(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5412902号
(P5412902)
(45)発行日　平成26年2月12日(2014.2.12)　(24)登録日　平成25年11月22日(2013.11.22)

(51)Int.Cl.　F I
G06F 3/06 (2006.01)　G06F 3/06
G06F 3/06 301Z

請求項の数 7　(全 18 頁)

(21)出願番号　特願2009-63903(P2009-63903)
(22)出願日　平成21年3月17日(2009.3.17)
(65)公開番号　特開2010-218193(P2010-218193A)
(43)公開日　平成22年9月30日(2010.9.30)
審査請求日　平成23年12月8日(2011.12.8)

(73)特許権者　000004237
日本電気株式会社
東京都港区芝五丁目７番１号
(74)代理人　100124811
弁理士　馬場　資博
(74)代理人　100088959
弁理士　境　廣巳
(74)代理人　100131428
弁理士　若山　剛
(72)発明者　鈴木　秀明
東京都港区芝五丁目７番１号　日本電気株式会社内

審査官　古河　雅輝

最終頁に続く

(54)【発明の名称】ストレージシステム

(57)【特許請求の範囲】
【請求項１】
ディスク装置毎に当該ディスク装置に対する単位時間あたりの稼働状況を計測して、稼働した前記ディスク装置を稼働ディスクと判定し、稼働していない前記ディスク装置を非稼働ディスクと判定すると共に、前記ディスク装置内の記憶領域が複数に分割された各区画における単位時間当たりの負荷を計測して、計測結果を記憶する性能監視部と、
前記性能監視部による計測結果に基づいて、稼働ディスクと判定された前記ディスク装置に形成された前記区画のうち、負荷が他の前記区画と比較して高い前記区画における負荷を軽減するよう、負荷が他の前記区画と比較して高い前記区画に格納されたデータの一部又は全部を、他の前記ディスク装置に格納して、データの再配置を行い、その後、前記性能監視部による計測結果に基づいて、稼働ディスクと判定された前記ディスク装置に形成された前記区画のうち、負荷が他の前記区画と比較して低い前記区画に格納されているデータを、稼働ディスクと判定された他の前記ディスク装置に格納するデータ再配置部と、
を備えたストレージシステム。
【請求項２】
請求項１に記載のストレージシステムであって、
前記データ再配置部は、前記性能監視部による計測結果に基づいて、負荷が低い順から予め設定された順番までの前記区画に格納されているデータを、稼働ディスクと判定された他の前記ディスク装置に格納する、

ストレージシステム。
【請求項３】
請求項１又は２に記載のストレージシステムであって、
前記データ再配置部は、非稼働ディスクと判定された前記ディスク装置を含めて、前記データの再配置を行う、
ストレージシステム。
【請求項４】
請求項１乃至３のいずれか一項に記載のストレージシステムであって、
前記データ再配置部は、前記性能監視部による計測結果に基づいて、負荷が高い順から予め設定された順番までの前記区画における負荷を軽減するよう、前記各区画に格納されたデータの再配置を行う、
ストレージシステム。
【請求項５】
請求項１乃至４のいずれか一項に記載のストレージシステムであって、
前記性能監視部は、前記各区画における単位時間あたりの入力開始から当該入力に対する出力完了までに要する時間であるターンアラウンドタイムの累積時間を、前記各区画における前記負荷として計測する、
ストレージシステム。
【請求項６】
コンピュータに、
ディスク装置毎に当該ディスク装置に対する単位時間あたりの稼働状況を計測して、稼働した前記ディスク装置を稼働ディスクと判定し、稼働していない前記ディスク装置を非稼働ディスクと判定すると共に、前記ディスク装置内の記憶領域が複数に分割された各区画における単位時間当たりの負荷を計測して、計測結果を記憶する性能監視部と、
前記性能監視部による計測結果に基づいて、稼働ディスクと判定された前記ディスク装置に形成された前記区画のうち、負荷が他の前記区画と比較して高い前記区画における負荷を軽減するよう、負荷が他の前記区画と比較して高い前記区画に格納されたデータの一部又は全部を、他の前記ディスク装置に格納して、データの再配置を行い、その後、前記性能監視部による計測結果に基づいて、稼働ディスクと判定された前記ディスク装置に形成された前記区画のうち、負荷が他の前記区画と比較して低い前記区画に格納されているデータを、稼働ディスクと判定された他の前記ディスク装置に格納するデータ再配置部と、
を実現させるためのプログラム。
【請求項７】
ディスク装置毎に当該ディスク装置に対する単位時間あたりの稼働状況を計測して、稼働した前記ディスク装置を稼働ディスクと判定し、稼働していない前記ディスク装置を非稼働ディスクと判定すると共に、前記ディスク装置内の記憶領域が複数に分割された各区画に単位時間当たりの負荷を計測して、計測結果を記憶して、性能監視を行い、
前記計測結果に基づいて、稼働ディスクと判定された前記ディスク装置に形成された前記区画のうち、負荷が他の前記区画と比較して高い前記区画における負荷を軽減するよう、負荷が他の前記区画と比較して高い前記区画に格納されたデータの一部又は全部を、他の前記ディスク装置に格納し、その後、前記計測結果に基づいて、稼働ディスクと判定された前記ディスク装置に形成された前記区画のうち、負荷が他の前記区画と比較して低い前記区画に格納されているデータを、稼働ディスクと判定された他の前記ディスク装置に格納して、データの再配置を行う、
データ格納方法。

【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にかかり、特に、複数の記憶装置を装備したストレージシステムに関する。

【背景技術】

【０００２】

近年、データ記憶容量、信頼性の向上を図るべく、多数のディスク装置を備えたストレージシステムが開発されている。例えば、ＮＡＳ（Network Attached Storage）装置や、グリッドストレージシステムといったものがある。

【０００３】

ここで、仮に、数百台のクライアントが起動時の特定処理で読み込むファイルが、ストレージシステム上に格納されているとする。すると、このファイルには、数百台のクライアントが同時に起動する際にアクセスするため、当該ファイルに負荷が集中し、ホットスポットとなる。ところが、このようなホットスポットが出現する時間は短く、１日平均のディスクビジー率でみると、システム全体のわずかなリソースしか占有しない。また、ストレージシステムでは、一般的に、容量がフルに近づくと性能低下が起きやすくなるという問題がある。これに対応すべく、また、突発的な容量増加に耐えられるよう、７０％以下など余裕をもって運用されるのが普通である。なお、ＩＬＭ（情報ライフサイクル管理：Information Lifecycle Management）の考え方では、大部分のデータはアクセス頻度が低いデータ（例えば１日～数ヶ月に１回以下）だと言われている。

【０００４】

しかしながら、上述したように、ストレージシステム全体の稼働率（負荷・容量）が低い場合であっても、多数のディスク装置を備えているストレージシステムでは、これら多数のディスク装置を絶えず駆動する必要がある。その結果、ストレージシステムの消費電力が下がりにくい、という問題がある。

