(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2008-152807
(P2008-152807A)
(43)公開日　平成20年7月3日(2008.7.3)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 3/06	(2006.01)	G06F	3/06	302J	5B065
		G06F	3/06	540	

審査請求 有　請求項の数 14　ＯＬ　（全 35 頁）

(21)出願番号　特願2008-63185(P2008-63185)
(22)出願日　平成20年3月12日(2008.3.12)
(62)分割の表示　特願2001-53458(P2001-53458)の分割
原出願日　平成13年2月28日(2001.2.28)
(31)優先権主張番号　特願2000-205510(P2000-205510)
(32)優先日　平成12年7月6日(2000.7.6)
(33)優先権主張国　日本国(JP)

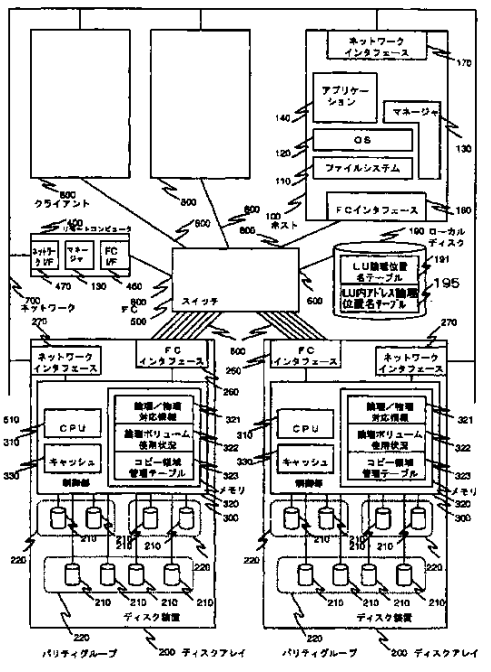
(71)出願人　000005108
株式会社日立製作所
東京都千代田区丸の内一丁目６番６号
(74)代理人　100079108
弁理士　稲葉　良幸
(74)代理人　100093861
弁理士　大賀　眞司
(72)発明者　荒川　敬史
神奈川県川崎市麻生区王禅寺１０９９番地　株式会社日立製作所システム開発研究所内
(72)発明者　茂木　和彦
神奈川県川崎市麻生区王禅寺１０９９番地　株式会社日立製作所システム開発研究所内

最終頁に続く

(54)【発明の名称】計算機システム

(57)【要約】
【課題】異なるストレージサブシステム間でのデータの再配置を適正に行えるようにする。
【解決手段】ディスクアレイ２００は、ホスト１００からのリード/ライトに対してディスク装置２１０の使用状況を取得する。ホスト１００は、複数のディスクアレイ２００から使用状況を収集して、再配置対象ＬＵに格納されているデータの再配置先のＬＵを決定する。そして、アプリケーションにとってのデータ位置であるデータの論理位置名とＬＵとの対応を決定するＬＵ論理位置名テーブル１９１を変更する。また、ディスクアレイ２００間で、再配置対象となったＬＵに格納されているデータを、再配置先のＬＵに移動することで、異なるディスクアレイ２００間でのデータの再配置を行う。
【選択図】図１

図1



【特許請求の範囲】
【請求項１】
計算機と、
前記計算機に接続される複数のストレージサブシステムとを有し、
前記複数のストレージサブシステムの各々は、
該ストレージサブシステムが有する記憶領域の各々について、前記記憶領域の使用状況を取得する取得手段を有し、
前記計算機は、
前記複数のストレージサブシステムの各々から、前記取得手段によって取得された記憶領域各々の使用状況を収集する収集手段を有する
ことを特徴とする計算機システム。
【請求項２】
前記記憶領域は、ディスク装置を用いて形成される論理的な記憶領域である
ことを特徴とする請求項１記載の計算機システム。
【請求項３】
前記複数のストレージサブシステムの各々は、
前記計算機の指示にしたがって、他の前記複数のストレージサブシステムのうちのいずれかからデータの移動を行う移動手段と、前記移動手段が完了したことを前記計算機に通知する手段とを有し、
前記計算機は、
前記データの論理的な位置と、前記ストレージサブシステムにおいて前記データが格納される記憶領域との対応を記憶する対応記憶部と、
前記通知手段によって前記ストレージサブシステム間でのデータの移動が完了した通知を受け取った場合に、前記対応記憶部を更新する更新手段とを有する
ことを特徴とする請求項２記載の計算機システム。
【請求項４】
前記計算機は、
前記複数のストレージサブシステム各々から収集された各々の前記記憶領域の使用状況を表示する表示部と、
該計算機システムの使用者に、前記表示部に表示された各々の前記記憶領域のいずれかを選択させる選択手段とを有し、
前記指示には、前記選択手段によって選択された記憶領域を指定する情報が含まれている
ことを特徴とする請求項３記載の計算機システム。
【請求項５】
前記計算機は、前記複数のストレージサブシステム各々から収集された各々の前記記憶領域の使用状況に基づいて、所定の条件を満たす前記記憶領域を選択する手段を有し、
前記指示には、前記選択手段によって選択された記憶領域を指定する情報が含まれている
ことを特徴とする請求項３記載の計算機システム。
【請求項６】
前記所定の条件とは、前記記憶領域へのアクセスの頻度が一定値以上であるという条件である
ことを特徴とする請求項５記載の計算機システム。
【請求項７】
前記計算機は、前記データと前記計算機で実行されるファイルシステムが管理するファイルとの対応関係を保持するファイル対応関係記録部と、
前記ファイル対応関係記録部を使用して前記ファイルごとに前記記憶領域の使用状況を収集する手段とを有し、
前記複数のストレージサブシステムの各々は、

前記指示にしたがって、指定された前記ファイルを他の前記複数のストレージサブシステムのうちのいずれかから移動する手段と
前記移動する手段が完了したことを前記計算機に通知する手段を有する
ことを特徴とする請求項３記載の計算機システム。
【請求項８】
前記計算機は、前記通知する手段によって、ストレージサブシステム間のファイルの移動が完了したことを通知されると、前記ファイル対応関係記録部を更新する手段を有する
ことを特徴とする請求項７記載の計算機システム。
【請求項９】
前記使用状況には、データを読み出す際に前記記憶領域が占有される時間が含まれる
ことを特徴とする請求項１～８記載の計算機システム。
【請求項１０】
ホストコンピュータと接続される接続部と、
複数の記憶装置と、
前記複数の記憶装置を制御する制御部と、
前記複数の記憶装置の使用状況を示す情報が格納される記憶部とを有し、
前記制御部は、
前記接続部を介して、前記複数の記憶装置の使用状況を示す情報の転送を要求する命令を受け取った場合には、前記記憶部から前記複数の記憶装置の使用状況を示す情報を読み出し、前記ホストコンピュータに転送する
ことを特徴とするストレージサブシステム。
【請求項１１】
複数のストレージサブシステムと接続される接続部と、
前記複数のストレージサブシステムから取得した前記複数のストレージサブシステム各々が有する記憶領域の使用状況が格納される記憶部と、
中央演算処理装置とを有し、
前記中央演算処理装置は、
前記記憶部に格納された前記複数のストレージサブシステムの使用状況の情報に基づいて、移動すべきデータを決定する決定手段と、
前記決定手段によって決定されたデータの情報を前記接続部を介して、移動先となる前記複数のストレージサブシステムのうちのいずれか一つに通知する通知手段とを有する
ことを特徴とする計算機。
【請求項１２】
複数のストレージサブシステム間においてデータを移動する方法であって、
前記複数のストレージサブシステムにおいて収集された情報を前記複数のストレージシステムに接続される計算機で収集し、
移動対象となるデータを前記計算機で決定し、
前記決定されたデータの位置を、移動先となる前記複数のストレージサブシステムのうちの一つに通知し、
通知を受けたストレージサブシステムで前記通知されたデータの移動を行う
ことを特徴とするデータの移動方法。
【請求項１３】
複数のストレージサブシステム間においてデータを移動する方法であって、
前記複数のストレージサブシステムにおいて収集された情報を前記複数のストレージシステムに接続される計算機で収集し、
移動対象となるデータを前記計算機で決定し、
前記決定されたデータの位置を、移動先となる前記複数のストレージサブシステムのうちの一つに通知し、
通知を受けたストレージサブシステムで前記通知されたデータの移動を行う
ことを特徴とするデータの移動方法。

【請求項１４】
複数のストレージサブシステム間においてデータを移動するコンピュータプログラムであって、
前記複数のストレージサブシステムにおいて収集された情報を前記複数のストレージシステムに接続される計算機で収集し、
移動対象となるデータを前記計算機で決定し、
前記決定されたデータの位置を、移動先となる前記複数のストレージサブシステムのうちの一つに通知するプログラムと、
通知を受けたストレージサブシステムで前記通知されたデータの移動を行うプログラムとから構成される
ことを特徴とするコンピュータプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、記憶装置に記憶されたデータを再配置する技術に関し、特に、複数の記憶装置を有する計算機システムでのデータの再配置に好適な技術に関する。
【背景技術】
【０００２】
計算機システムにおける、ストレージサブシステム内に記憶されたデータを再配置する技術として、特開平9-274544号公報記載のディスクアレイシステムがある。ここで、ストレージサブシステムとは、複数の記憶装置から構成される記憶装置システムのことを言う。
【０００３】
ディスクアレイシステムとは、複数のディスク装置をアレイ状に配置し、各ディスク装置を並列に動作させることで、各ディスク装置に分割格納されるデータのリード/ライトを高速に行うシステムのことである。D. A. Patterson, G. Gibson, and R. H. Kats, ”A Case for Redundant Arrays of Inexpensive Disks （ＲＡＩＤ）”（in Proc. ACM SIGMOD, pp. 109-116, June 1988）に記載されているように、冗長性を付加したディスクアレイシステムには、その冗長構成に応じてレベル１からレベル５の種別が与えられている。これらの種別に加えて、冗長性無しのディスクアレイシステムをレベル０と呼ぶこともある。
【０００４】
ディスクアレイシステムを構成するためのコスト、ディスクアレイシステムの性能や特性等は、ディスクアレイシステムのレベルによって異なる。そこで、システムの使用目的に応じて、ディスクアレイシステムの構築の際にレベルの異なる複数のアレイ（ディスク装置の組）を混在させることも多い。このようにレベルの異なるアレイを混在させたディスクアレイシステムにおいて、各レベルのアレイは、パリティグループと呼ばれる。また、ディスク装置についても性能や容量等によりコストが異なるので、最適のコストパフォーマンスを実現するシステムを構築するために、性能や容量の異なる複数種のディスク装置を用いることがある。
【０００５】
ディスクアレイシステムでは、データが複数のディスク装置に分散して配置されるため、ディスクアレイシステムに接続されるホストコンピュータが、論理記憶領域とディスク装置の記憶領域を示す物理記憶領域との対応付け（アドレス変換）を行っている。
【０００６】
特開平9-274544号公報には、ディスクアレイシステム内において、物理記憶領域間におけるデータの再配置を実行し、データの論理記憶領域に対する物理記憶領域への対応付けを、再配置前の物理記憶領域から再配置後の物理記憶領域に変更する技術が開示されている。また、ディスクアレイシステムがホストコンピュータからの各論理記憶領域に対するアクセスによる負荷状況を管理し、その実績に応じて再配置後にデータが適正配置となる

ように、再配置の内容を決定するという技術も開示されている。
【０００７】
ホストコンピュータおよびディスクアレイシステム等のストレージサブシステム間におけるデータ転送の技術としては、M. T. O'Keefe, "Shared File Systems and Fibre Channel"（in Proc. Sixth Goddard Conference on Mass Storage Systems and Technologies, pp. 1-16, March 1998）に開示された技術がある。
【０００８】
この技術は、高速のネットワークであるFibre Channel（以下「ＦＣ」と称する。）で複数のホストコンピュータと複数のストレージサブシステムとを接続し、ＦＣ経由でデータ共有を実現するストレージ環境、いわゆるStorage Area Network（ＳＡＮ）を実現するための技術である。このように、ＦＣ経由でデータ転送を行うことにより、一般的なネットワーク経由に比べ、ホストコンピュータおよびネットワークの負荷が削減される。
【０００９】
高速なＦＣを使用しない、一般的なネットワークに接続されたストレージサブシステムに保持されているファイル等のデータを、複数のコンピュータで共有する技術としては、ＮＦＳ（Network File System）が広く知られている。
【００１０】
ＮＦＳを用いてネットワーク間でデータ共有を行う場合には、ＦＣを使用する場合に比べ、ファイルを共有しているコンピュータや、コンピュータとストレージサブシステムをつなぐネットワークに対する負荷が大きくなる。しかし、ＮＦＳを用いると、既存のネットワークを使用できることから、新たにＦＣのネットワークを敷設することと比較すると、新規設備コストを抑えられ、またファイル共有等の管理が容易である等の利点がある。
【発明の開示】
【発明が解決しようとする課題】
【００１１】
上述したように、特開平9-274544号公報に開示された技術では、１つのストレージサブシステム内におけるデータの再配置が可能となる。しかしながら、複数のストレージサブシステムを有する計算機システムにおいて、異なるストレージサブシステム間でのデータの再配置を行うことはできない。また、ディスクアレイシステムはファイルを認識できないため、ファイル単位でデータの再配置を行うことができない。
【００１２】
一方、ＳＡＮの技術を用いれば、異なるストレージサブシステムにおける高速なデータ転送が可能となる。しかしながら、従来技術においては、ＳＡＮの技術を利用してストレージサブシステム間でデータの再配置を行うことは出来なかった。
【００１３】
つまり、従来技術においては、ＳＡＮを構成する各ストレージサブシステムの各記憶領域の負荷状況等、データの適正な配置を決定するために必要な情報を、ホストコンピュータやシステムを使用するユーザが得ることについて、何ら考えられていない。このため、ホストコンピュータや、そのユーザが、どのようにしてデータの再配置を行えば、データの適正な配置を実現することができるのかを判断できない。
【００１４】
さらに、仮にユーザが自らストレージサブシステム間におけるデータの再配置を行おうとしても、データを再配置するための未使用領域の管理等を、全てユーザが詳細に検討して行わなければならず、ユーザに対する負担は大きいものがあった。
【００１５】
また、異なるストレージサブシステム間でデータを転送すると、アプリケーションが認識するデータの位置、すなわち、アプリケーションが同じデータにアクセスするために指定すべきアクセス先が再配置の前後で変化してしまうが、この変化についても従来技術では何ら考慮されていない。
【００１６】

