(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5084197号
(P5084197)
(45)発行日　平成24年11月28日(2012.11.28)　(24)登録日　平成24年9月14日(2012.9.14)

(51)Int.Cl.　F I
G06F 15/173 (2006.01)　G06F 15/173 640M

請求項の数 12　(全 22 頁)

(21)出願番号　特願2006-217953(P2006-217953)
(22)出願日　平成18年8月10日(2006.8.10)
(65)公開番号　特開2008-41027(P2008-41027A)
(43)公開日　平成20年2月21日(2008.2.21)
審査請求日　平成21年7月31日(2009.7.31)

(73)特許権者　310021766
株式会社ソニー・コンピュータエンタテインメント
東京都港区港南１丁目７番１号
(74)代理人　100105924
弁理士　森下　賢樹
(74)代理人　100109047
弁理士　村田　雄祐
(74)代理人　100109081
弁理士　三木　友由
(74)代理人　100134256
弁理士　青木　武司

最終頁に続く

(54)【発明の名称】プロセッサノードシステムおよびプロセッサノードクラスタシステム

(57)【特許請求の範囲】
【請求項１】
プロセッサと、前記プロセッサの入出力バスと周辺デバイスが接続されるＰＣＩエクスプレスとの間でデータを中継するブリッジとが搭載されたプロセッサ基板を複数含み、
前記ブリッジのポートは当該プロセッサがホストとなるルートコンプレックスモードまたは当該プロセッサが周辺デバイスとなるエンドポイントモードに設定可能に構成され、
一のプロセッサ基板のブリッジのルートコンプレックスモードに設定されたポートを、別のプロセッサ基板のブリッジのエンドポイントモードに設定されたポートに接続することにより、前記複数のプロセッサ基板間を相互結合してなることを特徴とするプロセッサノードシステム。
【請求項２】
前記一のプロセッサ基板のブリッジのルートコンプレックスモードに設定されたポートに設けられるＰＣＩエクスプレスコネクタと、前記別のプロセッサ基板のブリッジのエンドポイントモードに設定されたポートに設けられるＰＣＩエクスプレスコネクタとがフレキシブル基板により配線接続されてなることを特徴とする請求項１に記載のプロセッサノードシステム。
【請求項３】
前記一のプロセッサ基板のブリッジのルートコンプレックスモードに設定されたポートに設けられるＰＣＩエクスプレスコネクタと、前記別のプロセッサ基板のブリッジのエンドポイントモードに設定されたポートに設けられるＰＣＩエクスプレスコネクタとを相互

接続するための一枚のバックプレーン基板をさらに設けたことを特徴とする請求項１に記載のプロセッサノードシステム。

【請求項４】

プロセッサと、前記プロセッサの入出力バスと周辺デバイスが接続されるＰＣＩエクスプレスとの間でデータを中継するブリッジとが２組搭載されたプロセッサ基板を４枚含み、

各ブリッジは、当該プロセッサがホストとなるルートコンプレックスモードに設定されたポートと当該プロセッサが周辺デバイスとなるエンドポイントモードに設定されたポートを有し、

一のプロセッサ基板のルートコンプレックスモードに設定されたポートは、別のプロセッサ基板のブリッジのエンドポイントモードに設定されたポートに接続されることを条件として、各プロセッサ基板の合計４個のポートの内、３個のポートを用いて、前記４枚のプロセッサ基板の内、任意の２枚のプロセッサ基板間を相互結合してなることを特徴とするプロセッサノードシステム。

【請求項５】

各プロセッサ基板の前記３個のポートを第１ポート、第２ポート、第３ポートとし、当該プロセッサ基板以外の３枚のプロセッサ基板を、第１プロセッサ基板、第２プロセッサ基板、第３プロセッサ基板とした場合、前記第１ポートは前記第１プロセッサ基板に接続され、前記第２ポートは前記第２プロセッサ基板に接続され、前記第３ポートは前記第３プロセッサ基板に接続されることを特徴とする請求項４に記載のプロセッサノードシステム。

【請求項６】

前記プロセッサ基板内の２つのプロセッサは入出力バスを介して接続されており、前記プロセッサ基板間のポートの接続により、４枚の前記プロセッサ基板の合計８個のプロセッサの内、任意の２個のプロセッサが互いに通信可能に相互接続されることを特徴とする請求項５に記載のプロセッサノードシステム。

【請求項７】

前記一のプロセッサ基板のブリッジのルートコンプレックスモードに設定されたポートに設けられるＰＣＩエクスプレスコネクタと、前記別のプロセッサ基板のブリッジのエンドポイントモードに設定されたポートに設けられるＰＣＩエクスプレスコネクタとがフレキシブル基板により配線接続されてなることを特徴とする請求項４から６のいずれかに記載のプロセッサノードシステム。

【請求項８】

筐体内に前記４枚のプロセッサ基板面を互いに平行に設置し、筐体背面に各プロセッサ基板のブリッジのポートに設けられるＰＣＩエクスプレスコネクタが配置されるように構成し、筐体背面に配置された各プロセッサ基板のＰＣＩエクスプレスコネクタ間をフレキシブル基板により接続してなることを特徴とする請求項７に記載のプロセッサノードシステム。

【請求項９】

筐体内に前記４枚のプロセッサ基板面を互いに平行に設置し、筐体背面に各プロセッサ基板のブリッジのポートに設けられるＰＣＩエクスプレスコネクタが配置されるように構成し、前記一のプロセッサ基板のブリッジのルートコンプレックスモードに設定されたポートに設けられるＰＣＩエクスプレスコネクタと、前記別のプロセッサ基板のブリッジのエンドポイントモードに設定されたポートに設けられるＰＣＩエクスプレスコネクタとを相互接続するための一枚のバックプレーン基板をさらに設けたことを特徴とする請求項４から６のいずれかに記載のプロセッサノードシステム。

【請求項１０】

各プロセッサのメモリ空間に、相互接続された他のプロセッサの共有領域がＩ／Ｏアドレス空間としてメモリマッピングされることにより、各プロセッサは前記他のプロセッサの共有領域にアクセス可能に構成されることを特徴とする請求項４から９のいずれかに記

載のプロセッサノードシステム。
【請求項１１】
請求項４から１０のいずれかに記載のプロセッサノードシステムを複数含み、各プロセッサノードシステムをクラスタとして相互接続した、クラスタの接続形態を有するプロセッサノードクラスタシステムであって、
隣接する２つのプロセッサノードシステム間で当該プロセッサノードシステム内のプロセッサ基板間の接続に使用されていない空きポートを互いに接続することにより、前記複数のプロセッサノードシステム間を相互結合してなることを特徴とするプロセッサノードクラスタシステム。
【請求項１２】
前記空きポートにＰＣＩエクスプレスコネクタが設けられ、前記空きポートのＰＣＩエクスプレスコネクタ間がフレキシブル基板により配線接続されてなることを特徴とする請求項１１に記載のプロセッサノードクラスタシステム。
【発明の詳細な説明】
【技術分野】
【０００１】
この発明は、複数のプロセッサを相互接続したプロセッサノードシステムおよびプロセッサノードクラスタシステムに関する。
【背景技術】
【０００２】
パーソナルコンピュータやサーバには、ＰＣＩ（Peripheral Component Interconnect）バスを介して各種の周辺デバイスが接続され、情報処理システムが構成される。プロセッサの入出力バスと、周辺デバイスの入出力バスであるＰＣＩバスとは規格が異なるため、通常、ブリッジを介してプロセッサと周辺デバイスとが接続される。
【０００３】
情報処理システムの機能拡張や性能強化を図るために、グラフィックプロセッサや高速なメモリデバイスをＰＣＩデバイスとして接続することがあり、より多くの周辺デバイスをＰＣＩバスで接続できるようにすることが要請されている。そのため、ＰＣＩエクスプレス（PCI Express）（商標または登録商標）スイッチを用いて、一つのプロセッサに対して複数のデバイスを接続することが行われている。また、複数のプロセッサノードを相互接続したり、プロセッサノードとデバイスを相互接続するために、Ｉｎｆｉｎｉｂａｎｄと呼ばれる超高速インタフェース技術が用いられることがある。
【発明の開示】
【発明が解決しようとする課題】
【０００４】
１０ギガビットイーサネット（商標または登録商標）やＩｎｆｉｎｉｂａｎｄ技術を用いて複数のプロセッサノードを相互接続したクラスタシステムでは、プロセッサ間の高速な通信を実現することができるという利点があるが、スイッチが未だ高価であるため、クラスタシステムを低価格で提供することは難しく、クラスタ内のプロセッサノード数を増やしていくには限界がある。さらに、イーサネット（商標または登録商標）やＩｎｆｉｎｉｂａｎｄでは、パケットの生成、プロトコル処理などソフトウェアのオーバーヘッドが大きいというデメリットがある。
【０００５】
本発明はこうした課題に鑑みてなされたものであり、その目的は、複数のプロセッサを安価な手段により結合して、高速なプロセッサ間通信を実現する技術およびその技術を利用したプロセッサノードシステムやプロセッサノードクラスタ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】
【０００６】
上記課題を解決するために、本発明のある態様のプロセッサノードシステムは、プロセ

ッサと、前記プロセッサの入出力バスと周辺デバイスが接続されるＰＣＩエクスプレスとの間でデータを中継するブリッジとが搭載されたプロセッサ基板を複数含む。前記ブリッジのポートは当該プロセッサがホストとなるルートコンプレックスモードまたは当該プロセッサが周辺デバイスとなるエンドポイントモードに設定可能に構成され、一のプロセッサ基板のブリッジのルートコンプレックスモードに設定されたポートを、別のプロセッサ基板のブリッジのエンドポイントモードに設定されたポートに接続することにより、前記複数のプロセッサ基板間が相互結合される。

【０００７】

前記一のプロセッサ基板のブリッジのルートコンプレックスモードに設定されたポートに設けられるＰＣＩエクスプレスコネクタと、前記別のプロセッサ基板のブリッジのエンドポイントモードに設定されたポートに設けられるＰＣＩエクスプレスコネクタとがフレキシブル基板により配線接続されてもよい。

【０００８】

前記一のプロセッサ基板のブリッジのルートコンプレックスモードに設定されたポートに設けられるＰＣＩエクスプレスコネクタと、前記別のプロセッサ基板のブリッジのエンドポイントモードに設定されたポートに設けられるＰＣＩエクスプレスコネクタとを相互接続するための一枚のバックプレーン基板をさらに設けてもよい。

【０００９】

本発明の別の態様もまた、プロセッサノードシステムである。このプロセッサノードシステムは、プロセッサと、前記プロセッサの入出力バスと周辺デバイスが接続されるＰＣＩエクスプレスとの間でデータを中継するブリッジのセットが２組搭載されたプロセッサ基板を４枚含む。各ブリッジは、当該プロセッサがホストとなるルートコンプレックスモードに設定されたポートと当該プロセッサが周辺デバイスとなるエンドポイントモードに設定されたポートを有する。一のプロセッサ基板のルートコンプレックスモードに設定されたポートは、別のプロセッサ基板のブリッジのエンドポイントモードに設定されたポートに接続されることを条件として、各プロセッサ基板の合計４個のポートの内、３個のポートを用いて、前記４枚のプロセッサ基板の内、任意の２枚のプロセッサ基板間を相互結合される。

