(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第3601955号
(P3601955)
(45)発行日　平成16年12月15日(2004.12.15)　(24)登録日　平成16年10月1日(2004.10.1)

(51)Int.Cl.[7]　F I
G06F 15/17　G06F 15/17
G06F 15/173　G06F 15/173

請求項の数 9　(全 25 頁)

(21)出願番号　特願平9-290597
(22)出願日　平成9年10月23日(1997.10.23)
(65)公開番号　特開平11-126196
(43)公開日　平成11年5月11日(1999.5.11)
審査請求日　平成16年3月30日(2004.3.30)

(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　100068504
弁理士　小川　勝男
(74)代理人　100061893
弁理士　高橋　明夫
(74)代理人　100086656
弁理士　田中　恭助
(72)発明者　保田　淑子
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　藤井　啓明
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
最終頁に続く

(54)【発明の名称】データ転送方法およびそれに適した計算機システム

(57)【特許請求の範囲】
【請求項１】
相互結合ネットワークで接続された複数の要素プロセッサを有し、各要素プロセッサ内には、ユーザプロセスと交信するメッセージパッシングライブラリと、そのメッセージパッシングライブラリと交信するメモリ間直接転送ライブラリとが組み込まれ、各要素プロセッサは、プロセッサと、メモリと、上記相互結合ネットワークとの間でメッセージを交換するためのネットワークインタフェース回路とを有する計算機システムにおいて、
送信側の要素プロセッサで走行中のユーザプロセスにより、その要素プロセッサ内のメッセージパッシングライブラリに対して発行されたデータ送信要求が要求する送信データと、そのデータ送信要求が要求する、メッセージパッシングライブラリにより定められた、受信側の要素プロセッサに送信されるべき付加情報との送信を、そのメッセージパッシングライブラリからその要素プロセッサ内のメモリ間直接転送ライブラリに対して要求し、
そのメモリ間直接転送ライブラリにより、上記要求された送信データおよび付加情報と、それらの受信の完了を示す制御情報を書き込むべき領域を指定するための、受信側の要素プロセッサがあらかじめ決定した受信側アドレス情報とを含むメッセージの送信を上記ネットワークインタフェース回路に対して要求し、
上記ネットワークインタフェース回路により、上記メッセージを組立て、受信側の要素プロセッサに宛てて上記相互結合ネットワークに送信し、
上記メッセージの送信後に、そのメモリ間直接転送ライブラリが決定したアドレスを有する、上記メモリ内の記憶位置に送信完了を示す制御情報を上記ネットワークインタフェー

ス回路により書き込むステップからなるデータ転送方法。
【請求項２】
受信側の要素プロセッサ内のネットワークインタフェース回路により、上記メッセージ内の送信データと付加情報を、受信側の要素プロセッサのメモリ内の、上記受信側アドレス情報により定まる領域に書き込み、
上記書き込みの終了後に、受信側の要素プロセッサ内のネットワークインタフェース回路により、受信完了を示す制御情報を、上記メモリ内の、上記受信側アドレス情報により定まる領域に書き込むステップをさらに有する請求項１記載のデータ転送方法。
【請求項３】
上記データ送信要求は、上記送信データに関する第１のアドレス情報とデータ長情報と上記付加情報を指定し、
上記メッセージの送信を要求するステップは、
送信側の要素プロセッサ内の上記メモリ間直接転送ライブラリにより、上記要求された付加情報を、そのメモリ間直接転送ライブラリが付加情報と送信完了を示す制御情報を書き込むための領域として決定した、上記メモリ内の領域に書き込み、
そのメモリ間直接転送ライブラリにより、上記第１のアドレス情報、上記送信データ長情報、上記付加情報が書き込まれた上記領域を指定する第２のアドレス情報および上記受信側アドレス情報とを指定するメッセージ送信要求を送信側の要素プロセッサ内の上記ネットワークインタフェース回路に対して発行するステップからなり、
上記メッセージを送信するステップは、
上記ネットワークインタフェース回路により、そのメッセージ送信要求で指定された上記第１のアドレス情報と上記データ長情報との組および上記第２のアドレス情報にそれぞれ基づいて、上記送信データおよび上記付加情報を送信側の要素プロセッサのメモリより読み出し、
上記ネットワークインタフェース回路により、そのメッセージ送信要求が指定した上記受信側アドレス情報と上記データ長情報、上記読み出された送信データおよび付加情報を含むメッセージを生成し、上記ネットワークを介して受信側の要素プロセッサに送信するステップを有し、
上記送信完了を示す制御情報を書き込むステップは、上記第２のアドレス情報に基づいて、上記付加情報が記憶された領域と異なる記憶位置に送信完了を示す制御情報を書き込むステップを有する請求項１記載のデータ送信方法。
【請求項４】
受信側の要素プロセッサ内のネットワークインタフェース回路により、上記メッセージ内の送信データと付加情報を、受信側の要素プロセッサのメモリ内の、上記受信側アドレス情報により定まる領域に書き込み、
上記書き込みの終了後に、受信側の要素プロセッサ内のネットワークインタフェース回路により、受信完了を示す制御情報を、上記メモリ内の、上記受信側アドレス情報により定まる領域に書き込むステップをさらに有し、
上記受信側のアドレス情報は、上記メッセージ内の送信データを書き込むべき第３のアドレス情報と受信完了を示す制御情報を書き込む第４のアドレス情報からなり、
上記メッセージ内の送信データと付加情報を書き込むステップは、上記第３、第４のアドレス情報に基づいて、上記送信データおよび上記付加情報を、受信側の要素プロセッサのメモリに書き込むステップからなり、
上記受信完了を示す制御情報を書き込むステップは、上記第４のアドレス情報に基づいて、受信側の要素プロセッサのメモリ内の、上記付加情報が書き込まれた領域と異なる領域に書き込むステップからなる請求項３記載のデータ転送方法。
【請求項５】
送信側の要素プロセッサで走行中の他のユーザプロセスによりその要素プロセッサ内の上記メモリ間直接転送ライブラリに対して発行された他のデータ送信要求が要求する送信データと、受信側の要素プロセッサがあらかじめ決定した、受信データと受信完了を示す制

御情報を書き込むべき領域を指定するための受信側アドレス情報とを含むメッセージの送信を、そのメモリ間直接転送ライブラリにより、送信側の要素プロセッサの上記ネットワークインタフェース回路に対して要求し、
上記メッセージを上記ネットワークインタフェース回路により組立て、受信側の要素プロセッサに宛てて上記相互結合ネットワークに送信し、
上記メッセージの送信後に、そのメモリ間直接転送ライブラリが決定したアドレスを有する、上記メモリ内の記憶位置に送信完了を示す制御情報を上記ネットワークインタフェース回路により書き込み、
受信側の要素プロセッサ内のネットワークインタフェース回路により、上記メッセージ内の送信データを、受信側の要素プロセッサのメモリ内の、上記受信側アドレス情報により定まる領域に書き込み、
上記書き込みの終了後に、受信側の要素プロセッサ内のネットワークインタフェース回路により、受信完了を示す制御情報を、上記メモリ内の、上記受信側アドレス情報により定まる領域に書き込むステップをさらに有する請求項２記載のデータ転送方法。

【請求項６】
相互結合ネットワークで接続された複数の要素プロセッサを有し、
各要素プロセッサは、プロセッサと、メモリと、上記相互結合ネットワークとの間でメッセージを交換するためのネットワークインタフェース回路とを有し、
上記ネットワークインタフェース回路は、
転送すべきメッセージに関する情報として上記プロセッサにより供給される、第１のアドレス情報とデータ長情報との組および第２のアドレス情報とに基づいて、それぞれ送信されるべき第１のデータおよびそのデータとともに送信されるべき第２のデータを上記メモリから読み出すメモリアクセス回路と、
読みだされた第１，第２のデータと、転送すべきメッセージに関する他の情報として上記プロセッサにより供給される第３，第４のアドレス情報を含む一つのメッセージを生成し、上記相互結合ネットワークに送信する回路とを有し、
上記メモリアクセス回路は、上記メッセージの送信後に、上記第２のアドレス情報に基づいて、上記送信完了を示す制御情報を上記メモリ内の、上記第２のデータが記憶されている記憶位置と異なる記憶位置に書き込み、
該第１のアドレス情報は、上記第１のデータを保持する、上記メモリ内の領域のアドレスを指示し、上記データ長情報は、上記第１のデータの長さを指定し、該第２のアドレス情報は、上記第２のデータを保持し、かつ、該第１のデータの送信完了を示す制御情報をさらに保持すべき、上記メモリ内の領域のアドレスを指示し、該第３のアドレス情報は、第１のデータを書き込むべき、受信側の要素プロセッサのメモリ内の領域のアドレスを指示し、該第４のアドレス情報は、上記第１のデータおよび該第１のデータの受信完了を示す制御情報とを格納するための、受信側の要素プロセッサのメモリ内の領域のアドレスを指示する計算機システム。

【請求項７】
上記ネットワークインタフェース回路は、他の要素プロセッサから送信されたメッセージを上記相互結合ネットワークから受信する回路をさらに有し、
上記メモリアクセス回路は、受信されたメッセージ内の上記第３のアドレス情報に基づいて、受信されたメッセージ内の上記第１のデータを上記メモリに書き込み、受信されたメッセージ内の上記第４のアドレス情報に基づいて、受信されたメッセージ内の上記第２のデータを上記メモリに書き込み、上記受信されたメッセージ内の上記第１，第２のデータの書き込みの終了後に、上記第４のアドレス情報に基づいて、上記第１のデータの受信完了を示す制御情報を、上記メモリ内の、受信された上記第２のデータが記憶されている記憶位置と異なる記憶位置に書き込む回路を有する請求項６記載の計算機システム。

【請求項８】
上記転送制御情報はモードビットをさらに有し、
上記メモリアクセス回路は、上記モードビットが第１の値の時に、上記第２のデータの読

み出しを実行し、上記モードビットが第２の値の時に、上記第２のデータの読み出しを実行せず、
上記メッセージ送信回路は、上記モードビットが第１の値の時に、上記第２のデータを含むメッセージを生成し、上記モードビットが第２の値の時に、上記第２のデータを含まないメッセージを生成し、
上記メモリアクセス回路は、上記モードビットが第１の値の時には、上記メッセージの送信後に、上記第２のアドレス情報に基づいて、上記送信完了を示す制御情報を、上記メモリ内の、上記第２のデータが記憶されている記憶位置と異なる記憶位置に書き込み、上記モードビットが第２の値の時には、上記メッセージの送信後に、上記送信完了を示す制御情報を、上記第２のアドレス情報に依存する、上記メモリ内の記憶位置に書き込む回路を有する請求項６記載の計算機システム。
【請求項９】
上記メッセージ送信が生成するメッセージは、上記モードビットを含み、
上記ネットワークインタフェース回路は、他の要素プロセッサから送信されたメッセージを上記相互結合ネットワークから受信する回路をさらに有し、
上記メモリアクセス回路は、
受信されたメッセージ内のモードビットが第１の値の時には、受信されたメッセージ内の上記第３のアドレス情報に基づいて、受信されたメッセージ内の上記第１のデータを上記メモリに書き込み、上記受信されたメッセージ内の上記第４のアドレス情報に基づいて、受信されたメッセージ内の上記第２のデータを上記メモリに書き込み、上記受信されたメッセージ内の上記第１，第２のデータの書き込みの終了後に、上記第４のアドレス情報に基づいて、上記第１のデータの受信完了を示す制御情報を、上記メモリ内の、受信された上記第２のデータが記憶されている記憶位置と異なる記憶位置に書き込み、
受信されたメッセージ内のモードビットが第２の値の時には、受信されたメッセージ内の上記第１のデータを、受信されたメッセージ内の上記第３のアドレス情報に基づいて、上記メモリ内の記憶位置に書き込み、上記受信されたメッセージ内の上記第１のデータの書き込みの終了後に、上記第１のデータの受信完了を示す制御情報を、上記第４のアドレス情報に依存する、上記メモリ内の記憶位置に書き込む回路を有する請求項８記載の計算機システム。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】
本発明は、相互結合ネットワークを介して接続された複数の要素プロセッサ間でのデータ転送方法およびそれに適した計算機システムに関する。
【０００２】
【従来の技術】
従来、並列計算機は、ローカルメモリと命令プロセッサから構成される複数の要素プロセッサを相互結合ネットワークで結合した構成をとっている。一般に、このような形態の並列計算機は、分散メモリ型並列計算機と呼ばれる。各要素プロセッサは、相互結合ネットワークを介して個々のローカルメモリに格納されているデータの授受を行い、並列に処理を実行する。
【０００３】
一般に、分散メモリ型の並列計算機では、メッセージパッシングと呼ぶプログラミングモデルを用いてデータ転送を実現する。メッセージパッシングモデルでは、ユーザが並列プログラム中に明示的に送信（ＳＥＮＤ）手続きおよび受信（ＲＥＣＥＩＶＥ）手続きを記して、要素プロセッサ間で必要になるデータの授受をメッセージのやりとりという形で行う。命令プロセッサはこれらの通信手続きを解析して、相互結合ネットワークにデータを送信したり、相互結合ネットワークからデータを受信しながら処理を進める。送信元の要素プロセッサは、転送先の要素プロセッサ番号を指定してメッセージを転送し、転送先の要素プロセッサでメッセージをバッファリングする。メッセージパシングモデルでは、メ

