット情報記憶手段を参照し、読み出し対象の前記データユニットに対応付けられた前記重複排除アドレスに示される前記ディスクノードから、前記重複排除アドレスに示される前記重複排除ユニットの識別情報を指定して前記重複排除対象データを取得する読み出しアクセス手段と、

を有することを特徴とするデータ管理装置。

【請求項９】

複数のディスクノードそれぞれで管理されるストレージ装置にデータを分散格納するマルチノードストレージシステムを構成する１つの前記ディスクノードで実行すべきデータ管理処理をコンピュータで実行するデータ管理方法において、

前記コンピュータが、

仮想的な記憶領域を定義した論理ボリュームを複数に分割して得られるデータユニットを指定した書き込み要求があると、前記コンピュータに接続されたストレージ装置内の単位記憶領域の１つを指定された前記データユニットに割り当て、前記データユニットに割り当てられた前記単位記憶領域に対してデータの書き込みを行い、

書き込み対象となった前記データユニットに対応付けて、現在の時刻を最終書き込み時刻としてデータユニット情報記憶手段に格納し、

前記データユニット情報記憶手段を参照し、最終書き込み時刻から重複排除化猶予期間以上経過している前記データユニットを検出し、

仮想的な記憶領域を定義した重複排除ボリュームを複数に分割して得られる重複排除ユニットのうち重複排除化の対象とされた重複排除対象データの格納に使用されている重複排除ユニットの識別情報および前記重複排除ユニットを管理している前記ディスクノードの識別子を含む重複排除アドレスと、前記重複排除対象データに所定の演算を行うことで得られる前記重複排除対象データの固有値とを対応付けて管理するインデックスサーバから、検出された前記データユニットに割り当てられた前記単位記憶領域内のデータの前記固有値に対応付けられた前記重複排除アドレスを取得し、

検出された前記データユニットに対応付けて、取得した前記重複排除アドレスを前記データユニット情報記憶手段に格納すると共に、検出された前記データユニットへの前記単位記憶領域の割り当てを解除し、

前記データユニットを指定した読み出し要求があると、読み出し対象の前記データユニットに割り当てられた前記単位記憶領域がある場合、割り当てられた前記単位記憶領域からデータを読み出し、読み出し対象の前記データユニットに割り当てられた前記単位記憶領域がない場合、前記データユニット情報記憶手段を参照し、読み出し対象の前記データユニットに対応付けられた前記重複排除アドレスに示される前記ディスクノードから、前記重複排除アドレスに示される前記重複排除ユニットの識別情報を指定して前記重複排除対象データを取得する、

ことを特徴とするデータ管理方法。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明はデータ管理プログラム、データ管理装置、およびデータ管理方法に関する。

【背景技術】

【０００２】

データを複数のコンピュータで分散管理するシステムとして、マルチノードストレージシステムがある。マルチノードストレージシステムは、ネットワークに接続した複数のディスクノードと制御ノードを有している。マルチノードストレージシステムでは、制御ノードの管理の下、仮想的なディスク（論理ボリューム）に格納するデータを複数のディスクノードに分散格納される。

【０００３】
具体的には、マルチノードストレージシステムでは、例えば論理ボリュームがセグメント単位に分割されている。この場合、ディスクノードが有するストレージ装置の記憶領域はスライス単位に分割される。このスライスは、固定長であり、セグメントと同サイズとなる。制御ノードによって、論理ボリュームの各セグメントに対して、ストレージ装置のスライスが割り当てられる。セグメントへのスライスの割り当て関係は、制御ノードからデータアクセスを行うコンピュータ（アクセスノード）に通知される。そして、アクセスノードから、セグメントを格納先として指定されたデータが、そのセグメントに割り当てられたスライスを管理するディスクノードに送られ、そのディスクノードが有するストレージ装置に格納される。

【０００４】
このようなマルチノードストレージシステムによれば、ネットワークにディスクノードを追加することで管理可能なデータ容量を増やすことができる。そのため、システムの拡張が容易になる。

【０００５】
なお、コンピュータシステム内で同じ内容のデータを複数持つことは、ストレージの使用効率を悪化させることとなる。例えば、データのバックアップを定期的に実行する場合、前回のバックアップ内容と比較してユニークなデータは、バックアップ対象の一部であることが多い。そこで、バックアップなどで移動させるデータの冗長性を低減する技術が考えられている。

【先行技術文献】
【特許文献】
【０００６】
【特許文献１】国際公開第２００４／１０４８４５号
【特許文献２】特開２００７－２３４０２６号公報
【発明の概要】
【発明が解決しようとする課題】
【０００７】
ところで、コンピュータが保持するデータを細分化して見てみると、運用中のシステム内でも同じ内容のデータが複数同時に存在する可能性がある。例えば同じファイルを添付した電子メールが、メールサーバを共有する複数のユーザに送信されると、そのメールサーバにおいて同じ内容のデータが異なる場所に格納される。

【０００８】
特にマルチノードストレージシステムでは、論理ボリュームごとに異なるユーザにサービスを提供する場合がある。この場合、複数の論理ボリュームそれぞれに同じアプリケーションプログラムがインストールされることもあり得る。同じアプリケーションプログラムが異なる論理ボリュームにインストールされると、マルチノードストレージシステム全体としては、同じ内容のデータが複数の場所に格納されることとなる。

【０００９】
しかし、従来は、マルチノードストレージシステムの異なるディスクノードに同じ内容のデータが重複して存在する場合、それらのデータの冗長性を低減させることができなかった。そのため、異なるディスクノードに同じ内容のデータが複数格納され、ストレージが効率的に利用されていなかった。

【００１０】
本発明はこのような点に鑑みてなされたものであり、マルチノードストレージシステムにおけるデータの冗長性を低減することができるデータ管理プログラム、データ管理装置、およびデータ管理方法を提供することを目的とする。

【課題を解決するための手段】
【００１１】
上記課題を解決するために、複数のディスクノードそれぞれで管理されるストレージ装

置にデータを分散格納するマルチノードストレージシステムを構成する１つのディスクノードで実行すべきデータ管理処理をコンピュータに実行させるデータ管理プログラムが提供される。このデータ管理プログラムは、コンピュータに以下の処理を実行させる。

【００１２】

コンピュータは、仮想的な記憶領域を定義した論理ボリュームを複数に分割して得られるデータユニットを指定した書き込み要求があると、コンピュータに接続されたストレージ装置内の単位記憶領域の１つを指定されたデータユニットに割り当て、データユニットに割り当てられた単位記憶領域に対してデータの書き込みを行う。コンピュータは、書き込み対象となったデータユニットに対応付けて、現在の時刻を最終書き込み時刻としてデータユニット情報記憶手段に格納する。コンピュータは、データユニット情報記憶手段を参照し、最終書き込み時刻から重複排除化猶予期間以上経過しているデータユニットを検出する。コンピュータは、仮想的な記憶領域を定義した重複排除ボリュームを複数に分割して得られる重複排除ユニットのうち重複排除化の対象とされた重複排除対象データの格納に使用されている重複排除ユニットの識別情報および重複排除ユニットを管理しているディスクノードの識別子を含む重複排除アドレスと、重複排除対象データに所定の演算を行うことで得られる重複排除対象データの固有値とを対応付けて管理するインデックスサーバから、検出されたデータユニットに割り当てられた単位記憶領域内のデータの固有値に対応付けられた重複排除アドレスを取得する。コンピュータは、検出されたデータユニットに対応付けて、取得した重複排除アドレスをデータユニット情報記憶手段に格納すると共に、検出されたデータユニットへの単位記憶領域の割り当てを解除する。

【００１３】

また、上記課題を解決するために、上記データ管理プログラムを実行するコンピュータと同様の機能を有するデータ管理装置が提供される。また、上記課題を解決するために、上記データ管理プログラムに基づいてコンピュータに実行する処理によるデータ管理方法が提供される。

【００１４】

さらに、上記課題を解決するために、複数のディスクノードそれぞれで管理されるストレージ装置にデータを分散格納するマルチノードストレージシステムの記憶領域管理処理をコンピュータに実行させるデータ管理プログラムが提供される。このデータ管理プログラムは、コンピュータに以下の処理を実行させる。

【００１５】

仮想的な記憶領域を定義した重複排除ボリュームを複数に分割して得られる重複排除ユニットのうちディスクノードで管理している重複排除ユニットの使用の有無、使用されている重複排除ユニットに割り当てられている単位記憶領域に格納された重複排除対象データに所定の演算を行うことで得られる固有値、および重複排除ユニットの識別情報を含む重複排除ユニット情報が、重複排除ユニットを管理しているディスクノードの識別子に対応付けて重複排除ユニット情報記憶手段に格納されている。コンピュータは、ディスクノードから、重複排除化データユニットに割り当てられた単位記憶領域内のデータに所定の演算を実行して得られた固有値を含む重複排除アドレス照会要求を受け取ると、重複排除ユニット情報記憶手段から、重複排除アドレス照会要求に示される固有値を含む重複排除ユニット情報を検索する。コンピュータは、検索により該当する重複排除ユニット情報が検出された場合、検出された重複排除ユニット情報に含まれる重複排除ユニットの識別情報と、検出された重複排除ユニット情報に対応する重複排除ユニットを管理しているディスクノードの識別子とを含む重複排除アドレスを、重複排除アドレス照会要求の送信元であるディスクノードに応答する。コンピュータは、検索により該当する重複排除ユニット情報が検出されなかった場合、重複排除ユニット情報記憶手段を参照し、未使用の重複排除ユニットを選択し、選択した重複排除ユニットを管理するディスクノードに対して選択した重複排除ユニットへの単位記憶領域の割り当て要求を送信し、選択した重複排除ユニットの単位記憶領域割り当て後の重複排除ユニット情報を重複排除ユニット情報記憶手段に格納し、選択した前記重複排除ユニットの識別情報と、選択した前記重複排除ユニット

を管理する前記ディスクノードの識別子とを含む重複排除アドレスを、重複排除アドレス照会要求の送信元であるディスクノードに応答する。

【発明の効果】

【００１６】

マルチノードストレージシステムにおけるディスクノード間のデータの冗長性を低減することができる。

【図面の簡単な説明】

【００１７】

【図１】第１の実施の形態のシステム構成を示すブロック図である。

【図２】第２の実施の形態のシステム構成例を示す図である。

【図３】本実施の形態に用いるディスクノードのハードウェア構成例を示す図である。

【図４】論理ボリュームと重複排除ボリュームへの記憶領域の割り当て関係を示す図である。

【図５】ストレージ装置に格納されている情報の例を示す図である。

【図６】論理ボリューム用スライスメタデータのデータ構造例を示す図である。

【図７】重複排除ボリューム用スライスメタデータのデータ構造例を示す図である。

【図８】制御ノードとアクセスノードとの機能を示すブロック図である。

【図９】制御ノードのスライスメタデータ記憶部のデータ構造例を示す図である。

【図１０】制御ノードの論理ボリューム構成情報記憶部のデータ構造例を示す図である。

【図１１】制御ノードの重複排除ボリューム構成情報記憶部のデータ構造例を示す図である。

【図１２】アクセスノードの論理ボリューム構成情報記憶部のデータ構造例を示す図である。

【図１３】ディスクノードとインデックスサーバとの機能を示すブロック図である。

【図１４】ディスクノードが有する重複排除ボリューム構成情報記憶部のデータ構造例を示す図である。

【図１５】重複排除ユニット情報収集処理を示すシーケンス図である。

【図１６】重複排除ユニット情報記憶部のデータ構造例を示す図である。

【図１７】重複排除ボリュームからのデータ読み出し処理の手順を示すシーケンス図である。

【図１８】状態が「Blank」のデータユニットへの書き込み処理を示すシーケンス図である。

【図１９】状態が「DeDup」のデータユニットの一部への書き込み処理を示すシーケンス図である。

【図２０】ディスクノードにおける書き込み処理を示すフローチャートである。

【図２１】状態が「Normal」のデータユニットへのパトロール処理を示すシーケンス図である。

【図２２】状態が「DeDup」のデータユニットへのパトロール処理を示すシーケンス図である。

【図２３】パトロール処理の手順を示すフローチャートである。

【図２４】状態が「Normal」のデータユニットのパトロール処理の手順を示すフローチャートである。

【図２５】重複排除ボリュームのパトロール処理の手順を示すフローチャートである。

【図２６】インデックスサーバにおける重複排除アドレス検索処理の手順を示すフローチャートである。

【図２７】未使用ユニット情報反映処理の手順を示すフローチャートである。

【図２８】パトロールによるデータの保存先変更状況を示す図である。

【図２９】重複排除ユニットへの関連付けの解消状況を示す図である。

【図３０】重複排除ユニットの未使用状態への変更状況を示す図である。

【発明を実施するための形態】

【００１８】
以下、本実施の形態について図面を参照して説明する。
〔第１の実施の形態〕
第１の実施の形態に係るマルチノードストレージシステムにおいて、データ実体を特別な仮想ボリューム（重複排除ボリューム）に保持し、ユーザの論理ボリュームには、そのデータ実体を指し示すインデックスを保持させる。このような処理を、重複排除（DeDup: Deduplication）処理と呼ぶ。これにより、同じ内容のデータを重複して保持することが抑止され、ストレージ容量の使用効率を向上する。
【００１９】
図１は、第１の実施の形態のシステム構成を示すブロック図である。第１の実施の形態にかかるマルチノードストレージシステムでは、複数のディスクノード１～３それぞれで管理されるストレージ装置５～７にデータを分散格納する。このマルチノードストレージシステムでは、インデックスサーバ４によって、重複排除処理における重複排除化されたデータの格納場所が管理される。
【００２０】
ディスクノード１は、書き込みアクセス手段１ａ、最終書き込み時刻更新手段１ｂ、データユニット情報記憶手段１ｃ、重複排除化データユニット検出手段１ｄ、重複排除アドレス取得手段１ｅ、単位記憶領域割当解除手段１ｆ、読み出しアクセス手段１ｇ、重複排除ユニット割当手段１ｈ、および重複排除データアクセス応答手段１ｉを有している。
【００２１】
書き込みアクセス手段１ａは、仮想的な記憶領域を定義した論理ボリュームを複数に分割して得られるデータユニットを指定した書き込み要求を、図示していないコンピュータから取得する。すると書き込みアクセス手段１ａは、ディスクノード１に接続されたストレージ装置５内の単位記憶領域の１つ（図１の例では単位記憶領域５ａ）を指定されたデータユニットに割り当てる。さらに書き込みアクセス手段１ａは、データユニットに割り当てられた単位記憶領域５ａに対してデータの書き込みを行う。
【００２２】
最終書き込み時刻更新手段１ｂは、書き込み対象となったデータユニットに対応付けて、現在の時刻を最終書き込み時刻としてデータユニット情報記憶手段１ｃに格納する。
データユニット情報記憶手段１ｃは、論理ボリュームに含まれるデータユニットごとに、最終書き込み時刻、または重複排除アドレスを記憶する記憶機能である。例えばデータユニット情報記憶手段１ｃは、ディスクノード１内のメモリの記憶領域の一部が使用される。また、ストレージ装置５の記憶領域の一部を、データユニット情報記憶手段１ｃとして使用することもできる。
【００２３】
ここで重複排除アドレスとは、重複排除ユニットのうち重複排除化の対象とされた重複排除対象データの格納に使用されている重複排除ユニットの識別情報、および重複排除ユニットを管理しているディスクノードの識別子を含む情報である。重複排除ユニットは、仮想的な記憶領域を定義した重複排除ボリュームを複数に分割して得られる仮想的な記憶領域である。
【００２４】
重複排除化データユニット検出手段１ｄは、データユニット情報記憶手段１ｃを参照し、最終書き込み時刻から重複排除化猶予期間以上経過しているデータユニット（重複排除化データユニット）を検出する。重複排除化猶予期間は、重複排除化データユニット検出手段１ｄに予め設定されている。検出されたデータユニットに割り当てられた単位記憶領域５ａに格納されたデータは、重複排除化の対象となる。
【００２５】
重複排除アドレス取得手段１ｅは、インデックスサーバ４から、検出されたデータユニットに割り当てられた単位記憶領域５ａ内のデータの固有値に対応付けられた重複排除アドレスを取得する。ここで、データの固有値とは、データに所定の演算を行うことで得ら

れる値である。この場合の演算は、入力されるデータが異なれば異なる固有値を生成するような演算である。例えばデータに対してハッシュ関数による演算を行うことで得られるハッシュ値が、固有値として使用できる。この場合のハッシュ関数としては、例えば、入力されるデータが異なれば異なるハッシュ値を生成するようなハッシュ関数が用いられる。

【００２６】

単位記憶領域割当解除手段１ｆは、検出されたデータユニットに対応付けて、取得した重複排除アドレスをデータユニット情報記憶手段１ｃに格納する。また単位記憶領域割当解除手段１ｆは、検出されたデータユニットへの単位記憶領域５ａの割り当てを解除する。割当が解除された単位記憶領域５ａは、未使用の状態となり、他のデータユニットまたは重複排除ユニットへの割り当てが可能となる。なお重複排除アドレス取得時にインデックスサーバ４において重複排除ユニット割当処理が行われる場合がある。この場合、単位記憶領域割当解除手段１ｆは、単位記憶領域５ａの割り当てを解除する前に、単位記憶領域５ａに格納されているデータを、重複排除アドレスで示されるディスクノードに転送する。

【００２７】

読み出しアクセス手段１ｇは、データユニットを指定した読み出し要求があると、読み出し対象のデータユニットに割り当てられた単位記憶領域５ａがある場合、割り当てられた単位記憶領域５ａからデータを読み出す。また読み出しアクセス手段１ｇは、読み出し対象のデータユニットに割り当てられた単位記憶領域５ａがない場合、データユニット情報記憶手段１ｃを参照する。そして読み出しアクセス手段１ｇは、読み出し対象のデータユニットに対応付けられた重複排除アドレスに示されるディスクノードから、重複排除アドレスに示される重複排除ユニットの識別情報を指定して重複排除対象データを取得する。

【００２８】

重複排除ユニット割当手段１ｈは、インデックスサーバ４から重複排除対象データの重複排除ユニット割当要求を受信すると、未使用の重複排除ユニットに単位記憶領域の１つを割り当て、割り当てた単位記憶領域に重複排除対象データを格納する。

【００２９】

重複排除データアクセス応答手段１ｉは、重複排除ユニットの識別情報を指定したデータの読み出し要求を受信すると、指定された重複排除ユニットに割り当てられた単位記憶領域内の重複排除対象データを応答する。

【００３０】

ディスクノード２，３もディスクノード１と同様の機能の要素を有する。なお図１では、ディスクノード２の有する要素のうち、重複排除ユニット割当手段２ｈと重複排除データアクセス応答手段２ｉが示されている。また図１では、ディスクノード３の有する要素のうち、データユニット情報記憶手段３ｃ、重複排除化データユニット検出手段３ｄ、重複排除アドレス取得手段３ｅ、および単位記憶領域割当解除手段３ｆが示されている。ディスクノード２，３に示される各要素は、ディスクノード１内の同名の要素と同じ機能を有する。

【００３１】

インデックスサーバ４は、重複排除ユニット情報記憶手段４ａ、重複排除ユニットアドレス検索手段４ｂ、および重複排除ユニット割当要求手段４ｃを有する。

重複排除ユニット情報記憶手段４ａは、重複排除ユニット情報を、その重複排除ユニット情報に対応する重複排除ユニットを管理するディスクノードの識別子に対応付けて記憶する。重複排除ユニット情報には、重複排除ボリュームの重複排除ユニットのうち、ディスクノード１～３で管理している重複排除ユニットの使用の有無が示される。また重複排除ユニット情報には、使用されている重複排除ユニットに割り当てられている単位記憶領域に格納された重複排除対象データの固有値、および重複排除ユニットの識別情報が含まれる。

【００３２】
重複排除ユニットアドレス検索手段４ｂは、ディスクノード１～３から、重複排除化データユニットに割り当てられた単位記憶領域内のデータに所定の演算を実行して得られた固有値を含む重複排除アドレス照会要求を受け取る。すると重複排除ユニットアドレス検索手段４ｂは、重複排除ユニット情報記憶手段４ａから、重複排除アドレス照会要求に示される固有値を含む重複排除ユニット情報を検索する。重複排除ユニットアドレス検索手段４ｂは、検索により該当する重複排除ユニット情報が検出された場合、重複排除アドレスを、重複排除アドレス照会要求の送信元であるディスクノードに応答する。この重複排除アドレスには、検出された重複排除ユニット情報に含まれる重複排除ユニットの識別情報と、検出された重複排除ユニット情報に対応する重複排除ユニットを管理しているディスクノードの識別子とが含まれる。

【００３３】
また重複排除ユニットアドレス検索手段４ｂは、検索により該当する重複排除ユニット情報が検出されなかった場合、重複排除ユニット割当要求手段４ｃに、重複排除ユニットへの単位記憶領域の割り当て要求を依頼する。そして重複排除ユニットアドレス検索手段４ｂは、重複排除ユニット割当要求手段４ｃによる割り当て結果に応じた重複排除アドレスを、重複排除アドレス照会要求の送信元であるディスクノードに応答する。この重複排除アドレスには、重複排除ユニット割当要求手段４ｃが選択した重複排除ユニットの識別情報と、重複排除ユニット割当要求手段４ｃが選択した重複排除ユニットを管理するディスクノードの識別子とが含まれる。このとき、重複排除ユニットアドレス検索手段４ｂは、応答内容に、重複排除ユニット割り当て処理が実行されたことを示す情報を含める。この情報を含めることで、重複排除アドレスを取得したディスクノードにおいて、重複排除化の対象となったデータユニットに割り当てられた記憶領域内のデータの転送が行われる。

【００３４】
重複排除ユニット割当要求手段４ｃは、重複検索により該当する重複排除ユニット情報が検出されなかった場合、重複排除ユニット情報記憶手段４ａを参照し、未使用の重複排除ユニットを選択する。次に重複排除ユニット割当要求手段４ｃは、選択した重複排除ユニットを管理するディスクノードに対して選択した重複排除ユニットへの単位記憶領域の割り当て要求を送信する。さらに重複排除ユニット割当要求手段４ｃは、単位記憶領域割り当て後の重複排除ユニット情報を重複排除ユニット情報記憶手段４ａに格納する。そして重複排除ユニット割当要求手段４ｃは、選択した重複排除ユニットの識別情報と、選択した重複排除ユニット情報を管理するディスクノードの識別子とを含む重複排除アドレスを、重複排除ユニットアドレス検索手段４ｂに渡す。

【００３５】
このような構成のマルチノードストレージシステムにおいて、まずディスクノード１に対してデータ（data[a]）の書き込み要求が入力されたものとする。その場合、書き込みアクセス手段１ａにより、書き込み要求で指定されたデータユニットに単位記憶領域５ａが割り当てられる。そして、書き込みアクセス手段１ａにより、単位記憶領域５ａにデータ（data[a]）が書き込まれる。このとき、最終書き込み時刻更新手段１ｂにより、書き込み対象となったデータユニットに対応付けて、現在の時刻が最終書き込み時刻としてデータユニット情報記憶手段１ｃに格納される。

