(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2011-128917
(P2011-128917A)
(43)公開日　平成23年6月30日(2011.6.30)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 3/06	(2006.01)	G06F	3/06	304B	5B065
		G06F	3/06	305F	

審査請求 未請求　請求項の数 7　ＯＬ　（全 49 頁）

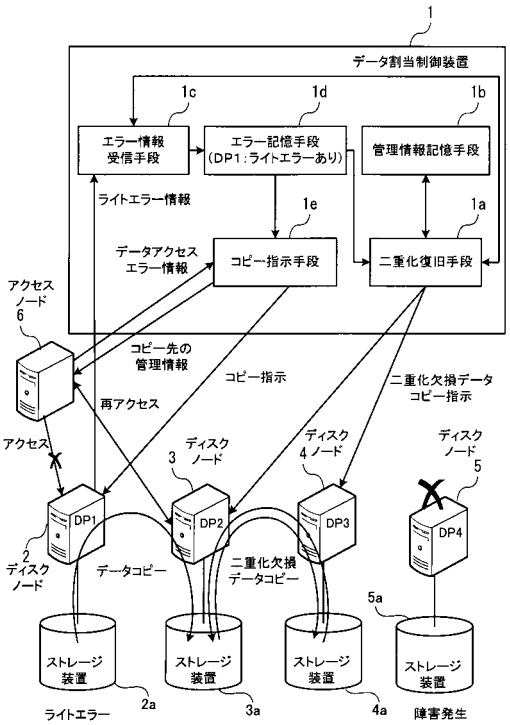
(21)出願番号　特願2009-287068 (P2009-287068)
(22)出願日　平成21年12月18日 (2009.12.18)

(71)出願人　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　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　櫻井　英樹
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

Ｆターム(参考)　5B065 BA01　EA02　EA24　EA31

(54)【発明の名称】データ割当制御プログラム、データ割当制御方法、およびデータ割当制御装置

(57)【要約】
【課題】ディスクノードの障害発生時のデータの喪失を抑制する。
【解決手段】二重化復旧手段１ａは、二重化欠損データのコピーをディスクノードに指示し、二重化復旧処理を行う。エラー情報受信手段１ｃは、二重化復旧処理中に、ストレージ装置２ａにおけるライトエラーの発生を示すライトエラー情報を受け取ると、ライトエラーが発生したストレージ装置２ａの識別情報をエラー記憶手段１ｄに格納する。コピー指示手段１ｅは、ライトエラーが発生したストレージ装置２ａ内のアクセスが行われたデータに冗長データが存在しない場合、ディスクノード２に対してアクセスが行われたデータのコピーを指示する。
【選択図】図１



【特許請求の範囲】
【請求項１】
接続されたストレージ装置内のデータを管理する複数のディスクノードに対する管理対象データの割り当て指示を、コンピュータに実行させるデータ割当制御プログラムにおいて、
前記コンピュータに、
複数のストレージ装置に格納されたデータのうち同一内容の冗長データが存在しない二重化欠損データのコピーを、二重化欠損データを管理するディスクノードに指示する二重化復旧処理を行い、
前記二重化復旧処理中に、ストレージ装置に対するデータのライトエラーを示すライトエラー情報を受け取ると、ライトエラーが発生したストレージ装置の識別情報をエラー記憶手段に格納し、
前記エラー記憶手段を参照してライトエラーが発生したストレージ装置を判断し、ライトエラーが発生したストレージ装置に格納されているデータをコピー対象データとし、ライトエラーが発生していないストレージ装置内に前記コピー対象データの冗長データが存在しない場合、前記コピー対象データを管理するディスクノードに対して、ライトエラーが発生していないストレージ装置への前記コピー対象データのコピーを指示する、
処理を実行させることを特徴とするデータ割当制御プログラム。
【請求項２】
前記二重化復旧処理では、二重化欠損データを順次選択し、選択した二重化欠損データを管理するディスクノードに対し、選択した二重化欠損データが格納されているストレージ装置と異なるストレージ装置であり、かつライトエラーが発生していないストレージ装置への、選択した二重化欠損データのコピーを指示することを特徴とする請求項１記載のデータ割当制御プログラム。
【請求項３】
前記二重化復旧処理では、前記複数のストレージ装置内の二重化欠損データを調査し、二重化欠損データの格納場所を示す管理情報を管理情報記憶手段に格納し、前記管理情報記憶手段内の管理情報に示される二重化欠損データを順次選択して、選択した二重化欠損データを管理するディスクノードに対して選択した二重化欠損データのコピーを指示し、
前記コピー対象データのコピー指示の際には、前記コピー対象データのコピーが完了すると、コピー先のデータの管理情報を二重化欠損データとして前記管理情報記憶手段に格納する、
ことを特徴とする請求項１記載のデータ割当制御プログラム。
【請求項４】
前記二重化復旧処理では、前記管理情報記憶手段内の管理情報で示されるすべての二重化欠損データのコピーが完了すると前記エラー記憶手段を参照してライトエラーが発生したストレージ装置の有無を判断し、ライトエラーが発生したストレージ装置がある場合、ライトエラーが発生したストレージ装置へのアクセスを停止することにより生じる二重化欠損データの管理情報を前記管理情報記憶手段に追加し、二重化復旧処理を続行する、
ことを特徴とする請求項３記載のデータ割当制御プログラム。
【請求項５】
前記コピー対象データのコピー指示では、ライトエラーが発生しているストレージ装置内のデータに対するアクセスで発生したエラーを示すデータアクセスエラー情報を受け取ると、エラー発生時のアクセス対象となっていたデータをコピー対象データとし、前記コピー対象データのコピーが完了すると、コピー先のデータの管理情報を、新たなアクセス先として前記データアクセスエラー情報を送信した装置に応答することを特徴とする請求項１記載のデータ割当制御プログラム。
【請求項６】
接続されたストレージ装置内のデータを管理する複数のディスクノードに対する管理対象データの割り当て指示をコンピュータで実行するデータ割当制御方法において、

前記コンピュータが、
複数のストレージ装置に格納されたデータのうち同一内容の冗長データが存在しない二重化欠損データのコピーを、二重化欠損データを管理するディスクノードに指示する二重化復旧処理を行い、
前記二重化復旧処理中に、ストレージ装置に対するデータのライトエラーを示すライトエラー情報を受け取ると、ライトエラーが発生したストレージ装置の識別情報をエラー記憶手段に格納し、
前記エラー記憶手段を参照してライトエラーが発生したストレージ装置を判断し、ライトエラーが発生したストレージ装置に格納されているデータをコピー対象データとし、ライトエラーが発生していないストレージ装置内に前記コピー対象データの冗長データが存在しない場合、前記コピー対象データを管理するディスクノードに対して、ライトエラーが発生していないストレージ装置への前記コピー対象データのコピーを指示する、
ことを特徴とするデータ割当制御方法。
【請求項７】
接続されたストレージ装置内のデータを管理する複数のディスクノードに対する管理対象データの割り当て指示を行うデータ割当制御装置において、
複数のストレージ装置に格納されたデータのうち同一内容の冗長データが存在しない二重化欠損データのコピーを、二重化欠損データを管理するディスクノードに指示する二重化復旧処理を行う二重化復旧処理手段と、
前記二重化復旧処理中に、ストレージ装置に対するデータのライトエラーを示すライトエラー情報を受け取ると、ライトエラーが発生したストレージ装置の識別情報をエラー記憶手段に格納するエラー情報受信手段と、
前記エラー記憶手段を参照してライトエラーが発生したストレージ装置を判断し、ライトエラーが発生したストレージ装置に格納されているデータをコピー対象データとし、ライトエラーが発生していないストレージ装置内に前記コピー対象データの冗長データが存在しない場合、前記コピー対象データを管理するディスクノードに対して、ライトエラーが発生していないストレージ装置への前記コピー対象データのコピーを指示するコピー指示手段と、
を有することを特徴とするデータ割当制御装置。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は複数のディスクノードに対する管理対象データの割り当てを行うデータ割当制御プログラム、データ割当制御方法、およびデータ割当制御装置に関する。
【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とを有している。マルチノードストレージシステムでは、例えば制御ノードの管理の下、仮想的なディスク（論理ディスク）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】
マルチノードストレージシステムでは、例えば論理ディスク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デバイス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、例えば、セグメントと同サイズとされる。制御ノードによって、論理ディスクの各セグメントに対して、ストレージデバイス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したデータが、そのセグメントに割り当てられたスライスを有するディスクノードに送られる。ディスクノードは、自己の管理するスライスが割り

当てられたセグメントのデータを受け取り、ストレージ装置内の該当スライスに格納する。

【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】

また、マルチノードストレージシステムでは、セグメントに対して複数のスライスを割り当てることができる。１つのセグメントに２つのスライスを割り当てた場合、一方をプライマリスライス、他方をセカンダリスライスとする。プライマリスライスは、アクセスノードが直接リードまたはライトするスライスである。セカンダリスライスは、プライマリスライスへのライト時にライトデータのミラーリング先となるスライスである。このように、２つのスライスを用いたミラーリングによりデータを二重化することで、データの冗長性が確保される。

【０００６】

なお、ストレージ装置の故障などでディスクノードの異常が検出されると、データの二重化復旧処理が実行される。例えば、１つのノードが異常となった場合、まず異常ノードがシステムから切り離される。次に異常ノードが持っていたデータと同じデータを持つ別のノードをコピー元とし、別のノードをコピー先として、データの冗長性が復元される。

【先行技術文献】

【特許文献】

【０００７】

【特許文献１】国際公開第２００４／１０４８４５号

【特許文献２】特開２００５－３０１５９４号公報

【発明の概要】

【発明が解決しようとする課題】

【０００８】

しかし、システムから切り離されたディスクノードへは、他のノードからアクセスできなくなる。そのためストレージ装置の故障が発生したディスクノードの切り離しを常に行うと、データの二重化復旧処理中に新たに別のディスクノードに接続されたストレージ装置で故障が発生した場合に、データを失う可能性がある。例えば、最初に故障したストレージ装置と、二重化復旧処理中に新たに故障したストレージ装置とのそれぞれに二重化して格納されていたデータは、新たに故障したストレージ装置が接続されたディスクノードの切り離しにより失われる可能性がある。

【０００９】

本発明はこのような点に鑑みてなされたものであり、ディスクノードの障害発生時のデータの喪失を抑制することができるデータ割当制御プログラム、データ割当制御方法、およびデータ割当制御装置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】

【００１０】

上記課題を解決するために、接続されたストレージ装置内のデータを管理する複数のディスクノードに対する管理対象データの割り当て指示を、コンピュータに実行させるデータ割当制御プログラムが提供される。このデータ割当制御プログラムは、コンピュータに以下の処理を実行させる。

【００１１】

コンピュータは、複数のストレージ装置に格納されたデータのうち同一内容の冗長データが存在しない二重化欠損データのコピーを、二重化欠損データを管理するディスクノードに指示する二重化復旧処理を行う。コンピュータは、二重化復旧処理中に、ストレージ装置に対するデータのライトエラーを示すライトエラー情報を受け取ると、ライトエラーが発生したストレージ装置の識別情報をエラー記憶手段に格納する。コンピュータは、エ

ラー記憶手段を参照してライトエラーが発生したストレージ装置を判断し、ライトエラーが発生したストレージ装置に格納されているデータをコピー対象データとし、ライトエラーが発生していないストレージ装置内にコピー対象データの冗長データが存在しない場合、コピー対象データを管理するディスクノードに対して、ライトエラーが発生していないストレージ装置へのコピー対象データのコピーを指示する。

【００１２】

また上記課題を解決するために、上記データ割当制御プログラムに基づいてコンピュータが実行する処理と同様の処理を行うデータ割当制御方法が提供される。さらに上記課題を解決するために、上記データ割当制御プログラムを実行するコンピュータが有する機能と同様の機能を有するデータ割当制御装置が提供される。

【発明の効果】

【００１３】

二重化復旧処理中にライトエラーとなったストレージ装置内のデータの喪失を抑制することができる。

【図面の簡単な説明】

【００１４】

【図１】実施の形態の概要を示す図である。

【図２】第２の形態のマルチノードストレージシステム構成の一例を示す図である。

【図３】第２の形態に用いる制御ノードのハードウェアの一構成例を示す図である。

【図４】論理ディスクのデータ構造の一例を示す図である。

【図５】第２の実施の形態に係るマルチノードストレージシステムの各装置の機能を示すブロック図である。

【図６】ストレージ装置のデータ構造の一例を示す図である。

【図７】メタデータ記憶部のデータ構造例を示す図である。

【図８】論理ディスクメタデータ記憶部のデータ構造の一例を示す図である。

【図９】エラー記憶部のデータ構造の一例を示す図である。

【図１０】ディスクノードでのライトエラー検出時に装置間で送受信される情報の例を示す図である。

【図１１】故障したディスクノードの切り離し処理の一例を示す図である。

【図１２】リカバリ処理手順の一例を示すシーケンス図である。

【図１３】リカバリ処理実行中のライトエラー発生時の処理手順の一例を示すシーケンス図である。

【図１４】ディスクノードでのライトエラー検出および通知処理の手順の一例を示すフローチャートである。

【図１５】エラーメッセージを受信したときのエラー処理手順の一例を示すフローチャートである。

【図１６】リカバリ処理手順の一例を示すフローチャートである。

【図１７】ライトエラーが発生したディスクノードへのアクセス要求時の処理手順を示すシーケンス図である。

【図１８】アクセス処理手順の一例を示すフローチャートである。

【図１９】ライトエラーが発生したディスクノードへのミラーライト処理の手順を示すシーケンス図である。

【図２０】ライトエラーが発生したディスクノードへのスライスコピー処理時の処理手順を示すシーケンス図である。

【図２１】スライスコピーのデータを受信したディスクノードの処理手順の一例を示すフローチャートである。

【図２２】メタデータ照会要求時のスライス割当処理手順の一例を示すフローチャートである。

【図２３】スライスコピーエラー時のスライス割り当て処理手順の一例を示すフローチャートである。

【図２４】プライマリスライスとセカンダリスライスとが割り当てられているセグメントの割り当て変更処理手順の一例を示すフローチャートである。
【図２５】プライマリスライスとセカンダリスライスとが割り当てられているセグメントの割り当て変更処理例を示す図である。
【図２６】プライマリスライスとリザーブスライスとが割り当てられているセグメントの割り当て変更処理手順の一例を示すフローチャートである。
【図２７】プライマリスライスとリザーブスライスとが割り当てられているセグメントの割り当て変更処理例を示す図である。
【図２８】シングルプライマリスライスが割り当てられているセグメントの割り当て変更処理手順の一例を示すフローチャートである。
【図２９】シングルプライマリスライスが割り当てられているセグメントの割り当て変更処理例を示す図である。
【図３０】スライスコピーを伴うスライス割当処理手順の一例を示すシーケンス図である。
【図３１】スライスコピーを伴わないスライス割当処理手順の一例を示すシーケンス図である。
【発明を実施するための形態】
【００１５】
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〕
図１は、実施の形態の概要を示す図である。複数のディスクノード２～５には、それぞれストレージ装置２ａ，３ａ，４ａ，５ａが接続されている。ディスクノード２～５は、接続されたストレージ装置２ａ，３ａ，４ａ，５ａ内のデータを管理する。データ割当制御装置１は、ディスクノード２～５に対する管理対象データの割り当て指示を行う。例えば、データ割当制御装置１は、ネットワークを介してディスクノード２～５に接続されている。
【００１６】
データ割当制御装置１は、既に割り当てられているデータの割り当て先を変更する場合、データのコピーをディスクノードに指示する。データ割り当て先の変更は、例えば、ストレージ装置に対するデータ書き込み時のライトエラーに応じて行われる。ライトエラー発生時のデータコピーを適切に指示するため、データ割当制御装置１は、二重化復旧手段１ａ、管理情報記憶手段１ｂ、エラー情報受信手段１ｃ、エラー記憶手段１ｄ、およびコピー指示手段１ｅを有している。
【００１７】
二重化復旧手段１ａは、複数のストレージ装置２ａ，３ａ，４ａ，５ａに格納されたデータのうち同一内容の冗長データが存在しない二重化欠損データのコピーを、二重化欠損データを管理するディスクノードに指示する二重化復旧処理を行う。例えば二重化復旧手段１ａは、複数のストレージ装置２ａ，３ａ，４ａ，５ａ内の二重化欠損データを調査し、二重化欠損データの格納場所を示す管理情報を管理情報記憶手段１ｂに格納する。二重化復旧手段１ａは、管理情報記憶手段１ｂ内の管理情報に示される二重化欠損データを順次選択して、選択した二重化欠損データを管理するディスクノードに対して選択した二重化欠損データのコピーを指示する。選択した二重化欠損データのコピーの指示は、例えば、選択した二重化欠損データが格納されているストレージ装置と異なるストレージ装置であり、かつライトエラーが発生していないストレージ装置が、コピー先として指定される。
【００１８】
なお二重化復旧手段１ａは、管理情報記憶手段１ｂ内の管理情報で示されるすべての二重化欠損データのコピーが完了すると、ライトエラーが発生したストレージ装置を切り離し、その結果生じる二重化欠損データのコピーを行うことができる。例えば、二重化復旧手段１ａは、管理情報記憶手段１ｂ内の管理情報で示されるすべての二重化欠損データの

コピーが完了すると、エラー記憶手段１ｄを参照してライトエラーが発生したストレージ装置の有無を判断する。ライトエラーが発生したストレージ装置がある場合、二重化復旧手段１ａは、ライトエラーが発生したストレージ装置へのアクセスを停止することにより生じる二重化欠損データの管理情報を管理情報記憶手段１ｂに追加する。そして、二重化復旧手段１ａは、管理情報記憶手段１ｂに追加された二重化復旧処理の管理情報に基づいて、二重化復旧処理を続行する。

【００１９】

管理情報記憶手段１ｂは、少なくとも二重化欠損データの格納場所を示す管理情報を記憶する。

エラー情報受信手段１ｃは、二重化復旧処理中にストレージ装置に対するデータのライトエラーを示すライトエラー情報を受け取ると、ライトエラーが発生したストレージ装置の識別情報をエラー記憶手段１ｄに格納する。またエラー情報受信手段１ｃは、二重化復旧処理中以外の期間に、ストレージ装置に対するデータのライトエラーを示すライトエラー情報を受け取る場合もある。この場合、エラー情報受信手段１ｃは、二重化復旧手段１ａに、ライトエラーが発生したストレージ装置へのアクセスを停止することにより生じる二重化欠損データの二重化復旧処理を開始させる。

【００２０】

エラー記憶手段１ｄは、ライトエラーが発生したストレージ装置の識別情報を記憶する。なお、１台のディスクノードに１台のストレージ装置が接続されている場合、ディスクノードの識別情報を、ストレージ装置の識別情報として用いることもできる。また１台のディスクノードに複数のストレージ装置が接続されている場合、ディスクノードの識別情報と、ディスクノード内でのローカルなストレージ装置の識別情報とを組み合わせて、ストレージ装置の識別情報とすることもできる。

【００２１】

コピー指示手段１ｅは、エラー記憶手段１ｄを参照してライトエラーが発生したストレージ装置を判断する。そしてコピー指示手段１ｅは、ライトエラーが発生したストレージ装置に格納されているデータをコピー対象データとする。コピー指示手段１ｅは、ライトエラーが発生していないストレージ装置内にコピー対象データの冗長データが存在しない場合、コピー対象データを管理するディスクノードに対して、コピー対象データのコピーを指示する。コピー指示では、ライトエラーが発生していないストレージ装置がコピー先として指定される。例えば、コピー指示手段１ｅは、ライトエラーが発生したストレージ装置内のデータに対するアクセスで発生したエラーを示すデータアクセスエラー情報を他の装置から受け取ると、アクセス対象であったデータをコピー対象データとする。なお、コピー指示手段１ｅは、コピー対象データのコピーが完了すると、コピー先のデータの管理情報を二重化欠損データとして管理情報記憶手段１ｂに格納する。

【００２２】

なおコピー指示手段１ｅは、例えば、コピー対象データのコピーが完了すると、コピー先のデータの管理情報を、新たなアクセス先としてデータアクセスエラー情報を送信した他の装置に応答する。図１の例では、アクセスノード６が設けられており、アクセスノード６からデータアクセスエラー情報が送信されている。そのため、コピー指示手段１ｅは、アクセスノード６に対してコピー先のデータの管理情報を送信する。

【００２３】

各ディスクノード２～５は、ライトエラー発生時に、データ割当制御装置１へライトエラー情報を送信する。また図１の例では、アクセスノード６が各ディスクノード２～５で管理されているデータにアクセスする。

【００２４】

ここで、ディスクノード５が故障すると、所定のタイミングで二重化復旧手段１ａにより二重化復旧処理が開始される。例えば、二重化復旧手段１ａは、ディスクノード２～４で管理されているストレージ装置２ａ，３ａ，４ａ内のデータのうち、冗長データが存在しない二重化欠損データを調べる。二重化復旧手段１ａは、二重化欠損データの管理情報