ッセージ通信に伴ってデータのバッファリングやフロー制御が必要となり、送受信オーバヘッドが大きくなってしまう。

【０００４】

この送受信オーバヘッドを削減するために、近年複数の並列計算機において、要素プロセッサのローカルメモリの内容を、メッセージの生成あるいは受信を行う送受信回路が直接アクセスするメモリ間直接転送方法が使用されている。この方法を実行する代表例は、ＰＵＴ／ＧＥＴ通信である。例えば、”情報処理学会並列処理シンポジウムＪＳＰＰ’９５、”ＰＰ．　２３３－２４０（１９９５年５月）参照。メモリ間直接転送方法では、各要素プロセッサのローカルメモリからの送信データの読み出しあるいはそのメモリへの受信データの書き込みをメッセージの生成あるいは受信を行う送受信回路が直接実行するため、これらのデータをＯＳ管理の領域にコピーする必要がなく、このコピーに由来するオーバーヘッドを削減できる。

【０００５】

しかしながら、このようなＰＵＴ／ＧＥＴ通信を実際に実行する通信ライブラリは、各並列計算機メーカーや研究機関が独自に開発しているため、それを用いて作成された並列プログラムを他機種へ移植することは困難であった。この問題点を解決するために、ＭＰＩ（Ｍｅｓｓａｇｅ　Ｐａｓｓｉｎｇ　Ｉｎｔｅｒｆａｃｅ）に代表される、メッセージパッシングライブラリの標準化が進みつつある。ＭＰＩは、米国各大学および並列計算機メーカーがメッセージパッシングインタフェース標準化団体ＭＰＩ　Ｆｏｒｕｍを組織し、その研究成果をまとめた仕様である。この仕様に基づいて作成されたライブラリ（以下ＭＰＩライブラリと呼ぶことがある）は、今後の並列プログラム開発支援ライブラリの主流になると考えられる。ＭＰＩライブラリを用いて記述された並列プログラムは、異機種間で変更なしに走らせることができる。各計算機メーカーは、自社の並列計算機上で高性能を達成するようにＭＰＩライブラリを開発している。

【０００６】

上記ＭＰＩ仕様は、ＰＵＴ／ＧＥＴ通信に関する仕様を含んでいない。しかし、データ転送の高速化のためには、ＰＵＴ／ＧＥＴ通信を併用することが重要である。このために、並列計算機メーカー等は、ＰＵＴ／ＧＥＴ通信ライブラリを使用可能にしたＭＰＩライブラリを開発している。たとえば、本出願人による、「並列計算機ＳＲ２２Ｄ支援ライブラリ」参照。したがって、ＭＰＩライブラリは、各計算機メーカーごとに異なるものであるが、ユーザプログラムから見れば、ＭＰＩライブラリとの間のインタフェースは、いずれの計算機メーカのＭＰＩライブラリに対しても同じである。従って、そのユーザプログラムは、いずれの計算機のメーカの上でも実行できることになる。

【０００７】

ＭＰＩライブラリを用いるユーザプログラムは、データを送信する時点で、ＭＰＩライブラリをコールする。従来は、ユーザプログラムがＰＵＴ／ＧＥＴ通信ライブラリを使用するときには、ユーザプログラムは送信すべきユーザデータおよびデータ長をこのコール文の引数でもって指定すればよい。しかし、ＭＰＩライブラリを使用するときには、ユーザプログラムは、このコール文の中でこのＭＰＩライブラリにより定められた付加情報を引数としてさらに指定する必要がある。この付加情報は、送信先のプロセスの識別子、プロセスグループ識別子等を含む固定長のデータであり、メッセージの送信先の要素プロセッサにおいて、受信したメッセージがそこで実行中のユーザプロセスが発行する受信要求が要求したメッセージか否かの識別に使用される。以下、この情報をＭＰＩ付加情報とも呼ぶ。従来のＭＰＩライブラリとＰＵＴ／ＧＥＴライブラリを併用した通信方法では、ユーザデータおよびＭＰＩ付加情報を異なる二つのメッセージにより転送していた。

【０００８】

【発明が解決しようとする課題】

従来のように、同じ転送先プロセッサに対しユーザデータおよびＭＰＩ付加情報を２つのメッセージとして転送する場合、各メッセージの転送に異なる転送制御情報が必要になる。その結果、これらの情報の生成も２度行わなければならない。このために、従来のＭＰ

Ｉライブラリを使用したデータ転送では、ユーザプログラムがデータの転送を要求してから、実際に転送が開始されるまでの遅延時間（転送レイテンシと呼ばれる）が大きい。

【０００９】

さらに、これらの２種のデータに対して別々のメッセージとして送信処理、受信処理を行うと、メッセージ数に比例してローカルメモリに対するアクセス回数（転送制御情報の読み出し、転送データの読み出し、フラグの書き込み）が増加してしまう。

【００１０】

本発明の目的は、以上の問題を減少させ、より高速にデータを転送できるデータ転送方法を提供することにある。

【００１１】

【発明を解決するための手段】

上記の目的を達成するために、本発明によるデータ転送方法は、

送信側の要素プロセッサで走行中のユーザプロセスから発行されたデータ送信要求が要求する、送信データとそれに関連する付加情報とを、送信元のメッセージパッシングライブラリから送信元のメモリ間直接転送ライブラリに通知し、

そのメモリ間直接転送ライブラリにより、上記送信データおよび付加情報と、それらの受信の完了を示す制御情報を書き込むべき領域を指定するための、受信側の要素プロセッサがあらかじめ決定した受信側アドレス情報とを含むメッセージの送信をネットワークインタフェース回路に対して要求し、

上記ネットワークインタフェース回路により、上記メッセージを組立て、受信側の要素プロセッサに宛てて上記相互結合ネットワークに送信し、

上記メッセージの送信後に、そのメモリ間直接転送ライブラリが決定したアドレスを有する、上記メモリ内の記憶位置に送信完了を示す制御情報を上記ネットワークインタフェース回路により書き込む。

【００１２】

より具体的には、上記メモリ間直接転送ライブラリは、上記付加情報を上記メモリ内の領域に書き込み、ユーザが指定した送信データの記憶位置を示す第１のアドレスとそのデータの長さと、上記付加情報の記憶位置を示す第２のアドレスと、送信先の要素プロセッサのメモリにおける、送信データの記憶位置を指定する第３のアドレスと、上記付加情報を記憶する位置を示す第４のアドレス等を指定し、これらの情報を含むメッセージの送信を上記ネットワークインタフェース回路に要求する。

【００１３】

この回路は、上記第１のアドレスと上記データ長により送信データを読み出し、上記第２のアドレスにより付加情報を読み出し、これらの送信データ、付加情報および上記第３，第４のアドレスを含むメッセージを生成し、送信先にあてて送信する。この送信の完了後に、上記第２のアドレスを使用して、送信完了を示す制御情報をメモリ内の、上記付加情報の書き込み位置と異なる位置、具体的には、付加情報が書き込まれた記憶位置の次の記憶位置に書き込む。

【００１４】

さらに、受信側の要素プロセッサにおいても、ネットワークインタフェース回路が、上記メッセージ内の上記送信データと付加情報とをそれぞれ上記第３、第４のメモリアドレスが指定する記憶位置に書き込むとともに、この書き込みの完了後に、受信完了を示す制御情報を上記第４のアドレスの基づいて上記付加情報の書き込み位置と異なる記憶位置に書き込む。

【００１５】

本発明のより具体的な態様では、ユーザプロセスが、メッセージパッシングライブラリを介して行う上記の転送とともに、他のユーザプロセスはメッセージパッシングライブラリを介さないでメモリ間直接転送ライブラリにデータの送信要求を発行することができるようになっている。この場合には、メモリ間直接転送ライブラリとネットワークインタフェース回路は、上に述べた処理における、付加情報が存在しない場合の処理と基本的に同じ

処理を行う。
【００１６】
【発明の実施の形態】
以下、本発明に係る計算機システム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。
【００１７】
＜発明の実施の形態＞
図１に、本発明における並列計算機の概略構成を示す。図中、１０１～１０４は並列計算機を構成する要素プロセッサ、１０５は相互結合ネットワークである。要素プロセッサ１０１～１０４は、相互結合ネットワーク１０５に接続し、相互結合ネットワーク１０５を介して、要素プロセッサ間でデータの授受を行う。相互結合ネットワーク１０５の構成方法（トポロジ）は、クロスバ結合、格子結合、リング結合、多段結合等多種存在するが、本発明は、これらのいずれにも適用可能であり、特定の相互結合ネットワークトポロジに限定されない。図３に要素プロセッサ１０１の概略構成を示す。図中、３０１は命令プロセッサ、３０２はキャッシュローカルメモリ、３０３はストレージコントローラ、３０４はローカルメモリ、３０５はネットワークインタフェース回路、３０６はＩ／Ｏインタフェース回路である。この並列計算機は、個々の要素プロセッサがそれぞれ固有のローカルメモリ３０４を有する分散ローカルメモリ型の並列計算機である。
【００１８】
各要素プロセッサは、他の要素プロセッサとの間でメッセージパッシングによる通信を実行するように構成されている。すなわち、各要素プロセッサは、標準のメッセージパッシングインタフェース、たとえばＭＰＩを有するメッセージパッシングライブラリ（以下、ＭＰＩライブラリと呼ぶ）と、このライブラリと交信して自要素プロセッサ内のローカルメモリとの間で直接データの授受を行うメモリ間直接転送を実行するライブラリとして、ＰＵＴ／ＧＥＴ型通信を実行するためのライブラリ（以下、ＰＵＴ／ＧＥＴ型ライブラリと呼ぶ）と、ＰＵＴ／ＧＥＴ型ライブラリからのコマンドにより起動されるネットワークインタフェース回路３０５を有している。なお、本発明は、この特定のメッセージパッシングライブラリに限定されるのではなく、他のメッセージパッシングライブラリたとえばＰＶＭ、ＰＡＲＭＡＣＳとして知られているライブラリも適用できる。
【００１９】
本実施の形態では、各要素プロセッサ内のユーザプロセスがＭＰＩライブラリに対してデータ送信要求を発行したときに、ＭＰＩライブラリ、ＰＵＴ／ＧＥＴ型ライブラリおよびネットワークインタフェース回路３０５は、協同してユーザデータとＭＰＩ付加情報を一つのメッセージにて転送し、さらに他の要素プロセッサからユーザデータとＭＰＩ付加情報を含むメッセージを受信したときに、ネットワークインタフェース回路３０５は、これらのデータを区分してローカルメモリに書き込むところに特徴がある。この付加情報は、ＭＰＩライブラリを介した通信のために使用されるもので、ＭＰＩライブラリが指定した形式の、データ送信に関連する複数の情報からなり、それぞれの情報は、ユーザプロセスにより指定される。具体的には、既に例示したように、この付加情報は、受信側のユーザプロセスの識別子、プロセスグループの識別子等を含む。
【００２０】
より具体的には、各要素プロセッサで実行中のユーザプロセスからＭＰＩライブラリに対する送信要求が発行されたときに、ＭＰＩライブラリは、その送信要求が指定するユーザデータと付加情報の送信をＰＵＴ／ＧＥＴ型ライブラリに要求する。ＰＵＴ／ＧＥＴ型ライブラリにその送信要求が指定する付加情報をローカルメモリに書き込み、そのユーザデータとＭＰＩ付加情報の両方を一つのメッセージとして転送するための転送制御情報を生成し、ローカルメモリ３０４に書き込み、その後転送制御情報によるユーザデータおよび付加情報の送信をネットワークインタフェース回路３０５に要求する。
【００２１】
ネットワークインタフェース回路３０５は、この送信要求に応答して、転送制御情報に従

