(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4391819号
(P4391819)
(45)発行日　平成21年12月24日(2009.12.24)　(24)登録日　平成21年10月16日(2009.10.16)

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

請求項の数 20 (全 19 頁)

(21)出願番号　特願2003-536899(P2003-536899)
(86)(22)出願日　平成14年8月9日(2002.8.9)
(65)公表番号　特表2005-505855(P2005-505855A)
(43)公表日　平成17年2月24日(2005.2.24)
(86)国際出願番号　PCT/US2002/025278
(87)国際公開番号　WO2003/034239
(87)国際公開日　平成15年4月24日(2003.4.24)
審査請求日　平成17年8月9日(2005.8.9)
(31)優先権主張番号　09/978,349
(32)優先日　平成13年10月15日(2001.10.15)
(33)優先権主張国　米国(US)

前置審査

(73)特許権者　591016172
アドバンスト・マイクロ・ディバイシズ・インコーポレイテッド
ＡＤＶＡＮＣＥＤ　ＭＩＣＲＯ　ＤＥＶＩＣＥＳ　ＩＮＣＯＲＰＯＲＡＴＥＤ
アメリカ合衆国、９４０８８－３４５３　カリフォルニア州、サニィベイル、ピィ・オゥ・ボックス・３４５３、ワン・エイ・エム・ディ・プレイス、メイル・ストップ・６８（番地なし）
(74)代理人　100099324
弁理士　鈴木　正剛
(74)代理人　100111615
弁理士　佐野　良太

最終頁に続く

(54)【発明の名称】コンピュータ・システムの入出力ノード

(57)【特許請求の範囲】
【請求項１】
コンピュータシステムの入出力ノードであって、
第１ノードからの第１コマンドを第１通信経路を通じて受信するための第１受信ユニットと、
前記第１コマンドに対応する第１対応コマンドを第２通信経路を通じて第２ノードに送信するように結合された第１送信ユニットと、
前記第２ノードからの第２コマンドを第３通信経路を通じて受信するための第２受信ユニットと、
前記第２コマンドに対応する第２対応コマンドを第４通信経路を通じて前記第１ノードに送信するように結合された第２送信ユニットと、
を有し、前記第１受信ユニット、第２受信ユニット、第１送信ユニット及び第２送信ユニットのそれぞれは、コマンドを格納するための１つ以上のバッファを備え、
前記第１受信ユニット及び前記第２受信ユニットから選択されたコマンドを受信するよう結合されるとともに前記選択されたコマンドに対応するコマンドを周辺バスに送信するためのブリッジユニットと、を有し、前記ブリッジユニットは、更に、前記選択されたコマンドを対応するコマンドに変換するよう構成されたものであり、
前記第１受信ユニットと前記第１送信ユニットとの間に形成された第１通信リンクと、前記第２受信ユニットと前記第２送信ユニットとの間に形成された第２通信リンクと、を有し、前記第１通信リンク及び第２通信リンクは、パケットを伝送するための通信トンネ

ルを形成するものである、入出力ノード。
【請求項２】
前記第１通信経路から前記第２通信経路及び前記周辺バスへのコマンド伝達と、前記第３通信経路から前記第４通信経路及び前記周辺バスへのコマンド伝達と、を制御するように結合された制御ユニットを更に有し、
前記制御ユニットは、トランザクションの種類、送信装置および宛先のバッファの空きの有無、トランザクションが転送トラフィックか注入トラフィックかであるか、のいずれかに基づいて、当該制御ユニットのバッファにあるトランザクション間の調停を行う、
請求項１記載の入出力ノード。
【請求項３】
前記制御ユニットは、更に、前記周辺バスから前記第２通信経路及び前記第４通信経路へのコマンド伝達を制御するためのものである、
請求項２記載の入出力ノード。
【請求項４】
前記ブリッジユニットは、更に、前記周辺バスから受信したコマンドに対応するコマンドを前記第１の送信ユニット及び前記第２送信ユニットに選択的に与える、
請求項３記載の入出力ノード。
【請求項５】
前記制御ユニットは、更に、前記第１受信ユニット、前記第２受信ユニット及び前記ブリッジユニットから受信した複数の制御コマンドに基づいて前記コマンドの伝達を選択的に制御するためのものである、
請求項４記載の入出力ノード。
【請求項６】
前記制御コマンドのそれぞれは、前記第１受信ユニット、第２受信ユニット及び前記ブリッジユニットによって受信した対応するコマンドの一部を含む、
請求項５記載の入出力ノード。
【請求項７】
前記制御ユニットは、更に、前記制御コマンドを制御コマンド・バスを介して受信するためのものである、請求項６記載の入出力ノード。
【請求項８】
前記周辺バスは、ＰＣＩバス（ＰＣＩ：Peripheral Component Interconnect）である、
請求項１記載の入出力ノード。
【請求項９】
前記周辺バスは、グラフィックス・バスである、
請求項１記載の入出力ノード。
【請求項１０】
前記第１通信経路及び第３通信経路及び前記第２通信経路及び第４通信経路は、HyperTransport（登録商標）リンクである、
請求項１記載の入出力ノード。
【請求項１１】
一つ以上のプロセッサと、
相互に接続されるとともに前記一つ以上のプロセッサの所定の一つにチェーン式に接続された１つ以上の入出力ノードと、を有し、各入出力ノードは当該入出力ノードに接続された他の入出力ノードと互いにパケットを送受信するようにされており、前記入出力ノードは、
前記１つ以上のプロセッサの所定の一つからの第１コマンドを第１通信経路を通じて受信するための第１受信ユニットと、
前記第１コマンドに対応する第１対応コマンドを前記１つ以上の入出力ノードの前記チェーンにおける次のノードへと第２通信経路を通じて送信するように結合された第１送信

ユニットと、
前記１つ以上の入出力ノードの前記チェーンにおける次のノードから第３通信経路を通じて第２コマンドを受信するための第２受信ユニットと、
を有し、前記第１受信ユニット、第２受信ユニット、第１送信ユニット及び第２送信ユニットのそれぞれは、コマンドを格納するための１つ以上のバッファを有し、
前記第２コマンドに対応する第２対応コマンドを前記１つ以上のプロセッサの所定の一つへと第４通信経路を通じて送信するように結合された第２送信ユニット、
前記第１受信ユニット及び前記第２受信ユニットから選択されたコマンドを受信するよう結合されるとともに前記選択されたコマンドに対応するコマンドを周辺バスに送信するためのブリッジユニットと、を有し、前記ブリッジユニットは、更に、前記選択されたコマンドを対応するコマンドに変換するよう構成されたものであり、
前記第１受信ユニットと前記第１送信ユニットとの間に形成された第１通信リンクと、前記第２受信ユニットと前記第２送信ユニットとの間に形成された第２通信リンクと、を有し、前記第１通信リンク及び第２通信リンクは、パケットを伝送するための通信トンネルを形成するものである、コンピュータシステム。
【請求項１２】
前記入出力ノードは、前記第１通信経路から前記第２通信経路及び前記周辺バスへのコマンド伝達と、前記第３通信経路から前記第４通信経路及び前記周辺バスへのコマンド伝達と、を制御するように結合された制御ユニットを更に有し、
前記制御ユニットは、トランザクションの種類、送信装置および宛先のバッファの空きの有無、トランザクションが転送トラフィックか注入トラフィックかであるかに基づいて、当該制御ユニットのバッファにあるトランザクション間の調停を行う、
請求項１１記載のコンピュータシステム。
【請求項１３】
前記制御ユニットは、更に、前記周辺バスから前記第２通信経路及び前記第４通信経路へのコマンド伝達を制御するためのものである、
請求項１２記載のコンピュータシステム。
【請求項１４】
前記ブリッジユニットは、更に、前記周辺バスから受信したコマンドに対応するコマンドを前記第１の送信ユニット及び前記第２送信ユニットに選択的に与える、
請求項１３記載のコンピュータシステム。
【請求項１５】
前記制御ユニットは、更に、前記第１受信ユニット、前記第２受信ユニット及び前記ブリッジユニットから受信した複数の制御コマンドに基づいて前記コマンドの伝達を選択的に制御するためのものである、
請求項１４記載のコンピュータシステム。
【請求項１６】
前記制御コマンドのそれぞれは、前記第１受信ユニット、第２受信ユニット及び前記ブリッジユニットによって受信した対応するコマンドの一部を含む、
請求項１５記載のコンピュータシステム。
【請求項１７】
前記制御ユニットは、更に、前記制御コマンドを制御コマンド・バスを介して受信するためのものである、
請求項１６記載のコンピュータシステム。
【請求項１８】
前記周辺バスは、ＰＣＩバス（ＰＣＩ：Peripheral Component Interconnect）である、
請求項１１記載のコンピュータシステム。
【請求項１９】
前記周辺バスは、グラフィックス・バスである、

