(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5315209号
(P5315209)
(45)発行日　平成25年10月16日(2013.10.16)　(24)登録日　平成25年7月12日(2013.7.12)

(51)Int.Cl.　F I
G06F 13/10 (2006.01)　G06F 13/10 330C

請求項の数 9　(全 28 頁)

(21)出願番号　特願2009-246376(P2009-246376)
(22)出願日　平成21年10月27日(2009.10.27)
(65)公開番号　特開2010-140471(P2010-140471A)
(43)公開日　平成22年6月24日(2010.6.24)
審査請求日　平成24年5月11日(2012.5.11)
(31)優先権主張番号　12/332957
(32)優先日　平成20年12月11日(2008.12.11)
(33)優先権主張国　米国(US)

(73)特許権者　390009531
インターナショナル・ビジネス・マシーンズ・コーポレーション
ＩＮＴＥＲＮＡＴＩＯＮＡＬ　ＢＵＳＩＮＥＳＳ　ＭＡＣＨＩＮＥＳ　ＣＯＲＰＯＲＡＴＩＯＮ
アメリカ合衆国１０５０４　ニューヨーク州　アーモンク　ニュー　オーチャード　ロード
(74)代理人　100108501
弁理士　上野　剛史
(74)代理人　100112690
弁理士　太佐　種一
(74)代理人　100091568
弁理士　市位　嘉宏

最終頁に続く

(54)【発明の名称】冗長構成を生成するための周辺機器相互接続入出力仮想化デバイスの使用

(57)【特許請求の範囲】
【請求項１】
冗長システム構成を生成するためのデータ処理システム実施方法であって、
各仮想機能について仮想機能パス許可テーブルを生成するステップであって、前記仮想機能パス許可テーブルは、当該仮想機能パス許可テーブルに関連付けられた仮想機能が複数のシステムにおける１組のアドレス範囲へアクセスする許可を得ているかを定義するエントリを複数有し、前記複数のエントリそれぞれが、無効な機能間アクセスを防止する境界をさらに定義し、前記仮想機能が、シングルルートまたはマルチルート周辺機器相互接続デバイスによって実行される、前記生成するステップと、
データをディスクに書き込むためのディスク書き込み動作のリクエスタからのディスク書き込み要求を受信して、要求されたデータを前記仮想機能から提供するステップと、
前記１組のアドレス範囲のうちの選択された第１のアドレス範囲内に第１の受信バッファを生成し、且つ、前記１組のアドレス範囲のうちの選択された第２のアドレス範囲内に第２の受信バッファを生成するステップであって、前記選択された第２のアドレスは、前記選択された第１のアドレスを有するシステムと異なるシステム上にある、前記生成するステップと、
前記仮想機能のための仮想機能作業待ち行列エントリを生成するステップであって、前記仮想機能作業待ち行列エントリは、前記第１の受信バッファ及び前記第２の受信バッファの各アドレスを含む、前記生成するステップと、
前記仮想機能パス許可テーブルにおいて、前記各選択されたアドレス範囲の使用を前記

仮想機能が許可されているかどうかを判定するステップと、
前記仮想機能は許可されていることに応答して、前記要求されたデータを、前記第１の受信バッファ及び前記第２の受信バッファに書き込むステップと、
前記書き込みに応答して、前記リクエスタに完了通知を発行するステップと
を含む、前記方法。
【請求項２】
前記仮想機能パス許可テーブル内にエントリが存在していることが、１次パスおよび対応する２次パスによる前記１つまたは複数のシステム内のシステムへの機能のアクセスを許可することである、請求項１に記載の方法。
【請求項３】
前記仮想機能パス許可テーブル内にエントリが存在していないことが、１次パスおよび対応する２次パスによる前記１つまたは複数のシステム内のシステムへの機能のアクセスを許可しないことである、請求項１に記載の方法。
【請求項４】
前記仮想機能パス許可テーブルが、各仮想機能と１組の仮想階層の間の対応、又は、各仮想機能と１組の論理パーティション内の１組のアドレス範囲の間の対応を含む、請求項１～３のいずれか一項に記載の方法。
【請求項５】
前記複数のエントリそれぞれが、前記仮想機能と前記複数のシステムのうちの一つのシステムにおけるアドレス範囲との間の１次通信パスと当該１次通信パスの代替通信パスとを有する、請求項１～４のいずれか一項に記載の方法。
【請求項６】
前記１次通信パスが優先パスであり、当該優先通信パスが利用不可能であることに応答して、前記代替通信パスの１つを使用する、請求項１～５のいずれか一項に記載の方法。
【請求項７】
冗長システム構成を生成するためのデータ処理システムであって、
バスと、
前記バスに接続されたコンピュータ実行可能命令を含むメモリと、
前記コンピュータ実行可能命令を実行する中央プロセッサ・ユニットであって、前記コンピュータ実行可能命令が、請求項１～６のいずれか一項に記載の方法の各ステップを実行するように前記データ処理システムに指令する、前記中央プロセッサ・ユニットと、
を備えている、前記データ処理システム。
【請求項８】
冗長システム構成を生成するためのデータ処理システムであって、
バスと、
前記バスに接続されたコンピュータ実行可能命令を含むメモリと、
前記コンピュータ実行可能命令を実行する中央プロセッサ・ユニットと
を備えており、
前記メモリが、
各仮想機能について仮想機能パス許可テーブルであって、前記仮想機能パス許可テーブルは、当該仮想機能パス許可テーブルに関連付けられた仮想機能が複数のシステムにおける１組のアドレス範囲へアクセスする許可を得ているかを定義するエントリを複数有し、前記複数のエントリそれぞれが、無効な機能間アクセスを防止する境界をさらに定義し、前記仮想機能が、シングルルートまたはマルチルート周辺機器相互接続デバイスによって実行される、前記仮想機能パス許可テーブルと、
前記仮想機能のための仮想機能作業待ち行列エントリであって、前記仮想機能作業待ち行列エントリは、前記１組のアドレス範囲のうちの選択された第１のアドレス範囲内に生成された第１の受信バッファ及び前記１組のアドレス範囲のうちの選択された第２のアドレス範囲内に生成された第２の受信バッファの各アドレスを含み、前記選択された第２のアドレスは、前記選択された第１のアドレスを有するシステムと異なるシステム上にあ

る、前記仮想機能作業待ち行列エントリと、
を記憶し、
前記コンピュータ実行可能命令が、
データをディスクに書き込むためのディスク書き込み動作のリクエスタからのディスク書き込み要求を受信して、要求されたデータを前記仮想機能から提供することと、
前記仮想機能パス許可テーブルにおいて、前記各選択されたアドレス範囲の使用を前記仮想機能が許可されているかどうかを判定することと、
前記仮想機能は許可されていることに応答して、前記要求されたデータを、前記第１の受信バッファ及び前記第２の受信バッファに書き込むことと、
前記書き込みに応答して、前記リクエスタに完了通知を発行すること
を実行するように前記データ処理システムに指令する、
前記データ処理システム。
【請求項９】
冗長システム構成を生成するためのコンピュータ・プログラムであって、
コンピュータに、請求項１～６のいずれか一項に記載の方法の各ステップを実行させる、前記コンピュータ・プログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、一般に、改良データ処理システムに関し、より具体的には、周辺機器相互接続入出力仮想化構成（peripheral component interconnect input/output virtualizationconfiguration）を使用して冗長構成を生成するための、コンピュータ実施方法、データ処理システム、およびコンピュータ・プログラム製品に関する。
【背景技術】
【０００２】
典型的なコンピューティング・デバイスは、元々は１９９０年代にインテル・コーポレーションによって作成され、現在はＰＣＩ－ＳＩＧによって管理される、周辺機器相互接続（ＰＣＩ：Peripheral Component Interconnect）規格のあるバージョンまたは実施を利用する、入出力（Ｉ／Ｏ）アダプタおよびバスを使用する。周辺機器相互接続（ＰＣＩ）規格は、周辺デバイスをコンピュータ・マザーボードに接続するためのコンピュータ・バスを規定する。ＰＣＩエクスプレス（PCIExpress）すなわちＰＣＩｅは、既存のＰＣＩプログラミング概念を使用するＰＣＩコンピュータ・バスの実施であるが、コンピュータ・バスは、まったく異なるはるかに高速なシリアル物理レイヤ通信プロトコルに基づいている。物理レイヤは、複数のデバイス間で共用できる双方向バスから成るのではなく、正確に２つのデバイスに接続される単一の一方向リンクから成る。
【０００３】
図１を参照すると、周辺機器相互接続エクスプレス仕様に従った周辺機器相互接続エクスプレス（ＰＣＩｅ：peripheral component interconnect express）バスを含むシステムを示す例示的な図が提示されている。図１に示された具体的なシステムは、複数のサーバ・ブレード（serverblade）１０１～１０４が提供されるブレード・エンクロージャ（blade enclosure）である。サーバ・ブレードは、高密度システム用に設計された内蔵型のコンピュータ・サーバである。サーバ・ブレードは、スペース、電力、および他の考慮点のために多くのコンポーネントが除去されているが、それでも、コンピュータと見なされるためのすべての機能コンポーネントを有している。ブレード・エンクロージャ１００は、電力、冷却、ネットワーキング、様々な相互接続、およびブレード・エンクロージャ１００内の様々なサーバ・ブレード１０１～１０４の管理などのサービスを提供する。サーバ・ブレード１０１～１０４およびブレード・エンクロージャ１００は、一緒になってブレード・システムを形成する。
【０００４】
図１に示されるように、周辺機器相互接続エクスプレスが、サーバ・ブレード１０１～

１０４の各々に実装され、周辺機器相互接続エクスプレス・デバイス１０５～１１２の１つに接続するために使用される。次にこれらのサーバ・ブレード１０１～１０４の各々が、ブレード・エンクロージャ１００内のスロットに差し込まれ、次にブレード・エンクロージャが、周辺機器相互接続エクスプレス・イーサネット・デバイス１０５、１０７、１０９、１１１の出力を、ブレード・エンクロージャ１００内のバックプレーン（backplane）を介して、イーサネット・スイッチ１１３に接続し、次にイーサネット・スイッチが、例えば、ブレード・エンクロージャ１００の外部のデバイスへの通信接続などの、外部接続用のイーサネット接続１１５を生成する。同様に、周辺機器相互接続エクスプレス・ストレージ・デバイス１０６、１０８、１１０、１１２の各々は、ブレード・エンクロージャ１００内のバックプレーンを介して、ストレージ・エリア・ネットワーク・スイッチ１１４に接続され、次にストレージ・エリア・ネットワーク・スイッチが、外部接続用のストレージ・エリア・ネットワーク接続１１６を生成する。

【０００５】

したがって、図１に示されるシステムは、周辺機器相互接続仕様または周辺機器相互接続エクスプレス仕様あるいはその両方が実施された、データ処理システムの例示的な１つのタイプである。周辺機器相互接続仕様または周辺機器相互接続エクスプレス仕様あるいはその両方を使用するデータ処理システムの他の構成も知られている。これらのシステムは、アーキテクチャが多様であり、したがって、その各々を本明細書で詳しく説明することはできない。周辺機器相互接続および周辺機器相互接続エクスプレスに関するより多くの情報については、周辺機器相互接続スペシャル・インタレスト・グループ（ＰＣＩ－ＳＩＧ：peripheral component interconnect special interest group）のウェブサイトwww.pcisig.comから入手可能な周辺機器相互接続仕様および周辺機器相互接続エクスプレス仕様を参照されたい。

【０００６】

周辺機器相互接続仕様および周辺機器相互接続エクスプレス仕様に加えて、周辺機器相互接続スペシャル・インタレスト・グループは、いくつかの論理パーティション（ＬＰＡＲ：logical partition）によって共用できる入出力アダプタ（ＩＯＡ：input/output adapter）をいかに設計するかについて規定するための、入出力仮想化（ＩＯＶ：input/outputvirtualization）規格も規定している。論理パーティションは、コンピュータのプロセッサ、メモリ、およびストレージを、リソースの多数の組に分割して、リソースの各組を、独自のオペレーティング・システム・インスタンスおよびアプリケーションとともに独立に動作させ得るようにする。生成できる論理パーティションの数は、システムのプロセッサ・モデルおよび利用可能なリソースに依存する。一般に、パーティションは、データベース動作、クライアント／サーバ動作、またはテスト環境と生産環境の分離など、様々な目的で使用される。各パーティションは、他のパーティションがあたかも別個のマシンであるかのように、他のパーティションと通信することができる。

【０００７】

論理パーティションをサポートする現代のシステムでは、いくつかのリソースは、論理パーティション間で共用することができる。上で言及されたように、周辺機器相互接続仕様および周辺機器相互接続エクスプレス仕様では、共用できるそのようなリソースの１つは、入出力仮想化メカニズムを使用する入出力アダプタである。

【０００８】

さらに、周辺機器相互接続スペシャル・インタレスト・グループは、多数のシステム間で入出力アダプタを共用するための入出力仮想化（ＩＯＶ）規格も規定している。この機能は、マルチルート（ＭＲ：multi-root）入出力仮想化と呼ばれる。

【０００９】

図２を参照すると、周辺機器相互接続エクスプレス・マルチルート入出力仮想化を含むシステムを示す例示的な図が提示されている。具体的には、図２は、多数のシステム間で周辺機器相互接続エクスプレス・デバイスを共用するために、図１に示されたアーキテクチャをどのように変更できるかを示している。

