(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4224077号
(P4224077)
(45)発行日　平成21年2月12日(2009.2.12)　(24)登録日　平成20年11月28日(2008.11.28)

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

請求項の数 7 (全 30 頁)

(21)出願番号　特願2006-103256(P2006-103256)
(22)出願日　平成18年4月4日(2006.4.4)
(65)公開番号　特開2007-279898(P2007-279898A)
(43)公開日　平成19年10月25日(2007.10.25)
審査請求日　平成18年4月4日(2006.4.4)

(73)特許権者　000003078
株式会社東芝
東京都港区芝浦一丁目１番１号
(73)特許権者　301063496
東芝ソリューション株式会社
東京都港区芝浦一丁目１番１号
(74)代理人　100058479
弁理士　鈴江　武彦
(74)代理人　100091351
弁理士　河野　哲
(74)代理人　100088683
弁理士　中村　誠
(74)代理人　100108855
弁理士　蔵田　昌俊

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】
物理ボリュームを有する少なくとも２つのノードを含む複数のノードから構成され、前記少なくとも２つのノードが有する物理ボリュームを利用して、ホストに対して複数の論理エクステントから構成される論理ボリュームを提供するストレージシステムにおいて、
前記複数のノードの少なくとも２つは、前記ホストに対して前記論理ボリュームを提供し、当該ホストからのアクセスを受け付けるコントローラを有し、
前記少なくとも２つのノードは、前記論理ボリュームを構成する論理エクステントを物理エクステントにマッピングして前記物理ボリューム内に格納するストレージを有し、
前記複数のノードの少なくとも２つは、前記論理ボリュームを構成する論理エクステントのうち、自身が担当すべき論理エクステントを格納しているストレージの所在に関する前記コントローラからの問い合わせに対して回答する所在管理サービスを実行する所在管理サーバを有し、
前記コントローラは、前記論理ボリュームを構成する論理エクステントに対応する所在管理サーバを示す第１のマッピングテーブルを有し、
前記所在管理サーバは、前記論理ボリュームを構成する論理エクステントのうち、当該サーバが担当すべき論理エクステントについて、その論理エクステントがどのノードのストレージに格納されているかを示す第２のマッピングテーブルを有し、
前記ストレージは、論理エクステントを、当該ストレージが有する物理ボリュームのどの物理エクステントに対応させて格納したかを示す第３のマッピングテーブルを有し、

前記コントローラは、前記ホストから前記論理ボリューム内の任意の論理エクステントをアクセス先として指定するアクセス要求が与えられた場合、当該コントローラが有する前記第１のマッピングテーブルを参照することにより当該アクセス先の論理エクステントに対応する所在管理サーバを特定し、特定された所在管理サーバに対して、当該論理エクステントがどのストレージに格納されているかを問い合わせ、

前記所在管理サーバは前記コントローラから問い合わせを受けた場合、当該サーバが有する前記第２のマッピングテーブルを参照することにより問い合わせられたストレージを特定して、当該ストレージを示す情報を前記問い合わせに対する回答として問い合わせ元の前記コントローラに通知し、

前記コントローラは、前記所在管理サーバから前記問い合わせに対する回答として通知された情報の示すストレージに対して前記ホストから要求された論理エクステントへのアクセスを要求し、

前記ストレージは、前記コントローラから論理エクステントへのアクセスが要求された場合、当該ストレージが有する前記第３のマッピングテーブルを参照することにより、当該ストレージが有する前記物理ボリューム内で前記要求された論理エクステントに対応させられている物理エクステントを特定して、当該特定された物理エクステントに対してアクセスしてアクセス要求元の前記コントローラに応答し、

論理エクステントを構成する物理エクステント内のデータを前記ストレージをまたがって別の物理エクステントに移動またはコピーするエクステント移動またはコピーが発生した場合、移動元及び移動先またはコピー先のストレージが有する前記第３のマッピングテーブルは当該エクステント移動またはコピーが反映されるように更新され、当該論理エクステントを格納しているストレージの所在の問い合わせに対して回答する所在管理サーバが有する前記第２のマッピングテーブルは、当該論理エクステントを格納しているストレージとして移動先のストレージ、またはコピー元及びコピー先のストレージを示すように更新される

ことを特徴とするストレージシステム。

【請求項２】

前記所在管理サーバが有する前記第２のマッピングテーブルによって示されている、ある論理エクステントがどのノードのストレージに格納されているかを示す情報を別の所在管理サーバが有する前記第２のマッピングテーブルに移動する所在管理サービス移動が発生した場合、移動元及び移動先の前記第２のマッピングテーブル、及び全ての前記コントローラが有する前記第１のマッピングテーブルは、当該所在管理サービス移動が反映されるように更新される

ことを特徴とする請求項１記載のストレージシステム。

【請求項３】

前記所在管理サーバが有する前記第２のマッピングテーブルによって示されている、ある論理エクステントがどのノードのストレージに格納されているかを示す情報を別の所在管理サーバが有する前記第２のマッピングテーブルにコピーする所在管理サービスコピーが発生した場合、コピー先の前記第２のマッピングテーブル、及び全ての前記コントローラが有する前記第１のマッピングテーブルは、当該所在管理サービスコピーが反映されるように更新される

ことを特徴とする請求項１記載のストレージシステム。

【請求項４】

前記第３のテーブルは、前記論理ボリュームの作成時の初期状態において当該論理ボリュームを構成する論理エクステントが格納される物理ボリュームのみを示し、

前記ストレージは、前記コントローラから論理エクステントへのアクセスが要求されて、当該ストレージが有する前記第３のマッピングテーブルを参照した結果、当該ストレージが有する前記物理ボリューム内で前記要求された論理ボリュームに物理エクステントが対応付けられていないことが判別された場合、前記物理ボリューム内の空き物理エクステントを当該論理エクステントに割り付けて前記第３のマッピングテーブルを更新する

ことを特徴とする請求項１記載のストレージシステム。
【請求項５】
前記第２のテーブルは、前記論理ボリュームの作成時の初期状態において、当該サーバに対応する論理エクステントに対してストレージの未割り付けの状態にあることを示し、
前記第３のテーブルは、前記論理ボリュームの作成時の初期状態において、当該論理ボリュームを構成する論理エクステントに対して物理ボリューム及び物理エクステントの未割り付けの状態にあることを示し、
前記所在管理サーバは、前記コントローラからの問い合わせに応じて当該サーバが有する前記第２のマッピングテーブルを参照した結果、前記アクセス先の論理エクステントに対してストレージの未割り付けの状態にあることが判別された場合、空き物理エクステントを有するストレージを当該論理エクステントに割り付けて、前記第２のマッピングテーブルを更新し、
前記ストレージは、前記コントローラから論理エクステントへのアクセスが要求されて、当該ストレージが有する前記第３のマッピングテーブルを参照した結果、当該ストレージが有する前記物理ボリューム内で前記要求された論理ボリュームに物理エクステントが対応付けられていないことが判別された場合、当該ストレージの前記物理ボリューム内の空き物理エクステントを当該論理エクステントに割り付けて、前記第３のマッピングテーブルを更新する
ことを特徴とする請求項１記載のストレージシステム。
【請求項６】
前記コントローラは、前記問い合わせ及び当該問い合わせに対する前記所在管理サーバからの回答に関する情報を格納するキャッシュを有し、
前記コントローラは、前記所在管理サーバに論理エクステントがどのストレージに格納されているかを問い合わせる前に当該問い合わせに対する回答に関する情報が前記キャッシュに格納されているかを調べ、格納されている場合には当該キャッシュに格納されている回答に関する情報の示すストレージに対して前記ホストから要求された論理エクステントへのアクセスを要求する
ことを特徴とする請求項１記載のストレージシステム。
【請求項７】
前記所在管理サーバが有する前記第２のマッピングテーブルの示す論理エクステントが格納されているストレージと、当該所在管理サーバが有するキャッシュに格納されている回答に関する情報の示す当該論理エクステントが格納されているストレージとが一致しているかを調べて、不一致の場合に当該キャッシュに格納されている当該情報を無効化する無効化手段を更に有することを特徴とする請求項６記載のストレージシステム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、物理ボリュームを有する少なくとも１つのノードを含む複数のノードから構成されるストレージシステムに関する。
【背景技術】
【０００２】
近年、相互に接続された複数のストレージ装置（例えばディスクストレージ装置）を有するストレージシステムが開発されている。この種のストレージシステムの１つとして、ストレージクラスタ（ストレージクラスタシステム）が知られている。
【０００３】
ストレージクラスタの特徴は、複数のストレージ装置（以下、ノードと称する）の各々に存在する記憶領域（以下、物理ボリュームと称する）を、当該ノードを超えて統合する（束ねる）ことで、ノード間にまたがった横断的・仮想的な少なくとも１つの記憶領域（以下、論理ボリュームと称する）をホストに提供することにある。ここでのボリュームは、ブロック単位でアクセスされるブロックボリュームであるとする。

【０００４】
ストレージクラスタを構成するノードは、常時稼動のストレージ装置である。つまり、ノードの参加・離脱は、運用中の計画的参加・離脱や、故障による突然の離脱による程度である。このためストレージクラスタでは、例えば毎日の電源ＯＮ／ＯＦＦによるような、頻繁なノードの参加・離脱はない。
【０００５】
ストレージクラスタにおいては、データ（の断片）を物理ボリューム内や、別のノードの物理ボリュームをまたがって、（自動／手動を問わず）適切に移動／コピーすることは、頻繁に起きる。これは、ホストからの論理ボリュームへのアクセス透過性を保障しながら、ストレージクラスタを構成するノードの突発的停止やディスク故障などに備えたデータの冗長化や、アクセス性能・信頼性の向上などの目的による。
【０００６】
一方、ホストは、ストレージクラスタ内の任意の１つのノードに対してパスを張ってアクセスするだけで、複数のノードをまたがって構成された論理ボリューム全体に対してアクセスすることができる。ここで、ホストがパスを張っているノードで障害が発生して、当該障害が発生したノード（障害ノード）が停止したものとする。この場合でも、障害ノードに格納されているデータがストレージクラスタ内の他のノード（生存ノード）に冗長化されていれば、ホストは任意の生存ノードに対してパスを張り直すことで、論理ボリュームに対するアクセスを継続することができる。
【０００７】
このようなストレージクラスタを構築するには、論理ボリューム上のデータが、どのノードのどの物理ボリューム上の、どの位置に配置されているかを管理するための仕組みが必要になる。
【０００８】
一般に、ストレージクラスタのノードは、機能的にコントローラとストレージとから構成される。コントローラはホストとのパスを管理しアクセスを受け付ける。ホストはどのコントローラに対してパスを張っても、同じ論理ボリュームにアクセスすることができる。ストレージはデータを蓄える物理ボリュームを有する。
【０００９】
ボリューム（ブロックボリューム）は、エクステントと呼ばれる、固定長のブロックの集合により構成される。各ノードのストレージは１つの物理ボリュームを管理する。一方、各ノードのコントローラは、物理ボリュームを構成する物理エクステントと、論理ボリュームの論理エクステントとをマッピングテーブルにより対応付けて管理する。これにより、論理ボリューム（論理ボリューム空間）が構成される。このようなマッピングテーブルによる物理エクステントと論理エクステントとの対応付けは、例えば特許文献１にも記載されている。
【００１０】
上述のマッピングテーブルを、論理エクステント－物理エクステントマッピングテーブル（ＬＰＭＴ：Logical Extent to Physical Extent Mapping Table）と呼ぶ。ＬＰＭＴは、論理ボリュームの論理エクステントが、どのノード（ストレージ）の物理ボリュームの、どの物理エクステントから構成されているか、その対応付けを管理するのに用いられる。
【００１１】
ストレージクラスタ内の全てのノードのコントローラは、それぞれ同一内容のＬＰＭＴを保持する。また各ノードは、ＬＰＭＴにより全論理ボリューム分を管理する。これらの理由は、ホストからコントローラへのパスが、別のコントローラに切り替わった際に、直ちに同一の論理ボリュームへのアクセスを継続可能とするためである。
【特許文献１】特開平９－６２４５２号公報
【発明の開示】
【発明が解決しようとする課題】

【００１２】
上記したように従来のストレージクラスタ（ストレージシステム）では、当該クラスタ内の各ノードは、ＬＰＭＴ（論理エクステント－物理エクステントマッピングテーブル）により全論理ボリューム分を管理する必要がある。ところが、ＬＰＭＴのサイズは物理ボリュームの数と、その物理ボリュームを構成するエクステントの量によってはとても大きくなり、コントローラ上の大量のメモリを消費することになる。
【００１３】
ストレージクラスタでは、当該クラスタ内のあるノード（以下、第１のノードと称する）の例えば第１の物理エクステントに格納されていた第１の論理エクステントを、別のノード（以下、第２のノードと称する）の例えば第２の物理エクステントにマイグレーション（移動）することが発生する。このようなマイグレーションは、第１のノードのストレージの負荷を軽減するなどの目的で行われる。また、上記特許文献１に記載されているように、あるノードのストレージに対するアクセス性能を向上させるために、当該ノードのある物理エクステントに格納されていた第１の論理エクステントを、当該ノードの別の物理エクステントにマイグレーションすることもある。このマイグレーションにより、ストレージクラスタのアクセス性能及び信頼性を向上することが可能となる。
【００１４】
このようにストレージクラスタでは、ストレージクラスタのアクセス性能・信頼性向上のために、論理エクステントを同じストレージ内の別の物理エクステントへ、また別のストレージの物理エクステントへの移動／コピーが発生する。この場合、ストレージクラスタの各コントローラは互いに同期をとって、自身が管理するＬＰＭＴを一斉（同時）に更新する必要がある。しかし、各コントローラのＬＰＭＴが同一となるように同時に更新することは、ＬＰＭＴ管理の処理の複雑さを招き、アクセス性能に悪影響を及ぼす。
【００１５】
本発明は上記事情を考慮してなされたものでその目的は、各ノードが管理する情報の更新作業量を抑えてアクセス性能を向上させ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】
【００１６】
本発明の１つの観点によれば、物理ボリュームを有する少なくとも１つのノードを含む複数のノードから構成され、前記少なくとも１つのノードが有する物理ボリュームを利用して、ホストに対して複数の論理エクステントから構成される論理ボリュームを提供するストレージシステムが提供される。このシステムにおいて、前記複数のノードの少なくとも２つは、前記ホストに対して前記論理ボリュームを提供し、当該ホストからのアクセスを受け付けるコントローラを有する。また、前記少なくとも１つのノードは、前記論理ボリュームを構成する論理エクステントを物理エクステントにマッピングして前記物理ボリューム内に格納するストレージを有し、前記複数のノードの少なくとも１つは、前記論理ボリュームを構成する論理エクステントを格納しているストレージの所在に関する前記コントローラからの問い合わせに対して回答する所在管理サービスを実行する所在管理サーバを有する。また、前記コントローラは、前記論理ボリュームを構成する論理エクステントに対応する所在管理サーバを示す第１のマッピングテーブルを有する。前記所在管理サーバは、当該サーバに対応する論理エクステントについて、その論理エクステントがどのノードのストレージに格納されているかを示す第２のマッピングテーブルを有し、前記ストレージは、論理エクステントを、当該ストレージが有する物理ボリュームのどの物理エクステントに対応させて格納したかを示す第３のマッピングテーブルを有する。前記コントローラは、前記ホストから前記論理ボリューム内の任意の論理エクステントをアクセス先として指定するアクセス要求が与えられた場合、当該コントローラが有する前記第１のマッピングテーブルを参照することにより当該アクセス先の論理エクステントに対応する所在管理サーバを特定し、特定された所在管理サーバに対して、当該論理エクステントがどのストレージに格納されているかを問い合わせる。前記所在管理サーバは前記コントロ