請求項１１記載のコンピュータシステム。
【請求項２０】
前記第１通信経路及び第３通信経路及び前記第２通信経路及び第４通信経路は、HyperTransport（登録商標）リンクである、
請求項１１記載のコンピュータシステム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、コンピュータ・システムの入出力（Ｉ／Ｏ）に関し、より詳細には、Ｉ／Ｏノードにおけるトランザクション処理に関する。
【背景技術】
【０００２】
ＷＯ－Ａ－９９／１６１９５には、ファイバーチャンネルデータリンクに接続されたＮＬ－ポート、ＮＬ－ポートとトランスミットプロトコルエンジンとの間に結合されたトランスミットフレームＦＩＦＯバッファ、及び前記ＮＬ－ポートと分離した受信プロトコルエンジンとの間に結合された分離受信フレームＦＩＦＯバッファを備えたファイバーチャンネル通信システムが開示されている。ホストメモリーインタフェースによって、ホストコンピュータが、トランスミットプロトコルエンジンと受信プロトコルエンジンとを備えた全二重通信プロセッサに結合される。受信プロトコルエンジンによって、上述のホストメモリーインターフェースを通じてホストメモリーへとフレームが送られる。
「ファイバーチャンネル調停ループ」（1997, Solution Technology, Boulder Creek, CA, US, RW Kembel entitled “In-Depth Fibre Channel Arbitrated Loop”)の２７０頁第５段落～２７５頁第２段落には、デュアルループコンフィグレーションをポートバイパス回路とともに用いることが開示されている。デュアルループコンフィグレーションによって、冗長パスが各デバイスに対してループ状態で提供され、これにより、上述のループのうち一つが故障しても、他のループもしくは冗長ループを用いてデバイスのアクセスを可能としている。
典型的なコンピュータ・システムでは、１個以上のプロセッサが、１つ以上のバスを介して入出力（Ｉ／Ｏ）デバイスと通信し得る。Ｉ／Ｏデバイスは、そのＩ／Ｏデバイスに接続された周辺バスと、プロセッサに接続された共有バスとの間の情報の転送を管理するＩ／Ｏブリッジを介してプロセッサに結合されてもよい。さらに、Ｉ／Ｏブリッジは、システム・メモリとＩ／Ｏデバイスとの間、あるいはシステム・メモリとプロセッサとの間の情報の転送を管理することもある。
【発明の開示】
【発明が解決しようとする課題】
【０００３】
しかし、バス・システムの多くは幾つもの難点を抱えている。例えば、バスに複数のデバイスを接続すると、バスに信号を送るデバイスの静電容量が比較的増大することがある。さらに、共有バスに複数の接続ポイントが存在すると、高い信号周波数において信号の反射が発生し、信号の整合性が失われる。このため、通常はバスの信号周波数を比較的低く保ち、信号の整合性を許容可能なレベルに維持している。信号周波数を比較的低くすると、信号帯域幅が減少し、バスに接続されたデバイスの性能が低下する。
【０００４】
共有バス・システムの別の欠点として、多数のデバイスを接続できず拡張性に欠けることがある。共有バスの利用可能な帯域幅はほぼ固定されている（また、デバイスをさらに接続すると、バスの信号周波数の低下を招く場合は、利用可能な帯域幅が減少することがある）。バスに（直接的もしくは間接的に）接続されているデバイスが要求する帯域幅が、バスの利用可能な帯域幅を超えると、デバイスがバスにアクセスしようとしたときに、デバイスのストールが頻繁に発生し、その共有バスを含むコンピュータ・システムの全体的な性能が大幅に低下する可能性がある。Ｉ／Ｏデバイスによって使用される共有バスの

例に、周辺機器インタフェース（peripheral component interconnect：ＰＣＩ）バスがある。

【０００５】

Ｉ／Ｏブリッジング・デバイスの多くが、何らかのバッファリング・メカニズムを使用して、ＰＣＩバスから最終転送先バスへの転送を待っているトランザクションをバッファしている。しかし、バッファリングが、ＰＣＩバス上でストールを発生させることもある。ストールは、一連のトランザクションがキューにバッファされて、転送先バスへの転送を待っているときに転送先バスでストールが発生してトランザクションが転送できなくなるという場合に発生する。この場合、これらの転送待ちのトランザクションを完了させるトランザクションがキューに到着し、このトランザクションは他のトランザクションの後に格納される。このストールを解消するには、何らかの方法によってキュー内のトランザクションを並び替えて（リオーダー）、新しく到着したトランザクションが、転送を待っているトランザクションよりも先に転送されるようにする必要がある。このように、上記のような状況を回避するためには、ＰＣＩバスの仕様に、ＰＣＩバス・トランザクションの処理およびリオーダーについて定めた一連のリオーダー・ルールが規定されている必要がある。

【０００６】

一部のコンピュータ・システムでは、共有バスの難点の一部を解消するため、デバイス間すなわちノード間でパケット・ベースの通信を行っている。この種のシステムでは、ノードは、情報のパケットを交換することによって、相互に通信することができる。一般に、「ノード」とは、相互接続においてトランザクションに参加することが可能なデバイスを指す。例えば、この相互接続はパケット・ベースであってもよく、ノードはパケットを送受信するように構成されてもよい。一般に、「パケット」とは、パケットを送信する発信元すなわち「ソース」ノードと、このパケットを受信する宛先すなわち「ターゲット」ノードとの２つのノード間で行われる通信のことである。パケットがターゲット・ノードに到着すると、ターゲット・ノードは、パケットが伝達する情報を受け取って、その情報を内部的に処理する。ソース・ノードとターゲット・ノードとの間の通信経路に存在するノードが、ソース・ノードからターゲット・ノードにパケットを中継すなわち転送することもある。

【０００７】

さらに、パケット・ベースの通信とバス・ベースの通信とを併用しているシステムもある。例えば、あるシステムが、ＰＣＩバスとＡＧＰなどのグラフィック・バスとを接続してもよい。ＰＣＩバスはパケット・バス・インタフェースに接続されており、パケット・バス・インタフェースは、ＰＣＩバス・トランザクションをパケット・トランザクションに変換して、パケット・バスで伝送されるようにする。同様に、グラフィックス・バスが、ＡＧＰトランザクションをパケット・トランザクションに変換するＡＧＰインタフェースに接続されていてもよい。各インタフェースは、プロセッサのうちの１つに、または場合によっては別の周辺機器に接続しているホスト・ブリッジと通信し得る。

【０００８】

ＰＣＩデバイスがトランザクションを開始すると、パケット・ベースのトランザクションは、ＰＣＩローカル・バス仕様に規定されているものと同じ順序付けルールによって制約され得る。これは、ＰＣＩバス宛のパケット・トランザクションにも当てはまることがある。パケット・バス・インタフェースで発生し得るトランザクションのストールによって、パケット・バス・インタフェースでデッドロックが発生する可能性があるため、このような順序付けルールがパケット・ベースのトランザクションにも適用されている。このデッドロックが、逆にパケット・バスの側に、更なるストールを発生させる可能性がある。さらに、ＡＧＰトランザクションが、データが正しく配達されるように、トランザクションの順序付けに関する一連のルールに従うこともある。

【０００９】

Ｉ／Ｏノードの構成によっては、あるノードから別のノードへ、トランザクションが、

ホスト・ブリッジに向かう方向かホスト・ブリッジから出る方向に転送され得る。あるいは、特定のノードにおいてトランザクションがパケット・トラフィックに注入され得る。いずれの場合でも、トランザクションが通信経路を送信される際に、トランザクションを制御することができるＩ／Ｏノードのアークテクチャが望ましい。

【課題を解決するための手段】

【００１０】

コンピュータ・システムの入出力ノードにおけるタグ付けおよび調停メカニズムの種々の実施形態が開示される。一実施形態においては、コンピュータ・システムの入出力ノードであって第１通信経路からの第１コマンドを受信するための第１受信ユニットと、前記第１コマンドに対応する第１対応コマンドを第２通信経路に送信するように結合された第１送信ユニット（１４０）と、を有する。

【００１１】

この入出力ノードは、第３通信経路から第２コマンドを受信するための第２受信ユニットと、前記第２コマンドに対応する第２対応コマンドを第４通信経路に送信するように結合された第２送信ユニットと、を有する。

【００１２】

ある実施形態において、上述の通信経路は、ポイント・ツー・ポイント通信リンク、例えばハイパートランスポート（ＨｙｐｅｒＴｒａｎｓｐｏｒｔ™）としてもよい。更に、入出力ノードは、前記第１受信ユニット及び前記第２受信ユニットから選択されたコマンドを受信するよう結合されるとともに前記選択されたコマンドに対応するコマンドを周辺バス（１５２）に送信するためのブリッジユニット（１５０）を備えてもよい。