ってユーザデータとＭＰＩ付加情報を含む一つのメッセージを組み立て、受信側の要素プロセッサに相互結合ネットワーク１０５を介して転送する。受信側の要素プロセッサでは、ネットワークインタフェース回路３０５は、このメッセージを受信すると、メッセージに含まれたユーザデータおよびＭＰＩ付加情報をメッセージのヘッダ内の転送制御情報が指定するローカルメモリ内の二つのアドレスに書き込み、それぞれを受信側のユーザプロセスおよび受信側のＭＰＩライブラリに引き渡す。

【００２２】

図２に転送制御情報の例を示す。転送制御情報２００には、ＧＥＴあるいはＰＵＴ動作の場合に使用される転送先プロセッサ番号２０１、ＰＵＴ動作の場合に送信されるユーザデータが格納されているローカルメモリ領域の先頭アドレスである送信データアドレス２０３、ＰＵＴ動作の場合に送信完了フラグを書き込むローカルメモリ領域の先頭アドレスである送信フラグアドレス２０４、ＧＥＴあるいはＰＵＴ動作の場合に使用される、転送されるデータの長さである転送データ長２０５、受信側の要素プロセッサにおいて受信データを書き込むローカルメモリ領域の先頭アドレスである受信データアドレス２０６、その要素プロセッサにおいて、その受信データに対する受信完了フラグを書き込むローカルメモリ領域の先頭アドレスである受信フラグアドレス２０７、その他通信処理に必要な情報２０８等を格納する。

【００２３】

さらに、本実施の形態では、モードビット２０２がセットされていない場合、送信フラグアドレス２０４および受信フラグアドレス２０７は、それぞれＰＵＴ動作時の送信完了フラグおよびＧＥＴ動作時の受信完了フラグを書き込むローカルメモリアドレスを指定する。しかし、モードビット２０２がセットされている場合、送信フラグアドレスフィールド２０４は、ＰＵＴ動作時にＭＰＩ付加情報を読み出すべきローカルメモリアドレスを指定するのに使用され、受信フラグアドレスフィールド２０７は、ＧＥＴ動作時に受信したＭＰＩ付加情報を書き込むべきローカルメモリアドレスを指定するのに使用される。

【００２４】

この結果、ＰＵＴ動作時に送信完了フラグを書き込むべきローカルメモリアドレスが転送制御情報２００により指定されなくなるが、本実施の形態では、あらかじめセットされたＭＰＩ付加情報サイズを送信フラグアドレスに加算し、その結果得られるアドレスにユーザデータおよびＭＰＩ付加情報という二つのデータの送信完了フラグを書き込む。同様に、ＧＥＴ動作時には、ＧＥＴ動作時には、あらかじめセットされたＭＰＩ付加情報サイズを受信フラグアドレスフィールドに加算し、その結果得られるアドレスに２種類のデータの受信完了フラグを書き込む。これにより、１つの転送制御情報２００を用いてユーザデータとＭＰＩ付加情報という２つの種類のデータを１つのメッセージで送信または受信し、従来と同様に送信完了フラグあるいは受信完了フラグもローカルメモリに書き込むことができる。

【００２５】

命令プロセッサ３０１は、プログラム処理を行うユニットである。キャッシュローカルメモリ３０２は、命令プロセッサ３０１に付随する、高速かつ小容量のローカルメモリである。ネットワークインタフェース回路３０５は、相互結合ネットワーク１０５に接続し、命令プロセッサ３０１からの指示に従って、ローカルメモリ分散型の並列計算機の特徴であるデータ転送処理を命令プロセッサ３０１のプログラム処理とは独立して行うユニットである。ストレージコントローラ３０３は、命令プロセッサ３０１、ネットワークインタフェース回路３０５およびＩ／Ｏインタフェース回路３０６から発行されるデータアクセス要求に従って、適当な記憶媒体にアクセスする。ローカルメモリ３０４は、ストレージコントローラ３０３で制御され、データ等を格納する。命令プロセッサ３０１およびネットワークインタフェース回路３０５は独立に動作するため、ストレージコントローラ３０３は、命令プロセッサ３０１からローカルメモリ３０４へのアクセス要求を処理するのと同時に、ネットワークインタフェース回路３０５からのデータ転送に伴うローカルメモリ３０４へのアクセスも処理する。Ｉ／Ｏインタフェース回路３０６は、ストレージコント

ローラ３０３からのアクセス要求に従って、Ｉ／Ｏ装置にアクセスする。Ｉ／Ｏインタフェース回路３０６は、要素プロセッサの構成によっては、存在しない場合もある。

【００２６】

図４に示すように、ストレージコントローラ３０３は、命令プロセッサインタフェース回路４０１、アドレス解析部４０２、メモリアクセスインタフェース回路４０３およびデータ転送インタフェース回路４０４で構成される。命令プロセッサインタフェース回路４０１は、命令プロセッサ３０１からローカルメモリ３０４へのアクセスおよび命令プロセッサ３０１からネットワークインタフェース回路３０５へのコマンド発行というトランザクションを線４０１Ｓから受け取る。通常、このコマンドは、ネットワークインタフェース回路３０５内部の制御レジスタへのアクセス要求である。命令プロセッサインタフェース回路４０１は、このトランザクションへの返答、ストレージコントローラ３０３やネットワークインタフェース回路３０５で検出した割り込み要因を線４０２Ｓを介して命令プロセッサ３０１へ伝える。このトランザクションの応答は、たとえば、ローカルメモリからの読み出しデータである。

【００２７】

アドレス解析部４０２は、命令プロセッサ３０１が発行した、ローカルメモリアクセス要求およびネットワークインタフェース回路３０５へのコマンドを線４０３Ｓを介して受け取り、そのアクセス要求あるいはコマンドが指定するアクセス先アドレスを解析する。ローカルメモリアクセス要求は線４０４Ｓを介してメモリアクセスインタフェース回路４０３に伝えられる。また、ネットワークインタフェース回路３０５へのコマンドは、線４０６Ｓを介してデータ転送インタフェース回路４０４に伝達される。

【００２８】

メモリアクセスインタフェース回路４０３は、アドレス解析部４０２からのローカルメモリアクセス要求を線４０４Ｓを介して受け、線４０７Ｓを介してローカルメモリ３０４に伝達する。ローカルメモリアクセス要求がローカルメモリからの読み出し要求であった場合、この読み出し要求が指定するデータがローカルメモリ３０４から線４０８Ｓを介して伝達される。読み出しデータは、メモリアクセスインタフェース回路４０３から線４０９Ｓを介して命令プロセッサインタフェース回路４０１に伝達され、線４０２Ｓを介して命令プロセッサ３０１に伝達される。また、ローカルメモリアクセスインターフェース４０３は、データ転送に関わるローカルメモリアクセスも処理する。データ転送処理に関わるローカルメモリアクセスは、データ転送インタフェース回路４０４から線４１０Ｓを介して伝達される。メモリアクセスインタフェース回路４０３は、アドレス解析部４０２からローカルメモリアクセス要求が伝達された時と同様に、ローカルメモリアクセス要求をローカルメモリ３０４に対して発行し、読み出しアクセスに対しては、読み出しデータを線４１１Ｓを介してデータ転送インタフェース回路４０４に返送する。

【００２９】

データ転送インタフェース回路４０４は、アドレス解析部４０２から線４０６Ｓを介して伝達されるネットワークインタフェース回路３０５へのコマンドを受け取り、線４１２Ｓを介してネットワークインタフェース回路３０５に伝達する。ネットワークインタフェース回路３０５からは、線４１３Ｓを介してコマンドに対する返答およびデータ転送に関わるローカルメモリアクセス要求が伝達される。データ転送インタフェース回路４０４は、前記コマンドに対する返答を線４１４Ｓを介して命令プロセッサインタフェース回路４０１に伝達し、線４０２Ｓを介して命令プロセッサ３０１に伝達する。ローカルメモリアクセス要求は、線４１０Ｓを介してメモリアクセスインタフェース回路４０３に伝達する。ローカルメモリ読み出しデータは、線４１１Ｓを介してデータ転送インタフェース回路４０４に伝達され、データ転送インタフェース回路４０４から線４１２Ｓを介してネットワークインタフェース回路３０５に伝達される。データ転送インタフェース回路４０４は、ネットワークインタフェース回路３０５内部で発生した割り込み伝達要求を受ける場合もある。この場合、割り込み伝達要求は線４１４Ｓを介して命令プロセッサインタフェース回路４０１に伝達され、さらに命令プロセッサ３０１に伝達される。

【００３０】
ネットワークインタフェース回路３０５は、コマンド受信部４０５、コマンド処理部４０６、メッセージ生成部４０７、メッセージ送信部４０８、メッセージ受信部４０９、メッセージ分解部４１０およびコマンド送信部４１１で構成される。コマンド受信部４０５は、ストレージコントローラ３０３から線４１２Ｓを介して、ネットワークインタフェース回路３０５内部の制御レジスタへのアクセスあるいはネットワークインタフェース回路３０５が要求したローカルメモリ３０４から読み出されたデータ等を受け取る。このデータは、線４１５Ｓを介してコマンド処理部４０６に伝達され、転送データとして使用されたり、ネットワークインタフェース回路３０５の動作制御用データとしてネットワークインタフェース回路内部の制御レジスタに設定されたり、データ送信時にメッセージ生成用データ（転送先プロセッサ番号、送信データアドレス、送信フラグアドレス、転送データ長、受信データアドレス、受信フラグアドレス等）として使用される。