【０００５】

一方で、特許文献１では、ＮＡＳ装置などのストレージシステムにおいて、省エネルギー化を図ることを目的とした技術を開示している。具体的には、ストレージシステムを構成する仮想ストレージ毎の負荷を計測し、負荷が高い仮想ストレージの処理を、他のストレージ処理部にて分割し、負荷が低い仮想ストレージの処理を他のストレージ処理部にて統合する。これにより、仮想ストレージシステムを制御しているストレージ処理部への電力供給を制御している。

【先行技術文献】

【特許文献】

【０００６】

【特許文献１】特開２００３－２９６１５３号公報

【発明の概要】

【発明が解決しようとする課題】

【０００７】

ところが、上述した特許文献１に開示の技術では、仮想ストレージ単位で負荷を計測しており、当該仮想ストレージを制御するストレージ処理部に対する電力供給を制御しているだけであるため、物理的なディスク装置に対する電力供給を適切に制御していない。従って、ディスク装置の消費電力の低下を図ることができず、依然としてシステム全体の消費電力が下がりにくい、という問題が生じる。

【０００８】

このため、本発明の目的は、上述した課題である、消費電力の低下が困難であることを解決することができるストレージシステムを提供することにある。

【課題を解決するための手段】

【０００９】

かかる目的を達成するため本発明の一形態であるストレージシステムは、

ディスク装置毎に当該ディスク装置に対する単位時間あたりの稼働状況を計測して、稼

働した上記ディスク装置を稼働ディスクと判定し、稼働していない上記ディスク装置を非稼働ディスクと判定すると共に、単位時間当たりの上記ディスク装置における負荷を計測して、計測結果を記憶する性能監視部と、
上記性能監視部による計測結果に基づいて、稼働ディスクと判定された上記ディスク装置に格納されているデータを、当該ディスク装置における負荷よりも高い負荷の他の上記ディスク装置に格納するデータ再配置部と、
を備える。
【００１０】
また、本発明の他の形態であるプログラムは、
コンピュータに、
ディスク装置毎に当該ディスク装置に対する単位時間あたりの稼働状況を計測して、稼働した上記ディスク装置を稼働ディスクと判定し、稼働していない上記ディスク装置を非稼働ディスクと判定すると共に、単位時間当たりの上記ディスク装置における負荷を計測して、計測結果を記憶する性能監視部と、
上記性能監視部による計測結果に基づいて、稼働ディスクと判定された上記ディスク装置に格納されているデータを、当該ディスク装置における負荷よりも高い負荷の他の上記ディスク装置に格納するデータ再配置部と、
を実現させるためのプログラムである。
【００１１】
また、本発明の他の形態であるデータ格納方法は、
ディスク装置毎に当該ディスク装置に対する単位時間あたりの稼働状況を計測して、稼働した上記ディスク装置を稼働ディスクと判定し、稼働していない上記ディスク装置を非稼働ディスクと判定すると共に、単位時間当たりの上記ディスク装置における負荷を計測して、計測結果を記憶して、性能監視を行い、
上記計測結果に基づいて、稼働ディスクと判定された上記ディスク装置に格納されているデータを、当該ディスク装置における負荷よりも高い負荷の他の上記ディスク装置に格納して、データの再配置を行う、
という構成を採る。
【発明の効果】
【００１２】
本発明は、以上のように構成されることにより、ストレージシステム全体のさらなる省電力化を図ることができる。
【図面の簡単な説明】
【００１３】
【図１】実施形態１におけるストレージシステムの構成を示す機能ブロック図である。
【図２】図１に開示したストレージシステムにおけるディスク装置の負荷の計測の一例を示す図である。
【図３】図１に開示した再配置候補リスト記憶部に記憶されているデータの一例を示す図である。
【図４】図１に開示したストレージシステムに設定されているデータの一例を示す図である。
【図５】図１に開示したストレージシステムの動作を示すフローチャートである。
【図６】図１に開示したストレージシステムの動作を示すフローチャートである。
【図７】図１に開示したストレージシステムにおけるデータ再配置時の様子を示す図である。
【図８】図１に開示したストレージシステムの動作を示すフローチャートである。
【図９】図１に開示したストレージシステムにおけるデータ再配置時の様子を示す図である。
【図１０】図１に開示したストレージシステムの動作を示すフローチャートである。
【図１１】図１に開示したストレージシステムにおけるデータ再配置時の様子を示す図で

ある。
【図１２】実施形態１におけるストレージシステムの構成を示す機能ブロック図である。
【発明を実施するための形態】
【００１４】
＜実施形態１＞
本発明の第１の実施形態を、図１乃至図１１を参照して説明する。図１は、本実施形態におけるストレージシステムの構成を示す機能ブロック図である。図２は、ディスク装置の負荷の計測の一例を示す図である。図３乃至図４は、ストレージシステムに記憶されているデータの一例を示す図である。図５乃至図１１は、ストレージシステムの動作を示す図である。
【００１５】
［構成］
本実施形態におけるストレージシステム１０は、図１に示すように、複数のディスク装置６，７，・・・，ｎを備えており、例えば、ＮＡＳ装置やグリッドストレージシステムによって実現されている。なお、図１では、ストレージシステムが、ディスク装置６等と一体的に構成されているよう図示しているが、本発明におけるストレージシステムは、ディスク装置６等と分離されており、ディスク装置６等に対するデータの記録再生制御を行うコンピュータにて構成されていてもよい。
【００１６】
そして、ストレージシステム１０は、装備された演算装置にプログラムが組み込まれることによって構築された、ファイル管理部１と、性能監視部２と、再配置候補リスト記憶部３と、ファイル再配置制御部４と、ファイル群選別制御部５と、を備えている。なお、上記各部１～５による処理に付随して記憶することが必要なデータを記憶するフラッシュメモリなどの記憶装置も備えている。以下、各構成について詳述する。
【００１７】
上記ファイル管理部１は、ディスク装置６等内に格納されたファイルデータの格納位置、つまり、どのディスクのどの区画に格納されているかということを管理している。なお、ディスク装置６等は、図１の符号６１，７１，ｎ１等に示すように、その記憶容量が複数の区画に分割されている。そして、各区画あるいは区画間をまたがって、ファイルデータ６２，７２，ｎ２等が格納されている。例えば、一つのディスク装置の記憶容量は、１０００程度の区画に分割されている。
【００１８】
また、上記性能監視部２は、上述した各区画をユニークな区画番号により識別する。そして、性能監視部２は、一定期間（例えば、１０分）における区画毎のＩ／Ｏ（Input/Output）度数分布と、ディスク全体のビジー率を観測する。具体的に、性能監視部２は、該当するディスク装置の単位時間当たりの稼働率、つまり、一定期間あたりの稼働している時間の割合を、上記ディスクビジー率として算出する。そして、ディスクビジー率が「０」ではない、つまり、一定期間（単に時間）内に稼働したディスク装置を、稼働ディスクとして判定し、ディスクビジー率が「０」、つまり、一定期間（単位時間）内に稼働しなかったディスク装置を、非稼働ディスクとして判定する。さらに、性能監視部２あるいはストレージシステム１０に装備された他の機能により、非稼働ディスクとして判定したディスク装置は停止される。つまり、非稼働ディスクであるディスク装置への電源供給は停止される。なお、非稼働ディスクは未使用という意味ではない。上述した単位時間あたりに稼働していないだけであって、データが格納されている可能性はある。
【００１９】
なお、上述したように計測対象となるディスク装置は、ディスク装置単体であることに限定されず、複数のディスク装置であってもよい。つまり、グループ化された複数のディスク装置を１つのディスク装置として、そのディスクビジー率を計測してもよい。
【００２０】
また、性能監視部２は、上述したように計測したディスクビジー率と、各区画毎のＩ／