【００１３】

他の実施形態において、上記入出力ノードは、前記第１通信経路から前記第１通信経路から前記第２通信経路及び前記周辺バスへのコマンド伝達と、前記第３通信経路から前記第４通信経路及び前記周辺バスへのコマンド伝達と、を制御するように結合された制御ユニットを更に有する。

【００１４】

また、この制御ユニットは、前記周辺バスから前記第２通信経路及び前記第４通信経路へのコマンド伝達を制御するためのものとしてもよい。更に、この制御ユニットは、前記第１受信ユニット、前記第２受信ユニット及び前記ブリッジユニットから受信した複数の制御コマンドに基づいて前記コマンドの伝達を選択的に制御するためのものとしてもよい。

【発明を実施するための最良の形態】

【００１５】

本発明は、様々に変形されたり他の形態とすることができるが、そのうちの特定の形態を、例として図中に図示され、かつ本明細書に詳細に記載する。しかし、図面および詳細な説明は、開示の実施形態に本発明を限定することを意図するものではなく、本発明が添付の特許請求の範囲によって規定される本発明の趣旨ならびに範囲に含まれる全ての変形例、均等物および代替例を含むことが意図にあることが理解されるべきである。

【００１６】

図１を参照すると、コンピュータ・システムの一実施形態のブロック図が示される。このコンピュータ・システムは、各々がコヒーレント・パケット・バス１５に相互接続されたプロセッサ１０Ａ～１０Ｄを備える。コヒーレント・パケット・バス１５の各々の部分は、各プロセッサ１０Ａ～Ｄとの間でポイント・ツー・ポイント・リンクを形成し得る。４個のプロセッサがポイント・ツー・ポイント・リンクを使用しているように図示されているが、プロセッサは４個でなくてもよいほか、他のタイプのバスをプロセッサ間の相互接続に使用してもよいという点が留意される。また、このコンピュータ・システムは、符号２０，３０，４０によって示される３つのＩ／Ｏノードも備え、３つのＩ／Ｏノードは各々Ｉ／Ｏパケット・バス５０Ｂ，５０Ｃによって鎖状（チェーン方式）に接続されている。Ｉ／Ｏパケット・バス５０Ａは、ホスト・ノード／プロセッサ１０ＡとＩ／Ｏノード

２０との間に結合されている。プロセッサ１０Ａは、Ｉ／Ｏパケット・バス５０Ａと通信を行うためのホスト・ブリッジを有することができるホスト・ノードとして示されている。プロセッサ１０Ｂ～Ｄが、他のＩ／Ｏパケット・バス（図示なし）と通信を行うホスト・ブリッジを備えていてもよい。Ｉ／Ｏパケット・バス５０Ａ～Ｃによって形成される通信リンクは、ポイント・ツー・ポイント・リンクと呼ばれることもある。Ｉ／Ｏノード２０は、一対の周辺バス２５Ａ，２５Ｂと接続されている。Ｉ／Ｏノード３０はグラフィック・バス３５と接続されており、Ｉ／Ｏノード４０は別の周辺バス４５と接続されている。

【００１７】

プロセッサ１０Ａ～１０Ｄの各々の例に、Ａｔｈｌｏｎ（登録商標）マイクロプロセッサなどのｘ８６マイクロプロセッサがある。さらに、Ｉ／Ｏパケット・バス５０Ａ～５０Ｃなどのパケット・バスの一例として、非コヒーレントなＨｙｐｅｒＴｒａｎｓｐｏｒｔ（登録商標）がある。周辺バス２５Ａ，２５Ｂおよび周辺バス４５の例として、周辺機器相互接続（ＰＣＩ）バスなどの一般的な周辺バスがある。グラフィックス・バス３５の例に、例えばアクセラレイティッド・グラフィックス・ポート（ＡＧＰ）がある。しかし、上記以外のマイクロプロセッサおよび周辺バスを使用してもよいということが理解される。

【００１８】

３つのＩ／Ｏノードがホスト・プロセッサ１０Ａに接続されているが、ノードの数は３でなくてもよいほか、別のトポロジによってノード同士が接続されていてもよいという点が留意される。図１には、理解しやすいようにチェーン方式のトポロジが示されている。

【００１９】

図中の実施形態においては、プロセッサ１０Ａのホスト・ブリッジが、Ｉ／Ｏノード２０，３０または４０などの下りノードから、上りパケット・トランザクションを受信し得る。あるいは、プロセッサ１０Ａのホスト・ブリッジが、例えば周辺バス２５Ａに接続され得る周辺機器（図示なし）などのデバイスに向けて、下り方向にパケットを送信することもある。

【００２０】

動作時に、Ｉ／Ｏノード２０，４０は、ＰＣＩバス・トランザクションを、Ｉ／Ｏストリーム内を伝わる上りパケット・トランザクションに変換し得るほか、下りパケット・トランザクションをＰＣＩバス・トランザクションに変換し得る。プロセッサ１０Ａのホスト・ブリッジ以外のノードから送信されるパケットは、全てプロセッサ１０Ａのホスト・ブリッジに向かって上り方向に向かった後に、他のノードに転送され得る。プロセッサ１０Ａのホスト・ブリッジから送信されるパケットは、全てＩ／Ｏノード２０，３０または４０などの他のノードに下り方向に流れる。本明細書中において、「上り」とは、プロセッサ１０Ａのホスト・ブリッジに向かうパケット・トラフィックの流れを指し、「下り」とは、プロセッサ１０Ａのホスト・ブリッジから離れる方向に向かうパケット・トラフィックの流れを指す。各Ｉ／Ｏストリームは、ユニットＩＤという識別子によって識別され得る。ユニットＩＤはパケット・ヘッダに含まれることもあれば、１つ以上のパケットに含まれる所定数のビットのこともあるということが考察される。本明細書中において、「Ｉ／Ｏストリーム」とは、同じユニットＩＤを有し、このため同じノードから送信された全てのパケット・トランザクションを指す。

【００２１】

説明のため、周辺バス４５にある周辺機器が、周辺バス２５の周辺機器宛のトランザクションを開始したとする。トランザクションはまず、一意的なユニットＩＤを有する１つ以上のパケットに変換されて、次に上り方向に送られ得る。ここで、パケットの各々には、そのパケットを識別する固有の情報が符号化されている。例えば、ユニットＩＤが、パケット・ヘッダ内に符号化されていてもよい。さらに、トランザクションの種類も、パケット・ヘッダ内に符号化されていてもよい。各パケットは、発信元のノードを識別するユニットＩＤを割り当てられ得る。Ｉ／Ｏノード２０が、周辺バス２５にある周辺機器に下

りからパケットを転送しない可能性があるため、パケットが、プロセッサ１０Ａのホスト・ブリッジに向けて上り方向に送信され得る。プロセッサ１０Ａのホスト・ブリッジは、プロセッサ１０Ａのホスト・ブリッジのユニットＩＤを付けてこのパケットを下り方向に送信し、Ｉ／Ｏノード２０は、このパケットを認識して、周辺バス２５にある周辺機器のためにこれを要求する。次に、Ｉ／Ｏノード２０は、パケットを周辺バス・トランザクションに変換して、このトランザクションを周辺バス２５にある周辺機器に送信し得る。

【００２２】

パケット・トランザクションが上り方向または下り方向に伝わる際に、パケットは、１つ以上のＩ／Ｏノードを通過し得る。時として、このような通過点はトンネルと呼ばれることがあり、通過するＩ／Ｏノードはトンネル・デバイスと呼ばれることがある。上り方向から下り方向に、あるいは下り方向から上り方向に送信されるパケットは「転送（forwarded）」トラフィックと呼ばれることがある。さらに、特定のＩ／Ｏノードから発せられ、上りトラフィックに挿入されるパケット・トラフィックは、「注入（injected）」トラフィックと呼ばれることがある。

【００２３】

下記に詳述するように、Ｉ／Ｏノードは、Ｉ／Ｏノードに接続され得る種々のバスの順序付けルールを守るために、パケットのバッファリングのほかにトランザクションのリオーダーを行ってもよい。Ｉ／Ｏノードは、転送トラフィックと注入トラフィックとの両方による、トンネルに入るパケットおよびトンネルから出るパケットの流れを制御する制御ロジックも備え得る。