【００３６】
その後、重複排除化データユニット検出手段１ｄにより、最終書き込み時刻から重複排除化猶予期間以上経過しているデータユニット（重複排除化データユニット）が検出される。ここで、データ（data[a]）が重複排除猶予期間以上更新されなかったものとする。この場合、データ（data[a]）が格納された単位記憶領域５ａが割り当てられたデータユニットについて、最終書き込み時刻から重複排除化猶予期間以上経過したことが検出される。

【００３７】

重複排除化データユニットが検出されると、重複排除アドレス取得手段１ｅにより、インデックスサーバ４から、重複排除化データユニットに割り当てられた単位記憶領域５ａ内のデータの固有値に対応付けられた重複排除アドレスが取得される。例えば、重複排除アドレス取得手段１ｅからインデックスサーバ４へ、重複排除化データユニットに割り当てられた単位記憶領域５ａ内のデータ（data[a]）に所定の演算を実行して得られた固有値を含む重複排除アドレス照会要求が送信される。

【００３８】

インデックスサーバ４の重複排除ユニットアドレス検索手段４ｂでは、重複排除アドレス照会要求に応答し、重複排除ユニット情報記憶手段４ａから、重複排除アドレス照会要求に示される固有値を含む重複排除ユニット情報が検索される。この時点では、データ（data[a]）の固有値を有する重複排除ユニット情報は、重複排除ユニット情報記憶手段４ａに存在しないものする。この場合、検索により該当する重複排除ユニット情報が検出されない。

【００３９】

検索により該当する重複排除ユニット情報が検出されなかった場合、重複排除ユニット割当要求手段４ｃにより、重複排除ユニット情報記憶手段４ａから未使用の重複排除ユニットが選択される。この例では、ディスクノード２が管理する重複排除ユニットが選択されたものとする。次に重複排除ユニット割当要求手段４ｃにより、選択した重複排除ユニットを管理するディスクノード２に対して選択した重複排除ユニットへの単位記憶領域の割り当て要求が送信される。すると、ディスクノード２の重複排除ユニット割当手段２ｈにより、未使用の重複排除ユニットに単位記憶領域６ａが割り当てられる。さらに重複排除ユニット割当要求手段４ｃにより、選択した重複排除ユニットの単位記憶領域割り当て後の重複排除ユニット情報が重複排除ユニット情報記憶手段４ａに格納される。そして重複排除ユニット割当要求手段４ｃにより、選択した重複排除ユニットの識別情報と、選択した重複排除ユニット情報を管理するディスクノード２の識別子とを含む重複排除アドレスが重複排除ユニットアドレス検索手段４ｂに渡される。すると、重複排除ユニットアドレス検索手段４ｂにより、重複排除ユニット割当要求手段４ｃから渡された重複排除アドレスが、重複排除アドレス照会要求の送信元であるディスクノード１に応答される。

【００４０】

ディスクノード１では、重複排除アドレスを重複排除アドレス取得手段１ｅが受け取る。すると単位記憶領域割当解除手段１ｆにより、検出されたデータユニットに対応付けて、取得した重複排除アドレスがデータユニット情報記憶手段１ｃに格納される。また、単位記憶領域割当解除手段１ｆにより、単位記憶領域５ａに格納されているデータ（data[a]）が、重複排除アドレスで示されるディスクノード２に転送される。すると、ディスクノード２の重複排除データアクセス応答手段２ｉによって、割り当てた単位記憶領域６ａにデータ（data[a]）が格納される。

【００４１】

その後、単位記憶領域割当解除手段１ｆにより、検出されたデータユニットへの単位記憶領域５ａの割り当てが解除される。割当が解除された単位記憶領域５ａは、未使用の状態となり、他のデータユニットまたは重複排除ユニットへの割り当てが可能となる。

【００４２】

このようにして、ディスクノード１で管理されていたデータ（data[a]）が、重複排除対象データとしてディスクノード２で管理されるようになる。

ディスクノード１では、重複排除アドレスに基づいてデータ（data[a]）を読み出すことができる。例えば、ディスクノード１に対して読み出し要求があったとき、読み出しアクセス手段１ｇにより、読み出し対象のデータユニットに割り当てられた単位記憶領域があれば、割り当てられた単位記憶領域５ａからデータが読み出される。他方、単位記憶領域５ａの割り当てが解除されたデータユニットへの読み出し要求が出された場合、ストレージ装置５内には既にデータが保存されていない。この場合、読み出しアクセス手段１ｇにより、データユニット情報記憶手段１ｃに基づいて、読み出し対象のデータユニットに

対応付けられた重複排除アドレスに示されるディスクノード２からデータ（data[a]）が取得される。ディスクノード２からデータを取得する際には、重複排除アドレスに示される重複排除ユニットの識別情報が指定される。これにより、取得対象のデータが一意に決定される。

【００４３】

その後、ディスクノード３において、重複排除対象データと同じ内容のデータ（data[a]）がストレージ装置７に格納され、重複排除化猶予期間以上更新されなかったものとする。この場合、ディスクノード３の重複排除化データユニット検出手段３ｄにより、データ（data[a]）が格納された単位記憶領域７ａが割り当てられたデータユニットについて、最終書き込み時刻から重複排除化猶予期間以上経過したことが検出される。

【００４４】

重複排除化データユニットが検出されると、重複排除アドレス取得手段３ｅにより、インデックスサーバ４から、重複排除化データユニットに割り当てられた単位記憶領域７ａ内のデータの固有値に対応付けられた重複排除アドレスが取得される。例えば、重複排除アドレス取得手段３ｅからインデックスサーバ４へ、重複排除化データユニットに割り当てられた単位記憶領域７ａ内のデータ（data[a]）の固有値を含む重複排除アドレス照会要求が送信される。

【００４５】

インデックスサーバ４の重複排除ユニットアドレス検索手段４ｂでは、重複排除アドレス照会要求に応答し、重複排除ユニット情報記憶手段４ａから、重複排除アドレス照会要求に示される固有値を含む重複排除ユニット情報が検索される。このとき単位記憶領域７ａ内のデータと同じ内容のデータ（data[a]）が単位記憶領域６ａに格納されている。そのため、単位記憶領域６ａが割り当てられている重複排除ユニットの重複排除ユニット情報が、検索結果として抽出される。すると重複排除ユニットアドレス検索手段４ｂにより、重複排除アドレスが、重複排除アドレス照会要求の送信元であるディスクノード３に応答される。

【００４６】

ディスクノード３では、重複排除アドレスを重複排除アドレス取得手段３ｅが受け取る。すると単位記憶領域割当解除手段３ｆにより、検出されたデータユニットに対応付けて、取得した重複排除アドレスがデータユニット情報記憶手段３ｃに格納される。その後、単位記憶領域割当解除手段３ｆにより、検出されたデータユニットへの単位記憶領域７ａの割り当てが解除される。なお、ディスクノード３からもディスクノード１と同様に、ストレージ装置６に格納されたデータ（data[a]）を読み出すことができる。

【００４７】

このようにして、マルチノードストレージシステムにおいて、複数のディスクノードで重複して格納されている同一内容の複数のデータが、１つのデータに統合される。すなわち、マルチノードストレージシステムにおいて、ノード間にまたがって重複するデータを除去して保持することで冗長性が低減され、ストレージ容量の使用効率を上げることができる。

【００４８】

また、上記実施の形態では、重複排除猶予期間以上更新されないデータのみが重複排除の対象となる。換言すると、書き込み頻度が高いデータに関しては重複排除が行われない。これにより、性能影響を低く抑えつつ重複するデータの除去が可能となる。

【００４９】

〔第２の実施の形態〕

次に、第２の実施の形態について説明する。第２の実施の形態は、論理ボリュームの記憶領域を一定の記憶容量のスライスに分割して管理するマルチノードストレージシステムに、第１の実施の形態で示したような重複排除技術を適用したものである。

【００５０】

第２の実施の形態では、論理ボリュームがセグメントに分けて管理されており、各セグ

メントにストレージ装置のスライスが割り当てられている。第２の実施の形態のシステムでは、このスライスをユニットに細分化し、ユニット単位でデータ実体を格納するための記憶領域（単位記憶領域）を割り当て、ユニット単位で重複データが存在するかどうかの管理を行う。

【００５１】

また、第２の実施の形態に係るシステムでは定期的にパトロールが実行され、アクセス頻度が低くなったユーザの論理ボリュームのユニット（データユニット）を重複排除される。逆に重複排除されたデータユニットに書き込みが発生すると、重複排除状態が解除され、再度、論理ボリューム内に単位記憶領域が割り当てられる。また、論理ボリューム割り当て直後の書き込みがなされていない状態のユニットには単位記憶領域は割り当てられず、書き込みが発生して初めて単位記憶領域が割り当てられる。

【００５２】

重複排除ボリュームにある重複排除ユニットの情報は、サーバ起動時にインデックスサーバに集められる。ディスクノードにおいてデータユニットを重複排除化する際には、既に同じデータが存在するかが検索され、存在していなければ重複排除対象のデータを格納する新たな重複排除ユニットが設定される。

【００５３】

重複排除ユニットには保存期限満了時刻が定められており、一定時間参照がなかった重複排除ユニットが保持するデータは廃棄される。システム内部では定期的にパトロールが実行され、すべてのデータユニットは指定された時間内に必ずパトロールが実施される。重複排除化されたデータユニットのパトロールの際には、そのデータユニットの重複排除化によるデータの参照先の重複排除ユニットに対して、参照が行われる。ユーザからの読み出し要求またはパトロールにより参照処理が行われた重複排除ユニットの保存期限満了時刻は、所定の保存期間だけ延長される。延長される保存期間より短い間隔でパトロールを実行することで、少なくとも１つのデータユニットからの参照先となっている重複排除ユニットについては、継続的にデータが保持される。他方、保存期限満了時刻を過ぎた重複排除ユニットは、単位記憶領域の割り当てが解除され、保持していたデータが消去される。

【００５４】

図２は、第２の実施の形態のシステム構成例を示す図である。ネットワーク１０には複数のディスクノード１００，２００，３００，４００、インデックスサーバ５００、制御ノード６００、およびアクセスノード７００，８００が接続されている。

【００５５】

ディスクノード１００，２００，３００，４００は、それぞれ「ＤＰ－Ａ」、「ＤＰ－Ｂ」、「ＤＰ－Ｃ」、「ＤＰ－Ｄ」のディスクノードＩＤが設定されており、このディスクノードＩＤによって一意に識別される。ディスクノード１００，２００，３００，４００は、それぞれストレージ装置１１０，２１０，３１０，４１０を有している。各ストレージ装置１１０，２１０，３１０，４１０は、例えばＲＡＩＤ（Redundant Arrays of Inexpensive Disks）５によりデータを管理するストレージシステムである。ディスクノード１００，２００，３００，４００は、制御ノード６００からの指示に従って、論理ボリュームのデータを管理する。また、ディスクノード１００，２００，３００，４００は、インデックスサーバ５００からの指示に従って、重複排除ボリューム内のデータを管理する。そして、ディスクノード１００，２００，３００，４００は、アクセスノード７００，８００からの要求に応じて、論理ボリュームにおける管理対象のデータを、ストレージ装置１１０，２１０，３１０，４１０へ入出力する。さらに、ディスクノード１００，２００，３００，４００は、他のディスクノードからの要求に応じて、重複排除ボリュームにおける管理対象のデータを、ストレージ装置１１０，２１０，３１０，４１０へ入出力する。

【００５６】

インデックスサーバ５００は、重複排除ボリュームに格納するデータのストレージ装置

１１０，２１０，３１０，４１０への割り当てを行う。また、既に重複排除ボリュームに格納されているデータと同じ内容のデータが重複排除対象となった場合、該当するデータのアドレスを、重複排除対象となったデータを管理するディスクノードに通知する。さらに、インデックスサーバ５００は、重複排除ボリューム内のデータのうち、いずれのディスクノードからも利用されていないデータを定期的に探索する。

【００５７】

制御ノード６００は、論理ボリュームや重複排除ボリュームを作成する。そして制御ノード６００は、論理ボリュームや重複排除ボリューム内のセグメントへのストレージ装置１１０，２１０，３１０，４１０の記憶領域の割り当てを行う。論理ボリュームと重複排除ボリュームへの記憶領域の割り当て結果は、制御ノード６００からディスクノード１００，２００，３００，４００に通知される。同様に論理ボリュームへの記憶領域の割り当て結果は、アクセスノード７００，８００に通知される。

【００５８】

アクセスノード７００，８００は、ユーザからの論理ボリュームへのアクセス要求に応答し、制御ノード６００から通知された論理ボリュームへの記憶領域の割り当て関係に基づいて、アクセス対象のデータを管理するディスクノードを判断する。さらにアクセスノード７００，８００は、該当するディスクノードに対してデータのアクセス要求を送信する。そして、アクセスノード７００，８００は、アクセス要求に対するディスクノードからの応答をユーザに返す。

【００５９】

このようなシステムにより、論理ボリュームで管理されているデータに重複排除対象データがあれば、そのデータが、適宜、重複排除ボリュームに移動される。本実施の形態では、重複排除ボリューム内のデータは、ハッシュ値によって論理ボリューム内のデータと関連付けられる。重複排除ボリューム内のデータに対して、論理ボリューム内の複数のデータを関連付けることで、同じ内容のデータが重複して格納されることを排除できる。

【００６０】

図３は、本実施の形態に用いるディスクノードのハードウェア構成例を示す図である。ディスクノード１００は、ＣＰＵ（Central Processing Unit）１０１によって装置全体が制御されている。ＣＰＵ１０１には、バス１０９を介してＲＡＭ（Random Access Memory）１０２と複数の周辺機器が接続されている。

【００６１】

ＲＡＭ１０２は、ディスクノード１００の主記憶装置として使用される。ＲＡＭ１０２には、ＣＰＵ１０１に実行させるＯＳ（Operating System）のプログラムやアプリケーションプログラムの少なくとも一部が一時的に格納される。また、ＲＡＭ１０２には、ＣＰＵ１０１による処理に必要な各種データが格納される。

【００６２】

バス１０９に接続されている周辺機器としては、ハードディスクドライブ（ＨＤＤ:Hard Disk Drive）１０３、グラフィック処理装置１０４、入力インタフェース１０５、光学ドライブ装置１０６、通信インタフェース１０７、およびストレージインタフェース１０８がある。

【００６３】

ＨＤＤ１０３は、内蔵したディスクに対して、磁気的にデータの書き込みおよび読み出しを行う。ＨＤＤ１０３は、ディスクノード１００の二次記憶装置として使用される。ＨＤＤ１０３には、ＯＳのプログラム、アプリケーションプログラム、および各種データが格納される。なお、二次記憶装置としては、フラッシュメモリなどの半導体記憶装置を使用することもできる。

【００６４】

グラフィック処理装置１０４には、モニタ１１が接続されている。グラフィック処理装置１０４は、ＣＰＵ１０１からの命令に従って、画像をモニタ１１の画面に表示させる。モニタ１１としては、ＣＲＴ（Cathode Ray Tube）を用いた表示装置や液晶表示装置など

がある。

【００６５】

入力インタフェース１０５には、キーボード１２とマウス１３とが接続されている。入力インタフェース１０５は、キーボード１２やマウス１３から送られてくる信号をＣＰＵ１０１に送信する。なお、マウス１３は、ポインティングデバイスの一例であり、他のポインティングデバイスを使用することもできる。他のポインティングデバイスとしては、タッチパネル、タブレット、タッチパッド、トラックボールなどがある。

【００６６】

光学ドライブ装置１０６は、レーザ光などを利用して、光ディスク１４に記録されたデータの読み取りを行う。光ディスク１４は、光の反射によって読み取り可能なようにデータが記録された可搬型の記録媒体である。光ディスク１４には、ＤＶＤ（Digital Versatile Disc）、ＤＶＤ－ＲＡＭ、ＣＤ－ＲＯＭ（Compact Disc Read Only Memory）、ＣＤ－Ｒ（Recordable）／ＲＷ（ReWritable）などがある。

【００６７】

通信インタフェース１０７は、ネットワーク１０に接続されている。通信インタフェース１０７は、ネットワーク１０を介して、他のコンピュータとの間でデータの送受信を行う。

【００６８】

ストレージインタフェース１０８にはストレージ装置１１０が接続されている。ストレージインタフェース１０８は、ＣＰＵ１０１からの指示に従ってストレージ装置１１０へのデータの入出力を行う。

【００６９】

以上のようなハードウェア構成によって、第２の形態の処理機能を実現することができる。なお図３には、ディスクノード１００のハードウェア構成例を示したが、他のディスクノード２００，３００，４００、インデックスサーバ５００、制御ノード６００，およびアクセスノード７００，８００も同様のハードウェアで実現することができる。

【００７０】

次に、論理ボリュームと重複排除ボリュームとに対するストレージ装置の記憶領域の割り当てについて説明する。

図４は、論理ボリュームと重複排除ボリュームへの記憶領域の割り当て関係を示す図である。図４の例では、２つの論理ボリューム２０，３０と１つの重複排除ボリューム４０とが設けられている。

【００７１】

論理ボリューム２０には、識別子として論理ボリュームＩＤ「ＬＶＯＬ－Ｘ」が付与されている。論理ボリューム２０は、複数のセグメント２１，２２，２３，・・・を有している。論理ボリューム３０には、識別子として論理ボリュームＩＤ「ＬＶＯＬ－Ｙ」が付与されている。論理ボリューム３０は、複数のセグメント３１，３２，３３，・・・を有している。重複排除ボリューム４０は、識別子として重複排除ボリュームＩＤ「ＤｅＤｕｐ－Ｕ」が付与されている。重複排除ボリューム４０は、複数のセグメント４１，４２，４３，・・・を有している。

【００７２】

論理ボリューム２０，３０や重複排除ボリューム４０の各セグメントは、所定の容量の仮想的なデータ記憶領域である。例えば１セグメント当たり１ＧＢの記憶領域が設けられる。各セグメント２１，２２，２３，・・・に対して、ストレージ装置１１０，２１０，３１０，４１０の記憶領域が割り当てられる。

【００７３】

ストレージ装置１１０には、複数のスライス１１１，１１２，１１３，１１４，・・・が定義されている。各スライス１１１，１１２，１１３，１１４，・・・は、論理ボリューム２０，３０のセグメントと同じ容量の記憶領域である。なおスライス１１１，１１２，１１３，１１４，・・・それぞれの記憶領域は、ストレージ装置１１０内の不連続（連

続しても良い）の記憶領域である。例えば１スライスの記憶容量が１ＧＢであれば、スライス１１１の記憶領域として、１ＧＢ分の記憶領域が所定データ長のデータユニット単位で、ストレージ装置１１０内に確保される。

【００７４】

同様に、ストレージ装置２１０には、複数のスライス２１１，２１２，２１３，２１４，・・・が定義されている。またストレージ装置３１０にも、複数のスライス３１１，３１２，３１３，３１４，・・・が定義されている。さらにストレージ装置４１０にも、複数のスライス４１１，４１２，４１３，４１４，・・・が定義されている。

【００７５】

各ストレージ装置１１０，２１０，３１０，４１０のスライスが、論理ボリューム２０，３０または重複排除ボリューム４０のセグメントに割り当てられる。図４の例では、論理ボリューム２０のセグメント２１～２３のセグメントＩＤを、「Ｘ１」、「Ｘ２」、「Ｘ３」としている。また論理ボリューム３０のセグメント３１～３３のセグメントＩＤを、「Ｙ１」、「Ｙ２」、「Ｙ３」としている。さらに重複排除ボリューム４０のセグメント４１～４３のセグメントＩＤを、「Ｕ１」、「Ｕ２」、「Ｕ３」としている。そして図４では、ストレージ装置１１０，２１０，３１０，４１０内のスライスに対して、割り当て先のセグメントのセグメントＩＤを示している。

【００７６】

ストレージ装置１１０，２１０，３１０，４１０には、各スライスの定義情報であるメタデータが格納されている。

図５は、ストレージ装置に格納されている情報の例を示す図である。ストレージ装置１１０の記憶領域は、メタデータ領域１１０ａとデータ実体領域１１０ｂとに分かれている。メタデータ領域１１０ａは、各スライスを管理するためのメタデータが格納される。メタデータのうち、論理ボリュームに割り当てられたスライスのメタデータを、論理ボリューム用スライスメタデータ５１，５２，５３，・・・とする。また、重複排除ボリュームに割り当てられたスライスのメタデータを、重複排除ボリューム用スライスメタデータ６１，６２，６３，・・・とする。

【００７７】

データ実体領域１１０ｂは、論理ボリュームまたは重複排除ボリュームに格納されるデータ実体を記憶する記憶領域である。データ実体領域１１０ｂは、ユニットと同じデータサイズの単位記憶領域７１～７６に分けられている。ここで、論理ボリュームに割り当てられたスライスの記憶領域となるユニット（データユニット）に対して、単位記憶領域７１～７３が割り当てられる。重複排除ボリュームに割り当てられたスライスのユニット（重複排除ユニット）に対して、単位記憶領域７４～７６が割り当てられる。

【００７８】

メタデータ領域１１０ａ内の各メタデータには、そのメタデータに対応するスライス内の各ユニットに割り当てられた単位記憶領域の位置が示されている。また各メタデータには、割当先となるセグメントが属する論理ボリュームまたは重複排除ボリュームの識別子と、セグメントＩＤとが設定されている。これにより、メタデータを介して、各セグメントと実際のデータを格納する単位記憶領域とが関連付けられることとなる。すなわち、メタデータ領域１１０ａは、第１の実施の形態におけるディスクノード１のデータユニット情報記憶手段１ｃの機能を含んでいる。

【００７９】

メタデータの総数は、データ実体領域１１０ｂをスライスの記憶容量で除算したときの商で示される数（最少スライス数）以上となる。例えばデータ実体領域の記憶容量が３００ＧＢであり、メタデータのサイズが１ＧＢであれば、少なくとも３００個のスライスを定義することができ、そのスライスに対応するメタデータが作成できる。本実施の形態ではデータユニット単位でのシステム内での同一データの重複が排除される。そのため、最少スライス数よりも多くのスライスを、定義することが可能である。なお、メタデータは、ストレージ装置１１０内のスライスを論理ボリュームまたは重複排除ボリュームのセグ

メントに割り当てたときに制御ノード６００により作成される。そして、制御ノード６００からディスクノード１００に対して新たに割り当てられたスライスのメタデータが送信され、ディスクノード１００によってストレージ装置１１０のメタデータ領域１１０ａにメタデータが格納される。

【００８０】

次に、論理ボリューム用スライスメタデータと重複排除ボリューム用スライスメタデータとのデータ構造について説明する。

まず論理ボリューム用スライスメタデータのデータ構造について説明する。

【００８１】

図６は、論理ボリューム用スライスメタデータのデータ構造例を示す図である。論理ボリューム用スライスメタデータ５０には、論理ボリュームＩＤ、セグメントＩＤ、および最終パトロール時刻のフィールドが設けられている。論理ボリュームＩＤのフィールドには、論理ボリューム用スライスメタデータ５０で定義されるスライスが割り当てられたセグメントが属する論理ボリュームの論理ボリュームＩＤが設定される。セグメントＩＤのフィールドには、論理ボリューム用スライスメタデータ５０で定義されるスライスが割り当てられたセグメントのセグメントＩＤが設定される。最終パトロール時刻のフィールドには、論理ボリューム用スライスメタデータ５０で定義されるスライスに対するパトロールを最後に実行した時刻が設定される。