Ｏ度数分布とから、各区画毎のビジー率を算出して、この値を各区画の負荷とする。このとき、各区画に対するRead回数とWrite回数とに基づいて、各区画に対するRead処理時におけるReadビジー率、Write処理時におけるWriteビジー率、をそれぞれ算出する。

【００２１】

ここで、図２を参照して、所定のディスク装置６の各区画におけるビジー率算出の一例を説明する。まず、図２（Ａ）に示すように、ディスク装置６が区画＃１～＃７に分割されているとする。そして、このときのディスク装置６全体のビジー率が８０％であると計測され、また、各区画＃１～＃７に対するRead回数とWrite回数とがそれぞれ図２（Ｂ）に示す回数であったとする。つまり、ディスク装置６を７つの区画＃１～＃７に分けた時の各区画におけるＩ／Ｏ度数が、図２（Ｂ）の上側の図に示すように、区画＃１からReadが、４回、１回、８回、０回、０回、０回、４回であり、図２（Ｂ）の下側の図に示すように、区画＃１からWriteが、０回、２回、０回、０回、２回、０回、０回であるとする。この場合に、ディスクビジー率（８０％）を、各区画のＩ／Ｏ度数分布で比例分配することにより、各区画のビジー率を算出する。すると、上記の場合には、全Ｉ／Ｏ数を合計すると２０回であるため、Ｉ／Ｏ一回で全体のビジー率８０％うちの４％に貢献することとなる。従って、図２（Ｃ）に示すように、各区画のReadビジー率は、それぞれ１２％、４％、３２％、０％、０％、０％、１６％となり、各区画のWriteビジー率は、それぞれ０％、８％、０％、０％、８％、０％、０％となる。このようにして、各区画のReadビジー率、Writeビジー率を算出し、当該各区画の負荷とする。

【００２２】

なお、上述した各区画のビジー率の算出は、ディスクビジー率が「０」ではなく稼働ディスクと判定されたディスク装置に対して行う。そして、算出した区画ビジー率が「０」でなければ、性能監視部２は、当該算出した区画ビジー率（Readビジー率、Writeビジー率）を再配置候補リスト記憶部３に入力して記憶する。

【００２３】

そして、上記再配置候補リスト記憶部３は、上述したように性能監視部２から受け取った各区画のビジー率、つまり、Readビジー率、Writeビジー率を、それぞれビジー率が高い順と低い順にソートして、それらの優先度を管理する。具体的には、図３に示すように、Read・Writeの二種類の処理において、それぞれ高ビジー率・低ビジー率の二種類の、計４種類の優先度リストを管理する。そして、各優先度リストにおいては、ストレージシステムにある全区画の中から、その値が高い順、及び、低い順に、予め設定された個数（例えば、５つ）ずつ保持している。このリストに登録される区画は、後述のファイル再配置制御部４にてファイルが再配置されるか、さらにより高い、もしくは、低い区画がリストに登録され、各リストの上位に位置する設定個数内から追い出されるまで、当該リスト中に存在し続ける。ただし、一つのリストに同じ区画番号は２度出現しない。例えば、リストをヒープ構造でもち、定期的に各エリアのビジー率をリストの最下位の値と比較して、リストの圏内であれば最下位の値と交換後にソートすることで、このような優先度リストを実現できる。

【００２４】

ここで、再配置候補リスト記憶部３における入力された各区画におけるRead・Writeビジー率のリストのソート方法について説明する。まず、すでにリストの最大個数を越えて区画が記憶されている場合には、新たに入力された区画のビジー率と、記憶しているリストの最下位の区画のビジー率と比較し、これよりも高い値（区画が高い順に並んでいる場合）あるいは低い値（区画が低い順に並んでいる場合）であれば、区画番号とビジー率の値を交換する。つまり、区画番号をリストに追加する。その後、交換が発生したリストをソートし、重複する区画が既に登録されていれば、その区画の登録が古い方を削除する。

【００２５】

例えば、最大リスト個数が「３」の高Readビジー率リストがあり、現在空である場合、前述のディスクの区画Readビジー率１２％（＃１）、４％（＃２）、３２％（＃３）、１６％（＃７）を挿入するケースを考える。まず、リストが空なので、＃１、＃２、＃３が

リストに登録されソートされる。リストの状況は、＃３→＃１→＃２の順番となる。次に＃７（１６％）を挿入すると、リストは既に最大個数（３）なので、最下位の区画＃２（４％）と追加区画＃７（１６％）のビジー率を比較する。ここでは、＃７のビジー率が高いため、＃２と＃７とを交換する。そして、再度ソートを行い、新たなリスト＃３→＃７→＃１を得る。なお、＃７の前後に他に＃７はないため、挿入した区画＃７は重複しておらず、重複削除の必要は無い。

【００２６】

なお、区画ビジー率のリストは、図３に示すように、ReadとWriteを別々に形成されていることに限定されない。例えば、Readだけ、あるいは、Writeだけに対応するリストを記憶していてもよく、あるいは、Read/Writeの区別無しで、上述した性能監視部２で区画毎のビジー率を計測して、再配置候補リスト記憶部３で管理してもよい。

【００２７】

さらに、上記では、各区画の負荷を表す値として、区画毎のビジー率を利用したが、他の計測した値や算出した値を、各区画の負荷とし、これに基づいて、図３に示すような各区画のリストを管理してもよい。例えば、上記性能監視部２は、各区画における単位時間あたりの入力開始から当該入力に対する出力完了までに要する時間であるターンアラウンドタイム（ＴＡＴ：Turn Around Time）を計測し、これらの累積時間を、各区画における負荷としてもよい。そして、この負荷の値が高い順や低い順に、ファイルの再配置を行う区画を選択するリストの並び替えを行ってもよい。このように、区画の負荷を表す値として、ＴＡＴを用いた場合には、当該ＴＡＴの値は故障気味のディスク装置で高くなることがある。従って、後述するように、ＴＡＴの値に基づいてファイルの再配置を行う区画を選択することで、故障気味のディスクを選択でき、当該ディスクからデータを退避するなどの用途に利用することができる。

