(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第3553634号
(P3553634)
(45)発行日　平成16年8月11日(2004.8.11)　(24)登録日　平成16年5月14日(2004.5.14)

| (51)Int.Cl.[7] | F I | |
|---|---|---|
| G06F 13/00 | G06F | 13/00　353B |
| G06F 15/177 | G06F | 15/177　670F |
| H04L 12/66 | H04L | 12/66 |

請求項の数 12　(全 20 頁)

(21)出願番号　特願平6-67584
(22)出願日　平成6年3月11日(1994.3.11)
(65)公開番号　特開平6-309252
(43)公開日　平成6年11月4日(1994.11.4)
審査請求日　平成13年1月22日(2001.1.22)
(31)優先権主張番号　042,957
(32)優先日　平成5年4月5日(1993.4.5)
(33)優先権主張国　米国(US)

(73)特許権者　398038580
ヒューレット・パッカード・カンパニー
HEWLETT-PACKARD COMPANY
アメリカ合衆国カリフォルニア州パロアルト　ハノーバー・ストリート　3000
(74)代理人　100081721
弁理士　岡田　次生
(72)発明者　アンドリュー・ジョン・ウィルクス
アメリカ合衆国カリフォルニア州パロアルト　パーク・ブルーヴァード　3788

審査官　石井　茂和

最終頁に続く

(54)【発明の名称】相互接続インターフェース

(57)【特許請求の範囲】
【請求項１】
相互接続組織に接続する複数のホストコンピュータを含み該ホストコンピュータが該相互接続組織上で互いに通信するマルチコンピュータデータ処理システムにおける、ローカルメモリを有する該ホストコンピュータと該相互接続組織とを接続する相互接続インターフェースであって、該ホストコンピュータのそれぞれが、メッセージを該相互接続組織に送信することによって同様の該相互接続インターフェースを有する他のホストコンピュータと通信し、該メッセージのそれぞれが、パケット本体、パケット識別情報、保護キー、およびスロット識別値をそれぞれ含む１つまたは複数のパケットを有し、
前記相互接続インターフェースは、
前記相互接続組織と前記ホストコンピュータへの接続を提供するインターフェース手段と、
前記相互接続組織からメッセージを受信するために、前記インターフェース手段に接続された受信ターミナルと、を含み、
前記受信ターミナルは、複数のレコードを格納するスロットテーブル手段を含み、該レコードのそれぞれが、前記スロット識別値の１つおよび前記相互接続組織上のメッセージの送り手の１つに対応し、該スロット識別値によって該送り手が識別され、該レコードのそれぞれが、前記ホストコンピュータのメモリにおける対応領域を識別する情報、前記保護キー、および入来するメッセージの構成を特定する情報を含んでおり、
また、前記受信ターミナルは、前記相互接続インターフェースに入来するメッセージを処

理する、前記インターフェース手段に接続された受信制御手段を含み、該受信制御手段は、前記パケットの保護キーが前記レコードに含まれた保護キーと合致する場合、該入来するメッセージにおけるスロット識別値に対応するレコードで識別された前記ホストコンピュータのメモリ領域に前記パケット本体を記憶させるよう構成されている、
前記相互接続インターフェース。
【請求項２】
前記受信制御手段は、前記ホストコンピュータによる処理を待っているメッセージのリストを格納する手段と、
入来するメッセージにおける全パケットが前記ホストコンピュータの前記メモリ領域に記憶されたときに、当該メッセージを、前記処理を待っているメッセージのリストに登録させるための手段と、を含む請求項１に記載の相互接続インターフェース。
【請求項３】
前記受信制御手段は、入来メッセージの全パケットが前記ホストコンピュータの前記メモリ領域に記憶されているとき、該ホストコンピュータが前記メッセージのリストにより特定されるメッセージの処理中でなければ該ホストコンピュータに割込む手段を含む請求項２に記載の相互接続インターフェース。
【請求項４】
前記相互接続組織上でメッセージを伝送する送信用ターミナルを含み、該送信用ターミナルは、
前記ターミナルにより送られるメッセージの全パケットに含まれる送信保護キーを格納する保護レジスタと、
前記相互接続組織上で送られるメッセージの位置および該メッセージの意図される受け手を特定するメッセージ識別手段と、
前記特定されたメッセージを含むパケットを前記相互接続組織上で生成および送信する送信制御手段と、を含み、
前記パケットのそれぞれが前記送信保護キーおよびスロット識別値を含み、該スロット識別値が、該メッセージの受け手に取り付けられた前記相互接続インターフェース上で前記スロットテーブル手段におけるレコードを特定するよう構成されている、
請求項１に記載の相互接続インターフェース。
【請求項５】
前記保護レジスタの内容が、前記相互接続インターフェースに接続された前記ホストコンピュータのオペレーティングシステムによってのみ変更することができる請求項４に記載の相互接続インターフェース。
【請求項６】
前記メッセージ識別手段は、複数のワーク・ディスクリプタ・レコードを含む入力ジョブキューの位置を特定する手段を含み、該ワーク・ディスクリプタ・レコードのそれぞれが、意図されるメッセージの受け手を特定する情報、前記スロットテーブル手段におけるレコード、および前記ホストコンピュータにおける該メッセージの位置を含んでおり、
また、前記メッセージ識別手段は、前記入力ジョブキューにおける前記ワーク・ディスクリプタ・レコードの１つに対するポインタを格納するＤＭＡポインタ手段を含む請求項４に記載の相互接続インターフェース。
【請求項７】
前記入力ジョブキューの位置特定手段は、前記ホストコンピュータにおける該入力ジョブキューのアドレスを格納する手段を含み、該アドレスが該ホストコンピュータのオペレーティングシステムによってのみ変更することができる請求項６に記載の相互接続インターフェース。
【請求項８】
前記メッセージ識別手段は、意図されるメッセージの受け手、前記スロットテーブル手段におけるレコード、および該メッセージが格納されるバッファを特定する情報を格納する受信識別手段を、含む請求項４に記載の相互接続インターフェース。

【請求項９】
１つまたは複数のＦＩＦＯバッファを含み、前記受信識別手段により特定されるバッファの場所が、該ＦＩＦＯバッファの１つを含む請求項８に記載の相互接続インターフェース。
【請求項１０】
プロセッサノードおよび相互接続組織を有し、該プロセッサノードが該相互接続組織上でメッセージを送受信して互いに通信するマルチコンピュータデータ処理システムにおける、プロセッサノード間での通信方法であって、該プロセッサノード間での通信方法は、
（ａ）第１のプロセッサノードから第２のプロセッサノードに少なくとも１つのパケットを有するメッセージを送信するステップであって、該パケットがパケット本体、送信用保護キー、および該第１のプロセッサノード上での処理を識別するスロット識別値を含む該送信ステップと
（ｂ）第１のプロセッサノードから入来したメッセージのパケットに含まれる送信用保護キーが、第２のプロセッサノードに格納された保護キーと合致するかどうかを、第２のプロセッサノードで判断するステップと、
（ｃ）前記判断ステップ（ｂ）で送信用保護キーが格納された保護キーと合致すると判断されたとき、第１のプロセッサノードから入来したメッセージのパケット本体を、第２のプロセッサノードのローカルメモリ記憶領域に記憶するステップと、を含み、
第１のプロセッサノードから入来したメッセージのパケットに含まれるスロット識別値を使用して、該ローカルメモリ記憶領域が識別される前記プロセッサノード間での通信方法。
【請求項１１】
スロット識別値が第２のプロセッサノード内のスロットテーブルにおけるレコードを識別し、該レコードが、ローカルメモリ記憶領域および格納された保護キーを識別する情報を少なくとも含む請求項１０に記載の方法。
【請求項１２】
前記レコードが、入来するメッセージの構成を特定する情報を含む請求項１１に記載の方法。
【発明の詳細な説明】
【０００１】
【産業上の利用分野】
この発明は、データ処理システムに関するものであり、より詳細にはマルチプロセッサ式のデータ処理システムにおいて用いるための改良された通信システムにおける相互接続インターフェースに関するものである。
【０００２】
【従来の技術】
一つのクラスのマルチプロセッサ式のデータ処理システムは、ある一つの相互接続組織によって接続された複数のプロセッサノードからなるものである。各プロセッサのノードに典型的に含まれているものは、一つのプロセッサ及びローカルＲＡＭである。高速の相互接続組織によれば、ノード間での通信が維持される。ある一つの演算上の問題が複数のノード間で分割されて、異なるノードにおいて利用可能な特別の資源を用いることによってその効率性を改善し、又は、ある一つの結果が利用可能であるのに先だって経過されねばならないリアルタイムを改善するようにされる。従って、一つのノード上で実行するプロセスは、該当のシステムの他のノードにおいて実行される演算に依存することになる。種々のプロセスが相互接続組織を介して通信され、情報の交換がなされるとともにプロセスの同期がとられる。
【０００３】
このようなシステムの実行能力のレベルは、第１のノード上を実行するプロセスから第２のノード上を実行するプロセスへと、メッセージを伝送することができる速度に依存している。各ノードに典型的に含まれているインターフェース回路は、メッセージの送信及び

