(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第3849951号
(P3849951)
(45)発行日　平成18年11月22日(2006.11.22)　(24)登録日　平成18年9月8日(2006.9.8)

(51)Int.Cl.　F I
G06F 12/08 (2006.01)　G06F 12/08 531B
G06F 12/08 531E

請求項の数 3　(全 24 頁)

(21)出願番号　特願平9-59914
(22)出願日　平成9年2月27日(1997.2.27)
(65)公開番号　特開平10-240707
(43)公開日　平成10年9月11日(1998.9.11)
審査請求日　平成16年2月18日(2004.2.18)

(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　100099298
弁理士　伊藤　修
(72)発明者　垂井　俊明
東京都国分寺市東恋ケ窪一丁目２８０番地　株式会社日立製作所　中央研究所内
(72)発明者　岡澤　宏一
神奈川県川崎市幸区鹿島田８９０番地　株式会社日立製作所　情報・通信開発本部内
(72)発明者　岡田　康行
神奈川県海老名市下今泉８１０番地　株式会社日立製作所　オフィスシステム事業部内

最終頁に続く

(54)【発明の名称】主記憶共有型マルチプロセッサ

(57)【特許請求の範囲】

【請求項１】

１つ以上のＣＰＵ、キャッシュと、主記憶を備える複数のノードと、ノード間を結ぶネットワークとからなり、ネットワークを用いてノード間でキャッシュコヒーレントチェックのためのコマンドとその応答のやり取りを行ってキャッシュコヒーレント制御を行う主記憶共有型マルチプロセッサにおいて、

前記各ノードは、自ノードの主記憶の各ページ対応に、該当するページが他のノードからアクセスされたかどうかを記憶する第１のビットが１ビット割り当てられ、該第１のビットはシステムの初期化時にリセットされ、主記憶の該当するページが他のノードからアクセスされた場合に、ハードウェアによりセットされるテーブルと、

自ノードのＣＰＵが自ノードの主記憶をアクセスする際に、アクセスするページに該当する前記テーブルの前記第１のビットを検査し、該第１のビットがセットされていた場合には他のノードへのキャッシュコヒーレントチェックのためのコマンドの送出を行い、該第１のビットがセットされていなかった場合には他のノードへのキャッシュコヒーレントチェックのためのコマンドの送出を行わない手段を備えることを特徴とする主記憶共有型マルチプロセッサ。

【請求項２】

請求項１記載の主記憶共有型マルチプロセッサにおいて、

システムソフトウェアが主記憶のページをアロケートする際に、アロケートするページに対応する前記テーブルのビットを該システムソフトウェアがリセットすることを特徴と

する主記憶共有型マルチプロセッサ。
【請求項３】
請求項１記載の主記憶共有型マルチプロセッサにおいて、
前記テーブルに、前記主記憶の各ページ対応に該ページに対してキャッシュコヒーレント制御が必要で無いことを記憶する第２のビットを１ビット割り当て、
前記手段は、自ノードのＣＰＵが自ノードの主記憶をアクセスする際に該第２のビットを検査し、該ビットがセットされていない場合には、前記第１のビットの値にしたがって他のノードへのキャッシュコヒーレント制御の要否を判断し、前記第２のビットがセットされていた場合には他のノードへのキャッシュコヒーレントチェックのためのコマンドの送出を行わないことを特徴とする主記憶共有型マルチプロセッサ。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】
本発明は情報処理装置、特に、パーソナルコンピュータ（ＰＣ）、ワークステーション（ＷＳ）、サーバ機等に用いられる主記憶共有型の並列計算機システムに係り、特に、主記憶の制御方式に関する。
【０００２】
【従来の技術】
近年ＰＣ，ＷＳの上位機種及びサーバ機などでは、主記憶共有型のマルチプロセッサ（ＳＭＰ）構成が広まっており、性能を向上させるために、２０～３０台以上の多数台のプロセッサの間で主記憶を共有することが重要な課題になってきている。
主記憶共有型のマルチプロセッサの構成方法として広く使われている方式として共有バスがあるが、バスではバスのスループットがネックになるため、接続可能なプロセッサの数は高々８台程度である。従って、多数台のプロセッサを接続する方式としては適さない。
【０００３】
現在行われている、多数台のプロセッサを接続する主記憶共有マルチプロセッサの構成方法には、大きく２つの方式がある。
その一つに、クロスバスイッチによる構成があり、例えば、「進化したシステム・アーキテクチャ」（ＳｕｎＷｏｒｌｄ誌１９９６年１月号第２９頁～第３２頁）に開示されている。
この方式では、プロセッサと主記憶を持つ各ボードを、高速なクロスバスイッチで接続し、プロセッサ間のキャッシュ一貫性を保持する。
この方式では、キャッシュ一貫性の保持が高速にできるという利点が有る。
しかし、キャッシュの一貫性を保持するためのトランザクションが全プロセッサにブロードキャストされるため、クロスバスイッチにかかるトラフィックが非常に高く、性能的に隘路になるとともに、高速なスイッチが必要になるためコスト高を招くという欠点がある。
さらに、キャッシュ一貫性保持のためのトランザクションをブロードキャストしなければならないため、台数が非常に多いシステムを実現することは困難であり、数十台程度が限度である。
以下ではこの方式をスイッチ型ＳＭＰ（Ｓｙｍｍｅｔｒｉｃａｌ Ｍｕｌｔｉ Ｐｒｏｃｅｓｓｏｒ）と呼ぶ。
【０００４】
それに対して、ディレクトリ方式によるマルチプロセッサの構成があり、たとえば「Ｔｈｅ Ｓｔａｎｆｏｒｄ ＦＬＡＳＨ Ｍｕｌｔｉｐｒｏｃｅｓｓｏｒ」（第２１回ＩＳＣＡ Ｐｒｏｃｅｅｄｉｎｇｓ）に開示されている。
この方式では、主記憶のデータライン毎に、そのデータラインがどこのプロセッサのキャッシュに接続されているかを示すビットマップであるディレクトリを設けることにより、必要なプロセッサにのみキャッシュ一貫性を保持するためのトランザクションを送る。
それにより、スイッチにかかるトラフィックを大幅に削減することができ、スイッチのハ

ードウェアコストを削減することができる。
しかし、キャッシュ一貫性を保持するためのトランザクションを出す際には、必ず、主記憶に置かれたディレクトリの内容をチェックしなければならないため、アクセスレーテンシが大幅に増えるという欠点がある。
さらに、ディレクトリを置くためのメモリのコストが高くなると言う欠点を持つ。
【０００５】
上記のようにスイッチ型ＳＭＰとディレクトリ方式は一長一短である。
一般にスイッチ型ＳＭＰの方がハードウェア規模は大きくなり、台数が多くなった場合のスケーラビリティは良くないが、高性能を達成できる。
したがって、ＰＣやサーバ機等の、台数のそれほど多くない（３０台程度までの）システムでは、スイッチ型ＳＭＰで実現する方が得策である。
【０００６】
ここで、主記憶共有型マルチプロセッサを構成する上でのもう一つの問題点として、信頼性の問題がある。
従来の主記憶共有型マルチプロセッサは、システム全体で一つのＯＳを持つ。この方式は、システムの全てのプロセッサを一つのＯＳで管理できるため、柔軟なシステム運用（負荷分散等）をできるという利点を持つ。
しかし、多数台のプロセッサを主記憶共有のマルチプロセッサ構成で接続した場合、システムの信頼性が低下するという欠点を持つ。
複数のプロセッサをネットワークで接続したクラスタ構成のサーバや、ＭＰＰ（Ｍａｓｓｉｖｅｌｙ　Ｐａｒａｌｌｅｌ　Ｐｒｏｃｅｓｓｏｒｓ）では、ノード毎にＯＳは別なので、ＯＳなどのバグ等のためにシステムダウンしても、システムダウンするのは当該のノードのみである。
それに対して主記憶共有型のマルチプロセッサで、システム全体を１つのＯＳで制御する場合、あるプロセッサがシステムバグ等のためにダウンすると、ＯＳがダウンしてしまうため、全てのプロセッサが影響を受けてしまう。
【０００７】
上記の問題を避けるために、主記憶共有型のマルチプロセッサにおいて、複数のＯＳを走らせる方式が「Ｈｉｖｅ：Ｆａｕｌｔ　Ｃｏｎｔａｉｎｍｅｎｔ　ｆｏｒ　Ｓｈａｒｅｄ－Ｍｅｍｏｒｙ　Ｍｕｌｔｉｐｒｏｃｅｓｓｏｒｓ」（第１５回　ＡＣＭ　Ｓｙｍｐｏｓｉｕｍ　ｏｎ　Ｏｐｅｒａｔｉｎｇ　Ｓｙｓｔｅｍｓ　Ｐｒｉｎｃｉｐｌｅｓ）において開示されている。
この方式では、ディレクトリ方式の主記憶共有のマルチプロセッサに、以下の２つの機構を持たせる。
（１）システム全体を複数のセル（パーティション）にわけ、各パーティション毎に、独立したＯＳを走らせる。全体のアドレス空間は一つであり、ＯＳごとに担当するアドレス範囲が異なる。
（２）主記憶のページ毎に書き込みアクセス可能なプロセッサを表すビットマップを設け、書き込みアクセスをビットマップが１であるプロセッサのみに許す。
各プロセッサの主記憶に書き込みが行われる場合（Ｆｅｔｃｈ＆Ｉｎｖａｌｉｄａｔｅ命令によりデータがキャッシングされる場合、もしくはＷｒｉｔｅＢａｃｋ要求が到来した場合）、ビットマップの内容がチェックされ、ビットマップが１であるプロセッサからのアクセスのみが許される。
上記の（１）の機構により、たとえ、１つのパーティションのＯＳがダウンしても、他のパーティションがダウンする事を避けることができる。
さらに、（２）の機構を設けることにより、バグにより暴走したパーティションのプロセッサが、他のパーティションが使用するデータを破壊することを防止することができる。
以上のように、主記憶共有型のマルチプロセッサ内を複数のパーティションに分けることにより、システムの信頼性を大幅に向上させることができる。
【０００８】

