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

特許第4290556号
(P4290556)

(45)発行日　平成21年7月8日(2009.7.8)　(24)登録日　平成21年4月10日(2009.4.10)

(51)Int.Cl.		F I		
G06F 13/36	(2006.01)	G06F	13/36	310A
G06F 13/14	(2006.01)	G06F	13/14	310H
H04L 12/56	(2006.01)	H04L	12/56	Z

請求項の数 31　(全 18 頁)

(21)出願番号	特願2003-550076 (P2003-550076)
(86)(22)出願日	平成14年9月12日 (2002.9.12)
(65)公表番号	特表2005-512194 (P2005-512194A)
(43)公表日	平成17年4月28日 (2005.4.28)
(86)国際出願番号	PCT/US2002/028985
(87)国際公開番号	WO2003/048951
(87)国際公開日	平成15年6月12日 (2003.6.12)
審査請求日	平成17年9月12日 (2005.9.12)
(31)優先権主張番号	09/998,758
(32)優先日	平成13年11月30日 (2001.11.30)
(33)優先権主張国	米国(US)

(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】マルチプロセッサ・コンピュータ・システムにおける接続のためのスイッチングＩ／Ｏノード

(57)【特許請求の範囲】

【請求項１】

マルチプロセッサ・コンピュータ・システム用の入出力（Ｉ／Ｏ）ノード・スイッチであって、前記入出力ノード・スイッチは、

集積回路チップに実装されており、周辺バスからの複数の周辺トランザクションを受信するように結合され、かつ前記複数の周辺トランザクションに対応する複数の上りパケット・トランザクションを送信するように構成されたブリッジ・ユニットと、

前記集積回路チップに実装されており、前記複数の上りパケット・トランザクションを内部ポイントツーポイント・パケット・バス・リンクで受信するように結合され、かつ前記複数の上りパケット・トランザクションの各々の宛先を判定するように構成されたパケット・バス・スイッチ・ユニットと、を備え、

前記パケット・バス・スイッチ・ユニットは、前記複数の上りパケット・トランザクションの前記各々の宛先を判定すると、前記複数の上りパケット・トランザクションのうちの選択されたパケット・トランザクションを第１のポイントツーポイント・パケット・バス・リンクに結合された第１のプロセッサ・インタフェースにルーティングすると共に、前記複数の上りパケット・トランザクションのうちの他のパケット・トランザクションを第２のポイントツーポイント・パケット・バス・リンクに結合された第２のプロセッサ・インタフェースにルーティングするようにさらに構成されており、

前記第１のポイントツーポイント・パケット・バス・リンク及び前記第２のポイントツーポイント・パケット・バス・リンクは、各々一方向ワイヤの第１の組と一方向ワイヤの

第２の組とを有するハイパートランスポート（HyperTransport：登録商標）互換のリンクであり、前記一方向ワイヤの第１の組および前記一方向ワイヤの第２の組の各々は、制御パケットとデータ・パケットとを含むパケットを伝送するように構成されており、前記制御パケットは、コマンド・パケット、情報パケットおよび応答パケットを含み、

前記パケット・バス・スイッチ・ユニットは、更に、前記第１ポイントツーポイント・パケット・バス・リンクおよび前記第２ポイントツーポイント・パケット・バス・リンクに結合された上流デバイスのＩ／Ｏ負荷又はトランザクション・タイプに基づいて、前記複数の上りパケット・トランザクションのそれぞれの送信先を決定するように構成されている、

入出力ノード・スイッチ。

【請求項２】

前記複数の上りパケット・トランザクションのうちの前記選択されたパケット・トランザクションを受信して、前記複数の上りパケット・トランザクションのうちの前記選択されたパケット・トランザクションを前記第１のポイントツーポイント・パケット・バス・リンクで送信するように結合された第１のトランシーバ・ユニットをさらに備える、

請求項１に記載の入出力ノード・スイッチ。

【請求項３】

前記複数の上りパケット・トランザクションのうちの前記他のパケット・トランザクションを受信して、前記複数の上りパケット・トランザクションのうちの前記他のパケット・トランザクションを前記第２のポイントツーポイント・パケット・バス・リンクで送信するように結合された第２のトランシーバ・ユニットをさらに備える、

請求項２に記載の入出力ノード・スイッチ。

【請求項４】

前記パケット・バス・スイッチ・ユニットは、前記複数の上りパケット・トランザクションの各々に関連付けられたアドレスをデコードするようにさらに構成されている、

請求項３に記載の入出力ノード・スイッチ。

【請求項５】

前記パケット・バス・スイッチ・ユニットは、前記アドレスに基づいて、前記複数の上りパケット・トランザクションのうちの追加のパケット・トランザクションをブロックするようにさらに構成されている、

請求項４に記載の入出力ノード・スイッチ。

【請求項６】

前記アドレスに基づいて、前記複数のパケット・トランザクションのうちの前記追加のパケット・トランザクションを受信して、前記パケット・トランザクションを第３のポイントツーポイント・パケット・バス・リンクで他のノードに送信するように結合された第３のトランシーバ・ユニットをさらに備える、

請求項５に記載の入出力ノード・スイッチ。

【請求項７】

前記第３のトランシーバ・ユニットは、第２の複数の上りパケット・トランザクションを前記第３のポイントツーポイント・パケット・バス・リンクで受信し、前記第２の複数の上りパケット・トランザクションを前記パケット・バス・スイッチに送信するようにさらに構成されている、

請求項６に記載の入出力ノード・スイッチ。

【請求項８】

前記パケット・バス・スイッチ・ユニットは、前記第２の複数の上りパケット・トランザクションの各々に関連付けられたアドレスをデコードし、前記第２の複数の上りパケット・トランザクションのうちの選択されたパケット・トランザクションを前記第１のポイントツーポイント・パケット・バス・リンクに結合された前記第１のプロセッサ・インタフェースにルーティングすると共に、前記第２の複数の上りパケット・トランザクションの各々に関連付けられた前記アドレスに基づいて、前記複数の上りパケット・トランザク

ションのうちの他のパケット・トランザクションを前記第２のポイントツーポイント・パケット・バス・リンクに結合された前記第２のプロセッサ・インタフェースにルーティングするようにさらに構成されている、
請求項７に記載の入出力ノード・スイッチ。
【請求項９】
前記パケット・バス・スイッチ・ユニットは、前記第１のプロセッサ・インタフェースによって送信された第１の複数の下りパケット・トランザクションと、前記第２のプロセッサ・インタフェースによって送信された第２の複数の下りパケット・トランザクションとを受信するようにさらに構成されている、
請求項３に記載の入出力ノード・スイッチ。
【請求項１０】
前記パケット・バス・スイッチ・ユニットは、前記第１の複数の下りパケット・トランザクションと前記第２の複数の下りパケット・トランザクション間を調停するようにさらに構成されている、
請求項９に記載の入出力ノード・スイッチ。
【請求項１１】
前記パケット・バス・スイッチ・ユニットは、前記第１の複数の下りパケット・トランザクションと、前記第２の複数の下りパケット・トランザクションとを前記第１の内部ポイントツーポイント・パケット・バス・リンクにルーティングするようにさらに構成されている、
請求項１０に記載の入出力ノード・スイッチ。
【請求項１２】
前記ブリッジ・ユニットは、前記第１複数の下りパケット・トランザクションと第２の複数の下りパケット・トランザクションの各々のアドレスをデコードし、前記アドレスに基づいて、前記第１複数の下りパケット・トランザクションと第２の複数の下りパケット・トランザクションの選択されたパケット・トランザクションに対応するバス・サイクルを始めるようにさらに構成されている、
請求項１１に記載の入出力ノード・スイッチ。
【請求項１３】
前記パケット・バス・スイッチ・ユニットは、プログラム可能なルックアップ・テーブルを使用して、前記複数の上りパケット・トランザクションの各々の前記宛先を判定するようにさらに構成されている、
請求項１に記載の入出力ノード・スイッチ。
【請求項１４】
前記パケット・バス・スイッチ・ユニットは、前記第１の外部パケット・バス・リンクに結合された上流のデバイスおよび第２の外部パケット・バス・リンクに結合された上流のデバイスに対応する使用可能なバッファ空間を示す値を使用して、前記複数の上りパケット・トランザクションの各々の前記宛先を判定するようにさらに構成されている、
請求項３に記載の入出力ノード・スイッチ。
【請求項１５】
前記第３のポイントツーポイント・パケット・バス・リンクは、一方向ワイヤの第１の組と一方向ワイヤの第２の組とを有するハイパートランスポート（HyperTransport：登録商標）互換のリンクであり、前記制御パケットと前記データパケットは、同じワイヤを共有する、
請求項１に記載の入出力ノード・スイッチ。
【請求項１６】
マルチプロセッサ・コンピュータ・システムであって、
プログラムされた命令を実行するように各々構成された第１のプロセッサおよび第２のプロセッサと、
集積回路チップに実装されており、第１のポイントツーポイント・パケット・バス・リ