【００２８】

また、上記ファイル郡選別制御部４は、ある区画に含まれるファイルを特定する機能を有する。具体的には、ファイルシステムのレイアウト情報を利用し、区画番号から、その区画を利用しているファイル群やメタデータ群を選別できる。このような逆マップ情報は、昨今のファイルシステムでは内部情報として普通に存在している。

【００２９】

また、上記ファイル再配置制御部５は、定期的に（例えば１日１回）、上述した再配置候補リスト記憶部３に記憶されたリストに基づいて、各区画に格納されているファイルの再配置を計画して、実行する機能を有する。具体的には、大きく分けて、以下の３つの処理Ａ，Ｂ，Ｃを実行する。

【００３０】

（処理Ａ）

まず、再配置候補リスト記憶部３に基づいて、ビジー率の高い区画（一番高い区画、あるいは、高い順から所定の順番までの区画）を選択し、現在稼働中のディスクに負荷を分散した場合の効果を予測して計算する。具体的な負荷の分散方法としては、例えば、ファイル群の一部を別ディスクに移動する、高いWrite要求のある単一ファイルをRAID5（Redundant Arrays of Inexpensive Disks 5）などの分散方式で複数ディスクに分散する、高いRead要求のあるファイルの複製を複数ディスクに作成する、などの再配置処理が考えられる。また、予測計算では、高いRead要求の区画のみのファイルを分散、高いWrite要求の区画のみのファイルを分散、高いRead/Write両方の区画のファイルを分散した場合のそれぞれのケースで、予め設定された評価関数により効果を数値化し、この数値に基づいて適したファイル再配置処理を決定する。

【００３１】

そして、データ再配置後の予測した評価等の値（例えば、評価関数による評価値や各区画のビジー率）が、予め設定された範囲内となる場合には、ファイルの再配置処理を実行する。その後、再配置に伴い、ファイル管理部１の情報を適切に変更し、再配置候補リスト記憶部３から当該区画を削除して、処理を終了する。

【００３２】
（処理Ｂ）
上記処理Ａで、ファイル再配置後の負荷が軽減されない場合には、この処理Ｂを行う。この処理Ｂでは、まず、再配置候補リスト記憶部３から、ビジー率の高い区画を選択し、現在、非稼働であるディスクを含めて負荷を分散した場合の効果を、ビジー率と消費電力に関する評価関数ＥＶ（Ｎ）を用いて、予測計算する。なお、この評価関数ＥＶ（Ｎ）におけるＮは非稼働ディスクを含めた使用するディスク台数であるため、すでに現在の稼働ディスク数が下限値となる。
【００３３】
そして、上記評価関数（Ｎ）は、ディスク台数が増えることでビジー率が下がる効果と、消費電力が増える効果がバランスした最適のディスク稼働台数が得られるものを使う。例えば、ビジー率とディスク数の定数倍を足す、図４に示すような評価関数ＥＶ（Ｎ）が用いられ、この評価関数に基づく評価値は、値が低い方が評価がよいこととなる。従って、図４の例では、Ｎｍａｘが最も効率のよいディスク台数を示している。なお、性能や消費電力が予め決めた閾値を外れる場合は、ＥＶ（Ｎ）を０とすることで、システムの動作をコントロールできる。
【００３４】
そして、上述した評価関数（Ｎ）を利用した評価の結果、一台以上の非稼働ディスクを利用してファイルの再配置を行うべきと判断された場合、ファイルの再配置を行う。その後、再配置に伴い、ファイル管理部１の情報を適切に変更し、再配置候補リスト記憶部３から当該区画を削除して、処理を終了する。
【００３５】
なお、上述した処理Ａ、Ｂによる高ビジー率の区画に対して、負荷を軽減するファイルの再配置処理は、上述した処理に限定されない。
【００３６】
（処理Ｃ）
上記処理Ａ，Ｂを実行した後に、この処理Ｃを実行する。この処理Ｃでは、再配置候補リスト記憶部３から、ビジー率の低い区画（一番低い区画、あるいは、低い順から所定の順番までの区画）を選択し、その区画上のファイル群を別の稼働ディスクに移動した場合の効果を表す評価の値を、予測計算する。そして、計算後のビジー率などの値が予め決められた範囲に入る場合には、ファイルの再配置を行い、再配置候補リスト記憶部３から当該区画を削除する。このとき、選択された区画のファイルの再配置は、稼働ディスクであって、選択された区画のビジー率よりも高いビジー率の区画に移動することで実行し、特に、ファイルが移動された後の区画が形成されたディスク装置を停止できる可能性が高いディスク装置であるとよい。
【００３７】
なお、上述したストレージシステムは、図１に示すように、各処理を行う機能が１台のコンピュータに装備されている場合を説明したが、各機能が複数台のコンピュータに装備されていてもよい。例えば、上記ファイル管理部１が独立したサーバコンピュータに存在し、クライアントからファイルの物理位置を隠蔽するメタデータサーバの役割を果たす、分散ファイルシステムとして構成されていてもよい。
【００３８】
また、上述したストレージシステムでは、各ディスク装置を、独立したファイルシステムとし、ファイル管理部１が全体を統括構成する構成も考えられる。例えば、画像検索システムを考えた場合に、ファイル管理部１を検索処理部と考え、アクセスパターンに最適な画像データのディスク配置を行う事ができる。
【００３９】
［動作］
次に、上述したストレージシステムの動作を説明する。はじめに、性能監視部２にて一定時間ごと（例えば、１０分毎）に実行される動作について、図５のフローチャートを参

照して説明する。

【００４０】

まず、各ディスク装置のビジー率と、各区画のRead・WriteのＩ／Ｏ度数分布を取得して記録する（ステップＳ１）。続いて、ディスク装置のビジー率が「０」ではなく、稼働ディスクと判定された場合には（ステップＳ２でＹｅｓ）、そのディスク装置に形成された区画毎のビジー率を、ディスクのビジー率を各区画のＩ／Ｏ度数分布で比例分配することにより算出する（ステップＳ３）。そして、算出した区画ビジー率が、「０」でなければ、再配置候補リスト記憶部３に入力する（ステップＳ４）。このとき、再配置候補リスト記憶部３では、上述したように、各リストを高い順、あるいは、低い順に、ソートする。以上の処理が、一定時間ごとに実行される。

【００４１】

続いて、ファイル再配置制御部４にて定期的に（例えば１日１回）実行される動作について、図６乃至図１１に示すフローチャート及びファイル再配置の様子を示す図を参照して、説明する。

【００４２】

はじめに、上述した処理Ａを実行するが、このときの動作を、図６及び図７を参照して説明する。まず、再配置候補リスト記憶部３から、ビジー率の高い区画を選択する（ステップＳ１１）。このとき、例えば、図７（Ａ）の符号６に示すディスク装置内の区画＃１１が選択されたとする。そして、この区画における負荷を、現在稼働中の他のディスク装置（区画）に分散した場合の効果を、ビジー率にて予測計算する（ステップＳ１２）。

