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

特許第6081293号
(P6081293)

(45)発行日　平成29年2月15日(2017.2.15)　(24)登録日　平成29年1月27日(2017.1.27)

(51)Int.Cl.		F I		
G06F 13/14	(2006.01)	G06F	13/14	330A
G06F 3/06	(2006.01)	G06F	3/06	301A
G06F 13/10	(2006.01)	G06F	13/10	340A
		G06F	13/10	330C

請求項の数 20　(全 23 頁)

(21)出願番号　特願2013-107281 (P2013-107281)
(22)出願日　平成25年5月21日 (2013.5.21)
(65)公開番号　特開2013-246821 (P2013-246821A)
(43)公開日　平成25年12月9日 (2013.12.9)
審査請求日　平成28年5月9日 (2016.5.9)
(31)優先権主張番号　13/479,118
(32)優先日　平成24年5月23日 (2012.5.23)
(33)優先権主張国　米国(US)

早期審査対象出願

(73)特許権者　510149482
ヴイエムウェア　インコーポレイテッド
ＶＭｗａｒｅ，Ｉｎｃ．
アメリカ合衆国　９４３０４　カリフォルニア州　パロ　アルト　ヒルビュー　アベニュー　３４０１
(74)代理人　100105957
弁理士　恩田　誠
(74)代理人　100068755
弁理士　恩田　博宣
(74)代理人　100142907
弁理士　本田　淳

最終頁に続く

(54)【発明の名称】ファブリック分散リソースのスケジューリング

(57)【特許請求の範囲】
【請求項１】
ストレージ・エリア・ネットワーク（ＳＡＮ）における複数の仮想マシン（ＶＭ）向けの、集中型の入出力（Ｉ／Ｏ）負荷ベースパス選択用システムであって、
コンピュータ装置に関連した記憶エリアであって、前記記憶エリアは前記ＳＡＮのトポロジ記述を記憶し、前記トポロジ記述は少なくとも前記ＳＡＮ用の閾値容量を含む、記憶エリアと、
プロセッサと、
を備え、前記プロセッサは、
前記仮想マシンを実行する複数のホストのうちの１つから、前記ホストが認識可能な少なくとも１つの論理的記憶装置にそれぞれ至る複数の記憶パスを識別することによって、初期の記憶パス選択を遂行する工程であって、前記識別には最大フロー・アルゴリズムを繰り返し用い、同最大フロー・アルゴリズムに用いる、前記ホストと前記論理的記憶装置との間の各エッジのウェイトを、前記識別を繰り返す毎に増大させる、工程と
前記遂行する工程において識別された前記記憶パスにおける複数の記憶パスが形成するリンクに基づき、識別された前記記憶パスのそれぞれについてＩ／Ｏ負荷を求める工程と、
前記記憶パスのそれぞれについての求められた前記Ｉ／Ｏ負荷を前記閾値容量から減じて、識別された前記記憶パスのそれぞれに対して、それぞれの識別された前記記憶パスに適合する最小限の帯域幅を見いだす工程と、

前記最小限の帯域幅に基づき、前記複数のホストのそれぞれに対して、識別された前記記憶パスのうちの１つを選択する工程と、
を行うようにプログラムされる、システム。
【請求項２】
前記プロセッサはさらに、
前記ＳＡＮのトポロジの変更を検出し、
検出された前記変更に応じて、前記初期の記憶パス選択を再度遂行するようにプログラムされる、
請求項１に記載のシステム。
【請求項３】
前記プロセッサはさらに、求められた前記Ｉ／Ｏ負荷に基づき前記ＳＡＮのトポロジの変更を推奨するようにプログラムされる、
請求項１に記載のシステム。
【請求項４】
前記プロセッサはさらに、前記ホストのうち少なくとも１つに対して前記複数の記憶パスから別の記憶パスを選択することと、別のコントローラに対して前記論理的記憶装置を再マッピングするように、記憶プロセッサ・アレイに命じることと、前記論理的記憶装置から別の論理的記憶装置へとデータを移動することと、前記複数のホストのうちの第１のホストから前記複数のホストのうちの第２のホストへとデータストアを移動することと、のうち１または複数を前記ＳＡＮの管理者からの入力なしで遂行することによって、前記ＳＡＮの前記トポロジに対する推奨された前記変更を開始するようにプログラムされる、
請求項３に記載のシステム。
【請求項５】
前記記憶エリアはさらに、前記複数の記憶パスが形成するリンクに関連したリンク容量を記憶する、
請求項１に記載のシステム。
【請求項６】
前記プロセッサは、前記最小限の帯域幅および前記記憶エリアに記憶された前記リンク容量に基づき、識別された前記記憶パスのうち前記１つを選択するようにプログラムされる、
請求項５に記載のシステム。
【請求項７】
前記システムはさらに、
前記記憶パスの輻輳を改善するためにトポロジ変更を識別する手段と；
前記記憶パスの輻輳を改善するためにトポロジ変更を自動的に開始する手段とを備える、
請求項１に記載のシステム。
【請求項８】
分散型共有リソース記憶システムにおける集中型の入出力（Ｉ／Ｏ）負荷ベースパス選択の方法であって、
前記分散型共有リソース記憶システムにおいて動作するプロセッサが、複数のホストのうちの１つから、前記ホストが認識可能な少なくとも１つの論理的記憶装置にそれぞれ至る複数の記憶パスを検索する工程であって、前記検索には最大フロー・アルゴリズムを繰り返し用い、同最大フロー・アルゴリズムに用いる前記ホストと前記論理的記憶装置との間の各エッジのウェイトを、前記検索を繰り返す毎に増大させる、工程と
前記検索する工程において検索された前記記憶パスが形成する複数のリンクに基づき、検索された前記記憶パスのそれぞれについてＩ／Ｏ負荷を求める工程と；
前記記憶パスのそれぞれについての求められた前記Ｉ／Ｏ負荷を閾値容量から減じて、前記記憶パスのそれぞれに対して、それぞれの前記記憶パスについての最小限の帯域幅を見いだす工程と；

前記複数のホストのそれぞれに対して、前記記憶パスのうちの１つを前記最小限の帯域幅に基づき選択する記憶パス選択工程と
を有する、方法。
【請求項９】
求められた前記Ｉ／Ｏ負荷に基づき、前記分散型共有リソース記憶システムにおけるトポロジ変更を選択する工程を含む、
請求項８に記載の方法。
【請求項１０】
選択された前記トポロジ変更を開始する工程を含む、
請求項９に記載の方法。
【請求項１１】
変更を検出するために時間を通じて前記Ｉ／Ｏ負荷を監視する工程を含む、
請求項８に記載の方法。
【請求項１２】
前記記憶パスのうちの１つに関して、監視された前記Ｉ／Ｏ負荷における変更を検出すると、
前記記憶パスのうちの前記１つに対する更新されたＩ／Ｏ負荷を求める工程と、
前記記憶パスのそれぞれに対して、それぞれの前記記憶パスに適合する最小限の帯域幅を算定する工程と、
前記複数のホストのそれぞれに対して、前記記憶パスのうちの１つを算定された前記最小限の帯域幅に基づき選択する工程と、
を含む、
請求項１１に記載の方法。
【請求項１３】
前記記憶パス選択工程はさらに、前記最小限の帯域幅に基づきバイナリ・サーチを遂行する工程を含む、
請求項８に記載の方法。
【請求項１４】
対応する前記ホストに対して選択された記憶パスのそれぞれを識別する工程を含む、
請求項８に記載の方法。
【請求項１５】
前記複数のホストのうちの１つ、ホスト・バス・アダプタ、アレイ・ポート、および論理装置番号（ＬＵＮ）を識別することによって、前記複数の記憶パスのそれぞれを定義する工程を含む、
請求項８に記載の方法。
【請求項１６】
コンピュータ実行可能命令を含む１または複数のコンピュータ可読記憶媒体であって、前記コンピュータ実行可能命令が実行されると、少なくとも１つのプロセッサは、
分散型共有リソース記憶システムにおいて動作するプロセッサが、複数のホストのうちの１つから、前記ホストが認識可能な少なくとも１つの論理的記憶装置にそれぞれ至る複数の記憶パスを検索する工程であって、前記検索には最大フロー・アルゴリズムを繰り返し用い、同最大フロー・アルゴリズムに用いる前記ホストと前記論理的記憶装置との間の各エッジのウェイトを、前記検索を繰り返す毎に増大させる、工程と、
検索された前記記憶パスの少なくとも１つにおける少なくとも１つの記憶パスに対して上限リンク容量を定義する工程と、
前記記憶パスのそれぞれについてＩ／Ｏ負荷を求める工程と、
前記記憶パスのそれぞれについての求められた前記Ｉ／Ｏ負荷を閾値容量から減じて、前記記憶パスのそれぞれに対して、それぞれの前記記憶パスおよび定義された前記上限リンク容量に適合する最小限の帯域幅を見いだす工程と、
前記複数のホストのそれぞれに対して、前記記憶パスのうちの１つを最小限の帯域幅に

基づき選択する工程と、
を行うことによって集中型の入出力（Ｉ／Ｏ）負荷ベースパス選択を遂行する、コンピュータ可読記憶媒体。
【請求項１７】
前記複数のホストは１または複数の仮想マシン（ＶＭ）を実行する、請求項８に記載の方法。
【請求項１８】
前記コンピュータ実行可能命令によって、前記プロセッサは、前記記憶パスにおけるリンクの過剰予約を防止するように前記上限リンク容量を定義する、
請求項１６に記載のコンピュータ可読記憶媒体。
【請求項１９】
前記コンピュータ実行可能命令によって、前記プロセッサは、前記記憶パスの内部のスイッチに問い合わせて前記スイッチの最大帯域幅容量を得ることによって、前記上限リンク容量を定義する、
請求項１６に記載のコンピュータ可読記憶媒体。
【請求項２０】
前記コンピュータ実行可能命令によって、前記プロセッサはさらに、前記最小限の帯域幅および前記スイッチの前記最大帯域幅容量に基づき、前記複数のホストのそれぞれに対して前記記憶パスのうち１つを選択する、
請求項１９に記載のコンピュータ可読記憶媒体。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明概念は、一般にデータ記憶装置の分野に関する。より詳細には、本開示は、共有記憶システムおよび関連方法に関する。
【背景技術】
【０００２】
共有記憶システムでは、複数のホストは、記憶装置の同一の組および／または記憶入出力（Ｉ／Ｏ）経路の同一の組を共有してよい。共有記憶システムの実例の１つには、多くの仮想マシン（ＶＭ：ｖｉｒｔｕａｌ　ｍａｃｈｉｎｅ）を実行する仮想データセンタがある。或る現存システムでは、ファブリック輻輳（ｃｏｎｇｅｓｔｉｏｎ）を防止するのが困難である。この困難は、部分的には、ファブリック・トポロジ、Ｉ／Ｏ負荷、およびＩ／Ｏ経路選択の複雑さによる場合がある。また、ファブリック・トポロジ、Ｉ／Ｏ負荷、およびＩ／Ｏ経路選択の変更による困難が存在することもある。例えば、仮想マシン、仮想ディスク、ホスト、記憶装置（または他の記憶領域）、およびファイバ・チャネル・リンクが、動的に付加されたり除去されたりする可能性がある。
【０００３】
固体ディスク（ＳＳＤ：ｓｏｌｉｄ－ｓｔａｔｅ　ｄｉｓｋ）などの先進のディスク技術は、他のタイプのディスクよりも優れたランダムＩ／Ｏ性能をもたらす。ＳＳＤは、キャッシュ記憶装置として、フロントエンド段として、および／または完全なスピンドル交換として使用されている。ＳＳＤを用いると、３ギガバイト／秒ものランダム・ディスクのＩ／Ｏトラフィックを達成することが可能になり得て、記憶装置または論理装置番号（ＬＵＮ：ｌｏｇｉｃａｌ　ｕｎｉｔ　ｎｕｍｂｅｒ）当たりのＩ／Ｏ帯域幅の向上をもたらす。さらに、バックアップ、クローニング、およびテンプレート配置など、高スループットの連続した入出力操作によって、ファブリック・リンクの飽和状態および／または故障が生じる恐れがある。
【先行技術文献】
【特許文献】
【０００４】
【特許文献１】米国特許出願公開第２０１２／００５４３２９号明細書