【００１０】
今度はサーバ・ブレード２０１～２０４は、図１のサーバ・ブレード１０１～１０４で行われたように、サーバ・ブレード２０１～２０４上に周辺機器相互接続エクスプレス・デバイス自体を含む代わりに、周辺機器相互接続エクスプレス・ルート・ポート（root port）２０５～２１２を生成し、ブレード・エンクロージャ２００のバックプレーンを横断する周辺機器相互接続エクスプレス接続を駆動する。その場合、各サーバ・ブレード２０１～２０４からの周辺機器相互接続エクスプレス・リンクは、マルチルート周辺機器相互接続エクスプレス・スイッチ２１３～２１４の一方に接続され、次にマルチルート周辺機器相互接続エクスプレス・スイッチは、周辺機器相互接続エクスプレス・イーサネット／ストレージ・デバイス２１７～２２０に接続される。周辺機器相互接続エクスプレス・イーサネット／ストレージ・デバイス２１７～２２０は、外部接続２１５、２１６を介して、外部のイーサネット・デバイスおよびストレージ・デバイスに接続する。そのようにして、周辺機器相互接続エクスプレス・デバイスを、ブレード・エンクロージャ２００内で使用することができる。これは、サーバ・ブレード２０１～２０４間で周辺機器相互接続エクスプレス・デバイス２１７～２２０が共用されるので、それらの数を最低限に抑えることができる点で、全体的コストを低減する。さらに、これは、周辺機器相互接続エクスプレス・デバイス２１７～２２０の組み込みを必要としないことによって、サーバ・ブレード２０１～２０４自体の複雑さおよびコストを低減することができる。
【００１１】
周辺機器相互接続スペシャル・インタレスト・グループは、いくつかの論理パーティションによって共用できる入出力アダプタをいかに設計するかについて規定するための規格を提供するが、その仕様は、入出力アダプタをいかにホスト・システムに接続するかについて規定していない。さらに、その規格は、各機能を単一システムにいかに割り当てることができるかについて規定するに過ぎない。
【発明の概要】
【発明が解決しようとする課題】
【００１２】
本発明は、周辺機器相互接続入出力仮想化構成を使用して冗長構成を生成するための、コンピュータ実施方法、データ処理システム、およびコンピュータ・プログラムを提供する。
【課題を解決するための手段】
【００１３】
本発明の一実施形態によれば、冗長システム構成を生成するためのコンピュータ実施方法が提示される。コンピュータ実施方法は、信用できるエンティティによって１組の仮想機能パス許可テーブル（virtual function path authorization table）を生成することであって、エントリが、１つまたは複数のシステム内の１組のアドレス範囲への機能のアクセスを定義し、エントリが、無効な機能間アクセス（crossfunction access）を防止する境界をさらに定義し、仮想機能が、シングルルートまたはマルチルート周辺機器相互接続デバイスによって実行される、生成することと、リクエスタから要求を受信して、要求されたデータを仮想機能から提供することと、１組のアドレス範囲において選択されたアドレス範囲内に受信バッファ（receivebuffer）を生成することと、選択されたアドレス範囲内の受信バッファのアドレスを含む、仮想機能のための仮想機能作業待ち行列エントリ（virtualfunction work queue entry）を生成することと、を含む。さらに、コンピュータ実施方法は、１組の仮想機能パス許可テーブルにおいて、選択されたアドレス範囲の使用を仮想機能が許可されているかどうかを判定することと、仮想機能は許可されているという判定に応答して、要求されたデータを、１つまたは複数のシステム内の各アドレス範囲の受信バッファに書き込むことと、要求されたデータの書き込みに応答して、リクエスタに完了通知を発行することと、を含む。
【００１４】
別の実施形態によれば、冗長システム構成を生成するためのデータ処理システムが提示

される。データ処理システムは、バスと、バスに接続されたコンピュータ実行可能命令を含むメモリと、中央プロセッサ・ユニットとを含む。中央プロセッサ・ユニットは、コンピュータ実行可能命令を実行し、コンピュータ実行可能命令は、信用できるエンティティによって１組の仮想機能パス許可テーブルを生成することであって、エントリが、１つまたは複数のシステム内の１組のアドレス範囲への機能のアクセスを定義し、エントリが、無効な機能間アクセスを防止する境界をさらに定義し、仮想機能が、シングルルートまたはマルチルート周辺機器相互接続デバイスによって実行される、生成することと、リクエスタから要求を受信して、要求されたデータを仮想機能から提供することと、１組のアドレス範囲において選択されたアドレス範囲内に受信バッファを生成することと、選択されたアドレス範囲内の受信バッファのアドレスを含む、仮想機能のための仮想機能作業待ち行列エントリを生成することと、１組の仮想機能パス許可テーブルにおいて、選択されたアドレス範囲の使用を仮想機能が許可されているかどうかを判定することと、仮想機能は許可されているという判定に応答して、要求されたデータを、１つまたは複数のシステム内の選択されたアドレス範囲の受信バッファに書き込むことと、要求されたデータの書き込みに応答して、リクエスタに完了通知を発行することと、を実行するようにデータ処理システムに指令する。

【００１５】

別の実施形態によれば、冗長システム構成を生成するためのコンピュータ・プログラム製品が提示される。コンピュータ・プログラム製品は、コンピュータ実行可能命令が保存されたコンピュータ可読媒体を含む。コンピュータ実行可能命令は、信用できるエンティティによって１組の仮想機能パス許可テーブルを生成するためのコンピュータ実行可能命令であって、エントリが、１つまたは複数のシステム内の１組のアドレス範囲への仮想機能のアクセスを定義し、エントリが、無効な機能間アクセスを防止する境界をさらに定義し、仮想機能が、シングルルートまたはマルチルート周辺機器相互接続デバイスによって実行される、コンピュータ実行可能命令と、リクエスタから要求を受信して、要求されたデータを仮想機能から提供するためのコンピュータ実行可能命令と、１組のアドレス範囲において選択されたアドレス範囲内に受信バッファを生成するためのコンピュータ実行可能命令と、を含む。コンピュータ実行可能命令は、選択されたアドレス範囲内の受信バッファのアドレスを含む、仮想機能のための仮想機能作業待ち行列エントリを生成するためのコンピュータ実行可能命令と、１組の仮想機能パス許可テーブルにおいて、選択されたアドレス範囲の使用を仮想機能が許可されているかどうかを判定するためのコンピュータ実行可能命令と、仮想機能は許可されているという判定に応答して、要求されたデータを、１つまたは複数のシステム内の選択されたアドレス範囲の受信バッファに書き込むためのコンピュータ実行可能命令と、要求されたデータの書き込みに応答して、リクエスタに完了通知を発行するためのコンピュータ実行可能命令と、をさらに含む。

【図面の簡単な説明】

【００１６】

【図１】周辺機器相互接続エクスプレス規格を実施するシステム・アーキテクチャのブロック図である。

【図２】周辺機器相互接続マルチルート入出力仮想化を含む図１のシステムのブロック図である。

【図３】周辺機器相互接続マルチルート入出力ファブリックを利用する分散コンピューティング・システムのブロック図である。

【図４】本発明の例示的な実施形態が実施できる、多数の論理パーティションを使用するシステム・リソースの仮想化のブロック図である。

【図５】例示的な一実施形態による、周辺機器相互接続エクスプレス・マルチルート入出力仮想化対応エンドポイントのブロック図である。

【図６】周辺機器相互接続エクスプレス・マルチルート対応周辺機器相互接続エクスプレス・スイッチのブロック図である。

【図７】例示的な一実施形態による、仮想機能作業待ち行列エントリのブロック図である

。
【図８】例示的な一実施形態による、マルチルート・デバイス内の任意の与えられた仮想階層にアクセスするための仮想機能の権限を確認するためのテーブルのブロック図である。
【図９】例示的な一実施形態による、マルチルート・デバイスの冗長パス実施のための代替ルート仮想階層を指定するためのテーブルのブロック図である。
【図１０】例示的な一実施形態による、許可アドレスと仮想機能の関係を指定するためのテーブルのブロック図である。
【図１１】例示的な一実施形態による、図１０のアドレスを使用する仮想機能作業待ち行列エントリのブロック図である。
【図１２】例示的な一実施形態による、マルチルート・デバイスおよびマルチルート・スイッチを使用する冗長システムの構成のブロック図である。
【図１３】例示的な一実施形態による、シングルルート・デバイスを使用する冗長論理パーティションの構成のブロック図である。
【図１４】例示的な一実施形態による、マルチルート・マルチシステム構成のマルチルート・ファブリック構成のプロセスのフローチャートである。
【図１５】例示的な一実施形態による、パートナ・システムに伝達するために必要とされる仮想階層番号をシステムが決定するためのプロセスのフローチャートである。
【図１６】例示的な一実施形態による、仮想機能作業待ち行列エントリを設定するためのプロセスのフローチャートである。
【図１７】例示的な一実施形態による、入出力ファブリック・パス動作ステータスおよび代替パス使用を動的に決定するためのプロセスのフローチャートである。
【図１８】例示的な一実施形態による、データ・レプリケーションのためのデュアル・パスを使用して書き込み動作を実行するプロセスのフローチャートである。
【発明を実施するための形態】
【００１７】
当業者によって理解されるように、本発明は、システム、方法、またはコンピュータ・プログラム製品として実施することができる。したがって、本発明は、全面的にハードウェアの実施形態、（ファームウェア、常駐ソフトウェア、マイクロコードなどを含む）全面的にソフトウェアの実施形態、またはソフトウェア態様とハードウェア態様を組み合わせた実施形態の形態を取ることができ、本明細書では、それらのすべてを一般に、「回路」、「モジュール」、または「システム」と呼ぶことができる。さらに、本発明は、媒体内に具体化されたコンピュータ使用可能プログラム・コードを有する表現である、任意の有形な媒体内に具体化されたコンピュータ・プログラム製品の形態を取ることができる。
【００１８】
１つまたは複数のコンピュータ使用可能またはコンピュータ可読媒体の任意の組合せを利用することができる。コンピュータ使用可能またはコンピュータ可読媒体は、例えば、電子、磁気、光、電磁気、赤外線、または半導体のシステム、装置、デバイス、または伝播媒体とすることができるが、それらに限定されない。コンピュータ可読媒体のより具体的な例（非網羅的なリスト）は、１つもしくは複数の電線を有する電気接続、ポータブル・コンピュータ・ディスケット、ハード・ディスク、ランダム・アクセス・メモリ（ＲＡＭ）、リード・オンリ・メモリ（ＲＯＭ）、消去可能プログラマブル・リード・オンリ・メモリ（ＥＰＲＯＭもしくはフラッシュ・メモリ）、光ファイバ、ポータブル・コンパクト・ディスク・リード・オンリ・メモリ（ＣＤＲＯＭ）、光記憶デバイス、インターネットもしくはイントラネットをサポートするような伝送媒体、または磁気記憶デバイスを含む。コンピュータ使用可能またはコンピュータ可読媒体は、例えば、紙または他の媒体の光学的スキャニングによって、プログラムが電子的に獲得でき、次に必要であれば、コンパイル、インタプリテーション、または他の適切な方法で処理でき、次にコンピュータ・メモリ内に保存できるように、プログラムが印刷された紙または別の適切な媒体とすることさえできることに留意されたい。本文書との関連では、コンピュータ使用可能またはコ

ンピュータ可読媒体は、命令実行システム、装置、もしくはデバイスによって使用するため、またはそれらに関連して使用するために、プログラムを格納、保存、伝達、伝播、または転送できる任意の媒体とすることができる。コンピュータ使用可能媒体は、ベースバンド内または搬送波の一部として、コンピュータ可読プログラム・コードが具体化された伝播データ信号を含むことができる。コンピュータ使用可能プログラム・コードは、無線、有線、光ファイバ・ケーブル、ＲＦなどを含むが、それらに限定されない、任意の適切な媒体を使用して、伝送することができる。

【００１９】

本発明の動作を実施するためのコンピュータ・プログラム・コードは、Ｊａｖａ、Ｓｍａｌｌｔａｌｋ、またはＣ＋＋などのオブジェクト指向プログラミング言語、および「Ｃ」プログラミング言語または類似のプログラミング言語などの従来の手続き型プログラミング言語を含む、１つまたは複数のプログラミング言語の任意の組合せで書くことができる。プログラム・コードは、スタンドアロン・ソフトウェア・パッケージとして、全面的にユーザのコンピュータ上で、部分的にユーザのコンピュータ上で、一部はユーザのコンピュータ上、一部はリモート・コンピュータ上で、または全面的にリモート・コンピュータもしくはサーバ上で実行することができる。後者のシナリオでは、リモート・コンピュータは、ローカル・エリア・ネットワーク（ＬＡＮ）もしくはワイド・エリア・ネットワーク（ＷＡＮ）を含む任意のタイプのネットワークを介して、ユーザのコンピュータに接続することができ、または接続は、（例えば、インターネット・サービス・プロバイダを使用してインターネットを介して）外部コンピュータに対して行うことができる。

【００２０】

本発明は、本発明の実施形態による方法、装置（システム）、およびコンピュータ・プログラム製品のフローチャート図またはブロック図あるいはその両方を参照しながら以下で説明される。フローチャート図またはブロック図あるいはその両方の各ブロック、およびフローチャート図またはブロック図あるいはその両方におけるブロックの組合せは、コンピュータ・プログラム命令によって実施することができることが理解されよう。

