(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第4141391号
(P4141391)

(45)発行日　平成20年8月27日(2008.8.27)　(24)登録日　平成20年6月20日(2008.6.20)

(51)Int.Cl.　F I
G06F 12/08 (2006.01)　G06F 12/08 557
G06F 12/08 523E
G06F 12/08 507Z

請求項の数 4　(全 23 頁)

(21)出願番号　特願2004-29028(P2004-29028)
(22)出願日　平成16年2月5日(2004.2.5)
(65)公開番号　特開2005-222274(P2005-222274A)
(43)公開日　平成17年8月18日(2005.8.18)
審査請求日　平成18年12月25日(2006.12.25)

(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　110000279
特許業務法人ウィルフォート国際特許事務所
(74)代理人　100095371
弁理士　上村　輝之
(74)代理人　100089277
弁理士　宮川　長夫
(74)代理人　100104891
弁理士　中村　猛
(72)発明者　星野　幸子
神奈川県小田原市中里３２２番２号　株式会社日立製作所ＲＡＩＤシステム事業部内
最終頁に続く

(54)【発明の名称】ストレージサブシステム

(57)【特許請求の範囲】
【請求項１】
上位装置とのデータ授受をそれぞれ制御する複数のチャネルアダプタと、
論理的な記憶領域をそれぞれ提供する複数の記憶デバイス群と、
前記各記憶デバイス群とのデータ授受をそれぞれ制御する複数のディスクアダプタと、
前記各チャネルアダプタ及び前記各ディスクアダプタによりそれぞれ使用されるキャッシュメモリと、
前記キャッシュメモリを論理的に分割して構成される複数のキャッシュ分割領域と、
前記各キャッシュ分割領域を管理するための管理情報を記憶する制御メモリと、
を備え、
前記管理情報は、前記各キャッシュ分割領域毎にそれぞれ設けられる分割管理情報と、前記各キャッシュ分割領域の全体に適用される共通管理情報とから構成されており、さらに、
前記管理情報は、未使用状態のキャッシュ管理単位が接続されるフリーキュー及びこのフリーキューに関連付けられるフリーキューカウンタと、前記記憶デバイス群に反映前のダーティ状態のデータを格納するキャッシュ管理単位が接続されるダーティキュー及びこのダーティキューに関連付けられるダーティキューカウンタと、前記記憶デバイス群に反映済のクリーン状態のデータを格納するキャッシュ管理単位が接続されるクリーンキュー及びこのクリーンキューに関連付けられるクリーンキューカウンタと、使用中のキャッシュ管理単位の総量をカウントする使用中カウンタとを含んで構成され、

前記フリーキューカウンタと、前記クリーンキューと、前記クリーンキューカウンタ及び前記使用中カウンタは、前記各キャッシュ分割領域毎にそれぞれ設けられて、前記分割管理情報をそれぞれ構成し、

前記フリーキューと、前記ダーティキュー及び前記ダーティキューカウンタは、前記共通管理情報として使用されるストレージサブシステム。

【請求項２】

前記各キューにはそれぞれキュー管理テーブルが関連付けられており、前記分割管理情報を構成するキューに関連付けられるキュー管理テーブルは、前記各キャッシュ分割領域毎にそれぞれ設けられる請求項１に記載のストレージサブシステム。

【請求項３】

前記各チャネルアダプタ毎に前記各キャッシュ分割領域をそれぞれ設定可能な請求項１に記載のストレージサブシステム。

【請求項４】

前記各キャッシュ分割領域のうち１つのキャッシュ分割領域を共用領域として設定し、この共用領域に属する資源を割り当てることにより、新たなキャッシュ分割領域を設定する請求項１に記載のストレージサブシステム。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、ストレージサブシステム及びストレージサブシステムの制御方法に関する。

【背景技術】

【０００２】

ストレージサブシステムは、ホストコンピュータに対して大容量かつ高性能のストレージサービスを提供可能である。ストレージサブシステムでは、例えば、多数のディスクドライブをアレイ状に配設し、RAID（Redundant Array of Independent Inexpensive Disks）に基づく記憶領域を構築している。各ディスクドライブが有する物理的な記憶領域上には、論理的な記憶領域である論理ボリュームが形成されている。論理ボリュームには、ＬＵＮ（Logical Unit Number）が予め対応付けられている。ホストコンピュータは、ＬＵＮやブロックアドレス等を特定することにより、ストレージサブシステムに対して所定形式の書込みコマンド又は読出しコマンドを発行する。これにより、ホストコンピュータは、ストレージサブシステムに対して所望のデータの読み書きを行うことができる。

【０００３】

ストレージサブシステムには複数のホストコンピュータを接続可能である。あるホストコンピュータまたはアプリケーションプログラムが管理するデータ群を、他のホストコンピュータまたはアプリケーションプログラムから読み書き可能である場合、不都合を生じることがある。そこで、例えば、ゾーニングやＬＵＮマスキング等のようなアクセス制御技術が用いられる。ゾーニングとは、ストレージサブシステムに１つまたは複数のゾーンを設定し、ゾーンに属する特定の通信ポートやWWN（World Wide Name）のみにデータ転送を許可する技術である。ＬＵＮマスキングとは、特定のホストコンピュータに対して特定のＬＵＮへのアクセスを許可する技術である。

【０００４】

ところで、ホストコンピュータとディスクドライブとの間のデータ授受は、キャッシュメモリを介して行われる。ホストコンピュータからライトコマンドが出力されると、例えば、データはいったんキャッシュメモリに保存された後で、ディスクドライブに書き込まれる。また、ホストコンピュータからリードコマンドが出力されると、例えば、ディスクドライブから読み出されたデータは、キャッシュメモリを介してホストコンピュータに提供される。そこで、ディスクドライブへのキャッシュメモリ割当量を適切に配分させる技術も知られている（特許文献１）。

【特許文献１】特開平４－２６４９４０号公報

【発明の開示】
【発明が解決しようとする課題】
【０００５】
従来のゾーニングやＬＵＮマスキングにより、ホストコンピュータからストレージサブシステムへのアクセス制限を設定することができる。しかし、従来技術では、論理ボリュームへの単純なアクセス制限が可能となるだけで、ストレージサブシステムが有する資源の分割管理を実現するものではないから、使い勝手が悪い。複数ユーザによってストレージサブシステムが共同で使用されている場合に、例えば、あるユーザが多量の入出力要求（IO要求）を出すと、キャッシュメモリの多くの領域がこの要求の処理のために使用される。従って、他のユーザからの入出力処理に十分なキャッシュメモリを使用することができなくなり、サブシステム全体としてのサービス性能が低下する。つまり、あるユーザへのサービスのために、他のユーザへのサービスに影響が及ぶ。
【０００６】
なお、上記特許文献では、各記憶装置にそれぞれキャッシュメモリを割り当てているが、これはキャッシュメモリに記憶されたデータを効率的に記憶装置に書き込むためのものであり、複数ユーザに提供するために資源（キャッシュメモリ）の分割管理を行うものではない。
【０００７】
そこで、本発明の一つの目的は、複数のユーザにそれぞれキャッシュ領域を割り当てて共同で使用させることができ、あるユーザへ提供するサービスのために他のユーザへ影響が及ぶのを低減させることができるストレージサブシステム及びストレージサブシステムの制御方法を提供することにある。本発明の一つの目的は、複数ユーザに資源を柔軟に割り当てることができ、かつ、より少ない情報量で、各ユーザの管理範囲が互いに影響しないように管理可能なストレージサブシステム及びストレージサブシステムの制御方法を提供することにある。本発明の一つの目的は、キャッシュメモリの管理単位に応じて管理情報を構成することにより、キャッシュ資源を論理的に分割して使用可能なストレージサブシステム及びストレージサブシステムの制御方法を提供することにある。本発明の他の目的は、後述する実施形態の記載から明らかになるであろう。
【課題を解決するための手段】
【０００８】
上記課題を解決すべく、本発明に従うストレージサブシステムは、上位装置とのデータ授受をそれぞれ制御する複数のチャネルアダプタと、論理的な記憶領域をそれぞれ提供する複数の記憶デバイス群と、各記憶デバイス群とのデータ授受をそれぞれ制御する複数のディスクアダプタと、各チャネルアダプタ及び各ディスクアダプタによりそれぞれ使用されるキャッシュメモリと、キャッシュメモリを論理的に分割して構成される複数のキャッシュ分割領域と、各キャッシュ分割領域を管理するための管理情報を記憶する制御メモリと、を備えている。そして、管理情報は、各キャッシュ分割領域毎にそれぞれ設けられる分割管理情報と、各キャッシュ分割領域の全体に適用される共通管理情報とから構成される。
【０００９】
キャッシュメモリと制御メモリとは、それぞれ別々のメモリ基板として実装することもできるし、両者を混載したメモリ基板を用いてもよい。また、メモリのある領域をキャッシュメモリとして使用し、他の領域を制御メモリとして使用してもよい。キャッシュメモリは、上位装置と記憶デバイス群との間で授受されるデータを一時的に（あるいは長期間にわたって）記憶するために用いられる。キャッシュメモリは、複数のキャッシュ分割領域に論理的に分割される。各キャッシュ分割領域は、管理情報によってそれぞれ独立して管理される。各キャッシュ分割領域は、例えば、各チャネルアダプタ毎にそれぞれ設ける（割り当てる）ことができる。そして、各キャッシュ分割領域は、例えば、それぞれ異なるユーザにより使用可能である。管理情報は、分割管理情報と共通管理情報とから構成される。分割管理情報は、各キャッシュ分割領域毎にそれぞれ設けられる情報である。共通

管理情報は、全てのキャッシュ分割領域に適用される情報である。分割管理情報と共通管理情報とから管理情報を構成することにより、制御メモリの記憶資源を効率的に使用して、複数のキャッシュ分割領域をそれぞれ個別に使用可能である。

【００１０】

各分割管理情報及び共通管理情報は、キャッシュ管理単位の属性に基づいて設定可能である。キャッシュ管理単位としては、例えば、スロットやセグメント等を挙げることができる。１つのスロットは、少なくとも１つ以上のセグメントから構成可能であり、１つのセグメントは、例えば、１６ＫＢ程度のデータサイズを有する。キャッシュ管理単位の属性としては、例えば、フリー状態、ダーティ状態、クリーン状態を挙げることができる。フリー状態とは、キャッシュ管理単位が未使用である状態を示す。ダーティ状態とは、キャッシュ管理単位に格納されているデータが記憶デバイスに書き込まれていない状態を示す。クリーン状態とは、キャッシュ管理単位に格納されているデータが記憶デバイスに書込み済である状態を示す。キャッシュ管理単位の属性によっては、各キャッシュ分割領域毎にそれぞれ個別の管理情報を設ける方が好ましい場合がある。逆に、キャッシュ管理単位の属性によっては、全てのキャッシュ分割領域に適用可能な管理情報もある。

【００１１】

例えば、管理情報が、複数種類のサブ管理情報から構成される場合は、一部のサブ管理情報を各キャッシュ分割領域にそれぞれ分割して設け、残りのサブ管理情報を全てのキャッシュ分割領域に適用させることができる。この場合、サブ管理情報のうちいずれを分割管理情報として使用し、いずれを共通管理情報として使用するかは、キャッシュ管理単位の属性に基づいて決定することができる。