【００８２】

また論理ボリューム用スライスメタデータ５０には、スライスに割り当てられたデータユニットごとに、データユニット情報５０ａが設けられている。データユニット管理情報は、スライスサイズをデータユニットサイズで除算したときの商で示される個数だけ設けられる。

【００８３】

各データユニット管理情報には、状態情報のフィールドが設けられている。状態情報のフィールドには、「Blank」、「Normal」、「DeDup」の３つの状態のいずれかが設定される。「Blank」は、書き込みがされていないデータユニットであり、データ実体を持たないことを示す。なおスライスをセグメントに割り当てた直後は、すべてのデータユニットの状態が「Blank」となっている。

【００８４】

「Normal」は、書き込みが行われたデータユニットであり、データユニットに割り当てられた単位記憶領域にデータ実体が格納されていることを示す。状態が「Normal」のデータユニット管理情報には、さらに最終書き込み時刻とデータ実体オフセットとのフィールドが設けられている。最終書き込み時刻のフィールドには、データユニットに対するデータの書き込みが最後に行われた時刻が設定される。データ実体オフセットのフィールドには、データ実体を格納している単位記憶領域の位置が、データ実体領域１１０ｂの先頭からのオフセットで示される。

【００８５】

「DeDup」は、書き込みが行われたデータユニットであり、データ実体が重複排除対象となって重複排除ボリューム４０で管理されていることを示す。状態が「DeDup」のデータユニット管理情報には、さらに重複排除ボリュームＩＤと重複排除オフセットとのフィールドが設けられている。重複排除ボリュームＩＤのフィールドには、データ実体を管理している重複排除ボリュームの重複排除ボリュームＩＤが設定される。重複排除オフセットのフィールドには、重複排除ボリューム内でのデータ実体を格納している単位記憶領域の位置が、重複排除ボリュームの先頭からのオフセットで示される。

【００８６】

次に重複排除ボリューム用スライスメタデータのデータ構造について説明する。

図７は、重複排除ボリューム用スライスメタデータのデータ構造例を示す図である。重複排除ボリューム用スライスメタデータ６０には、重複排除ボリュームＩＤ、セグメントＩＤ、および最終パトロール時刻のフィールドが設けられている。重複排除ボリュームＩ

Ｄのフィールドには、重複排除ボリューム用スライスメタデータ６０で定義されるスライスが割り当てられたセグメントが属する重複排除ボリュームの重複排除ボリュームＩＤが設定される。セグメントＩＤのフィールドには、重複排除ボリューム用スライスメタデータ６０で定義されるスライスが割り当てられたセグメントのセグメントＩＤが設定される。最終パトロール時刻のフィールドには、重複排除ボリューム用スライスメタデータ６０で定義されるスライスに対するパトロールを最後に実行した時刻が設定される。

【００８７】

また重複排除ボリューム用スライスメタデータ６０には、スライスに割り当てられた重複排除ユニットごとに、重複排除ユニット情報６０ａが設けられている。重複排除ユニット情報６０ａは、スライスサイズを重複排除ユニットサイズで除算したときの商で示される個数だけ設けられる。セグメントＩＤで示されるセグメントの領域の開始位置に近いユニットの重複排除ユニット情報６０ａほど、重複排除ボリューム用スライスメタデータ内で上位に設定されている。すなわち、セグメントの領域の開始位置を示すオフセットと、重複排除ユニット情報６０ａの重複排除ボリューム用スライスメタデータ６０内での順番により、その重複排除ユニット情報６０ａに対応するユニットの重複排除オフセットが求められる。

【００８８】

重複排除ユニット情報６０ａには、ハッシュ値、保存期限満了時刻（expire時刻）、およびデータ実体オフセットのフィールドが設けられている。ハッシュ値のフィールドには、重複排除ユニットに書き込まれたデータ実体を所定のハッシュ関数で演算することで得られるハッシュ値が設定される。保存期限満了時刻のフィールドには、重複排除対象となったデータの保存期限が満了する時刻が設定される。重複排除ユニットの保存期限満了時刻が経過すると、該当する重複排除ユニット内が初期化される。

【００８９】

なお、保存期限満了時刻は、重複排除ユニットへのデータ書き込み時に初期値が設定される。設定される保存期限満了時刻は、書き込み時刻を基準として、各スライスのパトロール間隔よりも長い時間経過後の時刻が設定される。

【００９０】

重複排除ユニットに対して参照要求が出されると、この保存期限満了時刻は随時延長される。その際にも再設定する保存期限満了時刻は、参照時刻を基準として、各スライスのパトロール間隔よりも長い所定の保存期間経過後の時刻が設定される。

【００９１】

なお、データの参照要求は、重複排除対象データに対してアクセスノードから参照要求が出された場合、および重複排除対象データと同じデータを有するスライスに対するパトロールが実行された場合に発生する。

【００９２】

データ実体オフセットのフィールドには、重複排除ユニットのデータ実体が保存されている単位記憶領域の位置が、データ実体領域１１０ｂの先頭からのオフセットで示される。

【００９３】

ストレージ装置１１０，２１０，３１０，４１０に図５～図７に示した情報が格納されている状態で、システム内の各機器が起動されると、各機器において分散ストレージシステムの運用に必要な機能が動作開始する。

【００９４】

図８は、制御ノードとアクセスノードとの機能を示すブロック図である。制御ノード６００は、スライスメタデータ収集部６１０、スライスメタデータ記憶部６２０、論理ボリューム管理部６３０、論理ボリューム構成情報記憶部６４０、および重複排除ボリューム構成情報記憶部６５０を有している。

【００９５】

スライスメタデータ収集部６１０は、制御ノード６００の起動時に、各ディスクノード

１００，２００，３００，４００からメタデータを収集する。すなわち、スライスメタデータ収集部６１０は、論理ボリューム用スライスメタデータと重複排除ボリューム用スライスメタデータとの取得要求を、各ディスクノード１００，２００，３００，４００に対して送信する。すると各ディスクノード１００，２００，３００，４００からストレージ装置１１０，２１０，３１０，４１０に格納されている論理ボリューム用スライスメタデータと重複排除ボリューム用スライスメタデータとが制御ノード６００に送信される。スライスメタデータ収集部６１０は、収集したメタデータを、スライスメタデータ記憶部６２０に格納する。

【００９６】

スライスメタデータ記憶部６２０は、メタデータ（論理ボリューム用スライスメタデータと重複排除ボリューム用スライスメタデータ）を記憶する記憶領域である。具体的には、スライスメタデータ記憶部６２０には、図６や図７に示したデータ構造の論理ボリューム用スライスメタデータと重複排除ボリューム用スライスメタデータとが、送信元のディスクノードの識別子（ディスクノードＩＤ）に関連付けて記憶される。スライスメタデータ記憶部６２０としては、例えば制御ノード６００内のＲＡＭやＨＤＤの記憶領域の一部が用いられる。

【００９７】

論理ボリューム管理部６３０は、スライスメタデータ記憶部６２０に格納された論理ボリューム用スライスメタデータに基づいて、論理ボリューム構成情報を生成する。具体的には、論理ボリューム管理部６３０は、論理ボリューム用スライスメタデータのうち、論理ボリュームＩＤの値が同じものをグループ化する。次に論理ボリューム管理部６３０は、グループ化した各論理ボリューム用スライスメタデータから、セグメントＩＤと、その論理ボリューム用スライスメタデータの送信元であるディスクノードのディスクノードＩＤとの組を作成する。そして、論理ボリューム管理部６３０は、セグメントＩＤとディスクノードＩＤとの組をセグメントＩＤでソートし、グループに共通の論理ボリュームＩＤで示される論理ボリュームの構成情報とする。論理ボリューム管理部６３０は、生成した論理ボリューム構成情報を、論理ボリューム構成情報記憶部６４０に格納する。

【００９８】

また論理ボリューム管理部６３０は、スライスメタデータ記憶部６２０に格納された重複排除ボリューム用スライスメタデータに基づいて、重複排除ボリューム構成情報を生成する。具体的には、論理ボリューム管理部６３０は、重複排除ボリューム用スライスメタデータのうち、重複排除ボリュームＩＤの値が同じものをグループ化する。次に論理ボリューム管理部６３０は、グループ化した各重複排除ボリューム用スライスメタデータから、セグメントＩＤと、その重複排除ボリューム用スライスメタデータの送信元であるディスクノードのディスクノードＩＤとの組を作成する。そして、論理ボリューム管理部６３０は、セグメントＩＤとディスクノードＩＤとの組をセグメントＩＤでソートし、グループに共通の重複排除ボリュームＩＤで示される重複排除ボリュームの構成情報とする。論理ボリューム管理部６３０は、生成した重複排除ボリューム構成情報を重複排除ボリューム構成情報記憶部６５０に格納する。

【００９９】

さらに論理ボリューム管理部６３０は、論理ボリュームのセグメントを指定したメタデータ要求をアクセスノード７００、８００から受け取ると、該当セグメントのメタデータをアクセスノード７００、８００に通知する。同様に論理ボリューム管理部６３０は、重複排除ボリュームのセグメントを指定したメタデータ要求をディスクノード１００，２００，３００，４００から受け取ると、該当セグメントのメタデータをディスクノード１００，２００，３００，４００に通知する。

【０１００】

アクセスノード７００は、論理ボリューム構成情報記憶部７１０、データアクセス要求部７２０を有している。

論理ボリューム構成情報記憶部７１０は、論理ボリュームのセグメントごとに、そのセ

グメントのデータを管理しているディスクノードの情報を記憶する。具体的には、論理ボリュームＩＤに対応付けて、各セグメントの論理ボリューム内での占有域を示すオフセットの範囲とディスクノードＩＤとの組が格納されている。論理ボリューム構成情報記憶部７１０として、例えばアクセスノード７００内のＲＡＭまたはＨＤＤの記憶領域の一部が使用される。

【０１０１】

データアクセス要求部７２０は、ユーザからのデータアクセス要求が発生すると、論理ボリューム構成情報記憶部７１０を参照して、アクセス対象のデータを管理しているディスクノードを判断する。そして、データアクセス要求部７２０は、アクセス対象のデータを管理しているディスクノードに対してアクセス要求を送信する。

【０１０２】

なお、データアクセス要求部７２０は、論理ボリューム構成情報記憶部７１０内にアクセス対象のデータが属するセグメントの情報が格納されていない場合、制御ノード６００に対して、該当セグメントのメタデータを要求する。データアクセス要求部７２０は、制御ノード６００から応答されたメタデータによって、アクセス対象となるデータを管理しているディスクノードを判断する。そして、データアクセス要求部７２０は、取得したメタデータの内容に基づいて論理ボリューム構成情報記憶部７１０に新たなセグメントの情報を登録すると共に、アクセス対象のデータを管理しているディスクノードに対してアクセス要求を送信する。

【０１０３】

データアクセス要求部７２０は、アクセス要求に対する応答をディスクノードから取得すると、応答されたアクセス結果をユーザからのデータアクセス要求に対する応答とする。

【０１０４】

なお、図８にはアクセスノード７００の機能を示したが、アクセスノード８００もアクセスノード７００と同様の機能を有している。

次に、制御ノード６００とアクセスノード７００，８００とが記憶する情報について詳細に説明する。

【０１０５】

図９は、制御ノードのスライスメタデータ記憶部のデータ構造例を示す図である。スライスメタデータ記憶部６２０には、メタデータの送信元であるディスクノードのディスクノードＩＤに対応付けて、そのディスクノードから送信された論理ボリューム用スライスメタデータと重複排除ボリューム用スライスメタデータとが格納されている。

【０１０６】

例えば、ディスクノード１００からは、図５に示したストレージ装置１１０のメタデータ領域１１０ａ内の論理ボリューム用スライスメタデータ５１，５２，５３，・・・と重複排除ボリューム用スライスメタデータ６１，６２，６３，・・・とが送信される。これらのメタデータが、ディスクノード１００のディスクノードＩＤ「ＤＰ－Ａ」に対応付けてスライスメタデータ記憶部６２０に格納される。

【０１０７】

スライスメタデータ記憶部６２０に格納された情報に基づいて、論理ボリューム管理部６３０によって論理ボリューム構成情報や重複排除ボリューム構成情報が作成される。

図１０は、制御ノードの論理ボリューム構成情報記憶部のデータ構造例を示す図である。論理ボリューム構成情報記憶部６４０には、論理ボリュームごとの論理ボリュームセグメント情報テーブル６４１，６４２が格納されている。論理ボリュームセグメント情報テーブル６４１には、論理ボリュームＩＤが設定されている。また論理ボリュームセグメント情報テーブル６４１には、セグメントＩＤ、オフセット、およびディスクノードＩＤの欄が設けられている。各欄の横方向に並べられた情報が関連付けられ、各セグメントのメタデータ（セグメントメタデータ６４１ａ）となる。

【０１０８】

セグメントＩＤの欄には、セグメントの識別番号が設定される。オフセットの欄には、セグメントに属する記憶領域の範囲が、論理ボリュームの記憶領域の先頭からのオフセットにより示されている。具体的には、セグメントに属する記憶領域の開始位置のオフセットと終了位置のオフセットとが、オフセットの欄に設定される。ディスクノードＩＤの欄には、セグメントに属するデータを管理するディスクノードのディスクノードＩＤが設定される。

【０１０９】

図１１は、制御ノードの重複排除ボリューム構成情報記憶部のデータ構造例を示す図である。重複排除ボリューム構成情報記憶部６５０には、重複排除ボリュームごとの重複排除ボリュームセグメント情報テーブル６５１が格納されている。重複排除ボリュームセグメント情報テーブル６５１には、重複排除ボリュームＩＤが設定されている。また重複排除ボリュームセグメント情報テーブル６５１には、セグメントＩＤ、オフセット、およびディスクノードＩＤの欄が設けられている。各欄の横方向に並べられた情報が関連付けられ、各セグメントのメタデータ（セグメントメタデータ６５１ａ）となる。各欄に設定される情報は、図１０に示した論理ボリュームセグメント情報テーブル６４１の同名の欄と同種の情報である。

【０１１０】

図１２は、アクセスノードの論理ボリューム構成情報記憶部のデータ構造例を示す図である。論理ボリューム構成情報記憶部７１０には、アクセスノード７００が使用する論理ボリュームの論理ボリュームセグメント情報テーブル７１１が格納されている。論理ボリュームセグメント情報テーブル７１１には、論理ボリュームＩＤが設定されている。また論理ボリュームセグメント情報テーブル７１１には、セグメントＩＤ、オフセット、およびディスクノードＩＤの欄が設けられている。各欄の横方向に並べられた情報が関連付けられ、各セグメントのメタデータ（セグメントメタデータ７１１ａ）となる。各欄に設定される情報は、図１０に示した論理ボリュームセグメント情報テーブル６４１の同名の欄と同種の情報である。

【０１１１】

なお、アクセスノード７００が有する論理ボリュームセグメント情報テーブル７１１には、例えば、過去にアクセスしたセグメントのメタデータのみを登録しておくことができる。その場合、メタデータが未登録のセグメントにアクセスする場合、データアクセス要求部７２０によって、該当するメタデータが制御ノード６００から取得される。また、アクセスノード７００の起動時またはその他の所定のタイミングで、データアクセス要求部７２０が、論理ボリューム内のすべてのセグメントのメタデータを、制御ノード６００から取得するようにしてもよい。その場合、論理ボリュームセグメント情報テーブル７１１には、過去にアクセスされたか否かに拘わらず、すべてのセグメントのメタデータが登録される。

【０１１２】

次に、ディスクノード１００，２００，３００，４００とインデックスサーバ５００とが有する機能について説明する。

図１３は、ディスクノードとインデックスサーバとの機能を示すブロック図である。ディスクノード１００は、データアクセス部１２１、アクセスユニット識別部１２２、データ実体領域管理部１２３、重複排除ユニット情報応答部１３１、重複排除ユニット割当部１３２，重複排除アドレス照会部１３３，重複排除ボリューム構成情報記憶部１４１、重複排除データアクセス要求部１４２、重複排除データアクセス応答部１４３、およびパトロール部１５０を有する。

【０１１３】

データアクセス部１２１は、アクセスノード７００，８００から要求された論理ボリュームへのリード／ライトのアクセスを実行する。具体的には、データアクセス部１２１は、アクセスノード７００，８００からのデータアクセス要求を受け取ると、アクセスユニット識別部１２２の機能を用いて、アクセス対象のユニットの状態を識別し、状態に応じ

た処理を行う。例えば状態が「Normal」であれば、データアクセス部１２１は、ストレージ装置１１０に格納されているデータユニットに対してアクセスする。また状態が「DeDup」であれば、データアクセス部１２１は、重複排除データアクセス要求部１４２の機能を用いて、アクセス対象のユニットに対応する重複排除ボリューム内の重複排除ユニットに対してアクセスする。

【０１１４】

なお新たなユニットへのデータ書き込みのアクセスの場合、データアクセス部１２１は、データ実体領域管理部１２３に対してデータ実体を格納するデータユニットの割り当てを依頼する。そして、データアクセス部１２１は、データ実体領域管理部１２３が新たに割り当てたデータユニットに対して、データの書き込みを行う。

【０１１５】

アクセスユニット識別部１２２は、データアクセス部１２１やパトロール部１５０からの要求に応じ、データユニットの状態を識別する。具体的には、アクセスユニット識別部１２２は、論理ボリューム用スライスメタデータを参照し、データアクセス部１２１やパトロール部１５０から指定された論理ボリューム内のアクセス位置に相当するデータユニットの状態を判断する。そして、アクセスユニット識別部１２２は、取得した状態をデータアクセス部１２１またはパトロール部１５０に応答する。

【０１１６】

データ実体領域管理部１２３は、ユニットにおけるデータ実体用領域の割り当て／解放を管理する。またデータ実体領域管理部１２３は、ディスクノード１００管理下のストレージ装置１１０のデータ実体領域１１０ｂにおける未使用領域の管理を行う。未使用領域の管理とは、データ実体領域１１０ｂ中の各単位記憶領域について、いずれかのスライスメタデータのユニットに割り当てられているか否かを管理する処理である。スライスメタデータにおいてユニットの管理情報中のデータ実体オフセットで指定された位置から１ユニット分の記憶領域が、該当ユニットに割り当てられているデータ実体領域１１０ｂ中の単位記憶領域である。例えばデータ実体領域管理部１２３は、データ実体領域１１０ｂ内の各単位記憶領域に対応するフラグの集合（ビットマップ）を保持し、そのスライスメタデータに割り当てられた単位記憶領域にフラグを立てる（フラグの値を「１」に設定する）。そしてデータ実体領域管理部１２３は、新たにスライスメタデータのユニットに対して単位記憶領域を割り当てる場合、フラグが立っていない（フラグの値が「０」）単位記憶領域を１つ選択して、スライスメタデータ中のユニットに割り当てる。

【０１１７】

重複排除ユニット情報応答部１３１は、インデックスサーバ５００からの重複排除ユニット情報取得要求に対して応答する。具体的には、重複排除ユニット情報応答部１３１は、インデックスサーバ５００から重複排除ユニット情報取得要求を受け取ると、メタデータ領域１１０ａから重複排除ボリューム用スライスメタデータ６１，６２，６３，・・・を取得する。そして重複排除ユニット情報応答部１３１は、取得した重複排除ボリューム用スライスメタデータ６１，６２，６３，・・・を、インデックスサーバ５００に送信する。

【０１１８】

重複排除ユニット割当部１３２は、インデックスサーバ５００からの指示に基づき、重複排除ユニットへのデータ実体領域１１０ｂ中の単位記憶領域の割り当て指示を行う。具体的には、重複排除ユニット割当部１３２は、重複排除ボリュームのセグメントに割り当てられたスライス内の重複排除ユニットを指定したデータ実体領域割り当て要求をインデックスサーバ５００から受け取る。すると、重複排除ユニット割当部１３２は、データ実体領域管理部１２３に対して、該当重複排除ユニットへのデータ実体領域の割り当てを依頼する。するとデータ実体領域管理部１２３により、未使用の単位記憶領域が選択され、重複排除ユニットに対して選択された単位記憶領域が割り当てられる。

【０１１９】

重複排除アドレス照会部１３３は、パトロール部１５０からの要求に応じて、インデッ

クスサーバ５００に対し、指定されたデータの重複排除アドレス（重複排除ボリュームのＩＤ＋重複排除オフセット）の照会を行う。具体的には、重複排除アドレス照会部１３３は、パトロール部１５０から、データのハッシュ値を受け取ると、インデックスサーバ５００に対して、ハッシュ値を含む重複排除アドレス照会要求を送信する。すると、インデックスサーバ５００から、ハッシュ値に対応付けられている重複排除アドレスが応答される。重複排除アドレス照会部１３３は、インデックスサーバ５００から送られた重複排除アドレスを、パトロール部１５０に渡す。

【０１２０】

重複排除ボリューム構成情報記憶部１４１は、重複排除ボリューム構成情報を記憶する。例えば、ＲＡＭ１０２またはＨＤＤ１０３の記憶領域の一部が、重複排除ボリューム構成情報記憶部１４１として使用される。重複排除ボリューム構成情報記憶部１４１には、図１１に示した重複排除ボリュームセグメント情報テーブル６５１と同様の情報が格納される。ディスクノード１００の重複排除ボリューム構成情報記憶部１４１には、例えば、重複排除ボリューム内のセグメントのうち、アクセスが行われたセグメントに関するメタデータのみが登録されている。なお、アクセスが行われたか否かに関係なく、すべてのセグメントに関するメタデータが重複排除ボリューム構成情報記憶部１４１に登録されている場合もある。

【０１２１】

重複排除データアクセス要求部１４２は、データアクセス部１２１からの要求に応じて、重複排除ユニットを管理するディスクノードに対し、その重複排除ユニットのデータを要求する。なお重複排除データアクセス要求部１４２は、重複排除ボリューム構成情報記憶部１４１を参照して、重複排除ユニットを管理するディスクノードを判断する。例えばディスクノード２００に対して重複排除ユニットのデータを要求した場合、ディスクノード２００の重複排除データアクセス応答部２４３によって、該当するデータが応答される。

【０１２２】

また、重複排除データアクセス要求部１４２は、重複排除ボリューム構成情報記憶部１４１内にアクセス対象の重複排除ユニットを含むセグメントのメタデータが格納されていない場合、該当するメタデータを制御ノード６００から取得する。そして重複排除データアクセス要求部１４２は、取得したメタデータを重複排除ボリューム構成情報記憶部１４１に格納する。

【０１２３】

重複排除データアクセス応答部１４３は、他のディスクノード２００，３００，４００から重複排除ユニットのデータが要求されると、ストレージ装置１１０から該当データを取得する。そして、重複排除データアクセス応答部１４３は、データを要求したディスクノードに対して取得したデータを送信する。