を管理情報記憶手段１ｂに格納する。そして、二重化復旧手段１ａは、二重化欠損データを順次選択し、二重化欠損データを管理しているディスクノードに対し、二重化欠損データが格納されているストレージ装置とは別のストレージ装置への二重化欠損データのコピーを指示する。二重化欠損データのコピー指示を受けたディスクノードでは、二重化欠損データのコピーが行われる。例えば、ディスクノード３が、ストレージ装置３ａ内の二重化欠損データを、ディスクノード４に接続されたストレージ装置４ａにコピーする。また、ディスクノード４が、ストレージ装置４ａ内の二重化欠損データを、ディスクノード３に接続されたストレージ装置３ａにコピーする。

【００２５】

このような二重化復旧処理中に、ディスクノード２に接続されたストレージ装置２ａでライトエラーが発生したものとする。例えば、ストレージ装置２ａ内での定期的なデータの書き込みチェックが行われ、書き込みができず、かつストレージ装置２ａ内でのリカバリができないような場合に、ストレージ装置２ａからディスクノード２にライトエラーが出力される。ディスクノード２は、ストレージ装置２ａでのライトエラーの発生を検出すると、データ割当制御装置１にライトエラー情報を送信する。ライトエラー情報には、例えば、システム内でストレージ装置２ａを一意に識別する識別情報が含まれる。

【００２６】

ディスクノード２が出力したライトエラー情報は、データ割当制御装置１のエラー情報受信手段１ｃで受信される。エラー情報受信手段１ｃは、二重化復旧処理中であることを確認し、ライトエラーが発生したストレージ装置の識別情報をエラー記憶手段１ｄに格納する。その結果、エラー記憶手段１ｄには、ストレージ装置２ａにライトエラーが発生したことを示す情報が記憶される。

【００２７】

ストレージ装置２ａでライトエラーが発生すると、二重化復旧手段１ａは、その後の二重化欠損データのコピー先からストレージ装置２ａを除外する。

また、ストレージ装置２ａでライトエラーが発生後、アクセスノード６がストレージ装置２ａ内のデータへのアクセス要求をディスクノード２に送信すると、ディスクノード２からアクセスノード６にエラーが応答される。するとアクセスノード６は、データ割当制御装置１に、ストレージ装置２ａへのアクセスでエラーが発生したことを示すデータアクセスエラー情報を送信する。データアクセスエラー情報には、エラー発生時のアクセス先であったアクセス対象データの識別情報が示される。

【００２８】

データ割当制御装置１では、コピー指示手段１ｅがデータアクセスエラー情報を受信する。コピー指示手段１ｅは、データアクセスエラー情報に基づいて、エラー発生時のアクセス対象のデータを認識し、コピー対象データとする。そして、コピー指示手段１ｅは、コピー対象データの冗長データが、ライトエラーが発生していないストレージ装置内になければ、ディスクノード２に対して、コピー対象データの他のストレージ装置へのコピーを指示する。この際、コピー対象データのコピー先は、ライトエラーが発生していないいずれかのストレージ装置である。図１の例では、ストレージ装置２ａ内のコピー対象データの、ストレージ装置３ａへのコピーが指示されている。ディスクノード２は、コピー対象データのコピー指示に従って、コピー対象データをコピーする。

【００２９】

コピー指示手段１ｅは、コピー対象データのコピーが完了すると、コピー先のデータの場所を示す管理情報を、新たなアクセス先としてアクセスノード６に通知する。アクセスノード６は、コピー指示手段１ｅからの管理情報の通知を受けて、コピー先となるストレージ装置３ａを管理するディスクノード３へ再アクセスを行うことができる。

【００３０】

このように、二重化復旧処理中にストレージ装置２ａでライトエラーが発生しても、ストレージ装置２ａのシステムからの切り離しは行われない。すなわち、ストレージ装置２ａにライトエラーが発生しても、データを読み出せる可能性があるため、ストレージ装置

２ａの即時切り離しは行われない。そして、ライトエラーが発生したストレージ装置内のコピー対象データのうち、ライトエラーが発生していないストレージ装置に冗長データが存在しないコピー対象データについては、他のストレージ装置にコピーされる。コピー対象データのコピーが完了すると、そのコピー対象データについてはデータの喪失を免れたこととなる。すなわち、二重化復旧処理中にライトエラーとなったストレージ装置２ａ内のデータの喪失が抑制されている。

【００３１】

またライトエラーが発生したストレージ装置２ａが切り離されていないため、ストレージ装置２ａ内に二重化欠損データがあれば、二重化復旧処理によって、ストレージ装置２ａ内の二重化欠損データが他のストレージ装置にコピーされる。ストレージ装置２ａから他のストレージ装置への二重化欠損データのコピーが完了すれば、その二重化欠損データについてはデータ喪失を免れたこととなる。すなわち、二重化復旧処理中にライトエラーとなったストレージ装置内のデータの喪失が抑制されている。

【００３２】

さらにストレージ装置５ａが故障したことによる二重化復旧処理が完了した場合、すべてのデータに冗長データが存在することとなる。データとそのデータの冗長データとは、異なるストレージ装置に格納される。すると、二重化復旧処理が完了した場合、ライトエラーが発生したストレージ装置２ａ内のすべてのデータは、同一内容のデータが他のストレージ装置にも格納されていることとなる。従って、ストレージ装置５ａが故障したことによる二重化復旧処理が完了後は、ストレージ装置２ａを切り離しても、データを喪失せずに済む。そこで、ストレージ装置２ａでライトエラー発生時に管理情報記憶手段１ｂ内に登録されていたすべての二重化欠損データのコピー完了後は、ライトエラーが発生したストレージ装置２ａが切り離される。すると、ストレージ装置２ａ内のデータと同一内容の他のストレージ装置３ａ，４ａ内のデータが二重化欠損データとなり、コピーされる。その結果、正常に動作するストレージ装置３ａ，４ａ内ですべてのデータが二重化され、システムの信頼性を向上させることができる。

【００３３】

〔第２の実施の形態〕

第２の実施の形態は、論理ディスクを用いてデータアクセスを管理するマルチノードシステムの例である。

【００３４】

図２は、第２の形態のマルチノードストレージシステム構成の一例を示す図である。本実施の形態では、ネットワーク１０を介して、複数のディスクノード１００，２００，３００，４００、制御ノード５００、およびアクセスノード６００が接続されている。ディスクノード１００，２００，３００，４００それぞれには、ストレージ装置１１０，２１０，３１０，４１０が接続されている。

【００３５】

ストレージ装置１１０には、複数のハードディスク装置（ＨＤＤ）１１１，１１２，１１３，１１４が実装されている。ストレージ装置２１０には、複数のＨＤＤ２１１，２１２，２１３，２１４が実装されている。ストレージ装置３１０には、複数のＨＤＤ３１１，３１２，３１３，３１４が実装されている。ストレージ装置４１０には、複数のＨＤＤ４１１，４１２，４１３，４１４が実装されている。各ストレージ装置１１０，２１０，３１０，４１０は、内蔵するＨＤＤを用いたＲＡＩＤシステムである。本実施の形態では、各ストレージ装置１１０，２１０，３１０，４１０のＲＡＩＤ５のディスク管理サービスを提供する。

【００３６】

ディスクノード１００，２００，３００，４００は、接続されたストレージ装置１１０，２１０，３１０，４１０に格納されたデータを管理し、管理しているデータをネットワーク１０経由で端末装置２１，２２，２３に提供する。また、ディスクノード１００，２００，３００，４００は、冗長性を有するデータを管理している。すなわち、同一のデー

タが、少なくとも２つのディスクノードで管理されている。

【００３７】

制御ノード５００は、ディスクノード１００，２００，３００，４００を管理する。例えば、制御ノード５００は、ディスクノード１００，２００，３００，４００から新たなストレージ装置の接続通知を受け取ると、論理ディスクを介して接続されたストレージ装置にアクセスできるようにする。

【００３８】

アクセスノード６００には、ネットワーク２０を介して複数の端末装置２１，２２，２３が接続されている。また、アクセスノード６００には、論理ディスクが定義されている。そして、アクセスノード６００は、端末装置２１，２２，２３からの論理ディスクのデータのアクセス要求に応答して、ディスクノード１００，２００，３００，４００内の対応するデータへアクセスする。

【００３９】

なお、図２に示すように第２の実施の形態では、各ディスクノード１００，２００，３００，４００に、１台ずつのストレージ装置１１０，２１０，３１０，４１０が接続されている。そのため、ストレージ装置でのライトエラーなどの障害の発生は、そのストレージ装置が接続されたディスクノードの障害とみなすことができる。そして、障害が発生したストレージ装置を切り離す場合には、ディスクノードの切り離しが行われることとなる。

【００４０】

図３は、第２の形態に用いる制御ノードのハードウェアの一構成例を示す図である。制御ノード５００は、ＣＰＵ（Central Processing Unit）５０１によって装置全体が制御されている。ＣＰＵ５０１には、バス５０８を介してＲＡＭ（Random Access Memory）５０２と複数の周辺機器が接続されている。

【００４１】

ＲＡＭ５０２は、制御ノード５００の主記憶装置として使用される。ＲＡＭ５０２には、ＣＰＵ５０１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、ＲＡＭ５０２には、ＣＰＵ５０１による処理に必要な各種データが格納される。

【００４２】

バス５０８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）５０３、グラフィック処理装置５０４、入力インタフェース５０５、光学ドライブ装置５０６、および通信インタフェース５０７がある。

【００４３】

ＨＤＤ５０３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。ＨＤＤ５０３は、制御ノード５００の二次記憶装置として使用される。ＨＤＤ５０３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。

【００４４】

グラフィック処理装置５０４には、モニタ１１が接続されている。グラフィック処理装置５０４は、ＣＰＵ５０１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。

【００４５】

入力インタフェース５０５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５０５は、キーボード１２やマウス１３から送られてくる信号をＣＰＵ５０１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。

【００４６】
光学ドライブ装置５０６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。
【００４７】
通信インタフェース５０７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、他の装置との間でデータの送受信を行う。
以上のようなハードウェア構成によって、本実施の形態の処理機能を実現することができる。なお、図３では制御ノード５００のハードウェア構成を示したが、ディスクノード１００，２００，３００，４００、およびアクセスノード６００も同様のハードウェア構成で実現することができる。ただし、ディスクノード１００，２００，３００，４００は、図３に示した機能に加え、ストレージ装置１１０，２１０，３１０，４１０を外部接続するためのインタフェースを有している。
【００４８】
次に、マルチノードストレージシステムにおいて定義される論理ディスクのデータ構造について説明する。
図４は、論理ディスクのデータ構造の一例を示す図である。第２の実施の形態では、論理ディスク６０には論理ディスク識別子「ＬＶＯＬ－Ｘ」が付与されている。ネットワーク経由で接続された４台のディスクノード１００，２００，３００，４００には、個々のノードの識別用にそれぞれ「ＤＰ１」、「ＤＰ２」、「ＤＰ３」、「ＤＰ４」というディスクノードＩＤが付与されている。そして、各ディスクノード１００，２００，３００，４００に接続されているストレージ装置１１０，２１０，３１０，４１０は、ディスクノードＩＤと、各ディスクノード内でのディスクＩＤとの組によってネットワーク１０で一意に識別される。
【００４９】
各ディスクノード１００，２００，３００，４００が有するストレージ装置１１０，２１０，３１０，４１０それぞれにおいてＲＡＩＤ５のストレージシステムが構成されている。各ストレージ装置１１０，２１０，３１０，４１０で提供される記憶機能は、複数のスライス１１５ａ～１１５ｃ，２１５ａ～２１５ｃ，３１５ａ～３１５ｃ，４１５ａ～４１５ｃに分割されて管理されている。
【００５０】
論理ディスク６０は、セグメント６１～６４という単位で構成される。セグメント６１～６４の記憶容量は、ストレージ装置１１０，２１０，３１０，４１０における管理単位であるスライスの記憶容量と同じである。例えば、スライスの記憶容量が１ギガバイトとするとセグメントの記憶容量も１ギガバイトである。論理ディスク６０の記憶容量はセグメント１つ当たりの記憶容量の整数倍である。セグメント６１～６４は、それぞれプライマリスライス６１ａ，６２ａ，６３ａ，６４ａとセカンダリスライス６１ｂ，６２ｂ，６３ｂ，６４ｂとの組（スライスペア）で構成される。
【００５１】
同一セグメントに属する２つのスライスは別々のディスクノードに属する。個々のスライスを管理する領域には論理ディスク識別子やセグメント情報や同じセグメントを構成するスライス情報の他にフラグがあり、そのフラグにはプライマリあるいはセカンダリなどを表す値が格納される。
【００５２】
図４の例では、論理ディスク６０内のスライスの識別子を、「Ｐ」または「Ｓ」のアルファベットと数字との組合せで示している。「Ｐ」はプライマリスライスであることを示している。「Ｓ」はセカンダリスライスであることを示している。アルファベットに続く数字は、何番目のセグメントに属するのかを表している。例えば、１番目のセグメント６

１のプライマリスライスが「Ｐ１」で示され、セカンダリスライスが「Ｓ１」で示される。

【００５３】

図５は、第２の実施の形態に係るマルチノードストレージシステムの各装置の機能を示すブロック図である。まずアクセスノード６００の機能について説明する。アクセスノード６００は、メタデータ照会部６１０、アクセス用メタデータ記憶部６２０、およびアクセス要求部６３０を有している。

【００５４】

メタデータ照会部６１０は、論理ディスク６０を定義するメタデータを、制御ノード５００から取得する。具体的には、メタデータ照会部６１０は、アクセスノード６００の起動時に制御ノード５００に対して全メタデータの照会要求を送信する。すると、制御ノード５００から論理ディスク６０に関する全メタデータが送られてくる。また、メタデータ照会部６１０は、アクセス要求部６３０による任意のスライスへのデータアクセスがエラーとなったとき、アクセス対象のスライスが割り当てられたセグメントに関するメタデータの照会要求を制御ノード５００に送信する。この場合、メタデータの照会要求が、データアクセスエラー情報の通知も兼ねることとなる。すると、制御ノード５００から、該当セグメントの最新のメタデータが送られてくる。なお、メタデータ照会部６１０は、制御ノード５００からメタデータを取得すると、そのメタデータをアクセス用メタデータ記憶部６２０に格納する。

【００５５】

アクセス用メタデータ記憶部６２０は、論理ディスク６０を定義するメタデータを記憶する。例えば、アクセスノード６００のＲＡＭの一部がアクセス用メタデータ記憶部６２０として使用される。なお、第２の実施の形態では、アクセスノード６００は常にプライマリスライス（シングルプライマリスライスも含む）にアクセスを行う。そのため、アクセス用メタデータ記憶部６２０には、論理ディスク６０のメタデータのうち、少なくともプライマリスライスに関するメタデータが格納されていればよい。

【００５６】

アクセス要求部６３０は、端末装置２１，２２，２３からの論理ディスク６０上でのデータのアクセス要求に応答して、ストレージ装置１１０，２１０，３１０，４１０のデータのアクセス要求（リード要求またはライト要求）を送信する。アクセス要求は、アクセス先のストレージ装置が接続されたディスクノードに対して送信される。具体的には、アクセス要求部６３０は、論理ディスク６０のアドレスを指定したアクセス要求を受け取ると、まず、アクセス用メタデータ記憶部６２０を参照し、アクセス対象のデータが属するセグメントを判断する。次に、アクセス要求部６３０は、該当するセグメントにプライマリスライスとして割り当てられたスライスを判断する。そして、アクセス要求部６３０は、該当するスライスを管理するディスクノードに対して、そのスライス内のデータのアクセス要求を送信する。ディスクノードからアクセス結果が応答されると、アクセス要求部６３０は、端末装置２１，２２，２３にアクセス結果を送信する。

【００５７】

なお、アクセス要求部６３０は、アクセス先のディスクノードからエラーが返された場合、エラーが発生したセグメントをメタデータ照会部６１０に通知する。その後、アクセス要求部６３０は、データアクセスのリトライを行う。リトライでは、アクセス用メタデータ記憶部６２０を参照して、新たにプライマリスライスとした割り当てられたスライスを判断する処理から再度実行される。すなわち、前回のアクセス要求後にアクセス用メタデータ記憶部６２０内のメタデータが更新されていれば、更新後のメタデータに基づいてリトライ時のアクセス先となるディスクノードが判断される。

【００５８】

次に、ディスクノード１００の機能について説明する。ディスクノードは、アクセス処理部１２０、エラー通知部１３０、エラー記憶部１４０、メタデータ記憶部１５０、およびメタデータ管理部１６０を有する。

【００５９】
アクセス処理部１２０は、アクセスノード６００からのアクセス要求に応答して、ストレージ装置１１０に対するデータアクセスを行う。具体的には、アクセス処理部１２０は、アクセスノード６００からアクセス要求を受け取ると、メタデータ記憶部１５０を参照し、アクセス対象となるセグメントに対して、割り当てられているストレージ装置１１０内のスライスを判断する。
【００６０】
次に、アクセス処理部１２０は、アクセス要求で指定されているスライス内のデータに対してアクセスする。例えば、データリードのアクセス要求であれば、アクセス処理部１２０は、該当するデータをストレージ装置１１０から読み出す。また、データライトのアクセス要求であれば、アクセス処理部１２０は、ストレージ装置１１０内の該当する記憶領域にアクセス要求に含まれるデータを書き込む。そして、アクセス処理部１２０は、アクセス結果をアクセスノード６００に送信する。データリードのアクセス要求の場合、ストレージ装置１１０から読み出したデータがアクセス結果に含まれる。
【００６１】
また、データライトのアクセス要求の場合、アクセス処理部１２０は、他のディスクノード２００，３００，４００との間でミラーリング処理を実行する。ミラーリング処理では、例えばアクセス処理部１２０は、メタデータ記憶部１５０を参照し、ライトアクセスの対象となるストレージ装置１１０内のスライス（プライマリスライス）と組となる、他のディスクノードのスライス（セカンダリスライス）を判断する。そして、アクセス処理部１２０は、セカンダリスライスを管理するディスクノードに対して、プライマリスライスに書き込まれたデータと同じデータの書き込み要求を送信する。すると、データの書き込み要求を受信したディスクノードにおいてセカンダリスライスへのデータの書き込みが行われる。
【００６２】
ミラーリング処理では、他のディスクノードのプライマリスライスに書き込まれたデータのセカンダリスライスへの書き込み要求が、他のディスクノードからディスクノード１００に送信されることがある。ディスクノード１００に送られたセカンダリスライスへのデータの書き込み要求は、アクセス処理部１２０で受け取られる。アクセス処理部１２０は、セカンダリスライスへのデータの書き込み要求を受け取ると、ストレージ装置１１０内の該当スライスにデータの書き込みを行う。
【００６３】
ストレージ装置１１０へのデータ書き込みの際には、ライトエラーが発生する場合がある。ライトエラーが発生すると、アクセス処理部１２０は、エラー記憶部１４０にライトエラーの発生を示すエラー情報を格納すると共に、エラー通知部１３０に対してライトエラーの発生を通知する。さらにアクセス処理部１２０は、アクセスノード６００からのアクセス要求に応じて実行したストレージ装置１１０へのライトアクセスがエラーとなった場合、アクセスノード６００に対してエラーを返す（エラーメッセージを送信する）。
【００６４】
エラー通知部１３０は、アクセス処理部１２０からライトエラー発生の通知を受け取ると、制御ノード５００へエラーの発生を示すエラーメッセージを送信する。例えばエラー通知部１３０は、ディスクノード１００の識別子（ディスクＩＤ）を含むエラーメッセージを制御ノード５００に送信する。
【００６５】
エラー記憶部１４０は、ストレージ装置１１０のエラー情報を記憶する。エラー情報は、ライトエラーの有無を示す情報である。例えばエラー情報として、「ライトエラーあり」または「ライトエラーなし」の２つの状態を有するフラグを用いることができる。エラー記憶部１４０としては、例えば、ディスクノード１００内のＲＡＭの記憶領域の一部が用いられる。
【００６６】

なお、図５の例ではディスクノード１００に対して１台のストレージ装置１１０が接続されている。そのため、エラー情報に、ストレージ装置の識別情報を含めなくても、エラー情報に対応するストレージ装置を一意に決定できる。ディスクノード１００に複数のストレージ装置が接続されていた場合、ストレージ装置ごとのエラー情報がエラー記憶部１４０に格納される。その場合、各エラー情報には、ストレージ装置をディスクノード１００内で一意に識別する識別子が含められる。

【００６７】

メタデータ記憶部１５０は、ディスクノード１００が管理しているスライスのメタデータの記憶機能である。例えば、ディスクノード１００のＲＡＭ内の一部の記憶領域がメタデータ記憶部１５０として使用される。

【００６８】

メタデータ管理部１６０は、ストレージ装置１１０内の各スライスのメタデータを管理する。具体的には、メタデータ管理部１６０は、ディスクノード１００起動時に、ストレージ装置１１０から各スライスのメタデータを読み出し、メタデータ記憶部１５０に格納する。また、メタデータ管理部１６０は、制御ノード５００からメタデータの収集要求があれば、メタデータ記憶部１５０に格納されているメタデータを制御ノード５００に送信する。さらにメタデータ管理部１６０は、制御ノード５００からメタデータの変更要求を受け取ると、その変更要求で指定されたメタデータの内容を変更する。この際、メタデータ管理部１６０は、メタデータ記憶部１５０内のメタデータと、ストレージ装置１１０内のメタデータとを変更する。