【００２４】

図２を参照すると、Ｉ／Ｏノードの一実施形態のブロック図が示される。このＩ／Ｏノードは、図１のＩ／Ｏノード２０，３０または４０の代表例であり、以下、説明を簡単にするためにＩ／Ｏノード２０と呼ぶ。図２のＩ／Ｏノード２０は、コマンド・バス１１１を介して送信装置１４０に結合され、コマンド・バス１１２を介して周辺インタフェース１５０に結合されたトランザクション受信装置１１０を備え得る。また、Ｉ／Ｏノード２０は、コマンド・バス１２１を介して送信装置１３０に結合され、コマンド・バス１２２を介して周辺インタフェース１５０に結合されたトランザクション受信装置１２０も備え得る。周辺インタフェース１５０は、コマンド・バス１５１を介して送信装置１３０，１４０に結合されているほか、周辺バス１５２にも結合されている。さらに、Ｉ／Ｏノード２０は、制御コマンド・バス１０１を介して各受信装置および各送信装置に結合されているほか、周辺インタフェースに結合されたトランザクション制御ユニット１００も備える。本明細書中において、コマンド・バスには、コマンド、制御およびデータ用の信号が含まれる。このため、それぞれのコマンド・バスを介してトランザクションまたはコマンドが送信されるといった場合、コマンド・ビットおよびデータ・ビットが含まれるということを指す。

【００２５】

図中の実施形態においては、受信装置１１０と送信装置１４０とは、Ｉ／Ｏトンネルの一通信経路を形成しており、受信装置１２０と送信装置１３０とは、Ｉ／Ｏトンネルの別の通信経路を形成している。この通信経路はそれぞれ一方向的であるため、いずれかの経路が上り経路か下り経路となる。このように、周辺インタフェース１５０からの注入トラフィックが、送信装置１３０，１４０のいずれかに提供される。

【００２６】

受信装置１１０，１２０の各々は、パケット・トランザクションを受け取って、受信バッファに入れる（図示なし）。各トランザクションが受信されると、受信したコマンドに含まれる情報の一部を含む制御コマンドが生成される。制御コマンドには、例えば、発信元ノードのユニットＩＤ、宛先の情報、データ・カウントおよびトランザクションの種類が含まれ得る。なお、制御コマンドは、他の情報を含んでいることもあれば、ここに挙げた情報の幾つかを含まないこともある。制御コマンドはトランザクション制御ユニット１００に送信される。

【００２７】
周辺インタフェース１５０が周辺バス１５２からトランザクションを受信すると、周辺インタフェース１５０も、上記した制御コマンドと類似した情報を含む制御コマンドを生成し得る。また、周辺インタフェース１５０は、１つ以上のバッファ内にトランザクションを記憶して、トランザクション制御ユニット１００に制御コマンドを送信し得る。
【００２８】
トランザクション制御ユニット１００は、受け取った各制御コマンドを、受信順に１つ以上のバッファ構造に記憶し得る。トランザクション制御ユニット１００は、トランザクション制御ユニット１００が自身のバッファに記憶している制御コマンドに基づいて、ソース・バッファ（すなわち受信装置および周辺インタフェースの少なくともいずれか）内で待機中の、対応するコマンドを送信する順序を決定し得る。図３～図６に関して下記に詳述するように、トランザクション制御ユニット１００は、トランザクションの種類、送信装置および宛先のバッファの空きの有無、トランザクションが転送トラフィックか注入トラフィックかであるかなどに基づいて、自身のバッファにあるトランザクション間を調停し得る。このように、トランザクション制御ユニット１１００は、Ｉ／Ｏノードのトンネルを通過するトランザクション全体の流れを調整する機能を果たすこともできる。
【００２９】
トランザクション制御ユニット１００が処理すべきトランザクション間を調停したら、トランザクション制御ユニット１００は、各々のソース・デバイスに対して、転送を待っているトランザクションを宛先デバイスに送信するよう指示する。例えば、トランザクション制御ユニット１００は、自身のバッファから、受信装置１１０から送信装置１４０に転送しようとしているトランザクションを表す制御コマンドを選択する。トランザクション制御ユニット１００は、コマンド・バス１１１を介して送信装置１４０にトランザクションを送信するように、受信装置１１０に対して通知する。次に、送信装置１４０は、トランザクションをチェーンの次のノードに送信する。次のノードは、上りまたは下りにある別のＩ／Ｏノードであることもあれば、図１のホスト・プロセッサ１０Ａなどのホスト・ノードのこともある。さらに、トランザクション制御ユニット１００と送信装置１４０とは、受信バッファに空きが存在するかどうかをもう１つのノードに知らせるロジック（図示なし）を備え得る。
【００３０】
図３を参照すると、トランザクション制御ユニットの一実施形態のブロック図が示される。簡潔を期すと共に説明を簡単にするために、図２と対応する回路部品には同じ符号が付されている。トランザクション制御ユニット１００は、各々１６０，１７０，１８０の符号が付された３つのスケジューラを有する。スケジューラ１６０，１７０，１８０の各々は、一対の仮想チャネル・コマンド・バッファと調停およびバッファ管理ユニットとを有する。スケジューラ１６０の仮想チャネル・コマンド・バッファは、Ｖ．Ｃ．ＦＩＦＯ１６６，１６７と示されており、調停およびバッファ管理ユニットは符号１６８によって示される。同様に、スケジューラ１７０の仮想チャネル・コマンド・バッファは、Ｖ．Ｃ．ＦＩＦＯ１７６，１７７と示されており、調停およびバッファ管理ユニットは符号１７８によって示され、スケジューラ１８０の仮想チャネル・コマンド・バッファはＶ．Ｃ．ＦＩＦＯ１８６，１８７として示されており、調停およびバッファ管理ユニットは符号１８８によって示される。
【００３１】
一般に、「仮想チャネル」とは、様々な処理ノード間でパケットを伝達する通信経路のことである。各仮想チャネルは、他の仮想チャネルの資源に依存しない（すなわち、ある仮想チャネル内で流れるパケットは、通常は物理的な伝送の点で、別の仮想チャネルの有無による影響を受けない）。パケットは、パケット・タイプに応じて仮想チャネルに割り当てられる。同じ仮想チャネル内のパケットは、相互の伝送が物理的に競合することがある（すなわち、同じ仮想チャネル内のパケット間での資源の競合が発生することがある）が、基本的には、別の仮想チャネルのパケットとの伝送は物理的に競合しない。

【００３２】
パケットによっては、他のパケットと論理的に競合する場合もある（すなわち、プロトコル、一貫性などの理由により、あるパケットが別のパケットと論理的に競合することがある）。論理的な理由やプロトコル上の理由により、１番目のパケットが、２番目のパケットよりも先に宛先ノードに到着しなければならない場合に、２番目のパケットが１番目のパケットの伝送を（競合する資源を占有することによって）物理的にブロックしてしまうと、コンピュータ・システムがデッドロックに陥る可能性がある。１番目のパケットと２番目のパケットとに別の仮想チャネルを割り当てると共に、異なる仮想チャネル内のパケットが相互の伝送をブロックできないような伝送媒体をコンピュータ・システムに実装することによって、デッドロックの発生しない運用を実現することができる。特に、異なる仮想チャネルのパケットが同じ物理リンクを介して伝送されるという点に留意されたい。しかし、伝送前に受信バッファを利用することが可能であるため、仮想チャネル同士がこの共有資源を使用している場合であっても、仮想チャネルが相互にブロックし合うことはない。

【００３３】
一観点から見ると、種々のパケット・タイプの各々（例えば、種々のコマンド符号化の各々）が独自の仮想チャネルを割り当てられてもよく、このため、一実施形態においては、各仮想チャネルに別個のバッファが割り当てられる。仮想チャネル毎に別のバッファが使用され得るため、ある仮想チャネルからのパケットが、別の仮想チャネルからのパケットと物理的に競合し得ない（これは、この種のパケットが、他のバッファに入れられるためである）。

【００３４】
各スケジューラは、特定の宛先と２つのソースとに対応している。図中の実施形態においては、スケジューラ１６０は、図２の送信装置１３０を宛先に、受信装置１２０および周辺インタフェース／ブリッジ１５０をソースに有するトランザクションを制御する。同様に、図３のスケジューラ１７０は、図２の送信装置１４０が宛先であり、受信装置１１０およびブリッジ１５０が送信元であるトランザクションを制御する。最後に、図３のスケジューラ１８０は、図２のブリッジ１５０が宛先であり、受信装置１１０および受信装置１２０が送信元であるトランザクションを制御する。図３において、各仮想チャネル・コマンド・バッファは、それぞれの受信装置またはブリッジから、その受信装置またはブリッジが受け取ったトランザクションに対応する制御コマンドを受け取る。制御コマンドには、どのスケジューラに制御コマンドを送信するかを指定する宛先ビットが含まれていてもよい。通常、制御コマンドには、宛先ビットが１つだけ設定されている。しかし、トランザクションがブロードキャスト・メッセージの場合には、複数のスケジューラが制御コマンドを受け取ることができるように、複数の宛先ビットが設定されていてもよい。

