(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5111236号
(P5111236)
(45)発行日　平成25年1月9日(2013.1.9)　(24)登録日　平成24年10月19日(2012.10.19)

(51)Int.Cl.　F I
G06F 13/36 (2006.01)　G06F 13/36 310D

請求項の数 4　(全 9 頁)

(21)出願番号　特願2008-134509(P2008-134509)
(22)出願日　平成20年5月22日(2008.5.22)
(65)公開番号　特開2009-282773(P2009-282773A)
(43)公開日　平成21年12月3日(2009.12.3)
審査請求日　平成22年6月8日(2010.6.8)

(73)特許権者　000005108
株式会社日立製作所
東京都千代田区丸の内一丁目６番６号
(74)代理人　110000442
特許業務法人　武和国際特許事務所
(72)発明者　関　辰一郎
神奈川県秦野市堀山下１番地　株式会社日立製作所　エンタープライズサーバ事業部内
(72)発明者　小国　哲
神奈川県秦野市堀山下１番地　株式会社日立製作所　エンタープライズサーバ事業部内

審査官　坂東　博司

最終頁に続く

(54)【発明の名称】データ転送システム

(57)【特許請求の範囲】
【請求項１】
PCI Express リンクを用いたデータ転送システムであって、
ルートコンプレックスと、PCI Express デバイスに接続されたアドインカードのそれぞれが接続される少なくとも１つのスロットと、パケットスイッチと、経路を選択する経路選択手段とを備え、
前記経路選択手段は、前記ルートコンプレックスと前記スロットとの間のデータ転送用の経路について、前記パケットスイッチを介する経路と、前記パケットスイッチを経由しない経路とを、前記アドインカードの性能に応じて選択することを特徴とするデータ転送システム。
【請求項２】
前記パケットスイッチと前記経路選択手段とがＦＰＧＡ上に実装されたことを特徴とする請求項１記載のデータ転送システム。
【請求項３】
前記パケットスイッチと前記経路選択手段とがＡＳＩＣ上に実装されたことを特徴とする請求項１記載のデータ転送システム。
【請求項４】
前記経路選択手段は、外部から与えられるセレクト信号により、前記経路の選択を行うことを特徴とする請求項１記載のデータ転送システム。
【発明の詳細な説明】

【技術分野】
【０００１】
本発明は、データ転送システムに係り、特に、用途に最適な経路を設定してデータ転送を制御するデータ転送システムに関する。
【背景技術】
【０００２】
近年、情報処理装置に採用されるＩ／Ｏインターフェースは、性能要件の向上に伴い、ＰＣＩ（ＴＭ）、ＰＣＩ－Ｘ（ＴＭ）等のパラレルバスから、PCI Express （ＴＭ）等の高速シリアルインターフェースに置き換わりつつある。
【０００３】
PCI Express は、従来のパラレルバスとは異なり、デバイス間を１対１で接続することができるため、ホストが備えるポート数以上のスロット数を実装したい場合等にしばしばPCI Express スイッチが使用される。
【０００４】
PCI Express スイッチは、スロット数の拡張だけではなく、複数のホストによるＩ／Ｏ資源の共有、Ｉ／Ｏの冗長構成、Ｉ／Ｏの仮想化等の幅広い応用が可能である。なお、この種のPCI Express スイッチの応用に関する各種の技術が、例えば、非特許文献１等に記載されて知られている。
【非特許文献１】“PLX PCI Express Presentation”, October, 26 2007, PLX Technology Inc
【発明の開示】
【発明が解決しようとする課題】
【０００５】
前述したように、PCI Express スイッチは、様々な応用が期待されるが、一方で、PCI Express カードやそれに関連するソフトウエアが、PCI Express スイッチを経由する使用方法に対応していない場合や、PCI Express スイッチを経由する際に増加するレイテンシが性能上無視できない場合等、PCI Express スイッチを経由したくない場合がある。
【０００６】
本発明の目的は、前述したような点に鑑み、PCI Express スイッチを含むシステムにおいて、用途に最適な経路を設定することを可能にしたデータ転送システムを提供することにある。
【課題を解決するための手段】
【０００７】
本発明によれば前記目的は、PCI Express リンクを用いたデータ転送システムであって、　ルートコンプレックスと、PCI Express デバイスに接続されたアドインカードのそれぞれが接続される少なくとも１つのスロットと、パケットスイッチと、経路を選択する経路選択手段とを備え、前記経路選択手段は、前記ルートコンプレックスと前記スロットとの間のデータ転送用の経路について、前記パケットスイッチを介する経路と、前記パケットスイッチを経由しない経路とを、前記アドインカードの性能に応じて選択することにより達成される。
【発明の効果】
【０００８】
本発明によれば、ユーザは、用途に応じた最適なデータ転送経路を選択して使用することが可能となる。
【発明を実施するための最良の形態】
【０００９】
以下、本発明によるデータ転送システムの実施形態を図面により詳細に説明する。
【００１０】
図１は本発明の第１の実施形態によるデータ転送システムの構成例を示すブロック図である。図１に示す本発明の第１の実施形態によるデータ転送システムは、３ポートのPCI