【００１０】

筐体内に前記４枚のプロセッサ基板面を互いに平行に設置し、筐体背面に各プロセッサ基板のブリッジのポートに設けられるＰＣＩエクスプレスコネクタが配置されるように構成し、筐体背面に配置された各プロセッサ基板のＰＣＩエクスプレスコネクタ間をフレキシブル基板により接続してもよい。

【００１１】

筐体内に前記４枚のプロセッサ基板面を互いに平行に設置し、筐体背面に各プロセッサ基板のブリッジのポートに設けられるＰＣＩエクスプレスコネクタが配置されるように構成し、前記一のプロセッサ基板のブリッジのルートコンプレックスモードに設定されたポートに設けられるＰＣＩエクスプレスコネクタと、前記別のプロセッサ基板のブリッジのエンドポイントモードに設定されたポートに設けられるＰＣＩエクスプレスコネクタとを相互接続するための一枚のバックプレーン基板をさらに設けてもよい。

【００１２】

本発明のさらに別の態様は、プロセッサノードクラスタシステムである。このプロセッサノードクラスタシステムは、プロセッサノードシステムを複数含む。隣接する２つのプロセッサノードシステム間で当該プロセッサノードシステム内のプロセッサ基板間の接続に使用されていない空きポートを互いに接続することにより、前記複数のプロセッサノードシステム間が相互結合される。

【００１３】

なお、以上の構成要素の任意の組合せ、本発明の表現を方法、装置、システム、コンピュータプログラム、データ構造、記録媒体などの間で変換したものもまた、本発明の態様として有効である。

【発明の効果】
【００１４】
本発明によれば、複数のプロセッサを相互接続して安価で高性能なシステムを構成することができる。
【発明を実施するための最良の形態】
【００１５】
実施の形態に係るクラスタシステムは、プロセッサが搭載された基板（ボード）をフレキシブル基板で密結合することにより構成される。図１を参照して、各プロセッサ基板の構成を説明し、図２を参照して、４つのプロセッサ基板をフレキシブル基板により密結合したノードの構成を説明する。図３を参照して、複数のノード間をフレキシブル基板により連結することにより構成されるクラスタシステムを説明する。また、図４～図８を参照して、プロセッサ基板間をフレキシブル基板により接続する形態について説明する。
【００１６】
図１は、プロセッサ基板５０の構成図である。プロセッサ基板５０には、２つのマルチコアプロセッサ（Multicore Processor）（以下、「ＭＣＰ」と呼ぶ）２０、２１が搭載されている。各ＭＣＰ２０、２１は、複数のプロセッサコアを１つのパッケージに集積したものであり、プロセッサコアとして、１つのプロセッシングエレメント（ＰＥ）と、複数のサブプロセッシングエレメント（ＳＰＥ）を含む。ＰＥは、キャッシュメモリを有し、ＤＲＡＭ１０から読み込んだデータをキャッシュしながら、情報処理を行う。また、ＰＥは、各ＭＣＰ２０、２１全体を統括的に制御する。各ＳＰＥはローカルメモリを内部にもち、ローカルメモリに対してデータを読み書きしながら、情報処理を行う。複数のＳＰＥは非同期で動作する。
【００１７】
２つのＭＣＰ２０、２１は入出力インタフェース（以下、「ＩＯＩＦ」と呼ぶ）６４を介して相互に接続されており、高速なデータ通信が可能である。さらに、各ＭＣＰ２０、２１は、ＩＯＩＦ６２、６３を介してブリッジ３０、３１の上流（アップストリーム）ポートに接続されている。ブリッジ３０、３１の下流（ダウンストリーム）ポートには、ＰＣＩエクスプレス６６、６７を介して各種の周辺（ペリフェラル）デバイスや他のプロセッサ基板が接続される。
【００１８】
ここで、ＰＣＩエクスプレス６６、６７は、ＰＣＩエクスプレス（PCI Express）（商標または登録商標）の仕様にしたがうものであるが、現行のＰＣＩエクスプレス規格に限定する趣旨ではなく、現行のＰＣＩエクスプレス規格に準拠するものや、現行のＰＣＩエクスプレス規格をさらに拡張したり、発展させた規格によるものであってもかまわない。ＰＣＩエクスプレス６６、６７で接続された周辺デバイスや他のプロセッサ基板を以下、「ＰＣＩデバイス」という。
【００１９】
ＩＯＩＦ６２、６３、６４は、上りと下りの２つのチャネルをもち、メモリバスに匹敵する高い帯域幅、たとえば、数十ギガバイト／秒を実現している。各ＭＣＰ２０、２１の所定のメモリ領域は、ＩＯＩＦ６２、６３、６４を介して参照可能なＩ／Ｏアドレス空間にメモリマッピングされる。各ＭＣＰ２０、２１は、ＩＯＩＦ６２、６３、６４を介してＩ／Ｏアドレス空間にマッピングされた他のＭＣＰのメモリ領域にアクセスすることが可能であり、高速なプロセッサ間通信が実現される。
【００２０】
各ブリッジ３０、３１は、ＩＯＩＦ６２、６３とＰＣＩエクスプレス６６、６７とを「橋渡し」することで、ＭＣＰ２０、２１とＰＣＩデバイスとを相互接続する。ＩＯＩＦ６２、６３と、ＰＣＩエクスプレス６６、６７とは、バスの規格が異なるため、ブリッジ３０、３１は、２つのバスの間でプロトコルの変換を行い、ＭＣＰ２０、２１とＰＣＩデバイスとがやりとりするデータのフォーマットを各バスの仕様に合わせる。
【００２１】

ＰＣＩエクスプレス６６、６７に接続されたＰＣＩデバイスの先にさらにＰＣＩエクスプレスを介してＰＣＩデバイスを接続していくと、ＭＣＰ２０、２１をルート（根）とし、リーフ（葉）にはＰＣＩデバイスが接続されたＰＣＩデバイスのツリー（木）構造が形成される。以下、このＰＣＩデバイスのツリー構造を「ＰＣＩツリー」という。

【００２２】

各ブリッジ３０、３１の下流ポートは２つ設けられており、一方は、ルートコンプレックス（ＲＣ；Root Complex）として、他方は、エンドポイント（ＥＰ；Endpoint）としてコンフィグレーションして用いることができる。１つのポートがルートコンプレックスモードとエンドポイントモードを切り替えられるように構成されてもよい。ブリッジ３０、３１の下流ポートをルートコンプレックスとして用いると、ＭＣＰ２０、２１は、ＰＣＩツリーのルートとなって、ＰＣＩデバイスを接続するホストとして機能する。ブリッジ３０、３１の下流ポートをエンドポイントとして用いると、ＭＣＰ２０、２１は、ホストに接続されるＰＣＩデバイスとして機能する。

【００２３】

ブリッジ３０、３１の下流ポートには、ルートコンプレックス用のコネクタ４０Ｒ、４１Ｒと、エンドポイント用のコネクタ４０Ｅ、４１Ｅとが設けられる。本実施の形態では、プロセッサ基板５０に設けられた合計４個のコネクタ４０Ｒ、４０Ｅ、４１Ｒ、４１Ｅの内、３個のコネクタを同一ノード内の他の３つのプロセッサ基板との接続に用い、残りの１個のコネクタを他のノードのプロセッサ基板との接続に用いる。

【００２４】

図２は、４つのプロセッサ基板が密結合されたノード１００の構成図である。ノード１００は、第１プロセッサ基板５０、第２プロセッサ基板５１、第３プロセッサ基板５２、および第４プロセッサ基板５３をフレキシブル基板で相互に接続したものである。各プロセッサ基板５０～５３の構成は、図１で説明した通りである。

【００２５】

以下、第１プロセッサ基板５０、第２プロセッサ基板５１、第３プロセッサ基板５２、第４プロセッサ基板５３をそれぞれ「プロセッサ基板０」、「プロセッサ基板１」、「プロセッサ基板２」、「プロセッサ基板３」と呼ぶ。

【００２６】

プロセッサ基板０に搭載された２つのＭＣＰ２０、２１をそれぞれ「ＭＣＰ０」、「ＭＣＰ１」と呼び、ＭＣＰ０、ＭＣＰ１に接続されたブリッジ３０、３１をそれぞれ「ブリッジ０」、「ブリッジ１」と呼ぶ。同様に、プロセッサ基板１に搭載された２つのＭＣＰ２２、２３をそれぞれ「ＭＣＰ２」、「ＭＣＰ３」と呼び、ＭＣＰ２、ＭＣＰ３に接続されたブリッジ３２、３３をそれぞれ「ブリッジ２」、「ブリッジ３」と呼ぶ。プロセッサ基板２に搭載された２つのＭＣＰ２４、２５をそれぞれ「ＭＣＰ４」、「ＭＣＰ５」と呼び、ＭＣＰ４、ＭＣＰ５に接続されたブリッジ３４、３５をそれぞれ「ブリッジ４」、「ブリッジ５」と呼ぶ。プロセッサ基板３に搭載された２つのＭＣＰ２６、２７をそれぞれ「ＭＣＰ６」、「ＭＣＰ７」と呼び、ＭＣＰ６、ＭＣＰ７に接続されたブリッジ３６、３７をそれぞれ「ブリッジ６」、「ブリッジ７」と呼ぶ。

【００２７】

各プロセッサ基板内の２つのＭＣＰを相互接続するＩＯＩＦを「ＩＯＩＦ０」と呼び、ＭＣＰとブリッジの上流ポート間のＩＯＩＦを「ＩＯＩＦ１」と呼ぶ。

【００２８】

ブリッジ０のＲＣ用コネクタ、ＥＰ用コネクタをそれぞれ「コネクタＲＣ０」、「コネクタＥＰ０」と呼ぶ。同様に、ブリッジ１～ブリッジ７のＲＣ用コネクタをそれぞれ「コネクタＲＣ１」～「コネクタＲＣ７」と呼び、ブリッジ１～ブリッジ７のＥＰ用コネクタをそれぞれ「コネクタＥＰ１」～「コネクタＥＰ７」と呼ぶ。

【００２９】

プロセッサ基板０のＭＣＰ０側のブリッジ０のコネクタＲＣ０は、プロセッサ基板１のＭＣＰ３側のブリッジ３のコネクタＥＰ３と接続される。この接続により、プロセッサ基

板０のＭＣＰ０から見た場合、ＭＣＰ０はルートコンプレックスとして機能し、プロセッサ基板１のＭＣＰ３はエンドポイントとして機能する。すなわち、プロセッサ基板０のＭＣＰ０はホストであり、プロセッサ基板１をＰＣＩデバイスとして接続した形態となり、ＭＣＰ０をルートとしてＭＣＰ３をつないだＰＣＩツリーが形成される。

【００３０】

プロセッサ基板０のＭＣＰ１側のブリッジ１のコネクタＲＣ１は、プロセッサ基板２のＭＣＰ４側のブリッジ４のコネクタＥＰ４と接続される。ルートコンプレックスであるプロセッサ基板０のＭＣＰ１から見た場合、プロセッサ基板０のＭＣＰ１をホスト、プロセッサ基板２をデバイスとするＰＣＩツリーが形成される。

【００３１】

プロセッサ基板０のＭＣＰ１側のブリッジ１のコネクタＥＰ１は、プロセッサ基板３のＭＣＰ６側のブリッジ６のコネクタＲＣ６と接続される。ルートコンプレックスであるプロセッサ基板３のＭＣＰ６から見た場合、プロセッサ基板３のＭＣＰ６をホスト、プロセッサ基板０をデバイスとするＰＣＩツリーが形成される。