【０１２４】

例えば、ディスクノード２００においてディスクノード１００が管理する重複排除ユニットのデータへのアクセスが生じた場合、ディスクノード２００内の重複排除データアクセス要求部２４２からデータが要求される。重複排除データアクセス応答部１４３は、その要求に応じてデータを取得し、取得したデータを重複排除データアクセス要求部２４２に送信する。

【０１２５】

パトロール部１５０は、ディスクノード１００内で定期的にパトロールを実行する。パトロール部１５０は、各スライスに設定されている最終パトロール時刻から指定した時刻が経過したスライスを随時パトロールする。またパトロール部１５０は、未使用な単位記憶領域も随時パトロールする。すなわちパトロールは、ディスクノード１００管理下のストレージ装置１１０内の情報（各種メタデータ／データユニット／重複排除ユニット／未使用領域）が対象となる。具体的にはパトロール部１５０は、パトロール処理において、データユニットに関しては、ディスク保守的な読み書きを行う。ここで、ディスク保守的

な読み書きとは、対象データを読み出してデータのエラーの有無を判断し、エラーがなければ読み出したデータを元の位置に書き込む処理である。

【０１２６】

またパトロール部１５０は、アクセスユニット識別部１２２を用いて、パトロールを実行するユニットの状態を識別し、それに応じた処理を行う。パトロール対象のユニットへの最後の書き込みから所定時間以上経過していれば、そのユニットを重複排除化対象と判断する。その場合、パトロール部１５０は、重複排除アドレス照会部１３３を用いて、重複対象ユニットの重複排除アドレスを入手する。

【０１２７】

パトロール部１５０は、重複排除ユニットのパトロールに関しては、ディスク保守的な読み書きを行う。またパトロール部１５０は、保存期限満了時刻が過ぎている重複排除ユニットがある場合、重複排除ボリューム用スライスメタデータからそのユニットの情報をクリアする。さらにパトロール部１５０は、各種メタデータ、および未使用領域のパトロールでは、ディスク保守的な読み書きを行う。

【０１２８】

インデックスサーバ５００は、重複排除ユニット情報収集部５１１、重複排除ユニット情報記憶部５１２、重複排除ユニット割当要求部５１３、および重複排除アドレス検索部５１４を有する。

【０１２９】

重複排除ユニット情報収集部５１１は、各ディスクノード１００，２００，３００，４００から重複排除ユニットの情報を収集する。例えば、ディスクノード１００から重複排除ユニットの情報を収集する場合、重複排除ユニット情報収集部５１１は、ディスクノード１００の重複排除ユニット情報応答部１３１に対して、重複排除ユニット情報取得要求を送信する。すると重複排除ユニット情報応答部１３１から、ディスクノード１００が管理するストレージ装置１１０内の重複排除ボリューム用スライスメタデータ６１，６２，６３，・・・が応答される。重複排除ユニット情報収集部５１１は、収集した重複排除ユニットの情報（重複排除ボリューム用スライスメタデータ）を重複排除ユニット情報記憶部５１２に格納する。

【０１３０】

重複排除ユニット情報記憶部５１２は、重複排除ユニット情報収集部５１１が収集した重複排除ユニットの情報（重複排除ボリューム用スライスメタデータ）を記憶する記憶機能である。例えば、インデックスサーバ５００のＲＡＭやＨＤＤの記憶領域の一部が、重複排除ユニット情報記憶部５１２として使用される。なお重複排除ユニット情報記憶部５１２には、使用中の重複排除ユニットだけでなく未使用重複排除ユニットの情報も格納される。

【０１３１】

重複排除ユニット割当要求部５１３は、重複排除アドレス検索部５１４からの要求に応じて、ディスクノード１００，２００，３００，４００に対して、指定した重複排除ユニットへの単位記憶領域の割り当てを要求する。具体的には重複排除ユニット割当要求部５１３は、重複排除アドレス検索部５１４が選択した重複排除ユニットが属するスライスを管理するディスクノードに対して、重複排除ユニットへの単位記憶領域の割り当てを要求する。ディスクノードから割り当て完了の応答を受け取ると、重複排除ユニット割当要求部５１３は、割り当てが完了したことを重複排除アドレス検索部５１４に通知する。

【０１３２】

重複排除アドレス検索部５１４は、ディスクノード１００，２００，３００，４００からの重複排除アドレス照会に対して応答する。例えば重複排除アドレス検索部５１４はディスクノード１００の重複排除アドレス照会部１３３からハッシュ値を含む重複排除アドレス照会要求を取得すると、そのハッシュ値に対応する重複排除ユニット情報を、重複排除ユニット情報記憶部５１２から検索する。重複排除アドレス検索部５１４は、検索により重複排除ユニット情報がヒットすれば、ヒットした情報の重複排除アドレス（重複排除

ボリュームＩＤ＋オフセット）を、重複排除アドレス照会部１３３に応答する。検索によりヒットする情報がなければ、重複排除アドレス検索部５１４は、重複排除ユニット情報記憶部５１２を参照し、未使用の重複排除ユニットを選択する。次に重複排除アドレス検索部５１４は、重複排除ユニット割当要求部５１３に対して、選択した重複排除ユニットへの単位記憶領域の割り当てを要求する。そして重複排除アドレス検索部５１４は、新たに割り当てられた重複排除アドレスを、ディスクノード１００の重複排除アドレス照会部１３３に応答する。この際、重複排除アドレス検索部５１４は、重複排除ユニット情報記憶部５１２内の選択した重複排除ユニットの重複排除ユニットデータを更新する。

【０１３３】

なお、図１３には複数のディスクノード１００，２００，３００，４００のうち、ディスクノード１００の機能のみを詳細に示しているが、他のディスクノード２００，３００，４００もディスクノード１００と同様の機能を有している。

【０１３４】

図１４は、ディスクノードが有する重複排除ボリューム構成情報記憶部のデータ構造例を示す図である。重複排除ボリューム構成情報記憶部１４１には、重複排除ボリュームセグメント情報テーブル１４１ａが格納されている。重複排除ボリュームセグメント情報テーブル１４１ａのデータ構造は、図１１に示した制御ノード６００が有する重複排除ボリューム構成情報記憶部６５０内の重複排除ボリュームセグメント情報テーブル６５１と同様である。重複排除ボリュームセグメント情報テーブル１４１ａには、例えば、ディスクノードが過去にアクセスしたセグメントのメタデータ（セグメントメタデータ１４１ｂ）のみが登録されている。

【０１３５】

以上のような構成において、システム起動時には、重複排除ユニット情報収集部５１１によって、重複排除ユニット情報の収集が行われる。重複排除ユニット情報は、各ストレージ装置１１０，２１０，３１０，４１０のメタデータ領域に保存されている。

【０１３６】

図１５は、重複排除ユニット情報収集処理を示すシーケンス図である。以下、図１５に示す処理をステップ番号に沿って説明する。なお、図１５には２台のディスクノード１００，２００とインデックスサーバ５００との間の処理を示しているが、他のディスクノード３００，４００とインデックスサーバ５００との間でも同様の処理が実行される。

【０１３７】

［ステップＳ１１」インデックスサーバ５００が起動されると、重複排除ユニット情報収集部５１１は、すべてのディスクノード１００，２００，３００，４００に対して、重複排除ユニット情報を要求する。

【０１３８】

［ステップＳ１２］ディスクノード１００の重複排除ユニット情報応答部１３１は、重複排除ユニット情報をインデックスサーバ５００に応答する。具体的には重複排除ユニット情報応答部１３１は、ストレージ装置１１０のメタデータ領域１１０ａから、重複排除ユニット情報を含む重複排除ボリューム用スライスメタデータ６１，６２，６３，・・・を取得する。そして重複排除ユニット情報応答部１３１は、取得した重複排除ボリューム用スライスメタデータ６１，６２，６３，・・・をインデックスサーバ５００に送信する。

【０１３９】

［ステップＳ１３］ディスクノード２００もディスクノード１００と同様に、重複排除ユニット情報を含む重複排除ボリューム用スライスメタデータをインデックスサーバ５００に応答する。なお、図示していない他のディスクノード３００，４００からも、インデックスサーバ５００へ重複排除ボリューム用スライスメタデータが応答される。

【０１４０】

［ステップＳ１４］インデックスサーバ５００の重複排除ユニット情報収集部５１１は、ディスクノード１００，２００，３００，４００から送られた重複排除ユニット情報を

、重複排除ユニット情報記憶部５１２に格納する。

【０１４１】

このようにして、インデックスサーバ５００において、重複排除ユニットの情報が収集され、重複排除ユニット情報記憶部５１２に格納される。

図１６は、重複排除ユニット情報記憶部のデータ構造例を示す図である。重複排除ユニット情報記憶部５１２には、各ディスクノードから収集した重複排除ボリューム用スライスメタデータが、収集元のディスクノードのディスクノードＩＤに対応付けて格納されている。重複排除ボリューム用スライスメタデータの内容は、図７に示す通りである。すなわち、重複排除ユニット情報には、重複排除ボリュームＩＤが含まれると共に、重複排除ユニットごとの、データ実体オフセット、ハッシュ値、保存期限満了時刻が含まれる。

【０１４２】

このような重複排除ユニット情報記憶部５１２内の重複排除ボリューム用スライスメタデータから、ハッシュ値を検索キーとして重複排除ユニット情報を検索することで、そのハッシュ値の生成元となったデータが重複排除対象となっているか否かを判断できる。検索を行う重複排除アドレス検索部５１４は、例えばハッシュテーブルとツリー構造を組み合わせた構成で、ハッシュ値の検索を行うことができる。

【０１４３】

また、重複排除ユニット情報記憶部５１２を参照するとこで、未使用の重複排除ユニットを検出できる。未使用の重複排除ユニットには、重複排除ボリューム用スライスメタデータにおいて「ＮＵＬＬ」のデータが設定されている。重複排除ユニット割当要求部５１３は、重複排除ユニット情報記憶部５１２から未使用の重複排除ユニットを検索し、見つけ出した重複排除ユニットの新たな重複排除対象データへの割り当て要求処理を行う。なお、重複排除ユニット割当要求部５１３は、未使用の重複排除ユニットについて、例えばキュー管理することができる。すなわち、重複排除ユニット割当要求部５１３は、未使用の重複排除ユニットを示す重複排除ユニット情報へのポインタをキューにキューイングしておく。そして重複排除ユニット割当要求部５１３は、新たに重複排除ユニットの重複排除対象データへの割り当てが必要となると、キューの先頭のポインタで示される重複排除ユニットを割り当ての対象として選択する。また重複排除ユニット割当要求部５１３は、使用していた重複排除ユニットが未使用になると、その重複排除ユニットに関する重複排除ユニット情報へのポインタを、キューイングする。

【０１４４】

なお、システム起動時には、制御ノード６００のスライスメタデータ収集部６１０により、論理ボリューム用スライスメタデータと重複排除ボリューム用スライスメタデータとが収集される。そして、論理ボリューム管理部６３０によって、論理ボリューム構成情報と重複排除ボリューム構成情報とが作成される。

【０１４５】

システム起動直後は、アクセスノード７００，８００が論理ボリュームにアクセスする際には、アクセス対象のデータを含むセグメントのセグメントメタデータを制御ノード６００から取得する。そして、アクセスノード７００，８００は、取得したセグメントメタデータに基づいて、アクセス対象のデータを管理しているディスクノードのディスクノードＩＤを判断する。

【０１４６】

同様にシステム起動直後は、ディスクノード１００，２００，３００，４００から他のディスクノードが管理する重複排除対象データにアクセスする場合、アクセス対象のデータを含むセグメントのセグメントメタデータを制御ノード６００から取得する。そして、ディスクノード１００，２００，３００，４００は、取得したセグメントメタデータに基づいて、アクセス対象のデータを管理しているディスクノードのディスクノードＩＤを判断する。

【０１４７】

＜読み出し処理＞

次に、読み出し要求時の処理について説明する。

図１７は、重複排除ボリュームからのデータ読み出し処理の手順を示すシーケンス図である。以下、図１７に示す処理をステップ番号に沿って説明する。

【０１４８】

［ステップＳ２１］アクセスノード７００のデータアクセス要求部７２０から読み出し要求がディスクノード１００に対して送信される。読み出し要求には、例えば、論理ボリュームＩＤ、論理ボリュームの先頭からのアクセス対象データのオフセット、およびデータサイズが含まれる。

【０１４９】

［ステップＳ２２］ディスクノード１００では、データアクセス部１２１が読み出し要求を受信する。するとデータアクセス部１２１は、アクセスユニット識別部１２２に対してアクセス対象のユニットの状態の識別を依頼する。アクセスユニット識別部１２２は、論理ボリューム用スライスメタデータを参照し、アクセス対象データを含むユニットの状態を判断する。

【０１５０】

例えばアクセスユニット識別部１２２は、読み出し要求に示される論理ボリュームＩＤで示される論理ボリュームに割り当てられたスライスの論理ボリューム用スライスメタデータを抽出する。論理ボリューム用スライスメタデータにはセグメントＩＤが含まれている。１セグメント当たりの記憶容量は一定（例えば１ＧＢ）である。そのため、セグメントＩＤによって、該当スライスが割り当てられたセグメントの記憶領域の範囲（開始位置のオフセットと終了位置のオフセット）が判断できる。そこでアクセスユニット識別部１２２は、読み出し要求内のオフセットとデータサイズで示される領域を包含するセグメントに割り当てられたスライスの論理ボリューム用スライスメタデータを抽出する。さらにアクセスユニット識別部１２２は、抽出した論理ボリューム用スライスメタデータ内のデータユニット情報から、アクセス対象データが属するユニットのデータユニット情報を検索する。そして、アクセスユニット識別部１２２は、検索で該当したデータユニット情報を選択し、そのデータユニット情報の状態情報に基づきユニットの状態を判断する。判断結果は、アクセスユニット識別部１２２からデータアクセス部１２１に通知される。

【０１５１】

図１７の例では、状態が「DeDup」であったものとする。すなわち、アクセス対象データは、すでに重複排除対象データとなり、重複排除ボリュームで管理されている。状態が「DeDup」のデータユニット情報には、重複排除ボリュームＩＤと重複排除オフセットとが含まれている。状態が「DeDup」の場合、データアクセス部１２１から重複排除データアクセス要求部１４２に対して、重複排除ユニットへのアクセスが要求される。

【０１５２】

［ステップＳ２３］重複排除データアクセス要求部１４２は、読み出し要求に対応するデータユニット情報の重複排除ボリュームＩＤと重複排除オフセットとに基づいて、重複排除ボリュームセグメント情報テーブル１４１ａ（図１４参照）からセグメントメタデータを検索する。具体的には、重複排除データアクセス要求部１４２は、選択したデータユニット情報の重複排除ボリュームＩＤが設定された重複排除ボリュームセグメント情報テーブル１４１ａを、重複排除ボリューム構成情報記憶部１４１から抽出する。次に、重複排除データアクセス要求部１４２は、抽出した重複排除ボリュームセグメント情報テーブル１４１ａから、選択したデータユニット情報の重複排除オフセットを包含するオフセットが設定されたセグメントメタデータを検索する。該当するセグメントメタデータが存在する場合、処理はステップＳ２７に進められる。該当するセグメントメタデータが存在しない場合、処理がステップＳ２４に進められる。

【０１５３】

［ステップＳ２４］重複排除データアクセス要求部１４２は、メタデータ要求を制御ノード６００に送信する。メタデータ要求には、例えば、選択したデータユニット情報の重複排除オフセットが含められる。

【０１５４】
［ステップＳ２５］制御ノード６００の論理ボリューム管理部６３０は、セグメントメタデータを応答する。具体的には、制御ノード６００は、重複排除ボリューム構成情報記憶部６５０内の重複排除ボリュームセグメント情報テーブル６５１から、メタデータ要求で示されるセグメントのセグメントメタデータを抽出する。例えば、メタデータ要求において重複排除オフセットが示されている場合、その値を包含するオフセットが設定されたセグメントメタデータが抽出される。そして、論理ボリューム管理部６３０は抽出したセグメントメタデータを、ディスクノード１００に送信する。
【０１５５】
［ステップＳ２６］ディスクノード１００の重複排除データアクセス要求部１４２は、重複排除ボリューム構成情報記憶部１４１内の重複排除ボリュームセグメント情報テーブル１４１ａに、取得したセグメントメタデータを追加登録する。
【０１５６】
［ステップＳ２７］重複排除データアクセス要求部１４２は、重複排除対象データの読み出し要求を、そのデータを管理しているディスクノード（図１７の例では、ディスクノード２００）に対して送信する。具体的には重複排除データアクセス要求部１４２は、ステップＳ２３～ステップＳ２６の処理で取得したセグメントメタデータに基づいて、重複排除対象データを管理するディスクノードを判断する。そして、重複排除データアクセス要求部１４２は、重複排除対象データが属する重複排除ボリュームの重複排除ボリュームＩＤ、および重複排除オフセットを指定した読み出し要求を、重複排除対象データを管理するディスクノード２００に送信する。なお、重複排除ボリュームＩＤ、および重複排除オフセットは、ステップＳ２２で選択したデータユニット情報から取得される。
【０１５７】
［ステップＳ２８］ディスクノード２００の重複排除データアクセス応答部２４３は、重複排除対象データの読み出し要求を受信すると、該当する重複排除対象データをディスクノード１００に応答する。具体的には、重複排除データアクセス応答部２４３は、読み出し要求で指定された重複排除ボリュームＩＤが設定された重複排除ボリューム用スライスメタデータを、ストレージ装置２１０から取得する。次に重複排除データアクセス応答部２４３は、取得した重複排除ボリューム用スライスメタデータから、読み出し要求に示される重複排除オフセットに示されるユニットの重複排除ユニット情報を抽出する。さらに、重複排除データアクセス応答部２４３は、抽出した重複排除ユニット情報に示されるデータ実体オフセットで示されるデータ実体領域内の単位記憶領域から、重複排除対象データを読み出す。そして、重複排除データアクセス応答部２４３は、読み出した重複排除対象データをディスクノード１００に送信する。
【０１５８】
［ステップＳ２９］ディスクノード１００の重複排除データアクセス要求部１４２は、ディスクノード２００から取得した重複排除対象データをデータアクセス部１２１に渡す。データアクセス部１２１は、受け取った重複排除対象データを、読み取り要求のアクセス対象データとして、アクセスノード７００に送信する。
【０１５９】
このようにして、データの読み出し処理が実行される。データの読み出し処理において、読み出し対象のデータを含むユニットの状態が「DeDup」であれば、対応する重複排除対象データを管理するディスクノード２００からデータが読み出される。なお図１７の例では、アクセスノード７００からの読み出し要求を受けたディスクノード１００とは別のディスクノード２００から重複排除対象データが読み出されているが、ディスクノード１００において重複排除対象データを管理している場合もある。その場合、重複排除対象データアクセス要求部１４２は、ストレージ装置１１０から重複排除対象データを読み出す。
【０１６０】
また、図１７の例では、読み出し対象のデータを含むユニットの状態が「DeDup」の場

合を示している。該当ユニットの状態が「Blank」の場合には、データアクセス部１２１はアクセスノード７００に対して「０」の値を送信する。

【０１６１】

また、該当ユニットの状態が「Normal」の場合には、データアクセス部１２１は、データユニット情報のデータ実体オフセットに基づいて、そのデータユニットに割り当てられている単位記憶領域を判断する。そしてデータアクセス部１２１は、その単位記憶領域からデータを読み出し、アクセスノード７００に送信する。

【０１６２】

＜書き込み処理＞

次にデータの書き込み処理について説明する。

図１８は、状態が「Blank」のデータユニットへの書き込み処理を示すシーケンス図である。以下、図１８に示す処理をステップ番号に沿って説明する。

【０１６３】

［ステップＳ３１］アクセスノード７００のデータアクセス要求部７２０から書き込み要求がディスクノード１００に対して送信される。書き込み要求には、例えば、論理ボリュームＩＤ、論理ボリュームの先頭からのアクセス対象データのオフセット、および書き込むべきデータが含まれる。

【０１６４】

［ステップＳ３２］ディスクノード１００では、データアクセス部１２１が書き込み要求を受信する。するとデータアクセス部１２１は、アクセスユニット識別部１２２にアクセス対象データを含むユニットの状態の判断を依頼し、判断結果を受け取る。状態判断処理の詳細は、図１７のステップＳ２２の処理と同様である。

【０１６５】

図１８の例では、状態が「Blank」であったものとする。

［ステップＳ３３］データアクセス部１２１は、データ実体領域管理部１２３に対してデータ実体領域内の単位記憶領域の割り当てを依頼する。するとデータ実体領域管理部１２３が、ストレージ装置１１０のデータ実体領域１１０ｂから未使用の単位記憶領域を選択し、アクセス対象のユニットに割り当てる。すなわち、論理ボリューム用スライスメタデータ内の該当ユニットのデータユニット情報に、選択した単位記憶領域のデータ実体オフセットを設定する。

【０１６６】

［ステップＳ３４］データアクセス部１２１は、アクセス対象のデータユニットのデータユニット情報の状態を「Normal」に変更する。

［ステップＳ３５］データアクセス部１２１は、データ実体領域１１０ｂ内のアクセス対象のデータユニットに割り当てられた単位記憶領域内に、データを書き込む。

【０１６７】

［ステップＳ３６］データアクセス部１２１は、現在の時刻を、アクセス対象のデータユニットのデータユニット情報に最終書き込み時刻として設定する。

［ステップＳ３７］データアクセス部１２１は、書き込み完了応答をアクセスノード７００に送信する。

【０１６８】

このようにして、状態が「Blank」のデータユニットへの書き込みが行われる。

図１９は、状態が「DeDup」のデータユニットの一部への書き込み処理を示すシーケンス図である。以下、図１９に示す処理をステップ番号に沿って説明する。

【０１６９】

［ステップＳ４１］アクセスノード７００のデータアクセス要求部７２０から書き込み要求がディスクノード１００に対して送信される。

［ステップＳ４２］ディスクノード１００では、データアクセス部１２１が書き込み要求を受信する。するとデータアクセス部１２１は、アクセスユニット識別部１２２にアクセス対象データを含むユニットの状態の判断を依頼し、判断結果を受け取る。