【特許文献２】米国特許出願公開第２０１０／００８３２６２号明細書
【特許文献３】米国特許出願公開第２０１１／０１１９４１３号明細書
【特許文献４】米国特許出願公開第２０１１／００７２２０８号明細書
【非特許文献】
【０００５】
【非特許文献１】ＧＵＬＡＴＩ ｅｔ ａｌ．，”ＰＡＲＤＡ：Ｐｒｏｐｏｒｔｉｏｎａｌ Ａｌｌｏｃａｔｉｏｎ ｏｆ Ｒｅｓｏｕｒｃｅｓ ｆｏｒ Ｄｉｓｔｒｉｂｕｔｅｄ Ｓｔｏｒａｇｅ Ａｃｃｅｓｓ”，７ｔｈ ＵＳＥＮＩＸ Ｃｏｎｆｅｒｅｎｃｅ ｏｎ Ｆｉｌｅ ａｎｄ Ｓｔｏｒａｇｅ Ｔｅｃｈｎｏｌｏｇｉｅｓ，２００９，１４ ｐａｇｅｓ．
【非特許文献２】ＵＮＫＮＯＷＮ，”ＶＭｗａｒｅ ＶＲＯＯＭ！ Ｂｌｏｇ：３５０，０００ Ｉ／Ｏ ｏｐｅｒａｔｉｏｎｓ ｐｅｒ Ｓｅｃｏｎｄ，Ｏｎｅ ｖＳｐｈｅｒｅ Ｈｏｓｔ”，Ｒｅｔｒｉｅｖｅｄ ｆｒｏｍ ＜＜ｈｔｔｐ：／／ｂｌｏｇｓ．ｖｍｗａｒｅ．ｃｏｍ／ｐｅｒｆｏｒｍａｎｃｅ／２００９／０５／３５００００－ｉｏ－ｏｐｅｒａｔｉｏｎｓ－ｐｅｒ－ｓｅｃｏｎｄ－ｏｎｅ－ｖｓｐｈｅｒｅ－ｈｏｓｔ－ｗｉｔｈ－３０－ｅｆｄｓ．ｈｔｍｌ＞＞，Ｍａｙ １８，２００９，２ ｐａｇｅｓ．
【非特許文献３】ＵＮＫＮＯＷＮ，”Ｖｅｒｉｔａｓ Ｓｔｏｒａｇｅ Ｆｏｕｎｄａｔｉｏｎ ５．０ Ｄｙｎａｍｉｃ Ｍｕｌｔｉ－Ｐａｔｈｉｎｇ”，Ｒｅｔｒｉｅｖｅｄ ｆｒｏｍ ＜＜ｈｔｔｐ：／／ｅｖａｌ．ｓｙｍａｎｔｅｃ．ｃｏｍ／ｍｋｔｇｉｎｆｏ／ｅｎｔｅｒｐｒｉｓｅ／ｗｈｉｔｅ＿ｐａｐｅｒｓ／ｅｎｔ－ｗｈｉｔｅｐａｐｅｒ＿ｖｓｆ＿５．０＿ｄｙｎａｍｉｃ＿ｍｕｌｔｉ－ｐａｔｈｉｎｇ＿０５－２００７．ｅｎ－ｕｓ．ｐｄｆ＞＞，Ｍａｙ ２００７，４１ ｐａｇｅｓ．
【発明の概要】
【発明が解決しようとする課題】
【０００６】
個々のホスト向け最適パスを手動で求めることによって動的事象に応答することは、困難であり、信頼性が低く、エラーが発生しやすく、効果的な負荷バランシングをもたらす可能性が低い。さらに、現存システムには、マルチパス化、入出力制限、ＬＵＮパス選択技法の遂行、１つのホストから別のホストへの作業負荷の移動、または１つのＬＵＮもしくはデータストアから別のＬＵＮへとデータを移動することによって、負荷バランスを試みるものもある。しかし、このような現存システムは、ＬＵＮ輻輳とリンク輻輳を区別しない。そのため、これらの現存システムは、現行パスが輻輳しているとき、ＬＵＮにアクセスするための代替パスを提案することができない。さらに、現存システムの多くはホスト・レベルで動作するものであり、したがって、全体最適をもたらすこと、あるいはボトルネックを改善するためのトポロジ変更または代替パスを推奨することができない。
【０００７】
当技術のこれらの困難または他の困難に対処するシステムおよび方法の提供が望まれており、これらのシステムおよび方法は、当業者に、本明細書の教示および議論を考慮した後に理解されるであろう。
【課題を解決するための手段】
【０００８】
本発明によれば、添付の特許請求の範囲で説明される装置および方法が提供される。本発明の他の特徴は、従属請求項およびそれに続く記述から明らかになるであろう。
この概要は、以下で、より詳細に説明される概念の選択を紹介する。この概要は、基本的な特徴を列挙したり、要求された対象の範囲を限定したりするようには意図されていない。
【０００９】
１例では、共有ストレージ・エリア・ネットワーク（ＳＡＮ：ｓｔｏｒａｇｅ ａｒｅａ ｎｅｔｗｏｒｋ）における、集中型の全体最適化された入出力（Ｉ／Ｏ）負荷ベースパス管理を有するシステムが提供される。複数のホストのうちの１つから、同ホストが認

識可能な少なくとも１つの論理的記憶装置にそれぞれ至る複数の記憶パスが識別される。少なくとも記憶パス上のＩ／Ｏ負荷および反復検索技法に基づき、それぞれのリンクまたは記憶パスに対する最小限の帯域幅が求められる。ホストと論理的記憶装置の各対に対して、記憶パスのうちの１つが最小限の帯域幅に基づき選択される。

【００１０】

本明細書で説明される例示的システムは、記憶パスに沿って、かつ／または論理的記憶装置において、輻輳の確率を低減するために、ＳＡＮトポロジおよびＩ／Ｏ負荷を監視して、トポロジの変更を識別し、変更を開始する。或る実例では、トポロジの変更には、ホスト上で動作している複数の仮想マシン（ＶＭ）と関連した移動する負荷および／またはデータストアが含まれる。

【００１１】

１例では、エリア・ネットワーク（ＳＡＮ）における複数の仮想マシン（ＶＭ）向けの、集中型の入出力（Ｉ／Ｏ）負荷ベースパス選択用システムが提供され、前記システムは、少なくともＳＡＮ用の閾値容量を含む同ＳＡＮのトポロジ記述を記憶する、コンピュータ装置に関連した記憶エリアと、ＶＭを実行する複数のホストのうちの１つから、同ホストが認識可能な少なくとも１つの論理的記憶装置にそれぞれ至る複数の記憶パスを識別することによって、初期のパス選択を遂行し、識別された記憶パスのそれぞれについてＩ／Ｏ負荷を求め、記憶パスのそれぞれについて求められたＩ／Ｏ負荷を閾値容量から減じて、識別された記憶パスのそれぞれに対して、それぞれの識別された記憶パスに適合する最小限の帯域幅を見いだし、複数のホストのそれぞれに対して、識別された記憶パスのうちの１つを最小限の帯域幅に基づき選択するようにプログラムされたプロセッサと、からなる。

【００１２】

１例では、このプロセッサは、ＳＡＮのトポロジの変更を検出し、検出された変更に応じて、初期のパス選択を再度遂行するようにさらにプログラムされる。

１例では、このプロセッサは、求められたＩ／Ｏ負荷に基づきＳＡＮのトポロジ変更を推奨するようにさらにプログラムされる。

【００１３】

１例では、このプロセッサは、ホストのうち少なくとも１つに対して複数の記憶パスから別のパスを選択する工程と、様々なコントローラに対して論理的記憶装置を再マッピングするように、記憶プロセッサ・アレイに命じる工程と、この論理的記憶装置から別の論理的記憶装置へとデータを移動する工程と、複数のホストの中の第１のホストから複数のホストの中の第２のホストへとデータストアを移動する工程と、のうち１または複数をＳＡＮの管理者からの入力なしで遂行することによって、ＳＡＮのトポロジに対する推奨された変更を開始するようにさらにプログラムされる。

【００１４】

１例では、複数の記憶パスのそれぞれが複数のリンクからなり、記憶エリアは、複数のリンクに関連したリンク容量をさらに記憶する。

１例では、このプロセッサは、最小限の帯域幅および記憶エリアに記憶されたリンク容量に基づき、識別された記憶パスのうち１つを選択するようにプログラムされる。

【００１５】

１例では、このシステムは、記憶パスの輻輳を修正するためにトポロジ変更を識別する手段と、記憶パスの輻輳を修正するためにトポロジ変更を自動的に開始する手段とを含む。

【００１６】

１例では、分散型共有リソース記憶システムにおける集中型入出力（Ｉ／Ｏ）負荷ベースパス選択の方法が提供され、前記方法は、複数のホストのうちの１つから、同ホストが認識可能な少なくとも１つの論理的記憶装置にそれぞれ至る複数の記憶パスに、該分散型共有記憶システムにおいて動作するプロセッサでアクセスする工程と、アクセスされた記憶パスのそれぞれについてＩ／Ｏ負荷を求める工程と、記憶パスのそれぞれについての求

められたＩ／Ｏ負荷を閾値容量から減じて、アクセスされた記憶パスのそれぞれに対して、それぞれのアクセスされた記憶パスについての最小限の帯域幅を見いだす工程と、複数のホストのそれぞれに対して、アクセスされた記憶パスのうちの１つを最小限の帯域幅に基づき選択する工程と、からなる。