【発明が解決しようとする課題】
上記従来技術で述べた、スイッチ型ＳＭＰを構成し、さらにＳＭＰ内をパーティションに分けようとする場合、以下に示す３つの問題点がある。
【０００９】
（Ａ）ローカル主記憶のアクセスが遅い
プロセッサが同一ボード内の主記憶をアクセスする場合、理想的にはクロスバスイッチを経由せずに高速にアクセスを行うことができるはずである。
しかし、実際は、必ず他のプロセッサへのキャッシュ一貫性保持のためのトランザクションを出し、他のプロセッサのキャッシュのチェック（以下ではこの処理をＣＣＣ：Ｃａｃｈｅ　Ｃｏｈｅｒｅｎｔ　Ｃｈｅｃｋと呼ぶ）を行わなければならない。
なぜなら、他のプロセッサのキャッシュの上に、アクセスされたデータのコピーがキャッシングされている可能性があるからである。
実際に他のプロセッサのキャッシュにデータがキャッシングされていた場合は、上記のＣＣＣは無駄にはならない。
しかし、アクセスされたデータが、他のプロセッサから一回もアクセスされていないローカルなデータの場合は、該当するデータが他のプロセッサのキャッシュ上にキャッシングされている可能性は０であり、本来は、ＣＣＣの処理は全く不要である。
そのため、無駄なＣＣＣのために、アクセスレーテンシを増加させてしまうのみならず、スイッチ上のトラフィックを増大させてしまうという問題点がある。
【００１０】
ディレクトリ方式では、キャッシュライン単位でどのプロセッサにキャッシングされているかが管理されているため、無駄なＣＣＣは生じない。
しかし、先に述べたように、ディレクトリのためのハードウェア量が多いのみならず、ディレクトリを管理するためのオーバヘッドが非常に大きいという欠点を持つ。
たとえば、１６プロセッサ、４ＧＢ主記憶、６４Ｂ／ラインのシステムのディレクトリとしては、
４ＧＢ／６４Ｂ＊１６ｂｉｔ＝１２８ＭＢ
もの主記憶が必要になる。
従って、大幅なハードウェア量の削減が必要になる。
【００１１】
（Ｂ）パーティションのアドレスが０番地から始まらない
上記の従来のパーティション管理機構では、システム全体でアドレス空間が一つである。
従って、各パーティションの分担するアドレスが０番地から始まらない。
例えば、パーティション数が２、各パーティションの主記憶容量が１ＭＢとすると、パーティション０は０番地から１Ｍ番地までのアドレス空間を持つのに対し、パーティション１は１Ｍ番地から２Ｍ番地までのアドレス空間を持たなければならないことになる。
既存のＯＳは、アドレスが０番地から主記憶が実装されていることを前提としているため、上記の制限は従来ＯＳを使用する場合、大きな障害になる。
【００１２】
（Ｃ）パーティション管理のためのハードウェア量が多い
上記従来例のパーティション管理機構を用いる場合、４ＫＢのページ毎に、各プロセッサが該当するページへのアクセスを許されるかどうかがビットマップで記憶されている。
したがって、該当するビットマップのハードウェア量が非常に大きいという問題点がある。
例えば、プロセッサの台数を１６台、システムの主記憶容量を４ＧＢとすると、
４ＧＢ／４ＫＢ×１６＝１６ＭＢ
ものメモリがパーティション管理のために必要になり、コストの増大を招く。
【００１３】
従って、
本発明の第一の目的は、他のプロセッサから全くアクセスされていないローカルなデータ

を、他のノードへのＣＣＣを行わずに高速にアクセスすることが可能な、主記憶共有型のマルチプロセッサを、少ないハードウェアオーバヘッドで実現することである。
本発明のもう一つの目的は、主記憶共有型のマルチプロセッサをパーティションに分けた際に、各パーティションの持つローカルな主記憶が独立したアドレス空間を持つことによりローカルな主記憶を０番地から始めることができ、かつ、必要な領域を共有することができる主記憶共有型のマルチプロセッサを構成することである。
本発明のさらなる目的は、上記のパーティション管理を少ないハードウェア量で実現することである。

【００１４】

【課題を解決するための手段】

上記目的を達成するため、本発明は、
１つ以上のＣＰＵ、キャッシュと、主記憶を備える複数のノードと、ノード間を結ぶネットワークとからなり、ネットワークを用いてノード間でキャッシュコヒーレント制御を行う主記憶共有型マルチプロセッサにおいて、
各ノードは、自ノードの主記憶の各ページ対応に、該当するページが他のノードからアクセスされたかどうかを記憶する第１のビットが１ビット割り当てられ、該第１のビットはシステムの初期化時にリセットされ、主記憶の該当するページが他のノードからアクセスされた場合に、ハードウェアによりセットされるテーブルと、自ノードのＣＰＵが自ノードの主記憶をアクセスする際に、アクセスするページに該当する前記テーブルの前記第１のビットを検査し、該第１のビットがセットされていた場合には他のノードへのキャッシュコヒーレント制御を行い、該第１のビットがセットされていなかった場合には他のノードへのキャッシュコヒーレント制御を行わない手段を備えるようにしている。

【００１５】

さらに、システムソフトウェアが主記憶のページをアロケートする際に、アロケートするページに対応する前記テーブルのビットを該システムソフトウェアがリセットするようにしている。

【００１６】

また、前記テーブルに、前記主記憶の各ページ対応に該ページに対してキャッシュコヒーレント制御が必要で無いことを記憶する第２のビットを１ビット割り当て、
前記手段は、自ノードのＣＰＵが自ノードの主記憶をアクセスする際に該第２のビットを検査し、該ビットがセットされていない場合には、前記第１のビットの値にしたがって他のノードへのキャッシュコヒーレント制御の要否を判断し、前記第２のビットがセットされていた場合には他のノードへのキャッシュコヒーレント制御を行わないようにしている。

【００１７】

１つ以上のＣＰＵ、キャッシュと、主記憶を備える複数のノードと、ノード間を結ぶネットワークとからなり、ネットワークを使ってノード間でキャッシュコヒーレント制御を行い、主記憶を共有している各ノードを１つ以上のノードからなる複数のパーティションに分けることが可能な主記憶共有型マルチプロセッサにおいて、
各ノードの主記憶を、全ノードからアクセス可能な共有領域と、パーティション内からのみアクセス可能なローカル領域に分割し、各々の領域について別個の開始アドレスを指定するようにしている。

【００１８】

さらに、前記各ノードは、アクセスされたアドレスがローカル領域であるか共有領域であるかを判定する手段と、パーティション内にどのノードが含まれているかを判定する手段を備え、
他のノードへキャッシュコヒーレンス制御のためのコマンドを出すとき、共有領域へのアクセスコマンドに関しては、システム内の全ノードにコマンドをブロードキャストし、ローカル領域へのアクセスコマンドに関しては、パーティション内のノードにのみコマンドをマルチキャストするようにしている。

【００１９】
また、前記各パーティションのローカル領域のアドレスが０番地から始まるようにしている。
また、他のノードからキャッシュコヒーレントコマンドが到来した際に、アクセスアドレスがローカル領域か共有領域かを判定する手段と、アクセス元のノードがパーティション内かパーティション外かを判定する手段を備え、
ローカル領域にパーティション外のノードからコマンドが到来したと判定された場合には、アクセスを抑止し、エラーを報告するようにしている。
また、前記各ノードは前記共有領域の構成情報を記憶するレジスタを備えるようにしている。
また、前記共有領域の構成情報は、共有領域の開始アドレスと、１プロセッサの担当する共有領域の大きさからなるようにしている。
また、前記共有領域の構成情報は、共有領域の開始アドレスと終了アドレスの組からなるようにしている。
また、前記各ノードは、パーティションの構成情報を記憶する手段として、パーティション内のノードの分布をビットマップで記憶する手段を備えるようにしている。
【００２０】
【発明の実施の形態】
以下、本発明に係る主記憶共有型マルチプロセッサを、図面に示した実施の形態を参照してさらに詳細に説明する。
【００２１】
（１）装置の概要
図１は本発明に係る主記憶共有型マルチプロセッサのブロック図である。
本システムは６４ノードのシステムであり、複数のノード、例えば１００、２００（これらはノード０、ノード６３と呼ぶことがある）が、ネットワーク９００により接続される。各ノードは同じ構造を有する。
すなわち、各ノードは、ＣＰＵ１１０～１１２、（部分）主記憶１６０、主記憶アクセス回路１３０、ネットワークコマンド送信／受信回路１８０／１９０を持つ。１８０、１９０、９００については、公知の技術であるので内部の説明は省略する。
主記憶１６０は、このシステムに共通の主記憶の一部を構成し、各ノードで実行されるプログラムおよびデータの一部を保持するもので、このシステムはいわゆる分散共有メモリ型の並列計算機システムである。
主記憶は、ローカル主記憶１６１、共有主記憶１６２に分けられ、ベースアドレスレジスタ１６１０、１６２０により、それぞれ別個のアドレスを指定できる。
ベースアドレスレジスタ１６１０、１６２０はシステムの立ち上げ時に、システムのアドレスマップに応じて、後に述べるパーティション／主記憶構成情報レジスタ１５０とともにセットされる。
図ではノード内のＣＰＵはバスにより接続されているが、バス以外の結合方式、例えば一対一結合、スイッチによる結合でもよい。
これらノード内の接続方法については公知の技術であるので内部構造の詳細な説明は行わない。
【００２２】
主記憶アクセス回路１３０は、ＣＰＵから主記憶アクセス命令が発行されたときに、他のノードへのＣＣＣのためのネットワークコマンドの発行、他ノードへの主記憶アクセスコマンドの発行、自ノード内の主記憶のアクセス、および、他のノードから送られてきたＣＣＣコマンド、主記憶アクセスコマンドを実行するための回路である。
先ず、送信側の回路を説明すると、１３１はＣＰＵから送られてきたバスコマンドの受信／バスコマンドの分類を行うための回路である。
１３２はＣＰＵからアクセスされたアドレスが内部（ノード内の主記憶のアドレス）かリモート（他のノードが持つ主記憶のアドレス）かを判断するための回路である。

