

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第5731108号
(P5731108)

(45) 発行日 平成27年6月10日 (2015. 6. 10)

(24) 登録日 平成27年4月17日 (2015. 4. 17)

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

請求項の数 34 (全 26 頁)

(21) 出願番号	特願2009-158812 (P2009-158812)	(73) 特許権者	000004237
(22) 出願日	平成21年7月3日 (2009. 7. 3)		日本電気株式会社
(65) 公開番号	特開2011-14023 (P2011-14023A)		東京都港区芝五丁目7番1号
(43) 公開日	平成23年1月20日 (2011. 1. 20)	(74) 代理人	100109313
審査請求日	平成24年6月18日 (2012. 6. 18)		弁理士 机 昌彦
審判番号	不服2014-6830 (P2014-6830/J1)	(74) 代理人	100124154
審判請求日	平成26年4月14日 (2014. 4. 14)		弁理士 下坂 直樹
		(72) 発明者	鈴木 順
			東京都港区芝五丁目7番1号
			日本電気株式会社内
		(72) 発明者	飛鷹 洋一
			東京都港区芝五丁目7番1号
			日本電気株式会社内

最終頁に続く

(54) 【発明の名称】 中継手段、中継システム、中継方法およびプログラム

(57) 【特許請求の範囲】

【請求項1】

デバイスと、

前記デバイスを共有し、それぞれが前記デバイスに対して設定要求を発行する複数の設定手段と、

前記複数の設定手段と前記デバイスとを接続し、前記複数の設定手段の前記デバイスに対する前記設定要求に含まれる設定情報に関する制限値の範囲に基づき、前記設定要求をすべて満たす制限値を算出し、算出した値を前記デバイスに関する設定値として保存手段に格納するとともに、前記設定要求に基づいて設定したことを表す応答を前記複数の設定手段に送信する中継手段と

を含む、中継システム。

【請求項2】

前記中継手段は、前記設定手段からの参照要求に対して、前記保存手段に保持した前記設定情報を元に、前記設定手段に応答することを特徴とする、請求項1に記載の中継システム。

【請求項3】

前記中継手段は、前記保存手段に保持した前記設定情報を元に、前記デバイスの入力仕様に従ったフォーマットの設定指示を生成し、前記デバイスに対して発行することを特徴とする、請求項1に記載の中継システム。

【請求項4】

前記中継手段は、前記デバイスが発行した通知を受信して前記保存手段に保持し、

- (1) 前記デバイス及び前記設定手段の状態を維持する、
- (2) 前記通知を元に通知情報を生成し、前記設定手段に発行する、
- (3) 前記通知を元に前記デバイスを設定する、

の何れか 1 つの処理を実行することを特徴とする、請求項 1 乃至 3 の何れか 1 項に記載の中継システム。

【請求項 5】

前記複数の設定手段がネットワーク接続手段を介して接続され、前記設定要求がカプセル化されたネットワークパケットを指定されたアドレスに配送するネットワーク手段と、
をさらに備え、

10

前記中継手段は、前記デバイスを前記ネットワーク手段に接続し、前記複数の設定手段による前記デバイスに対する設定情報を仲裁することを特徴とする請求項 1 乃至 4 の何れか 1 項に記載の中継システム。

【請求項 6】

前記中継手段は、

前記設定要求をネットワークパケットにカプセル化し、ネットワークパケットから前記設定要求をデカプセル化する手段と、

前記複数の設定手段から前記デバイスに対して発行された前記設定要求に含まれる設定情報を、前記保存手段に保持し、前記デバイスに対して前記設定情報を仲裁する手段と、

前記設定要求のアドレスと前記デバイスが発行した通知のアドレスと前記設定手段からの参照要求に対する応答のアドレスの少なくとも 1 つを変換する手段とを備えたことを特徴とする請求項 5 に記載の中継システム。

20

【請求項 7】

前記中継手段は、前記デバイスが前記複数の設定手段に対してサービスを開始する前に、前記デバイスの識別情報を取得し、前記識別情報を前記保存手段に保持することを特徴とする、請求項 5 又は 6 に記載の中継システム。

【請求項 8】

前記中継手段は、前記制限値が下限値である場合には前記下限値の最大値を、前記設定値として算出し、前記制限値が上限値である場合には前記上限値の最小値を、前記設定値として算出することを特徴とする、請求項 5 又は 6 に記載の中継システム。

30

【請求項 9】

前記中継手段は、前記複数の設定手段の電力制御の電力制限値を前記保存手段に保持し、前記設定手段からの参照要求に対して、前記デバイスが電力制御されているような返却値を前記設定手段に返すことを特徴とする、請求項 5 又は 6 に記載の中継システム。

【請求項 10】

前記中継手段は、前記設定手段が発行する前記デバイスをリセットするリセット要求に応じて、前記設定手段に割当てられた機能のみをリセットすることを特徴とする、請求項 5 又は 6 に記載の中継システム。

【請求項 11】

前記デバイスは、P C I エクスプレス (P C I e) が定めるシングルルート入出力仮想化及び共有化仕様 (S i n g l e - R o o t I / O V i r t u a l i z a t i o n a n d S h a r i n g S p e c i f i c a t i o n : S R - I O V) に準拠し、仮想関数 (V i r t u a l F u n c t i o n) を単位として前記複数の設定手段のそれぞれに入出力 (I / O : i n p u t a n d o u t p u t) 資源を割当てることによって、前記複数の設定手段から前記デバイスを共有することを特徴とする、請求項 1 乃至 6 の何れか 1 項に記載の中継システム。

40

【請求項 12】

デバイスと、前記デバイスを共有しそれぞれが前記デバイスに対して設定要求を発行する複数の設定手段とを接続する中継手段であって、

前記複数の設定手段の前記デバイスに対する前記設定要求に含まれる設定情報に関する

50

制限値の範囲に基づき、前記設定要求をすべて満たす制限値を算出し、算出した値を前記デバイスに関する設定値として保存手段に格納するとともに、前記設定要求に基づいて設定したことを表す応答を前記複数の設定手段に送信する制御手段

を含む、中継手段。

【請求項 1 3】

前記制御手段は、前記設定手段からの参照要求に対して、前記保存手段に保持した前記設定情報を元に、前記設定手段に応答することを特徴とする、請求項 1 2 に記載の中継手段。

【請求項 1 4】

前記制御手段は、前記保存手段に保持した前記設定情報を元に、前記デバイスの入力仕様に従ったフォーマットの設定指示を生成し、前記デバイスに対して発行することを特徴とする、請求項 1 2 に記載の中継手段。

【請求項 1 5】

前記制御手段は、前記デバイスが発行した通知を受信して前記保存手段に保持し、

- (1) 前記デバイス及び前記設定手段の状態を維持する、
- (2) 前記通知を元に通知情報を生成し、前記設定手段に発行する、
- (3) 前記通知を元に前記デバイスを設定する、

の何れか 1 つの処理を実行することを特徴とする、請求項 1 2 乃至 1 4 の何れか 1 項に記載の中継手段。

【請求項 1 6】

前記制御手段は、前記デバイスを、前記複数の設定手段がネットワーク接続手段を介して接続され前記設定要求がカプセル化されたネットワークパケットを指定されたアドレスに配送するネットワーク手段に接続することを特徴とする請求項 1 2 乃至 1 5 の何れか 1 項に記載の中継手段。

【請求項 1 7】

前記設定要求をネットワークパケットにカプセル化し、ネットワークパケットから前記設定要求をデカプセル化する手段と、

前記複数の設定手段から前記デバイスに対して発行された設定要求に含まれる設定情報を前記保存手段に保持し、前記デバイスに対して前記設定情報を仲裁する手段と、

前記設定要求のアドレスと前記デバイスが発行した通知のアドレスと前記設定手段からの参照要求に対する応答のアドレスの少なくとも 1 つを変換する手段とをさらに備えたことを特徴とする請求項 1 6 に記載の中継手段。

【請求項 1 8】

前記デバイスが前記複数の設定手段に対してサービスを開始する前に、前記デバイスの識別情報を取得し、前記識別情報を前記保存手段に保持することを特徴とする、請求項 1 6 又は 1 7 に記載の中継手段。

【請求項 1 9】

前記制限値が下限値である場合には前記下限値の最大値を、前記設定値として算出し、前記制限値が上限値である場合には前記上限値の最小値を、前記設定値として算出することを特徴とする、請求項 1 6 又は 1 7 に記載の中継手段。

【請求項 2 0】

前記複数の設定手段の電力制御の電力制限値を前記保存手段に保持し、前記設定手段からの参照要求に対して、前記デバイスが電力制御されているような返却値を前記設定手段に返すことを特徴とする、請求項 1 6 又は 1 7 に記載の中継手段。

【請求項 2 1】

前記設定手段が発行するデバイスをリセットするリセット要求に応じて、前記設定手段に割当てられた機能のみをリセットすることを特徴とする、請求項 1 6 又は 1 7 に記載の中継手段。

【請求項 2 2】

前記デバイスは、P C I エクスプレス (P C I e) が定めるシングルルート入出力仮想

10

20

30

40

50

化及び共有化仕様 (Single-Root I/O Virtualization and Sharing Specification: SR-IOV) に準拠し、仮想関数 (Virtual Function) を単位として前記複数の設定手段のそれぞれに入出力 (I/O: input and output) 資源を割当てることによって、前記複数の設定手段から前記デバイスを共有することを特徴とする、請求項 1 2 乃至 1 7 の何れか 1 項に記載の中継手段。

【請求項 2 3】

デバイスと、前記デバイスを共有し、それぞれが前記デバイスに対して設定要求を発行する複数の設定手段とを中継する中継方法であって、

前記複数の設定手段の前記デバイスに対する前記設定要求に含まれる設定情報を保持し、前記保持した前記設定情報に関する制限値の範囲に基づき、前記設定要求をすべて満たす制限値を算出し、算出した値を前記デバイスに関する設定値として保存手段に格納するとともに、前記設定要求に基づいて設定したことを表す応答を前記複数の設定手段に送信することを特徴とする中継方法。

10

【請求項 2 4】

前記設定手段からの参照要求に対して、前記保持した前記設定情報を元に、前記設定手段に応答することを特徴とする、請求項 2 3 に記載の中継方法。

【請求項 2 5】

前記保持した前記設定情報を元に、前記デバイスの入力仕様に従ったフォーマットの設定指示を生成し、前記デバイスに対して発行することを特徴とする、請求項 2 3 に記載の中継方法。

20

【請求項 2 6】

前記デバイスが発行した通知を受信して保持し、

(1) 前記デバイス及び前記設定手段の状態を維持する

(2) 前記通知を元に通知情報を生成し、前記設定手段に発行する、

(3) 前記通知を元に前記デバイスを設定する、

の何れか 1 つの処理を実行することを特徴とする、請求項 2 3 乃至 2 5 の何れか 1 項に記載の中継方法。

【請求項 2 7】

前記デバイスを、前記複数の設定手段がネットワーク接続手段を介して接続され前記設定要求がカプセル化されたネットワークパケットを指定されたアドレスに配送するネットワーク手段に接続することを特徴とする請求項 2 3 乃至 2 6 の何れか 1 項に記載の中継方法。