ーラから問い合わせを受けた場合、当該サーバが有する前記第２のマッピングテーブルを参照することにより問い合わせられたストレージを特定して、当該ストレージを示す情報を前記問い合わせに対する回答として問い合わせ元の前記コントローラに通知する。前記コントローラは、前記所在管理サーバから前記問い合わせに対する回答として通知された情報の示すストレージに対して前記ホストから要求された論理エクステントへのアクセスを要求する。前記ストレージは、前記コントローラから論理エクステントへのアクセスが要求された場合、当該ストレージが有する前記第３のマッピングテーブルを参照することにより、当該ストレージが有する前記物理ボリューム内で前記要求された論理エクステントに対応させられている物理エクステントを特定して、当該特定された物理エクステントに対してアクセスしてアクセス要求元の前記コントローラに応答する。

【発明の効果】

【００１７】

本発明によれば、第１乃至第３のマッピングテーブルを用いることにより、ノード（コントローラ）が管理する情報の更新作業量を抑えてアクセス性能を向上させることができる。

【発明を実施するための最良の形態】

【００１８】

以下、本発明の実施の形態につき図面を参照して説明する。

［第１の実施形態］

図１は本発明の第１の実施形態に係るストレージクラスタ構成のストレージシステムの概略構成を示すブロック図である。図１のストレージシステム（以下、ストレージクラスタと称する）は、複数、例えば３台のストレージ装置（以下、ノードと称する）１-1，１-2及び１-3（＃１，＃２及び＃３）から構成される。ノード１-1，１-2及び１-3は、ファイバチャネル、或いはｉＳＣＳＩ（Small Computer System Interface）のようなインタフェイス２によって、ストレージクラスタを利用するホスト（ホストコンピュータ）と接続されている。

【００１９】

ノード１-1，１-2及び１-3は、それぞれ、コントローラ１１-1，１１-2及び１１-3と、所在管理サーバ１２-1，１２-2及び１２-3と、ストレージ１３-1，１３-2及び１３-3とから構成される。

【００２０】

コントローラ１１-i（ｉ＝１，２，３）はホストに対して論理ボリュームを提供する。ここでは、ストレージ１３-1，１３-2及び１３-3の後述する物理ボリュームを統合することで、論理ボリュームＬＶｏｌ＝１がホストに提供されるものとする。コントローラ１１-iはホストからのアクセス要求を受け付ける。アクセス要求は、アクセスされるべき論理ボリューム内の論理ブロックを示すアドレス（論理ブロックアドレス）ＬＢＡを含む。コントローラ１１-iは、要求された論理ブロックを含む論理エクステントがどのノードのストレージに格納されているか、後述するＬＭＭＴ１１２ａを参照して所在管理サーバ１２-j（ｊ＝１，２，３）に問い合わせる。コントローラ１１-iは、問い合わせに対して所在管理サーバ１２-jから通知されたノードのストレージに対してアクセスを行う。

【００２１】

図２は、コントローラ１１-iの構成を示すブロック図である。コントローラ１１-iは、ターゲットインタフェイス１１１、サーバ特定部１１２及び入出力マネージャ（ＩＯマネージャ）１１３を含む。

【００２２】

ターゲットインタフェイス１１１は、ホストからのアクセス要求を受け付け、要求された論理ボリュームの論理ブロックアドレスＬＢＡ（の示す論理ブロック）が位置する論理エクステントを特定する。ここでは、論理エクステントを識別する論理エクステントＩＤ（ＬＥＩＤ）が、（ＬＥＩＤ＝ＬＢＡ／エクステントサイズ）の演算によって特定される。ターゲットインタフェイス１１１は、特定された論理ボリューム内の論理エクステント

がどのストレージに格納されているかをサーバ特定部１１２を介して所在管理サーバ１２-jに問い合わせる。この問い合わせは、論理ボリューム／論理エクステントを識別する論理ボリュームＩＤ（ＬＶｏｌ）／論理エクステントＩＤ（ＬＥＩＤ）を含む。ターゲットインタフェイス１１１は、所在管理サーバ１２-iに対する問い合わせに応じてサーバ特定部１１２を介して通知されたストレージに対してＩＯマネージャ１１３を介してアクセスし、ホストに対して応答する。

【００２３】

サーバ特定部１１２は、ターゲットインタフェイス１１１からの問い合わせを受け渡す所在管理サーバ１２-jを特定する。以下の説明では、ターゲットインタフェイス１１１からの問い合わせに含まれている論理ボリュームＩＤ及び論理エクステントＩＤで示される論理エクステントをターゲット論理エクステントと呼ぶこともある。所在管理サーバ１２-jは、このターゲット論理エクステントを格納するストレージの所在を管理するサーバとして特定される。

【００２４】

サーバ特定部１１２は、ターゲット論理エクステントの所在を管理する所在管理サーバを特定するために、第１のマッピングテーブルとしての論理エクステント－所在管理サーバマッピングテーブル（ＬＭＭＴ：LogicalExtent to ManagementServer Mapping Table）１１２ａを有する。ＬＭＭＴ１１２ａは、論理エクステントをどのストレージが格納しているか、その所在に関して責任を負う（つまり所在を管理する）所在管理サーバを提供するノードのＩＤ（ＭＳ）を保持する。サーバ特定部１１２は、コントローラ１１-1からの問い合わせを特定された所在管理サーバに受け渡す。サーバ特定部１１２は、ターゲットインタフェイス１１１からの上記問い合わせに対する所在管理サーバ１２-jからの応答（ターゲット論理エクステントが格納されているストレージのノードＩＤ（ＰＶｏｌ）を当該ターゲットインタフェイス１１１に受け渡す。サーバ特定部１１２はまた、キャッシュ１１２ｂを有する。このキャッシュ１１２ｂについては、後述する第５の実施形態で説明する。

【００２５】

ＩＯマネージャ１１３は、コントローラ１１-iとストレージ１３-jとの間のインタフェイスを提供する。

【００２６】

再び図１を参照すると、所在管理サーバ１２-i（ｉ＝１，２，３）は、コントローラ１１-j（ｊ＝１，２，３）によって問い合わせられた論理エクステントを格納しているストレージのノードＩＤを調べて、そのノードＩＤを当該コントローラ１１-jに通知する。所在管理サーバ１２-iは、自身が担当する論理エクステントについて、その論理エクステントがどのノードＩＤのストレージ（物理ボリューム）に格納されているかを管理するための、第２のマッピングテーブルとしての論理エクステント－ストレージマッピングテーブル（ＬＳＭＴ：Logical Extent to Storage Mapping Table）ＬＳＭＴ１２０-iを有する。つまり所在管理サーバ１２-1，１２-2及び１２-3は、それぞれＬＳＭＴ１２０-1，１２０-2及び１２０-3を有する。

【００２７】

ストレージ１３-1，１３-2及び１３-3は、それぞれ物理ボリューム１３０-1，１３０-2及び１３０-3（ＰＶｏｌ＝１，ＰＶｏｌ＝２及びＰＶｏｌ＝３）を有する。ストレージ１３-i（ｉ＝１，２，３）は論理エクステントを物理エクステントにマッピングして物理ボリューム１３０-i内に格納する。ストレージ１３-iは、コントローラ１１-iからの入出力要求（ＩＯ要求）を受け付ける。このＩＯ要求は、アクセスされるべき論理ボリューム内の論理エクステントのＩＤ（ＬＥＩＤ）を含む。ストレージ１３-iは、この論理エクステントと対応付けられている物理エクステントを調べる。ストレージ１３-iは、この物理エクステントに対してＩＯを行い、コントローラ１１-iに対して応答する。

【００２８】

図３は、ストレージ１３-iの構成を示す。ストレージ１３-iは、ＩＯドライバ１３１、

論理エクステント－物理エクステント管理部（ＬＰ管理部）１３２及びディスク１３３を含む。

【００２９】

ＩＯドライバ１３１は、ストレージ１３-iとコントローラ１１-iとのインタフェイスを提供する。

【００３０】

ＬＰ管理部１３２は、論理エクステントと物理エクステントとの対応を管理する。ＬＰ管理部１３２は、論理エクステントが、自身の物理ボリューム１３０-iのどの物理エクステントに対応させて格納されているかを示す、第３のマッピングテーブルとしての論理エクステント－物理エクステントマッピングテーブル（ＬＰＭＴ：LogicalExtent to PhysicalExtent Mapping Table）１３２ａを保持する。このように、本実施形態で適用されるＬＰＭＴ１３２ａは、［背景技術］で述べたＬＰＭＴと異なり、全論理ボリューム分のマッピング情報を持たないことに注意されたい。また、ＬＰＭＴ１３２ａの内容は、各ストレージ１３-i毎に異なることにも注意されたい。

【００３１】

ディスク１３３は、物理ボリューム１３０-i（ＰＶｏｌ＝ｉ）を有する。物理ボリューム１３０-iは物理エクステント単位で管理される。また、物理ボリューム１３０-iへのアクセスの単位はブロックであり、物理エクステントは、複数のブロックから構成される。

【００３２】

また、図１のストレージクラスタは、当該クラスタの構成の計画・管理を司る構成管理部（図示せぬ）を有する。この構成管理部は、ボリューム作成時に、論理ボリューム／論理エクステントの格納先となる物理ボリューム／物理エクステントを決定する。構成管理部はまた、エクステントや後述する所在管理サービスの移動／コピー（冗長化）のための移動元／移動先を決定する。構成管理部は更に、オンデマンドでエクステント割り付けの割り付け先を決定する。この構成管理部が、ノード１-1～１-3から独立に設けられても、当該ノード１-1～１-3のいずれかに設けられても構わない。

【００３３】

本実施形態において、構成管理部は、以下の手順で論理ボリュームを作成する。

（ａ）構成管理部は、論理ボリュームを構成する論理エクステントの各々を、どのノードＩＤのストレージに格納するか決定する。構成管理部は、論理エクステントの各々に、決定されたストレージの物理エクステントを割り付ける。構成管理部は、ストレージ１３-1～１３-3毎に、そのストレージの物理エクステントと当該物理エクステントが割り付けられた論理エクステントとの対応を示すＬＰＭＴ１３２ａを生成して、そのストレージに保持する。

【００３４】

（ｂ）構成管理部は、論理エクステントが割り付けられた物理エクステントを格納するストレージ（物理ボリューム）から、その論理エクステントの所在を管理するサービス（所在管理サービス）をノード１-1～１-3のうちのいずれのノードの所在管理サーバに担当させるかを決定する。構成管理部は、所在管理サーバ１２-1～１２-3毎に、その所在管理サーバが担当する論理エクステントと物理ボリュームとの対応を示すＬＳＭＴ１２０-1～１２０-3を生成して、その所在管理サーバ１２-1～１２-3内に設定する。なお、後述する第２の実施形態のように、ストレージクラスタ内の全ての所在管理サーバ１２-1～１２０-3にＬＳＭＴが設定される必要はない。

【００３５】

（ｃ）構成管理部は、ＬＭＭＴ１１２ａを生成して全コントローラ１１-1～１１-3に配布する。ＬＭＭＴ１１２ａは先に述べたように、論理ボリュームを構成する論理エクステント毎に、その論理エクステントが割り付けられている物理エクステントを格納するストレージ（物理ボリューム）をどの所在管理サーバに問い合わせればよいかを示す。

【００３６】

このようにして、論理ボリュームが作成される。すると、作成された論理ボリュームに

対してホストからアクセスすることが可能となる。ホストからの論理ボリュームに対するアクセスは、ホストからノード１-1～１-3のうちの任意のノード１-iのコントローラ１１-iに対してパスを張ることで行われる。

【００３７】

以下、例えばホストからノード１-1（ｉ＝１の場合）に対してパスが張られた後に、当該ホストからアクセス要求が発行された場合の処理の流れについて、図４及び図５を参照して説明する。図４は処理の手順を示すフローチャート、図５は情報の流れを示す図である。図５における矢印Ａ１～Ａ８は、図４のフローチャートにおけるステップＳ１～Ｓ８にそれぞれ対応する。

【００３８】

まず、ホストからノード１-1のコントローラ１１-1に含まれているターゲットインタフェイス１１１に対してアクセス要求（ＩＯ要求）が発行されたものとする。また、このアクセス要求が、ある論理ボリュームＬＶｏｌ＝ＬＶｏｌｘ（例えばＬＶｏｌ＝１）のある論理ブロックアドレスＬＢＡ＝ＬＢＡｘに対するアクセスを要求しているものとする。

【００３９】

ターゲットインタフェイス１１１は、ホストからのアクセス要求を受け付ける（ステップＳ０）。するとターゲットインタフェイス１１１は、ホストからのアクセス要求で指定されるアクセスされるべき論理ボリュームＬＶｏｌ＝ＬＶｏｌｘ上のＬＢＡ＝ＬＢＡｘから、当該ＬＢＡｘが位置する論理エクステントのＩＤ＝ＬＥＩＤｘを次式

ＬＥＩＤｘ＝ＬＢＡｘ／エクステントサイズ

により求める（ステップＳ１）。ここで、論理ボリュームＬＶｏｌ＝ＬＶｏｌｘ＝１上の論理エクステントＩＤがＬＥＩＤｘの論理エクステントを、論理エクステント（ＬＶｏｌｘ，ＬＥＩＤｘ）と表現する。