受信を管理するものである。先行技術の通信システムにおいては、メッセージを受信しているノードが、該当のメッセージが受信されたか否かの制御をする。第１のノード上のプロセスが第２のノード上のプロセスへのメッセージの伝送を所望するときには、第１のノード上のインターフェース回路が相互接続組織を介してメッセージを伝送し、そして、第２のノードにおけるインターフェース回路からの確認（acknowledgement ）を待つ。

【０００４】

【発明が解決しようとする課題】

該当のメッセージが第２のノード上のインターフェース回路に到達したときには、メッセージが受信されるか、又は、該メッセージが失われるかの二つの可能性がある。ノードがビジー過ぎるか又はメッセージを処理するのには満杯であるときには、このメッセージは失われることになる。受信側は、メッセージが失われたことを指示するメッセージを戻すか若しくは沈黙の状態に留まる。いずれの場合にしても、送信側はある時間の経過後に該当のメッセージを再送せねばならない。不都合なことに、送信側には別の通信を試行するための最適の時点を定める手段が備えられていない。その理由は、最適の時点で要求されることは受信側のノードにおいて実行しているプロセスについての知識であり、これらのプロセスは送信側には明白ではないことにある。その結果として、この送信は、典型的には、ある所定の時間だけ待機し、これに次いで別の通信を試行するようにプログラムされている。送信の通信開始が早すぎるときには、受信側はまだビジー状態にあり、メッセージは再び失われることになる。メッセージが伝送され、失われる度に、相互接続組織の使用の効率性が低下される。この低下した効率性により、他のノードで実行している他のプロセスがより遅くなるようにされるが、その理由は、これら他のプロセスは、相互接続組織における十分なバンド幅を最大の速度で実行するようにアクセスできないことにある。メッセージを再送する前の送信の待機時間が長過ぎるときには、第１のメッセージの主題として委任されたタスクの結果を含む第２のノードからの戻りのメッセージを待機して止められていることから、この送信上のプロセスは、より低い最大速度で実行することができる。

【０００５】

メッセージが第２のノードによって受け入れられるときであっても、この先行技術の通信技術はまだ非効率のものである。ある一つのメッセージが受信側に到達すると、受信側におけるインターフェース回路がプロセッサを中断して、入来するメッセージを扱う必要のあることが、このセッサに対して告知される。このメッセージが長いものであるときには、それは典型的には幾つかのブロックに分割される。各ブロックの到達により典型的には個別の中断が発生される。これらの中断を処理することにより、プロセッサが他のタスクに関してビジー状態にある場合の、このプロセッサのスループットに重大な低下が生じるようにされる。原理的にいえば、幾つかのブロックに分割された一つの長いメッセージに関連する中断は、その最大のメッセージを格納するのに十分なバッファを備えたインターフェース回路を用いることによって回避することができる。しかしながら、このようなシステムでも、まだメッセージ毎に１回だけプロセッサを中断させる。これに加えて、受信側では、入来するメッセージが受信側のローカルメモリに配置されるべき場所及びこのメッセージを移動させる際の受信側のローカルメモリにおけるそれらの宛先を定めるために、典型的に著しく長い処理時間を費やす。

【０００６】

また、先行技術によるこのタイプのシステムは、送信の観点からも非効率のものである。第１に、プロセスにおいてメッセージの伝送が所望される毎に、オペレーティングシステムの呼出しを求めることによって実行されねばならない。多くのプロセスが単一のノード上を実行しているときには、他のプロセスにより用いられて、一つのプロセスを悪化させる資源の可能性が常に存在する。一つのプロセスがインターフェース回路及び／又は相互接続組織に対する無制限のアクセスを有するときには、このような悪化は可能性があるものである。これに加えて、保護がなされねばならないものは、第２のノード上でプロセスを「アタック」し、又は、相互接続組織に優越する第１のノード上のプロセスに対してで

ある。先行技術においては、これらの保護は典型的にはオペレーティングシステム内で構築されている。通信の受信側においては、そのオペレーティングシステムにより、受信側のメモリに書き込むことの許容がテストされる。送信側においては、そのオペレーティングシステムにより、一つのプロセスが他のプロセスとして仮装することを停止される。不都合なことに、オペレーティングシステムの呼出しにより、通信のプロセスが低速にされる。

【０００７】

先行技術に係る別の問題は、初めてオンラインになったとき、システムが新規なプロセッサノードの弱点であることにある。新規なプロセッサがマルチプロセッサシステムに付加されたときには、それは保護されない状態で起動される。この時間中は、それは誤りをともなってロードされ、その結果として不可能化される。先行技術においては、この問題に対する解決策として利用可能なものはない。

【０００８】

先行技術の相互接続システムに係る更に別の問題は、規則的なメッセージを搬送する同じ相互接続組織を用いることにより、プロセスに待ち時間の短い（low-latency ）同期化を付与する技術が欠如していることである。異なるノードを実行させるプロセスを同期化させることは重要なことが多い。例えば、多くの適用において、プロセッサの全てによって用いられるデータ値を変更することは、一時に一つのプロセスだけに与えることができる。先行技術のシステムにおいては、個別の相互接続組織を通して、又は、オペレーティングシステムの呼出しを介してのいずれかによって同期が与えられる。前者の解決策によれば相互接続組織のハードウエアのコストが著しく上昇することは明かであり、また、後者の解決策は多くの適用に対して遅過ぎるものである。

【０００９】

広範な言い方をすれば、この発明の目的は、改良されたマルチコンピュータの相互接続インターフェースを提供することにある。

【００１０】

この発明の別の目的は、受信側がビジー状態にあるために失われるメッセージの数が、先行技術のシステムと比較して著しく減少される相互接続インターフェースを提供することにある。

【００１１】

この発明の更に別の目的は、既にビジー状態にあるときに、受信側が入来メッセージによって中断されることがない相互接続インターフェースを提供することにある。

【００１２】

この発明の更に別の目的は、メッセージを伝送する間にシステム上の他のプロセスからの偶発的な又は悪意のあるアタックから各プロセスを保護するための、オペレーティングシステムの呼び出しが必要とされない相互接続インターフェースを提供することにある。

【００１３】

この発明の更に別の目的は、入来メッセージを受信側のメモリに格納する場所を定めるために、受信側で消費される時間量が節減される相互接続インターフェースを提供することにある。

【００１４】

この発明の更に別の目的は、規則的なメッセージトラフィックのために用いられるのと同じ相互接続組織を用い、集積化され、保護され、待ち時間の短い同期化操作がなされる相互接続インターフェースを提供することにある。

【００１５】

【課題を解決するための手段】

この発明は、ある一つのホストコンピュータを複数のホストコンピュータを含むマルチコンピュータのデータ処理システム内の相互接続組織に接続させるための相互接続インターフェースである。ホストコンピュータは相互に通信する相互接続組織に接続されている。それぞれのホストコンピュータは、相互接続組織上でメッセージの送信及び受信をするこ