ンクによって前記第１のプロセッサに結合され、かつ第２のポイントツーポイント・パケット・バス・リンクによって前記第２のプロセッサに結合された入出力ノード・スイッチと、を備え、前記入出力ノード・スイッチは、

周辺バスから複数の周辺トランザクションを受信するように結合され、かつ前記複数の周辺トランザクションに対応する複数の上りパケット・トランザクションを送信するように構成されたブリッジ・ユニットと、

前記複数の上りパケット・トランザクションを内部ポイントツーポイント・パケット・バス・リンクで受信するように結合され、かつ前記複数の上りパケット・トランザクションの各々の宛先を判定するように構成されたパケット・バス・スイッチ・ユニットと、を備え、

前記パケット・バス・スイッチ・ユニットは、前記複数の上りパケット・トランザクションの前記宛先の各々を判定すると、前記複数の上りパケット・トランザクションのうちの選択されたパケット・トランザクションを前記第１のポイントツーポイント・パケット・バス・リンクを介して前記第１のプロセッサにルーティングすると共に、前記複数の上りパケット・トランザクションのうちの他のパケット・トランザクションを、前記第２のポイントツーポイント・パケット・バス・リンクを介して前記第２のプロセッサにルーティングするようにさらに構成されている、

前記第１のポイントツーポイント・パケット・バス・リンク及び前記第２のポイントツーポイント・パケット・バス・リンクは、各々一方向ワイヤの第１の組と一方向ワイヤの第２の組とを有するハイパートランスポート（HyperTransport：登録商標）互換のリンクであり、前記一方向ワイヤの第１の組および前記一方向ワイヤの第２の組の各々は、制御パケットとデータ・パケットとを含むパケットを伝送するように構成されており、前記制御パケットは、コマンド・パケット、情報パケットおよび応答パケットを含み、

前記パケット・バス・スイッチ・ユニットは、更に、前記第１ポイントツーポイント・パケット・バス・リンクおよび前記第２ポイントツーポイント・パケット・バス・リンクに結合された上流デバイスのＩ／Ｏ負荷又はトランザクション・タイプに基づいて、前記複数の上りパケット・トランザクションのそれぞれの送信先を決定するように構成されている、

マルチプロセッサ・コンピュータ・システム。

【請求項１７】

前記入出力ノード・スイッチは、前記複数の上りパケット・トランザクションのうちの前記選択されたパケット・トランザクションを受信して、前記複数の上りパケット・トランザクションのうちの前記選択されたパケット・トランザクションを前記第１のポイントツーポイント・パケット・バス・リンクで送信するように結合された第１のトランシーバ・ユニットをさらに備える、

請求項１６に記載のコンピュータ・システム。

【請求項１８】

前記入出力ノード・スイッチは、前記複数の上りパケット・トランザクションのうちの前記他のパケット・トランザクションを受信して、前記複数の上りパケット・トランザクションのうちの前記他のパケット・トランザクションを前記第２のポイントツーポイント・パケット・バス・リンクで送信するように結合された第２のトランシーバ・ユニット（４２０）をさらに備える、

請求項１６に記載のコンピュータ・システム。

【請求項１９】

前記パケット・バス・スイッチ・ユニットは、前記複数の上りパケット・トランザクションの各々に関連付けられたアドレスをデコードするようにさらに構成されている、

請求項１６に記載のコンピュータ・システム。

【請求項２０】

前記パケット・バス・スイッチ・ユニットは、前記アドレスに基づいて、前記複数の上りパケット・トランザクションのうちの追加のパケット・トランザクションをブロックす

るようにさらに構成されている、
請求項１９に記載のコンピュータ・システム。
【請求項２１】
前記入出力ノード・スイッチは、前記アドレスに基づいて、前記複数のパケット・トランザクションのうちの前記追加のパケット・トランザクションを受信して、前記パケット・トランザクションを前記第３のポイントツーポイント・パケット・バス・リンクで他のノードに送信するように結合された第３のトランシーバ・ユニットをさらに備える、
請求項２０に記載のコンピュータ・システム。
【請求項２２】
前記第３のトランシーバ・ユニットは、第２の複数の上りパケット・トランザクションを前記第３のポイントツーポイント・パケット・バス・リンクで受信して、前記第２の複数の上りパケット・トランザクションを前記パケット・バス・スイッチに送信するようにさらに構成されている、
請求項２１に記載のコンピュータ・システム。
【請求項２３】
前記パケット・バス・スイッチ・ユニットは、前記第２の複数の上りパケット・トランザクションの各々に関連付けられたアドレスをデコードし、前記第２の複数の上りパケット・トランザクションのうちの選択されたパケット・トランザクションを前記第１のポイントツーポイント・パケット・バス・リンク（４０１）に結合された前記第１のプロセッサにルーティングすると共に、前記第２の複数の上りパケット・トランザクションの各々に関連付けられた前記アドレスに基づいて、前記複数の上りパケット・トランザクションのうちの他のパケット・トランザクションを前記第２のポイントツーポイント・パケット・バス・リンク（４０２）に結合された前記第２のプロセッサにルーティングするようにさらに構成されている、
請求項２２に記載のコンピュータ・システム。
【請求項２４】
前記パケット・バス・スイッチ・ユニットは、前記第１のプロセッサによって送信された第１の複数の下りパケット・トランザクションと、前記第２のプロセッサによって送信された第２の複数の下りパケット・トランザクションとを受信するようにさらに構成されている、
請求項１８に記載のコンピュータ・システム。
【請求項２５】
前記パケット・バス・スイッチ・ユニットは、前記第１の複数の下りパケット・トランザクションと前記第２の複数の下りパケット・トランザクション間を調停するようにさらに構成されている、
請求項２４に記載のコンピュータ・システム。
【請求項２６】
前記パケット・バス・スイッチ・ユニットは、前記第１の複数の下りパケット・トランザクションと、前記第２の複数の下りパケット・トランザクションとを前記第１の内部ポイントツーポイント・パケット・バス・リンクにルーティングするようにさらに構成されている、
請求項２５に記載のコンピュータ・システム。
【請求項２７】
前記ブリッジ・ユニットは、前記第１複数の下りパケット・トランザクションおよび第２の複数の下りパケット・トランザクションの各々のアドレスをデコードし、前記アドレスに基づいて、前記第１の複数の下りパケット・トランザクションおよび第２の複数の下りパケット・トランザクションの選択されたパケット・トランザクションに対応するバス・サイクルを始めるようにさらに構成されている、
請求項２６に記載のコンピュータ・システム。
【請求項２８】