【００２１】

これらのコンピュータ・プログラム命令は、汎用コンピュータ、専用コンピュータ、またはマシンを生成するための他のプログラマブル・データ処理装置のプロセッサに提供することができ、コンピュータまたは他のプログラマブル・データ処理装置のプロセッサによって実行される命令は、フローチャートまたはブロック図あるいはその両方の１つまたは複数のブロックで指定される機能／動作を実施するための手段を生成する。特定の方法で機能するようにコンピュータまたは他のプログラマブル・データ処理装置に指令することができる、これらのコンピュータ・プログラム命令はまた、コンピュータ可読媒体内に保存することができ、コンピュータ可読媒体内に保存された命令は、フローチャートまたはブロック図あるいはその両方の１つまたは複数のブロックで指定される機能／動作を実施する命令手段を含む製造品を生成する。

【００２２】

コンピュータ・プログラム命令はまた、一連の動作ステップがコンピュータまたは他のプログラマブル装置上で実行されて、コンピュータ実施プロセスを生成するように、コンピュータまたは他のプログラマブル・データ処理装置にロードすることができ、コンピュータまたは他のプログラマブル装置上で実行される命令は、フローチャートまたはブロック図あるいはその両方の１つまたは複数のブロックで指定される機能／動作を実施するためのプロセスを提供する。

【００２３】

例示的な実施形態は、入出力仮想化機能から別個のシステムへの多数のパスを可能にする、マルチルート入出力仮想化（ＭＲ－ＩＯＶ）アダプタおよび入出力ファブリック（input/output fabric）の構成のためのメカニズムを提供する。例示的な実施形態は、周辺機器相互接続エクスプレス（ＰＣＩｅ）アダプタまたはエンドポイントに関して説明されるが、本発明は、そのようなものに限定されない。むしろ、例示的な実施形態のメカニズ

ムは、入出力アダプタ内で入出力仮想化をサポートする任意の入出力ファブリックにおいて実施することができる。

【００２４】

さらに、例示的な実施形態は、ハイパーバイザ（hypervisor）が利用される実施に関して説明されるが、本発明は、そのようなものに限定されない。反対に、ソフトウェア、ハードウェア、またはソフトウェアとハードウェアの任意の組合せのどれで実施されようと、現在知られている、または今後開発される、ハイパーバイザ以外の他のタイプの仮想化プラットフォームが、本発明の主旨および範囲から逸脱することなく使用できる。

【００２５】

ここで図を、特に図３を参照すると、周辺機器相互接続マルチルート入出力ファブリック（peripheralcomponent interconnect multi-root input output fabric）を利用する分散コンピューティング・システムのブロック図が、本発明の例示的な一実施形態に従って示されている。図３は、システム・ノードを共用入出力アダプタと接続するための周辺機器相互接続ファブリックを追加して、図１および図２の構成を増強する。図３に示されるように、分散コンピューティング・システム３００は、周辺機器相互接続マルチルート入出力ファブリック３４４に結合された複数のルート・ノード（rootnode）３６０～３６３を含み、次に周辺機器相互接続マルチルート入出力ファブリック３４４は、マルチルート入出力ファブリック構成マネージャ（multi-rootinput output fabric configuration manager）３６４と、周辺機器相互接続入出力アダプタまたはエンドポイント３４５～３４７に結合される。各ルート・ノード３６０～３６３は、１つまたは複数の対応するルート・コンプレックス（rootcomplex）３０８、３１８、３２８、３３８、３３９を含み、それらは、入出力リンク３１０、３２０、３３０、３４２、３４３をそれぞれ介して、周辺機器相互接続マルチルート入出力ファブリック３４４に接続され、さらにルート・ノード（ＲＮ）３６０～３６３のメモリ・コントローラ３０４、３１４、３２４、３３４に接続される。入出力ファブリック３４４は、リンク３５１、３５２、３５３を介して、入出力アダプタ３４５、３４６、３４７に接続される。入出力アダプタ３４５、３４６、３４７は、周辺機器相互接続エクスプレス入出力アダプタ３４５におけるような非入出力仮想化対応アダプタ、周辺機器相互接続エクスプレス入出力アダプタ３４６におけるようなシングルルート（ＳＲ）入出力仮想化アダプタ、または周辺機器相互接続エクスプレス入出力アダプタ３４７におけるようなマルチルート入出力仮想化アダプタとすることができる。

【００２６】

示されるように、ルート・コンプレックス３０８、３１８、３２８、３３８、３３９は、ルート・ノード３６０、３６１、３６２、３６３の一部である。ルート・ノード３６３に示されるように、ルート・ノード当たり２つ以上のルート・コンプレックスが存在してもよい。ルート・コンプレックスは、中央プロセッサ／メモリを入出力アダプタに接続する入出力階層（input/output hierarchy）のルートである。ルート・コンプレックスは、ホスト・ブリッジ（host bridge）と、ゼロまたは１つ以上のルート・コンプレックス組み込みエンドポイントと、ゼロまたは１つ以上のルート・コンプレックス・イベント・コレクタ（rootcomplex event collector）と、１つまたは複数のルート・ポートを含む。各ルート・ポートは、別個の入出力階層をサポートする。入出力階層は、ルート・コンプレックス、例えば、ルート・コンプレックス３０８と、ゼロまたは１つ以上の相互接続スイッチまたはブリッジ（スイッチもしくは周辺機器相互接続マルチルート入出力ファブリック３４４などの周辺機器相互接続エクスプレス・ファブリックを含む）あるいはその両方と、周辺機器相互接続エクスプレス入出力アダプタまたはエンドポイント３４５～３４７などの１つまたは複数のエンドポイントから成ることができる。

【００２７】

ルート・コンプレックスに加えて、各ルート・ノードは、１つまたは複数の中央処理装置３０１、３０２、３１１、３１２、３２１、３２２、３３１、３３２と、メモリ３０３、３１３、３２３、３３３と、メモリ・コントローラ３０４、３１４、３２４、３３４か

ら成る。メモリ・コントローラ３０４、３１４、３２４、３３４は、中央処理装置３０１、３０２、３１１、３１２、３２１、３２２、３３１、３３２を、バス３０５、３０６、３０７、３１５、３１６、３１７、３２５、３２６、３２７、３３５、３３６、３３７によって、メモリ３０３、３１３、３２３、３３３と接続し、バス３０９、３１９、３２９、３４０、３４１によって、入出力ルート・コンプレックス３０８、３１８、３２８、３３８、３３９と接続する。メモリ・コントローラは一般に、メモリのためのコヒーレンシ・トラフィック（coherency traffic）の処理などの機能を実行する。ルート・ノード３６０、３６１は、１つのコヒーレンシ・ドメイン（coherencydomain）を形成するために、それぞれのメモリ・コントローラ３０４、３１４を介して、接続３５９で互いに接続することができる。したがって、ルート・ノード３６０～３６１は、単一の対称マルチプロセッシング（ＳＭＲ：symmetricmulti-processing）システムとして動作することができ、またはルート・ノード３６２、３６３におけるように、別個のコヒーレンシ・ドメインを有する独立ノードとすることもできる。

【００２８】

マルチルート入出力ファブリック構成マネージャ３６４は、ルート・ノードの他の動作から分離されることができ、したがって、入出力ファブリック３４４に別個に接続されるように示されている。しかし、これはシステムに費用を追加し、したがって、本明細書で開示される実施形態は、１つまたは複数のルート・ノード３６０、３６１、３６２、３６３の一部として、この機能を含むことができる。構成マネージャ３６４は、マルチルート入出力ファブリック３４４のリソースを構成し、リソースをルート・ノード３６０、３６１、３６２、３６３に割り当てる。

【００２９】

図３に示されたハードウェアが様々であり得ることを当業者であれば理解されよう。例えば、光ディスク・ドライブなどの他の周辺デバイスを、示されたハードウェアに追加して、またはそれと交換して使用することができる。示された例は、本発明に関するアーキテクチャの限定を暗示することを意図していない。

【００３０】

図３の分散コンピューティング・システム３００の例を使用して、例示的な実施形態は、入出力仮想化デバイスの単一機能に、多数のシステムにアクセスする能力を提供する。その能力は、冗長パスを有する入出力サブシステムを構成マネージャ３６４によって構成することを可能にし、入出力仮想化デバイスの単一機能が、多数のシステム間の多数の通信パスを介して多数のシステムにアクセスすることを可能にする。

【００３１】

例示的な実施形態は、入出力（Ｉ／Ｏ）ファブリック３４４が、ルート・ノード３６０、３６１、３６２、３６３のシステムもしくは論理パーティション（ＬＰＡＲ）など、２つ以上のシステムによって共用される状況、各システムまたは論理パーティションが潜在的に、周辺機器相互接続エクスプレス入出力アダプタもしくはエンドポイント３４５～３４７などの入出力アダプタ（ＩＯＡ）を、他の論理パーティションと共用できる状況、または多数のシステムが、マルチルート入出力仮想化ファブリックの使用によって、入出力アダプタを共用できる状況に対処する。例示的な実施形態は、ルート・ノードの多数のシステムまたは論理パーティションにアクセスすることを周辺機器相互接続エクスプレス入出力アダプタ３４７などの入出力仮想化アダプタの単一機能に許可する一方で、またアクセスを許可すべきではないシステムへのアクセスを防止するための、メカニズムを規定する。したがって、エンドポイント３４５～３４７とマルチルート・ノードのメモリ３０３、３１３、３２３、３３３の間に冗長パスを確立する目的で、単一の入出力仮想化機能は、多数の仮想階層（ＶＨ：virtual hierarchy）またはパス、およびマルチルート入出力ファブリック３４４の仮想化機能へのアクセスを許可される。

【００３２】

図４を参照すると、本発明の例示的な実施形態が実施できる多数の論理パーティションを使用するシステム・リソースの仮想化のブロック図が提示されている。論理パーティシ

ョン化プラットフォーム（logical partitioned platform）４００内のハードウェアは、例えば、図３のルート・ノード３６０、３６１、３６２、３６３内で実施することができ、ルート・ノードに割り当てられた、マルチルート入出力ファブリック３４４および入出力アダプタ３４５～３４７の部分をさらに含むことができる。

【００３３】

論理パーティション化プラットフォーム４００は、パーティション化ハードウェア４３０と、オペレーティング・システム４０２、４０４、４０６、４０８と、プラットフォーム・ファームウェア４１０のパーティション管理を含む。オペレーティング・システム４０２、４０４、４０６、４０８は、論理パーティション化プラットフォーム４００上で同時に動作する、単一のオペレーティング・システムの多数のコピー、また多数の異種のオペレーティング・システムとすることができる。

【００３４】

オペレーティング・システム４０２、４０４、４０６、４０８は、パーティション４０３、４０５、４０７、４０９内に配置される。ハイパーバイザ・ソフトウェアまたはファームウェアは、プラットフォーム・ファームウェア４１０のパーティション管理を実施するために使用できるソフトウェアの一例である。ファームウェアは、電力がなくてもその内容を保持するメモリ・チップ内に、例えば、リード・オンリ・メモリ（ＲＯＭ）、プログラマブル・リード・オンリ・メモリ（ＰＲＯＭ）、消去可能プログラマブル・リード・オンリ・メモリ（ＥＰＲＯＭ）、電気的消去可能プログラマブル・リード・オンリ・メモリ（ＥＥＰＲＯＭ）、および不揮発性ランダム・アクセス・メモリ（ＮＶＲＡＭ）内に保存された「ソフトウェア」である。

【００３５】

加えて、パーティション４０３、４０５、４０７、４０９は、パーティション・ファームウェア４１１、４１３、４１５、４１７も含む。パーティション・ファームウェア４１１、４１３、４１５、４１７は、初期ブート・ストラップ・コード、例えば、電気電子学会（ＩＥＥＥ）１２７５規格のオープン・ファームウェア（Open Firmware）、およびランタイム・アブストラクション・ソフトウェア（ＲＴＡＳ：runtime abstractionsoftware）を使用して、実施することができる。パーティション４０３、４０５、４０７、４０９がインスタンス化される場合、ブート・ストラップ・コードのコピーが、プラットフォーム・ファームウェア４１０によって、パーティション４０３、４０５、４０７、４０９にロードされる。その後、制御は、ブート・ストラップ・コードに移され、ブート・ストラップ・コードを用いて、オープン・ファームウェアおよびランタイム・アブストラクション・ソフトウェアをロードする。その後、パーティション・ファームウェア４１１、４１３、４１５、４１７を実行するために、パーティション４０３、４０５、４０７、４０９に関連する、または割り当てられたプロセッサが、パーティションのメモリにディスパッチされる。

【００３６】

パーティション化ハードウェア４３０は、複数のプロセッサ４３２、４３４、４３６、４３８と、複数のシステム・メモリ・ユニット４４０、４４２、４４４、４４６と、複数の入出力アダプタ４４８、４５０、４５２、４５４、４５６、４５８、４６０、４６２と、ストレージ・ユニット４７０と、不揮発性ランダム・アクセス・メモリ・ストレージ４９８を含む。プロセッサ４３２、４３４、４３６、４３８、メモリ・ユニット４４０、４４２、４４４、４４６、不揮発性ランダム・アクセス・メモリ・ストレージ４９８、および入出力アダプタ４４８、４５０、４５２、４５４、４５６、４５８、４６０、４６２、またはそれらの部分の各々は、論理パーティション化プラットフォーム４００内の多数のパーティションの１つに割り当てることができ、その各々は、オペレーティング・システム４０２、４０４、４０６、４０８の１つに対応する。