Express スイッチ（以下、単に、スイッチという）を、メインボード上に搭載して構成した例である。なお、PCI Express スイッチは、一般に、マルチプレクサ等のスイッチとの区別のために、パケットスイッチと呼ばれている。

【００１１】

そして、図１（ａ）に示す構成例は、メインボード１００におけるルートコンプレックス１１０からスロット１４０ａ及びスロット１４０ｂまでの経路が、スイッチ１３０を経由するように設定が施された構成例を示し、図１（ｂ）に示す構成例は、メインボード１００におけるルートコンプレックス１１０からスロット１４０ａ及びスロット１４０ｂまでの経路が、スイッチ１３０を迂回するように設定が施された構成例を示している。また、図１（ａ）、図１（ｂ）に示す例は、スイッチ１３０を経由する経路を作成するか否かを制御するために、ルートコンプレックス１１０とスロット１４０ａ及びスロット１４０ｂとの間に、マルチプレクサ１２０ａ、スイッチ１３０、マルチプレクサ１２０ｂ、１２０ｃが設けられて図示のように構成されている。

【００１２】

前述において、ルートコンプレックス１１０は、ＰＣＩ規格における最上位の親となるコントローラであり、一般には、図示しないホストコンピュータ等の情報処理装置に接続される。また、スロット１４０ａ、１４０ｂには、図１（ａ）の場合、アドインカード２００ａ、２００ｂが経路１５０ｊ、１５０ｉを介して接続され、図１（ｂ）の場合、アドインカード２００ｃ、２００ｄが経路１５０ｊ、１５０ｉを介して接続されている。これらのアドインカードには、ネットワーク、記憶装置、Ｉ／Ｏ等を含む図示しないPCI Express デバイスが接続され、また、これらのアドインカードは、接続されるPCI Express デバイスに対するコントローラの機能を有する。なお、ルートコンプレックス及びアドインカードの機能等については、後述する本発明の他の実施形態においても前述と同様である。

【００１３】

図１（ａ）に示す例において、ルートコンプレックス１１０のリンクは、１６レーン幅の１ポートに設定され、経路１５０ａを経てマルチプレクサ１２０ａに接続される。また、マルチプレクサ１２０ａは、ルートコンプレックス１１０とスイッチ１３０との接続を１６レーン幅で経路１５０ａ、１５０ｂを経てリンクするように設定する。マルチプレクサ１２０ｂは、スイッチ１３０とアドインカード２００ａとの接続をスロット１４０ａを介して、経路１５０ｆ、１５０ｈ、１５０ｊを経て１６レーン幅でリンクするように設定する。マルチプレクサ１２０ｃは、スイッチ１３０とアドインカード２００ｂとの接続をスロット１４０ｂを介して、経路１５０ｅ、１５０ｇ、１５０ｉを経て１６レーン幅でリンクするよう設定する。各マルチプレクサの設定は、共通のセレクト信号１６０を用いて行われる。このセレクト信号１６０は、図示しないボード管理用のコントローラを介して、ホストコンピュータから指示されるものであっても、また、図示しないサービスプロセッサを用いて利用者等により指示されるものであってもよく、後述する本発明の第２及び第４の実施形態の場合も同様である。