【００６９】

なお、図５には複数のディスクノード１００，２００，３００，４００のうち、代表的にディスクノード１００の機能を示したが、他のディスクノード２００，３００，４００も同様の機能を有している。

【００７０】

次に制御ノード５００の機能について説明する。制御ノード５００は、エラー受信部５１０、エラー記憶部５２０、論理ディスク管理部５３０および論理ディスクメタデータ記憶部５４０を有している。

【００７１】

エラー受信部５１０は、ディスクノードからのライトエラーを示すエラーメッセージを受信する。エラー受信部５１０は、エラーメッセージを受信すると、二重化復旧のリカバリ処理中でなければ、エラーメッセージを送信したディスクノードでライトエラーが発生したことを示すエラー情報を、エラー記憶部５２０に格納する。また、エラー受信部５１０は、エラーメッセージを受信したときリカバリ処理中であれば、論理ディスク管理部５３０内のリカバリ処理部５３３にリカバリ処理の開始を指示する。

【００７２】

エラー記憶部５２０は、ディスクノードでのライトエラーの有無を示すエラー情報を記憶する。例えば、ＲＡＭ５０２またはＨＤＤ５０３の記憶領域の一部が、エラー記憶部５２０として使用される。

【００７３】

論理ディスク管理部５３０は、マルチノードストレージシステム内の各ディスクノード１００，２００，３００，４００のメタデータを管理する。論理ディスク管理部５３０は、メタデータ収集部５３１、スライス割当処理部５３２、およびリカバリ処理部５３３を有する。

【００７４】

メタデータ収集部５３１は、所定のタイミングで、各ディスクノード１００，２００，３００，４００からメタデータを収集する。例えば、メタデータ収集部５３１は、制御ノード５００が起動した場合にメタデータを収集する。メタデータを収集するタイミングになると、メタデータ収集部５３１は、各ディスクノード１００，２００，３００，４００にメタデータの収集要求を送信する。すると各ディスクノード１００，２００，３００，

４００からメタデータが応答される。メタデータ収集部５３１は、収集したメタデータを論理ディスクメタデータ記憶部５４０に格納する。

【００７５】

スライス割当処理部５３２は、論理ディスク６０の各セグメントへのスライスの割り当て処理を行う。例えば、スライス割当処理部５３２は、新たな論理ディスクが定義された場合などに、セグメントへのスライスの割り当て処理を行う。スライスの割り当て処理では、例えば、各セグメントに対して、異なるディスクノードで管理されている２つのスライスが割り当てられる。セグメントに割り当てるスライスとしては、他のセグメントに割り当てられておらず、かつ異常スライスではないスライス（フリースライス）が選択される。スライス割当処理部５３２は、新たにセグメントに割り当てられたスライスに関する論理ディスクメタデータ記憶部５４０内のメタデータを更新する。また、スライス割当処理部５３２は、更新されたメタデータを含むメタデータ変更要求を、そのメタデータに対応するスライスを管理しているディスクノードに送信する。

【００７６】

さらにスライス割当処理部５３２は、アクセスノード６００からのメタデータ照会要求を受け取ると、照会要求で指定されたセグメントに割り当てられたスライスのメタデータを、アクセスノード６００に送信する。なお照会要求で指定されたセグメントに割り当てられたスライスが、ライトエラーが発生したディスクノードで管理されている場合がある。その場合、スライス割当処理部５３２は、指定されたセグメントに他のスライスを割り当て、その後、新たに割り当てられたスライスのメタデータをアクセスノード６００に送信する。

【００７７】

リカバリ処理部５３３は、所定のタイミングで、二重化復旧のリカバリ処理を実行する。例えば、リカバリ処理部５３３は、エラー受信部５１０からリカバリ指示を受け取ったときにリカバリ処理を開始する。また、リカバリ処理部５３３は、システム管理者の操作入力によるリカバリ処理開始指示があった場合に、リカバリ処理を開始することもできる。

【００７８】

論理ディスクメタデータ記憶部５４０は、論理ディスク６０を構成するセグメントへのスライスの割り当て関係を示すメタデータを記憶する記憶機能である。例えば、ＲＡＭ５０２またはＨＤＤ５０３の記憶領域の一部が論理ディスクメタデータ記憶部５４０として使用される。

【００７９】

次に、各ノードで管理されているデータの構造について説明する。

図６は、ストレージ装置のデータ構造の一例を示す図である。ストレージ装置１１０には、スライス１１５ａ，１１５ｂ，１１５ｃ，・・・とは別に複数のメタデータ１１７ａ，１１７ｂ，１１７ｃ，・・・が格納されている。

【００８０】

ストレージ装置１１０に格納されたメタデータ１１７ａ，１１７ｂ，１１７ｃ，・・・は、ディスクノード１００の起動時にメタデータ管理部１６０によって読み出され、メタデータ記憶部１５０に格納される。

【００８１】

なお、図６の例では、各スライスに隣接した記憶領域に、それぞれのスライスに関するメタデータが格納されているが、メタデータの格納領域は図６に示した場所には限らない。例えば、１つのメタデータ記憶領域内に、すべてのスライスのメタデータを記憶させることもできる。

【００８２】

また、１つのスライスのデータを不連続の記憶領域に格納することもできる。その場合、例えば、各スライスが、所定のデータ長の複数のデータユニットに分割される。そして、各スライスのメタデータには、各データユニットに対応するデータの格納先となる記憶

領域を特定する情報が設定される。データの格納先となる記憶領域を特定する情報は、例えば、ストレージ装置の記憶領域の先頭からのオフセット値である。

【００８３】

図７は、メタデータ記憶部のデータ構造例を示す図である。メタデータ記憶部１５０には、メタデータテーブル１５１が格納されている。メタデータテーブル１５１には、ディスクノードＩＤ、ディスクＩＤ、スライスＩＤ、状態、論理ディスクＩＤ、セグメントＩＤ、論理ディスクアドレス、ペアのディスクノードＩＤ、ペアのディスクＩＤ、およびペアのスライスＩＤの欄が設けられている。メタデータテーブル１５１内の横方向に並べられた情報同士が互いに関連付けられ、メタデータを示す１つのレコードを構成している。

【００８４】

ディスクノードＩＤの欄は、ストレージ装置１１０を管理しているディスクノード１００の識別情報（ディスクノードＩＤ）が設定される。

ディスクＩＤの欄には、ストレージ装置１１０の識別情報（ディスクＩＤ）が設定される。ディスクノード１００に複数のストレージ装置が接続される場合、各ストレージ装置に異なるディスクＩＤが設定される。

【００８５】

スライスＩＤの欄には、メタデータに対応するスライスのストレージ装置１１０内での識別情報（スライスＩＤ）が設定される。

状態の欄には、スライスの状態を示す状態フラグが設定される。スライスが論理ディスク６０のセグメントに割り当てられていない場合、状態フラグ「Ｆ」が設定される。論理ディスク６０のセグメントのプライマリストレージに割り当てられている場合、状態フラグ「Ｐ」または「ＳＰ」が設定される。種別フラグ「Ｐ」は、ペアとなるセカンダリセグメント（ミラースライス）が存在するプライマリスライスを示している。種別フラグ「ＳＰ」は、ミラースライスが障害などによって存在しなくなったプライマリスライス（シングルプライマリスライス）を示している。なお、種別フラグ「ＳＰ」は、そのスライスが割り当てられているセグメントの二重化状態が損なわれていることを示す欠損フラグでもある。論理ディスク６０のセグメントのセカンダリストレージに割り当てられている場合、状態フラグ「Ｓ」が設定される。異常スライスと判定された場合、異常であることを示す状態フラグ「Ｂ」が設定される。異常スライスは、セグメントへの割り当て候補から除外される。

【００８６】

なお、シングルプライマリスライスが割り当てられているセグメントは冗長性が確保されていない。以下、シングルプライマリスライスのみが割り当てられているセグメントを、欠損セグメントと呼ぶ。欠損セグメントに対しては、セカンダリスライスとして割り当てるべきスライスが予約される。予約されたスライスは、リザーブスライスと呼ばれる。欠損セグメントのリザーブスライスに対してプライマリスライスからデータがコピーされる。これにより、欠損セグメントのデータの二重化状態が復旧する。このような二重化復旧処理中のリザーブスライスのメタデータに対しては、データのコピーが完了するまで種別フラグ「Ｒ」が設定される。リザーブスライスに対するプライマリスライスからのデータのコピーが完了すると、リザーブスライスはセカンダリスライスに変更される。

【００８７】

論理ディスクＩＤの欄には、スライスに対応するセグメントが属する論理ディスク６０を識別するための識別情報（論理ディスクＩＤ）が設定される。

セグメントＩＤの欄には、スライスが割り当てられたセグメントの識別情報（セグメントＩＤ）が設定される。

【００８８】

論理ディスクアドレスの欄には、スライスが割り当てられているセグメントの先頭を示す論理ディスク６０内でのアドレスが設定される。

ペアのディスクノードＩＤの欄には、ペアのスライス（同じセグメントに属する別のスライス）を有するストレージ装置を管理するディスクノードの識別情報（ディスクノード

ＩＤ）が設定される。

【００８９】

ペアのディスクＩＤの欄には、ペアのスライスを有するストレージ装置の識別情報（ディスクＩＤ）が設定される。

ペアのスライスＩＤの欄には、ペアのスライスを、そのスライスが属するストレージ装置内で識別するための識別情報（スライスＩＤ）が設定される。

【００９０】

図７にはディスクノード１００のメタデータ記憶部１５０の内容を示しているが、他のディスクノード２００，３００，４００も同様のメタデータ記憶部を有している。そして、各ディスクノード１００，２００，３００，４００のメタデータ記憶部に格納されたメタデータは、制御ノード５００からの要求に応じて制御ノード５００に送信される。制御ノード５００では、ディスクノード１００，２００，３００，４００から収集したメタデータは、メタデータ収集部５３１により論理ディスクメタデータ記憶部５４０に格納される。

【００９１】

図８は、論理ディスクメタデータ記憶部のデータ構造の一例を示す図である。論理ディスクメタデータ記憶部５４０には、論理ディスクメタデータテーブル５４１が格納されている。論理ディスクメタデータテーブル５４１には、ディスクノードＩＤ、ディスクＩＤ、スライスＩＤ、状態、論理ディスクＩＤ、セグメントＩＤ、論理ディスクアドレス、ペアのディスクノードＩＤ、ペアのディスクＩＤ、およびペアのスライスＩＤの欄が設けられている。論理ディスクメタデータテーブル５４１内の横方向に並べられた情報が互いに関連付けられ、メタデータを示す１つのレコードを構成している。論理ディスクメタデータテーブル５４１の各欄に設定される情報は、メタデータテーブル１５１の同名の欄と同種の情報である。

【００９２】

論理ディスクメタデータテーブル５４１に格納されたメタデータは、アクセスノード６００からの照会要求に応答して、アクセスノード６００に送信される。アクセスノード６００は、取得したメタデータを記憶する。具体的には、アクセスノード６００のアクセス用メタデータ記憶部６２０にメタデータが格納される。

【００９３】

アクセス用メタデータ記憶部６２０のデータ構造は、論理ディスクメタデータ記憶部５４０と同様である。なお、本実施の形態では、アクセスノード６００は常にプライマリスライスまたはシングルプライマリスライスにアクセスする。そのため、アクセス用メタデータ記憶部６２０には、少なくともプライマリスライスおよびプライマリスライスに関するメタデータ（状態フラグが「Ｐ」または「ＳＰ」のメタデータ）が格納されていればよい。また、アクセス用メタデータ記憶部６２０には、各メタデータにおけるペアのディスクノードＩＤ、およびペアのスライスＩＤの各欄のデータは無くてもよい。

【００９４】

図９は、エラー記憶部のデータ構造の一例を示す図である。エラー記憶部５２０には、各ディスクノード１００，２００，３００，４００のライトエラーの有無を示す情報が格納される。図９の例では、エラー記憶部５２０内にエラー管理テーブル５２１が設けられている。

【００９５】

エラー管理テーブル５２１には、ディスクノードＩＤ、ディスクＩＤ、およびライトエラーの欄が設けられている。ディスクノードＩＤの欄には、各ディスクノードの識別子が設定される。ディスクＩＤの欄には、各ディスクノードに接続されているストレージ装置のディスクＩＤが設定される。ライトエラーの欄には、対応するディスクノードのストレージ装置でライトエラーが発生したか否かを示す情報が設定される。例えば、ライトエラーの欄には、ライトエラーに関して「あり」を示す状態と、「なし」を示す状態とを有するフラグが設定される。

【００９６】
以上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、いずれかのディスクノードに故障が発生すると、リカバリ処理が実行される。ディスクノードの故障は、例えば、制御ノードで検知される。制御ノード５００は、ディスクノードからのハートビートの途絶や、ディスクノードからのライトエラーの発生を示すエラーメッセージの受信によって、故障を検知することができる。
【００９７】
以下、ディスクノード４００で最初にライトエラーが発生したことに起因するリカバリ処理中に、他のディスクノードでライトエラーが発生した場合を想定し、第２の実施の形態の処理を詳細に説明する。
【００９８】
図１０は、ディスクノードでのライトエラー検出時に装置間で送受信される情報の例を示す図である。図１０では、各装置の機能のうち、他の装置と通信を行う機能が示されている。
【００９９】
制御ノード５００の論理ディスク管理部５３０から各ディスクノード１００，２００，３００のメタデータ管理部１６０，２６０，３６０それぞれへは、メタデータ変更要求、スライスコピー要求などが送信される。ディスクノード１００のエラー通知部１３０から制御ノード５００のエラー受信部５１０へは、エラーメッセージが送信される。ディスクノード１００，２００，３００のアクセス処理部１２０，２２０，３２０間では、ミラーリングによる書き込みデータや、スライスコピー時の書き込みデータなどが互いに送信される。
【０１００】
アクセスノード６００のアクセス要求部６３０からディスクノード１００のアクセス処理部１２０へは、リードまたはライトのアクセス要求が送信される。なお、図１０には示していないが、アクセス要求部６３０からディスクノード２００，３００のアクセス処理部２２０，３２０へも、リードまたはライトのアクセス要求が送信される。アクセスノード６００のメタデータ照会部６１０から制御ノード５００の論理ディスク管理部５３０へは、メタデータ照会要求が送信される。
【０１０１】
このような装置間の通信による装置の協働動作により、故障したディスクノードの切り離し処理、欠損セグメントの二重化を回復するリカバリ処理、およびリカバリ処理中のライトエラーに起因するエラー処理が進行する。そこで、まず故障したディスクノードの切り離し処理について説明する。
【０１０２】
第２の実施の形態では、リカバリ処理の実行中を除き、故障したディスクノードはマルチノードストレージシステムから切り離される。故障したディスクノードの切り離しとは、アクセスノード６００からのアクセス先、またはミラーリングによるコピー先から、故障したディスクノードを排除することである。
【０１０３】
図１１は、故障したディスクノードの切り離し処理の一例を示す図である。以下、図１１に示す処理をステップ番号に沿って説明する。なお、図１１の例には、リカバリ処理が実行されていない状況下でディスクノード４００が故障した場合の処理手順が示されている。
【０１０４】
［ステップＳ１１］ディスクノード４００は、ストレージ装置４１０へのライトアクセス時にライトエラーを検出すると、ライトエラーの発生を示すエラーメッセージを制御ノード５００に送信する。
【０１０５】
［ステップＳ１２］制御ノード５００のエラー受信部５１０は、エラーメッセージを受

信すると、リカバリ処理が実行されていないことを確認後、論理ディスク管理部５３０にリカバリ指示を出す。

【０１０６】

［ステップＳ１３］論理ディスク管理部５３０内のスライス割当処理部５３２は、ライトエラーメッセージを送信したディスクノード４００を切り離すため、メタデータ更新処理を行う。メタデータ更新処理は、故障通知処理（ステップＳ１３ａ）と論理ディスクメタデータ更新処理（ステップＳ１３ｂ）とに分かれる。

【０１０７】

［ステップＳ１３ａ］論理ディスク管理部５３０は、まずシングルプライマリ化を指示するメタデータ変更要求をディスクノード１００，２００，３００に送信する。メタデータ変更要求は、ディスクノード４００が管理しているスライスとペアとなるスライスについて、シングルプライマリスライスに変更すること（シングルプライマリ化）の要求である。メタデータ変更要求には、故障したディスクノード４００のディスクノードＩＤが含まれる。

【０１０８】

［ステップＳ１４］シングルプライマリ化のメタデータ変更要求を受信したディスクノード１００では、メタデータ管理部１６０がシングルプライマリ化の処理を実行する。例えば、メタデータ管理部１６０は、メタデータ記憶部１５０内のメタデータを参照し、ペアのディスクノードＩＤの欄に、ディスクノード４００のディスクノードＩＤ「ＤＰ４」が設定されているメタデータを検索する。そして、メタデータ管理部１６０は、検索で合致したメタデータテーブル１５１内のメタデータについて、状態をシングルプライマリスライスを示す「ＳＰ」に変更する。またメタデータ管理部１６０は、検索で合致したメタデータテーブル１５１内のメタデータについて、ペアのディスクノードＩＤ、およびペアのスライスＩＤの欄に該当する領域の情報を削除する。情報が削除された領域の内容は、例えば「ＮＵＬＬ」となる。

【０１０９】

メタデータ管理部１６０は、メタデータ記憶部１５０内のメタデータの更新が完了すると、更新されたメタデータに対応するストレージ装置１１０内のメタデータの記憶領域に、更新されたメタデータを書き込む。これにより、ストレージ装置１１０内のメタデータが更新され、ストレージ装置１１０内のメタデータとメタデータ記憶部１５０内のメタデータとの同一性が保たれる。

【０１１０】

メタデータ管理部１６０は、メタデータ記憶部１５０とストレージ装置１１０とのメタデータの更新が完了すると、制御ノード５００に対してシングルプライマリ化完了応答を送信する。制御ノード５００では、ディスクノード１００からのシングルプライマリ化完了応答を受信することで、ディスクノード１００でのシングルプライマリ化処理が完了したことを認識する。

【０１１１】

なお、他のディスクノード２００，３００もディスクノード１００と同様にシングルプライマリ化処理を実行する。そしてディスクノード２００，３００は、シングルプライマリ化処理が完了すると、制御ノード５００に対してシングルプライマリ化完了応答を送信する。

【０１１２】

［ステップＳ１３ｂ］スライス割当処理部５３２は、各ディスクノード１００，２００，３００からシングルプライマリ化完了応答を受け取ると、論理ディスクメタデータ記憶部５４０内のメタデータを更新する。具体的にはスライス割当処理部５３２は、ディスクノード４００が管理しているスライスとペアとなるスライスのメタデータの状態を、シングルプライマリスライスに変更する。またスライス割当処理部５３２は、ディスクノード４００が管理しているスライスとペアとなるスライスのメタデータから、ペアのディスクノードＩＤ、およびペアのスライスＩＤの欄の情報を削除する。さらにスライス割当処理

部５３２は、ディスクノード４００が管理しているスライスのメタデータを論理ディスクメタデータ記憶部５４０から削除するか、あるいは該当メタデータの状態をすべて異常であることを示す状態フラグ「Ｂ」に変更する。

【０１１３】

スライス割当処理部５３２は、故障したディスクノード４００を除くすべてのディスクノード１００，２００，３００からメタデータを収集し、論理ディスクメタデータテーブル５４１を再作成することもできる。ディスクノード１００，２００，３００内のメタデータは故障通知に応答して更新されているため、最新のメタデータを収集すれば、ディスクノード４００がシステムに含まれない状態での論理ディスクメタデータテーブル５４１が作成できる。

【０１１４】

制御ノード５００で更新された論理ディスクメタデータテーブル５４１内のメタデータは、所定のタイミングでアクセスノード６００に通知される。例えば、制御ノード５００からの指示により、切り離されたディスクノード４００のアクセス処理部の機能を停止させる。するとアクセスノード６００がディスクノード４００にアクセスすると、アクセスエラーとなる。この場合、アクセスノード６００は、制御ノード５００にメタデータ照会要求を送信する。制御ノード５００は、メタデータ照会要求に応答して、最新のメタデータをアクセスノード６００に通知する。また制御ノード５００は、論理ディスクメタデータテーブル５４１内のメタデータを更新した場合、アクセスノード６００からのメタデータ照会要求を待たずに、更新されたメタデータをアクセスノード６００に送信してもよい。

【０１１５】

ディスクノード４００を切り離した後の論理ディスクメタデータテーブル５４１では、ディスクノード４００が管理しているスライスとペアとなるスライスがすべてシングルプライマリスライスに変更されている。これにより、その後のアクセスノード６００からのアクセスは、シングルプライマリ化されたスライスに対して実行されることとなる。また、ディスクノード４００が管理しているスライスとペアとなるスライスがすべてシングルプライマリスライスに変更されていることで、ディスクノード４００がミラーリング処理におけるコピー先となることもなくなる。さらに、ディスクノード４００が管理しているスライスのメタデータが削除されているか、あるいは状態フラグ「Ｂ」に変更される。これにより、その後のセグメントへのスライス割り当てにおいて、ディスクノード４００のスライスがセグメントに割り当てられることもなくなる。