さらに、一般的なネットワークで接続されているコンピュータ同士で、ＮＦＳを使用してデータ共有を行う場合にも、以下の問題がある。

【００１７】

すなわち、従来、ＮＦＳによるデータ共有を実現するために使用されるホストコンピュータ（ＮＦＳサーバ）が、複数のストレージサブシステムを管理している場合、ＮＦＳサーバ自身は、その複数のストレージサブシステム間でのデータの物理的再配置等を行うことはできなかった。このため、ＮＦＳサーバを用いてデータ共有を行うコンピュータ毎に、共有されるデータの物理的位置を変更するといった、ストレージサブシステムの有する記憶領域についてのより細かい区別および管理を行うことができなかった。

【００１８】

本発明は上記事情に鑑みてなされたものであり、本発明の目的は、ＮＦＳサーバを含めたホストコンピュータがデータの適正配置の決定に必要な情報をストレージサブシステムから取得できるようにし、ＳＡＮ環境において、データの再配置を実現することにある。また、異なる目的としては、ユーザのストレージサブシステムの管理負担を軽減することにある。また、異なるストレージサブシステム間におけるデータの再配置を、アプリケーションが認識するデータの位置が、再配置の前後で変化しないようにして、行えるようにすることにある。さらにまた、ファイルを対象とするデータの再配置を可能とすることにある。

【課題を解決するための手段】

【００１９】

前記の課題を解決するため、本発明は以下の構成とする。すなわち、計算機と、前記計算機に接続される複数のストレージサブシステムとを有し、ストレージサブシステムは、ストレージサブシステムが有する記憶領域の各々について記憶領域の使用状況を取得する取得手段を有し、計算機は、複数のストレージサブシステムから、記憶領域各々の使用状況を取得する取得手段を有すること計算機システムという構成とする。記憶領域は、論理的な記憶領域であってもよい。

また、ストレージサブシステムには、計算機の指示にしたがって、データの移動を行う移動手段を付加し、計算機には、データの論理的な位置と、ストレージサブシステムにおいてデータが格納される記憶領域との対応を規定する対応テーブルと、移動手段によってストレージサブシステム間でデータが移動した場合に、対応テーブルを更新する更新手段とを付加した構成とすることもできる。

【００２０】

ここで、各記憶領域の使用状況とは、例えば、その記憶領域の物理的な記憶空間の使用状況やその記憶空間へのアクセス処理に消費された処理時間等である。

【００２１】

また、ストレージサブシステムで移動されるデータの単位及び計算機で管理されるデータの単位がファイル単位であることも考えられる。

【００２２】

さらに、全てのストレージサブシステムが有する論理的な記憶領域全体を管理する手段と、記憶装置の特徴と論理的な記憶領域との対応関係を管理する手段とを計算機に付加した構成も考えられる。

【００２３】

また、計算機に、ストレージサブシステムに格納されているデータをネットワーク間で共有する手段を付加した構成も考えられる。

【発明を実施するための最良の形態】

【００２４】

図１は、本発明が適用された計算機システムの第１実施形態の構成を示す図である。

【００２５】

本実施形態の計算機システムは、ホストコンピュータ（ホスト）１００、ディスクアレイ２００、スイッチ５００、クライアント８００及びローカルディスク１９０を有する。

【００２６】
ホスト１００は、ネットワークインタフェース１７０により、ネットワーク７００を介して、クライアント８００及びディスクアレイ２００に接続される。ホスト１００は、また、ＦＣインタフェース１６０、スイッチ５００及びＦＣ６００を介して、ディスクアレイ２００及びローカルディスク１９０に接続される。
【００２７】
ホスト１００は、ファイルシステム１１０、オペレーティングシステム（以下「ＯＳ」と称する。）１２０、マネージャ１３０及びアプリケーション１４０をホスト１００自身が有する記憶領域に有する。
【００２８】
アプリケーションプログラム（以下、単にアプリケーションと呼ぶ）１４０は、ＯＳ１２０およびファイルシステム１１０を介してディスクアレイ２００に対してリードやライトの要求を出す。
【００２９】
ホスト１００及びクライアント８００としては、一般的な電子計算機が用いられる。ファイルシステム１１０等のプログラムは、ホスト１００の外部にあるローカルディスク１９０に記憶され、必要に応じてホスト１００に読み込まれて実行される。
【００３０】
ホスト１００がその内部に記憶装置を有する場合には、当該記憶装置をローカルディスク１９０として使用することも考えられる。
【００３１】
ローカルディスク１９０には、ＯＳ１２０及びファイルシステム１１０が使用する論理ユニット（以下、「ＬＵ」と称する）論理位置名テーブル１９１及びＬＵ内アドレス論理位置名テーブル１９５等の各種管理情報が格納されている。ＬＵ論理位置名テーブル１９１は、アプリケーション１４０がディスクアレイシステム２００のデータにアクセスするときに指定する論理位置名と、論理位置名により特定されるデータを格納するＬＵとの対応を示す情報を保持したテーブルである。
【００３２】
ＬＵ内アドレス論理位置名テーブル１９５は、アプリケーション１４０がディスクアレイシステム２００のデータにアクセスするときに指定する論理位置名と、論理位置名により特定されるデータのＬＵ内アドレスとの対応を示す情報を保持したテーブルである。
【００３３】
ディスクアレイ２００は、制御部３００、複数のディスク装置２１０、ネットワークインタフェース２７０及びＦＣインタフェース２６０を有する。
【００３４】
制御部３００は、処理を実行するためのＣＰＵ３１０、メモリ３２０及びキャッシュメモリ３３０を有する。
【００３５】
メモリ３２０には、論理/物理対応情報３２１、論理ボリューム使用状況３２２及びコピー領域管理テーブル３２３が格納される。これらの情報の詳細については後述する。
【００３６】
本実施形態では、ｎ台（ｎは２以上の整数）のディスク装置２１０でアレイ（以下「ＲＡＩＤ」と称する。）が構成されており、このｎ台のディスク装置２１０によるＲＡＩＤをパリティグループ２２０と呼ぶ。
【００３７】
ＲＡＩＤとしては、１つのパリティグループ２２０に含まれるｎ台のディスク装置２１０のうち、ｎ－１台のディスク装置２１０の格納内容から生成される冗長データ（以下「パリティ」と称する。）を残りの１台に格納する構成や、ｎ/２台に格納されている内容を残りのｎ/２台がコピーしたミラーディスク（ＲＡＩＤレベル１）構成が考えられる。また、各パリティグループ２２０を、ホスト１００からのアクセス対象の１単位とみなす

ことができる。

【００３８】

本実施形態においては、ディスクアレイ２００を構築する各パリティグループ２２０の性能、信頼性、特性などの属性は、同一であってもよいし、あるいは、異なっていてもかまわない。

【００３９】

制御部３００は、ホスト１００がリード/ライトする論理ボリュームとディスク装置２１０の記憶領域を示す物理アドレスとの対応付け（アドレス変換）を行い、ホスト１００に論理ボリュームを提供する。ディスクアレイ２００は、アドレス変換において、複数の論理ボリュームを結合して１つのＬＵとしてホスト１００に提供することもできる。すなわち、ディスクアレイ２００は、少なくとも１つの論理ボリュームからなるＬＵをホスト１００に提供する。ホスト１００は、ＬＵに対してリード/ライトを行う。

【００４０】

本実施形態では、複数のディスクアレイ２００間における、ディスクアレイ２００の使用状況を考慮したデータの物理的再配置を可能とする。具体的には、ディスクアレイ２００は、ホスト１００からのリード/ライトに対するディスク装
置２１０の使用状況を取得する。ホスト１００は、複数のディスクアレイ２００各々が取得した使用状況を収集し、ユーザに提示する。さらに、ホスト１００は、ディスクアレイ２００の使用状況の提示を受けたユーザからの指示等に応じ、ローカルディスク１９０内のＬＵ論理位置名テーブル１９１を変更すると共に、ディスクアレイ２００がＬＵに格納しているデータをコピーする。これにより、複数のディスクアレイ２００間におけるＬＵの再配置が行われる。そして、ディスクアレイ２００の使用状況を考慮したデータの再配置を可能とすることにより、データの適正配置が可能となる。

【００４１】

図２は、ディスクアレイ２００が、ホスト１００からのリード/ライト要求に応答して行うリード/ライト処理、及びディスクアレイ２００が、ディスク装置２１０の使用状況を取得する使用状況取得の処理の手順を示すフロー図である。使用状況取得の処理は、随時、又はホスト１００からのリード/ライト要求時に行われる。

【００４２】

ホスト１００のアプリケーション１４０は、ファイル論理位置名によってファイルを指定し、ファイルに対するリード/ライトをＯＳ１２０に要求する。ＯＳ１２０は、ファイルシステム１１０にファイルのリード/ライトを要求する。

【００４３】

ファイルシステム１１０は、ＦＣインタフェース１６０を介してローカルディスク１９０にアクセスし、指定されたファイルが格納されているＬＵ番号をＬＵ論理位置名テーブル１９１から求める。ファイルシステム１１０は、指定されたファイルが格納されているＬＵ内アドレス等を、ＬＵ内アドレス論理位置名テーブル１９５から求める。

【００４４】

ホスト１００は、ＦＣインタフェース１６０を介して、求めたＬＵ番号のＬＵを提供するディスクアレイ２００に対し、ＬＵ番号やＬＵ内アドレスを伴うSmall Computer System Interface（ＳＣＳＩ）規格のリードコマンド、あるいはライトコマンドを発行する。

【００４５】

アプリケーション１４０が、論理ドライブ名、ディレクトリ名及びファイル名によるファイルの論理位置までのパスの記述によりファイルを指定するシステムでは、論理位置（論理ドライブやディレクトリやファイル）へのパスの記述が、ファイルの論理位置の論理位置名となる。一般的には、論理位置名とは、アプリケーションがアクセス対象の指定に使用する論理位置の情報である。

【００４６】

ファイルシステム１１０は、各論理位置を管理するため、ディレクトリ構造などの各論理位置間の階層的な論理構造を管理する他、各論理位置の論理位置名とＬＵ番号との対応

をＬＵ論理位置名テーブル１９１に記述し管理する。また、各論理位置の論理位置名とＬＵ内アドレスとの対応をＬＵ内アドレス論理位置名テーブル１９５に記述し管理する。なお、ＬＵ番号は、そのＬＵ番号のＬＵを提供するディスクアレイ２００も表す（ステップ１０００）。

【００４７】

ホスト１００からリード/ライトのコマンドを受領した制御部３００は、メモリ３２０内の論理/物理対応情報３２１を用いて、コマンドで指定されているＬＵ番号により特定されるＬＵを構成する論理ボリュームを特定する。制御部３００は、論理ボリュームに対応するディスク装置２１０の領域を求め、コマンドで指定されているＬＵ内アドレスの物理アドレスへの変換を行う。

【００４８】

論理/物理対応情報３２１は、図３に示すように、ＬＵとディスク装置２１０の物理アドレスとの対応関係についての情報を保持するテーブルである。

【００４９】

図中、ＬＵ番号５００１およびＬＵ内アドレス５００２は、ホスト１００のファイルシステム１１０がリード/ライト処理で指定するＬＵ番号及びＬＵ内アドレスを示す。論理ボリューム番号５００３は、ＬＵ番号５００１で特定されるＬＵに対応する論理ボリュームの番号である。論理ボリュームアドレス５００４は、ＬＵ内アドレス５００２に対応する論理ボリューム内のアドレスである。

【００５０】

物理アドレスは、データとパリティが格納されるディスク装置２１０上の領域を示すアドレスである。物理アドレスは、パリティグループ番号５００５、データおよびパリティ各々に対するディスク装置番号５００６及びディスク装置内アドレス５００７を有する。パリティグループ番号５００５は、個々のパリティグループ２２０を示す。ディスク装置番号５００６は、個々のディスク装置２１０を示す。ディスク装置内アドレス５００７は、ディスク装置２１０内での領域を示すアドレスである（ステップ１０１０）。

【００５１】

データのリードの場合、制御部３００は、アドレス変換で得た物理アドレスに基づいて、ディスク装置２１０のデータを読み出し、ホスト１００に転送する。データのライトの場合、制御部３００は、ホスト１００から転送されたデータ及びデータに関連して生成したパリティを、アドレス変換で得たディスク装置２１０の物理アドレスの位置に格納する（ステップ１０２０）。

【００５２】

リード/ライト処理を終了した制御部３００は、使用状況取得処理を実行する。この処理では、制御部３００は、リード/ライト処理でのリード/ライト種別やシーケンシャル/ランダムアクセス種別を判別し、メモリ３２０のリード/ライト対象となった論理ボリュームの論理ボリューム使用状況３２２を更新する。論理ボリューム使用状況３２２は、ディスクアレイ２００に含まれるＬＵの使用状況についての情報を保持したテーブルである。論理ボリューム使用状況３２２の一例を、図４に示す。

【００５３】

論理ボリューム使用状況３２２には、論理ボリューム毎に、論理ボリューム番号５１０１及びリード/ライト種別およびシーケンシャル/ランダムアクセス種別毎のディスク使用時間（マイクロ秒単位）５１０２が記述される。ここでは、リード/ライト対象となった論理ボリュームの論理ボリューム番号５１０１に対応する、ディスク使用時間５１０２に、リード/ライトに要した時間が加算される（ステップ１０３０）。

【００５４】

図５は、ホスト１００が、各ディスクアレイ２００からディスク装置２１０の使用状況を収集する使用状況収集処理の手順を示すフロー図である。この処理は、随時行われる。

【００５５】

ホスト１００のマネージャ１３０は、ＦＣインタフェース１６０を介し、コマンドボリ

ュームに対して、情報収集用のパラメータをライトデータとするＳＣＳＩ規格のライトコマンドを発行する。コマンドボリュームは、ディスクアレイ２００が有する情報転送用のＬＵであって、対応する物理領域が指定されない論理ボリュームである。（ステップ１１００）。

【００５６】

制御部３００は、発行されたコマンドがコマンドボリュームに対するライトコマンドであることを確認すると、ホスト１００から転送された情報収集用のパラメータに含まれるオペレーションコードから、要求された情報を判別する。制御部３００は、要求された情報をメモリ３２０上に用意する（ステップ１１１０）。制御部３００は、ＦＣインタフェース２６０を介して、ホスト１００にライトの完了を報告する（ステップ１１２０）。