とにより、同種の相互接続インターフェースを有する他のホストコンピュータと通信する。それぞれのメッセージには一つ又は複数のパケットが含まれており、それぞれのパケットには、パケット本体、パケット識別情報、保護キー及びスロット識別値である。相互接続インターフェースに含まれているものは、一つの受信用ターミナル及び一つ又は複数の送信用ターミナルである。受信用ターミナルにおいては、相互接続組織に接続されている他のコンピュータからのメッセージを受信する。受信用端末には複数のレコードを格納するためのスロットテーブルが含まれており、それぞれのレコードは、相互接続組織上を伝送されるメッセージに含まれているスロット識別値の可能性のある一つのものに対応している。各レコードに含まれている情報は、ホストコンピュータのメモリにおける対応領域の位置及びサイズを識別するものである。メモリのこの領域は、対応するスロット識別値を有する受信したメッセージを格納するために用いられる。各レコードには、保護キー及び入来メッセージの構成を特定する情報も含まれている。ある一つの特定のレコードを参照する入来メッセージ内の保護キーは、レコードに格納されている保護キーに合致せねばならない。ある一つの受信用ターミナルは、パケットにおける保護キーがレコードに格納されている保護キーと合致したときには、相互接続組織を介して受信されたパケットを、入来メッセージ内のスロット識別値に対応するレコードにおいて識別されるホストコンピュータのメモリ領域に格納する。受信用ターミナルには、ホストコンピュータによる処理を待機している受信メッセージを特定する情報を格納するための複数のリストも含まれている。

【００１６】

送信用ターミナルは、相互接続組織上でメッセージを伝送するために用いられる。各送信用ターミナルに含まれている保護レジスタは、送信用ターミナルによって伝送されるメッセージのパケットに含まれるべき送信用保護キーを格納するためのものである。各送信用ターミナルに設けられるものは、相互接続組織上をメッセージの意図された受信側に伝送されるべき、メッセージの位置上の情報である。送信用ターミナルにおいてはメッセージのためのパケットが発生され、そして、相互接続組織を介してパケットが伝送される。各パケットには送信用の保護キーが含まれている。この発明の好適な実施例においては、保護キーが格納されるレジスタはオペレーティングシステムによってのみ書き込むことが可能にされる。

【００１７】

【実施例】

図１は、この発明による相互接続インターフェースシステムが用いられる典型的なマルチプロセッサシステムのブロック図である。データ処理システム１０には複数のプロセッサノード（以下、ノードという）が含まれている。例示的なノードは１２及び１４で示されている。各ノードにはローカルＲＡＭ１６を有するホストコンピュータであるプロセッサ１５が含まれている。ある一つの特定のノードにおけるプロセッサは、相互接続組織１１を介して他のノードにおけるプロセッサとの通信をする。該当する技術においては多様なインターフェース組織の設計が知られている。ここでの検討の目的のためには、相互接続組織１１がある一つの送信用ノードからある一つの受信用ノードまでの経路を提供することを注意すれば十分である。それぞれの処理用ノードは、相互接続インターフェース１７を介して相互接続組織１１に接続されている。

【００１８】

この発明と先行技術の相互接続インターフェースとの間の一つのキーの差異は、送信側と受信側との間の通信をいずれの相互接続インターフェースが制御するかの識別をすることにある。先行技術の設計においては、受信側が制御することは、受信側に対するメッセージの伝送が成功したかどうか、及び、受信側のメモリのどこにこのメッセージが配置されたかということである。一般的にいって、各インターフェース回路は、そのノードにおいてメモリのブロックを制御する。典型的には、メモリの一部はインターフェース回路内にあり、メモリの一部はローカルＲＡＭ内にある。先行技術のシステムにおいて、ある一つの特定のノード上でメモリの割当てをすることは、問題のノードにおける相互接続インタ

ーフェース及び／又はプロセッサによって制御される。利用可能なメモリがフルであるとき、又は、インターフェース上を到来するメッセージを処理するために、プロセッサがより優先度の高いタスクによってビジー過ぎるときには、先行技術においてはメッセージが失われることになる。

【００１９】

この発明に係る相互接続インターフェースを用いるシステムにおいては、各潜在的な送信は、受信側でのメモリの部分を管理する。このメモリは送信側によって伝送されるデータを格納するために用いられる。ノード１２上でのプロセスに対する特定のタスクを実行するためにノード１４上でのプロセスに命令を出す、ノード１２からノード１４へのメッセージについて考察する。ノード１２上での相互接続インターフェース１７はメッセージが挿入されるべきスペースのブロックを管理していることから、ノード１２上での相互接続インターフェースは、ノード１４にメッセージのためのスペースがあることを「知る」。ノード１４上でのプロセスがそのタスクを完了したときには、受信側のプロセスは通常は「タスク処理完了（done with task）」メッセージを伝送する。このメッセージは、また、該当のタスクを開始させたメッセージのために用いられたメモリが現在はフリーであることを、ノード１２上での相互接続インターフェースに信号を出すためにも用いられる。このために、ノード１２上での相互接続インターフェース１７は、各潜在的な受信側に割り当てられるメモリの状態を常に知っていることになる。送信側としては受信側での送信の資源の状態を知っていることから、メッセージを受信するために受信側で割り当てられたメモリが不十分であるときには、送信側はメッセージの伝送を回避することができる。このために、受信側でのスペースの欠如のためにメッセージが失われる問題は、この発明によって回避される。

【００２０】

また、この発明に係る相互接続インターフェースによれば、ホストがビジー過ぎて、新規なメッセージが到来した旨の信号を出して中断をするサービスができないときに、ホストの中断の防止もする。この発明を用いるデータ処理システムにおいては、相互接続インターフェースで導出されたタスクについて作業していないときにのみ、相互接続インターフェースによりそのノードにおけるプロセッサの中断がなされる。ホストプロセッサが相互接続インターフェースを介して伝送されたタスクについて既に作業をしているときには、ホストプロセッサを中断することなく、新規のメッセージが作業キューに対して付加される。この発明の実施例では、実行されているタスクよりも優先度が低い相互接続インターフェース導出のタスクについてホストプロセッサが作業しているときに、中断がホストプロセッサに伝送されることも、当業者には明かなことである。

【００２１】

ここで、この発明に係る相互接続インターフェース１００の部分のより詳細なブロック図である図２を参照すると、データ及びプログラムの格納のためのローカルＲＡＭ１０３がホストコンピュータである中央演算装置（以下、ＣＰＵと記す）１０２が含まれているプロセッサに対する相互接続組織からのメッセージを受信するようにされている。送信用ターミナル、受信用ターミナルとしての機能を含む相互接続インターフェース１００に含まれているインターフェースである相互接続組織インターフェース１０４は、相互接続組織上でメッセージを送受信するためのものである。相互接続組織インターフェース１０４によれば、相互接続組織上でデータを実際に伝送することが必要とされる種々のプロトコルが管理される。このような相互接続組織インターフェースは当業者には周知のことであるから、ここではインターフェース１０４について詳述しない。また、相互接続インターフェース１００には受信制御手段であるコントローラ１０５も含まれていて、下記に説明される種々の論理的な動作がなされる。コントローラ１０５も、該当の技術においては通常のものである。

【００２２】

これも相互接続インターフェース１００に含まれている固有のローカルメモリ１０６は、入来するメッセージに関連した情報を格納するために用いられる。スロット１０７のレコ

ードを格納するためのスロットテーブルとしてのローカルメモリ１０６には、「スロット」として参照される複数のレコードが含まれている。例示的なスロットは１０７において示されている。各スロットに格納されている情報は、相互接続組織上の相互接続インターフェース１００に接続されているある一つの特定のプロセッサに割り当てられたローカルＲＡＭ１０３におけるメッセージ領域を記述するためのものである。ある一つのスロットに格納された情報にはローカルＲＡＭ１０３におけるアドレスが含まれており、問題のプロセッサからのメッセージはここで格納され、また、その量のスペースが割り当てられることになる。

【００２３】

これも各スロットに含まれている保護キーは、誤指向のメッセージに対して保護を与えるためのものである。メッセージ内の各入来パケットが、このメッセージにおいて特定されたスロットに格納されたのと同じ保護キーを有していることを要求することにより、保護を与えることがなされる。