【０１７０】
図１９の例では、状態が「DeDup」であったものとする。また、データの書き込み領域は、ユニットの一部（全体でない）であるものとする。
［ステップＳ４３］データアクセス部１２１は、重複排除データアクセス要求部１４２に対して重複排除対象データの読み出し要求を、そのデータを管理しているディスクノード（図１９の例では、ディスクノード２００）に対して送信する。この処理の詳細は、図１７のステップＳ２７の処理と同様である。
【０１７１】
［ステップＳ４４］ディスクノード２００の重複排除データアクセス応答部２４３は、重複排除対象データの読み出し要求を受信すると、該当する重複排除対象データをディスクノード１００に応答する。この処理の詳細は、図１７のステップＳ２８の処理と同様である。
【０１７２】
［ステップＳ４５］ディスクノード１００の重複排除データアクセス要求部１４２は、ディスクノード２００から取得した重複排除対象データをデータアクセス部１２１に渡す。ディスクノード１００のデータアクセス部１２１は、データ実体領域管理部１２３に対してデータ実体領域内の単位記憶領域の割り当てを依頼する。するとデータ実体領域管理部１２３が、ストレージ装置１１０のデータ実体領域１１０ｂから未使用の単位記憶領域を選択し、アクセス対象のユニットに割り当てる。
【０１７３】
［ステップＳ４６］データアクセス部１２１は、アクセス対象のデータユニットのデータユニット情報の状態を「Normal」に変更する。
［ステップＳ４７］データアクセス部１２１は、取得した重複排除対象データを書き込み要求に応じて更新し、データ実体領域１１０ｂ内のアクセス対象のデータユニットに割り当てられた単位記憶領域内に、更新後のデータ実体を書き込む。
【０１７４】
［ステップＳ４８］データアクセス部１２１は、現在の時刻を、アクセス対象のデータユニットのデータユニット情報に最終書き込み時刻として設定する。
［ステップＳ４９］データアクセス部１２１は、書き込み完了応答をアクセスノード７００に送信する。
【０１７５】
このようにして、状態が「DeDup」のデータユニットへの書き込みが行われる。
図１８，図１９に示した状態が「Blank」または「DeDup」のデータユニットへの書き込み処理以外に、状態が「Normal」のデータユニットへの書き込み処理がある。状態が「Normal」のデータユニットへの書き込み処理では、既にデータユニットに割り当てられているデータ実体領域内の単位記憶領域に対して、データの書き込みが行われる。
【０１７６】
また状態が「DeDup」のデータユニットへの書き込みの場合、データユニット内の全データの更新書き込みが行われる場合がある。その場合、既存の重複排除対象データの読み出しを行う必要はない。書き込み処理におけるこれらの条件判断を含むディスクノードでの書き込み処理を以下に示す。
【０１７７】
図２０は、ディスクノードにおける書き込み処理を示すフローチャートである。以下、図２０に示す処理をステップ番号に沿って説明する。この処理は、アクセスノード７００からの書き込み要求を受信した際に、ディスクノード１００で実行される。
【０１７８】
［ステップＳ５１］データアクセス部１２１は、アクセスユニット識別部１２２の機能を用い、アクセス対象のデータユニットの状態を判断する。状態が「Normal」であれば、処理がステップＳ５５に進められる。状態が「Blank」であれば、処理がステップＳ５３に進められる。状態が「DeDup」であれば、処理がステップＳ５２に進められる。

【０１７９】
［ステップＳ５２］データアクセス部１２１は、書き込みデータがユニット全体に対するデータか否かを判断する。ユニット全体に対するデータであれば、処理がステップＳ５３に進められる。ユニット全体に対するデータでなければ、処理がステップＳ５６に進められる。
【０１８０】
［ステップＳ５３］データアクセス部１２１は、データ実体領域管理部１２３に対してアクセス対象のデータユニットに対する単位記憶領域の割り当てを依頼する。するとデータ実体領域管理部１２３がデータ実体領域１１０ｂから未使用の単位記憶領域を選択し、データユニットに割り当てる。
【０１８１】
［ステップＳ５４］データアクセス部１２１は、アクセス対象のデータユニットのデータユニット情報の状態を「Normal」に変更する。
［ステップＳ５５］データアクセス部１２１は、書き込み要求に含まれるデータを、データ実体領域１１０ｂ内のアクセス対象のデータユニットに割り当てられた単位記憶領域内に書き込む。その後、処理がステップＳ６０に進められる。
【０１８２】
［ステップＳ５６］データアクセス部１２１は、重複排除データアクセス要求部１４２を介してデータユニット全体の重複排除対象データを取得する。
［ステップＳ５７］データアクセス部１２１は、データ実体領域管理部１２３に対してアクセス対象のデータユニットに対する単位記憶領域の割り当てを依頼する。するとデータ実体領域管理部１２３がデータ実体領域１１０ｂから未使用の単位記憶領域を選択し、データユニットに割り当てる。
【０１８３】
［ステップＳ５８］データアクセス部１２１は、アクセス対象のデータユニットのデータユニット情報の状態を「Normal」に変更する。
［ステップＳ５９］データアクセス部１２１は、取得した重複排除対象データを書き込み要求に応じて更新する。そしてデータアクセス部１２１は、データ実体領域１１０ｂ内のアクセス対象のデータユニットに割り当てられた単位記憶領域内に、更新後のデータ実体を書き込む。
【０１８４】
［ステップＳ６０］データアクセス部１２１は、現在の時刻を、アクセス対象のデータユニットのデータユニット情報に最終書き込み時刻として設定する。
［ステップＳ６１］データアクセス部１２１は、書き込み完了応答をアクセスノード７００に送信する。
【０１８５】
以上のようにして、アクセス対象のデータユニットの状態に応じた適切な書き込み処理が実行される。
＜パトロール処理＞
次に、パトロール処理について詳細に説明する。パトロール処理では、論理ボリューム内のデータユニットから重複排除対象データを検出し、その重複排除対象データを重複排除ボリュームに移動する処理が行われる。さらにパトロール処理では、重複排除ボリューム内の重複排除ユニットから、論理ボリューム内のデータユニットとの関連性を失った重複排除ユニットを検出し、その重複排除対象データを削除する。なおパトロール処理は、スライス内のデータユニット、およびデータ実体領域１１０ｂ内の未使用の単位記憶領域に対して、所定のパトロール時間ごとに定期的に実行される。
【０１８６】
図２１は、状態が「Normal」のデータユニットへのパトロール処理を示すシーケンス図である。以下、図２１に示す処理をステップ番号に沿って説明する。なお、図２１のシーケンス図は、スライス内のデータユニットに対するパトロールの例である。

【０１８７】
［ステップＳ７１］パトロール部１５０は、パトロール対象のデータユニットに割り当てられた単位記憶領域からデータ実体を読み出す。
［ステップＳ７２］パトロール部１５０は、最終書き込み時刻からの経過時間を判定する。具体的には、パトロール部１５０は、パトロール対象のユニットのデータユニット情報に設定されている最終書き込み時刻から、現在時刻までの経過時間を算出する。そしてパトロール部１５０は、経過時間が予め設定されている所定の重複排除化猶予期間を超えているか否かを判断する。最終書き込みから重複排除化猶予期間を経過している場合、パトロール対象のデータユニットへのデータの書き込み頻度が低いことを示す。その場合、パトロール部１５０は、パトロール対象のデータユニット内のデータを重複排除対象データとする。図２１の例では、経過時間が重複排除化猶予期間以上経過しているものとする。
【０１８８】
［ステップＳ７３］パトロール部１５０は、所定のハッシュ関数を用いて、ステップＳ７１で読み出したデータのハッシュ値を算出する。
［ステップＳ７４］パトロール部１５０は、インデックスサーバ５００に対して重複排除アドレス照会を行う。具体的には、パトロール部１５０は、重複排除アドレス照会部１３３に対して、ステップＳ７３で算出したハッシュ値を指定した重複排除アドレスの照会を依頼する。すると、重複排除アドレス照会部１３３は、インデックスサーバ５００に対して、ハッシュ値を含む重複排除アドレス照会要求を送信する。
【０１８９】
［ステップＳ７５］インデックスサーバ５００の重複排除アドレス検索部５１４は、重複排除アドレス照会要求に示されるハッシュ値が設定された重複排除ユニット情報を、重複排除ユニット情報記憶部５１２から検索する。図２１の例では該当する重複排除ユニット情報が未登録であり、検索により何もヒットしないものとする。
【０１９０】
［ステップＳ７６］インデックスサーバ５００の重複排除アドレス検索部５１４は重複排除ユニット情報記憶部５１２から未使用の重複排除ユニットを選択し、重複排除ユニット割当要求部５１３に対して、選択した重複排除ユニットへの単位記憶領域の割当を依頼する。すると、重複排除ユニット割当要求部５１３は、選択された重複排除ユニットが属するスライスを管理するディスクノード（図２１の例ではディスクノード２００）に対して、単位記憶領域の割当を要求する。
【０１９１】
［ステップＳ７７］ディスクノード２００は、インデックスサーバ５００からの要求に応じて重複排除ユニットに対して単位記憶領域を割り当てる。
［ステップＳ７８］ディスクノード２００は、重複排除ユニットへの単位記憶領域の割当完了を応答する。
【０１９２】
［ステップＳ７９］インデックスサーバ５００の重複排除ユニット割当要求部５１３は、割当が完了したことを重複排除アドレス検索部５１４に伝える。すると重複排除アドレス検索部５１４は、重複排除ユニット情報記憶部５１２内の新たに単位記憶領域を割り当てた重複排除ユニットの重複排除ユニット情報を更新する。更新された重複排除ユニット情報には、少なくともディスクノード１００から送られたハッシュ値が設定される。
【０１９３】
［ステップＳ８０］重複排除アドレス検索部５１４は、重複排除アドレス（論理ボリュームＩＤと重複排除オフセット）をディスクノード１００に応答する。この際、重複排除アドレス検索部５１４は、新たに重複排除ユニット割当処理（ステップＳ７６～Ｓ７９）が行われたことを示す情報を、重複排除アドレスに付加する。
【０１９４】
［ステップＳ８１］ディスクノード１００の重複排除アドレス照会部１３３は、インデ

ックスサーバ５００から送られた重複排除アドレスをパトロール部１５０に渡す。パトロール部１５０は、重複排除ユニット割当処理が行われたことを認識し、重複排除データアクセス要求部１４２に対して、取得した重複排除アドレスを指定して重複排除対象データの書き込みを要求する。すると、重複排除データアクセス要求部１４２が、ディスクノード２００に対して、重複排除オフセットを書き込み位置として指定して、重複排除対象データを送信する。

【０１９５】

［ステップＳ８２］ディスクノード２００の重複排除データアクセス応答部２４３は、ディスクノード１００から送られた重複排除対象データを、指定された重複排除対象オフセットに対応する重複排除ユニットに割り当てられた単位記憶領域に書き込む。この時、重複排除データアクセス応答部２４３は、重複排除ユニットの重複排除ユニット情報に、現在の時刻から所定の保存期間経過後の保存期限満了時刻を設定する。

【０１９６】

［ステップＳ８３］ディスクノード２００の重複排除データアクセス応答部２４３は、ディスクノード１００に対して書き込み完了を応答する。

［ステップＳ８４］ディスクノード１００の重複排除データアクセス要求部１４２は、データ書き込みの完了をパトロール部１５０に伝える。すると、パトロール部１５０は、重複排除対象データが格納されていたデータユニットのデータユニット情報における状態を、「DeDup」に変更する。

【０１９７】

［ステップＳ８５］パトロール部１５０は、重複排除対象データが格納されていたデータユニットに割り当てられていたデータ実体領域１１０ｂ内の単位記憶領域を解放する。すなわち、そのデータユニットのデータユニット情報に設定されていた最終書き込み時刻とデータ実体オフセットが削除され、代わりに重複排除アドレス（重複排除ボリュームＩＤと重複排除オフセット）が設定される。

【０１９８】

［ステップＳ８６］パトロール部１５０は、重複排除対象データが格納されていたデータユニットに割り当てられていたデータ実体領域１１０ｂ内の単位記憶領域に対して、初期データを書き込む。

【０１９９】

このようにして、重複排除化猶予期間以上書き込みが行われていないデータユニット内のデータは、重複排除対象データとして重複排除ボリュームに移される。

なお図２１の例では、ステップＳ７５の処理で該当する重複排除ユニット情報が未登録であると判断されたため、ステップＳ７６～ステップＳ７９の重複排除ユニット割当処理が実行されている。そして重複排除ユニット割当処理が実行されたことにより、ステップＳＳ８１～ステップＳ８２のデータ転送処理が行われている。他方、ステップＳ７５の検索で重複排除ユニット情報がヒットし、重複排除ユニット割当処理が行われない場合には、ディスクノード１００からディスクノード２００へのデータ転送も行われない。

【０２００】

図２２は、状態が「DeDup」のデータユニットへのパトロール処理を示すシーケンス図である。以下、図２２に示す処理をステップ番号に沿って説明する。

［ステップＳ９１］ディスクノード１００のパトロール部１５０は、パトロール対象の単位記憶領域が割り当てられているデータユニットの状態が「DeDup」の場合、重複排除データアクセス要求部１４２に対して重複排除対象データの参照を要求する。すると重複排除データアクセス要求部１４２は、重複排除対象データを管理するディスクノード２００に対して重複排除対象データの参照要求を送信する。

【０２０１】

［ステップＳ９２］ディスクノード２００の重複排除データアクセス応答部２４３は、参照要求に応じて重複排除対象データを取得する。この際、重複排除データアクセス応答部２４３は、重複排除対象データが格納された重複排除ユニットの保存期限満了時刻を、

現在の時刻からら所定の保存期間経過後の保存期限満了時刻に更新する。

【０２０２】

［ステップＳ９３］重複排除データアクセス応答部２４３は、ディスクノード１００に対して取得した重複排除対象データを応答する。

このようにして、状態が「DeDup」のデータユニットに対するパトロールが行われるごとに、対応する重複排除対象データが格納された重複排除ユニットの保存期限満了時刻が更新される。パトロール間隔は、重複排除対象データの保存期間よりも短い。そのため、重複排除ユニットの重複排除オフセットが設定されたデータユニットが少なくとも１つ存在する間は、重複排除対象データは削除されない。

【０２０３】

以上のようなパトロールを実現するためのパトロール部１５０の処理を、以下に具体的に説明する。なお、同一のスライスに属するユニットに割り当てられた単位記憶領域については、そのスライスの最終パトロール時刻から所定のパトロール周期に相当する時間経過後に実行される。スライス内のパトロールが終了すると、パトロール部１５０によってそのスライスの最終パトロール時刻が更新される。

【０２０４】

図２３は、パトロール処理の手順を示すフローチャートである。以下、図２３に示す処理をステップ番号に沿って説明する。なお、以下の処理は、各スライスのデータユニット、およびデータ実体領域１１０ｂ内の未使用の単位記憶領域それぞれに対して、所定の時間間隔で実行される。

【０２０５】

［ステップＳ９４］パトロール部１５０は、パトロール対象がスライス内の使用領域として定義されたデータユニットか、未使用の単位記憶領域かを判断する。スライス内の領域として定義されたデータユニットに対するパトロールであれば、処理がステップＳ９５に進められる。未使用の単位記憶領域に対するパトロールであれば、処理がステップＳ１００に進められる。

【０２０６】

［ステップＳ９５］パトロール部１５０は、パトロール対象のユニットが属するスライスが割り当てられたボリュームのタイプ（論理ボリュームか重複排除ボリュームか）を判断する。論理ボリュームに割り当てられている場合、処理がステップＳ９６に進められる。重複排除ボリュームに割り当てられている場合、処理がステップＳ９９に進められる。

【０２０７】

［ステップＳ９６］パトロール対象のデータユニットが属するスライスが、論理ボリュームに割り当てられていれば、パトロール部１５０はデータユニットの状態を判断する。データユニットの状態は、そのデータユニットに対応するデータユニット情報内に設定されている。データユニットの状態が「Blank」であれば、処理が終了する。データユニットの状態が「DeDup」であれば、処理がステップＳ９７に進められる。データユニットの状態が「Normal」であれば、処理がステップＳ９８に進められる。

【０２０８】

［ステップＳ９７］データユニットの状態が「DeDup」の場合、パトロール部１５０は、重複排除対象データを管理しているディスクノードに対して、重複排除対象データの参照要求を送信する。具体的にはパトロール部１５０は、重複排除データアクセス要求部１４２に対して、パトロール対象のデータユニットに対応する重複排除対象データの取得を依頼する。すると、重複排除データアクセス要求部１４２により、データユニット情報に示される重複排除ボリュームＩＤと重複排除オフセットとに基づいて、重複排除対象データを管理するディスクノードから重複排除対象データが取得される。そして、取得した重複排除対象データが、重複排除データアクセス要求部１４２からパトロール部１５０に渡される。その後、処理が終了する。

【０２０９】

［ステップＳ９８］データユニットの状態が「Normal」の場合、パトロール部１５０は

、Normalユニットパトロール処理を実行する。この処理の詳細は後述する。Normalユニットパトロール処理の後、パトロール処理が終了する。

【０２１０】

［ステップＳ９９］パトロール対象のデータユニットが重複排除ボリュームに割り当てられたスライスのユニット（重複排除ユニット）である場合、重複排除ユニットパトロール処理が実行される。重複排除ユニットパトロール処理の後、パトロール処理が終了する。

【０２１１】

［ステップＳ１００］パトロール対象が未使用の単位記憶領域である場合、パトロール部１５０は、その単位記憶領域に初期データ（例えば「０」の連続）を書き込む。その後、パトロール処理が終了する。

【０２１２】

図２４は、状態が「Normal」のデータユニットのパトロール処理の手順を示すフローチャートである。以下、図２４に示す処理をステップ番号に沿って説明する。

［ステップＳ１０１］パトロール部１５０は、データユニットに割り当てられた単位記憶領域からデータを読み出す。

【０２１３】

［ステップＳ１０２］パトロール部１５０は、データユニットの最終書き込み時刻から重複排除化猶予期間以上経過したか否かを判断する。重複排除化猶予期間以上経過している場合、処理がステップＳ１０４に進められる。重複排除化猶予期間を経過していなければ、処理がステップＳ１０３に進められる。

【０２１４】

［ステップＳ１０３］パトロール部１５０は、重複排除化猶予期間を経過していなければ、読み出したデータを、データユニットに割り当てられた単位記憶領域に書き込む。その後、処理が終了する。

【０２１５】

［ステップＳ１０４］パトロール部１５０は、重複排除化猶予期間を経過していれば、ステップＳ１０１で読み出したデータのハッシュ値を算出する。

［ステップＳ１０５］パトロール部１５０は、重複排除アドレス照会部１３３を用いて重複排除アドレスの照会を行う。これにより、パトロール部１５０は、重複排除アドレスを取得する。

【０２１６】

［ステップＳ１０６］パトロール部１５０は、ステップＳ１０５の重複排除アドレスの照会に応答して、インデックスサーバ５００において重複排除ユニット割当処理が実行されたか否かを判断する。重複排除ユニット割当処理の実行の有無は、重複排除アドレスの照会に対するインデックスサーバ５００からの応答で示される。重複排除ユニット割当処理が新たに実行された場合、処理がステップＳ１０７に進められる。重複排除ユニット割当処理が実行されていない場合、処理がステップＳ１０８に進められる。

【０２１７】

［ステップＳ１０７］パトロール部１５０は、重複排除データアクセス要求部１４２を用いて、ステップＳ１０１で読み出したデータを、ステップＳ１０５で取得した重複排除アドレスで示される重複排除ユニットに転送する。

【０２１８】

［ステップＳ１０８］パトロール部１５０は、データユニットの状態を「DeDup」に変更する。

［ステップＳ１０９］パトロール部１５０は、データユニットに割り当てられた単位記憶領域を解放する。

【０２１９】

［ステップＳ１１０］パトロール部１５０は、解放した単位記憶領域に初期データを書き込む。

このようにして、論理ボリューム内の状態が「Normal」のデータユニットのパトロールが行われる。

【０２２０】

図２５は、重複排除ボリュームのパトロール処理の手順を示すフローチャートである。以下、図２５に示す処理をステップ番号に沿って説明する。

［ステップＳ１１１］パトロール部１５０は、重複排除ユニットに割り当てられた単位記憶領域からデータを読み出す。

【０２２１】

［ステップＳ１１２］パトロール部１５０は、現在の時刻が重複排除ユニットの保存期限満了時刻を過ぎているか否かを判断する。保存期限満了時刻を過ぎている場合、処理がステップＳ１１４に進められる。保存期限満了時刻を過ぎていない場合、処理がステップＳ１１３に進められる。

【０２２２】

［ステップＳ１１３］パトロール部１５０は、保存期限満了時刻を過ぎていなければ、ステップＳ１１１で読み出したデータを、重複排除ユニットに割り当てられた単位記憶領域に書き込む。その後、重複排除ボリュームパトロール処理が終了する。

【０２２３】

［ステップＳ１１４］パトロール部１５０は、保存期限満了時刻を過ぎている場合、重複排除ユニットを未使用（ＮＵＬＬ）にする。すなわち、パトロール部１５０は、保存期限満了時刻を過ぎている重複排除ユニットの重複排除ユニット情報に対し、「ＮＵＬＬ」を設定する。

【０２２４】

［ステップＳ１１５］パトロール部１５０は、重複排除ユニットに割り当てられている単位記憶領域を解放する（割り当てを解除する）。

［ステップＳ１１６］パトロール部１５０は、重複排除ユニットに割り当てられていた単位記憶領域に、初期データを書き込む。その後、重複排除ボリュームパトロール処理が終了する。

【０２２５】

このようにして、パトロール部１５０によるパトロール処理が実行される。

次に、パトロール処理の過程で重複排除アドレス照会要求が出された場合の、インデックスサーバ５００の処理を詳細に説明する。

【０２２６】

図２６は、インデックスサーバにおける重複排除アドレス検索処理の手順を示すフローチャートである。以下、図２６に示す処理をステップ番号に沿って説明する。

［ステップＳ１２１］重複排除アドレス検索部５１４は、ハッシュ値を含む重複排除アドレス照会要求を受信する。

【０２２７】

［ステップＳ１２２］重複排除アドレス検索部５１４は、重複排除ユニット情報記憶部５１２から、重複排除アドレス照会要求に示されるハッシュ値と一致するハッシュ値を含む重複排除ユニットを検索する。

【０２２８】

［ステップＳ１２３］重複排除アドレス検索部５１４は、該当する重複排除ユニットがあるか否かを判断する。該当する重複排除ユニットがあれば、処理がステップＳ１２７に進められる。該当する重複排除ユニットがなければ、処理がステップＳ１２４に進められる。

【０２２９】

［ステップＳ１２４］重複排除アドレス検索部５１４は、現在定義されている重複排除ボリューム内に未使用の重複排除ユニットがあるか否かを判断する。未使用の重複排除ユニットがあれば、処理がステップＳ１２６に進められる。未使用の重複排除ユニットがなければ、処理がステップＳ１２５に進められる。