【００３２】

同様に、プロセッサ基板１のＭＣＰ２側のブリッジ２のコネクタＲＣ２は、プロセッサ基板２のＭＣＰ５側のブリッジ５のコネクタＥＰ５と接続され、ブリッジ２のコネクタＥＰ２は、プロセッサ基板３のＭＣＰ７側のブリッジ７のコネクタＲＣ７と接続される。プロセッサ基板２のＭＣＰ４側のブリッジ４のコネクタＲＣ４は、プロセッサ基板３のＭＣＰ７側のブリッジ７のコネクタＥＰ７と接続される。

【００３３】

ノード１００内のプロセッサ基板間の接続に用いられないブリッジのコネクタ、すなわち、プロセッサ基板０のブリッジ０のコネクタＥＰ０、プロセッサ基板１のブリッジ３のコネクタＲＣ３、プロセッサ基板２のブリッジ５のコネクタＲＣ５、およびプロセッサ基板３のブリッジ６のコネクタＥＰ６は、空きスロットとして、他のノードのプロセッサ基板との接続に利用される。

【００３４】

図３は、複数のノードを連結したクラスタシステム２００の構成図である。クラスタシステム２００は、図２で説明した構成のノード１００～１０２、１１０～１１２、１２０～１２０を上下左右に連結したものである。たとえば、ノード１００の右にはノード１０１が接続され、ノード１０１のさらに右にはノード１０２が接続される。ノード１００の下にはノード１１０が接続され、ノード１１０のさらに下にはノード１２０が接続される。

【００３５】

同図に示すように、左右に並ぶ２つのノードは、左側のノードのプロセッサ基板３のコネクタＥＰ６と、右側のノードのプロセッサ基板１のコネクタＲＣ３とを接続することにより、結合される。上下に並ぶ２つのノードは、上側のノードのプロセッサ基板２のコネクタＲＣ５と、下側のノードのプロセッサ基板０のコネクタＥＰ０とを接続することにより、結合される。

【００３６】

クラスタシステム２００において、端部に位置するノードの隣接ノードが存在しない側のコネクタは空きスロットになるが、この空きスロットには各種の周辺デバイスを接続したり、さらにノードを接続することにより、システムを拡張することができる。

【００３７】

このように、クラスタシステム２００では、ノードを上下左右に結合する平面上の配置により、ノード数を自由自在に増やしていくことができるという利点がある。

【００３８】

クラスタシステム２００において、各ノード内の４枚のプロセッサ基板間の接続、およびノード間の接続には、フレキシブル基板が用いられる。以下、図４～図８を参照して、フレキシブル基板を用いた接続形態を説明する。

【００３９】
図４は、プロセッサ基板５０の裏面の配線の模式図である。同図において、ＭＣＰ０と複数のＤＲＡＭ１０の間の配線、ＭＣＰ０とブリッジ０の間の配線、ブリッジ０とコネクタＲＣ０、ＥＰ０の間の配線が示されている。また、ＭＣＰ１と複数のＤＲＡＭ１１の間の配線、ＭＣＰ１とブリッジ１の間の配線、ブリッジ１とコネクタＲＣ１、ＥＰ１の間の配線が示されている。各コネクタＲＣ０、ＥＰ０、ＲＣ１、ＥＰ１はＰＣＩ－Ｅｘｐｒｅｓｓ×１６コネクタであり、フレキシブル基板を接続することができる。
【００４０】
図５は、ノード１００内の４枚のプロセッサ基板５０～５３間をフレキシブル基板によって接続した構成を示す図である。フレキシブル基板は、プリント配線基板の一種であり、ＦＰＣ（Flexible Printed Circuit）とも呼ばれ、薄くて屈曲性がある。
【００４１】
図２で説明したプロセッサ基板５０～５３（プロセッサ基板０～３）を、フレキシブル基板による接続がしやすいように、プロセッサ基板１（符号５１）、プロセッサ基板０（符号５０）、プロセッサ基板２（符号５２）、プロセッサ基板３（符号５３）の順に、基板面を互いに平行にして配置する。
【００４２】
プロセッサ基板１のコネクタＲＣ２は、フレキシブル基板２０１によりプロセッサ基板２のコネクタＥＰ５と接続される。プロセッサ基板１のコネクタＥＰ２は、フレキシブル基板２０２によりプロセッサ基板３のコネクタＲＣ７と接続される。プロセッサ基板１のコネクタＥＰ３は、フレキシブル基板２０３によりプロセッサ基板０のコネクタＲＣ０と接続される。
【００４３】
プロセッサ基板０のコネクタＲＣ１は、プロセッサ基板２のコネクタＥＰ４とフレキシブル基板２０４によって接続される。プロセッサ基板０のコネクタＥＰ１は、プロセッサ基板３のコネクタＲＣ６とフレキシブル基板２０５によって接続される。プロセッサ基板２のコネクタＲＣ４は、プロセッサ基板３のコネクタＥＰ７とフレキシブル基板２０６によって接続される。
【００４４】
プロセッサ基板１のコネクタＥＰ２とプロセッサ基板３のコネクタＲＣ７をつなぐフレキシブル基板２０２は、プロセッサ基板０のコネクタＲＣ１とプロセッサ基板２のコネクタＥＰ４をつなぐフレキシブル基板２０４の上側をまたいでいる。このようにフレキシブル基板を用いれば、配線の上に別の配線が通るような接続形態も可能であり、４枚のプロセッサ基板を平行に並べて相互に密結合させ、省スペース化を図ることができる。
【００４５】
また、汎用品のＰＣＩエクスプレスコネクタとフレキシブル基板を用いてプロセッサ基板間を接続する構成であるため、ＰＣＩエクスプレススイッチなどでプロセッサ基板間を相互接続した構成に比べて、はるかに安価であり、製造コストを削減することができる。
【００４６】
さらに、プロセッサ基板の部品実装密度を高くし、プロセッサ基板を小型化することによって、より短いフレキシブル基板でプロセッサ基板を相互接続することができ、高速信号を扱うことが可能になる。ＰＣＩ－Ｅｘｐｒｅｓｓは高速通信を前提としており、ケーブル接続では信号の伝搬が遅く、ケーブル接続によってプロセッサ基板間の密結合を実現することは困難である。本実施の形態では、フレキシブル基板でプロセッサ基板間を配線するため、高速信号の伝搬が可能である。
【００４７】
図６は、クラスタシステム２００内の複数のノード間をフレキシブル基板によって接続した構成を示す図である。同図では、図３の４つの隣接するノード１００、１０１、１１０、１１１の接続形態が示されている。各ノード１００、１０１、１１０、１１１内の４枚のプロセッサ基板間は、図５で説明したようにフレキシブル基板で接続されている。た

だし、ノード１１０については、ノード間の接続形態を把握しやすくするため、ノード内のプロセッサ基板間を接続するフレキシブル基板を図示していない。

【００４８】

ノード１００のプロセッサ基板３のコネクタＥＰ６は、ノード１０１のプロセッサ基板１のコネクタＲＣ３とフレキシブル基板２１１によって接続される。これにより２つのノード１００、１０１が左右方向に結合する。同様にノード１１０のプロセッサ基板３のコネクタＥＰ６は、ノード１１１のプロセッサ基板１のコネクタＲＣ３とフレキシブル基板２２１によって接続され、２つのノード１１０、１１１が左右方向に結合する。

【００４９】

ノード１００のプロセッサ基板２のコネクタＲＣ５は、ノード１１０のフレキシブル基板０のコネクタＥＰ０とフレキシブル基板２１４によって接続される。これにより２つのノード１００、１１０が上下方向に結合する。同様に、ノード１０１のプロセッサ基板２のコネクタＲＣ５は、ノード１１１のフレキシブル基板０のコネクタＥＰ０とフレキシブル基板２２４によって接続され、２つのノード１０１、１１１が上下方向に結合する。

【００５０】

クラスタシステム２００では、ノード間の接続にもフレキシブル基板が用いられ、省スペース化とコストダウンを図ることができる。クラスタシステム２００は、複数のノードを平面上で上下左右に配置して接続する形態であるため、隣り合うノード間の距離を短くすることができ、ノード間接続に用いるフレキシブル基板の長さを十分に短くすることができ、ＰＣＩ－Ｅｘｐｒｅｓｓの高速信号を扱うことができる。

【００５１】

図７は、ノード１００の筐体を説明する図である。ノード１００の筐体には、４枚のプロセッサ基板５０から５３が収納されており、背面のコネクタ間は図５で説明したように６個のフレキシブル基板２０１～２０６で接続されている。さらに、図６で説明したように、プロセッサ基板０には、上方向に隣接するノードのプロセッサ基板２と接続するためのフレキシブル基板２１３が設けられ、プロセッサ基板２には、下方向に隣接するノードのプロセッサ基板０と接続するためのフレキシブル基板２１４が設けられる。一方、プロセッサ基板１には、左方向に隣接するノードのプロセッサ基板３と接続するためのフレキシブル基板２１２が設けられ、プロセッサ基板３には、右方向に隣接するノードのプロセッサ基板１と接続するためのフレキシブル基板２１１が設けられる。

【００５２】

図８は、クラスタシステム２００の筐体を説明する図である。図７のノード１００の筐体を上下左右に並べ、図７で説明したフレキシブル基板２１１、２１２によって左右方向にノード間を接続し、フレキシブル基板２１３、２１４によって上下方向にノード間を接続する。このように、クラスタシステム２００は、ノードの筐体を平面に配置してフレキシブル基板で接続することで容易に構成することができる。また、ノードの追加がしやすく、スケーラビリティがあり、多数のノードを結合したノードクラスタを省スペースで安価に提供することができる。

【００５３】

図４～図８では、プロセッサ基板にフレキシブル基板用コネクタが設けられ、フレキシブル基板用コネクタ間をフレキシブル基板で接続する形態を説明した。このように汎用ＰＣＩエクスプレスコネクタをフレキシブル基板で接続する形態は、接続形態の一例に過ぎず、これ以外の接続形態も考えられる。別の接続形態として、プロセッサ基板のカードエッジを差し込むための汎用のＰＣＩエクスプレスコネクタを搭載したバックプレーン基板を一枚用意して、４枚のプロセッサ基板をバックプレーン基板に差し込むことで図５で説明したＰＣＩエクスプレスコネクタ間の接続をバックプレーン基板上で実現してもよい。また、さらに別の接続形態として、プロセッサ基板に差動信号用コネクタペアであるＺＤコネクタを設け、バックプレーン基板上でＺＤコネクタを接続するように構成してもよい。このようなバックプレーン基板を用いた接続形態もフレキシブル基板を用いた接続形態と同様、安価な高速通信を実現することができ、また、省スペース化を図ることができる

。

【００５４】

図９Ａ～図９Ｄを参照して、図２で説明したノード１００内の４枚のプロセッサ基板のフルメッシュ型の結合により形成されるＰＣＩツリーを説明する。ＰＣＩツリーは、ノード内の各ＭＣＰがＰＣＩエクスプレスで接続されたＰＣＩデバイスを検索することにより得られる。

【００５５】

図９Ａは、ＭＣＰ０またはＭＣＰ１を中心に置いた場合のＰＣＩツリーを説明する図である。同図ではＰＣＩツリー構造において同じ階層にあるＭＣＰを水平に配置し、ルートに近い方を上に、リーフに近い方を下に配置している。