【００５７】

完了報告を受けたホスト１００のマネージャ１３０は、ＦＣインタフェース１６０を介して、ディスクアレイ２００のコマンドボリュームに、ＳＣＳＩ規格のリードコマンドを発行する（ステップ１１３０）。

【００５８】

制御部３００は、コマンドボリュームに対するリードコマンドを受領すると、メモリ３２０上に用意した情報を、ＦＣインタフェース２６０を介してホスト１００に転送する（ステップ１１４０）。制御部３００は、ＦＣインタフェース２６０を介してホスト１００にリードの完了を報告する（ステップ１１５０）。

【００５９】

ステップ１１００でライトされる情報収集用のパラメータ及びステップ１１１０で用意される情報には、論理ボリューム情報、パリティグループ情報及び使用状況情報の３種類の情報が含まれる。

【００６０】

ステップ１１００でライトされる情報収集用のパラメータが、図６に示すような論理ボリューム情報のパラメータである場合、制御部３００は、その０～１バイト目で指定された論理ボリューム番号５２０１で特定される論理ボリュームについて、図７に示すような論理ボリューム情報（ディスクアレイ２００内のその論理ボリュームの構成を示す情報）を用意する。

【００６１】

図７に示す論理ボリューム情報において、８～４７バイト目には、その０～１バイト目に記述されている論理ボリューム番号５２０１で特定される論理ボリュームの各種情報５２０２が記述される。４９～１２１バイト目には、その論理ボリュームが属するＬＵを構成する各論理ボリュームの情報５２０３が記述される。

【００６２】

情報収集用のパラメータが、パリティグループ情報のパラメータの場合、制御部３００は、パラメータで指定された論理ボリュームが属するパリティグループ２２０のパリティグループ情報（ＲＡＩＤの構成、ディスク装置２１０の型名等、ディスクアレイ２００内のそのパリティグループ２２０の構成を示す情報）を用意する。

【００６３】

情報収集用のパラメータが、ディスク装置２１０の使用状況を確認するためのパラメータの場合、制御部３００は、パラメータで指定された論理ボリュームの使用状況情報（ディスクアレイ２００内のリソースの使用状況、例えば論理ボリュームが占有される時間、論理ボリュームの各種コマンド受領回数やキャッシュメモリ３３０のヒット回数等の情報、プロセッサ３１０の占有時間及び内部バスの占有時間等の情報等）を用意する。

【００６４】

制御部３００は、あらかじめ、論理ボリューム毎に、各種コマンド受領回数やキャッシュメモリ３３０のヒット回数やプロセッサ３１０の占有時間や内部バスの占有時間等を取得している。マネージャ１３０は、例えば複数回取得した占有時間の平均を取得間隔で割ることにより、単位時間あたりの占有時間率を求めることができる。

【００６５】
制御部３００は、論理ボリューム情報やパリティグループ情報を生成する際に、論理/物理対応情報３２１の一部あるいは全部を使用する。マネージャ１３０は、各型のディスク装置２１０の性能に関する情報を保持しており、ディスク装置２１０の型名を基に、パリティグループ２２０を構成するディスク装置２１０の性能を得ることができる。
【００６６】
また、ホスト１００のマネージャ１３０は、ＬＵに対しＳＣＳＩ規格のINQUIRYコマンドを発行して応答データを得ることで、この応答データからＬＵに属する論理ボリューム番号を得ることもできる。
【００６７】
図８は、ホスト１００が再配置すべきデータを決定する再配置対象決定処理の手順を示すフロー図である。本処理は、ユーザが再配置すべきデータを検索する際に使用するアプリケーションが実行された時に実行される。
【００６８】
ホスト１００のマネージャ１３０は、ＯＳ１２０が使用しているＬＵ及び使用していないＬＵ（空きＬＵ）を、例えばローカルディスク１９０に格納されているＬＵ論理位置名テーブル１９１から判定する。マネージャ１３０は、ＯＳ１２０が使用している各ＬＵについて、ＬＵが属するディスクアレイ２００における各論理ボリュームの使用状況や、ＬＵに対応する論理ボリュームの使用状況等を計算する。この計算には、INQUIRYコマンドを発行して得られるＬＵに属する論理ボリューム番号、使用状況収集処理で得られる各ディスクアレイ２００における論理ボリューム情報、パリティグループ情報および論理ボリュームの使用状況等が使用される（ステップ１２００）。
【００６９】
マネージャ１３０は、使用状況等の計算結果を、各論理ボリュームが属するパリティグループ２２０の属性（ＲＡＩＤ構成、ディスク装置２１０の型名又はディスク装置２１０の性能等）等と共にユーザに提示する（ステップ１２１０）。
【００７０】
マネージャ１３０は、各ＬＵについて、INQUIRYコマンドを発行して得られた各ＬＵに属する論理ボリューム番号、使用状況収集処理で得られた各ディスクアレイ２００における論理ボリューム情報、パリティグループ情報及び論理ボリュームの使用状況等とから、各空きＬＵが対応する各論理ボリュームの使用状況等を計算する（ステップ１２２０）。この計算結果が、各空きＬＵに関連するパリティグループ２２０の属性等と共にユーザに分類されて提示される（ステップ１２３０）。
【００７１】
使用状況等の情報は、ホスト１００あるいはホスト１００にネットワーク接続された他の計算機で表示することもできる。
【００７２】
ユーザは、各ディスクアレイ２００の各ＬＵについての情報を参照し、データを再配置すべきＬＵ（再配置元ＬＵ）及びデータの再配置先のＬＵを決定する。ただし、ユーザではなく、ホスト１００のマネージャ１３０が、各ＬＵについての情報から自動的にデータ再配置元又は再配置先を決定してもよい。再配置の決定は、たとえば、再配置後に、ディスクアレイ２００間での負荷分散、パリティグループ２２０間での負荷分散、高性能を要求するファイルが存在するＬＵの高性能パリティグループ２２０への配置等が実現されるように行なわれる。再配置先ＬＵのサイズは、再配置元ＬＵのサイズ以上でなければならない。各ＬＵのサイズは、ＳＣＳＩ規格のREAD CAPACITYコマンドで取得することができる（ステップ１２４０）。
【００７３】
図９は、ホスト１００が行う、データの再配置処理の手順を示すフロー図である。ホスト１００は、再配置を決定したユーザの指示、例えば再配置を指示する実行コマンドの入力等があった場合に本処理を実行する。

【００７４】
ユーザからの指示が入力されたホスト１００のマネージャ１３０は、ファイルシステム１１０に再配置元ＬＵのロックを指示する（ステップ１３００）。ファイルシステム１１０は、ロック指示に応じて、再配置元ＬＵへのリード/ライト要求の受付を禁止する（ステップ１３１０）。
【００７５】
次に、マネージャ１３０は、ファイルシステム１１０に、再配置元ＬＵについてのキャッシュメモリのフラッシュを指示する（ステップ１３２０）。ファイルシステム１１０は、再配置元ＬＵに格納されるデータであって、ホスト１００上のメモリにキャッシュされていて且つディスクアレイ２００に未だライトされていないデータを、ディスクアレイ２００の再配置元ＬＵにライトする（ステップ１３３０）。
【００７６】
マネージャ１３０は、ファイルシステム１１０に、再配置元ＬＵについてのキャッシュの無効化を指示する（ステップ１３４０）。ファイルシステム１１０は、再配置元ＬＵに格納されるデータであってホスト１００上のメモリにキャッシュされているデータを無効にする（ステップ１３５０）。
【００７７】
ＬＵのロック、キャッシュメモリのフラッシュ及び無効化の処理は、ＬＵのアンマウントの処理に相当する。
【００７８】
マネージャ１３０は、再配置先ＬＵが存在するディスクアレイ２００に、再配置元ＬＵから再配置先ＬＵへのデータのコピーを指示する。この指示は、使用状況収集処理と同様、再配置先ＬＵが存在するディスクアレイ２００のコマンドボリュームに、コピー指示オペレーションコードや再配置元ＬＵや再配置先ＬＵ等のコピー指示のパラメータを含んだライトコマンドを発行することで行われる（ステップ１３６０）。ディスクアレイ２００は、後述のコピー処理を開始し、コピー指示の受領をマネージャ１３０に通知する（ステップ１３７０）。
【００７９】
マネージャ１３０は、ローカルディスク１９０に格納されているファイルシステム１１０が使用するＬＵ論理位置名テーブル１９１を書き換え、再配置元ＬＵと再配置先ＬＵとの論理位置名を入れ替える。入れ替えられるＬＵ論理位置名テーブル１９１の例を、図１０及び図１１に示す。
【００８０】
図中、ディスクアレイ番号、ＩＤおよびＬＵＮは、ＬＵ番号６００１を特定するために必要な情報である。図１０は、論理位置名をディレクトリ形式で示したものであり、図１１は、論理位置名をドライブ形式で示したものである。いずれも、アプリケーション１４０が使用する記憶領域としてのＬＵの論理位置を示している（ステップ１３８０）。マネージャ１３０は、ファイルシステム１１０に、ＬＵ論理位置名テーブル１９１の更新（再読み込み）及びステップ１３００で指示したロックの解除を指示する（ステップ１３９０）。
【００８１】
ファイルシステム１１０は、ＬＵ論理位置名テーブル１９１を再度読み込んで情報を更新する（ステップ１４００）。ファイルシステム１１０は、ロックを解除してリード/ライト要求の受け付けを再開する（ステップ１４１０）。
【００８２】
ステップ１４００及び１４１０の処理は、ＬＵのマウント処理に相当する。
【００８３】
ステップ１４１０の処理が実行された後は、ファイルシステム１１０のリード/ライトの対象となるＬＵが再配置の対象であるＬＵであれば、ファイルシステム１１０のリード/ライト処理は、ステップ１３８０において情報が入れ替えられた再配置先ＬＵに対して

行われる。

【００８４】

図１２は、再配置処理において、ディスクアレイ２００が、ホスト１００からコピー指示を受けた際に行うコピー処理の手順を示すフロー図である。

【００８５】

再配置先ＬＵが存在するディスクアレイ２００が、ＦＣインタフェース２６０を介してホスト１００からコピー指示を受け取ると、制御部３００は、コピー指示で指定された再配置先ＬＵについてのコピー領域管理テーブル３２３をメモリ３２０上に用意する。

【００８６】

図１３は、コピー領域管理テーブル３２３の内容を示す図である。コピー領域管理テーブル３２３は、コピーされるデータの範囲、大きさ等の情報が登録されているテーブルである。

【００８７】

図中、コピー先ＬＵ番号６１０１及びコピー元ＬＵ番号６１０２は、ＦＣ６００のネットワーク内において再配置先ＬＵと再配置元ＬＵを一義的に示す番号を格納する領域である。具体的には、ホスト１００からコピー指示のパラメータとして指定された８バイトの番号（WORLD WIDE NAME）、３バイトの番号（N_PORT ID）、ＳＣＳＩ規格のターゲットＩＤもしくはＬＵＮが格納される。コピーブロック数６１０３には、コピーする領域のブロック（最小リード/ライト単位）の数であり、コピー領域の大きさを示すデータが格納される。ビットマップ６１０４のビットには、ＬＵのコピー対象領域の各ブロックが割り当てられる。ビットが「１」である場合は未コピーを示し、「０」である場合はコピー済を示す。初期時は、コピー対象領域に対応するすべてのビットが１に設定される（ステップ１５００）。

【００８８】

制御部３００は、コピー指示の受領をホスト１００に通知する。この通知は、コピー指示を実際に受領してから、コピー領域管理テーブル３２３の設定後、実際にコピーを行う前の時点で行われる。このため、コピー指示の受領から当該通知までの時間は短い（ステップ１５１０）。

【００８９】

制御部３００は、ＦＣインタフェース２６０を介して再配置元ＬＵから格納すべきデータをリードし、再配置先ＬＵに格納するコピーを行う（ステップ１５２０）。

【００９０】

制御部３００は、ＬＵのコピー対象領域について、コピー済の領域に対応するビットマップ６１０４のビットを順次０に変更する（ステップ１５３０）。制御部３００は、対象となるＬＵ全体のコピーが終了したら、コピー処理を終了する（ステップ１５４０）。

【００９１】

再配置元ＬＵが存在するディスクアレイ２００と再配置先ＬＵが存在するディスクアレイ２００とが同一の場合には、ディスクアレイ２００内でＬＵのコピーが行われる。

【００９２】

ホスト１００からの再配置対象ＬＵへのリード/ライトは、再配置対象ＬＵのデータがコピー中であっても、再配置先ＬＵ、すなわち再配置先ＬＵの存在するディスクアレイ２００に対して行われる。

【００９３】

図１４は、再配置先ＬＵの存在するディスクアレイ２００が、データの再配置におけるコピー処理の最中に、再配置の対象となるＬＵに対するリード/ライトコマンドを受けた場合における処理の手順について示すフロー図である。

【００９４】

ディスクアレイ２００が、ＦＣインタフェース２６０を介してリードコマンドを受け取ると、制御部３００は、リード対象範囲とテーブル３２３のビットマップ６１０４とを比較する（ステップ１６１０）。リード対象領域に未コピーの領域が含まれている場合には

、制御部３００は、リード対象領域のデータを優先して読み出してコピーを済ませる（ステップ１６３０）。制御部３００は、ビットマップ６１０４のリード対象領域に対応するビットをコピー済みに更新する（ステップ１６４０）。制御部３００は、ディスクアレイ２００内のコピーしたデータをホスト１００に転送する（ステップ１６５０）。リード対象領域がすべてコピー済であれば、制御部３００は、ディスクアレイ２００内のコピー済みのデータをホスト１００に転送する（ステップ１６５０）。

【００９５】

制御部３００は、ＦＣインタフェース２６０を介してライトコマンドを受け取ると、ホスト１００から転送されたデータについて、ライト対象領域にライトを行う（ステップ１６７０）。制御部３００は、コピー領域管理テーブル３２３のビットマップ６１０４のライト対象領域に対応するビットをコピー済に更新する（ステップ１６８０）。制御部３００は、残りの未コピー領域のコピーを継続する（ステップ１６９０）。

【００９６】

以上の処理により、再配置先ＬＵが存在するディスクアレイ２００は、データ再配置におけるコピー処理中であっても、ホスト１００からのリード/ライトを処理することができる。