【０１１６】

このようにして、マルチノードストレージシステムの各装置からのディスクノード４００へのアクセスを停止することができる。ディスクノード４００をアクセス先とする処理がなくなることで、マルチノードストレージシステムからのディスクノード４００の切り離しが完了する。

【０１１７】

スライス割当処理部５３２は、ディスクノード４００の切り離しが完了すると、リカバリ処理部５３３にリカバリ処理の開始を指示する。すると、リカバリ処理部５３３の制御によりリカバリ処理が実行される。

【０１１８】

図１２は、リカバリ処理手順の一例を示すシーケンス図である。以下、リカバリ処理の手順をステップ番号に沿って説明する。

［ステップＳ２１］制御ノード５００のリカバリ処理部５３３は、故障したディスクノード４００の切り離しが完了するとリカバリ処理を開始する。リカバリ処理は、リザーブスライスの割り当て処理（ステップＳ２１ａ）、スライスコピー処理（ステップＳ２１ｂ）、およびスライス状態更新処理（ステップＳ２１ｃ）に分かれる。

【０１１９】

［ステップＳ２１ａ］リカバリ処理部５３３は、論理ディスクメタデータテーブル５４

１を参照し、シングルプライマリスライス（状態フラグが「ＳＰ」のスライス）を選択する。図１２の例では、ディスクノード２００が管理するスライスが選択されている。シングルプライマリスライスが割り当てられているセグメントは、欠損セグメントである。そこでリカバリ処理部５３３は、論理ディスクメタデータテーブル５４１を参照し、欠損セグメントに対してセカンダリスライスとして割り当てるスライスを、フリースライス（状態が「Ｆ」のスライス）の中から選択する。選択されるフリースライスは、シングルプライマリスライスとは異なるディスクノードであり、かつライトエラーが発生していないディスクノードで管理されているスライスである。図１２の例では、ディスクノード３００で管理されているスライスが選択されている。

【０１２０】

リカバリ処理部５３３は、選択したシングルプライマリスライスを管理しているディスクノード２００に対して、スライスのプライマリ化を指示するメタデータ変更要求を送信する。プライマリ化を指示するメタデータ変更要求には、欠損セグメントに割り当てられているシングルプライマリスライスのスライスＩＤが含まれる。

【０１２１】

またリカバリ処理部５３３は、選択したフリースライスを管理するディスクノード３００に対して、リザーブスライスへの変更を指示するメタデータ変更要求を送信する。このメタデータ変更要求には、変更対象のメタデータに対応するスライスのスライスＩＤ、選択したシングルプライマリ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、および選択したシングルプライマリスライスのスライスＩＤが含まれる。

【０１２２】

［ステップＳ２２］ディスクノード２００のメタデータ管理部２６０は、シングルプライマリ化を指示するメタデータ変更要求を受け取ると、メタデータ変更要求で指定されたスライスをシングルプライマリスライスに変更する。具体的には、メタデータ管理部２６０は、メタデータ変更要求で示される変更対象のスライスのメタデータを、メタデータ記憶部内から選択する。次にメタデータ管理部２６０は、選択したメタデータの状態の欄の種別フラグを、「ＳＰ」に変更する。またメタデータ管理部２６０は、ストレージ装置２１０内の変更対象のスライスのメタデータについても、状態がプライマリスライスになるように変更する。そして、メタデータ管理部２６０は、メタデータの更新が完了すると、更新完了応答を制御ノード５００に送信する。

【０１２３】

［ステップＳ２３］リザーブスライスへの変更を指示するメタデータ変更要求を受信したディスクノード３００では、メタデータ管理部３６０が指定されたフリースライスの状態を、リザーブスライスに変更する。具体的には、メタデータ管理部３６０は、メタデータ変更要求で示される変更対象のスライスのメタデータを、メタデータ記憶部内から選択する。次にメタデータ管理部３６０は、選択したメタデータの状態の欄の種別フラグを、「Ｒ」に変更する。またメタデータ管理部３６０は、ストレージ装置３１０内の変更対象のスライスのメタデータについても、状態がリザーブスライスとなるように変更する。そして、メタデータ管理部３６０は、メタデータの更新が完了すると、更新完了応答を制御ノード５００に送信する。

【０１２４】

なお、リカバリ処理部５３３は、ステップＳ２２，Ｓ２３のメタデータ更新が完了すると、論理ディスクメタデータ記憶部５４０内のメタデータを、ディスクノード２００，３００で更新されたメタデータと同じ内容に更新する。

【０１２５】

［ステップＳ２１ｂ］リカバリ処理部５３３は、ディスクノード２００，３００からメタデータ更新完了の応答を受け取ると、欠損セグメントのプライマリスライスを管理しているディスクノード２００に対して、スライスコピー要求を送信する。スライスコピー要求には、欠損セグメントのプライマリスライスのスライスＩＤ、リザーブ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、およびリザーブスライスのスライスＩＤが含

まれる。

【０１２６】

［ステップＳ２４］スライスコピー要求を受信したディスクノード２００のアクセス処理部２２０は、指定されたスライスのペアのスライスを管理するディスクノード３００に対して、指定されたスライス内のデータを送信する。データを送信する際には、データの格納先として、スライスコピー要求で指定されたスライスのペアのスライスのスライスＩＤが指定される。

【０１２７】

［ステップＳ２５］ディスクノード３００のアクセス処理部３２０は、ディスクノード２００から送られたデータを受信し、指定されたスライスにデータを格納する。アクセス処理部３２０は、データの格納が完了すると、書き込み完了応答をディスクノード２００に送信する。

【０１２８】

ディスクノード２００のアクセス処理部２２０は、ディスクノード３００からの書き込み完了応答を受け取ると、制御ノード５００に対してスライスコピー完了の応答を送信する。

【０１２９】

［ステップＳ２１ｃ］制御ノード５００のリカバリ処理部５３３は、ディスクノード２００からスライスコピー完了の応答を受け取ると、ディスクノード２００に対してメタデータへのペアのスライスの設定を指示するメタデータ変更要求を送信する。このメタデータの変更要求には、欠損セグメントのプライマリスライスのスライスＩＤ、リザーブ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、およびリザーブスライスのスライスＩＤが含まれる。

【０１３０】

また、リカバリ処理部５３３は、ディスクノード２００からスライスコピー完了の応答を受け取ると、ディスクノード３００に対してスライスのセカンダリ化を指示するメタデータ変更要求を送信する。セカンダリ化を指示するメタデータ変更要求には、欠損セグメントに割り当てられているリザーブスライスのスライスＩＤが含まれる。

【０１３１】

［ステップＳ２６］メタデータへのペアのスライスの設定を指示するメタデータ変更要求を受信したディスクノード２００では、メタデータ管理部２６０がペア相手設定処理を行う。具体的には、メタデータ管理部２６０は、メタデータ変更要求で示される変更対象のスライスのメタデータを、メタデータ記憶部内から選択する。次にメタデータ管理部２６０は、選択したメタデータのペアのディスクノードＩＤ、ペアのスライスＩＤの欄に、メタデータ変更要求に含まれているリザーブスライスのディスクノードＩＤとリザーブスライスのスライスＩＤとを設定する。またメタデータ管理部２６０は、ストレージ装置２１０内の変更対象のスライスのメタデータに対しても、ペアのディスクノードＩＤとペアのスライスＩＤとを設定する。そして、メタデータ管理部２６０は、メタデータの更新が完了すると、更新完了応答を制御ノード５００に送信する。

【０１３２】

［ステップＳ２７］ディスクノード３００のメタデータ管理部３６０は、セカンダリ化を指示するメタデータ変更要求を受け取ると、メタデータ変更要求で指定されたスライスをセカンダリスライスに変更する。具体的には、メタデータ管理部３６０は、メタデータ変更要求で示される変更対象のスライスのメタデータを、メタデータ記憶部内から選択する。次にメタデータ管理部３６０は、選択したメタデータの状態の欄の種別フラグを「Ｓ」に変更する。またメタデータ管理部３６０は、ストレージ装置３１０内の変更対象のスライスのメタデータについても、状態がセカンダリスライスになるように変更する。そして、メタデータ管理部３６０は、メタデータの更新が完了すると、更新完了応答を制御ノード５００に送信する。

【０１３３】

なお、リカバリ処理部５３３は、ステップＳ２６，Ｓ２７のメタデータ更新が完了すると、論理ディスクメタデータ記憶部５４０内のメタデータを、ディスクノード２００，３００で更新されたメタデータと同じ内容に更新する。

【０１３４】

以後、制御ノード５００のリカバリ処理部５３３は、シングルプライマリスライスが無くなるまで、ステップＳ２１ａ，Ｓ２１ｂ，Ｓ２１ｃの処理を繰り返し実行する。

ところで、図１１、図１２の例では、リカバリ処理が実行されていない状況下においてディスクノード４００でエラーが発生したため、ディスクノード４００の切り離し処理と、リカバリ処理とが実行されている。一方、リカバリ処理実行中にいずれかのディスクノードでライトエラーが発生した場合、データの喪失を抑制するため、エラーが発生したディスクノードの即時の切り離しは行われない。以下、リカバリ処理実行中のライトエラーの発生によるエラー処理について説明する。

【０１３５】

図１３は、リカバリ処理実行中のライトエラー発生時の処理手順の一例を示すシーケンス図である。図１３には、リカバリ処理部５３３によるリカバリ処理実行中に、ディスクノード１００においてストレージ装置１１０へのライトエラーが検出された場合の処理が示されている。以下、図１３に示す処理をステップ番号に沿って説明する。

【０１３６】

［ステップＳ３１］ディスクノード１００のアクセス処理部１２０は、ストレージ装置１１０へのライトアクセス時にエラーを検出すると、ライトエラーの発生をエラー通知部１３０に伝える。するとエラー通知部１３０は、ライトエラーを示すエラーメッセージを制御ノード５００に送信する。

【０１３７】

［ステップＳ３２］制御ノード５００では、エラー受信部５１０がディスクノード１００からのエラーメッセージを受け取る。図１３の例では、リカバリ処理部５３３によるリカバリ処理実行中であるため、エラー受信部５１０はエラー記憶部５２０に、ディスクノード１００でライトエラーが発生したことを示すエラー情報を格納する。

【０１３８】

このように、リカバリ処理が継続して実行されている間にディスクノード１００でライトエラーが検出されると、ディスクノード１００の切り離し処理に代えて、ディスクノード１００でライトエラーが発生したことを示すエラー情報の記録が行われる。

【０１３９】

次に、ディスクノード１００におけるライトエラー検出および通知処理例について説明する。

図１４は、ディスクノードでのライトエラー検出および通知処理の手順の一例を示すフローチャートである。以下、図１４に示す処理をステップ番号に沿って説明する。

【０１４０】

［ステップＳ４１］ディスクノード１００のアクセス処理部１２０は、ライトエラーを検出する。例えば、アクセス処理部１２０は、アクセスノード６００からのデータライトのアクセス要求に応じてストレージ装置１１０に書き込みを行う際に、書き込みに失敗すると、ライトエラーが発生したものと判断する。また、アクセス処理部１２０は、他のディスクノード２００，３００からのミラーリング処理におけるライト要求に応じてストレージ装置１１０に書き込みを行う際に、書き込みに失敗したとき、ライトエラーが発生したものと判断することもできる。さらに、アクセス処理部１２０は、他のディスクノード２００，３００からのスライスデータのコピー要求に応じてストレージ装置１１０に書き込みを行う際に、書き込みに失敗すると、ライトエラーが発生したものと判断することもできる。

【０１４１】

なお、ストレージ装置１１０内で定期的なライトエラーチェック処理が自動実行されている場合もある。この場合、エラーチェック処理においてライトエラーが検出されたこと

の通知をストレージ装置１１０から受け取ることで、アクセス処理部１２０がライトエラーの発生を認識できる。

【０１４２】

ところで、ストレージ装置は、多くの場合、データの書き込みができない単位記憶領域（ブロック）が発生すると、予備に用意されていたブロックにデータの書き込みを行う。このような書き込み先を変更する処理により、ストレージ装置内でライトエラーに対するリカバリが完了する。ストレージ装置ないでリカバリが完了した場合、ストレージ装置は、ディスクノードからのデータのライト要求に対してエラー応答をせずにすむ。ただし、ストレージ装置における予備のブロックの数には限りがある。予備のブロックがすべて使用された後にデータの書き込みができないブロックが発生すると、そのブロックに書き込むべきデータの書き込み先が確保できない。この場合、ストレージ装置はディスクノードに対してライトエラーを応答する。アクセス処理部１２０は、ストレージ装置１１０内部でのリカバリが不能となった場合に出力されるライトエラーについてのみ、制御ノードへのエラー通知の対象のライトエラーと判断することができる。

【０１４３】

［ステップＳ４２］ライトエラーを検出したアクセス処理部１２０は、ライトエラーが発生したことを示すエラー情報を、ライトエラーが発生したストレージ装置のディスクＩＤに対応付けてエラー記憶部１４０に格納する。例えばアクセス処理部１２０は、エラー記憶部１４０に格納されているライトエラーの発生の有無を示すフラグを、「ライトエラー有り」の状態に変更する。

【０１４４】

［ステップＳ４３］アクセス処理部１２０は、エラー通知部１３０に対して、ライトエラーの発生を通知する。エラー通知部１３０は、アクセス処理部１２０からのエラー発生の通知を受けて、制御ノード５００に対して、ライトエラーの発生を示すエラーメッセージを送信する。

【０１４５】

このようにして、ディスクノード１００において、ストレージ装置１１０のライトエラー発生を検出すると、ライトエラーを示すエラー情報が記録されると共に、エラーメッセージが制御ノード５００に送信される。エラーメッセージを受信した制御ノード５００では、リカバリ処理を開始するか、あるいはエラー情報の記録が行われる。

【０１４６】

図１５は、エラーメッセージを受信したときのエラー処理手順の一例を示すフローチャートである。以下、図１５に示す処理をステップ番号に沿って説明する。

［ステップＳ５１］制御ノード５００のエラー受信部５１０は、ディスクノード１００からのライトエラーを示すエラーメッセージを受信する。

【０１４７】

［ステップＳ５２］エラー受信部５１０は、リカバリ処理中か否かを判断する。例えば、リカバリ処理部５３３の実行プロセスが起動されている場合、リカバリ処理中と判断される。また、論理ディスクメタデータ記憶部５４０を参照し、シングルプライマリスライスが存在している間はリカバリ処理中であると判断することもできる。リカバリ処理中であれば、処理がステップＳ５３に進められる。リカバリ処理中でなければ、処理がステップＳ５４に進められる。

【０１４８】

［ステップＳ５３］エラー受信部５１０はリカバリ処理中であれば、エラー記憶部５２０に、ディスクノード１００でライトエラーが発生したことを示すエラー情報を記録する。例えばエラー受信部５１０は、エラー記憶部５２０内のエラー管理テーブル５２１から、エラーメッセージを送信したディスクノードのディスクノードＩＤを検索する。そしてエラー受信部５１０は、検索で該当したディスクノードＩＤに対応付けられたライトエラーの欄に、ライトエラー有りを示すフラグを設定する。その後、エラー処理が終了する。

【０１４９】

［ステップＳ５４］エラー受信部５１０はリカバリ処理中でなければ、リカバリ処理の開始を、リカバリ処理部５３３に指示する。これにより、リカバリ処理部５３３によるリカバリ処理が開始される。その後、エラー処理が終了する。

【０１５０】

このようにして、ライトエラーのエラーメッセージ受信時には、リカバリ処理中か否かによって処理が分かれる。すなわち、リカバリ処理中であればエラー情報の記録が行われ、リカバリ処理中でなければリカバリ処理が開始される。

【０１５１】

図１６は、リカバリ処理手順の一例を示すフローチャートである。以下、図１６に示す処理をステップ番号に沿って説明する。

［ステップＳ６１］リカバリ処理部５３３は、エラーメッセージを送信したディスクノードを切り離し対象とする。このとき、リカバリ処理部５３３は、エラー管理テーブル５２１から、切り離し対象のディスクノードに対応するレコードを削除する。

【０１５２】

［ステップＳ６２］リカバリ処理部５３３は、切り離し対象のディスクノードが管理するスライスとペア相手のスライスを検出する。例えば、リカバリ処理部５３３は、論理ディスクメタデータテーブル５４１のペアのディスクノードＩＤの欄から、切り離し対象のディスクノードのディスクノードＩＤを検索する。検索により該当したメタデータに対応するスライスが、切り離し対象のディスクノードが管理するスライスとペア相手である。

【０１５３】

［ステップＳ６３］リカバリ処理部５３３は、ペア相手のスライスを管理するディスクノードに対して、ペア相手のスライスのシングルプライマリ化を指示する。例えば、リカバリ処理部５３３は、ペア相手のスライスのスライスＩＤを含むシングルプライマリ化を指示するメタデータ変更要求を、ペア相手のスライスを管理するディスクノードに送信する。

【０１５４】

［ステップＳ６４］リカバリ処理部５３３は、すべてのペア相手のスライスのシングルプライマリ化が完了したか否かを判断する。例えばリカバリ処理部５３３は、シングルプライマリ化を指示するメタデータ変更要求のすべてに対して、シングルプライマリ化完了の応答が返された場合に、シングルプライマリ化が完了したものと判断する。シングルプライマリ化が完了した場合、処理がステップＳ６５に進められる。シングルプライマリ化が未完了であれば、ステップＳ６４が繰り返される。

【０１５５】

なお、ディスクノードでのシングルプライマリ化のメタデータ変更が完了すると、リカバリ処理部５３３は、論理ディスクメタデータテーブル５４１内のペア相手のスライスのメタデータについても、状態をシングルプライマリスライスに変更する。

【０１５６】

［ステップＳ６５］リカバリ処理部５３３は、シングルプライマリスライスが存在するか否かを判断する。例えばリカバリ処理部５３３は、論理ディスクメタデータテーブル５４１から、状態欄の種別フラグが「ＳＰ」のメタデータを検索する。少なくとも１つのメタデータが検索で見つかった場合、シングルプライマリスライスが存在すると判断される。種別フラグが「ＳＰ」のメタデータが検出されなければ、シングルプライマリスライスは存在しないと判断される。シングルプライマリスライスが存在する場合、処理がステップＳ６６に進められる。シングルプライマリスライスが存在しない場合、処理がステップＳ７１に進められる。

【０１５７】

［ステップＳ６６］リカバリ処理部５３３は、シングルプライマリスライスの１つを、リカバリ対象として選択する。例えばリカバリ処理部５３３は、論理ディスクメタデータテーブル５４１から種別フラグが「ＳＰ」のメタデータを１つ抽出し、抽出したメタデータに対応するスライスをリカバリ対象として選択する。

【０１５８】
［ステップＳ６７］リカバリ処理部５３３は、リザーブ化対象のスライスを選択する。例えばリカバリ処理部５３３は、論理ディスクメタデータテーブル５４１から種別フラグが「Ｆ」であり、かつリカバリ対象のスライスと異なるディスクノードで管理されているスライスのメタデータを１つ抽出する。そしてリカバリ処理部５３３は、抽出したメタデータに対応するスライスを、リザーブ化対象として選択する。
【０１５９】
この際、リカバリ処理部５３３は、エラー管理テーブル５２１を参照してライトエラーが発生しているディスクノードを判断し、ライトエラーが発生しているディスクノードで管理されているスライスを、リザーブ化対象として選択しないようにしてもよい。すなわち、ライトエラーが発生しているディスクノードが管理するスライスは、リザーブ対象として選択しても、スライスコピー時にエラーとなり、リザーブ化対象の再選択を行うこととなる可能性がある。ライトエラーが発生しているディスクノードで管理されているスライスをリザーブ化対象から除外することで、リザーブ化対象の再選択処理の発生を抑制できる。
【０１６０】
［ステップＳ６８］リカバリ処理部５３３は、リカバリ対象のスライスとリザーブ化対象のスライスに関するメタデータを変更する。例えば、リカバリ対象スライスは、シングルプライマリスライスからプライマリスライスに変更される。リザーブ化対象のスライスは、フリースライスからリザーブスライスに変更される。具体的にはリカバリ処理部５３３は、各スライスを管理しているディスクノードに対してメタデータ変更要求を送信することで、ディスクノード内のメタデータを変更する。また、リカバリ処理部５３３は、論理ディスクメタデータテーブル５４１内のリカバリ対象のスライスとリザーブ化対象のスライスとのそれぞれのメタデータを変更する。
【０１６１】
［ステップＳ６９］リカバリ処理部５３３は、プライマリスライスのデータのリザーブスライスへのコピーを指示するスライスコピー要求を、シングルプライマリスライスを管理しているディスクノードに送信する。コピーが完了すると、ディスクノードからスライスコピー完了の応答が返される。
【０１６２】
［ステップＳ７０］リカバリ処理部５３３は、リカバリ対象のスライスにペア相手のスライスを設定するメタデータの変更を行う。またリカバリ処理部５３３は、リザーブ化したスライスをセカンダリ化するメタデータの変更を行う。具体的には、リカバリ処理部５３３は、プライマリスライスを管理しているディスクノードに対して、ペア相手のスライスとしてリザーブスライスを指定したメタデータ変更要求を送信する。また、リカバリ処理部５３３は、リザーブスライスを管理するディスクノードに対して、リザーブスライスからセカンダリスライスへの変更を指示するメタデータ変更要求を送信する。さらにリカバリ処理部５３３は、論理ディスクメタデータテーブル５４１内のプライマリスライスとリザーブスライスとのそれぞれのメタデータを変更する。その後、処理がステップＳ６５に進められる。
【０１６３】
［ステップＳ７１］リカバリ処理部５３３は、シングルプライマリスライスが無くなると、切り離されていないライトエラー発生ディスクノードの有無を判断する。例えばリカバリ処理部５３３は、エラー記憶部５２０を参照し、ライトエラー有りのディスクノードがある場合、切り離されていないライトエラー発生ディスクノードがあると判断する。切り離されていないライトエラー発生ディスクノードがある場合、処理がステップＳ７２に進められる。切り離されていないライトエラー発生ディスクノードがない場合、処理が終了する。
【０１６４】
［ステップＳ７２］リカバリ処理部５３３は、切り離されていないライトエラー発生デ