【００４３】

ここでは、図７（Ａ）の矢印に示すように、選択された区画＃１１内のファイルを、稼働中の他のディスク装置７内の区画＃１２に移動して、負荷を分散する場合を考える。このような負荷分散処理を行った場合の効果を、予め記憶された評価関数などの評価情報に基づいて評価値として算出する。そして、この評価値が、再配置により効果が上がると判定できる予め設定された基準を満たすか否かを判断する（ステップＳ１３）。基準を満たす場合には（ステップＳ１３でＹｅｓ）、予測した負荷分散処理を実際に実行する（ステップＳ１４）。すると、図７（Ｂ）に示すように、区画＃１１内に格納された１つのファイルが、稼働中の他のディスク装置７内の区画＃１２に移動して格納された状態となる。

【００４４】

その後は、ファイルの再配置に伴い、ファイル管理部１の情報を適切に変更する。また、再配置候補リスト記憶部３内の高ビジー率リストから当該区画を削除して（ステップＳ１５）、処理を終了する。

【００４５】

ここで、上述したステップＳ１２で予測した分散処理の効果が基準を満たさなかった場合には（ステップＳ１３でＮｏ）、上記処理Ｂを実行する。このときの動作を、図８及び図９を参照して説明する。

【００４６】

まず、再配置候補リスト記憶部３から、ビジー率の高い区画を選択する（ステップＳ２１）。このとき、例えば、図９（Ａ）の符号６に示すディスク装置内の区画＃１１が選択されたとする。そして、この区画における負荷を、上記処理Ａとは異なり、今度は、現在、非稼働であるディスク装置も含めて、負荷を分散した場合の効果を予測計算する（ステップＳ２２）。

【００４７】

ここでは、図９（Ａ）の矢印に示すように、選択された区画＃１１内のファイルを、非稼働である他のディスク装置９内の区画＃１４に移動して、負荷を分散する場合を考える。このような負荷分散処理を行った場合の効果を、予め記憶された評価関数などの評価情報に基づいて評価値として算出する。そして、この評価値が、再配置により効果が上がると判定できる予め設定された基準を満たすか否かを判断する（ステップＳ２３）。基準を満たす場合には（ステップＳ２３でＹｅｓ）、予測した負荷分散処理を実際に実行する（

ステップＳ２４）。すると、図９（Ｂ）に示すように、区画＃１１内に格納された１つのファイルが、非稼働の他のディスク装置９内の区画＃１４に移動して格納された状態となり、また、非稼働であったディスク装置９は稼働状態となる。

【００４８】

その後は、ファイルの再配置に伴い、ファイル管理部１の情報を適切に変更する。また、再配置候補リスト記憶部３内の高ビジー率リストから当該区画を削除して（ステップＳ２５）、処理を終了する。

【００４９】

以上のように、本実施形態におけるストレージシステムでは、負荷の高い区画に格納されたファイルを分散したり、あるいは複製して、他のディスク装置に格納するなどして、データの再配置を行っている。これにより、該当する区画が形成されたディスク装置の負荷が軽減されることとなる。その結果、ストレージシステム全体において電力が効率的に利用され、省電力化を図ることができる。特に、非稼働ディスクを増やして負荷分散するさいには、実際の分散処理の前に評価関数などを用いて性能予測を行うことで、性能及び省電力化のバランスを維持することができる。

【００５０】

続いて、上記処理Ａあるいは処理Ｂが実行された後に、上記処理Ｃを実行する。このときの動作を、図１０及び図１１を参照して説明する。

【００５１】

まず、再配置候補リスト記憶部３から、ビジー率の低い区画を選択する（ステップＳ３１）。このとき、例えば、図１１（Ａ）の符号８に示すディスク装置内の区画＃１３が選択されたとする。そして、この区画に配置されているファイルを、稼働中の他のディスク装置（区画）に移動する場合の効果を予測計算する（ステップＳ３２）。

【００５２】

ここでは、図１１（Ａ）の矢印に示すように、選択された区画＃１３内のファイルを、稼働中の他のディスク装置７内の区画＃１２に移動する場合を考える。このような負荷分散処理を行った場合の効果を、予め記憶された評価関数などの評価情報に基づいて評価値として算出して、この評価値が、再配置により効果が上がると判定できる予め設定された基準を満たすか否かを判断する（ステップＳ３３）。基準を満たす場合には（ステップＳ３３でＹｅｓ）、予測したファイルの移動を実際に実行する（ステップＳ３４）。すると、図１１（Ｂ）に示すように、区画＃１３内に格納されたファイルが、稼働中の他のディスク装置７内の区画＃１２に移動して格納された状態となる。

【００５３】

その後は、ファイルの再配置に伴い、ファイル管理部１の情報を適切に変更する。また、再配置候補リスト記憶部３内の低ビジー率リストから当該区画を削除して（ステップＳ３５）、処理を終了する。

【００５４】

そして、その後は、上述したようにもともと負荷の低かった区画＃１３内のファイルが他の区画に移動されることから、当該区画＃１３にはその後の負荷がさらに低減される。従って、後に性能監視部２にてディスク装置のビジー率が計測されたときに、区画＃１３が位置するディスク装置８のビジー率が「０」であった場合には、当該ディスク装置８を非稼働ディスクと判定でき、その後は、電源供給が停止される。

【００５５】

以上により、本発明によると、負荷の低い区画内のファイルを他のディスク装置に移動することで、非稼働ディスクを効率的に増加させることができる。従って、非稼働となったディスク装置に対する電源供給を停止することができ、ストレージシステム全体の省電力化を図ることができる。

【００５６】

ここで、上述したように、区画ビジー率を計算し、ソート済みリストを管理するための処理量を見積もる。

単位時間あたりの処理：1000台の1TBディスク装置を1000個のエリアに分け、全ディスクが稼働ディスクで、10分に一回ビジー率を集計した場合
ディスクビジー率取得回数：1K個/10分＝1.67個/秒
各エリアのサイズ：1GB
エリア数：1M個
ビジー率の計算：1M個/10分＝1.67K個/秒
上記のうち、10％がリストの圏内に入ったとすると、リストへのエリア挿入と再ソートが167回/秒である。そして、リストの長さが短ければ、昨今のCPUにとって大きな負荷ではない。また、稼働ディスク数が1/10になれば、管理対象の区画も1/10になり管理処理は大幅に削減される。このようにディスクが1000台つながっていても、少ない計算と統計情報で、再配置処理を実行することができる。

【００５７】

＜実施形態２＞

本発明の第１の実施形態を、図１２を参照して説明する。図１２は、ストレージシステムの構成を示す機能ブロック図である。なお、本実施形態では、ストレージシステムの概略を説明する。