１３２はパーティション構成情報１５０の内容を用いて内部／リモートを判断する。
１３２、１５０は本実施の形態に特有の回路である。
【００２３】
１３８は自ノード内の主記憶の各ページの属性（他のノードからアクセスされたか否か、及び他のノードへのＣＣＣが不要であるかどうか）を、記憶するためのテーブルＲＡＴ（Ｒｅｍｏｔｅ　Ａｃｃｅｓｓ　Ｔａｂｌｅ）、１３３はＣＰＵからアクセスされたアドレスのＲＡＴの値をチェックし、必要な動作を起動するための回路、１４８は他のノードからアクセスされたページのＲＡＴの値を変更するための回路である。
これらの１３８、１３３、１４８は本実施の形態に特有の回路である。
【００２４】
１３７はＲＡＴを初期化等のためにＣＰＵからアクセスするための回路である。
１３４はＣＣＣコマンド、他ノード主記憶アクセスコマンド、他ノードへの返答コマンド等のネットワークコマンドを生成するための回路である。
１３９はアクセスアドレス、アクセスコマンドより、ネットワークコマンドをどのノードに出すかを判断する回路である。
１３９は本実施の形態に特有の動作を行う。
１３５はＣＰＵが自ノードの主記憶上のデータをアクセスする際に、他ノードへ出したＣＣＣの内容を記憶し、他のノードから帰ってきたＣＣＣに対する返答コマンドを集計するためのＣＣＣ待ち合わせ回路Ａ、１３６はＣＰＵがアクセスした自ノードの主記憶１６０の内容をアクセスするための主記憶アクセス回路Ａである。
【００２５】
次に受信側の回路を説明すると、１４１はネットワークコマンドのアドレス、送信元ノード番号と、パーティション構成情報１５０の内容の整合性をチェックするための回路であり、本実施の形態に特有の回路である。
１４２は他のノードから送られてきたネットワークコマンドの分類を行う回路、１４７は他のノードから送られてきたコマンドのアクセスアドレスが内部かリモートかを判断する回路である。１４７は、本実施の形態に特有の回路である。１４３は他のノードから送られてきたＣＣＣコマンドなどを、ノード内のバスに出力するための回路である。
１４４は、他のノードへのデータフェッチ要求の回答を集計し、ＣＰＵに返送するデータを選択するための回路である。
１４５は、他のノードが自ノード内の主記憶をアクセスした際に、自ノードのバスに出したＣＣＣコマンドの内容を記憶し、自ノードのバスからの返答を待ち合わせるためのＣＣＣ待ち回路Ｂ、１４６は、他のノードからの要求に応じて自ノード内の主記憶１６０をアクセスするための回路である。
１４９は他のノードからのアクセスコマンドが来た際に、アクセス元のノード番号を一時的に記憶するためのラッチである。返答先のノード番号を知るために用いられる。
【００２６】
（２）バス、ネットワークコマンドの説明
ノード内のバス上では、以下の６つのコマンドが使われる。括弧内はこの実施例で使われる略号である。
・Ｆｅｔｃｈ（Ｆ）
データのライン転送を要求する。
ＣＰＵの読み出しコマンドがミスした場合に出される。ＣＣＣコマンドの一つである。
・Ｆｅｔｃｈ＆Ｉｎｖａｌｉｄａｔｅ（ＦＩ）
データのライン転送と同時に、他のキャッシュ上のデータの無効化を要求する。ＣＰＵの書き込みコマンドがミスした場合に出される。ＣＣＣコマンドの一つである。
・Ｉｎｖａｌｉｄａｔｅ（Ｉ）
他のキャッシュ上のデータの無効化を要求する。
ＣＰＵが、他のキャッシュと共有されているキャッシュラインに対して書き込み要求を出した場合に出される。ＣＣＣコマンドの一つである。

・ＷｒｉｔｅＢａｃｋ（ＷＢ）
キャッシュラインの書き戻しを要求する。リプレースにより、データが追い出されたときに生じる。
・Ｄａｔａ（Ｄ）
データ転送を要求する。
Ｆ，ＦＩコマンドへの返答。
・ＮｏＤａｔａ（ＮＤ）
Ｆ，ＦＩコマンドに対し、どのＣＰＵもＤコマンドを出さない状態。
これは、バス上では明示的なコマンドとしては存在しないが、便宜的にコマンドとして扱う。
バス上では、コマンドに付随して、アドレスが転送され、さらにＷＢ，Ｄコマンドではデータ（キャッシュライン）が転送される。
【００２７】
ネットワークコマンドは以下の７種類が存在する。
バスコマンドと同じものについては意味の説明を略する。
・Ｆｅｔｃｈ（Ｆ）
・Ｆｅｔｃｈ＆Ｉｎｖａｌｉｄａｔｅ（ＦＩ）
・Ｉｎｖａｌｉｄａｔｅ（Ｉ）
・ＷｒｉｔｅＢａｃｋ（ＷＢ）
・Ｄａｔａ（Ｄ）
Ｆ，ＦＩコマンドに対し、キャッシュ上のデータを返送するためのコマンド。
・ＤａｔａＭｅｍ（ＤＭ）
Ｆ，ＦＩコマンドに対し、主記憶上のデータを返送するためのコマンド。
ここで、他のいずれかのノードのキャッシュよりＤコマンドが来た場合は、ＤＭコマンドで返送されたデータは無視されなければならない。
・ＮｏＤａｔａ（ＮＤ）
Ｆ，ＦＩコマンドに対し、該当するノードではどのＣＰＵもＤコマンドを出さないことを示す返答。
【００２８】
図６、図７にバスコマンドのフォーマットを示す。
図６はＦ，ＦＩ，Ｉ，ＮＤコマンドであり、コマンドの他に、宛先ノード番号を表すビットマップ、アクセス元ノード番号、アクセスアドレスを含む。
ここで、宛先ノード番号はビットマップで表されているため、複数のビットを立てることにより、特定の複数のノード（例えばパーティション内のノード）に向けたマルチキャスト、さらには、全ビットを立てることにより、システムの全プロセッサに向けたブロードキャストを容易に実現することができる。
図７はＷＢ，Ｄ，ＤＭコマンドのフォーマットであり、上記の情報の他にキャッシュラインのデータを含む。
【００２９】
（３）ＣＰＵのからのアクセスに対する動作
以下では、ＣＰＵからのアクセスに対する主記憶をアクセス回路１３０の動作を場合に分けて順に説明する。
共有バス上にコマンドが出されると、バスコマンド受信／バスコマンド分類回路１３１は、アクセスされたアドレスをリモート判定回路１３２に送り、アクセスされたコマンドが自ノードの主記憶のアドレスをアクセスしているか（内部）、他ノードの主記憶のアドレスをアクセスしているか（リモート）を判断する。
リモート判定回路１３２では、パーティション／主記憶構成情報１５０の内容を使用し、内部／リモートの判定を行う。
【００３０】
図５にパーティション／主記憶構成情報１５０の内容を示す。

パーティション／主記憶構成情報１５０は、パーティション内にどのノードが属しているかを示すビットマップ１５００、パーティション内のノードの数を表すレジスタ１５０６を持ち、さらに、共有領域に関しては、共有領域の先頭アドレス（ベースアドレス）１５０１、１ノードあたりの共有領域の大きさ１５０２を記憶する。
ここでは、各ノードの共有領域は、全て同じ大きさである。
さらに、ローカル領域に関しては、パーティション内の各ノードに対して、ローカル領域の開始アドレス１５０３と終了アドレス１５０４を記憶する。
パーティションレジスタ１５００の該当するビットが１であるノードに対してのみこの情報は有効である。
各ノード毎に別個のレジスタを持つことにより、ローカル領域に関しては各ノードの主記憶容量は異なっても良い。