前記パケット・バス・スイッチ・ユニットは、プログラム可能なルックアップ・テーブルを使用して、前記複数の上りパケット・トランザクションの各々の前記宛先を判定するようにさらに構成されている、
請求項１６に記載のコンピュータ・システム。
【請求項２９】
前記パケット・バス・スイッチ・ユニットは、前記第１プロセッサおよび第２プロセッサの各々に対応する使用可能なバッファ空間を示す値を使用して、前記複数の上りパケット・トランザクションの各々の前記宛先を判定するようにさらに構成されている、
請求項１６に記載のコンピュータ・システム。
【請求項３０】
前記パケット・バス・スイッチ・ユニットは、前記複数の上りパケット・トランザクションの各々を使用可能なバッファ空間を示す値が比較的高い値を有する所定のプロセッサにルーティングするようにさらに構成されている、
請求項１６に記載のコンピュータ・システム。
【請求項３１】
前記第３のポイントツーポイント・パケット・バス・リンクは、一方向ワイヤの第１の組と一方向ワイヤの第２の組とを有するハイパートランスポート（HyperTransport：登録商標）互換のリンクであり、前記一方向ワイヤの第１の組および前記一方向ワイヤの第２の組の各々は、前記制御パケットと前記データパケットは、同じワイヤを共有する、
請求項１６に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】
本発明はマルチプロセッサ・コンピュータ・システムのＩ／Ｏノードに関し、より詳細には、スイッチングＩ／Ｏノードに関する。
【背景技術】
【０００２】
米国特許第６，０８５，２７６Ａ号は、マルチプロセッサ・コンピュータ・システムを開示している。このシステムは、結合されたＩ／Ｏバスを介して外部デバイスとマルチプロセッサ・ノードの間で制御データを転送するために、プロセッサ・モジュール、メモリ・モジュール、およびＩ／Ｏプロセッサ（ＩＯＰ）モジュールを結合しているローカルの９ポートのスイッチを備える。インタフェース・モジュールは、対応する複数の双方向クロック転送データ・リンクによってローカル・スイッチに結合されている。ローカル・スイッチは、ＱＳＡ（クワド・スイッチ・アドレス：Quad Switch Address）制御チップとＱＳＤ（クワド・スイッチ・データ：Quad Switch Data）・スライス・チップとを備える。ＱＳＡチップは、プロセッサ・モジュール、ＩＯＰモジュールおよびメモリ間でアドレス・パスを制御するためのアービタを備える。さらに、ＱＳＡチップは、ローカル・スイッチを介したデータの流れを制御するためにＱＳＤチップに対する制御を付与する。複数のトランザクションを含む転送メッセージに仮想チャネルが使用されており、各トランザクションは複数の別個の仮想チャネルに割り当てられる。仮想チャネルを使用することにより、データの一貫性（コヒーレンシー）が維持される。
処理ユニットを複数使用したコンピュータ・システムは、現行のシングルプロセッサ・ベースのシステムよりも費用面で満足できる高性能を実現すると期待されている。シングル・プロセッサではアプリケーションの全処理が１つのプロセッサに集中するが、マルチプロセッシング環境では、タスクをグループ分けして、複数のプロセッサで処理させることができる。これにより、処理負荷全体が複数のプロセッサに分散され、分散させたタスクの同時並列処理が可能となる。オペレーティング・システムのソフトウェアは、プログラム・コードの各種部分を、別個に実行可能なスレッドに分割して、通常は各スレッドに優先度レベルを割り当てる。
【０００３】

パーソナルコンピュータ（ＰＣ）やその他のコンピュータ・システムは、メモリにアクセスするための共有バス・システムを中心として設計されてきた。共有バスを介して、１つ以上のプロセッサと１つ以上の入出力（Ｉ／Ｏ）デバイスとがメモリに結合され得る。Ｉ／Ｏデバイスは、共有バスとＩ／Ｏデバイス間の情報の転送を管理するＩ／Ｏブリッジを介して共有バスに結合され得る。これに対し、プロセッサは、通常は共有バスに直接結合されるか、あるいはキャッシュ階層を介して共有バスに結合される。代表的なマルチプロセッサ・コンピュータ・システムについて、従来技術を図示した図１を参照して以下に記載する。

【０００４】

図１を参照すると、マルチプロセッサ・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。マルチプロセッサ・コンピュータ・システムは、プロセッサ・ユニット１００Ａ，１００Ｂ、システム・バス１０５を介してプロセッサ・ユニット１００Ａ，１００Ｂに結合されたシステム・コントローラ１１０、およびメモリ・バス１２５を介してシステム・コントローラ１１０に結合されたシステム・メモリ１２０を備える。さらに、システム・コントローラ１１０はＩ／Ｏバス１３５を介してＩ／Ｏハブ１３０に結合されている。

【０００５】

図１のマルチプロセッサ・コンピュータ・システムは、全処理ユニット１００Ａ，１００Ｂが同じ記憶空間（すなわちシステム・メモリ１２０）を共有し、同じアドレス・マッピングを使用して記憶空間にアクセスするという点において対称的であり得る。さらに、このマルチプロセッシング・システムは、全処理ユニット１００Ａ，１００ＢがＩ／Ｏハブ１３０に平等にアクセスできるという点においても対称的であり得る。

【０００６】

一般に、システム・メモリ１２０内には、オペレーティング・システム・ソフトウェアの１つのコピーが、各ユーザー・アプリケーションのファイルの１つのコピーとともに格納され得る。処理ユニット１００Ａ，１００Ｂの各々は、オペレーティング・システム・ファイルおよびユーザー・アプリケーション・ファイルのこれら１つのコピーから、オペレーティング・システムとユーザー・アプリケーションとを実行し得る。処理コア（図示せず）が同時にコードを実行することができるものの、所定のある時点において、システム・バス１０５の制御権(mastership)を取得するのは、処理ユニット１００Ａ，１００Ｂのいずれか一方だけである。このため、処理ユニット１００Ａ，１００Ｂが同時にバス要求を行った場合に、所定の調停アルゴリズムに基づいてこの要求を調停して、処理ユニット１００Ａ，１００Ｂのいずれかに制御権を付与するバス調停メカニズムがシステム・コントローラ１１０に設けられることになる。各種のバス調停技術が公知となっている。

【０００７】

図１のコンピュータ・システムに使用される共有バス（例えばシステム・バス１０５）は、システム・バスの調停に起因し得る制約のほかに、帯域幅が狭いなどの欠点を有することがある。共有バスに追加のプロセッサが接続されると、プロセッサが複数接続されたことにより、バスで信号を駆動しているデバイスに対する容量性負荷が上昇する。また、接続ポイントが複数存在することにより、高い周波数に対応した伝送線のモデルが比較的複雑となる。このため、動作周波数が低下する可能性がある。

【発明の開示】

【発明が解決しようとする課題】

【０００８】

一部のコンピュータ・システムでは、共有バスの欠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。

【０００９】

図２を参照すると、複数本の下りパケット・バス・リンクを１本の上りパケット・バス・リンクに切り替えているマルチプロセッサ・コンピュータ・システムが示される。マルチプロセッサ・コンピュータ・システム２００は、プロセッサ２０１Ａおよびプロセッサ２０１Ｂを備え、これらはシステム・バス２０２によって相互接続されている。プロセッサ２０１Ｂは、パケット・バス・リンク２０５によってＩ／Ｏノード・スイッチ２１０に接続されている。Ｉ／Ｏノード・スイッチ２１０はさらに、別のパケット・バス・リンク２１５を介してＩ／Ｏノード２２０に接続されている。さらに、Ｉ／Ｏノード・スイッチ２１０は、パケット・バス・リンク２２５を介して別のＩ／Ｏノード２３０に接続されている。