【００４０】

ターゲットインタフェイス１１１は、論理ボリューム名（ＩＤ：ＬＶｏｌ＝ＬＶｏｌｘ＝１）及び求められた論理エクステントＩＤ＝ＬＥＩＤｘをサーバ特定部１１２に渡すことにより、当該論理エクステントＩＤ＝ＬＥＩＤｘの論理エクステント（ＬＶｏｌｘ，ＬＥＩＤｘ）が、どのノードのストレージに格納されているかを問い合わせる（ステップＳ２）。

【００４１】

サーバ特定部１１２は、ターゲットインタフェイス１１１によって渡された論理エクステントＩＤ＝ＬＥＩＤｘの論理エクステント（ＬＶｏｌｘ，ＬＥＩＤｘ）がどのノードのストレージに格納されているかを管理している所在管理サーバを提供するノードのＩＤ（ＭＳ＝ＭＩＤｘ）を、ＬＭＭＴ１１２ａに基づいて特定する（ステップＳ３）。ここでは、サーバ特定部１１２と同じノード１-1、つまり所在管理サーバ１２-1を提供するノード１-1のＩＤ（ＭＳ＝ＭＩＤｘ＝１）が特定されたものとする。本実施形態において、ノード１-1～１-3のストレージ１３-1～１３-3は物理ボリューム１３０-1～１３０-3を有する。つまり、ノード１-1～１-3のストレージ１３-1～１３-3は、それぞれ１つの物理ボリュームを有する。このため本実施形態では、所在管理サーバ１２-1～１２-3を提供するノード１-1～１-3のＩＤ（ＭＳ）と当該ノード１-1～１-3のストレージ１３-1～１３-3が有する物理ボリューム１３０-1～１３０-3のＩＤ（ＰＶｏｌ）とは一致し、１～３である。

【００４２】

サーバ特定部１１２は、特定されたノード１-1の所在管理サーバ１２-1に対して、ターゲットインタフェイス１１１からの問い合わせを受け渡す（ステップＳ４）。即ち、サーバ特定部１１２は所在管理サーバ１２-1に対して、論理エクステントＩＤ＝ＬＥＩＤｘの論理エクステント（ターゲット論理エクステント）（ＬＶｏｌｘ，ＬＥＩＤｘ）に対応する物理エクステントを格納しているストレージのノードＩＤ（ＰＶｏｌ）＝ＳＩＤｘを問い合わせる。

【００４３】

所在管理サーバ１２-1は、どのノードＩＤ＝ＳＩＤｘのストレージが、論理エクステン

ト（ＬＶｏｌｘ，ＬＥＩＤｘ）に対応する物理エクステントを格納しているかを、ＬＳＭＴ１２０-1に基づいて調べて、そのノードＩＤ＝ＳＩＤｘを、問い合わせ元のサーバ特定部１１２に対して通知する（ステップＳ５）。

【００４４】

サーバ特定部１１２は、所在管理サーバ１２-1から通知されたノードＩＤ＝ＳＩＤｘを、ターゲットインタフェイス１１１からの問い合わせに対する回答として当該ターゲットインタフェイス１１１に渡す（ステップＳ６）。ここでは、ストレージ１３-1のノードＩＤ＝ＳＩＤｘがターゲットインタフェイス１１１に渡されたものとする。

【００４５】

ターゲットインタフェイス１１１は、サーバ特定部１１２によって渡されたノードＩＤ＝ＳＩＤｘで示されるノード１-1のストレージ１３-1に対し、論理エクステント（ＬＶｏｌｘ，ＬＥＩＤｘ）に対応する物理エクステントへのアクセスをＩＯマネージャ１１３を介して要求する（ステップＳ７）。

【００４６】

ストレージ１３-1のＩＯドライバ１３１は、要求された論理エクステント（ＬＶｏｌｘ，ＬＥＩＤｘ）に、当該ストレージ１３-1（物理ボリューム１３０-1）内のどのＩＤ（＝ＰＥＩＤｘ）の物理エクステント（以下、物理エクステントＰＥＩＤｘと表現する）が対応付けられているかを、ＬＰＭＴ１３２ａに基づいて調べる（ステップＳ８）。このステップＳ８において、ＩＯドライバ１３１は、論理エクステント（ＬＶｏｌｘ，ＬＥＩＤｘ）に対応する物理エクステントＰＥＩＤｘに対してアクセスし、コントローラ１１-1のＩＯマネージャ１１３に対して応答する。ターゲットインタフェイス１１１は、ストレージ１３-1のＩＯドライバ１３１からの応答をＩＯマネージャ１１３から受け取って、ホストに返す（ステップＳ９）。

【００４７】

以上の説明では、ホストからのアクセス要求を受け付けたターゲットインタフェイスと、ターゲット論理エクステントがどのストレージに格納されているかを管理している所在管理サーバと、当該ターゲット論理エクステントを物理エクステント内に格納しているストレージとが、いずれも同一のノード１-1に属する場合を想定している。

【００４８】

図６は、上述のターゲットインタフェイス、所在管理サーバ及びストレージが、それぞれ別々のノードに属する場合の情報の流れを示す。図６における矢印Ｂ１～Ｂ８は、図４のフローチャートにおけるステップＳ１～Ｓ８にそれぞれ対応する。

【００４９】

図７は、図１のシステムにおいて、論理ボリュームＬＶｏｌがＬＶｏｌ＝１であって、当該論理ボリュームＬＶｏｌ＝１を、ＬＥＩＤがＬＥＩＤ＝１～ＬＥＩＤ＝４の４つの論理エクステントから構成した場合のシステム構成を示す。また、図８は図７の状態におけるＬＭＭＴ１１２ａのデータ構造例を、図９は図７の状態におけるＬＳＭＴ１２０-i、例えばノード１-1のＬＳＭＴ１２０-1のデータ構造例を、図１０は図７の状態における例えばノード１-1のＬＰＭＴ１３２ａのデータ構造例を示す。

【００５０】

ＬＭＭＴ１１２ａは、論理ボリュームＬＶｏｌ＝１を構成する論理エクステントの数に一致するエントリを有する。ＬＭＭＴ１１２ａの各エントリは、論理ボリュームのＩＤ（論理ボリューム番号ＬＶｏｌ）と、当該ＬＶｏｌで示される論理ボリュームに含まれる論理エクステントのＩＤ（ＬＥＩＤ）と、当該ＬＶｏｌ及びＬＥＩＤで示される論理エクステント（ＬＶｏｌ，ＬＥＩＤ）をどのストレージ（物理ボリューム）が格納しているか、その所在を管理する所在管理サーバ１２-jを提供するノードのＩＤ（ＭＳ）とを保持する。図８の例では、論理ボリュームＬＶｏｌ＝１を構成する論理エクステントが、ＬＥＩＤ＝１～ＬＥＩＤ＝４の４つであることが示されている。

【００５１】

ＬＳＭＴ１２０-i（１２０-1）は、所在管理サーバ１２-i（１２-1）が管理する論理エ

クステントの数に一致するエントリを有する。ＬＳＭＴ１２０-i（１２０-1）の各エントリは、論理ボリュームのＩＤ（ＬＶｏｌ）と、当該ＬＶｏｌで示される論理ボリュームに含まれる論理エクステントのＩＤ（ＬＥＩＤ）と、当該ＬＶｏｌ及びＬＥＩＤで示される論理エクステントが格納されているストレージ（物理ボリューム）のノードＩＤ（物理ボリューム番号ＰＶｏｌ）とを保持する。

【００５２】

ＬＰＭＴ１３２ａは、物理ボリューム１３０-i（１３０-1）を構成する物理エクステントの数に一致するエントリを有する。ＬＰＭＴ１３２ａの各エントリは、物理ボリューム１３０-iを構成する物理エクステントのＩＤ（ＰＥＩＤ）と、当該ＰＥＩＤで示される物理エクステントが割り付けられる論理エクステントを含む論理ボリュームのＩＤ（ＬＶｏｌ）と、当該論理エクステントのＩＤ（ＬＥＩＤ）と、当該論理エクステントをどのストレージ（物理ボリューム）が格納しているか、その所在を管理する所在管理サービス（所在管理サーバ１２-j）を提供するノードのＩＤ（ＭＳ）とを保持する。このＬＰＭＴ１３２ａ内のエントリのＭＳは、所在管理サービスの逆引きに用いられる。図１０の例では、物理ボリューム１３０-i（１３０-1）を構成する物理エクステントが、ＰＥＩＤ＝１～ＰＥＩＤ＝３の３つであることが示されている。

【００５３】

図７において、コントローラ１１-1のＬＭＭＴ１１２ａのエントリから所在管理サーバ１２-1～１２-3のＬＳＭＴ１２０-1～１２０-3へ向かう一点鎖線の矢印は、ＬＭＭＴ１１２ａのエントリの情報によって参照されるＬＳＭＴ１２０-1～１２０-3のエントリを指す。同様に、ＬＳＭＴ１２０-1～１２０-3のエントリからストレージ１３-1～１３-2のＬＰＭＴ１３２ａへ向かう一点鎖線の矢印は、ＬＳＭＴ１２０-1～１２０-3のエントリの情報によって参照されるＬＰＭＴ１３２ａのエントリを指す。同様に、ストレージ１３-1～１３-2のＬＰＭＴ１３２ａのエントリからストレージ１３-1～１３-2の物理ボリューム１３０-1～１３０-3へ向かう一点鎖線の矢印は、ＬＰＭＴ１３２ａの情報に基づいてアクセスされる物理エクステントを指す。

【００５４】

次に、図７を参照して、ホストからノード１-1に対してパスが張られた後に、当該ホストからの、論理ボリュームＬＶｏｌ＝１／論理エクステントＬＥＩＤ＝１で指定されるブロックに対するアクセスの流について説明する。

【００５５】

まず、図７において矢印Ｂ１で示されるように、ホストからノード１-1に対して、論理ボリュームＬＶｏｌ＝１内のあるブロックへのアクセス要求が発行されたものとする。ノード１-1のコントローラ１１-1（に含まれているターゲットインタフェイス１１１）は、アクセスされるべきブロックがＬＥＩＤ＝１の論理エクステント（論理エクステントＬＥＩＤ＝１）に格納されていることを求める。

【００５６】

コントローラ１１-1のターゲットインタフェイス１１１は、論理エクステントＬＥＩＤ＝１がどのノードのストレージに格納されているかを、当該コントローラ１１-1内のサーバ特定部１１２に問い合わせる。サーバ特定部１１２は、自身のＬＭＭＴ１１２ａ内のＬＥＩＤ＝１が保持されているエントリ（図８の例では、１行目のエントリ）を参照する。これによりサーバ特定部１１２は、論理ボリュームＬＶｏｌ＝１／論理エクステントＬＥＩＤ＝１に対応する所在管理サーバがノード１-1に存在することを認識する。

【００５７】

コントローラ１１-1のサーバ特定部１１２は、図７において矢印Ｂ２で示されるように、ノード１-1に存在する所在管理サーバ１２-1に対して、論理ボリュームＬＶｏｌ＝１／論理エクステントＬＥＩＤ＝１がどのノードのストレージ（物理ボリューム）に格納されているかを問い合わせる。

【００５８】

所在管理サーバ１２-1は、自身のＬＳＭＴ１２０-1内でＬＶｏｌ＝１／ＬＥＩＤ＝１が

保持されているエントリ（図９の例では、１行目のエントリ）を参照して、論理ボリュームＬＶｏｌ＝１／論理エクステントＬＥＩＤ＝１の実体がノード１-1のストレージ（物理ボリューム）に格納されていることを認識する。所在管理サーバ１２-1は、ノード１-1のＩＤ（ＳＩＤｘ＝ＰＶｏｌ＝１）を、図７において矢印Ｂ３で示されるように、コントローラ１１-1のサーバ特定部１１２に回答する（ステップＢ３）。これを受けてサーバ特定部１１２は、ノード１-1のＩＤ（ＳＩＤｘ＝ＰＶｏｌ＝１）をターゲットインタフェイス１１１に通知する。

【００５９】

ターゲットインタフェイス１１１（コントローラ１１-1のターゲットインタフェイス１１１）は、図７において矢印Ｂ４で示されるように、サーバ特定部１１２によって通知されたノードＩＤで指定されるストレージ、即ちノード１-1のストレージ１３-1に対して、論理ボリュームＬＶｏｌ＝１／論理エクステントＬＥＩＤ＝１へのアクセスをＩＯマネージャ１１３を介して要求する。

【００６０】

ノード１-1のストレージ１３-1は、自身のＬＰＭＴ１３２ａ内の論理ボリュームＬＶｏｌ＝１／論理エクステントＬＥＩＤ＝１が保持されているエントリ（図１０の例では、１行目のエントリ）を参照する。これによりストレージ１３-1は、論理ボリュームＬＶｏｌ＝１／論理エクステントＬＥＩＤ＝１に対応する物理エクステントのＰＥＩＤがＰＥＩＤ＝１であることを認識する。ストレージ１３-1は、このＰＥＩＤ＝１の物理エクステント（物理エクステントＰＥＩＤ＝１）にアクセスして、図７において矢印Ｂ５で示されるように、コントローラ１１-1に対して応答する。

【００６１】

本実施形態において、各ノード１-1～１-3のコントローラ１１-1～１１-3が有するＬＭＭＴ１１２ａは同一の内容である。つまり各ノード１-1～１-3（コントローラ１１-1～１１-3）はＬＭＭＴ１１２ａを共有する。このためホストは、ストレージクラスタを構成するノード１-1～１１-3のいずれのノード（のコントローラ）に対してもパスを張ることができる。例えば、ホストがパスを張っている先のコントローラが異常停止した場合、或いは当該コントローラの負荷が高まったときなどは、ホストは別のコントローラに対してパスを張り替えることができる。

【００６２】

ＬＭＭＴ１１２ａなど、全ノード１-1～１１-3が共有する情報（共有情報）は、当該ノード１-1～１１-3のコントローラ１１-1～１１-3が有する不揮発性メモリの記憶領域、当該ノード１-1～１１-3のストレージ１３-1～１３-3の記憶領域、或いはストレージ１３-1～１３-3のうちの任意のストレージの記憶領域に格納される。コントローラ１１-1～１１-3は、この記憶領域を参照することにより共有情報を管理する。この管理のために、コントローラ１１-1～１１-3は上記記憶領域のアドレスを共有する。共有情報はＬＭＭＴ１１２ａの他に、論理ボリューム構成情報を含む。論理ボリューム構成情報は、図１のストレージクラスタがホストに対してどのような論理ボリュームを提供するかを示す。上述した各ノード１-1～１１-3が情報を共有する構成により、従来技術と比較して、当該各ノード１-1～１１-3が管理しなければならない情報量を減らすことができる。