【００１２】

例えば、管理情報は、複数種類のキューと、これら各キューにそれぞれ関連付けられたカウンタとを含んで構成可能である。この場合、キャッシュ管理単位の属性に基づいて、各キュー及び各カウンタの一部を各キャッシュ分割領域毎にそれぞれ設けることにより各分割管理情報を構成することができる。そして、各キューの残り及び各カウンタの残りを共通管理情報として使用可能である。「キャッシュ管理単位の属性に基づいて」とは、「キュー（及びカウンタ）の属性に基づいて」と言い換えることも可能である。

【００１３】

ここで、互いに関連付けられているキューとカウンタのうち、一方は分割管理情報を構成し、他方は共通管理情報として使用することもできる。即ち、例えば、互いに関連付けられているキューとカウンタのうち、カウンタのみを分割して各キャッシュ分割領域にそれぞれ割り当てることができる。

【００１４】

さらに、各キューにはそれぞれキュー管理テーブルを関連付けることができる。そして、分割管理情報を構成するキューに関連付けられるキュー管理テーブルは、各キャッシュ分割領域毎にそれぞれ設けることができる。即ち、キューとキュー管理テーブルとは常に一体的に使用されるもので、キューが各キャッシュ分割領域毎にそれぞれ設けられる場合は、キュー管理テーブルも各キャッシュ分割領域にそれぞれ設けられる。

【００１５】

管理情報は、未使用状態のキャッシュ管理単位が接続されるフリーキュー及びこのフリーキューに関連付けられるフリーキューカウンタと、記憶デバイス群に反映前のダーティ状態のデータを格納するキャッシュ管理単位が接続されるダーティキュー及びこのダーティキューに関連付けられるダーティキューカウンタと、記憶デバイス群に反映済のクリーン状態のデータを格納するキャッシュ管理単位が接続されるクリーンキュー及びこのクリーンキューに関連付けられるクリーンキューカウンタと、使用中のキャッシュ管理単位の総量をカウントする使用中カウンタとを含んで構成可能である。そして、フリーキューカウンタと、クリーンキューと、クリーンキューカウンタ及び使用中カウンタは、分割管理情報として、各キャッシュ分割領域毎にそれぞれ設けることができる。また、フリーキューと、ダーティキュー及びダーティキューカウンタは、共通管理情報として使用すること

ができる。

【００１６】

各キャッシュ分割領域のうち１つのキャッシュ分割領域を共用領域として設定し、この共用領域に属する資源を割り当てることにより、新たなキャッシュ分割領域を設定することもできる。例えば、初期状態では、全てのキャッシュ領域を共用領域として使用する。そして、新たにキャッシュ分割領域を生成する場合は、共用領域から一部のキャッシュ領域を切り取って、新しいキャッシュ分割領域に割り当てる。なお、キャッシュ分割領域を削除する場合は、このキャッシュ分割領域に割り当てられていたキャッシュ領域を、共用領域に戻せばよい。

【発明を実施するための最良の形態】

【００１７】

以下、図１～図１８に基づき、本発明の実施の形態を説明する。本実施形態では、上位装置とのデータ授受をそれぞれ制御する複数の上位インターフェース制御部と、論理的な記憶領域をそれぞれ提供する複数の記憶デバイス群と、各記憶デバイス群とのデータ授受をそれぞれ制御する複数の下位インターフェース制御部と、各上位インターフェース制御部及び各下位インターフェース制御部によりそれぞれ使用されるメモリ部と、を備えたストレージサブシステムが開示される。そして、本実施例では、メモリ部が提供するキャッシュ領域を複数のキャッシュ分割領域に分割する。さらに、本実施例では、メモリ部でデータを管理するためのキャッシュ管理単位の属性に応じて、管理情報を各キャッシュ分割領域毎にそれぞれ分割し、各管理情報に基づいて、各キャッシュ分割領域毎にそれぞれデータを管理させる。

【実施例１】

【００１８】

図１は、ストレージサブシステム１０の外観構成を示す概略斜視図である。ストレージサブシステム１０は、例えば、基本筐体１１と複数の増設筐体１２とから構成することができる。基本筐体１１は、ストレージサブシステム１０の最小構成単位であり、記憶機能及び制御機能の両方を備えている。増設筐体１２は、ストレージサブシステム１０のオプションであり、基本筐体１１の有する制御機能により制御される。例えば、最大４個の増設筐体１２を基本筐体１１に接続可能である。

【００１９】

基本筐体１１には、複数の制御パッケージ１３と、複数の電源ユニット１４と、複数のバッテリユニット１５と、複数のディスクドライブ８０とがそれぞれ着脱可能に設けられている。増設筐体１２には、複数のディスクドライブ８０と、複数の電源ユニット１４及び複数のバッテリユニット１５が着脱可能に設けられている。また、基本筐体１１及び各増設筐体１２には、複数の冷却ファン１６がそれぞれ設けられている。

【００２０】

制御パッケージ１３は、後述するチャネルアダプタ２０、ディスクアダプタ３０及びキャッシュメモリ４０等をそれぞれ実現するためのモジュールである。即ち、基本筐体１１には、複数のチャネルアダプタパッケージ、複数のディスクアダプタパッケージ及び１つ以上のメモリパッケージがそれぞれ着脱可能に設けられ、パッケージ単位で交換可能となっている。

【００２１】

図２は、ストレージサブシステム１０を含む記憶システムの全体概要を示すブロック図である。ストレージサブシステム１０は、通信ネットワークＣＮ１を介して、複数のホストコンピュータ１Ａ～１Ｃ（以下、特に区別しない場合は、「ホストコンピュータ１」と呼ぶ）と双方向通信可能にそれぞれ接続されている。ここで、通信ネットワークＣＮ１は、例えば、ＬＡＮ（Local Area Network）、ＳＡＮ（Storage Area Network）、インターネットあるいは専用回線等である。ＬＡＮを用いる場合、ホストコンピュータ１とストレージサブシステム１０との間のデータ転送は、TCP/IP（Transmission Control Protocol/Internet Protocol）プロトコルに従って行われる。ＳＡＮを用いる場

合、ホストコンピュータ１とストレージサブシステム１０とは、ファイバチャネルプロトコルに従ってデータ転送を行う。また、ホストコンピュータ１がメインフレームの場合は、例えば、FICON（Fibre Connection：登録商標）、ESCON（Enterprise System Connection：登録商標）、ACONARC（Advanced Connection Architecture：登録商標）、FIBARC（Fibre Connection Architecture：登録商標）等の通信プロトコルに従ってデータ転送が行われる。

【００２２】

各ホストコンピュータ１は、例えば、サーバ、パーソナルコンピュータ、ワークステーション、メインフレーム等として実現されるものである。例えば、各ホストコンピュータ１は、図外に位置する複数のクライアント端末と別の通信ネットワークを介して接続されている。各ホストコンピュータ１は、例えば、各クライアント端末からの要求に応じて、ストレージサブシステム１０にデータの読み書きを行うことにより、各クライアント端末へのサービスを提供する。図中では、１つのみ図示してあるが、ストレージサブシステム１０内には、複数の仮想筐体（SLPR：Storage Logical Partition）を設定可能である。

【００２３】

SLPRとは、ストレージサブシステム１０内の各種物理的資源及び論理的資源を各ユーザ毎に（あるいは各アプリケーションプログラム毎に）分割して割り当てることにより構成された領域である。即ち、例えば、SLPRには、それぞれ専用のチャネルアダプタ２０と、それぞれ専用のキャッシュ領域４０と、それぞれ専用の仮想的な論理デバイス（VDEV）７０等とを備えている。即ち、SLPRは、仮想的な小型のストレージサブシステムであるかのように振る舞う。

【００２４】

SLPRには、少なくとも１つ以上のキャッシュ分割領域（CLPR：Cache Logical Partiton）を設けることができる。図２中では、左側にSLPRを１つだけ示してあるが、複数のSLPRを設けることができる。そして、１つのSLPR内には、１つまたは複数のCLPRを設定することができる。

【００２５】

CLPRとは、キャッシュメモリ４０を論理的に複数の領域に分割したものである。CLPRは、各チャネルアダプタ２０毎にそれぞれ設定可能である。例えば、チャネルアダプタ２０をｎ個実装する場合は、ｎ個のCLPRを設けることができる。例えば、ｎ＝３２に設定可能であるが、本発明はこれに限定されない。各CLPR0～CLPRnは、それぞれ互いに独立して使用されるものであり、各ホストコンピュータ１は、それぞれが利用可能なCLPRのみを独占的に使用することができる。そして、あるCLPRに対するホストコンピュータ１からの操作は、他のCLPRにできるだけ影響を与えないように構成されている。即ち、特定のホストコンピュータ１からのアクセスが集中した場合でも、そのホストコンピュータ１が利用可能なCLPRのみで必要なキャッシュ容量が確保され、他のCLPRの資源（キャッシュ領域）を奪うことがないように構成されている。

【００２６】

なお、図中右側に示すCLPR0は、SLPRが定義されていない共用領域であり、この共用領域CLPR0には、ストレージサブシステム１０全体で使用される各種の情報が記憶される。また、後述のように、初期状態では、キャッシュメモリ４０の全領域が共用領域CLPR0に割り当てられている。そして、共用領域CLPR0からキャッシュ領域を所定量だけ切り出すことにより、新たなCLPRを設定するようになっている。

【００２７】

図２に示す例では、ホストコンピュータ１Ａは、共用領域CLPR0のみにアクセスして、データの入出力を行うことができる。ホストコンピュータ１ＢはCLPR1のみに、ホストコンピュータ１ＣはCLPR2のみに、ホストコンピュータ１ＮはCLPRnのみに、それぞれアクセ

ス可能であり、他のCLPRの利用や参照等を行うことはできない。

【００２８】

SVP（Service Processor）９０は、ストレージサブシステム１０の管理や監視を行うためのコンピュータ装置であり、管理用のサーバ機能を提供する。SVP９０は、例えば、ストレージサブシステム１０内に設けられたＬＡＮ等の内部ネットワークＣＮ３（図３参照）を介して、各チャネルアダプタ２０や各ディスクアダプタ３０等から各種の環境情報や性能情報等を収集する。SVP９０が収集する情報としては、例えば、装置構成、電源アラーム、温度アラーム、入出力速度（IOPS）等が挙げられる。SVP９０と各管理端末２Ａ～２Ｎ，３とは、例えば、ＬＡＮ等の通信ネットワークＣＮ２を介して接続されている。管理者は、管理端末を介してSVP９０にログインすることにより、権限のある範囲内において、例えば、ＲＡＩＤ構成の設定、各種パッケージ（チャネルアダプタパッケージ、ディスクアダプタパッケージ、メモリパッケージ、ディスクドライブ等）の閉塞処理や各種設定変更等を行うことができる。

【００２９】

SVP９０には、複数の管理端末２Ａ～２Ｎ，３を接続可能である。ここで、管理端末２Ａ～２Ｎは、各SLPR毎に設けられる端末であり、管理端末３は、ストレージサブシステム１０の全体を管理するために設けられる端末である。以下の説明では、１つのSLPRに１つのCLPRを設ける場合を例に挙げて説明する。従って、管理端末２Ａ～２Ｎは、各CLPRをそれぞれ管理する管理者（以下、分割管理者と呼ぶ）によりそれぞれ操作されるサブ端末である。管理端末３は、ストレージサブシステム１０の全体を管理するシステム管理者（あるいは全体管理者と呼ぶこともできる）により操作される全体端末である。

【００３０】

各サブ端末２Ａ～２Ｎの分割管理者は、自己が管理権限を有するCLPRについてのみ各種設定変更等を行うことができ、他のCLPRの構成等を参照したり変更することは許可されない。これに対し、システム管理者は、各CLPRを含めてストレージサブシステム１０の全体の各種設定変更等を行うことができる。