【００１０】

プロセッサ２０１Ａ，２０１Ｂについては、図１のプロセッサ１０１Ａ，１０１Ｂとほぼ同じように動作し得る点が留意される。しかし、図２ではＩ／Ｏの接続が異なっている。Ｉ／Ｏノード・スイッチ２１０は、プロセッサ２０１Ａまたは２０１ＢからＩ／Ｏノード２２０，２３０のいずれへ通信を行うためのスイッチング・メカニズムを提供し得る。このタイプのシステムでは、プロセッサ２０１Ｂは、Ｉ／Ｏノード２２０，２３０と通信するためのホスト・ブリッジ（図示なし）を備え得る。さらに、プロセッサ２０１Ａは、プロセッサ２０１Ｂを介してＩ／Ｏノード２２０，２３０と通信し得る。このように結合されたシステムは、図２のパケット・バスを使用することによって、図１に示されるマルチプロセッサ・システムよりも優れたマルチプロセッシング・ソリューションとなり得るが、それでも欠点が解消されない可能性がある。例えば、プロセッサ２０１Ａが生成したトランザクション、またはプロセッサ２０１Ａ宛のトランザクションは、最初にプロセッサ２０１Ｂを通過するため、レイテンシ・ペナルティ(latency penalties)を発生させる可能性がある。

【課題を解決するための手段】

【００１１】

マルチプロセッサ・コンピュータ・システムにおける接続のためのスイッチングＩ／Ｏノードの各種実施形態が開示される。一実施形態においては、マルチプロセッサ・コンピュータ・システム用の入出力ノード・スイッチは、集積回路チップに実装されたブリッジ・ユニットを備える。このブリッジ・ユニットは、ＰＣＩバスなどの周辺バスから、複数の周辺トランザクションを受信するように結合され得、この複数の周辺トランザクションに対応する複数の上りパケット・トランザクションを送信するように構成され得る。入出力ノード・スイッチは、集積回路チップに実装されたパケット・バス・スイッチ・ユニットをも備える。このパケット・バス・スイッチ・ユニットは、複数の上りパケット・トランザクションを内部ポイントツーポイント・パケット・バス・リンクで受信するように結合され得、この複数の上りパケット・トランザクションの各々の宛先を判定するように構成され得る。パケット・バス・スイッチ・ユニットは、複数の上りパケット・トランザクションの宛先を判定すると、複数の上りパケット・トランザクションのうちの選択されたパケット・トランザクションを第１のポイントツーポイント・パケット・バス・リンクに結合された第１のプロセッサ・インタフェースにルーティングすると共に、複数の上りパケット・トランザクションのうちの他のパケット・トランザクションを第２のポイントツーポイント・パケット・バス・リンクに結合された第２のプロセッサ・インタフェースにルーティングするようにさらに構成され得る。

【００１２】

特定の一実装において、入出力ノード・スイッチは、集積回路チップに実装された第１のトランシーバ・ユニットおよび第２のトランシーバ・ユニットをさらに有する。第１のトランシーバ・ユニットは、複数の上りパケット・トランザクションのうちの選択されたパケット・トランザクションを受信して、この選択されたパケット・トランザクションを第１のポイントツーポイント・パケット・バス・リンクで送信するように結合され得る。第２のトランシーバ・ユニットは、複数の上りパケット・トランザクションのうちの他の選択されたパケット・トランザクションを受信して、この他の選択されたパケット・トランザクションを第２のポイントツーポイント・パケット・バス・リンクで送信するように結合され得る。ポイントツーポイント・パケット・バス・リンクの各々は、ハイパートランスポート(HyperTransport)バス・リンクであり得る。

【００１３】

特定の一実装において、パケット・バス・スイッチ・ユニットは、プログラム可能なルックアップ・テーブルを使用して複数の上りパケット・トランザクションの各々の宛先を判定するように構成され得る。

【００１４】

他の特定の実装において、パケット・バス・スイッチ・ユニットは、プロセッサなどの、第１の外部パケット・バス・リンク上流のデバイスおよび第２の外部パケット・バス・リンクに結合された上流のデバイスに対応する使用可能なバッファ空間を示す値を使用して、複数の上りパケット・トランザクションの宛先を判定するようにさらに構成され得る。

【００１５】

さらに別の特定の実装において、パケット・バス・スイッチ・ユニットは、前記複数の上りパケット・トランザクションの各々に関連付けられたアドレスをデコードするように構成され得る。さらに別の特定の実装において、パケット・バス・スイッチ・ユニットは、このアドレスに基づいて、複数の上りパケット・トランザクションのうちの追加のトランザクションをブロックするように構成され得る。

【発明を実施するための最良の形態】

【００１６】

本発明は、様々に変形されたり代替形態を取り得るが、その特定の実装形態が、例として図中に図示され、かつ本明細書に詳細に記載される。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を意図したものであることが理解されるべきである。

【００１７】

図３を参照すると、複数本の上りパケット・バス・リンクを有するマルチプロセッサ・コンピュータ・システムのブロック図が示される。マルチプロセッサ・コンピュータ・システム３００は、プロセッサ３０１Ａ～３０１Ｄを備え、プロセッサ３０１Ａ～３０１Ｄは、それぞれ一対のＩ／Ｏパケット・バス・リンク３１５Ａ，３１５Ｂを介してＩ／Ｏノード・スイッチ・デバイス３１０に結合されている。プロセッサ３０１Ａ～Ｄは、コヒーレント・パケット・バス・リンク３０５Ａ～Ｄを介して結合されている。Ｉ／Ｏノード・スイッチ３１０が、周辺バス３４０とＩ／Ｏリンク３４５とに接続されていることが示されている。Ｉ／Ｏノード・スイッチ３１０はさらに、Ｉ／Ｏパケット・バス・リンク３２５を介してＩ／Ｏノード３２０に結合されている。

Ｉ／Ｏノード３２０は別のＩ／Ｏパケット・バス・リンク３３０に結合されており、このＩ／Ｏパケット・バス・リンク３３０は、他のＩ／Ｏノード（図示なし）に接続されることがある。なお、マルチプロセッサ・コンピュータ・システム３００の各要素の構成が異なる実施形態もとり得る。例えば、別の実施形態では、プロセッサの数が異なっており、追加のプロセッサが追加のＩ／Ｏパケット・バス・リンク（図示なし）を介してＩ／Ｏノード・スイッチ３１０に結合される構成をとっても良い。さらに、Ｉ／Ｏリンク３４５が、マルチプロセッサ・コンピュータ・システム３００と同じネットワーク内または異な

るネットワーク内でデバイスに接続されていてもよい。

【００１８】

図中の実施形態においては、コヒーレント・パケット・バス・リンク３０５の各々は、一方向のラインの組として実施される（例えば、ライン３０５Ｂはプロセッサ３０１Ａからプロセッサ３０１Ｂにパケットを送信するために使用され、ライン３０５Ｃはプロセッサ３０１Ｂからプロセッサ３０１Ｃにパケットを送信するために使用される）。図３に示すように、これ以外のラインの組３０５Ａ，Ｄは、他のプロセッサ間でのパケットの送信に使用される。コヒーレント・パケット・インタフェース３０５は、処理ノード間の通信について、キャッシュ・コヒーレントな方法で動作し得る（「コヒーレント・リンク」）。

さらに、Ｉ／Ｏパケット・バス３１５は、Ｉ／Ｏノード間、およびＩ／Ｏノードとプロセッサ間（プロセッサ３０１Ａなど）の通信について、非コヒーレントな方法で動作し得る（「非コヒーレントなリンク」）。また、この非コヒーレント・リンクは、一方向のラインの組として実装されてもよい（例えば、ライン３１５Ａはプロセッサ３０１ＡからＩ／Ｏノード・スイッチ３１０にパケットを送信するために使用され、ライン３１５Ｂはプロセッサ３０１ＢからＩ／Ｏノード・スイッチ３１０にパケットを送信するために使用される）。コヒーレント・リンクを介した２つ以上のノード間の相互接続は、「コヒーレント・ファブリック」と呼ばれることがある。