【０２３０】
［ステップＳ１２５］重複排除アドレス検索部５１４は、重複排除ボリュームを拡張する。例えば、重複排除アドレス検索部５１４は、制御ノード６００に対して重複排除ボリュームへのセグメントの追加を依頼する。制御ノード６００では、論理ボリューム管理部６３０が重複排除ボリュームにセグメントを追加し、そのセグメントにいずれかのディスクボリュームで管理されているスライスを割り当て、重複排除ボリューム用スライスメタデータを生成する。生成された重複排除ボリューム用スライスメタデータは、割り当てられたスライスを管理するディスクノードに送信されると共に、インデックスサーバ５００に送信される。インデックスサーバ５００に送信される重複排除ボリューム用スライスメタデータには、そのスライスを管理するディスクノードのディスクノードＩＤが付与されている。インデックスサーバ５００は、取得した重複排除ボリューム用スライスメタデータを重複排除ユニット情報記憶部５１２に格納する。新たに作成された重複排除ボリューム用スライスメタデータに含まれる重複排除ユニット情報は、すべて未使用である。すなわち、未使用の重複排除ユニットが生成されたこととなる。
【０２３１】
［ステップＳ１２６］重複排除アドレス検索部５１４は、未使用の重複排除ユニットを、新たな重複排除対象データの格納先として割り当てる。
［ステップＳ１２７］重複排除アドレス検索部５１４は、重複排除アドレスをディスクノードに応答する。ステップＳ１２３で該当する重複排除ユニットが見つかった場合、その重複排除ユニットの重複排除アドレスが送信される。ステップＳ１２３で該当する重複排除ユニットが見つからなかった場合、ステップＳ１２６で割り当てた重複排除ユニットの重複排除アドレスが送信される。
【０２３２】
このようにして、重複排除アドレス照会要求に応じた重複排除アドレス検索処理が行われる。ここで、未使用の重複排除ユニットが存在しなければ、適宜、重複排除ボリュームの拡張が行われる。すなわち、重複排除ボリュームの記憶容量は、初期状態では必要最小限の容量にとどめておき、重複排除ユニットの不足が発生した場合に拡張される。その結果、システム全体における資源の有効利用が可能となる。
【０２３３】
以上のようにしてパトロールが実行される。その結果、ディスクノード１００において最後の書き込みから重複排除化猶予期間を経過した論理ボリューム内のデータについては、データ実体が重複排除ボリュームに移される。論理ボリュームには、移行先の重複排除ユニットを示す重複排除アドレスが残される。
【０２３４】
またパトロールにより重複排除ボリュームから、保存期限満了時刻を過ぎた重複排除ユニットが検出されると、その重複排除ユニットのデータは削除され、データ実体領域内の単位記憶領域が解放される。
【０２３５】
＜未使用ユニット情報反映処理＞
パトロールの実行により、各ディスクノード１００，２００，３００，４００で管理されている重複排除ユニットについては、データユニットからの関連付けが無くなった場合に未使用に変更される。重複排除ユニットが未使用に変更されたことは、インデックスサーバ５００の重複排除ユニット情報記憶部５１２にも反映される。
【０２３６】
図２７は、未使用ユニット情報反映処理の手順を示すフローチャートである。以下、図２７に示す処理をステップ番号に沿って説明する。なお、以下の処理は、重複排除ユニット情報記憶部５１２に格納されている重複排除ユニット情報ごとに実行される。
【０２３７】
［ステップＳ１３１］インデックスサーバ５００の重複排除ユニット情報収集部５１１は、現在の時刻が保存期限満了時刻を過ぎているか否かを判断する。保存期限満了時刻を

経過した場合、処理がステップＳ１３２に進められる。保存期限満了時刻を経過していなければ、ステップＳ１３１の処理が繰り返される。

【０２３８】

［ステップＳ１３２］重複排除ユニット情報収集部５１１は、処理対象の重複排除ユニット情報の取得元であるディスクノードに対して、その重複排除ユニット情報で示される重複排除ユニットの現在の状況を問い合わせる。なお問い合わせ対象の重複排除ユニットは、重複排除ボリュームＩＤと、重複排除ボリューム内でのオフセット（重複排除オフセット）によって一意に指定される。この問い合わせに対して、ディスクノードの重複排除ユニット情報応答部が、現在の状況を応答する。応答には、現在の重複排除ユニット情報が含まれる。

【０２３９】

［ステップＳ１３３］重複排除ユニット情報収集部５１１は、ディスクノードから受け取った応答内容により、重複排除ユニットの状態を判断する。具体的には、重複排除ユニット情報収集部５１１は、取得した重複排除ユニット情報が「ＮＵＬＬ」であれば、該当する重複排除ユニットは未使用であると判断する。重複排除ユニット情報収集部５１１は、取得した重複排除ユニット情報が「ＮＵＬＬ」でなければ、該当する重複排除ユニットは使用中であると判断する。重複排除ユニットが未使用であれば、処理がステップＳ１３４に進められる。重複排除ユニットが使用中であれば、処理がステップＳ１３６に進められる。

【０２４０】

［ステップＳ１３４］重複排除ユニット情報収集部５１１は、ディスクノードにおける重複排除ユニットの状態が未使用であれば、重複排除ユニット情報記憶部５１２内の対応する重複排除ユニットの状態も未使用にする。具体的には、重複排除ユニット情報収集部５１１は、重複排除ユニット情報をＮＵＬＬに変更する。

【０２４１】

［ステップＳ１３５］重複排除ユニット情報収集部５１１は、ステップＳ１３４で未使用に変更した重複排除ユニット情報へのポインタを、未使用ユニットキューに追加する。その後、処理が終了する。

【０２４２】

［ステップＳ１３６］重複排除ユニット情報収集部５１１は、ディスクノードにおける重複排除ユニットの状態が使用中であれば、ディスクノードから取得した重複排除ユニット情報から保存期限満了時刻を取得する。

【０２４３】

［ステップＳ１３７］重複排除ユニット情報収集部５１１は、重複排除ユニット情報記憶部５１２内の対応する重複排除ユニットの保存期限満了時刻を更新する。その後、処理が終了する。

【０２４４】

このようにして、ディスクノード１００，２００，３００，４００における重複排除ユニット情報の更新内容が、インデックスサーバ５００に反映される。すなわちインデックスサーバにおいて重複排除ユニット情報の保存期限満了時刻が過ぎた場合、重複排除ユニットを管理しているディスクノードから最新の重複排除ユニット情報が取得される。そして、インデックスサーバ５００では、取得した重複排除ユニット情報の内容が、重複排除ユニット情報記憶部５１２に反映される。

【０２４５】

次に、パトロールによるデータの保存先の変更例について説明する。

図２８は、パトロールによるデータの保存先変更状況を示す図である。図２８の例では、論理ボリューム２０のセグメントＩＤ「2」のセグメントにディスクノード１００が管理するスライス１１１が割り当てられている。また論理ボリューム３０のセグメントＩＤ「3」のセグメントにディスクノード２００が管理するスライス２１１が割り当てられている。さらに重複排除ボリューム４０のセグメントＩＤ「1」のセグメントにディスクノ

ード３００が管理するスライス３１１が割り当てられている。

【０２４６】

スライス１１１の先頭から３つのデータユニットに割り当てられた単位記憶領域には、それぞれ「data[a]」、「data[b]」、「data[c]」が格納されている。スライス２１１の先頭から３つのデータユニットに割り当てられた単位記憶領域には、それぞれ「data[u]」、「data[a]」、「data[v]」が格納されている。すなわち、スライス１１１の先頭のデータユニットと、スライス２１１の２番目のデータユニットとは、格納されているデータの内容が同じである。

【０２４７】

また、スライス１１１の２番目のデータユニットとスライス２１１の１番目のデータユニットとは、頻繁に書き込みが行われ、最後の書き込みから重複排除化猶予期間が経過していないものとする。その他のデータユニットは、重複排除化猶予期間以上書き込みが行われていない。

【０２４８】

このような状況下でパトロールが実行されると、重複排除化猶予期間以上書き込みが行われていないデータユニット内のデータが重複排除対象データとして選択され、重複排除ボリューム４０に移動される。図２８の例では、「data[a]」が重複排除ボリューム４０の先頭の重複排除ユニットに移動され、「data[c]」が重複排除ボリューム４０の２番目の重複排除ユニットに移動され、「data[v]」が重複排除ボリューム４０の３番目の重複排除ユニットに移動されている。

【０２４９】

データの移動が行われた移動元のデータユニットには、移動先の重複排除ユニットを一意に示す重複排除アドレスが設定される。図２８では、重複排除ボリュームＩＤの後に、重複排除ユニットのオフセットを示している。図２８の例では、重複排除ボリュームの最初のセグメント（セグメントＩＤ「１」）の重複排除ユニットが使用されているため、そのセグメント内での各重複排除ユニットの順番がオフセット値となる。

【０２５０】

ハッシュ値は、データの内容が同一であれば同じ値となる。図２８の例では、「data[a]」のハッシュ値は「hash[A]」、「data[c]」のハッシュ値は「hash[C]」、「data[v]」のハッシュ値は「hash[V]」である。

【０２５１】

データを移動する際に算出したハッシュ値が既に重複排除ユニットに設定されていれば、その重複排除ユニットの重複排除アドレスがデータ移動元のデータユニットに設定される。図２８の例では「data[a]」については、パトロール前は２つの論理ボリューム２０，３０の両方に重複して格納されていたが、パトロール後は、「data[a]」は重複排除ボリューム４０に１つだけ格納されている。「data[a]」が格納されていたデータユニットには、同じ重複排除ユニットを示す重複排除アドレスが設定される。これにより、システム全体で保持すべきデータ量を削減し、資源の効率的利用が可能となる。

【０２５２】

論理ボリュームにおいて重複排除対象データと判断され重複排除ボリュームに移行されたデータであっても、アクセスノードからのアクセスによりデータが更新されることがある。その場合には、更新後のデータがデータユニットに割り当てられた単位記憶領域に格納され、重複排除ユニットとの関連付けは解消される。

【０２５３】

図２９は、重複排除ユニットへの関連付けの解消状況を示す図である。図２９の例では、論理ボリューム３０のセグメントＩＤ「３」のセグメントに対する書き込みが行われた場合を想定している。具体的には、そのセグメントに割り当てられたスライス２１１の２番目と３番目のデータユニットのデータが、それぞれ「data[x]」、「data[y]」に更新されている。

【０２５４】

この場合、更新されたデータユニットに割り当てられた単位記憶領域に「data[x]」、「data[y]」が格納され、各データユニットの重複排除アドレスは削除される。重複排除アドレスが削除されたことで、重複排除ユニットとの関連付けは解消される。

【０２５５】

さらに、すべてのデータユニットからの関連付けが解消された重複排除ユニットについては、保持していたデータが削除され未使用の状態に変更される。

図３０は、重複排除ユニットの未使用状態への変更状況を示す図である。図３０の例では、ディスクノード１００でパトロールが行われた後に、ディスクノード３００でパトロールが行われた状況を示している。ディスクノード１００でパトロールが行われると、ディスクノード１００が、データユニットとの関連付けを有する重複排除ユニットの参照（データ取得）を行う。すると参照された重複排除ユニットの保存期限満了時刻が更新される。

【０２５６】

その後、ディスクノード３００でパトロールが行われると、データユニットとの関連付けを有していない重複排除ユニットについて、保存期限満了時刻を過ぎていることが検出される。該当する重複排除ユニットは、未使用に変更される。図３０の例では、重複排除ボリューム４０のセグメントＩＤ「１」のセグメントに割り当てられたスライス３１１の３番目の重複排除ユニットが未使用に変更されている。

【０２５７】

このようにして、重複排除ボリューム４０から不要なデータを削除し、資源の効率的な利用が促進される。

〔その他の応用例〕

第２の実施の形態では、インデックスサーバ５００と制御ノード６００とが個別に設けられているが、１つの装置にインデックスサーバ５００と制御ノード６００との機能を搭載することもできる。またインデックスサーバ５００の機能を、いずれかのディスクノード１００，２００，３００，４００に搭載することもできる。

【０２５８】

第２の実施の形態では、重複排除ユニットに参照があれば、保存期限満了時刻が更新される。そこで、状態が「DeDup」のデータユニットのパトロールでは、関連付けられた重複排除ユニットの参照を実行することで、重複排除ユニットの保存期限満了時刻を更新させている。この処理では、重複排除ユニットの保存期限満了時刻を更新させることができれば、参照以外の機能を利用してもよい。例えば、データユニットを管理するディスクノードから重複排除ユニットを管理するディスクノードへ、パトロールごとに保存期限満了時刻更新要求を送信するようにしてもよい。重複排除ユニットを管理するディスクノードでは、保存期限満了時刻更新要求に応答して、重複排除ユニットの保存期限満了時刻を更新する。

【０２５９】

また、ハッシュ関数によるハッシュ値の算出に代えて、例えば暗号化処理による暗号データを用いることもできる。

なお、上記の処理機能は、コンピュータによって実現することができる。その場合、ディスクノード１００，２００，３００，４００、インデックスサーバ５００、制御ノード６００，およびアクセスノード７００，８００が有すべき機能の処理内容を記述したプログラムが提供される。そのプログラムをコンピュータで実行することにより、上記処理機能がコンピュータ上で実現される。処理内容を記述したプログラムは、コンピュータで読み取り可能な記録媒体に記録しておくことができる。コンピュータで読み取り可能な記録媒体としては、磁気記憶装置、光ディスク、光磁気記録媒体、半導体メモリなどがある。磁気記憶装置には、ハードディスク装置（ＨＤＤ）、フレキシブルディスク（ＦＤ）、磁気テープなどがある。光ディスクには、ＤＶＤ、ＤＶＤ－ＲＡＭ、ＣＤ－ＲＯＭ／ＲＷなどがある。光磁気記録媒体には、ＭＯ（Magneto-Optical disc）などがある。

【０２６０】

プログラムを流通させる場合には、例えば、そのプログラムが記録されたＤＶＤ、ＣＤ－ＲＯＭなどの可搬型記録媒体が販売される。また、プログラムをサーバコンピュータの記憶装置に格納しておき、ネットワークを介して、サーバコンピュータから他のコンピュータにそのプログラムを転送することもできる。

【０２６１】

プログラムを実行するコンピュータは、例えば、可搬型記録媒体に記録されたプログラムもしくはサーバコンピュータから転送されたプログラムを、自己の記憶装置に格納する。そして、コンピュータは、自己の記憶装置からプログラムを読み取り、プログラムに従った処理を実行する。なお、コンピュータは、可搬型記録媒体から直接プログラムを読み取り、そのプログラムに従った処理を実行することもできる。また、コンピュータは、サーバコンピュータからプログラムが転送されるごとに、逐次、受け取ったプログラムに従った処理を実行することもできる。

【０２６２】

また、上記の処理機能の少なくとも一部を、ＤＳＰ（Digital Signal Processor）、ＡＳＩＣ（Application Specific Integrated Circuit）、ＰＬＤ（Programmable Logic Device）などの電子回路で実現することもできる。

【０２６３】

以上、実施の形態を例示したが、実施の形態で示した各部の構成は同様の機能を有する他のものに置換することができる。また、他の任意の構成物や工程が付加されてもよい。さらに、前述した実施の形態のうちの任意の２以上の構成（特徴）を組み合わせたものであってもよい。

【０２６４】

以上の実施の形態に開示された技術には、以下の付記に示す技術が含まれる。

（付記１）　複数のディスクノードそれぞれで管理されるストレージ装置にデータを分散格納するマルチノードストレージシステムを構成する１つの前記ディスクノードで実行すべきデータ管理処理をコンピュータに実行させるデータ管理プログラムにおいて、

前記コンピュータに、

仮想的な記憶領域を定義した論理ボリュームを複数に分割して得られるデータユニットを指定した書き込み要求があると、前記コンピュータに接続されたストレージ装置内の単位記憶領域の１つを指定された前記データユニットに割り当て、前記データユニットに割り当てられた前記単位記憶領域に対してデータの書き込みを行い、

書き込み対象となった前記データユニットに対応付けて、現在の時刻を最終書き込み時刻としてデータユニット情報記憶手段に格納し、

前記データユニット情報記憶手段を参照し、前記最終書き込み時刻から重複排除化猶予期間以上経過している前記データユニットを検出し、

仮想的な記憶領域を定義した重複排除ボリュームを複数に分割して得られる重複排除ユニットのうち重複排除化の対象とされた重複排除対象データの格納に使用されている重複排除ユニットの識別情報および前記重複排除ユニットを管理している前記ディスクノードの識別子を含む重複排除アドレスと、前記重複排除対象データに所定の演算を行うことで得られる前記重複排除対象データの固有値とを対応付けて管理するインデックスサーバから、検出された前記データユニットに割り当てられた前記単位記憶領域内のデータの前記固有値に対応付けられた前記重複排除アドレスを取得し、

検出された前記データユニットに対応付けて、取得した前記重複排除アドレスを前記データユニット情報記憶手段に格納すると共に、検出された前記データユニットへの前記単位記憶領域の割り当てを解除する、

処理を実行させることを特徴とするデータ管理プログラム。

【０２６５】

（付記２）　前記コンピュータに、さらに、

前記インデックスサーバから、前記重複排除ユニットの識別情報を指定した重複排除ユニット割当要求を受信すると、指定された前記重複排除ユニットに前記単位記憶領域の１

つを割り当て、指定された前記重複排除ユニットをデータの読み出し先とする前記ディスクノードから前記重複排除対象データを取得し、指定された前記重複排除ユニットに割り当てた前記単位記憶領域に格納し、
前記重複排除ユニットの識別情報を指定したデータの読み出し要求を受信すると、指定された前記重複排除ユニットに割り当てられた前記単位記憶領域内の前記重複排除対象データを応答する、
処理を実行させることを特徴とする付記１記載のデータ管理プログラム。
【０２６６】
（付記３） 前記コンピュータに、さらに、
前記重複排除ユニットを前記重複排除対象データの読み出し先とする前記ディスクノードが存在しなくなった場合、前記重複排除ユニットへの前記単位記憶領域の割り当てを解除する、
処理を実行させることを特徴とする付記２記載のデータ管理プログラム。
【０２６７】
（付記４） 前記コンピュータにさらに、
前記重複排除ユニットに対応付けて、前記重複排除ユニットに割り当てられた前記単位記憶領域から最後にデータの読み出しが行われた時刻から保存期間経過後の保存期限満了時刻を重複排除ユニット情報記憶手段に格納し、
前記重複排除ユニット情報記憶手段を参照し、現在の時刻が前記保存期限満了時刻を過ぎている前記重複排除ユニットを検出し、検出した前記重複排除ユニットへの前記単位記憶領域の割り当てを解除する、
処理を実行させることを特徴とする付記３記載のデータ管理プログラム。
【０２６８】
（付記５） 前記コンピュータに、さらに、
前記インデックスサーバから取得した前記重複排除アドレスが、新たに割り当てられた前記重複排除ユニットの前記重複排除アドレスの場合、前記データユニットに割り当てられた前記単位記憶領域内のデータを前記重複排除アドレスに示される前記ディスクノードへ送信後、前記データユニットへの前記単位記憶領域の割り当てを解除する、
処理を実行させることを特徴とする付記２記載のデータ管理プログラム。
【０２６９】
（付記６） 前記コンピュータに、さらに、
定期的に前記データユニット情報記憶手段を参照し、前記重複排除アドレスが対応付けられた前記データユニットを検出し、前記重複排除アドレスに示される前記ディスクノードから、前記重複排除アドレスで示される前記重複排除ユニットの識別情報を指定してデータを取得する、
処理を実行させることを特徴とする付記１記載のデータ管理プログラム。
【０２７０】
（付記７） 前記固有値は、前記重複排除対象データに対して所定のハッシュ関数による演算を行うことで得られるハッシュ値であることを特徴とする付記１記載のデータ管理プログラム。
【０２７１】
（付記８） 前記重複排除ユニットの識別情報は、前記重複排除ボリューム内での前記重複排除ユニットの位置を示すオフセットであることを特徴とする付記１記載のデータ管理プログラム。
【０２７２】
（付記９） 前記コンピュータに、さらに、
前記データユニットを指定した読み出し要求があると、読み出し対象の前記データユニットに割り当てられた前記単位記憶領域がない場合、前記データユニット情報記憶手段を参照し、読み出し対象の前記データユニットに対応付けられた前記重複排除アドレスに示される前記ディスクノードから、前記重複排除アドレスに示される前記重複排除ユニット

の識別情報を指定して前記重複排除対象データを取得する、
処理を実行させることを特徴とする付記１記載のデータ管理プログラム。
【０２７３】
（付記１０） 前記コンピュータに、さらに、
前記データユニットを指定した読み出し要求があると、読み出し対象の前記データユニットに割り当てられた前記単位記憶領域がある場合、割り当てられた前記単位記憶領域からデータを読み出す、
処理を実行させることを特徴とする付記１記載のデータ管理プログラム。
【０２７４】
（付記１１） 複数のディスクノードそれぞれで管理されるストレージ装置にデータを分散格納するマルチノードストレージシステムの記憶領域管理処理をコンピュータに実行させるデータ管理プログラムにおいて、
前記コンピュータに、
仮想的な記憶領域を定義した重複排除ボリュームを複数に分割して得られる重複排除ユニットのうち前記ディスクノードで管理している前記重複排除ユニットの使用の有無、使用されている前記重複排除ユニットに割り当てられている単位記憶領域に格納された重複排除対象データに所定の演算を行うことで得られる固有値、および前記重複排除ユニットの識別情報を含む重複排除ユニット情報が、前記重複排除ユニットを管理している前記ディスクノードの識別子に対応付けて重複排除ユニット情報記憶手段に格納されており
前記ディスクノードから、重複排除化データユニットに割り当てられた単位記憶領域内のデータに前記所定の演算を実行して得られた固有値を含む重複排除アドレス照会要求を受け取ると、前記重複排除ユニット情報記憶手段から、前記重複排除アドレス照会要求に示される固有値を含む前記重複排除ユニット情報を検索し、
検索により該当する前記重複排除ユニット情報が検出された場合、検出された前記重複排除ユニット情報に含まれる前記重複排除ユニットの識別情報と、検出された前記重複排除ユニット情報に対応する前記重複排除ユニットを管理している前記ディスクノードの識別子とを含む重複排除アドレスを、前記重複排除アドレス照会要求の送信元である前記ディスクノードに応答し、
検索により該当する前記重複排除ユニット情報が検出されなかった場合、前記重複排除ユニット情報記憶手段を参照し、未使用の前記重複排除ユニットを選択し、選択した前記重複排除ユニットを管理する前記ディスクノードに対して選択した重複排除ユニットへの前記単位記憶領域の割り当て要求を送信し、選択した前記重複排除ユニットの前記単位記憶領域割り当て後の前記重複排除ユニット情報を前記重複排除ユニット情報記憶手段に格納し、選択した前記重複排除ユニットの識別情報と、選択した前記重複排除ユニットを管理する前記ディスクノードの識別子とを含む重複排除アドレスを、前記重複排除アドレス照会要求の送信元である前記ディスクノードに応答する、
処理を実行させることを特徴とするデータ管理プログラム。
【０２７５】
（付記１２） 前記コンピュータに、
前記重複排除ユニット情報を複数の前記ディスクノードそれぞれから取得し、取得元の前記ディスクノードの識別子に対応付けて重複排除ユニット情報記憶手段に格納する、
処理を実行させることを特徴とする付記１１記載のデータ管理プログラム。
【０２７６】
（付記１３） 前記コンピュータに、さらに、
前記重複排除ユニット情報には、前記重複排除ユニットに割り当てられた前記単位記憶領域から最後にデータの読み出しが行われた時刻から保存期間経過後の保存期限満了時刻が含められており、前記重複排除ユニット情報記憶手段を参照し、現在の時刻が前記保存期限満了時刻を過ぎている前記重複排除ユニットを検出し、
検出された前記重複排除ユニット情報の取得元の前記ディスクノードから最新の前記重複排除ユニット情報を取得し、前記重複排除ユニット情報記憶手段に反映する、