【００３１】

システム管理者は、全体端末３を介してSVP９０にログインし、ストレージサブシステム１０の有する物理的資源及び論理的資源を適宜分割することにより、各ユーザ毎にSLPR（CLPR）を設定することができる。また、システム管理者は、各分割管理者に対して、ユーザID等を発行することもできる。分割管理者は、システム管理者から発行された専用のユーザIDを用いてSVP９０にログインすることができる。分割管理者は、サブ端末２を操作することにより、自己の管理下にあるCLPR内の設定を変更することができる。

【００３２】

図３は、ストレージサブシステム１０の論理的構成に着目したブロック図である。ストレージサブシステム１０は、複数のチャネルアダプタ（以下、CHA）２０と、複数のディスクアダプタ（以下、DKA）３０と、少なくとも１つ以上のキャッシュメモリ４０及び共有メモリ５０と、スイッチ部６０と、複数の仮想的論理デバイス（VDEV）７０と、複数のディスクドライブ８０と、SVP９０とを備えている。

【００３３】

各CHA２０は、各ホストコンピュータ１との間のデータ転送を制御するもので、通信ポート２１を備えている。ストレージサブシステム１０には、例えば３２個のCHA２０を設けることができる。CHA２０は、例えば、オープン系用CHA、メインフレーム系用CHA等のように、ホストコンピュータ１の種類に応じて用意される。

【００３４】

各CHA２０は、それぞれに接続されたホストコンピュータ１から、データの読み書きを要求するコマンド及びデータを受信し、ホストコンピュータ１から受信したコマンドに従って動作する。DKA３０の動作も含めて先に説明すると、例えば、CHA２０は、ホストコンピュータ１からデータの読出し要求を受信すると、読出しコマンドを共有メモリ５０に記

憶させる。DKA３０は、共有メモリ５０を随時参照しており、未処理の読出しコマンドを発見すると、ディスクドライブ８０からデータを読み出して、キャッシュメモリ４０に記憶させる。CHA２０は、キャッシュメモリ４０に移されたデータを読み出し、コマンド発行元のホストコンピュータ１に送信する。

【００３５】

CHA２０は、ホストコンピュータ１からデータの書込み要求を受信すると、書込みコマンドを共有メモリ５０に記憶させる。また、CHA２０は、受信したデータ（ユーザデータ）をキャッシュメモリ４０に記憶させる。ここで、ホストコンピュータ１から書込みを要求されたデータは、ディスクドライブ８０に書き込まれていない「ダーティデータ」であるため、例えば複数箇所にそれぞれ記憶されて多重化される。

【００３６】

CHA２０は、キャッシュメモリ４０にデータを記憶した後、ホストコンピュータ１に対して書込み完了を報告する。そして、DKA３０は、共有メモリ５０に記憶された書込みコマンドに従って、キャッシュメモリ４０に記憶されたデータを読出し、所定のディスクドライブ８０に記憶させる。ディスクドライブ８０に書き込まれたデータは、「ダーティデータ」から「グリーンデータ」に属性が変化し、キャッシュメモリ４０による多重管理から解放される。なお、本明細書において、「ダーティデータ」とは、ディスクドライブ８０に書き込まれていない状態のデータを意味する。また、「クリーンデータ」とは、ディスクドライブ８０に書き込み済のデータを意味する。

【００３７】

各DKA３０は、ストレージサブシステム１０内に例えば４個や８個等のように複数個設けることができる。各DKA３０は、各ディスクドライブ８０との間のデータ通信を制御する。各DKA３０と各ディスクドライブ８０とは、例えば、ＳＡＮ等の通信ネットワークＣＮ４を介して接続されており、ファイバチャネルプロトコルに従ってブロック単位のデータ転送を行う。各DKA３０は、ディスクドライブ８０の状態を随時監視しており、この監視結果は内部ネットワークＣＮ３を介してSVP９０に送信される。各CHA２０及び各DKA３０は、例えば、プロセッサやメモリ等が実装されたプリント基板と、メモリに格納された制御プログラム（いずれも不図示）とをそれぞれ備えており、これらのハードウェアとソフトウェアとの協働作業によって、それぞれ所定の機能を実現するようになっている。

【００３８】

キャッシュメモリ４０は、例えば、ユーザデータ等を記憶するものである。キャッシュメモリ４０は、例えば不揮発メモリから構成される。キャッシュメモリ４０は、複数のメモリから構成することができ、ダーティデータを多重管理することができる。本実施例では、キャッシュメモリ４０が有する全キャッシュ領域を複数個に分割することにより、各CLPR0～nを設定している。

【００３９】

共有メモリ（あるいは制御メモリ）５０は、例えば不揮発メモリから構成される。共有メモリ５０には、例えば、制御情報や管理情報等が記憶される。これらの制御情報等の情報は、複数の共有メモリ５０により多重管理することができる。共有メモリ５０及びキャッシュメモリ４０は、それぞれ複数個設けることができる。また、同一のメモリ基板にキャッシュメモリ４０と共有メモリ５０とを混在させて実装することもできる。あるいは、メモリの一部をキャッシュ領域として使用し、他の一部を制御領域として使用することもできる。

【００４０】

スイッチ部６０は、各CHA２０と、各DKA３０と、キャッシュメモリ４０と、共有メモリ５０とをそれぞれ接続するものである。これにより、全てのCHA２０，DKA３０は、キャッシュメモリ４０及び共有メモリ５０にそれぞれアクセス可能である。スイッチ部６０は、例えば超高速クロスバスイッチ等として構成することができる。

【００４１】

ストレージサブシステム１０は、多数のディスクドライブ８０を実装可能である。各デ

ィスクドライブ８０は、例えば、ハードディスクドライブ（ＨＤＤ）や半導体メモリ装置等として実現可能である。ディスクドライブ８０は、物理的な記憶デバイス（PDEV）である。そして、RAID構成等によっても相違するが、例えば、４個１組のディスクドライブ８０が提供する物理的な記憶領域上には、仮想的な論理デバイス（VDEV）７０が構築される。なお、VDEV７０上にさらに仮想的な論理デバイスを設定可能である。また、ストレージサブシステム１０により使用される記憶資源は、全てストレージサブシステム１０内に設ける必要はなく、ストレージサブシステム１０の外部に存在する記憶資源を利用することもできる。例えば、外部の他社製ストレージサブシステムが有する記憶デバイスを特定のVDEV７０に割り付け、他社製記憶デバイスへのアクセスパス等を管理することにより、他社の記憶デバイスをあたかも自己の記憶デバイスであるかのように使用可能である。

【００４２】

図４は、各CLPRが資源を分割して利用する様子を模式的に示す説明図である。図中では、共用領域CLPR0と２個の専用CLPR1，2とによって、キャッシュメモリ４０を分割して使用する場合を示してある。各CLPRには、それぞれキャッシュ領域が割り当てられている。各CLPRがそれぞれ使用可能なキャッシュ領域の最大値（最大割当量）は、システム管理者により設定される。各CLPRにそれぞれ割り当てられたキャッシュ領域には、各CLPRを利用するホストコンピュータ１により使用されるデータ（ライトデータ、リードデータ）が格納される。図４中では、データが格納された領域に斜線を付し、「使用中」と表記してある。各CLPRは、それぞれが使用可能な最大割当量までのキャッシュ領域を使用可能であり、他のCLPRのキャッシュ領域を使用することはできない。

【００４３】

さらに、各CLPRは、それぞれのキャッシュ領域に加えて、他のキャッシュ関連データを管理している。キャッシュ関連データとしては、例えば、DCR（Ｆ１）と、サイドファイルＦ２と、ＰＣＲ（Ｆ３）とを挙げることができる。

【００４４】

DCR（Dynamic Cache Residency）１００とは、キャッシュ常駐化機能であり、VDEV７０の特定の領域に存在するデータをキャッシュ領域に常駐化させるものである。これにより、重要なデータ群に対するアクセス性能を高めることができる。

【００４５】

サイドファイルＦ２とは、図外のリモートサイトにデータを転送してコピーさせるためのデータを記憶するファイルである。例えば、リモートコピー対象のデータがサイドファイルＦ２に所定量以上蓄積された場合は、通信ネットワークを介して、距離的に離れた場所にある別のストレージサブシステムにデータが転送され、保持される。リモートコピーの初期コピーでは、指定されたデータ群を丸ごとリモートサイトにコピーし、それ以後に更新されたデータは、差分ファイルとしてリモートコピーされる。

【００４６】

PCR（Partical Cache Residence）１０２とは、例えば、ストレージサブシステム１０にNAS（Network Attached Storage）機能を設ける場合に使用されるもので、データの種別毎にキャッシュ領域に常駐化させる機能である。DCR（Ｆ１）は、特定の記憶空間に存在するデータを常駐化させるものであるのに対し、ＰＣＲ（Ｆ３）は、特定種類のデータを常駐化させるものである。なお、これらキャッシュ関連データDCR（Ｆ１）、サイドファイルＦ２及びＰＣＲ（Ｆ３）は、それを必要とするCLPRのみに設けられる。即ち、常駐化機能が設定されていないCLPRには、キャッシュ常駐化に関するデータは記憶されない。同様に、リモートコピーが設定されていないCLPRには、サイドファイルＦ２は記憶されない。

【００４７】

そして、図４に示すように、各CLPRには、少なくとも１つ以上のVDEV７０がそれぞれ割り当てられている。各CLPRのキャッシュ領域に書き込まれたデータは、それぞれに割り当てられたVDEV７０の所定領域に書き込まれる。また、VDEV７０から読み出されたデータは

、対応するCLPRのキャッシュ領域に保持される。

【００４８】

図５は、キャッシュ領域に記憶されているデータを管理するための方法を示す説明図である。ストレージサブシステム１０では、キャッシュメモリ４０に記憶されているデータを効率的に検索するために、以下のような階層構造で管理している。

【００４９】

まず、ホストコンピュータ１からデータの入出力要求があると、この入出力要求に含まれるLBA（Logical
Block Address）に基づいて、VDEVSLOT番号（VDEVSLOT＃）を求める。そして、VDEVSLOT番号に基づいて、VDSLOT－PAGEテーブルＴ１を参照することにより、次の階層へのポインタを取得する。VDSLOT－PAGEテーブルＴ１には、PAGE－DIRテーブルＴ２へのポインタが含まれている。PAGE－DIRテーブルＴ２には、PAGE－GRPPテーブルＴ３へのポインタが含まれている。さらに、PAGE－GRPPテーブルＴ３には、GRPT１テーブルＴ４へのポインタが含まれている。GRPT１テーブルＴ４には、GRPT２テーブルＴ５へのポインタが含まれている。GRPT２テーブルＴ５には、SLCB（スロット制御テーブル）Ｔ６へのポインタが含まれている。

【００５０】

このように、LBAに基づいて、各テーブルＴ１～Ｔ５を順番に参照していくことにより、SLCBテーブルＴ６に到達する。SLCBテーブルＴ６には、少なくとも１つ以上のSGCB（セグメント制御ブロック）テーブルＴ７が関連付けられている。SGCBテーブルＴ７は、キャッシュ管理の最小単位であるセグメントに関する制御情報を格納している。１つのスロットには、１～４個のセグメントを関連付けることができる。１つのセグメントには、例えば、４８ＫＢまたは６４ＫＢのデータを格納することができる。

【００５１】

キャッシュ管理の最小単位はセグメントであるが、ダーティデータ（物理ディスクへの書込み前の状態）と、クリーンデータ（物理ディスクへの書込み後の状態）の各状態間の遷移は、スロット単位で行われる。また、キャッシュ領域の確保（リザーブ）及び解放（リリース）も、スロット単位またはセグメント単位で行われる。