【００５６】

ルートコンプレックスであるＭＣＰ０のすぐ下の階層には、ＭＣＰ３がエンドポイントとして接続されている。ＭＣＰ２がＭＣＰ３と同階層にあって、ＭＣＰ３に接続されている。これにより、ＭＣＰ０をルートとする第１のＰＣＩツリーが形成される。ＭＣＰ１はＭＣＰ０と同階層にあって、ＭＣＰ０に接続されている。ルートコンプレックスであるＭＣＰ１のすぐ下の階層には、ＭＣＰ４がエンドポイントとして接続されている。ＭＣＰ５がＭＣＰ４と同階層にあって、ＭＣＰ４に接続されている。これにより、ＭＣＰ１をルートとする第２のＰＣＩツリーが形成される。ルートコンプレックスであるＭＣＰ６は、ＭＣＰ１のすぐ上の階層にあって、ＭＣＰ１をエンドポイントとして接続している。ＭＣＰ７がＭＣＰ６と同階層にあって、ＭＣＰ６に接続されている。これにより、ＭＣＰ６をルートとする第３のＰＣＩツリーが形成される。

【００５７】

図９Ｂは、ＭＣＰ２またはＭＣＰ３を中心に置いた場合のＰＣＩツリーを説明する図である。ルートコンプレックスであるＭＣＰ２のすぐ下の階層には、ＭＣＰ５がエンドポイントとして接続されている。ＭＣＰ４がＭＣＰ５と同階層にあって、ＭＣＰ５に接続されている。これにより、ＭＣＰ２をルートとする第１のＰＣＩツリーが形成される。ルートコンプレックスであるＭＣＰ７は、ＭＣＰ２のすぐ上の階層にあって、ＭＣＰ２をエンドポイントとして接続している。ＭＣＰ６がＭＣＰ７と同階層にあって、ＭＣＰ７に接続されている。これにより、ＭＣＰ７をルートとする第２のＰＣＩツリーが形成される。ＭＣＰ３はＭＣＰ２と同階層にあって、ＭＣＰ２に接続されている。ルートコンプレックスであるＭＣＰ３のすぐ下の階層には、他のノードのエンドポイントが接続される。ルートコンプレックスであるＭＣＰ０は、ＭＣＰ３のすぐ上の階層にあって、ＭＣＰ３をエンドポイントとして接続している。ＭＣＰ１がＭＣＰ０と同階層にあって、ＭＣＰ０に接続されている。これにより、ＭＣＰ０をルートとする第３のＰＣＩツリーが形成される。

【００５８】

図９Ｃは、ＭＣＰ４またはＭＣＰ５を中心に置いた場合のＰＣＩツリーを説明する図である。ルートコンプレックスであるＭＣＰ４のすぐ下の階層には、ＭＣＰ７がエンドポイントとして接続されている。ＭＣＰ６がＭＣＰ７と同階層にあって、ＭＣＰ７に接続されている。これにより、ＭＣＰ４をルートとする第１のＰＣＩツリーが形成される。ルートコンプレックスであるＭＣＰ１は、ＭＣＰ４のすぐ上の階層にあって、ＭＣＰ４をエンドポイントとして接続している。ＭＣＰ０がＭＣＰ１と同階層にあって、ＭＣＰ１に接続されている。これにより、ＭＣＰ１をルートとする第２のＰＣＩツリーが形成される。ＭＣＰ５はＭＣＰ４と同階層にあって、ＭＣＰ４に接続されている。ルートコンプレックスであるＭＣＰ５のすぐ下の階層には、他のノードのエンドポイントが接続される。ルートコンプレックスであるＭＣＰ２は、ＭＣＰ５のすぐ上の階層にあって、ＭＣＰ５をエンドポイントとして接続している。ＭＣＰ３がＭＣＰ２と同階層にあって、ＭＣＰ２に接続されている。これにより、ＭＣＰ２をルートとする第３のＰＣＩツリーが形成される。

【００５９】

図９Ｄは、ＭＣＰ６またはＭＣＰ７を中心に置いた場合のＰＣＩツリーを説明する図である。ルートコンプレックスであるＭＣＰ６のすぐ下の階層には、ＭＣＰ１がエンドポイ

ントとして接続されている。ＭＣＰ０がＭＣＰ１と同階層にあって、ＭＣＰ１に接続されている。これにより、ＭＣＰ６をルートとする第１のＰＣＩツリーが形成される。ＭＣＰ７はＭＣＰ６と同階層にあって、ＭＣＰ６に接続されている。ルートコンプレックスであるＭＣＰ７のすぐ下の階層には、ＭＣＰ２がエンドポイントとして接続されている。ＭＣＰ３がＭＣＰ２と同階層にあって、ＭＣＰ２に接続されている。これにより、ＭＣＰ７をルートとする第２のＰＣＩツリーが形成される。ルートコンプレックスであるＭＣＰ４は、ＭＣＰ７のすぐ上の階層にあって、ＭＣＰ７をエンドポイントとして接続している。ＭＣＰ５がＭＣＰ４と同階層にあって、ＭＣＰ４に接続されている。これにより、ＭＣＰ４をルートとする第３のＰＣＩツリーが形成される。

【００６０】

このように、ノード１００内の４つのプロセッサ基板間で図２で説明したようにＲＣコネクタとＥＰコネクタを接続することにより、あるＭＣＰをルートとするＰＣＩツリーが複数形成される。ノード１００内のＭＣＰ０～ＭＣＰ７はそれぞれ、自己をルートとするＰＣＩツリー内で、もしくは異なるＰＣＩツリーをまたぐことで他のＭＣＰとの間でデータ通信を行うことができる。ノード１００の空きスロットのコネクタと接続された隣接ノードのプロセッサ基板上のＭＣＰは、同一ＰＣＩツリー内にあるため、ノードをまたいでデータ通信が可能である。しかし、ノード１００の空きスロットのコネクタと接続されていない他のノードのＭＣＰとデータ通信をする場合は、同一ＰＣＩツリー内にないため、ルーティングが必要となる。このため、ノード１００内の各ＭＣＰは、ソフトウェアでルーティングを実行して、他のＰＣＩツリー内のＭＣＰとの通信を可能にする。

【００６１】

ノード１００内の各ＭＣＰが、自己のＰＣＩツリー内で、もしくは異なるＰＣＩツリーをまたぐことで他のＭＣＰの所定の共有領域にアクセスすることができるように、各ＭＣＰのメモリ空間には他のＭＣＰの所定の共有領域がメモリマッピングされる。図１０～図１７を参照して、このメモリマッピングを説明する。

【００６２】

図１０は、ノード１００内の各ＭＣＰのメモリ空間３００を説明する図である。メモリ空間３００には、コヒーレントなローカルメモリ領域３５１とノンコヒーレントな共有メモリ領域３５２がある。コヒーレントなローカルメモリ領域３５１は、メモリアクセスのアトミック性が保証され、同期制御がなされる領域であり、他のＭＣＰからはアクセスすることはできない。ノンコヒーレントな共有メモリ領域３５２は、他のＭＣＰのメモリ空間にマッピングされ、他のＭＣＰからアクセスされる。メモリ空間３００には、さらに各ＭＣＰのＳＰＥおよびＰＥのレジスタやＳＰＥのローカルストアがマッピングされたノンコヒーレント領域３５３がある。このノンコヒーレント領域３５３の少なくとも一部は、他のＭＣＰのメモリ空間にマッピングされ、他のＭＣＰからアクセスされる。

【００６３】

メモリ空間３００には、ＩＯＩＦ０を介してアクセス可能なＩ／Ｏアドレス空間がＩＯＩＦ０領域３６０としてメモリマッピングされる。また、ＩＯＩＦ１を介してアクセス可能なＩ／Ｏアドレス空間がＩＯＩＦ１領域３７０としてメモリマッピングされる。

【００６４】

各ＭＣＰは、自分のメモリ空間３００内のノンコヒーレント領域３５３に含まれるＳＰＥ／ＰＥのレジスタやＳＰＥのローカルストア、およびノンコヒーレントな共有メモリ領域３５２を共有領域（shared area）として、ＩＯＩＦ０を介して他のＭＣＰに開放してアクセスを許可する。各ＭＣＰは、他のＭＣＰにアクセスを許可する共有領域の情報をＩＯＩＦ０用のＩ／Ｏページテーブル（以下、「ＩＯＰＴ」という）３１０に格納する。他のＭＣＰは、このＩＯＰＴ３１０を参照して、共有領域を自分のメモリ空間にマッピングしてアクセス可能にする。

【００６５】

図１１は、ＭＣＰ０の共有領域がＭＣＰ１のメモリ空間３０１にマッピングされ、ＭＣＰ１の共有領域がＭＣＰ０のメモリ空間３００にマッピングされる様子を説明する図であ

る。ＭＣＰ０のＩＯＩＦ０用のＩＯＰＴ３１０（「ＩＯＰＴ０」）がＩＯＩＦ０経由でＭＣＰ１に提示されると、ＩＯＰＴ０により指定されたＭＣＰ０の共有領域３２１がＭＣＰ１のメモリ空間３０１のＩＯＩＦ０領域３６１にマッピングされる。ＭＣＰ０の共有領域３２１には、ＭＣＰ０のＳＰＥ／ＰＥのレジスタ、ＭＣＰ０のＳＰＥのローカルストア、およびＭＣＰ０の共有メモリが含まれる。

【００６６】

一方、ＭＣＰ１のＩＯＩＦ０用のＩＯＰＴ３１１（「ＩＯＰＴ１」）がＩＯＩＦ０経由でＭＣＰ０に提示されると、ＩＯＰＴ１により指定されたＭＣＰ１の共有領域３２０がＭＣＰ０のメモリ空間３００のＩＯＩＦ０領域３６０にマッピングされる。ＭＣＰ１の共有領域３２０には、ＭＣＰ１のＳＰＥ／ＰＥのレジスタ、ＭＣＰ１のＳＰＥのローカルストア、およびＭＣＰ１の共有メモリが含まれる。

【００６７】

このように、ＩＯＩＦ０を介して接続されたＭＣＰ０とＭＣＰ１は、互いに相手の共有領域が自分のメモリ空間３００、３０１にメモリマッピングされているため、相手の共有領域にアクセスすることができる。

【００６８】

図１２は、ＩＯＩＦ０で相互接続されたＭＣＰ０およびＭＣＰ１のそれぞれの共有領域がＩＯＩＦ１経由で接続された他のＭＣＰのメモリ空間にマッピングされる様子を説明する図である。

【００６９】

図１１で説明したように、ＭＣＰ１のメモリ空間のＩＯＩＦ０領域３６１には、ＭＣＰ０の共有領域３２１がマッピングされている。ＭＣＰ１は、自分の共有領域とともにＭＣＰ０の共有領域３２１をＩＯＩＦ１経由で接続された他のＭＣＰに開放してアクセスを許可する。ＭＣＰ１は、ＩＯＩＦ１用のＩＯＰＴ３３１に、自分の共有領域、すなわちＭＣＰ１のＳＰＥ／ＰＥのレジスタ、ＭＣＰ１のローカルストア、およびＭＣＰ１の共有メモリの情報を格納する。さらにＭＣＰ１は、ＩＯＩＦ１のＩＯＰＴ３３１に、ＭＣＰ０の共有領域３２１、すなわちＭＣＰ０のＳＰＥ／ＰＥのレジスタ、ＭＣＰ０のローカルストア、およびＭＣＰ０の共有メモリの情報を格納する。