【００３１】
リモート判定回路１３２では、アクセスアドレスをＡとすると、Ａが自ノードのローカル領域の開始アドレス１５０３と終了アドレス１５０４の間にあるかを調べる。
さらに、自ノードの番号をＮ、共有領域ベースアドレスレジスタ１５０１の内容をＢ、共有領域サイズ／ノード レジスタ１５０２の内容をＳとすると、
Ｂ＋Ｓ×Ｎ≦Ａ＜Ｂ＋Ｓ×（Ｎ＋１）
であるかどうかを調べる。
いずれかが成り立つ場合はバスコマンド受信／バスコマンド分類回路１３１に、アクセスされたアドレスが内部であることを、そうでない場合はリモートの主記憶であることを知らせる。
その後、バスコマンド受信／バスコマンド分類回路１３１は、アクセスコマンド、アクセスアドレスが内部であるかリモートであるかによって、異なった動作をする。

【００３２】
（Ａ）内部アドレスへのＦ、ＦＩコマンドの場合
まず、１３１ａを通してＲＡＴチェック回路１３３にアクセスアドレス、コマンドが送られる。
ＲＡＴチェック回路１３３では、アクセスされたページに該当するＲＡＴ１３８の内容をチェックし、他のノードへのＣＣＣが要るかどうかを判断する。

【００３３】
図３にＲＡＴの詳細を示す。
ＲＡＴでは、ページ毎にＡ，Ｎの２つのビットを記憶するメモリである。
Ａビットは他のノードへのＣＣＣが必要であるかどうかを示す。
Ａが０の場合、該当するページが他のノードからアクセスされていない（従って他のノードへのＣＣＣは不要である）ことを示し、Ａが１の場合該当するページが他のノードからアクセスされたこと（従って、他のノードへのＣＣＣが必要であること）を示す。
Ａビットは立ち上げ時に０に初期化されるほか、ソフトウェアがあるページを割り当てる際にソフトウェアにより０にリセットされる。

【００３４】
図９に本実施例におけるページ割り当てのアルゴリズムを示す。
これにより、ページがシステムソフトウェアにより回収され、再利用された場合にも、ＲＡＴが正しく動作することを保証することができる。
Ａビットはハードウェア（ＲＡＴ変更回路１４８）で１にセットされる。これについては後ほど述べる。

【００３５】
Ｎビットは通常のキャッシュコヒーレント管理が不要であるページを示すのに用いる。
Ｎビットが０の場合該当するページは通常のキャッシュコヒーレント管理の対象になることを示し、Ｎが１の場合は、該当するページは（パージ、フラッシュ等の命令を用いて）ソフトウェアによりコヒーレント管理が行われるため通常のキャッシュコヒーレント管理は不要であることを示す。

Ｎビットが１のページに対しては、他のノードへのＣＣＣのためのコマンドは出されない。
Ｉ／Ｏ領域や、数値計算の配列がおかれる領域などで有効である。
Ｎビットはアプリケーションソフトウェアが定義したアドレス情報に基づき、システムソフトウエアが管理する。
【００３６】
図２にＲＡＴチェック回路１３３の詳細を示す。
先ず、１３３３、１３３４でアクセスされたページに該当するＲＡＴの値が読み出された後、ゲート１３３５により、Ａ＝１かつＮ＝０であるかどうか（すなわち、他のノードへのＣＣＣが必要であるかどうか）を判断する。
他のノードへのＣＣＣが必要である場合、スイッチ１３３０、信号１３３ａを通してネットワークコマンド生成回路１３４に他ノードへのコマンドを出すことを依頼するとともに、他ノードへのＣＣＣが必要で、かつコマンドがＦかＦＩの場合（コマンドデコード回路１３３２およびゲート１３３６により判断する）スイッチ１３３１、信号１３３ｂを通してＣＣＣ待ち回路ＡにＣＣＣの結果を待ちあわせることを依頼する。
それに対して、他ノードへのＣＣＣが必要でない場合（Ａ＝０もしくはＮ＝１の場合）は、ゲート１３３７でコマンドがＦかＦＩかどうかが判断され、ＦかＦＩで外部へのＣＣＣが不要な場合には、スイッチ１３３８、信号１３３ｃを通じて主記憶アクセス回路Ａ１３６に主記憶の内容を読み出すことを依頼する。
すなはち、外部へのＣＣＣを省略し、即座に主記憶をアクセスできる。
【００３７】
（Ａ１）他のノードへのＣＣＣが不要の場合
この場合、信号１３３ｃを通じて主記憶アクセス回路Ａ１３６にアクセスアドレスが伝えられる。
主記憶アクセス回路Ａ１３６はアクセスされたデータラインを自ノード内の主記憶１６０から読み出し（ローカル領域１６１、共有領域１６２の何れの場合も同じである）、読み出されたデータを１３６ａ、１４３を通じて自ノード内のＣＰＵに返す。
【００３８】
（Ａ２）他のノードへのＣＣＣが必要な場合
先ず、信号１３３ａを通じてネットワークコマンド生成回路１３４に他のノードにＦ，ＦＩコマンドを送出することを依頼する。
それと同時に信号１３３ｂを通じてＣＣＣ待ち回路Ａ１３５に対して他のノードからの返答を待ち合わせることを依頼する。
ネットワークコマンド生成回路１３４は、宛先生成回路１３９により判断されたノードにコマンドを送付する。
ここで、注意しなければならないのは、宛先ノードは図６、図７に示すように全ての（６４個の）ノードに対応するビットマップで表されるため、複数のノードを指定すればブロードキャスト、マルチキャストを指定することができる。
【００３９】
図１１に宛先生成回路１３９の構成を示す。
宛先生成回路では、先ず、アクセスアドレスが内部であるかリモートであるか（リモート判定回路１３９１）が判断され、さらに、アクセスアドレスが共有メモリであるかどうかが（共有メモリ判定回路１３９２）が判断される。
１３９１の動作は１３２と同一である。
共有メモリ判定回路１３９２では、アクセスアドレスをＡ、共有領域ベースアドレスレジスタ１５０１の内容をＢ、共有領域サイズ／ノード レジスタ１５０２の内容をＳとすると、
$B \leqq A < B + S \times 64$
で有るかどうかが調べられる（アクセスアドレスが上記の範囲内に入っていれば共有メモリである）。

さらに、回路１３９４により、アクセスアドレスのｈｏｍｅノードが求められる。ｈｏｍｅノードとは、アクセスアドレスを主記憶上に持つノードである。
具体的には以下の手順で求められる（以下ではアクセスアドレスをＡとする）。
・Ａと全てのノードに対応するローカル領域レジスタ１５０３、１５０４の値が比較され、Ａが何れかのノードＨＬの開始アドレスと終了アドレスの間に有る場合、ＨＬがｈｏｍｅノードである（Ａはローカル領域である）。
・さらに、共有領域ベースアドレスレジスタ１５０１の内容をＢ、共有領域サイズ／ノード レジスタ１５０２の内容をＳとすると、以下の計算を行う。
ＨＳ＝（Ａ－Ｂ）／Ｓ　（小数点以下切り捨て）
ＨＳが０以上６４未満の場合、ＨＳがｈｏｍｅノード番号である（Ａは共有領域である）。

【００４０】
セレクタ１３９０は、上記の判断結果、及び、アクセスコマンドにより、図１２に示す動作をする。
つまり、コマンドがＦ、ＦＩ、Ｉの場合は、共有領域の場合は全ノードへブロードキャストするためオール１のビットマップ１３９３ａが選ばれ（ここでは説明を簡単にするため６４個全ノードが実装されているとするが、実装されているノードが６４未満でも、ビットマップを変えることにより対応できる）、ローカル領域の場合は、パーティション内のノードにのみマルチキャストするため、パーティションレジスタ１５００の内容が選ばれる。
Ｄ、ＮＤコマンドの場合、及びリモート主記憶からのＤＭコマンドの場合、アクセス元ノード番号レジスタ１４９の値をデコードした（ビットマップにした）結果が選ばれ、アクセス元のノードに結果が返送される。
リモートの主記憶へのＷＢコマンドの場合は、回路１３９４で求められたｈｏｍｅノード番号をデコードした結果が選ばれ、データをｈｏｍｅノードに書き戻す。
内部のアドレスへのＷＢ、ＮＤコマンドはエラーである（あり得ないはずである）。
この場合、Ｆ，ＦＩコマンドであるので、ローカル領域の場合はパーティション内の全ノードにマルチキャストされ、、共有領域の場合はシステム内の全ノードにブロードキャストされる。
他ノードへＦ，ＦＩコマンドを出した結果は、ＤもしくはＮＤコマンドにより、返送される。
他のノードより返送されたコマンドはネットワークコマンドチェック回路１４１でパーティション／主記憶構成情報１５０との整合性をチェックされた後、ネットワークコマンド分類回路１４２に送付される。

【００４１】
ネットワークコマンド分類回路１４２は、コマンドの種類、アクセスされたアドレスが内部（自ノードの主記憶）かリモート（他ノードの主記憶）か（リモート判定回路１４７を用いてを判断される、リモート判定回路１４７の動作は１３２と全く同一である）に応じ、該当する出力にコマンドを出す。
内部アドレスへのＮ、ＮＤコマンドの場合は、信号１４２ａを通じてＣＣＣ待ち回路Ａ（１３５）にコマンドを送る。

【００４２】
図１４にネットワークコマンドチェック回路１４１の詳細を示す。
ネットワークコマンドチェック回路１４１では、ローカル領域へのパーティション外のノードへのアクセスをエラーとして検出するための回路である（ローカル領域を他のパーティションからの不正なアクセスから守るために、ローカル領域へはパーティション内のノードからのみしかアクセスが許されない）。
まず、アクセスアドレスが共有メモリか否かが共有メモリ判定回路１４１０により判断される（共有メモリ判定回路１４１０の動作は、１３９２と同一である）。
さらに、パーティション内外判定回路１４１１によりアクセス元のノードがパーティショ