【００３５】
議論を簡略化するために、スケジューラ１６０についてのみ詳述する。制御コマンドが受信されてＶ．Ｃ．ＦＩＦＯ１６６または１６７に格納されると、トランザクションの種類に応じてそれぞれのＦＩＦＯ部に格納される。Ｖ．Ｃ．ＦＩＦＯ１６６と１６７とは同一であるため、Ｖ．Ｃ．ＦＩＦＯ１６６のみについて詳述する。Ｖ．Ｃ．ＦＩＦＯ１６６は、ポスト済み、未ポストおよび応答の３種類のトランザクションに対応する３つのＦＩＦＯ部を有する。制御コマンドは、受信順にそれぞれのＦＩＦＯに格納される。しかし、トランザクションの種類によっては、元のコマンドを生成した種々のバスまたはデバイスのタイプに関連する順序付けルールを保持するために、トランザクションが受信順に処理されないこともある。

【００３６】
図４～図６に関して下記に詳述するように、調停およびバッファ管理ロジック１６８は、１番目、２番目、それ以降に処理されるトランザクションを、Ｖ．Ｃ．ＦＩＦＯ１６６またはＶ．Ｃ．ＦＩＦＯ１６７のトランザクション間で調停するように構成され得る。例えば、応答コマンドより前にＶ．Ｃ．ＦＩＦＯ１６６に到着したポスト済み命令は、順序

付けルールにより、この応答コマンドより後に処理する必要がある。さらに、調停およびバッファ管理ロジック１６８は、公平ルールの組と次のＩ／Ｏノードまたはホスト・ブリッジの受信バッファの空きの有無とに基づいて、どのＶ．Ｃ．ＦＩＦＯのトランザクションを処理すべきか調停し得る。宛先が図２のブリッジ１５０である場合、その調停ルールは、上記の調停ルールとは異なっていてもよい。

なお、上述の例は、トランザクション制御ユニットの一実装例を示す物である。追加機能を実行することができる他の実装例用いることも勿論可能である。

【００３７】

図４を参照すると、スケジューラの一実施形態のブロック図が示される。簡潔を期すと共に説明を簡単にするために、図３の回路部品と対応するものには同一の符号を付している。トランザクション・スケジューラ４００は、調停および公平ロジック４５０に結合された仮想チャネルＦＩＦＯバッファ４１０を備える。トランザクション・スケジューラ４００は、これも調停および公平ロジック４５０に結合された仮想チャネルＦＩＦＯバッファ４２０も備える。調停および公平ロジック４５０はＦＩＦＯバッファ４６０に接続されており、ＦＩＦＯバッファ４６０はバッファ管理ロジック４７０に接続されている。バッファ管理ロジックの出力は、出力レジスタ４８０によってラッチされる。

【００３８】

図３に関して上記したように、仮想チャネルＦＩＦＯバッファ４１０，４２０の各々は、例えば、図２の受信装置１１０またはブリッジ１５０などのそれぞれの送信元から制御コマンドを受信し得る。制御コマンドは、その制御コマンドが表すトランザクションの種類によって、仮想チャネルＦＩＦＯバッファ４１０，４２０に格納され得る。具体的にいえば、制御コマンドは、ポスト済みコマンド、未ポスト・コマンド、応答コマンドのいずれかを表し得、このためポスト済み、未ポストまたは応答の各キューを表し得る。

【００３９】

図中の実施形態においては、調停および公平ロジック４５０は、調停ユニット４３０，４４０および公平ユニット４４５を備える。調停ユニット４３０は、仮想チャネルＦＩＦＯバッファ４１０に記憶されている制御コマンドを１つ選択するように構成され得る。下記に詳述するように、この選択処理には、所定の調停アルゴリズムによって１つを「ウィナー（勝者）」として選択することが含まれ得る。同様に、調停ユニット４４０は、調停ユニット４３０と同様のアルゴリズムを使用して、仮想チャネルＦＩＦＯバッファ４２０に記憶されている制御コマンドを１つ選択するように構成され得る。次に、公平ユニット４４５は、調停ユニット４３０，４４０によってウィナーとして選択されたトランザクションのうちの１つを選択し得る。公平ユニット４４５は、トランザクションが転送トランザクションか、注入トランザクションかに基づいて、公平アルゴリズムを使用し得る。また、調停ユニット４３０，４４０は、スケジューラの出力先に応じて、次のＩ／Ｏノードあるいは図１に示したプロセッサ１０Ａのホスト・ブリッジの受信バッファなど、対応するトランザクションの宛先のバッファにあるバッファ空間を監視するロジック（図示なし）を有していてもよい。

【００４０】

図中の実施形態においては、制御コマンドがスケジューラ４００内を伝わるのに３クロック・サイクルの遅延が生じ得る。仮想チャネルＦＩＦＯバッファ４１０および４２０をそれぞれ迂回するためのバイパス４１５およびバイパス４２５が図示されている。スケジューラ４００がソースから制御コマンドを受け取り、ある仮想チャネルＦＩＦＯバッファの各キューが空である場合、仮想チャネルＦＩＦＯバッファをバイパスすることによって、１クロック・サイクル節約することが可能となる。例えば、未ポスト制御コマンドが、現在空である仮想チャネルＦＩＦＯバッファ４１０に格納される場合を考える。トランザクションの宛先のバッファに使用可能なバッファ空間が存在することが調停ユニット４３０によって示される場合、調停ユニット４３０にあるロジックによって、未ポスト制御コマンドが仮想チャネルＦＩＦＯバッファ４１０を迂回して、直接ＦＩＦＯバッファ４６０に格納され得る。さらに、上記したように、公平ユニット４４５は、公平アルゴリズムに

応じてバイパスを許可し得る。このように、上記の例においては、１クロック・サイクル分だけ遅延を短縮することができる。本実施形態においては３クロック・サイクルの遅延が発生していたが、遅延が３クロック・サイクルよりも短い実施形態や長い実施形態も可能であるということが考察される。さらに、バイパス４１５，４２５によって達成される実際の遅延の短縮が、本例よりも大きい場合もあれば小さい場合もある。

【００４１】

図４のＦＩＦＯバッファ４６０は、ウィナーの制御コマンドを受信し得る。図中の実施形態においては、ＦＩＦＯバッファ４６０は２階層のバッファであるが、他の実施形態においては、ＦＩＦＯバッファ４６０の階層数が多い場合もあれば少ない場合もあるということが考察される。

【００４２】

バッファ管理ロジック４７０は、図２の送信装置１３０または１４０、あるいはブリッジ１５０にあるバッファ空間を監視するように構成される。ＦＩＦＯバッファ４６０にトランザクションが記憶されると、バッファ管理ロジック４７０は、次のバッファの空きの有無を確認して、バッファ空間に空きが生じるまで制御コマンドを保持するか、あるいは出力レジスタ４８０に出力する。制御コマンドが出力レジスタ４８０によってラッチされると、それぞれのトランザクションの送信元は、図２の送信装置１３０または１４０、あるいはブリッジ１５０に制御コマンドに対応するトランザクションを送信可能であることを通知される。

【００４３】

図５を参照すると、タグ付けロジックを備えたスケジューラの一実施形態のブロック図が示される。スケジューラ５００は、仮想チャネルＦＩＦＯバッファ５０５に結合されたタグ付けロジック５１０を備える。仮想チャネルＦＩＦＯバッファ５０５は、ポスト済み、未ポストおよび応答の３種類のトランザクションに対応する３つのキューを有する。タグ比較／調停ロジック・ユニット５２０は、仮想チャネルＦＩＦＯバッファ５０５に結合されている。このほか、仮想チャネルＦＩＦＯバッファ５０５の分解図も示されている。この分解図には、未ポスト・キューと応答キューの各々は、対応するタグを１つ有することが示される。しかし、ポスト済みキューは、未ポストキューと応答キューとに対応する２つのタグを有する。

【００４４】

タグ付けロジック５１０は、各制御コマンドが仮想チャネルＦＩＦＯバッファ５０５に記憶される前に、その制御コマンドにタグを割り当て得る。制御コマンドには、図２の受信装置１１０またはブリッジ１５０などのソース・ユニットから制御コマンドを受信した順に、タグが割り当てられてもよい。タグは、制御コマンドの最後に追加されてもよい。

【００４５】

制御コマンドが仮想チャネルＦＩＦＯバッファ５０５の一番上の場所に達すると、タグ比較／調停ロジック・ユニット５２０は、３つの仮想チャネル間を調停して、ウィナーの制御コマンドを選択するように構成され得る。ウィナーは、順序付けルールのセットに基づいたアルゴリズムを使用して選択され、この順序付けルールのセットは、Ｉ／Ｏノードに接続されている周辺バスが従っている順序付けルールに対応したものであってもよい。一実施形態においては、この順序付けルールは、ＰＣＩの順序付けルールに対応したものであってもよい。別の実施形態においては、この順序付けルールは、ＡＧＰの順序付けルールに対応したものであってもよい。