【００１４】

図１（ｂ）に示す例において、ルートコンプレックス１１０のリンクは、８レーン幅の２ポートに分割され、経路１５０ｋ、１５０ｌを経てマルチプレクサ１２０ａに接続される。また、マルチプレクサ１２０ａは、ルートコンプレックス１１０とスロット１４０ａとを、マルチプレクサ１２０ａ及びマルチプレクサ１２０ｂを介して８レーン幅で経路１５０ｌ、１５０ｄ、１５０ｈを経てスイッチ１３０を介することなく接続するように直接リンクさせ、ルートコンプレックス１１０とスロット１４０ｂとを、マルチプレクサ１２０ａ及びマルチプレクサ１２０ｃを介して８レーン幅で経路１５０ｋ、１５０ｃ、１５０ｇを経てスイッチ１３０を介することなく接続するように直接リンクさせている。

【００１５】

前述で説明した図１（ａ）、図１（ｂ）に示す本発明の第１の実施形態によれば、ルートコンプレックス１１０に接続された図示しないホストコンピュータは、用途に応じて、

スイッチ１３０を介したデータ転送経路を経由して、アドインカードに接続されたPCI Express デバイスを使用する経路と、スイッチ１３０を介することのないデータ転送経路を経由して、アドインカードに接続されたPCI Express デバイスを使用する経路とを使い分けることができる。

【００１６】

図２は図１に示すPCI Express スイッチ１３０の内部構成を示すブロック図である。

【００１７】

図２に示すように、PCI Express スイッチ１３０は、図１（ａ）、図１（ｂ）に示しているルートコンプレックス１１０とリンクするポート１３１ａと、スロット１４０ａ、１４０ｂとリンクし、アドインカードに接続されるPCI Express デバイスとリンクするポート１３１ｂ、１３１ｃと、これらのポート１３１ａ～１３１ｃ相互間の接続を制御するスイッチ論理１３５とを備えている。また、図２に示すスイッチ１３０は、前述のポート１３１ａ～１３１ｃのそれぞれと、スイッチ論理１３５との間には、物理層１３２ａ～１３２ｃ、データリンク層１３３ａ～１３３ｃ、トランザクション層１３４ａ～１３４ｃが設けられている。

【００１８】

前述において、物理層１３２ａ～１３２ｃは、転送データのシリアル化・非シリアル化、８ビットデータを１０ビットデータの時間幅で転送するようにＰＣＩ規格による１０ｂ／８ｂ変換等の処理を行う。データリンク層１３３ａ～１３３ｃは、主にリンクの管理とエラー検出・訂正との処理を行う。トランザクション層１３４ａ～１３４ｃは、トランザクションレイヤパケット（ＴＬＰ）の分解・生成の処理を行うと共に、相手側のデバイスとのパケット交換のフローコントロールの処理を担う。スイッチ論理１３５は、各ポートのトランザクション層間のパケットをルーティングする。そして、スイッチ論理１３５は、ポート１３１ａとポート１３１ｂとの間、ポート１３１ａとポート１３１ｃとの間のデータ転送に限らず、ポート１３１ｂとポート１３１ｃとの間のいわゆるpeer-to-peer転送も行うことが可能である。

【００１９】

前述したように構成されるスイッチ１３０は、一般に、このスイッチによるパケットのルーティングを行う際に、前述した各階層の処理に起因するレイテンシが増加することが知られている。