【００９７】

なお、このリード/ライトの処理の際、制御部３００は、同時に、先に説明した使用状況取得処理も行う。

【００９８】

また、ホスト１００のマネージャ１３０は、データ再配置におけるコピー処理中に、ディスクアレイ２００のコマンドボリュームにコピー進捗取得のためのパラメータを含むデータのライトコマンドを発行し、ディスクアレイ２００が用意したデータをリードすることで、コピーの進捗情報等をディスクアレイ２００に問い合わせることができる。

【００９９】

具体的には、コマンドボリュームに対するライトコマンドを受け付けた制御部３００は、コマンドボリュームにライトされたパラメータを確認する。制御部３００は、コピー領域管理テーブル３２３を参照してパラメータに対応するコピーの進捗率などの情報をメモリ３２０上に用意し、ライト完了をホスト１００に通知する。マネージャ１３０は、コマンドボリュームに対するリードを行う。制御部３００は、ホスト１００のリードに対して、メモリ３２０上に用意したデータを転送することによって、コピーの進捗等の問い合わせに答える。

【０１００】

本実施形態によれば、複数のディスクアレイ２００間におけるＬＵの再配置によるデータの適正配置を、アプリケーション１４０にとって再配置前後で論理的に等価となるように、すなわち、アクセス対象のアクセスにアプリケーションが使用すべき論理位置名が変化しないようにしつつ実現できる。

【０１０１】

また、本態様によれば、計算機は、ディスクアレイから取得した、各記憶領域の物理的な記憶装置資源の使用状況を、例えば記憶装置資源の負荷分散等の観点による、データの適正配置の決定に用いることができる。したがって、この情報を用いて、例えば異なるストレージサブシステム間でデータを再配置することにより、データの適正配置を行うことができる。

【０１０２】

なお、本実施形態では、複数のディスクアレイ２００間におけるデータの再配置について説明した。しかし、再配置対象データを格納するストレージサブシステムは、ディスクアレイサブシステムでなくてもよい。磁気ディスク装置、光磁気ディスク装置、磁気テープ装置又は半導体ディスク装置などを用いた他の種類のストレージサブシステムであってもよい。

【０１０３】

尚、マネージャ１３０は、ＦＣ６００経由ではなく、ネットワーク７００経由で、例えばSimple Network Management Protocol（ＳＮＭＰ）で規定されているプロトコルを用いて情報の収集や指示を行ってもよい。

【０１０４】

本実施形態では、ディスクアレイ２００の制御部３００が取得する論理ボリューム使用状況３２２が使用時間の累積値である場合について説明した。しかし、制御部３００が単位時間毎の使用時間を使用率の形式にしてメモリ３２０に蓄積し、これを論理ボリューム使用状況３２２として、ホスト１００のマネージャ１３０が収集するようにしてもよい。

【０１０５】

図１５は、本発明が適用された計算機システムの第２実施形態の構成を示す図である。

【０１０６】

図示するように、本実施形態の計算機システムは、ローカルディスク１９０にＬＵ領域範囲テーブル１９２を格納し、スイッチ５００にコピー制御部５１０を設けた構成を有している点が、第１実施形態の計算機システムと異なる。

【０１０７】

本実施形態では、ディスクアレイ２００がディスク装置２１０の使用状況を取得し、ホスト１００が複数のディスクアレイ２００から使用状況を収集し、使用状況を計算機システムのファイルに基づく分析も含めてユーザに提示する。ホスト１００は、ファイル管理のためのデータ（以下「メタデータ」と称する）を変更する。スイッチ５００は、ホスト１００の指示に基づいて、ディスクアレイ２００に格納されているデータをコピーする。これにより、複数のディスクアレイ２００間におけるファイルの再配置を可能とし、データの適正配置を行えるようにする。

【０１０８】

第１実施形態においては、ホスト１００のファイルシステム１１０は、各ＬＵを、使用中のものと使用していないものとに区別して管理した。本実施形態では、ファイルシステム１１０は、全てのＬＵを使用し、全てのＬＵの領域の集合を単一領域（以下、「統合領域」と称する。）として管理する。また、統合領域上のファ
イルを、後述するメタデータで管理する。メタデータは、統合領域の既定の位置に格納される。

【０１０９】

図１６は、本実施形態において、ファイルシステム１１０が統合領域を管理するために用いるＬＵ領域範囲テーブル１９２の例を示した図である。ＬＵ領域範囲テーブル１９２は、統合領域の範囲と各ＬＵ内領域の範囲との対応を示す情報を保持している。

【０１１０】

図中、領域内アドレス６３０１には、統合領域内でのアドレスが格納される。ＬＵ番号６３０２は、ディスクアレイ番号、ＩＤ及びＬＵＮを含み、領域内アドレス６３０１に格納されるＬＵを示す。ＬＵ内アドレス６３０３は、対応するＬＵ番号６３０２で特定されるＬＵ内でのアドレスが格納される

図１７は、ホスト１００がリード/ライトを行う場合の処理の手順を示すフロー図である。

【０１１１】

前提として、ホスト１００のアプリケーション１４０は、ファイルシステム１１０が管理するファイルの論理位置を指定して、ディスクアレイ２００が格納するデータにリードやライトを行うものとする。また、ファイルシステム１１０は、データをファイルとして管理するために、メタデータをディスクアレイ２００に格納している。

【０１１２】

なお、メタデータはディスクアレイ２００に格納されているが、ファイルシステム１１０の管理に基づき、ホスト１００上のメモリにキャッシュされている場合もある。以下、メタデータがホスト１００上のメモリにキャッシュされている場合で説明する。

【０１１３】

図１８は、メタデータの内容を示す図である。

【０１１４】

図示するように、メタデータには、各ファイルの作成日時、更新日時、アクセス日時、属性、ファイル論理位置名、セキュリティ情報、及びファイル位置等が含まれる。各ファイルに対応する統合領域内の範囲は、ファイル位置６４０７に格納された情報で示される。

【０１１５】

ホスト１００のアプリケーション１４０は、ファイル論理位置名によってファイルを指定し、ファイルに対するリード/ライトをＯＳ１２０に要求する（ステップ１７００）。ＯＳ１２０は、ファイルシステム１１０に、ファイルのリード/ライトを要求する（ステップ１７１０）。ファイルシステム１１０は、キャッシュメモリされているメタデータを参照し、メタデータ及びＬＵ領域範囲テーブル１９２の情報から、指定されたファイルの位置（ＬＵおよびＬＵ内アドレス）を得る（ステップ１７２０）。

【０１１６】

要求がライト要求である場合、ファイルシステム１１０は、さらにメタデータの更新を行う（ステップ１７４０）。ファイルシステム１１０は、ステップ１７２０で得たファイルの位置が示す領域内のリード/ライトをディスクアレイ２００に対して行い（ステップ１７５０）、キャッシュされたメタデータ及びディスクアレイ２００のメタデータを更新する（ステップ１７６０）。

【０１１７】

ステップ１７４０及び１７６０でのメタデータの更新は、アクセスされたファイルについて、作成日時６４０１、更新日時６４０２、アクセス日時６４０３、属性６４０４、ファイル論理位置名６４０５、セキュリティ情報６４０６０、及びファイル位置６４０７等に格納された情報を、アクセス内容に応じて更新することで行われる。例えば、ライトによりファイルサイズが増減する場合は、これに合わせて、メタデータのファイル位置６４０７が示す領域内の範囲が増減される。また、ファイルが新規に作成される場合は、メタデータに新規ファイルのエントリが追加され、ファイルが削除される場合は対応するエントリが削除される。

【０１１８】

本実施形態において、制御部３００は、第１実施形態と同様の使用状況取得処理を行う。また、ホスト１００のマネージャ１３０は、第１実施形態と同様の使用状況収集処理を行う。

【０１１９】

図１９は、ホスト１００が行うファイル単位の再配置対象決定処理の手順を示すフロー図である。

【０１２０】

ホスト１００のマネージャ１３０は、統合領域に存在する各ファイルについて、ファイルとＬＵとの対応を、ファイルシステム１１０に問い合わせる（ステップ１８００）。ファイルシステム１１０は、キャッシュされたメタデータ及びＬＵ領域範囲テーブル１９２を用いて、問い合わせに答える（ステップ１８１０）。

【０１２１】

マネージャ１３０は、ディスクアレイ２００毎の各論理ボリュームの使用状況、各ＬＵの各論理ボリュームの使用状況及びファイル毎の各論理ボリュームの使用状況等を計算する。この計算には、INQUIRYコマンドによって得られた各ディスクアレイ２００における各ＬＵに属する論理ボリューム番号、使用状況収集処理で得られた各ディスクアレイ２００における論理ボリューム情報及びパリティグループ情報及び論理ボリュームの使用状況等が使用される（ステップ１８２０）。マネージャ１３０は、計算結果を、各論理ボリュームが属するパリティグループ２２０の属性等と共にユーザに提示する。すなわち、ホスト１００は、使用状況に関する情報を、ディスクアレイ２００、論理ボリューム、ＬＵ、ファイルといった各種の視点でユーザに提供する（ステップ１８３０）。

【０１２２】
マネージャ１３０は、各ディスクアレイ２００が提供するＬＵや論理ボリュームについて利用可能な空き領域を計算し、ユーザに提示する（ステップ１８４０）。マネージャ１３０は、各ディスクアレイ２００が提供するＬＵや論理ボリュームについて利用可能な空き領域を、ファイルシステム１１０に問い合わせる（ステップ１８５０）。ファイルシステム１１０は、キャッシュされたメタデータ及びＬＵ領域範囲テーブル１９２を参照して、ファイルが存在しない空き領域を特定し、マネージャ１３０に答える（ステップ１８６０）。マネージャ１３０は、使用状況収集処理で得た各種使用状況等から、空き領域の論理ボリュームの使用状況等を、論理ボリュームやパリティグループ２２０の属性等と共にユーザに分類して提示する（ステップ１８７０）。

【０１２３】
使用状況や空き領域の情報は、ホスト１００またはホスト１００にネットワークで接続された他の計算機で表示することができる。ユーザは、これらの情報より再配置すべきファイルと再配置先の空き領域とを決定する。マネージャ１３０は、これらの情報から、自動的に同様の再配置対象や空き領域を決定してもよい（ステップ１８８０）。

【０１２４】
ホスト１００のファイルシステム１１０が、ＯＳ１２０やアプリケーション１４０からの各ファイルへのリード/ライト要求頻度（アクセス頻度）を監視して統計情報を生成し、ステップ１８３０でユーザに提示するようにしてもよい。

【０１２５】
これにより、ユーザは、ホスト１００での各ファイルのアクセス頻度を勘案して再配置すべきファイルを決定することができる。

【０１２６】
図２０は、ホスト１００が、再配置対象決定処理の結果を受けて行う再配置処理の手順を示すフロー図である。本処理は、基本的には、図９に示すＬＵ単位の再配置決定処理の手順において、ＬＵをファイルに、ディスクアレイ２００をスイッチ５００に読み替えた処理と同じである。以下、図９とは異なる部分についてのみ説明する。

【０１２７】
マネージャ１３０は、ファイルシステム１１０に、再配置先の空き領域についての領域の使用予約を指示する（ステップ１９４０）。ファイルシステム１１０は、指定された再配置先領域が確保されるよう、キャッシュされたメタデータを更新する（ステップ１９５０）。マネージャ１３０は、ファイルシステム１１０に、メタデータのキャッシュメモリのフラッシュを指示する（ステップ１９６０）。ファイルシステム１１０は、ホスト１００上のメモリにキャッシュメモリしてあるメタデータを、ディスクアレイ２００にライトする（ステップ１９７０）。

【０１２８】
マネージャ１３０は、メタデータを書き換え、指定されたファイルの位置を、再配置元領域から再配置先領域へ入れ替える。これにより、再配置元の領域を空き領域とする（ステップ２０１０）。マネージャ１３０は、ファイルシステム１１０に、メタデータについて、キャッシュの無効化を指示する（ステップ２０２０）。ファイルシステム１１０は、ホスト１００上のメモリにキャッシュしてあるメタデータを無効にする（ステップ２０３０）。

【０１２９】
以降、ファイルシステム１１０が、ファイルにリード/ライトする場合には、再配置先領域にコピーされたデータに対して正常にリード/ライトを行うことができる。

【０１３０】
本実施形態によれば、複数のディスクアレイ２００間でのファイルの適正配置を、アプリケーション１４０にとって再配置前後で論理的に等価となるように行うことが可能となる。

【０１３１】

図２１は、本発明が適用された計算機システムの第３実施形態の構成を示す図である。

【０１３２】

本実施形態の計算機システムは、クライアント８００が、ＦＣインタフェース８６０及びネットワークインタフェース８７０を有する。そして、クライアント８００がＦＣインタフェース８６０を介してＦＣ６００経由でホスト１００、ディスクアレイ２００及びスイッチ５００に接続され、かつネットワークインタフェース８７０を介してネットワーク７００経由でホスト１００およびディスクアレイ２００に接続される点が、第２実施形態の計算機システムと異なる。本実施形態では、複数のクライアント８００とホスト１００とが、ディスクアレイ２００上のファイルを共有する。クライアント８００は、ＯＳ８２０とアプリケーション８４０を有する。クライアント８００は一般的な電子計算機である。

【０１３３】

第２実施形態と同様に、本実施形態のファイルシステム１１０は、全てのＬＵを使用し、全てのＬＵの領域を集合して単一の統合領域として管理する。そして、統合領域上のファイルを、第２実施形態と同様にメタデータにより管理する。

【０１３４】

クライアント８００が、ディスクアレイ２００に格納されているファイルへアクセスする処理について説明する。

【０１３５】

図２２は、クライアント８００がディスクアレイ２００に格納されているファイルのリードを行う場合の処理の手順を示すフロー図である。

【０１３６】

クライアント８００のアプリケーション８４０は、ＯＳ８２０にファイルのリードを要求する（ステップ２１００）。ＯＳ８２０は、ネットワークインタフェース８７０あるいはＦＣインタフェース８６０を介して、ホスト１００のファイルシステム１１０にファイルのリードを通知する（ステップ２１１０）。