【００２４】

これも各スロットに含まれているスペースは、当該スロット内のデータによって制御される受信メッセージの構成を特定する情報を格納するためのものである。ある一つの特定のメッセージに対して、このスペースは２個のサブグループの値に分割されているものとして観察される。その第１のサブグループにおいては、ある一つの受信メッセージについての事項が特定される。これに含まれるものは、該当するメッセージの長さ、及び、このスロットによって制御されるメモリブロックの始まりに関してメッセージが始まる位置である。その第２のサブグループは、マルチパケットメッセージの受信を制御するために用いられる。

【００２５】

上述された実施例ではローカルメモリ１０６に格納されたスロット「テーブル」が用いられているけれども、当業者には明かであるように、実際のテーブルをローカルメモリとしてのローカルＲＡＭ１０３に格納することができる。この場合において、スロットテーブルからのエントリは、キャッシュメモリにおいて用いられるのと類似の態様で、ローカルＲＡＭ１０３との間で移動することができる。

【００２６】

長文のメッセージは典型的にはパケットに分割されて、相互接続組織を介して個別に伝送される。多くのインターフェース組織はパケットからパケットへと変化する予測不能な通過時間を有しているために、ある特定のメッセージに対するパケットの到着は順不同になる。ある時点において唯一のメッセージだけがある一つの特定のスロットに伝送されている、という簡単な場合について考察してみる。コントローラ１０５は、該当のメッセージが完全である時点を決定できねばならない。このタスクにおいてコントローラ１０５を助けるためには、各々のスロットにはメッセージ識別フィールド及びカウンタ用のフィールドが含まれており、メッセージ識別フィールドに格納されている識別値を有するメッセージに対して受信されるべく残留しているパケットのトラックが維持される。送信側がメッセージの伝送をセットアップすると、メッセージ識別及びパケットカウントがこれらのフィールドにロードされる。

【００２７】

新規なメッセージの到着は、残っているパケットのカウントが０であることによって検知される。新規なメッセージの到着では、メッセージ内のパケットのカウントは０ではない。このパケットのために、パケット残留カウンタは、該新規なメッセージ及び再ロードされるべきメッセージ識別フィールドに対応するカウントに適切にセットされるようになる。このメッセージに対する各連続するパケットはカウントが減少するようになる。より詳細に後述されるように、いずれのパケットでも相互接続組織において遅れる可能性があることから、相互接続インターフェースは、メッセージのパケットが、いかなる既知の順序でも受信されることを予測することはできない。このために、各々のパケットは、スロット情報のセットアップのために必要な、全ての決定的な情報を反復する。例えば各々のパ

ケットに含まれているものは、メッセージ内のパケットの数、問題のパケットのパケットナンバー、メッセージ識別値及び保護キーである。

【００２８】

メッセージ識別フィールドは、メッセージの再伝送の間のエラーを回避するために用いられる。メッセージの伝送の間にエラーが生じたときには、送信側はメッセージの再伝送をするように指示される。受信側は、再伝送の間に到着する先行の試行からパケットを防護し、パケット残留カウンタが誤って減少することがないようにせねばならない。先行するメッセージのパケットの一つが相互接続組織内で遅れるときには、このような状況が生起する可能性がある。このような状況を回避するために、メッセージ識別フィールドは、それぞれの新規なメッセージの始めにおいて再ロードされる。先行するメッセージからの遅いパケットは、そのメッセージ識別フィールドにおいて異なる値を有していることから、遅いパケットは無視される。

【００２９】

上記の説明では、スロット毎に進行しているメッセージは一つだけであるとされた。各スロットに対して多くのメッセージが進行することが許容される場合には、各スロットに含まれるものは、許容されたメッセージ毎の、複数のメッセージ識別フィールド及びパケットカウントフィールドである。ローカルＲＡＭ１０３におけるオフセット及びメッセージ長フィールドも、可能性のある多くのメッセージの各々に対して含まれる。

【００３０】

これに加えて、多くのメッセージが存在するときには、各スロットに対してメッセージカウントが格納される。このメッセージカウントにより、ホストコンピュータとしてのＣＰＵ１０２に対するメッセージ到着の告知が、指示された数のメッセージが到着した後でのみなされることを、送信側に対して特定化するために許容するようにされる。このフィールドに対するデフォルト値は０である。即ち、各入来するメッセージにおける受信側に告知するようにされる。カウントが０よりも大きい値にセットされたときには、カウントは各完了したメッセージ上で減少し、カウントが０に到達したときにのみ告知がなされる。

【００３１】

メッセージが完全に受信されたときには、コントローラ１０５は、告知のレコードを作業キュー１０８に配置する。ＣＰＵ１０２が先行のメッセージに応答して既に満杯であるときには、中断が生起することはない。この場合において、ＣＰＵ１０２は、作業キュー１０８を介して受け入れたタスクが終了した後で、作業キュー１０８をチェックするようにプログラムされている。メッセージが完了したときに、ＣＰＵ１０２が相互接続インターフェースで導出されたタスクについて作業していなかったとすると、ＣＰＵ１０２はＣＰＵ１０２に対する中断を発生させて、メッセージが待機している旨をこのＣＰＵに告知するようにされる。

【００３２】

コントローラ１０５は、また、種々のプロトコルエラーに対する入来メッセージのチェックもする。例えば、プロセッサに対して割り当てられたメモリ領域に対して長過ぎるメッセージの伝送要求、誤りの保護キーを有するパケット等に対するチェックをする。エラーの状態は、好適には、送信側よりもメッセージを受信する側で処理される。エラーが受信側において検出されたときには、受信側のオペレーティングシステムは、更に別の問題が送信によって発生されることを防止するための動作をするように告知される。誤りのあるメッセージを発生させる問題を送信側が有しているときに好適であることは、該メッセージの受信側が、マルチプロセッサの環境を管理するオペレーティングシステムの部分に対して、このような状態の存在を指示することである。オペレーティングシステムのこの部分は、これに次いで、欠点のあるノードを再ブートさせるための動作をすることができる。送信側がこのようなエラーをしているときには、それ自体でそれらを訂正することに依存することはできない。実際のところ、エラーは伝送ノードにおける悪意のあるプロセスの結果である可能性がある。後の場合においては、伝送ノードから離れるような制御がされねばならない。

【００３３】
上述された実施例では相互接続インターフェースに格納された単一の作業キューが用いられているけれども、当業者には明かであるように、異なる優先度を有する多くの作業キューを用いることができる。これに加えて、実際の作業キューはホストメモリ、すなわちローカルＲＡＭ１０３に配置することができる。この場合においては、作業キューに対するポインタが相互接続インターフェースに格納されることになる。
【００３４】
この発明による相互接続インターフェースの受信部分についての上述された実施例は、小さいメッセージに対して最も良好に使用される。その理由は、この実施例においては、ローカルＲＡＭ１０３における物理的メモリが、送信側がメッセージを伝送できる全体的な時間的周期にわたり、それぞれに有効な可能性のある送信側に対して割り当てられることにある。メモリの大きさは、対応の送信から予測される最長のメッセージを受信するのに十分なものでなければならない。
【００３５】
大きいデータの伝送のために、この発明による相互接続インターフェースにおいては、伝送と同時にスペースが割り当てられるプロトコルが好適に用いられる。この状況においては、送信側は、受信する側が要求するスペースにおいて、オペレーティングシステムに対してメッセージを伝送する。受信側のオペレーティングシステムがスペースを割当て、そして、それを記述した情報を新規なスロットに配置する。これに次いで、スロットナンバー及び保護キーが、伝送を実行する際に用いられる希望の送信に対して戻される。データが受信側において処理されるときには、スロットが再指定されて、メモリ（ローカルメモリ）が他の使用に戻される。
【００３６】
その開始において、各ノードには、システム内の他のノードと互いにメッセージのサービスをするための最低限の大きさのメモリが割り当てられ、これらのメモリブロックを指摘するためにローカルメモリ１０６における対応のスロットが初期化される。このために、システムがパワーアップされたときには、各ノードは全ての他のノードからのメッセージを受信できるようにされる。これらの初期的なスロットは、ブートストラップスロットととして参照される。
【００３７】
ブートストラップスロットに対して選択されるキーは、スロットが割り当てられるノードアドレスを伝送する機能である。初期のキーは、いずれのノードでもキーを算出できるように選択される。しかしながら、可能性のあるキーから用いられるキーの数は、劣性のあるノードが偶発的にキーを発生させる公算が小さいように、十分に小さいものとされる。新規なノードの始めの動作は、より低い予測性の態様において導出されるものに対して、これらのキーにおける変更を協議することである。このようにすることで、あるノードにおける悪意のある処理（例えば、ウイルス）に対する防護が付与されて、他のノードに対するメッセージの伝送が可能にされる。
【００３８】
実行中のシステムに対するノードの動的な付加は、通常のブートストラップと同じやり方で達成される。ある一つのノードは、それと通信していないノードと、システムを排除された（又は付加されなかった）ものとの間の差異を決定することができない。このために、各ノードがローカルメモリとシステム内に潜在的に存在する可能性があるすべてのノードに対するスロットとを保留するときには、新規なノードは常に付加されることができる。
【００３９】
最大数のノードは大きいけれども、活動的なノードの数が小さいときには、この手順により、各ノードにおけるメモリについて、受け入れ難い大きさの無駄が生じる可能性がある。この無駄を回避するために、幾つかのノードは「ブートサーバ」として指定されることができる。全ての可能性のあるノードのために、メモリはこれらのノードに割り当てられ

