(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7179947号
(P7179947)

(45)発行日　令和4年11月29日(2022.11.29)

(24)登録日　令和4年11月18日(2022.11.18)

(51)国際特許分類

		F I		
G06F 3/06 (2006.01)		G06F	3/06	304B
G06F 13/10 (2006.01)		G06F	3/06	301X
G06F 11/20 (2006.01)		G06F	3/06	301Z
G06F 16/188 (2019.01)		G06F	13/10	340A
		G06F	11/20	671

請求項の数　6　（全32頁）　最終頁に続く

(21)出願番号　特願2021-179972(P2021-179972)
(22)出願日　令和3年11月4日(2021.11.4)
(62)分割の表示　特願2018-177578(P2018-177578)の分割
原出願日　平成30年9月21日(2018.9.21)
(65)公開番号　特開2022-20744(P2022-20744A)
(43)公開日　令和4年2月1日(2022.2.1)
審査請求日　令和3年11月4日(2021.11.4)

(73)特許権者　000005108
株式会社日立製作所
東京都千代田区丸の内一丁目６番６号
(74)代理人　110001689青稜弁理士法人
(72)発明者　佐藤 賢太
東京都千代田区丸の内一丁目６番６号
株式会社日立製作所内
(72)発明者　出口 彰
東京都千代田区丸の内一丁目６番６号
株式会社日立製作所内
(72)発明者　川口 智大
東京都千代田区丸の内一丁目６番６号
株式会社日立製作所内
審査官　吉田 歩

最終頁に続く

(54)【発明の名称】　ストレージシステム及びストレージ制御方法

(57)【特許請求の範囲】

【請求項１】

クラスタを構成する複数のストレージノードを有するストレージシステムにおいて、

データを記憶する記憶装置と、

ストレージシステムの制御を行うクラスタ制御部と、

各ストレージノードに設けられ、前記記憶装置を利用してボリュームという単位で記憶領域をホスト装置に提供し、ホスト装置からのＩＯ要求に応じて記憶装置にデータを保存するストレージ制御部と、を有し、

前記ストレージ制御部は、前記クラスタ内の他のストレージノードのストレージ制御部とストレージ制御部グループを構成し、前記ストレージ制御部グループの少なくとも１のストレージ制御部がアクティブモードのストレージ制御部としてホスト装置からのＩＯ要求を処理し、前記ストレージ制御部グループの他のストレージ制御部が前記アクティブモードのストレージ制御部の処理を引き継ぐことが可能であるよう、構成され、

前記クラスタ制御部は、

減設または障害によって、稼働するストレージノードが減少する時に、第１のストレージ制御部グループが担当するボリュームの識別情報を取得し、取得した識別情報に対応する各ボリュームに対し、複数の第２のストレージ制御部グループを決定し、第１のストレージ制御部グループを構成するストレージ制御部から第２のストレージ制御部グループのストレージ制御部へ、前記第１のストレージ制御部グループが担当するボリュームを移動させ、

前記ボリュームを移動させてボリュームがなくなったストレージ制御グループを削除することを特徴とするストレージシステム。
【請求項２】
ストレージノードを減設する場合に、前記ボリュームの移動を行なうものであり、
減設するノードにアクティブモードのストレージ制御部が配置されるストレージ制御部グループのボリュームを、他のストレージ制御部グループに移動させることを特徴とする請求項１に記載のストレージシステム。
【請求項３】
減設するノードに、前記アクティブモードのストレージ制御部の処理を引き継ぐことが可能なスタンバイモードのストレージ制御部が配置される場合、ストレージ制御部グループのボリュームを他のノードに移動させてスタンバイモードのストレージ制御部を起動させて、同じストレージ制御部グループのスタンバイモードのストレージ制御部とすることを特徴とする請求項１に記載のストレージシステム。
【請求項４】
ストレージノードに障害が発生した場合に、前記ボリュームの移動を行なうものであり、
前記障害が発生したノードにアクティブモードのストレージ制御部が配置されるストレージ制御部グループは、他のノードに配置されるストレージ制御部をアクティブに切り替え、
前記障害が発生したノードにストレージ制御部が配置される少なくとも一つのストレージ制御部グループは、そのボリュームを他のストレージ制御部グループに移動させて削除されることを特徴とする請求項１に記載のストレージシステム。
【請求項５】
前記障害が発生したノードにストレージ制御部が配置され、前記ボリュームを有する少なくとも一つのストレージ制御部グループは、前記削除されるストレージ制御部グループのストレージ制御部が配置されて障害が発生していないノードに、前記ボリュームをコピーしてストレージ制御部を起動してストレージ制御部グループを形成することを特徴とする請求項４に記載のストレージシステム。
【請求項６】
クラスタを構成する複数のストレージノードを有するストレージシステムの制御方法において、
前記ストレージシステムは、
データを記憶する記憶装置と、
ストレージシステムの制御を行うクラスタ制御部と、
各ストレージノードに設けられ、前記記憶装置を利用してボリュームという単位で記憶領域をホスト装置に提供し、ホスト装置からのＩＯ要求に応じて記憶装置にデータを保存するストレージ制御部と、を有し、
前記ストレージ制御部は、前記クラスタ内の他のストレージノードのストレージ制御部とストレージ制御部グループを構成し、前記ストレージ制御部グループの少なくとも１のストレージ制御部がアクティブモードのストレージ制御部としてホスト装置からのＩＯ要求を処理し、前記ストレージ制御部グループの他のストレージ制御部が前記アクティブモードのストレージ制御部の処理を引き継ぐことが可能であるよう、構成され、
前記クラスタ制御部は、減設または障害によって、稼働するストレージノードが減少する時に、第１のストレージ制御部グループが担当するボリュームの識別情報を取得し、取得した識別情報に対応する各ボリュームに対し、複数の第２のストレージ制御部グループを決定し、第１のストレージ制御部グループを構成するストレージ制御部から第２のストレージ制御部グループのストレージ制御部へ、前記第１のストレージ制御部グループが担当するボリュームを移動させ、
前記ボリュームを移動させてボリュームがなくなったストレージ制御グループを削除することを特徴とするストレージシステムの制御方法。
【発明の詳細な説明】

【技術分野】
【０００１】
本発明は、ストレージシステム及びその制御方法に関する。
【背景技術】
【０００２】
高い信頼性を求められる情報処理システムでは、複数のサーバを用いてシステムを冗長化することが一般的である。しかしながら、このような冗長化構成の場合、サーバ障害発生後に冗長度を回復させるために、障害が発生したサーバの代替となるスペアサーバを用意しておく必要がある。通常時にスペアのサーバは処理を行わないため、サーバの利用効率が低下する。
【０００３】
一方、近年では、仮想化技術を用いて、サーバを仮想化することで、物理サーバの利用効率を向上させ、物理サーバの台数を削減する構成も増えている。仮想マシンの冗長化に関する発明が、例えば特許文献１に開示されている。特許文献１では、複数台の現用系の仮想マシンと、これら現用系の仮想マシンを冗長化するために設けられた予備系の仮想マシンとを、複数の物理サーバに配置する技術が開示されている。このような仮想マシンを配置する技術によれば、物理サーバ障害によって冗長化した仮想マシンの片方を喪失した場合に、喪失した仮想マシンを別の物理サーバ上にコピーして冗長化構成を再構築することで、スペアの物理サーバを用意することなく冗長度を回復させること可能としている。
【先行技術文献】
【特許文献】
【０００４】
【文献】特開２０１４－７５０２７号公報
【発明の概要】
【発明が解決しようとする課題】
【０００５】
高い信頼性が求められる情報処理システムの中には、例えば、冗長化動作を行うための動作基盤といった、システム内部の基盤処理を安定動作させるために、システムが処理する情報量とは関係なく、一定量のCPUコアやメモリなどの情報処理資源を必要とするものがある。例えば、仮想化技術を用いたストレージシステムが安定動作するためには、ボリューム数などとは関係なく、一定量の情報処理資源を必要とする。
【０００６】
また、仮想マシン、コンテナ、マルチプロセスなどによって、1台のサーバ上で複数の独立したシステムを動作させる構成において、このような一定量の情報処理資源を最低限必要とするシステムを動作させる場合、同一サーバ上で動作する他のシステムの影響を受けないようにする必要がある。そのため、当該システムに必要な情報処理資源を予約し、当該システムに固定的に割り当てることが一般的である。
【０００７】
しかしながら、このようなシステムに特許文献１の配置技術を適用する場合、冗長度が低下したシステムに、当該システムが必要とする最低限の情報処理資源が再構築先の物理サーバに余っている必要がある。そのため、確実に冗長度を回復させるためには、予め物理サーバに冗長度回復用の情報処理資源を予約しておく必要がある。冗長度回復用に予約された情報処理資源は、障害等により冗長度の低下する場合以外は、利用されないため物理サーバの利用効率が低下し、システム構築コストが高くなる。
【０００８】
また、仮想化技術を用いたストレージシステムである、Software Defined Storage(SDS)は高い信頼性が求められる一方、比較的安価なサーバを用いて、低コストで情報処理システムを構築することが求められている。
【０００９】
本発明の目的は、システムの可用性を担保しつつ、低コストのストレージシステム、お

よびストレージ制御方法を提供することにある。
【課題を解決するための手段】
【００１０】
上記課題を解決するため、本発明の望ましい態様の一つのストレージシステムは、クラスタを構成する複数のストレージノードを有するストレージシステムにおいて、各ストレージノードは、データを記憶する記憶装置と、ストレージシステム全体の制御を行うクラスタ制御部と、記憶装置を利用してボリュームという単位で記憶領域をホスト装置に提供し、ホスト装置からのIO要求に応じて前記記憶装置にデータを保存するストレージ制御部とを有し、ストレージ制御部は、クラスタ内の他のストレージノードのストレージ制御部とストレージ制御部グループを構成し、ストレージ制御部グループの一つのストレージ制御部がアクティブモードのストレージ制御部としてホスト装置からのIO要求を処理し、ストレージ制御部グループの残りストレージ制御部がスタンバイモードのストレージ制御部として、アクティブモードのストレージ制御部が失われた場合に、アクティブモードに切り替わり、アクティブモードのストレージ制御部の処理を引き継ぐよう構成され、複数のストレージノードの内の一台のストレージノードをストレージシステムから取り除く場合、残存する他のストレージノードのクラスタ制御部は、取り除かれるストレージノードのストレージ制御部を用いて構成されるストレージ制御部グループが担当する複数のボリュームを取得し、取得した複数のボリュームの各ボリュームに対し退避先のストレージ制御部グループを決定し、取り除かれるストレージノードのストレージ制御部を用いて構成されるストレージ制御部グループを構成するストレージ制御部から退避先の複数のストレージ制御部グループのストレージ制御部へ、取り除かれるストレージノードのストレージ制御部を用いて構成されるストレージ制御部グループが担当する複数のボリュームを分散して退避させる。
【発明の効果】
【００１１】
本発明により、冗長度の回復可能性を保証するための予約情報処理資源が不要となり、物理サーバの利用効率が向上する。
【図面の簡単な説明】
【００１２】
【図１】実施例１による情報処理システムの全体構成を示すブロック図である。
【図２】ストレージノードの詳細構成を示すブロック図である。
【図３】実施例１によるストレージシステムの論理的な構成を示す図である。
【図４】実施例１によるデータ管理を説明する図である。
【図５】ストレージノード管理表の一例を示す図である。
【図６】ストレージ制御部管理表の一例を示す図である。
【図７】ボリューム管理表の一例を示す図である。
【図８】論理チャンク管理表の一例を示す図である。
【図９】物理チャンク管理表の一例を示す図である。
【図１０】本発明の課題を説明する図である。
【図１１】メモリに格納されるプログラムと管理情報の一例を示す図である。
【図１２】障害回復プログラム（１）の処理の一例を示す図である。
【図１３】障害回復プログラム（２）の処理の一例を示す図である。
【図１４】ボリューム退避プログラム（１）の処理の一例を示す図である。
【図１５】ボリューム退避先決定プログラム（１）の処理の一例を示す図である。
【図１６】ストレージ制御部ペア削除プログラム（１）の処理の一例を示す図である。
【図１７】ストレージノード減設プログラム（１）の処理の一例を示す図である。
【発明を実施するための形態】
【００１３】
以下、図面を参照して、本発明の実施形態について詳述する。ただし、以下の記載および図面は、本発明を説明するための例示であって、説明の明確化のために、適宜省略およ

び簡略化が行われており、本発明の技術的範囲を限定するものではない。

【００１４】

以後の説明では「テーブル」、「表」、「リスト」、「キュー」などの表現にて各種情報を説明するが、各種情報はこれら以外のデータ構造で表現されていてもよい。そのため、データ構造に依存しないことを示すために、単に「情報」と呼ぶことがある。各種情報の内容を説明する際に、「識別情報」、「識別子」、「名」、「名前」、「ID」、「番号」などの表現を用いるが、これらについては相互に置換が可能である。