【００７０】

図９Ａで説明したように、ＭＣＰ１とＭＣＰ６の接続関係は、ＭＣＰ６がルートコンプレックス、ＭＣＰ１がエンドポイントの関係であるから、エンドポイントであるＭＣＰ１が自分の共有領域の情報をルートコンプレックスであるＭＣＰ６に提示する。ＭＣＰ１は、ＩＯＩＦ１用のＩＯＰＴ３３１をＩＯＩＦ１経由で接続されたＭＣＰ６に提示する。ＭＣＰ６は、ＩＯＩＦ１用のＩＯＰＴ３３１で指定されたＭＣＰ０とＭＣＰ１の両方の共有領域３４２を自分のメモリ空間３０６のＩＯＩＦ１領域３７６にマッピングする。

【００７１】

図１３（ａ）、（ｂ）は、ＭＣＰ０がＩＯＩＦ１経由で接続された他のＭＣＰからＩＯＩＦ１用のＩＯＰＴの提示を受けた場合に、ＭＣＰ０のメモリ空間３００に他のＭＣＰの共有領域がマッピングされる様子を説明する図である。

【００７２】

図１３（ａ）に示すように、ルートコンプレックスであるＭＣＰ０は、ＩＯＩＦ１を経由してエンドポイントであるＭＣＰ３に接続されている。ＭＣＰ３はＩＯＩＦ０によりＭＣＰ２と相互接続されるから、ＭＣＰ３のメモリ空間のＩＯＩＦ０領域にはＭＣＰ２の共有領域がマッピングされる。図１２で説明したＭＣＰ１からＭＣＰ６へのＩＯＩＦ１用のＩＯＰＴの提示と同様に、エンドポイントであるＭＣＰ３は、自分の共有領域とＭＣＰ２の共有領域の情報をＩＯＩＦ１用のＩＯＰＴに格納してルートコンプレックスであるＭＣＰ０に提示する。

【００７３】

ＭＣＰ０は、ＭＣＰ３からＩＯＩＦ１用のＩＯＰＴの提示を受けて、図１３（ｂ）に示すように、メモリ空間３００のＩＯＩＦ１領域３７０にＭＣＰ２およびＭＣＰ３の共有領

域３４０をマッピングする。

【００７４】

図１４（ａ）、（ｂ）は、ＭＣＰ１がＩＯＩＦ１経由で接続された他のＭＣＰからＩＯＩＦ１用のＩＯＰＴの提示を受けた場合に、ＭＣＰ１のメモリ空間３０１に他のＭＣＰの共有領域がマッピングされる様子を説明する図である。

【００７５】

図１４（ａ）に示すように、ルートコンプレックスであるＭＣＰ１は、ＩＯＩＦ１を経由してエンドポイントであるＭＣＰ４に接続されている。ＭＣＰ４はＩＯＩＦ０によりＭＣＰ５と相互接続されるから、ＭＣＰ４のメモリ空間のＩＯＩＦ０領域にはＭＣＰ５の共有領域がマッピングされる。エンドポイントであるＭＣＰ４は、自分の共有領域とＭＣＰ５の共有領域の情報をＩＯＩＦ１用のＩＯＰＴに格納してルートコンプレックスであるＭＣＰ１に提示する。ＭＣＰ１は、ＭＣＰ４からＩＯＩＦ１用のＩＯＰＴの提示を受けて、図１４（ｂ）に示すように、メモリ空間３０１のＩＯＩＦ１領域３７１にＭＣＰ４およびＭＣＰ５の共有領域３４６をマッピングする。

【００７６】

同様に、ＭＣＰ６は、自分の共有領域とＭＣＰ７の共有領域の情報をＩＯＩＦ１用のＩＯＰＴに格納してＭＣＰ１に提示し、ＭＣＰ１は、ＭＣＰ６からＩＯＩＦ１用のＩＯＰＴの提示を受けて、図１４（ｂ）に示すように、メモリ空間３０１のＩＯＩＦ１領域３７１にＭＣＰ６およびＭＣＰ７の共有領域３４８をマッピングする。

【００７７】

次に、ＭＣＰ０とＭＣＰ１は、図１３（ｂ）、図１４（ｂ）のメモリ空間３００、３０１のＩＯＩＦ１領域３７０、３７１にマッピングされた、ＩＯＩＦ１経由で接続された他のＭＣＰの共有領域の情報を互いに交換する。

【００７８】

図１５は、ＭＣＰ０とＭＣＰ１間でメモリマッピングされた共有領域の情報をやりとりする様子を説明する図である。ＭＣＰ０は、メモリ空間３００のＩＯＩＦ１領域３７０にマッピングされたＭＣＰ２およびＭＣＰ３の共有領域３４０の情報をＩＯＩＦ０を介してＭＣＰ１に与える。ＭＣＰ１は、ＭＣＰ０から与えられた情報にもとづき、ＭＣＰ２およびＭＣＰ３の共有領域を自分のメモリ空間３０１のＩＯＩＦ０領域３６１にマッピングする。

【００７９】

一方、ＭＣＰ１は、メモリ空間３０１のＩＯＩＦ１領域３７１にマッピングされたＭＣＰ４およびＭＣＰ５の共有領域３４６の情報と、ＭＣＰ６およびＭＣＰ７の共有領域３４８の情報とをＩＯＩＦ０を介してＭＣＰ０に与える。ＭＣＰ０は、ＭＣＰ１から与えられた情報にもとづき、ＭＣＰ４およびＭＣＰ５の共有領域とＭＣＰ６およびＭＣＰ７の共有領域を自分のメモリ空間３００のＩＯＩＦ０領域３６０にマッピングする。

【００８０】

図１１、図１３（ｂ）、図１４（ｂ）、および図１５で説明した手順でメモリ空間に他のＭＣＰの共有領域がメモリマッピングされることにより、ＭＣＰ０は、図９Ａで説明した第１～第３ＰＣＩツリー内にあるＭＣＰ１～ＭＣＰ７の共有領域にアクセスすることができるようになる。なぜなら第１～第３ＰＣＩツリーをまたがって一つのアドレスマップが構築されているからである。同様に、ＭＣＰ１は、図９Ａで説明した第１～第３ＰＣＩツリー内にあるＭＣＰ０、ＭＣＰ２～ＭＣＰ７の共有領域にアクセスすることができるようになる。

【００８１】

このように、ノード１００内の各ＭＣＰは、第１～第３ＰＣＩツリー内の他のＭＣＰの共有領域を自分のメモリ空間にメモリマッピングしており、第１～第３ＰＣＩツリー内の他のＭＣＰの共有領域にアクセスしたり、第１～第３ＰＣＩツリー内の他のＭＣＰと共有領域を介したデータ通信や同期制御を実行することができる。ノード１００内のプロセッサ基板間はフレキシブル基板で接続され、高速なＰＣＩ－Ｅｘｐｒｅｓｓによる通信が可

能なハードウェア構成が採用されている。したがって、ノード１００内の各ＭＣＰは、メモリマッピングされた共有領域を高速にアクセスすることができ、他のＭＣＰとデータのやりとりを効率良く行うことができる。

【００８２】

図１６は、連接ノードのＭＣＰとの接続も含めたＰＣＩツリーを説明する図である。ＭＣＰ０のブリッジ０のコネクタＥＰ０は、隣接ノードのブリッジ５’のコネクタＲＣ５と接続され、ＭＣＰ５’がＭＣＰ０に対してルートコンプレックスとなる。ＭＣＰ４’はＭＣＰ５’と同階層にあって、ＭＣＰ５’と接続されている。ＭＣＰ０は、隣接ノードのＭＣＰ５’からＩＯＩＦ１用ＩＯＰＴの提示を受けて、ＭＣＰ５’およびＭＣＰ４’の共有領域３４９をメモリ空間３００のＩＯＩＦ１領域３７０にマッピングする。

【００８３】

図１７は、図１６のＰＣＩツリーの場合におけるＭＣＰ０のメモリ空間３００を説明する図である。図１７に示すように、ＩＯＩＦ１領域３７０には、ＭＣＰ２およびＭＣＰ３の共有領域３４０の他、ＭＣＰ４’およびＭＣＰ５’の共有領域３４０がメモリマッピングされる。また、ＩＯＩＦ０領域３６０には、ＭＣＰ４およびＭＣＰ５の共有領域３２６、ＭＣＰ６およびＭＣＰ７の共有領域３２８、およびＭＣＰ１の共有領域３２０がメモリマッピングされる。

【００８４】

まとめると、ＰＣＩのメモリマップは、ＰＣＩツリーのルートにあるホストプロセッサが、デバイスやスイッチのベースアドレスを設定することで構成される。エンドポイントであるデバイスは、自分が要求するアドレス領域のサイズをホストプロセッサに通知し、ホストプロセッサは、デバイスが要求したサイズにしたがってメモリマップを構築する。具体的には、要求するアドレスレンジのサイズは、コンフィグレーションレジスタのＢＡＲフィールドに実装するビット数で指定される。

【００８５】

本実施の形態のブリッジデバイスは、エンドポイントとして動作する場合、外部からアクセス可能なコンフィグレーションレジスタと内部からアクセス可能なコンフィグレーションレジスタをそれぞれ別々のレジスタとして実装し、それぞれのレジスタについて要求されるアドレスレンジのサイズ、すなわちＢＡＲの実装ビット数を設定することが可能である。これにより、システム初期化時に設定したサイズのアドレスレンジにより、ＰＣＩのアドレスマップがそれぞれのホストプロセッサによって構築される。ここで、それぞれのホストプロセッサとは、ルートコンプレックスとなるプロセッサと、エンドポイントとして動作するプロセッサのことである。

【００８６】

一方、ＩＯＩＦのメモリマップは、ＩＯＰＴに共有領域の情報を格納して他のＭＣＰに提示することにより設定される。この作業は外部からメモリアクセスがあった場合、その先にマッピングされる領域を設定するものである。この設定作業は、自分がルートコンプレックスとして動作する場合でも、自分がエンドポイントとして動作する場合でも、ＰＣＩからトランザクションを受け、それをメモリアクセスとして許可する場合は必要となる。

【００８７】

実際の運用としては、ＰＣＩで構築するアドレスサイズは余裕をもたせたサイズにしてＰＣＩメモリマップを構築し、その中で実際にメモリをマップする範囲は、ＩＯＰＴによって設定することになる。また、ＰＣＩメモリマップのアドレスレンジに関しては、ＰＣＩエクスプレスの規格にしたがい、エンドポイントが通知し、ルートコンプレックスがアドレス構築するということになるが、その中で、どの範囲がメモリやローカルストアにマッピングされているかについての情報は、図１５で説明したように、共有メモリを介したオリジナルプロトコルでやりとりする必要がある。

【００８８】

以上説明したように、本実施の形態によれば、プロセッサ基板の汎用のＰＣＩエクスプ

レスコネクタ間を安価なフレキシブル基板やバックプレーン基板で直接接続することにより、ＰＣＩエクスプレススイッチを必要としない、安価でかつ高性能なクラスタシステムを構築することができる。

【００８９】

以上、本発明を実施の形態をもとに説明した。実施の形態は例示であり、それらの各構成要素や各処理プロセスの組合せにいろいろな変形例が可能なこと、またそうした変形例も本発明の範囲にあることは当業者に理解されるところである。