【００３１】
コマンド処理部４０６は、線４１５Ｓを介してコマンド受信部４０５から伝達されるネットワークインタフェース回路３０５内部の制御レジスタへのアクセスを行う。制御レジスタ読み出しアクセスを受けた場合、コマンド処理部４０６は、読み出し結果を線４１７Ｓを介してコマンド送信部４１１に伝達する。また、制御レジスタ書き込みアクセスを受けた場合には、コマンド処理部４０６は、その書き込みを実行する。メッセージの送信処理は、上記の制御レジスタへのアクセスがメッセージ送信起動用レジスタへの書き込み要求である場合に開始する。メッセージの送信処理では、データを宛先の要素プロセッサに転送するのに必要な情報であるヘッダを作成したり、転送データが存在するローカルメモリアドレスを知るために、ローカルメモリ３０４に格納されている転送制御情報２００を読み出すローカルメモリアクセス要求が発生される。このアクセス要求は線４１７Ｓを介してコマンド送信部４１１に伝達され、線４１３Ｓ、データ転送インタフェース回路４０４、線４１０Ｓ、メモリアクセスインタフェース回路４０３、線４０７Ｓを介してローカルメモリ３０４から読み出される。ローカルメモリ３０４からの読み出し結果は、線４０８Ｓ、メモリアクセスインタフェース回路４０３、線４１１Ｓ、データ転送インタフェース回路４０４、線４１２Ｓ、コマンド受信部４０５を介してコマンド処理部４０６に伝達され、メッセージ生成部４０７に伝達される。

【００３２】
メッセージ生成部４０７は、本実施の形態での特徴的な回路の一つであり、４１８Ｓを介して伝達された転送データと転送制御情報２００を含むヘッダからメッセージを生成し、線４１９Ｓを介してメッセージ送信部４０８に送出する。転送制御情報２００内のモードビット２０２が１である場合、メッセージ生成部４０７は、転送制御情報２００から生成されるヘッダと、転送制御情報２００内の送信データアドレス２０３に従ってローカルメモリ３０４から読み出した送信データと、送信フラグアドレス２０４に従ってローカルメモリ３０４から読み出したＭＰＩ付加情報からメッセージを組み立て、メッセージ送信部４０８に送出する。

【００３３】
図５に、メッセージ生成部４０７の内部構成を示す。メッセージ生成部４０７は、ローカルメモリ３０４から読み出した転送制御情報２００内の、送信データアドレス、送信フラグアドレスフィールド内のアドレス、転送データ長をそれぞれ保持するレジスタ５０１、５０２、５０３を有する。メッセージ生成部４０７は、さらに、ローカルメモリ３０４から生成中のメッセージのためにローカルメモリからすでに読みだされたデータの量を保持するレジスタ５０４と、本実施の形態に特徴的な回路として、ＭＰＩ付加情報のサイズを保持するレジスタ５０５と、転送制御情報２００内モードビットを保持するレジスタ５０６を有する。レジスタ５０４に保持された読み出し済みのデータの量はデータがローカルメモリ３０４から読み込まれるたびにカウントアップされ、送信すべきすべてのデータがメッセージ送信部４０８に伝達された後リセットされる。従って、この読み出し済みのデータの量は、送信済みのデータの総量と考えることができる。ＭＰＩ付加情報のサイズは

、ユーザプロセスからの初期化要求によりＭＰＩライブラリとＰＵＴ／ＧＥＴライブラリが初期化される時にレジスタ５０５にあらかじめセットされる。レジスタ５０６には、転送制御情報２００の読み出し時にその情報内のモードビットがセットされ、すべてのデータがメッセージ送信部４０８に伝達された後このレジスタ内のモードビットがリセットされる。メッセージ生成部４０７は、さらに、ローカルメモリ読み出し要求発行部５１１と、ローカルメモリ書込み要求発行部５１２と、メッセージ組み立て部５１３の他に、本実施の形態に特徴的な回路として、アンドゲート５０７と、加算器５０８および５０９と比較回路５１０とを有する。

【００３４】

アンドゲート５０７は、モードビットが１である場合には、レジスタ５０５内のＭＰＩ付加情報サイズを出力し、０である場合にはモードビットの値０を出力する。加算器５０９は、アンドゲート５０７の出力とレジスタ５０２に保持されている送信フラグアドレスフィールドの値を加算する。ＰＵＴ動作時にモードビットが１にセットされている場合、レジスタ５０２に保持されている送信フラグアドレスフィールドには、ローカルメモリ３０４内の、ＭＰＩ付加情報７１４の先頭アドレスが含まれているので、この加算の結果アドレスは、そのＭＰＩ付加情報の次のアドレスを指すことになり、ＰＵＴ動作時の送信完了フラグの書込みアドレスとして使用される。加算器５０８は、アンドゲート５０７の出力と、レジスタ５０３に保持された転送データ長を加算する。この加算の結果は、モードビットに１がセットされた場合、ＰＵＴ動作時に送信すべきメッセージに含まれるべき、ローカルメモリ３０４から読み出すべきデータの総量を示す。

【００３５】

比較回路５１０は、レジスタ５０４に保持された読み出し済みのデータ量を、レジスタ５０３に保持された転送データ量と加算器５０８から出力される転送データの総量とを比較する。比較の結果として、レジスタ５０４に保持された読み出し済みデータ量がレジスタ５０３内の転送データ長を越えていない場合には、そのことを示す比較結果信号をローカルメモリ読み出し要求発行部５１１に出力する。レジスタ５０４に保持された読み出し済みデータ量がレジスタ５０３内の転送データ長を越えているが、加算器５０８から与えられる、ローカルメモリ３０４から読み出すべきデータの総量を越えていない場合には、そのことを示す比較結果信号をローカルメモリ読み出し要求発行部５１１に出力する。レジスタ５０４に保持された読み出し済みデータ量が読み出すべきデータの総量に達した場合、そのことを示す比較結果信号をローカルメモリ書込み要求発行部５１２に出力する。

【００３６】

ローカルメモリ読み出し要求発行部５１１は、比較回路５１０からの比較結果信号と、レジスタ５０１内の送信データアドレスおよびレジスタ５０２内の送信フラグアドレスとからローカルメモリ読み出し要求を生成し、コマンド送信部４１１に送信する。すなわち、比較結果信号が、読み出し済みのデータが転送データ長を超えていないことを示すときには、送信データアドレスを元に後続の未読み出しのユーザデータを読み出すためのローカルメモリ読み出し要求を生成し、読み出し済みのデータが転送データ長を越えているが、読み出すべきデータの総量を超えていないときには、送信フラグアドレスを元にして未読み出しのＭＰＩ付加情報を読み出すためのローカルメモリ読み出し要求を生成する。

【００３７】

ローカルメモリ書込み要求発行部５１２は、比較回路５１０からの比較結果信号と、加算器５０９から出力される送信完了フラグアドレスとを元にローカルメモリ書込み要求をコマンド送信部４１１に送信する。すなわち、比較結果信号が、読み出し済みデータ量が読み出すべきデータの総量に達したことを示す場合、加算器５０８より与えられる送信完了フラグアドレスに送信完了フラグを書き込むことを要求する書き込み要求を生成する。

【００３８】

メッセージ組み立て部５１３では、本実施の形態に特徴的なセレクタ５１４がモードビットの値に従ってメッセージを組み立て、メッセージ送信部４０８にそのメッセージの送信要求を送付する。モードビットが１である場合、ヘッダとデータとＭＰＩ付加情報を含む

メッセージ５１５を組立て、モードビットが０である場合には、ヘッダとデータのみからなるメッセージ５１６を組立る。

【００３９】

図４において、メッセージ送信部４０８は、メッセージ生成部４０７からのメッセージ送信要求を受けて線４２０Ｓを介して相互結合ネットワーク１０５にメッセージを送出する。送出されたメッセージは、相互結合ネットワーク１０５を介してそのヘッダ情報に従って宛先に転送される。メッセージ生成部４０７における送信処理は、ネットワークインタフェース回路３０５へ送信される複数のメッセージ送信要求に対してそれらの送信要求の到着順に順次行われる。

【００４０】

次に、メッセージ受信部４０９について説明する。メッセージ受信部４０９は、線４２１Ｓを介して、相互結合ネットワーク１０５からメッセージを順次受け取り、メッセージ分解部４１０に線４２２Ｓを介して転送する。メッセージ分解部４１０はこのメッセージのヘッダ部に含まれる転送制御情報に従ってこのメッセージをデータ部とヘッダ部に分解し、ローカルメモリ３０４へこのデータ部や受信完了フラグの書き込みを要求する書き込み要求を線４２３Ｓを介してコマンド送信部４１１に伝達する。

【００４１】

図６にメッセージ分解部４１０の内部構成を示す。メッセージ分解部４１０には、メッセージヘッダ内の受信データアドレス、受信フラグアドレス、転送データ長、モードビットをそれぞれ保持するレジスタ６０１、６０２、６０３、６０６が設けられている。レジスタ６０６は本実施の形態で特徴的なレジスタであり、レジスタ６０６には、ヘッダ受信時にヘッダ内のモードビットがセットされ、メッセージ内の全データを受信したときにそのモードビットがリセットされる。さらに、受信したメッセージ内のデータの内、ローカルメモリに書き込み済みのデータの総量を保持するレジスタ６０４と、ＭＰＩ付加情報のサイズを保持する、本実施の形態に特徴的なレジスタ６０５が設けられている。レジスタ６０５には、ＭＰＩ初期化時あるいはジョブ起動時にあらかじめ定められたＭＰＩ付加情報サイズがセットされ、このサイズ情報はメッセージ内の全データが受信されたときにリセットされる。レジスタ６０４に保持された書き込み済みのデータの量は受信されたデータがローカルメモリ３０４に書き込まれるたびにカウントアップされ、受信すべきすべてのデータがローカルメモリ３０４に書き込まれた後リセットされる。従って、この書き込み済みのデータの総量は、受信済みのデータの総量であるとも考えることができる。

【００４２】

メッセージ分解部４１０には、ローカルメモリ書込み要求発行部６１１の他に、本実施の形態で特徴的な、アンドゲート６０７、加算器６０８、６０９と比較回路６１０とがさらに設けられている。アンドゲート６０７は、レジスタ６０６内のモードビットが１である場合にレジスタ６０５内のＭＰＩ付加情報サイズを出力し、モードビットが０である場合には０を出力する。加算器６０８および６０９の動作は、加算器５０８および５０９（図５）と同様である。比較回路６１０は、レジスタ５０４内の受信済みデータの総量を、転送データ長、アンドゲート６０７の出力結果データと比較し、比較結果信号をローカルメモリ書込み要求発行部６１１に出力する。すなわち、この比較回路は、受信済みのデータが転送データ長よりも短いか、受信済みのデータの総量が転送データ長より大きいが、転送データ長とアンドゲート６０７から出力されるＭＰＩ付加情報のサイズの和より小さいか、あるいは受信済みのデータの総量が転送データ長とＭＰＩ付加情報サイズの和より大きいかを判別する。

【００４３】

メモリ書込み要求発行部６１１は、比較回路６１０の比較結果信号と、レジスタ６０１内の受信データアドレスと、加算器６０９の出力とレジスタ６０２内の受信フラグアドレスフィールドの値とから、ローカルメモリ書込み要求を生成し、コマンド送信部４１１に線４２Ｓを介して伝達する。加算器６０９は、モードビットが１である場合に、レジスタ６０２内の受信フラグアドレスフィールドの値にＭＰＩ付加情報サイズを加算したアドレス

を受信完了フラグを書き込むべきローカルメモリアドレスとして出力する。

【００４４】

比較回路６１０の出力が、受信済みのデータの総量が転送データ長よりも大きいことを示す場合、メモリ書き込み要求発行部６１１は、受信されたデータをそれまでに受信したデータの書き込み位置に続けて書き込むためのローカルメモリ書き込み要求をレジスタ６０１内の受信データアドレスに基づいて生成する。