【００６３】

次に、論理エクステントを構成する物理エクステントのデータを、ストレージをまたがって別の物理エクステントにマイグレート（移動）する場合の動作について、図１１を参照して説明する。

【００６４】

まず、例えば構成管理部によってデータの移動元及び移動先が決定される。ここでは、図１のストレージクラスタが図１０に示す状態にあるときに、データの移動元として、論理エクステントＬＥＩＤ＝４に対応付けられている、ノード１-1のストレージ１３-1内の物理ボリュームＰＶｏｌ＝１（１３０-1）／物理エクステントＰＥＩＤ＝２が、移動先として、ノード１-2のストレージ１３-2内の物理ボリュームＰＶｏｌ＝２（１３０-2）／物

理エクステントＰＥＩＤ＝１が、それぞれ決定されたものとする。

【００６５】

この場合、ノード１-1のストレージ１３-1内の物理ボリュームＰＶｏｌ＝１／物理エクステントＰＥＩＤ＝２に格納されているデータが、図１１おいて矢印Ｃ１で示されるように、ノード１-2のストレージ１３-2内の物理ボリュームＰＶｏｌ＝２／物理エクステントＰＥＩＤ＝１に移動される。

【００６６】

また、移動元のストレージ１３-1及び移動先のストレージ１３-2の各々のＬＰＭＴ１３２ａが書き換えられる。ここでは、ストレージ１３-1のＬＰＭＴ１３２ａ内のＰＥＩＤ＝２が保持されているエントリのＬＶｏｌ，ＬＥＩＤ，ＭＳがそれぞれ「１」→「ＮＵＬＬ」，「４」→「ＮＵＬＬ」，「１」→「ＮＵＬＬ」に書き換えられ、ストレージ１３-2のＬＰＭＴ１３２ａ内のＰＥＩＤ＝１が保持されているエントリのＬＶｏｌ，ＬＥＩＤ，ＭＳが、図１１おいて矢印Ｃ２で示されるように、それぞれ「ＮＵＬＬ」→「１」，「ＮＵＬＬ」→「４」，「ＮＵＬＬ」→「１」に書き換えられる。

【００６７】

また、論理エクステントＬＥＩＤ＝４を参照していた（つまり論理エクステントＬＥＩＤ＝４に関する所在管理サービスを担当している）所在管理サーバ１２-1のＬＳＭＴ１２０-1が書き換えられる。ここでは、ＬＳＭＴ１２０-1内のＬＥＩＤ＝４が保持されているエントリのＰＶｏｌが、図１１おいて矢印Ｃ３で示されるように、「１」→「２」に書き換えられる。

【００６８】

このように本実施形態において、論理エクステント（ＬＥＩＤ＝４）のマイグレーションに係わる通知（更新）は、その論理エクステント（ＬＥＩＤ＝４）を管理する所在管理サーバ（１２-1）に対して行われるだけで良い。つまり、上記の通知はストレージクラスタを構成する全てのノード１-1～１-3に対して行われる必要がない。このため、通知コストを小さくすることができる。

【００６９】

次に、論理エクステントを構成する物理エクステントのデータを、ストレージをまたがって別の物理エクステントにコピーする場合の動作について、図１２を参照して説明する。このようなコピー処理は、論理エクステントを構成する物理エクステントを冗長化し、耐障害性を高め、また並列処理によりリードアクセスを高速化するために行われる。

【００７０】

まず、例えば構成管理部によってコピー元及びコピー先が決定される。ここでは、図１のストレージクラスタが図１１に示す状態にあるときに、データのコピー元として、論理エクステントＬＥＩＤ＝４に対応付けられている、ノード１-2のストレージ１３-2内の物理ボリュームＰＶｏｌ＝２（１３０-2）／物理エクステントＰＥＩＤ＝１が、コピー先として、ノード１-3のストレージ１３-3内の物理ボリュームＰＶｏｌ＝１（１３０-3）／物理エクステントＰＥＩＤ＝２が、それぞれ決定されたものとする。

【００７１】

この場合、ノード１-2のストレージ１３-2内の物理ボリュームＰＶｏｌ＝２／物理エクステントＰＥＩＤ＝１に格納されている論理ボリュームＬＶｏｌ＝１／論理エクステントＬＥＩＤ＝４のデータが、図１２において矢印Ｄ１で示されるように、ノード１-3のストレージ１３-3内の物理ボリュームＰＶｏｌ＝１／物理エクステントＰＥＩＤ＝２にコピーされる。このコピーは、例えばストレージ１３-2及び１３-3のＩＯドライバ１３１によって行われる。

【００７２】

また、コピー先のストレージ１３-3のＬＰＭＴ１３２ａが書き換えられる。ここでは、ストレージ１３-2のＬＰＭＴ１３２ａ内のＰＥＩＤ＝１が保持されているエントリのＬＶｏｌ＝１，ＬＥＩＤ＝４，ＭＳ＝１が、図１２において矢印Ｄ２で示されるように、ストレージ１３-3のＬＰＭＴ１３２ａ内のＰＥＩＤ＝２が保持されているエントリのＬＶｏｌ

，ＬＥＩＤ，ＭＳにコピーされる。

【００７３】

また、論理エクステントＬＥＩＤ＝４を参照していた所在管理サーバ１２-1のＬＳＭＴ１２０-1が、例えばストレージ１３-1のＩＯドライバ１３１によって書き換えられる。ここでは、ノード１-1の所在管理サーバ１２-1が管理するＬＳＭＴ１２０-1内の、ＬＥＩＤ＝４が保持されているエントリのＰＶｏｌに、図１２において矢印Ｄ３で示されるように、「３」が追加される。このＬＳＭＴ１２０-1内のエントリの書き換えにより、論理エクステントＬＥＩＤ＝４が、ノード１-2の物理ボリュームＰＶｏｌ＝２に加えて、新たにノード１-3の物理ボリュームＰＶｏｌ＝３にも対応付けられる。つまり、論理エクステントＬＥＩＤ＝４を構成する物理エクステントが、ノード１-2の物理ボリュームＰＶｏｌ＝２及びノード１-3の物理ボリュームＰＶｏｌ＝３に冗長化される。

【００７４】

論理エクステントを構成する物理エクステントの冗長化構成後に、ホストからコントローラ１１-iに対して、当該論理エクステントへのライトアクセスの要求が発生したものとする。この場合、コントローラ１１-iは、上記論理エクステントに対応する複数の物理エクステントの全て対してライトを行う。したがって、論理エクステントＬＥＩＤ＝４へのライトアクセスが発生した場合であれば、ノード１-2の物理ボリュームＰＶｏｌ＝２／物理エクステントＰＥＩＤ＝１及びノード１-3の物理ボリュームＰＶｏｌ＝３／物理エクステントＰＥＩＤ＝２に対してデータライトが行われる。リードの場合はこの限りではない。つまり、物理ボリュームＰＶｏｌ＝２／物理エクステントＰＥＩＤ＝１または物理ボリュームＰＶｏｌ＝３／物理エクステントＰＥＩＤ＝２の一方からのデータリードが行われる。

【００７５】

次に、論理エクステントに関する所在管理サービスを、ノードをまたがってマイグレート（移動）する場合の動作について、図１３を参照して説明する。この所在管理サービスのマイグレーションは、当該所在管理サービスの負荷分散、或いはノードを構成するハードウェアの交換／更新などのために行われる。

【００７６】

論理エクステントに関する所在管理サービスのマイグレーションに際し、例えば構成管理部により移動元及び移動先が決定される。ここでは、図１のストレージクラスタが図１２に示す状態にあるときに、所在管理サービスの移動元として、論理ボリュームＬＶｏｌ＝１／論理エクステントＬＥＩＤ＝４に関するネームサービスを担当する、ノード１-1の所在管理サーバ１２-1が決定されたものとする。また、そのネームサービスの移動先として、ノード１-2の所在管理サーバ１２-2が決定されたものとする。

【００７７】

すると、移動元ノード１-1の論理ボリュームＬＶｏｌ＝１／論理エクステントＬＥＩＤ＝４に関するＬＳＭＴ１２０-1内エントリの情報が、図１３において矢印Ｅ１で示されるように、移動先ノード１-2のＬＳＭＴ１２０-2にコピーされる。

【００７８】

移動元ノード１-1の例えば所在管理サーバ１２-1または移動先ノード１-2の例えば所在管理サーバ１２-2は、論理ボリュームＬＶｏｌ＝１／論理エクステントＬＥＩＤ＝４に関する所在管理サービスがノード１-2に移動されたことを、全てのコントローラ１１-1～１１-3に通知する。

【００７９】

コントローラ１１-1～１１-3の各々は、論理ボリュームＬＶｏｌ＝１／論理エクステントＬＥＩＤ＝４に関するＬＭＭＴ１１２ａ内のエントリのＭＳを、図１３おいて矢印Ｅ２で示されるように、ノード１-1（の所在管理サーバ１２-1）のＩＤ（ＭＳ＝１）からノード１-2（の所在管理サーバ１２-2）のＩＤ（ＭＳ＝２）に書き換える（張り替える）。

【００８０】

また、論理ボリュームＬＶｏｌ＝１／論理エクステントＬＥＩＤ＝４を管理するノード

１-2及び１-3のストレージ１３-2及び１３-3は、自身のＬＰＭＴ１３２ａ中の論理ボリュームＬＶｏｌ＝１／論理エクステントＬＥＩＤ＝４に関する所在管理サービスの逆引きを、図１３おいて矢印Ｅ３で示されるように、ノード１-1（の所在管理サーバ１２-1）のＩＤ（ＭＳ＝１）からノード１-2（の所在管理サーバ１２-2）のＩＤ（ＭＳ＝２）に変更する。ストレージ１３-2及び１３-3は、この変更を所在管理サービス移動の通知元ノードに応答する。

【００８１】

ここで、全コントローラ１１-1～１１-3がＬＭＭＴ１１２ａを書き換えるには、Ａ）所在管理サーバ（ここでは所在管理サーバ１２-1または１２-2）によるブロードキャスト（一斉通知）による方法、Ｂ）コントローラのサーバ特定部（ここではコントローラ１１-1または１１-2のサーバ特定部１１２）によるブロードキャストによる方法、Ｃ）移動元からコントローラ１１-1～１１-3の各々に対する個々の更新通知による方法などが適用可能である。

【００８２】

ノード１-1の例えば所在管理サーバ１２-1は、全ノード１-1～１-3からの応答を確認して、当該所在管理サーバ１２-1の論理ボリュームＬＶｏｌ＝１／論理エクステントＬＥＩＤ＝４に関するＬＳＭＴ１２０-1内のエントリを削除する。これにより、移動元ノード１-1の論理ボリュームＬＶｏｌ＝１／論理エクステントＬＥＩＤ＝４に関するＬＳＭＴ１２０-1内エントリのデータが、移動先ノード１-2のＬＳＭＴ１２０-2に移動されたことになる。つまり、論理ボリュームＬＶｏｌ＝１／論理エクステントＬＥＩＤ＝４に関する移動元ノード１-1のＬＳＭＴ１２０-1及び移動先ノード１-2のＬＳＭＴ１２０-2内のエントリが、それぞれ書き換えられたことになる。

【００８３】

次に、論理エクステントに関する所在管理サービスを、ノードをまたがってコピーする場合の動作について、図１４を参照して説明する。

【００８４】

所在管理サービスの多重化がなされていない状態では、当該所在管理サービスを担当するあるノードが故障した場合、ホストからのアクセス要求が実行できなくなる。そこで、所在管理サービスをノードをまたがってコピーして当該所在管理サービスを多重化（冗長化）することにより、耐障害性向上及び負荷分散を図ることが必要となる。

【００８５】

所在管理サービスをノードをまたがってコピー（冗長化）するに際し、例えば構成管理部によりコピー元及びコピー先が決定される。ここでは、図１のストレージクラスタが図１３に示す状態にあるときに、所在管理サービスのコピー元として、論理ボリュームＬＶｏｌ＝１／論理エクステントＬＥＩＤ＝４に関するネームサービスを担当する、ノード１-2の所在管理サーバ１２-2が決定されたものとする。また、そのネームサービスのコピー先として、ノード１-3の所在管理サーバ１２-3が決定されたものとする。

【００８６】

すると、コピー元ノード１-2の論理ボリュームＬＶｏｌ＝１／論理エクステントＬＥＩＤ＝４に関するＬＳＭＴ１２０-2内エントリの情報が、図１４において矢印Ｆ１で示されるように、コピー先ノード１-3のＬＳＭＴ１２０-3にコピーされる。

【００８７】

コピー元ノード１-2の例えば所在管理サーバ１２-2またはコピー先ノード１-3の例えば所在管理サーバ１２-3は、論理ボリュームＬＶｏｌ＝１／論理エクステントＬＥＩＤ＝４に関する所在管理サービスがノード１-2からノード１-3にも機能分担されたことを、全てのコントローラ１１-1～１１-3に通知する。

【００８８】

コントローラ１１-1～１１-3の各々は、論理ボリュームＬＶｏｌ＝１／論理エクステントＬＥＩＤ＝４に関するＬＭＭＴ１１２ａ内のエントリのＭＳ、つまりノード１-2（の所在管理サーバ１２-2）のＩＤ（ＭＳ＝２）が設定されているＭＳに、図１４おいて矢印Ｆ

２で示されるように、ノード１-3（の所在管理サーバ１２-3）のＩＤ（ＭＳ＝３）を加える。これにより、論理ボリュームＬＶｏｌ＝１／論理エクステントＬＥＩＤ＝４に関するＬＭＭＴ１１２ａ内のエントリが、ノード１-2だけでなくノード１-3にも向けられる。

【００８９】

また、論理ボリュームＬＶｏｌ＝１／論理エクステントＬＥＩＤ＝４を新たに管理するノード１-3のストレージ１３-3は、自身のＬＰＭＴ１３２ａ中の論理ボリュームＬＶｏｌ＝１／論理エクステントＬＥＩＤ＝４に関する所在管理サービスの逆引き（ＭＳ）を、図１４において矢印Ｆ３で示されるように、ノード１-2（の所在管理サーバ１２-2）だけでなく、ノード１-3（の所在管理サーバ１２-3）にも向ける。ストレージ１３-3は、この変更を所在管理サービスコピーの通知元ノードに応答する。全コントローラ１１-1～１１-3がＬＭＭＴ１１２ａを書き換えるには、前述した所在管理サービスの移動の場合と同様の方法Ａ），Ｂ）またはＣ）が適用可能である。