同様に、非コヒーレント・リンクを介した２つ以上のノード間の相互接続は、「非コヒーレント・ファブリック」と呼ばれることがある。１つのプロセッサから別のプロセッサに送信されるパケットが、１つ以上の中間ノードを通過することがあり得る点が留意される。例えば、プロセッサ３０１Ａからプロセッサ３０１Ｃに送信されるパケットは、図３に示すようにプロセッサ３０１Ｂまたはプロセッサ３０１Ｄを通過することがあり得る。適したルーティング・アルゴリズムであれば、どのようなルーティング・アルゴリズムでも使用され得る。プロセッサ３０１Ｃ，Ｄを囲む破線によって示されるように、図３の実施形態とはプロセッサの数が異なるマルチプロセッサ・コンピュータ・システム３００の実施形態も可能である。

【００１９】

ここに示す実施形態においては、プロセッサ３０１Ａ～Ｄの各々の例として、アスロン[Athlon（登録商標）]マイクロプロセッサなどのｘ８６アーキテクチャ・プロセッサがある。しかし、適したプロセッサであれば、別のプロセッサを使用してもよい。さらに、Ｉ／Ｏパケット・バス・リンク３１５Ａ，Ｂおよび３２５は、高速ポイントツーポイント・パケット・インタフェースの代表的なリンクであって、ハイパートランスポート[HyperTransport（登録商標）]技術と互換性を有する。さらに、Ｉ／Ｏリンク３４５は、イーサネット(Ethernet)接続またはインフィニバンド[Infiniband（登録商標）]接続などの代表的な接続である。周辺バス３４０は、例えば、ＰＣＩ（Peripheral Component Interconnect：周辺機器相互接続）またはＰＣＩ－Ｘ（Extended Peripheral Component Interconnect：拡張周辺機器相互接続）バスなどの、適切な任意の周辺バスの一例である。

【００２０】

マルチプロセッサ・コンピュータ・システム３００などのマルチプロセッサ・コンピュータ・システムにおいて、システム・メモリ内には、オペレーティング・システム・ソフトウェアの１つのコピーと、各ユーザー・アプリケーションのファイルの１つのコピーとが格納され得る。処理ユニット３０１Ａ～３０１Ｂの各々は、オペレーティング・システム・ファイルおよびユーザー・アプリケーション・ファイルのこれら１つのコピーから、オペレーティング・システムとユーザー・アプリケーションとを実行し得る。

【００２１】

図中の実施形態においては、プロセッサ３０１Ａ，３０１Ｂの各々は、それぞれＩ／Ｏパケット・バス・リンク３０５Ａ，３０５Ｂに対するインタフェースを有するホスト・ブリッジ（図示なし）を備え得る。他の実施形態においては、プロセッサ３０１Ｃ，３０１Ｄもホスト・ブリッジを備えており、追加のＩ／Ｏパケット・バス・リンクに接続されて

いてもよいことが考察される。図４および図５の説明で詳細に後述するように、Ｉ／Ｏノード・スイッチ３１０は下流のソース・ノードから複数のパケット・トランザクションを受信して、プロセッサ３０１Ａ，Ｂなど、複数の上流の宛先にこのトランザクションをルーティングし得る。本明細書中で使用されるように、「上り（上流）」との用語は、プロセッサ３０１Ａなどのプロセッサ、またはプロセッサ内のホスト・ブリッジに向かう方向に流れるトランザクションを指す。「下り（下流）」との用語は、プロセッサまたはプロセッサ内のホスト・ブリッジから離れる方向に流れるトランザクションを指す。

【００２２】

一般に、パケットとは、２つのノード間（パケットを送信する開始ノード、およびこのパケットを受信する宛先ノード）で行われる通信である。開始ノードおよび宛先ノードは、パケットがその一部であるトランザクションのソース・ノードおよびターゲット・ノードと一致しないこともあれば、いずれがソース・ノードまたはターゲット・ノードであることもある。制御パケットとは、トランザクションに関する制御情報を伝達するパケットのことである。一部の制御パケットは、後にデータ・パケットが続くことを指定している。データ・パケットは、トランザクションに対応するデータと、指定を行っているこの制御パケットに対応するデータとを伝達する。一実施形態においては、制御パケットは、コマンド・パケット、情報パケットおよび応答パケットを含んでいてもよい。別の実施形態においては、これ以外の種類のパケットを含むことがあることが考察される点が留意される。

【００２３】

Ｉ／Ｏノード・スイッチ３１０は、例えば周辺バス３４０、ネットワーク・リンク３４５およびＩ／Ｏパケット・バス３２５など、複数の下流のソース・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。通常、各パケット・トランザクションは、アドレスがエンコードされているヘッダを有し得る。Ｉ／Ｏノード・スイッチ３１０は、各パケット・トランザクションの宛先アドレスをデコードして、デコードされたアドレスに基づいてトランザクションをルーティングし得る。また、Ｉ／Ｏノード・スイッチ３１０は、プロセッサ３０１Ａ～Ｄが発信した下りパケット・トランザクションも受信し得る。Ｉ／Ｏノード・スイッチ３１０は、各パケット・トランザクションの宛先アドレスをデコードし、そのアドレスに従ってパケット・トランザクションをルーティングし得る。

【００２４】

図４に、Ｉ／Ｏノード・スイッチの一実施形態のブロック図を示す。Ｉ／Ｏノード・スイッチ４００は、Ｉ／Ｏパケット・バス４０１，４０２にそれぞれ結合された１対の上流のトランシーバ４１０，４２０を有する。Ｉ／Ｏパケット・バス・リンク４０１，４０２は、例えば、図１のプロセッサ３０１Ａ，Ｂなどの上流のデバイスに結合され得る。また、図４のトランシーバ４１０，４２０は、それぞれ内部パケット・バス・リンク４１５，４２５を介してパケット・バス・スイッチ４３０に結合されている。

パケット・バス・スイッチ４３０は、内部パケット・バス・リンク４３５によってインタフェース・バス４４０に結合されている。インタフェース・バス４４０は、内部パケット・バス・リンク４４５を介して下流のトランシーバ４８０に結合されている。トランシーバ４８０はＩ／Ｏパケット・バス・リンク４８５に結合されており、Ｉ／Ｏパケット・バス・リンク４８５は他のＩ／Ｏノードまたは他のデバイス（図示なし）に接続され得る。また、インタフェース・バス４４０は、Ｉ／Ｏインタフェース４７０に結合されると共に、周辺インタフェース４５０，４６０に結合されている。

周辺インタフェース４５０，４６０は周辺バス４５５，４６５にそれぞれ結合されている。Ｉ／Ｏインタフェース４７０はＩ／Ｏリンク４７５に結合されている。本実施形態では２つの上流のトランシーバが２つのＩ／Ｏパケット・バス・リンクに接続されていることが図示されているが、別の実施形態では、上流のトランシーバおよびＩ／Ｏパケット・バス・リンクの数が他の適切な数であってもよいこととしてもよい。

【００２５】

Ｉ／Ｏインタフェース４７０は、集積化されたＩ／Ｏコントローラであり得、例えば、

ギガビットイーサネット(Gigabit Ethernet)コントローラまたはインフィニバンドポート・コントローラなど、特定のＩ／Ｏデバイスを実装する回路を備え得る点が留意される。このような実施形態では、Ｉ／Ｏリンク４７５は、イーサネット接続またはインフィニバンド接続などの代表的な接続であり得る。周辺バス４５５，４６５の例として、例えば、ＰＣＩ（Peripheral Component Interconnect：周辺機器相互接続）またはＰＣＩ－Ｘ（Extended Peripheral Component Interconnect：拡張周辺機器相互接続）バスなどの、適した任意の周辺バスがある。