【００１５】

以後の説明では、「プログラム」を主語として説明を行う場合があるが、プログラムはプロセッサ（例えばCPU（Central Processing Unit）やＧＰＵ（Graphics Processing Unit））によって実行されることで定められた処理を、記憶資源（例えばメモリ）やインタフェースデバイス（例えば通信装置）などを適宜用いながら行うため、プロセッサを主語とした説明としてもよい。同様に、プログラムを実行して行う処理の主体が、プロセッサを有する例えばコントローラ、装置、システム、計算機、ノード、ストレージ装置、サーバ、クライアント、又はホストであってもよい。また、プログラムの一部または全ては、ハードウェア回路を用いて処理してもよい。

【００１６】

各種プログラムは、プログラム配布サーバや記憶メディアによって各計算機にインストールされてもよい。また、以後の説明において、２以上のプログラムが１つのプログラムとして実現されてもよく、逆に１つのプログラムが２以上のプログラムとして実現されてもよい。

【実施例１】

【００１７】

以下、図１～図１６を参照して、実施例１について詳述する。

【００１８】

図１は、実施例１による情報処理システムの物理的な構成を示す図である。この情報処理システムは、1以上のホスト装置100と、1以上の管理端末110と、2以上のストレージノード200から構成されるマルチノード構成のストレージシステム200とを備えて構成される。各ホスト装置100及び管理端末110、各ストレージノード210間は、例えばファイバーチャネル（Fibre Channel）、イーサネット（登録商標）、無線ＬＡＮ（Local Area Network）又はInfiniBandなどから構成されるネットワーク300を介して接続される。図には示していないが、ネットワーク300は、ネットワークスイッチやゲートウェイといった各種中継装置を含んでいてもよい。なお、各ストレージノード210間専用のネットワークを別途備えていてもよく、また、各ホスト装置100及び管理端末110、各ストレージノード210がそれら以外のネットワークに接続されていてもよい。

【００１９】

ホスト装置100は、インストールされたアプリケーションプログラムを実行することで各種業務処理を行うためのサーバ装置である。ホスト装置100は実行しているアプリケーションプログラムからの要求に応じて、ネットワーク300を介してストレージノード210に対してデータの読み込み要求又は書き込み要求を送信する。なお、ホスト装置100は仮想マシンやコンテナのような仮想的なサーバ装置であってもよい。

【００２０】

管理端末110は、ストレージシステムの管理者がストレージシステム200に対して各種の設定操作や状態監視を行うためのクライアント装置である。管理端末110はスマートフォンやタブレット端末のような携帯端末でもよく、一部のホスト装置100が管理端末を兼ねていてもよい。

【００２１】

ストレージシステム200は、ホスト装置100に対してデータの読み書きをするための記憶領域を提供するサーバ装置である。なお、ストレージシステム200を構成するストレージノード210は仮想マシンやコンテナのような仮想的なサーバ装置であってもよく、ホス

ト装置100の仮想的なサーバ装置とストレージノード210の仮想的なサーバ装置を同一の物理サーバ装置に配置する構成でもよい。

【００２２】

図２は、ストレージノード210の詳細構成を示す図である。ストレージノード210は、CPU211、メモリ212、記憶装置213、通信装置214とを備えており、これらが内部ネットワーク215を介して接続されたサーバ装置により構成される。ただし、図２はストレージノードの一例であり、本発明は図の構成に限定されるものではなく、CPU211、メモリ212、記憶装置213、通信装置214の全て或いは何れかが複数であっても良い。

【００２３】

CPU211は、ストレージノード210全体の動作制御を司る制御装置であり、メモリ212に格納された各種プログラムを実行することで、各種処理を実行する。メモリ212は、例えば、ストレージノード210で使用される制御情報、CPU211が実行するプログラム、ホスト装置がアクセスするデータなどを格納する。メモリ212は一般にＤＲＡＭ（Dynamic RAM（Random Access Memory））で構成するが、例えば、ＭＲＡＭ（Magnetoresistive RAM），ReRAM（Resistive RAM）、ＰＣＭ（Phase Change Memory）、ＮＡＮＤなど、ＤＲＡＭ以外の記憶メディアで構成されてもよい。

【００２４】

記憶装置213は、物理的に記憶領域を有する装置であり、例えば、HDD（Hard Disk Drive）、SSD（Solid State Drive）、SCM（Storage Class Memory）、又は光ディスクなどの不揮発性の記憶装置から構成される。記憶装置213へアクセスするためのインタフェースとして、SAS（Serial Attached SCSI）とNVMe（Non-Volatile Memory Express）を記載しているが、例えば、SATA（Serial ATA）、USB（Universal Serial Bus）など、それ以外のインタフェースであってもよい。

【００２５】

一般に、マルチノード構成のストレージシステムでは、ノード障害に備え、別のストレージノード210にデータの複製を格納してデータを保護する。ノード内で複数の記憶装置213を束ねてRAID（Redundant Arrays of Independent Disks）のような高信頼化技術を使用してもよい。

【００２６】

通信装置214は、ネットワーク300を介してホスト装置100や他のストレージノード210、ストレージシステム200を管理するための管理端末110等と接続され、ホスト装置100や管理端末110、他のストレージノード210との通信を仲介する。図２では、通信装置214をホスト装置100向けの通信と管理端末110向けの通信、他のストレージノード210向けの通信とで共有しているが、それぞれの通信のために異なる通信装置を設けてもよい。

【００２７】

図３は、実施例１によるストレージシステムの論理的な構成を示す図である。クラスタ制御部216は、複数のストレージノードから構成されるストレージシステム全体の制御を司るソフトウェアである。クラスタ制御部216には、マスタとワーカの２種類の動作ロールがある。ワーカロールのクラスタ制御部216bはマスタロールのクラスタ制御部216aの指示に従ってストレージノード内の各種制御や状態監視を行い、マスタロールのクラスタ制御部216aはクラスタ全体での排他制御や一貫性制御が必要な処理、管理端末110を介した各種設定操作の処理や障害等発生時の通知等を行う。なお、マスタはワーカの機能を内包している。

【００２８】

マスタロールとして動作するクラスタ制御部はクラスタ内に常に１つ存在し、その他のクラスタ制御部はワーカロールで動作する。マスタロールのクラスタ制御部とワーカロールのクラスタ制御部は、ストレージノード間通信などによって、互いに生死監視を行っている。マスタロールのクラスタ制御部は、ストレージノード障害などによって、クラスタ制御部が失われた場合は、ストレージノード障害が発生したと判断し、障害回復処理を行う。障害回復処理の詳細に関しては、図を用いて後述する。

【００２９】
マスタロールのクラスタ制御部が失われた場合は、クラスタ内のワーカロールのクラスタ制御部の何れか１つがマスタロールに切り替わる。複数のワーカロールのクラスタ制御部からマスタロールに切り替わるクラスタ制御部の選出に関しては、一般に「リーダ選出」と呼ばれる技術及び機能を利用するため、説明は省略する。
【００３０】
ストレージ制御部219は、記憶領域としてホスト装置に提供するボリュームに関する各種制御を司るソフトウェアにより実現される。ストレージ制御部219は、記憶装置を利用してボリュームという単位で記憶領域をホスト装置に提供し、ホスト装置からのIO（Input／Output）要求に応じて記憶装置にデータを保存する機能を有する。また、あるストレージ制御部が担当していたボリュームを他のストレージ制御部に移動させる機能（マイグレーション機能）も有する。
【００３１】
ストレージ制御部219には、アクティブとスタンバイの２種類の動作モードがある。あるストレージノードに配置されたアクティブモードのストレージ制御部219aは、クラスタ内の異なるストレージノードに配置されたスタンバイモードのストレージ制御部219bとペア（ストレージ制御部ペア217と呼ぶ）を構成して動作する。ストレージ制御部ペアの他、一つのアクティブモードのストレージ制御部に複数のスタンバイモードのストレージ制御部を対応させる場合、ストレージ制御部グループとする。通常時は、アクティブモードのストレージ制御部219aがホスト装置からのIO要求を処理する。スタンバイモードのストレージ制御部219bは、ストレージノード障害などによるアクティブモードのストレージ制御部219aの喪失に備えて待機しておく。アクティブモードのストレージ制御部219aが失われた場合は、スタンバイモードのストレージ制御部219bがアクティブモードに切り替わり、IO要求などの処理を引き継ぐ。なお、ストレージ制御部ペアを構成する二つのストレージ制御部の両方がアクティブモードで動作していてもよく、ストレージ制御部グループを
構成する二つ以上のストレージ制御部がアクティブモードで動作していてもよい。この場合は、アクティブモードのストレージ制御部間で排他制御などの追加処理が必要になる。また、ストレージノード障害などによってアクティブモードのストレージ制御部が失われた場合に、残存しているアクティブモードのストレージ制御部が、失われたアクティブモードのストレージ制御部が担当していたIO要求などの処理を引き継いでもよい。
【００３２】
図３に示すように、１つのストレージノードに、２以上のストレージ制御部を配置してもよい。また、１つのストレージノードに配置するアクティブモードとスタンバイモードのストレージ制御部を同数に揃えることで、ストレージノード間での、例えばCPUやメモリなどの情報処理資源の利用率を均等にすることができる。
【００３３】
データ冗長化部218は、複数のストレージノード210間でデータを冗長化して記憶装置に保存することで、ストレージノード障害によるデータ喪失を防止するためのソフトウェアにより実現される。データ冗長化の方法として、例えば、異なるストレージノード210にデータの複製を格納する方法や、パリティを複数のストレージノード210に分散して格納する方法などが考えられる。図には示していないが、ストレージノード内の記憶装置障害に備えて、ストレージノード間のデータ冗長化に加えて、ノード内でRAIDなどのデータ冗長化を行ってもよい。
【００３４】
このように、実施例１は、サーバを仮想化する仮想化技術の応用例として、複数台の物理サーバをストレージノードとして利用するストレージシステムに関する。このようなストレージシステムでは、ホスト装置に記憶領域としてボリュームを提供するストレージ制御部のアクティブ（現用系）とスタンバイ（予備系）を、異なるストレージノード間に配置して冗長化する。さらに、ストレージシステム全体の処理性能を向上させるために、一

つのストレージノードには、アクティブとスタンバイから成るストレージ制御部ペアを複数備える。

【００３５】

図４は、実施例1によるデータ管理の概要を説明する図である。図４は、ホスト装置からの書き込み要求を処理する場合を示している。

【００３６】

データ冗長化部218が、チャンクを用いて複数のストレージノード間でデータを複製する。物理チャンク222は、ストレージノード内の記憶装置を１以上の所定容量の小領域(例えば、42MB)に分割して作成した物理的な記憶領域である。論理チャンク221は１以上の物理チャンクが対応付けられた論理的なチャンクである。論理チャンク221は後述するボリューム220のブロック223に対応付けられ、ホスト装置の書き込みデータが格納される。１つの論理チャンク221に、それぞれ異なるストレージノードに作成した２以上の物理チャンク222を対応付け、論理チャンク221に書き込まれたデータを、対応付けられた全ての物理チャンク222に保存することでノード間のデータ冗長化を実現する。図４では、アクティブモードとスタンバイモードのストレージ制御部219が配置された各ストレージノード210の物理チャンク222にデータを保存している。このように、アクティブモードとスタンバイモードのストレージ制御部219が配置されるストレージノード210にデータを保存（データのローカリティを確保）しているので、ホスト装置にボリュームを提供するストレージ装置に対し、データのリード要求があった場合、他のストレージノードからデータを読み出す必要がなく、高い応答性を確保できる。

【００３７】

データのローカリティを担保しなければ、データを任意の２つのストレージノードの物理チャンクに保存してもよい。例えば、ストレージ制御部219が配置されたストレージノードの記憶装置の空き容量が不足した場合に、記憶装置の空き容量に余裕のあるストレージノードの物理チャンクに保存する、といった処理を行ってもよい。

【００３８】

ボリューム220は、ストレージ制御部219がホスト装置100に提供する仮想的な記憶領域であり、ホスト装置100はボリュームに対してデータの書き込み要求を行う。ボリューム220は、ストレージシステム200の管理者が、管理端末110を介してストレージシステム200に対してボリューム作成指示を行うことによって作成される。ボリューム220の作成先となるストレージ制御部219は、ボリューム作成時に管理者が指定してもよく、ボリューム作成指示を受けたマスタロールのクラスタ制御部216aが、各ストレージノードの空き記憶容量や各ストレージ制御部のCPU利用率等を基に選択してもよい。