30

【請求項 2 8】

前記設定要求をネットワークパケットにカプセル化し、ネットワークパケットから前記設定要求をデカプセル化するステップと、

前記複数の設定手段から前記デバイスに対して発行された設定情報を保持し、前記デバイスに対して前記設定情報を仲裁するステップと、

前記設定要求のアドレスと前記デバイスが発行した通知のアドレスと前記設定手段からの参照要求に対する応答のアドレスの少なくとも 1 つを変換するステップとをさらに備えたことを特徴とする請求項 2 7 に記載の中継方法。

40

【請求項 2 9】

前記デバイスが前記複数の設定手段に対してサービスを開始する前に、前記デバイスの識別情報を取得し、前記識別情報を保持することを特徴とする、請求項 2 7 又は 2 8 に記載の中継方法。

【請求項 3 0】

前記制限値が下限値である場合には前記下限値の最大値を、前記設定値として算出し、前記制限値が上限値である場合には前記上限値の最小値を、前記設定値として算出することを特徴とする、請求項 2 7 又は 2 8 に記載の中継方法。

【請求項 3 1】

50

前記複数の設定手段の電力制御の電力制限値を保持し、前記設定手段からの参照要求に対して、前記デバイスが電力制御されているような返却値を前記設定手段に返すことを特徴とする、請求項 27 又は 28 に記載の中継方法。

【請求項 32】

前記設定手段が発行するデバイスをリセットするリセット要求に応じて、前記設定手段に割当てられた機能のみをリセットすることを特徴とする、請求項 27 又は 28 に記載の中継方法。

【請求項 33】

前記デバイスは、P C I エクスプレス (P C I e) が定めるシングルルート入出力仮想化及び共有化仕様 (S i n g l e - R o o t I / O V i r t u a l i z a t i o n a n d S h a r i n g S p e c i f i c a t i o n : S R - I O V) に準拠し、仮想関数 (V i r t u a l F u n c t i o n) を単位として前記複数の設定手段のそれぞれに入出力 (I / O : i n p u t a n d o u t p u t) 資源を割当てることによって、前記複数の設定手段から前記デバイスを共有することを特徴とする、請求項 23 乃至 28 の何れか 1 項に記載の中継方法。

【請求項 34】

デバイスと、前記デバイスを共有し、それぞれが前記デバイスに対して設定要求を発行する複数の設定手段とを中継する中継プログラムであって、

前記複数の設定手段の前記デバイスに対する前記設定要求に含まれる設定情報に関する制限値の範囲に基づき、前記設定要求を全て満たす制限値を算出し、算出した値を前記デバイスに関する設定値として保存手段に格納するとともに、前記設定要求に基づいて設定したことを表す応答を前記複数の設定手段に送信する機能

をコンピュータに行わせることを特徴とする中継プログラム。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は中継手段、中継システム、中継方法およびプログラムに関し、特に入出力 (I / O) デバイスを複数のコンピュータ間で共有する入出力 (I / O) システムおよび入出力 (I / O) 制御方法に関する。

【背景技術】

【0002】

現在、コンピュータ内部の各パーツ間を接続する、入出力 (I / O : i n p u t / o u t p u t) に関する標準のバス規格として、ペリフェラル コンポーネント インタコネク (P C I) が広く使用されている。

【0003】

ハードウェア資源の効率的な利用を目的として、システムイメージ (S I) の多重化を可能にしてこの P C I ハードウェア資源を共有するよう拡張した次世代の規格が、シングルルート入出力仮想化及び共有化仕様 (S i n g l e R o o t I / O V i r t u a l i z a t i o n a n d S h a r i n g S p e c i f i c a t i o n : S R - I O V) により、P C I エクスプレス (P C I e) として規定されている。

【0004】

P C I e においては、ハードウェアとシステムイメージ (S I) との間に仮想化仲裁 (V i r t u a l i z a t i o n I n t e r m e d i a r y : V I) と呼ばれるソフトウェアが仲介する。これにより、ハードウェア側からシステムイメージ (S I) が仮想マシン (V i r t u a l M a c h i n e : V M) とみなされる。

【0005】

このようなシステムにおいて、プラットフォーム資源のオーバーヘッドを削減するためのシステム構成が、非特許文献 1 に提案されている。

【0006】

図 22 は、このような入出力 (I / O) システム構成の一例を示すブロック図である。

10

20

30

40

50

【0007】

図22において、非特許文献1に記載のI/Oシステム900は、ホスト9とI/Oデバイス5からなる。I/Oデバイス5は、PCIeが定めるSR-IOVに対応する仕様である。ホスト9は、演算処理を行うCPU903と、プログラムやデータを格納するメモリ905と、ブリッジ904とを含む。ブリッジ904は、I/Oデバイス5と、メモリ905と、CPU903を相互に接続する。ホスト9は、さらに、CPU903上で動作するN個の仮想マシン(VM)902-1~902-N、及び、これらのN個の仮想マシンを管理する、管理仮想マシン(管理VM)901とを含む。

【0008】

I/Oデバイス5は、単一のホスト内で仮想マシン902-1~902-NがI/Oデバイス5を共有する際、仮想マシン902-1~902-Nが直接アクセスするインタフェース503を備える。直接アクセスすることにより、I/Oデバイス5に関するソフトウェアによる共有処理を省くことができ、共有に関するオーバーヘッドが削減される。

【0009】

I/Oデバイス5は、I/Oデバイス5の全体の設定を管理仮想マシン901から受け付ける物理関数(Physical Function: PF)501と、仮想マシン902-1~902-Nに個別に割り当てられて、仮想マシン902-1~902-Nのそれぞれの設定を受け付ける仮想関数(Virtual Function: VF)502-1~502-Nとからなる。ここで、仮想関数502-1~502-Nは、仮想マシン902-1~902-Nにそれぞれ割り当てられている。物理関数501と仮想関数502-1~502-Nのそれぞれは、PCIeに基づいて割り当てられるID番号のサブフィールドであるファンクション番号で区別される。物理関数501と仮想関数502-1~502-Nは、それぞれ、PCIeに準拠したコンフィグレーション(CFG)レジスタである物理関数コンフィグレーション(PF CFG)レジスタ5011と、仮想関数コンフィグレーション(VF CFG)レジスタ5021-1~5021-Nを備える。

【0010】

PF CFGレジスタ5011は、管理仮想マシン901からI/Oデバイス5に関する設定を受け付ける。また、VF CFGレジスタ5021-1~5021-Nは、それぞれ仮想マシン902-1~902-NからI/Oデバイス5に関する設定を受け付ける。仮想マシン902-1~902-Nは、それぞれに割り当てられた仮想関数502-1~502-Nにアクセスし、I/O要求を発行する。これにより、I/Oデバイス5の内部でI/O機能の使用に関する仲裁が行われる。すなわち、I/Oデバイス5の機能が仮想関数502-1~502-Nを通して仮想マシン902-1~902-Nの間で共有される。

【先行技術文献】

【非特許文献】

【0011】

【非特許文献1】PCI-SIG, Single Root I/O Virtualization and Sharing Specification, Revision 1.0, 2007年9月11日, pp.13-15

【発明の概要】

【発明が解決しようとする課題】

【0012】

しかしながら、このようなI/Oデバイスは、単一ホスト内で管理され、単一ホスト内の複数の仮想マシンの間で共有されることを想定して設計されている。このようなI/Oデバイスを複数のホストにより共有する使用形態の要請が近年増えているが、このようなシステムでは、複数のホストからの設定要求が、I/Oデバイスに対して発せられる。この際に、これらの設定要求が競合すると、I/Oデバイスの設定に障害が発生し、I/Oデバイスの制御ができなくなる。このため、この設定要求の競合は、このようなI/Oデバイスの複数のホストによる共有には大きな障壁となっている。

10

20

30

40

50

【0013】

さらに、ホスト上で動作するソフトウェアも、複数のホストがI/Oデバイスを共有する場合には、I/Oデバイスの設定要求の競合などの障害には対応できず、このような使用形態の実現の障壁となっている。

【0014】

これらの問題を解決するためには、このようなI/Oデバイスの構成と、I/Oデバイスを制御するホスト側のソフトウェアを、複数のホストがI/Oデバイスを共有するような変更が必要となる。しかし、このような変更作業には、I/Oデバイスの構成要素のそれぞれを複数のホストから共有されるように機能を変更・追加する必要があり、膨大な作業を要する。また、I/Oデバイスの設計時から、複数のホストから共有されるように機能

10

【0015】

従って、通常単一のホストで管理されているI/Oデバイスを、必要に応じて複数のホストから共有されるようにするため、I/Oデバイスの外部で、複数のホストからI/Oデバイスへの設定要求を仲裁する処理を行うシステムを構築することが必要である。

【0016】

本発明の目的は、単一のホストで管理されるI/Oデバイスに改変を加えることなく、I/Oデバイスが複数のホストから共有されるシステムを提供することである。

20

【課題を解決するための手段】

【0017】

本発明の中継システムは、デバイスと、デバイスを設定する設定手段と、設定手段とデバイスとを接続する中継手段とを含む中継システムであって、中継手段は設定手段のデバイスに対する設定要求に含まれる設定情報を保存手段に保持し、保存手段に保持した設定情報を元にデバイスを設定することを特徴とする。

【0018】

本発明の中継手段は、デバイスと、デバイスを設定する設定手段とを接続する中継手段であって、設定手段のデバイスに対する設定要求に含まれる設定情報を保持する保存手段と、保存手段に保持した設定情報を元にデバイスを設定する制御手段とを含むことを特徴とする。

30

【0019】

本発明の中継方法は、デバイスと、デバイスを設定する設定手段とを中継する中継方法であって、設定手段のデバイスに対する設定要求に含まれる設定情報を保持し、保持した設定情報を元にデバイスを設定することを特徴とする。

【0020】

本発明の中継プログラムは、デバイスと、デバイスを設定する設定手段とを中継する中継プログラムであって、設定手段のデバイスに対する設定要求に含まれる設定情報を保持する保持処理と、保持した設定情報を元にデバイスを設定するデバイス設定処理とをコンピュータに行わせることを特徴とする。

40

【発明の効果】

【0021】

本発明は単一のホスト内で管理されるI/Oデバイスを複数のホストにより共有されるシステムに関する。本発明によれば、ホストに不要な処理を実行させることなく、I/Oデバイスの状態に応じた最適な設定が可能になる。