【００２６】

トランシーバ４１０，４２０は、それぞれＩ／Ｏパケット・バス・リンク４０１，４０２で下りパケット・トランザクションを受信すると共に、上りパケット・トランザクションを送信するように構成され得る。トランシーバ４１０，４２０は、処理待ちのパケット・トランザクションを記憶するための送受信バッファ回路（図示なし）を備え得る。また、トランシーバ４１０，４２０は、Ｉ／Ｏパケット・バス・リンク４０１，４０２でパケット・トランザクションを送信するためのＩ／Ｏドライバ回路（図示なし）も備え得る。

【００２７】

パケット・スイッチ・ユニット４３０は、内部パケット・バス・リンク４３５で上りパケット・トランザクションを受信し、各トランザクションのアドレスをデコードして、各トランザクションをルーティングするための上りパスが存在すれば、そのパスを決定するように構成され得る。下記に詳細に記載するように、あるトランザクションが、パケット・バス・スイッチ４３０の上り内部パケット・バス・リンクのいずれにも関連付けられていないアドレスを含んでいる場合、そのパケット・トランザクションは下り方向に戻される。別法として、アドレスがパケット・バス・スイッチ４３０の上り内部パケット・バス・リンクに対応しているかどうかに関わらず、パケット・トランザクションが上り方向に送信されてもよい。

【００２８】

インタフェース・バス４４０は、内部パケット・バス・リンク４３５，４４５の各々、周辺インタフェース４５５，４６５、およびＩ／Ｏインタフェース４７０を相互接続するのに必要な内部パケット・バス・アーキテクチャと回路（図示なし）とを備え得る。例えば、インタフェース・バス４４０は、内部パケット・バス・リンク４３５、Ｉ／Ｏインタフェース４７０、周辺インタフェース４５０，４６０用のインタフェース・ポート（図示なし）と、内部パケット・バス・リンク４４５用のインタフェース・ポート（図示なし）とを備え得る。

各インタフェース・ポートは、特定のフィルタと一致するアドレスを有するパケット・トランザクションを要求する機能を備えたアドレス・フィルタを備え得る。このように、インタフェース・バス４４０は、周辺インタフェースの各々、Ｉ／Ｏインタフェース、および内部パケット・バス・リンク４４５の間でのピアツーピアのトラフィックを処理できるように構成され得る。

周辺インタフェース４６０宛のパケット・トランザクションが、周辺インタフェース４５０によって、インタフェース・バス４４０へ上り方向に送信される例を考える。インタフェース・バス４４０に接続されている各インタフェース・ポートによって、このパケット・トランザクションのヘッダがデコードされ得る。デコードされたアドレスは、周辺インタフェース４６０のアドレス・フィルタのみに一致し、従って、周辺インタフェース４６０によってアドレスが要求される。

【００２９】

あるインタフェースによって上り方向に送信されたパケット・トランザクションが、インタフェース・ポートのどのアドレス・フィルタとも一致しないことがある。そのような場合、パケット・バス・スイッチ４３０は、要求されなかったパケット・トランザクションを、トランザクションの発信元へ下り方向に戻し得る。さらに、存在しないアドレスがトランザクションに含まれていたことを示すエラー・メッセージが、トランザクションに付され得る。

【００３０】
別の実施形態においては、内部パケット・バス・リンク４３５用のインタフェース・ポートがアドレス・フィルタを有さないこともある。この場合、パケット・バス・スイッチ４３０は、要求されなかった全パケット・トランザクションを上りパケット・バス・リンク４１５，４２５のいずれかを介して上り方向に送信し得る。
【００３１】
パケット・バス・スイッチ４３０は、特定のトランザクションをどの上りパケット・バス・リンクに送信すべきかを判定し得る。この判定は、幾つかの要因によって決まる。一例を挙げると、上りトランザクションが、図１のプロセッサ３０１Ａ～Ｄなどのデバイスからの前に行われた下り要求に対する応答である場合、応答は、要求側のプロセッサに対応したアドレスを有し得る。このため、パケット・バス・スイッチ４３０は、上りトランザクションを正しいプロセッサにルーティングできる。
【００３２】
別の例では、上りトランザクションが、メモリ読込みを実行しているＰＣＩ－ＸデバイスなどのＩ／Ｏマスタからのものである場合、パケット・バス・スイッチ４３０は、この要求をどのプロセッサにルーティングすべきかを判定し得る。このため、パケット・バス・スイッチ４３０は、プロセッサが処理するトランザクションの種類を示すルックアップ・テーブルまたは類似するデータ構造を格納することができるプログラム可能な記憶回路（図示なし）を備え得る。この場合、オペレーティング・システムはこの種のテーブルの内容を判定して、パケット・バス・スイッチ４３０をプログラムする命令を、１つ以上のプロセッサに実行させ得る。したがって、Ｉ／Ｏマスタ・トランザクションがパケット・バス・スイッチ４３０によって受信されると、このテーブルがパケット・バス・スイッチ４３０によって使用され、パケット・トランザクションをどのプロセッサにルーティングすべきかが判定され得る。
【００３３】
ルックアップ・テーブルの情報が破損したか、その内容が最新ではなくなった場合、オペレーティング・システムによって上りパケット・トランザクションが別のプロセッサに間違って送信されることがある。それでも、誤ってルーティングされたパケット・トランザクションは、最終的に目的の宛先に到着することができる。ただし、このトランザクションが最初に別のプロセッサを経由してルーティングされ得るので、遅延が発生するおそれがある。誤ルーティングの再発を防止するため、トランザクションを受信したプロセッサは、エラー・メッセージを発生させることができる。このエラー・メッセージは、ルックアップ・テーブルを更新する必要があることをオペレーティング・システムに通知するものであり得る。次に、オペレーティング・システムは、自身の優先順序付け方式に従ってルックアップ・テーブルの更新をスケジューリングし得る。
【００３４】
他の形態においては、パケット・バス・スイッチ４３０は、各プロセッサのＩ／Ｏ負荷を判定して、Ｉ／Ｏ負荷が最も低いと判定されたプロセッサにトランザクションをルーティングするように構成されてもよい。パケット・バス・スイッチ４３０は、特定のプロセッサのホスト・ブリッジの受信バッファ内の記憶空間の使用可能度（availability）を確認するか、あるいは他の受信ロジックによってプロセッサのＩ／Ｏ負荷を判定し得る。続いて、パケット・バス・スイッチ４３０は、利用可能な受信バッファ空間が最も多いプロセッサにパケット・トランザクションをルーティングし得る。
【００３５】
また、パケット・バス・スイッチ４３０は、内部パケット・バス・リンク４１５，４２５の両方から下りパケット・トランザクションを受信するようにも構成され得る。内部パケット・バス・リンク４１５，４２５から同時にトランザクションが受信されることがある。さらに、パケット・バス・スイッチ４３０が、下りのイベントを待機している処理待ちのトランザクションを有していることがあり得る。このため、パケット・バス・スイッチ４３０は、例えばトランザクション間を調停するために、下りの内部パケット・バス・

リンク・インタフェースにおいて調停回路（図示なし）を備え得、この調停回路はラウンド・ロビン方式などの周知の調停技術を１つ以上使用し得る。別の実施形態においては、この調停回路は、優先順位に基づいたものであってもよく、宛先の優先順位は、調停における主要な属性の１つである。