【００５２】

SLCBテーブルＴ６には、図８に示すように、後方ポインタと、前方ポインタと、キューステータスと、VDSLOT＃と、スロットステータスと、CLPR情報と、少なくとも１つ以上のSGCBポインタとを含めて構成することができる。キューステータスは、そのSLCBテーブルＴ６に関連付けられているキューの種別やキュー番号を含む。スロットステータスは、そのSLCBテーブルＴ６に対応するスロットの状態を含む。CLPR情報は、そのSLCBテーブルＴ６が属するCLPRに関する情報（CLPR番号等）を含む。SGCBポインタは、そのSLCBテーブルＴ６に関連付けられているSGCBテーブルＴ７を特定するためのポインタが含まれている。

【００５３】

なお、SLCBテーブルＴ６及びSGCBテーブルＴ７は、例えば、ポインタ変更等の競合が発生して整合性が失われないように、VDEV７０単位で排他制御が行われる。即ち、SLCBテーブルＴ６またはSGCBテーブルＴ７が属するVDEV７０について、ロックを取得した場合にのみ、SLCBテーブルＴ６またはSGCBテーブルＴ７を操作できるようになっている。本明細書では、このVDEV単位の排他制御を「VDEVロック」と呼ぶ。VDEVロックは、論理ボリューム単位の排他制御であると言い換えることもできる。

【００５４】

図６は、SLCBテーブルＴ６の状態遷移を示す説明図である。SLCBテーブルＴ６は、例えば、未使用状態ＳＴ１、デバイス反映済み状態ＳＴ２、パリティ無し＆デバイス反映前状態ＳＴ３、パリティ生成中状態ＳＴ４及びパリティ有り＆ドライブ反映前状態ＳＴ５の５つのステータスを取ることができる。

【００５５】

未使用状態ＳＴ１とは、そのスロットが未使用であることを示す状態である。未使用状

態にあるスロット（SLCB）は、フリーSGCBキュー（以下、フリーキューと略記）により管理される。デバイス反映済み状態ＳＴ２とは、そのスロットに格納されているデータが物理デバイス（ディスクドライブ８０）に書込み済みであることを示す状態である。デバイス反映済み状態ＳＴ２にあるスロットは、クリーンキューにより管理される。

【００５６】

パリティ無し＆デバイス反映前状態ＳＴ３とは、例えばRAID５等のようにパリティデータを必要とする場合において、そのスロットに格納されているデータについて、未だパリティデータが生成されていないことを示す状態である。パリティ無し＆デバイス反映前状態ＳＴ３にあるスロットは、中間ダーティキューにより管理される。パリティ生成中状態ＳＴ４とは、そのスロットに格納されているデータについて、パリティデータを生成している最中であることを示す状態である。パリティ生成中状態ＳＴ４にあるスロットは、ホストダーティキューにより管理される。

【００５７】

パリティ有り＆ドライブ反映前状態状態ＳＴ５とは、そのスロットに格納されているデータについて、パリティデータが生成されているが、物理デバイスに未だ書き込まれていないことを示す状態である。パリティ有り＆ドライブ反映前状態ＳＴ５にあるスロットは、物理ダーティキュー（以下、ダーティキューと略記）により管理される。

【００５８】

ホストコンピュータ１からライトアクセスがあった場合、ＳＴ１からＳＴ２を経てＳＴ３に状態が遷移し、パリティデータの生成が開始される。そして、書込みを要求されたデータについてパリティデータが生成されると、状態はＳＴ３からＳＴ４を経てＳＴ５に変化する。ホストコンピュータ１からCLPRのキャッシュ領域の１つまたは複数のスロットにデータが書き込まれた時点で、ホストコンピュータ１に対して書込み完了報告が送信される。この時点では、書込みを要求されたデータは、キャッシュ領域上にのみ存在するので、ダーティ状態（ＳＴ５）となっている。

【００５９】

そして、ダーティ状態にあるスロットのデータが所定の物理デバイスに書き込まれると（デステージ）、そのスロットは、ダーティ状態からクリーン状態に変化する（ＳＴ５→ＳＴ２）。クリーン状態にあるデータは、物理デバイスに書き込まれているので、もしも必要ならばキャッシュ領域から消去しても構わない。クリーン状態にあるスロットに格納されているデータをキャッシュ上から消去すると、そのスロットは、クリーン状態から未使用状態へ戻る（ＳＴ２→ＳＴ１）。

【００６０】

ホストコンピュータ１からリードアクセスがあった場合は、図５と共に述べた各階層テーブルＴ１～Ｔ６を順番に辿ることにより、要求されたデータが特定される。そして、ホストコンピュータ１が要求しているデータを物理デバイスから読み出して、キャッシュ領域の１つまたは複数のスロットに格納させる。物理デバイスから読み出されたデータを格納するスロットは、クリーン状態を有する。

【００６１】

図７は、キュー構造を示す説明図である。キュー１００は、同一状態にあるSLCBテーブルＴ６を一方向または双方向につなげる待ち行列として構成される。SLCBテーブルＴ６の状態が変化すると、例えば、FIFO（First-In First-Out：先入れ先出し）やLRU（Least Recently Used：最長未使用時間）等のような予め設定されたアルゴリズムに基づいて、キュー１００の構成は変化する。

【００６２】

SLCBテーブルＴ６への操作が競合して整合性が失われるのを防止するために、各キュー１００毎に排他制御を行う。即ち、操作対象のキューのロックを取得した場合にのみ、そのキューへの操作を行うことができる。本明細書では、このキュー単位の排他制御を「キューロック」と呼ぶ。従って、上述したVDEVロックとキューロックとの２段階の排他制御

の下で、SLCBテーブルＴ６は操作可能となっている。つまり、操作しようとするSLCBテーブルＴ６が属するVDEV７０のロックと、そのSLCBテーブルＴ６が接続されているキューのロックとの２つのロックを取得した場合のみ、そのSLCBテーブルＴ６の内容を変更することができる。

【００６３】

このように、キュー１００への操作は排他制御されるため、ロック取得時のオーバーヘッドを考慮して、同一種類のキューを複数本設けている。クリーンキュー、ダーティキュー、フリーキュー等は、それぞれ複数本のキューから構成される。

【００６４】

キューセットを構成する個々のキューには、それぞれカウンタ２００が関連付けられている。カウンタ２００は、そのキューに属するSLCBテーブルＴ６及びSGCBテーブルＴ７の数を管理するものである。カウンタ２００に管理されているSLCB数等に基づいて、後述の過負荷制御等を行うことができる。

【００６５】

また、キューセットを構成する個々のキュー１００には、キュー管理テーブル３００も関連付けられる。キュー管理テーブル３００は、キュー１００の構造等を管理するものである。キュー１００、カウンタ２００及びキュー管理テーブル３００は、それぞれ共有メモリ５０に記憶される。

【００６６】

図９は、各CLPRのキャッシュ領域の使用状態と各キューとの関係を模式的に示す説明図である。各CLPRのキャッシュ領域は、その使用状態に応じて、「使用中領域」と「未使用領域」とに分けることができる。使用中領域とは、データを格納しているスロット群（セグメント群）を意味する。未使用領域とは、空いているスロット群（セグメント群）を意味する。

【００６７】

使用されているセグメントの総数と１つのセグメントあたりのデータサイズとを乗算したものが、そのCLPRにおける使用中領域のサイズとなる。使用されているセグメントには、クリーン状態にあるセグメントとダーティ状態にあるセグメントとが含まれる。従って、そのCLPRに割り当てられているキャッシュ領域のうち現在使用されている領域のサイズは、（クリーン状態のセグメント数＋ダーティ状態のセグメント数）で表すことができる。未使用領域のサイズは、フリー状態にあるセグメント数で表すことができる。

【００６８】

フリーキュー１０１は、フリー状態（未使用状態）にあるSLCBテーブルＴ６を管理するものである。フリーキュー１０１に接続されているSLCBテーブルＴ６（及びSGCBテーブルＴ７）の数は、フリーキュー１０１に関連付けられているフリーカウンタ２０１により把握されている。フリーキュー１０１の制御単位はVDEVであり、キューロックの単位もVDEV単位である。そして、フリーキュー１０１は、例えば、FIFOに従って使用される。フリーキュー１０１は、例えば、VDEV７０の実装数の２倍設けている。従って、実装可能なVDEV７０の最大数を５１２とすると、フリーキュー１０１の総数は１０２４となる。

【００６９】

クリーンキュー１０２は、クリーン状態にあるSLCB管理テーブルＴ６を管理するものである。クリーンキュー１０２に接続されているSLCBテーブルＴ６（及びSGCBテーブルＴ７）の数は、クリーンキュー１０２に関連付けられているクリーンカウンタ２０２によって把握されている。クリーンキュー１０２の制御単位はVDEVであり、キューロックの単位もVDEV単位である。クリーンキュー１０２は、例えば、LRUに従って使用される。クリーンキュー１０２は、例えば、VDEV７０の実装数の４倍設けている（５１２×４＝２０４８）。

【００７０】

ダーティキュー１０３は、ダーティ状態にあるSLCB管理テーブルＴ６を管理するものである。ダーティキュー１０３に接続されているSLCBテーブルＴ６（及びSGCBテーブルＴ７

）の数は、ダーティキュー１０３に関連付けられているダーティカウンタ２０３によって把握されている。ダーティキュー１０３の制御単位は物理ドライブである。ダーティキュー１０３は、例えば、ディスクドライブ８０の実装数の３２倍設けている。ディスクドライブ８０を２０４８個実装する場合、ダーティキュー１０３は６５５３６個設けることができる。ダーティキュー１０３は、物理デバイスに書込み前のデータ（セグメント）を管理するためのものである。従って、論理デバイス（VDEV７０）よりも物理デバイス（ディスクドライブ８０）の方に関連が深い。そこで、ダーティキュー１０３は、物理デバイスの数に依存して設定するようになっている。

【００７１】

使用中カウンタ２０６は、各CLPRにおける使用中のセグメント数（ダーティ状態のセグメント数＋クリーン状態のセグメント数）をカウントするものである。

【００７２】

図１０は、各CLPRと各キュー及びカウンタ等の関係を示す説明図である。上述した各キュー、カウンタ及び管理テーブルは、キャッシュ管理単位（セグメントまたはスロット）を管理するための管理情報を構成する。本実施例では、セグメントまたはスロットの性質（ステータス）に応じて、各CLPR毎にそれぞれ設ける管理情報（分割管理情報）と、全CLPRに適用する共通管理情報とに区別する。

【００７３】

図１０に示すように、各CLPRには、クリーンキュー１０２と、クリーンカウンタ（正確にはクリーンキューカウンタ）２０２と、フリーキューカウンタ２０１と、色分けキューカウンタ２０４と、使用中カウンタ２０６と、BINDキューカウンタ２０５とが、それぞれ設けられている。また、図示を省略しているが、クリーンキュー１０２の構成を管理するためのクリーンキュー管理テーブルも各CLPR毎にそれぞれ設けられている。以上のように、各CLPR毎にそれぞれ設けられる管理用の情報が分割管理情報を構成する。

【００７４】

これに対し、フリーキュー１０１と、ダーティキュー１０３と、ダーティカウンタ（正確にはダーティキューカウンタ）２０３と、色分けキュー１０４と、BINDキュー１０５と、これら各キュー１０１，１０３～１０５を管理するためのキュー管理テーブル３０１とは、全てのCLPRに共通に適用される共通管理情報を構成する。