【００１７】

１例では、この方法は、求められたＩ／Ｏ負荷に基づき、分散型共有リソース記憶システムにおけるトポロジ変更を選択する工程を含む。

１例では、この方法は、選択されたトポロジ変更を開始する工程を含む。

【００１８】

１例では、この方法は、変更を検出するために長時間にわたってＩ／Ｏ負荷を監視する工程を含む。

１例では、この方法は、記憶パスのうちの１つに関して、監視されたＩ／Ｏ負荷における変更を検出すると、記憶パスのうちの１つに対する更新されたＩ／Ｏ負荷を求める工程と、記憶パスのそれぞれに対して、それぞれの記憶パスに適合する最小限の帯域幅を算定する工程と、複数のホストのそれぞれに対して、記憶パスのうちの１つを、算定された最小限の帯域幅に基づき選択する工程と、を含む。

【００１９】

１例では、複数のホストのそれぞれに対して記憶パスのうちの１つを選択する工程は、最小限の帯域幅に基づきバイナリ・サーチを遂行する工程をさらに含む。

１例では、この方法は、各繰返しに対してエッジ・ウェイトを増す最大フロー・アルゴリズムを繰り返し遂行することによって、複数の記憶パスを検索する工程を含む。

【００２０】

１例では、この方法は、対応するホストに対して選択されたパスのそれぞれを識別する工程を含む。

１例では、この方法は、複数のホスト、ホスト・バス・アダプタ、アレイ・ポート、および論理装置番号（ＬＵＮ）のうちの１つを識別することによって、複数の記憶パスのそれぞれを定義する工程を含む。

【００２１】

１例では、コンピュータ実行可能命令を含むコンピュータ可読記憶媒体が提供され、これらのコンピュータ実行可能命令が実行されると、少なくとも１つのプロセッサが、複数のホストのうちの１つから、同ホストが認識可能な少なくとも１つの論理的記憶装置にそれぞれ至る複数の記憶パスに、分散型共有記憶システムにおいて動作するプロセッサでアクセスする工程と、アクセスされた記憶パスの少なくとも１つにおける少なくとも１つのリンクに対して上限リンク容量を定義する工程と、アクセスされた記憶パスのそれぞれに対するＩ／Ｏ負荷を求める工程と、記憶パスのそれぞれについての求められたＩ／Ｏ負荷を閾値容量から減じて、記憶パスのそれぞれに対して、それぞれの記憶パスおよび定義された上限リンク容量に適合する最小限の帯域幅を見いだす工程と、複数のホストのそれぞれに対して、記憶パスのうちの１つを、算定された最小限の帯域幅に基づき選択する工程と、によって集中型の入出力（Ｉ／Ｏ）負荷ベースパス選択を遂行する。

【００２２】

１例では、これらの工程は、複数の記憶パスの中のリンクの過剰な予約を防止するために上限リンク容量を定義する工程を含む。

１例では、これらの工程は、記憶パスの内部のスイッチに問い合わせてスイッチの最大帯域幅容量を得ることによって、上限リンク容量を定義する工程を含む。

【００２３】

１例では、これらの工程は、最小限の帯域幅およびスイッチの最大帯域幅容量に基づき、複数のホストのそれぞれに対して記憶パスのうちの１つを選択する工程を含む。

１例では、ストレージ・エリア・ネットワーク（ＳＡＮ）における複数の仮想マシン（ＶＭ）向けの、集中型の入出力（Ｉ／Ｏ）負荷ベースパス管理の方法が提供され、前記方法は、ＶＭを実行する複数のホストのうちの１つから、同ホストが認識可能な少なくとも

１つの論理的記憶装置にそれぞれ至る複数の記憶パスに、プロセッサでアクセスする工程と、アクセスされた記憶パスのそれぞれについてＩ／Ｏ負荷を監視する工程と、監視されたＩ／Ｏ負荷が記憶パスの少なくとも１つにおける輻輳を表していると判断すると、１または複数のトポロジ変更を識別する工程と、識別された１または複数のトポロジ変更を、プロセッサによって開始する工程と、からなる。

【００２４】

１例では、識別された１または複数のトポロジ変更を開始する工程は、ＳＡＮの管理者からの入力なしでトポロジ変更を開始する工程を含む。

１例では、識別された１または複数のトポロジ変更を開始する工程は、複数の記憶パスから、ホストのうち少なくとも１つ向けに別のパスを選択する工程と、ホストのうち少なくとも１つに対して選択された別のパスを用いるように指示する工程とをさらに含む。

【００２５】

１例では、識別された１または複数のトポロジ変更を開始する工程は、論理的記憶装置を管理している記憶プロセッサ・アレイに対して、論理的記憶装置を別のコントローラに再マッピングするように命令を送る工程を含む。

【００２６】

１例では、識別された１または複数のトポロジ変更を開始する工程は、この論理的記憶装置から別の論理的記憶装置へとデータを移動する工程と、複数のホストの中の第１のホストから複数のホストの中の第２のホストへとデータストアを移動する工程と、のうち１または複数を含む。

【図面の簡単な説明】

【００２７】

【図１】例示のホスト・コンピュータ装置のブロック図。

【図２】図１に示されたホスト・コンピュータ装置などのコンピュータ装置上にインスタンスが生成される仮想マシンのブロック図。

【図３】ネットワーク・スイッチによって記憶パスを介してディスク・アレイにアクセスするホストを含む例示のストレージ・エリア・ネットワーク（ＳＡＮ）構成のブロック図。

【図４】ネットワーク・フロー解析に対する最適の記憶パス選択の例示のマッピングのブロック図。

【図５】集中型の入出力（Ｉ／Ｏ）記憶パスの選択および管理を実施するための例示のコンピュータ装置のブロック図。

【図６】ＳＡＮの中の最適の記憶パスを選択するように、複数のホストを管理する集中型システムによって遂行される例示の方法の流れ図。

【図７】リンクまたは論理装置番号（ＬＵＮ）の輻輳に応答してトポロジ変更を開始するように、複数のホストを管理する集中型システムによって遂行される例示の方法の流れ図。

【発明を実施するための形態】

【００２８】

対応する参照文字は、各図面を通じて該当部分を示す。

本明細書で説明される実施形態は、共有ストレージアレイを有し、ＳＡＮ　５１２などのストレージ・エリア・ネットワーク（ＳＡＮ）における複数のホスト（例えばクラスタ）にわたって集中型の入出力（Ｉ／Ｏ）パス選択および管理を実施する。パス推奨は、ファブリック（例えばリンク）のＩ／Ｏ負荷状態に基づくものであり、例えばネットワーク・フロー・アルゴリズム（例えば最大フロー・アルゴリズム）の最上位で条件付き反復検索を用いて選択される。結果として、本開示の態様は、ホスト５１４からアレイ３０２までのパス（経路）上のどの単一ポイントにも過負荷をかけず、したがって全体最適をもたらすように、各ホストとＬＵＮの対に対してＩ／Ｏパスを選択する。

【００２９】

或る実施形態では、初期のパス選択推奨は、ディスク・アレイ３０２などのストレージ

アレイにおけるホスト・バス・アダプタ（ＨＢＡ：ｈｏｓｔ　ｂｕｓ　ａｄａｐｔｅｒ）およびポート３０４にわたってパスの数のバランスをとるように、ホスト５１４ごとに記憶装置当たりの均一なＩ／Ｏ負荷を想定して、本明細書に説明される動作を遂行することによって求められる。ホスト５１４ごとに記憶装置当たりのＩ／Ｏ帯域幅消費（例えば負荷または要求）が長時間にわたって監視されながら、パス選択推奨は、再構成される。ＳＡＮ　５１２トポロジが変更されるときも、パス選択推奨が再構成されてよい。或る実施形態では、リンク当たりの帯域幅制限が（既知のとき）強制される。

【００３０】

本開示の態様は、ファブリックにおけるＩ／Ｏ輻輳の可能性を低減し、トポロジ変更および動的Ｉ／Ｏ状態に適合する、または対応する。集中的管理は、所与のトポロジに対して、各ホスト５１４上で認識可能な個々のＬＵＮ用のＩ／Ｏパスを、負荷バランスをとるやり方で計算し、トポロジ変更または他の改善を開始して、トポロジの変更およびリンク輻輳の他の徴候につながる事象に対処する。例えば、集中型のパス選択は、記憶プロセッサ３０６または他の記憶コントローラにわたってＬＵＮまたはＬＵＮ負荷の不平衡などの輻輳シナリオを識別するばかりでなく、特定のホスト５１４用パスの変更、ＬＵＮ侵入（例えばＬＵＮとコントローラのマッピングの変更）の開始、および／またはＩ／Ｏ負荷の移動などの改善も提案する。

【００３１】

或る実施形態は、共有ストレージアレイを有する多くの仮想マシン（ＶＭ）を実行する仮想データセンタにおいて実施される。共有ストレージアレイでは、複数のＬＵＮは、複数の別々のホスト５１４によって同一の記憶パスを介してアクセスされる可能性がある。そのため、合計のスループットが単一パス上で利用可能なリンクの帯域幅を超えることがあり、ボトルネックが生じる。１例として、ＬＵＮが、８Ｇｂ／秒のファイバ・チャネルのリンク速度によって２５０ＭＢ／秒の能力がある場合、１０００ＭＢ／秒の理論上の最大帯域幅は、４つのＬＵＮ上の同時の連続したＩ／Ｏによって飽和状態になる可能性がある。このような同時の連続したＩ／Ｏ動作は、例えば、バックアップ、クローニング、およびテンプレート配置の間に起こる可能性がある。本開示の態様により、パス選択は、リンク上で、またＬＵＮにおいて、このような輻輳を低減することができる。

【００３２】

さらに、集中型パス選択によって、ホスト５１４からアレイへのマルチホップパス上のすべてのリンクが、その最大のリンク容量で動作することが可能になり、ストレージアレイ・ポート３０４および／またはファブリックのリンクのいくつかが、ホスト５１４上のＨＢＡよりも低いリンク速度で動作するときでさえ、最大のエンドツーエンド帯域幅が使用可能になる。例えば、ＨＢＡが８ギガビット／秒で動作する一方で、ストレージアレイ・ポート３０４は、４ギガビット／秒で動作しうる。

【００３３】

さらに、本開示の態様は、複数のパスが単一の記憶ポート３０４に集中する可能性がある、かつ／または不均衡なＬＵＮ－コントローラのマッピング（例えば「インキャスト」問題）がある環境で動作可能である。例えば、複数のホスト５１４が、ディスク・アレイ３０２上のより少量の記憶ポート３０４に接続していることがある。本開示の態様は、このようなシステムの性能を改善する。