【００３７】

プラットフォーム・ファームウェア４１０は、論理パーティション化プラットフォーム４００のパーティショニングを生成および実施するために、パーティション４０３、４０

５、４０７、４０９のための多数の機能およびサービスを実行する。プラットフォーム・ファームウェア４１０は、基礎をなすハードウェアと同じファームウェア実施仮想マシンを含むことができる、パーティション管理ファームウェアを含むことができる。したがって、プラットフォーム・ファームウェア４１０内のパーティション管理ファームウェアは、論理パーティション化プラットフォーム４００のハードウェア・リソースを仮想化することによって、独立のオペレーティング・システム・イメージ４０２、４０４、４０６、４０８の同時実行を可能にする。

【００３８】

サービス・プロセッサ４９０は、パーティション４０３、４０５、４０７、４０９におけるプラットフォーム・エラーの処理など、様々なサービスを提供するために使用することができる。これらのサービスは、エラーをベンダに報告するためのサービス・エージェントとしても動作することができる。パーティション４０３、４０５、４０７、４０９の動作は、ハードウェア管理コンソール４８０などのハードウェア管理コンソールを介して制御することができる。ハードウェア管理コンソール４８０は、異なるパーティションへのリソースの再アロケーションを含む様々な機能をシステム管理者がそこから実行できる、別個の分散コンピューティング・システムである。制御され得る動作は、パーティションが動作中であってもなくても、パーティションに割り当てられるコンポーネントに関連したパーティションの構成などを含む。

【００３９】

論理パーティション（ＬＰＡＲ）環境では、１つのパーティション内のリソースまたはプログラムが別のパーティションにおける動作に影響を及ぼすことは許可されない。さらに、有用なこととして、リソースの割り当ては、細かい粒度（fine-grained）とする必要がある。例えば、特定の周辺機器相互接続ホスト・ブリッジ（ＰＨＢ：peripheralcomponent interconnect host bridge）下のすべての入出力アダプタを同じパーティションに割り当てることは、パーティション間でリソースを動的に移動させる能力を含むシステムの構成可能性を制限するので、しばしば許可されない。

【００４０】

したがって、個々の入出力アダプタまたは入出力アダプタの部分を別個のパーティションに割り当てることができ、同時に、他のパーティションのリソースにアクセスすることなどによって、割り当てられたリソースが他のパーティションに影響を及ぼすことを防止することができるように、いくつかの機能が、入出力アダプタを入出力バスに接続するブリッジ内で必要とされる。

【００４１】

図５を参照すると、周辺機器相互接続エクスプレス・マルチルート入出力仮想化対応エンドポイントのブロック図が提示されている。図５に示されるように、図３のマルチルート周辺機器相互接続エクスプレス入出力アダプタ３４７などの周辺機器相互接続エクスプレス・マルチルート入出力仮想化エンドポイント５００は、周辺機器相互接続エクスプレス・ファブリックの周辺機器相互接続エクスプレス・スイッチなどとの通信がそれを介して実行され得る周辺機器相互接続エクスプレス・ポート５０１を含む。内部ルーティング５０２は、構成管理機能５０３および構成管理機能５０９、ならびに複数の仮想機能（ＶＦ：virtual function）５０４～５０６への通信経路を提供する。構成管理機能５０３は、仮想機能５０４～５０６とは反対に、物理機能（ＰＦ：physicalfunction）とすることができ、構成管理機能５０９は、基本機能（ＢＦ：base function）とすることができる。周辺機器相互接続仕様において使用される用語としての物理「機能」は、単一の構成空間によって表される１組のロジックである。言い換えると、物理「機能」は、例えば、分離不能リソース５０７において提供され得るような、メモリ内の機能関連構成空間に保存されたデータに基づいて構成可能な回路ロジックである。基本「機能」５０９についても同様のことが言える。

【００４２】

構成管理機能５０３は、仮想機能５０４～５０６を構成するために使用することができ

る。仮想機能は、周辺機器相互接続エクスプレス・マルチルート入出力仮想化エンドポイント５００の共用可能リソース・プール５０８において提供され得る、例えば、リンクなどの１つまたは複数の物理エンドポイント・リソースを、例えば、別の機能と共用する、入出力仮想化対応エンドポイント内の機能である。仮想機能は、ハイパーバイザによるランタイム介入なしに、直接的に、システム・イメージからの入出力およびメモリ動作のためのシンクとすることができ、システム・イメージへのダイレクト・メモリ・アクセス（ＤＭＡ：direct memory access）、完了、および割り込み動作のソースとすることができる。

【００４３】

マルチルート入出力仮想化エンドポイント５００は、例えば、図３のルート・ノード３６０～３６３など、多数のルート・ノード間で共用することもできる。基本機能としての構成管理機能５０９は、例えば、どのルート・ノードが各物理機能にアクセスするかなど、物理機能の特性を構成するために使用することができる。

【００４４】

周辺機器相互接続エクスプレス・エンドポイントは、周辺機器相互接続エクスプレス・エンドポイントによってサポートされる「機能」に関して、多くの異なるタイプの構成を有することができる。例えば、エンドポイントは、単一の物理機能、多数の独立物理機能をサポートすることができ、または多数の従属物理機能さえもサポートすることができる。ネイティブ入出力仮想化をサポートするエンドポイントでは、エンドポイントによってサポートされる各物理機能は、１つまたは複数の仮想機能に関連付けることができ、仮想機能自体は、他の物理機能に関連付けられた仮想機能に従属することができる。ルート・ノードに割り当てられる入出力仮想化エンドポイントの単位は、物理機能であり、マルチルート入出力仮想化対応エンドポイントは、多数の物理機能を含む。

【００４５】

一実施形態では、仮想機能（ＶＦ）－仮想階層（ＶＨ）許可テーブル（virtualfunction (VF) to virtual hierarchy (VH) authorization table）５１０は、図３の構成マネージャ３６４が、多数の仮想階層へのアクセスを各機能に与えることを可能にする。この態様は後で説明される。やはりさらに説明される仮想機能作業待ち行列５１１が、仮想機能のために、デバイス・ドライバ・ソフトウェアによって設定され、仮想機能によって実行される動作を指定する。テーブル内の仮想機能作業待ち行列エントリは、要求された特定の動作のために使用する１つまたは複数の仮想階層番号も含む。

【００４６】

図６を参照すると、周辺機器相互接続エクスプレス・マルチルート対応周辺機器相互接続エクスプレス・スイッチのブロック図が提示されている。周辺機器相互接続エクスプレス・スイッチ５２０は、例えば、図３の周辺機器相互接続マルチルート入出力ファブリック３４４において、周辺機器相互接続マルチルート入出力仮想化仕様によって規定されるように、使用することができる。スイッチ５２０は論理的に、ルート・ノードに接続されるポート当たり１つの、多数の仮想プレーン（virtual plane）から成る。例えば、ルート・ノード５２１は、周辺機器相互接続エクスプレス・リンク５２４によって、論理周辺機器相互接続－周辺機器相互接続（Ｐ２Ｐ：peripheralcomponent interconnect to peripheral component interconnect）ブリッジ５２７に接続され、それは論理的に、周辺機器相互接続－周辺機器相互接続ブリッジ５３６～５３８に対するスイッチに内部的に接続される。同様に、ルート・ノード５２２は、周辺機器相互接続エクスプレス・リンク５２５によって、論理周辺機器相互接続－周辺機器相互接続ブリッジ５２８に接続され、それは論理的に、周辺機器相互接続－周辺機器相互接続ブリッジ５３０～５３２に対するスイッチに内部的に接続され、またルート・ノード５２３は、周辺機器相互接続エクスプレス・リンク５２６によって、論理周辺機器相互接続－周辺機器相互接続ブリッジ５２９に接続され、それは論理的に、周辺機器相互接続－周辺機器相互接続ブリッジ５３３～５３５に対するスイッチに内部的に接続される。

【００４７】

その場合、周辺機器相互接続－周辺機器相互接続ブリッジ５３０、５３３、５３６は、マルチルート周辺機器相互接続エクスプレス・デバイス５４２のリソースを共用できるように、周辺機器相互接続エクスプレス・マルチルート・リンク５３９を共用する。同様に、その場合、周辺機器相互接続－周辺機器相互接続ブリッジ５３１、５３４、５３７は、マルチルート周辺機器相互接続エクスプレス・デバイス５４３のリソースを共用できるように、周辺機器相互接続エクスプレス・マルチルート・リンク５４０を共用し、その場合、周辺機器相互接続－周辺機器相互接続ブリッジ５３２、５３５、５３８は、マルチルート周辺機器相互接続エクスプレス・デバイス５４４のリソースを共用できるように、周辺機器相互接続エクスプレス・マルチルート・リンク５４１を共用する。

【００４８】

スイッチ５２０を設定するための制御点が、基本機能（ＢＦ）５４５である。この入出力仮想化構成メカニズム、例えば、基本機能５４５は、マルチルート周辺機器相互接続マネージャ（ＭＲ－ＰＣＩＭ：multi-root peripheral component interconnect manager）・プログラムが、スイッチ５２０内の論理構造を決定することを可能にする。例えば、図６は、各ルート・ノード５２１～５２３が、各周辺機器相互接続エクスプレス・マルチルート・デバイス５４２～５４４の一部にアクセスする、かなり対称な構成を示している。通常のシステムでは、システムを使用するユーザの必要を満たすために、システム管理者は、あまり対称的ではない方法で、入出力を設定することを望むことがある。

【００４９】

基本機能５４５、５０９は、マルチルート周辺機器相互接続マネージャ・プログラムによってアクセスされる。このプログラムが存在する場所は、周辺機器相互接続スペシャル・インタレスト・グループ入出力仮想化仕様によって規定されていない。プログラムは、例えば、ルート・ノード５２１～５２３の１つによって示されるような、マルチルート周辺機器相互接続マネージャにもっぱら専用され、ルート・ポート・ノードの１つに接続されるノード内に存在することができ、または別個のプロセッサ、例えば、図４の４９０のようなサービス・プロセッサが接続される、ベンダ独自のポートを介して提供することができる。マルチルート周辺機器相互接続マネージャがどこで実行されるかに関わらず、主要な要件は、このプログラムが堅牢であり、システム内の他のアプリケーションの動作または動作失敗によって影響され得ないことである。

【００５０】

例示的な実施形態は、図５に示された周辺機器相互接続エクスプレス・マルチルート入出力仮想化エンドポイント５００など、２つ以上のシステムにアクセスするための入出力仮想化アダプタの構成のためのメカニズムを提供する。例示的な実施形態のメカニズムは、図６の周辺機器相互接続エクスプレス・スイッチ５２０など、１つまたは複数の周辺機器相互接続エクスプレス・スイッチを含むことができる入出力ファブリックが、２つ以上のシステム、例えば、図３のルート・ノード３６２および３６３によって共用される状況に対処する。

【００５１】

図７を参照すると、例示的な一実施形態による、仮想機能（ＶＦ）作業待ち行列エントリのブロック図が提示されている。提供された例は、図５における仮想機能作業待ち行列エントリ５１１を表している。仮想機能作業待ち行列エントリ６０１のフィールド６０５、６０７は、周辺機器相互接続エクスプレス・ファブリック仮想階層番号を含む。フィールド６０５、６０７は、動作のためのダイレクト・メモリ・アクセス・データを、どのシステムに送信すべきか、またはどのシステムから受信するべきかを、仮想機能に指示する。このフィールドは、デバイス・ドライバ・ソフトウェアが、同じデータを多数のシステムに送信することを可能にする。例えば、ネットワーク・アダプタ上で仮想機能ネットワーク接続から受信したパケットは、多数のシステムのシステム・メモリに書き込むことができる。例えば、データがディスクなどの不揮発性ストレージ媒体にコミットされる前にシステムが故障した場合にデータのコピーを維持するために、図３のシステム３６２、３６３内のシステム・メモリ３２３、３３３に書き込むことができる。冗長システムではデ

ータは安全であるので、冗長性を有することによって、システムは、例えば、データが実際にディスクに書き込まれる前に、データがディスクにコミットされたことを通知することができる。データが物理的にディスクに書き込まれる前に書き込み動作の完了を通知することは、より速い応答、より短い待ち時間を、ディスク書き込み動作のリクエスタに提供する。

【００５２】

仮想機能作業待ち行列エントリ６０１の他のフィールドは、動作タイプ６０２と、転送長６０３と、動作アドレス６０４、６０６を含む。動作タイプ６０２は、どの動作を実行すべきかを仮想機能に指示する。例えば、ネットワーク・アダプタの場合、動作は、受信バッファの設定とすることができる。この場合、受信バッファは、各システムにつき１つの対の、動作アドレスと周辺機器相互接続エクスプレス・ファブリック仮想階層番号のフィールド対を２つ以上使用して、２つ以上のシステムにおいて設定することができる。例えば、６０４と６０５、６０６と６０７など、これらのフィールドの対は、受信データを送信すべき各システムにつき１つ存在する。この場合、転送長６０３は、バッファ長となるように設定される。

【００５３】

動作タイプおよびフィールド・タイプは、アダプタによって提供される機能によって様々になり得ることを当業者であれば理解されよう。正しいシステムにデータを送るために、周辺機器相互接続エクスプレス・ファブリック仮想階層番号が、各アドレスに提供される。