ンの中であるかどうかが判断される。
具体的には、ネットワークコマンドのアクセス元ノード番号をデコーダ１４１２でデコードした結果と、パーティションレジスタ１５００（パーティション内のノードを表すビットマップ）をＡＮＤ－ＯＲゲートでゲートした結果により、アクセス元のノード番号に相当するビットがパーティションレジスタ１５００の中で立っているかどうかがチェックされる。
その結果、正しいアクセスの場合、つまりアクセスアドレスが共有メモリであるか、アクセス元のノードがパーティション内である場合は、スイッチ１４１５を通じてアクセスコマンドがネットワークコマンド分類回路１４２に送られる。
不正なアクセスの場合は、ゲート１４１６により、エラーが報告される（スイッチ１４１５によりアクセスは抑止される）。
これにより、ローカル領域への他のパーティションからの不正なアクセスを防止することができる。
【００４３】
内部のアドレスに対して他のノードへＦ，ＦＩコマンドを出した結果としては、全てのノードからＮＤが返る場合と、１つ以上のノードからＤコマンドが返る（それ以外のノードからはＮＤが返る）場合に分けられる。
他のノードからの返答はＣＣＣ待ち回路Ａ（１３５）で集計され、該当する動作がとられる。
【００４４】
図１０にＣＣＣ待ち回路Ａ（１３５）の詳細を示す。
信号１３３ｂを通じて与えられたＣＣＣアドレスはレジスタ１３５０６に蓄えられる。
後に他のノードからＣＣＣの返答（ＤもしくはＮＤコマンド）が来ると、先ず、コンパレータ１３５０７でＣＣＣアドレス１３５０６と比較される。
比較した結果が一致した場合は、コマンドがＤであるかＮＤであるかが、デコーダ１３５００とゲート１３５０１、１３５０２で判断される。
Ｄコマンドが（１つでも）到来した場合（つまり他のキャッシュから最新のデータが送られて来た場合）は、フリップフロップ（ＦＦ）１３５０４がセットされ、Ｄコマンドが到来したことが記憶されると同時に、ラッチ１３５１４にＤコマンドの内容が記憶される（Ｄコマンドにより送られてきた値は、後にＣＰＵに返される）。
それとは別に、Ｄ、ＮＤコマンドが１個到来する毎に、ゲート１３５０３を通じてカウンタ１３５０５がカウントアップされ、今までに何個の返答が来たかが数えられる。
カウンタの値が信号１３５１１ａで示される、期待される返答数と一致すると、全ての返答が返ったとして、ＣＣＣの待ち合わせが終了する。
【００４５】
その際にフリップフロップ１３５０４が１の場合、スイッチ１３５１５がＯＮになり、ラッチ１３５１４に記憶されていたＤコマンドにより返送されたデータが信号１３５ｂ、バスコマンド生成回路１４３を通じてＣＰＵに返される。
それに対して、フリップフロップ１３５０４が０の場合（全てのノードからＮｏＤａｔａコマンドが返された場合）は、スイッチ１３５１６がＯＮになり、信号１３５ａを通じて、ＣＣＣアドレスレジスタ１３５０６に記憶されていたアクセスアドレスが主記憶アクセス回路Ａ（１３６）に伝えられ、主記憶上のデータをアクセスすることが依頼される。
【００４６】
信号１３５１１ａで示される期待される返答数は、次の手順で求められる。
まず、ＣＣＣアドレスレジスタ１３５０６で記憶されているアクセスアドレスが共有主記憶であるかどうかが、共有メモリ判定回路１３５０８で判断され（共有メモリ判定回路１３５０８の動作は１３９２と同一である）、セレクタ１３５１１に伝えられる。
セレクタ１３５１１は、アクセスアドレスが共有メモリの場合は６４（システムの全ノード数、ただし、実装されているノード数が６４未満の場合は実装されているノード数を指定する）を出力し、アクセスアドレスが共有メモリでない（ローカルメモリの場合）はパ

ーティション内ノード数レジスタ１５０６の値を出力する。
これにより、該当するデータをアクセスするノードの数を求めることができる。
【００４７】
（Ａ２ａ）他のノードからの回答が全てＮＤの場合
ＣＣＣ待ち回路Ａ（１３５）より信号１３５ａを通じて主記憶アクセス回路Ａ（１３６）に、自ノード上の主記憶１６０をアクセスすることが依頼され、アクセスされたデータがＣＰＵに返される。
【００４８】
（Ａ２ｂ）何れかのノードよりＤコマンドが返ってきた場合
ＣＣＣ待ち回路Ａ（１３５）より、信号１３５ｂ、バスコマンド生成回路１４３を通じて、Ｄコマンドにより返送された最新のデータが、ＣＰＵに返される。
【００４９】
（Ｂ）内部アドレスへのＩコマンドの場合
この場合、ＲＡＴチェック回路１３５により、他のノードへのＣＣＣの要不要がチェックされるところまでは（Ａ）と同様である。
【００５０】
（Ｂ１）他のノードへのＣＣＣが必要な場合
信号１３３ａを通じてネットワークコマンド生成回路１３４に他のノードへのＩコマンドの生成を依頼する（Ｉコマンドに返答はない）。
ネットワークコマンド生成回路１３４は宛先生成回路１３９の指定するノードにコマンドを送る。
つまり、ローカル領域へのＩコマンドの場合、パーティション内の全ノードにマルチキャストされ、共有領域へのＩコマンドの場合、システムの全ノードにブロードキャストされる。
【００５１】
（Ｂ２）他のノードへのＣＣＣが不要な場合
Ｉコマンドの場合、主記憶のデータアクセスは不要なので何も起こらない。
すなわち、コマンドデコード回路１３３２は０を出力するため、ゲート１３３７の出力は０であり、信号１３３ｃには何も出力されない。
【００５２】
（Ｃ）内部アドレスへのＷＢコマンドの場合
この場合、バスコマンド受信／バスコマンド分類回路１３１は信号１３１ｃを通じて主記憶アクセス回路１３６に、書き戻されたデータを主記憶１６０に書き込むことを依頼する。
【００５３】
（Ｄ）リモートアドレスへのＦ，ＦＩコマンドの場合
この場合、信号１３１ｂを通じてネットワークコマンド生成回路に、他ノードへＦ，ＦＩコマンドを生成する事を要求する。
ネットワークコマンド生成回路１３４は宛先生成回路１３９の指定するノードにコマンドを送る。
つまり、ローカル領域へのＦ，ＦＩコマンドの場合、パーティション内の全ノードにマルチキャストされ、共有領域へのＦ，ＦＩコマンドの場合、システムの全ノードにブロードキャストされる。
リモートアドレスへのＦ，ＦＩに対しては、Ｄ，ＤＭ，ＮＤの３種のコマンドが到来する。
ネットワークコマンド分類回路１４２は、リモートアドレスに対するＤ，ＤＭ，ＮＤコマンドが到来した場合は、信号１４２ｅを通じて返答集計回路１４４に送出する。
【００５４】
この場合、返答されるコマンドの組み合わせとしては、
（あ）ｈｏｍｅノードからＤＭコマンドが、その他のノードからは全てＮＤが返る場合

（い）ｈｏｍｅノードからＤＭコマンドが、他の何れか一つ以上のノードからＤコマンドが返る（その他のノードからはＮＤコマンドが返る）場合
（う）ｈｏｍｅノードを含む何れか１つ以上のノードからＤコマンドが返る（その他のノードからはＮＤコマンドが返る）場合
に分けられる。
（あ）ではＤＭコマンドより送られたｈｏｍｅノードの主記憶から読み出された値が使われるのに対し、（い）では、他のノードのキャッシュの上のデータを優先しなければならず（なぜならキャッシュ上のデータは変更されている可能性があるため）、ＤＭコマンドでｈｏｍｅノードの主記憶から送られたデータは捨てられ、Ｄコマンドによりキャッシュから送られたデータが活用される。
（う）では、Ｄコマンドにより何れかのノードのキャッシュから送られてきたデータが用いられる。
各ノードからの返答は返答集計回路１４４で集計され、（あ）（い）（う）の何れかの場合であるかが判断され、アクセスされた結果（データライン）がアクセス元のＣＰＵに返される。

【００５５】

図１３に返答集計回路１４４の詳細を示す。
先ず、他ノードから送られてきたコマンドが１４４００によりデコードされる。
Ｄコマンドの場合はラッチ１４４０１によりＤコマンドの内容（アドレス、データ）が記憶されるとともに、フリップフロップ１４４０３によりＤコマンドが１つ以上到来したことが記憶される。
ＤＭコマンドの場合はラッチ１４４０２によりＤＭコマンドの内容が記憶される。それと並行して、Ｄ、ＤＭ、ＮＤコマンドが１つ到来する毎に、ゲート１４４０４を通じてカウンタ１４４０５がカウントアップされ、到着した返答の数を数える。
カウンタの値が信号１４４１３ａで示される期待される返答数と一致した場合（コンパレータ１４４０６で判断される）、全ての返答が到着したと判断し、Ｄコマンドが一つでも来ていた場合（フリップフロップ１４４０３が１の時）はゲート１４４０８、スイッチ１４４１０によりＤコマンドの内容が、信号１４４ａを通じてバスコマンド生成回路１４３に供給される。
それに対しＤコマンドが一つも来ていなかった場合は、ゲート１４４０７、スイッチ１４４０９により、ＤＭコマンドの内容がバスコマンド生成回路に供給される。
これにより、Ｄコマンドが一つ以上来ていた場合は、Ｄコマンドにより返送された他のノードのキャッシュ上のデータがアクセスを行ったＣＰＵに返され、Ｄコマンドが一つも来ていなかった場合は、ＤＭコマンドにより返送された、ｈｏｍｅノードの主記憶上の値がアクセスを行ったＣＰＵに返される。
１４１１、１４１３等の期待される返答数を求める回路は、アドレスラッチ１４４１４にラッチされたアクセスアドレスに対して、ＣＣＣの返答の数を求める。
詳細な動作は、ＣＣＣ待ち回路Ａ（１３５）の回路（１３５０８、１３５１１等）と全く同一である。