【００４５】

比較回路６１０の出力が、受信済みのデータの総量が転送データ長よりも大きく転送データ長とアンドゲート６０７から出力されるＭＰＩ付加情報のサイズの和より小さいことを示す場合、新たに受信されたデータは、ＭＰＩ付加情報である。従って、この場合には、メモリ書き込み要求発行部６１１は、受信されたデータをそれまでに受信したＭＰＩ付加情報の書き込み位置に続けて書き込むためのローカルメモリ書き込み要求を、レジスタ６０２内の受信フラグアドレスフィールドに含まれる、ＭＰＩ付加情報の書き込みアドレスに基づいて生成する。

【００４６】

比較回路６１０の出力が、受信済みのデータの総量が転送データ長とアンドゲート６０７から出力されるＭＰＩ付加情報のサイズの和より大きいことを示す場合、すべてのデータが受信されたこといなる。従って、この場合には、メモリ書き込み要求発行部６１１は、受信完了フラグをＭＰＩ付加情報の書き込み位置に続けて書き込むためのローカルメモリ書き込み要求を、加算器６０９の出力に基づいて生成する。

【００４７】

コマンド送信部４１１は、これらの書き込み要求に従って、受信されたデータあるいは受信完了フラグをローカルメモリ３０４に書き込む。コマンド送信部４１１は、線４１７Ｓを介して伝達される、ネットワークインタフェース回路３０５内部の制御レジスタからの読み出しデータ、線４２４Ｓを介して伝達される、メッセージ送信処理において使用される送信データのローカルメモリ３０４からの読み出し要求、線４２４Ｓを介して伝達される、メッセージ送信処理の完了に伴う送信完了フラグのローカルメモリ３０４への書き込み要求、線４２３Ｓを介して伝達される、メッセージ受信処理に伴う受信データあるいは受信完了フラグのローカルメモリ３０４への書き込み要求およびネットワークインタフェース回路３０５内で発生した割り込み要求を、線４１３Ｓを介してストレージコントローラ３０３内のデータ転送インタフェース回路４０４に伝達する。また、コマンド送信部４１１は、ネットワークインタフェース回路３０５の動作制御に関わる情報のローカルメモリ３０４からの読み出し要求、メッセージ送信処理に使用する転送制御情報２００のローカルメモリ３０４からの読み出し要求を線４１３Ｓを介してデータ転送インタフェース回路４０４に伝達する。

【００４８】

次に本実施の形態におけるメッセージ転送の流れを説明する。最初に自プロセッサのローカルメモリに格納されているユーザデータおよびＭＰＩ付加情報を転送先の要素プロセッサのローカルメモリに直接書き込むＰＵＴ処理について図４、５、７を用いて説明する。まず、本実施の形態に係わる通信方式では、送信側のユーザプロセスおよび受信側のユーザプロセスは、ＭＰＩライブラリを使用する前に、ＭＰＩライブラリ内の初期化ルーチンたとえばＭＰＩ＿ｉｎｉｔをコールするコマンド発行する。この初期化ルーチンのコールを受けると、ＭＰＩライブラリは、ＰＵＴ／ＧＥＴライブラリ内のいくつかの通信準備手続きをコールする。これらの通信準備手続きは、使用するＰＵＴ／ＧＥＴライブラリにより予め定められているが、以下では、後の説明に関連する部分および本実施の形態で新規に行われる処理のみを説明する。

【００４９】

本実施の形態では、送信側のＰＵＴ／ＧＥＴライブラリおよび受信側のＰＵＴ／ＧＥＴライブラリは、いずれもこれらの通信準備手続において以下の処理をすると仮定する。すなわち、ローカルメモリ７０１をユーザ空間にあらかじめマップし、さらに、通信領域７０

３、７０４（図７）を確保する。さらに、それぞれの通信領域内に送信データ領域およびそれに対応する送信完了フラグ領域を確保する。図７では、７１３は送信データ領域の例を示す。図では、受信側の通信領域７０４内の送信データ領域は図示していない。本実施の形態では、送信完了フラグ領域として、ＭＰＩ付加情報７１４および送信完了フラグ７１５の両方を格納する連続した領域を確保する点で従来と異なる。同様に、各通信領域内に、受信データ領域および受信完了フラグ領域を確保する。本実施の形態では図７では、７２１は受信データ領域の例を示す。図では、送信側の通信領域７０３内の受信データ領域は図示していない。本実施の形態では、受信完了フラグ領域として、ＭＰＩ付加情報７２２および受信完了フラグ７２３の両方を格納する連続した領域を確保する点で従来と異なる。なお、ＭＰＩ付加情報７１４の長さは例えば６４バイト程度である。なお、送信完了フラグあるいは受信完了フラグを書き込む領域７１５，７２３は例えば８バイトである。

【００５０】

ＰＵＴ／ＧＥＴ型ライブラリとして、送信すべきユーザデータを、ローカルメモリ７０１に常駐させることを前提とする場合とそうでない場合とがある。前者の場合には、送信データ領域７１３は、この常駐されたユーザデータの領域と一致するように、送信データ領域７１３が決定される。一方、後者の場合には、送信データ領域７１３は、送信側のユーザプロセスが使用するユーザデータに割り当てられたローカルメモリ内の領域とは独立に決定される。本発明はいずれの構造のＰＵＴ／ＧＥＴ型ライブラリにも適用可能である。しかし、後者の場合には、後に述べるように、ユーザプロセスが使用しているユーザデータに割り当てられたローカルメモリ内の領域のデータを送信データ領域７１３にコピーする処理が必要となる。しかし、前者の場合にはこのコピー動作が必要でなく、それだけデータ転送動作が高速化される。

【００５１】

以上のようにして、通信準備手続きが実行された後に、送信側のユーザプロセスの処理が進むと、そのユーザプロセスは、送信すべきデータを送信側の通信領域７０３内のユーザデータ領域７１３に書き込んだ後に、データ送信要求コマンド、たとえばＭＰＩ＿ｓｅｎｄを送信側のＭＰＩライブラリに対して発行する。このコマンドの名称は、使用するメッセージパッシングライブラリにより定まり、それが指定する引数も同様にそのライブラリにより定められた複数の種類の情報からなる。ここで仮定するＭＰＩライブラリの場合には、このコマンドの引数は、送信すべきユーザデータの先頭アドレス、ユーザデータ長と付加情報からなり、この先頭アドレスは、ユーザプロセスに割り当てられた仮想メモリ空間内での、そのユーザデータに対する仮想アドレスである。この付加情報は、受信側のユーザプロセスの識別子、プロセスグループの識別子等を含む。最初の二つの引数は、ＭＰＩライブラリを介さないでデータ転送をユーザプロセスがＰＵＴ／ＧＥＴ型ライブラリに直接要求するためのデータ送信要求が指定する引数と同じであり、付加情報がＭＰＩライブラリに対するデータ転送要求が新たに指定する引数である。

【００５２】

送信側のＭＰＩライブラリは、このデータ送信要求コマンドＭＰＩ＿ｓｅｎｄに応答して、送信側のＰＵＴ／ＧＥＴ型ライブラリに、この送信要求が指定するデータの送信を要求する。送信側のＭＰＩライブラリは、この要求を、ＭＰＩライブラリとＰＵＴ／ＧＥＴ型ライブラリにより予め定められた一つまたは複数のコマンドの形で発行する。以下では、それらのコマンドの内、本実施の形態で使用すると仮定する主なコマンドのみを説明する。

【００５３】

まず、送信側のＭＰＩライブラリは、送信権の取得を要求するコマンドを発行する。送信側のＰＵＴ／ＧＥＴ型ライブラリは、このコマンドに応答して、受信側のユーザプロセスと交信してそのプロセスに対するデータの送信権を得る。受信側のユーザプロセスおよびそのプロセスを実行している要素プロセッサの番号は、上記データ送信要求コマンドＭＰＩ＿ｓｅｎｄが指定する付加情報中の、受信側のプロセス識別番号とプロセスグループ識

別番号とにより決定される。

【００５４】

送信側のＭＰＩライブラリは、さらに、受信側のユーザプロセスの受信データ領域および受信フラグ領域のそれぞれの先頭位置を示す受信データアドレスおよび受信フラグアドレスを受信する。但し、この後に再度同じデータ送信要求コマンドＭＰＩ＿ｓｅｎｄを送信側のユーザプロセスが発行したときには、このコマンドを実行する必要はない。

【００５５】

既に述べたように、ＰＵＴ／ＧＥＴ型ライブラリが、送信すべきユーザデータがローカルメモリ７０１に常駐されることを前提としない場合には、送信側のＭＰＩライブラリは、送信すべきユーザデータを、ユーザデータ領域７１３にコピーすることを要求するコマンドを発行し、送信側のＰＵＴ／ＧＥＴ型ライブラリにより、このコマンドの引数で指定されるユーザデータのアドレスとデータ長で指定されるユーザデータに割り当てられた、ローカルメモリ内の領域のデータを、先に決定された送信データ領域７０３にコピーする。ＰＵＴ／ＧＥＴ型ライブラリが、送信すべきユーザデータをローカルメモリ７０１に常駐させることを前提とする場合には、このコピー動作は不要である。

【００５６】

次に、送信側のＭＰＩライブラリは、付加情報のローカルメモリへの書き込みを要求するコマンドを発行する。送信側のＰＵＴ／ＧＥＴ型ライブラリは、このコマンドに応答して、このコマンドの引数で指定されるＭＰＩ付加情報を、先に決定された送信データ領域７０３に対応して決定された送信フラグ領域の先頭の領域７１４に書き込む。

【００５７】

送信側のＭＰＩライブラリは、付加情報のローカルメモリへの書き込みを要求するコマンドを発行する。送信側のＰＵＴ／ＧＥＴ型ライブラリは、このコマンドに応答して、このコマンドの引数で指定されるＭＰＩ付加情報を、先に決定された送信データ領域７０３に対応して決定された送信フラグ領域の先頭の領域７１４に書き込む。

【００５８】

送信側のＭＰＩライブラリは、転送制御情報を生成することを要求するコマンドを発行する。送信側のＰＵＴ／ＧＥＴ型ライブラリは、このコマンドに応答して、転送制御情報７００を生成して、ローカルメモリ７００内の適当な領域に書き込む。この転送制御情報７００に含まれた情報は以下の通りである。転送先プロセッサ番号２０１は、受信側のユーザプロセスが実行されているプロセッサの番号であり、この番号は、すでに述べたように、ＭＰＩ初期化ルーチンにおいて決定されている。モードビット２０２は、送信フラグアドレスフィールド２０４を本実施の形態に従って拡張して使用するか否かを示すビットである。送信側のＰＵＴ／ＧＥＴ型ライブラリは、送信側の　ＭＰＩライブラリから、データ転送を要求されたときに、ＰＵＴ／ＧＥＴ型ライブラリはこのモードビットを１にセットする。モードビット２０２が１であることは、ユーザデータとＭＰＩ付加情報とを一つのメッセージで送信することを指示する。なお、本実施例の形態では、送信側のユーザプロセスが、ＭＰＩライブラリに対してでなく、ＰＵＴ／ＧＥＴ型ライブラリに対して直接データ送信要求コマンドを発行した場合には、ＰＵＴ／ＧＥＴ型ライブラリは、そのデータ要求に対して、図７に示す転送制御情報７００と同じ構造を有し、モードビット２０２の値が０である転送制御情報を生成する。

【００５９】

送信データアドレス２０３は、送信側のユーザプロセスにより転送が要求されたユーザデータまたはそのコピーを保持する送信データ領域７１３の先頭アドレスである。転送データ長２０５は、データ送信要求コマンドＭＰＩ＿ｓｅｎｄが指定した、ユーザデータのデータ長であり、そのコマンドを受けたＭＰＩライブラリがＰＵＴ／ＧＥＴ型ライブラリに通知する。