る。ブートサーバにより、いずれのノードが活動的であるかが決定される。他のノードに保留されるスロットは、活動的であるとして知られているノードに対するだけのものである。ある一つの新規なノードがオンラインにされると、この新規なノードは、それに応答するブートサーバに対するシステムを探索する。ブートサーバは、これに次いで、ノードにおけるオペレーティングシステム間での通信のためのスロットを割り当てることを開始する他のノードに対して、新規なノードの存在を通知する。

【００４０】

一つのノードにおいて実行中のプロセスが、他のノードにおいて動作しているプロセスに対してメッセージの伝送を所望したとすると、この希望の伝送プロセスにより、そのノードにおけるオペレーティングシステムに対するセットアップの要求がなされる。オペレーティングシステムにより、メッセージの戻りに対する送信側のノードにおけるスロットが設定され、これに次いで、受信側ノードにおけるオペレーティングシステムに対する開始の間に、オペレーティングシステムのスロットセットアップに対して、オペレーティングシステムを介してメッセージが伝送される。受信側ノードにおけるオペレーティングシステムは、これに次いで、伝送側のプロセスに対するスロットを設定し、受信側のプロセスに対して、制御される送信側におけるスロットの数を告知する。

【００４１】

この発明による相互接続インターフェースが他の相互接続インターフェースからのメッセージを受信するやり方について上記のように説明してきたが、これからは、送信及び受信の双方において相互接続インターフェースによって用いられるデータ構成及びコントロールレジスタについて、より詳細な検討をする。この発明を用いるデータプロセッサにおいては、全ての通信は、ターミナルとして参照される端点を介して実行される。図３のブロック図は、この発明による相互接続インターフェース３００の好適な実施例を示すものである。相互接続インターフェース３００に備えられているものは、一つの受信用ターミナル３０２及び一つ又は複数の送信用ターミナル３０３である。ターミナルの全ては、ホストコンピュータのメモリにおける参照位置に対する単一のメモリマッピングテーブルを共有しており、また、この発明において用いられる種々のレジスタが、ホストコンピュータにおけるメモリアドレスとして現れている。かくして、ホストコンピュータは、ロード命令及び格納命令をそれぞれに介して、種々のレジスタの読み取り及び書き込みをすることができる。また、メモリマッピングテーブルは、種々のレジスタに対するアクセスを制限する手段も付与するようにされる。オペレーティングシステムにのみアクセス可能なアドレスにレジスタを指定することにより、レジスタを駄目にする誤りのあるプロセスの可能性が著しく低下される。インターフェース手段としてのインターフェース・ハードウェア３０８により、相互接続組織に対するインターフェースが付与される。

【００４２】

一般的に、ターミナルで格納又は抽出されるメッセージは、それぞれに、ホスト・コンピュータ・メモリ３０４で受け入れられ、又は、これから送出されるものである。この配列により、種々のターミナルにおける中間バッファに対するメッセージの転送コストが軽減される。インターフェース・ハードウェアは図２においては１００で示されているものであり、上記されたインターフェース・ハードウェアは受信用ターミナルの一例である。

【００４３】

この発明による好適な実施例において用いられる送信用ターミナルは、ＤＭＡ（ダイレクトメモリアクセス）モード及びダイレクトモードとして参照される二つの異なるモードで動作することができる。図４のブロック図は、送信用ターミナル４００を示すものである。送信用ターミナル４００によって実行される種々の動作は、受信制御手段であるコントローラ４０１によって制御される。ＤＭＡモードにおいて、送信用ターミナルは、好適にはホストコンピュータシステムメモリの一部である入力キュー４０８からの作業をする。より詳細に後述されるように、入力キューに含まれているディスクリプタのリストは、相互接続組織を介して伝送される用意のある、ホスト・コンピュータ・メモリにおけるメッセージの位置を特定するものである。これらのメッセージは適当な順序でコントローラ４

０１により処理される。ダイレクトモードにおいては、相互接続組織を介して伝送されるべき単一のメッセージが、ダイレクト送信レジスタ４０６の内容によって識別される。メッセージはバッファ４００が典型的である一つ又は複数のＦＩＦＯ（先入れ先出し）バッファ４０９に格納されている。ダイレクトモードは典型的には短い待ち時間を必要とする短いメッセージのために用いられる。

【００４４】

先に記載されたように、多くのプロセスがシステムの各ノードで実行することができるマルチコンピュータシステムにおいて、防護の事項は極めて重要なことである。システムは、異なるプロセスを装い、また、システムの他のノードにおいて格納されているデータを駄目にする一つのノードにおける誤りのあるプロセスから保護されねばならない。通信の防護を確実にするこの発明によって用いられる一つの方法は、各メッセージに保護キーを含ませることである。このメッセージにおける保護キーの値は、受信側ノードの受信用ターミナルにおける対応のエントリの値と合致せねばならない。このレジスタは、送信用ターミナルに接続されたノードにおけるオペレーティングシステムによって書き込まれるだけである。保護キーのレジスタは「信託された」オペレーティングシステムソフトウエアによってのみアクセスされることから、誤りのあるメッセージを伝送することによって他のノードにおけるプロセスを駄目にできる、一つのノードにおいて実行しているプロセスの可能性が著しく低減される。

【００４５】

ＤＭＡモードで動作しているときには、ターミナルのメッセージ伝送は、ホストメモリから相互接続組織に対して直接的になされる。保護レジスタ、レジスタセット４０２は、このモードの動作に必要とされる種々の値を格納するために設けられている。ターミナルによって完成されるべき伝送タスクのリストは、ノードのローカルＲＡＭ内に配置されたＤＭＡ入力キュー４０８に格納されている。各タスクに含まれているものは、ターミナルを介して伝送されるべきメッセージの記述である。この記述は、以下の検討においてはセンド・ワーク・ディスクリプタ（send-work-descriptor）として参照される。キューの位置及びサイズを特定するデータは、レジスタセットとしてのＤＭＡレジスタ４０２におけるレジスタに格納されている。この情報は、オペレーティングシステムによってのみ書き込まれるものである。ＤＭＡレジスタ４０２に含まれているポインタは、ＤＭＡ入力キューにおいてターミナルによって調査された最後のセンド・ワーク・ディスクリプタの物理アドレスに対するものである。

【００４６】

この発明の好適な実施例におけるセンド・ワーク・ディスクリプタの内容は４個のフィールドに分割される。その第１のフィールドでは、意図された受信が識別される。これに含まれているものは、このーバのノードの識別、受信側のメモリ受け入れターミナルにおけるスロットの数、及び、問題のメッセージが受信側のメッセージスペースに配置されるべきアドレスを特定するオフセットである。その第２のフィールドで識別されるものは、送信側のメモリにおける物理アドレスに対するポインタ、及び、このポインタによって特定された送信のメッセージスペースにおけるメッセージのオフセットである。その第３のフィールドはメッセージタイプを特定するために用いられる。この発明の好適な実施例において、メッセージは、伝送メッセージ又は「リモートゲット（remote-get）」メッセージのいずれかである。ある一つのリモートゲットメッセージのために、受信側のメモリにおいて特定された領域のコピーが、送信側に戻るように伝送される。このプロトコルは、データを含む戻りメッセージによって追従される、受信側からのデータを要求するメッセージと同等のものである。しかしながら、この場合においては、受信側におけるインターフェース・ハードウェアにより、受信側を擾乱させることなく戻りメッセージが発生される。終りに、最後のフィールドで特定されるものは、メッセージの長さ、及び、受信側の告知のために用いられる識別コードである。