【００５６】

（Ｅ）リモートアドレスへのＩコマンドの場合
この場合はＲＡＴのチェックは行われず、信号１３１ｂにより直接ネットワークコマンド生成回路１３４に他のノードへのコマンドの送出の依頼が行われる。
その後の動作は（Ｂ１）と同様である。

【００５７】

（Ｆ）リモートアドレスへのＷＢコマンドの場合
この場合、バスコマンド受信／バスコマンド分類回路１３１は信号１３１ｂを通じてネットワークコマンド生成回路１３４に、ＷＢコマンドを送付することを依頼する。
宛先生成回路が出す宛先は、ｈｏｍｅノードであり、ＷＢコマンドはｈｏｍｅノードに送られる。

【００５８】
（４）他のノードからのアクセスに対する動作
ここでは、主記憶アクセス回路１３０が、ネットワーク９００を通じて送られてきた、他のノードからのコマンドに対してどのように動作するかを述べる。
ここで、Ｄ，ＤＭ，ＮＤコマンドに対する動作は既に（３）で述べてあるので、その他のＦ，ＦＩ，Ｉ，ＷＢに対する動作を述べる。
他のノードから送られてきたコマンドは、ネットワークコマンドチェック回路１４１によりチェックされた後、ネットワークコマンド分類回路１４２に送られる。
ネットワークコマンド分類回路１４２では、アクセスコマンドの種類、アクセスアドレスが内部かリモートか（リモート判定回路１４７により判断される）により、該当する出力にコマンドを送る。
また、Ｆ，ＦＩ，Ｉコマンドの場合は、アクセス元ノード番号レジスタ１４９にアクセス元ノード番号をセットする。
【００５９】
（Ａ）内部アドレスへのＦ，ＦＩコマンドの場合
ネットワークコマンド分類回路１４２は、信号１４２ｄを通じてバスコマンド生成回路１４３にバスコマンドの生成を依頼する。
バスコマンド生成回路１４３は、ノード内の共有バス１２０を通じてノード内のＣＰＵにＦ，ＦＩコマンドを出す。
それと同時に、信号１４２ｃを通じてＣＣＣ待ち回路Ｂ（１４５）に、ノード内の共有バスに出されたコマンドを待ち合わせることを依頼する。
【００６０】
図１５にＣＣＣ待ち回路Ｂ（１４５）の詳細を示す。
信号１４２ｃを通じて送られてきた、ＣＣＣを行うアドレスは、ＣＣＣアドレスレジスタ１４５１に記憶される。
後にノード内の共有バス１２０からバスコマンド受信／バスコマンド分類回路１３１を通じて、ＣＣＣの返答（ＤもしくはＮＤコマンド）が来ると、先ず、コンパレータ１４５２でＣＣＣアドレス１４５１と比較される。
比較した結果が一致した場合は、コマンドがＮＤであるかどうかが、デコーダ１４５０とゲート１４５３で判断される。
ＮＤコマンドが返送された場合は、スイッチ１４５５を通じて、ＣＣＣアドレスレジスタ１４５１に記憶されていたアクセスアドレスが、信号１４５ａを通して主記憶アクセス回路Ｂ１４６に伝えられ、主記憶上のデータをアクセスすることが依頼される。
【００６１】
ノード内の共有バスに出したＦ，ＦＩコマンドの結果は、Ｄ（ノード内のいずれかのＣＰＵのキャッシュに該当するデータがある場合）、もしくはＮＤ（ノード内の何れのＣＰＵのキャッシュにも該当するデータが無い場合）により返送される。
Ｄコマンドの場合は、データも同時に返送される。
バスコマンド受信／バスコマンド分類回路１３１は、コマンド、および、アクセスアドレスが内部であるか、リモートであるか（リモート判定回路１３２により判定される）に応じて、該当する出力にコマンドを送る。
【００６２】
（Ａａ）内部アドレスへのアクセスに対し、Ｄコマンドが返送された場合
自ノードのキャッシュ上にアクセスデータがあった場合、バスコマンド受信／バスコマンド分類回路１３１は、Ｄコマンド（コマンド、アドレス、データ）を、信号１３１ｂを通じて、ネットワークコマンド生成回路１３４に送る。
ネットワークコマンド生成回路１３４は、アクセス元のノードにアクセスデータをＤコマンドを用いて返送する（宛先生成回路１３９は、アクセス元ノード番号１４９を選択する）。
これにより、自ノードのＣＰＵのキャッシュ上のデータがアクセス元のノードに返送され

る。
【００６３】
（Ａｂ）内部アドレスへのアクセスに対し、ＮＤコマンドが返送された場合。
【００６４】
自ノードのキャッシュ上にはアクセスデータが無かった場合、バスコマンド受信／バスコマンド分類回路１３１は、ＮＤコマンドを信号１３１ｄを通じてＣＣＣ待ち回路Ｂ（１４５）に送付する。
ＣＣＣ待ち回路Ｂは信号１４５ａを通じて、主記憶アクセス回路Ｂ（１４６）に主記憶１６０上のアクセスアドレスの内容（キャッシュライン）を読み出すことを依頼する。
主記憶アクセス回路Ｂ（１４６）は、信号１６０ａを通じてアクセスラインを読み出すと、信号１４６ａを通じてネットワークコマンド生成回路１３４に送る。
ネットワークコマンド生成回路１３４は、アクセス元のノードにアクセスデータをＤＭコマンドを用いて返送する（宛先生成回路１３９は、アクセス元ノード番号１４９を選択する）。
これにより、自ノードの主記憶上のデータがアクセス元のノードに返送される。
それと同時に、主記憶アクセス回路Ｂ（１４６）は、信号１４６ｂを通じてＲＡＴ変更回路１４８にアクセスされたページに該当する、ＲＡＴ１３８のＡビットに１を設定することを依頼する。
【００６５】
図４にＲＡＴ変更回路１４８の詳細を示す。
アクセスされたアドレスのページ番号に対応するＲＡＴのＡビットに１が書き込まれる。
【００６６】
（Ｂ）内部アドレスへのＩコマンドの場合
ネットワークコマンド分類回路１４２は、信号１４２ｄを通じてバスコマンド生成回路１４３にバスコマンドの生成を依頼する。
バスコマンド生成回路１４３は、ノード内の共有バス１２０を通じてノード内のＣＰＵにＩコマンドを出す（Ｉコマンドには返答は無い）。
【００６７】
（Ｃ）内部アドレスへのＷＢコマンドの場合
ネットワークコマンド分類回路１４２は、信号１４２ｂを通じて、主記憶アクセス回路Ｂ（１４６）に、ＷＢされたデータを主記憶に書き込むことを依頼する。
主記憶アクセス回路Ｂ（１４６）は、信号１６０ａを通じて主記憶１６０へＷＢされたデータを書き込む。
それと同時に、主記憶アクセス回路Ｂ（１４６）は、信号１４６ｂを通じてＲＡＴ変更回路１４８にアクセスされたページに該当する，ＲＡＴ１３８のＡビットに１を設定することを依頼する。
【００６８】
（Ｄ）リモートアドレスへのＦ，ＦＩコマンドの場合
ネットワークコマンド分類回路１４２は、信号１４２ｄを通じてバスコマンド生成回路１４３にバスコマンドの生成を依頼する。
バスコマンド生成回路１４３は、ノード内の共有バス１２０を通じてノード内のＣＰＵにＦ，ＦＩコマンドを出す。
ノード内の共有バスに出したＦ，ＦＩコマンドの結果は、Ｄ（ノード内のいずれかのＣＰＵのキャッシュに該当するデータがある場合）、もしくはＮＤ（ノード内の何れのＣＰＵのキャッシュにも該当するデータが無い場合）により返送される。
Ｄコマンドの場合は、データも同時に返送される。
【００６９】
リモートアドレスに対するＤ，ＮＤコマンドの場合、何れの場合も、バスコマンド受信／バスコマンド分類回路１３１は、Ｄコマンド（コマンド、アドレス、データ）もしくはＮＤコマンド（コマンド、アドレス）を、信号１３１ｂを通じて、ネットワークコマンド生