【００３６】

図５を参照すると、Ｉ／Ｏノード・スイッチの別の実施形態のブロック図が示される。Ｉ／Ｏノード・スイッチ５００は、Ｉ／Ｏパケット・バス５０１，５０２にそれぞれ結合された１対の上流のトランシーバ５１０，５２０を有する。Ｉ／Ｏパケット・バス・リンク５０１，５０２は、例えば、図１のプロセッサ３０１Ａ，Ｂなどの上流のデバイスに結合され得る。また、図５のトランシーバ５１０，５２０は、それぞれ内部パケット・バス・リンク５１５，５２５を介してパケット・バス・スイッチ５３０に結合されている。

パケット・バス・スイッチ５３０は、内部パケット・バス・リンク５３５によってインタフェース・バス５４０に結合されている。パケット・バス・スイッチ５３０は、内部パケット・バス・リンク５４５を介して下流のトランシーバ５８０に結合されている。トランシーバ５８０はＩ／Ｏパケット・バス・リンク５８５に結合されており、Ｉ／Ｏパケット・バス・リンク４８５は他のＩ／Ｏノードまたは他のデバイス（図示なし）に接続され得る。インタフェース・バス５４０は、Ｉ／Ｏインタフェース５７０に結合されると共に、周辺インタフェース５５０，５６０に結合されている。

周辺インタフェース５５０，５６０は周辺バス５５５，５６５にそれぞれ結合されている。Ｉ／Ｏインタフェース５７０はＩ／Ｏリンク５７５に結合されている。本実施形態では２つの上流のトランシーバが２つのＩ／Ｏパケット・バス・リンクに接続されているように図示されているが、別の実施形態では、上流のトランシーバおよびＩ／Ｏパケット・バス・リンクの数がこれ以外の適切な数であってもよいことが考察される。

【００３７】

トランシーバ５１０，５２０は、図４の実施形態と似た機能を備え、ほぼ同じように動作する。このため、トランシーバ５１０，５２０の動作については、図４に関して上記したトランシーバ４１０，４２０の説明を参照されたい。

【００３８】

Ｉ／Ｏインタフェース５７０は、集積化されたＩ／Ｏコントローラであり得、例えば、ギガビットイーサネットコントローラまたはインフィニバンドポート・コントローラなどの特定のＩ／Ｏデバイスを実装する回路を備え得るという点が留意される。

このような実施形態では、Ｉ／Ｏリンク５７５は、イーサネット接続またはインフィニバンド接続などの代表的な接続であり得る。周辺バス５５５，５６５の例として、例えば、ＰＣＩ（Peripheral Component Interconnect：周辺機器相互接続）またはＰＣＩ－Ｘ（Extended Peripheral Component Interconnect：拡張周辺機器相互接続）バスなどの、適した任意の周辺バスがある。

【００３９】

パケット・スイッチ・ユニット５３０は、内部パケット・バス・リンク５３５で上りパケット・トランザクションを受信し、各トランザクションのアドレスをデコードして、各トランザクションをルーティングするための上りパスが存在すれば、そのパスを決定するように構成され得る。下記に詳細に記載するように、あるトランザクションが、パケット・バス・スイッチ５３０の上り内部パケット・バス・リンクのいずれにも関連付けられていないアドレスを含んでいる場合、そのパケット・トランザクションは下り方向に送信され得る。

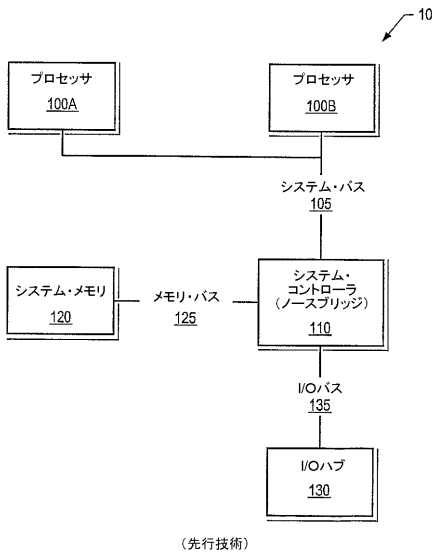
【００４０】

インタフェース・バス５４０は、内部パケット・バス・リンク５３５，５４５の各々、周辺インタフェース５５５，５６５、およびＩ／Ｏインタフェース５７０を相互接続するのに必要な内部パケット・バス・アーキテクチャと回路（図示なし）とを備え得る。例えば、インタフェース・バス５４０は、内部パケット・バス・リンク５３５、Ｉ／Ｏインタフェース５７０、および周辺インタフェース５５０，５６０用のインタフェース・ポート

（図示なし）を備え得る。
各インタフェース・ポートは、特定のフィルタと一致するアドレスを有するパケット・トランザクションを要求する機能を備えたアドレス・フィルタを備え得る。このように、インタフェース・バス５４０は、周辺インタフェースの各々、およびＩ／Ｏインタフェースの間でのピアツーピアのトラフィックを処理できるように構成され得る。周辺インタフェース５６０宛のパケット・トランザクションが、周辺インタフェース５５０によって、インタフェース・バス５４０へ上り方向に送信される例を考える。インタフェース・バス５４０に接続されている各インタフェース・ポートによって、このパケット・トランザクションのヘッダがデコードされ得る。
デコードされたアドレスは、周辺インタフェース５６０のアドレス・フィルタに一致し、このため周辺インタフェース５６０によって要求される。しかし、下記に詳述するように、内部パケット・バス・リンク５４５を介してトランシーバ５８０に接続されたＩ／Ｏノードまたは他のデバイス宛てのパケット・トランザクションは、最初に上り方向にルーティングされ得る。
【００４１】
あるインタフェースによって上り方向に送信されたパケット・トランザクションが、インタフェース・ポートのどのアドレス・フィルタとも一致しないことがある。この場合、パケット・バス・スイッチ５３０は、要求されなかった全パケット・トランザクションを上りパケット・バス・リンク５１５，５２５のいずれかを介して上り方向に送信し得る。
このように、内部パケット・バス・リンク５４５を介してトランシーバ５８０に接続されたＩ／Ｏノードまたは他のデバイス宛てのパケット・トランザクションは、パケット・バス・スイッチ５３０によって、最初に図３のプロセッサ３０１Ａ，Ｂのいずれかに関連付けられたホスト・ブリッジにルーティングされ得る。次に、ホスト・ブリッジはこのパケットを図５のパケット・バス・スイッチ５３０に宛てて下り方向に送信し、パケット・バス・スイッチ５３０は、このパケット・トランザクションを、トランシーバ５８０に接続されたＩ／Ｏノードまたは他のデバイスにルーティングし得る。
【００４２】
別の実施形態においては、パケット・バス・スイッチ５３０が、要求されなかったパケット・トランザクションの宛先アドレスが、内部パケット・バス・リンク５４５を介してトランシーバ５８０に接続されたＩ／Ｏノードまたは他のデバイスに関連付けられたアドレスであると認識してもよい。このため、パケット・バス・スイッチがこのようなピアツーピアのトラフィックを処理できるように構成されてもよい。
【００４３】
パケット・バス・スイッチ５３０は、図４のパケット・バス・スイッチ４３０とほぼ同じように、特定のトランザクションをどの上りパケット・バス・リンクに送信すべきかを判定し得る。このため、図５のパケット・バス・スイッチ５３０のこれ以外の動作については、図４のパケット・バス・スイッチ４３０の説明を参照されたい。
【００４４】
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】
【００４５】
本発明は、一般にマルチプロセッサ・コンピュータ・システムのＩ／Ｏノードに利用することができる。
【図面の簡単な説明】
【００４６】
【図１】共有バスを使用するマルチプロセッサ・コンピュータ・システムの一実施形態のブロック図である。
【図２】複数本の下りパケット・バス・リンクを１本の上りパケット・バス・リンクに切