ィスクノードを１台選択し、切り離し対象とする。このとき、リカバリ処理部５３３は、エラー管理テーブル５２１から、切り離し対象のディスクノードに対応するレコードを削除する。その後、処理がステップＳ６２に進められる。

【０１６５】

このように、リカバリ処理中にディスクノードでライトエラーが発生した場合、そのディスクノードの切り離しは、シングルプライマリスライスがなくなるまで延期される。シングルプライマリスライスがなくなるということは、すべてのセグメントのデータの二重化が復旧されている。シングルプライマリスライスがなくなると、ライトエラーが発生したディスクノードの切り離しと、その切り離しに伴って発生したシングルプライマリスライスの二重化復旧が行われる。

【０１６６】

ライトエラーが発生したディスクノードに対して外部からのアクセスがあった場合、アクセス対象のスライスが割り当てられたセグメントに関しては、ライトエラーが発生したディスクノードの切り離しを待たずに欠損セグメントに変更される。その結果、アクセス対象のスライスが割り当てられたセグメントに対する二重化復旧処理が実行される。ディスクノードに対する外部からのアクセスには、例えば、アクセスノード６００からのリードまたはライトのアクセス要求、他のディスクノードからのミラーライト要求、リカバリ処理中のスライスデータコピー時のスライスコピー要求または書き込み要求などがある。ライトエラーが発生したディスクノードに対して外部からアクセスがあると、アクセス対象のスライスが制御ノード５００に通知されることとなる。

【０１６７】

ただし、ライトエラーが発生したディスクノードが管理するデータに対する外部からのアクセスであっても、制御ノード５００からのデータリードのアクセスに関しては、その要求に従ってディスクノードでデータリードのアクセス処理が実行される。例えば、リカバリ処理やスライス割当処理におけるスライスコピー要求が、ライトエラーが発生したディスクノードに出された場合、そのディスクノードはスライスコピーのコピー元となる。コピー元はコピー対象のスライスのデータを読み出せればよいため、ライトエラーが発生したディスクノードでも実行可能である。

【０１６８】

ディスクノードに対してどのようなアクセスが行われたのかによって、アクセス対象のスライスを制御ノード５００に通知する手順が異なる。以下、アクセス対象のスライスを制御ノード５００に通知する手順について説明する。

【０１６９】

図１７は、ライトエラーが発生したディスクノードへのアクセス要求時の処理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。なお図１７の例では、ディスクノード１００ではライトエラーの発生を示すエラー情報が記憶されているものとする。また、制御ノード５００のリカバリ処理部５３３はリカバリ処理を実行中であるものとする。

【０１７０】

［ステップＳ８１］アクセスノード６００のアクセス要求部６３０は、端末装置からの要求に応じてディスクノード１００にリードまたはライトのアクセス要求を送信する。

［ステップＳ８２］アクセスノード６００からのアクセス要求は、ディスクノード１００のアクセス処理部１２０で受け取られる。アクセス処理部１２０は、エラー記憶部１４０を参照し、ライトエラーが発生していることを認識する。そして、アクセス処理部１２０は、アクセスノード６００に対してエラー応答を送信する。この処理の詳細は後述する（図１８参照）。

【０１７１】

［ステップＳ８３］エラー応答を受信したアクセスノード６００のアクセス要求部６３０は、メタデータ照会部６１０にメタデータの照会を依頼する。メタデータ照会部６１０は、アクセス要求のアクセス対象であったセグメントを指定して、該当セグメントのメタ

データ照会要求を制御ノード５００に送信する。

【０１７２】

［ステップＳ８４］制御ノード５００では、スライス割当処理部５３２がメタデータ照会要求を受け取る。スライス割当処理部５３２は、メタデータ照会要求を受信したことで、受信したメタデータ照会要求で示されたセグメントのプライマリスライスへのアクセスがエラーとなったことを認識する。またスライス割当処理部５３２はエラー記憶部５２０を参照し、エラーとなったスライスが、ライトエラーが検出されたディスクノード１００で管理されているスライスであることを認識する。そこで、スライス割当処理部５３２は、スライス割当処理を実行する。なお図１７の例では、ディスクノード２００が管理しているスライスが、メタデータ照会要求で指定されたセグメントのプライマリスライスとして割り当てられたものとする。スライス割当処理の詳細は後述する（図２２参照）。

【０１７３】

スライス割当処理が完了すると、スライス割当処理部５３２はメタデータ照会要求への応答として、メタデータ照会要求で指定されたセグメントのプライマリスライスのメタデータをアクセスノード６００に送信する。

【０１７４】

［ステップＳ８５］アクセスノード６００のアクセス要求部６３０は、取得したメタデータに従って、アクセスをリトライする。例えばアクセス要求部６３０は、取得したメタデータをアクセス用メタデータ記憶部６２０に格納する。またアクセス要求部６３０は、取得したメタデータに基づいて、アクセス対象のセグメントに割り当てられたスライスを判断し、該当スライス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。図１７の例では、ディスクノード２００に対してアクセス要求を送信する。

【０１７５】

［ステップＳ８６］ディスクノード２００のアクセス処理部２２０は、アクセス要求に応じてストレージ装置２１０へデータアクセスを行う。そしてアクセス処理部２２０は、アクセス結果をアクセスノード６００に応答する。

【０１７６】

次に、アクセスノード６００からのアクセス要求に応じたディスクノードにおけるアクセス処理手順について詳細に説明する。

図１８は、アクセス処理手順の一例を示すフローチャートである。以下、図１８に示す処理をステップ番号に沿って説明する。

【０１７７】

［ステップＳ９１］ディスクノード１００のアクセス処理部１２０は、アクセスノード６００からのデータリードまたはデータライトのアクセス要求を受信する。

［ステップＳ９２］アクセス処理部１２０は、エラー記憶部１４０を参照し、ライトエラーの有無を判断する。ライトエラーがある場合、処理がステップＳ９３に進められる。ライトエラーがなければ、処理がステップＳ９４に進められる。

【０１７８】

［ステップＳ９３］アクセス処理部１２０は、ライトエラーがある場合、アクセスノード６００に対してエラーを通知する。その後、アクセス処理が終了する。

［ステップＳ９４］アクセス処理部１２０は、ライトエラーがない場合、ストレージ装置１１０に対して、アクセス要求に応じたデータリードまたはデータライトのアクセスを行う。アクセス処理部１２０は、ストレージ装置１１０へのアクセスの結果を、アクセスノード６００に応答する。例えばデータリードのアクセスであれば、アクセス処理部１２０は、読み出したデータをアクセスノード６００に送信する。またデータライトのアクセスであれば、アクセス処理部１２０は、書き込み完了の応答メッセージをアクセスノード６００に送信する。その後、アクセス処理が終了する。

【０１７９】

このようにして、アクセス要求に応じてディスクノード１００でデータアクセスが実行される。

ディスクノード１００へのアクセスは、アクセスノード６００からのアクセス要求以外に、他のディスクノードからミラーライトのアクセスがある。以下、ミラーライト処理の手順について説明する。

【０１８０】

図１９は、ライトエラーが発生したディスクノードへのミラーライト処理の手順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。図１９の例では、ディスクノード３００がプライマリスライスを管理しており、ディスクノード１００がセカンダリスライスを管理しているセグメントに関して、アクセスノード６００がデータライトのアクセス要求を出力した場合を想定している。

【０１８１】

［ステップＳ１０１］アクセスノード６００のアクセス要求部６３０は、ディスクノード３００に対してデータライトのアクセス要求を送信する。

［ステップＳ１０２］ディスクノード３００のアクセス処理部３２０は、データライトのアクセス要求を受け取ると、ストレージ装置３１０にデータを書き込む。またアクセス処理部３２０は、セカンダリスライスを管理しているディスクノード１００に対してデータのミラーライト要求を送信する。

【０１８２】

［ステップＳ１０３］ディスクノード１００のアクセス処理部１２０は、ミラーライト要求を受け取ると、ライトエラーの有無を判断する。図１９の例では、すでにディスクノード１００においてライトエラーが発生している。アクセス処理部１２０は、ライトエラーがある場合、ミラーライトエラーをディスクノード３００に応答する。

【０１８３】

［ステップＳ１０４］ディスクノード３００のアクセス処理部３２０は、ディスクノード１００からのミラーライト応答を受け取ると、ライトエラーを示すエラーメッセージをアクセスノード６００に送信する。

【０１８４】

［ステップＳ１０５］エラー応答を受信したアクセスノード６００のアクセス要求部６３０は、メタデータ照会部６１０にメタデータの照会を依頼する。メタデータ照会部６１０は、アクセス要求のアクセス対象であったセグメントを指定して、該当セグメントのメタデータ照会要求を制御ノード５００に送信する。

【０１８５】

［ステップＳ１０６］制御ノード５００のスライス割当処理部５３２は、メタデータ照会要求に応答し、スライス割り当て処理を行う。ステップＳ１０６およびその後の処理は、図１７のステップＳ８４～Ｓ８６と同様である。

【０１８６】

このようにして、ライトエラーが発生したディスクノードへのミラーライトがエラーとなり、アクセスノード６００からメタデータ照会要求が送信される。制御ノード５００では、メタデータ照会要求に応答して、アクセス対象のセグメントに割り当てられたスライスが調査され、ライト要求のアクセス先であったスライスの割り当てが解除される。

【０１８７】

ライトエラーが発生したディスクノード１００への他のディスクノードからのアクセスとしては、スライスコピーがある。例えばリカバリ処理によりディスクノード１００が管理するスライスをリザーブスライスとしていた場合、プライマリスライスのデータがリザーブスライスにコピーされる。このようなスライス全体のデータのコピー処理が、スライスコピーである。ディスクノード１００をコピー先としてスライスコピーを実行している最中にディスクノード１００でライトエラーが発生する場合もある。そのような場合、スライスコピーがエラーとなる。

【０１８８】

図２０は、ライトエラーが発生したディスクノードへのスライスコピー処理時の処理手順を示すシーケンス図である。以下、図２０の処理をステップ番号に沿って説明する。

［ステップＳ１１１］ディスクノード３００のアクセス処理部３２０は、制御ノード５００からのスライスコピー要求に応答して、自己の管理するプライマリスライスのデータをリザーブスライスにコピーする。図２０の例では、ディスクノード１００が管理するスライスがリザーブスライスである。

【０１８９】

［ステップＳ１１２］ディスクノード１００のアクセス処理部１２０は、ディスクノード３００からスライスコピー対象のデータを取得する。このときアクセス処理部１２０は、ライトエラーの有無を判断する。ライトエラーが発生している場合、アクセス処理部１２０は、ディスクノード３００に対してエラーを応答する。なお、スライスコピー対象のデータ受信時のディスクノード１００の処理の詳細は後述する（図２１参照）。

【０１９０】

［ステップＳ１１３］ディスクノード３００のアクセス処理部３２０は、ディスクノード１００からのエラー応答を受け取ると、制御ノード５００に対してエラーメッセージを送信する。エラーメッセージには、スライスコピーでエラーとなったスライスを管理しているディスクＩＤや、スライスコピーでエラーとなったスライスのスライスＩＤが含まれる。

【０１９１】

［ステップＳ１１４］制御ノード５００のスライス割当処理部５３２は、スライスコピーでエラーとなったスライスが割り当てられていたセグメントのスライス割り当て処理を行う。この処理の詳細は後述する（図２３参照）。

【０１９２】

このようにして、ライトエラーが発生したディスクノード１００へのスライスコピーが行われると、スライスコピーがエラーとなり、制御ノード５００でスライス割り当て処理が開始される。

【０１９３】

図２１は、スライスコピーのデータを受信したディスクノードの処理手順の一例を示すフローチャートである。以下、図２１に示す処理をステップ番号に沿って説明する。

［ステップＳ１２１］ディスクノード１００のアクセス処理部１２０は、スライスコピー対象のデータを受信する。

【０１９４】

［ステップＳ１２２］アクセス処理部１２０は、ライトエラーの記録の有無を判断する。例えばアクセス処理部１２０は、エラー記憶部１４０を参照し、ライトエラーの有無を判断する。ライトエラーがある場合、処理がステップＳ１２３に進められる。ライトエラーがない場合、処理がステップＳ１２４に進められる。

【０１９５】

［ステップＳ１２３］アクセス処理部１２０は、スライスコピーによるデータの送信元であるディスクノードに、エラーメッセージを送信する。その後、処理が終了する。

［ステップＳ１２４］アクセス処理部１２０は、受信したデータをストレージ装置１１０内のスライスコピー先のスライスに書き込む。その後、処理が終了する。

【０１９６】

次に、スライス割当処理について詳細に説明する。スライス割当処理は、メタデータ照会要求に応じて実行される場合と、スライスコピー時のエラーメッセージに応じて実行される場合とがある。

【０１９７】

図２２は、メタデータ照会要求時のスライス割当処理手順の一例を示すフローチャートである。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Ｓ１３１］制御ノード５００のスライス割当処理部５３２は、アクセスノード６００からのメタデータ照会要求を受信する。

【０１９８】

［ステップＳ１３２］スライス割当処理部５３２は、メタデータ照会要求で指定された

セグメントを変更対象セグメントとして、変更対象セグメントに割り当てられているスライスを検査する。例えばスライス割当処理部５３２は、論理ディスクメタデータ記憶部５４０を参照し、変更対象セグメントに割り当てられているスライスのメタデータを抽出する。そして、スライス割当処理部５３２は、抽出したメタデータの状態の欄を参照し、スライスの種別を判別する。プライマリスライスとセカンダリスライスとが割り当てられている場合、処理がステップＳ１３３に進められる。プライマリスライスとリザーブスライスとが割り当てられている場合、処理がステップＳ１３４に進められる。シングルプライマリスライスのみが割り当てられている場合、処理がステップＳ１３５に進められる。

【０１９９】

［ステップＳ１３３］スライス割当処理部５３２は、プライマリスライスとセカンダリスライスとの割り当て変更処理を実行する。この処理の詳細は後述する（図２４参照）。

［ステップＳ１３４］スライス割当処理部５３２は、プライマリスライスとリザーブスライスとの割り当て変更処理を実行する。この処理の詳細は後述する（図２６参照）。

【０２００】

［ステップＳ１３５］スライス割当処理部５３２は、シングルプライマリスライスの割り当て変更処理を実行する。この処理の詳細は後述する（図２８参照）。

図２３は、スライスコピーエラー時のスライス割り当て処理手順の一例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。

【０２０１】

［ステップＳ１４１］制御ノード５００のスライス割当処理部５３２は、ディスクノードからスライスコピーエラーを受信する。

［ステップＳ１４２］スライス割当処理部５３２は、スライスコピー処理が行われたセグメントを変更対象セグメントとして、プライマリスライスとリザーブスライスとの割り当て変更処理を実行する。この処理の詳細は後述する（図２６参照）。

【０２０２】

次に、セグメントに割り当てられているスライスの種別に応じた割り当て変更処理の手順について詳細に説明する。

図２４は、プライマリスライスとセカンダリスライスとが割り当てられているセグメントの割り当て変更処理手順の一例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。

【０２０３】

［ステップＳ１５１］スライス割当処理部５３２は、変更対象セグメントに割り当てられている各スライスを管理しているディスクノードでのライトエラーの発生の有無を検査する。例えばスライス割当処理部５３２は、ステップＳ１３２で抽出したメタデータのディスクノードＩＤの欄を参照し、変更対象セグメントに割り当てられているスライスを管理しているディスクノードを判断する。次に、スライス割当処理部５３２は、エラー記憶部５２０を参照し、変更対象セグメントに割り当てられているスライスを管理しているディスクノードにライトエラーが発生している否かを判断する。プライマリスライスとセカンダリスライスとのそれぞれを管理しているディスクノード共にライトエラーが発生していれば、処理がステップＳ１５２に進められる。プライマリスライスを管理しているディスクノードのみにライトエラーが発生していれば、処理がステップＳ１５５に進められる。セカンダリスライスを管理しているディスクノードのみにライトエラーが発生していれば、処理がステップＳ１５６に進められる。

【０２０４】

［ステップＳ１５２］スライス割当処理部５３２は、変更対象セグメントに割り当てられているプライマリスライスとセカンダリスライスとに関するメタデータを変更する。具体的には、スライス割当処理部５３２は、プライマリスライスとセカンダリスライスとのいずれか一方をフリースライスに変更すると共に、リザーブスライスの新規割り当てを行う。なおライトエラーが発生したディスクノードが管理するスライスは、セグメントへの割り当ては解除される。より詳細には、スライス割当処理部５３２は、セカンダリスライ

スを管理しているディスクノードへ、フリースライスへの変更を指示するメタデータ変更要求を送信する。またスライス割当処理部５３２は、フリースライスを管理しているディスクノードへ、リザーブスライスへの変更を指示するメタデータ変更要求を送信する。スライス割当処理部５３２は、各ディスクノードからメタデータ変更完了の応答を受け取ると、論理ディスクメタデータ記憶部５４０内のメタデータを、ディスクノードにおける変更内容と同様に変更する。なお第２の実施の形態では、セカンダリスライスをフリースライスに変更するものとする。また新規に割り当てられるリザーブスライスを選択する際には、スライス割当処理部５３２は、エラー記憶部５２０を参照し、ライトエラーが発生しているディスクノードを判断する。そしてスライス割当処理部５３２は、エラーが発生していないディスクノードが管理するフリースライスの中から、リザーブスライスとするスライスを選択する。

【０２０５】

［ステップＳ１５３］スライス割当処理部５３２は、プライマリスライスからリザーブスライスへのスライスのデータのコピーを、プライマリスライスを管理しているディスクノードに対して指示する。

【０２０６】

［ステップＳ１５４］スライス割当処理部５３２は、変更対象セグメントに割り当てられているプライマリスライスとリザーブスライスとに関するメタデータを変更する。具体的には、スライス割当処理部５３２は、プライマリスライスをフリースライスに変更し、リザーブスライスをシングルプライマリスライスに変更する。より詳細には、スライス割当処理部５３２は、プライマリスライスを管理しているディスクノードへ、プライマリスライスからフリースライスへの変更を指示するメタデータ変更要求を送信する。またスライス割当処理部５３２は、リザーブスライスを管理しているディスクノードへ、リザーブスライスからシングルプライマリスライスへの変更を指示するメタデータ変更要求を送信する。スライス割当処理部５３２は、各ディスクノードからメタデータ変更完了の応答を受け取ると、論理ディスクメタデータ記憶部５４０内のメタデータを、ディスクノードにおける変更内容と同様に変更する。その後、処理が終了する。

【０２０７】

［ステップＳ１５５］スライス割当処理部５３２は、変更対象セグメントに割り当てられているプライマリスライスとセカンダリスライスとに関するメタデータを変更する。具体的には、スライス割当処理部５３２は、プライマリスライスをフリースライスに変更すると共に、セカンダリスライスをシングルプライマリスライスに変更する。より詳細には、スライス割当処理部５３２は、プライマリスライスを管理しているディスクノードへ、プライマリスライスからフリースライスへの変更を指示するメタデータ変更要求を送信する。またスライス割当処理部５３２は、セカンダリスライスを管理しているディスクノードへ、シングルプライマリスライスへの変更を指示するメタデータ変更要求を送信する。スライス割当処理部５３２は、各ディスクノードからメタデータ変更完了の応答を受け取ると、論理ディスクメタデータ記憶部５４０内のメタデータを、ディスクノードにおける変更内容と同様に変更する。その後、処理が終了する。

【０２０８】

［ステップＳ１５６］スライス割当処理部５３２は、変更対象セグメントに割り当てられているプライマリスライスとセカンダリスライスとに関するメタデータを変更する。具体的には、スライス割当処理部５３２は、プライマリスライスをシングルプライマリスライスに変更すると共に、セカンダリスライスをフリースライスに変更する。より詳細には、スライス割当処理部５３２は、プライマリスライスを管理しているディスクノードへ、シングルプライマリスライスへの変更を指示するメタデータ変更要求を送信する。またスライス割当処理部５３２は、セカンダリスライスを管理しているディスクノードへ、セカンダリスライスからフリースライスへの変更を指示するメタデータ変更要求を送信する。スライス割当処理部５３２は、各ディスクノードからメタデータ変更完了の応答を受け取ると、論理ディスクメタデータ記憶部５４０内のメタデータを、ディスクノードにおける