【００７５】

色分けキュー１０４とは、PCR機能によってキャッシュ領域に常駐化されるデータを管理するためのキューである。色分けキューカウンタ２０４は、PCRにより常駐化されているセグメント数をカウントする。

【００７６】

BINDキュー１０５とは、DCR機能によってキャッシュ領域に常駐化されるデータを管理するためのキューである。BINDキューカウンタ２０５は、DCRにより常駐化されているセグメント数（またはスロット数）をカウントする。

【００７７】

常駐化されるデータを格納するセグメントを管理するための色分けキュー１０４及びBINDキュー１０５をそれぞれ設けることにより、各CLPR内において、常駐化すべきデータを格納するセグメントがフリー状態に変更されるのを防止することができる。

【００７８】

次に、管理情報を構成する各サブ管理情報（キュー、カウンタ、キュー管理テーブル）のうち、いずれを各CLPR毎にそれぞれ設け、いずれを全CLPRに適用するのかの理由を説明する。

【００７９】

クリーンキュー１０２と、クリーンカウンタ２０２及びクリーンキュー管理テーブルは、それぞれ各CLPR毎に設けられている。フリー状態のセグメントが枯渇すると、クリーン状態のスロットが解放されてフリースロット化される。従って、クリーンキュー１０２は、最も使用頻度の高いキューである。もしも、全てのCLPRにおいて、１セットのクリーン

キューを用いる場合は、あるCLPRにおいてフリーセグメントの枯渇が生じると、他のCLPRで利用されているデータを格納するクリーン状態のスロットが解放されてしまう可能性がある。即ち、一方のCLPRにおけるキャッシュ利用状態が他方のCLPRにおけるキャッシュ利用状態に影響を与えることになる。そこで、本実施例では、クリーンキュー１０２と、クリーンカウンタ２０２とを各CLPR毎にそれぞれ設けることができるようにしている。なお、キュー管理テーブルは、キューと一体的に取り扱われるものである。従って、クリーンキュー１０２を各CLPR毎にそれぞれ設ける場合、クリーンキュー管理テーブルも各CLPR毎にそれぞれ設けられる。

【００８０】

使用中カウンタ２０６は、各CLPR毎にそれぞれ設けられる。使用中カウンタ２０６は、クリーン状態のセグメント数とダーティ状態のセグメント数とを加算した値をカウントするものである。従って、使用中カウンタ２０６のカウント値からクリーンカウンタ２０２のカウント値を減算して得られる値は、そのCLPRにおけるダーティ状態のセグメント数に他ならない。従って、クリーンカウンタ２０２及び使用中カウンタ２０６が各CLPR毎にそれぞれ設けられていれば、ダーティカウンタ２０３を各CLPR毎にそれぞれ設ける必要はない。

【００８１】

ダーティキュー１０３は、物理デバイスに関連するキューであり、制御単位も物理デバイス単位となっている。従って、もしも、ダーティキュー１０３を各CLPR毎にそれぞれ設ける場合は、ダーティキュー管理テーブルを記憶するための記憶容量が増大し、共有メモリ５０の記憶領域を圧迫する。例えば、キュー１本あたりに必要となる情報量は、キュー管理テーブルが１６バイト、カウンタが４バイトである。キューの分割はキュー管理テーブルの分割を意味するため、共有メモリ５０の多くの記憶領域が分割管理情報の記憶のために消費される。そこで、本実施例では、共有メモリ５０の記憶資源を効率的に使用するために、ダーティキュー１０３及びダーティカウンタ２０３は、各CLPR毎に設けるのではなく、全てのCLPRについて適用することとした。

【００８２】

また、共有メモリ５０の記憶資源を有効利用する観点から、フリーキュー１０１も各CLPR毎にそれぞれ設けることなく、バイト数の少ないフリーカウンタ２０１のみを各CLPR毎にそれぞれ設けることとした。

【００８３】

フリーキュー１０１と同様の理由により、色分けキュー１０４，BINDキュー１０５も、それぞれのカウンタ２０４，２０５のみを各CLPR毎にそれぞれ設け、キュー自体は全てのCLPRに適用することとした。

【００８４】

このように、本実施例では、サブ管理情報の属性に応じて、即ち、サブ管理情報により管理されるキャッシュ管理単位の属性に応じて、サブ管理情報を分配する。従って、共有メモリ５０の記憶資源を効率的に使用し、各CLPR間の独立性を担保できる。

【００８５】

図１１は、キャッシュ管理構造と排他制御の関係を示す説明図である。図１１では、図１０と共に述べた管理情報の一部のみが示されている。図１１中の「VDEVロックGr」とは、VDEVロックのグループを示す。ディレクトリを操作するためにはVDEVロックの取得が必要となり、キューを操作するためにはキューロックの操作が必要となる。つまり、CLPRに管理されているデータを移動させるためには、そのデータを格納するセグメントを管理するキューへの操作権限と、そのセグメントが属するディレクトリへの操作権限との両方が必要となる。

【００８６】

図１１の右側に示すように、各CLPRには、１つまたは複数のVDEV７０を関連付けることができる。あるCLPRに関連付けられているVDEV７０を、別のCLPRに関連付けることもできる。図１２は、CLPR0に関連付けられているVDEV＃1をCLPR1へ付け替える様子を示す説明

図である。

【００８７】

VDEV７０の所属先CLPRを変更する場合は、VDEV７０とCLPRとの関連付けを変更し、さらに、移動対象のVDEV７０に属するデータ群を新たなCLPRに関連付ける必要がある。即ち、移動対象のセグメント（またはスロット）を、移動元CLPRに関連付けられたキューから外し、移動先CLPRに関連付けられたキューに接続する。この操作に際して、VDEVロックとキューロックの両方が必要とされる。

【００８８】

図１３は、CLPRの構成を変更する場合の処理の一例を示す。図１３には、移動元CLPRに関連付けられているVDEV７０を、移動先CLPRに移動させる場合の概略が示されている。

【００８９】

まず、管理端末３を介して、システム管理者がVDEV７０の移動を要求すると（S1：YES）、CHA２０やDKA３０の代表プロセッサは、移動対象として指定されたVDEV７０の固有情報を変更する（Ｓ２）。即ち、そのVDEV７０が所属するCLPRに関する情報（帰属先CLPR情報）が、移動先のCLPRに関する情報に書き換えられる。

【００９０】

次に、代表プロセッサは、移動対象のVDEV７０に対応するSLCBテーブルＴ６を検索して抽出し、この移動対象となるSLCBテーブルＴ６について、CLPR情報を変更する（Ｓ３）。即ち、図８と共に上述したように、移動対象となっている各SLCBテーブルＴ６について、その所属先CLPRに関する情報を移動先CLPRに関する情報に変更する。

【００９１】

そして、代表プロセッサは、移動対象のSLCBテーブルＴ６のスロットステータスがクリーン状態であるか否かを検査する（Ｓ４）。クリーン状態のスロットである場合（S4：YES）、即ち、移動対象のSLCBテーブルＴ６がクリーンキュー１０２に接続されている場合、代表プロセッサは、そのSLCBテーブルＴ６のキューステータスを、移動先CLPRに合わせて変更する（Ｓ５）。つまり、移動先CLPRに設けられているの何番目のクリーンキュー１０２に接続するか等の情報を書き換える。

【００９２】

代表プロセッサは、キューステータスを変更後、移動対象のSLCBテーブルＴ６を現在のクリーンキュー１０２から外す（Ｓ６）。そして、代表プロセッサは、移動させるクリーンスロットの数に応じて、移動元CLPRに設けられている使用中カウンタ２０６及びクリーンカウンタ２０２のカウント値をそれぞれ減じさせる（Ｓ７）。次に、代表プロセッサは、移動対象のSLCBテーブルＴ６を、移動先CLPRのクリーンキュー１０２に接続する（Ｓ８）。そして、代表プロセッサは、移動先CLPRにおける使用中カウンタ２０６及びクリーンカウンタ２０２のカウント値をそれぞれ増加させて（Ｓ９）、本処理を終了する。

【００９３】

一方、移動対象のVDEV７０に対応するSLCBテーブルＴ６がクリーン状態ではない場合（S4：NO）、代表プロセッサは、その移動対象のSLCBテーブルＴ６がフリー状態であるか否かを検査する（Ｓ１０）。移動対象のSLCBテーブルＴ６がフリー状態の場合（S10：YES）、代表プロセッサは、移動元及び移動先のフリーカウンタ２０１をそれぞれ変更する（Ｓ１１）。即ち、移動させるフリースロットの数に応じて、移動元のフリーカウンタ２０１を減算し、移動先のフリーカウンタ２０１を加算し、処理を終了する。移動対象のSLCBテーブルＴ６がダーティ状態の場合（S10：NO）、何もせずに処理を終了する。ダーティキュー１０３及びダーティカウンタ２０３は、全CLPRについて共通に適用されるので、キュー移動やカウンタの増減を行う必要はない。

【００９４】

なお、図１３中では省略しているが、色分けキュー１０４及びBINDキュー１０５に接続されるSLCBテーブルＴ６は、それぞれフリーキュー１０１と同様に処理される。色分けキュー１０４及びBINDキュー１０５は、各CLPR毎にそれぞれ分割して設けるのではなく、そのカウンタ２０４，２０５のみを各CLPRに設けるようになっている。従って、キュー間の

移動は生じず、移動元と移動先でのカウンタ値の変更のみが発生する。

【００９５】

図１４は、CLPRのキャッシュ割当量を変更する場合の概略処理を示すフローチャートである。システム管理者によるキャッシュ割当処理の変更操作が行われると（S31：YES）、代表プロセッサは、キャッシュ割当の変更に関連するSLCBテーブルＴ６を検出する（Ｓ３２）。そして、代表プロセッサは、変更に関連するSLCBテーブルＴ６のCLPR情報を、キャッシュ割当の変更に応じて、変更させる（Ｓ３３）。

【００９６】

例えば、CLPR1のキャッシュ割当量を減らして、その分だけCLPR２のキャッシュ割当量を増加させる場合、この変動領域に位置するSLCBテーブルＴ６の帰属先CLPRは、CLPR1からCLPR2に変更される。そして、キャッシュ領域の割当変更に応じて、クリーンカウンタ２０２等の値を増減させて処理を終了する（Ｓ３４）。

【００９７】

図１５は、フリーセグメントを確保するためのフリーセグメント収集処理の概略を示すフローチャートである。ここで、フリーセグメントとは、フリー状態にあるセグメントを意味する。本処理は、定期的に、あるいは、フリーセグメントの不足時（過負荷時）に、行うことができる。

【００９８】

まず、クリーン状態のSLCBテーブルＴ６のうち最も使用頻度の低いものを検出し（Ｓ４１）、検出されたSLCBテーブルＴ６のステータスを、クリーン状態からフリー状態に変更する（Ｓ４２）。このステータス変更に伴って、SLCBテーブルＴ６は、クリーンキュー１０２からフリーキュー１０１に付け替えられる。また、クリーンカウンタ２０２及びフリーカウンタ２０１のカウント値もそれぞれ変更される。Ｓ４１及びＳ４２の処理は、所定量のフリーセグメントが確保されるまで、繰り返し行われる（Ｓ４３）。

【００９９】

図１６は、CLPRを定義する様子を示す模式図である。本実施例では、CLPR0を共用のプール領域として使用する。CLPR0には、ストレージサブシステム１０内で共通に使用される各種情報が記憶される。なお、上述のように、各種の管理情報は共有メモリ５０に記憶されている。