【図面の簡単な説明】

【0022】

【図1】本発明の第1の実施形態の構成を示すブロック図と流れ図である。

【図2】本発明の第1の実施形態の動作を示す流れ図である。

【図3】本発明の第1の実施形態の動作を示す流れ図である。

50

【図 4】本発明の第 2 の実施形態の構成を示すブロック図である。

【図 5】本発明の第 2 の実施形態におけるネットワーク処理部の構成を示すブロック図である。

【図 6】本発明の第 2 の実施形態における I / O パケット転送部の構成を示すブロック図である。

【図 7】本発明の第 2 の実施形態における I / O 仮想化部の構成を示すブロック図である。

【図 8】本発明の第 2 の実施形態におけるアドレス変換部の構成を示すブロック図である。

【図 9】本発明の第 2 の実施形態の動作を示す流れ図である。 10

【図 10】本発明の第 2 の実施形態の動作を示す流れ図である。

【図 11】本発明の第 2 の実施形態の動作を示す流れ図である。

【図 12】本発明の第 2 の実施形態の動作を示す流れ図である。

【図 13】本発明の第 3 の実施形態の構成を示すブロック図である。

【図 14】本発明の第 3 の実施形態における I / O 仮想化部の構成を示すブロック図である。

【図 15】本発明の第 3 の実施形態の動作を示す流れ図である。

【図 16】本発明の第 4 の実施形態の構成を示すブロック図である。

【図 17】本発明の第 4 の実施形態における I / O 仮想化部の構成を示すブロック図である。 20

【図 18】本発明の第 4 の実施形態の動作を示す流れ図である。

【図 19】本発明の第 5 の実施形態における I / O 仮想化ブリッジの構成を示すブロック図である。

【図 20】本発明の第 5 の実施形態における I / O 仮想化部の構成を示すブロック図である。

【図 21】本発明の第 5 の実施形態の動作を示す流れ図である。

【図 22】関連技術を説明するための図である。

【発明を実施するための形態】

【0023】

(第 1 の実施形態)

次に、発明を実施するための形態について、図面を参照して、詳細に説明する。 30

【0024】

本発明の第 1 の実施形態による中継システムについて、図面を参照して詳細に説明する。図 1 は本発明の実施形態による中継システムの構成を示すブロック図である。

【0025】

図 1 を参照すると、本発明の第 1 の実施形態による中継システム 1000 は、設定手段 1001 と、設定手段 1001 が設定しようとするデバイス 1002 と、設定手段 1001 とデバイス 1002 を接続する中継手段 1003 とを含む。

【0026】

中継手段 1001 は、制御手段 1004 と、記憶手段 1005 とを含む。 40

【0027】

設定手段 1001 は、デバイス 1002 に対して設定要求を送信する。中継手段 1003 は、設定手段 1001 が発行した設定要求を受信し、制御手段 1004 により設定要求を処理する。設定要求は設定情報を含む。ここで、設定情報は、例えば、デバイスに対して何らかの状態を設定するための情報である。

【0028】

設定手段 1001 はまた、参照要求をデバイス 1002 に対して送信し、デバイス 1002 は、受信した参照要求に対して応答を返す。

【0029】

本発明の第 1 の実施形態による中継手段 1003 は、設定手段 1001 が発信した設定 50

情報を受信し、また、設定手段1001が発信した参照要求を受信する。さらに、中継手段1003は、設定手段1001に対して、デバイス1002が設定手段1001に返す応答と同等の応答を設定手段1001に返す。

【0030】

さらにまた、中継手段1003は、デバイス1002の設定に関する情報を用いてデバイス1002を設定する。或は、中継手段1003は、設定手段1001がデバイス1002に対して発行する設定要求と同等の設定指示を、デバイス1002に対して発行してもよい。

【0031】

中継手段1003が備える制御手段1004は、設定手段1001が発行した設定要求を受信して、設定要求に含まれる設定情報を記憶手段1005に保持することができる。また、記憶手段1005に保持した設定情報を元に、デバイス1002に設定指示を発行することができる。

10

【0032】

次に、図1乃至3を参照して、本発明を実施するための第1の実施形態における中継システム1000の動作を説明する。

【0033】

設定手段1001は、デバイス1002宛に信号を発行する(図2のステップS101)。

【0034】

設定手段1001が発行した信号がデバイス1002に対する設定要求であれば(ステップS102のYES)、中継手段1003は、設定要求を受け、制御手段1004により設定要求に含まれる設定情報を記憶手段1005に保持する(ステップS103)。

20

【0035】

設定手段1001が発行した信号が、デバイス1002への参照要求であれば(ステップS102のNO)、制御手段1004は記憶手段1005に保持した設定情報を元に、設定手段1001に応答を返す(ステップS104)。設定手段1001への応答は、例えば、デバイス1002の実際の設定状態にかかわらず、デバイス1002が設定手段1001の発行した設定要求の通りに設定されている、という内容を含んでも良い。

【0036】

制御手段1004は、デバイス1002を設定する場合(ステップS105のYES)は、記憶手段1005の保持した設定情報を元に設定に関する情報を生成し、この設定に関する情報に基づいてデバイス1002を設定する(ステップS106)。また、デバイス1002の入力の仕様によっては、設定手段1001がデバイス1002に対して発行する設定要求と同等の設定指示を、中継手段1003がデバイス1002に対して発行することができる。この場合、制御手段1004は、デバイス1002の設定に関する情報を含む設定指示を、設定手段1001が発行する設定要求のフォーマットに変換して、デバイス1002に対して発行する。

30

【0037】

また、上述のデバイス1002の設定処理によりデバイス1002に問題が生じた場合(図3のステップS107のYES)は、デバイス1002から設定手段1001に宛てた通知を中継手段1003が受信する(ステップS108)。

40

【0038】

中継手段1003の制御手段1004は通知を受けて、この通知を制御する。すなわち、この通知を破棄して放置する(ステップS110)か、デバイスを再設定するかを判断する(ステップS109)。

【0039】

また、制御手段1003は、デバイス1002を制御手段1003により再設定する(ステップS111)か、デバイス1002からの通知を設定手段1001に送信してデバイス1002の再設定を要求するか(ステップS112)を判断する(ステップS113

50

）。なお、デバイス1002の再設定は、問題が生じる設定要求を受信する前の状態に戻してもよい。

【0040】

第1の実施形態において、中継手段1003は、設定手段1001がデバイス1002に対して発行した設定要求を元に設定情報を記憶手段1005に保持する。中継手段1003は、この設定情報に基づいて、デバイス1002の設定を制御する。設定手段1001に対しては、照会に対して、デバイス1002が設定済みであると応答する。これにより、設定手段1001に不要な処理を実行させることなく、デバイス1002の状態に応じた最適な設定ができる。また、通信回線への不要な負荷を抑制することができる。

【0041】

(第2の実施形態)

本発明の第2の実施形態による入出力(I/O)システム(中継システム)について図面を参照して詳細に説明する。図4は本発明に実施形態によるI/Oシステムの構成を示すブロック図である。

【0042】

なお、I/Oデバイスの構成は、図22と同じであるので、各部材に同じ参照番号を付して、説明を省略する。

【0043】

第2の実施形態では、複数のホスト(設定手段)がI/Oデバイス(デバイス)を共有する。

【0044】

図4を参照すると、本発明の第2の実施形態によるI/Oシステム100は、ホスト1-1~1-Nと、ホスト1-1~1-Nによって共有されるI/Oデバイス5と、ホスト1-1~1-NとI/Oデバイス5とを相互に接続するネットワーク3と、ホスト1-1~1-NのI/Oデバイス5に対する設定要求を仲裁するI/O仮想化ブリッジ4(中継手段)とを含む。I/O仮想化ブリッジ4は、I/Oデバイス5をネットワーク3に接続する。このI/Oシステムは、さらに、ホスト1-1~1-Nをネットワーク3に接続するホストブリッジ2-1~2-Nを含む。

【0045】

ホスト1-1~1-Nは、それぞれ、演算処理を行うCPU101-1~101-Nと、プログラムやデータを格納するメモリ103-1~103-Nと、ブリッジ102-1~102-Nとを含む。ブリッジ102-1~102-Nは、ホストブリッジ2-1~2-Nと、メモリ103-1~103-Nと、CPU101-1~101-Nを相互に接続する。

【0046】

ホストブリッジ2-1~2-Nは、それぞれホスト1-1~1-Nが発行するI/Oパケットをネットワーク3で定められたパケット(以下ネットワークパケットと称する)の仕様(フォーマット)にカプセル化し、ネットワーク3に送信する。なお、PCIeでは特に、I/Oパケットをトランザクションレイヤパケット(Transaction Layer Packet: TLP)と呼ぶ。以下では、このようなI/OパケットをTLPと呼ぶ。

【0047】

また、ホストブリッジ2-1~2-Nは、TLPがカプセル化されたネットワークパケットを受信して、TLPをデカプセル化する。このTLPはI/Oデバイス5から送信され、ホストブリッジ2-1~2-Nがそれぞれ接続するホスト1-1~1-N宛である。デカプセル化されたTLPは、ホスト1-1~1-Nのうちの送信先として指定されたホストに送信される。

【0048】

ネットワーク3は、TLPがカプセル化されたネットワークパケットを、宛先として指定されたホストに接続されるノードに送信する。

10

20

30

40

50

【 0 0 4 9 】

I/O仮想化ブリッジ4は、ネットワーク処理部401と、I/Oパケット転送部402と、I/O仮想化部403と、アドレス変換部404と、仮想CFGレジスタ405-1~405-Nを含む。ネットワーク処理部401は、TLPのカプセル化およびデカプセル化処理を行う。I/Oパケット転送部402は、TLPの種類を判定し、指定された宛先にTLPを転送する。I/O仮想化部403は、ホスト1-1~1-NによるI/Oデバイス5への設定を仲裁する。アドレス変換部404は、TLPの宛先アドレスを変換する。仮想CFGレジスタ405-1~405-Nは、ホスト1-1~1-Nに対して、I/Oデバイス5の仮想関数コンフィグレーション(VF CFG)レジスタを仮想的に提供する。

10

【 0 0 5 0 】

図5乃至8を参照して、I/O仮想化ブリッジ4を構成する各部の働きを次に説明する。

【 0 0 5 1 】

図5に示すネットワーク処理部401は、ホスト1-1~1-Nから送信された設定要求を含むTLPがカプセル化されたネットワークパケットを、ネットワークパケット送受信部413によりネットワーク3から受信する。デカプセル化部411は、このネットワークパケットからTLPをデカプセル化し、このTLPをI/Oパケット転送部402に転送する。また、ネットワーク処理部401は、I/Oパケット転送部402からホスト1-1~1-N宛のTLPを受信し、カプセル化部412によりTLPをカプセル化する。ネットワークパケット送受信部413は、カプセル化したTLPをネットワーク3に送信する。ネットワーク処理部401は、アドレス保持部414に保持されている、宛先ホスト1-1~1-Nが接続するホストブリッジ2-1~2-Nのネットワークアドレスを用いてTLPのカプセル化をする。

20

【 0 0 5 2 】

図6に示すI/Oパケット転送部402は、ホスト1-1~1-Nが発行したTLPをネットワーク処理部401から受信し、パケット選択部421により、これらのTLPから、I/O仮想化部403において仲裁されるI/Oデバイス5の設定パケットを選択する。第1転送部422は、選択されたTLP(以後、設定TLPと言う。)をI/O仮想化部403に転送する。また、第1転送部422は、ホスト1-1~1-Nが発行したTLPのうち、選択されなかったTLPを、アドレス変換部404に転送する。さらに、I/Oパケット転送部402は、I/O仮想化部403とアドレス変換部404とから、ホスト1-1~1-NとI/Oデバイス5のいずれかに宛てたTLPを受信する。第2転送部423は、受信したTLPを、指定された宛先へ転送する。