【００３４】

次に、１または複数の仮想マシンを実行するホスト・コンピュータ装置１００を含む例示の動作環境が説明される。しかし、本開示の態様は、このような環境に限定されることなく、非ＶＭ環境にも適用可能である。

【００３５】

図１は、例示のホスト・コンピュータ装置１００のブロック図である。ホスト・コンピュータ装置１００は、命令を実行するためのプロセッサ１０２を含み、ホスト５１４のうちの１つと称されてよい。或る実施形態では、実行可能命令は、記憶装置１０４に記憶される。記憶装置１０４は、実行可能命令および／または他のデータなどの情報が記憶され

、かつ取り出され得る任意の装置である。例えば、記憶装置１０４は、１または複数のランダム・アクセス・メモリ（ＲＡＭ）モジュール、フラッシュ・メモリ・モジュール、ハード・ディスク、固体ディスク、および／または光ディスクを含んでよい。

【００３６】

ホスト・コンピュータ装置１００は、ユーザ１０８からデータを受け取るため、および／またはユーザ１０８にデータを表示するためのユーザ・インターフェース装置１１０を含んでよい。ユーザ１０８は、ＶＭｗａｒｅのｖＣｅｎｔｅｒ　Ｓｅｒｖｅｒまたは他の管理装置など別のコンピュータ装置を介してホスト・コンピュータ装置１００と間接的に対話してよい（ｉｎｔｅｒａｃｔ）。ユーザ・インターフェース装置１１０は、例えばキーボード、ポインティング・デバイス、マウス、スタイラス、タッチ・パネル（例えばタッチ・パッドまたはタッチ・スクリーン）、ジャイロスコープ、加速度計、位置検出器、および／またはオーディオ入力装置を含んでよい。或る実施形態では、ユーザ・インターフェース装置１１０は、ユーザ１０８からデータを受け取るように動作し、別の装置（例えばプレゼンテーション装置）は、ユーザ１０８にデータを見せるように動作する。他の実施形態では、ユーザ・インターフェース装置１１０は、ユーザ１０８へのデータ出力およびユーザ１０８からのデータ受取りのどちらにも機能するタッチ・スクリーンなどの単一の構成要素を有する。このような実施形態では、ユーザ・インターフェース装置１１０は、ユーザ１０８に情報を表示するためのプレゼンテーション装置として動作する。このような実施形態では、ユーザ・インターフェース装置１１０は、ユーザ１０８に情報を伝えることができる任意の構成要素を表す。例えば、ユーザ・インターフェース装置１１０は、限定されることなく、表示装置（例えば液晶ディスプレイ（ＬＣＤ）、有機発光ダイオード（ＯＬＥＤ）ディスプレイ、または「電子インク」ディスプレイ）および／または音声出力装置（例えばスピーカまたはヘッドホン）を含んでよい。或る実施形態では、ユーザ・インターフェース装置１１０は、ビデオ・アダプタおよび／またはオーディオ・アダプタなどの出力アダプタを含む。出力アダプタは、プロセッサ１０２に対して動作可能に結合され、表示装置または音声出力装置などの出力装置に対して動作可能に結合されるように構成される。

【００３７】

ホスト・コンピュータ装置１００は、ネットワーク通信インターフェース１１２も含み、これによって、有線または無線のパケット・ネットワークなどの通信媒体を介して遠隔装置（例えば別のコンピュータ装置）と通信することができる。例えば、ホスト・コンピュータ装置１００は、ネットワーク通信インターフェース１１２を介してデータを送受してよい。ユーザ・インターフェース装置１１０および／またはネットワーク通信インターフェース１１２は、まとめて入力インターフェースと称されてよく、ユーザ１０８から情報を受け取るように構成されてよい。

【００３８】

ホスト・コンピュータ装置１００はさらに、記憶インターフェース１１６を含み、これによって、仮想ディスク・イメージ、ソフトウェア・アプリケーション、および／または本明細書に説明された方法とともに用いるのに適切なその他のデータを記憶する１または複数のデータストア３１６と通信することができる。例示的実施形態では、記憶インターフェース１１６は、ホスト・コンピュータ装置１００を、（例えばパケット・ネットワークを介して）ＳＡＮ５１２などのストレージ・エリア・ネットワーク（例えばファイバ・チャネル・ネットワーク）におよび／またはネットワークに接続された記憶（ＮＡＳ：ｎｅｔｗｏｒｋ－ａｔｔａｃｈｅｄ　ｓｔｏｒａｇｅ）システムに結合する。記憶インターフェース１１６は、ネットワーク通信インターフェース１１２と一体化されてよい。

【００３９】

図２は、ホスト・コンピュータ装置１００上にインスタンスが生成される仮想マシン２３５$_1$、２３５$_2$、．．．、２３５$_N$のブロック図を示す。ホスト・コンピュータ装置１００は、ｘ８６アーキテクチャのプラットフォームなどのハードウェア・プラットフォーム２０５を含む。ハードウェア・プラットフォーム２０５は、プロセッサ１０２、記憶装

置１０４、ネットワーク通信インターフェース１１２、ユーザ・インターフェース装置１１０、およびプレゼンテーション装置１０６（図１に示されている）など他の入出力（Ｉ／Ｏ）装置を含んでよい。以下でハイパーバイザ２１０とも称される仮想化ソフトウェア層が、ハードウェア・プラットフォーム２０５の最上位にインストールされる。

【００４０】

仮想化ソフトウェア層は、複数の仮想マシン（ＶＭ　２３５$_1$～２３５$_N$）が同時にインスタンスを生成され、かつ実行され得る仮想マシン実行空間２３０に対応する。ハイパーバイザ２１０は、装置ドライバ層２１５を含む。ハイパーバイザ２１０は、ハードウェア・プラットフォーム２０５の物理的リソース（例えばプロセッサ１０２、記憶装置１０４、ネットワーク通信インターフェース１１２、および／またはユーザ・インターフェース装置１１０）を、「仮想」リソースＶＭ　２３５$_1$～２３５$_N$のそれぞれに対して、マッピングする。それによって、ＶＭ　２３５$_1$～２３５$_N$のそれぞれが、それ自体の仮想ハードウェア・プラットフォーム（例えば仮想ハードウェア・プラットフォーム２４０$_1$～２４０$_N$の対応するもの）を有し、各仮想ハードウェア・プラットフォームが、それ自体のエミュレートされたハードウェア（ＶＭ　２３５$_1$のプロセッサ２４５、記憶装置２５０、ネットワーク通信インターフェース２５５、ユーザ・インターフェース装置２６０および他のエミュレートされた入出力デバイスなど）を有する。ハイパーバイザ２１０は、「ＶＭ　２３５$_1$～２３５$_N$は、ハイパーバイザ２１０の予想外の終了および／または初期設定に際して自動的に再起動されること」などのハイパーバイザ２１０に関する方針に従って、ＶＭ　２３５$_1$～２３５$_N$の実行を管理してよい（例えば、監視、開始、および／または終了してよい）。それに加えて、またはその代わりに、ハイパーバイザ２１０は、ホスト・コンピュータ装置１００とは別の装置から受け取った要求に基づきＶＭ　２３５$_1$～２３５$_N$の実行を管理してよい。例えば、ハイパーバイザ２１０は、管理装置から、ネットワーク通信インターフェース１１２を介して第１のＶＭ　２３５$_1$の実行開始を指示する実行指示を受け取り、これを実行して第１のＶＭ　２３５$_1$の実行を開始してよい。

【００４１】

或る実施形態では、第１の仮想ハードウェア・プラットフォーム２４０$_1$の記憶装置２５０は、ホスト・コンピュータ装置１００の記憶装置１０４（例えばハード・ディスクまたは固体ディスク）に記憶された１または複数の仮想ディスク・イメージに関連づけられた、すなわち「マッピングされた」仮想ディスクを含む。仮想ディスク・イメージは、第１のＶＭ　２３５$_1$によって、単一のファイルまたはそれぞれがファイル・システムの一部分を含んでいる複数のファイルの中で用いられるファイル・システム（例えばディレクトリおよびファイルの階層）を表す。それに加えて、またはその代わりに、仮想ディスク・イメージは、ＳＡＮ構成の中などの１または複数の遠隔コンピュータ装置１００の記憶装置１０４に記憶されてよい。このような実施形態では、いかなる量の仮想ディスク・イメージも遠隔コンピュータ装置１００によって記憶され得る。

【００４２】

例えば、装置ドライバ層２１５は、例えばホスト・コンピュータ装置１００に接続されたローカル・エリア・ネットワーク（ＬＡＮ）との間でデータを送受するネットワーク通信インターフェース１１２と相互作用する通信インターフェース・ドライバ２２０を含む。通信インターフェース・ドライバ２２０には、１つの通信インターフェース（例えばネットワーク通信インターフェース１１２）から他の通信インターフェース（例えばＶＭ　２３５$_1$～２３５$_N$の仮想通信インターフェース）が受け取る、物理的ネットワークにおけるデータ・パケットの同送をシミュレートする仮想ブリッジ２２５も含まれる。第１のＶＭ　２３５$_1$用ネットワーク通信インターフェース２５５など、各ＶＭ　２３５$_1$～２３５$_N$用の各仮想通信インターフェースには、仮想ブリッジ２２５がネットワーク通信インターフェース１１２からの着信データ・パケットの転送をシミュレートすることを可能にする、独自の仮想媒体アクセス制御（ＭＡＣ：Ｍｅｄｉａ　Ａｃｃｅｓｓ　Ｃｏｎｔｒｏｌ）アドレスが割り当てられてよい。１実施形態では、ネットワーク通信インターフェ

ース１１２は、（それ自体の物理的ＭＡＣアドレス宛てのイーサネット（登録商標）・パケットばかりでなく）受け取ったイーサネット（登録商標）・パケットのすべてを仮想ブリッジ２２５に渡す「プロミキャス・モード」に構成されたイーサネット（登録商標）・アダプタであり、仮想ブリッジ２２５は、同様に、これらのイーサネット（登録商標）・パケットをＶＭ　２３５$_1$～２３５$_N$へさらに転送することができる。この構成によって、宛先アドレスとして仮想ＭＡＣアドレスを有するイーサネット（登録商標）・パケットは、このような仮想ＭＡＣアドレスに対応する仮想通信インターフェースを有するホスト・コンピュータ装置１００の中のＶＭに適切に到達することができる。