【００３９】

ボリューム自体は物理的な記憶領域を有しておらず、ホスト装置100からの書き込み要求に応じて論理チャンク221を割り当て、論理チャンク221にデータを論理的に書き込む。ボリューム220は、記憶領域を先頭から１以上の所定容量のブロック223に分割して管理される。このブロックは、例えば論理チャンクと一対一で対応付けられる。ボリューム作成直後は、どのブロックに対しても論理チャンクの対応付けは行われておらず、ホスト装置100がボリューム220に対してデータの書き込みを行った際に、データを書き込んだ領域に対応するブロック223に論理チャンク221が対応付けられてない場合に、論理チャンク221の作成と、ブロック223と論理チャンク221とを対応付ける処理が行われる。

【００４０】

ホスト装置100からのIO要求の処理はアクティブモードのストレージ制御部219aが担当する。新しい論理チャンクを作成し、ブロックと論理チャンクとの対応付けを行った場合は、その対応関係を表す情報をスタンバイモードのストレージ制御部219bに転送する。アクティブモードのストレージ制御部219aとスタンバイモードのストレージ制御部219bとで、一つのストレージ制御ペア217を構成する。図４に示したように、ストレージノード０のスタンバイモードのストレージ制御部219は、ストレージノード１以外のストレージノードのストレージ制御部とストレージ制御ペア217を構成し、ストレージノード１

のスタンバイモードのストレージ制御部219は、ストレージノード０以外のストレージノードのストレージ制御部とストレージ制御ペア217を構成する。

【００４１】

論理チャンク221に書き込まれたデータは、データ冗長化部218が論理チャンク221と物理チャンク222の対応関係に従って、物理チャンク222に書き込む。図４の例では、物理チャンクを複製（二重化）することでデータを冗長化する場合を示しており、ホスト装置100から書き込まれたデータは、「ストレージノード０」と「ストレージノード１」の物理チャンクに書き込まれる。物理チャンクを三重化する場合や、ストレージノード間でRAIDやErasure Codingを利用して冗長化する場合なども、データ冗長化部がその冗長化方式に応じて、物理チャンクの複製やパリティの生成を行う。図４では、ブロックと論理チャンクは同容量かつ一対一で対応付けられており、以後においても、ブロックと論理チャンクは一対一に付けられているものとして説明を進める。但し、例えば、１以上のボリュームからなる、２以上ブロックが１つの論理チャンクに対応付けられるものとしてもよい。

【００４２】

図４には示していないが、「ストレージノード０」の障害時に、「ストレージノード１」のストレージ制御部とデータ冗長化部が処理を引き継ぐため、「ストレージノード１」もボリュームに関する情報、ブロックに関する情報、論理チャンクに関する情報を有しており、どちらか一方のストレージノードで情報を更新すると、それに同期して、もう一方のストレージノードに更新内容が転送されて情報が更新される。それぞれの情報の詳細に関しては管理情報が記載されている図を用いて説明する。

【００４３】

次に、実施例１によるストレージシステムを制御するための管理情報（管理表）について説明する。なお、各種管理情報は、管理端末110を介してストレージシステム200の管理者が参照及び設定できるようにしてもよい。

【００４４】

図５は、ストレージノード管理表256の一例である。ストレージノード管理表は、表形式以外のデータ構造で表現されていてもよい。そのため、データ構造に依存しないことを示すために、単に「情報」と呼ぶことがある。ストレージノード管理表は、ストレージノードの動作状況と、ストレージノードが持つ各種情報処理資源を管理する情報である。ストレージノード管理表は、マスタロールのクラスタ制御部が動作するストレージノードのメモリに格納される。ストレージノード管理表256は、ストレージノードID2561、ロール2562、動作状態2563、CPUコア数2564、メモリ量2565、通信帯域利用率2566、記憶装置総容量2567、記憶装置総使用量2568を含むレコードを管理する。

【００４５】

ストレージノードID2561は、ストレージノードを一意に識別するIDであり、ストレージシステム全体でユニークなIDである。ロール2562は、当該ストレージノードで動作するクラスタ制御部の動作ロール（マスタ、ワーカ）を示す情報である。動作状態2563は、当該ストレージノードが正常に動作しているか否かを示す情報である。CPUコア数2564とメモリ量2565は、それぞれ、当該ストレージノードに搭載されたCPUのコア数とメモリの容量を示す情報である。通信帯域利用率2566は、当該ストレージノードに搭載された通信装置の帯域利用率を示す情報である。記憶装置総容量2567は、当該ストレージノードに搭載された記憶装置の容量の合計である。記憶装置総使用量2568は、当該ストレージノードに搭載された記憶装置の容量のうち、実際に利用している容量の合計である。ストレージノードIDが１のストレージノードは、マスタロールのクラスタ制御部として動作していることを示す。

【００４６】

通信帯域利用率2566、記憶装置総使用量2568は、定期的にマスタロールのクラスタ制御部が、各ストレージ制御部で動作するワーカロールのクラスタ制御部から取得した情報である。省略しているが、各ストレージノードは、マスタロールのクラスタ制御部が収集

するストレージノードの情報を管理している。また、ストレージノード障害などによって、当該ストレージノードのクラスタ制御部が動作していない場合は、マスタロールのクラスタ制御部が当該ストレージノードで障害が発生していると判断し、ストレージノード管理表の動作状態を障害に変更する。図４では、ストレージノードIDが０のストレージノードで障害が発生していることを示す。またストレージノード障害により、通信帯域利用率と記憶装置総使用量を取得できなかった場合は“ＮＡ”と示している。

【００４７】

図６は、ストレージ制御部管理表257の一例である。ストレージ制御部管理表257は、表形式以外のデータ構造で表現されていてもよい。そのため、データ構造に依存しないことを示すために、単に「情報」と呼ぶことがある。ストレージ制御部管理表257は、ストレージ制御部のペア関係と、ストレージ制御部とストレージノードの関係、ストレージ制御部の動作状況を管理する情報である。ストレージ制御部管理表257は、マスタロールのクラスタ制御部が動作するストレージノードのメモリに格納される。ストレージ制御部管理表257は、ストレージ制御部ID2571、ストレージ制御部ペアID2572、ストレージノードID2573、動作モード2574、割り当てCPUコア数2575、割り当てメモリ量2576、CPU利用率2577、メモリ利用量2578を含むレコードを管理する。

【００４８】

ストレージ制御部ID2571は、ストレージ制御部を一意に識別するIDであり、ストレージシステム全体でユニークなIDである。ストレージ制御部ペアID2572は、当該ストレージ制御部が属しているストレージ制御部ペアを一意に識別するためのIDである。ストレージノードID2573は、当該ストレージ制御部が配置されているストレージノードのIDを一意に識別するためのIDである。動作モード2574は、当該ストレージ制御部の動作モードがアクティブかスタンバイかの状態を示す情報である。

【００４９】

図６では、ストレージ制御部毎に一定量のCPUコアとメモリを固定的に割り当てる構成となっており、割り当てCPUコア数2575と割り当てメモリ量2576は、それぞれ、ストレージノードから当該ストレージ制御部に割り当てられているCPUコア数とメモリ量を示す情
報である。CPU利用率2577は、当該ストレージ制御部に割り当てられているCPUコアそれぞれの利用率の平均値を示す情報である。メモリ利用量2578は、当該ストレージ制御部に割り当てられているメモリのうち、実際に利用しているメモリ量を示す情報である。

【００５０】

図６では、ストレージ制御部ID2571が「０」のストレージ制御部は、ストレージノードIDが「０」のストレージノードにおいてアクティブモードで動作しており、ストレージノードIDが「１」のストレージノードにおいてスタンバイモードで動作するストレージ制御部IDが「１」のストレージ制御部とストレージ制御部ペアIDが「０」のストレージ制御部ペアを構成していることを示す。

【００５１】

CPU利用率2577とメモリ利用量2578は、定期的にマスタロールのクラスタ制御部216aが、各ストレージノードで動作するワーカロールのクラスタ制御部216bを介して、各ストレージ制御部から取得する情報である。図６では、ストレージノード障害などによって、当該ストレージノードのクラスタ制御部が動作していない場合は、“ＮＡ”と示している。

【００５２】

図７は、ボリューム管理表261の一例である。ボリューム管理表261は、表形式以外のデータ構造で表現されていてもよい。そのため、データ構造に依存しないことを示すために、単に「情報」と呼ぶことがある。ボリューム管理表261は、ボリュームとストレージ制御部ペアの関係と、ボリューム内のブロックと論理チャンクの関係、各ボリュームに対する単位時間当たりのIO量を管理する。ボリューム管理表261は、各ストレージノードのメモリに格納される。ボリューム管理表は、ボリュームID2611、容量2612、使用容量2613、ストレージ制御部ID2614、ブロックID2615、論理チャンクID2616、IO量2617

を含むレコードを管理する。ボリューム管理表261は、ストレージ制御部により参照することができる。

【００５３】

ボリュームID2611は、ボリュームを一意に識別するためのIDである。ボリュームはホスト装置に提供される資源であり、ストレージシステム全体でユニークなIDである。容量2612は当該ボリュームの容量を示す情報である。使用容量2613は、当該ボリュームが実際に利用している物理的な記憶領域の容量を示す情報である。使用容量2613は、論理チャンクが割り当てられているブロック数にブロックサイズを積算することでも計算可能である。ストレージ制御部ペアID2614は、ホスト装置から当該ボリュームへのIO要求の処理を担当するストレージ制御部ペアを一意に識別するためのIDである。ブロックID2615は当該ボリューム先頭からのブロック位置情報である。

【００５４】

論理チャンクID2616は、当該ボリュームの当該ブロックに対応付けられた論理チャンクを一意に識別するためのIDである。ストレージ制御部ペアID2614と論理チャンクID2616を組み合わせることで、当該ボリュームの当該ブロックに対応付けられた論理チャンクを一意に識別することが可能となる。IO量2617は、各ボリュームに対する単位時間当たりのIO量を表す情報である。

【００５５】

図７では、ボリュームID2611が「０」のボリュームは、ストレージ制御部ペアID2614が「０」のストレージ制御部ペアがホスト装置からのIO要求の処理を担当し、ブロックID2615が「０」のブロックに論理チャンクID2616が「０」の論理チャンクが対応付けられている。

【００５６】

このように、ボリューム管理表261は、各ボリュームとストレージ制御部ペアとを対応付けて管理している。一つのボリュームと当該ボリュームに対するホスト装置からのIO要求を担当するストレージ制御部ペアが一対一で対応している。ストレージ制御部ペアを構成するストレージ制御部は、図６のストレージ制御部管理表にて特定され、ストレージ制御部ペアを構成するストレージ制御部の内、アクティブのストレージ制御部がボリュームに対するホスト装置からのIOを処理するストレージ制御部として対応する構成となる。

【００５７】

図７では、ブロックと論理チャンクを一対一で対応付けた場合を示しているが、論理チャンクを分割し、１つの論理チャンクに複数のブロックを対応付ける場合は、分割した論理チャンクを識別するためのIDの列が追加される。

【００５８】

図８は、論理チャンク管理表271の一例である。論理チャンク管理表271は、表形式以外のデータ構造で表現されていてもよい。そのため、データ構造に依存しないことを示すために、単に「情報」と呼ぶことがある。論理チャンク管理表271は、論理チャンクと物理チャンクの関係と、論理チャンクに対応づいているストレージ制御部ペアを管理する情報である。論理チャンク管理表271は各ストレージノードのメモリに格納される。論理チャンク管理表271は、論理チャンクID2711、ストレージ制御部ペアID2712、ストレージノードID（マスタ）2713、物理チャンクID（マスタ）2714、ストレージノードID（ミラー）2715、物理チャンクID（ミラー）2716を含むレコードを管理する。論理チャンク管理表は、データ冗長化部218により参照することができる。

【００５９】

論理チャンクID2711は、論理チャンクを一意に識別するためのIDである。論理チャンクはストレージ制御部ペアに対応付けられる資源であり、ストレージ制御部ペア内でユニークなIDである。ストレージ制御部ペアID2712は当該論理チャンクと対応付けられているストレージ制御部ペアを一意に識別するためのIDである。ストレージノードID（マスタ）2713は、ストレージノードを一意に識別するためのIDである。物理チャンクID（マスタ）2714は、物理チャンクを一意に識別するためのIDである。ストレージノードIDと物

理チャンクIDを組み合わせることで、当該論理チャンクに対応付けられた物理チャンクを一意に識別することが可能となる。ストレージノードID（ミラー）2715と物理チャンクID（ミラー）2716には、障害に備えて冗長化（ミラー）された物理チャンクを識別するための情報である。図８では、論理チャンクID2711が「０」の論理チャンクは、ストレージ制御部ペア「０」に対応付けられており、ストレージノードID(マスタ)2713が「０」のストレージノードの物理チャンクID2714が「０」の物理チャンクと、ストレージノードID（ミラー）2715が「１」のストレージノードの物理チャンクID2716が「１」の物理チャンクに対応付けられている。