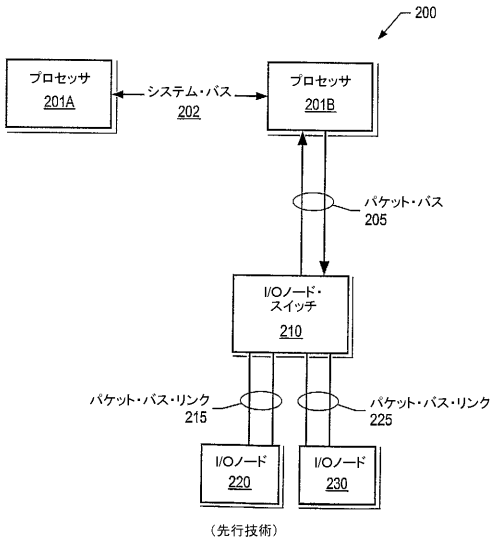
り替えているマルチプロセッサ・コンピュータ・システムの一実施形態のブロック図である。
【図３】複数本の上りパケット・バス・リンクを有するマルチプロセッサ・コンピュータ・システムの一実施形態のブロック図である。
【図４】Ｉ／Ｏノード・スイッチの一実施形態を示すブロック図である。
【図５】Ｉ／Ｏノード・スイッチの別の実施形態を示すブロック図である。

【図１】



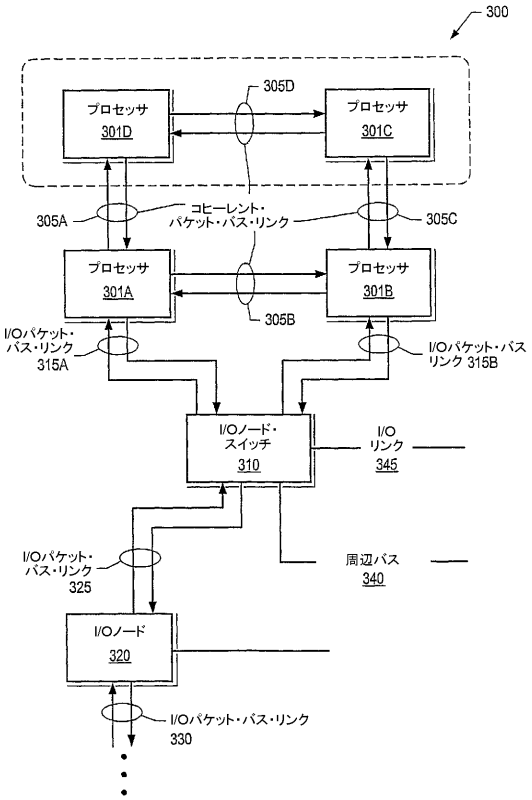
（先行技術）

【図２】

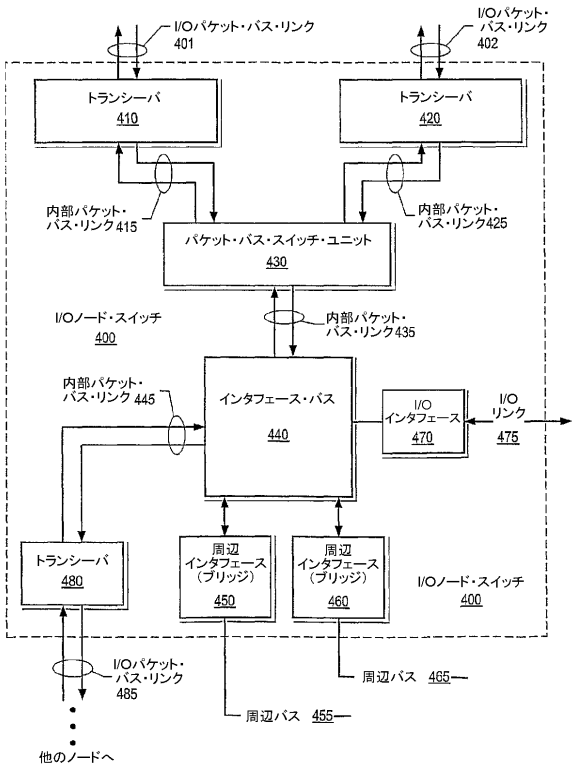


（先行技術）

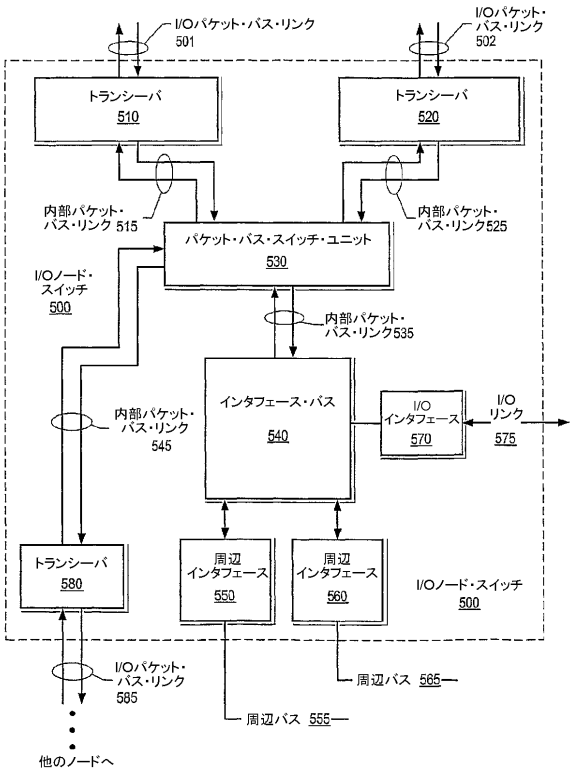
【図３】


300
プロセッサ 301D
プロセッサ 301C
305D
305A
コヒーレント・パケット・パス・リンク
305C
プロセッサ 301A
プロセッサ 301B
305B
I/Oパケット・パス・リンク 315A
I/Oパケット・パス・リンク 315B
I/Oノード・スイッチ 310
I/O リンク 345
I/Oパケット・パス・リンク 325
周辺バス 340
I/Oノード 320
I/Oパケット・パス・リンク 330

【図４】


I/Oパケット・パス・リンク 401
I/Oパケット・パス・リンク 402
トランシーバ 410
トランシーバ 420
内部パケット・パス・リンク415
内部パケット・パス・リンク425
パケット・パス・スイッチ・ユニット 430
I/Oノード・スイッチ 400
内部パケット・パス・リンク435
内部パケット・パス・リンク 445
インタフェース・バス 440
I/O インタフェース 470
I/O リンク 475
トランシーバ 480
周辺インタフェース（ブリッジ） 450
周辺インタフェース（ブリッジ） 460
I/Oノード・スイッチ 400
I/Oパケット・パス・リンク 485
周辺バス 465
周辺バス 455
他のノードへ

【図５】


I/Oパケット・パス・リンク 501
I/Oパケット・パス・リンク 502
トランシーバ 510
トランシーバ 520
内部パケット・パス・リンク 515
内部パケット・パス・リンク525
パケット・パス・スイッチ・ユニット 530
I/Oノード・スイッチ 500
内部パケット・パス・リンク535
インタフェース・バス 540
I/O インタフェース 570
I/O リンク 575
内部パケット・パス・リンク 545
トランシーバ 580
周辺インタフェース 550
周辺インタフェース 560
I/Oノード・スイッチ 500
I/Oパケット・パス・リンク 585
周辺バス 565
周辺バス 555
他のノードへ

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 デイル イー. ガリック
アメリカ合衆国、テキサス州 ７８７３８、オースティン、アストリア １１７１５

審査官 西出 隆二

(56)参考文献 特開２００１－３２５２０７（ＪＰ，Ａ）
特開平０９－１３８７８２（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
G06F 13/36
G06F 13/14
H04L 12/56