【００２０】

図３はスイッチ１３０のソフトウェアから見たトポロジを示す図である。

【００２１】

PCI Express は、ソフトウエアの互換性を維持するため、ＰＣＩ互換のコンフィグレーションメカニズムが採用される。各ポート１３１ａ～１３１ｃは、仮想ＰＣＩバス１３７ａ～１３７ｃを介して仮想PCI to PCIブリッジ１３６ａ～１３６ｃに接続される。仮想PCI to PCIブリッジ相互間の通信は、仮想ＰＣＩバス１３８上で行われる。コンフィグレーションソフトウェアは、各仮想ＰＣＩバス１３７ａ～１３７ｃ、１３８のそれぞれが異なるバス番号となるように、各仮想PCI to PCIブリッジのコンフィグレーション空間レジスタを設定する。

【００２２】

図１（ａ）に示して説明した構成例では、ルートコンプレックス１１０とスイッチ１３０との間、及び、２つのアドインカード２００ａ、２００ｂとスイッチ１３０との間が、全て１６レーン幅でリンクしている。ルートコンプレックスと両アドインカードとの間の合計スループットは、ルートコンプレックス１１０とスイッチ１３０との間の１６レーン幅に制限されるものの、アドインカード２００ａとルートコンプレックス１１０との間、アドインカード２００ｂとルートコンプレックス１１０との間の通信が重ならない場合、それぞれ１６レーン幅のスループットを享受することができる。また、図１（ａ）に示して説明した構成例の場合の最大スループットは、図１（ｂ）に示して説明した構成例の場合における８レーン幅と比べて、アドインカード同士がpeer-to-peer転送をサポートする

場合のアドインカード２００ａとアドインカード２００ｂとがスイッチ１３０を介して通信を行う１６レーン幅となる。

【００２３】

図１（ｂ）に示して説明した構成例では、ルートコンプレックス１１０とアドインカード２００ｃ、２００ｄが直接リンクするため、図１（ａ）に示した構成例の場合のように、スイッチ１３０を経由する際のレイテンシを増加させることがない。このため、アドインカードの性能がレイテンシに左右されやすい場合には、図１（ａ）に示す構成例のものを使用する場合に比較して、図１（ｂ）に示す構成例のものを使用する方が性能面で有利となる。また、アドインカードを扱うソフトウェアがスイッチ１３０による仮想PCI-to-PCIブリッジを介した使用をサポートしていないような場合には、システムの構成を図１（ｂ）に示す構成例とすることによりアドインカードを使用することが可能となる。

【００２４】

前述した本発明の第１の実施形態によるデータ転送システムは、３ポートのPCI Express スイッチを用いるとしたが、この実施形態は、２ポートのPCI Express スイッチを用いても、あるいは、３ポート以上のPCI Express スイッチを用いてもよい。

【００２５】

図４は本発明の第２の実施形態によるデータ転送システムの構成例を示すブロック図である。図４に示す本発明の第２の実施形態は、Ｉ／Ｏ仮想化機構を含む２ポートのスイッチをライザーカードに搭載した場合の構成例である。

【００２６】

一般に、仮想化環境での性能を向上させる手段として、Ｉ／Ｏ仮想化機構が知られている。そして、Ｉ／Ｏ仮想化機構に対応していない既存のハードウエアを最小限の変更で対応させる方法として、メインボード上に搭載するライザーカードの追加あるいは変更という形態がある。本発明の第２の実施形態は、この場合の構成例である。

【００２７】

そして、図４（ａ）に示す構成例は、メインボード３００上に備えられたルートコンプレックス３１０とスロット３２０とがリンクするように経路３３０ａにより接続され、ライザカード４００上に備えられたマルチプレクサ４１０ａからスロット４３０までの経路が、スイッチ４２０を経由するように設定が施された構成例を示し、図４（ｂ）に示す構成例は、メインボード３００上に備えられたルートコンプレックス３１０とスロット３２０とがリンクするように経路３３０ａにより接続され、ライザカード４００上に備えられたマルチプレクサ４１０ａからスロット４３０までの経路が、スイッチ４２０を迂回するように設定が施された構成例を示している。また、図４（ａ）、図４（ｂ）に示す例は、スイッチ４２０を経由する経路を作成するか否かを制御するために、メインボード３００に搭載されるライザカード４００上にマルチプレクサ４１０ａ、スイッチ４２０、マルチプレクサ４１０ｂ、スロット４３０が設けられて図示のように構成されている。