【００４６】

図中の実施形態においては、仮想チャネルＦＩＦＯバッファ５０５には最大で１６の場所を含むことができるため、タグは４ビットであり得る。しかし、他の実施形態においては、仮想チャネルＦＩＦＯバッファ５０５に含まれる場所の数が１６以外であってもよく、このためタグのビット数が適宜変わってもよいという点が留意される。タグ付けロジック５１０は、仮想チャネルＦＩＦＯバッファ５０５に格納されている未ポスト制御コマンドおよび応答制御コマンドについて、現在のタグを監視するカウンタ・ロジック（図示な

し）を備え得る。このタグは、ポスト済み制御コマンドを受信したとき、現在のポスト済み制御コマンドより先で、かつ以前のポスト済み制御コマンドより後に、ブロック可能な未ポストの制御コマンドまたは応答制御コマンドを少なくとも１つ受信している場合に、未ポスト・カウンタまたは応答カウンタをそれぞれインクリメントするアルゴリズムに従って割り当てられ得る。本明細書中において、未ポスト制御コマンドまたは応答制御コマンドがブロック可能であるとは、仮想チャネルＦＩＦＯバッファ５０５内で、未ポスト制御コマンドまたは応答制御コマンドが、ポスト済み制御コマンドを追い越すことができることを示す特別なビットが、各制御コマンド内でセットされている場合を指す。一実施形態においては、この特別なビットをＰａｓｓＰＷビットと呼ぶ。

【００４７】

タグ付け・アルゴリズムの使用を示すために、表１に、制御コマンドを受信する順序と、ポスト済み、未ポストおよび応答の３つのキューに入るタグを示す。１番目の列は、９個の制御コマンドを受信する順序を示す。２番目の列は、受信コマンドの種類を示す。３番目の列は、未ポスト・コマンドおよび応答コマンドに割り当てられたタグを、４番目および５番目の列は、ポスト済み制御コマンドによってインクリメントされ得る後の未ポスト・コマンドおよび応答コマンドのカウンタ値を示す。ポスト済み制御コマンドには２つのタグが割り当てられるため、未ポスト・カウンタおよび応答カウンタの各々の現在のカウンタ値に示される両方のタグが割り当てられる。仮想チャネルＦＩＦＯバッファ５０５の分解図は、表１のコマンドが格納される様子を示す。

【表１】

受信順	制御コマンド	ＰａｓｓＰＷビット	タグ値	未ポスト・カウンタ	応答カウンタ
１	ポスト済み１	０		０	０
２	応答１	０	０	０	０
３	ポスト済み２	０		０	１
４	未ポスト１	０	０	０	１
５	応答２	０	１	０	１
６	未ポスト２	０	０	０	１
７	ポスト済み３	０		１	２
８	応答３	０	２	１	２
９	未ポスト３	０	１	１	２

【００４８】

表２に、ＰａｓｓＰＷビットにより、表１に示した受信制御コマンドのタグがどのように変わるかを示す。未ポスト制御コマンドまたは応答制御コマンドにＰａｓｓＰＷビットがセットされている場合、次のポスト済み制御コマンドが到着しても、それぞれのカウンタがインクリメントされない。例えば、表２において、応答１にＰａｓｓＰＷビットがセットされているため、ポスト済み２の制御コマンドが受信されたときに、応答カウンタがインクリメントされない。しかし、未ポスト２の制御コマンドにＰａｓｓＰＷビットがセットされており、ポスト済み３の制御コマンドによって、未ポスト・カウンタと応答カウンタとがインクリメントされる。これは、最後のポスト済み制御コマンドより後で、かつ現在のポスト済み制御コマンドより前に、ＰａｓｓＰＷビットがクリアされた状態で未ポスト１制御コマンドが受信され、このためカウンタをインクリメントするタグ付けルールを満たしているためである。ＰａｓｓＰＷビットがセットされると、未ポスト制御コマンドまたは応答制御コマンドがポスト済み制御コマンドを追い越し可能であることを示すと記載したが、このロジックが逆転している実施形態も可能であるということが考察される。

【表２】

受信順	制御コマンド	Ｐａｓｓ PWビット	タグ値	未ポスト・カウンタ	応答カウンタ
1	ポスト済み１	0		0	0
2	応答１	1	0	0	0
3	ポスト済み２	0		0	0
4	未ポスト１	0	0	0	0
5	応答２	0	0	0	0
6	未ポスト２	1	0	0	0
7	ポスト済み３	0		1	1
8	応答３	0	1	1	1
9	未ポスト３	0	1	1	1

【００４９】

図５に戻ると、タグ比較および調停ロジック・ユニット５２０は、調停中に、各制御コマンドの後に追加されたタグを比較することによって、仮想チャネルＦＩＦＯバッファ５０５からウィナーの制御コマンドを選択するように構成される。さらに、タグ比較および調停ロジック・ユニット５２０は、タグを比較する前に、各仮想チャネルについて、次のＩ／Ｏノードのバッファ空間に空きがあるかどうかを判定してもよい。ブロックされている仮想チャネルがあれば、そのチャネルは、このサイクルの調停の対象とならない。３つのチャネル全てが調停の対象となる場合は、未ポスト・チャネルと応答チャネルとの間でラウンド・ロビン方式で調停が行われ、ポスト済みチャネルと未ポスト・チャネルとの間の比較、およびポスト済みチャネルと応答チャネルとの間の比較が個別に行われる。最後に、タグ比較および調停ロジック・ユニット５２０は、公平アルゴリズムを使用してウィナーを決定し得る。この公平アルゴリズムについては、図６を参照して下記に詳述する。

【００５０】

次に図６を参照すると、欠乏回避ロジックを備えたトランザクション・スケジューラの一実施形態のブロック図が示される。トランザクション・スケジューラ６００は、調停回路６５０に結合された仮想チャネルＦＩＦＯバッファ６１０を備える。トランザクション・スケジューラ６００は、これも調停回路６５０に結合された仮想チャネルＦＩＦＯバッファ６２０も備える。調停回路６５０はＦＩＦＯバッファ６７０に接続されており、ＦＩＦＯバッファ６７０はバッファ管理ロジック６８０に接続されている。バッファ管理ロジックの出力は、出力レジスタ６９０によってラッチされる。

【００５１】

上記の図３，図４と同様に、仮想チャネルＦＩＦＯバッファ６１０，６２０の各々は、例えば、図２の受信装置１１０またはブリッジ１５０などの入力元から制御コマンドをそれぞれ受信し得る。制御コマンドは、その制御コマンドが表すトランザクションの種類に従って、仮想チャネルＦＩＦＯバッファ６１０，６２０に格納され得る。例えば、制御コマンドは、ポスト済みコマンド、未ポスト・コマンド、応答コマンドのいずれかを表し得、このためポスト済み、未ポストまたは応答の各キューを表し得る。

【００５２】

図中の実施形態においては、調停回路６５０は、調停ユニット６３０，６４０および公平回路６４５を備える。調停ユニット６３０，６４０は、調停サイクルの間に、仮想チャネルＦＩＦＯバッファ６１０，６２０のそれぞれに記憶されている制御コマンドを選択するように構成され得る。さらに、公平回路６４５は、調停ユニット６３０か６４０かのいずれが、ウィナーのトランザクションを選択するかを決定し得る選択条件を提供してもよい。図７に関して下記に詳述するように、公平回路６４５は公平アルゴリズムを使用して、帯域幅を振り分ける調停の優先順位を確立し得る。公平アルゴリズムは、所定数の調停サイクルの間、あるトランザクションがブロックされていることを判定する欠乏回避ロジックを使用してもよい。

【００５３】

図７を参照すると、図６に示した公平回路の一実施形態のブロック図が示される。公平回路６４５は、３２個の３ビット・カウンタ０～３１を備えた公平ユニット７００を有し、３ビット・カウンタ０～３１は８ビット・カウンタ７０５に結合されている。ラッチ７１０がカウンタ７０５に結合されている。挿入レート・ロジック７１５がラッチ７１０に結合されている。また、公平回路６４５は、３つの仮想チャネル・カウンタ７５５～７５７を備えた欠乏ユニット７５０も有し、仮想チャネル・カウンタ７５５～７５７は、欠乏閾値レジスタ７６０およびトランザクション選択ユニット７７５に結合されており、トランザクション選択ユニット７７５は公平ユニット７００および欠乏ユニット７５０に結合されている。