【０１００】

図１６（ａ）に示すように、ユーザ専用のCLPRを１つも設定していない初期状態では、キャッシュメモリ４０の有する全てのキャッシュ領域は、CLPR0に所属する。図１６（ｂ）に示すように、CLPR0に割り当てられたキャッシュ領域は、共通の諸情報等を記憶する「使用中領域」と「未使用領域」とに区別することができる。ユーザ専用のCLPR1を新たに設定する場合は、図１６（ｃ）に示すように、プール領域であるCLPR0から必要なだけのキャッシュ領域を切り取って、CLPR1に割り当てる。さらに、別のユーザ用にCLPR2を新たに設定する場合は、図１６（ｄ）に示すように、プール領域として使用されるCLPR0から必要な量だけキャッシュ領域を切り取り、新たに生成するCLPR2に割り当てる。ユーザ専用のCLPR1またはCLPR2を削除する場合は、削除されるCLPRに割り当てられているキャッシュ領域は、CLPR0に戻される。

【０１０１】

図１７は、上述した処理を示すフローチャートである。初期状態では、キャッシュメモリ４０の有する全てのキャッシュ領域が、プール領域として使用される共用領域CLPR0に割り当てられる（Ｓ５１）。ストレージサブシステム１０全体で使用する諸情報は、CLPR0に記憶され、運用が開始される（Ｓ５２）。

【０１０２】

CLPRを新たに追加する場合は（S53：YES）、CLPRの構成変更を行う（Ｓ５４）。このCLPR構成変更処理では、CLPR0からキャッシュ領域を切り取り、新たなCLPRに割り当てる等の処理を行う。一方、既に生成されたCLPRを削除する場合（S53：NO、S55：YES）、削除するCLPRに割り当てられていたキャッシュ領域をCLPR0に戻す等の処理を行う（Ｓ５６）

。なお、原則として、CLPR0を削除することはできない。

【０１０３】

図１８は、各サブ端末２毎に、それぞれのサブ端末２が管理するCLPRに関する情報を表示等する様子を示す説明図である。

【０１０４】

SVP９０は、認証部９１と、構成管理部９２とを備えている。各CLPRの管理者は、サブ端末２を介してSVP９０にアクセスし、ユーザID等を入力する。認証部９１によるユーザ認証の結果、正当なユーザであると認められた場合、CLPRの管理者は、自己の管理下にあるCLPRについて構成変更を行うことができる。この構成変更の操作は、構成管理部９２によって各CLPRに反映される。また、構成管理部９２から取得されたCLPRに関する構成情報は、サブ端末２に表示させることができる。

【０１０５】

サブ端末２に表示可能な情報としては、例えば、CLPRに割り当てられている最大キャッシュ量、CLPRに割り当てられているVDEVに関する情報（VDEV番号やボリュームサイズ等）、使用中のキャッシュサイズ、未使用のキャッシュサイズ、CLPRへのIO頻度等を挙げることができる。なお、ストレージサブシステム１０の全体を管理するシステム管理者は、全体端末３を介して、全てのCLPRに関する情報を取得することができ、各CLPRの構成変更等を行うことができる。

【０１０６】

このように、本実施例によれば、ストレージサブシステム１０内に設定した各CLPR毎に、それぞれ管理用の情報を分割する構成とした。従って、各CLPR間の干渉を抑制することができ、使い勝手を高めることができる。即ち、一方のCLPRに対するアクセスが増大した場合でも、他のCLPRに影響が及ぶのを極力防止することができる。

【０１０７】

また、本実施例では、セグメントまたはスロットの属性に基づいて、管理情報を分配する構成とした。従って、全ての管理情報を単純に各CLPRにそれぞれ設ける構成と比較して、管理情報全体のデータサイズを低減することができ、共有メモリ５０を効率的に使用することができる。

【０１０８】

なお、本発明は、上述した実施の形態に限定されない。当業者であれば、本発明の範囲内で、種々の追加や変更等を行うことができる。

【図面の簡単な説明】

【０１０９】

【図１】本発明の実施例に係わるストレージサブシステムの外観図である。

【図２】ストレージサブシステムの論理的な概略構成を示すブロック図である。

【図３】ストレージサブシステムのより詳細なブロック図である。

【図４】CLPR毎に割り当てられるキャッシュ領域とVDEVの関係を示す説明図である。

【図５】階層構造を有するキャッシュ管理方法を示す説明図である。

【図６】SLCBテーブルの状態遷移図である。

【図７】キュー、カウンタ及びキュー管理テーブルの関係を示す説明図である。

【図８】SLCBテーブルの概略構造を示す説明図である。

【図９】キャッシュ領域と各キュー及びカウンタとの関係を示す説明図である。

【図１０】CLPRの管理情報を分配する様子を示す模式図である。

【図１１】VDEVロック及びキューロックによってデータ移動が可能となる様子を示す説明図である。

【図１２】CLPR0からCLPR1にVDEVを移動させる様子を示す説明図である。

【図１３】VDEVを移動させる場合の処理を示すフローチャートである。

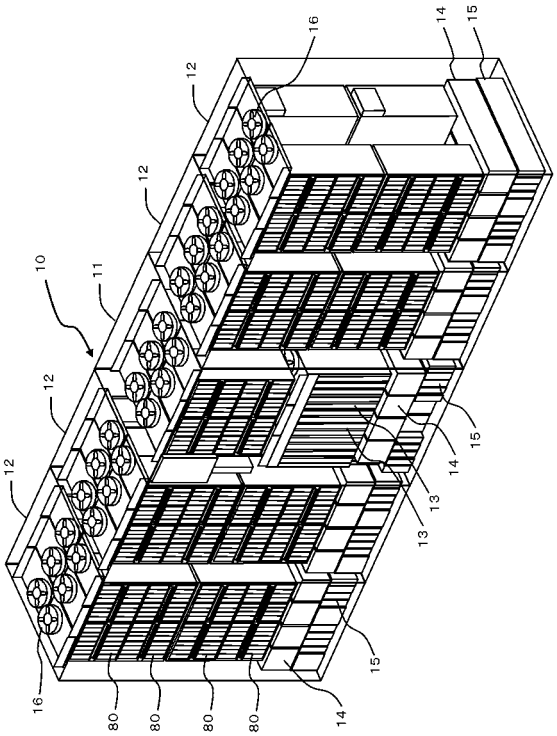
【図１４】CLPRへのキャッシュ割当を変更する場合のフローチャートである。

【図１５】フリーセグメントの収集処理を示すフローチャートである。

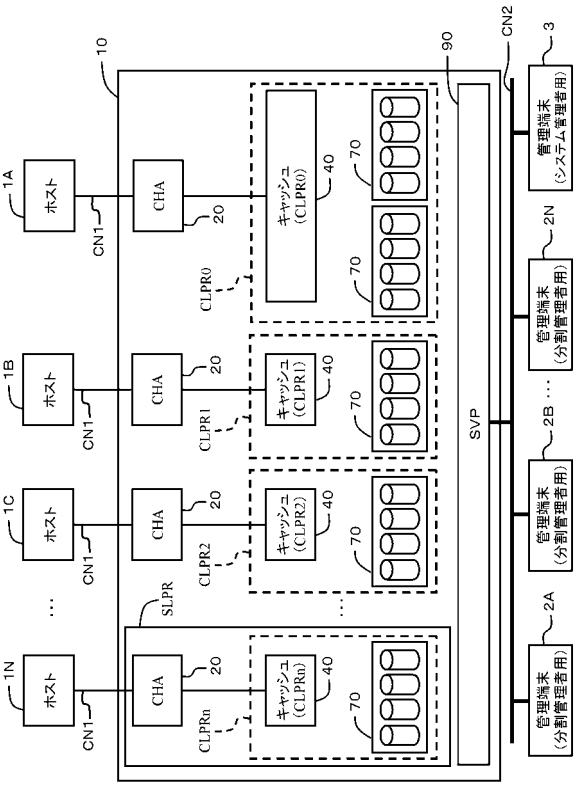
【図１６】プール領域として用いられるCLPR0からキャッシュ領域を切り取って、新たなC

LPRを設定する様子を示す説明図である。
【図１７】CLPRを定義する場合の処理を示すフローチャートである。
【図１８】各CLPRに関する情報をサブ管理端末に表示させる様子を示す説明図である。
【符号の説明】
【０１１０】
１…ホストコンピュータ、２…サブ端末、３…全体端末、１０…ストレージサブシステム、１１…基本筐体、１２…増設筐体、１３…制御パッケージ、１４…電源ユニット、１５…バッテリユニット、１６…冷却ファン、２０…チャネルアダプタ、２１…通信ポート、３０…ディスクアダプタ、４０…キャッシュメモリ、５０…共有メモリ、６０…スイッチ部、７０…論理デバイス、８０…ディスクドライブ、９０…サービスプロセッサ、９１…認証部、９２…構成管理部、１００…キュー、１０１…フリーキュー、１０２…クリーンキュー、１０３…ダーティキュー、１０４…色分けキュー、１０５…BINDキュー、２００…カウンタ、２０１…フリーキューカウンタ、２０２…クリーンキューカウンタ、２０３…ダーティキューカウンタ、２０４…色分けキューカウンタ、２０５…BINDキューカウンタ、２０６…使用中カウンタ、３００…キュー管理テーブル、CLPR…論理的キャッシュ分割領域、SLPR…仮想筐体、SLCB…スロット制御テーブル、SGCB…セグメント制御ブロック、Ｆ１…ＤＣＲ、Ｆ２…サイドファイル、Ｆ３…ＰＣＲ、ＣＮ…通信ネットワーク、ＳＴ…ステータス、Ｔ１～Ｔ７…テーブル