【００６０】

図８の例は、物理チャンクを複製（二重化）することでデータを冗長化する場合の論理チャンクテーブルの一例を示している。つまり、一つの論理チャンクに対し、一組のストレージ制御部ペアが対応し、マスタとミラーの複数のストレージノードにおいて、それぞれ物理チャンクが対応している。物理チャンクを３重化する場合や、ストレージノード間でRAIDやErasure Codingを適用する場合など、データ冗長化の方式に合わせて論理チャンク管理表の構造は変えてもよい。

【００６１】

図９は、物理チャンク管理表272の一例である。物理チャンク管理表272は、表形式以外のデータ構造で表現されていてもよい。そのため、データ構造に依存しないことを示すために、単に「情報」と呼ぶことがある。物理チャンク管理表272は物理チャンクに対応する記憶装置のアドレスを管理する。物理チャンク管理表は各ストレージノードのメモリに格納される。物理チャンク管理表は、物理チャンクID2721、記憶装置ID2722、記憶装置内オフセット2723を含むレコードを管理する。物理チャンク管理表272は、データ冗長化部218により参照することができる。

【００６２】

物理チャンクID2721は、物理チャンクを一意に識別するためのIDである。物理チャンクはストレージノード内の資源であり、ストレージノード内でユニークなIDである。記憶装置ID2722はストレージノード内の各記憶装置を識別するためのIDである。記憶装置内オフセット2723は、物理チャンクIDで識別される物理チャンクの先頭が対応する記憶装置のア
ドレスである。図９では、物理チャンクID2721が「０」の物理チャンクは、記憶装置ID2722が「０」の記憶装置に保存されており、物理チャンクの先頭アドレスは、記憶装置内オフセット2723で示される「0x0000」である。

【００６３】

図１０は、本発明の課題を説明する概念図である。図１０では、図３の構成において、「ストレージノード０」で障害が発生した場合を示している。

【００６４】

マスタロールのクラスタ制御部216aは、各ストレージノードで動作するワーカロールのクラスタ制御部216bとの定期的通信などにより、各ストレージノードの生死監視を行っている。

【００６５】

ストレージノードの障害を検出した場合は、まず、管理端末110を介してストレージシステム200の管理者に障害が発生したことを通知する。続いて、ホスト装置から当該ストレージ制御部が担当していたボリュームへのIO要求の処理を引き継ぐために、当該ストレージノードで動作していたアクティブモードのストレージ制御部219aとペアを構成していたスタンバイモードのストレージ制御部219bに対して、アクティブモードへの切り替えを指示する。図１０では、「ストレージ制御部０」219aとペアを構成していた「ストレージ制御部１」219bの動作モードがスタンバイからアクティブに切り替わっている。

【００６６】

次に、ストレージ制御部ペア217の冗長度を回復させるために、正常なストレージノードにスタンバイモードのストレージ制御部219cを再構築する。マスタロールのクラスタ制

御部216aは、新しいストレージ制御部を動作させるのに必要な、CPUコアやメモリなどの情報処理資源に空きのあるストレージノードを選択し、当該ストレージノードのクラスタ制御部に対して、ストレージ制御部の再構築を指示する。図１０の例では、「ストレージノード２」210に、機能を喪失した「ストレージ制御部０」の代替として「ストレージ制御部２２」219cが再構築され、１つのストレージノードに３つのストレージ制御部が配置された状態となっている。また、図示していないが、「ストレージ制御部２１」の代替となるストレージ制御部の再構築も行われている。

【００６７】

ストレージ制御部再構築先を選択する際に、情報処理資源に空きのあるストレージノードが存在しなかった場合は、ストレージ制御部を再構築することができず、冗長度を回復することができない。各ストレージノードに予めストレージ制御部再構築用の情報処理資源を予約しておけば、ストレージノード障害時に確実にストレージ制御部の冗長度を回復可能なことを保証できるが、通常時は予約した情報処理資源を利用することができず、ストレージノードの利用効率が低下し、システム構築コストが高騰することになる。

【００６８】

図１１～１６を用いて、ストレージノードの利用効率が高く、システム構築コストを低く抑えるための技術を説明する。

【００６９】

図１１は、ストレージノードのメモリ212に格納されている制御情報(管理表)256、257、261、271、272とプログラム250～255、258、260、270を示している。なお、実際のメモリには、これら以外のプログラムや管理情報も格納され得るが、図１１は本発明の説明に必要となるものを示している。例えば、ホストからのIO要求を処理するためのプログラムやキャッシュ管理表などは省略されている。また、プログラムは各ストレージノードの記憶装置にも格納されており、ストレージシステム起動時やプログラム実行時などにメモリにロードされるものとする。電源障害などに備えて、管理表256、257、261、271、272を記憶装置に保存してもよく、メモリを記憶装置に保存した管理表のキャッシュとして用いてもよい。

【００７０】

障害回復プログラム250、ボリューム退避プログラム251、ボリューム退避先決定プログラム252、ストレージ制御部ペア作成プログラム253、ストレージ制御部ペア削除プログラム254、ストレージノード減設プログラム255は、ストレージ制御部ペア再構築プログラム258、クラスタ制御部216を構成するプログラムの一部である。障害回復プログラム250、ボリューム退避プログラム251、ボリューム退避先決定プログラム252、ストレージノード減設プログラム255、ストレージ制御部ペア再構築プログラム258は、クラスタ制御部216がマスタロールで動作する際に実行され得るプログラムである。ストレージ制御部ペア作成プログラム253、ストレージ制御部ペア削除プログラム254は、クラスタ制御部がワーカロールで動作する際に実行され得るプログラムである。

【００７１】

ストレージノード管理表256とストレージ制御部管理表257はマスタロールのクラスタ制御部216のメモリに格納される管理情報である。ストレージノード管理表256とストレージ制御部管理表257は、図５及び図６にそれぞれ内容が示されている。これらの管理表は１以上のワーカロールのクラスタ制御部に複製を保持されるものとする。これは、マスタロールのクラスタ制御部216が配置されたストレージノードに障害が発生した際に、ワーカロールのクラスタ制御部がマスタロールに昇格して処理を引き継ぐためである。後述する説明において、これらの管理表を更新する際は、クラスタ制御部間の通信などによって複製された管理表も同時に更新するものとする。

【００７２】

ボリューム移動プログラム260は、ストレージ制御部219を構成するプログラムの一部である。ボリューム管理表261はストレージ制御部219のメモリに格納される管理情報で

あり、図７にその内容は示されている。ボリューム管理表261は、ストレージ制御部ペアを構成するストレージ制御部間で複製される。これは、ストレージノード障害などによってアクティブモードのストレージ制御部が動作不能に陥った際にスタンバイモードのストレージ制御部が処理を引き継ぐためである。後述する説明において、ボリューム管理表261を更新する際は、ストレージ制御部間の通信などによって複製された管理表も同時に更新するものとする。

【００７３】

物理チャンク再配置プログラム270は、データ冗長化部218を構成するプログラムの一部である。論理チャンク管理表271と物理チャンク管理表272はデータ冗長化部のメモリに格納される管理情報である。論理チャンク管理表271と物理チャンク管理表272は、それぞれ図８及び図９にその内容が示されている。論理チャンク管理表271は、各ストレージノードのデータ冗長化部が当該ストレージノードに配置されたストレージ制御部ペアに関連するレコードのみを保持する。図８の例では、物理チャンクはマスタとミラーに複製されるため、1つのレコードは２つのストレージノードのデータ冗長化部間で複製されていることになる。物理チャンク管理表272は、ストレージノード内の記憶装置のアドレスを管理する情報のため、複製等を行う必要はない。

【００７４】

図１２と図１３は、それぞれ障害回復プログラム250の処理の一例を示す図である。ストレージノード障害などによってストレージシステム内に冗長度が低下したストレージ制御部グループが存在する状態から、すべてのストレージ制御部グループの冗長度が保たれている状態に回復させる処理をおこなうものである。ストレージノード障害に代えて、ストレージノード減設の場合にも、障害を減設に読み替えて適用可能である。以下、説明の容易化のため、ストレージ制御部グループを、一台のアクティブモードのストレージ制御部と一台のスタンバイモードのストレージ制御部からなるストレージ制御部ペアのケースで説明する。但し、ペアに限らず、3台以上のストレージ制御部から構成されるストレージ制御部グループの場合であっても、基本的には同一の処理を行うこととなる。ストレージ制御部ペア又はグループが複数のアクティブモードのストレージ制御部を含んでいる場合、回復処理のなかで生存しているアクティブモードのストレージ制御部はスタンバイモードのストレージ制御部と基本的には同様に扱ってもよい。

【００７５】

例えば、このプログラム250はストレージノード障害によるストレージ制御部ペアの冗長度低下によって低下したボリュームの冗長度を回復させるための処理を行う。また、このプログラムは、クラスタ制御部216のメモリに格納され、当該クラスタ制御部が配置されたストレージノードのCPUによって実行される。このプログラム250はマスタロールで動作するクラスタ制御部216がストレージノードの障害を検出した際に起動され、当該クラスタ制御部が実行する。なお、マスタロールのクラスタ制御部が配置されていたストレージノードで障害が発生した場合は、ワーカロールのクラスタ制御部から新たに選出されたマスタロールのクラスタ制御部が配置されたストレージノードのCPUによって、このプログラムは実行される。

【００７６】

図１２は、ストレージノード障害が発生した場合、ストレージ制御部ペアの片方のストレージ制御部を喪失したストレージ制御部ペアを全て削除し、解放された情報処理資源を用いて、新たにストレージ制御部ペアを作成している。図１０の例によると、障害が発生したストレージノード「０」には、２つのストレージ制御部ペア０とストレージ制御部ペア１０が存在するため、これら２つのストレージ制御部ペアを削除し、一つの制御部ペアを作成する。

【００７７】

図１３では、ペアの片方を喪失したストレージ制御部ペアのうち、アクティブモードのストレージ制御部を喪失したストレージ制御部ペアのみを削除し、解放された情報処理資源を用いて、スタンバイモードのストレージ制御部を喪失したストレージ制御部ペアの新

しいペア相手を再構築する。図１０の例によると、障害が発生したストレージノード０にあるアクティブなストレージ制御部219aが属するストレージ制御部ペア０を削除し、１つのストレージ制御部ペア１０の冗長度を回復させる。つまり、ストレージ制御部ペア０（及び、ストレージ制御部１）を削除したことで、ストレージノード１にストレージ制御部1つ分の資源が解放される。これを使ってストレージノード１にストレージ制御部２１を再構築して、ストレージ制御部ペア１０の冗長度を回復する。

【００７８】

まず、図１２の処理の例について説明する。図１２の障害回復プログラム（１）は、ストレージノード障害によってストレージ制御部ペアの片方を喪失したストレージ制御部ペアを特定する(S100)。この処理は、ストレージ制御部管理表257を参照し、障害となったストレージノードID2573で検索し、ストレージ制御ペアID2572を特定することで実現する。この際、特定したストレージ制御部ペアへの新規ボリューム作成を抑止する処理を追加してもよい。この処理は、図６のストレージ制御部管理表257にボリューム作成抑止フラグの列を追加し、特定したストレージ制御部に対してこのフラグをONにすることで実現できる。

【００７９】

続けて、障害回復プログラム（１）250は特定したストレージ制御ペアから処理対象となるストレージ制御ペアを選択する(S101)。実施例１(図１０)の場合、１つのストレージノードにアクティブモードとスタンバイモードのストレージ制御部が1つずつ配置されているため、ペアの片方を喪失したストレージ制御部ペアは２つ存在し、残存しているストレージ制御部の一方はアクティブモードで動作しており、他方はスタンバイモードで動作している。

【００８０】

障害回復プログラム（１）250は、残存ストレージ制御部の動作モードがスタンバイか否かを判定し(S102)、スタンバイモードでなければステップS103をスキップしステップS104へ進む。スタンバイモードであれば、ステップS103で当該ストレージ制御部に対して、アクティブモードへの切り替えを指示する。動作モードの切り替えに関する処理は、以下の通り実現される。障害回復プログラム（１）250は、ストレージ制御部管理表257を参照し、ステップS101で選択したストレージ制御ペアIDで検索することにより、ステップS101で選択したストレージ制御ペアを構成するストレージ制御部IDを絞り込む。さらに、スト
レージノード管理表256を参照して、絞り込んだストレージ制御部IDで検索し、動作状態が正常であるストレージノードに対応するストレージ制御部IDを特定することで残存ストレージ制御部IDを特定できる。特定した残存ストレージ制御部ID2571に対応するストレージ制御部管理表257の動作モード2574を取得することで、動作モードを取得することができる。