【００９０】

以上、複数のノード（ストレージ）にまたがるエクステント移動／コピー及び所在管理サービス移動／コピーについて説明した。しかし、エクステント移動／コピーは１つのノード（ストレージ）の物理ボリューム内でも行うことができる。この場合、そのノード（ストレージ）のＬＰＭＴ１３２ａを更新するだけで良い。また、例えば所在管理サーバ１２-iのＬＳＭＴ１２０-i及びストレージ１３-iのＬＰＭＴ１３２ａを冗長化のためにコピーすることも可能である。

【００９１】

次に、図１のストレージクラスタに、新たにノードを参加させる（追加する）場合の動作について説明する。この場合、既にストレージクラスタを構成しているノード１-1～１-1のいずれかから、当該ストレージクラスタに新たに追加されるノード（新たなノード）に対して、少なくともＬＭＭＴ１１２ａをコピーする。これにより、新たなノードはＩＯを実行することができる。

【００９２】

新たなノードにノード１-1～１-1のいずれかの所在管理サーバの処理（所在管理サービス）を分担させるには、例えば構成管理部と連携して、前述の所在管理サービスをマイグレートする処理を行えば良い。ノードの参加で新たに増えたストレージ内の物理ボリューム／物理エクステントも、構成管理部と連携して、適宜分散して利用することができる。

【００９３】

上述したように本実施形態（第１の実施形態）においては、複数のノード（実施形態では３つのノード）を用いてストレージクラスタを構成したとき、各々のノードが管理しなければならない情報量を減らすことができる。また本実施形態においては、あるノード（のストレージ）の物理ボリューム内で、或いは複数のノード（のストレージ）の物理ボリュームをまたがって、論理エクステントに対応する物理エクステントを移動／コピーする場合に、ストレージクラスタに含まれる全てのコントローラが情報を書き換える必要はないため（所在管理サーバの情報を書き換えるだけで良いため）、コントローラ間での同期更新処理が必要なく、エクステント移動／コピー時の処理を簡単にし、総じて性能を向上させることができる。

【００９４】

［第２の実施形態］

前記第１の実施形態では、ストレージクラスタを構成する各ノード１-1～１-3は、それぞれコントローラ１１-1～１-3、所在管理サーバ１２-1～１２-3及びストレージ１３-1～１３-3を備えている。しかし、各ノードが、コントローラ、所在管理サーバ及びストレージの全てを必ずしも備える必要はない。そこで、ストレージクラスタを構成するノードの幾つかが、コントローラ、所在管理サーバ及びストレージの１つまたは２つのみを備える本発明の第２の実施形態について説明する。

【００９５】

図１５は本発明の第２の実施形態に係るストレージクラスタ構成のストレージシステム

の構成を示すブロック図である。図１５において、図７と同様の要素には便宜的に同一参照符号を付してある。図１５のストレージシステム（ストレージクラスタ）は、前記第１の実施形態と同様に、３つのノード１-1～１-3から構成される。図１５のストレージクラスタが第１の実施形態と異なるのは、ノード１-1～１-3のうちノード１-2及び１-3がコントローラ、所在管理サーバ及びストレージの１つまたは２つのみを備える点にある。ここでは、ノード１-1は、第１の実施形態と同様に、コントローラ１１-1、所在管理サーバ１２-1及びストレージ１３-1を備える。これに対し、ノード１-2はコントローラ１１-2のみを備え、ノード１-3はコントローラを除く所在管理サーバ１２-3及びストレージ１３-3を備えている。この図１５に示すストレージクラスタの動作自体は、第１の実施形態と同様である。

【００９６】

本実施形態では、ストレージクラスタが３つのノードから構成されているものの、コントローラ、所在管理サーバ及びストレージの数が２つである場合を想定している。しかし、所在管理サーバ及びストレージの数はストレージクラスタ全体で１つであっても構わない。つまり複数のノードから構成されるストレージクラスタにおいて、コントローラの数は少なくとも２つであれば良く、所在管理サーバ及びストレージの数は少なくとも１つであれば良い。この制約のもとで、コントローラ、所在管理サーバ及びストレージの数は適宜増減可能である。つまり本実施形態においては、ストレージクラスタを構成するとき、状況に応じて、必要な機能のみを拡張することができる。

【００９７】

［第３の実施形態］

次に、本発明の第３の実施形態について説明する。この第３の実施形態の特徴は、実際にホストからある論理エクステントへの初回のアクセスが要求されるまで、ストレージ側で当該論理エクステントに対応する物理エクステントの割り付けを行わず、ディスクの有効利用を図ることにある。

【００９８】

図１６は、本発明の第３の実施形態に係るストレージクラスタ構成のストレージシステムの構成を示すブロック図である。図１６において、図７と同様の要素には便宜的に同一参照符号を付してある。図１６のストレージシステム（ストレージクラスタ）は、前記第１の実施形態と同様に、３つのノード１-1～１-3から構成される。図１６のストレージクラスタが第１の実施形態と異なるのは、ホストから論理エクステントへの初回のアクセス、例えばライトアクセスが要求されるまで、ストレージ１３-i側で当該論理エクステントに対応する物理エクステントの割り付けを行わない点にある。

【００９９】

図１６の例では、論理ボリュームＬＶｏｌ＝１／論理エクステントＬＥＩＤ＝１及び論理ボリュームＬＶｏｌ＝１／論理エクステントＬＥＩＤ＝４に対して、ノード１-1における物理ボリューム１３０-1の物理エクステントの割り付けが行われていない状態が、ストレージ１３-1のＬＰＭＴ１３２ａによって示されている。同様に、論理ボリュームＬＶｏｌ＝１／論理エクステントＬＥＩＤ＝２に対して、ノード１-2内の物理ボリューム１３０-2の物理エクステントの割り付けが行われていない状態が、ストレージ１３-2のＬＰＭＴ１３２ａによって示されている。また、論理ボリュームＬＶｏｌ＝１／論理エクステントＬＥＩＤ＝３に対して、ノード１-3内の物理ボリューム１３０-3の物理エクステントの割り付けが行われていない状態が、ストレージ１３-3のＬＰＭＴ１３２ａによって示されている。

【０１００】

ここでは、未割り付けの状態は、ＬＰＭＴ１３２ａ内で列項目ＰＥＩＤに“ＮＡ(Not Assignment）”が設定されているエントリにより示される。但し、ＬＰＭＴ１３２ａ内で列項目ＰＥＩＤに“ＮＡ”が設定されているエントリの他の列項目ＬＶｏｌ，ＬＥＩＤ及びＭＳには、それぞれ論理ボリューム、論理エクステント及びノードを指定する情報が設定されている。このことは、論理ボリューム／論理エクステントに対するストレージ（物

理ボリューム）の割り付けは行われていることを示す。換言するならば、論理ボリューム／論理エクステントに対するＬＰＭＴ１３２ａ内のエントリ（ＬＶｏｌ，ＬＥＩＤ）で指定される論理ボリューム／論理エクステントに必要な領域が、当該エントリの列項目ＭＳで指定されるノード上で予約されていることを示す。なお、図１６では、ＬＰＭＴ１３２ａ内のエントリにおける各列項目の並びが便宜的に変更されている点に注意されたい。

【０１０１】

ＬＰＭＴ１３２ａ内のエントリ（ＬＶｏｌ，ＬＥＩＤ）で指定される論理ボリューム／論理エクステントに対して物理エクステントの割り付けが行われると、当該エントリ中の列項目ＰＥＩＤが、“ＮＡ”から当該割り付けられた物理エクステントのＩＤで書き換えられる。

【０１０２】

この図１６に示すストレージクラスタの基本的な動作は、第１の実施形態と同様である。但し、論理ボリューム作成時と、論理エクステントへの初回のアクセス（ライトアクセス）時の挙動は、第１の実施形態と異なる。

【０１０３】

そこで、論理ボリューム作成時において第１の実施形態と異なる点について説明する。第１の実施形態では、前述した（ａ），（ｂ），（ｃ）の手順で論理ボリュームが作成される。これに対して第３の実施形態では、論理エクステントへの初回のライトアクセス時に当該論理エクステントに対する物理エクステントの割り付けが行われるため、上記（ａ）の動作が次の（ａ１）のようになる。

【０１０４】

（ａ１）構成管理部は、作成されるべき論理ボリュームの容量を確保するために、当該論理ボリュームを構成する論理エクステントの各々を、どのノードＩＤ（ＰＶｏｌ）のストレージ（物理ボリューム）に格納するか決定する。但し、上記（ａ）とは異なり、各論理エクステントへの物理エクステントの割り付けは行われない。構成管理部は、ストレージ１３-1～１３-3毎に、各論理エクステントに対して物理エクステントが未割り付け状態にあることを示すＬＰＭＴ１３２ａを生成して、そのストレージに保持する。

【０１０５】

次に、図１６のストレージクラスタにおける、ホストから論理エクステントへのライトアクセス時の動作について、図１乃至図３を援用しながら、図１７のフローチャートを参照して説明する。図１７において、図４のフローチャートと同様の部分には同一符号を付してある。

【０１０６】

図１７のフローチャートが、図４のフローチャートと異なるのは、ある論理エクステントへの初回のアクセスがライトアクセスの場合に、当該論理エクステントに物理エクステントが割り付けられていないため、割り付け先の物理エクステント先を決定する処理（ステップＳ１８）が追加されている点である。

【０１０７】

今、第１の実施形態と同様に、ホストからノード１-1のコントローラ１１-1に含まれているターゲットインタフェイス１１１に対してアクセス（ライトアクセス）が要求されたものとする。この場合、図４のフローチャートと同様に、まずステップＳ０～Ｓ７が実行される。但し図１７では、ステップＳ０～Ｓ７のうちのステップＳ１～Ｓ６が省略されている。ステップＳ７では、ノードＩＤ＝ＳＩＤｘで示されるノードのストレージ、例えばノード１-1のストレージ１３-1に対し、論理エクステント（ＬＶｏｌｘ，ＬＥＩＤｘ）に対応する物理エクステントへのアクセスが要求される。するとストレージ１３-1では、ステップＳ１８が次のように実行される。

【０１０８】

まず、ストレージ１３-1のＩＯドライバ１３１は、論理エクステント（ＬＶｏｌｘ，ＬＥＩＤｘ）に対応付けられているＬＰＭＴ１３２ａ内のエントリの列項目ＰＥＩＤを参照する（ステップＳ１８ａ）。ストレージ１３-1は、このＰＥＩＤが非ＮＡであるか否かに

より、論理エクステント（ＬＶｏｌｘ，ＬＥＩＤｘ）に物理エクステントが割り付けられているか否かを判定する（ステップＳ１８ｂ）。

【０１０９】

もし、論理エクステント（ＬＶｏｌｘ，ＬＥＩＤｘ）に物理エクステントが割り付けられていないならば、ＩＯドライバ１３１は構成管理部と連携して、論理エクステント（ＬＶｏｌｘ，ＬＥＩＤｘ）にストレージ１３-1（物理ボリューム１３０-1）内の適当な物理エクステントＰＥＩＤｘを割り付ける（ステップＳ１８ｃ）。このステップＳ１８ｃにおいて、ＩＯドライバ１３１は、論理エクステント（ＬＶｏｌｘ，ＬＥＩＤｘ）に対応付けられているＬＰＭＴ１３２ａ内のエントリの列項目ＰＥＩＤを、“ＮＡ”から物理エクステントＰＥＩＤｘのＩＤ（＝ＰＥＩＤｘ）に更新する。そしてＩＯドライバ１３１は、物理エクステントＰＥＩＤｘに対してアクセスし、コントローラ１１-1のＩＯマネージャ１１３に対して応答する（ステップＳ１８ｄ）。

【０１１０】

一方、論理エクステント（ＬＶｏｌｘ，ＬＥＩＤｘ）に物理エクステント（例えば物理エクステントＰＥＩＤｘ）が既に割り付けられているならば（ステップＳ１８ｂ）、ＩＯドライバ１３１はステップＳ１８ｃをスキップしてステップＳ１８ｄを実行する。

【０１１１】

このように第３の実施形態においては、ある論理エクステントに対する初回のアクセスがライトアクセスの場合に、当該論理エクステントに物理エクステントが割り付けられる。これに対し、ある論理エクステントに対する初回のアクセスがリードアクセスの場合、つまりデータがライトされていない論理エクステントに対してリードアクセスが発生した場合には、当該論理エクステントに物理エクステントを割り付けることも、割り付けないことも可能である。割り付ける場合は、ＩＯドライバ１３１によって上記ステップＳ１８が実行されれば良い。いずれの場合にも、リード要求に対する応答として、全てゼロなどの適切なデータが返されるようにすれば良い。なお、本実施形態（第３の実施形態）で適用される物理エクステントの割り付け方法は、第２の実施形態にも同様に適用可能である。本実施形態においては、各ストレージにおいて、ホストからアクセスが要求された際の状態に応じて、最適な物理エクステントの割り付けが可能になる。

【０１１２】

［第４の実施形態］

次に本発明の第４の実施形態について説明する。この第４の実施形態の特徴は、実際にホストからある論理エクステントへの初回のアクセスが要求されるまで、当該論理エクステントの領域の予約が行われない点にある。

【０１１３】

図１８は、本発明の第４の実施形態に係るストレージクラスタ構成のストレージシステムの構成を示すブロック図である。図１８において、図７と同様の要素には便宜的に同一参照符号を付してある。図１８のストレージシステム（ストレージクラスタ）は、前記第１の実施形態と同様に、３つのノード１-1～１-3から構成される。図１８のストレージクラスタが第１の実施形態と異なるのは、実際にホストからある論理エクステントへの初回のアクセスが要求されるまで、当該論理エクステントへの物理エクステントの割り付けだけでなく、当該論理エクステントへのストレージ（物理ボリューム）の割り付け（予約）も行われない点にある。