【００２８】

図４（ａ）に示す構成例は、メインボード３００と、ライザーカード４００と、アドインカード２００ｅとから構成される。ライザーカード４００上のマルチプレクサ４１０ａ、４１０ｂは、スイッチ４２０を経由してメインボード３００上のスロット３２０とライザーカード４００上のスロット４３０とを接続するように経路３３０ｂ、４４０ａ、４４０ｃ、４４０ｄを設定する。また、図４（ｂ）に示す構成例は、メインボード３００と、ライザーカード４００と、アドインカード２００ｆとから構成される。ライザーカード４００上のマルチプレクサ４１０ａ、４１０ｂは、スイッチ４２０を迂回してメインボード３００上のスロット３２０とライザーカード４００上のスロット４３０とを直接接続するように経路３３０ｂ、４４０ｂ、４４０ｄを設定する。

【００２９】

本発明の第２の実施形態は、前述のようにしてアドインカード２００ｅ、２００ｆにスイッチ４２０を経由する経路、あるいは、迂回する経路を提供することにより、Ｉ／Ｏ仮想化機構を使用する場合に図４（ａ）に示す構成とし、不要な場合やアドインカードやソ

フトウエアの制約により使用不可である場合に図４（ｂ）に示す構成とすることにより、従来のシステムと互換性を保ちつつＩ／Ｏ仮想化機構を導入できるという効果を得ることができる。

【００３０】

前述した本発明の第２の実施形態によるデータ転送システムは、２ポートのPCI Express スイッチを用いるとしたが、この実施形態は、３ポート以上のPCI Express スイッチを用いてもよい。

【００３１】

図５は本発明の第３の実施形態によるデータ転送システムの構成例を示すブロック図である。図５に示す本発明の第３の実施形態は、図４により説明した本発明の第２の実施形態と同等の機能をＦＰＧＡにより実現した構成例である。

【００３２】

ＦＰＧＡは、論理を読み込んで論理を再構成することが可能なＬＳＩであり、起動される毎に、その都度その機能を任意に変更可能なものである。本発明の第３の実施形態は、ライザーカード５００上に、図４に示したスイッチ４２０と同等のスイッチ５１１を含むＦＰＧＡ５１０を搭載して構成される。そして、この本発明の第３の実施形態は、ＦＰＧＡ５１０の初期化時に、読み込むＲＯＭの変更により、図５（ａ）に示すように、スイッチ５１１を経由する経路と、図５（ｂ）に示すように、スイッチ５１１を迂回する経路との切り替えが可能である。

【００３３】

前述したように構成される本発明の第３の実施形態においても、図４により説明した本発明の第２の実施形態と同等な効果を得ることができる。また、この本発明の第３の実施形態においても、スイッ５１１として、３ポート以上のPCI Express スイッチを用いることができる。

【００３４】

図６は本発明の第４の実施形態によるデータ転送システムの構成例を示すブロック図である。図６に示す本発明の第４の実施形態は、図５により説明した本発明の第３の実施形態と同等の機能をＡＳＩＣにより実現した構成例である。

【００３５】

本発明の第４の実施形態は、ライザーカード６００上にＡＳＩＣ６１０を搭載し、このＡＳＩＣ６１０上に、第２の実施形態の場合と同様に、マルチプレクサ６１１ａ、６１１ｂ、及び、スイッチ６１２を実装して構成した例であり、セレクト信号６１４により、スイッチ６１２を経由する経路６１３ａ、６１３ｂ、６１３ｄ、６１３ｅと、スイッチ６１２を迂回する経路６１３ａ、６１３ｃ、６１３ｅとの選択を行うことが可能である。また、この本発明の第４の実施形態においても、スイッ６１２として、３ポート以上のPCI Express スイッチを用いることができる。