【００８１】

障害回復プログラム(1)は、ストレージ制御部からアクティブモードの切り替え完了を受領すると、ストレージ制御部管理表257の当該ストレージ制御部の動作モードを「アクティブ」に更新する(S103)。なお、スタンバイモードのストレージ制御部がペア相手のアクティブモードの監視を行い、障害が発生したことを検知すると、自律的にアクティブモードに動作を切り替える構成でもよい。この場合、ステップS103は、スタンバイモードのストレージ制御部がアクティブに切り替わるのを待機し、切り替え完了後にストレージ制御部管理表257を更新する処理となる。いずれにせよ、このステップにより、障害が発生したストレージノードで動作していたアクティブモードのストレージ制御部が実行していたボリュームへのIO要求の処理などが、スタンバイモードであったストレージ制御部に引き継がれる。

【００８２】

次に、障害回復プログラム（1）は、ステップS100で特定した全てのストレージ制御部ペアについて、ステップS101～S103の処理が完了したか否かを判定する(S104)。完了

した場合はステップS105に進み、完了していない場合はステップS101に戻る。

【００８３】

なお、図には示していないが、ステップS105に進む前に残存している各ストレージノードの空き記憶容量がストレージノード障害を回復するのに十分であるか否かを確認し、空き記憶容量不足による障害回復処理の失敗を抑止してもよい。この場合、空き記憶容量が不足すると判断した場合は、管理端末110を介してストレージシステム200の管理者に空き記憶容量不足を通知し、プログラムの処理を終了する。その後、ストレージシステム200の管理者が記憶装置の増設やストレージノードの追加といった対処を行ったうえで、管理端末110を介して、マスタロールのクラスタ制御部216aに障害回復プログラム250の再実行を指示する。

【００８４】

物理チャンクをストレージ制御部が配置されたストレージノード以外のストレージノードにも保存する構成としている場合や、物理チャンクを三重化している場合、ストレージノード間でRAIDやErasure Codingを適用して物理チャンクを冗長化している場合は、ストレージノード障害により喪失した物理チャンクによって、喪失したストレージ制御部と関係のないストレージ制御部ペアに対応付けられた論理チャンクの冗長度が低下している可能性がある。そのため、ステップS105に進む前に、障害回復プログラム（１）が各ストレージノードのデータ冗長化部に対して、喪失したストレージ制御部と関係のないストレージ制御部ペアに対応付けられた論理チャンクの冗長度低下の確認と冗長度の回復を指示する。指示を受けたデータ冗長化部は、冗長化方式に応じて、冗長度の回復処理を実行する。

【００８５】

例えば、物理チャンクを複製（二重化）する冗長化方式の場合、データ冗長化部は論理チャンク管理表271から障害が発生したストレージノードに物理チャンクを保存している論理チャンクが存在するか否かを確認する。存在する場合は、当該論理チャンクに対応付けられているストレージ制御部ペアID2712が喪失したストレージ制御部を含むストレージ制御部ペアか否かを確認する。喪失したストレージ制御部を含むストレージ制御部ペアでなければ、新しい物理チャンクを確保し、当該論理チャンクを構成する喪失していないほうの物理チャンクから、新たに確保した物理チャンクにデータをコピーし、論理チャンクと物理チャンクの対応付けを更新する。全てのストレージノードで論理チャンクの冗長度低下の確認と冗長度の回復が完了すると、ステップS105に進む。

【００８６】

次に、障害回復プログラム（１）はステップS100で特定したストレージ制御ペアから処理対象となるストレージ制御ペアを選択する(S105)。続けて、障害回復プログラム（１）は、当該ストレージ制御部ペアが担当していた全ボリュームをボリューム毎に、ストレージノード障害の影響を受けていない複数の正常なストレージ制御部ペアに、それぞれ退避させる処理を実行する(S106)。つまり、当該ストレージ制御部ペアが担当していた全ボリュームをボリューム毎に、異なるストレージ制御部ペアに分散して退避させる。このボリューム退避処理の詳細については、図１４、１５を用いて後述するが、この処理によって当該ストレージ制御部ペアが担当していた全ボリュームのデータが退避先のストレージ制御部ペアにコピーされ、以後、当該ボリュームへのホスト装置からのIO要求は退避先のストレージ制御部ペアが処理を担当する。

【００８７】

全ボリュームの退避完了後、障害回復プログラム（1）は、残存ストレージ制御部が動作しているストレージノードのクラスタ制御部に対して、当該ストレージ制御部ペアの削除を指示する(S107)。ストレージ制御部ペアの削除処理の詳細については、図１６を用いて後述する。ストレージ制御部ペアの削除によって、当該ストレージ制御部ペアに割り当てられていたCPU、メモリといった情報処理資源や、論理チャンク、物理チャンクなどの記憶資源が解放される。削除完了後、障害回復プログラム（1）は、ストレージ制御部管理表257を更新し、当該ストレージ制御部ペア2572のレコードを削除する(S108)。

【００８８】
次に、障害回復プログラム（1）は、ステップS100で特定したストレージ制御部ペアの削除が全て完了したか否かを判定する(S109)。完了した場合はステップS110に進み、完了していない場合はステップS105に戻る。
【００８９】
ペアを片方喪失したストレージ制御部ペアの削除が完了すると、障害回復プログラム（1）は、ペアを片方喪失したストレージ制御部ペアの残存ストレージ制御部が配置されていたストレージノードのクラスタ制御部に対して、ストレージ制御部及びストレージ制御部ペアの作成を指示する(S110)。指示を受けたクラスタ制御部(ストレージ制御部ペア作成プログラム253)は、CPUコアやメモリなどの情報処理資源を確保し、記憶装置からストレージ制御部を構成するプログラムをメモリにロードして、ストレージ制御部を起動する。新しいストレージ制御部及びストレージ制御部ペアは、削除によって解放された情報処理資源を用いて作成される。障害回復プログラム（１）は、ストレージ制御部及びストレージ制御部ペアの作成完了後、ストレージ制御部管理表257を更新し、レコードを追加する。
【００９０】
ストレージ制御部ペア作成プログラム253によるストレージ制御部及びストレージ制御部ペア作成処理については、ストレージシステム構築時と同様のため詳細は省略する。
【００９１】
次に、図１３の処理の例について説明する。図１３の障害回復プログラム（２）250は、図１２のステップS100を実行して、ストレージ制御部ペアの片方を喪失したストレージ制御部ペアを特定する(S200)。この際、特定したストレージ制御部ペアへの新規ボリューム作成を抑止する処理を追加してもよい。この処理は、図６のストレージ制御部管理表257にボリューム作成抑止フラグの列を追加し、特定したストレージ制御部に対してこのフラグをONにすることで実現できる。
【００９２】
続けて、障害回復プログラム（２）は、図１２のステップS101～S104を実行して、障害が発生したストレージノードで動作していたアクティブモードのストレージ制御部が実行していたボリュームへのIO要求等の処理を、ペアを構成するスタンバイモードのストレージ制御部に引き継ぐ(S201)。
【００９３】
なお、図には示していないが、ステップS202に進む前に残存している各ストレージノードの空き記憶容量がストレージノード障害を回復するのに十分であるか否かを確認し、空き記憶容量不足による障害回復処理の失敗を抑止してもよい。この場合、空き記憶容量が不足すると判断した場合は、管理端末110を介してストレージシステム200の管理者に空き記憶容量不足を通知し、プログラムの処理を終了する。その後、ストレージシステム200の管理者が記憶装置の増設やストレージノードの追加といった対処を行ったうえで、管理端末110を介して、マスタロールのクラスタ制御部216aに障害回復プログラム250の再実行を指示する。
【００９４】
なお、物理チャンクをストレージ制御部が配置されたストレージノード以外のストレージノードにも保存する構成としている場合や、物理チャンクを三重化している場合、ストレージノード間でRAIDやErasure Codingを適用して物理チャンクを冗長化している場合は、ストレージノード障害によって喪失した物理チャンクによって、喪失したストレージ制御部と関係のないストレージ制御部ペアに対応付けられた論理チャンクの冗長度が低下している可能性がある。そのため、ステップS202に進む前に、障害回復プログラム（２）が各ストレージノードのデータ冗長化部に対して、喪失したストレージ制御部と関係のないストレージ制御部ペアに対応付けられた論理チャンクの冗長度低下の確認と冗長度の回復を指示する。指示を受けたデータ冗長化部は、冗長化方式に応じて、冗長度の回復処理を実行する。例えば、物理チャンクを複製（二重化）する冗長化方式の場合、データ冗長

化部は論理チャンク管理表271から障害が発生したストレージノードに物理チャンクを保存している論理チャンクが存在するか否かを確認する。存在する場合は、当該論理チャンクに対応付けられているストレージ制御部ペアID2712が喪失したストレージ制御部を含むストレージ制御部ペアか否かを確認する。喪失したストレージ制御部を含むストレージ制御部ペアでなければ、新しい物理チャンクを確保し、当該論理チャンクを構成する喪失していないほうの物理チャンクから、新たに確保した物理チャンクにデータをコピーし、論理チャンクと物理チャンクの対応付けを更新する。全てのストレージノードで論理チャンクの冗長度低下の確認と冗長度の回復が完了すると、ステップS202に進む。

【００９５】

次に、障害回復プログラム（２）は、処理対象となるストレージ制御ペアを選択する(S202)。図１２と同様に、図１０で示した例によるとペアの片方を喪失したストレージ制御部ペアは２つ存在し、残存しているストレージ制御部の一方はアクティブモードで動作しており、他方はスタンバイモードで動作している。

【００９６】

障害回復プログラム（２）は、残存ストレージ制御部の動作モードがスタンバイか否かを判定し(S203)、スタンバイモードでなければ、ステップS204をスキップしステップS205へ進む。スタンバイモードであれば、ステップS204で図１２のステップS106～S108を実行し、当該ストレージ制御部ペアが担当する全ボリュームの退避、当該ストレージ制御部ペアの削除、ストレージ制御部管理表257から当該ストレージ制御部ペアに関するレコード削除を行う。

【００９７】

次に、障害回復プログラム（２）は、ステップS200で特定したストレージ制御部ペアのうち、スタンバイモードのストレージ制御部が残存しているストレージ制御部ペアの削除が完了したか否かを判定する(S205)。完了した場合はステップS206へ進み、完了していない場合はステップS202に戻る。

【００９８】

スタンバイモードのストレージ制御部が残存しているストレージ制御部ペアの削除が完了すると、障害回復プログラム（２）は、当該残存ストレージ制御部が配置されていたストレージノードのクラスタ制御部（ストレージ制御部ペア再構築プログラム258）に対して、アクティブモードのストレージ制御部が残存しているストレージ制御部ペアの再構築を指示する(S206)。

【００９９】

指示を受けたクラスタ制御部は、CPUコアやメモリなどの情報処理資源を確保し、記憶装置からストレージ制御部を構成するプログラムをメモリにロードして、ストレージ制御部を起動する。起動後にアクティブモードのストレージ制御部からボリューム管理表261をコピーする。ストレージ制御部ペアの再構築は、削除によって解放された情報処理資源を用いて行われる。障害回復プログラム（２）は、ストレージ制御部及びストレージ制御部ペアの作成完了後、ストレージ制御部管理表257を更新し、当該ストレージ制御部ペアの情報を更新する。この再構築処理は、本発明を適用しない場合のストレージ制御部ペアの冗長度回復と同じ処理のため、詳細は省略する。ステップS200において、特定したストレージ制御部ペアへの新規ボリューム作成を抑止する処理を追加した場合は、このステップの後に新規ボリューム作成の抑止を解除する処理を追加する。この処理は、図６のストレージ制御部管理表257において、ストレージ制御部ペアID2572が特定したストレージ制御部ペアとなっているストレージ制御部に対して、追加したボリューム作成抑止フラグをOFFにすることで実現できる。

【０１００】

図１２による処理の場合、ペアの片方を喪失した２つのストレージ制御部ペアが順番にボリューム退避を行っているが、実際には、２つのストレージ制御部ペアが同時にボリューム退避を行ってもよい。図１３による処理の場合、スタンバイモードのストレージ制御部が残存しているストレージ制御部ペアからのボリューム退避とストレージ制御部ペア削

除の完了まで、アクティブモードのストレージ制御部が残存しているストレージ制御部ペアのペア相手となるストレージ制御部の再構築を行うことができない。したがって、ボリュームの冗長度が低下している時間は、図１２による処理方が短く、可用性が高い。一方で、図１２による処理の場合、新しく作成したストレージ制御部ペアは、障害回復処理の完了直後はボリュームを1つも担当していない状態となり、ストレージ制御部に割り当てられている情報処理資源の利用効率が悪い状態になってしまう。