【００４７】

ＤＭＡレジスタ４０２には、ＤＭＡ入力キュー４０８の開始の物理アドレス及びキューの

長さを格納するためのレジスタグループも含まれている。これらのレジスタはオペレーティングシステムによって書き込まれる。ＤＭＡレジスタセット４０２には、相互接続インターフェースがキューにおいて処理を開始したという、最後のセンド・ワーク・ディスクリプタの物理アドレスを特定する現在位置ポインタも含まれている。

【００４８】

これもＤＭＡレジスタ４０２に含まれているものは、メッセージの処理の間に多くの例外のいずれかが生じたときのエラーデータを格納するためのステイタスレジスタ、問題のターミナルに対する優先レベルを格納するためのレジスタ、及び、現在のタスクの状態を指示するためのレジスタである。それぞれの送信用ターミナルは関連した優先度を有している。この発明の好適な実施例においては二つの優先レベルが用いられている。その一つはコントロールメッセージのような高い優先度のトラフィックであり、他の一つはバルクデータ転送のようなより低い優先度のトラフィックである。優先度はパケットレベルにおいて賦課される。即ち、より高い優先度のメッセージは、次に続くパケット境界において、より低い優先度のメッセージに先立つことになる。この発明の好適な実施例において、ターミナルの優先度を特定するレジスタは、オペレーティングシステムによる書き込みが可能なものだけである。

【００４９】

ＤＭＡレジスタ４０２には、その中に格納されている動作を開始させるためのコントロールレジスタも含まれている。ある適切なコードがこのレジスタに書き込まれると、コントローラ４０１においては、このコードによって識別された伝送動作が開始される。

【００５０】

この発明の好適な実施例においては、ＤＭＡ入力キュー４０８におけるワークアイテムのリストが、物理的メモリの隣接ブロックにおけるサーキュラーベクトル（circular vector）としてフォーマット化される。送信用ターミナルは始めのセンド・ワーク・ディスクリプタを開始させることにより動作され、その処理がなされ、そして、これに次いで次続のものに進行される。送信用ターミナルが０のノード識別を有するセンド・ワーク・ディスクリプタに到達すると、作業が停止され、継続に先立つノード内のプロセッサからの入力を待機するようにされる。送信用ターミナルがベクトルの端部に到達すると、受信用ターミナルはベクトルにおける始めのアドレスに対して再び巻き込まれる。

【００５１】

当業者には明かであるように、リスト構造を用いて各ターミナルとホストプロセッサとの間で通信を体系化するためには、多くの異なる可能性のある配列がある。相互接続インターフェースにおける種々のターミナルの状態を定めることをホストプロセッサが許容するいかなる構成のものでも、この発明から離れることなく使用することができる。

【００５２】

先に注意されたように、ダイレクトモードは、短い待ち時間を要する短いメッセージの伝送のために用いられる。この発明の好適な実施例においては、これらのメッセージは、短いコントロールメッセージ又はフェッチアンドオピメッセージ（Fetch-and-Opmessage）のいずれかである。ある一つのフェッチアンドオピメッセージによれば、受信側のメモリにおいて特定された位置の内容が戻され、また、メッセージの一部である一つ又は複数のオペランドにおいて、受信側である種の動作をするようにされる。直接的なターミナルに含まれている付加的なコントロールレジスタセット４０４は、アプリケーションのアドレススペースにマッピングされ、そして、ロード／格納命令を用いて、アプリケーションによって直接的にアクセスされる。付加的なレジスタによれば、センド・ワーク・ディスクリプタにおいて正常に生成された情報が格納される。メッセージが格納されるローカルＲＡＭにおけるアドレスを特定するのに代えて、メッセージが格納されるＦＩＦＯバッファ４０９の識別が付与される。ＤＭＡレジスタには、進行中の活動のタイプ、例外が生じたときのエラーデータ、及び、ある一つの結果がフェッチアンドオピオペレーションから利用可能であるか否かを指示するレジスタを指示するような、ステイタスレジスタのセットも含まれている。ＤＭＡレジスタに含まれている一つのレジスタは、最後のフェッチアン

ドオピからの戻りの値を保持するためのものである。

【００５３】

直接的なターミナルを用いてメッセージを伝送するために、ローカルプロセッサは、ターミナルがアイドル状態になるまで待機する。そして、受信側のノード、スロット、開始アドレス及び告知データが、ダイレクトターミナルにおけるコントロールレジスタセットに書き込まれる。これに次いで、伝送されるべきパケットがＦＩＦＯバッファ４０９に書き込まれる。このことは、単一アドレスに対する単一ワード格納動作のシーケンス、又は、幾つかのバイトを一時に発するためのキャッシュラインフラッシュ（cache-line-flush）動作のいずれかとして達成できる。最後に、ある一つの伝送コマンドが送信用ターミナルにおけるコントロールレジスタセットに書き込まれる。

【００５４】

リモートのフェッチアンドオピが、伝送動作と同様な態様で実行される。ただし、メッセージ本体に代えてオペランドが付与されている。そして、この動作の結果は、通常、動作が一度完了されると検索されるべきことが必要とされる。オペランドがＦＩＦＯバッファ４０９に書き込まれ、ＯＰコマンドがコントロールレジスタセットに書き込まれることを除き、その手順は上述されたことと同様である。フェッチアンドオピメッセージに供されるステイタスレジスタが、結果が用意されていることを指示したときには、結果を求めている受信側に対してメッセージが伝送される。

【００５５】

この発明による相互接続インターフェースにおける活動的な送信用ターミナルの数は、ハードウエアによってセットされる最大数まで動的に変動させることができる。これで付与される方法は、システムの実行能力に対するハードウエアのコストの妥協をさせることである。大きなシステム又は短い待ち時間の相互接続が特に重要であるシステムは、より多くのターミナルを用いることが可能であり、付加的なハードウエアのコストを受ける。要求される厳しさがより緩やかなより小さいシステムでは、僅かなターミナルで十分な実行能力を達成することができる。ハードウエアにおいて備えられているよりも多くのターミナルがシステムに求められたときには、そのオペレーティングシステムは、ターミナルの数を効果的に多重化させるソフトウエアを用いることができる。しかしながら、このような増大では、ノードにおける実行能力に鑑みて遥かに高いコストが加えられる。

【００５６】

ターミナルは、ダイレクトモード及びＤＭＡモードの双方で同時に活動的になることができる。ダイレクトモードの転送は、それらがＤＭＡキューの先頭に挿入されているかのようにして扱われる。この発明の好適な実施例においては、各モードに対して、個別の保護キーレジスタがコントロールレジスタセット４０４において備えられる。

【００５７】

基本的なターミナルの動作は非同期のものである。初期的には伝送が要求され、必要であるときには、幾らかの時間後に結果との集合（rendezvous）がなされる。「送信」動作に対しては、集合を明白にすることを必要としない。

【００５８】

「リモートゲット」動作では、ターミナルがより同期的なモードで動作することが要求される。この場合においては、送信用ターミナルは、戻りを「得る」までブロックされる。センド・ワーク・ディスクリプタはメッセージの本体として組み立てられ、そのメッセージは「リモートゲット」の動作タイプをもって伝送される。メッセージが受信側に到着したときには、それは送信側からの戻りメッセージを発生させるために用いられる。メッセージが受信側に到着したときには、それはこのようなメッセージのために保留されたバッファに入力され、適当な優先レベルが付与される。この発明の好適な実施例においては、要求側の優先レベルが用いられる。メッセージの機能フィールドは、値「送信」に強制される。より高い優先度のメッセージが存在しないときには、送信は実際にはローカルの相互接続インターフェースによって処理される。