30

【 0 0 5 3 】

図7に示すI/O仮想化部403は、ホスト1-1~1-NによるI/Oデバイス5への設定を仲裁する。すなわち、I/O仮想化部403は、I/Oデバイス5にそのまま設定されると各ホストからの設定が競合するような設定を仲裁する。I/O仮想化部403は、ホスト1-1~1-Nが発行したI/Oデバイス5の設定TLPをI/Oパケット転送部402から受信する。パケット仲裁部431は、設定TLPを元にして得られた設定値を仮想CFGレジスタ405-1~405-Nに登録する。

40

【 0 0 5 4 】

I/O仮想化部403は、仮想CFGレジスタに登録された設定値を元にI/Oデバイス5への設定を仲裁する。すなわち、以下の3つの少なくともいずれか1つの処理を行う。

(1) パケット仲裁部431は、I/Oデバイス5に対しては設定処理を行わないが、ホスト1-1~1-Nに対してはI/Oデバイス5が設定された状態にする。

(2) I/Oデバイス設定値生成・変換部432は、ホスト1-1~1-Nから仮想CFGレジスタ405-1~405-Nに登録した設定情報に対し、これらの設定情報が要求する全ての条件を満たす設定値を求め、得られた設定値に基づいてI/Oデバイス5の物

50

理関数コンフィグレーション (P F C F G) レジスタ 5 0 1 1 を設定する。例えば、設定指示を生成し、I / O デバイス 5 に対して発行する。

(3) I / O デバイス 5 の物理関数 5 0 1 が仮想関数 5 0 2 - 1 ~ 5 0 2 - N を制御するインタフェースを実装している場合には、I / O デバイス設定値生成・変換部 4 3 2 は、(2) で求めた設定値を、I / O デバイス 5 の P F C F G レジスタ 5 0 1 1 を設定する設定 T L P (設定指示) に変換し、変換した設定 T L P を I / O デバイス 5 に発行する。なお、この設定 T L P は、ホスト 1 - 1 ~ 1 - N が発行した設定 T L P (設定要求) とは異なったものになる。

【 0 0 5 5 】

なお、上記 (1) での、ホスト 1 - 1 ~ 1 - N に対して I / O デバイス 5 が設定された状態は、ホスト 1 - 1 ~ 1 - N の少なくとも何れか 1 つが I / O デバイス 5 にリード要求を発行したときに、仮想 C F G レジスタが I / O デバイス 5 の設定状態を返却値としてホストに返すことにより実現する。

【 0 0 5 6 】

I / O デバイス 5 に設定をしているか否かによらず、ホスト 1 - 1 ~ 1 - N は I / O デバイス 5 が設定されているという返却値を受ける。

【 0 0 5 7 】

I / O 仮想化部 4 0 3 はまた、ホスト 1 - 1 ~ 1 - N を設定する、I / O デバイス 5 がホスト 1 - 1 ~ 1 - N へ発行した設定 T L P (通知) を仲裁する。ホスト 1 - 1 ~ 1 - N からの設定要求により、I / O デバイス 5 でエラーなどが発生し、I / O デバイス 5 を共有する全てのホスト 1 - 1 ~ 1 - N に影響を与えるような場合に、I / O デバイス 5 がホスト 1 - 1 ~ 1 - N に対してこのような設定 T L P を発行する。I / O 仮想化部 4 0 3 は、この設定 T L P を仲裁する。

【 0 0 5 8 】

すなわち、I / O 仮想化部 4 0 3 は、以下の 3 つの少なくともいずれか 1 つの処理を行い、ホスト 1 - 1 ~ 1 - N への設定 T L P を仲裁する。

(1 ') パケット仲裁部 4 3 1 は、I / O デバイス 5 が発行した設定 T L P を I / O パケット転送部 4 0 2 から受信し、受信した T L P を破棄して放置する。

(2 ') ホスト通知生成部 4 3 4 は、全てのホスト 1 - 1 ~ 1 - N 宛にイベントを通知する T L P を作成し、作成した T L P を I / O パケット転送部 4 0 2 を経由してホスト 1 - 1 ~ 1 - N に送信する。

(3 ') I / O デバイス設定値生成・変換部 4 3 2 は、問題を生じる設定の代わりに、その設定を基にして I / O デバイス 5 を制御する設定 T L P を独自に作成し、作成した T L P を基に I / O デバイス 5 の P F C F G レジスタ 5 0 1 1 を設定する。なお、この再設定は、I / O デバイスの設定を、ホスト 1 - 1 ~ 1 - N が設定 T L P を発行する前の状態に戻してもよい。

【 0 0 5 9 】

また、I / O 仮想化部 4 0 3 は、デバイス情報登録部 4 3 3 により、I / O デバイス 5 の物理関数を含む P F C F G レジスタ 5 0 1 1 のみが保持しているデバイス情報を取得し、このデバイス情報を仮想 C F G レジスタ 4 0 5 - 1 ~ 4 0 5 - N に登録してもよい。これにより、ホスト 1 - 1 ~ 1 - N が仮想 C F G レジスタ 4 0 5 - 1 ~ 4 0 5 - N をリード (R E A D) した場合に、デバイス 5 のデバイス情報が取得できる。

【 0 0 6 0 】

仮想 C F G レジスタ 4 0 5 - 1 ~ 4 0 5 - N は、I / O デバイス 5 の各 V F C F G レジスタを、ホスト 1 - 1 ~ 1 - N に、仮想的に提供する。すなわち、ホスト 1 - 1 ~ 1 - N からのリード要求に対して、仮想 C F G レジスタが I / O デバイス 5 の V F C F G レジスタの情報を返すことができる。仮想 C F G レジスタ 4 0 5 - 1 ~ 4 0 5 - N は I / O デバイス 5 が含む仮想関数 5 0 2 - 1 ~ 5 0 2 - N にそれぞれ対応し、ホスト 1 - 1 ~ 1 - N からのアクセスをそれぞれ受け付ける。仮想 C F G レジスタ 4 0 5 - 1 ~ 4 0 5 - N は、ホスト 1 - 1 ~ 1 - N が I / O デバイス 5 に設定した設定情報を、例えばデバイス情

10

20

30

40

50

報登録部433を経由して登録し、保持する。また、仮想CFGレジスタ405-1~405-Nは、ホスト1-1~1-Nからリード要求があれば、保持されている値を返す。仮想CFGレジスタ405-1~405-Nはまた、ホスト1-1~1-NがI/Oデバイス5に割当てたアドレスとメモリ領域とI/O領域の情報を、例えばデバイス情報登録部433を経由して登録し、保持する。

【0061】

図8に示すアドレス変換部404は、ホスト1-1~1-NがI/Oデバイス5に発行した設定要求を含むTLPのうち、I/O仮想化部403において仲裁されるI/Oデバイス5の設定パケットとして選択されなかった設定TLPを受信する。上述のように、仮想CFGレジスタ405-1~405-Nには、ホスト1-1~1-NがI/Oデバイス5に割当てたアドレスとメモリ領域とI/O領域の情報が登録されている。アドレス参照部442は、これらの登録情報を参照し、ヘッダアドレス変換部441により、TLPの宛先アドレスを、設定されているI/Oデバイス5のアドレスとメモリ領域とI/O領域を元に変換する。アドレス変換部404は、宛先アドレスが変換された設定TLPを、I/Oパケット転送部402を経由して、I/Oデバイス5に送信する。また、アドレス変換部404は、I/Oデバイス5がホスト1-1~1-Nに対して発行した通知を含むTLPを受信する。アドレス参照部442は、仮想CFGレジスタ405-1~405-Nに登録された情報を参照し、ヘッダアドレス変換部441は、TLPの宛先アドレスを変換する。アドレス変換部404は、I/Oパケット転送部402を経由して、変換したTLPを、ホスト1-1~1-Nのうち、TLPの宛先に指定されているホストに送信する。

【0062】

I/Oデバイス5は、仮想関数502-1~502-Nを、それぞれホスト1-1~1-Nに割当てた。これにより、ホスト1-1~1-Nは、I/Oデバイス5の機能を共有する。

【0063】

次に、図4、図9及び図10を参照して、本発明を実施するための第2の実施形態において、ホスト1-1からI/Oデバイス5へTLPを発行する動作を詳細に説明する。なお、この他のホスト1-2~1-Nの何れか1つからI/Oデバイス5へTLPを発行する場合の動作も同様である。

【0064】

ホスト1-1は、I/Oデバイス5宛のTLPを発行する(図9のステップS201)。ホストブリッジ2-1は、ホスト1-1が発行したTLPをI/O仮想化ブリッジ4のネットワークアドレスを宛先としてカプセル化し(ステップS202)、カプセル化したTLPを含むネットワークパケットをネットワーク3に送信する(ステップS203)。

【0065】

ネットワーク処理部401は、TLPがカプセル化されたネットワークパケットを受信し、TLPをデカプセル化し(ステップS204)、TLPをI/Oパケット転送部402に転送する。I/Oパケット転送部402は、TLPがI/O仮想化部403が仲裁する設定TLPかどうかを判定し(ステップS205)、そのTLPがI/O仮想化部403が仲裁する設定TLPであれば、I/O仮想化部403へTLPを転送する。I/Oパケット転送部402は、TLPがI/O仮想化部403が仲裁する設定TLPでなければ、TLPをアドレス変換部404に転送する。

【0066】

仮想CFGレジスタ405-1には、登録されているホスト1-1がI/Oデバイス5に割当てたアドレスとメモリ領域とI/O領域の情報が登録されている。アドレス変換部404は、この仮想CFGレジスタ405-1に登録されている情報を参照し、TLPのヘッダのアドレスを、I/Oデバイス5の仮想関数502-1に割当てられたアドレスに変換する(ステップS205)。アドレス変換部404は、このTLPを、I/Oデバイス5へ送信する(ステップS207)。

【 0 0 6 7 】

一方、ステップ S 2 0 5 において、T L P が I / O 仮想化部 4 0 3 が仲裁する設定 T L P であった場合の I / O 仮想化部 4 0 3 の処理を、図 1 0 を参照して詳しく説明する。

【 0 0 6 8 】

I / O 仮想化部 4 0 3 は、ホスト 1 - 1 ~ 1 - N からの T L P がライト (W R I T E) 、すなわち設定要求であれば (図 1 0 のステップ S 2 1 1) 、次の 3 つのうちの少なくとも一つの処理をする。

(1) I / O デバイス 5 を設定するかどうか判定し (ステップ S 2 1 3) 、設定しない場合は、仮想 C F G レジスタ 4 0 5 - 1 に設定情報をライトし、I / O デバイス 5 に対して設定処理を行わないが、ホスト 1 - 1 に対して I / O デバイス 5 が設定された状態にする (ステップ S 2 1 4) 。