変更内容と同様に変更する。その後、処理が終了する。

【０２０９】

図２５は、プライマリスライスとセカンダリスライスとが割り当てられているセグメントの割り当て変更処理例を示す図である。図２５の例では、変更対象セグメントのプライマリスライスをディスクノード１００（ＤＰ１）が管理しており、変更対象セグメントのセカンダリスライスをディスクノード２００（ＤＰ２）が管理しているものとする。またディスクノード３００（ＤＰ３）には、ライトエラーが発生していないものとする。このような状況において、ディスクノード１００が管理するプライマリスライス、ディスクノード２００が管理するセカンダリスライス、およびディスクノード３００が管理するフリースライスに着目する。

【０２１０】

図２５では、いずれのディスクノードにおいてもライトエラーが検出されていない状態を初期状態としている。変更対象セグメントに割り当てられているスライスを管理しているディスクノード１００，２００のライトエラーの有無を検査した結果に応じて、メタデータの変更内容が異なる。

【０２１１】

両方のディスクノード１００，２００でライトエラーが発生している場合、各ディスクノード１００，２００，３００の処理は以下の通りである。

ディスクノード１００が管理しているプライマリスライス（状態フラグ「Ｐ」）は、最初のメタデータ変更では変更されず、スライスコピーのコピー元とされる。スライスコピーが完了後、ディスクノード１００が管理しているプライマリスライスは、フリースライス（状態フラグ「Ｆ」）に変更される。ディスクノード２００が管理しているセカンダリスライス（状態フラグ「Ｓ」）は、最初のメタデータ変更でフリースライス（状態フラグ「Ｆ」）に変更される。ディスクノード３００が管理しているフリースライス（状態フラグ「Ｆ」）は、最初のメタデータ変更でリザーブスライス（状態フラグ「Ｒ」）に変更され、その後のスライスコピーのコピー先とされる。スライスコピーが完了後、ディスクノード３００が管理しているリザーブスライスは、シングルプライマリスライス（状態フラグ「ＳＰ」）に変更される。

【０２１２】

ディスクノード１００のみでライトエラーが発生している場合、各ディスクノード１００，２００，３００の処理は以下の通りである。

ディスクノード１００が管理しているプライマリスライス（状態フラグ「Ｐ」）は、フリースライス（状態フラグ「Ｆ」）に変更される。ディスクノード２００が管理しているセカンダリスライス（状態フラグ「Ｓ」）は、シングルプライマリスライス（状態フラグ「ＳＰ」）に変更される。ディスクノード３００が管理しているフリースライス（状態フラグ「Ｆ」）は変化しない。

【０２１３】

ディスクノード２００のみでライトエラーが発生している場合、各ディスクノード１００，２００，３００の処理は以下の通りである。

ディスクノード１００が管理しているプライマリスライス（状態フラグ「Ｐ」）は、シングルプライマリスライス（状態フラグ「ＳＰ」）に変更される。ディスクノード２００が管理しているセカンダリスライス（状態フラグ「Ｓ」）は、フリースライス（状態フラグ「Ｆ」）に変更される。ディスクノード３００が管理しているフリースライス（状態フラグ「Ｆ」）は変化しない。

【０２１４】

図２６は、プライマリスライスとリザーブスライスとが割り当てられているセグメントの割り当て変更処理手順の一例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。

【０２１５】

［ステップＳ１６１］スライス割当処理部５３２は、変更対象セグメントに割り当てら

れている各スライスを管理しているディスクノードでのライトエラーの発生の有無を検査する。プライマリスライスとリザーブスライスとのそれぞれを管理しているディスクノード共にライトエラーが発生している場合、およびプライマリスライスを管理しているディスクノードのみにライトエラーが発生している場合、処理がステップＳ１６２に進められる。リザーブスライスを管理しているディスクノードのみにライトエラーが発生していれば、処理がステップＳ１６５に進められる。

【０２１６】

［ステップＳ１６２］スライス割当処理部５３２は、変更対象セグメントに割り当てられているプライマリスライスとリザーブスライスとに関するメタデータを変更する。具体的には、スライス割当処理部５３２は、リザーブスライスをフリースライスに変更し、かつリザーブスライスの新規割り当てを行う。より詳細には、スライス割当処理部５３２は、リザーブスライスを管理しているディスクノードへ、リザーブスライスからフリースライスへの変更を指示するメタデータ変更要求を送信する。またスライス割当処理部５３２は、フリースライスを管理しているディスクノードへ、フリースライスからリザーブスライスへの変更を指示するメタデータ変更要求を送信する。スライス割当処理部５３２は、各ディスクノードからメタデータ変更完了の応答を受け取ると、論理ディスクメタデータ記憶部５４０内のメタデータを、ディスクノードにおける変更内容と同様に変更する。

【０２１７】

［ステップＳ１６３］スライス割当処理部５３２は、プライマリスライスからリザーブスライスへのスライスのデータのコピーを、プライマリスライスを管理しているディスクノードに対して指示する。

【０２１８】

［ステップＳ１６４］スライス割当処理部５３２は、変更対象セグメントに割り当てられているプライマリスライスとリザーブスライスとに関するメタデータを変更する。具体的には、スライス割当処理部５３２は、プライマリスライスをフリースライスに変更し、リザーブスライスをシングルプライマリスライスに変更する。より詳細には、スライス割当処理部５３２は、プライマリスライスを管理しているディスクノードへ、プライマリスライスからフリースライスへの変更を指示するメタデータ変更要求を送信する。またスライス割当処理部５３２は、リザーブスライスを管理しているディスクノードへ、シングルプライマリスライスへの変更を指示するメタデータ変更要求を送信する。スライス割当処理部５３２は、各ディスクノードからメタデータ変更完了の応答を受け取ると、論理ディスクメタデータ記憶部５４０内のメタデータを、ディスクノードにおける変更内容と同様に変更する。その後、処理が終了する。

【０２１９】

［ステップＳ１６５］スライス割当処理部５３２は、変更対象セグメントに割り当てられているプライマリスライスとリザーブスライスとに関するメタデータを変更する。具体的には、スライス割当処理部５３２は、プライマリスライスをシングルプライマリスライスに変更すると共に、リザーブスライスをフリースライスに変更する。より詳細には、スライス割当処理部５３２は、プライマリスライスを管理しているディスクノードへ、シングルプライマリスライスへの変更を指示するメタデータ変更要求を送信する。またスライス割当処理部５３２は、リザーブスライスを管理しているディスクノードへ、リザーブスライスからフリースライスへの変更を指示するメタデータ変更要求を送信する。スライス割当処理部５３２は、各ディスクノードからメタデータ変更完了の応答を受け取ると、論理ディスクメタデータ記憶部５４０内のメタデータを、ディスクノードにおける変更内容と同様に変更する。その後、処理が終了する。

【０２２０】

図２７は、プライマリスライスとリザーブスライスとが割り当てられているセグメントの割り当て変更処理例を示す図である。図２７の例では、変更対象セグメントのプライマリスライスをディスクノード１００（ＤＰ１）が管理しており、変更対象セグメントのリザーブスライスをディスクノード２００（ＤＰ２）が管理しているものとする。またディ

スクノード３００（ＤＰ３）には、ライトエラーが発生していないものとする。このような状況において、ディスクノード１００が管理するプライマリスライス、ディスクノード２００が管理するリザーブスライス、およびディスクノード３００が管理するフリースライスに着目する。

【０２２１】

図２７では、いずれのディスクノードにおいてもライトエラーが検出されていない状態を初期状態としている。変更対象セグメントに割り当てられているスライスを管理しているディスクノード１００，２００のライトエラーの有無を検査した結果に応じて、メタデータの変更内容が異なる。

【０２２２】

両方のディスクノード１００，２００でライトエラーが発生している場合、各ディスクノード１００，２００，３００の処理は以下の通りである。

ディスクノード１００が管理しているプライマリスライスは変更されず、スライスコピーのコピー元とされる。スライスコピーが完了後、ディスクノード１００が管理しているプライマリスライスは、フリースライス（状態フラグ「Ｆ」）に変更される。ディスクノード２００が管理しているリザーブスライス（状態フラグ「Ｒ」）は、最初のメタデータ変更でフリースライス（状態フラグ「Ｆ」）に変更される。ディスクノード３００が管理しているフリースライス（状態フラグ「Ｆ」）は、最初のメタデータ変更でリザーブスライス（状態フラグ「Ｒ」）に変更され、その後のスライスコピーのコピー先とされる。スライスコピーが完了後、ディスクノード３００が管理しているリザーブスライスは、シングルプライマリスライス（状態フラグ「ＳＰ」）に変更される。

【０２２３】

ディスクノード１００のみでライトエラーが発生している場合の処理は、両方のディスクノード１００，２００でライトエラーが発生している場合と同様である。なお、この場合、ディスクノード２００にはライトエラーは発生していないため、ディスクノード２００のリザーブスライスを、そのままスライスコピー先とすることもできる。

【０２２４】

ディスクノード２００のみでライトエラーが発生している場合、各ディスクノード１００，２００，３００の処理は以下の通りである。

ディスクノード１００が管理しているプライマリスライス（状態フラグ「Ｐ」）は、シングルプライマリスライス（状態フラグ「ＳＰ」）に変更される。ディスクノード２００が管理しているリザーブスライス（状態フラグ「Ｒ」）は、フリースライス（状態フラグ「Ｆ」）に変更される。ディスクノード３００が管理しているフリースライス（状態フラグ「Ｆ」）は変化しない。

【０２２５】

図２８は、シングルプライマリスライスが割り当てられているセグメントの割り当て変更処理手順の一例を示すフローチャートである。以下、図２８に示す処理をステップ番号に沿って説明する。

【０２２６】

［ステップＳ１７１］スライス割当処理部５３２は、変更対象セグメントに割り当てられているプライマリスライスに関するメタデータを変更する。具体的には、スライス割当処理部５３２は、シングルプライマリスライスをプライマリスライスに変更し、リザーブスライスの新規割り当てを行う。より詳細には、スライス割当処理部５３２は、シングルプライマリスライスを管理しているディスクノードへ、プライマリスライスへの変更を指示するメタデータ変更要求を送信する。またスライス割当処理部５３２は、フリースライスを管理しているディスクノードへ、リザーブスライスへの変更を指示するメタデータ変更要求を送信する。スライス割当処理部５３２は、各ディスクノードからメタデータ変更完了の応答を受け取ると、論理ディスクメタデータ記憶部５４０内のメタデータを、ディスクノードにおける変更内容と同様に変更する。

【０２２７】

［ステップＳ１７２］スライス割当処理部５３２は、プライマリスライスからリザーブスライスへのスライスのデータのコピーを、プライマリスライスを管理しているディスクノードに対して指示する。

【０２２８】

［ステップＳ１７３］スライス割当処理部５３２は、変更対象セグメントに割り当てられているプライマリスライスとリザーブスライスとに関するメタデータを変更する。具体的には、スライス割当処理部５３２は、プライマリスライスをフリースライスに変更し、リザーブスライスをシングルプライマリスライスに変更する。より詳細には、スライス割当処理部５３２は、プライマリスライスを管理しているディスクノードへ、フリースライスへの変更を指示するメタデータ変更要求を送信する。またスライス割当処理部５３２は、リザーブスライスを管理しているディスクノードへ、シングルプライマリスライスへの変更を指示するメタデータ変更要求を送信する。スライス割当処理部５３２は、各ディスクノードからメタデータ変更完了の応答を受け取ると、論理ディスクメタデータ記憶部５４０内のメタデータを、ディスクノードにおける変更内容と同様に変更する。その後、処理が終了する。

【０２２９】

図２９は、シングルプライマリスライスが割り当てられているセグメントの割り当て変更処理例を示す図である。図２９の例では、変更対象セグメントのシングルプライマリスライスをディスクノード１００（ＤＰ１）が管理しているものとする。またディスクノード２００（ＤＰ２）には、ライトエラーが発生していないものとする。このような状況において、ディスクノード１００が管理するシングルプライマリスライス、ディスクノード２００が管理するフリースライスに着目する。

【０２３０】

図２９では、いずれのディスクノードにおいてもライトエラーが検出されていない状態を初期状態としている。シングルプライマリスライスのみが割り当てられているセグメントに対してスライスの割り当て処理が実行されるのは、シングルプライマリスライスを管理しているディスクノード１００でライトエラーが発生した場合である。その場合、各ディスクノード１００，２００において以下の処理が行われる。

【０２３１】

ディスクノード１００が管理しているシングルプライマリスライス（状態フラグ「ＳＰ」）は、最初のメタデータ変更ではプライマリスライスに変更され、その後、スライスコピーのコピー元とされる。スライスコピーが完了後、ディスクノード１００が管理しているプライマリスライスは、フリースライス（状態フラグ「Ｆ」）に変更される。ディスクノード２００が管理しているフリースライス（状態フラグ「Ｆ」）は、最初のメタデータ変更でリザーブスライス（状態フラグ「Ｒ」）に変更され、その後のスライスコピーのコピー先とされる。スライスコピーが完了後、ディスクノード２００が管理しているリザーブスライスは、シングルプライマリスライス（状態フラグ「ＳＰ」）に変更される。

【０２３２】

以上のようなメタデータ割当変更処理により、変更対象セグメントには、ライトエラーが発生していないディスクノードが管理するスライスが、シングルプライマリスライスとして割り当てられる。

【０２３３】

次に、スライス割当処理のうち、スライスコピーを伴う処理における制御ノード５００とディスクノードとの連係動作について説明する。スライスコピーを伴う処理は、ライトエラーが発生したディスクノードが管理するスライスのセグメントへの割り当てを解除すると、セグメントのデータが失われてしまう場合に行われる。

【０２３４】

以下の例では、代表的に、処理対象セグメントに割り当てられているプライマリスライスとセカンダリスライスとのそれぞれを管理するディスクノードの両方でライトエラーが発生している場合の処理について説明する。

【０２３５】
図３０は、スライスコピーを伴うスライス割当処理手順の一例を示すシーケンス図である。この処理は、例えば図２４のステップＳ１５１からステップＳ１５２に処理が遷移した場合に実行される。以下、図３０に示す処理をステップ番号に沿って説明する。
【０２３６】
［ステップＳ１８１］スライスコピーを伴うスライス割当処理は、リザーブスライスの割り当て処理（ステップＳ１８１ａ）、スライスコピー処理（ステップＳ１８１ｂ）、およびシングルプライマリ化処理（ステップＳ１８１ｃ）に分かれる。
【０２３７】
［ステップＳ１８１ａ］スライス割当処理部５３２は、各ディスクノード１００，２００，３００に対してメタデータ変更要求を送信する。例えば、ディスクノード１００へは、変更対象セグメントのプライマリスライスのペアとなるセカンダリスライスの情報の削除を指示するメタデータ変更要求が送信される。またディスクノード２００へは、変更対象セグメントのセカンダリスライスのフリースライスへの変更を指示するメタデータ変更要求が送信される。ディスクノード３００へは、フリースライスを変更対象セグメントのリザーブスライスに変更することを指示するメタデータ変更要求が送信される。
【０２３８】
［ステップＳ１８２］ディスクノード１００のメタデータ管理部１６０は、メタデータ変更要求に応じてメタデータ記憶部１５０とストレージ装置１１０とのメタデータを更新し、メタデータ変更完了の応答を制御ノード５００に送信する。
【０２３９】
［ステップＳ１８３］ディスクノード２００のメタデータ管理部２６０は、メタデータ変更要求に応じてメタデータ記憶部とストレージ装置２１０とのメタデータを更新し、メタデータ変更完了の応答を制御ノード５００に送信する。
【０２４０】
［ステップＳ１８４］ディスクノード３００のメタデータ管理部３６０は、メタデータ変更要求に応じてメタデータ記憶部とストレージ装置３１０とのメタデータを更新し、メタデータ更新完了の応答を制御ノード５００に送信する。
【０２４１】
スライス割当処理部５３２は、ステップＳ１８２～Ｓ１８３のメタデータ更新が完了すると、論理ディスクメタデータ記憶部５４０内のメタデータを、ディスクノード１００，２００，３００で更新されたメタデータと同じ内容に更新する。
【０２４２】
［ステップＳ１８１ｂ］スライス割当処理部５３２は、ディスクノード１００，２００，３００からメタデータ更新完了の応答を受け取ると、プライマリスライスを管理しているディスクノード１００に対して、スライスコピー要求を送信する。スライスコピー要求には、変更対象セグメントのプライマリスライスのスライスＩＤ、リザーブ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、およびリザーブスライスのスライスＩＤが含まれる。
【０２４３】
［ステップＳ１８５］スライスコピー要求を受信したディスクノード１００のアクセス処理部１２０は、指定されたスライスのペアのスライスを管理するディスクノード３００に対して、指定されたスライス内のデータを送信する。データを送信する際には、データの格納先として、スライスコピー要求で指定されたスライスのペアのスライスのスライスＩＤが指定される。
【０２４４】
［ステップＳ１８６］ディスクノード３００のアクセス処理部３２０は、ディスクノード１００から送られたデータを受信し、指定されたスライスにデータを格納する。アクセス処理部３２０は、データの格納が完了すると、書き込み完了応答をディスクノード１００に送信する。

【０２４５】
ディスクノード１００のアクセス処理部１２０は、ディスクノード３００からの書き込み完了応答を受け取ると、制御ノード５００に対してスライスコピー完了の応答を送信する。
【０２４６】
［ステップＳ１８１ｃ］制御ノード５００のスライス割当処理部５３２は、ディスクノード１００からスライスコピー完了の応答を受け取ると、ディスクノード１００に対してプライマリスライスのフリースライス化を指示するメタデータ変更要求を送信する。フリースライス化を指示するメタデータ変更要求には、例えば、変更対象セグメントに割り当てられているプライマリスライスのスライスＩＤが含まれる。
【０２４７】
また、スライス割当処理部５３２は、ディスクノード１００からスライスコピー完了の応答を受け取ると、ディスクノード３００に対してリザーブスライスのシングルプライマリ化を指示するメタデータ変更要求を送信する。シングルプライマリ化を指示するメタデータ変更要求には、例えば変更対象セグメントに割り当てられているリザーブスライスのスライスＩＤが含まれる。
【０２４８】
［ステップＳ１８７］ディスクノード１００のメタデータ管理部１６０は、プライマリスライスのフリースライスへの変更を指示するメタデータ変更要求に応答して、メタデータ記憶部１５０とストレージ装置１１０とのメタデータを更新する。メタデータ管理部１６０は、メタデータの更新が完了すると、更新完了応答を制御ノード５００に送信する。
【０２４９】
［ステップＳ１８８］ディスクノード３００のメタデータ管理部３６０は、リザーブスライスのシングルプライマリスライスへの変更を指示するメタデータ変更要求に応答して、メタデータ記憶部とストレージ装置１１０とのメタデータを更新する。メタデータ管理部３６０は、メタデータの更新が完了すると、更新完了応答を制御ノード５００に送信する。
【０２５０】
なお、スライス割当処理部５３２は、ステップＳ１８７，Ｓ１８８のメタデータ更新が完了すると、論理ディスクメタデータ記憶部５４０内のメタデータを、ディスクノード１００，２００，３００で更新されたメタデータと同じ内容に更新する。
【０２５１】
このような手順によって、スライスコピーを伴うスライス割当処理が完了する。次に、スライスコピーを伴わないスライス割当処理における制御ノード５００とディスクノードとの連係動作について説明する。
【０２５２】
図３１は、スライスコピーを伴わないスライス割当処理手順の一例を示すシーケンス図である。この処理は、例えば図２４のステップＳ１５１からステップＳ１５３に処理が遷移した場合に実行される。以下、図３１に示す処理をステップ番号に沿って説明する。
【０２５３】
［ステップＳ１９１］スライス割当処理部５３２は、各ディスクノード１００，２００に対してメタデータ変更要求を送信する。例えば、ディスクノード１００へは、変更対象セグメントのプライマリスライスのフリースライスへの変更を指示するメタデータ変更要求が送信される。またディスクノード２００へは、変更対象セグメントのセカンダリスライスのシングルプライマリスライスへの変更を指示するメタデータ変更要求が送信される。
【０２５４】
［ステップＳ１９２］ディスクノード１００のメタデータ管理部１６０は、メタデータ変更要求に応じてメタデータ記憶部１５０とストレージ装置１１０とのメタデータを更新し、メタデータ変更完了の応答を制御ノード５００に送信する。