【００５４】

図８を参照すると、例示的な一実施形態による、マルチルート・デバイス内の任意の与えられた仮想階層にアクセスするための仮想機能の権限を確認するためのテーブルのブロック図が提示されている。仮想機能－仮想階層許可テーブル６１０は、図５の仮想機能－仮想階層許可テーブル５１０の一例である。マルチルート・デバイスでは、アダプタは、異なるシステムによってアクセスされ得る機能間にファイヤウォールの等価物を提供する。図７の周辺機器相互接続エクスプレス・ファブリック仮想階層番号フィールド６０５、６０７は、異なるシステムによって制御される異なる機能に通常は割り当てられる仮想階層番号を使用して、ファイヤウォールをトンネリングするためのメカニズムを提供する。図７の周辺機器相互接続エクスプレス・ファブリック仮想階層番号フィールド６０５、６０７は、１つのシステム内のデバイス・ドライバ・ソフトウェアによって設定されるので、アダプタ上の許可されたファイヤウォールだけをトンネリングするよう、システムが関連する仮想機能を設定できるように、使用される仮想階層番号が確認されることは重要である。必要な機能は、仮想機能－仮想階層許可テーブル６１０を介して提供される。アダプタ内の各仮想機能について仮想機能－仮想階層番号許可テーブル（virtual function to virtual hierarchy number authorization table）６１１、６１５が存在する。この例では、テーブルは、多数のエントリ６１２～６１４、６１６～６１８を含むことができ、テーブルが適用される仮想機能がアクセスを許可された各仮想階層につき１つのエントリを含むことができる。仮想機能が仮想機能作業待ち行列エントリ６０１を処理できるようにする前に、仮想機能が仮想階層番号にアクセスする許可を得ていることを確認するために、周辺機器相互接続エクスプレス・ファブリック仮想階層番号フィールド６０５、６０７は、適切な仮想機能－仮想階層許可テーブルと照合される。許可されていない場合、仮想機能作業待ち行列エントリ６０１の処理は許可されず、エラーが、デバイス・ドライバ・ソフトウェアに通知される。仮想機能－仮想階層許可テーブル６１０は、信用できるソフトウェアによって設定される。例えば、信用できるソフトウェアは、図３のマルチルート入出力ファブリック構成マネージャまたはマルチルート周辺機器相互接続マネージャ３６４とすることができる。テーブルは、システム内のデバイス・ドライバ・ソフトウェアによって変更することができず、したがって、トンネリング・プロセスの制御を安全にする。これらのテーブルの使用のさらなる説明は後で説明される。

【００５５】

図９を参照すると、例示的な一実施形態による、マルチルート・デバイスの冗長パス実施のための代替ルート仮想階層を指定するためのテーブルのブロック図が提示されている。仮想機能に許可された仮想階層番号のテーブル（authorized virtual hierarchy number for virtual function table）６２０は、図５の内部ルーティング５０２内に維持される１次および２次パス・エントリの対応である。仮想機能－仮想階層許可テーブル６１０において仮想階層番号によって指定されたパスの１つが利用不可能になった場合、冗長かつ堅牢な構成は、動作のために所望されるシステムへの代替パスを使用するための能力を提供する。この場合、仮想機能－仮想階層許可テーブル６１０の代わりに、仮想機能に許可された仮想階層番号の拡張テーブル６２０が使用できる。仮想機能に許可された仮想階層番号のテーブル６２０における相違は、各エントリ６２１、６２３、６２５に対して、動作不能の仮想階層番号の代わりに使用する代替仮想階層番号を指定する代替エントリ６２２、６２４、６２６が存在することである。例えば、エントリ６２１が仮想階層番号「１」を指定し、エントリ６２２が仮想階層番号「３」を指定した場合、仮想機能作業待ち行列エントリ６０１が仮想階層番号「１」を指定し、仮想階層番号「１」が動作不能として検出されたときは、仮想階層番号「１」を用いて利用可能だった同じシステム内の同じシステム・メモリにアクセスするために、仮想階層番号「３」が使用できる。したがって、例えば、同じ動作のための多数のシステムに同じデータを送信することによって、全システム故障を回避する方法が存在するばかりでなく、冗長性を通して入出力ファブリック故障を回避する方法も存在する。仮想階層番号は、パス識別子として使用することができる。

【００５６】

図１０を参照すると、例示的な一実施形態による、許可アドレスと仮想機能の関係を指定するためのテーブルのブロック図が示されている。仮想機能－アドレス範囲許可テーブル（virtual function to address range authorization table）６２８は、許可を必要とする各仮想機能のためのテーブルを含む。各機能について、関連する仮想機能がアクセスを許可されたアドレスの範囲を表すテーブル６３０、６４０内のエントリを用いて、１組の許可アドレスが提供される。この例では、第１の仮想機能６３０のためのテーブルは、１組の関連するアドレスを有する。第１の仮想機能が使用を許可されたアドレス６３０は、許可アドレス６３２～６３８として列挙される。同様に、最後の仮想機能「ＶＦｎ」も、テーブル・エントリ６４０によって示される１組のエントリを有する。仮想機能－アドレス範囲許可テーブル６２８の機能は、例えば、同じルート・ノードの異なる論理パーティション内のアドレス範囲などのリソースへの仮想機能によるアクセスを許可する点において、図８の仮想機能－仮想階層許可テーブル６１０の機能と同様である。仮想機能－アドレス範囲許可テーブル６２８は、仮想階層を使用する代わりにアドレスを用いて、仮想機能－仮想階層許可テーブル６１０と同様の能力を提供する。

【００５７】

図１１を参照すると、例示的な一実施形態による、図１０のアドレスを使用する仮想機能（ＶＦ）作業待ち行列エントリのブロック図が提示されている。提供された例は、図５の仮想機能作業待ち行列５１１のさらなる例としての、図７の仮想機能作業待ち行列エントリ６０１を表している。この例では、仮想機能作業待ち行列エントリ６４２は、以前のように、動作タイプ６４４および転送長６４６を含む多数のフィールドを含む。図７の先の仮想機能作業待ち行列エントリとの相違は、仮想階層番号が存在しないことである。仮想階層番号の代わりに、動作アドレス６４８から動作アドレス６５０が見出される。動作アドレスは、データに関連付けられたロケーション、例えば、同じルート・ノードの異なる論理パーティション内のアドレスを指定する。

【００５８】

図１２を参照すると、例示的な一実施形態による、マルチルート・デバイスおよびマルチルート・スイッチを使用する冗長システムの構成のブロック図が提示されている。図３の分散コンピューティング・システム３００の例を使用して、マルチルート・デバイスと、コンピュータ電子コンプレックス（ＣＥＣ：computer electronic complex）を使用し

てマルチルート・デバイス１　７２７およびマルチルート・デバイス２　７２８などのコンピュータ電子コンプレックス通信デバイスに接続されるマルチルート・スイッチとを使用するシステム７００の構成が定義される。

【００５９】

コンピュータ電子コンプレックス１　７０１およびコンピュータ電子コンプレックス２　７０２を含む２つのコンピュータ・システムが示されているが、３ウェイ以上の冗長システムも構成できることは当業者であれば理解されよう。コンピュータ電子コンプレックスは、図３のルート・ノードに対応し、周辺機器相互接続ホスト・ブリッジ（ＰＨＢ）は、図３のルート・コンプレックスに対応する。

【００６０】

２つのコンピュータ電子コンプレックスは、論理パーティションの組を形成するために、図４におけるように、パーティショニングすることもできる。２つのコンピュータ電子コンプレックスは、システム・メモリ７０３、７０４と、各々３つの周辺機器相互接続ホスト・ブリッジ７０５～７０７、７０８～７１０から成る。マルチルート周辺機器相互接続マネージャ７１１は、図３の構成マネージャ３６４に対応する。これは冗長性の高いシステムであるので、１次マルチルート周辺機器相互接続マネージャ７１１の故障、コンピュータ電子コンプレックス１の故障、周辺機器相互接続ホスト・ブリッジ１（ＰＨＢ１）７０５の故障、またはマルチルート周辺機器相互接続マネージャ７１１がマルチルート入出力ファブリック動作を制御することを妨げる他の任意の故障の場合には、１次マルチルート周辺機器相互接続マネージャ７１１を引き継ぐことができる、バックアップ・マルチルート周辺機器相互接続マネージャ７１２も存在する。マルチルート周辺機器相互接続マネージャのフェイルオーバ・プロセスは、本発明の範囲を超えている。

【００６１】

マルチルート周辺機器相互接続マネージャ７１１、７１２は、周辺機器相互接続ホスト・ブリッジ１（ＰＨＢ１）７０５およびマルチルート・スイッチ１　７１９への周辺機器相互接続エクスプレス・リンク７１３を介して、また周辺機器相互接続ホスト・ブリッジ６（ＰＨＢ６）７１０およびマルチルート・スイッチ２　７２０への周辺機器相互接続エクスプレス・リンク７１６を介して、周辺機器相互接続エクスプレス・マルチルート入出力仮想化仕様によって管理仮想階層として定義されるマルチルート・ファブリックの仮想階層０に接続される。他の周辺機器相互接続ホスト・ブリッジは、マルチルート・ファブリックへの１次仮想階層接続および２次仮想階層接続を形成する。具体的には、コンピュータ電子コンプレックス１の１次仮想階層は、仮想階層１であり、コンピュータ電子コンプレックス１は、周辺機器相互接続ホスト・ブリッジ２（ＰＨＢ２）７０６、さらにマルチルート・スイッチ１　７１９への周辺機器相互接続エクスプレス・リンク７１４を介して、仮想階層１に接続する。コンピュータ電子コンプレックス１の２次仮想階層接続は、周辺機器相互接続ホスト・ブリッジ３（ＰＨＢ３）７０７、さらにマルチルート・スイッチ２　７２０への周辺機器相互接続エクスプレス・リンク７１８を介して仮想階層３に接続する。同様に、コンピュータ電子コンプレックス２の１次仮想階層は、周辺機器相互接続ホスト・ブリッジ５（ＰＨＢ５）７０９、さらにマルチルート・スイッチ２　７２０への周辺機器相互接続エクスプレス・リンク７１５を介して仮想階層４に接続する。コンピュータ電子コンプレックス２の２次仮想階層接続は、周辺機器相互接続ホスト・ブリッジ４（ＰＨＢ４）７０８、さらにマルチルート・スイッチ１　７１９への周辺機器相互接続エクスプレス・リンク７１７を介して仮想階層２に接続する。

【００６２】

「２次」リンクは、必ずしもバックアップ目的だけとは限らず、デバイスがその下に配置されるスイッチに応じて、デバイスとの通信用にも使用される。一般に、デバイスからコンピュータ電子コンプレックスへの最短パスが使用され、それは、動作待ち時間を短縮するために最小数のスイッチを通過するパスである。その場合、多数のスイッチを通過するパスは、バックアップ目的で残して置かれる。周辺機器相互接続エクスプレス・リンク７２１、７２２は、代替パスを提供するためにスイッチ間接続を提供する。

【００６３】
各マルチルート・スイッチの下には、２つのマルチルート・デバイスが示されている。マルチルート・デバイス１　７２７は、接続７３８によってネットワークに接続し、周辺機器相互接続エクスプレス・リンク７２３を介してマルチルート・スイッチ１　７１９に接続するネットワーク・デバイスとして示されている。同様に、マルチルート・デバイス２　７２８は、接続７３９によってネットワークに接続し、周辺機器相互接続エクスプレス・リンク７２６を介してマルチルート・スイッチ２に接続するネットワーク・デバイスとして示されている。
【００６４】
冗長性の理由で、２つのネットワーク・アダプタ接続７３８、７３９は、外部ネットワーク・スイッチに接続する可能性が最も高く、両方のデバイスは、同じネットワークにアクセスする。そのようにして、一方のネットワーク・アダプタが故障した場合でも、両方の中央電子コンプレックスは依然として、残りのアダプタを介してネットワークにアクセスする。ネットワーク・アダプタ接続７２７、７２８に加えて、２つのディスク・アダプタが、マルチルート・デバイス３　７２９およびマルチルート・デバイス４　７３０として存在し、これらのデバイスの両方は、同じ１組のディスク・ドライブ７３１にアクセスする。マルチルート・デバイス３　７２９は、周辺機器相互接続エクスプレス・リンク７２４を介してマルチルート・スイッチ１　７１９に接続し、マルチルート・デバイス４　７３０は、周辺機器相互接続エクスプレス・リンク７２５を介してマルチルート・スイッチ２　７２０に接続する。
【００６５】
この例では、マルチルート・デバイス１　７２７は、４つの仮想階層、すなわち、仮想階層１　７３２、仮想階層２　７３３、仮想階層３　７３４、および仮想階層４　７３５にアクセスする。これらの仮想階層の各々は、通常は別個の周辺機器相互接続エクスプレス機能に関連付けられる。例えば、１つの仮想機能が別の仮想機能の仮想階層にアクセスできないように機能の各々がファイヤウォール７３７によって分離されることになる仮想機能。ファイヤウォール・トンネル７３６を仮想階層１　７３２と仮想階層２　７３３の間（例えば、マルチルート・デバイス１　７２７の仮想機能１と仮想機能２の間）に生成して、マルチルート・デバイス１　７２７が、仮想階層の異なる組に接続された両方のコンピュータ電子コンプレックス内のメモリ７０３およびメモリ７０４にダイレクト・メモリ・アクセスによってデータを送ることを可能にすることができる。
【００６６】
マルチルート・デバイス１　７２７は、図５に示された周辺機器相互接続エクスプレス・マルチルート入出力仮想階層エンドポイント５００と論理的に同様である。そのようなものとして、マルチルート・デバイス１　７２７は、図５および図８の仮想機能－仮想階層許可テーブル５１０、６１０と、図７の仮想機能作業待ち行列エントリ６０１を有する図５の仮想機能作業待ち行列５１１を含む。マルチルート周辺機器相互接続マネージャ７１１におけるような信用できるソフトウェアが、仮想機能が仮想階層１　７３２および仮想階層２　７３３の両方にアクセスすることを、基本的にファイヤウォール・トンネル７３６を抜けるトンネルを形成して可能にするように、仮想機能－仮想階層許可テーブルを設定する。
【００６７】
ファイヤウォールを抜けるトンネルの他の実施形態も使用することができる。例えば、１つの仮想機能が、何らかの手段によって別の仮想機能への通信パスを生成し、情報を、データ上で実行する動作とともに、他の仮想機能へ渡す能力が提供できる。他の手段は、ファイヤウォールを抜けるトンネルが信用できるコードによって制御され得るように、説明されたメカニズムのような手段を設定する安全な方法も必要とする。
【００６８】
以下では、ディスクに書き込まれる予定のデータをネットワーク・リンク７３８から受信する動作について説明する。仮想機能の処理を担当するコンピュータ電子コンプレック