【０１０１】

このように、図１２の図１３の２つの処理は、可用性と情報処理資源の利用効率のトレードオフの関係にある。実際のストレージシステムにおいては、図１２と図１３の２つの処理方式のうち、どちらか一方のみを備えていてもよく、両方とも備えておき、クラスタ制御部が何らかの判断基準によってどちらの処理方式を実行するかを決定してもよく、ストレージシステム管理者が事前に設定した方式を実行するようにしてもよい。なお、本発明はストレージ制御部を再構築するための予約リソースを不要とすることで、情報処理資源の利用効率を改善するものだが、例えば、ストレージノード1台までの障害回復に必要な情報処理資源を予約するという構成にしてもよい。この場合、ストレージノード1台までの障害であれば、ボリューム退避を行わずに、予約しておいた情報処理資源を用いてストレージ制御部の再構築を行うことができる。

【０１０２】

図１２の処理による場合、新しく作成したストレージ制御部ペアは、障害回復処理の完了直後はボリュームを1つも担当していない状態となるため、情報処理資源の利用効率を高めるために、図６に示したストレージ制御部管理表257のCPU利用率2577等を参照して、負荷の高いストレージ制御部及びストレージ制御部ペアが担当するボリュームを新しく作成したストレージ制御部ペアに移動させ、システム全体の負荷の分散を図る処理を行っても良い。

【０１０３】

ところで、ストレージシステム全体の空き記憶容量が不足した場合、SDSでは新しいストレージノードを追加することで、記憶容量不足を解消するという運用が一般的である。この場合、新しく追加したストレージノードの物理チャンクを論理チャンクに割り当てることで、記憶容量不足が解消される。一方、Erasure Codingの中には、他のストレージノードの物理チャンクにアクセスすることなくリード処理を行えるという特徴(リードローカリティ)を持ったものがある。リードローカリティを持つErasure Codingを適用したストレージシステムにおいて、記憶容量が不足を解消するために、新しいストレージノードを追加し、追加したストレージノードの物理チャンクを論理チャンクに割り当てた場合、リードローカリティという特徴を失われてしまうという問題がある。障害回復処理後に、新しく作成したストレージ制御部ペアにボリュームを移動させるのと同様に、新しく追加したノードに新しく作成したストレージ制御部ペアにボリュームを移動させることで、リードローカリティを持つErasure Codingの特徴を失うことなく、空き記憶容量不足を解消することができる。

【０１０４】

図１４は、ボリューム退避プログラム251の処理の一例を示す図である。このプログラム251は、障害回復プログラム250から指示された、ストレージノード障害によってペアの片方を喪失したストレージ制御部ペアが担当する全ボリュームを、冗長度が低下していない正常なストレージ制御部ペアに退避させる処理を行う。つまり、ペアの片方を喪失したストレージ制御部ペアが担当する全ボリュームを、冗長度が低下していない正常な複数のストレージ制御部ペアに分散して退避させる処理を行う。このプログラム251は、クラスタ制御部216のメモリに格納されており、当該クラスタ制御部が配置されたストレージノードのCPUによって実行される。このプログラム251は、マスタロールのクラスタ制御部216が実行する障害回復プログラム250から起動され、当該クラスタ制御部が実行する。マスタロールのクラスタ制御部が障害により機能しない場合には、クラスタ内の他のワーカロールのクラスタ制御部がマスタロールとなり実行する。

【０１０５】
ボリューム退避プログラム251は、障害回復プログラム250から指定された退避元ストレージ制御部ペアが担当するボリューム一覧を取得する(S300)。つまり、ペアの片方を喪失したストレージ制御部ペアが担当する全ボリュームの情報を取得する。前述したように、ボリューム管理表261は、ストレージ制御部のメモリに存在する情報である。そのため、実際には当該ストレージ制御部ペアの残存ストレージ制御部が配置されているストレージノードのクラスタ制御部を介して、情報を取得することになる。ボリューム退避プログラム251は、残存ストレージ制御部が配置されているストレージノードのクラスタ制御部を介して、ストレージ制御部ペアの残存ストレージ制御部からボリューム管理表261を受け取り、ストレージ制御部ペアID2614と対応する全てのボリュームID2611を取得する。
【０１０６】
次に、ボリューム退避プログラム251は、取得したストレージ制御部ペアID2614と対応する全てのボリュームから未退避のボリュームを一つずつ選択し(S301)、ボリューム退避先のストレージ制御部ペアを決定するために、選択された各ボリュームに対してボリューム退避先決定処理を実行する(S302)。ボリューム退避先決定処理の詳細については、図１５を用いて後述する。
【０１０７】
ボリュームの退避先が決定後、ボリューム退避プログラム251は、退避元のストレージ制御部と退避先にストレージ制御部に対して、当該ボリュームの移動を指示する(S303)。このボリューム移動処理自体は、一般にボリュームマイグレーションなどと呼ばれる技術及び機能と同様のため、説明は省略する。当該ボリュームの退避が完了後、ボリューム退避プログラム251は、ステップS300で特定した全てのボリュームを退避が完了したか否かを判断する(S304)。全てのボリュームの退避が完了した場合、ボリューム退避プログラム251は、処理を終了する。一方、全てのボリュームの退避が完了していない場合、ボリューム退避プログラム251は、ステップS301に戻り別のボリュームに対してステップS302～S304を実行する。
【０１０８】
図１４の例では、ボリュームの移動完了を待ちながら１つずつ退避させているが、移動完了を待たずに次のボリューム退避を開始し、複数ボリュームを同時並行的に退避してもよい。ただし、この場合は退避先ストレージ制御部ペアを決定する際に、現在同時並行して退避処理を行っているボリュームの影響を考慮する必要がある。
【０１０９】
また、ボリューム作成時などに、ボリューム毎にGold（高ランク）、Silver（中ランク）、Bronze（低ランク）のようなランク付けを行えるようにしておき、ランクに応じてボリューム移動を行う順番を変えるようにしてもよい。同様に、ボリュームのランクに応じて、ボリューム移動を行う際の処理速度を変えてもよい。例えば、ランクが高い程、先にボリューム移動を行い、処理速度も速くすることで、ランクの高いボリュームの冗長度が低下している期間を短くすることができる。これらの処理は、図７のボリューム管理表261に、ボリュームのランクを示す列を追加し、ステップS301において移動対象のボリュームを選択する際や、ステップS303においてボリューム移動の指示を行う際に、当該列を参照することで実現できる。
【０１１０】
また、図１４の例では、ボリューム毎に独立して退避先ストレージ制御部ペアを決定しているが、あるボリュームのスナップショットから作成したボリュームのように依存関係のある複数ボリュームに対して、まとめて１つの退避先ストレージ制御部ペアを決定してもよい。複数のボリューム間で重複しているデータを取り除く重複排除機能を有するストレージシステムにおいては、データの重複度合いが高い複数のボリュームに対して、まとめて１つの退避先ストレージ制御部ペアを決定してもよい。
【０１１１】
図１５は、ボリューム退避先決定プログラム252の処理の一例である。このプログラム

252はボリューム退避プログラム251から指示されたボリュームの退避先として最適なストレージ制御部ペアを決定する処理を行う。即ち、図１５で示された処理は、ペアの片方を喪失したストレージ制御部ペアが担当する全ボリュームに対して、ボリューム毎に実行される。このプログラム252は、マスタロールのクラスタ制御部216のメモリに格納されており、当該クラスタ制御部が配置されたストレージノードのCPUによって実行される。このプログラム252は、マスタロールのクラスタ制御部が実行するボリューム退避プログラムから起動され、当該クラスタ制御部216が実行する。マスタロールのクラスタ制御部が障害により機能しない場合には、クラスタ内の他のワーカロールのクラスタ制御部がマスタロールとなり実行する。

【０１１２】

ボリューム退避先決定プログラム252は、ストレージ制御部管理表257を参照し、退避先候補となるストレージ制御部ペアの一覧を取得する(S400)。つまり、ストレージ制御部管理表257のストレージ制御部ペアID2572の全ての情報を取得する。

【０１１３】

取得したストレージ制御部ペアの一覧から、処理対象のストレージ制御部ペアを1つずつ選択し(S401)、さらに、ストレージ制御部ペアを構成するストレージ制御部から、処理対象のストレージ制御部を1つずつ選択する(S402)。ボリューム退避先決定プログラム252は、当該ストレージ制御部が配置されたストレージノードの動作状態が正常か否かを判定する(S403)。正常でなければ、ステップS401に戻り、別のストレージ制御部ペアに対してS402～S407を実行する。

【０１１４】

正常であれば、当該ストレージ制御部が配置されたストレージノードに当該ボリュームを退避させた際に、当該ストレージ制御部が配置されたストレージノードの記憶装置の空き容量が閾値以上であるか否かを確認する(S404)。閾値未満であれば、ステップS401に戻り、別のストレージ制御部ペアに対してS402～S407を実行する。

【０１１５】

閾値以上であれば、当該ストレージ制御部が配置されたストレージノードの通信帯域利用率が所定の閾値以下であるか否かを判定する(S405)。閾値よりも大きければ、ステップS401に戻り、別のストレージ制御部ペアに対してS402～S407を実行する。

【０１１６】

閾値以下であれば、ストレージ制御部管理表を参照し、当該ストレージ制御部のCPU利用率が所定の閾値以下であるか否かを判定する(S406)。閾値よりも大きければ、ステップS401に戻り、別のストレージ制御部ペアに対してS402～S407を実行する。閾値以下であれば、ステップS407に進む。

【０１１７】

ボリューム退避先決定プログラム252がステップS403で判定した、ストレージ制御部が配置されたストレージノードの動作状態は、ストレージ制御部管理表257からステップS402で選択したストレージ制御部ID2571に対応するストレージノードID2573を取得し、ストレージノードID2573に対応するストレージノード管理表256の動作状態2563から取得できる。

【０１１８】

ステップS404で判定したボリューム退避後のストレージノードの空き容量は、ストレージノード管理表256のストレージノードID2561に対応する記憶装置総容量2567から記憶装置総使用量2568を減じ、さらに、ボリューム管理表261のボリュームID2611に対応する使用容量2613を減ずることで取得できる。

【０１１９】

ステップS405で判定したストレージノードの通信帯域利用率は、ストレージ制御部管理表257からステップS402で選択したストレージ制御部ID2571に対応するストレージノードID2573を取得し、ストレージノードID2573に対応するストレージノード管理表256の動作状態2563と通信帯域利用率2566から取得できる。

【０１２０】
ボリューム退避先決定プログラム252がステップS406で判定したストレージ制御部のCPU利用率は、ステップS402で選択したストレージ制御部IDに対応するストレージ制御部管理表257のCPU利用率2577から取得できる。
【０１２１】
ステップS404～S406での判定に用いた閾値は、ストレージシステム全体で固定の値としてもよいし、ストレージノード毎に設定できる値としてもよい。ストレージノード毎に設定可能とする場合は、ストレージノード管理表にそれぞれの閾値の列を追加しておき、ボリューム退避先決定プログラムがステップS404～S406を実行する際に取得する。
【０１２２】
物理チャンクをストレージ制御部が配置されたストレージノード以外のストレージノードにも保存する構成としている場合は、ステップS404を一律スキップして、ステップS405に進んでもよい。
【０１２３】
ボリューム退避先決定プログラム252は、当該ストレージ制御部ペアを構成する全てのストレージ制御部に対して、ステップS403～S406が完了しているか否かを判定する(S407)。完了していれば、当該ストレージ制御部ペアを退避先として決定する(S408)。完了していないストレージ制御部が残っていれば、ステップS402に戻り、ステップS403～S407を実行する。
【０１２４】
図１５の判定処理は一例であり、実際のストレージシステムの構造や特性などに応じて、任意の判定処理を行ってよい。例えば、ステップS403～ステップS406の判断は、退避先として求められるストレージシステムの特性に応じて、少なくとも一つのステップを実行することで判定されても良い。また、当該ボリュームを利用するホスト装置と、退避先候補のストレージ制御部が配置されたストレージノード間のネットワーク上の距離や、ネットワークを介した通信経路上に存在するネットワークスイッチの通信帯域利用率などによって判定してもよい。また、重複排除機能を有するストレージシステムにおいて、同一ストレージ制御部ペア内のボリューム間でのみ重複排除が適用可能といった制限がある場合、ボリュームの退避先によって、重複排除機能によるデータ削減量が変化する。この場合、退避先候補のストレージ制御部ペアにボリュームを退避させた際に重複排除機能によって削減されるデータ量を事前に見積もり、削減量が最も大きくなるストレージ制御部ペアに退避させるようにしてもよい。また、ボリューム管理表261(図７)で管理するIO量を利用し、退避するボリュームに対するIO量に応じて、ステップS405の通信帯域利用率、ステップS406のCPU利用率の閾値を変更するように構成しても良い。また、図５のストレージノード管理表256に、ストレージノードに搭載しているCPUの動作周波数や、搭載している各記憶装置の種別やIO性能などの情報を追加し、障害が発生したストレージノードに搭載されているCPUや記憶装置と同等以上のものが搭載されているストレージノード上に配置されたストレージ制御部を退避先として選択することで、ボリュームを退避した後も、退避前と同等のIO性能が得られるようにしてもよい。
【０１２５】
図１６は、ストレージ制御部ペア削除プログラム254の処理の一例である。このプログラム254は、障害回復プログラム250から指示されたストレージ制御部を削除する処理を行う。このプログラム254は、クラスタ制御部216のメモリ212に格納されており、当該クラスタ制御部が配置されたストレージノードのCPUによって実行される。このプログラム254は、マスタロールのクラスタ制御部が実行する障害回復プログラム250から起動される。そして、削除対象であり、ペアの片方を喪失しているストレージ制御部ペアの残存ストレージ制御部が動作しているストレージノードのクラスタ制御部が実行する。
【０１２６】
ストレージ制御部ペア削除プログラム254は、当該ストレージノードで動作している削除対象のストレージ制御部ペアの残存ストレージ制御部に対して、停止を指示する(S500)