【００６０】

送信フラグアドレスフィールド２０４には、従来では送信完了フラグを書き込むためのローカルメモリアドレスが格納されるが、本実施の形態では、モードビット２０２が１であ

る場合には、ローカルメモリ７０１内のＭＰＩ付加情報７１４の先頭アドレスをこの送信フラグアドレスフィールド２０４に格納する。なお、モードビット２０２が１の場合には、送信完了フラグを格納するローカルメモリ内の領域７１５のアドレスは送信フラグアドレスフィールド２０４によっては明には指定されないことになる。本実施の形態では、ネットワークインタフェース回路３０５が、ユーザプロセスとＭＰＩ付加情報とに対する共通の送信完了フラグを、ローカルメモリ７０１内のＭＰＩ付加情報７１４の最終のアドレスの次のアドレスの領域７１５に格納するようになっている。このため、送信型のＰＵＴ／ＧＥＴ型ライブラリは、ユーザプロセスからの先のデータ送信要求コマンドに対する応答として、この送信完了フラグが書き込まれた時点で、送信完了を送信側のユーザプロセスに通知するようになっている。なお、モードビット２０２が０である場合には、ＰＵＴ／ＧＥＴ型ライブラリは、送信フラグアドレスフィールド２０４に、そのライブラリが決定した送信完了フラグを書き込むアドレスをセットする。

【００６１】

受信データアドレス２０６は、受信側の要素プロセッサにおいて、受信したデータを格納するためのローカルメモリ領域７２１（図７）のアドレスである。受信フラグアドレスフィールド２０７には、この受信データとともに受信した付加情報と受信完了を示す受信完了フラグを格納する領域の先頭アドレスである。本実施の形態では、付加情報を記憶する領域７２２の後続の領域７２３に受信完了フラグを書き込む。したがって、受信フラグアドレスフィールド２０７には、この領域７２２の先頭アドレスが書き込まれる。受信データアドレス２０６と受信フラグアドレス２０７は、いずれもＭＰＩ初期化ルーチンにて送信先のＰＵＴ／ＧＥＴライブラリにより通知される。これらのアドレスは、ＧＥＴ動作時に送信側のＰＵＴ／ＧＥＴ型ライブラリにより使用される。

【００６２】

その他制御情報２０８は、ＰＵＴメッセージあるいはＧＥＴメッセージあるいはＧＥＴ要求メッセージの種別を示したり、１対１通信あるいは１対多通信等の通信形態を示すといった、その他通信処理に必要な情報を含む。

【００６３】

なお、転送制御情報２００の中に、ＭＰＩ付加情報を読み出すべきローカルメモリアドレスを指定するフィールドおよび受信したＭＰＩ付加情報を書き込むべきローカルメモリアドレスを指定するフィールドを別に設けることも可能であるが、本実施の形態のようにモードビット２０２を用いて、送信フラグアドレス２０４および受信フラグアドレス２０７が指定する二つのアドレスを切り替えることにより転送制御情報２００の構造と大きさを、ユーザプロセスがＰＵＴ／ＧＥＴ型ライブラリのみを使用して送信あるいは受信する場合と同じとすることができる。

【００６４】

こうして、転送制御情報７００が生成されると、送信側のＭＰＩライブラリは、生成された転送制御情報に従って、データの送信を行うことをネットワークインタフェース回路３０５に要求するコマンドを発行する。ＰＵＴ／ＧＥＴライブラリは、このコマンドに応答して、ネットワークインタフェース回路３０５内部の送信起動用レジスタ（図示せず）にその転送制御情報７００のアドレスを書き込むことを要求する書き込みコマンドを発行する。コマンド処理部４０６は、この書き込みコマンドを実行して、ネットワークインタフェース回路３０５内のメッセージ送信起動用レジスタ（図示せず）へ転送制御情報７００の先頭アドレスを書き込む。この書き込みによりネットワークインタフェース回路３０５はメッセージ送信処理を開始する。

【００６５】

図４に示すように、コマンド処理部４０６は、メッセージのヘッダを作成するためにローカルメモリ３０４に格納されている転送制御情報７００を読み出すローカルメモリアクセス要求を線４１７Ｓを介してコマンド送信部４１１に伝達する。線４１３Ｓ、データ転送インタフェース回路４０４、線４１０Ｓ、メモリアクセスインタフェース回路４０３を介して、送信起動用レジスタ（図示せず）に書き込まれたアドレスをもとにローカルメモリ

３０４から読み出された転送制御情報７００は、メモリアクセスインタフェース回路４０３、線４１１Ｓ、データ転送インタフェース回路４０４、線４１２Ｓ、コマンド受信部４０５を介してコマンド処理部４０６に伝達され、メッセージ生成部４０７に伝達される。

【００６６】

図５において、メッセージ生成部４０７は、転送制御情報７００内の送信データアドレス２０３、送信フラグアドレスフィールド２０４の値、転送データ長２０５、モードビット２０２をそれぞれレジスタ５０１、５０２、５０３、５０６にセットする。その他の情報は図示しないレジスタに保持される。レジスタ５０６内のモードビットは今の場合には１である。レジスタ５０６内のモードビットが１であるため、アンドゲート５０７がＯＮになり、レジスタ５０５内のＭＰＩ付加情報サイズを出力する。加算器５０８は、レジスタ５０３内の転送データ長とアンドゲート５０７から出力されたＭＰＩ付加情報を足しあわせ、ローカルメモリ７０１から読み出すべきデータの総量を出力する。比較回路５１０がレジスタ５０４に保持された、読み出し済みデータの総量がレジスタ５０３内の転送データ長よりも小さいと判断した場合には、ローカルメモリ７０１内の、レジスタ５０１に保持された送信データアドレスの記憶位置からユーザデータ７１３を読み出すために、コマンド送信部４１１にローカルメモリ読み出し要求を伝達する。このユーザデータは、コマンド送信部４１１からストレージコントローラ３０３内のメモリアクセスインタフェース回路４０３を介してローカルメモリ３０４から読み出される。読み出されたデータは、ストレージコントローラ３０３内のメモリアクセスインタフェース回路４０３、データ転送インタフェース回路４０４およびコマンド受信部４０５を介してコマンド処理部４０６に伝達され、線４１８Ｓを介してメッセージ生成部４０７に伝達される。

【００６７】

メッセージ生成部４０７では、比較回路５１０が、レジスタ５０４内の読み出し済みのデータの量が、レジスタ５０３内の転送データ長より大きいが、（転送データ長＋レジスタ５０５内のＭＰＩ付加情報サイズ）以下であると判断したときには、メモリ読み出し要求発行部５１１は、ローカルメモリ７０１内の、レジスタ５０２に保持された送信フラグアドレスを有する記憶位置からＭＰＩ付加情報７１４を読み出すためのローカルメモリ読み出し要求をコマンド送信部４１１に伝達する。この読み出し要求は、コマンド送信部４１１からストレージコントローラ３０３内の、データ転送インタフェース回路４０４およびメモリアクセスインタフェース回路４０３を介してローカルメモリ３０４に送られ、ＭＰＩ付加情報７１４がそこから読み出される。読み出されたＭＰＩ付加情報７１４は、ストレージコントローラ３０３内のメモリアクセスインタフェース回路４０３、データ転送インタフェース回路４０４およびコマンド受信部４０５を介してコマンド処理部４０６に伝達され、線４１８Ｓを介してメッセージ生成部４０７に伝達される。

【００６８】

メッセージ生成部４０７では、メッセージ組み立て部５１３内のセレクタ５１４は、モードビットが１であることから、ローカルメモリ３０４から読み出された転送制御情報７００内の送信データアドレス、送信フラグアドレス以外の部分をレジスタ５１５内のヘッダ部の格納する。同様に、ユーザデータ７１３およびＭＰＩ付加情報７１４をレジスタ５１５のデータ部に格納する。

【００６９】

比較回路６１０が、読み出したユーザデータの総量が、転送データ長とＭＰＩ付加情報サイズの和に等しくなったことを検出したとき、メモリ書き込み要求発行部５１２は、送信完了フラグを書き込むことを要求するローカルメモリ書込み要求をコマンド送信部４１１に伝達する。このコマンドは、加算器５０９により与えられる、レジスタ６０２内の送信フラグアドレスフィールドに保持されたＭＰＩ付加情報の先頭アドレスと、レジスタ５０５内のＭＰＩ付加情報サイズとの和に等しいアドレスにこのフラグを書き込むことを要求する。送信完了フラグ７１５は、コマンド送信部４１１からストレージコントローラ３０３内のメモリアクセスインタフェース回路４０３を介してローカルメモリ３０４に書き込まれる。本実施の形態では、メッセージに含まれるべきユーザデータとＭＰＩ付加情報の

読み出しが完了した時点で、メッセージの送信が完了したと見なして、送信完了フラグ７１５を書き込む。しかし、このメッセージが実際に相互結合ネットワーク１０５に送信された時点でこのフラグを書き込むようにしてもよい。

【００７０】

こうして、レジスタ５１５内にユーザデータとそれに関連するＭＰＩ付加情報を含む一つのメッセージ７０５が生成される。このメッセージ７０５には、転送制御情報７００に含まれていたのと同じ転送先プロセッサ番号７１６、モードビット７１７、転送データ長７１８、受信データアドレス７１９、受信フラグアドレス７２０、その他の制御情報２０８をそのまま含み、転送制御情報７００に含まれていた送信データアドレス２０３と送信フラグアドレス２０４に代えて、ユーザデータ７２６、付加情報７２７を含むことになる。送信データアドレスメッセージ生成部４０７はそのメッセージ７０５をメッセージ送信部４０８に送信する。メッセージ生成部４０７は、メッセージ７０５の生成に使用される上記３つの情報をメッセージ送信部４０８へ送出し終ると、メッセージ送信部４０８はレジスタ５０４と５０６をリセットする。メッセージ送信部４０８はそのメッセージ７０５を相互結合ネットワーク１０５に送出する。相互結合ネットワーク１０５はそのメッセージ内の転送先プロセッサ番号２０１により指定されるプロセッサにそのメッセージを転送する。

【００７１】

次に受信処理について説明する。相互結合ネットワーク１０５から転送されたメッセージ７０５は、まずメッセージ受信部４０９で受け取られ、メッセージ分解部４１０に転送される。メッセージ分解部４１０は、メッセージヘッダ内の受信データアドレス７１９、受信フラグアドレス７２０、転送データ長７１８、モードビット７１７をそれぞれレジスタ６０１、６０２、６０３、６０６（図６）に書き込む。レジスタ６０４はあらかじめ０にリセットされ、レジスタ６０５にはあらかじめＭＰＩ付加情報サイズがセットされている。メッセージ分解部４１０では、メモリ書き込み要求発行部６１１は、受信されたユーザデータをヘッダ内の受信データアドレス７１９が示すローカルメモリ領域に書き込むためのローカルメモリアクセス要求を生成し、線４２３Ｓを介してコマンド送信部４１１に伝達する。受信されたユーザデータは、メッセージヘッダ内の受信データアドレス７１９に従って、ストレージコントローラ３０３内のメモリアクセスインタフェース回路４０３を介して、ローカルメモリ３０４の通信領域７０４内の領域７２１に書き込まれる。レジスタ６０６内のモードビットは１であるため、アンドゲート６０７がＯＮになり、レジスタ６０５内のＭＰＩ付加情報サイズを加算器６０８に供給する。加算器６０８では、レジスタ６０３内の転送データ長とアンドゲート６０７から与えられるＭＰＩ付加情報サイズを足しあわせ、ローカルメモリに書込むべきデータの総量を得る。レジスタ６０４内の受信データ量は、メッセージ受信部４０９が相互結合ネットワーク１０５からメッセージ内のデータの異なる部分を受信するごとに更新される。