【０１１４】

このように、図１８のストレージクラスタでは、ある論理エクステントへの初回のアクセスが要求されるまで、当該論理エクステントの領域の予約が行われないため、論理ボリューム作成時には、ストレージ１３-1～１３-3のＬＰＭＴ１３２ａのエントリを予約する必要もない。論理エクステントに対するストレージ（物理ボリューム）及び物理エクステントの割り付けは、コントローラ１１-i（ｉ＝１，２，３）からの所在管理サーバ１２-jに対する問い合わせに応じて当該所在管理サーバ１２-j（ｊ＝１，２，３）によって行われる。この物理エクステントの割り付け計画は、例えば構成管理部により行われる。

【０１１５】
図１８の例では、論理ボリュームを構成する各論理エクステントに対するストレージ（物理ボリューム）の割り付けが未だ行われていない状態が示されている。つまり図１８の例では、論理ボリュームＬＶｏｌ＝１／論理エクステントＬＥＩＤ＝１及び論理ボリュームＬＶｏｌ＝１／論理エクステントＬＥＩＤ＝４に対して、ストレージ（物理ボリューム）の割り付けが行われていない状態が、所在管理サーバ１２-1のＬＳＭＴ１２０-1によって示されている。同様に、論理ボリュームＬＶｏｌ＝１／論理エクステントＬＥＩＤ＝２に対して、ストレージ（物理ボリューム）の割り付けが行われていない状態が、所在管理サーバ１２-2のＬＳＭＴ１２０-2によって示されている。また、論理ボリュームＬＶｏｌ＝１／論理エクステントＬＥＩＤ＝３に対して、ストレージ（物理ボリューム）の割り付けが行われていない状態が、所在管理サーバ１２-3のＬＳＭＴ１２０-3によって示されている。ここでは、ストレージ（物理ボリューム）の未割り付けの状態は、ＬＳＭＴ１２０-1～１２０-3内で列項目ＰＶｏｌに“ＮＡ”が設定されているエントリにより示される。図１８において、ＬＳＭＴ１２０-1～１２０-3のエントリから、ストレージ１３-1～１３-3のＬＰＭＴ１３２ａのエントリに向かう矢印が描かれていないことに注意されたい。
【０１１６】
この図１８に示すストレージクラスタの基本的な動作は、第１の実施形態と同様である。但し、論理ボリューム作成時と、論理エクステントへの初回のアクセス時の挙動は、第１の実施形態と異なる。
【０１１７】
そこで、論理ボリューム作成時において第１の実施形態と異なる点について説明する。第１の実施形態では、前述した（ａ），（ｂ），（ｃ）の手順で論理ボリュームが作成される。これに対して第４の実施形態では、論理エクステントへの初回アクセス時に当該論理エクステントを格納するストレージ（物理ボリューム）が決定され、この決定されたストレージ（物理ボリューム）内で当該論理エクステントに対して物理エクステントが割り付けられる。このため、上記（ａ），（ｂ）の動作が次の（ａ２），（ｂ２）のようになる。
【０１１８】
（ａ２）構成管理部は、作成されるべき論理ボリュームを構成する論理エクステントの各々を、どのノードＩＤ（ＰＶｏｌ）のストレージ（物理ボリューム）に格納するか決定しない。このため構成管理部は、ストレージ１３-1～１３-3毎に、図１８に示されるように空の状態のＬＰＭＴ１３２ａを生成して、そのストレージに保持する。
【０１１９】
（ｂ２）構成管理部は、どのノードの所在管理サーバに論理エクステントの所在管理サービスを担当させるかを決定する。但し、その論理エクステントをどのストレージ（物理ボリューム）に格納するかは決定しない。このため構成管理部は、所在管理サーバ１２-1～１２-3毎に、その所在管理サーバが担当する論理エクステントについてのエントリ情報であって、列項目ＰＶ０ｌが“ＮＡ”のエントリの情報を含むＬＳＭＴ１２０-1～１２０-3を生成して、その所在管理サーバ１２-1～１２-3内に設定する。なお、第２の実施形態と同様に、ストレージクラスタ内の全てのノード１-1～１-3に所在管理サーバ（ＬＳＭＴ）が設けられる必要はない。
【０１２０】
次に、図１８のストレージクラスタにおける、ホストから論理エクステントへのライトアクセス時の動作について、図１乃至図３を援用しながら、図１９のフローチャートを参照して説明する。図１９において、図４または図１７のフローチャートと同様の部分には同一符号を付してある。
【０１２１】
今、第１の実施形態と同様に、ホストからノード１-1のコントローラ１１-1に含まれているターゲットインタフェイス１１１に対してアクセスが要求されたものとする。この場合、図４のフローチャートと同様に、まずステップＳ０～Ｓ４が実行される。図１９では

、ステップＳ０～Ｓ４のうちのステップＳ１～Ｓ３が省略されている。ステップＳ４では、例えばコントローラ１１-1のサーバ特定部１１２が例えば所在管理サーバ１２-1に対して、論理エクステント（ターゲット論理エクステント）（ＬＶｏｌｘ，ＬＥＩＤｘ）に対応する物理エクステントを格納しているストレージのノードＩＤ＝ＳＩＤｘを問い合わせる。すると所在管理サーバ１２-1では、ステップＳ２５が次のように実行される。

【０１２２】

まず、所在管理サーバ１２-1は、どのノードＩＤ＝ＳＩＤｘのストレージ（物理ボリューム）が、論理エクステント（ＬＶｏｌｘ，ＬＥＩＤｘ）に対応する物理エクステントを格納しているかを調べるために、当該論理エクステント（ＬＶｏｌｘ，ＬＥＩＤｘ）に対応付けられているＬＳＭＴ１２０-1内のエントリの列項目ＰＶｏｌ（＝ＳＩＤｘ）を参照する（ステップＳ２５ａ）。

【０１２３】

所在管理サーバ１２-1は、このＰＶｏｌが非ＮＡであるか否かにより、論理エクステント（ＬＶｏｌｘ，ＬＥＩＤｘ）にストレージ（物理ボリューム）が割り付けられているか否かを判定する（ステップＳ２５ｂ）。

【０１２４】

もし、論理エクステント（ＬＶｏｌｘ，ＬＥＩＤｘ）にストレージ（物理ボリューム）が割り付けられていないならば、所在管理サーバ１２-1は、ストレージ１３-1～１３-3（の物理ボリューム１３０-1～１３０-3）における物理エクステントの空きを確認する（ステップＳ２５ｃ）。この確認処理は、例えば所在管理サーバ１２-1がストレージ１３-1～１３-3のＩＯドライバ１３１と通信を行うことにより実現される。

【０１２５】

もし、ストレージ１３-1～１３-3のいずれにも物理エクステントの空きがない（または要求に応えられないない）場合には、ＩＯエラーとなる（ステップＳ２５ｄ）。この場合、ホストからのアクセス要求に対する応答として、エラーが返される（ステップＳ９）。

【０１２６】

一方、ストレージ１３-1～１３-3のいずれかに物理エクステントの空きがあるならば、所在管理サーバ１２-1は物理エクステントに空きのあるストレージ（例えばストレージ１３-1）を選択して、論理エクステント（ＬＶｏｌｘ，ＬＥＩＤｘ）に当該ストレージを割り付ける（ステップＳ２５ｅ）。このステップＳ２５ｅにおいて、所在管理サーバ１２-1は、論理エクステント（ＬＶｏｌｘ，ＬＥＩＤｘ）へのストレージの割り付けがＬＳＭＴ１２０-内の対応するエントリに反映されるように、当該エントリを更新する。ここでは、論理エクステント（ＬＶｏｌｘ，ＬＥＩＤｘ）に対応付けられているＬＳＭＴ１２０-1内のエントリの列項目ＰＶｏｌが、“ＮＡ”から当該論理エクステント（ＬＶｏｌｘ，ＬＥＩＤｘ）に割り付けられるストレージ１３-1のノードＩＤ（物理ボリューム１３０-1のノードＩＤ＝ＰＶｏｌ＝１）に更新される。

【０１２７】

ステップＳ２５ｅにおいてＬＳＭＴ１２０-1が更新されると、第１の実施形態で図４のフローチャートのステップＳ４が実行された場合と同様に、ステップＳ５，Ｓ６が実行される。これにより、所在管理サーバ１２-1からコントローラ１１-1に対し、論理エクステント（ＬＶｏｌｘ，ＬＥＩＤｘ）に割り付けられるストレージ１３-1のノードＩＤ（物理ボリューム１３０-1のノードＩＤ＝ＰＶｏｌ）＝ＳＩＤｘ＝１が問い合わせ元のサーバ特定部１１２を介して通知される。なお、、論理エクステント（ＬＶｏｌｘ，ＬＥＩＤｘ）にストレージ（物理ボリューム）が割り付けられている場合には（ステップＳ２５ｂ）、直ちにステップＳ５，Ｓ６が実行される。

【０１２８】

ターゲットインタフェイス１１１は、サーバ特定部１１２を介して通知されたノードＩＤ＝ＳＩＤｘ＝１を受け取ると、第１及び第３の実施形態と同様に、ノードＩＤ＝ＳＩＤｘ＝１で示されるノード１-1のストレージ１３-1に対し、論理エクステント（ＬＶｏｌｘ，ＬＥＩＤｘ）に対応する物理エクステントへのアクセスをＩＯマネージャ１１３を介し

て要求する（ステップＳ７）。このステップＳ７以降の動作は、第３の実施形態と同様であり、ステップＳ１８ａ～Ｓ１８ｄからなるステップＳ１８がストレージ１３-1において実行される。これにより、論理エクステント（ＬＶｏｌｘ，ＬＥＩＤｘ）に物理エクステントが割り付けられていない場合には、物理ボリューム１３０-1内の空き物理エクステントが当該論理エクステント（ＬＶｏｌｘ，ＬＥＩＤｘ）に割り付けられる。ここでは、そして、物理エクステントＰＥＩＤｘが割り付けられたものとする。この場合、論理エクステント（ＬＶｏｌｘ，ＬＥＩＤｘ）に対応付けられているＬＰＭＴ１３２ａ内のエントリの列項目ＰＥＩＤを、“ＮＡ”から物理エクステントＰＥＩＤｘのＩＤ（＝ＰＥＩＤｘ）に更新される。

【０１２９】

本実施形態においては、各ストレージにおいて、ホストからアクセスが要求された際の状態に応じて、最適な物理エクステントの割り付けが可能になるだけでなく、総物理ボリューム容量の上限値を超える容量の論理ボリュームを作成することができる。

【０１３０】

［第５の実施形態］

次に本発明の第５の実施形態について説明する。この第５の実施形態の特徴は、コントローラ１１-i（ｉ＝１，２，３）のサーバ特定部１１２から所在管理サーバ１２-ｊ（ｊ＝１，２，３）への問い合わせと当該問い合わせに対する回答（結果）とに関する情報をキャッシュ１１２ｂに保持することにある。この情報は、論理ボリュームＬＶｏｌｘ／論理エクステントＬＥＩＤｘ、つまり論理エクステント（ＬＶｏｌｘ，ＬＥＩＤｘ）を示す情報（ＬＶｏｌｘ，ＬＥＩＤｘ）と、当該論理エクステント（ＬＶｏｌｘ，ＬＥＩＤｘ）に対応している物理エクステントを格納しているストレージのノードＩＤ＝ＳＩＤｘとを含む。このキャッシュ１１２ｂを利用することで、再度の問い合わせが必要となった場合に、当該問い合わせに対する回答に相当する情報をキャッシュ１１２ｂから速やかに取得することが可能となる。

【０１３１】

次に、第５の実施形態における、ホストから論理エクステントへのアクセス時の動作について、図１乃至図３及び図７を援用しながら、図２０のフローチャートを参照して説明する。

【０１３２】

今、第１の実施形態と同様に、ホストからノード１-1のコントローラ１１-1に含まれているターゲットインタフェイス１１１に対してアクセスが要求されたものとする。この場合、図４のフローチャートと同様に、まずステップＳ０～Ｓ２が実行される。図２０では、ステップＳ０～Ｓ２のうちのステップＳ０，Ｓ１が省略されている。ステップＳ２では、コントローラ１１-1のターゲットインタフェイス１１１からコントローラ１１-1のサーバ特定部１１２に対して、論理エクステント（ＬＶｏｌｘ，ＬＥＩＤｘ）が、どのノードのストレージに格納されているかが問い合わせられる。

【０１３３】

するとサーバ特定部１１２は、第１の実施形態とは異なって、自身が管理しているキャッシュ１１２ｂを参照して、論理エクステント（ＬＶｏｌｘ，ＬＥＩＤｘ）に対応している物理エクステントを格納しているストレージのノードＩＤ＝ＳＩＤｘを調べる（ステップＳ３１）。

【０１３４】

もし、目的のノードＩＤ＝ＳＩＤｘが（ＬＶｏｌｘ，ＬＥＩＤｘ）と対応付けてキャッシュ１１２ｂに格納されていないならば、つまりミスヒットならば、サーバ特定部１１２は、第１の実施形態におけるターゲットインタフェイス１１１からの問い合わせ（ステップＳ２）に対するのと同様に、ステップＳ３及びＳ４を実行する。

【０１３５】

本実施形態では、論理エクステント（ＬＶｏｌｘ，ＬＥＩＤｘ）に対応している物理エクステントが格納されているストレージの所在が、第１の実施形態と同様にノード１-1の

所在管理サーバ１２-1によって管理され、当該ストレージのノードＩＤ（＝ＰＶｏｌ）がＳＩＤｘ＝１であるものとする。このノードＩＤはＬＳＭＴ１２０-1に基づいて求められる。この場合、所在管理サーバ１２-1は、論理エクステント（ＬＶｏｌｘ，ＬＥＩＤｘ）に対応している物理エクステントが格納されているストレージのノードＩＤ＝ＳＩＤｘと、ＬＳＭＴ１２０-1のシリアル番号（ＬＳＭＴシリアル番号）ＳＮとを、問い合わせ元のサーバ特定部１１２に対して通知する（ステップＳ５ａ）。第１の実施形態のステップＳ５との違いは、ノードＩＤに加えて、当該ノードＩＤを求めるのに用いられたＬＳＭＴ１２０-1のシリアル番号ＳＮが通知される点である。このシリアル番号ＳＮは、ＬＳＭＴ１２０-1が更新される毎に、所在管理サーバ１２-1によって例えばインクリメントされる。このことは、ＬＳＭＴ１２０-2，１２０-3についても同様である。