。
【０１２７】
ストレージ制御部が停止すると、当該ストレージ制御部に割り当てられていたCPUコアやメモリなどの情報処理資源を解放する(S501)。なお、このプログラム254が実行された後に、解放した情報処理資源を利用して、新しいストレージ制御部を作成するため、実際には解放せずに、新しいストレージ制御部を作成する際に、解放する予定だった情報処理資源を再利用するようにしてもよい。次に、ストレージ制御部ペア削除プログラム254は、当該ストレージノードで動作するデータ冗長化部218に対して当該ストレージ制御部ペアに割り当てられていた論理チャンク及び、論理チャンクに対応付けられている物理チャンクの削除を指示する(S502)。データ冗長化部は、指示された論理チャンクと物理チャンクを削除し、論理チャンク管理表271と物理チャンク管理表272から関連するレコードを削除する。
【０１２８】
このように実施例１によれば、冗長度が低下したストレージ制御部の冗長度を回復させずに、ストレージ制御部が処理を担当していた複数のボリュームを正常なストレージ制御部に分散して退避させ、退避完了後に冗長度が低下したストレージ制御部自体を削除することで、冗長度の回復可能性を保証するための予約情報処理資源を不要とし、物理サーバの利用効率を向上させる。
【０１２９】
また、ストレージシステムを構成するストレージノードに障害が発生した場合、当該ストレージノードが管理している制御情報（各種管理表）と物理チャンクに格納されたデータを、予備リソースを確保することなく正常なストレージノードに引き継ぐことができる。正常なストレージノードに処理を引き継いだ後も、ストレージ制御部とボリュームの対応を管理し、ストレージ制御部が配置されるストレージノードにデータを保存（データのローカリティを確保）することで、ホスト装置からのIO要求に対し、高い応答性を維持することができる。つまり、ホスト装置にボリュームを提供するストレージ装置に対し、データのリード要求があった場合、他のストレージノードからデータを読み出す必要がない。
【０１３０】
また、予備リソースを確保する必要がないため、ストレージシステムの構築コストを低減し、仮想化技術を用いたSDSに求められる低コスト化に効率よく対応することができる。尚、予備リソースを確保した従来の技術と比べ、同程度の可用性を実現するため、必要なCPUコア数やメモリ容量を２／３程度削減することができる。これにより、ストレージシステム構築コストを２０パーセント低減することが可能となる。
【実施例２】
【０１３１】
以下、図１７を参照して、実施例２について詳述する。
【０１３２】
実施例１では、ストレージノード障害が発生した際の回復方法を示した。実施例２では、ストレージノード減設に適用する技術について説明する。つまり、ストレージシステムから、ストレージノードを取り除く処理をおこなうものである。以下、説明の容易化のため、ストレージ制御部グループを、一台のアクティブモードのストレージ制御部と一台のスタンバイモードのストレージ制御部からなるストレージ制御部ペアのケースで説明する。但し、ペアに限らず、3台以上のストレージ制御部から構成されるストレージ制御部グループの場合であっても、基本的には同一の処理を行うこととなる。ストレージ制御部ペア又はグループが複数のアクティブモードのストレージ制御部を含んでいる場合、減設処理のなかでアクティブモードのストレージ制御部はスタンバイモードのストレージ制御部と基本的には同様に扱ってもよい。
【０１３３】
図１７は、ストレージノード減設プログラム255の処理の一例である。このプログラム255は、ストレージシステムの管理者から指定されたストレージノードをストレージシス

テムから取り除くための処理を行う。このプログラム255は、管理端末110を介してストレージシステム200の管理者の指示によって起動され、マスタロールのクラスタ制御部が配置されたストレージノードのCPUによって実行される。マスタロールのクラスタ制御部218が定期的に行っている各ストレージノードの監視処理によって、当該ストレージノード障害の予兆を検出した場合に、このプログラム255を起動し、ストレージノード障害による冗長度低下を未然に防止するようにしてもよい。なお、図示していないが、減設対象のストレージノードに配置されたクラスタ制御部218がマスタロールの場合は、事前に他のストレージノードに配置されたワーカロールのクラスタ制御部の何れかをマスタロールに切り替える。

【０１３４】

このプログラム255は、新たにマスタロールに切り替わったクラスタ制御部が配置されたストレージノードのCPUによって実行される。マスタロールに切り替えるクラスタ制御部の選出方法は、ストレージノード障害によってマスタロールのクラスタ制御部が失われた場合と同様でもよい。

【０１３５】

また、減設対象のストレージノードに配置されたマスタロールのクラスタ制御部が、何らかの判定処理によって新しいマスタロールのクラスタ制御部を選出してもよい。なお、ストレージノード減設プログラム255は処理を開始する前に、減設対象のストレージノードを除いたストレージシステムを構成する全ストレージノードの記憶装置の総空き容量と、減設対象のストレージノードの記憶装置の総使用量とを比較して、ストレージノード減設中に記憶装置の容量が不足すると判断した場合は、処理を中止する。減設処理中に、ストレージノード減設中の容量不足を防止する処理を追加してもよい。

【０１３６】

また、ストレージノード減設プログラム255実行中に、ストレージシステムの管理者からボリュームの作成指示を受領した場合、ストレージノード減設によって、空き容量不足が発生する可能性があるか否かを判断し、発生する可能性が高いと判断した場合は、ボリューム作成を中止し、ストレージノード減設中の容量不足を防止するといった処理を追加してもよい。

【０１３７】

ストレージ減設プログラム255は、減設対象のストレージノードに配置されたストレージ制御部219を含むストレージ制御部ペア217を選択する(S600)。この処理はストレージ制御部管理表257を参照し、減設対象のストレージノードID2573で検索し、ストレージ制御ペアIDを特定することで実現する。この際、特定したストレージ制御部ペアへの新規ボリューム作成を抑止する処理を追加してもよい。この処理は、図６のストレージ制御部管理表257にボリューム作成抑止フラグの列を追加し、ストレージ制御部ペアID2572が特定したストレージ制御部ペアとなっているストレージ制御部に対してこのフラグをONにすることで実現できる。図１２及び図１３と同様に、図10に示した例によると減設対象のストレージノードに配置されたストレージ制御部を含むストレージ制御部ペアは２つ存在し、減設対象のストレージノードに配置されたストレージ制御部の一方はアクティブモードで動作しており、他方はスタンバイモードで動作している。

【０１３８】

物理チャンクをストレージ制御部217が配置されたストレージノード以外のストレージノードにも保存する構成としている場合や、物理チャンクを三重化している場合、ストレージノード間でRAIDやErasure Codingを適用して物理チャンクを冗長化している場合は、ステップS600で特定したストレージ制御部ペア以外のストレージ制御部ペアに対応付けられた論理チャンクを構成する物理チャンクが減設対象のストレージノードに保存されている可能性がある。そのため、ステップS601に進む前に、ストレージノード減設プログラム255は減設対象を除く全てのストレージノードのデータ冗長化部に対して、減設対象のストレージノードに配置されたストレージ制御部とは関係のないストレージ制御部ペアに対応付けられた論理チャンクを構成する物理チャンクが、減設対象のストレージノードに

配置されているか否かの確認と、他のストレージノードへの再配置を指示する。
【０１３９】
指示を受けたデータ冗長化部は、冗長化方式に応じて、物理チャンクの配置先の確認と再配置を実行する。例えば、物理チャンクを複製（二重化）する冗長化方式の場合、データ冗長化部は論理チャンク管理表271から減設対象のストレージノードに物理チャンクを保存している論理チャンクが存在するか否かを確認する。存在する場合は、当該論理チャンクに対応付けられているストレージ制御部ペアID2712が減設対象のストレージノードに配置されたストレージ制御部を含むストレージ制御部ペアか否かを確認する。喪失したストレージ制御部を含むストレージ制御部ペアでなければ、新しい物理チャンクを確保し、当該論理チャンクを構成する物理チャンクの何れかから、新たに確保した物理チャンクにデータをコピーする。コピー完了後、論理チャンク管理表271を更新し、論理チャンクを構成する物理チャンクのうち、減設対象ノードに保存された物理チャンクを、コピー先の物理チャンクに変更する。
【０１４０】
減設対象を除く全てのストレージノードで物理チャンクの配置先の確認と再配置が完了すると、ステップS601に進む。
【０１４１】
ストレージノード減設プログラム255は、特定したストレージ制御部ペアから処理対象となるストレージ制御部ペアを選択する(S601)。当該ストレージ制御部ペアのアクティブモードのストレージ制御部が減設対象ストレージノードに配置されているか否かを判定する(S602)。
【０１４２】
配置されていれば、図１２のステップS106～S108を実行し、当該ストレージ制御部ペアが担当している全ボリュームを退避させ、当該ストレージ制御部ペアを削除する(S603)。これにより、当該ストレージ制御部ペアのスタンバイモードが配置されていたストレージノードに、ストレージ制御部を１つ配置するのに必要な情報処理資源が解放されたことになる。アクティブモードのストレージ制御部が配置されていなければ、ステップS603をスキップし、ステップS604に進む。S602の処理は、ストレージ制御部管理表257を参照し、減設対象のストレージノードID2571とステップS601で選択したストレージ制御部ペアID2572のAND条件で検索して取得したストレージ制御部の動作モード2574を取得することで実
現する。
【０１４３】
ストレージノード減設プログラム255は、減設対象ストレージノードにアクティブモードのストレージ制御部が配置されている全ストレージ制御部ペアの削除が完了しているか否かを判定する(S604)。完了していれば、ステップS605に進み、完了していなければ、ステップS601に戻る。
【０１４４】
次にストレージノード減設プログラム255は、削除したストレージ制御部ペアを構成する減設対象ではないストレージノードのクラスタ制御部216に、減設対象ストレージノードに残っているストレージ制御部ペアのスタンバイモードのストレージ制御部のコピーを指示する(S605)。当該ストレージ制御部ペアのアクティブモードのストレージ制御部から再構築せずに、スタンバイモードのストレージ制御部からコピーするのは、アクティブモードのストレージ制御部が処理を担当しているボリュームのIO処理への影響を最小化するためである。
【０１４５】
指示を受けたクラスタ制御部216は、CPUコアやメモリなどの情報処理資源を確保し、記憶装置からストレージ制御部を構成するプログラムをメモリにロードして、ストレージ制御部を起動する。確保する情報処理資源は、削除によって解放されたものを用いる。起動後に減設対象ストレージノードに残っているストレージ制御部ペアのスタンバイモード

のストレージ制御部からボリューム管理表をコピーする。

【０１４６】

ストレージ制御部217のコピー完了後、ストレージノード減設プログラム255は、削除したストレージ制御部ペアを構成する減設対象ではないストレージノードのデータ冗長化部に対して、ステップS605の処理対象としたストレージ制御部ペアに対応付けられた論理チャンクを構成する物理チャンクの再配置を指示する(S606)。指示を受けたデータ冗長化部は、減設対象のストレージ制御部のデータ冗長化部の論理チャンク管理表271のレコードから当該ストレージ制御部ペアID2712に対応付けられたレコードをコピーする。

【０１４７】

その後、データ冗長化部218は、新しい物理チャンクを確保し、減設対象のストレージ制御部に保存されている物理チャンクのデータをコピーする。論理チャンク管理表を更新し、論理チャンクを構成する物理チャンクのうち、減設対象のストレージノードに保存された物理チャンクを、コピーした物理チャンクに変更する。