ス１　７０１内のデバイス・ドライバが、システム・メモリ７０３内に受信バッファを生成する。加えて、コンピュータ電子コンプレックス１　７０１は、例えば、２つのコンピュータ電子コンプレックス間のネットワーク接続を使用することによって、コンピュータ電子コンプレックス２　７０２内の対応するドライバと通信する。対応するコンピュータ電子コンプレックス２　７０２のドライバは、システム・メモリ７０４内に対応する受信バッファをアロケートし、その後、受信バッファのアドレスをコンピュータ電子コンプレックス１　７０１内のドライバに伝達する。その後、コンピュータ電子コンプレックス１　７０１内のドライバは、仮想階層１　７３２を介してコンピュータ電子コンプレックス１　７０１の受信バッファを、仮想階層２　７３３を介してコンピュータ電子コンプレックス２　７０２の受信バッファを指し示す仮想機能作業待ち行列エントリを、マルチルート・デバイス１　７２７の仮想機能内に設定する。仮想機能によって通信パケットを受信すると、仮想機能は、バッファがどこに配置されているかを識別するために仮想機能作業待ち行列エントリ内に与えられた情報を使用し、その後、仮想機能に対応する図８の仮想機能－仮想階層許可テーブル６１０を使用して、ファイヤウォールをトンネリングするための仮想機能の権限を検証する。権限が合格すると、マルチルート・デバイス１　７２７は、次に仮想階層１　７３２を介してコンピュータ電子コンプレックス１　７０１のシステム・メモリにダイレクト・メモリ・アクセスによってデータを送り、次に仮想階層２　７３３を介してコンピュータ電子コンプレックス２　７０２のシステム・メモリ７０４にダイレクト・メモリ・アクセスによってデータを送る。これらのダイレクト・メモリ・アクセスの正常完了時に、デバイス・ドライバは、マルチルート・デバイス１　７２７から割り込みによって通知され、両方のコンピュータ電子コンプレックスに対して正常に完了した動作を検出する。この時点で、データは、コンピュータ電子コンプレックスの一方における故障から安全であり、動作は、正常に完了したと見なすことができる。この時点で、例えば、システム・クラッシュからデータは安全であるので、データがまだディスク上になくても、ディスク書き込み動作のオリジネータは、動作が完了したと通知される。その後、コンピュータ電子コンプレックス１　７０１内のデバイス・ドライバは、マルチルート・デバイス３　７２９を介するディスク書き込み動作を待ち行列に入れる。

【００６９】

図１２をさらに参照すると、２つのマルチルート・スイッチから成るマルチルート・ファブリックを通る多数のパスが提示されている。例えば、周辺機器相互接続エクスプレス・リンク７１７の故障が存在する場合、マルチルート・デバイスは、システム・メモリ７０４に対する書き込みを、上で説明されたように実行することができない。マルチルート・デバイスが、図９に示された冗長テーブルを実装している場合、そのリンクを介するマルチルート・デバイス１　７２７からコンピュータ電子コンプレックス２　７０２へのパスが動作可能でないとき、仮想階層２　７３３の代わりの仮想階層４　７３５による代替パスが存在することを決定するために、図９に示されたテーブルを使用することができ、データは、リンク７２３、次にマルチルート・スイッチ１　７１９、次に周辺機器相互接続エクスプレス・リンク７２１、７２２、次にマルチルート・スイッチ２　７２０、次に周辺機器相互接続エクスプレス・リンク７１５、次に周辺機器相互接続ホスト・ブリッジ５（ＰＨＢ５）７０９を介して、システム・メモリ７０４まで流れる。

【００７０】

図１３を参照すると、例示的な一実施形態による、シングルルート・デバイスのみを使用する冗長論理パーティションの構成のブロック図が提示されている。図４の論理パーティション化プラットフォーム４００のさらなる一例として、多数の仮想階層という概念が存在しない構成８００が提示されている。別個の仮想階層を有する代わりに、仮想機能にダイレクト・メモリ・アクセス・アドレス範囲を割り当てるという概念が存在する。単一のシステム８０１は、多数の論理パーティション８０２～８０３から成り、その各々は、１つまたは複数の中央処理装置８０４～８０７を有し、各中央処理装置は、メモリ８０８～８０９を有する。論理パーティションは、１つまたは複数の周辺機器相互接続ホスト・ブリッジ（ＰＨＢ）８１０～８１１を共用し、シングルルート・デバイス８１４～８１５

は、周辺機器相互接続エクスプレス・リンク８１２～８１３を介して、周辺機器相互接続ホスト・ブリッジに接続される。シングルルート・デバイスは、ネットワーク・アダプタとして示されており、デバイスは、リンク８１６～８１７を介してネットワークに接続される。図１２におけるように、他方のリンクが故障した場合でも、両方のリンクが同じネットワークにアクセスできるように、２つのネットワーク・リンクが、ネットワーク・スイッチ（図示されず）に接続される。また図１２におけるように、仮想機能８１８～８２１は、ファイヤウォール８２３によって分離され、仮想機能が別の仮想機能の論理パーティション・メモリにアクセスすることを許可するために、ファイヤウォール・トンネル８２２が生成される。アクセスは、各仮想機能がただ１つだけの論理パーティションのメモリにアクセスすることを要求する、標準的な周辺機器相互接続エクスプレス入出力仮想化仕様とは異なる。

【００７１】

シングルルート・トンネリングを可能にするデータ構造は、図８に示されたマルチルートの場合に必要とされるものに類似している。仮想階層を含むテーブルの代わりに、許可された各仮想階層番号は、許可された周辺機器相互接続エクスプレス・ダイレクト・メモリ・アクセス・アドレス範囲によって置き換えられる。マルチルートの場合のマルチルート周辺機器相互接続マネージャに類似するシングルルート周辺機器相互接続マネージャ（図示されず）は、周辺機器相互接続エクスプレス・ダイレクト・メモリ・アクセス・アドレス範囲をアロケートし、図１０の仮想機能－アドレス範囲許可テーブル６２８を設定する。マルチルートの場合の仮想階層のためのように、明示的に設定されない限り、１つの論理パーティションが別の論理パーティションのメモリにアクセスできないように、論理パーティション内のソフトウェアは、テーブルにアクセスさせてもらえない。図１２のマルチルートの場合におけるように、２つの論理パーティションは、受信バッファを設定するために、マルチルートの場合のコンピュータ電子コンプレックスにおいてソフトウェアが行ったのと同じ方法で通信する。図１１の仮想機能作業待ち行列エントリ６４２は、この場合、仮想階層番号を含むが、図１１の６４８のような動作アドレスをパス識別子として使用することができる。

【００７２】

図１３には、２つの論理パーティションが示されているが、論理パーティションの３ウェイ以上の冗長組も構成できることは当業者であれば理解されよう。示されるように、仮想機能は、マルチルート周辺機器相互接続デバイス仮想機能およびシングルルート周辺機器相互接続デバイス仮想機能の一方とすることができる。

【００７３】

図１４を参照すると、例示的な一実施形態による、マルチルート・マルチシステム構成のマルチルート・ファブリック構成のプロセスのフローチャートが提示されている。構成プロセス９００は、図１２に示されたような構成を提供する、図３の構成マネージャ３６４の構成プロセスの一例である。構成プロセス９００が開始し（ステップ９０２）、マルチルート周辺機器相互接続マネージャが、マルチルート・ファブリックを構成する（ステップ９０４）。マルチルート・ファブリックの構成は、冗長性のための任意の所望の代替ルートを含む、デバイスからルート・コンプレックスまでの正しいルートを生成する。マルチルート周辺機器相互接続マネージャは、仮想階層番号－周辺機器相互接続ホスト・ブリッジ（ＰＨＢ）相関をルート・コンプレックスから利用可能にする（ステップ９０６）。マルチルート周辺機器相互接続マネージャは、任意の代替相関を含む仮想機能－仮想階層番号許可テーブルを設定するために、デバイス物理機能のためのデバイス・ドライバを起動し（ステップ９０８）、その後、構成プロセス９００は終了する。

【００７４】

図１５を参照すると、例示的な一実施形態による、パートナ・システムに伝達するために必要とされる仮想階層番号をシステムが決定することを可能にするプロセスのフローチャートが提示されている。プロセス１０００は、ルート・ノード３６０およびルート・ノード３６１によって図１２の構成を使用する、または図１３の構成ならびに図４の論理パ

ーティション４０３および論理パーティション４０５を使用するプロセスの一例である。
【００７５】
プロセス１０００が開始し（ステップ１００２）、それぞれのパートナおよびパートナに関連付けられた仮想階層番号を発見するために、コンピュータ電子コンプレックスが互いに通信し、または論理パーティションが使用される場合、論理パーティションが互いに通信する（ステップ１００４）。コンピュータ電子コンプレックスまたは論理パーティションの各々は、それぞれのコンプレックスまたはパーティションに関連付けられたデバイスを発見し、それぞれの発見デバイスのためのデバイス・ドライバをロードし、それぞれの仮想機能のための仮想機能－仮想階層番号許可テーブルを読む（ステップ１００６）。デバイス・ドライバは今では、図７の適切な仮想機能作業待ち行列エントリ６０１を設定するために必要な仮想階層番号を有している。その後、プロセス１０００は終了する（ステップ１００８）。
【００７６】
図１６を参照すると、例示的な一実施形態による、仮想機能作業待ち行列エントリを設定するためのプロセスのフローチャートが提示されている。プロセス１１００は、図１２のＣＥＣ１　７０１または図１３のＬＰＡＲ１　８０２などの中央電子コンプレックスによって、図５の仮想機能作業待ち行列エントリ５１１を確立するためのプロセスの一例である。プロセス１１００が開始し（ステップ９０２）、マスタ・コンピュータ電子コンプレックスまたは論理パーティションが、システム仮想機能内の仮想機能作業待ち行列エントリを設定する（ステップ１１０４）。生成されたエントリは、動作が適用可能なすべてのコンピュータ電子コンプレックスまたは論理パーティションに対して仮想階層番号を指定する。マスタ・コンピュータ電子コンプレックスまたは論理パーティションは、特定の動作のためのデバイス・ドライバが存在する場所である。すべてのコンピュータ電子コンプレックスまたは論理パーティションは、同時に実行するマスタ動作を有することができる。すなわち、様々なコンピュータ電子コンプレックスまたは論理パーティション間に作業負荷を分散させるために、１つのコンピュータ電子コンプレックスまたは論理パーティションは、作業負荷の一部を担い、その部分を制御することができ、別のコンピュータ電子コンプレックスまたは論理パーティションは、作業負荷の別の一部を担うことができる。
【００７７】
デバイスは、要求された動作を実行し、動作のための仮想機能作業待ち行列エントリ内の仮想階層番号およびアドレスを使用して、データをすべての適切なコンピュータ電子コンプレックスまたは論理パーティションに送信する（ステップ１１０６）。その後、プロセス１１００は終了する（ステップ１１０８）。
【００７８】
図１７を参照すると、例示的な一実施形態による、入出力ファブリック・パス動作ステータスおよび代替パス使用を動的に決定するためのプロセスのフローチャートが提示されている。構成プロセス１２００は、パス利用可能性を決定するための、図１２のＭＲデバイス１　７２７などのデバイスのプロセスの一例である。構成プロセス１２００が開始し（ステップ１２０２）、デバイスが、システム・メモリへの仮想階層パスの動作ステータスを定期的に決定し、仮想階層パスが利用可能でない場合にはフラグを設定する（ステップ１２０４）。例えば、デバイスは、ダイレクト・メモリ・アクセスを介してシステム・メモリ内のロケーションを読み、読み取り時に動作タイムアウトなどのエラーを受信した場合、パスを利用不能としてマーク付ける。一般に、デバイスは、１次パスが利用可能である場合、そのパス上で動作を開始し、それ以外の場合、デバイスは、代替パスを使用する（ステップ１２０６）。その後、プロセス１２００は終了する（ステップ１２０８）。
【００７９】
図１８を参照すると、例示的な一実施形態による、データ・レプリケーションのためのデュアル・パスを使用して書き込み動作を実行するプロセスのフローチャートが提示されている。構成プロセス１３００は、図１２のＣＥＣ１　７０１または図１３のＬＰＡＲ１