【０１３６】

サーバ特定部１１２は、所在管理サーバ１２-1からの通知を受けて、キャッシュ１１２ｂに、｛（ＬＶｏｌｘ，ＬＥＩＤｘ），ＳＩＤｘ，ＳＮ｝を含む情報（つまり、問い合わせと回答とに関する情報）を格納する（ステップＳＳ５ｂ）。これにより、サーバ特定部１１２は、論理エクステント（ＬＶｏｌｘ，ＬＥＩＤｘ）がどのノードのストレージに格納されているかをターゲットインタフェイス１１１によって再び問い合わせられた場合に、当該問い合わせに対する回答に相当する情報を上記ステップＳ３１の処理で速やかに取得することができる。つまり、サーバ特定部１１２から所在管理サーバへの問い合わせの回数を減らして、アクセス速度を向上させることができる。

【０１３７】

但し、所在管理サーバ１２-1～１２-3上のＬＳＭＴ１２０-1～１２０-3が変更された場合、その変更がキャッシュ１１２ｂに反映される必要がある。その理由は、例えば第１の実施形態のように、ストレージ間で物理エクステントのコピー（冗長構成）を管理する場合（図１２参照）、ＬＳＭＴ１２０-1～１２０-3の更新がキャッシュ１１２ｂに反映されていないと、ライトアクセスによって一方の物理エクステントのみが書き換えられてしまい、冗長構成が崩れてしまうからである。そのため、本実施形態では、以下のようなキャッシュ更新の仕組みが用いられる。

【０１３８】

まず、ＬＳＭＴ１２０-1～１２０-3の各々にシリアル番号ＳＮが付与される。このシリアル番号ＳＮは、ストレージ１３-1～１３-3のＩＯドライバ１３１によってＬＳＭＴ１２０-1～１２０-3が更新される都度、例えば当該ＩＯドライバ１３１によってインクリメント（更新）される。

【０１３９】

また、ステップＳ５ａにおいてノードＩＤ＝ＳＩＤｘを求めるのに用いられたＬＳＭＴ１２０-j（ｊ＝１，２，３）のシリアル番号ＳＮが、（ＬＶｏｌｘ，ＬＥＩＤｘ），ＳＩＤｘと組をなしてキャッシュ１１２ｂに格納される（ステップＳ５ｂ）。

【０１４０】

コントローラ１１-i（ｉ＝１，２，３）は、ストレージ１３-jへのアクセスを行うときに、このシリアル番号ＳＮを受け渡す。ストレージ１３-jのＩＯドライバ１３１は、このシリアル番号ＳＮと自身が認識しているＬＳＭＴ１２０-jの最新のシリアル番号ＳＮとを比較する。ＩＯドライバ１３１は、両シリアル番号ＳＮが一致していないならば、コントローラ１１-iのサーバ特定部１１２が管理するキャッシュ１１２ｂには、ＬＳＭＴ１２０-jの最新の更新内容が反映されていないと判定する。この場合、ストレージ１３-jのＩＯドライバ１３１はコントローラ１１-iに対してキャッシュ１１２ｂの該当する情報の無効化を通知する。つまりＩＯドライバ１３１は、キャッシュ無効化手段として機能する。コントローラ１１-iのサーバ特定部１１２は、この通知に応じて、ステップＳ３２でミスヒットが判定された場合と同様に、所在管理サーバ１２-jに対して再度の問い合わせを行う（ステップＳ３，Ｓ４）。これによりサーバ特定部１１２は、所在管理サーバ１２-jからＬＳＭＴ１２０-jの最新の情報を取得して、キャッシュ１１２ｂに反映させることができる。

【０１４１】
なお、以下に述べるように、所在管理サーバ１２-jがキャッシュ無効化手段として機能しても良い。つまり、所在管理サーバ１２-jがコントローラ１１-jのサーバ特定部１１２上のキャッシュ１１２ｂを監視して、自身が管理するＬＳＭＴ１２０-jとの不整合が生じているかを判別する構成とすることも可能である。もし不整合を判別した場合、所在管理サーバ１２-jからコントローラ１１-jのサーバ特定部１１２に、キャッシュ１１２ｂの該当するエントリ情報の破棄・更新が必要なことを通知すれば良い。
【０１４２】
なお、本発明は、上記各実施形態そのままに限定されるものではなく、実施段階ではその要旨を逸脱しない範囲で構成要素を変形して具体化できる。また、上記各実施形態に開示されている複数の構成要素の適宜な組み合せにより種々の発明を形成できる。例えば、各実施形態に示される全構成要素から幾つかの構成要素を削除してもよい。
【図面の簡単な説明】
【０１４３】
【図１】本発明の第１の実施形態に係るストレージクラスタ構成のストレージシステムの概略構成を示すブロック図。
【図２】図１に示されるコントローラの構成を示すブロック図。
【図３】図１に示されるストレージの構成を示すブロック図。
【図４】同第１の実施形態においてホストからアクセス要求が発行された場合の処理の手順を示すフローチャート。
【図５】同第１の実施形態においてホストからアクセス要求が発行された場合の情報の流れを示す図。
【図６】同第１の実施形態においてホストからアクセス要求が発行された場合の情報の流れの変形例を示す図。
【図７】図１のシステムにおいて、論理ボリュームＬＶｏｌがＬＶｏｌ＝１であって、当該論理ボリュームＬＶｏｌ＝１を、ＬＥＩＤがＬＥＩＤ＝１～ＬＥＩＤ＝４の４つの論理エクステントから構成した場合のシステム構成を示す図。
【図８】図７の状態におけるＬＭＭＴのデータ構造例を示す図。
【図９】図７の状態におけるＬＳＭＴのデータ構造例を示す図。
【図１０】図７の状態におけるＬＰＭＴのデータ構造例を示す図。
【図１１】同第１の実施形態において、論理エクステントを構成する物理エクステントのデータを、ストレージをまたがって別の物理エクステントにマイグレートする場合の動作を説明するための図。
【図１２】同第１の実施形態において、論理エクステントを構成する物理エクステントのデータを、ストレージをまたがって別の物理エクステントにコピーする場合の動作を説明するための図。
【図１３】同第１の実施形態において、論理エクステントに関する所在管理サービスを、ノードをまたがってマイグレートする場合の動作を説明するための図。
【図１４】同第１の実施形態において、論理エクステントに関する所在管理サービスを、ノードをまたがってコピーする場合の動作を説明するための図。
【図１５】本発明の第２の実施形態に係るストレージクラスタ構成のストレージシステムの構成を示すブロック図。
【図１６】本発明の第３の実施形態に係るストレージクラスタ構成のストレージシステムの構成を示すブロック図。
【図１７】同第３の実施形態においてホストからアクセス要求が発行された場合の処理の手順を示すフローチャート。
【図１８】本発明の第４の実施形態に係るストレージクラスタ構成のストレージシステムの構成を示すブロック図。
【図１９】同第４の実施形態においてホストからアクセス要求が発行された場合の処理の手順を示すフローチャート。

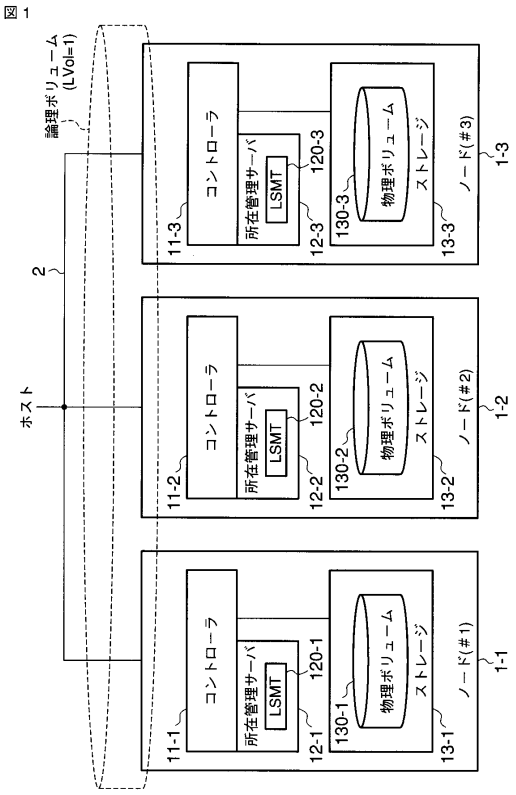
【図２０】本発明の第５の実施形態においてホストからアクセス要求が発行された場合の処理の手順を示すフローチャート。

【符号の説明】

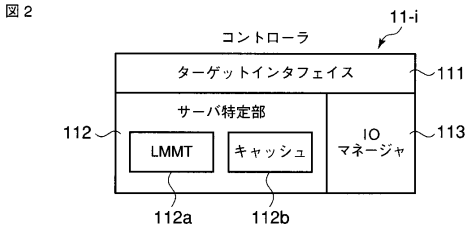
【０１４４】

１-1～１-3…ノード、１１-1～１１-3，１１-i…コントローラ、１２-1～１２-3…所在管理サーバ、１３-1～１３-3，１３-i…ストレージ、１１２ａ…ＬＭＭＴ（論理エクステント－所在管理サーバマッピングテーブル、第１のマッピングテーブル）、１１２ｂ…キャッシュ、１２０-1～１２０-3，１２０-i…ＬＳＭＴ（論理エクステント－ストレージマッピングテーブル、第２のマッピングテーブル）、１３０-1～１３０-3，１３０-i…物理ボリューム、１３２ａ…ＬＰＭＴ（論理エクステント－物理エクステントマッピングテーブル、第３のマッピングテーブル）。