【０１３７】

ファイルのリードの通知を受けたファイルシステム１１０は、ファイルが格納されているＬＵおよびＬＵ内アドレスを、メタデータとＬＵ領域範囲テーブル１９２とを参照して求める（ステップ２１２０）。ファイルシステム１１０は、ファイルが格納されているＬＵのＬＵ内アドレスを他のクライアント８００からのライトに対してロックする（ステップ２１３０）。ファイルシステム１１０は、ホスト１００のキャッシュメモリにあるメタデータをフラッシュする（ステップ２１４０）。ファイルシステム１１０は、クライアント８００のＯＳ８２０に、ファイルが格納されているＬＵおよびＬＵ内アドレスと、メタデータの格納されているＬＵおよびＬＵ内アドレスとを返答する（ステップ２１５０）。

【０１３８】

返答を受けたクライアント８００のＯＳ８２０は、リードの対象となるファイルが格納されているＬＵが存在するディスクアレイ２００に対し、ＦＣインタフェース８６０を介して、ファイルが格納されているＬＵ内アドレスに対するリードを行って、アプリケーション８４０からの要求を処理する（ステップ２１６０）。

【０１３９】

クライアント８００から要求されたデータのリード処理が終了したら、ＯＳ８２０は、ホスト１００のファイルシステム１１０から通知されたＬＵおよびＬＵ内アドレスにあるメタデータ上のファイルのアクセス日時を更新する（ステップ２１７０）。ＯＳ８２０は、ファイルシステム１１０に、ネットワークインタフェース８７０またはＦＣインタフェース８６０を介して、処理の完了を通知する（ステップ２１８０）。

【０１４０】

完了通知を受けたファイルシステム１１０は、ホスト１００上のメタデータのキャッシュメモリを無効化し（ステップ２１９０）、ステップ２１３０で行ったロックを解除する（ステップ２２００）。

【０１４１】
図２３は、ライトを行う場合の処理の手順を示すフロー図である。
【０１４２】
ライト処理は、図２２のリード処理において、リードをライトに置き換えた処理とほぼ同一である。以下、異なる部分について説明する。
【０１４３】
ファイルシステム１１０は、ライトで増加する可能性のあるファイル使用領域のための領域の予約をメタデータに記述する（ステップ２３４０）。ファイルシステム１１０は、クライアント８００のＯＳ８２０に、ファイルが格納されているＬＵおよびＬＵ内アドレス（ライトで増加する可能性のあるファイル使用領域のために予約した領域を含める）と、メタデータが格納されているＬＵ及びＬＵ内アドレスとを返答する。なお、ライトで増加する可能性のあるファイル使用領域の増加量は、クライアント８００のＯＳ８２０からのライトの通知に含まれているものとする（ステップ２３６０）。
【０１４４】
返答を受けたＯＳ８２０は、ライトの対象となるファイルが格納されているＬＵが存在するディスクアレイ２００に対し、ＦＣインタフェース８６０を介して、ファイルが格納されているＬＵ内アドレスに対するライトを行い、アプリケーション８４０からの要求を処理する（ステップ２３７０）。
【０１４５】
このようにして、クライアント８００のアクセスを処理することにより、クライアント８００およびホスト１００は、ディスクアレイ２００に格納されているファイルを矛盾なく共有して使用することができる。なお、ホスト１００自身のファイルアクセスも、クライアント８００によるファイルアクセスと同様に処理される。
【０１４６】
次に、本実施形態でのファイルの再配置について説明する。
【０１４７】
本実施形態でのファイルの再配置に関する処理（使用状況取得処理、使用状況収集処理、再配置対象決定処理および再配置処理）は、第２実施形態と同様である。ただし、アプリケーション８４０が要求するデータのリード/ライト処理でファイルがロックされている間、再配置処理は実行されない。また、図２０に示す再配置処理のステップ１９２０及び１９３０におけるファイルのキャッシュメモリのフラッシュと、ディスクアレイへ２００への書き戻しは、ファイルシステム１１０がそのファイルをキャッシュメモリしているクラインアント８００に対して指示し、これを行わせる。
【０１４８】
本実施形態によれば、ディスクアレイ２００に格納されているデータを共有して使用する環境においても、複数のディスクアレイ２００間におけるファイルの物理的な再配置を、アプリケーション１４０、８４０に対して、再配置前後で論理的に等価となるように行うことができる。
【０１４９】
本実施形態においても、ホスト１００のファイルシステム１１０が、ＯＳ１２０、８２０やアプリケーション１４０、８４０からの各ファイルへのリード/ライト要求頻度を監視し、統計情報を生成して、再配置対象決定処理においてユーザに提示するようにしてもよい。
【０１５０】
本実施形態において、クライアント８００上にマネージャ１３０のプログラムが格納され、そのマネージャ１３０が、ＦＣインタフェース８６０あるいはネットワークインタフェース８７０を用いて、使用状況等の情報の収集や指示などの処理を、ホスト１００のファイルシステム１１０やディスクアレイ２００に要求するようにしてもよい。
【０１５１】
図２４は、本発明が適用された計算機システムの第４実施形態の構成を示す図である。

【０１５２】
本実施形態の計算機システムは、ホスト１００がＬＵプールマネージャ９００及びＬＵ管理テーブル９１０を有する点で、第１実施形態の計算機システムと異なる。
【０１５３】
本実施形態によれば、ＬＵの再配置先の選択を容易にすることができる。
【０１５４】
図２５は、ＬＵ管理テーブル９１０を示す図である。
【０１５５】
ＬＵ管理テーブル９１０は、システム全体のＬＵの状態に関する情報が登録されているテーブルである。
【０１５６】
ＬＵ番号３３１０には、各ＬＵに一意に割り当てられた番号が登録される。この番号は、ＬＵプールマネージャ９００が各ＬＵを管理するために使用される。サイズ３３２０には、対応するＬＵの容量が登録される。構成３３３０には、ＲＡＩＤ構成の種別が格納される。構成３３３０には、ＬＵがキャッシュメモリ３３０や単体ディスクで構成されている場合には、その情報も格納される。
【０１５７】
状態３３４０には、ＬＵの状態を示す情報が格納される。その種別として、「オンライン」、「オフライン」、「未実装」及び「障害オフライン」が設けられている。「オンライン」は、ＬＵが正常な状態であり、ホスト１００からアクセス可能であることを示す。「オフライン」は空きＬＵ、すなわちＬＵは正常に存在するが、ホスト１００からはアクセス不能の状態におかれていることを示す。「未実装」は、このＬＵは定義されておらず、ホスト１００からアクセス不能であることを示す。「障害オフライン」は、ＬＵに障害が発生してホスト１００からのアクセスができないことを示す。
【０１５８】
ディスクアレイ番号３３５０には、対応するＬＵが存在するディスクアレイ２００を示す情報が格納される。
【０１５９】
パス３３６０には、各ディスクアレイ２００に複数接続するＦＣ６００のどれにＬＵが割り当てられているかを示す番号が格納される。ＩＤ３３７０及びＬＵＮ３３８０には、ＬＵを示す番号が格納される。
【０１６０】
ディスク性能３３９０には、対応するＬＵが現在配置されているディスク装置２１０の性能を示す指標が格納される。具体的には、図２５に示すとおり、ディスク装置２１０の平均シーク時間、平均回転待ち時間及び構成から、ディスク装置２１０の性能が高性能、中性能、低性能の指標に分類されて格納されている。キャッシュメモリ上のＬＵは、超高性能に分類される。
【０１６１】
エミュレーションタイプ３４００には、ディスクアレイ２００がホスト１００に提供する各ＬＵのディスク装置としての型を示す情報が格納される。
【０１６２】
再配置可能フラグ３４１０には、ＬＵの再配置を行う際に、ＬＵの再配置先として使用できるか否かを指定するためのフラグが格納される。ユーザは、このフラグを用いて再配置用のＬＵとその他のＬＵを区別することができる。ユーザはフラグのオン／オフを変更することができる。
【０１６３】
図２５は、ディスクアレイ番号０についてのＬＵ管理テーブルを示す図である。マネージャ１３０は、すべてのディスクアレイ２００についてのＬＵ管理テーブルを保持している。
【０１６４】

本実施形態における再配置対象の決定は、以下のようにして行われる。

【０１６５】

ユーザは、マネージャ１３０に対して、再配置元ＬＵの指定及び再配置先ＬＵとして必要とされる条件を指定する。具体的な条件としては、性能条件や信頼性レベル等がある。

【０１６６】

例えば、あるＬＵが過度に使用され、そのＬＵを含むディスク装置の能力を超えて負荷がかかっている場合、そのＬＵの再配置先としてより高性能のディスク装置を指定すれば、ＬＵの処理能力が増大し、計算機システムの性能向上が期待できる。

【０１６７】

又、重要なデータを格納しているＬＵが単体ディスクや冗長なしＲＡＩＤ（ＲＡＩＤ０）上に存在する場合、再配置先としてＲＡＩＤ５やＲＡＩＤ１を指定すれば、冗長性による耐障害性を確保できる。

【０１６８】

マネージャ１３０は、ＬＵ管理テーブル９１０に登録された情報を用いて再配置先のＬＵを決定し、ユーザに通知した上で、ＬＵの再配置を行う。

【０１６９】

図２６は、本実施形態における再配置対象決定処理の手順を示すフロー図である。本処理は、ユーザの指示に対応して実行される。

ユーザは、マネージャ１３０に対して再配置元ＬＵのディスクアレイ番号、パス、ＩＤ及びＬＵＮを指定する。この場合、パス及びＩＤ等の代わりに、ディスクアレイ番号及びＬＵ番号を指定してもよい（ステップ２５００）。

【０１７０】

ユーザは、マネージャ１３０に対して、再配置先についての要求条件として性能条件や信頼性レベルを指定する（ステップ２５１０）。

【０１７１】

マネージャ１３０は、再配置元ＬＵ、および再配置先についての要求条件をＬＵプールマネージャ９００に通知する（ステップ２５２０）。ＬＵプールマネージャ９００は、ＬＵ管理テーブル９１０内を検索して、要求された条件を満たすＬＵの有無を確認する（ステップ２５３０）。

【０１７２】

この場合、検索条件は、「状態がオフライン」かつ「サイズが再配置元ＬＵ以上」かつ「エミュレーションタイプが再配置元ＬＵと同じ」かつ「再配置可能フラグがオン（真）すなわち可能」かつ「性能条件が要求を満たす」かつ「信頼性レベルが要求を満たす」でなければならない。

【０１７３】

ステップ２５４０において条件を満たすＬＵが存在した場合、ＬＵプールマネージャは、該当するＬＵをマネージャ１３０に通知する（ステップ２５５０）。マネージャ１３０は、通知されたＬＵを再配置先ＬＵとして決定する（ステップ２５６０）。

【０１７４】

ステップ２５４０で条件を満たすＬＵが存在しなかった場合、ＬＵプールマネージャ９００は、ＬＵ管理テーブル９１０内を検索して「状態が未実装」のＬＵ番号３３１０を探す（ステップ２５７０）。

【０１７５】

未実装のＬＵ番号３３１０が存在しなかった場合は、ＬＵプールマネージャ９００は、マネージャ１３０に条件を満たすＬＵの利用不可を通知する（ステップ２５８０）。通知を受けたマネージャ１３０は、ユーザに再配置先ＬＵ決定不可を通知する（ステップ２５９０）。

【０１７６】

ステップ２５７０で未実装のＬＵが存在した場合は、ＬＵプールマネージャ９００は、未実装のＬＵ番号と再配置先ＬＵについての条件を指定して、該当するディスクアレイ２

００に再配置先ＬＵの構築を指示する（ステップ２６００）。

【０１７７】

この場合の再配置先ＬＵについての条件は、「サイズが再配置元ＬＵ以上」かつ「エミュレーションタイプが再配置元ＬＵと同じ」かつ「性能条件が要求を満たす」かつ「信頼性レベルが要求を満たす」である。

【０１７８】

ＬＵの構築を指示されたディスクアレイ２００は、ＬＵ構築処理を行う（ステップ２６１０）。構築が成功した場合は、ディスクアレイ２００は、ＬＵプールマネージャ９００に、構築したＬＵについてのディスクアレイ番号、パス、ＩＤ及びＬＵＮなどを含む一連の情報を通知する（ステップ２６２０）。構築が失敗した場合には、ディスクアレイ２００は、ＬＵプールマネージャ９００に構築不可の通知を行う（ステップ２６１０）。

【０１７９】

ＬＵプールマネージャ９００は、通知されたＬＵの情報をＬＵ管理テーブル９１０に登録し（ステップ２６３０）、マネージャ１３０に通知する（ステップ２５５０）。マネージャ１３０は、このＬＵを再配置先ＬＵとして決定する（ステップ２５６０）。

【０１８０】

構築不可の通知を受けたＬＵプールマネージャ９００は、マネージャ１３０に条件を満たすＬＵの利用不可を通知する（ステップ２５８０）。通知を受けたマネージャ１３０は、ユーザに再配置先ＬＵ決定不可を通知する（ステップ２５９０）。

【０１８１】

図２７は、ディスクアレイ２００が行うＬＵ構築処理の手順を示すフロー図である。この処理は、ＬＵプールマネージャ９００の指示を受けた時に行われる　ディスクアレイ２００は、ＬＵプールマネージャ９００からの指示により、未実装のＬＵ番号と再配置先ＬＵについての条件を受け取る（ステップ２７００）。

【０１８２】

ディスクアレイ２００は、ディスク装置２１０やキャッシュメモリ３３０などの内部資源割り当て状況等と受け取った条件を比較して、要求された条件のＬＵが構築可能かどうかを判断する（ステップ２７１０）。ＬＵが構築可能な場合は、ディスクアレイ２００は、内部資源を割り当て、フォーマット／初期化処理を行ってＬＵを構築する。ディスクアレイ２００は、構築したＬＵに、ＬＵプールマネージャ９００から受けとった未実装のＬＵに対応するＬＵ番号を割り当てる（ステップ２７２０）。

【０１８３】

ディスクアレイ２００は、ＦＣインタフェース２６０を設定し、ＬＵにパス、ＩＤ、ＬＵＮを割り当てる（ステップ２７３０）。ディスクアレイ２００は、構築したＬＵについての、ディスクアレイ番号、パス、ＩＤ及びＬＵＮ等を含む一連の情報をＬＵプールマネージャ９００に通知する（ステップ２７４０）。