【００５９】

受信ターミナルは、告知システム及びそれに関連のキューについて、殆ど単独で関係している。告知キューは、ＤＭＡの要求を処理するために用いられたものと類似の態様で管理されている。しかしながら、この場合においては、告知キユーにおいてアイテムを扱う際に進行される距離及び現在の告知のセットを処理することが終了したかどうかを、インターフェースに指示するものはローカルプロセッサである。このデータの交換は、受信用ターミナルの一部によりレジスタに書き込むことによって処理される。

【００６０】

メッセージが受信されると、このメッセージを記述したデータが、受信用ターミナルによって告知キュー３０１における次続のフリースロットに書き込まれる。メッセージエントリに含まれているものは、告知を発生した受信用ターミナルにおけるスロットのスロット数、メッセージ識別タグ、メッセージ領域におけるメッセージの位置、及び、このメッセージの長さである。告知キューにおけるエントリの処理の終了をプロセッサが指示したとき、そして、このときにのみ、この発明の好適な実施例においては中断が発生される。プロセッサによって告知が取り扱われると、告知キューのスロットにおける値がクリアされて、エントリの再使用が可能であることを、インターフェースに対して指示するようにされる。

【００６１】

この発明の好適な実施例においては、告知キュー３０１が複数のテーブルに分割される。各テーブルは異なる優先レベルを有している。プロセッサにより、優先度が最高のテーブルにおける動作が始めに完了される。

【００６２】

当業者には明かであるように、この発明は、通常のＡＳＩＣセル及び／又はゲートアレイを含むマイクロプロセッサ又は専用ハードウエアから構成することができる。

【００６３】

この発明に対する種々の修正は、これまでの説明や添付された図面から、当業者には明かなところである。従って、この発明は、特許請求の範囲の欄における記載事項によってのみ限定されるべきである。

【００６４】

以上、本発明の各実施例について詳述したが、以下、各実施例を要約する。

【００６５】

１．相互に通信する相互接続組織（１１）に接続されるように、複数のホストコンピュータ（１５，１０２）を含むマルチコンピュータデータ処理システム内の相互接続組織（１１）に対してローカルメモリ（１６，１０３，３０４）を有するホストコンピュータ（１５，１０２）を接続するための相互接続インターフェース（１７，３００）であり、各前記ホストコンピュータ（１５，１０２）が、前記相互接続組織（１１）上でメッセージを伝送することにより、同種の相互接続インターフェース（１７，３００）を有する他の前記ホストコンピュータ（１５，１０２）と通信し、各前記メッセージが１つ又は複数のパケットを有し、各パケットが、パケット本体、パケット識別情報、保護キー及びスロット識別値を有する相互接続インターフェースにおいて、前記相互接続組織（１１）及び前記ホストコンピュータ（１５，１０２）に対して接続を付与するためのインターフェース手段（１０４，３０８）と、前記相互接続組織（１１）からのメッセージを受信するために前記インターフェース手段（１０４，３０８）に接続され、夫々前記スロット識別値のある１つの可能性があるものに対応し、かつ前記ホストコンピュータ（１５，１０２）の前記ローカルメモリ（１６，１０３，３０４）の対応領域を識別する情報、保護用キー及び入来メッセージの構成を特定する情報を含む複数のレコードを格納するためのスロットテーブル手段（１０６）を有する受信ターミナル（１００，３０２，４００）と、前記相互接続インターフェースに対する入来メッセージを処理するために前記インターフェース手段（１０４，３０８）に接続され、前記パケットの前記保護キーが前記レコード（１０７）に格納された前記保護キーと合致しているときは、前記入来メッセージの前記スロット識別値に対応する前記レコード（１０７）内で識別された前記ホストコンピュータ（１

５，１０２）の前記ローカルメモリ（１６，１０３，３０４）領域に、前記パケット本体を格納させる受信制御手段（１０５，４０１）と、を具備した相互接続インターフェース（１７，３００）である。

【００６６】

２．前記受信制御手段（１０５，４０１）が、前記ホストコンピュータによる処理を待機するリストを格納するための手段と、入来メッセージの全てのパケットが前記ホストコンピュータ（１５，１０２）の前記ローカルメモリ（１６，１０３，３０４）領域に格納されるとき、処理を待機するメッセージの前記リスト（１０８，３０１）に登録させるための手段とを含む前記１に記載の相互接続インターフェース（１７，３００）である。

【００６７】

３．前記受信制御手段（１０５，４０１）が、入来メッセージの全てのパケットが前記ホストコンピュータ（１５，１０２）の前記ローカルメモリ（１６，１０３，３０４）領域に格納されたとき、そして前記ホストコンピュータ（１５，１０２）がメッセージの前記リストによって特定されたメッセージを処理していないときにのみ、前記ホストコンピュータ（１５，１０２）を中断させるための手段を含む前記２に記載の相互接続インターフェース（１７，３００）である。

【００６８】

４．前記相互接続組織（１１）上にメッセージを送信するための送信用ターミナル（３０３，４００）が、前記送信用ターミナルによって送信されるメッセージの全パケットに含まれるべき送信保護キーを格納するための保護レジスタ（４０４）と、前記相互接続組織（１１）上で伝送されるべきメッセージの位置及び前記メッセージについて意図される受信を特定するためのメッセージ識別手段と、前記相互接続組織（１１）上で前記特定されたメッセージを含むパケットを発生及び伝送するための送信制御手段（４０１）とを有し、各前記パケットが、前記送信保護キーと前記メッセージの前記受信側に取り付けられた前記相互接続インターフェース（１７，３００）上で、前記スロットテーブル手段（１０６）内のレコード（１０７）を特定するスロット識別値とを有する前記１に記載の相互接続インターフェース（１７，３００）である。

【００６９】

５．前記保護レジスタ（４０４）の内容が、前記相互接続インターフェース（１７，３００）に接続された前記ホストコンピュータ（１５，１０２）のオペレーティングシステムによってのみ変更可能である前記４に記載の相互接続インターフェース（１７，３００）である。

【００７０】

６．前記メッセージ識別手段が、入力ジョブキュー（４０８）の位置を特定するための手段と、前記入力ジョブキューの前記ワーク・ディスクリプタ・レコードの１つに対して、１つのポインタを格納するためのＤＭＡポインタ手段とを有し、前記入力ジョブキュー（４０８）が、夫々メッセージについて意図された受信、前記スロットテーブル手段（１０６）におけるレコード（１０７）及び前記ホストコンピュータ（１５，１０２）における前記メッセージの位置を特定する情報を含む複数のワーク・ディスクリプタ・レコードを有する前記４に記載の相互接続インターフェース（１７，３００）である。

【００７１】

７．前記入力ジョブキューの位置を特定するための前記手段が、前記ホストコンピュータ（１５，１０２）の前記入力ジョブキューのアドレスを格納するための手段と、前記ホストコンピュータ（１５，１０２）のオペレーティングシステムによってのみ変更可能である前記６に記載の相互接続インターフェース（１７，３００）である。

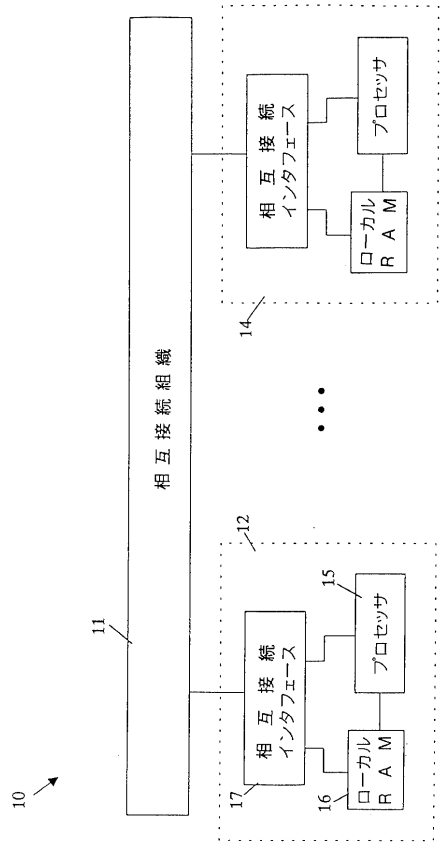
【００７２】

８．前記メッセージ識別手段が、メッセージについて意図された受信を特定する情報を格納するための受信識別手段、前記スロットテーブル手段（１０６）のレコード及び前記メッセージが格納されるバッファを含む前記４に記載の相互接続インターフェース（１７，３００）である。