【００４３】

仮想ハードウェア・プラットフォーム２４０$_1$は、任意のｘ８６互換デスクトップ・オペレーティング・システム（例えばマイクロソフトＷＩＮＤＯＷＳ（登録商標）のオペレーティング・システム、ＬＩＮＵＸ（登録商標）のオペレーティング・システム、ＳＯＬＡＲＩＳ（登録商標）のオペレーティング・システム、ＮＥＴＷＡＲＥ、またはＦＲＥＥＢＳＤ）が、第１のＶＭ　２３５$_1$などのインスタンスを生成されたＶＭ向けのアプリケーション２７０を実行するためのゲスト・オペレーティング・システム（ＯＳ）２６５としてインストールされるように、標準的なｘ８６ハードウェア・アーキテクチャ相当品として機能してよい。仮想ハードウェア・プラットフォーム２４０$_1$～２４０$_N$は、ハイパーバイザ２１０と対応するＶＭ　２３５$_1$～２３５$_N$との間の動作を調整するために仮想システム・サポートを実施する仮想マシン・モニタ（ＶＭＭ：ｖｉｒｔｕａｌ　ｍａｃｈｉｎｅ　ｍｏｎｉｔｏｒ）２７５$_1$～２７５$_N$の一部分であると見なされてよい。当業者は、図２の中の仮想化構成要素を説明するのに用いられた様々な用語、層、および分類は、それらの機能または本開示の趣旨もしくは範囲から逸脱することなく、別々に参照されてよいことを理解するであろう。例えば、仮想ハードウェア・プラットフォーム２４０$_1$～２４０$_N$はＶＭＭ　２７５$_1$～２７５$_N$と分離していると考えられてもよく、ＶＭＭ２７５$_1$～２７５$_N$はハイパーバイザ２１０と分離していると考えられてもよい。本開示の実施形態で用いられ得るハイパーバイザ２１０の実例の１つは、ＶＭｗａｒｅのＥＳＸ商標のソフトウェアに構成要素として含まれており、ＶＭｗａｒｅ社から市販されている。

【００４４】

図３は、ネットワーク・スイッチ５１６によって記憶パスを介してディスク・アレイ３０２にアクセスするホスト５１４を含む例示のストレージ・エリア・ネットワーク（ＳＡＮ）アーキテクチャのブロック図である。一般に、Ｉ／Ｏパスは、ホスト・バス・アダプタ、ケーブル、ＳＡＮスイッチ（例えばスイッチ５１６）、記憶ポート３０４、およびディスク・アレイ３０２におけるディスク・コントローラまたは記憶プロセッサ３０６を含んでよい。他の実施形態では、Ｉ／Ｏパスまたは他の記憶パスは、ホスト５１４、１つのホスト・バス・アダプタ、１つのアレイ・ポート３０４、およびＬＵＮのうちの１つを識別することによって定義されてよい。記憶プロセッサ３０６当たり複数のポート３０４があってもよい。記憶プロセッサ３０６は、１または複数のＬＵＮを構成して、これらを単一の記憶装置としてホスト５１４に表示することによって、アレイの中のディスクを仮想化する。ホスト５１４、スイッチ５１６、ポート３０４、コントローラ、およびＬＵＮのそれぞれの独特な組合せは、様々なＩ／Ｏパスを表す。

【００４５】

図３の実例では、ホストＡ、ホストＢ、およびホストＣを含む３つのホスト５１４は、ネットワーク・スイッチ１およびネットワーク・スイッチ２を含むネットワーク・スイッチ５１６を介してディスク・アレイ３０２に接続する。或る実施形態では、ネットワーク・スイッチ５１６は、ファイバ・チャネル（ＦＣ）のプロトコルに適合する。例示のネットワーク・スイッチ５１６は、それぞれが、４Ｇｂ／秒のリンクを提供し得る。

【００４６】

ホストＡ、Ｂ、およびＣのそれぞれは、１または複数の仮想マシンを実行する。これらのホストのそれぞれは、少なくとも１つのホスト・バス・アダプタを含む。ホストＡは、

ホスト・バス・アダプタＡ１およびホスト・バス・アダプタＡ２を含む。ホストＢは、ホスト・バス・アダプタＢ１およびホスト・バス・アダプタＢ２を含む。ホストＣは、ホスト・バス・アダプタＣ１およびホスト・バス・アダプタＣ２を含む。例示のＨＢＡは、それぞれが８Ｇｂ／秒のリンクを提供し得る。

【００４７】

ディスク・アレイ３０２は、１または複数のポート３０４、１または複数の記憶プロセッサ３０６、および１または複数のＬＵＮを含む。図３の実例では、ディスク・アレイ３０２は、３つのＬＵＮを管理する１つの記憶プロセッサ３０６をそれぞれ有する２つのポート３０４を含む。１方の記憶プロセッサ３０６がＬＵＮ　１、３、および５を管理し、他方の記憶プロセッサ３０６がＬＵＮ　０、２、および４を管理する。例示のＬＵＮのそれぞれが、２５０ＭＢ／秒のリンクを提供し得る。

【００４８】

本開示の態様は、任意のＩ／Ｏパス記述に対して動作可能である。１例として、以下で式（１）に示されるように、ホストH_1とＬＵＮ　L_1の間にＩ／Ｏパスが定義され得る。

【００４９】

【数１】

$$\{H_1, HBA_i, Sport_{k_i}, \ldots Sport_{k_n}, Aport_j, LUN_1\} \quad (1)$$

【００５０】

この実例では、HBA_iはホスト５１４における各ＨＢＡのうちの１つを示し、$Sport_{k_n}$はネットワーク・スイッチ５１６のうちの１つにおけるポートを示し、$Aport_j$はディスク・アレイ３０２におけるポート３０４と記憶プロセッサ３０６の組合せを示す。しかし、ネットワーク・スイッチ５１６に関する構造の細部が未知の場合（例えば、マルチパス層がスイッチ・ポートを認識できない場合）、Ｉ／Ｏパスは、次の式（２）に示されるようにエンドポイントの観点から記述されてよい。

【００５１】

【数２】

$$\{H_1, HBA_i, Aport_j, LUN_1\} \quad (2)$$

【００５２】

各ホスト５１４からＬＵＮへの複数のパスがあると、ディスク・アレイ３０２の全体の有用性が増す。複数のＩ／Ｏパスは、ホスト５１４から、ＨＢＡ、ＳＡＮスイッチ（例えばスイッチ５１６）、または様々な要素を接続するケーブルを含むディスク・アレイ３０２までのあらゆる単一故障に対して防護する。例えば、連続したＩ／Ｏ要求または作業負荷に関するキャッシュ・ヒットが一時的飽和をもたらした後に、ＬＵＮ性能が、Ｉ／Ｏパスの高帯域幅によって制限されることがある。ホスト５１４は、応答して、またはＩ／Ｏパスのいずれかの要素が故障したとき、ホスト５１４にアクセスしているアプリケーションに対して劣化した性能または要素の故障をさらすのでなく、別の利用可能なパスに切り換える、すなわち転換する（ｄｉｖｅｒｔ）ことを選択してよい。

【００５３】

マルチパス化は、様々な方法で実施されてよい。例えば、アクティブ－アクティブのディスク・アレイは、任意のポート３０４上の単一のＬＵＮに対するＩ／Ｏ要求を同時に実行する。別の実例として、アクティブ－パッシブのディスク・アレイは、１つの記憶プロセッサ３０６（アクティブ・コントローラまたは１次コントローラ）の１または複数のポート３０４の単一のＬＵＮに対するＩ／Ｏ要求を実行するが、これらのＩ／Ｏ要求を、別

の記憶プロセッサ３０６（パッシブ・コントローラまたは２次コントローラ）のポート３０４のＬＵＮへフェイルオーバすることもできる。このようなフェイルオーバの遂行は、ＬＵＮ侵入（ｔｒｅｓｐａｓｓ）と称されることがある。このようなディスク・アレイは、Ｉ／Ｏ要求が任意の所与の時間に記憶プロセッサ３０６のうちの１つを通って流れていること確認するために、Ｉ／Ｏ要求を受け取ったポート３０４に基づきＬＵＮに対するフェイルオーバを起動する。

【００５４】

ディスク・アレイを実施するさらに別の実例には、疑似アクティブ－アクティブがある。このようなディスク・アレイでは、ＬＵＮは、性能が不均一な（例えば様々な程度またはレベルの性能の）複数の記憶プロセッサ３０６を通じてアクセスされてよい。複数の記憶プロセッサ３０６を通るパスは、初期の装置発見段階の間に（例えばオペレーティング・システムによって）発見されてよい。発見されたパスは、別々の装置として出現し得るが、パス管理層は、パス間の共通の装置エンドポイントを識別してよく、動的なパス選択を可能にする。

【００５５】

パス管理層は、パス選択の管理およびエラー処理の方針を実施する。例示の方針には、ラウンドロビン、最小限のキュー長さ、バランスのとれたパス、適応型パス選択、および単一アクティブパスが含まれる。ラウンドロビンの方針は、特にＩ／Ｏ要求によって要求されたデータがほぼ同一サイズであって、パスがＨＢＡ、ポート３０４、およびコントローラにわたって均一に配置されているとき、Ｉ／Ｏ要求をそれぞれのアクティブパスへ均一に配分する。

【００５６】

最小限のキュー長さの方針は、ホスト５１４を要求しているホスト・バス・アダプタにおいて顕著な要求の量が最少のアクティブパスへ、各Ｉ／Ｏ要求をパス設定することによって、局所的負荷バランシングを実施する。バランスのとれたパスの方針は、各Ｉ／Ｏ要求を、開始ブロック・アドレスに対応するパスに割り当てることができるように、装置のブロック・アドレス空間を、それぞれが各アクティブパスの量の領域に分割する。適応型パス選択の方針は、より高いスループットまたはより優れた性能を最近配送したパスに対してより高い優先順位を割り当てるように、サービス時間および最近のスループットに基づき周期的なパス優先順位を計算する。このような各方針は、様々なパスに対して（厳密な優先順位でなく）重みを割り当て、割り当てられた重みに比例したＩ／Ｏ要求を送る。このような各方針は、高度に動的な作業負荷または非常に様々な性能もしくはホップ・カウントを有するパスに対して動作可能である。単一アクティブパスの方針は、１つのパスを故障するまで用いて、次いでフェイルオーバパスに切り換える。

【００５７】

本開示の態様は、本明細書に説明されたものなど、既存のマルチパス技法に対して動作可能であるが、パス選択の集中的管理はさらに、Ｉ／Ｏパスにおける輻輳を検出して対応する。或る実施形態では、集中型のパス選択は、パス選択を、ネットワークの最大フロー解析へマッピングすることによって実施される。次に例示のネットワークの最大フロー解析を参照しながら説明するが、本開示の態様は、本明細書に説明されるように改良された任意の最大フロー解析に対して動作可能である。

【００５８】

ネットワークの最大フロー解析は、グラフＧ＝（Ｖ，Ｅ）上で動作する。ここで、Ｖは頂点の組であり、Ｅはエッジと各エッジ（ｕ，ｖ）に対する容量ｃ（ｕ，ｖ）の組である。例示のネットワークの最大フロー問題は、以下のようにエッジ（ｕ，ｖ）当たりのフローｆ（ｕ，ｖ）を計算することによって、起点ｓと宛先ｔから最大フローを見つけるものと説明され得る。