【００９０】

上記の実施の形態では、プロセッサ基板にマルチコアプロセッサが搭載された場合を説明したが、これはシングルプロセッサであってもよい。また、実施の形態では、プロセッサ基板に２つのマルチコアプロセッサが搭載され、４枚のプロセッサ基板で１つのノードを構成する例を説明したが、プロセッサ基板に搭載されるプロセッサの数、１つのノード内のプロセッサ基板の数、ブリッジのコネクタ数などは、設計の自由度がある。ノード内の複数のプロセッサ基板をフレキシブル基板によって密結合し、ノード間をさらにフレキシブル基板で連結してノードクラスタを構成することができる限り、ノード内のプロセッサ基板の数と配置、ノードクラスタ内のノードの配置にはいろいろなパターンがありうる。いずれにしても安価、省スペース、高速通信の各要求を満足する接続形態が好ましい。

【００９１】

上記の実施の形態では、プロセッサ基板のブリッジのポートに他のプロセッサ基板のポートを接続したが、プロセッサ基板のブリッジのポートに周辺デバイスを接続し、プロセッサと各種周辺デバイスを相互結合したシステムを構成してもよい。また、ブリッジはプロセッサの入出力バスをＰＣＩエクスプレスに接続したが、他のプロセッサ基板や周辺デバイスが接続される外部インタフェースとしてＰＣＩエクスプレス以外のインタフェースが用いられてもよい。

【図面の簡単な説明】

【００９２】

【図１】プロセッサ基板の構成図である。

【図２】４つのプロセッサ基板が密結合されたノードの構成図である。

【図３】複数のノードを連結したクラスタシステムの構成図である。

【図４】プロセッサ基板の裏面の配線の模式図である。

【図５】ノード内の４枚のプロセッサ基板間をフレキシブル基板によって接続した構成を示す図である。

【図６】クラスタシステム内の複数のノード間をフレキシブル基板によって接続した構成を示す図である。

【図７】ノードの筐体を説明する図である。

【図８】クラスタシステムの筐体を説明する図である。

【図９Ａ】図２のノード内で形成されるＰＣＩツリーを説明する図である。

【図９Ｂ】図２のノード内で形成されるＰＣＩツリーを説明する図である。

【図９Ｃ】図２のノード内で形成されるＰＣＩツリーを説明する図である。

【図９Ｄ】図２のノード内で形成されるＰＣＩツリーを説明する図である。

【図１０】ノード内の各ＭＣＰのメモリ空間を説明する図である。

【図１１】あるＭＣＰのメモリ空間に他のＭＣＰの共有領域がマッピングされる様子を説明する図である。

【図１２】あるＭＣＰのメモリ空間に他のＭＣＰの共有領域がマッピングされる様子を説明する図である。

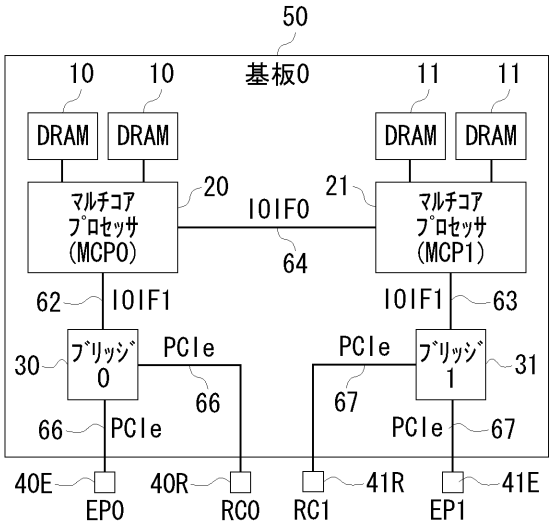
【図１３】あるＭＣＰのメモリ空間に他のＭＣＰの共有領域がマッピングされる様子を説明する図である。

【図１４】あるＭＣＰのメモリ空間に他のＭＣＰの共有領域がマッピングされる様子を説明する図である。

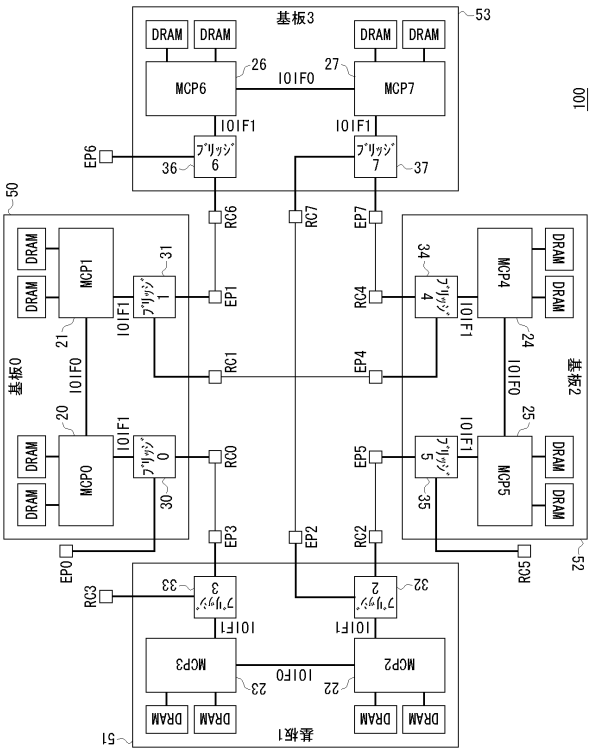
【図１５】２つのＭＣＰ間でメモリマッピングされた共有領域の情報をやりとりする様子

を説明する図である。
【図１６】連接ノードのＭＣＰとの接続も含めたＰＣＩツリーを説明する図である。
【図１７】図１６のＰＣＩツリーの場合におけるＭＣＰのメモリ空間を説明する図である。
【符号の説明】
【００９３】
１０　ＤＲＡＭ、　２０　マルチコアプロセッサ、　３０　ブリッジ、　５０　プロセッサ基板、　１００　ノード、　２００　クラスタシステム、　２０１～２０６、２１１～２１４　フレキシブル基板、　３００　メモリ空間、　３１０　ＩＯＰＴ。

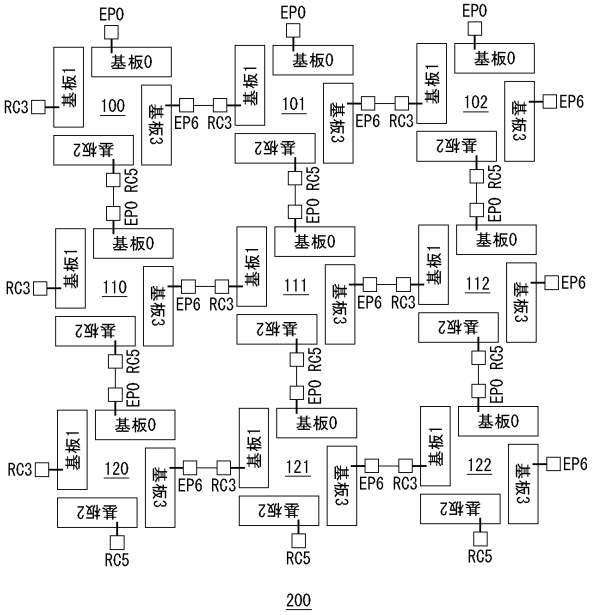
【図１】



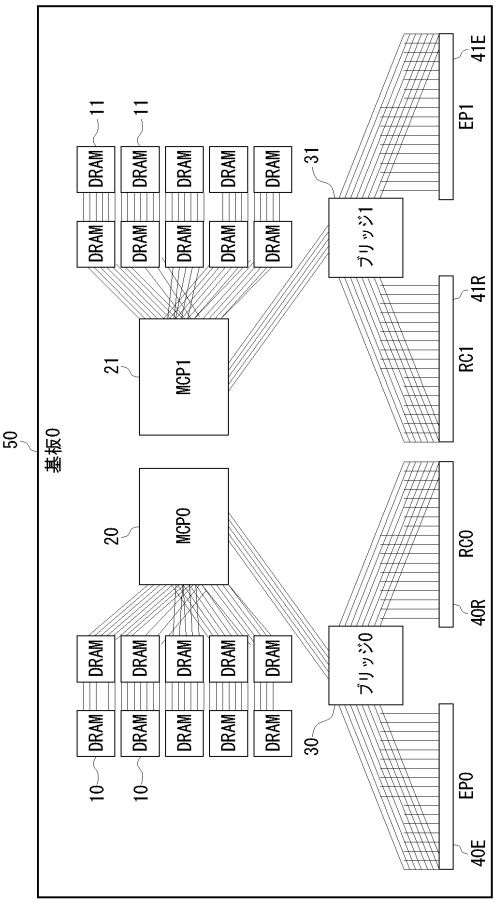
【図２】



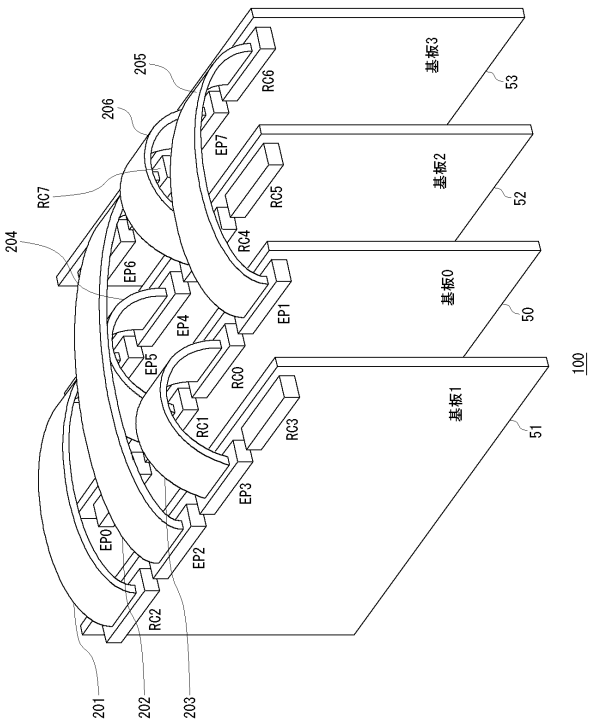
【図３】


EP0
基板0
基板1
基板2
基板3
RC3
EP6
RC5
100
101
102
110
111
112
120
121
122
200

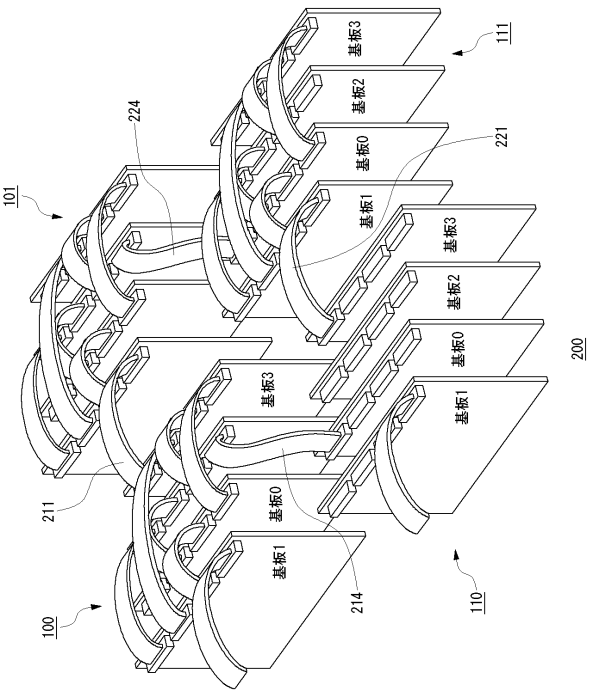
【図４】


50
基板0
DRAM
11
10
MCP1
MCP0
21
20
ブリッジ1
ブリッジ0
31
30
EP1
EP0
41E
41R
RC1
RC0
40R
40E