【００５８】

図１に示すように、本実施形態におけるストレージシステム１００は、
ディスク装置１１１，１１２毎に当該ディスク装置に対する単位時間あたりの稼働状況を計測して、稼働した上記ディスク装置を稼働ディスクと判定し、稼働していない上記ディスク装置を非稼働ディスクと判定すると共に、単位時間当たりの上記ディスク装置における負荷を計測して、計測結果を記憶する性能監視部１０１と、
上記性能監視部による計測結果に基づいて、稼働ディスクと判定された上記ディスク装置に格納されているデータを、当該ディスク装置における負荷よりも高い負荷の他の上記ディスク装置に格納するデータ再配置部１０２と、
を備える。

【００５９】

そして、上記ストレージシステムでは、
上記性能監視部は、上記ディスク装置内の記憶領域が複数に分割された各区画における負荷を計測して、計測結果を記憶し、
上記データ再配置部は、上記性能監視部による計測結果に基づいて、稼働ディスクと判定された上記ディスク装置に形成された上記区画のうち、負荷が他の上記区画と比較して低い上記区画に格納されているデータを、稼働ディスクと判定された他の上記ディスク装置に格納する、
という構成を採る。

【００６０】

また、上記ストレージシステムでは、
上記データ再配置部は、上記性能監視部による計測結果に基づいて、負荷が低い順から予め設定された順番までの上記区画に格納されているデータを、稼働ディスクと判定された他の上記ディスク装置に格納する、
という構成を採る。

【００６１】

上記発明によると、まず、ストレージシステムでは、ディスク装置毎に単位時間当たりの稼働状況を計測する。そして、単位時間あたりに稼働したディスク装置を稼働ディスクと判定し、単位時間あたりに稼働していないディスク装置を非稼働ディスクと判定する。これにより、例えば、非稼働ディスクと判定されたディスク装置を停止することで、ストレージシステム全体の省電力化を図ることができる。

【００６２】

また、ストレージシステムでは、上述したようにディスク装置毎の稼働状況を計測すると共に、ディスク装置における負荷、例えば、ディスク装置内に形成された各区画毎にお

ける負荷を計測して記憶している。そして、稼働ディスクであるディスク装置内に形成された区画のうち、負荷が他の区画と比較して低い区画に格納されているデータを、稼働ディスクである他のディスク装置に格納する。なお、上述したデータの移動は、負荷が低い区画順に実行してもよい。

【００６３】

これにより、データを稼働中の他のディスク装置に移動させたディスク装置は、もともと負荷が低かったため、その後、アクセスが減少して、非稼働ディスクとなる可能性が高くなる。そして、非稼働ディスクとなった場合には、上述したように、そのディスク装置を停止する。これにより、ストレージシステム全体のさらなる省電力化を図ることができる。

【００６４】

また、上記ストレージシステムでは、

上記データ再配置部は、上記性能監視部による計測結果に基づいて、稼働ディスクと判定された上記ディスク装置に形成された上記区画のうち、負荷が他の上記区画と比較して高い上記区画における負荷を軽減するよう、上記各区画に格納されたデータの再配置を行う、

という構成を採る。

【００６５】

また、上記ストレージシステムの他の態様において、

上記データ再配置部は、稼働ディスクと判定された上記ディスク装置に形成された上記区画のうち、負荷が他の上記区画と比較して高い上記区画に格納されたデータの一部又は全部を、他の上記ディスク装置に格納して、データの再配置を行う、

という構成を採る。

【００６６】

また、上記ストレージシステムでは、

上記データ再配置部は、非稼働ディスクと判定された上記ディスク装置を含めて、上記データの再配置を行う、

という構成を採る。

【００６７】

また、上記ストレージシステムでは、

上記データ再配置部は、上記性能監視部による計測結果に基づいて、負荷が高い順から予め設定された順番までの上記区画における負荷を軽減するよう、上記各区画に格納されたデータの再配置を行う、

という構成を採る。

【００６８】

また、上記ストレージシステムでは、

上記性能監視部は、上記各区画における単位時間あたりの入力開始から当該入力に対する出力完了までに要する時間であるターンアラウンドタイムの累積時間を、上記各区画における上記負荷として計測する、

という構成を採る。

【００６９】

上記発明によると、ストレージシステムでは、負荷の高い区画に格納されたデータを分散あるいは複製して他のディスク装置に格納するなどして、データの再配置を行う。これにより、該当する区画が形成されたディスク装置の負荷が軽減されるため、ストレージシステム全体において電力が効率的に利用され、省電力化を図ることができる。

【００７０】

そして、上述したストレージシステムは、コンピュータにプログラムが組み込まれることで実現できる。具体的に、本発明の他の形態であるプログラムは、コンピュータに、

ディスク装置毎に当該ディスク装置に対する単位時間あたりの稼働状況を計測して、稼働した上記ディスク装置を稼働ディスクと判定し、稼働していない上記ディスク装置を非

稼働ディスクと判定すると共に、単位時間当たりの上記ディスク装置における負荷を計測して、計測結果を記憶する性能監視部と、
上記性能監視部による計測結果に基づいて、稼働ディスクと判定された上記ディスク装置に格納されているデータを、当該ディスク装置における負荷よりも高い負荷の他の上記ディスク装置に格納するデータ再配置部と、
を実現させるためのプログラムである。
【００７１】
そして、上記プログラムでは、
上記性能監視部は、上記ディスク装置内の記憶領域が複数に分割された各区画における負荷を計測して、計測結果を記憶し、
上記データ再配置部は、上記性能監視部による計測結果に基づいて、稼働ディスクと判定された上記ディスク装置に形成された上記区画のうち、負荷が他の上記区画と比較して低い上記区画に格納されているデータを、稼働ディスクと判定された他の上記ディスク装置に格納する、
という構成を採る。
【００７２】
また、上記プログラムでは、
上記データ再配置部は、上記性能監視部による計測結果に基づいて、稼働ディスクと判定された上記ディスク装置に形成された上記区画のうち、負荷が他の上記区画と比較して高い上記区画における負荷を軽減するよう、上記各区画に格納されたデータの再配置を行う、
という構成を採る。
また、上述したストレージシステムが作動することにより実行される、本発明の他の形態であるデータ格納方法は、
ディスク装置毎に当該ディスク装置に対する単位時間あたりの稼働状況を計測して、稼働した上記ディスク装置を稼働ディスクと判定し、稼働していない上記ディスク装置を非稼働ディスクと判定すると共に、単位時間当たりの上記ディスク装置における負荷を計測して、計測結果を記憶して、性能監視を行い、
上記計測結果に基づいて、稼働ディスクと判定された上記ディスク装置に格納されているデータを、当該ディスク装置における負荷よりも高い負荷の他の上記ディスク装置に格納して、データの再配置を行う。
【００７３】
また、上述したストレージシステムが作動することにより実行される、本発明の他の形態であるデータ格納方法は、
上記性能監視時に、上記ディスク装置内の記憶領域が複数に分割された各区画における負荷を計測して、計測結果を記憶し、
上記データの再配置時に、上記性能監視による計測結果に基づいて、稼働ディスクと判定された上記ディスク装置に形成された上記区画のうち、負荷が他の上記区画と比較して低い上記区画に格納されているデータを、稼働ディスクと判定された他の上記ディスク装置に格納する、
という構成を採る。
【００７４】
そして、上記データ格納方法の他の態様において、
上記データ再配置時は、上記性能監視による計測結果に基づいて、稼働ディスクと判定された上記ディスク装置に形成された上記区画のうち、負荷が他の上記区画と比較して高い上記区画における負荷を軽減するよう、上記各区画に格納されたデータの再配置を行う、
という構成を採る。
【００７５】
上述した構成を有する、プログラム、又は、データ格納方法、の発明であっても、上記