【０２５５】
［ステップＳ１９３］ディスクノード２００のメタデータ管理部２６０は、メタデータ変更要求に応じてメタデータ記憶部とストレージ装置２１０とのメタデータを更新し、メタデータ変更完了の応答を制御ノード５００に送信する。
【０２５６】
スライス割当処理部５３２は、ステップＳ１９１，Ｓ１９２のメタデータ更新が完了すると、論理ディスクメタデータ記憶部５４０内のメタデータを、ディスクノード１００，２００で更新されたメタデータと同じ内容に更新する。
【０２５７】
以上説明したように、リカバリ処理中にディスクノードでライトエラーが発生した場合、そのディスクノードでのライトエラーの発生が、ディスクノードと制御ノードとに記録される。そしてリカバリ処理中に、ライトエラーが発生したディスクノードが管理するスライスにアクセスがあった場合、ライトエラーが発生したディスクノードが管理するスライスのセグメントへの割り当て解除によりセグメントのデータが喪失するか否かが判断される。そして、割り当て解除によりセグメントのデータを喪失する場合には、スライスコピー後にアクセスされたスライスの割り当てが解除される。その結果、データの喪失が抑制される。
【０２５８】
しかも、ライトエラーが発生したディスクノードが管理するスライスのセグメントへの割り当て解除によりセグメントのデータが喪失しない場合であっても、ペアとなるスライスをシングルプライマリスライスにすることで、リカバリ処理の対象となる。従って、ライトエラーが発生したディスクノードが管理するスライスのうち、アクセスがあったスライスは優先的にリカバリ処理が実行されることとなる。
【０２５９】
またリカバリ処理中にライトエラーが発生したディスクノードで管理されるスライスにアクセスがあると、アクセスされたスライスのセグメントへの割り当てが解除される。割り当てが解除されたセグメントは、シングルプライマリのみが割り当てられた欠損セグメントに変更される。一方、図１６に示すように、リカバリ処理中はシングルプライマリスライスが順次選択され、二重化が復旧される。従って、欠損セグメントに変更されたセグメントは、リカバリ処理によって二重化状態が復旧される。その結果、マルチノードストレージシステムの信頼性が向上する。
【０２６０】
ところで第２の実施の形態では、ライトエラーが発生したディスクノードのメタデータ割当処理の対象となったスライスは、メタデータ割当処理後にフリースライス（状態フラグ「Ｆ」）に変更している。フリースライスであっても、リカバリ処理およびスライス割当処理のいずれにおいても、ライトエラーが発生したディスクノード内のスライスがデータのコピー先（リザーブスライス）として選択されることはない。また、メタデータ割当処理後には、アクセスノード６００に新たなメタデータが送信される（図１７参照）。これにより、ライトエラーが発生したディスクノード内のメタデータ割当処理の対象となったスライスが、アクセスノード６００からのアクセス対象となることもなくなる。その結果、ライトエラーが発生したディスクノード内のメタデータ割当処理の対象となったスライスは、マルチノードストレージシステム内でのアクセス対象から除外される。すなわち、メタデータ割当処理の対象となったスライスが切り離される。そして、アクセスノード６００は、コピー先のスライスに対してアクセスを行うことができる。
【０２６１】
なおライトエラーが発生したディスクノードのメタデータ割当処理の対象となったスライスは、メタデータ割当処理後に異常スライス（状態フラグ「Ｂ」）に変更してもよい。
また第２の実施の形態では、各ディスクノードに１台ずつのストレージ装置を接続しているため、ストレージ装置が故障するとディスクノードの切り離しが行われている。１台のディスクノードに複数のストレージ装置が接続されている場合には、ストレージ装置ご

とに個別に切り離しを行うことができる。

【０２６２】

さらにライトエラーが発生したディスクノードにおいて、アクセスノード６００からのデータリードのアクセス要求に対しては、要求に従ったストレージ装置へのアクセスを行い、リードしたデータをアクセスノード６００に応答するようにしてもよい。すなわち第２の実施の形態では、ライトエラーが発生したディスクノードは、制御ノード５００からのスライスコピー要求に対しては、ストレージ装置に対するスライス内のデータリードを行い、コピー先に転送する。その一方で、ライトエラーが発生したディスクノードは、アクセスノード６００からのアクセス要求に対しては、リード要求であってもエラー応答を送信している。これは、アクセスがあったセグメントを欠損セグメントに変更し、リカバリ処理による二重化復旧を早期に行うためである。しかし、ライトエラーが発生していてもデータのリードは可能な場合がある。そこで、アクセスノード６００への早期のデータ応答を優先する場合、ライトエラーが発生したディスクノードは、アクセス要求のうちリード要求に対してはストレージ装置へのリードアクセスを実行し、読み出したデータを応答してもよい。これにより、アクセスノード６００は、スライス割当処理の完了を待たずにアクセス対象のデータを取得可能となる。

【０２６３】

〔その他の応用例〕

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、データ割当制御装置１、制御ノード５００、ディスクノード１００，２００，３００，４００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。

【０２６４】

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。

【０２６５】

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。

【０２６６】

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。

【０２６７】

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。

【０２６８】
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） 接続されたストレージ装置内のデータを管理する複数のディスクノードに対する管理対象データの割り当て指示を、コンピュータに実行させるデータ割当制御プログラムにおいて、
前記コンピュータに、
複数のストレージ装置に格納されたデータのうち同一内容の冗長データが存在しない二重化欠損データのコピーを、二重化欠損データを管理するディスクノードに指示する二重化復旧処理を行い、
前記二重化復旧処理中に、ストレージ装置に対するデータのライトエラーを示すライトエラー情報を受け取ると、ライトエラーが発生したストレージ装置の識別情報をエラー記憶手段に格納し、
前記エラー記憶手段を参照してライトエラーが発生したストレージ装置を判断し、ライトエラーが発生したストレージ装置に格納されているデータをコピー対象データとし、ライトエラーが発生していないストレージ装置内に前記コピー対象データの冗長データが存在しない場合、前記コピー対象データを管理するディスクノードに対して、ライトエラーが発生していないストレージ装置への前記コピー対象データのコピーを指示する、
処理を実行させることを特徴とするデータ割当制御プログラム。
【０２６９】
（付記２） 前記二重化復旧処理では、二重化欠損データを順次選択し、選択した二重化欠損データを管理するディスクノードに対し、選択した二重化欠損データが格納されているストレージ装置と異なるストレージ装置であり、かつライトエラーが発生していないストレージ装置への、選択した二重化欠損データのコピーを指示することを特徴とする付記１記載のデータ割当制御プログラム。
【０２７０】
（付記３） 前記二重化復旧処理では、前記複数のストレージ装置内の二重化欠損データを調査し、二重化欠損データの格納場所を示す管理情報を管理情報記憶手段に格納し、前記管理情報記憶手段内の管理情報に示される二重化欠損データを順次選択して、選択した二重化欠損データを管理するディスクノードに対して選択した二重化欠損データのコピーを指示し、
前記コピー対象データのコピー指示の際には、前記コピー対象データのコピーが完了すると、コピー先のデータの管理情報を二重化欠損データとして前記管理情報記憶手段に格納する、
ことを特徴とする付記１記載のデータ割当制御プログラム。
【０２７１】
（付記４） 前記二重化復旧処理では、前記管理情報記憶手段内の管理情報で示されるすべての二重化欠損データのコピーが完了すると前記エラー記憶手段を参照してライトエラーが発生したストレージ装置の有無を判断し、ライトエラーが発生したストレージ装置がある場合、ライトエラーが発生したストレージ装置へのアクセスを停止することにより生じる二重化欠損データの管理情報を前記管理情報記憶手段に追加し、二重化復旧処理を続行する、
ことを特徴とする付記３記載のデータ割当制御プログラム。
【０２７２】
（付記５） 前記コピー対象データのコピー指示では、ライトエラーが発生しているストレージ装置内のデータに対するアクセスで発生したエラーを示すデータアクセスエラー情報を受け取ると、エラー発生時のアクセス対象となっていたデータをコピー対象データとし、前記コピー対象データのコピーが完了すると、コピー先のデータの管理情報を、新たなアクセス先として前記データアクセスエラー情報を送信した装置に応答することを特徴とする付記１記載のデータ割当制御プログラム。
【０２７３】

（付記６）　前記二重化復旧処理中以外の期間に、ストレージ装置に対するデータのライトエラーを示すライトエラー情報を受け取ると、ライトエラーが発生したストレージ装置へのアクセスを停止することにより生じる二重化欠損データの二重化復旧処理を開始することを特徴とする付記１記載のデータ割当制御プログラム。

【０２７４】

（付記７）　接続されたストレージ装置内のデータを管理する複数のディスクノードに対する管理対象データの割り当て指示をコンピュータで実行するデータ割当制御方法において、

前記コンピュータが、

複数のストレージ装置に格納されたデータのうち同一内容の冗長データが存在しない二重化欠損データのコピーを、二重化欠損データを管理するディスクノードに指示する二重化復旧処理を行い、

前記二重化復旧処理中に、ストレージ装置に対するデータのライトエラーを示すライトエラー情報を受け取ると、ライトエラーが発生したストレージ装置の識別情報をエラー記憶手段に格納し、

前記エラー記憶手段を参照してライトエラーが発生したストレージ装置を判断し、ライトエラーが発生したストレージ装置に格納されているデータをコピー対象データとし、ライトエラーが発生していないストレージ装置内に前記コピー対象データの冗長データが存在しない場合、前記コピー対象データを管理するディスクノードに対して、ライトエラーが発生していないストレージ装置への前記コピー対象データのコピーを指示する、

ことを特徴とするデータ割当制御方法。

【０２７５】

（付記８）　接続されたストレージ装置内のデータを管理する複数のディスクノードに対する管理対象データの割り当て指示を行うデータ割当制御装置において、

複数のストレージ装置に格納されたデータのうち同一内容の冗長データが存在しない二重化欠損データのコピーを、二重化欠損データを管理するディスクノードに指示する二重化復旧処理を行う二重化復旧処理手段と、

前記二重化復旧処理中に、ストレージ装置に対するデータのライトエラーを示すライトエラー情報を受け取ると、ライトエラーが発生したストレージ装置の識別情報をエラー記憶手段に格納するエラー情報受信手段と、

前記エラー記憶手段を参照してライトエラーが発生したストレージ装置を判断し、ライトエラーが発生したストレージ装置に格納されているデータをコピー対象データとし、ライトエラーが発生していないストレージ装置内に前記コピー対象データの冗長データが存在しない場合、前記コピー対象データを管理するディスクノードに対して、ライトエラーが発生していないストレージ装置への前記コピー対象データのコピーを指示するコピー指示手段と、

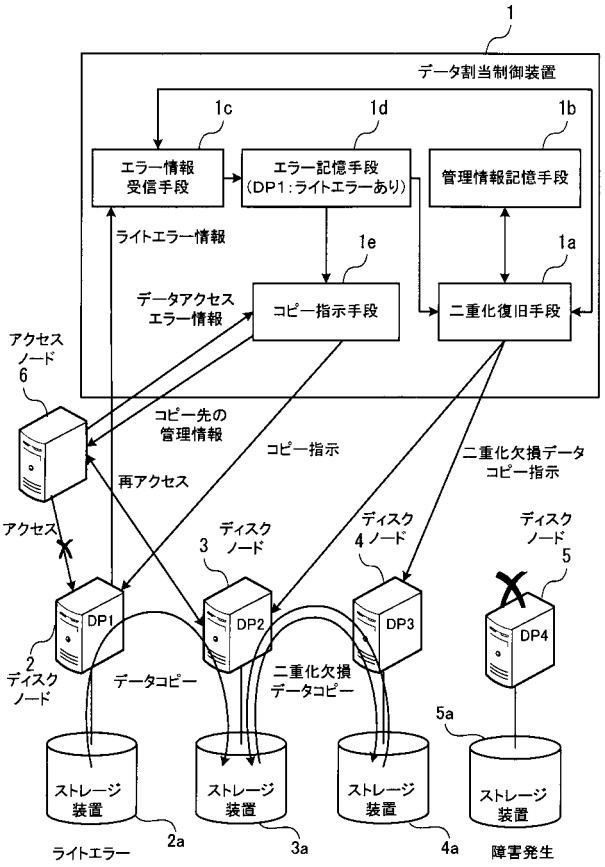
を有することを特徴とするデータ割当制御装置。

【符号の説明】

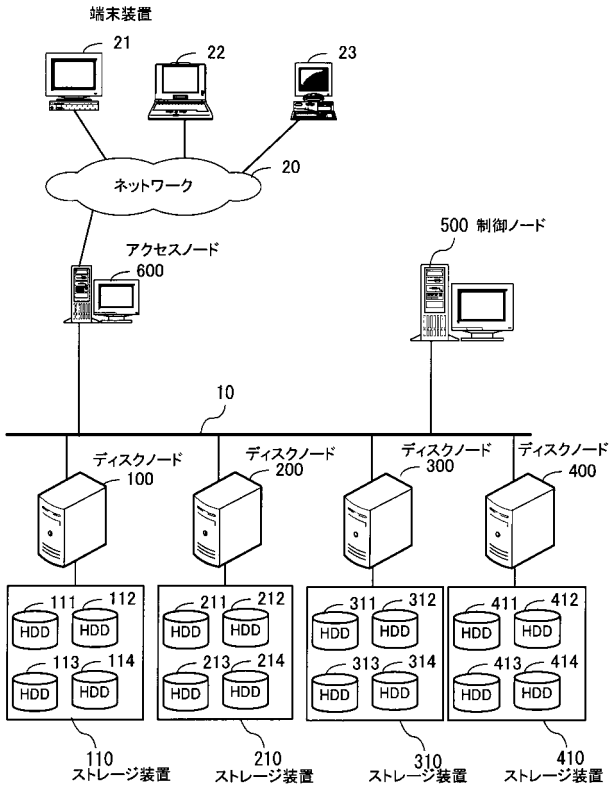
【０２７６】

１　データ割当制御装置
１ａ　二重化復旧手段
１ｂ　管理情報記憶手段
１ｃ　エラー情報受信手段
１ｄ　エラー記憶手段
１ｅ　コピー指示手段
２～５　ディスクノード
２ａ，３ａ，４ａ，５ａ　ストレージ装置
６　アクセスノード

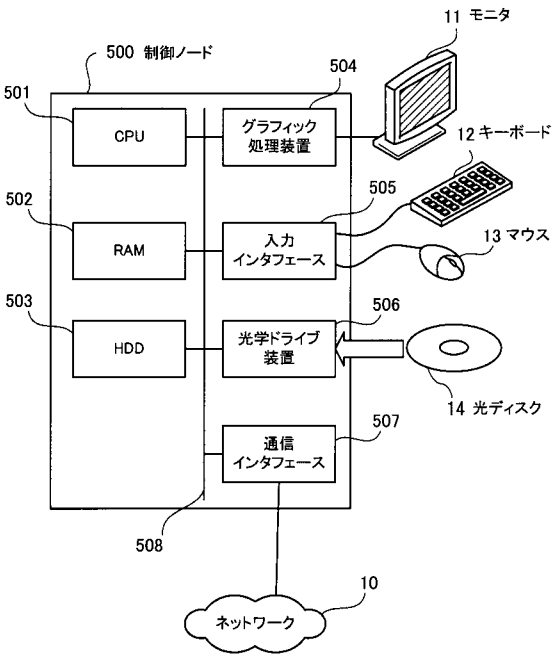
【図１】


1
データ割当制御装置
1c
1d
1b
エラー情報受信手段
エラー記憶手段（DP1：ライトエラーあり）
管理情報記憶手段
ライトエラー情報
1e
1a
データアクセスエラー情報
コピー指示手段
二重化復旧手段
アクセスノード 6
コピー先の管理情報
コピー指示
二重化欠損データコピー指示
再アクセス
アクセス
ディスクノード 3
ディスクノード 4
ディスクノード 5
DP1
DP2
DP3
DP4
2 ディスクノード
データコピー
二重化欠損データコピー
5a
ストレージ装置
2a
3a
4a
ライトエラー
障害発生

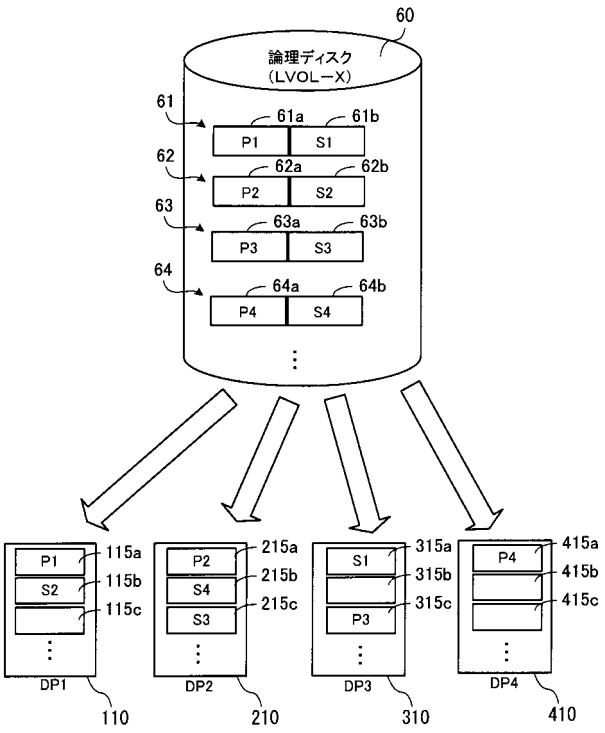
【図２】


端末装置
21
22
23
ネットワーク
20
500 制御ノード
アクセスノード
600
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
HDD
111 112 113 114
211 212 213 214
311 312 313 314
411 412 413 414
110 ストレージ装置
210 ストレージ装置
310 ストレージ装置
410 ストレージ装置

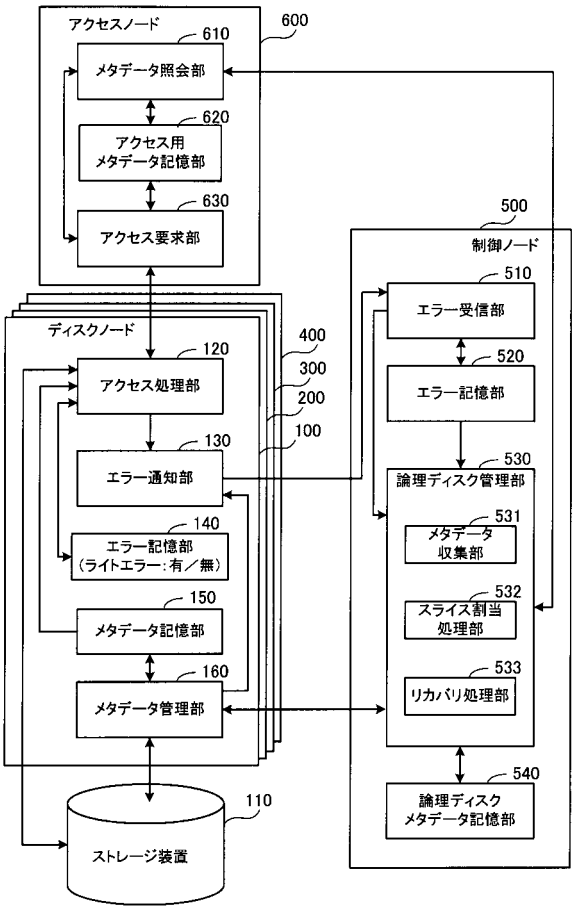
【図３】


11 モニタ
500 制御ノード
504
501
CPU
グラフィック処理装置
12 キーボード
505
502
RAM
入力インタフェース
13 マウス
503
HDD
506
光学ドライブ装置
14 光ディスク
507
通信インタフェース
508
10
ネットワーク

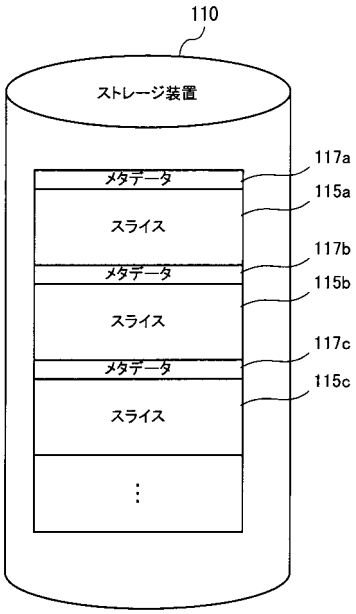
【図４】


60
論理ディスク（LVOL−X）
61
61a 61b
P1 S1
62
62a 62b
P2 S2
63
63a 63b
P3 S3
64
64a 64b
P4 S4
P1 S2
115a 115b 115c
DP1 110
P2 S4 S3
215a 215b 215c
DP2 210
S1 P3
315a 315b 315c
DP3 310
P4
415a 415b 415c
DP4 410

【図５】



【図６】



【図７】

150 メタデータ記憶部

151 メタデータテーブル

ディスクノードID	ディスクID	スライスID	状態	論理ディスクID	セグメントID	論理ディスクアドレス	ペアのディスクノードID	ペアのディスクID	ペアのスライスID
DP1	1	1	P	LVOL−X	1	ADDR1	DP3	1	1
		2	S	LVOL−X	2	ADDR2	DP2	1	
DP1	1	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…