【００３６】

前述した本発明の各実施形態によれば、PCI Express を採用したシステムにおいて、用途に応じてPCI Express スイッチを経由する経路と、PCI Express スイッチを経由しない経路とを選択することができるため、単一のシステムで両方の経路が有するメリットを享受することができる。

【図面の簡単な説明】

【００３７】

【図１】本発明の第１の実施形態によるデータ転送システムの構成例を示すブロック図である。

【図２】PCI Express スイッチの内部構成を示すブロック図である。

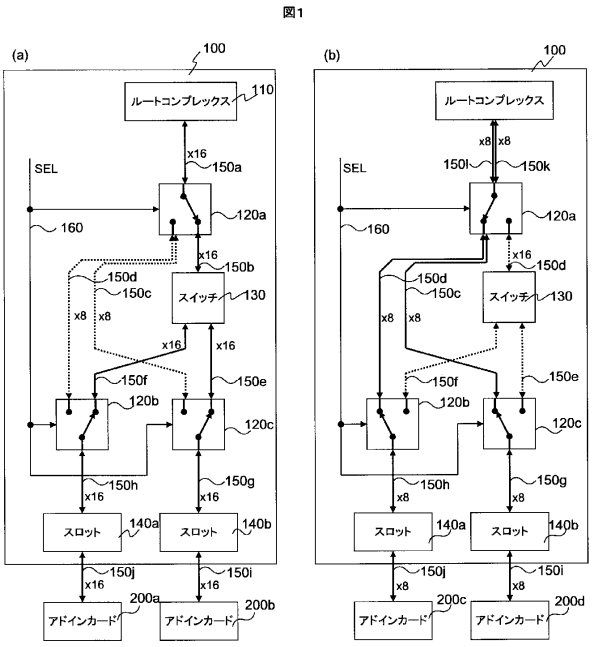
【図３】PCI Express スイッチのソフトウェアから見たトポロジを示す図である。

【図４】本発明の第２の実施形態によるデータ転送システムの構成例を示すブロック図である。

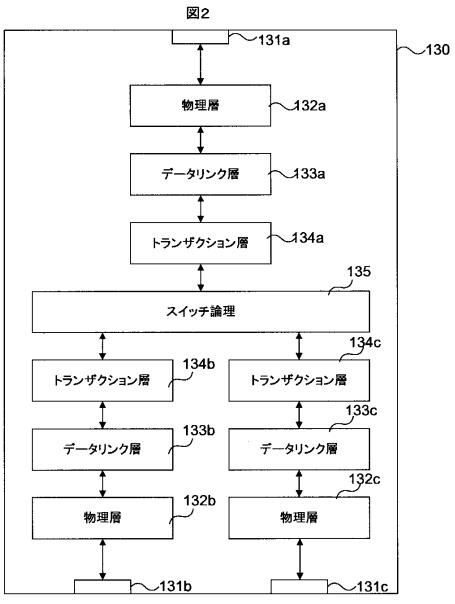
【図５】本発明の第３の実施形態によるデータ転送システムの構成例を示すブロック図で

ある。
【図６】本発明の第４の実施形態によるデータ転送システムの構成例を示すブロック図である。
【符号の説明】
【００３８】
１００、３００　メインボード
１１０、３０１、３１０　ルートコンプレックス
１２０ａ～１２０ｃ、４１０ａ、４１０ｂ、６１１ａ、６１１ｂ　マルチプレクサ
１３０、４２０、５１１、６１２　PCI Express スイッチ
１４０ａ、１４０ｂ、　３２０、４３０、５２０、６２０　スロット
２００ａ～２００ｆ　アドインカード
４００、５００、６００　ライザカード
５１０　ＦＰＧＡ
６１０　ＡＳＩＣ