(2) I / O デバイス 5 を設定する場合は、ホスト 1 - 1 ~ 1 - N が仮想 C F G レジスタ 4 0 5 - 1 ~ 4 0 5 - N に登録した設定情報に対し、これらの設定情報が要求する全ての条件を満たす設定値を求める。I / O デバイス 5 にインタフェースが実装されるかどうかを判定し (ステップ S 2 1 5) 、インタフェースが具備されない場合は、得られた設定値に基づいて I / O デバイス 5 の P F C F G レジスタ 5 0 1 1 に設定指示を送り、直接設定する (ステップ S 2 1 6) 。

(3) I / O デバイス 5 の物理関数 5 0 1 が仮想関数 5 0 2 - 1 ~ 5 0 2 - N を制御するインタフェースを実装している場合には、(2) で求めた設定値を、P F C F G レジスタ 5 0 1 1 を設定する設定 T L P に変換し、変換した設定 T L P を設定指示として I / O デバイス 5 に発行する (ステップ S 2 1 7) 。

【 0 0 6 9 】

一方、ホスト 1 - 1 ~ 1 - N からの T L P がリード (参照要求) であれば、I / O 仮想化部 4 0 3 は、仮想 C F G レジスタ 4 0 5 - 1 の所定の位置から設定情報をリードし、リードされた設定情報を元に、ホスト 1 - 1 に応答を返す (ステップ S 2 1 2) 。

【 0 0 7 0 】

次に、図 4 、図 1 1 及び図 1 2 を参照して、本発明を実施するための第 2 の実施形態において、I / O デバイス 5 からホスト 1 - 1 へ T L P を発行する動作を詳細に説明する。I / O デバイス 5 からこの他のホスト 1 - 2 ~ 1 - N の何れか 1 つへ T L P を発行する場合の動作も同様である。

【 0 0 7 1 】

ホスト 1 - 1 に対する処理は、I / O デバイス 5 では仮想関数 5 0 2 - 1 に割当てられている。この仮想関数 5 0 2 - 1 は、ホスト 1 - 1 宛の T L P を発行する (図 1 1 のステップ S 3 0 1) 。I / O 仮想化ブリッジ 4 の I / O パケット転送部 4 0 2 は、この T L P を受信し、この T L P が I / O 仮想化部 4 0 3 が仲裁する設定 T L P かどうかを判定する (ステップ S 3 0 2) 。この T L P が I / O 仮想化部 4 0 3 が仲裁する設定 T L P であれば、I / O パケット転送部 4 0 2 は、I / O 仮想化部 4 0 3 へ T L P を転送する。この T L P が I / O 仮想化部 4 0 3 が仲裁する設定 T L P でなければ、パケット転送部 4 0 2 は、アドレス変換部 4 0 4 に T L P を転送する。

【 0 0 7 2 】

ホスト 1 - 1 が I / O デバイス 5 に割当てたアドレスとメモリ領域と I / O 領域の情報は、仮想 C F G レジスタ 4 0 5 - 1 に登録されている。アドレス変換部 4 0 4 は、この登録された情報を参照し、転送された T L P のヘッダのアドレスを、この登録情報に基づいて変換する (ステップ S 3 0 3) 。

【 0 0 7 3 】

続いて、ネットワーク処理部 4 0 1 は、ホストブリッジ 2 - 1 のネットワークアドレスを宛先として T L P をカプセル化し (ステップ S 3 0 4) 、カプセル化した T L P を含むネットワークパケットをネットワーク 3 に送信する (ステップ S 3 0 5) 。

【 0 0 7 4 】

ホストブリッジ 2 - 1 は、T L P がカプセル化されたネットワークパケットを受信し、

10

20

30

40

50

T L Pをデカプセル化する(ステップS 3 0 6)。ホストブリッジ2 - 1はさらに、デカプセル化されたT L Pをホスト1 - 1に転送する(ステップS 3 0 7)。

【0075】

一方、ステップS 3 0 2において、受信したT L Pが、I / O仮想化部4 0 3が仲裁する設定T L Pであった場合のI / O仮想化部4 0 3の処理を、図12を参照して説明する。I / O仮想化部4 0 3は、次の3つのうちの少なくとも一つの処理を行う。

(1') I / Oデバイス5を再設定するかどうかを判定し(図12のステップS 3 1 1)、再設定しない場合は受信したT L Pを破棄して放置する(ステップS 3 1 2)。

(2') I / Oデバイス5を再設定する場合、設定処理をホスト1 - 1 ~ 1 - Nが担当するかI / O仮想化部4が担当するかを判断する(ステップS 3 1 3)。ホストが再設定の処理を担当する場合、全てのホスト1 - 1 ~ 1 - N宛の通知T L Pを作成し、この作成されたT L Pを、I / Oパケット転送部4 0 2を経由して、ホスト1 - 1 ~ 1 - Nに送信する(ステップS 3 1 4)。ホスト1 - 1 ~ 1 - Nは、これらの通知T L Pを受信して、I / Oデバイス5で発生している問題に対応して、I / Oデバイス5への設定を調整する。

(3') I / Oデバイス5の再設定をI / O仮想化部が担当する場合、受信した設定T L Pを元にしてI / Oデバイス5を制御する設定T L P(設定指示)を独自に作成し、作成したT L Pを基に、P F C F Gレジスタ5 0 1 1を再設定する(ステップS 3 1 5)。なお、この再設定は、I / Oデバイス5を、ホスト1 - 1 ~ 1 - Nが設定要求を含む設定T L Pを発行する前の状態に戻しても良い。

【0076】

次に、本発明を実施するための第2の実施形態の効果を説明する。

【0077】

第2の実施形態では、I / Oデバイスを共有する複数ホストからのI / Oデバイスに対する設定を、(1) I / Oデバイスへの設定処理を行わずに、ホストに対しては設定した状態にする、(2) 全てのホストからの設定要求を満たす設定値を、実際にI / Oデバイスを設定する設定値として採用する、或は、(3) I / Oデバイスが実装する管理インタフェースに対する設定指示に変換する、などの処理により、複数ホストからのI / Oデバイスに対する設定の競合を仲裁する。また、I / Oデバイスからホストに宛てた通知に対しては、(1') I / Oデバイスからホストへの通知を破棄する、(2') 通知をコピーして全てのホストに通知する、或は、(3) I / Oデバイスの設定を変更する、などの処理を行う。これらの処理により、I / Oデバイスとホストのそれぞれのソフトウェアを変更せずに、複数のホストがI / Oデバイスを共有する。

【0078】

これにより、単一ホスト内で管理され、単一ホスト内の複数の仮想マシンの中で共有されることを想定して設計されているI / Oデバイスを、複数のホストで共有するI / Oシステムにおいて、ホスト1 - 1 ~ 1 - Nに不要な処理を実行させること無く、I / Oデバイスの状態に応じた最適な設定ができる。また、通信回線への不要な負荷を抑制することができ、システムの稼働率や処理効率を高く維持することができる。

(第3の実施形態)

次に、本発明を実施するための第3の実施形態について図面を参照して詳細に説明する。

【0079】

図13に示す本発明を実施するための第3の実施形態によるI / O仮想化ブリッジ6は、第2の実施形態におけるI / O仮想化部4 0 3に換えて、I / O仮想化部6 0 1を備える。さらに、I / O仮想化ブリッジ6は、第2の実施形態における仮想C F Gレジスタ4 0 5 - 1 ~ 4 0 5 - Nに換えて、仮想C F Gレジスタ6 0 5 - 1 ~ 6 0 5 - Nを備える。このI / O仮想化部6 0 1における処理は、I / Oデバイス5のデバイスI Dと制限値に関する処理の仲裁を含む。

【0080】

仮想C F Gレジスタ6 0 5 - 1 ~ 6 0 5 - Nは、デバイスI D保持部6 0 5 1 - 1 ~ 6

10

20

30

40

50

051 - Nと、制限値保持部6052 - 1 ~ 6052 - Nとを含む。デバイスID保持部6051 - 1 ~ 6051 - Nは、I/Oデバイス5の製造者や種別を示すデバイスIDの値を保持する。制限値保持部6052 - 1 ~ 6052 - Nは、I/Oデバイス5が用いるTLPのパケットサイズやタイムアウト値等の上限値や下限値に対する制限値を保持する。この制限値は、ホスト1 - 1 ~ 1 - Nが設定する。

【0081】

図14を参照して、I/O仮想化部601の各部の構成を説明する。なお、第2の実施形態における構成と同じ部分については、説明を省略する。

【0082】

I/O仮想化部601は、第2の実施形態におけるデバイス情報登録部433に換えて、デバイス情報登録・仲裁部633を備える。

10

【0083】

図13乃至15を参照して、第3の実施形態におけるI/O仮想化ブリッジ6の動作を説明する。なお、第2の実施形態における動作と重複する部分については、説明を省略する。

【0084】

I/O仮想化部601は、I/Oデバイス5がホスト1 - 1 ~ 1 - Nに対するサービスを開始する前に、I/Oデバイス5のPF CFGレジスタ5011からデバイスID保持部6051 - 1 ~ 6051 - Nに表示する値を読み込み、デバイスID保持部6051 - 1 ~ 6051 - Nに反映させる(図15のステップS401)。ホスト1 - 1 ~ 1 - Nが設定TLPを発行し、デバイスIDをリードする(ステップS402)場合、I/O仮想化部601はデバイスID保持部6051 - 1 ~ 6051 - Nに登録した値をホスト1 - 1 ~ 1 - Nに返す(ステップS403)。

20

【0085】

I/O仮想化部601はまた、ホスト1 - 1 ~ 1 - NがI/Oデバイス5に設定するTLPのパケットサイズやタイムアウト値等の上限値や下限値を、制限値保持部6052 - 1 ~ 6052 - Nに登録する(ステップS404)。ホスト1 - 1 ~ 1 - Nからのリード要求に対しては、制限値保持部6052 - 1 ~ 6052 - Nに登録された値が返される。I/O仮想化部601のデバイス情報登録・仲裁部633は、ここに登録された値の中で、全てのホストからの設定要求が満たされる値(制限値が下限値であれば全ての制限値の最大値、制限値が上限値であれば全ての制限値の最小値)を実際の設定値として、これら制限値の最大値と最小値を仮想CFGレジスタ605 - 1 ~ 605 - Nに保持する(ステップS405)。これらの設定値は、I/Oデバイス設定値生成・変換部432とI/Oパケット転送部402を介してI/Oデバイス5に設定される(ステップS406)。ホスト1 - 1 ~ 1 - NからのTLPが仲裁を必要とする場合に、パケット仲裁部431は、仮想CFGレジスタ605 - 1 ~ 605 - Nからこれらの制限値を読み込んで、TLPの仲裁の制約条件とする。