【０１４８】

物理チャンクの再配置が完了すると、ストレージノード減設プログラム255は、当該ストレージ制御部ペアのスタンバイモードのストレージ制御部をコピーしたストレージ制御部に切り替える指示を行い、ストレージ制御部管理表257を更新する(S607)。ステップS600において、特定したストレージ制御部ペアへの新規ボリューム作成を抑止する処理を追加した場合は、このステップに新規ボリューム作成の抑止を解除する処理を追加する。この処理は、図６のストレージ制御部管理表257において、ストレージ制御部ペアID2572が特定したストレージ制御部ペアとなっているストレージ制御部に対して、追加したボリューム作成抑止フラグをOFFにすることで実現できる。

【０１４９】

次に、ストレージノード減設プログラム255は減設対象のストレージノードで動作しているクラスタ制御部とデータ冗長化部に停止を指示し(S608)、ストレージノード管理表256から当該ストレージノードに関するレコードを削除する(S609)。以上の処理が完了すると、減設対象のストレージノードは、ストレージシステムから完全に切り離され、物理的にストレージノードを取り除いても問題ない状態となる。

【０１５０】

このように実施例２によれば、減設対象のストレージノードで動作するストレージ制御部が担当していた複数のボリュームを減設対象外のストレージ制御部に分散して退避させ、退避完了後に減設対象のストレージノードで動作するストレージ制御部を削除することで、ストレージノードを減設するための予約予備情報資源を不要とし、物理サーバの利用効率を向上させることができる。

【０１５１】

また、減設対象のストレージノードで動作するストレージ制御部が担当していたボリュームに対するIO要求の処理を、他のストレージ制御部が引き継いだ後も、ストレージ制御部が配置されるストレージノードにデータを保存（データのローカリティを確保）することで、ホスト装置からのIO要求に対し、高い応答性を維持することができる。つまり、ホスト装置にボリュームを提供するストレージ装置に対し、データのリード要求があった場合、他のストレージノードからデータを読み出す必要がない。

【０１５２】

また、ストレージノードを減設するための予備リソースを確保する必要がなく、ストレージシステムをスケールインすることが可能となるため、仮想化技術を用いたSDSに求められるスケーラビリティを向上させることが可能となる。

【０１５３】

以上、実施例１では、障害が発生したストレージノードのストレージ制御部を用いて構成されるストレージ制御部ペアからボリュームを退避させているが、実施例２ではストレージ制御部にかかるストレージノードを減設させる場合を示した。

【０１５４】

本発明は、さらに、ストレージノードは障害発生も減設もせずに稼働を続け、ボリュームの担当変更させるストレージ制御部ペアが担当ボリュームを残して稼働を続けてもよい。また、一つのストレージ制御部ペアから複数のストレージ制御部ペアにボリューム担当が分散して移動しているが、これに解らず、１つのストレージ制御部ペアから１つのストレージ制御部ペアへ、複数のストレージ制御部ペアから１つのストレージ制御部ペアへ、複数のストレージ制御部ペアから複数のストレージ制御部ペアへ、ボリューム担当を移動させてもよい。

【符号の説明】

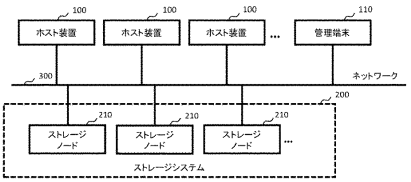
【０１５５】

100：ホスト装置、200：ストレージシステム、210：ストレージノード、211：CPU、212：メモリ、213：記憶装置、214：通信装置、215：、216：クラスタ制御部、217：ストレージ制御部ペア、218：データ冗長化部、219：ストレージ制御部、220：ボリューム、221：論理チャンク、222：物理チャンク、250：障害回復プログラム、251：ボリューム退避プログラム、252：ボリューム退避先決定プログラム、256：ストレージノード管理表、257：ストレージ制御部管理表、300：ネットワーク。

【図面】

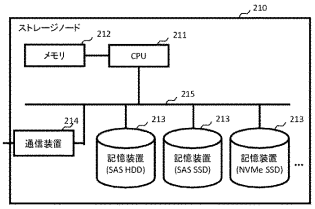
【図１】

図1



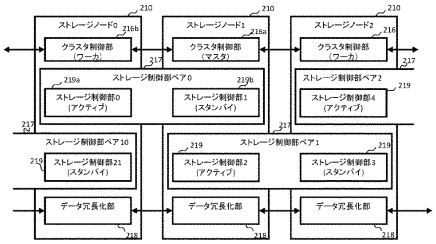
【図２】

図2



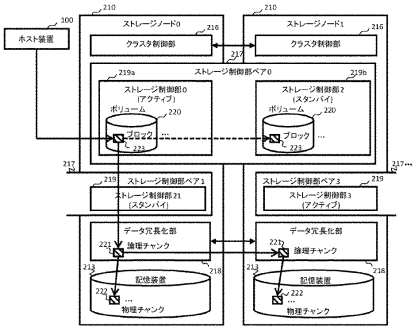
【図３】

図3



【図４】

図4



【図５】

図5

ストレージノード管理表 256

2561 ストレージノードID	2562 ロール	2563 動作状態	2564 CPUコア数	2565 メモリ量	2566 通信帯域利用率	2567 記憶装置総容量	2568 記憶装置総使用量
0	ワーカ	障害	20	96GB	NA	10TB	NA
1	マスタ	正常	20	96GB	70%	10TB	3TB
2	ワーカ	正常	20	96GB	40%	15TB	2TB
3	ワーカ	正常	20	96GB	30%	15TB	2TB
...	...	...	...	...	...	...	...

【図６】

図6

ストレージ制御部管理表 257

2571 ストレージ制御部ID	2572 ストレージ制御部ペアID	2573 ストレージノードID	2574 動作モード	2575 割り当てCPUコア数	2576 割り当てメモリ量	2577 CPU利用率	2578 メモリ利用量
0	0	0	アクティブ	8	30GB	NA	NA
1	0	1	スタンバイ	4	30GB	30%	28GB
2	1	1	アクティブ	8	30GB	50%	28GB
3	1	2	スタンバイ	4	30GB	40%	25GB
4	2	2	アクティブ	8	30GB	30%	24GB
5	2	3	スタンバイ	4	30GB	40%	25GB
6	3	3	アクティブ	8	30GB	30%	28GB
...	...	...	...	...	...	...	...
21	10	0	スタンバイ	4	30	NA	NA

【図７】

図7

ボリューム管理表　261

2611 ボリュームID	2612 容量	2613 使用容量	2614 ストレージ制御部ペアID	2615 ブロックID	2616 論理チャンクID	2617 IO量
0	200GB	50GB	0	0	0	10 kIOPS
				4	10	
				...	...	
1	300GB	80GB	0	0	1	30 kIOPS
				10	25	
				...	...	
...	...	...	...	...	...	

【図８】

図8

論理チャンク管理表　271

2711 論理チャンクID	2712 ストレージ制御部ペアID	2713 ストレージノードID（マスタ）	2714 物理チャンクID（マスタ）	2715 ストレージノードID（ミラー）	2716 物理チャンクID（ミラー）
0	0	0	0	1	1
1	0	0	2	1	5
...	...	...	...	...	...
0	1	1	1	2	0
1	1	1	3	2	4
...	...	...	...	...	...

【図９】

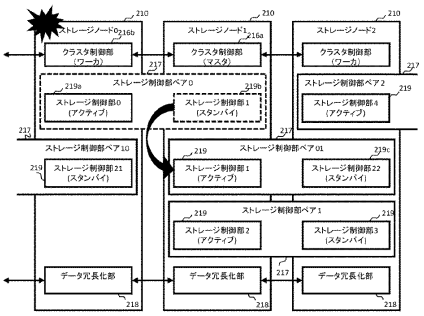
図9

物理チャンク管理表　272

2721 物理チャンクID	2722 記憶装置ID	2723 記憶装置内オフセット
0	0	0x0000
1	0	0x1000
2	1	0x0000
3	1	0x1000
...	...	...

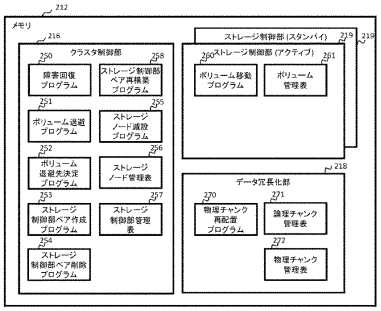
【図１０】

図10



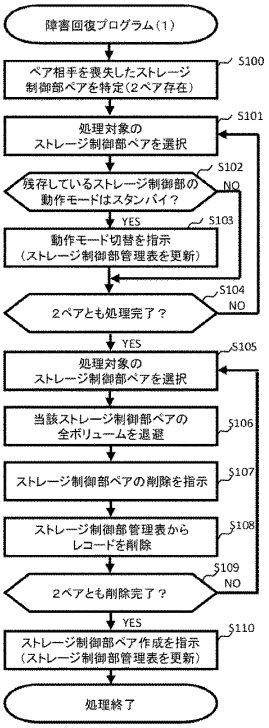
【図１１】

図11



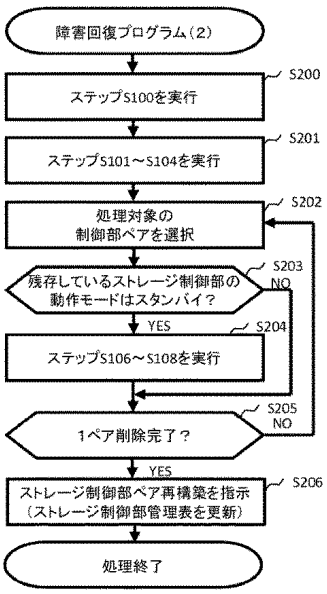
【図１２】

図12



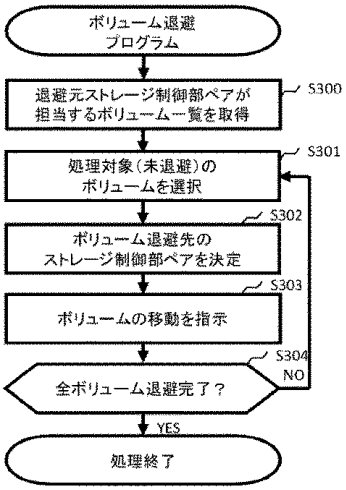
【図１３】

図13



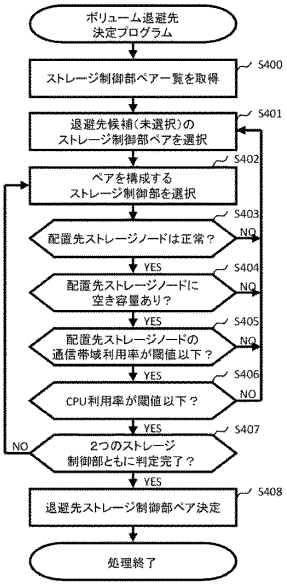
【図１４】

図14



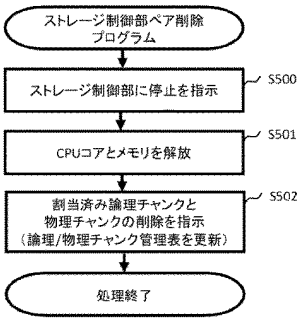
【図１５】

図15



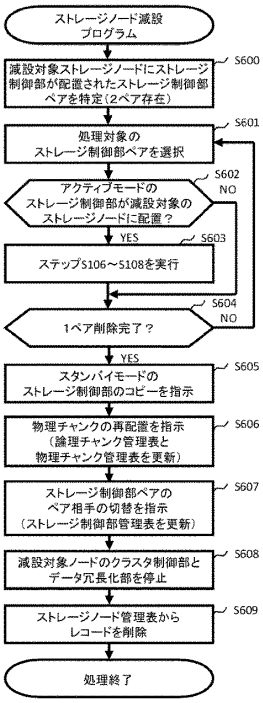
【図１６】

図16



【図１７】

図17



フロントページの続き

(51)国際特許分類

FI
G06F 11/20 689
G06F 16/188

(56)参考文献 国際公開第２０１５／０７２０２６（WO，Ａ１）
特開２００４－３４８７０１（ＪＰ，Ａ）
特開２０１０－０６６８６２（ＪＰ，Ａ）

(58)調査した分野 (Int.Cl.，ＤＢ名)

Ｇ０６Ｆ ３／０６
Ｇ０６Ｆ １３／１０
Ｇ０６Ｆ １１／２０
Ｇ０６Ｆ １６／１８８