【０１８４】

ステップ２７１０においてＬＵが構築不可能だった場合は、ディスクアレイ２００は、構築不可をＬＵプールマネージャ９００に通知する（ステップ２７５０）。

【０１８５】

再配置先ＬＵが決定されたら、マネージャ１３０は、第一の実施の形態と同様に再配置元ＬＵと再配置先ＬＵについての再配置処理を行う。

【０１８６】

図２８は、再配置元ＬＵのオフライン化処理の手順を示すフロー図である。

【０１８７】

マネージャ１３０は、第一の実施の形態で説明した方法でコピーの進捗を取得し、コピーが終了した場合は、ＬＵプールマネージャ９００に再配置元ＬＵのオフライン化を指示する（ステップ２８００）。

【０１８８】

オフライン化の指示を受けたＬＵプールマネージャ９００は、再配置元ＬＵのディスク

アレイ２００に、再配置元ＬＵのオフライン化を指示する（ステップ２８１０）。オフライン化の指示をうけたディスクアレイ２００は、ＦＣインタフェース２６０を設定してＬＵにロックをかけることで、ＬＵをオフラインにする（ステップ２８２０）。ディスクアレイ２００は、オフライン化したことをＬＵプールマネージャ９００に通知する（ステップ２８３０）。

【０１８９】

オフライン化の通知を受けたＬＵプールマネージャは、ＬＵ管理テーブル９１０のＬＵの状態３３４０の内容をオフラインに更新する（２８４０）。

【０１９０】

ここでは、マネージャ１３０がコピーの進捗情報を取得する例を説明したが、ディスクアレイ２００がコピー終了をマネージャ１３０に通知してもよい。

【０１９１】

また、マネージャ１３０がオフライン化を指示する代わりにディスクアレイ２００がコピー終了時点で再配置元ＬＵをオフライン化し、オフライン化したことをＬＵプールマネージャ９００に通知してもよい。

【０１９２】

本実施形態においては、マネージャ１３０が、ＳＣＳＩ規格のEXTENDED COPYコマンドを用いてスイッチ５００のコピー制御部５１０へのコピー指示を行う場合について説明したが、他のコマンドを用いてもよい。他のコマンドとは、例えば、コマンドボリュームへのライトコマンド等である。また、図１５に示すように、ディスクアレイ２００がコピー制御部５１０を有し、マネージャ１３０がディスクアレイ２００のコピー制御部５１０に、コピー指示を行って、ディスクアレイ２００がコピー処理を行うようにしてもよい。

【０１９３】

本実施形態では、再配置先ＬＵとして要求する条件などの情報はユーザが指定しているとしたが、マネージャ１３０が自動的に判断して指定してもよい。

【０１９４】

本実施形態では、ＬＵプールマネージャ９００とマネージャ１３０が同じホスト１００に存在するとしたが、ＬＵプールマネージャ９００がリモートコンピュータ４００といった、マネージャ１３０とは異なるコンピュータに存在してもよい。この場合、ＬＵプールマネージャ９００とマネージャ１３０は、ＦＣ６００やネットワーク７００を介してＳＣＳＩやＳＮＭＰや他のプロトコルやコマンド体系で指示や通知を行う。

【０１９５】

本実施形態によれば、ＬＵの再配置の処理において、再配置先のＬＵの管理や選択を容易にしてユーザの負荷を削減し、計算機システムの管理を容易にすることができる。

【０１９６】

図２９は、本発明を適用した計算機システムの第５実施形態を示す図である。本実施形態の計算機システムは、ＬＵ領域範囲テーブル１９２に新たな項目を付加したＬＵ領域範囲テーブル１９３を用いて、クライアント８００からのリード／ライト要求に基づき、ホスト１００がファイルをリード／ライトする。そして、クライアント８００との間でネットワーク７００を介してデータを転送する処理を行う点が、第３実施形態の計算機システムと異なる。

ネットワーク７００を経由したファイル共有のためのプロトコルとしては、Network File System（ＮＦＳ）やCommon Internet File System（ＣＩＦＳ）が広く用いられている。これらのプロトコルや広く普及しているネットワーク７００を用いることにより、容易にファイル共有環境を実現することができる。本実施形態においても、ＮＦＳ又はＣＩＦＳを使用することを考える。

【０１９７】

図３０は、ＬＵ領域範囲テーブル１９３を示す図である。ＬＵ領域範囲テーブル１９３には、ＬＵ領域に対するアクセスがネットワークを使用するか否かに関する情報が格納される。

【０１９８】
使用種別３５１０には、ＬＵ領域が、リード／ライトの処理が第３実施形態のようにＦＣ６００を介して行われるＬＵ領域であるか、本実施形態で説明するようにネットワーク７００を介してリード／ライトの処理が行われるＬＵ領域であるかを示す情報が格納される。
【０１９９】
使用種別３５１０には、ＬＵ領域が、第１実施形態のようにＬＵを再配置する構成および方法に用いる領域（この場合のリード／ライト要求はＦＣ６００を経由する）であるか、ＬＵへのリード／ライトの処理が、ネットワーク７００を介して行う領域であるかの情報を格納することができる。使用種別３５１０には、未使用領域の情報を格納することもできる。その他、領域内アドレス、ディスクアレイ番号、ＩＤ、ＬＵＮ、ＬＵ内アドレスは、第３実施形態で説明したものと同様なので、説明を省略する。
【０２００】
ＬＵ領域範囲テーブル１９３を用いてＬＵを集中して管理することで、ファイルシステム１１０は、ＬＵを少なくとも使用種別毎に区別された、複数の領域として管理することができる。
【０２０１】
ＬＵ領域範囲テーブル１９３が設定されることによって、ホスト１００は、クライアント８００からの要求が、第３実施形態で説明した方法でのアクセスか、ネットワークを介した形でのアクセスかを、要求で用いられるプロトコルなどで区別する。ホスト１００は、この種別に応じて、ＬＵ領域を使用種別３５１０毎に区別して扱う。
【０２０２】
ホスト１００は、第３実施形態の方法でアクセスされるファイルおよび領域と、本実施形態の方法でアクセスされるファイルおよび領域を区別して処理する。したがって、同一のファイルおよび領域へアクセスする方法が混在することはない。
【０２０３】
ホスト１００は、アクセス可能なファイルの検索においても、同様の区別を行う。つまり、同一のディスクアレイ２００に存在する各ファイルに、クライアント８００からのアクセス要求があった場合、クライアント８００からの使用種別を識別することにより、クライアント８００の使用種別とは異なる他の使用種別のファイルをクライアント８００に対して返答しない。したがって、クライアント８００には、自己が使用するアクセス方法でのみアクセス可能なファイルだけが通知される。このことにより、本システムにおいては、共有ファイルの管理を容易に行うことができる。
【０２０４】
さらに、第１実施形態のように、ＬＵを再配置する構成および方法に用いる領域（リード／ライトはＦＣ６００経由）と、ＬＵへのリード／ライトをホスト１００およびネットワーク７００を介して行う領域との区別を行うことによって、上述したような効果をこれら全ての使用種別に対して得ることができる。又、ユーザは、ホスト１００またはリモートコンピュータ４００を介して、ＬＵ領域範囲テーブル１９３を自由に設定することができる。
【０２０５】
本実施形態では、ＮＦＳやＣＩＦＳのようなファイル共有プロトコルをネットワーク７００経由で使用し、ホスト１００とクライアント８００間のデータ転送をネットワーク７００経由で行うとしたが、ネットワーク７００の代わりに、ＦＣ８００経由で行う処理も考えられる。更に、クライアント８００が行う各ＬＵへのリード／ライト要求を、ホスト１００およびネットワーク７００を介して行う処理も考えられる。この場合、ホスト１００は、クライアント８００が要求するリード／ライト対象領域を、ＬＵ領域範囲テーブル１９２を用いて求める。ホスト１００は、対象となるデータをリードしてクライアント８００にネットワーク７００経由で転送する。あるいは、ホスト１００は、クライアント８００からデータをネットワーク７００経由で受領してライトする。

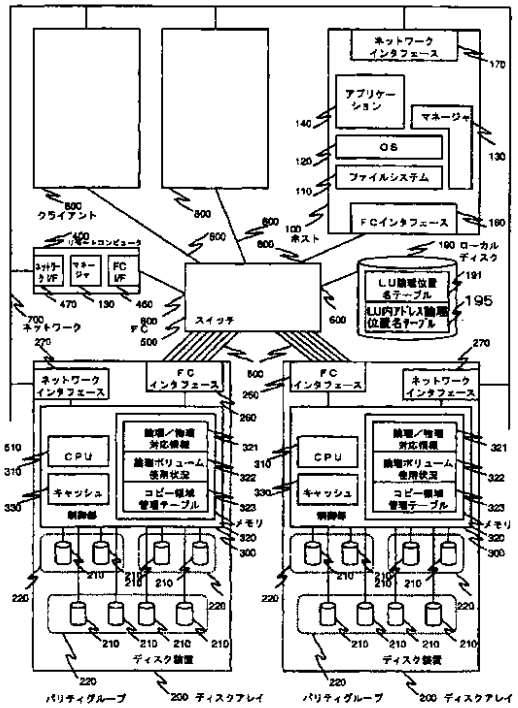
【０２０６】
図３１は、クライアント８００のアプリケーション８４０が、ディスクアレイ２００に格納されているファイルに対してリードを行う場合における、ホスト１００の処理の手順を示すフロー図である。
【０２０７】
第３実施形態と同様に、リード通知をうけたホスト１００のファイルシステム１１０は、ＬＵ領域範囲テーブル１９３とメタデータを参照することで、ファイルの格納されているＬＵ及びＬＵ内領域を求める（ステップ２９００）。ファイルシステム１１０は、他のライト要求に対してリード対象となるファイルをロックする（ステップ２９１０）。ファイルシステム１１０は、ファイル内のデータをリードして（ステップ２９２０）、クライアント８００にネットワーク７００を介してリードした内容を転送する（ステップ２９３０）。ファイルシステム１１０は、メタデータ上のファイルアクセス日時を更新する（ステップ２９４０）。ファイルシステム１１０は、ファイルのロックを解除し（ステップ２９５０）、リード処理の完了をクライアント８００に通知する（ステップ２９６０）。
【０２０８】
図３２は、アプリケーション８４０が、ライトを行う場合の処理の手順を示すフロー図である。
【０２０９】
ライト通知を受けたホスト１００は、ネットワーク７００経由でクライアント８００からライトデータを受け取る（ステップ３０００）。ホスト１００は、ＬＵ領域範囲テーブル１９３とメタデータを参照することで、ファイルの格納されているＬＵとＬＵ内領域を求める（ステップ３０１０）。ホスト１００は、ファイルをロックし（ステップ３０２０）、ファイルにデータをライトする。このとき、必要ならばメタデータを更新してファイル使用領域の追加を行う（ステップ３０３０）。
【０２１０】
ホスト１００は、メタデータ上のファイル更新日時とアクセス日時を更新する（ステップ３０４０）。ホスト１００は、ロックを解除し（ステップ３０５０）、ライト完了をクライアント８００に通知する（ステップ３０６０）。
【０２１１】
図３３は、クライアント８００のアプリケーション８４０またはＯＳ８２０が、アクセス可能なファイルの存在についてホスト１００に問い合わせた場合の処理の手順を示すフロー図である。
【０２１２】
アプリケーション８４０またはＯＳ８２０自身の要求があった時、ＯＳ８２０は、ネットワーク７００を介してホスト１００にアクセス可能なファイルの存在を問い合わせる（ステップ３１００）。
【０２１３】
通知をうけたホスト１００のファイルシステム１１０は、アクセス可能なファイルを、ＬＵ領域範囲テーブル１９３とメタデータを参照して求める（ステップ３１１０）。ファイルシステム１１０は、各ファイルのファイル名などの情報をクライアント８００に通知する（ステップ３１２０）。
【０２１４】
本実施形態では、クライアント８００およびホスト１００は、ディスクアレイ２００に格納されているファイルをホスト１００経由で共有して使用することができる。データの再配置の方法等は第３実施形態と同様である。ただし、再配置処理は各使用種別の領域内で行われる。
【０２１５】
本実施形態では、ディスクアレイ２００に格納されているデータを共有して使用する環境においても、アプリケーション１４０およびアプリケーション８４０が関与することなく、複数のディスクアレイ２００間でのファイルの物理的再配置を行うことができる。

【０２１６】
本発明は、各実施形態に限定されるものではなく、その要旨の範囲内で数々の変形が可能である。
【０２１７】
たとえば、図１、図１５および図２１に示すように、マネージャ１３０を、ネットワークインタフェース４７０とＦＣインタフェース４６０とを有するリモートコンピュータ４００上のプログラムとして、ホスト１００の外部に配置してもよい。ホスト１００外部のマネージャ１３０がＦＣ６００あるいはネットワーク７００経由で情報の収集や指示を行い、各実施形態と同様の処理を行って、複数のディスクアレイ２００間におけるＬＵの再配置によるデータの適正配置を、アプリケーション１４０に対して透過的に行うことができる。
【０２１８】
また、第１実施形態において、第３実施形態と同様に、ファイルの共有等を行うようにしてもよい。この場合も、記第１実施形態と同様に、複数のディスクアレイ２００間でのデータの物理的再配置を、アプリケーション１４０、８４０に対し、再配置前後で論理的に等価となるように透過的に行うことができる。
【０２１９】
以上説明したように、本発明によれば、ストレージサブシステム間におけるデータの再配置を容易に行うことが出来る。また、本発明によれば、ホストコンピュータが適正配置の決定に必要な情報を複数のストレージサブシステムから取得することができる。また、異なるストレージサブシステム間におけるデータの再配置を、アプリケーションにとってのデータ位置が、再配置の前後で変化しないように行うことができる。さらに、異なるストレージサブシステム間におけるファイルを対象とするデータの再配置を行うことができる。
【図面の簡単な説明】
【０２２０】
【図１】本発明の第１実施形態が適用された計算機システムの構成を示す図である。
【図２】本発明の第１実施形態でのリード/ライト処理および使用状況取得処理の手順を示すフロー図である。
【図３】本発明の第１実施形態で用いる論理/物理対応情報を示す図である。
【図４】本発明の第１実施形態で用いる論理ボリューム使用状況を示す図である。
【図５】本発明の第１実施形態での使用状況収集処理の手順を示すフロー図である。
【図６】本発明の第１実施形態で用いる論理ボリューム情報のパラメータを示す図である。
【図７】本発明の第１実施形態で用いる論理ボリューム情報を示す図である。
【図８】本発明の第１実施形態での再配置対象決定処理の手順を示すフロー図である。
【図９】本発明の第１実施形態での再配置処理の手順を示すフロー図である。
【図１０】本発明の第１実施形態で用いるＬＵ論理位置名テーブルを示す図である。
【図１１】本発明の第１実施形態で用いるＬＵ論理位置名テーブルを示す図である。
【図１２】本発明の第１実施形態でのコピー処理の手順を示すフロー図である。
【図１３】本発明の第１実施形態で用いるコピー領域管理テーブルを示す図である。
【図１４】本発明の第１実施形態での、図１６に示す処理によるコピー中における、再配置先ＬＵへのリード/ライトコマンドに対する処理の手順を示すフロー図である。
【図１５】本発明の第２実施形態が適用された計算機システムの構成を示す図である。
【図１６】本発明の第２実施形態で用いるＬＵ領域範囲テーブルを示す図である。
【図１７】本発明の第２実施形態でのリード/ライト処理の手順を示すフロー図である。
【図１８】本発明の第２実施形態で用いるメタデータを示す図である。
【図１９】本発明の第２実施形態での再配置対象決定処理の手順を示すフロー図である。
【図２０】本発明の第２実施形態での再配置処理の手順を示すフロー図である。
【図２１】本発明の第３実施形態が適用された計算機システムの構成を示す図である。