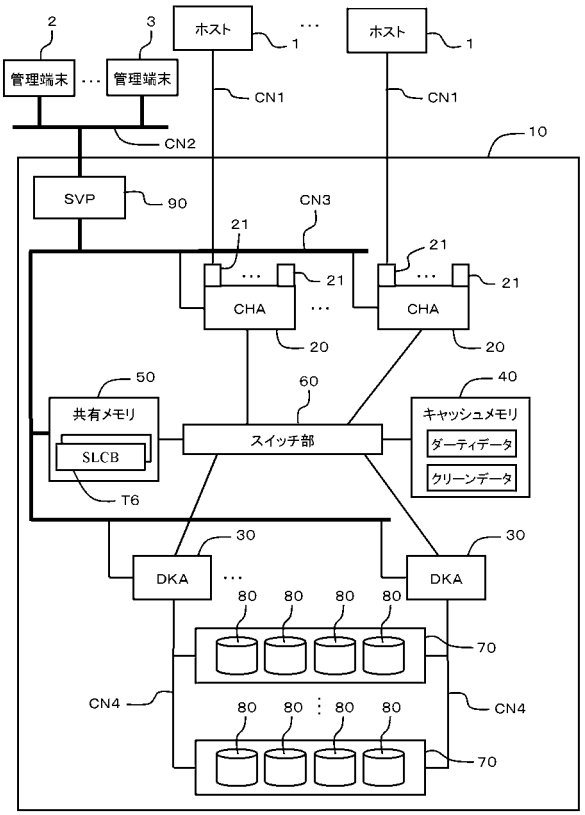
【図１】



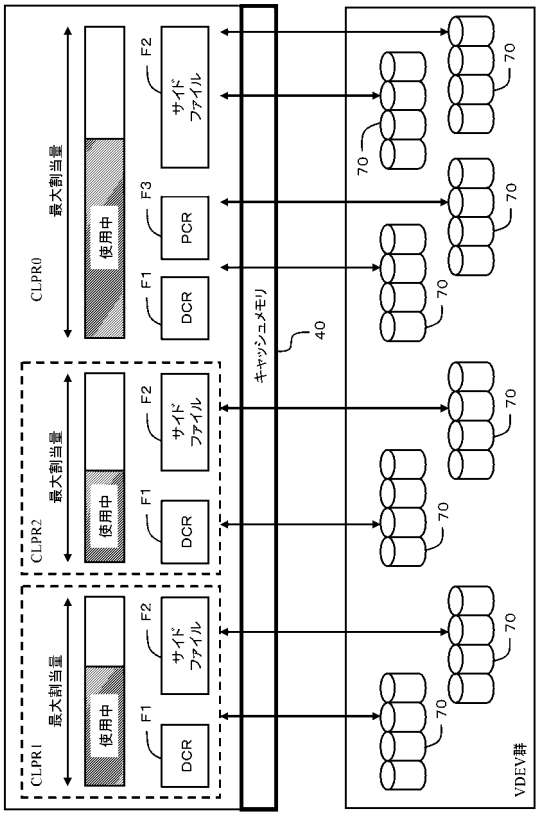
【図２】



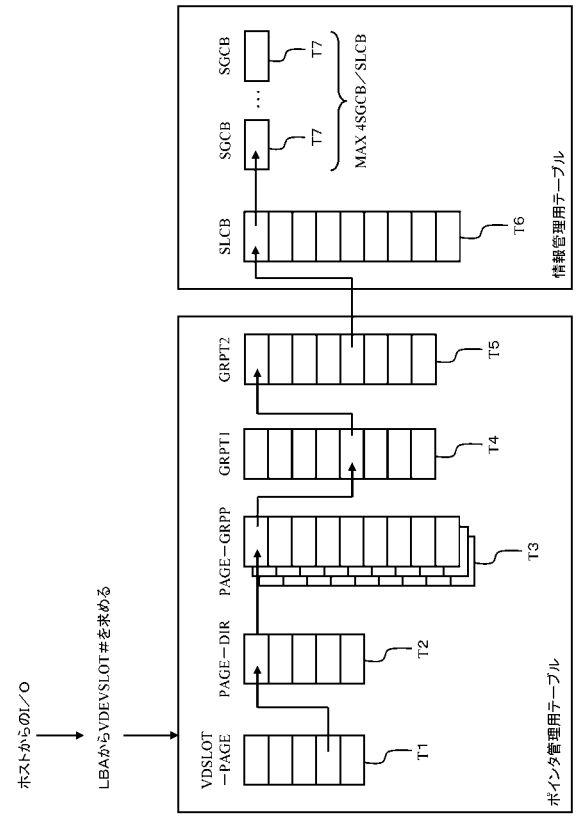
【図３】


ホスト
管理端末
CN1
CN2
SVP
90
10
CN3
21
CHA
20
50
共有メモリ
SLCB
T6
60
スイッチ部
40
キャッシュメモリ
ダーティデータ
クリーンデータ
30
DKA
80
70
CN4

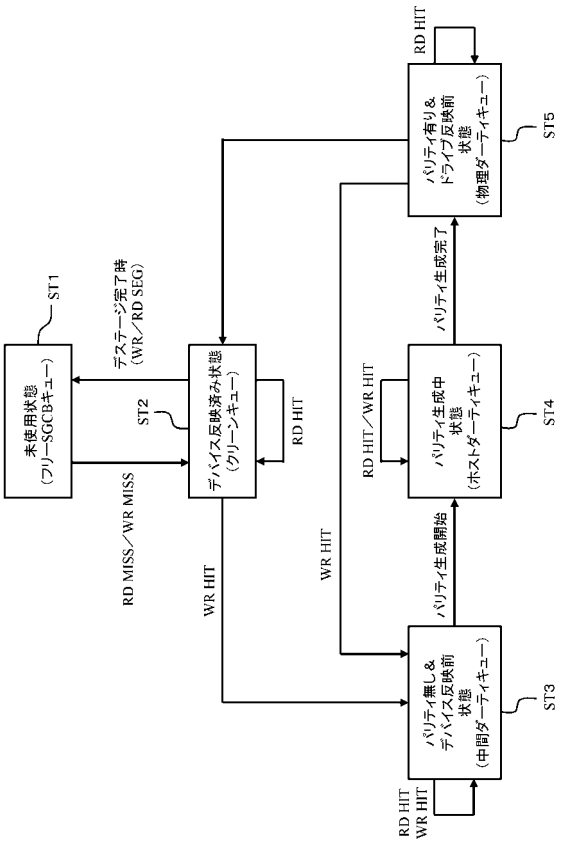
【図４】


CLPR0
CLPR1
CLPR2
最大割当量
使用中
F1
DCR
F2
サイド
ファイル
F3
PCR
キャッシュメモリ
40
70
VDEV群

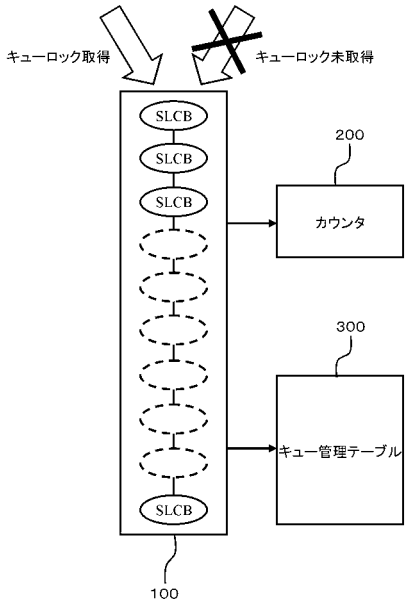
【図５】


ホストからのI／O
LBAからVDEVSLOTを求める
VDSLOT
−PAGE
PAGE−DIR
PAGE−GRPP
GRPT1
GRPT2
T1
T2
T3
T4
T5
ポインタ管理用テーブル
SLCB
SGCB
T6
T7
MAX 4SGCB／SLCB
情報管理用テーブル

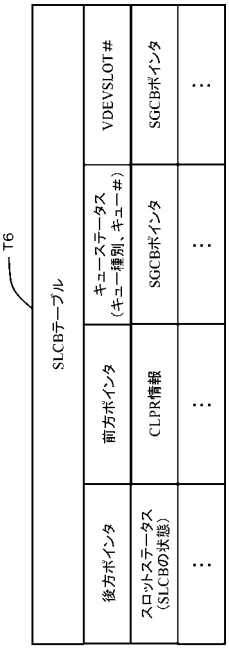
【図６】


ST1
未使用状態
（フリーSGCBキュー）
ST2
デステージ完了時
（WR／RD SEG）
デバイス反映済み状態
（クリーンキュー）
RD HIT
RD MISS／WR MISS
WR HIT
ST3
パリティ無し&
デバイス反映前
状態
（中間ダーティキュー）
パリティ生成開始
ST4
パリティ生成中
状態
（ホストダーティキュー）
RD HIT／WR HIT
パリティ生成完了
ST5
パリティ有り&
ドライブ反映前
状態
（物理ダーティキュー）
RD HIT
WR HIT

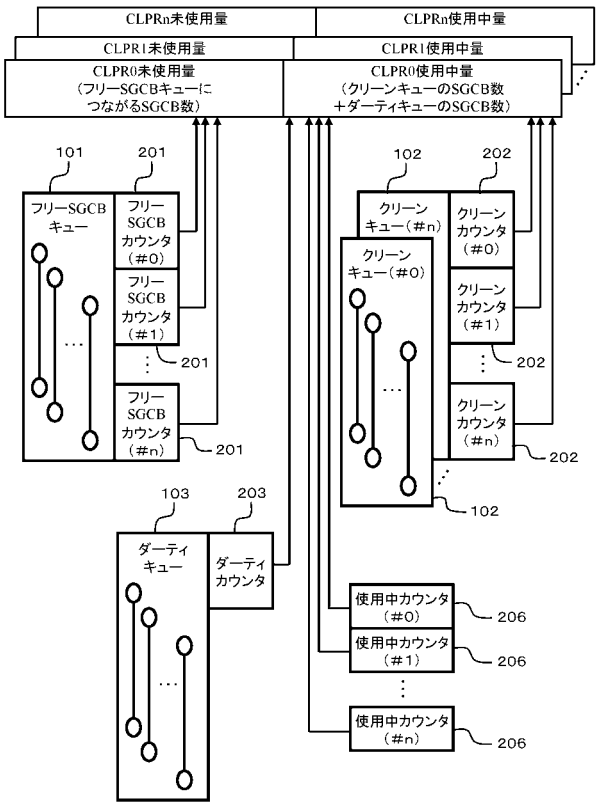
【図７】


キューロック取得
キューロック未取得
SLCB
SLCB
SLCB
SLCB
200
カウンタ
300
キュー管理テーブル
100

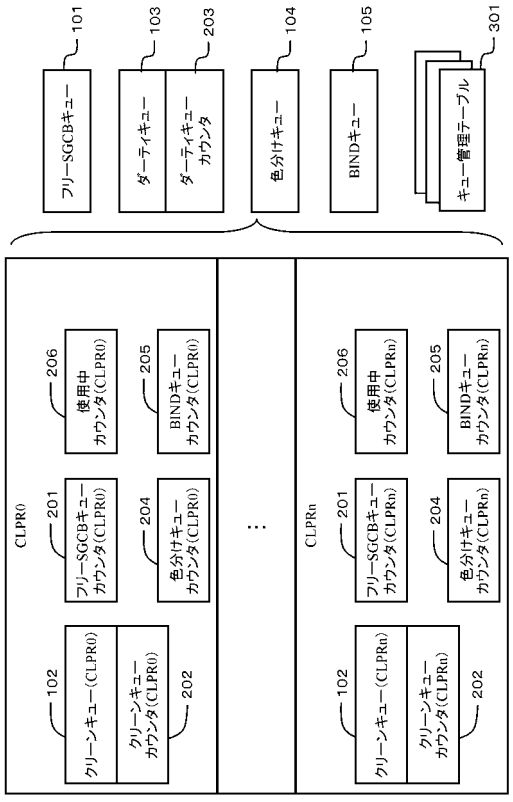
【図８】


T6
SLCBテーブル
後方ポインタ
スロットステータス
（SLCBの状態）
前方ポインタ
CLPR情報
キューステータス
（キュー種別、キュー#）
SGCBポインタ
VDEVSLOT#
SGCBポインタ