１．すべてのエッジに関してｆ（ｕ，ｖ）≦ｃ（ｕ，ｖ）

２．ノードｓおよびｔを除いて、あるノードに入るフローの和は、そのノードを出るフローの和に等しい。

３．起点ｓから宛先ｔまでの全フローは、可能な最大である。

【００５９】

本開示の態様は、フォード－フルカーソン（Ｆｏｒｄ－Ｆｕｌｋｅｒｓｏｎ）のアルゴリズムおよびエドモンド－カープ（Ｅｄｍｏｎｄ－Ｋａｒｐ）のアルゴリズムなどの、最大フロー問題を解決する、または最大フロー解析を遂行する様々な現存システムに対して動作可能である。或る実施形態では、集中型のパス選択のマッピングを可能にする複数起点、複数宛先の最大フロー問題の改良は、それぞれのホストH_iおよび$LUN\ L_j$に対して起点ノードH_iL_jを生成することと、各$LUN\ L_j$に対して宛先ノードL_jを生成することと、すべてのノードH_iL_jの前に単一の起点を付加し、エッジ上に無限の容量を有するすべてのL_jのノードに単一のシンクを付加することと、各ホスト５１４および各ＬＵＮに接続されたアレイ・ポート３０４においてホスト・バス・アダプタに対応する中間ノードを生成することとを含む。このようなマッピングの１例が、次に図４で示される。

【００６０】

図４は、ネットワーク・フロー解析に対する最適の記憶パス選択の例示のマッピングのブロック図である。この実例では、２つのホスト５１４（Ｈ１およびＨ２）がそれぞれ２つのＨＢＡを有し、ＨＢＡのそれぞれ（ＨＢＡ１、ＨＢＡ２、ＨＢＡ３、およびＨＢＡ４）が３つのＬＵＮ（ＬＵＮ１、ＬＵＮ２、およびＬＵＮ３）に接続されている。

【００６１】

初期のパス選択は、すべてのＬＵＮにわたってＩ／Ｏ負荷が均一であると想定し、各ＨＢＡおよび記憶プロセッサ・ポート３０４を通るパスのバランスをとることによって遂行される。動作においては、起点へのエッジ（例えば起点ノード）およびシンクへのエッジ（宛先ノード）を除いたすべてのエッジ（例えばリンク）に、１の容量が割り当てられる。すべてのH_iL_jノードがL_jノードへの非ゼロパスを有するかどうか判断するために、最大フロー・アルゴリズムが遂行される。見つかったそれぞれのパスに関して、パスのすべてのエッジによって１に等しい容量が消費される。すべてのノードにパスがあるわけではない場合、エッジ・ウェイトが１だけ増加され、すべてのノードがこのようなパスを有するようになるまで最大フロー・アルゴリズムが（例えば反復して）再実行される。すべてのノードがパスを有するようになった後、このような最大フロー・アルゴリズムの反復動作からもたらされる、すべてのホスト５１４からのすべてのＬＵＮ向けの初期パスは、各ＨＢＡおよびコントローラ・ポート３０４に対する最小ストレスを用いる。

【００６２】

次に、利用可能なパスの数または量ばかりでなく、各ホスト５１４によって各ＬＵＮのために消費される実際のＩ／Ｏ帯域幅を用いることにも基づくパス割当てを可能にするために、負荷ベースＩ／Ｏパス選択が遂行される。H_iL_jノードからＨＢＡノードへのエッジおよびポート・ノードからＬＵＮへのエッジのＩ／Ｏ帯域幅に関して、特定の量の容量Ｃが定義されている。例えば、エッジ当たりC_{min}およびC_{max}の２つの容量値が、それぞれ、ＨＢＡ当たりのホスト５１４ごとの平均的Ｉ／Ｏ帯域幅および最高の可能なネットワーク・スイッチ・リンク容量として初期化される。しかし、本開示の態様は、他の初期容量値を企図する。

【００６３】

すべてのH_iL_jノードがL_jノードへの非ゼロパスを有するかどうか判断するために、最大フロー・アルゴリズムが遂行される。パスが見つかるたびに、ホストH_iからL_jへの要求が、パスの容量から減じられる。すべてのパスを満たす最小限のＩ／Ｏ帯域幅または容量を見いだすために、バイナリ・サーチが遂行される。この最小限のＩ／Ｏ帯域幅に対応するパスは、選択されたパスすなわちホスト５１４に推薦されるパスを表す。

【００６４】

或る実施形態では、特定のパスが過剰に予約されることがある。リンクがその容量以上の要求に対応しないことを保証するために、ホスト・バス・アダプタ、ポート３０４、または中間リンクの帯域幅もしくは他の容量が既知であれば、このリンク制限を、パス選択

アルゴリズムにおけるリンク容量の上限として用いてもよい。最大のリンク容量値は、（例えばキャッシュに入った記憶位置に対してアイドル期間中に大きなＩ／Ｏ要求を用いて）リンク帯域幅を発見すること、および／またはネットワーク・スイッチ５１６に対して情報を問い合わせることによって、見いだされ得る。特定のホスト５１４からＬＵＮへの、リンク容量を超えないパスが見つからない場合には、トポロジ変更の提案および／またはリンク飽和の警報が、管理者５０２に送られてよい。図７を参照しながら以下でさらに説明されるように、トポロジ変更が開始されてよい。例えば、ＬＵＮ侵入、特定のＬＵＮからのデータ移動、およびＬＵＮからポートへの再構成マッピングが、本明細書の説明のように改良された最大フロー・アルゴリズムを繰り返し実行することによって評価されてよい。

【００６５】

図５は、ＳＡＮの複数の仮想マシンなどに対して集中型の入出力（Ｉ／Ｏ）記憶パスの選択および管理を実施するための例示のコンピュータ装置５０４のブロック図である。コンピュータ装置５０４は、コンピュータ装置５０４と関連した動作および機能を実施するための命令を（例えばアプリケーション・プログラム、オペレーティング・システムの機能または両方として）実行する任意の装置を表す。例えば、コンピュータ装置５０４は、本明細書に説明されたように、分散型リソース・スケジューリングを管理するための命令を実行する。コンピュータ装置５０４は、任意のコンピュータ装置または処理ユニットを含んでよい。例えば、コンピュータ装置５０４は、１群の処理ユニットまたはクラウド・コンピューティング構成など他のコンピュータ装置を表してよい。或る実施形態では、管理者５０２、またはユーザ１０８などの他のユーザは、コンピュータ装置５０４の動作の態様を管理するためにコンピュータ装置５０４と対話する。

【００６６】

コンピュータ装置５０４は、少なくとも１つのプロセッサ５０６および記憶エリア５０８を有する。プロセッサ５０６は、何らかの量の処理ユニットを含み、本開示の態様を実施するためのコンピュータ実行可能命令を実行するようにプログラムされる。これらの命令は、プロセッサ５０６もしくはコンピュータ装置５０４の内部で動作する複数のプロセッサによって遂行されてよく、またはコンピュータ装置５０４の外部のプロセッサによって遂行されてもよい。或る実施形態では、プロセッサ５０６は、分散型共有記憶システムの図に示されたものなどの命令を実行するようにプログラムされる。

【００６７】

記憶エリア５０８は、コンピュータ装置５０４に関連した、またはコンピュータ装置５０４がアクセス可能な、何らかの量のコンピュータ可読媒体を含む。記憶エリア５０８またはその一部分は、コンピュータ装置５０４の内部、コンピュータ装置５０４の外部、またはその両方にあってよい。

【００６８】

図５の実例では、記憶エリア５０８は、ＳＡＮ　５１２のトポロジ記述５１０を記憶する。トポロジ記述５１０は、ホスト５１４、スイッチ５１６などＳＡＮ　５１２の要素、およびディスク・アレイ３０２、ならびにその構成（例えばＨＢＡ、ポート３０４、記憶プロセッサ３０６など）およびその間のリンクを識別する。或る実施形態では、トポロジ記述５１０はさらに、ＳＡＮ　５１２に関する閾値容量を含む。閾値容量は、例えばＨＢＡ当たりのホスト５１４ごとのＩ／Ｏ帯域幅を表すが、他の閾値容量も本開示の範囲内である。記憶エリア５０８は、リンクと関連したリンク容量も記憶してよい。

【００６９】

或る実施形態では、コンピュータ装置５０４は、ネットワーク（図示せず）を介してＳＡＮ　５１２の要素にアクセスする。ネットワークは、コンピュータ装置５０４とＳＡＮ　５１２の要素の間を連絡するための任意の手段を表す。本開示の態様は、任意のネットワークのタイプまたは構成に対して動作可能である。

【００７０】

図６は、ＳＡＮ　５１２などのＳＡＮの中の最適の記憶パスを選択するように、複数の

ホスト５１４を管理する集中型システムによって遂行される例示の方法の流れ図である。方法６００は、コンピュータ装置５０４（図５に示されている）による実行を参照しながら説明されているが、任意のコンピュータ装置によって遂行され得るように企図されている。さらに、図６に示された動作の実行は、ＶＭ環境に限定されることなく、任意の複数起点および複数宛先の環境に適用可能である。また、コンピュータ実行可能命令を記憶する１または複数のコンピュータ可読記憶媒体が動作することにより、プロセッサ５０６は、図６に示された動作を遂行することによって集中型のＩ／Ｏ負荷ベースパス選択を実施してよい。

【００７１】

ステップ６０４で、複数のホスト５１４のうちの１つから、同ホストが認識可能な少なくとも１つの論理的記憶装置にそれぞれ至る複数の記憶パスを識別するように、複数のホスト５１４に対して初期のパス選択が遂行される（図４を参照されたい）。初期のパス選択は、新規のトポロジに対して、またはＩ／Ｏ負荷の変更に際して遂行されるが、或る実施形態は、パスを選択するのでなく、１組の選択されたパスにアクセスする。

【００７２】

一般に、初期のパス選択は、各繰返しに対してエッジ・ウェイトを増す最大フロー・アルゴリズムを繰り返し遂行することによって、記憶パスを繰り返し検索することを含む。各繰返しで、ホスト－ＬＵＮ対の何らかの量が発見されてよく、または発見されないこともあり、ホスト－ＬＵＮ対がすべて見つかるまでエッジ・ウェイトが増加される。例えば、ステップ６０２で、各リンクに対して１のエッジ・ウェイトを定義することによって初期のパス選択が開始する。ステップ６０６で各ホスト５１４用に記憶パスが見つからなければ、ステップ６０７でエッジ・ウェイトが増加され、ステップ６０４でパス選択が遂行される。このように、各ホスト５１４からそれぞれの接続されたＬＵＮ（例えば記憶装置）へのパスが見つかるまでパス選択を繰り返して、初期のパス選択を完了する。