【図２２】本発明の第３実施形態において、クライアントのアプリケーションがファイルのリードを行う際の処理の手順を示すフロー図である。
【図２３】本発明の第３実施形態において、クライアントのアプリケーションがファイルのライトを行う際の処理の手順を示すフロー図である。
【図２４】本発明の第４実施形態が適用された計算機システムの構成を示す図である。
【図２５】本発明の第４実施形態でのＬＵ管理テーブル９１０を示す図である。
【図２６】本発明の第４実施形態での再配置対象決定処理の手順を示すフロー図である。
【図２７】本発明の第４実施形態でのＬＵ構築処理の手順を示すフロー図である。
【図２８】本発明の第４実施形態での再配置元ＬＵオフライン化処理の手順を示すフロー図である。
【図２９】本発明の第５実施形態が適用された計算機システムの構成を示す図である。
【図３０】本発明の第５実施形態でのＬＵ領域範囲テーブル１９３を示す図である。
【図３１】本発明の第５実施形態でのクライアント８００のアプリケーション８４０がファイルのリードを行う際の処理の手順を示したフロー図である。
【図３２】本発明の第５実施形態でのクライアント８００のアプリケーション８４０がファイルのライトを行う際の処理の手順を示すフロー図である。
【図３３】本発明の第５実施形態でのアクセス可能ファイル応答処理の手順を示すフロー図である。
【符号の説明】
【０２２１】
１００…ホスト、１１０…ファイルシステム、１２０、８２０…ＯＳ、１３０…マネージャ、１４０、８４０…アプリケーション、１６０、２６０、４６０、８６０…ＦＣインタフェース、１７０、２７０、４７０、８７０…ネットワークインタフェース、１９０…ローカルディスク、１９１…ＬＵ論理位置名テーブル、１９２…ＬＵ領域範囲テーブル、２００…ディスクアレイ、２１０…ディスク装置、２２０…パリティグループ、３００…制御部、３１０…ＣＰＵ、３２０…メモリ、３２１…論理/物理対応情報、３２２…論理ボリューム使用状況、３２３…コピー領域管理テーブル、３３０…キャッシュメモリ、４００…リモートコンピュータ、５００…スイッチ、５１０…コピー制御部、６００…Fibre Channel（ＦＣ）、７００…ネットワーク、８００…クライアント。

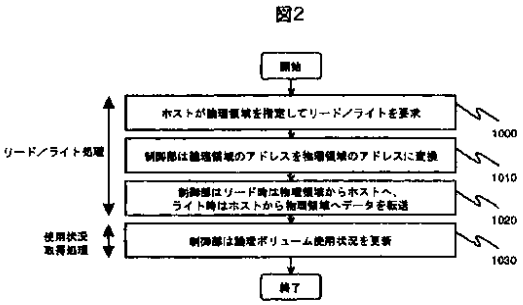
【図１】

図1



【図２】

図2



【図３】

図3

論理アドレス				物理アドレス							
						データ		パリティ			
5001	5002	5003	5004	5005	5006	5007	5006	5007			
LU番号	LU内アドレス	論理ボリューム番号	論理ボリュームアドレス	パリティグループ番号	ディスク装置番号	ディスク装置内アドレス	ディスク装置番号	ディスク装置内アドレス			
0	0~999	0	0~999	100	0	0~999	2	0~999			
0	1000~1999	0	1000~1999	100	1	0~999	2	0~999			
0	2000~2999	1	0~999	100	0	1000~1999	2	1000~1999			
0	3000~3999	1	1000~1999	100	1	1000~1999	2	1000~1999			
0	4000~4999	2	0~999	101	3	0~999	4	0~999			
1	0~999	3	0~999	100	0	0~999	6	0~999			
2	0~999	4	0~999	コマンドボリューム	-	-	-	-			

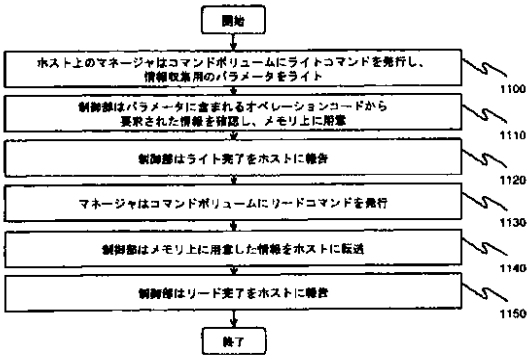
【図４】

図4

5101	5102					
論理ボリューム番号	ランダムリード処理によるバリティグループ占有時間(マイクロ秒単位)	ランダムライト(データ)処理によるバリティグループ占有時間(マイクロ秒単位)	ランダムライト(パリティ)処理によるバリティグループ占有時間(マイクロ秒単位)	シーケンシャルリード処理によるバリティグループ占有時間(マイクロ秒単位)	シーケンシャルライト(データ)処理によるバリティグループ占有時間(マイクロ秒単位)	シーケンシャルライト(パリティ)処理によるバリティグループ占有時間(マイクロ秒単位)
0	3141	101	121	59265	0	0
1	0	131	111	0	24360	17320

【図５】

図5



【図６】

図6

バイト	内容
0−1	論理ボリューム番号
2−4	Reserved
5	0xB6:オペレーションコード(論理ボリューム情報取得)
6−7	Reserved
8−511	Don't care

5201

【図７】

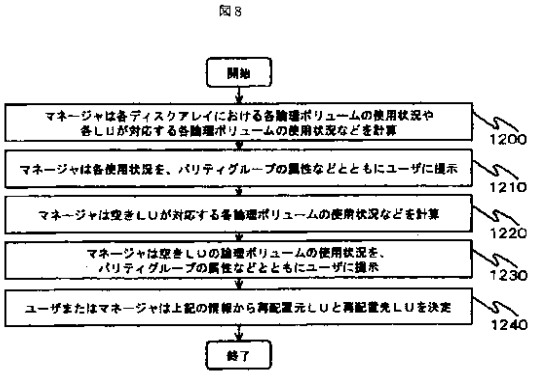
図7

バイト	内容
0−1	論理ボリューム番号
2−4	Don't care
5	0xB6:オペレーションコード(論理ボリューム情報取得)
6−7	Don't care
8−11	ディスクアレイ装置番号
12−43	ディスクアレイがホストに対しエミュレーションしてみせるボリューム形式(ボリュームエミュレーションタイプ)(コードで示す)
44−47	ボリューム容量(MB単位)
48	ボリュームがコピー先となれるか否かを示すフラグ
49	当該LUを構成する論理ボリューム数
51−121	当該LUを構成する論理ボリューム番号(2バイト毎)
122−511	Reserved

5201
5202
5203

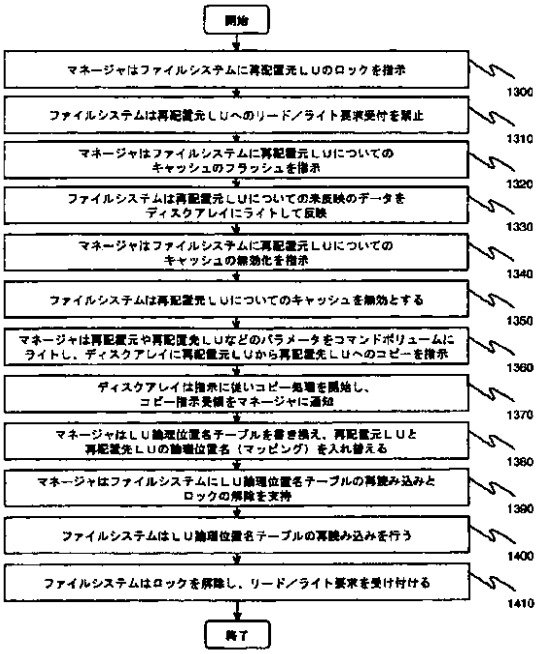
【図８】

図8



【図９】

図9



【図１０】

図10

6001			6002
LU番号			論理位置名（ディレクトリ）
ディスクアレイ番号	ID	LUN	
0	2	3	・・・/home1／prject1
1	4	5	・・・/temp
1	5	2	使用していない（空き）

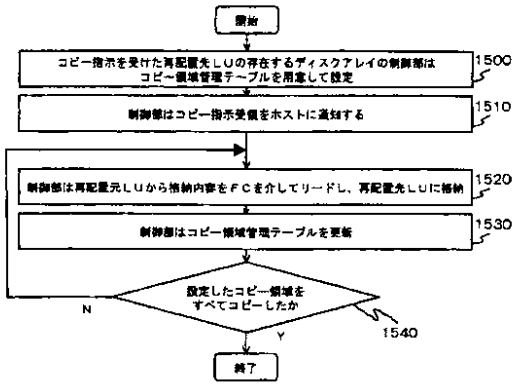
【図１１】

図11

6001			6002
LU番号			論理位置名（ドライブ）
ディスクアレイ番号	ID	LUN	
0	2	3	X:
1	4	5	Y:
1	5	2	使用していない（空き）

【図１２】

図12



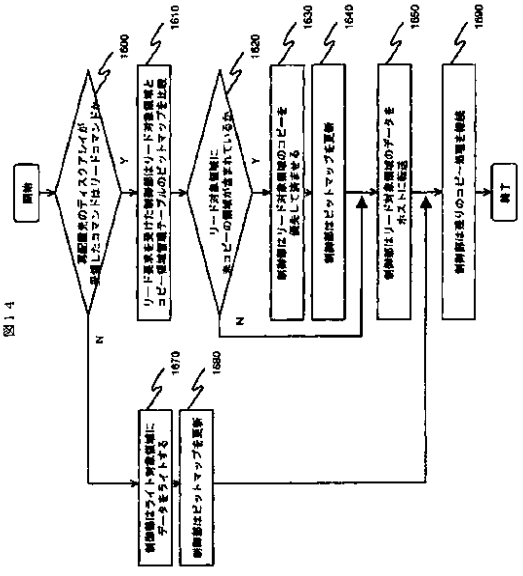
【図１３】

図13

6101	6102	6103	6104									
コピー元LU番号	コピー先LU番号	コピーブロック数	ビットマップ									
11:22:33:44:55:66:77:88	10:20:30:40:50:60:70:80	1000	1	1	1	1	1	1	1	1	1	1

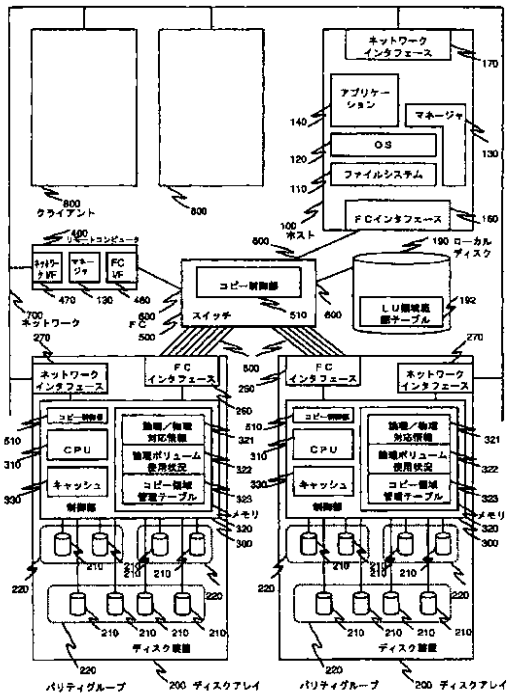
【図１４】

図14



【図１５】

図１５



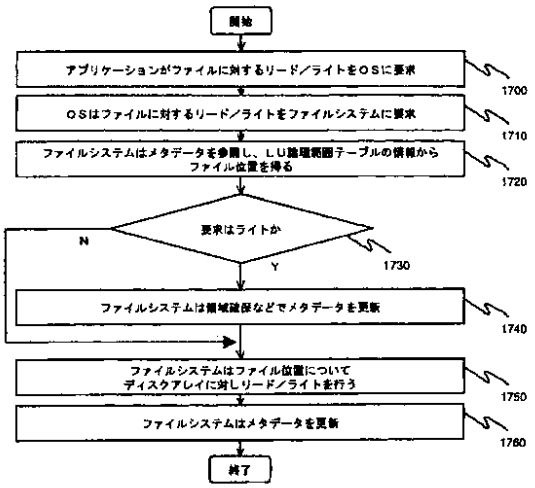
【図１６】

図１６

領域内アドレス	LU番号			LU内アドレス
	ディスクアレイ番号	ID	LUN	
0～999	0	2	3	0～999
1000～2999	1	4	5	0～1999
3000～3999	1	5	2	0～999