処理を実行させることを特徴とする付記１１記載のデータ管理プログラム。

【０２７７】

（付記１４）　複数のディスクノードそれぞれで管理されるストレージ装置にデータを分散格納するマルチノードストレージシステムを構成する１つの前記ディスクノードで実行すべきデータ管理処理をコンピュータによって実現するデータ管理装置において、

仮想的な記憶領域を定義した論理ボリュームを複数に分割して得られるデータユニットを指定した書き込み要求があると、前記コンピュータに接続されたストレージ装置内の単位記憶領域の１つを指定された前記データユニットに割り当て、前記データユニットに割り当てられた前記単位記憶領域に対してデータの書き込みを行う書き込みアクセス手段と、

書き込み対象となった前記データユニットに対応付けて、現在の時刻を最終書き込み時刻としてデータユニット情報記憶手段に格納する最終書き込み時刻更新手段と、

前記データユニット情報記憶手段を参照し、最終書き込み時刻から重複排除化猶予期間以上経過している前記データユニットを検出する重複排除化データユニット検出手段と、

仮想的な記憶領域を定義した重複排除ボリュームを複数に分割して得られる重複排除ユニットのうち重複排除化の対象とされた重複排除対象データの格納に使用されている重複排除ユニットの識別情報および前記重複排除ユニットを管理している前記ディスクノードの識別子を含む重複排除アドレスと、前記重複排除対象データに所定の演算を行うことで得られる前記重複排除対象データの固有値とを対応付けて管理するインデックスサーバから、検出された前記データユニットに割り当てられた前記単位記憶領域内のデータの前記固有値に対応付けられた前記重複排除アドレスを取得する重複排除アドレス取得手段と、

検出された前記データユニットに対応付けて、取得した前記重複排除アドレスを前記データユニット情報記憶手段に格納すると共に、検出された前記データユニットへの前記単位記憶領域の割り当てを解除する単位記憶領域割当解除手段と、

前記データユニットを指定した読み出し要求があると、読み出し対象の前記データユニットに割り当てられた前記単位記憶領域がある場合、割り当てられた前記単位記憶領域からデータを読み出し、読み出し対象の前記データユニットに割り当てられた前記単位記憶領域がない場合、前記データユニット情報記憶手段を参照し、読み出し対象の前記データユニットに対応付けられた前記重複排除アドレスに示される前記ディスクノードから、前記重複排除アドレスに示される前記重複排除ユニットの識別情報を指定して前記重複排除対象データを取得する読み出しアクセス手段と、

を有することを特徴とするデータ管理装置。

【０２７８】

（付記１５）　複数のディスクノードそれぞれで管理されるストレージ装置にデータを分散格納するマルチノードストレージシステムを構成する１つの前記ディスクノードで実行すべきデータ管理処理をコンピュータで実行するデータ管理方法において、

前記コンピュータが、

仮想的な記憶領域を定義した論理ボリュームを複数に分割して得られるデータユニットを指定した書き込み要求があると、前記コンピュータに接続されたストレージ装置内の単位記憶領域の１つを指定された前記データユニットに割り当て、前記データユニットに割り当てられた前記単位記憶領域に対してデータの書き込みを行い、

書き込み対象となった前記データユニットに対応付けて、現在の時刻を最終書き込み時刻としてデータユニット情報記憶手段に格納し、

前記データユニット情報記憶手段を参照し、最終書き込み時刻から重複排除化猶予期間以上経過している前記データユニットを検出し、

仮想的な記憶領域を定義した重複排除ボリュームを複数に分割して得られる重複排除ユニットのうち重複排除化の対象とされた重複排除対象データの格納に使用されている重複排除ユニットの識別情報および前記重複排除ユニットを管理している前記ディスクノードの識別子を含む重複排除アドレスと、前記重複排除対象データに所定の演算を行うことで得られる前記重複排除対象データの固有値とを対応付けて管理するインデックスサーバか

ら、検出された前記データユニットに割り当てられた前記単位記憶領域内のデータの前記固有値に対応付けられた前記重複排除アドレスを取得し、

検出された前記データユニットに対応付けて、取得した前記重複排除アドレスを前記データユニット情報記憶手段に格納すると共に、検出された前記データユニットへの前記単位記憶領域の割り当てを解除し、

前記データユニットを指定した読み出し要求があると、読み出し対象の前記データユニットに割り当てられた前記単位記憶領域がある場合、割り当てられた前記単位記憶領域からデータを読み出し、読み出し対象の前記データユニットに割り当てられた前記単位記憶領域がない場合、前記データユニット情報記憶手段を参照し、読み出し対象の前記データユニットに対応付けられた前記重複排除アドレスに示される前記ディスクノードから、前記重複排除アドレスに示される前記重複排除ユニットの識別情報を指定して前記重複排除対象データを取得する、

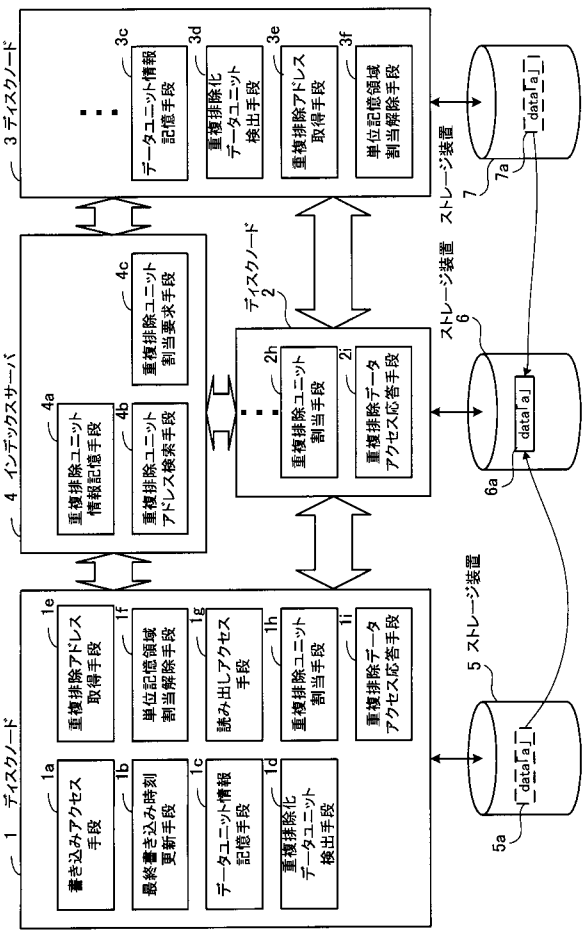
ことを特徴とするデータ管理方法。

【符号の説明】

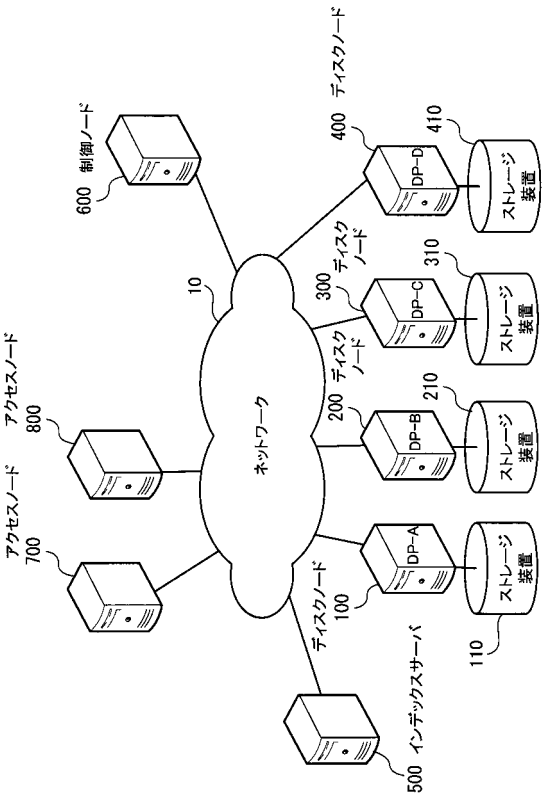
【０２７９】

１～３　ディスクノード
１ａ　書き込みアクセス手段
１ｂ　最終書き込み時刻更新手段
１ｃ，３ｃ　データユニット情報記憶手段
１ｄ，３ｄ　重複排除化データユニット検出手段
１ｅ，３ｅ　重複排除アドレス取得手段
１ｆ，３ｆ　単位記憶領域割当解除手段
１ｇ　読み出しアクセス手段
１ｈ，２ｈ　重複排除ユニット割当手段
１ｉ，２ｉ　重複排除データアクセス応答手段
４　インデックスサーバ
４ａ　重複排除ユニット情報記憶手段
４ｂ　重複排除ユニットアドレス検索手段
４ｃ　重複排除ユニット割当要求手段
５～７　ストレージ装置
５ａ，６ａ，７ａ　単位記憶領域

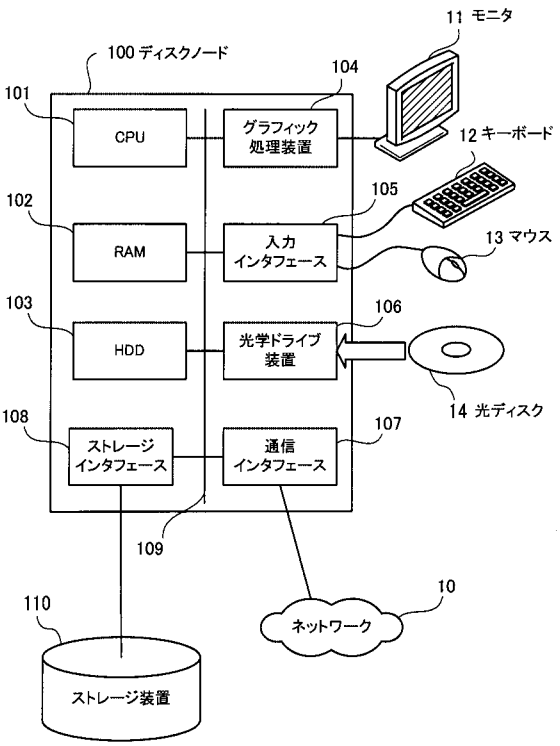
【図１】


1 ディスクノード
書き込みアクセス手段
最終書き込み時刻更新手段
データユニット情報記憶手段
重複排除化データユニット検出手段
重複排除アドレス取得手段
単位記憶領域割当解除手段
読み出しアクセス手段
重複排除ユニット割当手段
重複排除データアクセス応答手段
2 ディスクノード
3 ディスクノード
4 インデックスサーバ
重複排除ユニット情報記憶手段
重複排除ユニットアドレス検索手段
重複排除ユニット割当要求手段
5 ストレージ装置
6 ストレージ装置
7 ストレージ装置
data[a]

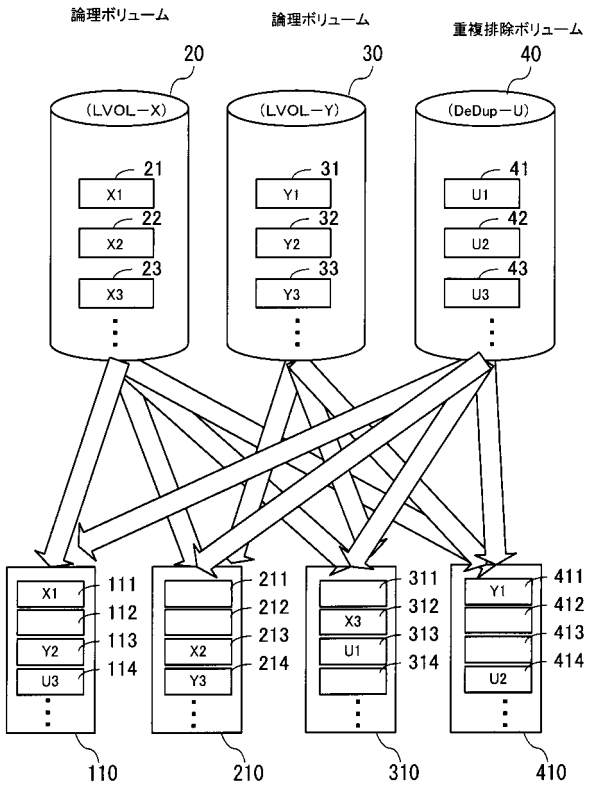
【図２】


制御ノード
600
アクセスノード
800
アクセスノード
700
ネットワーク
10
ディスクノード
100
ディスクノード
200
ディスクノード
300
ディスクノード
400
DP-A
DP-B
DP-C
DP-D
ストレージ装置
110
210
310
410
インデックスサーバ
500

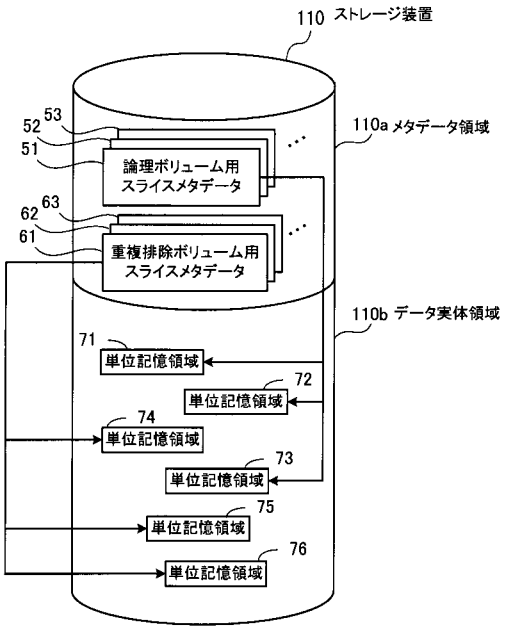
【図３】


100 ディスクノード
CPU
RAM
HDD
ストレージインタフェース
グラフィック処理装置
入力インタフェース
光学ドライブ装置
通信インタフェース
11 モニタ
12 キーボード
13 マウス
14 光ディスク
ネットワーク
10
ストレージ装置
109
110

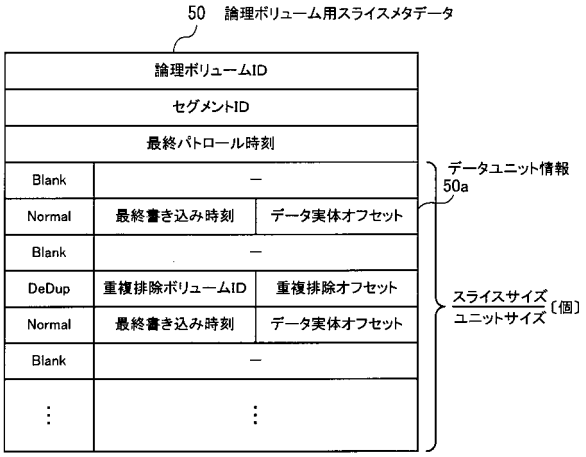
【図４】


論理ボリューム
論理ボリューム
重複排除ボリューム
(LVOL−X)
(LVOL−Y)
(DeDup−U)
X1
X2
X3
Y1
Y2
Y3
U1
U2
U3
110
210
310
410

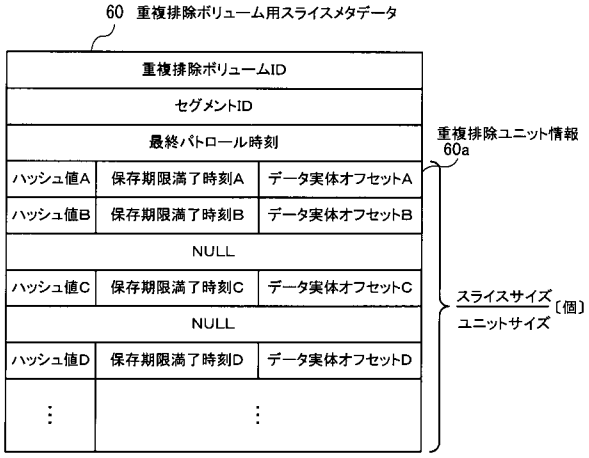
【図５】


110 ストレージ装置
110a メタデータ領域
52
53
51
論理ボリューム用
スライスメタデータ
62
63
61
重複排除ボリューム用
スライスメタデータ
110b データ実体領域
71
単位記憶領域
72
単位記憶領域
74
単位記憶領域
73
単位記憶領域
75
単位記憶領域
76
単位記憶領域

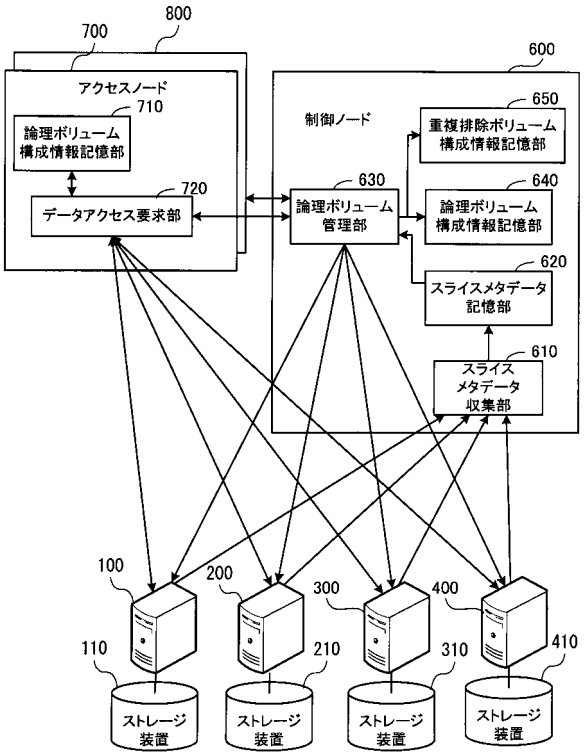
【図６】


50 論理ボリューム用スライスメタデータ
論理ボリュームID
セグメントID
最終パトロール時刻
Blank
−
Normal
最終書き込み時刻
データ実体オフセット
Blank
−
DeDup
重複排除ボリュームID
重複排除オフセット
Normal
最終書き込み時刻
データ実体オフセット
Blank
−
データユニット情報
50a
スライスサイズ
ユニットサイズ
〔個〕

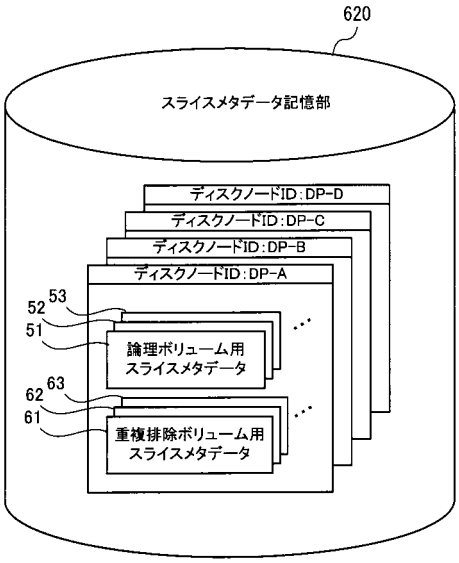
【図７】


60 重複排除ボリューム用スライスメタデータ
重複排除ボリュームID
セグメントID
最終パトロール時刻
ハッシュ値A
保存期限満了時刻A
データ実体オフセットA
ハッシュ値B
保存期限満了時刻B
データ実体オフセットB
NULL
ハッシュ値C
保存期限満了時刻C
データ実体オフセットC
NULL
ハッシュ値D
保存期限満了時刻D
データ実体オフセットD
重複排除ユニット情報
60a
スライスサイズ
ユニットサイズ
〔個〕

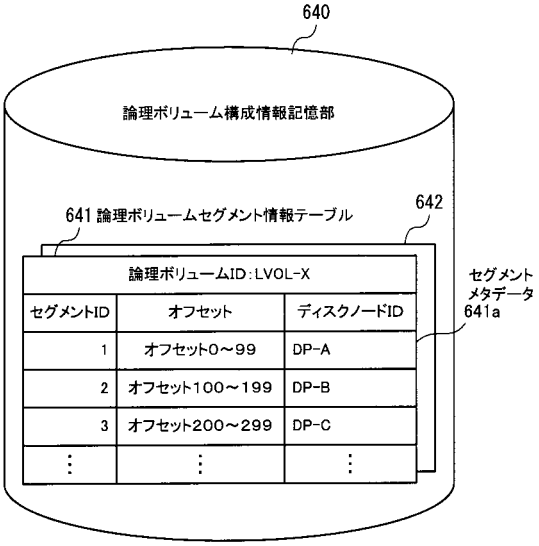
【図８】


700
800
600
アクセスノード
710
論理ボリューム
構成情報記憶部
720
データアクセス要求部
制御ノード
650
重複排除ボリューム
構成情報記憶部
630
論理ボリューム
管理部
640
論理ボリューム
構成情報記憶部
620
スライスメタデータ
記憶部
610
スライス
メタデータ
収集部
100
200
300
400
110
210
310
410
ストレージ
装置
ストレージ
装置
ストレージ
装置
ストレージ
装置