【００７３】

或る実施形態では、例えばリンクの過剰予約を防止するために、任意選択で、１または複数のリンクに対する上限リンク容量が定義されてよい。或る実施形態では、ネットワーク要素にリンク容量を問い合わせてよく、またはテストＩ／Ｏ要求によってリンク容量を発見することができる。例えば、記憶パスを選択するのに使用されるスイッチ５１６の最大容量を得るために、記憶パス内のスイッチ５１６に問い合わせてよい。

【００７４】

ステップ６０８では、見つかった各記憶パスについて、記憶パスに対するＩ／Ｏ負荷が求められる。ステップ６１０では、各記憶パスに対して求めたＩ／Ｏ負荷を閾値容量から減じる。ステップ６１４で、例えばバイナリ・サーチを遂行して、バイナリ・サーチからの各値に対して最大フロー・アルゴリズムを実行することによって、記憶パスのそれぞれに適合する、そうでなければ各ホスト５１４からそれぞれの接続されたＬＵＮへの少なくとも１つのパスを見つけるのに必要とされる、最小限の帯域幅が見いだされる。

【００７５】

ステップ６１６で、ホスト５１４のそれぞれに対して、記憶パスのうちの１つが最小限の帯域幅に基づき選択される。各ホスト５１４向けに選択された記憶パスが、少なくとも対応するホスト５１４に対して、表示される、識別される、推奨される、または示される。

【００７６】

図６に示されたパス選択動作は、ＳＡＮ　５１２のトポロジの変更またはパスに沿ったＩ／Ｏ負荷の変更に応答して繰り返される。例えば、ステップ６１８で、トポロジおよびＩ／Ｏの負荷が監視されてよい。ステップ６２０でトポロジ変更が検出された場合、または選択された記憶パスに沿ったＩ／Ｏの変更がステップ６２２で検出された場合、これに応答して、初期のパス選択が再度遂行されてよい。例えば、記憶パスのうちの１つに関して監視されたＩ／Ｏ負荷の変更が検出されるとすぐに、更新されたＩ／Ｏ負荷が求められ、記憶パスのそれぞれに適合する最小限の帯域幅が再度計算され、各ホスト５１４に対し

て、記憶パスのうちの１つが、再計算された最小の帯域幅に基づき選択される。

【００７７】

図７を参照しながら次に説明されるように、トポロジ変更が推奨されてよく、或る実施形態では自動的に開始される。

図７は、リンクまたは論理装置番号（ＬＵＮ）の輻輳に応答してトポロジ変更を開始するように、複数のホスト５１４を管理する集中型システムによって遂行される例示の方法の流れ図である。方法７００は、コンピュータ装置５０４（図５に示されている）による実行を参照しながら説明されているが、任意のコンピュータ装置によって遂行され得るように企図されている。さらに図７に示された動作の実行は、ＶＭ環境に限定されることなく、トポロジ変更が開始され得る任意の複数起点および複数宛先の環境に適用可能である。またコンピュータ実行可能命令を記憶する１または複数のコンピュータ可読記憶媒体が動作することにより、プロセッサ５０６は、図７に示された動作を遂行することによって集中型のＩ／Ｏ負荷ベースパス選択を実施してよい。

【００７８】

ホスト５１４（例えば仮想マシンを実行する）のうちの１つから、ＬＵＮまたはホスト５１４から認識可能な他の論理的記憶装置のうちの１つにそれぞれ至る複数の記憶パスが、識別される、またはアクセスされる。ステップ７０４で、記憶パスのそれぞれに対するＩ／Ｏ負荷が監視される。ステップ７０６で、記憶パスの少なくとも１つにおいて監視されたＩ／Ｏ負荷が輻輳を示していると判断すると、ステップ７０８で、１または複数のトポロジ変更が識別される。例示のトポロジ変更は、それだけではないが、１つの論理的記憶装置から別の論理的記憶装置へデータを移動することと、複数のホスト５１４の第１のホストから複数のホスト５１４の第２のホストへデータストアを移動することと、論理的記憶装置をコントローラへ再マッピングすることとを含む。

【００７９】

コンピュータ装置５０４がＳＡＮ　５１２の中のエンドポイントを再構成することができる実施形態では、ステップ７１０で、識別された１または複数のトポロジ変更は、自動的に、かつ／または管理者５０２の入力なしで開始されてよい。例えば、記憶装置を別の記憶プロセッサ３０６またはコントローラに再マッピングするために、記憶装置を管理する記憶プロセッサ３０６のうちの１つに命令が送られてよい。或る実施形態では、トポロジ変更を開始することはさらに、記憶パスを再構成することと、ホスト５１４のうち少なくとも１つに対して別の記憶パスを用いるように指示することとを含んでよい。

追加の実例

本開示の態様が、１つのＬＵＮ当たり１つのパスを用いるアクティブ－パッシブ・モードでアクセスされているＬＵＮを参照しながら本明細書で説明されてきたが、他の実施形態は、１つのＬＵＮ当たり複数のパスを用いるアクティブ－アクティブ・モードでＬＵＮがアクセスされる環境を企図する。このような実施形態では、最大フロー・アルゴリズムが、ｋ個のパスが許容されるＬＵＮに対してｋ個のパスを見つけるように変更される。パスが見つかるたびに、要求は、１ではなくｋだけインクリメントされる。

例示の動作環境

本明細書で説明された動作は、コンピュータ装置５０４などのコンピュータまたはコンピュータ装置によって遂行されてよい。コンピュータ装置は、メッセージおよび／または記憶したデータの交換を通じて互いに通信する。通信は、任意の有線または無線の接続にわたって、任意のプロトコルまたは機構を用いて起こり得る。コンピュータ装置は、（例えばネットワークおよび／またはデータバスの全体への）ブロードキャスト・メッセージ、（例えば複数の他のコンピュータ装置にアドレス指定された）マルチキャスト・メッセージ、および／または複数のユニキャスト・メッセージとしてメッセージを伝送してよく、これらのそれぞれが、個々のコンピュータ装置にアドレス指定される。さらに、或る実施形態では、メッセージは、ユーザ・データグラム・プロトコル（ＵＤＰ：Ｕｓｅｒ　Ｄａｔａｇｒａｍ　Ｐｒｏｔｏｃｏｌ）などの配信を保証しないネットワーク・プロトコルを用いて伝送される。したがってコンピュータ装置は、メッセージを伝送するとき、配信

不能のリスクを低減することができるように、メッセージの複数のコピーを伝送してよい。

【００８０】

例示のコンピュータ読取り可能媒体は、フラッシュ・メモリ・ドライブ、デジタル多用途ディスク（ＤＶＤ）、コンパクト・ディスク（ＣＤ）、フロッピー（登録商標）・ディスク、およびテープ・カセットを含む。限定的でない実例として、コンピュータ読取り可能媒体は、コンピュータ記憶媒体および通信媒体を含む。コンピュータ記憶媒体には、コンピュータ読取り可能命令、データ構造、プログラム・モジュールまたは他のデータなどの情報を記憶するための任意の方法または技術で実施された、揮発性媒体および不揮発性媒体、取外し可能媒体および固定型媒体が、含まれる。コンピュータ記憶媒体は、有形であり、伝搬されたデータ信号を除外し、通信媒体とは相互排他的である。或る実施形態では、コンピュータ記憶媒体はハードウェアで実施される。例示のコンピュータ記憶媒体は、ハード・ディスク、フラッシュ・ドライブ、および他の固体記憶装置を含む。それと対照的に、通信媒体は、一般に、コンピュータ読取り可能命令、データ構造、プログラム・モジュール、または搬送波もしくは他の移送機構などの変調データ信号における他のデータを具体化し、任意の情報配送媒体を含む。

【００８１】

本開示の実施形態は、例示のコンピュータ・システム環境と関連して説明されてきたが、多数の他の汎用または専用のコンピュータ・システムの環境または構成で動作可能である。本開示の態様とともに用いるのに適切であり得る周知のコンピュータ・システムの実例、環境、および／または構成には、それだけではないが、モバイル・コンピューティング装置、パーソナル・コンピュータ、サーバ・コンピュータ、ハンドヘルドまたはラップトップの装置、マルチプロセッサ・システム、ゲーム機、マイクロプロセッサ・ベースのシステム、セット・トップ・ボックス、プログラム可能な家電、携帯電話、ネットワークＰＣ、ミニコンピュータ、メインフレーム・コンピュータ、上記システムのうち任意ものを含む分散型コンピュータ環境または装置などが含まれる。

【００８２】

本開示の実施形態は、１または複数のコンピュータまたは他の装置によって実行されるプログラム・モジュールなどのコンピュータ実行可能命令の一般的な状況において説明されてよい。コンピュータ実行可能命令は、１または複数のコンピュータ実行可能な構成要素またはモジュールに体系化され得る。一般に、プログラム・モジュールは、それだけではないが、特定のタスクを遂行する、または特定の抽象データ型を実施するルーチン、プログラム、オブジェクト、構成要素、およびデータ構造を含む。本開示の態様は、このような構成要素またはモジュールの任意の数および構成を用いて実施され得る。例えば、本開示の態様は、特定のコンピュータ実行可能命令または各図面に示されて本明細書で説明された特定の構成要素もしくはモジュールに限定されない。本開示の他の実施形態は、本明細書に示され説明されたものよりも、程度の差はあるが機能性を有する様々なコンピュータ実行可能命令または構成要素を含んでよい。

【００８３】

本開示の態様は、本明細書に説明された命令を実行するようにプログラムされたとき、汎用コンピュータを専用コンピュータ装置に変換する。

本明細書で図示して説明した実施形態ならびに本明細書で具体的には説明されていないが本発明の態様の範囲内の実施形態は、記憶パスの輻輳を修正するためにトポロジ変更を識別する例示の手段と、記憶パスの輻輳を修正するためにトポロジ変更を自動的に開始する例示の手段とを構成する。例えば、本明細書で説明されたような動作を実行するようにプログラムされたプロセッサ５０６は、これらの例示の手段を構成する。

【００８４】

図に示された様々な要素の機能の少なくとも一部分は、図中の他の要素、また図には示されていないエンティティ（例えばプロセッサ、ウェブ・サービス、サーバ、アプリケーション・プログラム、コンピュータ装置など）によって遂行されてよい。

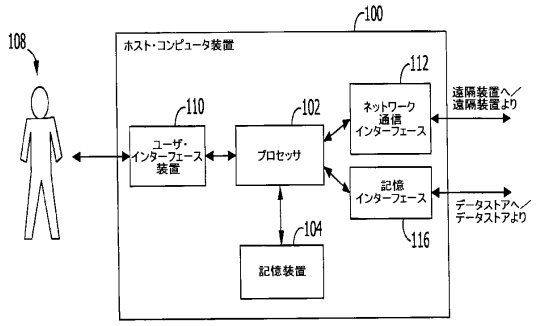
【００８５】
或る実施形態では、図に示された動作は、コンピュータ可読媒体に符号化されたソフトウェア命令として、この動作を遂行するようにプログラムされた、もしく設計されたハードウェアで、またはその両方で実施されてよい。例えば、本開示の態様は、チップまたは複数の相互接続された導電素子を含んでいる他の回路上のシステムとして実施されてよい。