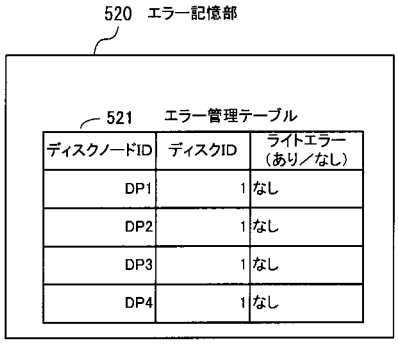
【図８】

540 論理ディスクメタデータ記憶部

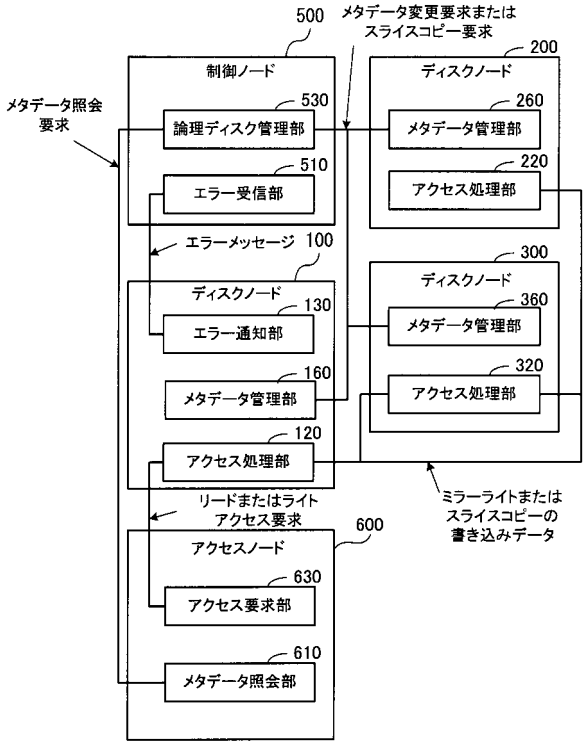
541 論理ディスクメタデータテーブル

ディスクノードID	ディスクID	スライスID	状態	論理ディスクID	セグメントID	論理ディスクアドレス	ペアのディスクノードID	ペアのディスクID	ペアのスライスID
DP1	1	1	P	LVOL−X	1	ADDR1	DP3	1	1
DP1	1	2	S	LVOL−X	2	ADDR2	DP2	1	
…	…	…	…	…	…	…	…	…	…
DP2	1	1	P	LVOL−X	2	ADDR2	DP1	1	2
DP2	1	2	S	LVOL−X	4	ADDR4	DP4	1	1
DP2	1	3	S	LVOL−X	3	ADDRn	DP3	1	3
…	…	…	…	…	…	…	…	…	…
DP3	1	1	S	LVOL−X	1	ADDR1	DP1	1	1
DP3	1	2	F	NULL	NULL	NULL	NULL	NULL	NULL
DP3	1	3	P	LVOL−X	3	ADDRn	DP2	1	3
…	…	…	…	…	…	…	…	…	…
DP4	1	1	P	LVOL−X	4	ADDR4	DP2	1	2
…	…	…	…	…	…	…	…	…	…

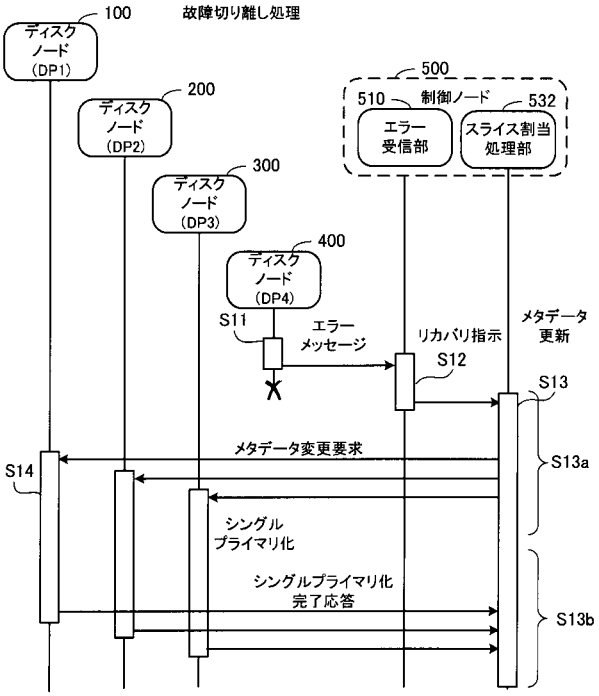
【図９】



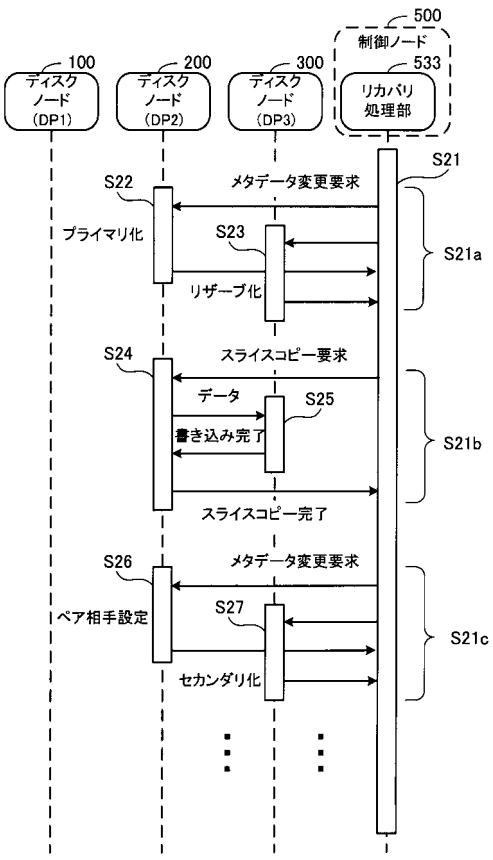
【図１０】



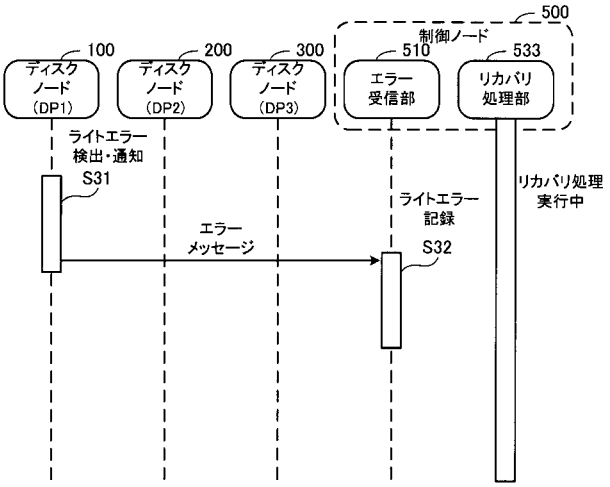
【図１１】



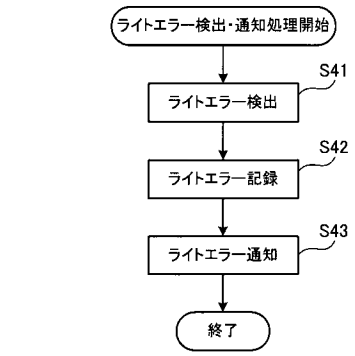
【図１２】



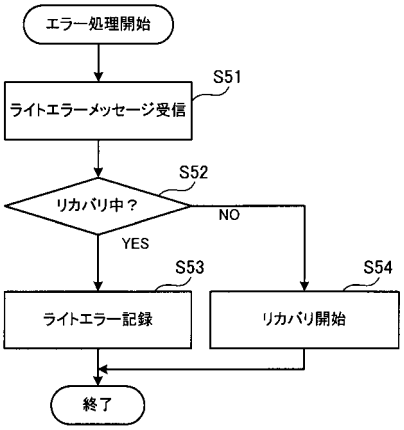
【図１３】



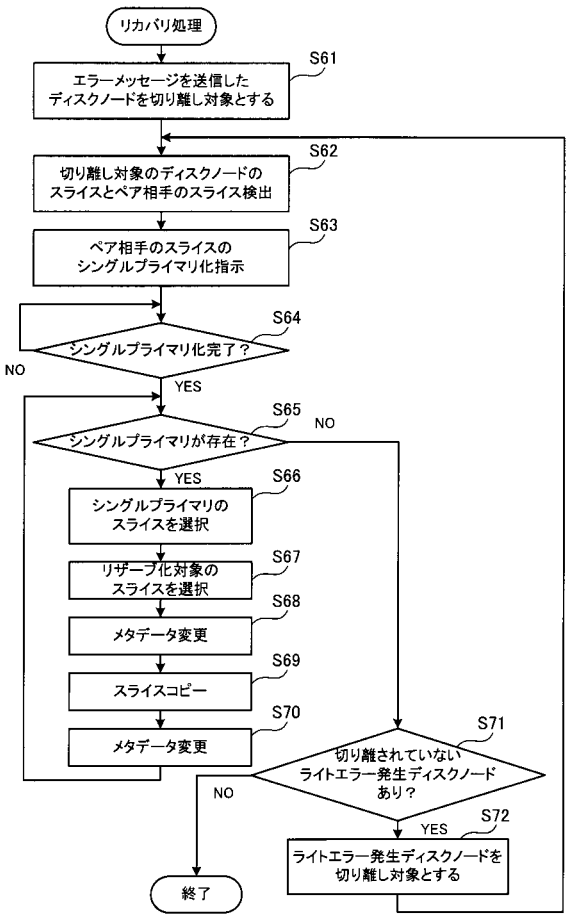
【図１４】



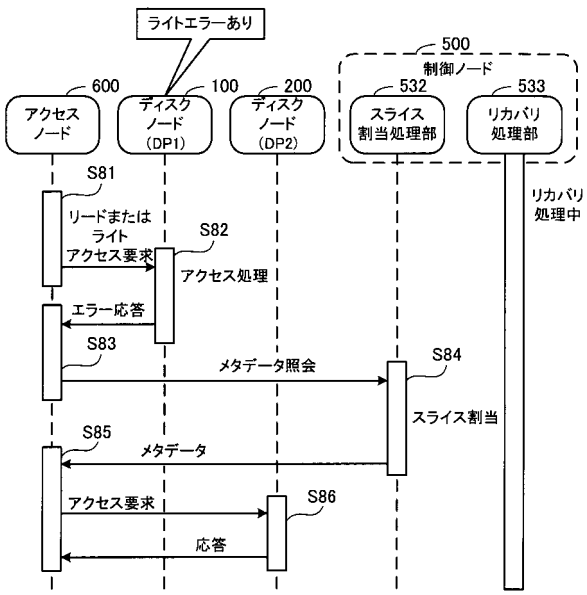
【図１５】



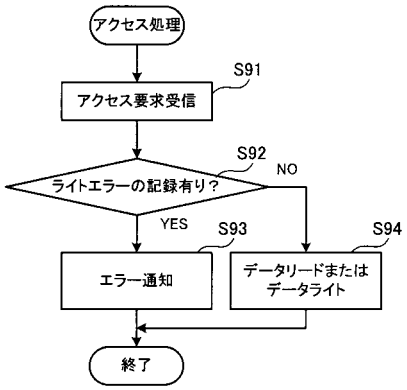
【図１６】



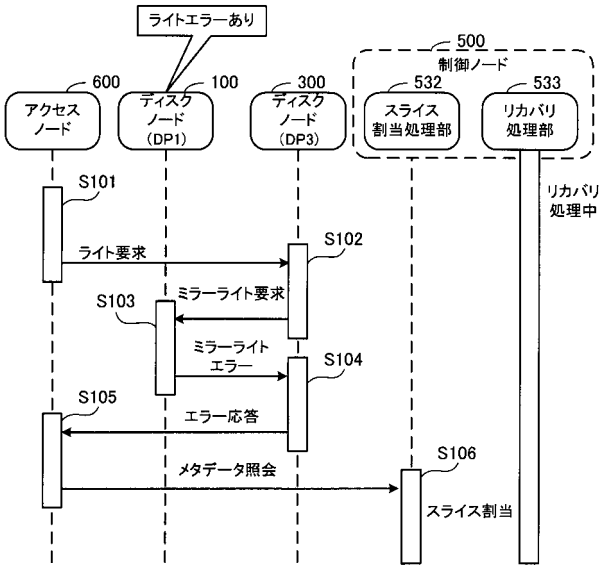
【図１７】



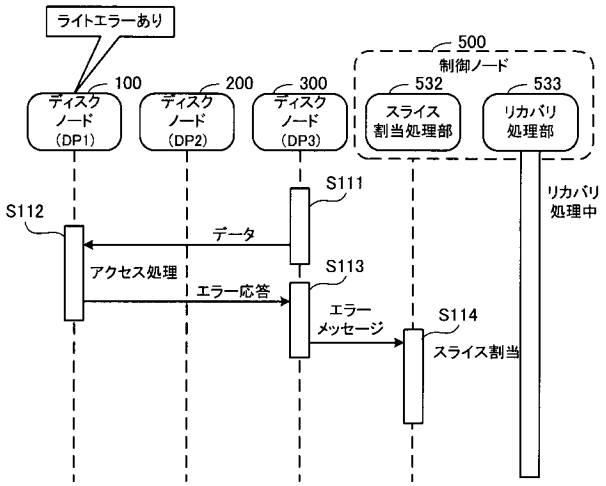
【図１８】



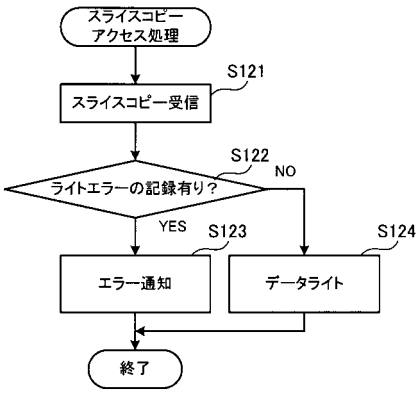
【図１９】



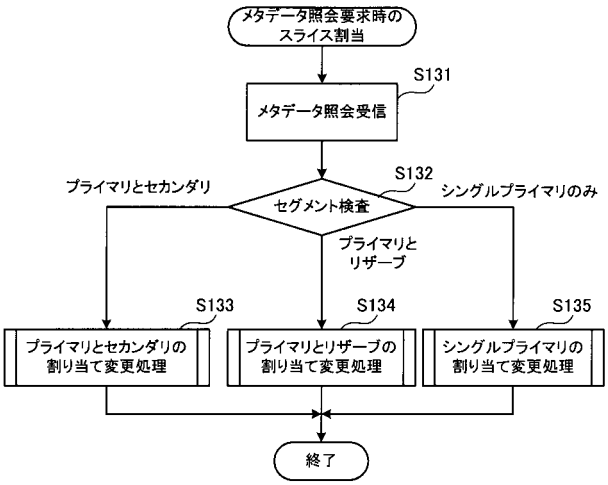
【図２０】



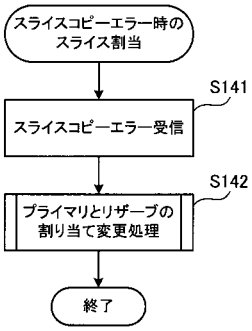
【図２１】



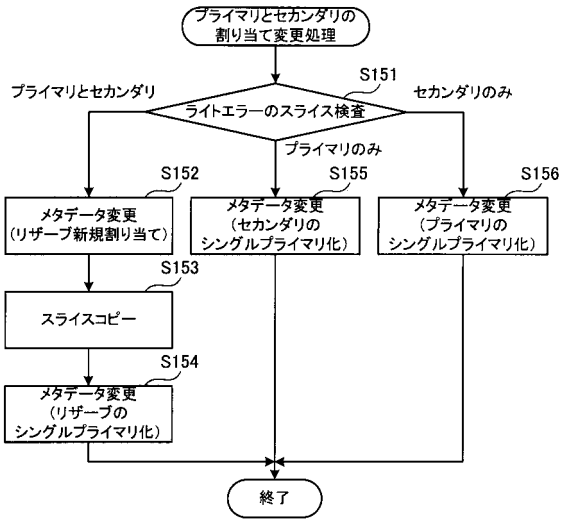
【図２２】



【図２３】



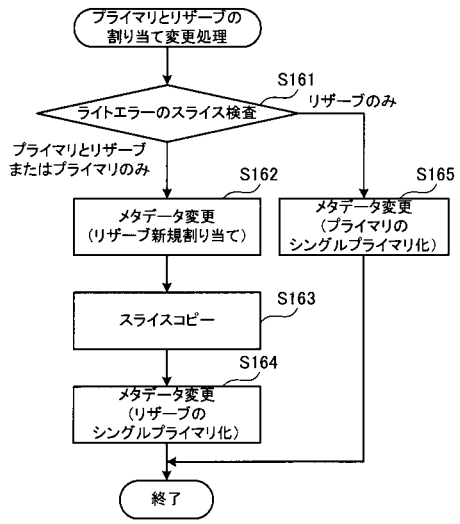
【図２４】



【図２５】

ライトエラー（DP1,DP2）	DP1	DP2	DP3	シーケンス
初期状態（なし、なし）	P	S	F	
（あり、あり）	1.P→P 2.コピー元 3.P→F	1.S→F	1.F→R 2.コピー先 3.R→SP	1.メタデータ変更 2.スライスコピー 3.メタデータ変更
（あり、なし）	1.P→F	1.S→SP	変化なし	1.メタデータ変更
（なし、あり）	1.P→SP	1.S→F	変化なし	1.メタデータ変更

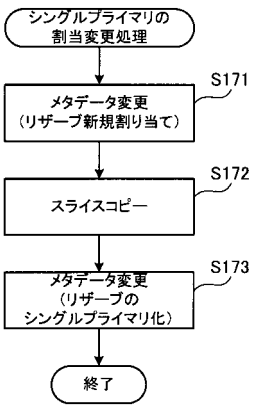
【図２６】



【図２７】

ライトエラー（DP1,DP2）	DP1	DP2	DP3	シーケンス
初期状態（なし、なし）	P	R	F	
（あり、あり）	1.P→P 2.コピー元 3.P→F	1.R→F	1.F→R 2.コピー先 3.R→SP	1.メタデータ変更 2.スライスコピー 3.メタデータ変更
（あり、なし）	1.P→P 2.コピー元 3.P→F	1.R→F	1.F→R 2.コピー先 3.R→SP	1.メタデータ変更 2.スライスコピー 3.メタデータ変更
（なし、あり）	1.P→SP	1.R→F	変化なし	1.メタデータ変更

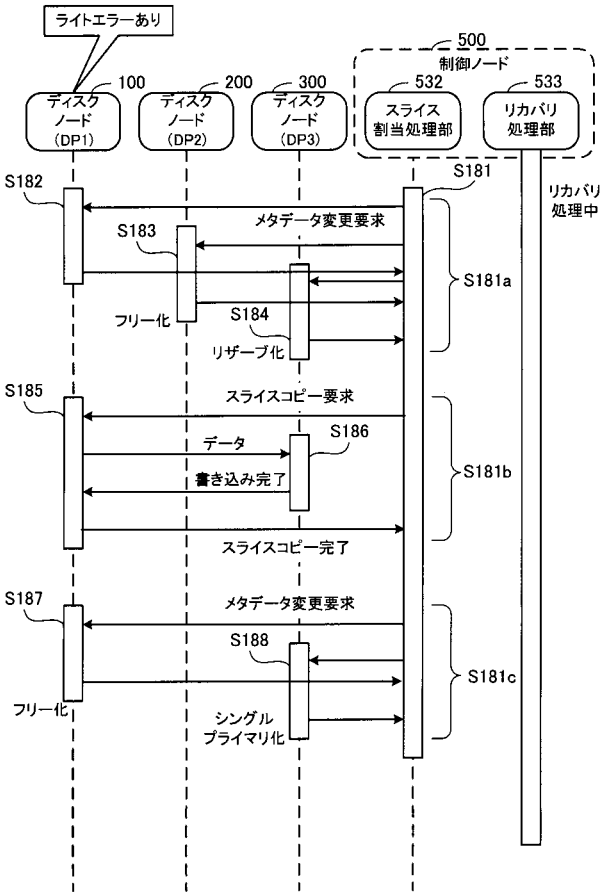
【図２８】



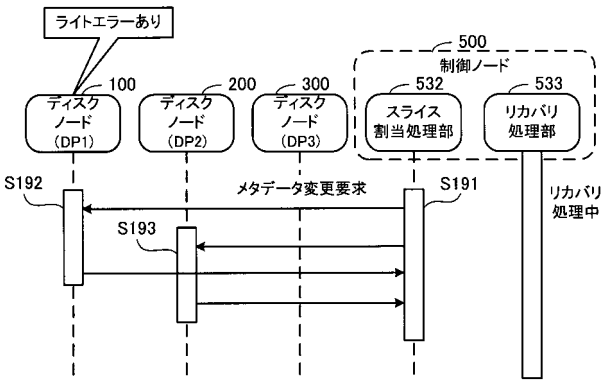
【図２９】

ライトエラー（DP1,DP2）	DP1	DP2	シーケンス
初期状態（なし、なし）	SP	F	
（あり、なし）	1.SP→P 2.コピー元 3.P→F	1.F→R 2.コピー先 3.R→SP	1.メタデータ変更 2.スライスコピー 3.メタデータ変更

【図３０】


ライトエラーあり
500
制御ノード
100
ディスク
ノード
(DP1)
200
ディスク
ノード
(DP2)
300
ディスク
ノード
(DP3)
532
スライス
割当処理部
533
リカバリ
処理部
リカバリ
処理中
S181
S182
メタデータ変更要求
S183
S181a
フリー化
S184
リザーブ化
S185
スライスコピー要求
データ
S186
書き込み完了
S181b
スライスコピー完了
S187
メタデータ変更要求
S188
S181c
フリー化
シングル
プライマリ化

【図３１】


ライトエラーあり
500
制御ノード
100
ディスク
ノード
(DP1)
200
ディスク
ノード
(DP2)
300
ディスク
ノード
(DP3)
532
スライス
割当処理部
533
リカバリ
処理部
リカバリ
処理中
S192
メタデータ変更要求
S191
S193