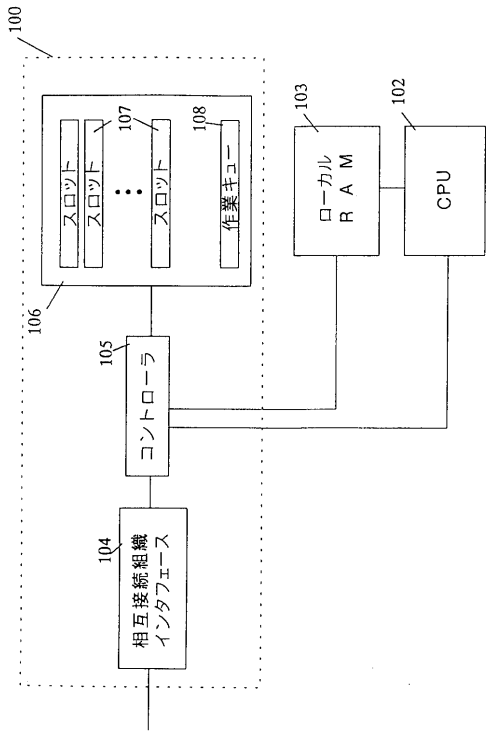
【００７３】
９．１つ又は複数のＦＩＦＯバッファ（４０９）を有し、前記受信識別手段によって特定される前記バッファが、前記ＦＩＦＯバッファの１つを有する前記８に記載の相互接続インターフェース（１７，３００）である。
【００７４】
１０．プロセッサノード及び該プロセッサノードが相互通信するためのメッセジを送受信する相互接続機構を有するマルチコンピュータデータ処理システムのプロセッサノード間を通信するための方法であって、（ａ）．第１のプロセッサノードから第２のプロセッサノードに、パケット本体、送信用保護キー及びスロット識別値を有する少なくとも一つのパケットを有するメッセージを伝送する工程と、（ｂ）．第１のプロセッサノードから入来したメッセージのパケットに含まれている送信用保護キーが、第２のプロセッサノードにおいて格納されている保護キーと合致したときに、第２のプロセッサノードにおいて決定する工程と、（ｃ）．送信用保護キーが格納された保護キーと合致したことが前記決定工程（ｂ）において決定されたとき、第１のプロセッサノードから到来したメッセージのパケット本体を第２のプロセッサノードのローカルメモリの格納領域に格納する工程であって、このローカルメモリの格納領域が第１のプロセッサノードから入来したメッセージのパケットに含まれているスロット識別値を用いて識別される工程と、を含む方法である。
【００７５】
１１．前記スロット識別値が第２のプロセッサノード内のスロットテーブルのレコードを識別し、前記レコードが少なくともローカルメモリの格納領域及び格納されている保護キーを識別する情報を含む前記１０に記載の方法である。
【００７６】
１２．前記レコードが、入来するメッセージの構成を特定する情報を含む前記１１に記載の方法である。
【００７７】
１３．前記スロットテーブルの各レコードが、前記相互接続組織に接続されたプロセッサノードの一つに関連される前記１１に記載の方法である。
【００７８】
【発明の効果】
以上詳細に説明したように、本発明によれば、ローカルメモリを有する複数のホストコンピュータが相互に相互接続組織を介して通信を行うに際に、パケット本体、パケット識別情報、保護キー、スロット識別値を有する１個又は複数のパケットを含むメッセージをインターフェース手段及びローカルメモリの対応領域を識別する情報、保護キー、入来メッセージを識別する情報を含むレコードを格納するスロットテーブルを有する受信ターミナルからなる相互接続インターフェースを介して相互接続組織に伝送し、パケットの保護キーがレコードに含まれる保護キーと合致している場合には、入来メッセージのスロット識別値に対応してレコードにおいて識別されたホストコンピュータのローカルメモリの領域にパケット本体を相互接続インターフェースの受信制御手段で格納するようにしたので、受信側がビジー状態でも、入来メッセージで中断されることがなく、且つメッセージの失われる数を減少できる。また、メッセージ伝送中に、他のプロセスからの偶発的あるいは悪意のあるアタックから各プロセスを保護するためのオペレーティングシステムの呼び出しが不要であるとともに、入来メッセージを受信側で格納する場所を決定するのに要する時間を短縮できる。さらに、規則的なメッセージを用いるのと同じ相互接続組織を使用して集積化とプロセスの保護及び待ち時間の短縮が可能となる。
【図面の簡単な説明】
【図１】本発明による相互接続インターフェースが用いられる典型的なマルチプロセッサのブロック図である。
【図２】メッセージを受信する本発明による相互接続インターフェース１００の部分についてのより詳細にされたブロック図である。

【図３】本発明による相互接続インターフェースの好適な実施例のブロック図である。
【図４】本発明による相互接続インターフェースにおける送信用ターミナルのブロック図である。
【符号の説明】
１０ データ処理システム
１１ 相互接続組織
１２，１４ ノード
１５ プロセッサ
１６，１０３ ローカルＲＡＭ
１７，１００，３００ 相互接続インターフェース
１０２ ＣＰＵ
１０４ 相互接続組織インターフェース
１０５，４０１ コントローラ
１０６ ローカル・メモリ
１０７ スロット
１０８ 作業キュー
３０２ 受信用ターミナル
３０３ 送信用ターミナル
３０４ ホスト・コンピュータ・メモリ
３０８ インターフェース・ハードウェア
４００ 送信用ターミナル
４０８ 入力キュー
４０９ ＦＩＦＯバッファ

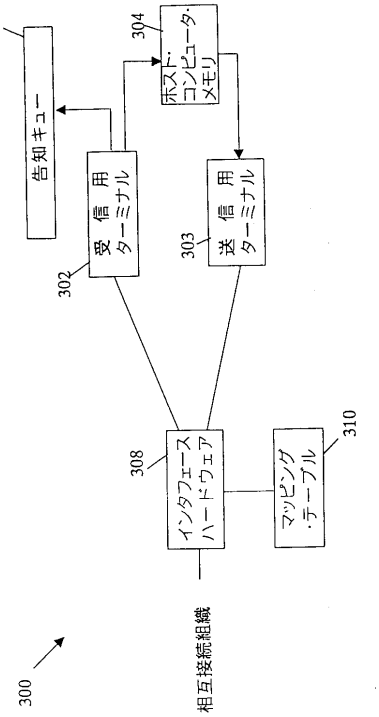
【図１】



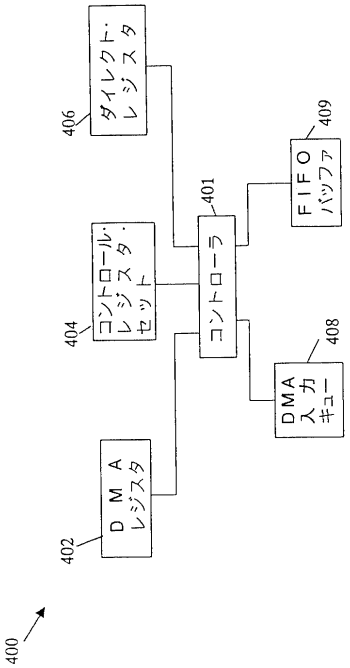

【図２】



【図３】


300
301
告知キュー
302
受信用ターミナル
303
送信用ターミナル
304
ホスト・コンピュータ・メモリ
308
インタフェースハードウェア
310
マッピング・テーブル
相互接続組織

【図４】


400
401
コントローラ
402
DMAレジスタ
404
コントロール・レジスタ・セット
406
ダイレクト・レジスタ
408
DMA入力キュー
409
FIFOバッファ

フロントページの続き

(56)参考文献 特開平０３－２０４０６５（ＪＰ，Ａ）
米国特許第０４７７４７０７（ＵＳ，Ａ）
米国特許第０５４４８６９８（ＵＳ，Ａ）

(58)調査した分野(Int.Cl.[7]，ＤＢ名)
G06F 13/00 353
G06F 15/177 670
H04L 12/66
WPI(DIALOG)