【００８６】
本明細書で図示して説明した本開示の実施形態の動作の実行または遂行の順序は、別段の定めがない限り必須ではない。すなわち、これらの動作は、別段の定めがない限り任意の順序で遂行されてよく、本開示の実施形態は、本明細書で開示されたものに対して追加の動作を含んでよく、より少ない動作しか含まなくてもよい。例えば、特定の動作を、別の動作の以前に、同動作と同時に、または同動作の後に、実行する、または遂行するということは、本開示の態様の範囲に入るように企図されている。

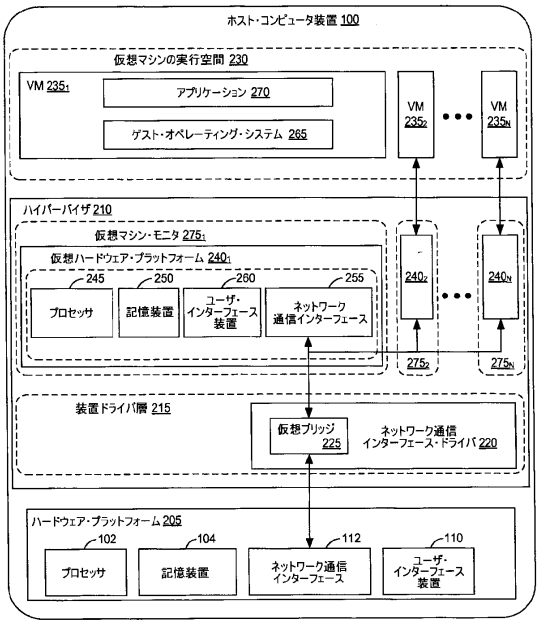
【００８７】
本開示の態様の要素またはその実施形態を紹介するとき、冠詞「１つの（ａ）」、「１つの（ａｎ）」、「この（ｔｈｅ）」、および「前記（ｓａｉｄ）」は、同要素の１または複数があることを意味するように意図されている。用語「備える」、「含む」、および「有する」は、包含的であるように意図されており、列挙された要素とは別に追加の要素が存在し得ることを意味する。

【００８８】
本開示の態様を詳細に説明してきたが、添付の特許請求の範囲で定義される本開示の態様の範囲から逸脱することなく、修正形態および変形形態が可能であることが明らかであろう。本開示の態様の範囲から逸脱することなく、上記の構成、結果、および方法に様々な変更を加えることができるので、上記の記述の中に含まれ、添付図面に示されていることは、すべて例示であって限定する意味ではないと解釈されるべきであることが意図されている。

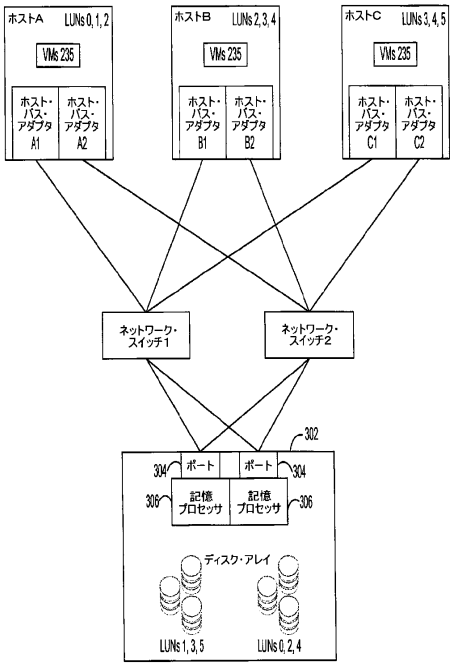
【図１】


100
108
ホスト・コンピュータ装置
112
110
102
ネットワーク通信インターフェース
遠隔装置へ／遠隔装置より
ユーザ・インターフェース装置
プロセッサ
記憶インターフェース
データストアへ／データストアより
104
116
記憶装置

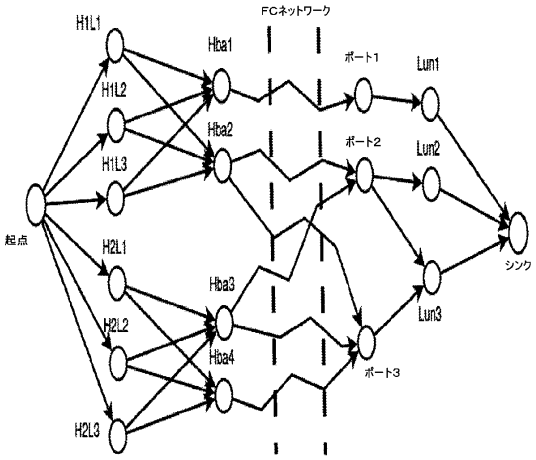
【図２】


ホスト・コンピュータ装置 100
仮想マシンの実行空間 230
VM 235₁
アプリケーション 270
ゲスト・オペレーティング・システム 265
VM 235₂
VM 235N
ハイパーバイザ 210
仮想マシン・モニタ 275₁
仮想ハードウェア・プラットフォーム 240₁
245
250
260
255
プロセッサ
記憶装置
ユーザ・インターフェース装置
ネットワーク通信インターフェース
240₂
240N
275₂
275N
装置ドライバ層 215
仮想ブリッジ 225
ネットワーク通信インターフェース・ドライバ 220
ハードウェア・プラットフォーム 205
102
104
112
110
プロセッサ
記憶装置
ネットワーク通信インターフェース
ユーザ・インターフェース装置

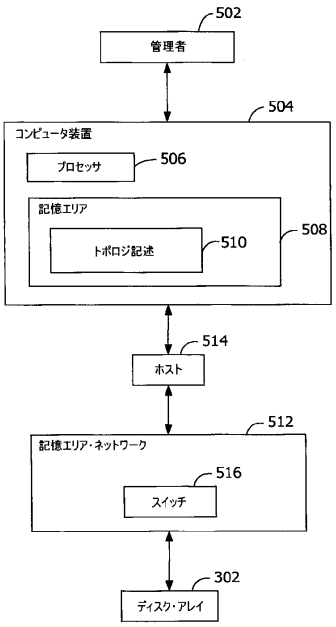
【図３】


ホストA LUNs 0, 1, 2
ホストB LUNs 2, 3, 4
ホストC LUNs 3, 4, 5
VMs 235
ホスト・バス・アダプタ A1
ホスト・バス・アダプタ A2
ホスト・バス・アダプタ B1
ホスト・バス・アダプタ B2
ホスト・バス・アダプタ C1
ホスト・バス・アダプタ C2
ネットワーク・スイッチ1
ネットワーク・スイッチ2
302
304
ポート
ポート
304
306
記憶プロセッサ
記憶プロセッサ
306
ディスク・アレイ
LUNs 1, 3, 5
LUNs 0, 2, 4

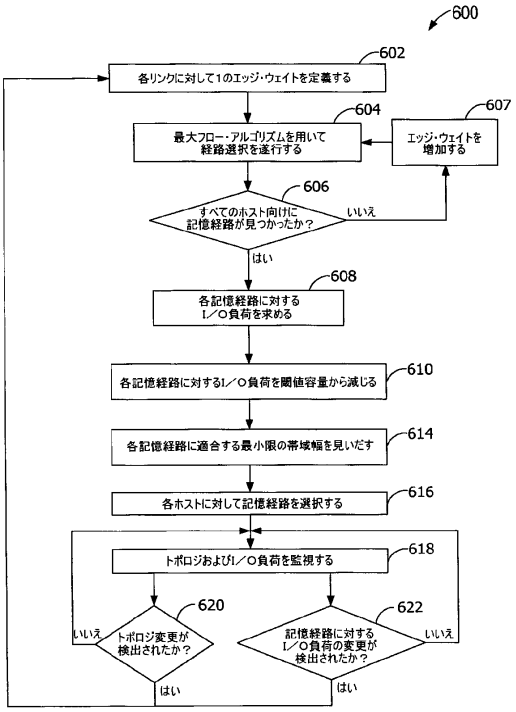
【図４】


FCネットワーク
H1L1
H1L2
H1L3
H2L1
H2L2
H2L3
起点
Hba1
Hba2
Hba3
Hba4
ポート1
ポート2
ポート3
Lun1
Lun2
Lun3
シンク

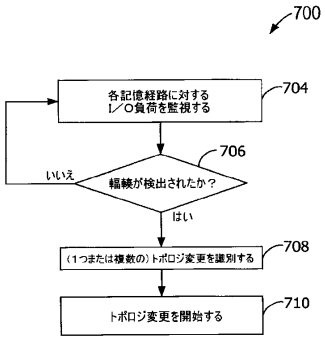
【図５】


502
管理者
504
コンピュータ装置
506
プロセッサ
記憶エリア
510
トポロジ記述
508
514
ホスト
512
記憶エリア・ネットワーク
516
スイッチ
302
ディスク・アレイ

【図６】


600
602
各リンクに対して1のエッジ・ウェイトを定義する
604
最大フロー・アルゴリズムを用いて経路選択を遂行する
607
エッジ・ウェイトを増加する
606
すべてのホスト向けに記憶経路が見つかったか？
いいえ
はい
608
各記憶経路に対するI／O負荷を求める
610
各記憶経路に対するI／O負荷を閾値容量から減じる
614
各記憶経路に適合する最小限の帯域幅を見いだす
616
各ホストに対して記憶経路を選択する
618
トポロジおよびI／O負荷を監視する
620
トポロジ変更が検出されたか？
いいえ
はい
622
記憶経路に対するI／O負荷の変更が検出されたか？
いいえ
はい

【図７】


700
704
各記憶経路に対するI／O負荷を監視する
706
いいえ
輻輳が検出されたか？
はい
708
（1つまたは複数の）トポロジ変更を識別する
710
トポロジ変更を開始する

フロントページの続き

(72)発明者　クリシュナ　ラージ　ラージャ
アメリカ合衆国　９５０５１　カリフォルニア州　サンタ　クララ　ギャンブリン　ドライブ　２６５８
(72)発明者　アージェイ　グラティ
アメリカ合衆国　９４３０４　カリフォルニア州　パロ　アルト　ヒルビュー　アベニュー　３４０１

審査官　田上　隆一

(56)参考文献　米国特許出願公開第２００９／０１７２６６６（ＵＳ，Ａ１）
米国特許出願公開第２００８／０２５０１７８（ＵＳ，Ａ１）
特開２００７－０９４６８１（ＪＰ，Ａ）

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