【００７２】

メッセージ分解部４１０では、ユーザデータの異なる部分がメッセージ受信部４０９により受信されるごとに、比較回路６１０が、レジスタ６０４内の受信されたデータの総量がレジスタ６０３内の転送データ長よりも小さいか否かを判断し、前者が後者より小さいと判断したときには、メモリ書き込み要求発行部６１１は、レジスタ６０１に保持された受信データアドレスにしたがって受信されたデータを受信側のローカルメモリ７０２に書き込むことを要求するコマンド送信部４１１にローカルメモリ書込み要求を伝達する。

【００７３】

その後比較回路６１０が、レジスタ６０４内の、受信されたデータの総量がレジスタ６０３内の転送データ長よりも大きいが、転送データ長とレジスタ６０５内のＭＰＩ付加情報サイズの和以下と判断したならば、メモリ書き込み要求発行部６１１は、レジスタ６０２に保持された受信フラグアドレスフィールドの値のアドレスに、受信されたＭＰＩ付加情報を書き込むことを要求するローカルメモリ書込み要求を伝達する。

【００７４】

加算器６０９は、レジスタ６０２内の受信フラグアドレスフィールドの値とアンドゲート６０７から与えられるＭＰＩ付加情報サイズを加算し、受信完了フラグを書き込むべきメモリアドレスを決定する。比較回路６１０が、レジスタ６０４内の受信データ数が、加算器６０８より与えられる、転送データ長＋ＭＰＩ付加情報サイズに等しくなったことを検出すると、メモリ書き込み要求発行部６１１は、加算器６０９により与えられるアドレスに受信完了フラグを書き込むことを要求するローカルメモリ書き込み要求をコマンド送信部４１１に伝達する。その結果、受信完了フラグ７２３は、コマンド送信部４１１から、ストレージコントローラ３０３内のメモリアクセスインタフェース回路４０３を介してローカルメモリ３０４に書き込まれる。こうして、データ受信処理が完了する。また、受信の完了でもってデータ転送処理が終了する。

【００７５】

なお、受信されたデータおよび付加情報は、以下のようにして受信側の要素プロセッサ７０２で使用される。受信側のユーザプロセスが、他の要素プロセッサから送信されたデータの受信を要求するコマンド、例えば、ＭＰＩ＿ｒｅｃｖを発行する。このコマンドは、受信すべきユーザデータを指定するアドレスと、そのデータの最大長、およびＭＰＩライブラリにより定められた付加情報とからなる引数を指定する。この付加情報は、送信元ユーザプロセスの識別子その他の情報からなる。このコマンドで指定される上記アドレスは、受信側のユーザプロセスに割り当てられたアドレス空間に属する仮想アドレスである。受信側のＭＰＩライブラリは、この受信コマンドの引数で指定されるユーザデータがローカルメモリ７０４に書き込み済みであるか否かを、ローカルメモリ７０４に書き込まれたユーザデータ７２１、付加情報７２２、受信完了フラグ７２３に基づいて判別する。もし、この要求されたデータがローカルメモリ７０４に書き込み済みであるときには、受信側のＭＰＩライブラリは、受信側のユーザプロセスに受信完了を通知する。この要求されたデータがローカルメモリ７０４に書き込み済みでないときには、ＭＰＩライブラリは、上記判別が成功するまでその判別を繰り返す。

【００７６】

なお、受信側のユーザプロセスが指定した仮想アドレスを有するデータ領域が、ローカルメモリ７０４に常駐していない場合には、受信側のＭＰＩライブラリは、上記通知を行う前に、受信されたユーザデータ７２１を、ユーザプロセスが指定するアドレスに割り当てられたローカルメモリ領域にコピーする。もし、受信側のユーザプロセスが指定した仮想アドレスを有するデータ領域が、ローカルメモリ７０４に常駐している場合には、このコピーは不要である。

【００７７】

受信側のユーザプロセスは、この受信完了の通知を受けると、受信データを読みだす命令を実行する。したがって、この受信側でのユーザデータの受信判別処理では、ユーザデータ７２１と付加情報７２２が書き込み済みであるか否かを検出するのに、共通の受信完了フラグ７２３を使用するところが従来と異なる。

【００７８】

以上は、ユーザプロセスがデータの送信要求コマンドをＭＰＩライブラリに対して発行した場合である。本実施の形態では、他のユーザプロセスは、ＰＵＴ／ＧＥＴライブラリに対してデータ送信要求を発行することもできるようになっている。この場合には、このデータ送信要求コマンドは、送信すべきユーザデータを示す仮想アドレスと、データ長を指定する。送信側のＭＰＩライブラリは、このコマンドに応答して、先に述べたと同じようにして、転送制御情報７００を生成する。但し、この情報の中のモードビット２０２の値は０である。さらに、送信フラグアドレスフィールド２０４は、送信完了フラグの書き込み領域７１５のアドレスを指定する。受信フラグアドレスフィールド７２３についても同じである。上記データ送信要求コマンドの場合には、付加情報記憶領域７１４，７２２は不要である。

【００７９】

転送制御情報７００内のモードビット２０２が０である場合、転送制御情報７００内のモ

ードビット２０２が１である場合と比べると、メッセージ生成部４０７およびメッセージ分解部４１０の動作が異なる。すなわち、レジスタ５０６内のモードビットが０であるため、加算器５０８の出力は、レジスタ５０３内の転送データ長に等しく、加算器５０９の出力は、レジスタ５０２内の送信完了フラグアドレスに等しい。したがって、メッセージ生成部４０７では、比較回路５１０が、レジスタ５０４内の読み出し済みのデータの総量がアンドゲート５０７より与えられる、レジスタ５０３内の転送データ長に等しくなったことを検出したときに、メモリ読み出し要求発行部５１１は、ローカルメモリ７０１からの送信すべきデータの読み出しを終了する。この読み出されたデータを含むメッセージがすべて相互結合ネットワーク１０５へ送出されると、メッセージ生成部４０７内のメモ書き込み要求発行部５１２は、加算器５０９より与えられる、レジスタ５０２内の送信フラグアドレスに示されるローカルメモリ領域に送信完了フラグを書き込むことを要求する書き込み要求をコマンド送信部４１１に伝達する。

【００８０】

受信側の要素プロセッサでは、相互結合ネットワーク１０５から転送されたメッセージ内のヘッダ内のモードビット７１７は０であるため、レジスタ６０６（図６）には０がセットされる。レジスタ６０６内のモードビットが０であるため、加算器６０８の出力はレジスタ６０３内の転送データ長に等しく、加算器６０９の出力は、レジスタ６０２にセットされる受信フラグアドレスに等しい。したがって、比較回路６１０が、レジスタ６０４内の受信済みのデータの総量が加算器６０８から出力される、転送データ長に等しくなったことを検出するまで、メモリ書き込み要求発行部６１１は、受信データのローカルメモリへの書き込みを要求する書き込み要求を発行する。すべての受信データがローカルメモリに書き込まれた後、メモリ書き込み要求発行部６１１は加算器６０９が出力する、レジスタ６０２内の受信フラグアドレスに従ってローカルメモリ３０４に受信完了フラグを書き込むための書き込み要求を発行する。こうして、ＭＰＩライブラリが要求したデータ転送が終了する。

【００８１】

なお、比較のために、従来のＰＵＴ処理では、ＭＰＩ＿ｓｅｎｄに示すように、送信側のＰＵＴ／ＧＥＴライブラリは、ユーザデータとそれに対する付加情報をそれぞれ転送するための転送制御情報Ａ、Ｂ（８００、８０１）を作成し、通信領域８０２内にユーザデータとそれに対する送信完了フラグＡ、付加情報とそれに対する送信完了フラグＢを記憶するように構成され、送信側のネットワークインタフェース回路はこえらの転送制御情報に基づいて、二つのメッセージＡ、Ｂ（８０４、８０５）を送信する。受信側のネットワークインタフェース回路は、通信領域８０３内にユーザデータおよびそれに対する受信完了フラグ、付加情報とそれに対する受信完了フラグを書き込むように構成される。なお、転送制御情報Ａ、Ｂには、本実施の形態で言うモードビットが存在しない。

【００８２】

これに対して、本実施の形態では、ユーザデータおよびその付加情報という異なる２つのデータをローカルメモリ間直接転送にしたがって１回のメッセージ転送で行なえる。したがって、２回のメッセージ転送が必要であった従来よりも転送レイテンシおよびローカルメモリアクセス回数を削減して並列処理効率を向上できる。さらに、ＭＰＩライブラリを介さない従来のメモリ間直接転送も実行できる。

【００８３】

さらに、送信側のネットワークインタフェース回路は、送信完了フラグアドレスを付加情報と送信完了を示す制御情報の書き込みの両方に使用しているので、ＰＵＴ／ＧＥＴライブラリがネットワークインタフェース回路に対して指定すべき情報量が少なくて済む。また、受信側においても、ネットワークインタフェース回路は、受信完了アドレスを、付加情報と受信完了を示す制御情報の書き込みの両方に使用しているので、ＰＵＴメッセージに含まれる情報量が少なくて済み、それだけネットワークの混雑を防ぐことができ、さらに、メッセージの、送信元の要素プロセッサでの送信時間、ネットワーク上の転送時間および送信先の要素プロセッサにおける受信時間が短くでき、全体としてメッセージの転送

時間が短くなる。

【００８４】

以上のように、従来の方式では、異なる２種類のデータ送信処理を行う場合、ネットワークインタフェース回路３０５がローカルメモリを６回アクセス（転送制御情報の読み出し×２と転送データの読み出し×２とフラグの書き込み×２）し、受信処理の場合も送信と同様に、ネットワークインタフェース回路３０５がローカルメモリを４回アクセス（転送データの書き込み×２とフラグの書き込み×２）することになり、本実施の形態に比べて処理オーバーヘッドが大きい。本実施の形態では、転送制御情報にモードビットを持つことで、転送制御情報の読み出し回数、およびフラグのローカルメモリへの書き込み回数を送信側４回、受信側３回に削減してローカルメモリアクセスに関わる処理オーバヘッドを小さくできる。また、従来は２回にわけて転送していたメッセージを１回で転送できるため、データ転送のレイテンシを削減できる。

【００８５】

以上ではＭＰＩライブラリを介したＰＵＴ処理について説明したが、他の通信として、送信元の要素プロセッサが宛先の要素プロセッサのローカルメモリに格納されているユーザデータを取ってくるＧＥＴ処理がある。本実施の形態は、ＭＰＩライブラリを介したＧＥＴ処理にも同様に適用できる。すなわち、ＭＰＩライブラリとＰＵＴ／ＧＥＴ処理を併用して、ユーザデータとそれに対する付加情報を同時に一つのメッセージでＧＥＴ（転送）できる。

【００８６】

まず、要求元のユーザプロセスは、ＧＥＴ要求コマンドをＭＰＩライブラリに対して発行する。このコマンドは、要求するデータのアドレスとデータ長およびＭＰＩライブラリが定めた付加情報等からなる点で、先に説明した送信要求コマンドと同様の引数を指定する。このコマンドを受けて、ＭＰＩライブラリとＰＵＴ／ＧＥＴライブラリとネットワークインタフェース回路は、ＰＵＴ動作の場合と同様にしてＧＥＴ要求メッセージを宛先要素プロセッサに対して転送する。このメッセージには、ＰＵＴのときのメッセージにおける受信データアドレス、受信完了フラグアドレスの代わりに、ＧＥＴすべきデータに対する送信データアドレスおよび送信完了フラグが含まれ、ユーザデータと付加情報は含まれない。このメッセージが従来のＧＥＴ要求メッセージと異なる点は、モードビットを有することである。