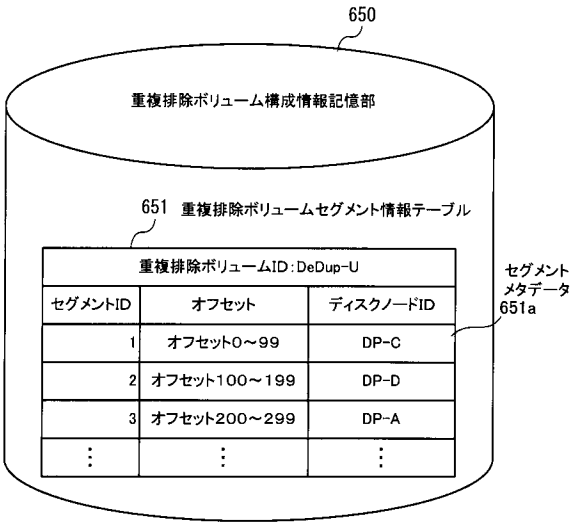
【図９】



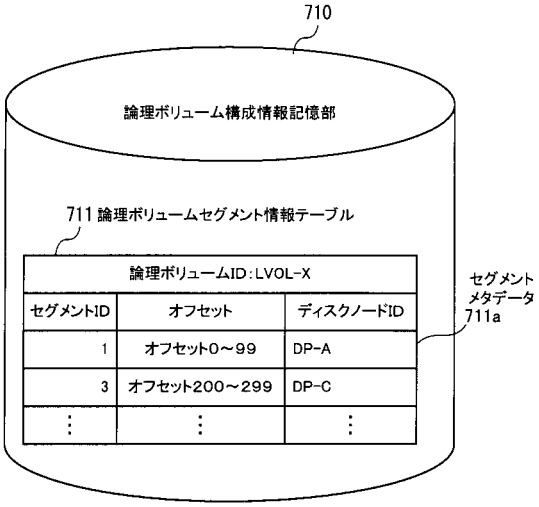
【図１０】



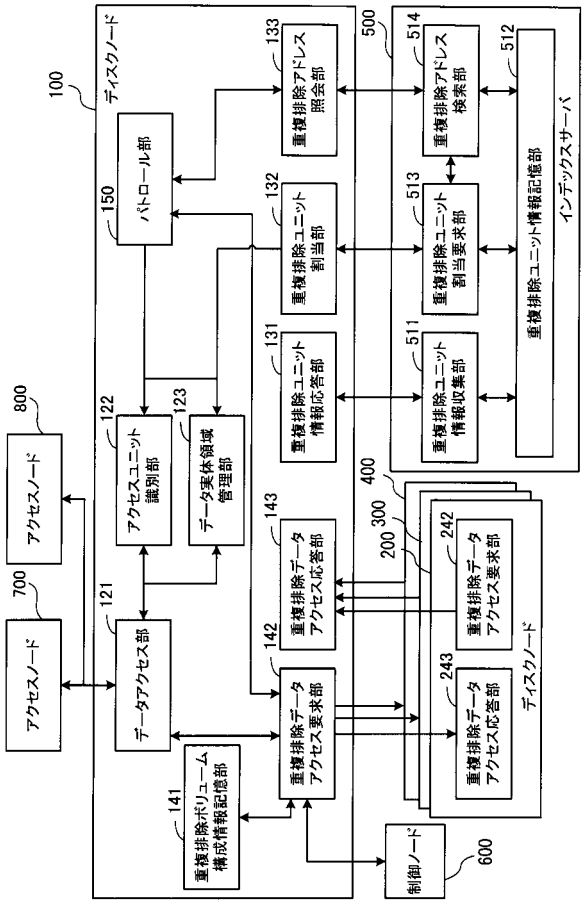
【図１１】



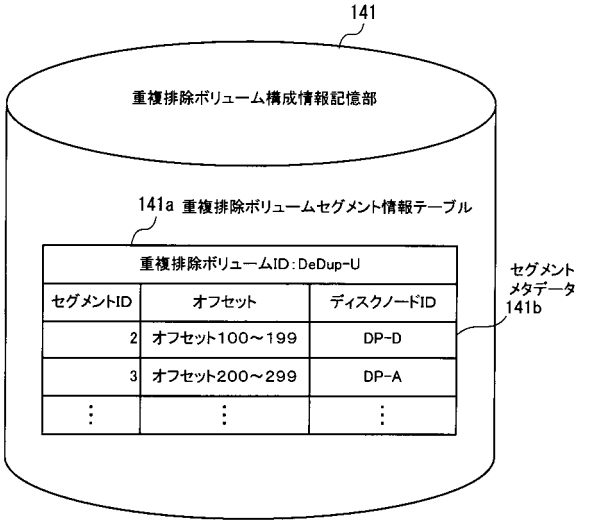
【図１２】



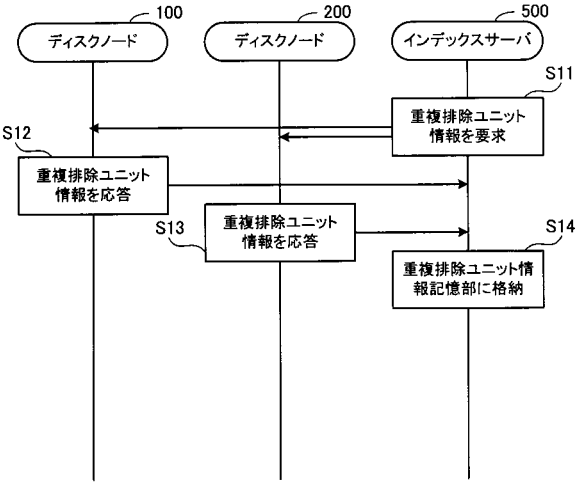
【図１３】



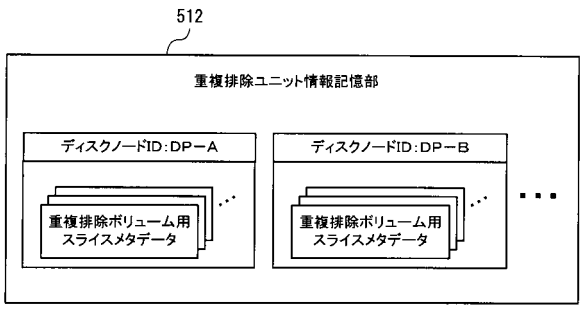
【図１４】



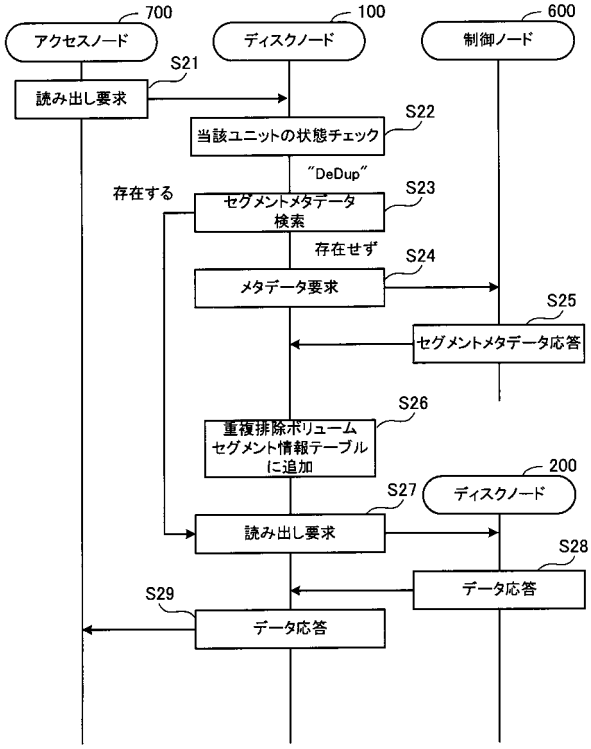
【図１５】



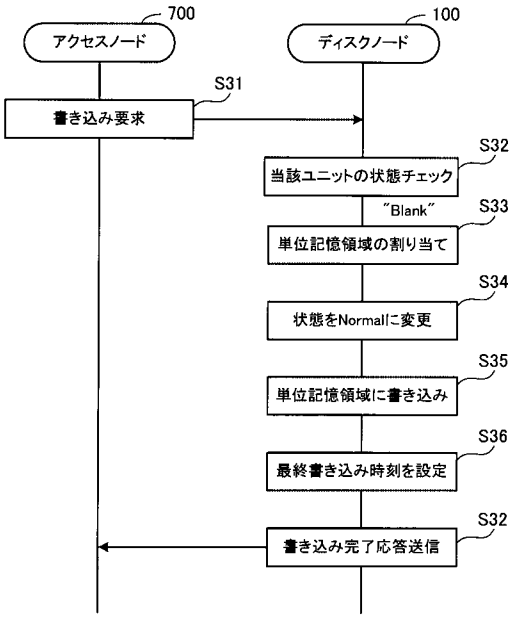
【図１６】



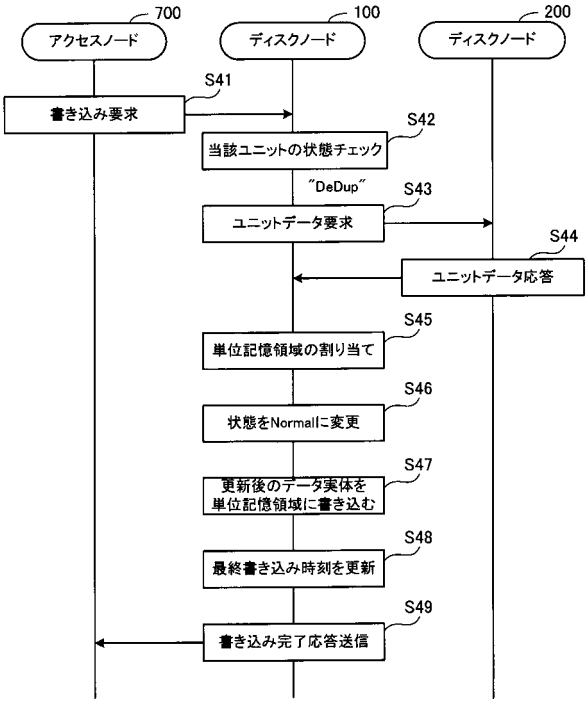
【図１７】



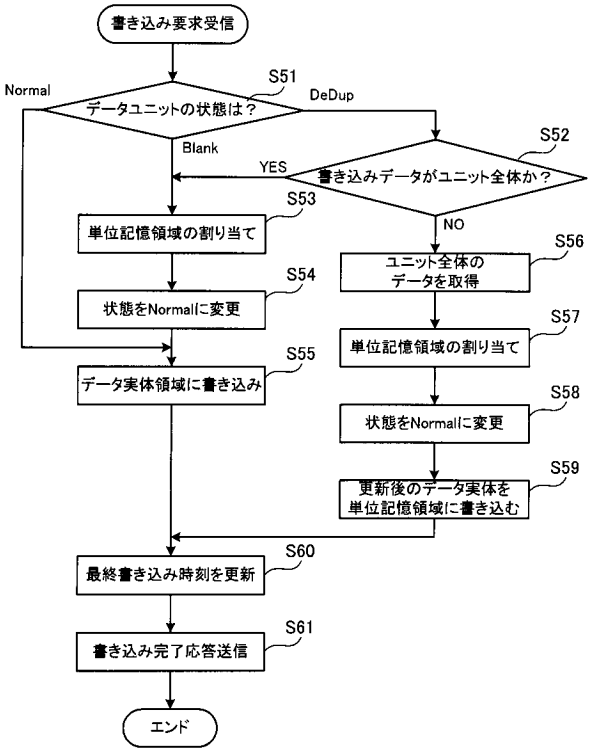
【図１８】



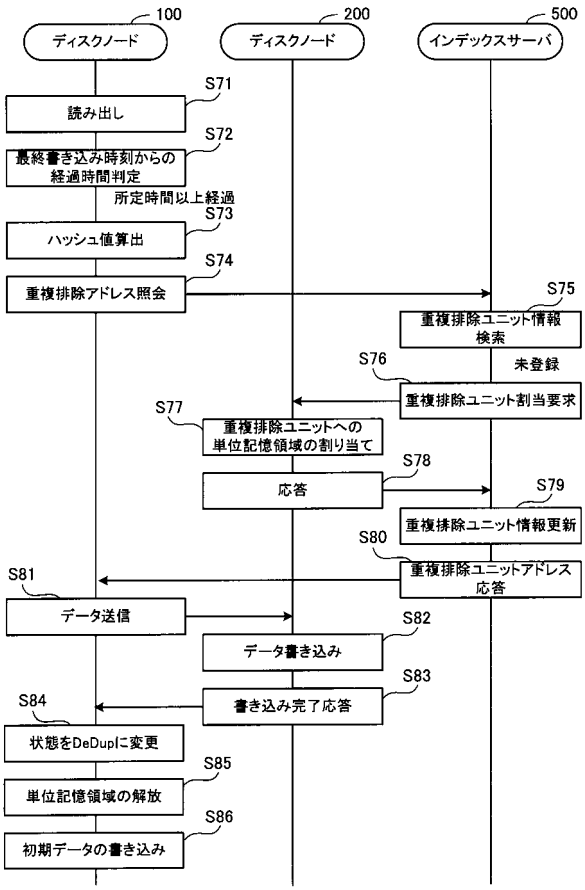
【図１９】



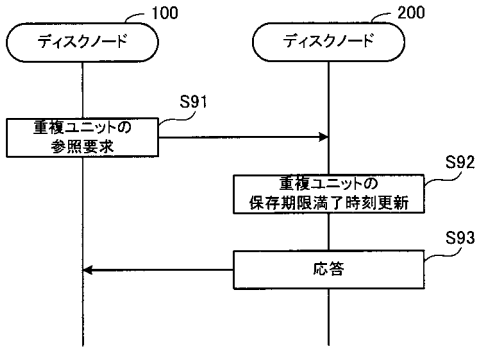
【図２０】



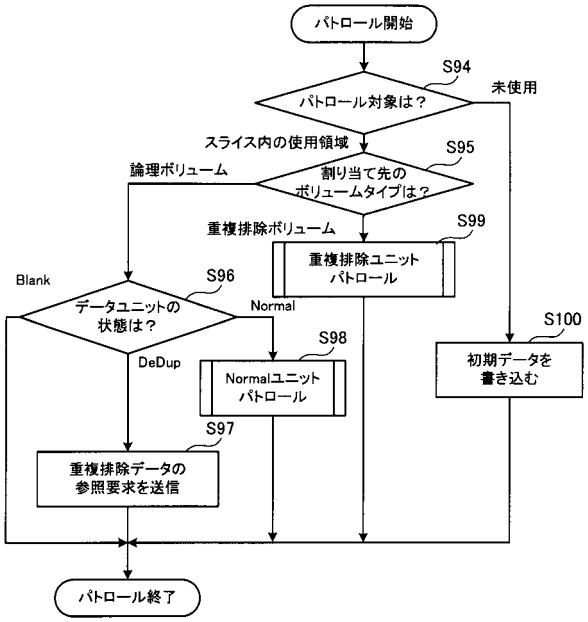
【図２１】



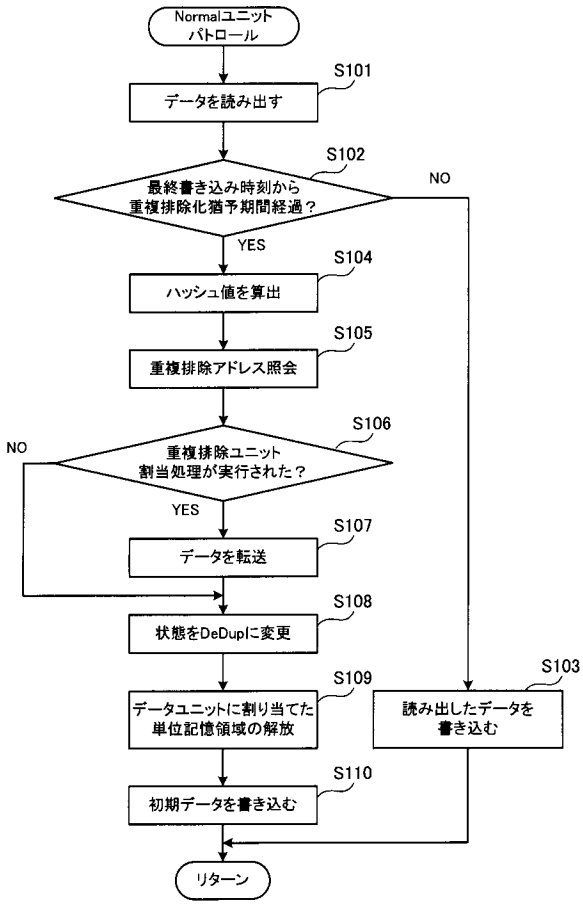
【図２２】



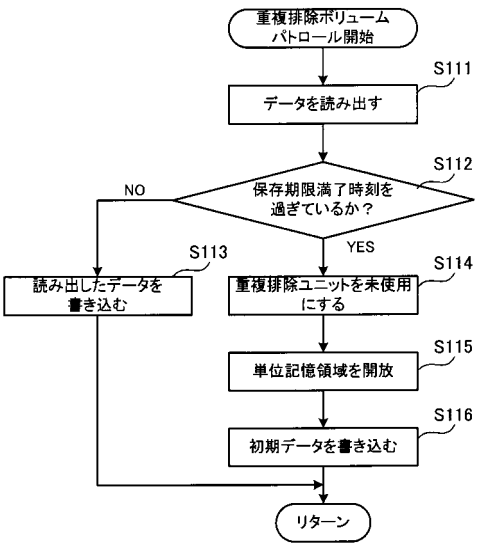
【図２３】



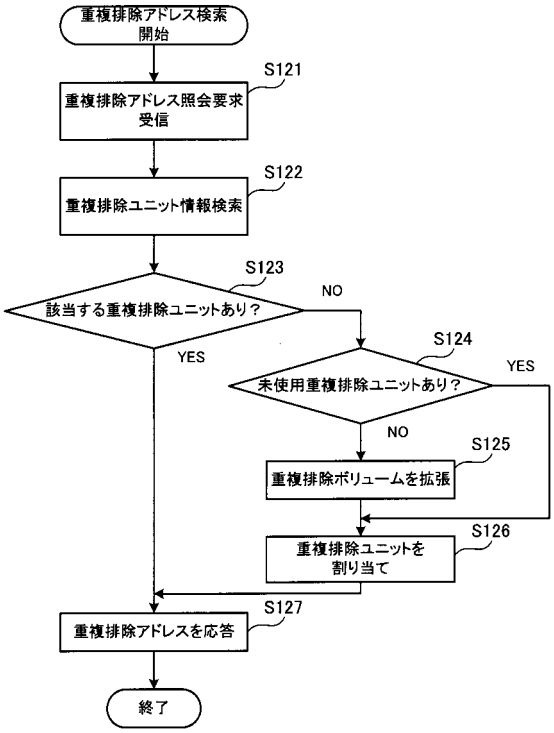
【図２４】



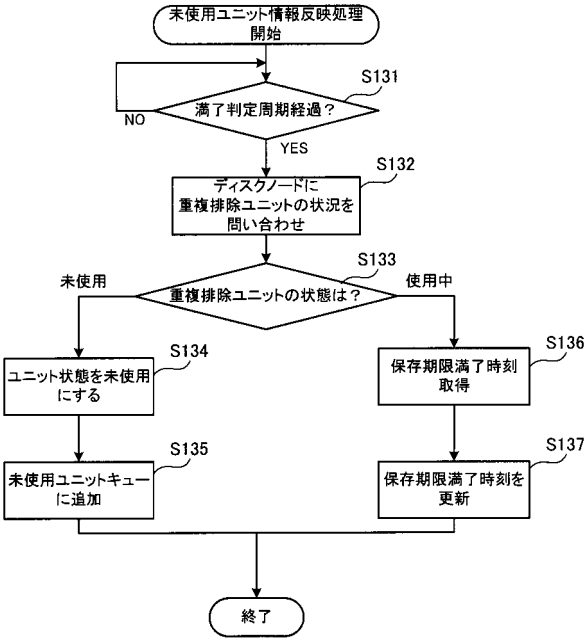
【図２５】



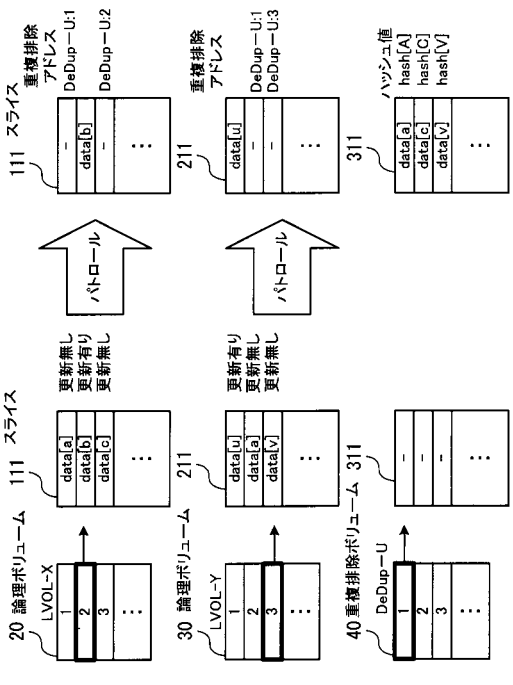
【図２６】



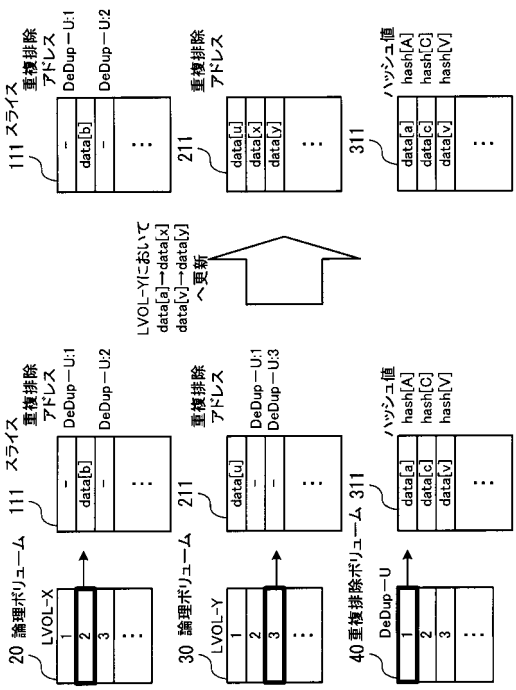
【図２７】



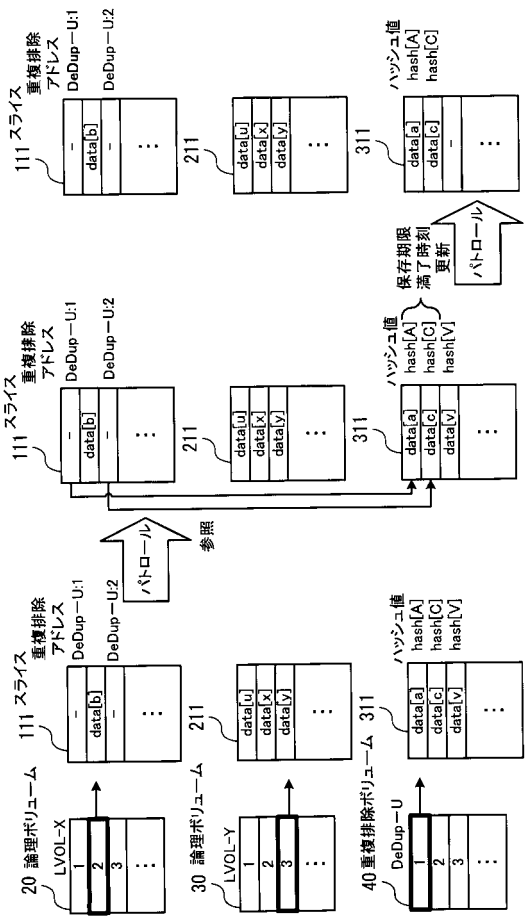
【図２８】



【図２９】



【図３０】


20 論理ボリューム
LVOL-X
30 論理ボリューム
LVOL-Y
40 重複排除ボリューム
DeDup-U
111 スライス
重複排除アドレス
DeDup-U:1
DeDup-U:2
211
311
ハッシュ値
hash[A]
hash[C]
hash[V]
data[a]
data[b]
data[c]
data[u]
data[x]
data[y]
data[v]
パトロール
参照
保存期限満了時刻更新

フロントページの続き

(72)発明者　丸山　哲太郎
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内
(72)発明者　土屋　芳浩
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内
(72)発明者　渡辺　高志
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内
(72)発明者　熊野　達夫
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内
(72)発明者　大江　和一
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内

審査官　辻本　泰隆

(56)参考文献　特開２００９－２３７９７９（ＪＰ，Ａ）
特開２００９－８７０２１（ＪＰ，Ａ）
特開２００９－１４６３８１（ＪＰ，Ａ）
特開２００９－２０５２０１（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１２／００
ＪＳＴＰｌｕｓ／ＪＭＥＤＰｌｕｓ／ＪＳＴ７５８０（ＪＤｒｅａｍＩＩ）
ＣｉＮｉｉ