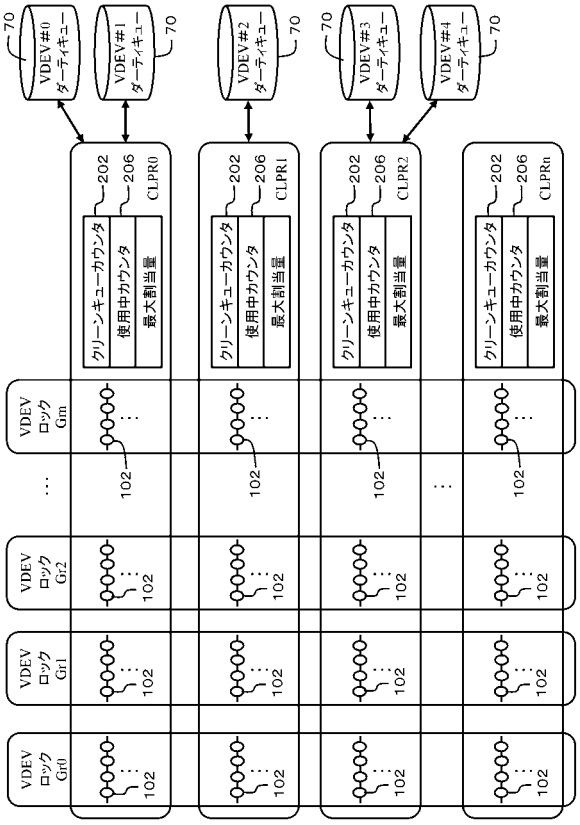
【図９】


CLPRn未使用量
CLPRn使用中量
CLPR1未使用量
CLPR1使用中量
CLPR0未使用量
（フリーSGCBキューに
つながるSGCB数）
CLPR0使用中量
（クリーンキューのSGCB数
＋ダーティキューのSGCB数）
101
201
フリーSGCB
キュー
フリー
SGCB
カウンタ
（#0）
フリー
SGCB
カウンタ
（#1）
フリー
SGCB
カウンタ
（#n）
102
202
クリーン
キュー（#n）
クリーン
キュー（#0）
クリーン
カウンタ
（#0）
クリーン
カウンタ
（#1）
クリーン
カウンタ
（#n）
103
203
ダーティ
キュー
ダーティ
カウンタ
使用中カウンタ
（#0）
使用中カウンタ
（#1）
使用中カウンタ
（#n）
206

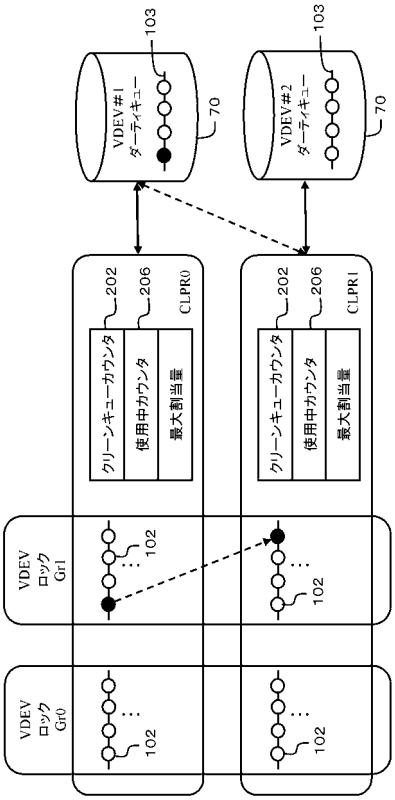
【図１０】


101
フリーSGCBキュー
103
ダーティキュー
203
ダーティキュー
カウンタ
104
色分けキュー
105
BINDキュー
301
キュー管理テーブル
CLPR0
206
使用中
カウンタ(CLPR0)
205
BINDキュー
カウンタ(CLPR0)
201
フリーSGCBキュー
カウンタ(CLPR0)
204
色分けキュー
カウンタ(CLPR0)
102
クリーンキュー(CLPR0)
クリーンキュー
カウンタ(CLPR0)
202
CLPRn
206
使用中
カウンタ(CLPRn)
205
BINDキュー
カウンタ(CLPRn)
201
フリーSGCBキュー
カウンタ(CLPRn)
204
色分けキュー
カウンタ(CLPRn)
102
クリーンキュー(CLPRn)
クリーンキュー
カウンタ(CLPRn)
202

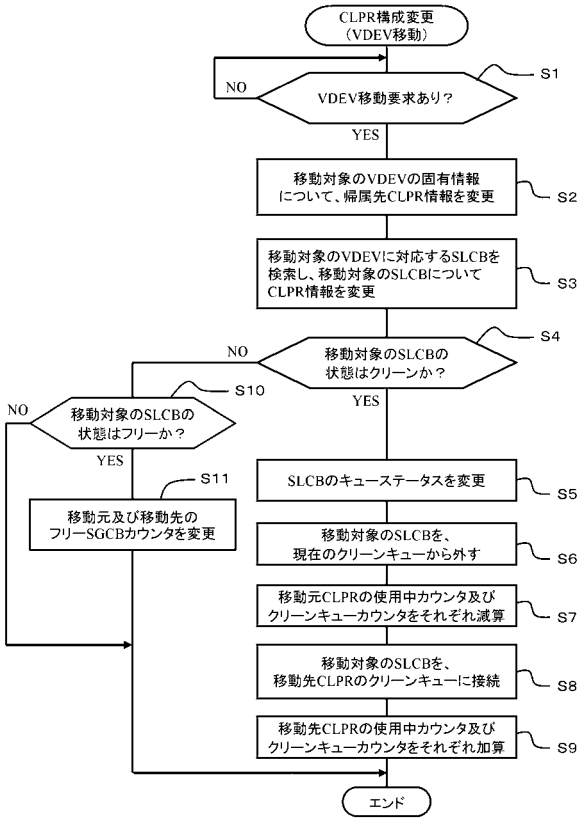
【図１１】



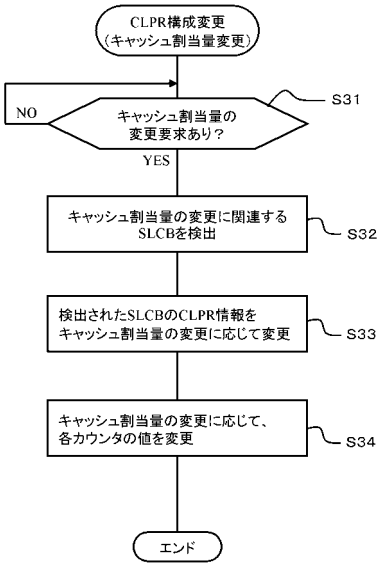
【図１２】



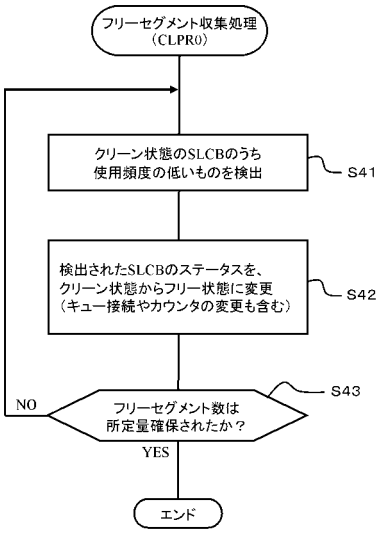
【図１３】



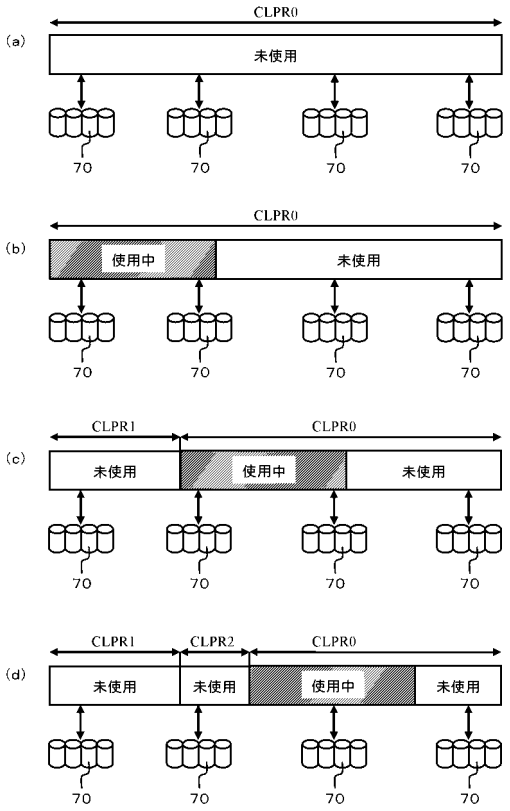
【図１４】



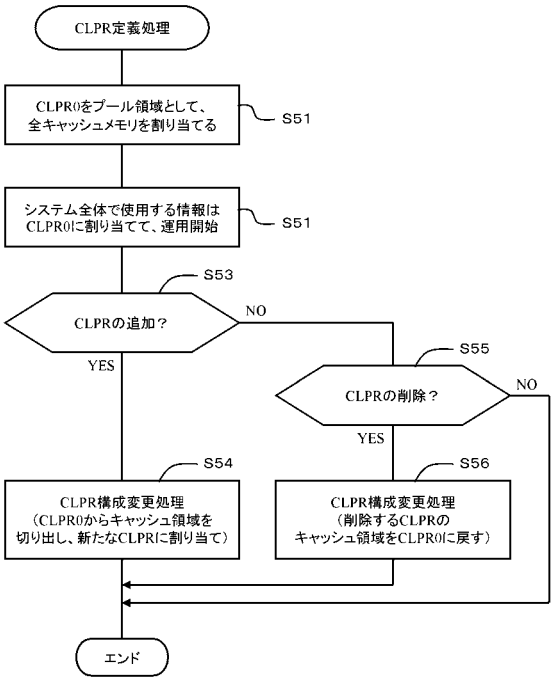
【図１５】



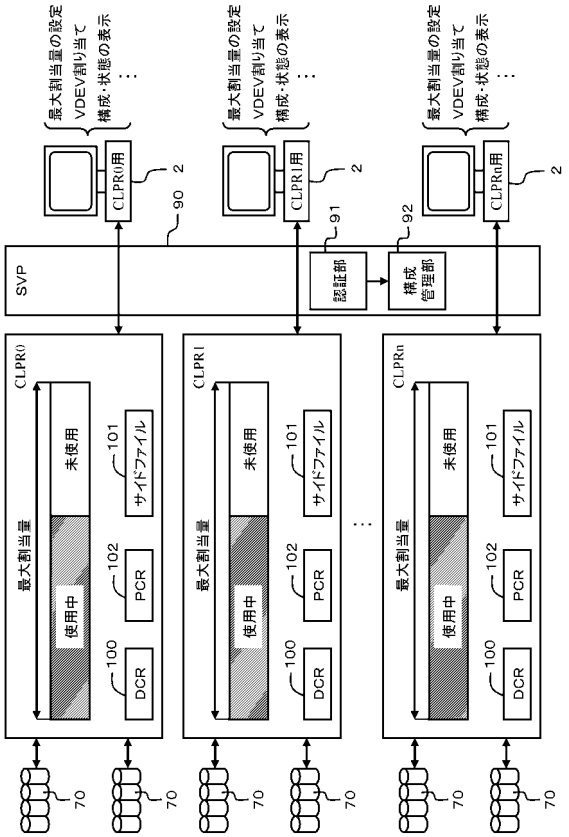
【図１６】



【図１７】



【図１８】



フロントページの続き

(72)発明者　坂口　孝
神奈川県小田原市中里３２２番２号　株式会社日立製作所ＲＡＩＤシステム事業部内
(72)発明者　長副　康之
神奈川県小田原市中里３２２番２号　株式会社日立製作所ＲＡＩＤシステム事業部内
(72)発明者　杉野　昇史
神奈川県小田原市中里３２２番２号　株式会社日立製作所ＲＡＩＤシステム事業部内

審査官　清木　泰

(56)参考文献　特開平０５－１２８００２（ＪＰ，Ａ）
特開平１０－１２４３８８（ＪＰ，Ａ）
特開平０９－１４６８４２（ＪＰ，Ａ）
特開２００５－１９００５７（ＪＰ，Ａ）
特開２００３－１３１９４６（ＪＰ，Ａ）
特開２００１－１８４２４２（ＪＰ，Ａ）
特開平０９－０４４４０６（ＪＰ，Ａ）
特開平０６－２４３０４２（ＪＰ，Ａ）
特開平０１－２２２３５２（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ１２／０８－１２／１２