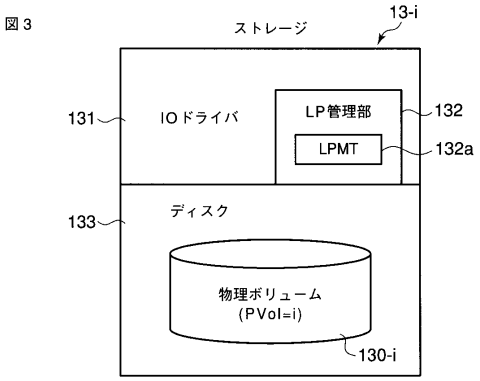
【図１】



【図２】

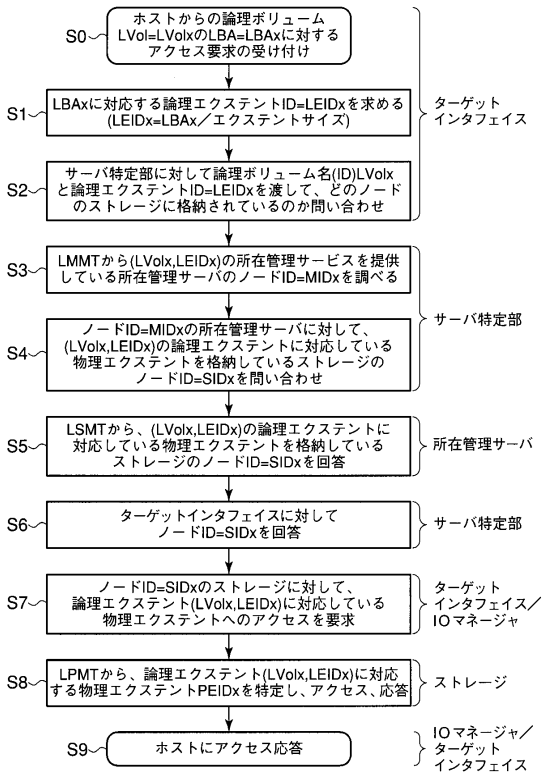


【図３】



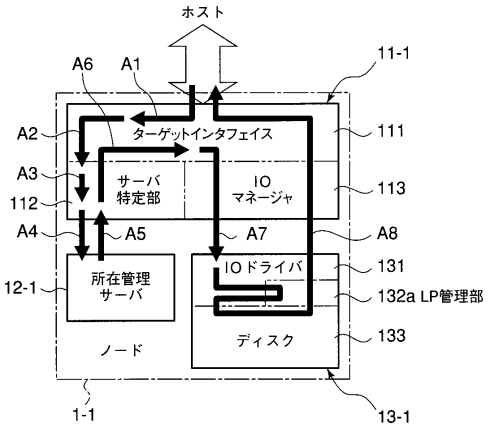
【図４】

図４



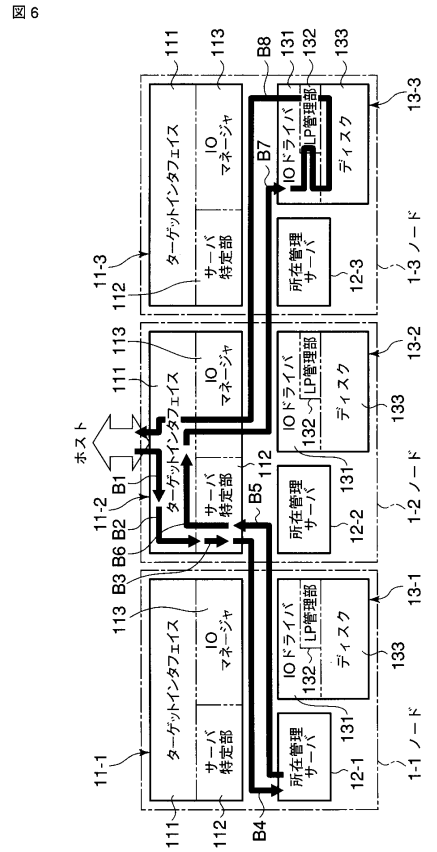
【図５】

図５



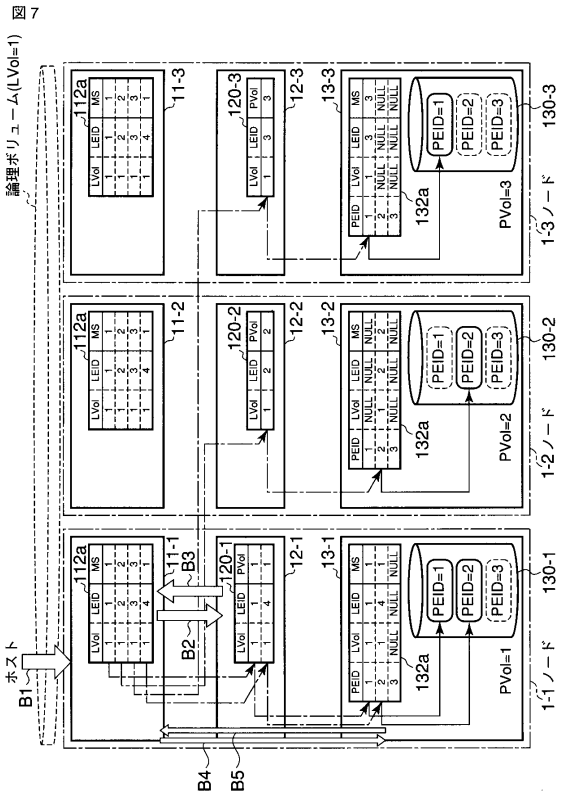
【図６】

図６



【図７】

図７



【図８】

図８

LMMT 112a

LVol	LEID	MS
1	1	1
1	2	2
1	3	3
1	4	1

【図９】

図９

LSMT 120-i

LVol	LEID	PVol
1	1	1
1	4	1

【図１０】

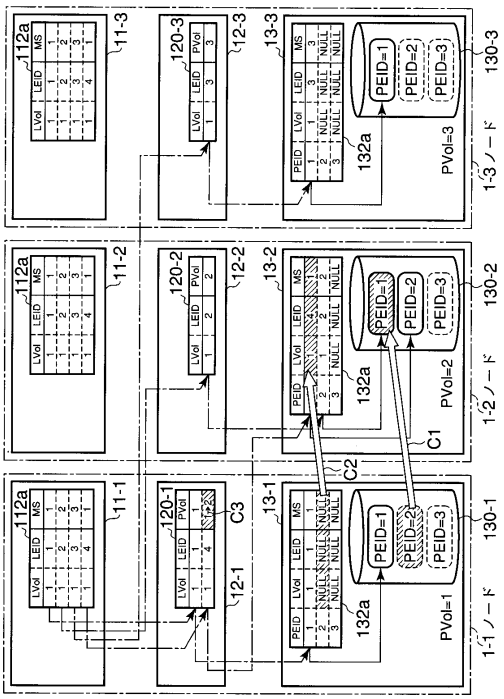
図10

LPMT 132a

PEID	VLol	LEID	MS
1	1	1	1
2	1	4	1
3	NULL	NULL	NULL

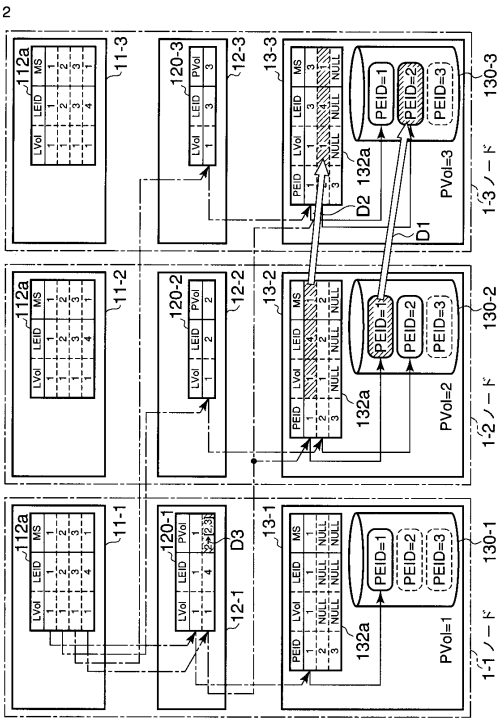
【図１１】

図11



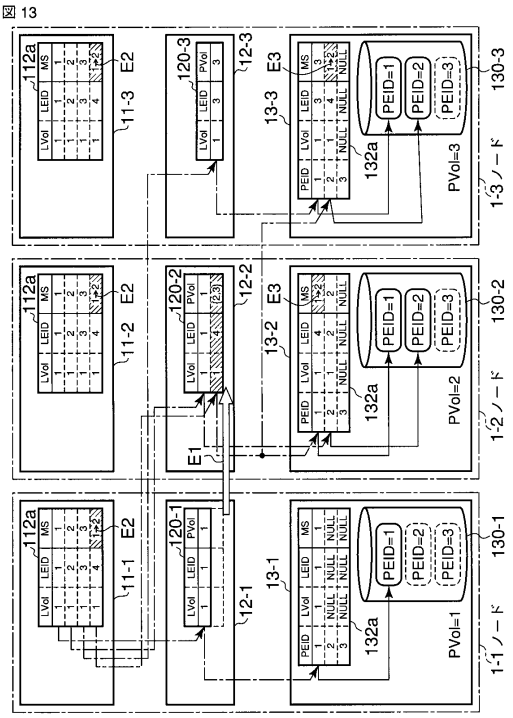
【図１２】

図12



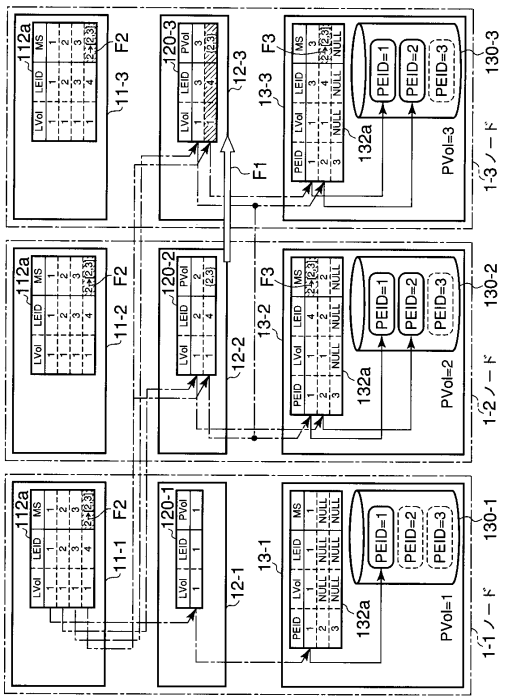
【図１３】

図13



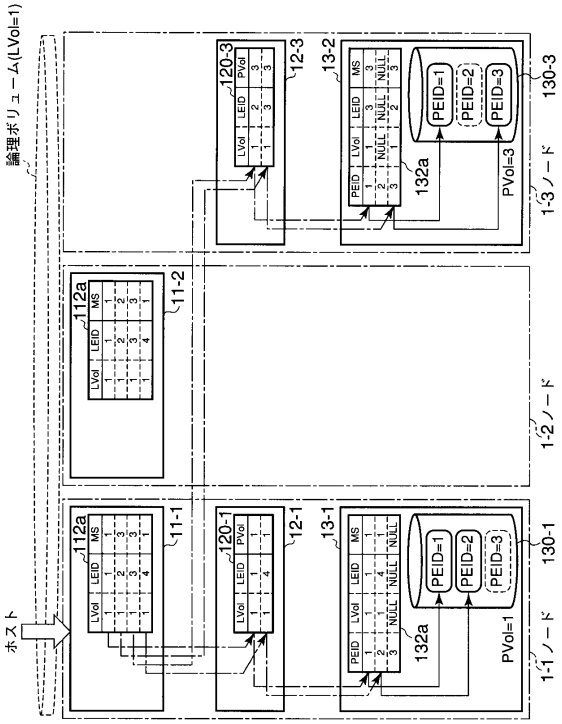
【図１４】

図 14



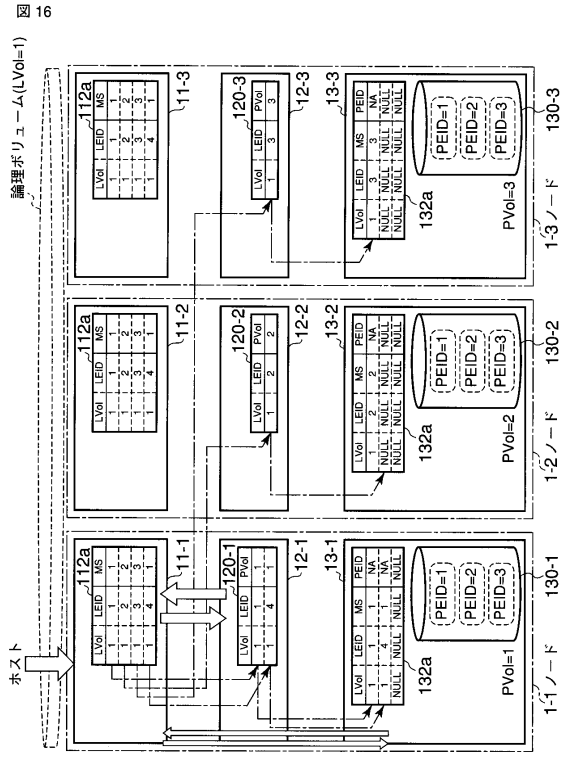
【図１５】

図 15



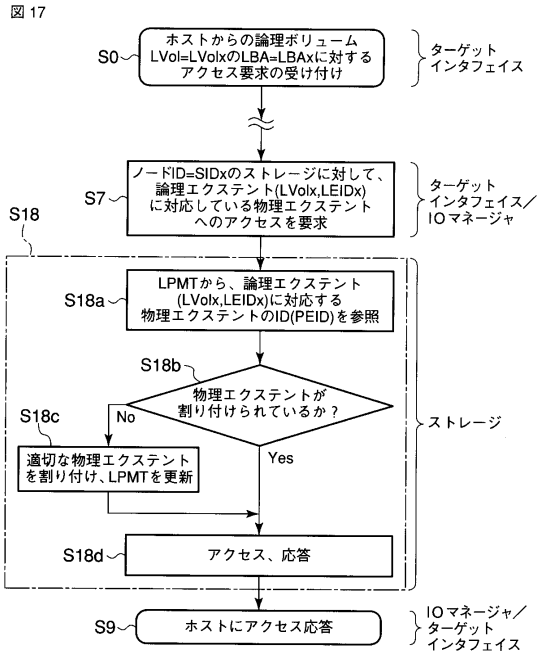
【図１６】

図 16



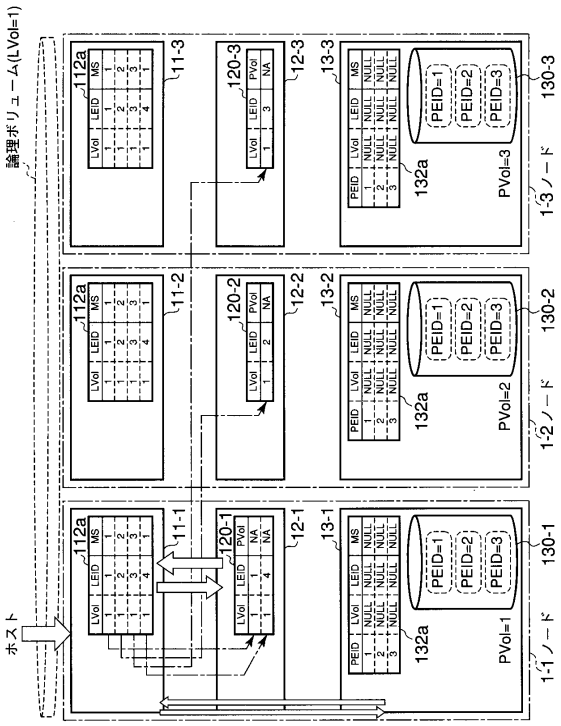
【図１７】

図 17



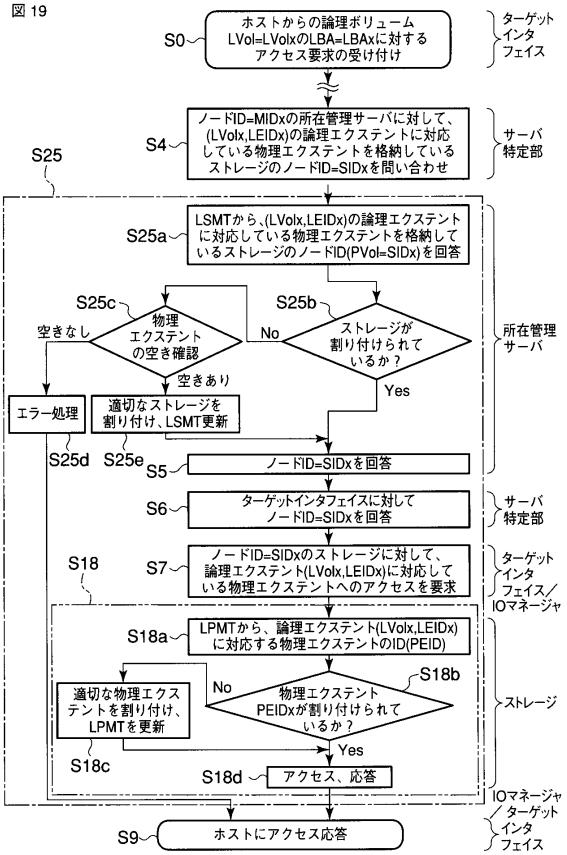
【図１８】

図 18



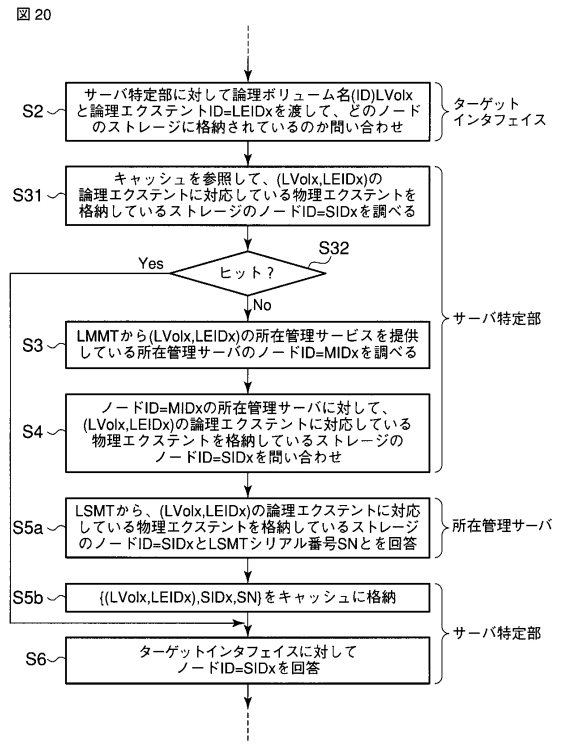
【図１９】

図 19



【図２０】

図 20



フロントページの続き

(74)代理人 100075672
弁理士 峰 隆司
(74)代理人 100109830
弁理士 福原 淑弘
(74)代理人 100084618
弁理士 村松 貞男
(74)代理人 100092196
弁理士 橋本 良郎
(72)発明者 小原 誠
東京都港区芝浦一丁目１番１号 東芝ソリューション株式会社内

審査官 木村 雅也

(56)参考文献 特開平０７－１５２４９１（ＪＰ，Ａ）
特開２００４－０７８３９８（ＪＰ，Ａ）
特開２００１－３１２３７２（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ 3／０６