30

【0086】

ホスト1 - 1 ~ 1 - Nのそれぞれは、その処理能力に応じてI/Oデバイス5を設定しようとして、設定要求をI/O仮想化ブリッジ4に送信する。例えば、最大パケットサイズを256Bと設定するホストと、最大パケットサイズを512Bと設定するホストが、それぞれ設定要求を送信するとする。この場合、I/Oデバイス5には最大パケットサイズが256Bに設定される。256Bより大きな値に設定されると、最大パケットサイズを256Bと設定しようとしたホストに不具合が生じる。一方、ホストからのリード要求に対しては、設定要求により保持された値を応答として返すので、最大パケットサイズを512Bと設定したホストは、実際にI/Oから受信するパケットが256B以下である動作であっても、最大パケットサイズが512Bと設定されているとの応答を受ける。

40

【0087】

第3の実施形態では、I/Oデバイスにおいてホスト1 - 1 ~ 1 - Nにそれぞれ割り当てられたデバイスのデバイスIDをI/O仮想化ブリッジ6が保持しており、ホストからの

50

リード要求に対して、デバイスIDを付して返却値をホストに返す。これにより、I/Oデバイスの個々の属性を考慮した設定処理が可能になる。また、ホスト1-1~1-NがI/Oデバイス5に対して設定した制限値の要求をすべて満足するような値を、最大値と最小値として求めて、I/O仮想化ブリッジ6が保持し、I/Oデバイス5に設定する。この最大値と最小値を制約条件とすることにより、TLPの仲裁を高速に処理することができる。

(第4の実施形態)

次に、本発明を実施するための第4の実施形態について図面を参照して詳細に説明する。

【0088】

図16に示す本発明を実施するための第4の実施形態によるI/O仮想化ブリッジ7は、第2の実施形態におけるI/O仮想化部403に換えて、I/O仮想化部701を備える。さらに、I/O仮想化ブリッジ7は、第2の実施形態における仮想CFGレジスタ405-1~405-Nに換えて、仮想CFGレジスタ705-1~705-Nを備える。このI/O仮想化部7における処理は、I/Oデバイス5に対する電力制御(パワーマネジメント)に関する処理の仲裁を含む。

10

【0089】

仮想CFGレジスタ405-1~405-Nは、ホスト1-1~1-NからI/Oデバイス5に対する電力制御の要求を受け付けるとともに、I/Oデバイス5の電力制御の状態をホスト1-1~1-Nに示す電力制御値保持部7053-1~7053-Nを含む。

20

【0090】

図17を参照して、I/O仮想化部701の各部の構成を説明する。なお、第2の実施形態における構成と同じ部分については、説明を省略する。

【0091】

I/O仮想化部701は、第2の実施形態におけるパケット仲裁部431に換えて、パケット仲裁・電力制御値調整部731を備える。

【0092】

図16乃至18を参照して、第4の実施形態におけるI/O仮想化ブリッジ7の動作を説明する。なお、第2の実施形態における動作と重複する部分については、説明を省略する。

30

【0093】

I/O仮想化部701のデバイス情報登録433は、ホスト1-1~1-NがI/Oデバイス5に対して発行するTLPを受け、TLPに含まれるI/Oデバイス5の電力制御に関する設定情報を電力制御値保持部7053-1~7053-Nに登録する(図18のステップS501)。

【0094】

I/O仮想化部701において、パケット仲裁・電力制御値調整部731は、ホストから受信したTLPを仲裁し、さらに、仲裁した結果を元に、仮想CFGレジスタ701-1~701-Nの電力制御値保持部7053-1~7053-Nに保持された電力制御値を調整する。すなわち、第2の実施例と同様に、

40

(1) I/Oデバイス5に電力制御を設定するかどうか判定し(ステップS502)、電力制御の設定をしない場合は、仮想CFGレジスタ705-1に保持されている電力制御値を調整する。I/O仮想化部701は、I/Oデバイス5に対して電力制御の設定を行わないが、ホスト1-1に対してI/Oデバイス5が電力制御されている状態にする(ステップS503)。すなわち、ホスト1-1~1-Nの何れかからリード要求があったときは、I/Oデバイスが電力制御されているような電力制御値を返却値としてホストに返す。

(2) I/Oデバイス5に電力制御を設定する場合は、ホスト1-1~1-Nが仮想CFGレジスタ705-1~705-Nに登録した設定情報とTLPの仲裁の結果を元にして、ホスト1-1~1-Nの要求するI/Oデバイスの動作に支障が生じないようにI/O

50

デバイス5の電力制御の設定を調整する。I/Oデバイス5にインタフェースが実装されるかどうかを判定し(ステップS504)、インタフェースが具備されない場合は、得られた電力制御値に基づいて設定指示を発行し、I/Oデバイス5のPF_CFGレジスタ5011を直接設定する(ステップS505)。

(3) I/Oデバイス5の物理関数501が仮想関数502-1~502-Nを制御するインタフェースを実装している場合には、(2)で求めた電力制御値を、PF_CFGレジスタ5011を設定する設定TLPに変換し、変換した設定TLPを設定指示として、I/Oデバイス5に発行する(ステップS506)。

【0095】

第4の実施形態では、I/Oデバイスにおいて、ホスト1-1~1-NからI/Oデバイス5に対して発行された電力制御(パワーマネジメント)要求を受けて、I/Oデバイスに電力制御を設定しない、或は、ホストから要求されるI/Oデバイスの動作に支障が生じないようにI/Oデバイス5に電力制御を設定する。これにより、複数のホストからの電力制御に起因した不用意な電力制御によるI/Oデバイスの動作の不具合を防止することができる。すなわち、一つのホストからの電力制御要求により、他のホストから要求された動作に支障が生じるといった不具合を回避し、要求されたI/Oデバイスの動作の稼働率を高く維持することができる。

(第5の実施形態)

次に、本発明を実施するための第5の実施形態について図面を参照して詳細に説明する。

【0096】

図19に示す本発明を実施するための第5の実施形態によるI/O仮想化ブリッジ8は、第2の実施形態において、I/O仮想化部403に換えてI/O仮想化部801を備える。このI/O仮想化ブリッジ8における処理は、I/Oデバイス5に対するデバイスリセット処理の仲裁を含む。

【0097】

図20を参照して、I/O仮想化部801の各部の構成を説明する。なお、第2の実施形態における構成と同じ部分については、説明を省略する。

【0098】

I/O仮想化部801は、第2の実施形態におけるパケット仲裁部431に換えて、パケット仲裁・リセット要求検出部831を備える。

【0099】

図19乃至21を参照して、第5の実施形態におけるI/O仮想化ブリッジ8の動作を説明する。なお、第2の実施形態における動作と重複する部分については、説明を省略する。

【0100】

I/O仮想化部801のパケット仲裁・リセット要求検出部831は、ホスト1-1~1-NがI/Oデバイス5に対して発行するTLPを受け、TLPを仲裁するとともに、TLPに含まれるI/Oデバイス5に対するリセットを要求する信号を検出する(ステップS601)。パケット仲裁・リセット要求検出部831は、リセット要求を含むTLPとそれ以外のTLPとを仲裁し、仮想関数502-1~502-Nのうち、リセットTLPを発行したホストに割当てられた仮想関数をリセットしても、ホストから要求されたI/Oデバイスの処理に、強制終了を受けて障害が生じないようにリセットのタイミングを求める(ステップS602)。

【0101】

パケット仲裁・リセット要求検出部831は、求めたリセットのタイミングを元に、リセットTLPをI/Oデバイス設定値生成・変換部432に転送する。I/Oデバイス設定値生成・変換部432は、I/Oデバイス5をリセットする要求を含むリセットTLPを、I/Oデバイス5のうちでTLPを発行したホストに割当てられた仮想関数のみをリセットするファンクションレベルリセットに変換する。I/O仮想化ブリッジ8は、変換

10

20

30

40

50

されたファンクションレベルリセットをI/Oデバイス5に送信する(ステップS603)。ファンクションレベルリセットを受けた仮想関数はリセットする(ステップS604)。

【0102】

第5の実施形態では、I/Oデバイスへのリセット要求を含むTLPについて、リセット要求をI/Oデバイスの仮想関数のリセットに部分的に限定する。そしてこのリセット要求を含むTLPを、その他のTLPと仲裁することによって、I/Oデバイスの仮想関数のリセットが他の仮想関数での動作に影響しないようにする。これにより、複数のホストからのI/Oデバイスに対する制御は、リセット要求を含むことができ、リセットによるシステムへの負担を抑制した、自由度の高いI/Oデバイス制御が可能になる。

10

【産業上の利用可能性】

【0103】

本発明によれば、コンピュータ装置やネットワーク装置、産業用機器やコンシューマ機器において、I/Oを複数のホスト、あるいはCPUやCPUに類する演算装置を含む情報処理装置の間で共有するといった用途に適用できる。

【符号の説明】

【0104】

- 1 - 1 ~ 1 - N ホスト
- 2 - 1 ~ 2 - N ホストブリッジ
- 3 ネットワーク
- 4 I/O仮想化ブリッジ
- 5 I/Oデバイス
- 6 I/O仮想化ブリッジ
- 7 I/O仮想化ブリッジ
- 8 I/O仮想化ブリッジ
- 9 ホスト
- 100 I/Oシステム
- 101 - 1 ~ 101 - N CPU
- 102 - 1 ~ 102 - N ブリッジ
- 103 - 1 ~ 103 - N メモリ
- 401 ネットワーク処理部
- 402 I/Oパケット転送部
- 403 I/O仮想化部
- 404 アドレス変換部
- 405 - 1 ~ 405 - N 仮想CFGレジスタ
- 411 デカプセル化部
- 412 カプセル化部
- 413 ネットワークパケット送受信部
- 414 アドレス保持部
- 421 パケット選択部
- 422 第1転送部
- 423 第2転送部
- 431 パケット仲裁部
- 432 I/Oデバイス設定値生成・変換部
- 433 デバイス情報登録部
- 434 ホスト通知生成部
- 441 ヘッダアドレス変換部
- 442 アドレス参照部
- 501 物理関数(PF)
- 502 - 1 ~ 502 - N 仮想関数(VF)

20

30

40

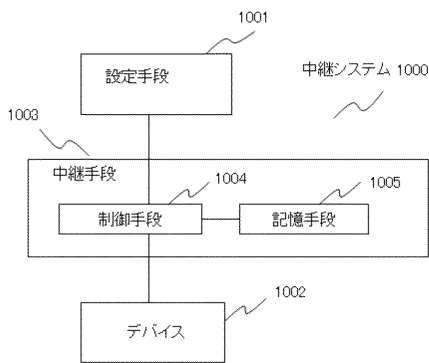
50

- 5 0 3 インタフェース
- 6 0 1 I / O 仮想化部
- 6 0 5 - 1 ~ 6 0 5 - N 仮想 C F G レジスタ
- 6 3 3 デバイス情報登録・仲裁部
- 7 0 1 I / O 仮想化部
- 7 0 5 - 1 ~ 7 0 5 - N 仮想 C F G レジスタ
- 7 3 1 パケット仲裁・電力制御値調整部
- 8 0 1 I / O 仮想化部
- 8 3 1 パケット仲裁・リセット要求検出部
- 9 0 0 I / O システム
- 9 0 1 管理仮想マシン (管理 V M)
- 9 0 2 - 1 ~ 9 0 2 - N 仮想マシン (V M)
- 9 0 3 C P U
- 9 0 4 ブリッジ
- 9 0 5 メモリ
- 5 0 1 1 P F C F G レジスタ
- 5 0 2 1 - 1 ~ 5 0 2 1 - N V F C F G レジスタ
- 6 0 5 1 - 1 ~ 6 0 5 1 - N デバイス I D 保持部
- 6 0 5 2 - 1 ~ 6 0 5 2 - N 制限値保持部
- 7 0 5 3 - 1 ~ 7 0 5 3 - N 電力制御値保持部
- 1 0 0 1 設定手段
- 1 0 0 2 デバイス
- 1 0 0 3 中継手段
- 1 0 0 4 制御手段
- 1 0 0 5 記憶手段