【図１７】

図１７



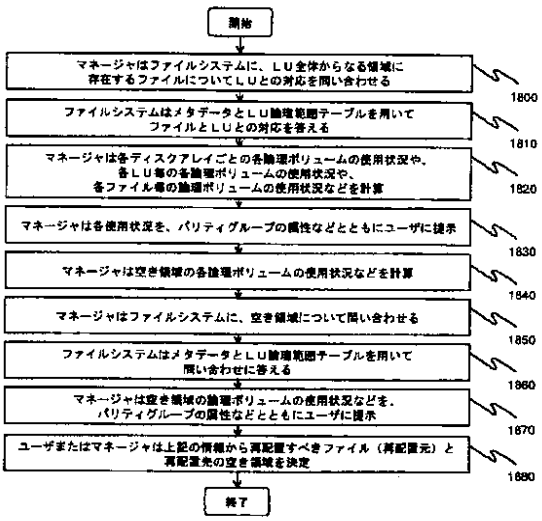
【図１８】

図１８

	6401	6402	6403	6404	6405	6406	6407
	作成日時	更新日時	アクセス日時	属性	ファイル論理位置名	セキュリティ情報	ファイル位置
	Jan 7, 2000 19:30	Jan 7, 2000 19:40	Jan 7, 2000 19:50	ノーマル	.../prjct1/doc1.txt	オーナのみのみリード／ライト可	100
	Jan 9, 2000 8:30	Jan 9, 2000 8:40	Jan 10, 2000 13:50	システムファイル	.../prjct3/voice1.mld	グループのみリード／ライト可	300

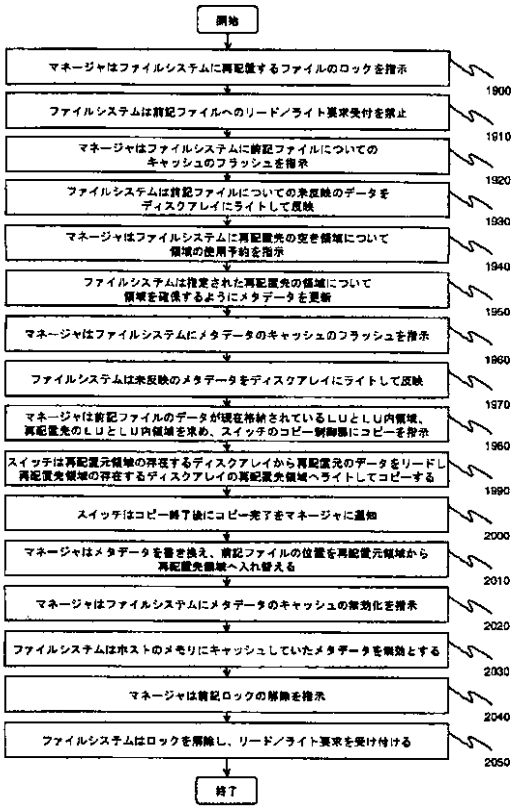
【図１９】

図１９



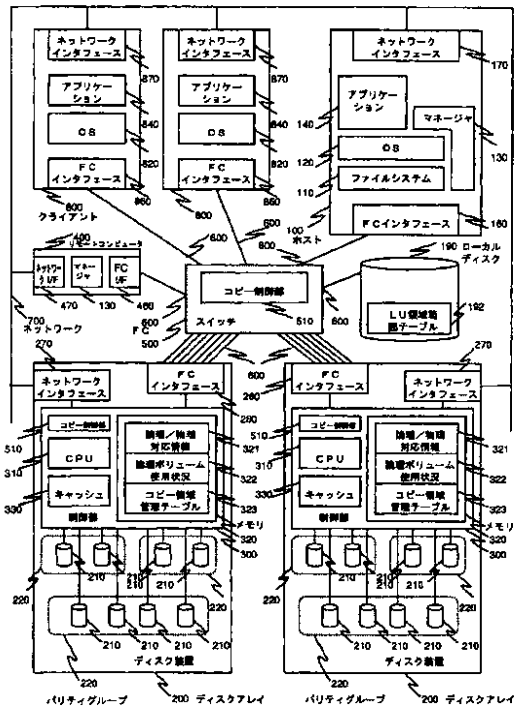
【図２０】

図２０



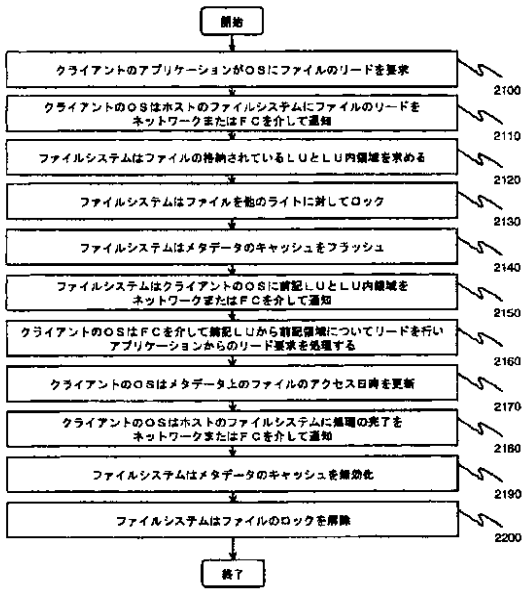
【図２１】

図２１



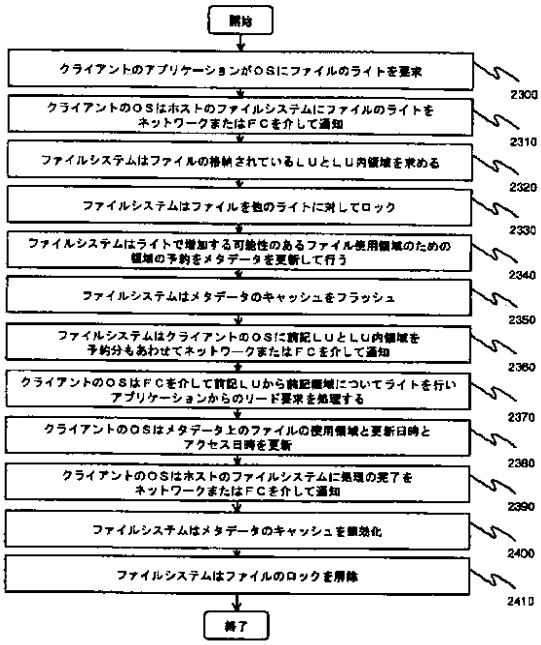
【図２２】

図２２



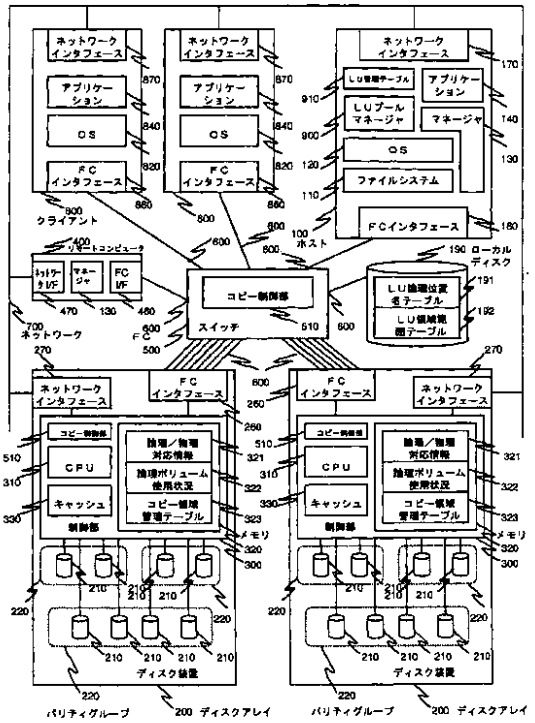
【図２３】

図２３



【図２４】

図２４



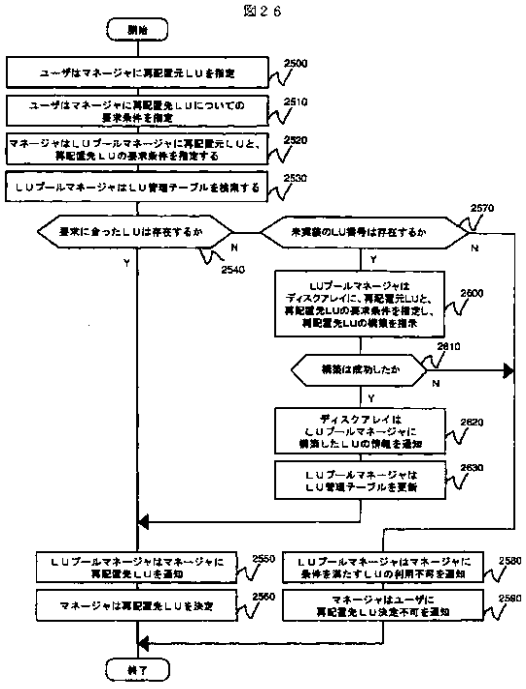
【図２５】

図２５

LU番号	サイズ	構成	状態	ディスクアレイ番号	バス	ID	LUN	ディスク性能	エミュレーションタイプ	再配置可能フラグ
0	1000	RAID1	オンライン	0	0	0	0	中性能	Open3	オフ
1	1000	RAID5	オフライン	0	-	-	-	高性能	Open3	オン
2	3500	キャッシュ	オンライン	0	0	0	1	超高性能	Open3	オフ
2	3500	キャッシュ	オンライン	0	1	0	0	超高性能	Open3	オフ
k	-	-	未実装	-	-	-	-	-	-	-
k+1	1000	単体ディスク	障害オフライン	0	0	1	0	低性能	3390-3	オフ
k+2	1000	RAID1	オフライン	0	-	-	-	高性能	3390-3	オン
n	-	-	未実装	-	-	-	-	-	-	-

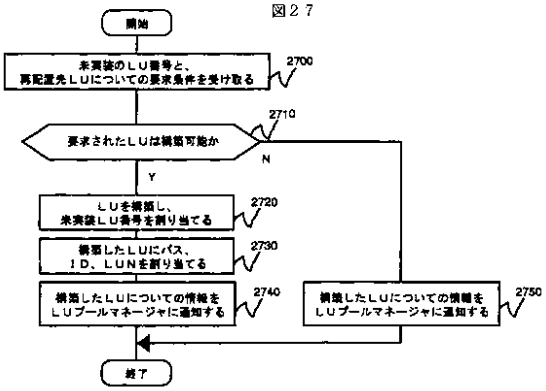
【図２６】

図２６



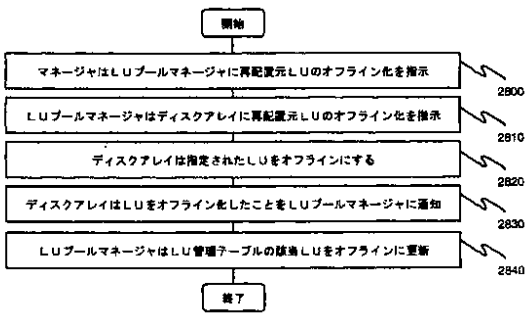
【図２７】

図２７



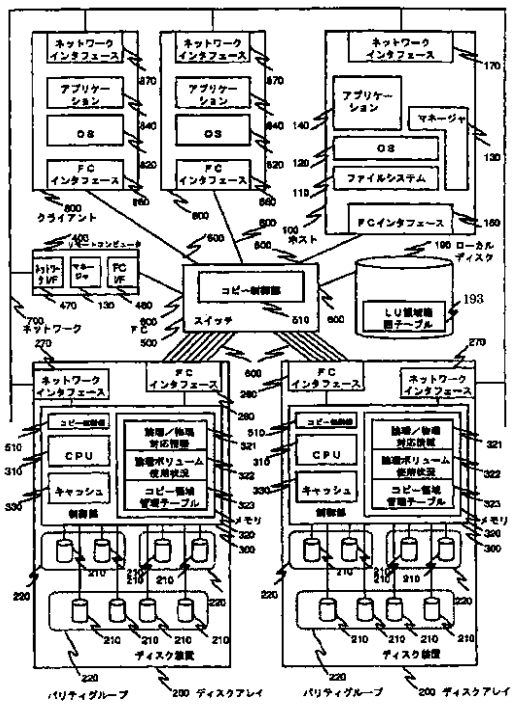
【図２８】

図２８



【図２９】

図２９



【図３０】

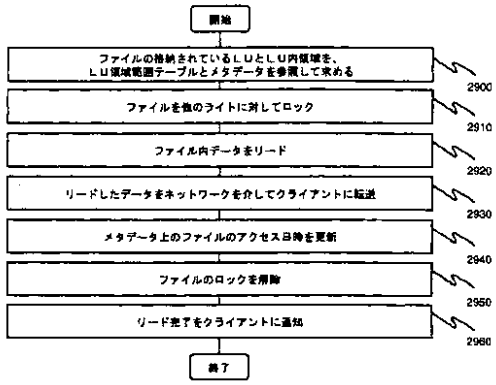
図３０

3510

使用種別	領域内アドレス	LU番号			LU内アドレス
		ディスクアレイ番号	ID	LUN	
FC経由データ転送によるファイル共有	0～999	0	2	3	0～999
	1000～2999	1	4	5	0～999
	3000～3999	1	5	2	0～999
ネットワーク経由データ転送によるファイル共有	0～999	0	2	4	0～999
	1000～2999	1	4	6	0～999
FC経由データ転送によるLU共有	0～999	1	1	1	0～999
ネットワーク経由データ転送によるLU共有	0～999	1	2	2	0～999
未使用領域	—	0	0	0	0～999
	—	0	0	1	0～999

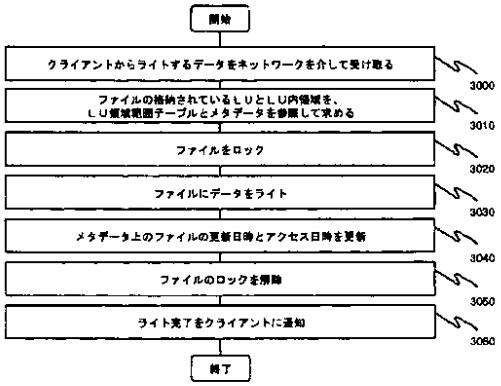
【図３１】

図３１



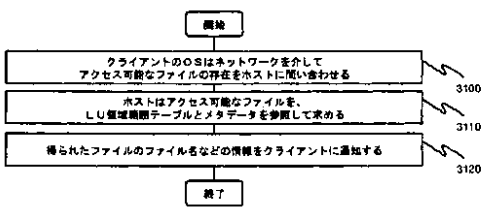
【図３２】

図３２



【図３３】

図３３



フロントページの続き

(72)発明者　江口　賢哲
神奈川県川崎市麻生区王禅寺１０９９番地　株式会社日立製作所システム開発研究所内
(72)発明者　荒井　弘治
神奈川県小田原市国府津２８８０番地　株式会社日立製作所ストレージシステム事業部内

Ｆターム(参考) 5B065 BA01　CA30　CH18