８０２などの中央電子コンプレックスのデバイス・ドライバ部分の一例である。プロセス１３００は、デュアル・パス・ディスク書き込み動作の一例であるが、より多くまたは多数のパスも同様に使用することができる。例えば、書き込み動作は、１組のパスを含むことができ、その組は３つ以上のパスを含む。ストレージ・サーバ・ディスク書き込み動作のプロセス１３００が開始し（ステップ１３０２）、デバイス・ドライバが、各中央電子コンプレックスに１つの受信バッファが配置される、２つの通信ネットワーク受信バッファを指し示すための仮想機能作業待ち行列エントリを設定する（ステップ１３０４）。ディスクに書き込まれるデータが、通信ネットワークを介してホストから受信され、両方の中央電子コンプレックス内のバッファに書き込まれる（ステップ１３０６）。各受信バッファへの書き込み動作は、同時またはほぼ同時に発生する。ディスク書き込み動作は、データがまだディスクに書き込まれていなくても、完了したものとしてホストに通知される（ステップ１３０８）。ディスク書き込みの正常完了の後、データは物理的にディスクに書き込まれ、両方の中央電子コンプレックスのメモリから廃棄される（ステップ１３１０）。その後、プロセス１３００は終了する（ステップ１３１２）。

【００８０】

したがって、例示的な実施形態は、入出力仮想化デバイスの単一の機能が、多数のシステム間の多数のパスを介して多数のシステムにアクセスするための能力を提供する。特に、単一の機能は、冗長通信パスを確立するために、入出力ファブリックの多数の仮想階層へのアクセスを許可されることができる。冗長システムの確立は、データ・インテグリティの理由で多数のシステムの２つ以上のシステムにデータが送信されること、または同じ入出力仮想化アダプタの同じ機能から２つ以上のシステム内のデータにアクセスすることを可能にする。例示的な一実施形態では、許可は、仮想機能－仮想階層許可対応テーブルまたは仮想機能－アドレス範囲許可テーブルの使用を通して確立される。対応は、機能が、リソースに関連付けられた別の仮想機能または仮想階層のリソースを使用するために、仮想機能または仮想階層を分離するファイヤウォールをトンネリングすることを特に許可する。

【００８１】

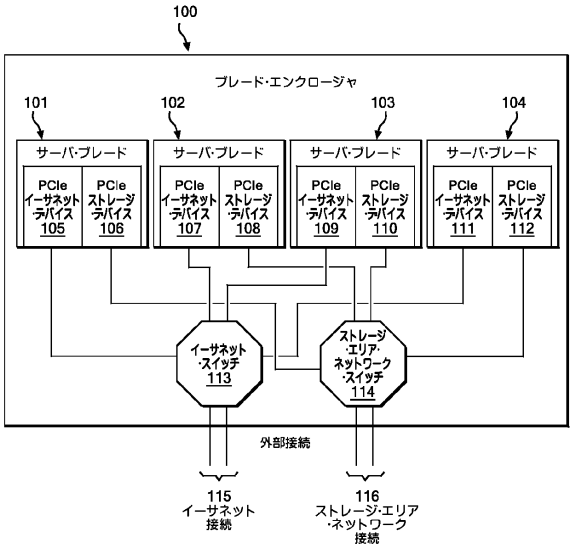
図のフローチャートおよびブロック図は、本発明の様々な実施形態による、システム、方法、およびコンピュータ・プログラム製品の可能な実施の、アーキテクチャ、機能、および動作を示す。この点で、フローチャートまたはブロック図の各ブロックは、規定された論理機能を実施するための１つまたは複数の実行可能命令を含む、モジュール、セグメント、またはコードの部分を表すことができる。いくつかの代替実施では、ブロック内で述べられた機能は、図内で述べられた順序とは異なる順序で発生し得ることにも留意されたい。例えば、連続して示される２つのブロックは、実際には実質的に同時に実行されることがあり、またはブロックは、含まれる機能に応じて、時には逆の順序で実行されることもある。ブロック図またはフローチャート図あるいはその両方の各ブロック、およびフローチャート図またはブロック図あるいはその両方におけるブロックの組合せは、指定された機能もしくは動作を実行する専用ハードウェア・ベースのシステムによって、または専用ハードウェアおよびコンピュータ・プログラム命令の組合せによって実施できることにも留意されたい。

【００８２】

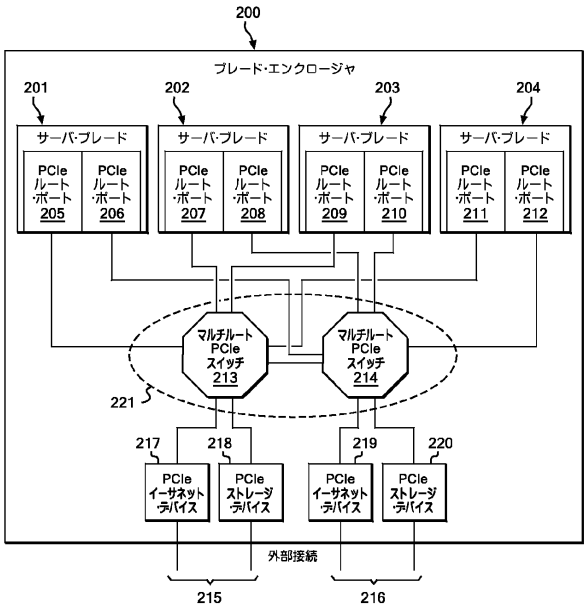
添付の特許請求の範囲におけるすべての手段またはステップおよび機能要素に対応する構造、材料、動作、および均等物は、具体的に特許請求されるように他の特許請求要素と組み合わせて機能を実行するための任意の構造、材料、または動作を含むことが意図されている。本発明の説明は、例示および説明の目的で提示されたものであり、網羅的であること、または開示された形態の本発明に限定することは意図していない。本発明の範囲および主旨から逸脱しない多くの修正および変形が、当業者には明らかであろう。本発明の原理および実際の応用を最もよく説明し、企図された特定の使用に適した様々な修正を有する様々な実施形態のために当業者が本発明を理解できるように、実施形態が選択され、説明された。

【００８３】
本発明は、全面的にハードウェアの実施形態、全面的にソフトウェアの実施形態、またはハードウェア要素とソフトウェア要素の両方を含む実施形態の形態を取ることができる。好ましい一実施形態では、本発明は、ファームウェア、常駐ソフトウェア、マイクロコードなどを含むが、それらに限定されない、ソフトウェアで実施される。
【００８４】
さらに、本発明は、コンピュータもしくは任意の命令実行システムによって使用される、またはそれらに関連して使用されるプログラム・コードを提供するコンピュータ使用可能またはコンピュータ可読媒体から取得可能なコンピュータ・プログラム製品の形態を取ることができる。この説明のため、コンピュータ使用可能またはコンピュータ可読媒体は、命令実行システム、装置、もしくはデバイスによって使用される、またはそれらに関連して使用されるプログラムを含み、保存し、伝達し、伝播し、または転送することができる任意の有形な装置とすることができる。
【００８５】
媒体は、電子、磁気、光、電磁気、赤外線、または半導体システム（または装置もしくはデバイス）、あるいは伝播媒体とすることができる。コンピュータ可読媒体の例は、半導体または固体メモリ、磁気テープ、着脱可能コンピュータ・ディスケット、ランダム・アクセス・メモリ（ＲＡＭ）、リード・オンリ・メモリ（ＲＯＭ）、固定磁気ディスク、および光ディスクを含む。光ディスクの現在の例は、コンパクト・ディスク－リード・オンリ・メモリ（ＣＤ－ＲＯＭ）、コンパクト・ディスク－リード／ライト（ＣＤ－Ｒ／Ｗ）、およびＤＶＤを含む。
【００８６】
プログラム・コードの保存または実行あるいはその両方を行うのに適したデータ処理システムは、システム・バスを介して直接的または間接的にメモリ要素に結合された少なくとも１つのプロセッサを含む。メモリ要素は、プログラム・コードの実際の実行中に利用される局所メモリ、大容量記憶装置、および実行中に大容量記憶装置からコードを取り出さなければならない回数を減らすために、少なくとも一部のプログラム・コードの一時記憶を提供するキャッシュ・メモリを含むことができる。
【００８７】
（キーボード、ディスプレイ、ポインティング・デバイスなどを含むが、それらに限定されない）入出力すなわちＩ／Ｏデバイスは、直接的に、または介在するＩ／Ｏコントローラを介してシステムに結合することができる。
【００８８】
データ処理システムが、介在する私設ネットワークまたは公共ネットワークを介して、他のデータ処理システムまたはリモート・プリンタもしくはストレージ・デバイスに結合できるように、ネットワーク・アダプタも、システムに結合することができる。モデム、ケーブル・モデム、およびイーサネット・カードは、現在利用可能なタイプのネットワーク・アダプタのうちのいくつかに過ぎない。
【００８９】
本発明の説明は、例示および説明の目的で提示されたものであり、網羅的であること、または開示された形態の本発明に限定することは意図していない。多くの修正および変形が、当業者には明らかであろう。本発明の原理および実際の応用を最もよく説明し、企図された特定の使用に適した様々な修正を有する様々な実施形態のために当業者が本発明を理解できるように、実施形態が選択され、説明された。

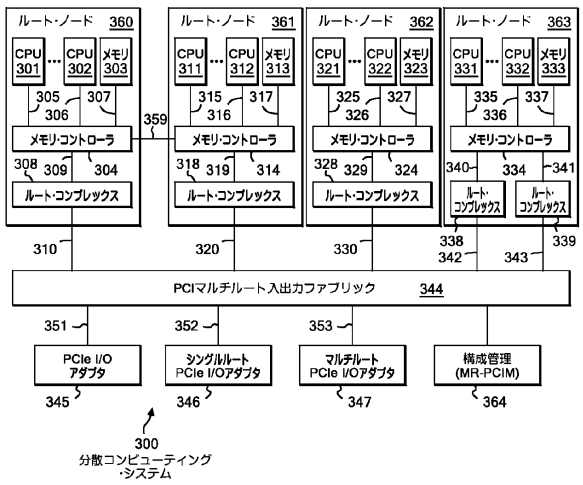
【図１】


100
ブレード・エンクロージャ
101
102
103
104
サーバ・ブレード
PCIe イーサネット・デバイス 105
PCIe ストレージ・デバイス 106
PCIe イーサネット・デバイス 107
PCIe ストレージ・デバイス 108
PCIe イーサネット・デバイス 109
PCIe ストレージ・デバイス 110
PCIe イーサネット・デバイス 111
PCIe ストレージ・デバイス 112
イーサネット・スイッチ 113
ストレージ・エリア・ネットワーク・スイッチ 114
外部接続
115 イーサネット接続
116 ストレージ・エリア・ネットワーク接続

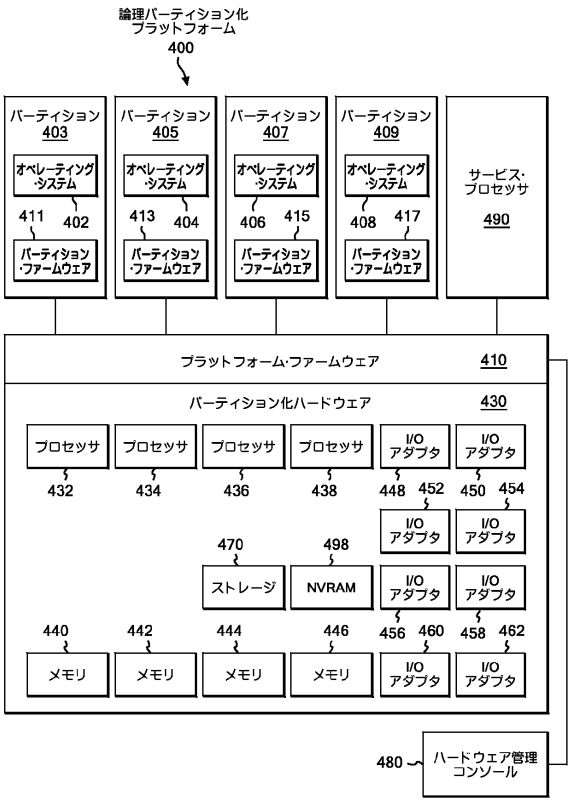
【図２】


200
ブレード・エンクロージャ
201
202
203
204
サーバ・ブレード
PCIe ルート・ポート 205
PCIe ルート・ポート 206
PCIe ルート・ポート 207
PCIe ルート・ポート 208
PCIe ルート・ポート 209
PCIe ルート・ポート 210
PCIe ルート・ポート 211
PCIe ルート・ポート 212
マルチルート PCIe スイッチ 213
マルチルート PCIe スイッチ 214
221
217
218
219
220
PCIe イーサネット・デバイス
PCIe ストレージ・デバイス
外部接続
215
216