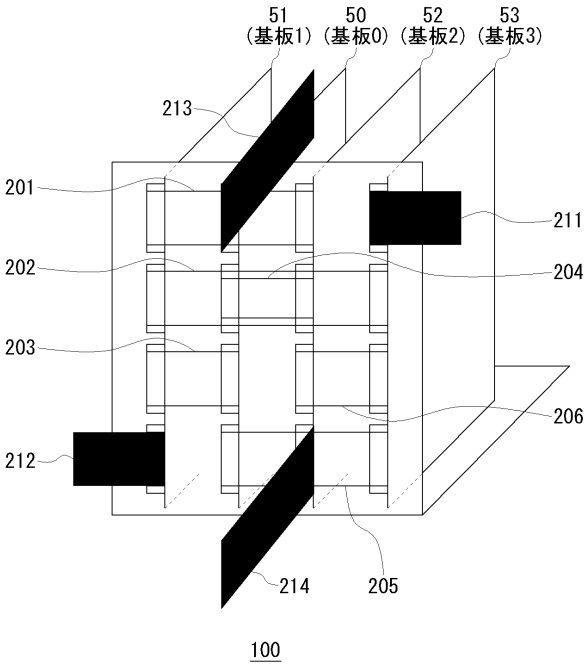
【図５】


基板3
基板2
基板0
基板1
53
52
50
51
100
205
206
204
RC7
RC6
RC5
RC4
RC0
RC1
RC3
RC2
EP7
EP6
EP5
EP4
EP3
EP2
EP1
EP0
201
202
203

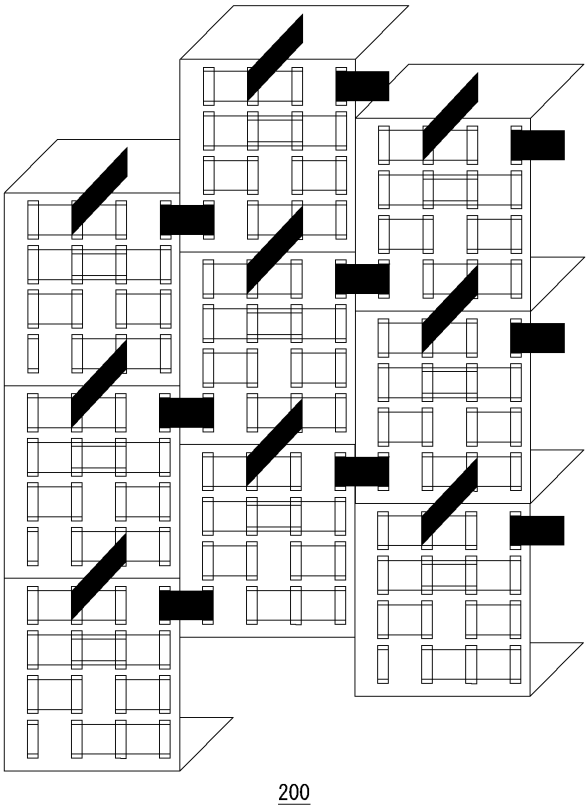
【図６】


111
101
224
221
211
214
110
100
200
基板3
基板2
基板0
基板1

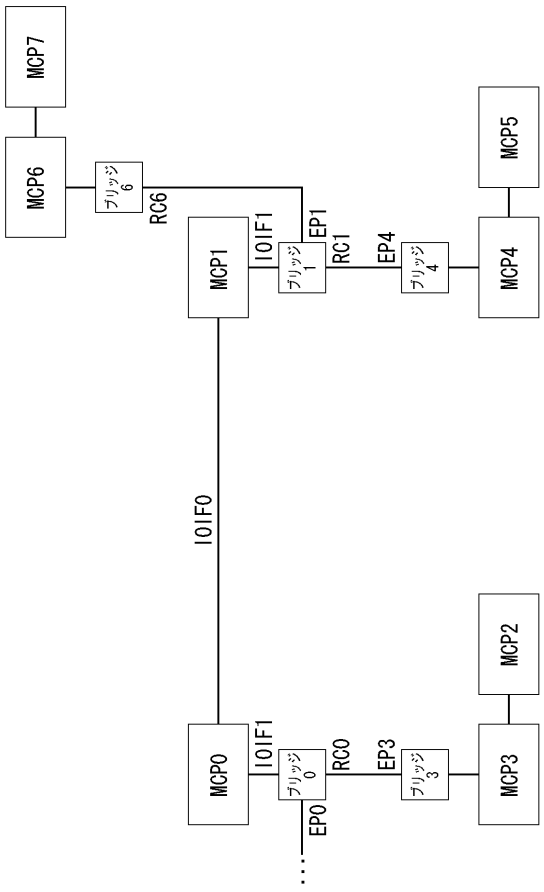
【図７】


51
(基板1)
50
(基板0)
52
(基板2)
53
(基板3)
213
201
202
203
212
211
204
206
214
205
100

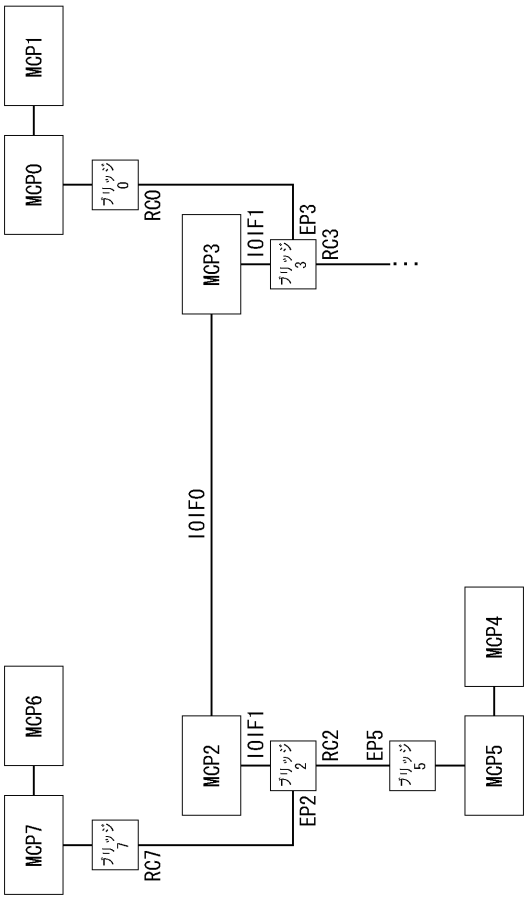
【図８】


200

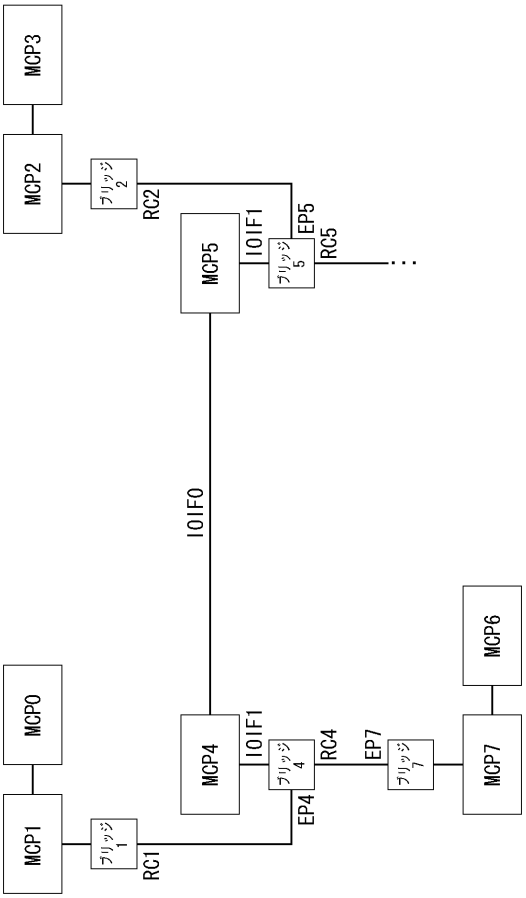
【図９Ａ】


MCP7
MCP6
ブリッジ 6
RC6
MCP1
IOIF1
ブリッジ 1
EP1
RC1
EP4
ブリッジ 4
MCP5
MCP4
IOIF0
MCP0
IOIF1
ブリッジ 0
EP0
RC0
EP3
ブリッジ 3
MCP2
MCP3

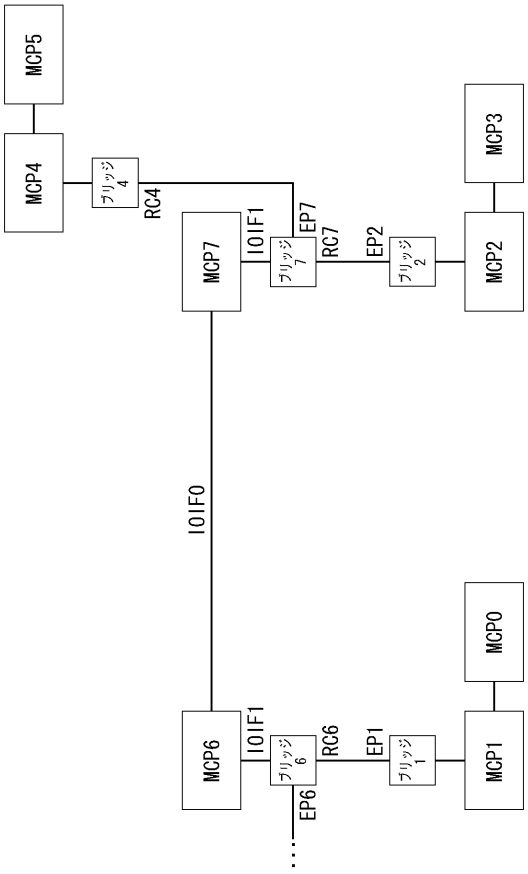
【図９Ｂ】


MCP1
MCP0
ブリッジ 0
RC0
MCP3
IOIF1
ブリッジ 3
EP3
RC3
IOIF0
MCP6
MCP7
ブリッジ 7
RC7
MCP2
IOIF1
ブリッジ 2
EP2
RC2
EP5
ブリッジ 5
MCP4
MCP5