【００８７】

宛先プロセッサは、ＧＥＴ要求メッセージに含まれたデータアドレスで指定されるローカルメモリ領域からユーザデータを読み出し、送信完了フラグアドレスで指定されるローカルメモリ領域から付加情報を読み出し、送信元プロセッサにそのユーザデータと付加情報を含むＧＥＴメッセージを送り返す。このメッセージは、受信データアドレスと受信完了フラグも含む。宛先プロセッサではそのメッセージを全て相互結合ネットワークに送出した後、送信完了フラグをローカルメモリ内の、上記付加情報の記憶領域の次のアドレス位置に書き込む。

【００８８】

送信元要素プロセッサは、ＧＥＴメッセージを受信し、メッセージ内の受信データアドレスに従ってローカルメモリにそのメッセージ内のデータを書き込み、そのメッセージ内の受信完了フラグアドレスに従って、メッセージ内の付加情報をローカルメモリに書き込む。これらのユーザデータと付加情報をすべて受信した後に受信完了フラグをローカルメモリに書き込む。

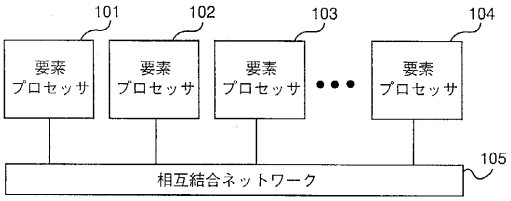
【００８９】

ＰＵＴ処理と同様に、ＧＥＴ処理でも、転送制御情報の読み出し回数およびフラグのローカルメモリへの書き込み回数を削減してローカルメモリアクセスに関わる処理オーバヘッドを小さくできる。また、従来はユーザデータとそれに対する付加情報とを２回にわけて転送していたが、本実施の形態では、一つのＧＥＴメッセージでこれらの二つのデータを転送できるため、データ転送のレイテンシを削減できる。

【００９０】
＜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、以下に示す変形例を含むいろいろの実施の形態により実施可能である。たとえば、ローカルメモリ３０４に格納する転送制御情報に拡張サイズフィールドを設ける。拡張サイズフィールドには、拡張したいフラグ領域のサイズをセットする。モードビットがセットされた場合、送信処理において、転送制御情報を含むヘッダと、送信データアドレスに従ってローカルメモリ領域からよみだされる転送データ長分のデータと、送信フラグアドレスに従ってローカルメモリ領域から読み出される拡張サイズフィールドに設定されたサイズ分の別のデータからメッセージを生成し、相互結合ネットワーク１０５に送出する。受信処理においては、メッセージをヘッダ部とデータ部に分解し、ヘッダ内の転送制御情報に含まれる受信データアドレスで示されるローカルメモリ領域にデータを転送データ長分書き込み、さらに、ヘッダ内の転送制御情報に含まれる受信フラグアドレスで示されるローカルメモリ領域に対し、別のデータを拡張サイズフィールドに設定されたサイズ分書き込む。モードビットがセットされていない場合、拡張サイズフィールドにセットされた値は無視される。
【００９１】
【発明の効果】
本実施の形態によれば、ＭＰＩのようなメッセージパッシングライブラリを介してメモリ間直接転送を行う場合におけるデータ転送処理をより高速に行うことができる。
【図面の簡単な説明】
【図１】本実施の形態が対象とする並列計算機の概略構成を示す図である。
【図２】本実施の形態で使用する転送制御情報の例を示す図である。
【図３】図１の装置に使用する要素プロセッサの概略構成を示す図である。
【図４】図３の要素プロセッサに使用するストレージコントローラおよびネットワークインタフェース回路の構成を示す図である。
【図５】図４のネットワークインタフェース回路内のメッセージ生成部の内部構成を示す図である。
【図６】図４のネットワークインタフェース回路内のメッセージ分解部の内部構成を示す図である。
【図７】本実施の形態におけるデータ転送処理の概要を説明する図である。
【図８】従来例のデータ転送処理の概要を説明する図である。
【符号の説明】
１０５…相互結合ネットワーク

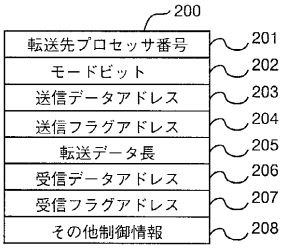
【図１】

図1



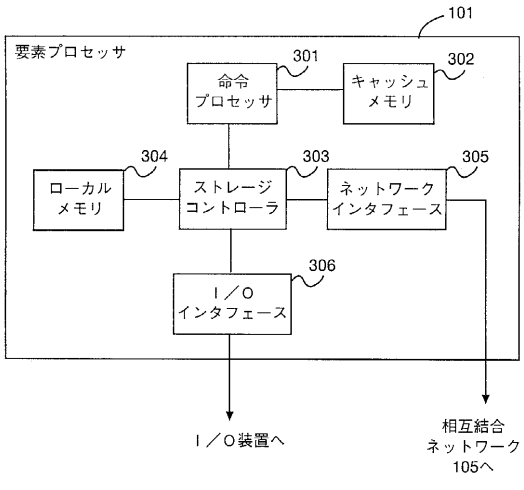
【図２】

図2



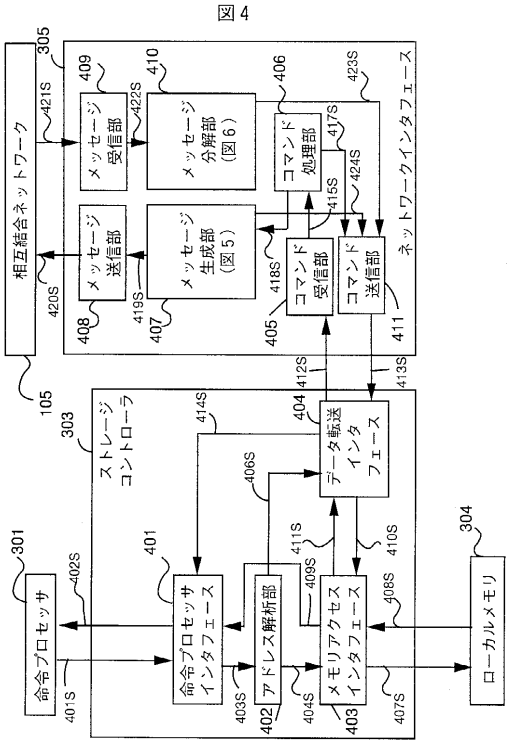
【図３】

図3



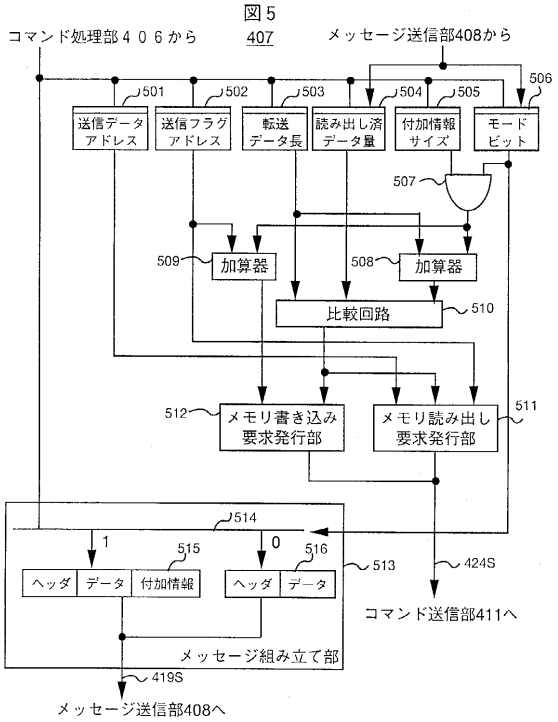
【図４】

図4



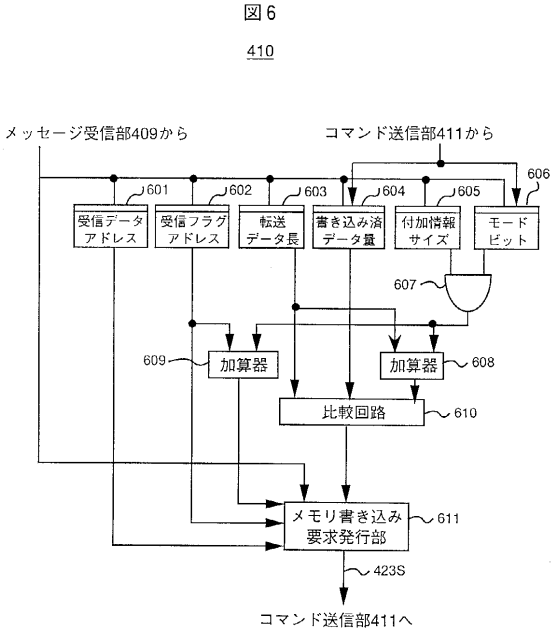
【図５】

図5



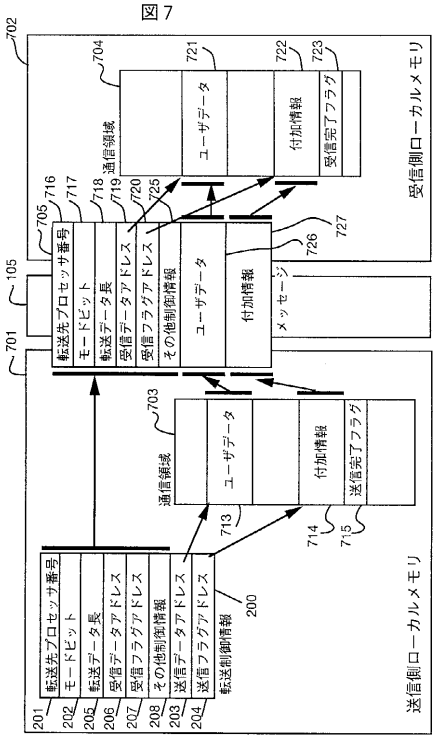
【図６】

図6



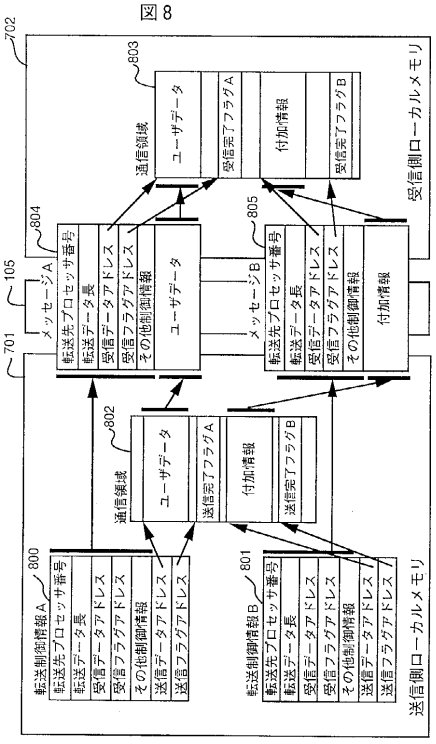
【図７】

図7



【図８】

図8



フロントページの続き

審査官　鳥居　稔

(56)参考文献　特開平７－３１１７５０（ＪＰ，Ａ）
特開平８－０３０５６６（ＪＰ，Ａ）
白木長武,小柳洋一,今村信貴,林憲一,清水俊幸,堀江健志,石畑宏明，高並列計算機ＡＰ１０００＋のメッセージハンドリング機構，情報処理学会論文誌　第３７巻　第７号，日本，社団法人情報処理学会，１９９６年　７月１５日，第37巻第7号，1388-1398

(58)調査した分野(Int.Cl.[7]，ＤＢ名)
G06F 15/16-173
G06F 13/00
G06F 9/46