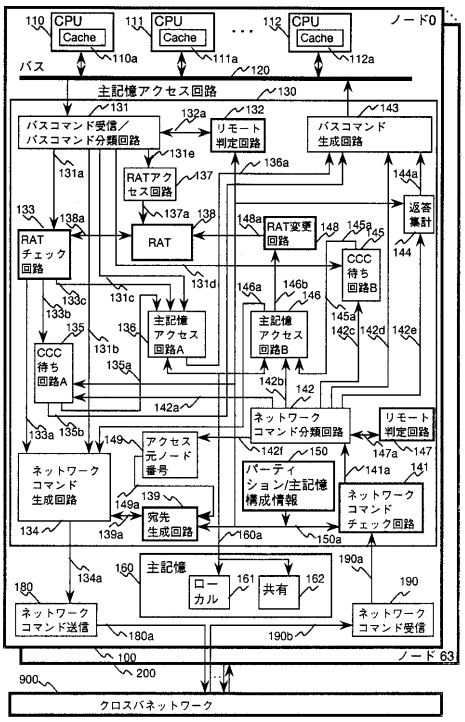
成回路１３４に送る。
ネットワークコマンド生成回路１３４は、アクセス元のノードにアクセス結果をＤ又はＮＤコマンドを用いて返送する（宛先生成回路１３９は、アクセス元ノード番号１４９を選択する）。
これにより、Ｄコマンドの場合、自ノードのＣＰＵのキャッシュ上のデータがアクセス元のノードに返送され、ＮＤコマンドの場合、自ノードのＣＰＵのキャッシュ上には該当するデータが無いことが伝えられる。
【００７０】
（Ｅ）リモートアドレスへのＩコマンドの場合
この場合、内部アドレスへのＩコマンドの場合（Ｂ）と全く同じ動作を行う。
【００７１】
（Ｆ）リモートアドレスへのＷＢコマンドの場合
他のノードからリモートアドレスへのＷＢコマンドが来ることはあり得ない。
【００７２】
ネットワークコマンド分類回路１４２はエラーを報告する。
【００７３】
以上述べた手順によりクロスバネットワークにより接続されたノード間でキャッシュコヒーレンスをとることができる。
その際に、リモートアクセステーブル（ＲＡＴ）１３８を用いて他のノードへのＣＣＣを削減することができる。
さらに、主記憶共有マルチプロセッサ内をパーティションに分ける際に、主記憶を、各パーティションのローカル領域／パーティション間共有領域の２つに分け、ローカル領域に対しては各パーティション間で独立したアドレス空間とすることにより、各パーティションのアドレスを０番地より始めるようにすることができる。
さらに、パーティション／主記憶構成情報１５０をレジスタにより記憶することにより、少ないハードウェア量でパーティションの管理を実現することができる。
【００７４】
＜変形例＞
本発明は以上の実施の形態に限定されるのではなくいろいろの変形例にも適用可能である。
例えば、
（１）以上においては、ノード内のＣＰＵはバス１２０により接続されているが、その他の接続形態（スイッチによる接続、主記憶アクセス回路１３０への一対一接続）も可能である。
（２）また、以上においては、ＲＡＴ１４８は、主記憶アクセス回路１３０に内蔵される、専用メモリにより、構成されているが、外付けのメモリにすることも可能である。
さらに、ＲＡＴを物理的にはローカル主記憶１６１上に置くことも可能である。
さらに、この場合、ＲＡＴの内容を主記憶アクセス回路１３０内にキャッシングする事も可能である。
【００７５】
（３）上記実施の形態において、パーティション／主記憶構成情報１５０では、共有領域の構成情報は、共有領域ベースアドレス１５０１及び、共有領域サイズ／ノード レジスタ１５０２の２つのレジスタにより記憶されている。
しかし、ローカル領域レジスタ１５０３、１５０４のように、ノード毎に開始アドレス、終了アドレスの組で覚えることも可能である（これにより、各ノードの共有領域の大きさを変えることができる）。
この場合、上記実施の形態においては共有領域ベースアドレス１５０１及び、共有領域サイズ／ノードレジスタ１５０２を用いて計算で求めていた、共有メモリ判定回路、ｈｏｍｅノード判定回路は、コンパレータにより構成することができる。
【００７６】

（４）以上においては、ノード内の各ＣＰＵ（１１０～１１２）は独立したキャッシュを持っているが、複数のＣＰＵに共有される外付けの２次キャッシュを持たせることも可能である。
各ノードの主記憶アクセス回路１３０に、ＣＰＵのキャッシュＴＡＧのコピーを持ち、他のノードから到来するキャッシュコヒーレントトランザクションをフィルタリングする事も可能である。
【００７７】
（５）以上においては、ノード間はクロスバネットワークにより接続されているが、他の形式のネットワーク（多段網等）により接続することも可能である。
（６）上記実施の形態においては、他のノードへのネットワークコマンドを出す際に、宛先のノードをビットマップにより指定することにより、ブロードキャスト、マルチキャストを指示しているが、ネットワークコマンド生成回路１３４が、宛先のノード毎に複数のコマンドを出すことにより、ブロードキャスト、マルチキャストを実現することも可能である。
【００７８】
【発明の効果】
本発明によれば、主記憶共有型のマルチプロセッサにおいて、リモートアクセステーブル（ＲＡＴ）を置くことにより、少ないハードウェア量の追加により、自ノード内の主記憶をアクセスする際に、他のノードからアクセスされていないページに関しては、ノード間のＣＣＣを省くことができる。
従って、アクセスレーテンシを削減するとともに、ノード間のネットワークにかかるトラフィックを削減することができる。
さらに、本発明によれば、主記憶共有マルチプロセッサ内を複数のパーティションに分ける際に、主記憶をパーティション毎のローカルな領域とパーティション間で共有される領域に分け、ローカル領域に対しては各ノードのアドレス空間を独立することにより、各パーティションの開始アドレスを０番地から始めることを可能にすることができる。
さらに、各パーティションのローカル領域、共有領域の範囲をレジスタにより記憶することにより、パーティション管理のためのハードウェア量を従来のページ毎の管理と比較して大幅に削減することができる。
【図面の簡単な説明】
【図１】本発明のキャッシュコヒーレンス保持機構を持つ主記憶共有型マルチプロセッサである。
【図２】各ノードのＲＡＴチェック回路のブロック図である。
【図３】各ノードのＲＡＴのブロック図である。
【図４】各ノードのＲＡＴ変更回路のブロック図である。
【図５】各ノードにおいて、パーティション／主記憶構成情報を表すレジスタ群の詳細図である。
【図６】ネットワーク上の、Ｆ，ＦＩ，Ｉ，ＮＤコマンドのパケットフォーマットである。
【図７】ネットワーク上の、ＷＢ，Ｄ，ＤＭコマンドのパケットフォーマットである。
【図８】各ノードの主記憶のブロック図である。
【図９】本発明のマルチプロセッサシステムにおいて、主記憶のページをアロケートする際のフロー図である。
【図１０】各ノードのＣＣＣ待ち回路Ａのブロック図である。
【図１１】各ノードの宛先生成回路のブロック図である。
【図１２】各ノードの宛先生成回路内の、宛先セレクタの入力と出力の関係の表を示す図である。
【図１３】各ノードの返答集計回路のブロック図である。
【図１４】各ノードのネットワークコマンドチェック回路のブロック図である。
【図１５】各ノードのＣＣＣ待ち回路Ｂのブロック図である。

【符号の説明】
１００、２００　ノード
１１０～１１２　ＣＰＵ
１１０ａ～１１２ａ　キャッシュ
１２０　バス
１３１　バスコマンド受信／バスコマンド分類回路
１３２、１４７　リモート判定回路
１３３　ＲＡＴチェック回路
１３４　ネットワークコマンド生成回路
１３５　ＣＣＣ待ち回路Ａ
１３６　主記憶アクセス回路Ａ
１３７　ＲＡＴアクセス回路
１３８　ＲＡＴ（Ｒｅｍｏｔｅ　Ａｃｃｅｓｓ　Ｔａｂｌｅ）
１３９　宛先生成回路
１４１　ネットワークコマンドチェック回路
１４２　ネットワークコマンド分類回路
１４３　バスコマンド生成回路
１４４　返答集計回路
１４５　ＣＣＣ待ち回路Ｂ
１４６　主記憶アクセス回路Ｂ
１４８　ＲＡＴ変更回路
１４９　アクセス元ノード番号レジスタ
１５０　パーティション／主記憶構成情報
１６０　主記憶
１６１　ローカル主記憶
１６２　共有主記憶
１８０　ネットワークコマンド送信回路
１９０　ネットワークコマンド受信回路
９００　ノード間クロスバネットワーク