ストレージシステムと同様の作用を有するために、上述した本発明の目的を達成することができる。

【産業上の利用可能性】

【００７６】

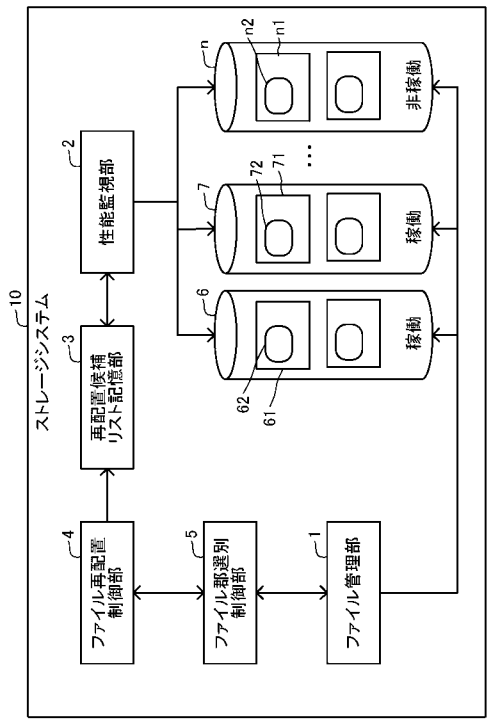
本発明は、ディスクアレイ装置、ＮＡＳ装置、アグリゲートＮＡＳ装置、グリッドストレージなど、複数のディスク装置を搭載したストレージシステムに適用することができ、産業上の利用可能性を有する。

【符号の説明】

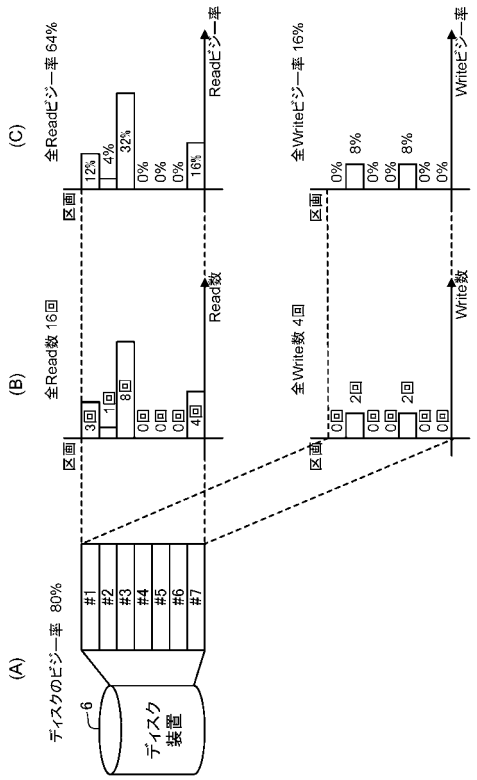
【００７７】

１　ファイル管理部
２　性能監視部
３　再配置候補リスト記憶部
４　ファイル再配置制御部
５　ファイル郡選別制御部
６，７，ｎ　ディスク装置
１０，１００　ストレージシステム
１０１　性能監視部
１０２　データ再配置制御部
１１１，１１２　ディスク装置

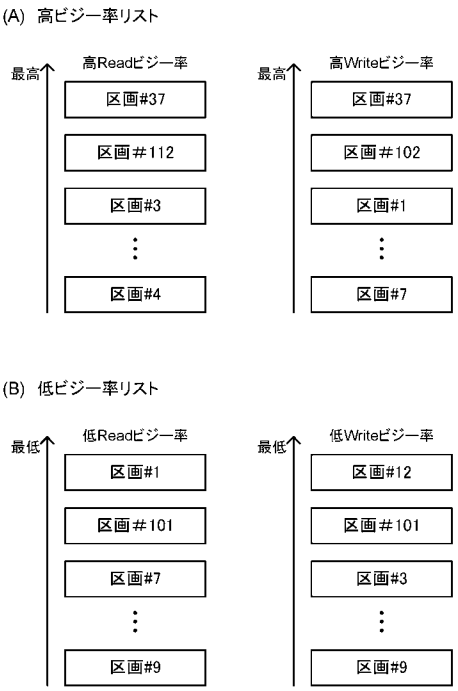
【図１】



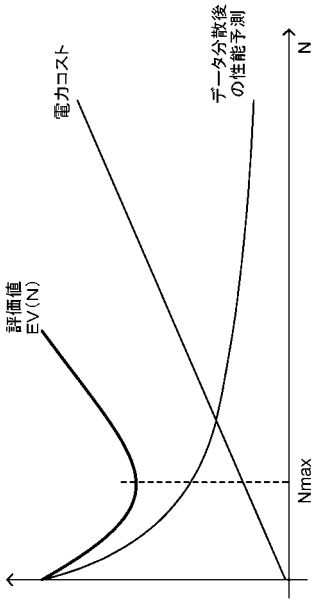
【図２】



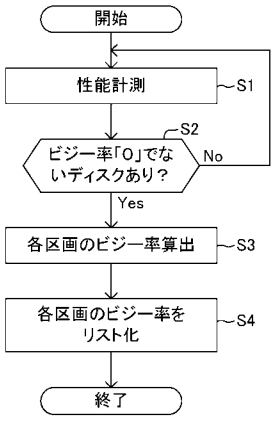
【図３】


(A) 高ビジー率リスト
最高
高Readビジー率
区画#37
区画＃112
区画#3
区画#4
最高
高Writeビジー率
区画#37
区画＃102
区画#1
区画#7
(B) 低ビジー率リスト
最低
低Readビジー率
区画#1
区画＃101
区画#7
区画#9
最低
低Writeビジー率
区画#12
区画＃101
区画#3
区画#9