【００５４】

一実施形態においては、トランザクションがトンネルを通って転送される度に、そのトランザクションを送信したＩ／Ｏノードに対応する３ビット・カウンタ０～３１がインクリメントされ得る。また、３ビット・カウンタ０～３１のいずれかがインクリメントされる度に、カウンタ７０５もインクリメントされ得る。３ビット・カウンタ０～３１のいずれか１つがオーバーフローすると、カウンタ７０５の値がラッチ７１０によって取得される。取得した値は、ある時点での下りノードのトランザクション要求の発生レート（rate）を表し得る。挿入レート・ロジック７１５は、取得した値を使用して、そのノードの許容可能な挿入レートを計算することができる。

【００５５】

トランザクション選択ユニット７７５によって、図６の調停ユニット６３０または６４０のポインタが、調停サイクル中に考慮されている仮想チャネル・バッファをポイントするようになる。転送用の仮想チャネル・バッファの仮想チャネルにも、挿入用のバッファの同じ仮想チャネルにもトランザクションが存在する場合があり得る。図７において、トランザクション選択ユニット７７５は、公平ユニット７００によって確立された優先順位に従って、この２つの仮想チャネル・バッファを交互に選択し得る。しかし、トランザクションの宛先のバッファに空きが存在しないなどの要因によって、ある仮想チャネルがブロックされている場合には、調停を続行している間、調停ロジックは、ブロックされているチャネルを避け、次の仮想チャネルを選択し得る。ブロックされているチャネルが利用可能になったときに、転送用のチャネルのみに空きが存在し、公平アルゴリズムによって転送用のチャネルは挿入用のチャネルよりも優先されるため、転送用のチャネルからのトランザクションが送信され得る。以前はブロックされていたチャネルが次に利用可能になると、現在の調停サイクルにおいて優先されないため、このチャネルは再度スキップされ得る。この状態はしばらく続き、このために挿入用の仮想チャネルが「欠乏」し得る。別の実施形態においては、ノードの構成によっては、同じような状況下で、注入チャネルのために転送チャネルが欠乏し得るということが考察される。

【００５６】

特定のチャネルが欠乏しないように、欠乏ユニット７５０は、トランザクションがブロックされた回数を監視し得る。トランザクション選択ユニット７７５が、調停の対象となっているがブロックされているトランザクションがどれであるかを判定する度に、トランザクション選択ユニット７７５は、対応する仮想チャネル・カウンタ７５５～７５７をインクリメントする。欠乏閾値レジスタ７６０は、調停中のチャネルのスキップ回数の最大数に対応する値を保持している。欠乏閾値レジスタ７６０に記憶される値は、特定の時点における計算によって求めた要求の発生レートによって動的に変更され得る。仮想チャネル・カウンタ７５５～７５７の値のいずれかが、欠乏閾値レジスタに記憶されている値に達すると、次の調停サイクルで、ブロックされているトランザクションがトランザクション選択ユニット７７５によって選択されるように、対応する仮想チャネルの優先順位が変更され得る。このように、優先順位を動的に変更することによって、特定のチャネルの欠乏を回避することができる。上記のロジック構成は実装の一例に過ぎないという点が留意される。別法による実施形態においては、カウンタの数が異なるか、カウンタのビット数が異なるロジック構成を使用して、上記した機能を実現してもよいということが考察され

る。

【００５７】

また、上記の実施形態は、スケジューラの一実装であることも留意される。別法による実施形態においては、別の機能および異なる機能の少なくともいずれかを実行する別の実装を有していてもよいということが考察される。

【００５８】

上記の開示が完全に理解されれば、数多くの変形例および変更例が当業者にとって自明となるであろう。添付の特許請求の範囲はこのような変更例および変更例を全て包含するものと解釈されることが意図される。

【産業上の利用可能性】

【００５９】

本発明は、一般にコンピュータ・システムの入出力（Ｉ／Ｏ）に適用可能であり、より詳細には、Ｉ／Ｏノードにおけるトランザクション処理に適用可能である。

【図面の簡単な説明】

【００６０】

【図１】コンピュータ・システムの一実施形態を示すブロック図である。

【図２】Ｉ／Ｏノードの一実施形態の一実施形態を示すブロック図である。

【図３】トランザクション制御ユニットの一実施形態を示すブロック図である。

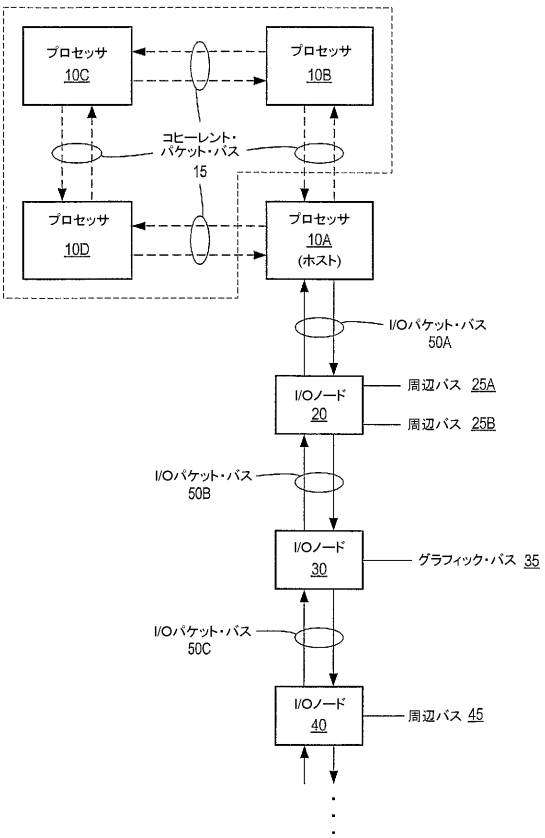
【図４】スケジューラの一実施形態を示すブロック図である。

【図５】タグ付けロジックを備えたスケジューラの一実施形態を示すブロック図である。

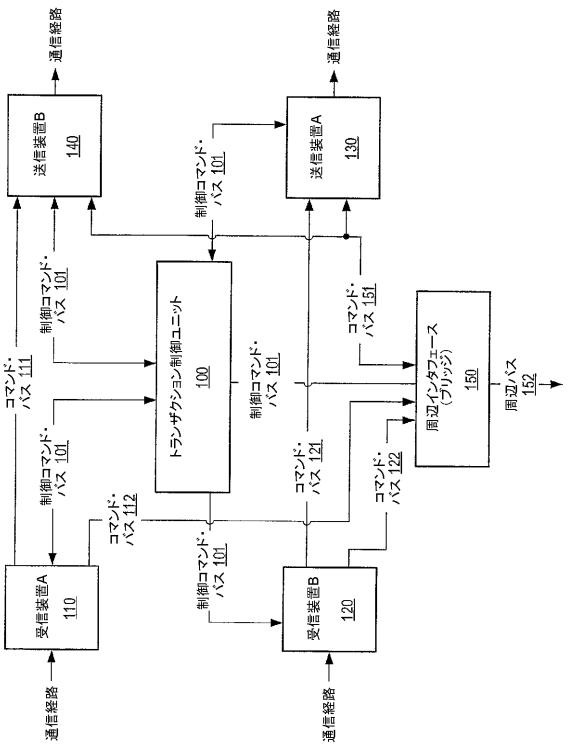
【図６】欠乏回避ロジックを備えたトランザクション・スケジューラの一実施形態を示すブロック図である。