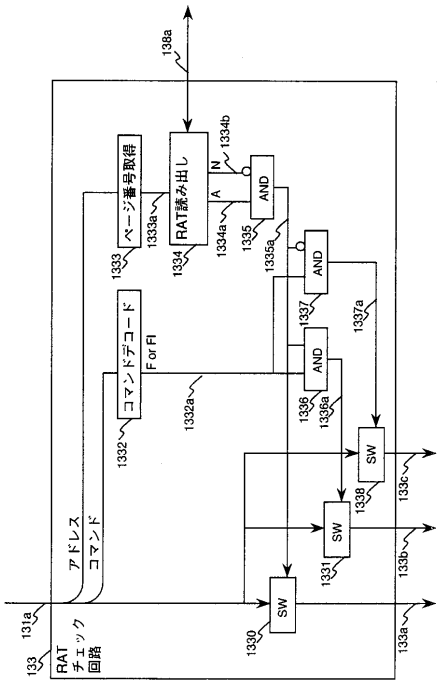
【図１】

（図１）



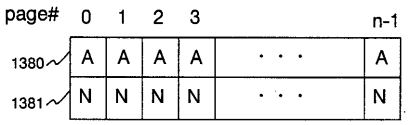
【図２】

（図２）



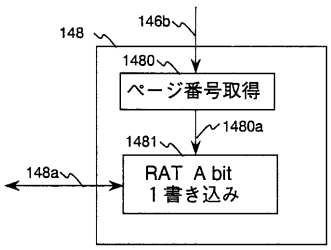
【図３】

（図３）



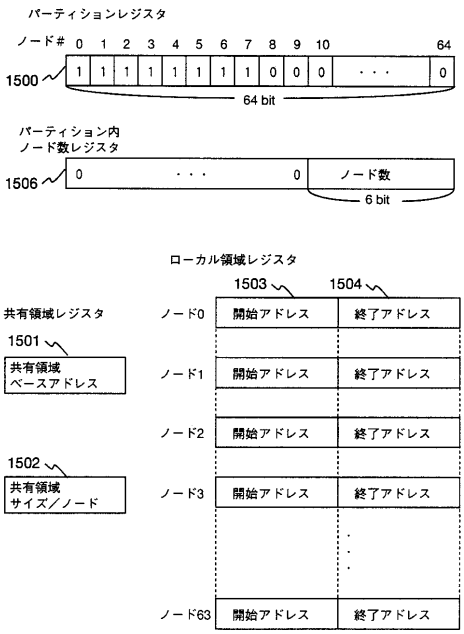
【図４】

（図４）



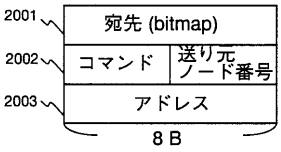
【図５】

（図５）



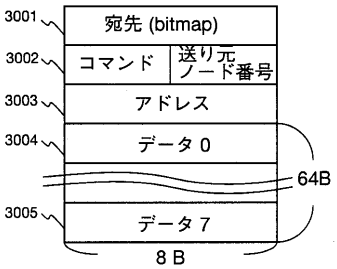
【図６】

(図６)



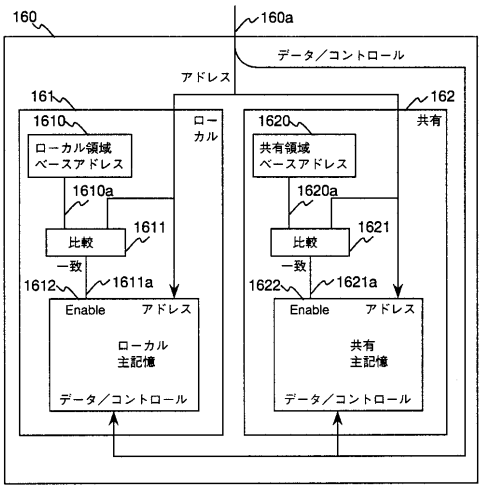
【図７】

(図７)



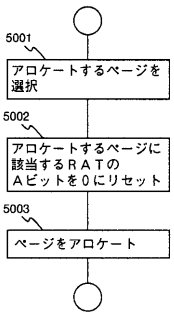
【図８】

(図８)



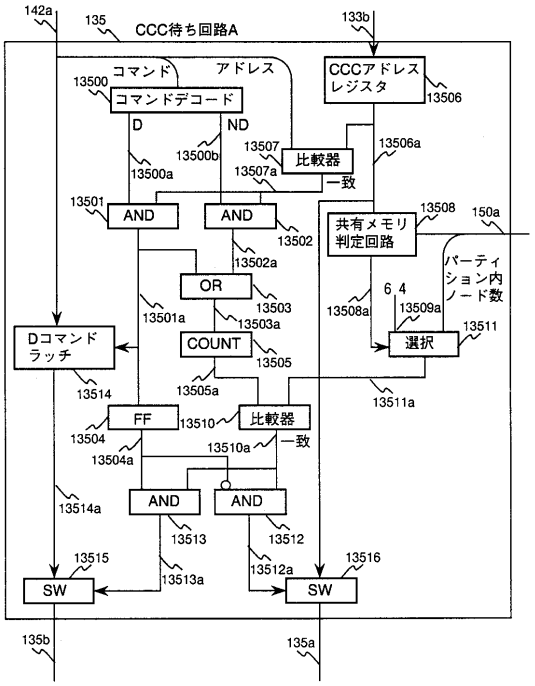
【図９】

(図９)



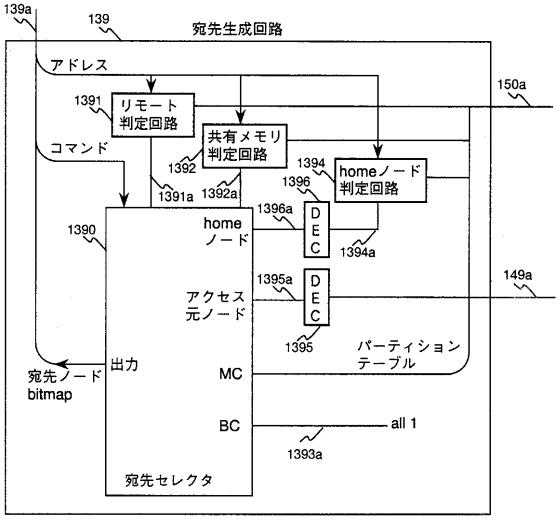
【図１０】

(図１０)



【図１１】

（図１１）



【図１２】

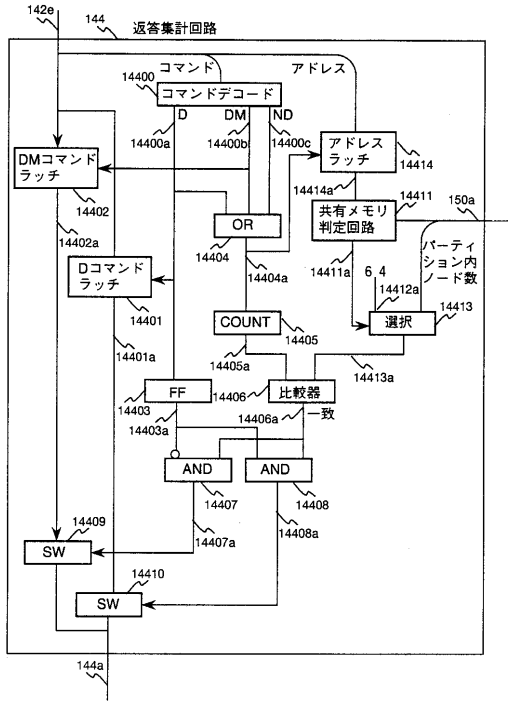
（図１２）

宛先セレクタの出力

アクセス コマンド	アクセスアドレス	
	内部	リモート
F	アクセスアドレスに応じて分岐 ・ローカル領域→MC ・共有領域 →BC	
FI		
I		
WB		homeノード
D、DM	アクセス元ノード	
ND		アクセス元ノード

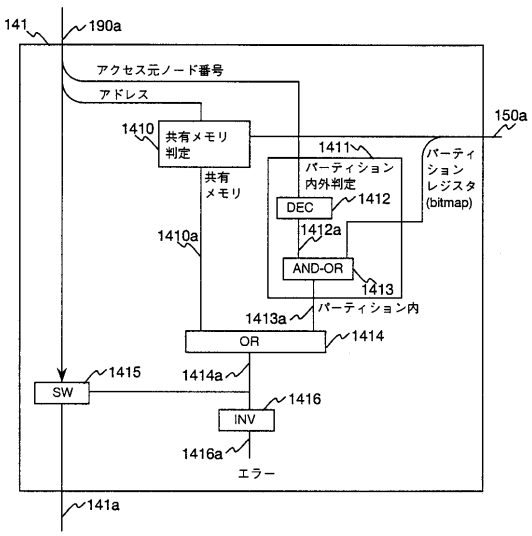
【図１３】

（図１３）



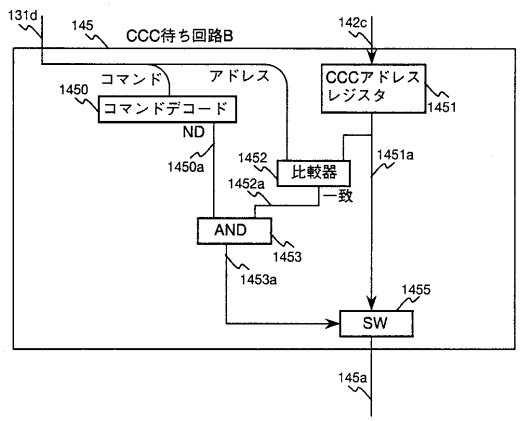
【図１４】

（図１４）



【図１５】

（図１５）


131d
145
CCC待ち回路B
142c
コマンド
アドレス
CCCアドレス
レジスタ
1451
1450
コマンドデコード
ND
1450a
1452
比較器
1451a
1452a
一致
AND
1453
1453a
1455
SW
145a

フロントページの続き

(72)発明者　庄内　　亨
東京都国分寺市東恋ケ窪一丁目２８０番地　株式会社日立製作所　中央研究所内
(72)発明者　大河内　俊夫
東京都国分寺市東恋ケ窪一丁目２８０番地　株式会社日立製作所　中央研究所内
(72)発明者　明石　英也
東京都国分寺市東恋ケ窪一丁目２８０番地　株式会社日立製作所　中央研究所内

審査官　鳥居　稔

(56)参考文献　特開平０１－１０９４６４（ＪＰ，Ａ）
特開平０６－２７４４６１（ＪＰ，Ａ）
特開平０７－２１０５２０（ＪＰ，Ａ）
特開平０７－１６０５８１（ＪＰ，Ａ）
特開平０８－０１６４６９（ＪＰ，Ａ）
特開平０８－０１６４７０（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
G06F 12/08
G06F 15/16-177