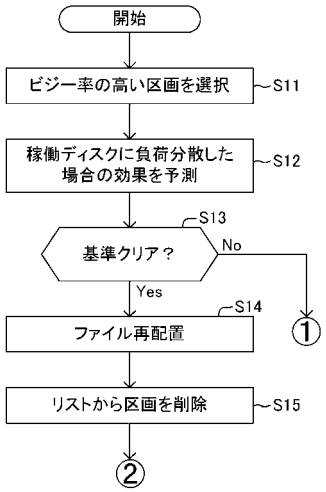
【図４】


電力コスト
データ分散後の性能予測
N
評価値EV(N)
Nmax

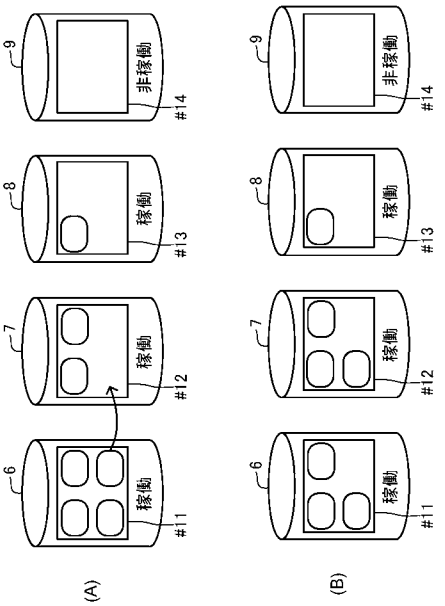
【図５】


開始
性能計測
S1
S2
ビジー率「0」でないディスクあり？
No
Yes
各区画のビジー率算出
S3
各区画のビジー率をリスト化
S4
終了

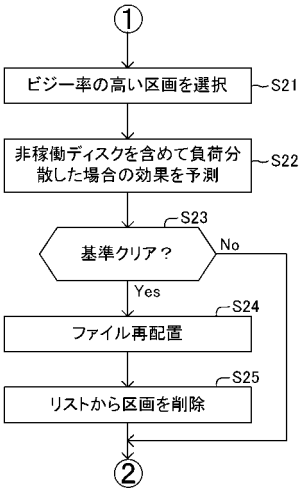
【図６】


開始
ビジー率の高い区画を選択
S11
稼働ディスクに負荷分散した場合の効果を予測
S12
S13
基準クリア？
No
Yes
S14
ファイル再配置
①
リストから区画を削除
S15
②

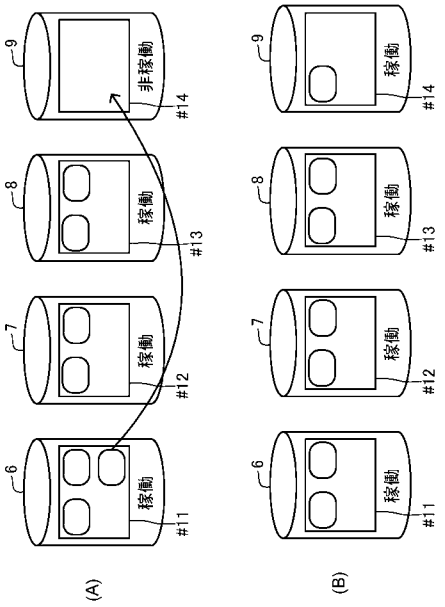
【図７】


9
非稼働
#14
8
稼働
#13
7
稼働
#12
6
稼働
#11
(A)
(B)

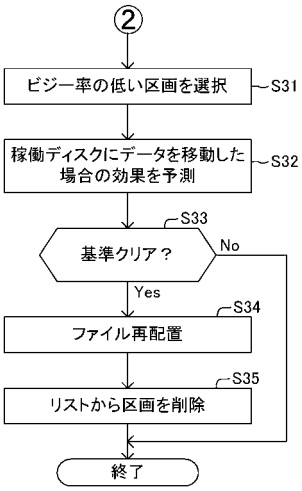
【図８】


①
ビジー率の高い区画を選択
S21
非稼働ディスクを含めて負荷分散した場合の効果を予測
S22
S23
基準クリア？
No
Yes
S24
ファイル再配置
S25
リストから区画を削除
②

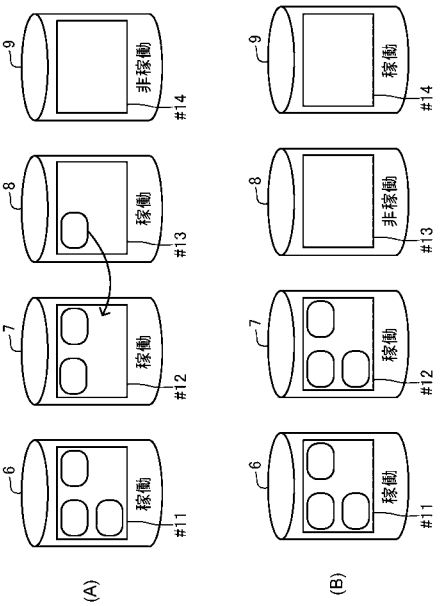
【図９】


9
非稼働
#14
8
稼働
#13
7
稼働
#12
6
稼働
#11
(A)
(B)

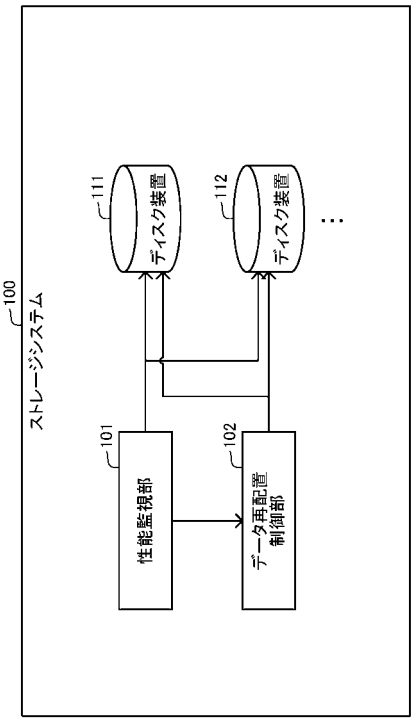
【図１０】


②
ビジー率の低い区画を選択
S31
稼働ディスクにデータを移動した場合の効果を予測
S32
S33
基準クリア？
No
Yes
S34
ファイル再配置
S35
リストから区画を削除
終了

【図１１】


9
非稼働
#14
8
稼働
#13
7
稼働
#12
6
稼働
#11
(A)
9
稼働
#14
8
非稼働
#13
7
稼働
#12
6
稼働
#11
(B)

【図１２】


100
ストレージシステム
111
ディスク装置
112
ディスク装置
…
101
性能監視部
102
データ再配置
制御部

フロントページの続き

(56)参考文献 特開２００１－０９３２２０（ＪＰ，Ａ）
特開２０００－２３１４５４（ＪＰ，Ａ）
特開２００８－２１７５７５（ＪＰ，Ａ）
特開２００８－１４６１４１（ＪＰ，Ａ）
特開２００８－００３７１９（ＪＰ，Ａ）
特開２００７－２９３４８６（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　３／０６－　３／０８
Ｇ０６Ｆ　１２／００
Ｇ０６Ｆ　１３／１０－１３／１４