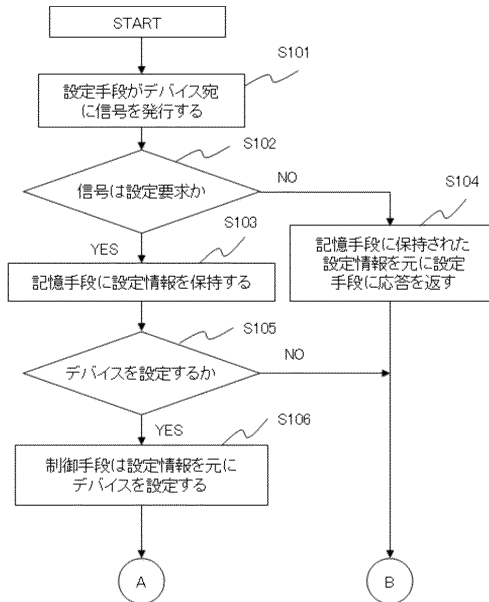
10

20

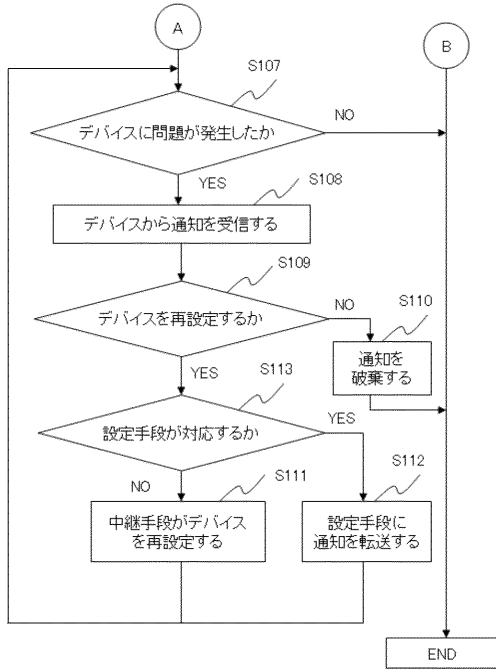
【 図 1 】



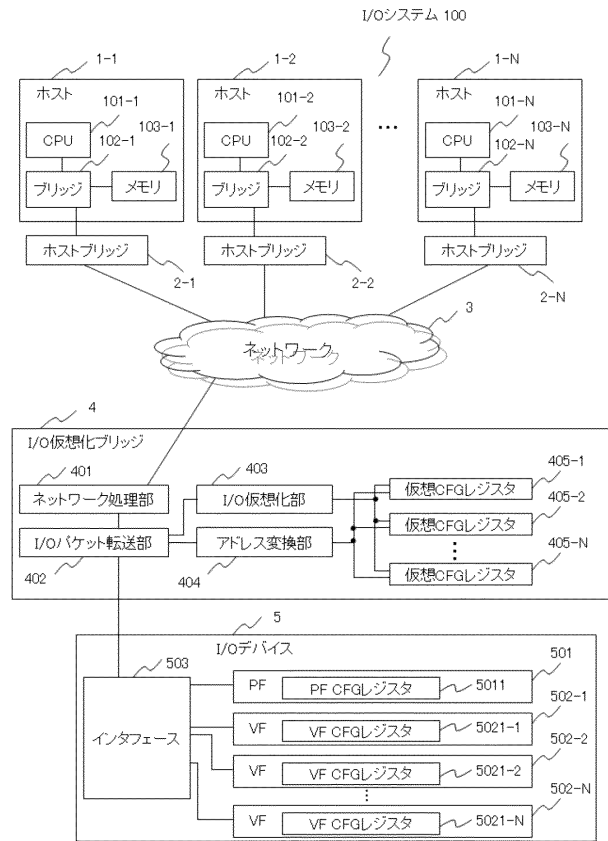
【 図 2 】



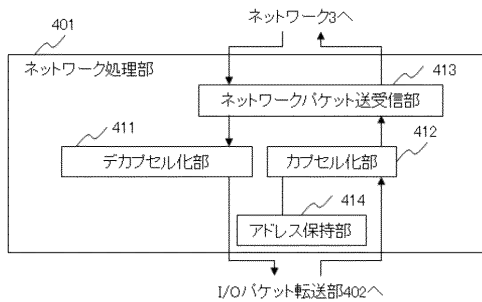
【図3】



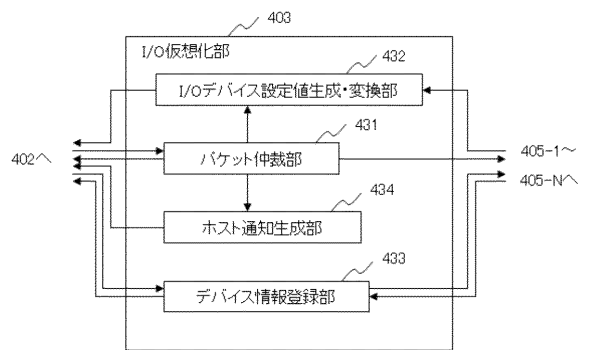
【図4】



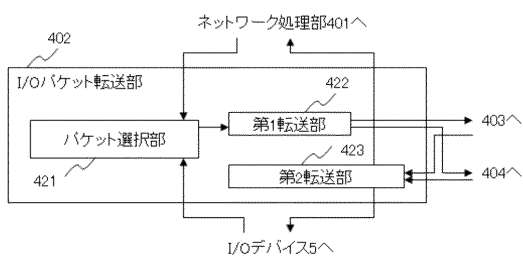
【図5】



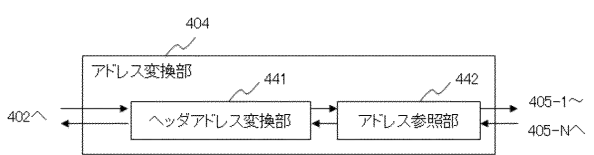
【図7】



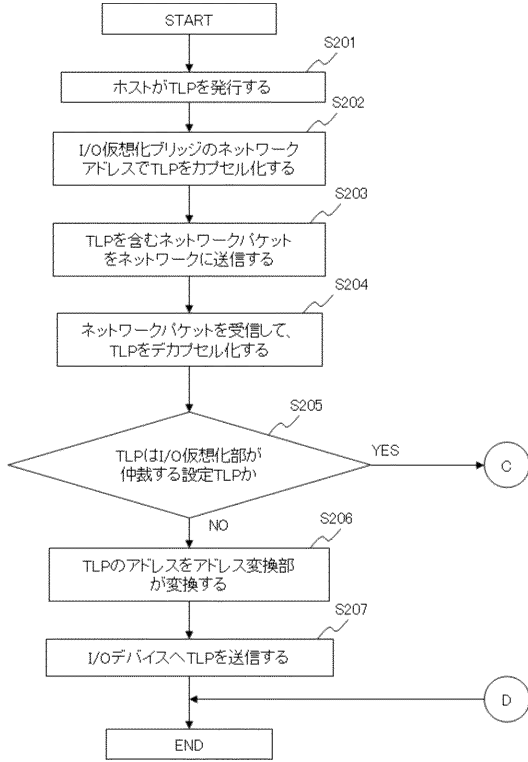
【図6】



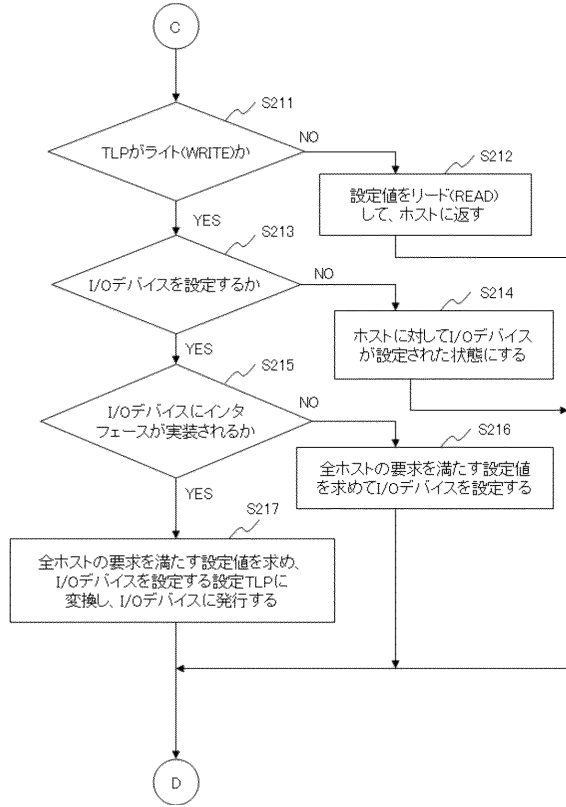
【図8】



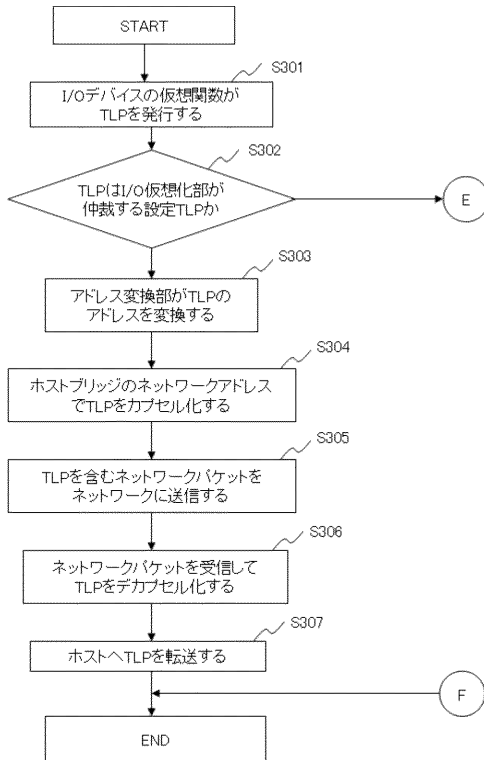
【図9】



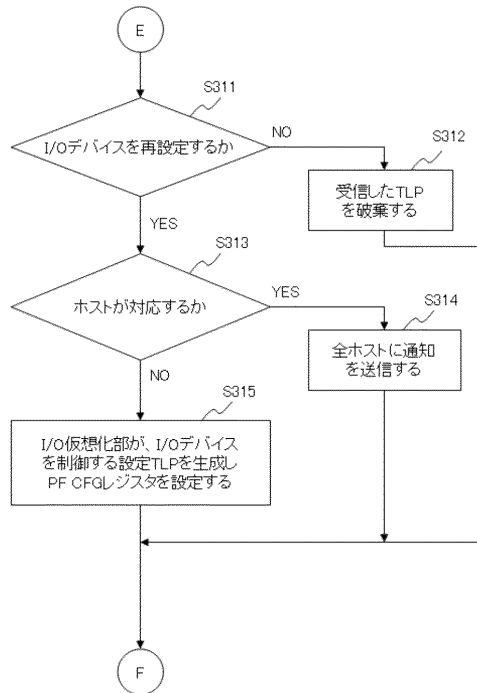
【図10】



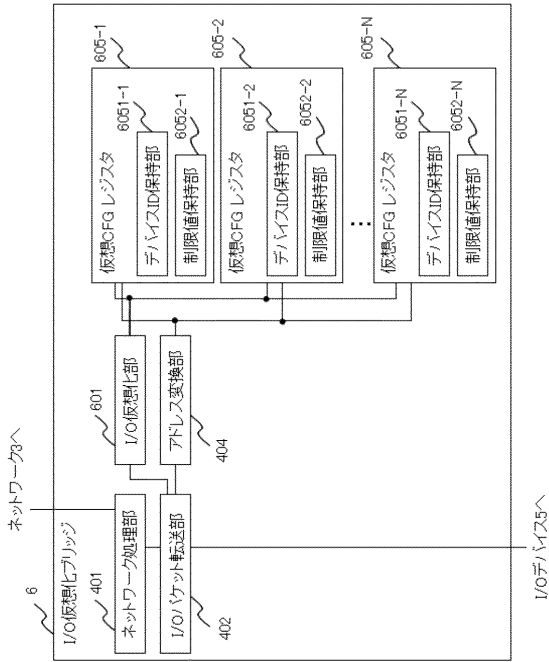
【図11】



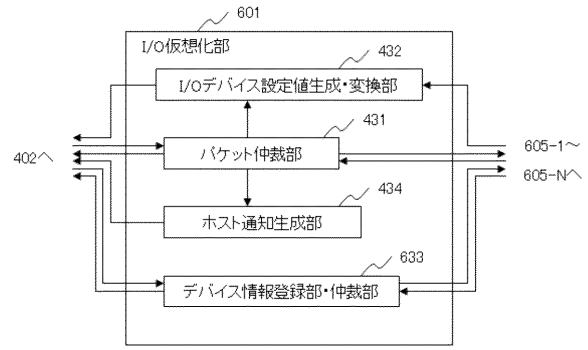
【図12】



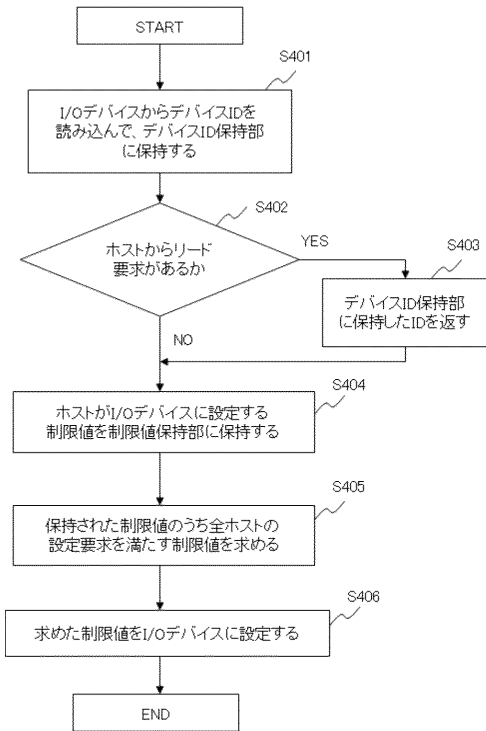