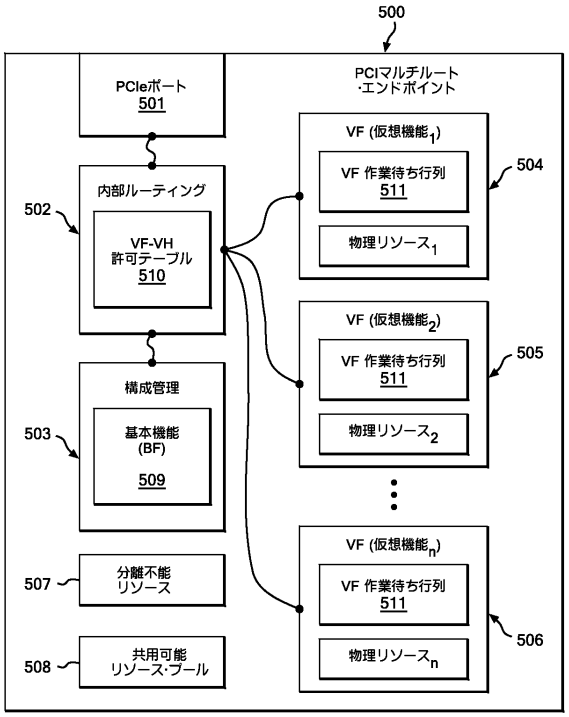
【図３】


ルート・ノード 360
ルート・ノード 361
ルート・ノード 362
ルート・ノード 363
CPU 301
CPU 302
メモリ 303
CPU 311
CPU 312
メモリ 313
CPU 321
CPU 322
メモリ 323
CPU 331
CPU 332
メモリ 333
メモリ・コントローラ
ルート・コンプレックス
359
PCIマルチルート入出力ファブリック 344
PCIe I/O アダプタ 345
シングルルート PCIe I/Oアダプタ 346
マルチルート PCIe I/Oアダプタ 347
構成管理 (MR-PCIM) 364
300 分散コンピューティング・システム

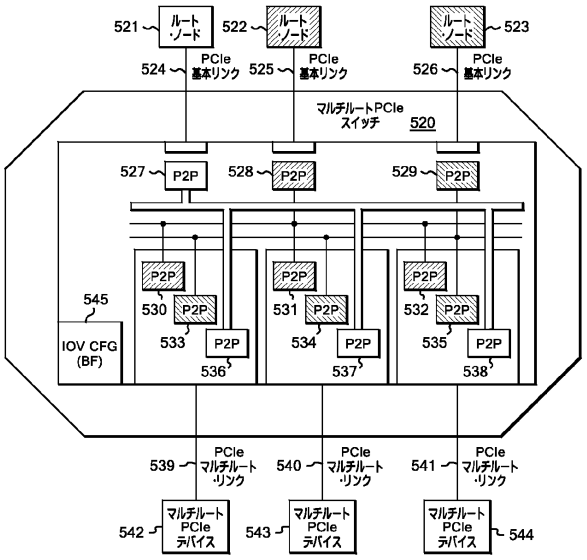
【図４】


論理パーティション化プラットフォーム 400
パーティション 403
パーティション 405
パーティション 407
パーティション 409
オペレーティング・システム
パーティション・ファームウェア
サービス・プロセッサ 490
プラットフォーム・ファームウェア 410
パーティション化ハードウェア 430
プロセッサ
I/O アダプタ
ストレージ
NVRAM
メモリ
ハードウェア管理コンソール 480

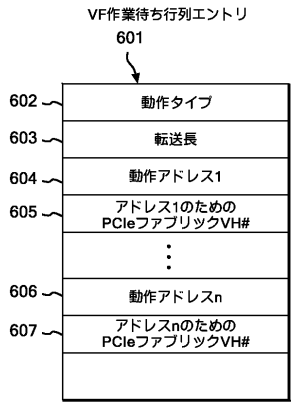
【図５】



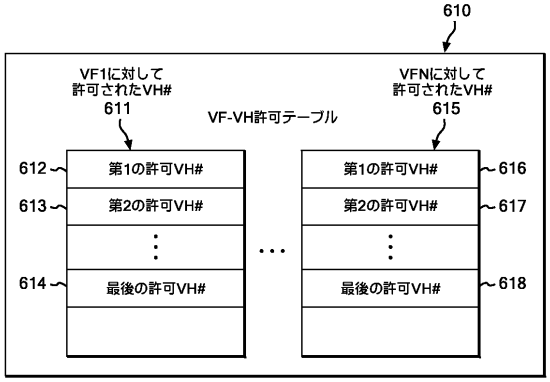
【図６】



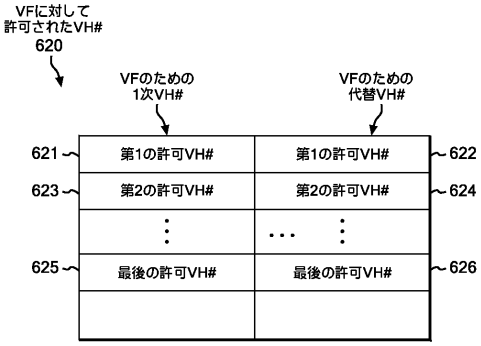
【図７】



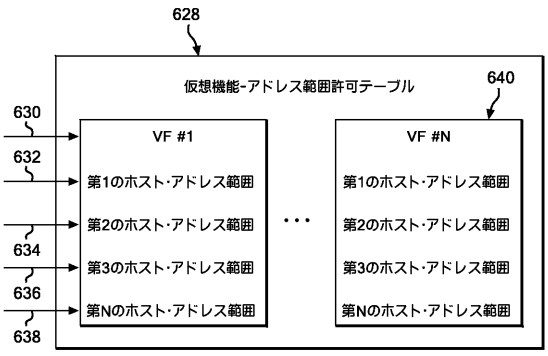
【図８】



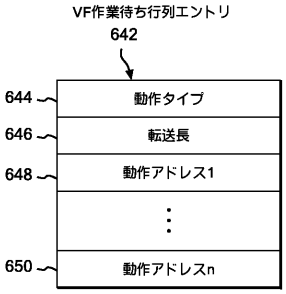
【図９】



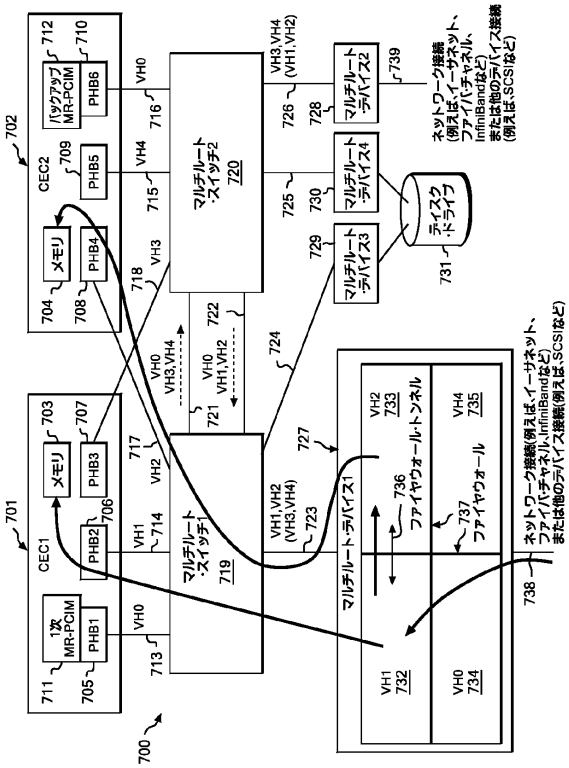
【図１０】



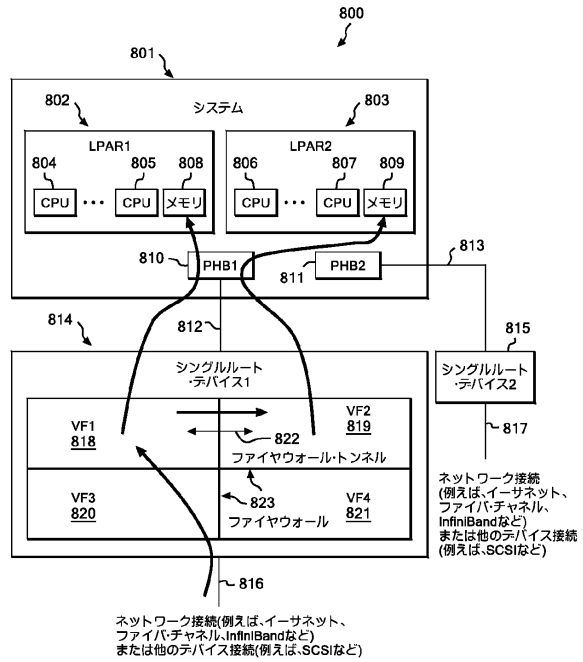
【図１１】



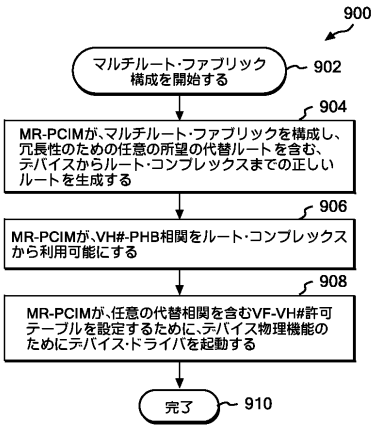
【図１２】



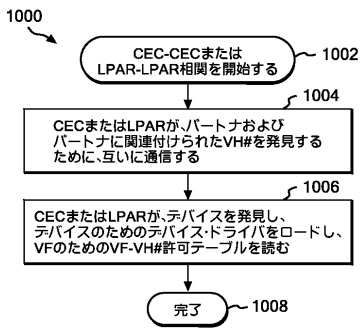
【図１３】



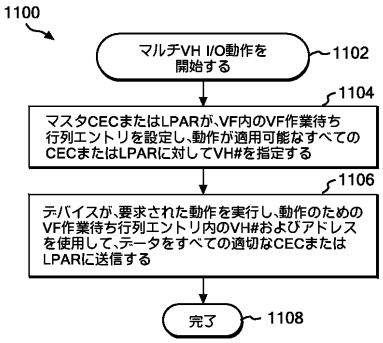
【図１４】



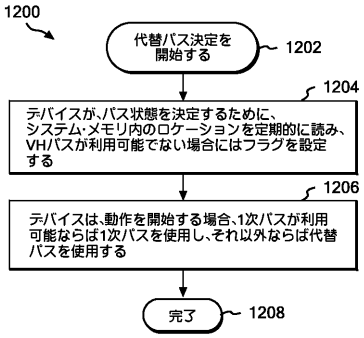
【図１５】



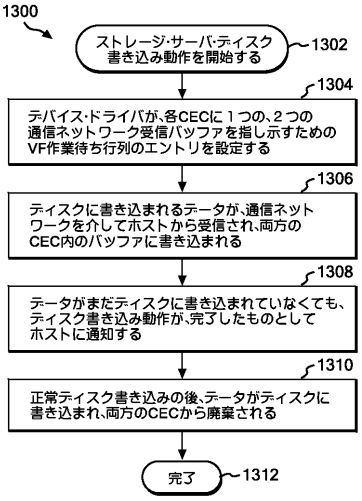
【図１６】


1100
マルチVH I/O動作を開始する
1102
1104
マスタCECまたはLPARが、VF内のVF作業待ち行列エントリを設定し、動作が適用可能なすべてのCECまたはLPARに対してVH#を指定する
1106
デバイスが、要求された動作を実行し、動作のためのVF作業待ち行列エントリ内のVH#およびアドレスを使用して、データをすべての適切なCECまたはLPARに送信する
完了
1108

【図１７】


1200
代替パス決定を開始する
1202
1204
デバイスが、パス状態を決定するために、システム・メモリ内のロケーションを定期的に読み、VHパスが利用可能でない場合にはフラグを設定する
1206
デバイスは、動作を開始する場合、1次パスが利用可能ならば1次パスを使用し、それ以外ならば代替パスを使用する
完了
1208

【図１８】


1300
ストレージ・サーバ・ディスク書き込み動作を開始する
1302
1304
デバイス・ドライバが、各CECに１つの、２つの通信ネットワーク受信バッファを指し示すためのVF作業待ち行列のエントリを設定する
1306
ディスクに書き込まれるデータが、通信ネットワークを介してホストから受信され、両方のCEC内のバッファに書き込まれる
1308
データがまだディスクに書き込まれていなくても、ディスク書き込み動作が、完了したものとしてホストに通知する
1310
正常ディスク書き込みの後、データがディスクに書き込まれ、両方のCECから廃棄される
完了
1312

フロントページの続き

(72)発明者　スティーブ・サーバー
アメリカ合衆国　７８７５８　テキサス州　オースティン　バーネットロード　１１５０１
(72)発明者　ダグラス・エム・フレイマス
アメリカ合衆国　１０５３２　ニューヨーク州　ホーソーン　スカイラインドライブ　１９

審査官　坂東　博司

(56)参考文献　特開２００８－１７１４１３（ＪＰ，Ａ）
特開２００８－０６５５６１（ＪＰ，Ａ）
特開２００４－０１３８９９（ＪＰ，Ａ）
特開２００８－００９９８０（ＪＰ，Ａ）
特開２００９－０９３３１６（ＪＰ，Ａ）
米国特許出願公開第２００５／０２８１１９１（ＵＳ，Ａ１）
米国特許出願公開第２００３／００８８６９８（ＵＳ，Ａ１）
米国特許出願公開第２００９／００９４４０３（ＵＳ，Ａ１）
米国特許第０７３９８３３７（ＵＳ，Ｂ１）
米国特許第０７４６４２１８（ＵＳ，Ｂ１）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１３／１０