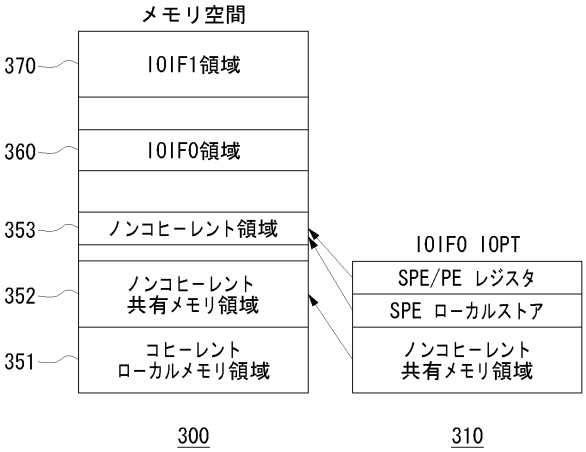
【図９Ｃ】



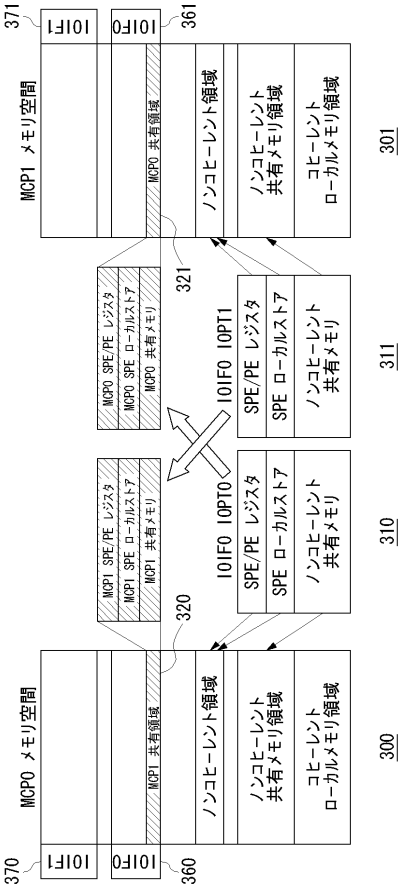
【図９Ｄ】



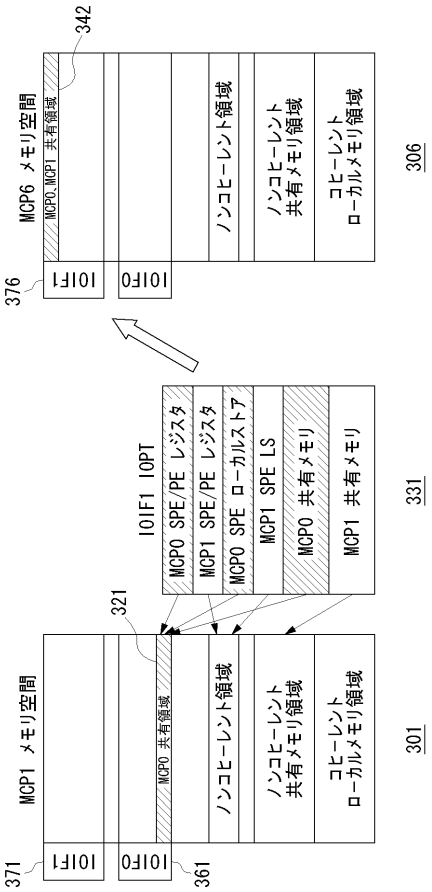
【図１０】



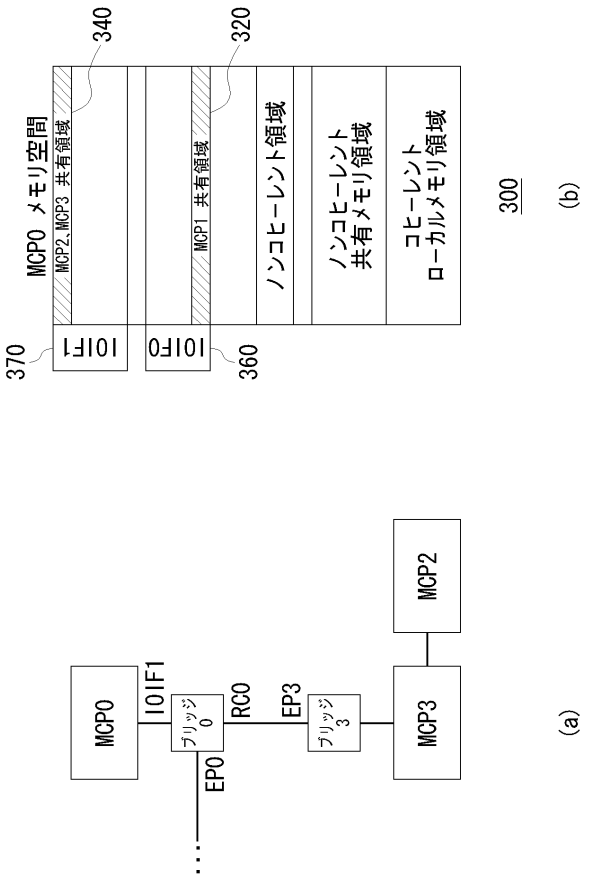
【図１１】



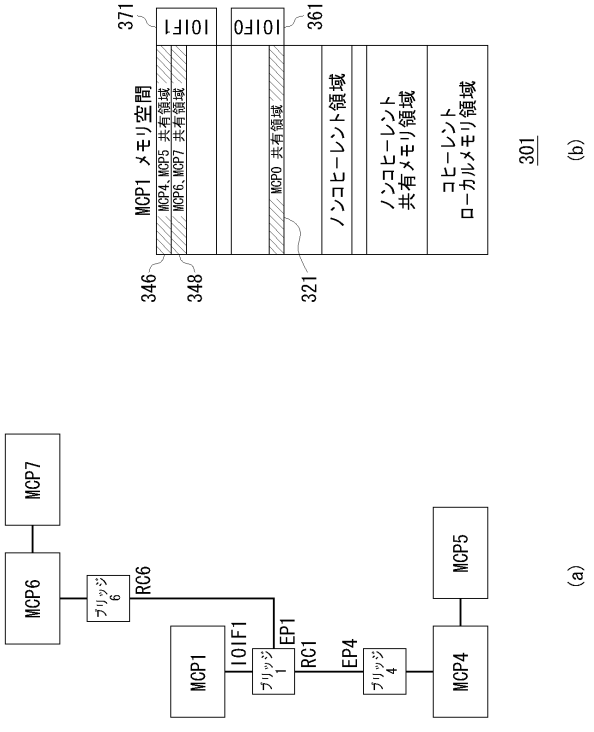
【図１２】



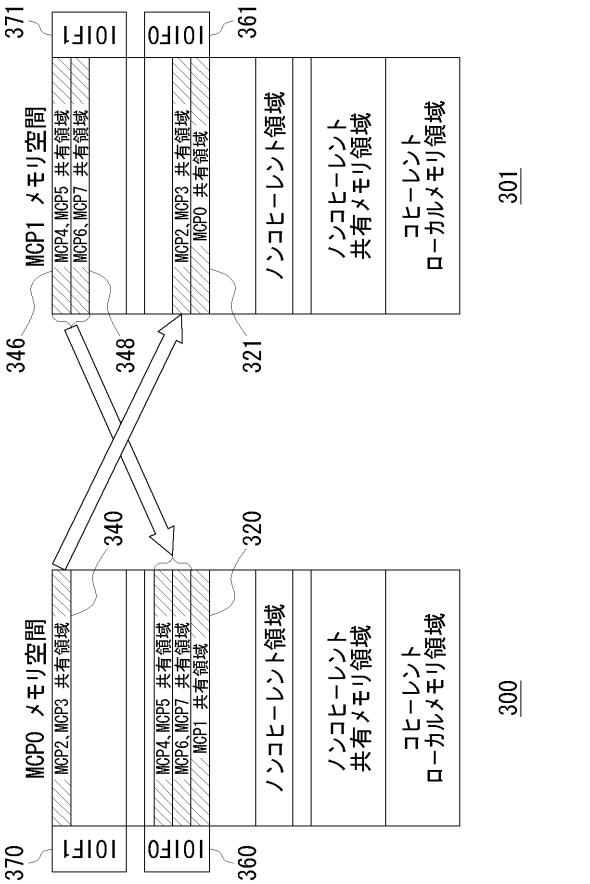
【図１３】



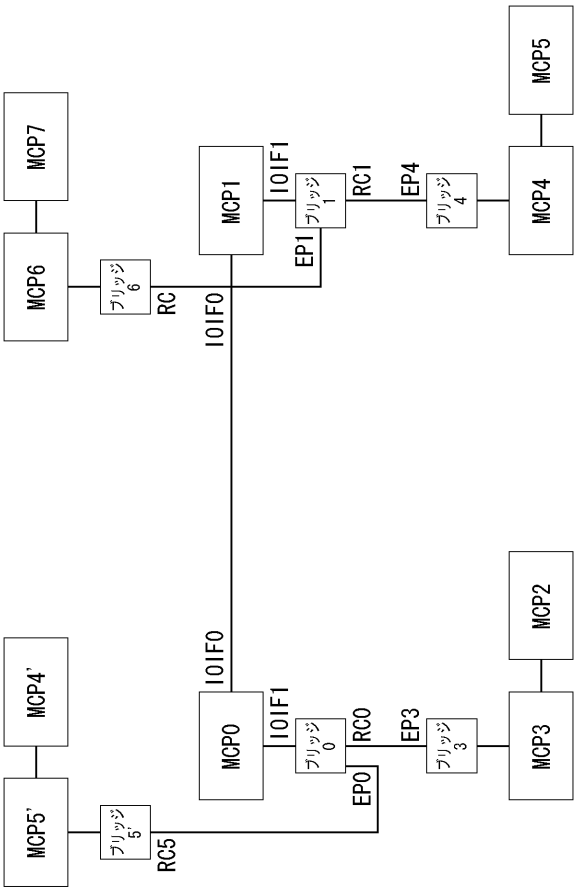
【図１４】



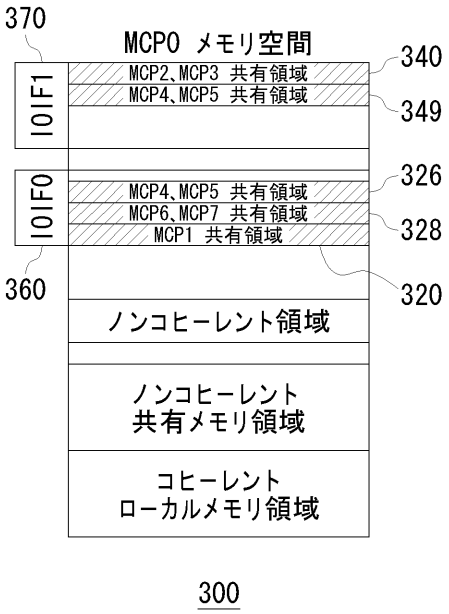
【図１５】



【図１６】


MCP7
MCP6
ブリッジ 6
RC
MCP1
IOIF1
ブリッジ 1
RC1
EP4
ブリッジ 4
MCP5
MCP4
EP1
IOIF0
MCP4'
MCP5'
ブリッジ 5'
RC5
MCP0
IOIF1
ブリッジ 0
RC0
EP3
ブリッジ 3
MCP2
MCP3
EP0
IOIF0

【図１７】


370
MCP0 メモリ空間
IOIF1
MCP2、MCP3 共有領域
MCP4、MCP5 共有領域
340
349
IOIF0
MCP4、MCP5 共有領域
MCP6、MCP7 共有領域
MCP1 共有領域
326
328
320
360
ノンコヒーレント領域
ノンコヒーレント
共有メモリ領域
コヒーレント
ローカルメモリ領域

300

フロントページの続き

(72)発明者　斎藤　英幸
東京都港区南青山２丁目６番２１号　株式会社ソニー・コンピュータエンタテインメント内
(72)発明者　堀江　和由
東京都港区南青山２丁目６番２１号　株式会社ソニー・コンピュータエンタテインメント内

審査官　三坂　敏夫

(56)参考文献　特開平０７－０９３２３６（ＪＰ，Ａ）
特表２０００－５０６６４５（ＪＰ，Ａ）
特開２００６－２０９４５６（ＪＰ，Ａ）
特開２００４－０７０９５４（ＪＰ，Ａ）
特開平０９－１４６８９５（ＪＰ，Ａ）
特開平０８－０４４４６０（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１５／１７３