【図１】

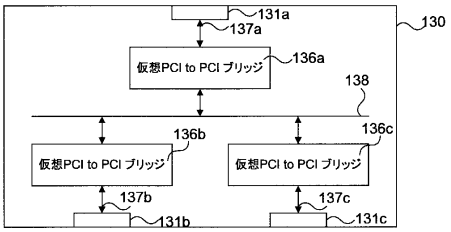


【図２】



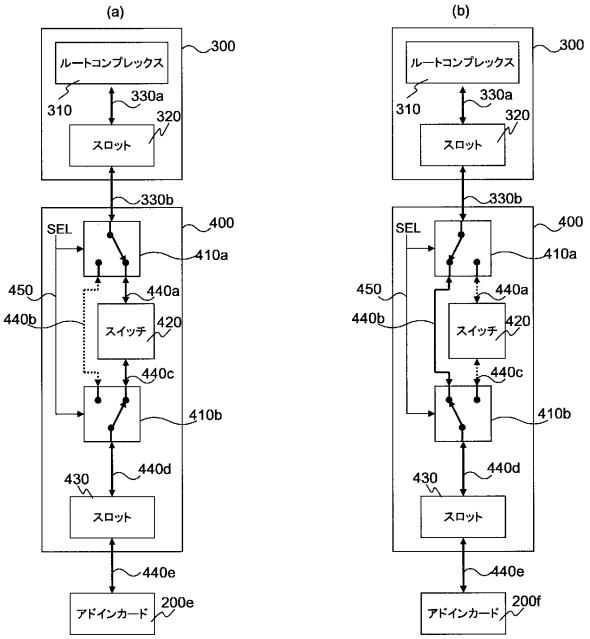
【図３】

図3



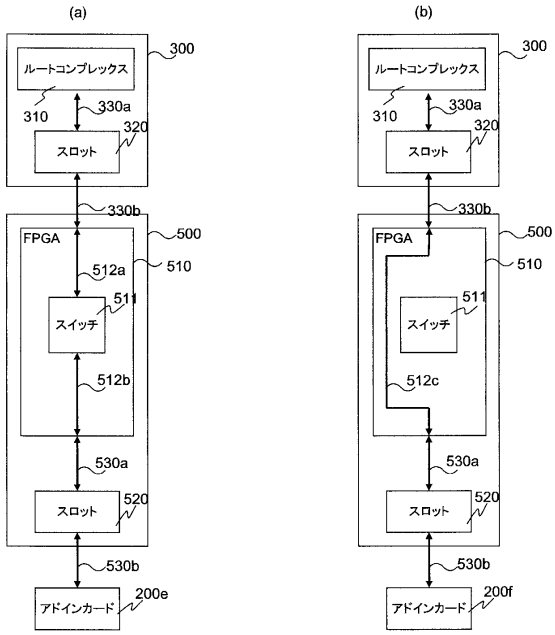
【図４】

図4



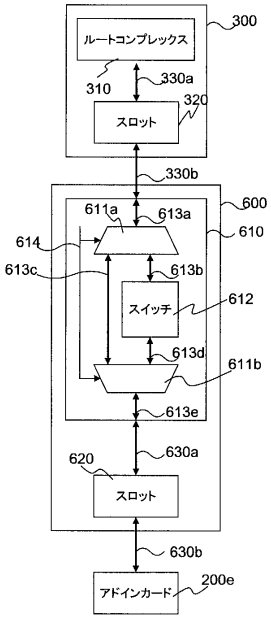
【図５】

図5



【図６】

図6



フロントページの続き

(56)参考文献 特開２００５－１６６０２８（ＪＰ，Ａ）
特開２００４－１３５１０６（ＪＰ，Ａ）
特開２００３－６９６１９（ＪＰ，Ａ）
特開２００３－６９６３７（ＪＰ，Ａ）
特開２００４－５６５９０（ＪＰ，Ａ）
特開２００１－３２０４２０（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１３／３６