【図13】



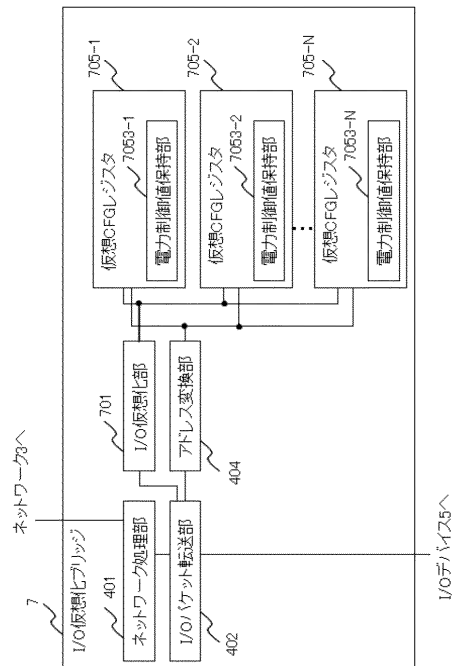
【図14】



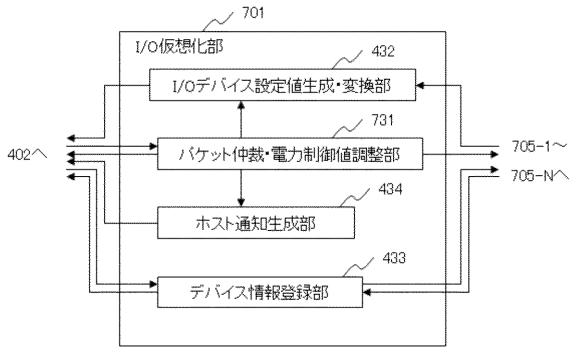
【図15】



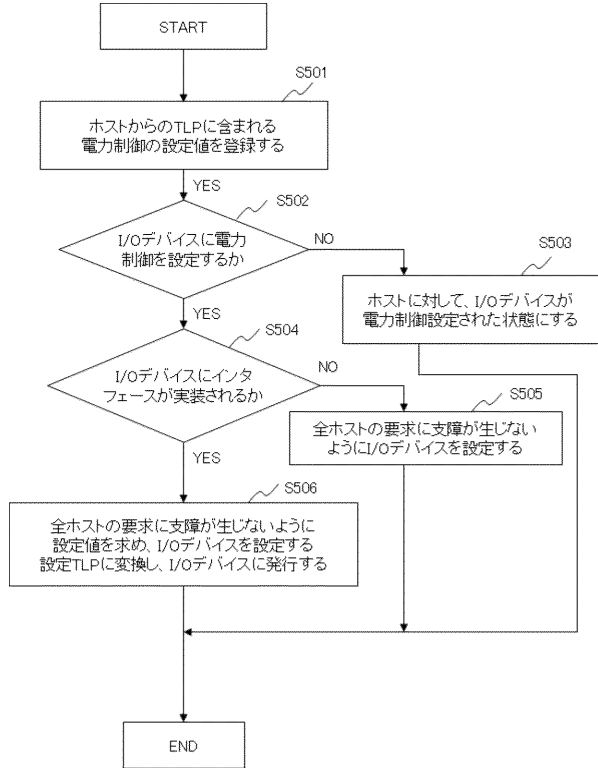
【図16】



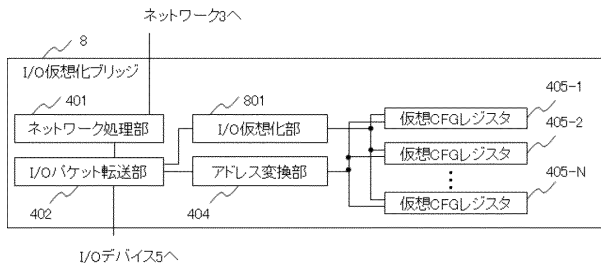
【図17】



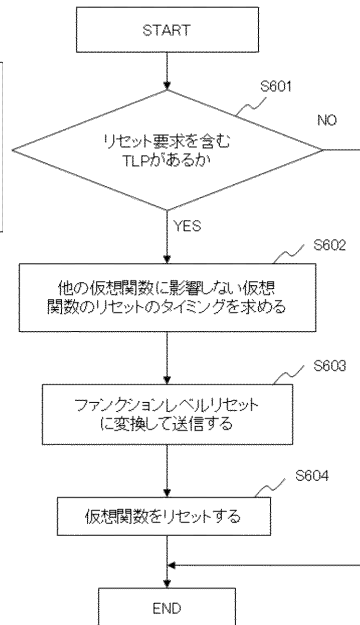
【図18】



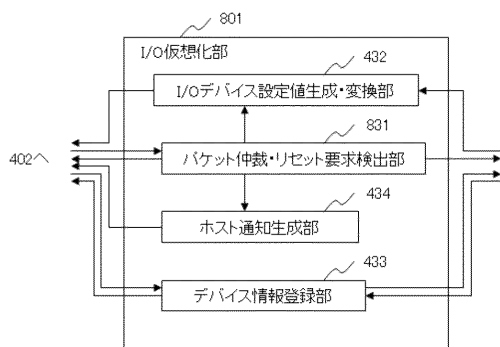
【図19】



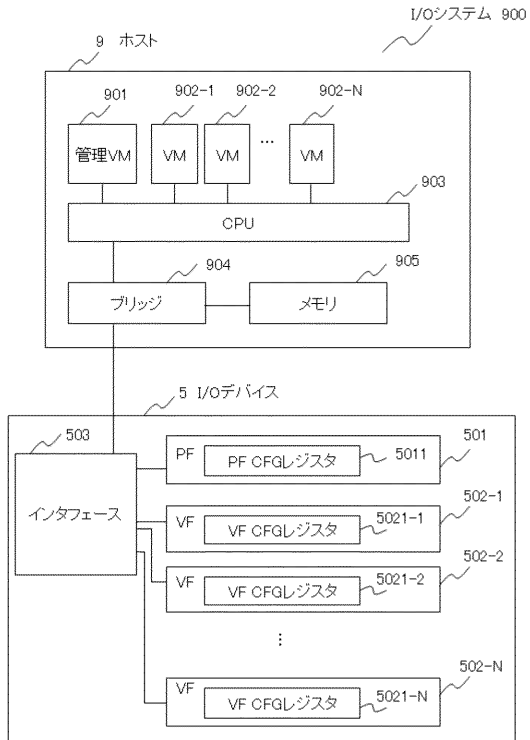
【図21】



【図20】



【図22】



フロントページの続き

- (72)発明者 樋口 淳一
東京都港区芝五丁目7番1号 日本電気株式会社内
- (72)発明者 吉川 隆士
東京都港区芝五丁目7番1号 日本電気株式会社内

合議体

- 審判長 小曳 満昭
審判官 白石 圭吾
審判官 千葉 輝久

- (56)参考文献 国際公開第2009/025381(WO, A1)
国際公開第2009/054525(WO, A1)
国際公開第2008/018485(WO, A1)
特開2003-167837(JP, A)

- (58)調査した分野(Int.Cl., DB名)
G06F 13/00 - 13/42