【図７】公平回路の一実施形態を示すブロック図である。

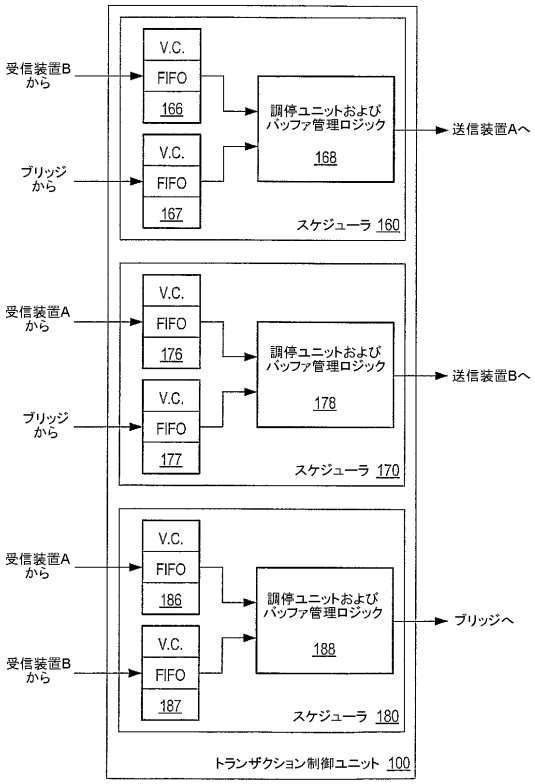
【図１】



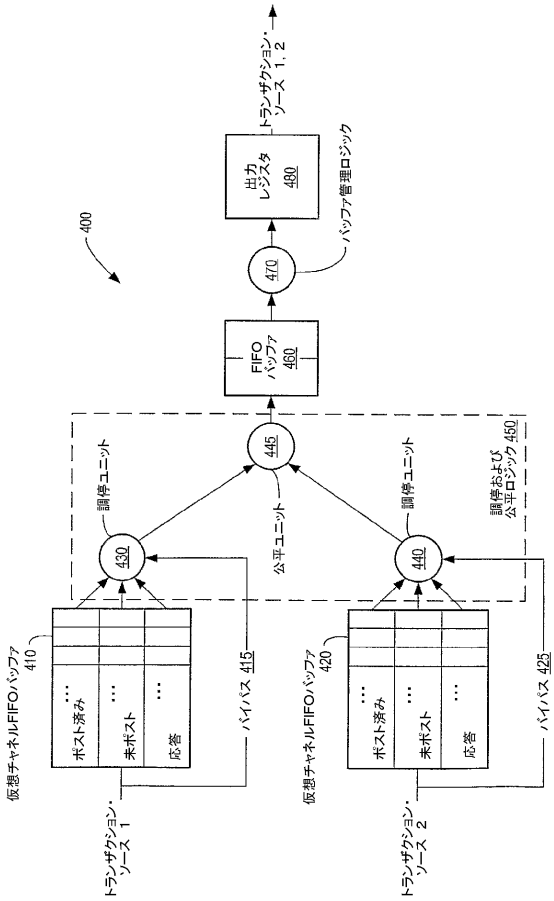
【図２】



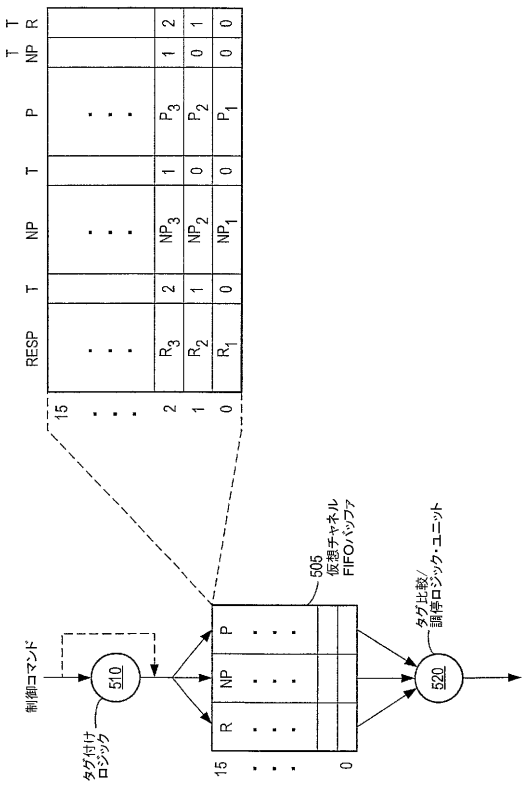
【図３】


受信装置Bから
ブリッジから
V.C.
FIFO
166
167
調停ユニットおよびバッファ管理ロジック
168
送信装置Aへ
スケジューラ 160
受信装置Aから
ブリッジから
176
177
178
送信装置Bへ
スケジューラ 170
186
187
188
ブリッジへ
スケジューラ 180
トランザクション制御ユニット 100

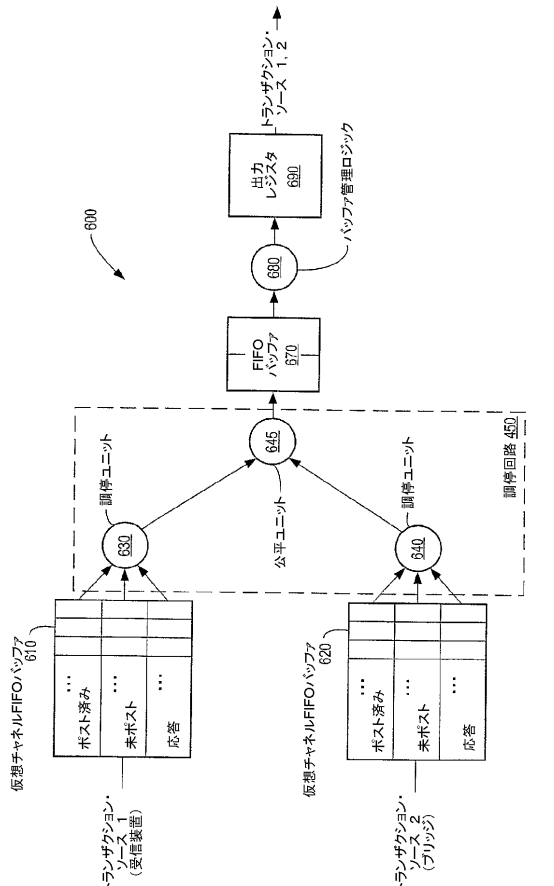
【図４】


400
トランザクション・ソース 1, 2
出力レジスタ 480
バッファ管理ロジック
470
FIFOバッファ 460
445
調停ユニット
公平ユニット
調停および公平ロジック 450
430
440
仮想チャネルFIFOバッファ 410
420
ポスト済み
未ポスト
応答
バイパス 415
バイパス 425
トランザクション・ソース 1
トランザクション・ソース 2

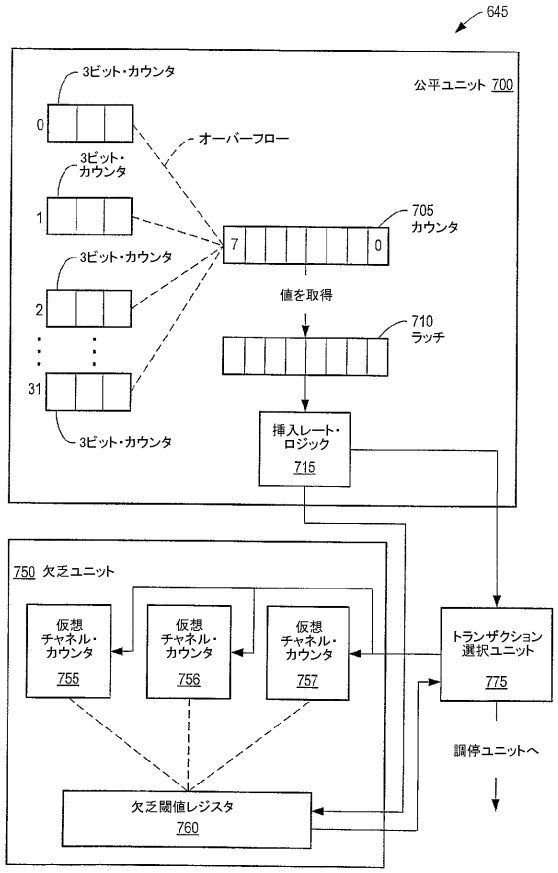
【図５】


RESP
NP
P
T
R
制御コマンド
タグ付けロジック
510
仮想チャネルFIFOバッファ
505
タグ比較/調停ロジック・ユニット
520

【図６】


600
トランザクション・ソース 1, 2
出力レジスタ 690
バッファ管理ロジック
680
FIFOバッファ 670
645
調停ユニット
公平ユニット
調停回路 450
630
640
仮想チャネルFIFOバッファ 610
620
ポスト済み
未ポスト
応答
トランザクション・ソース 1（受信装置）
トランザクション・ソース 2（ブリッジ）

【図７】


645
公平ユニット 700
3ビット・カウンタ
0
オーバーフロー
3ビット・
カウンタ
1
705
カウンタ
7
0
3ビット・カウンタ
2
値を取得
710
ラッチ
31
3ビット・カウンタ
挿入レート・
ロジック
715
750 欠乏ユニット
仮想
チャネル・
カウンタ
755
仮想
チャネル・
カウンタ
756
仮想
チャネル・
カウンタ
757
トランザクション
選択ユニット
775
調停ユニットへ
欠乏閾値レジスタ
760

フロントページの続き

(74)代理人 100108604
弁理士 村松 義人
(72)発明者 スティーブン シー. エニス
アメリカ合衆国、テキサス州 ７８７０４、オースティン、エアロール ウェイ ２１００－ビー
(72)発明者 ラリー ディー. ヒューイット
アメリカ合衆国、テキサス州 ７８７４６、オースティン、ベンド オリバー ドライブ ６１０３

審査官 木村 貴俊

(56)参考文献 特開平１１－１４３８４７（ＪＰ，Ａ）
特開平０８－２９７６３０（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
G06F 13/20-13/378