(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7645495号
(P7645495)

(45)発行日　令和7年3月14日(2025.3.14)

(24)登録日　令和7年3月6日(2025.3.6)

(51)国際特許分類
Ｇ０６Ｆ　15/173 (2006.01)

ＦＩ
Ｇ０６Ｆ　15/173　６６５Ｄ

請求項の数　3　（全15頁）

(21)出願番号　特願2022-530631(P2022-530631)
(86)(22)出願日　令和3年6月11日(2021.6.11)
(86)国際出願番号　PCT/JP2021/022268
(87)国際公開番号　WO2021/251479
(87)国際公開日　令和3年12月16日(2021.12.16)
審査請求日　令和5年12月27日(2023.12.27)
(31)優先権主張番号　特願2020-102484(P2020-102484)
(32)優先日　令和2年6月12日(2020.6.12)
(33)優先権主張国・地域又は機関
日本国(JP)

(73)特許権者　520211960
株式会社情報科学研究所
新潟県新潟市西区真砂一丁目１５番３６号
(73)特許権者　311009158
株式会社エヌアイエスプラス
東京都文京区小石川２－４－１０
(74)代理人　100121441
弁理士　西村 竜平
(74)代理人　100154704
弁理士　齊藤 真大
(74)代理人　100206151
弁理士　中村 惇志
(74)代理人　100218187
弁理士　前田 治子

最終頁に続く

(54)【発明の名称】　並列分散計算システム

(57)【特許請求の範囲】

【請求項１】

アドレス変換バッファ（ＴＬＢ）を有するプロセッサ、物理メモリ、及び、当該物理メモリに直接アクセス可能なネットワークインターフェースコントローラ（ＮＩＣ）を備えた複数の計算機を、複数のデータリンクを介して相互接続した並列分散計算システムであって、

送信元計算機のプロセス（以下、送信側プロセス）が、送信先計算機のプロセスを規定する操作対象プロセス（以下、受信側プロセス）の識別子と、当該受信側プロセスのメモリ領域を規定する操作対象アドレスと、書き込まれるデータサイズと、データ列とを含む操作要求パケットを送信し、

前記送信先計算機が、前記操作要求パケットを受信し、前記受信側プロセス及び前記操作対象アドレスにより規定されるメモリ領域に前記データ列を格納するものであり、

前記送信側プロセスは、前記受信側プロセス毎に前記操作要求パケットの送信数をカウントし、前記操作要求パケットに当該操作要求パケットのカウントアップ前又は後のカウント数を添付する送信数カウント部を有し、

前記受信側プロセスは、前記送信側プロセス毎に前記操作要求パケットの受信数をカウントする受信数カウント部を有し、

前記送信先計算機は、

前記受信数カウント部のカウント数と、前記操作要求パケットに添付されたカウント数とが連続している場合に、前記操作要求パケットの処理を行い、前記操作要求パケットに添付されたカウント数によって前記受信数カウント部のカウント数を更新し、

前記受信側プロセスへの到着順序が乱れて未着の操作要求パケットが存在する場合には、当該未着の操作要求パケットの前までしか前記受信数カウント部によるカウント数の更新を行わずに、前記未着の操作要求パケット以降の操作要求パケットの到着に関しては、操作要求パケットの処理を行い、前記受信数カウント部とは別に操作要求パケットに添付されたカウント数を既着カウント記録部に記録し、
未着だった操作要求パケットの到着によって前記既着カウント記録部を用いて前記受信側プロセスへの到着記録の記録内容が前記受信数カウント部から連続した場合に、その連続している範囲内で前記受信数カウント部を更新し、前記既着カウント記録部から更新済みの記録を削除し、
前記送信側プロセスは、自分が送ろうとする操作要求パケットがそれに先行する操作要求パケットの処理が済むまで操作されることを禁止したい場合は、確実に操作が終了していて欲しい操作要求パケットの送信数カウントを複数リンク同期情報として前記操作要求パケットに付加する、並列分散計算システム。

【請求項２】

前記送信先計算機は、前記複数リンク同期情報の付加された操作要求パケットを受信した場合に、前記受信側プロセスの前記受信数カウント部を参照して、前記複数リンク同期情報に記述されたカウント数が前記受信数カウント部のカウント数以下であれば、操作要求パケットの処理を実行し、前記受信数カウント部及び前記既着カウント記録部の更新作業を行う、請求項１記載の並列分散計算システム。

【請求項３】

前記送信先計算機は、前記複数リンク同期情報の付加された操作要求パケットを受信した場合に、前記受信側プロセスの前記受信数カウント部を参照して、前記複数リンク同期情報に記述されたカウント数が前記受信数カウント部のカウント数より大きければ、前記受信数カウント部のカウント数が増えて、前記複数リンク同期情報に記述されたカウント数以上になるまで、操作要求パケットの処理を遅延し、前記受信数カウント部及び前記既着カウント記録部の更新作業も遅延させる、請求項２記載の並列分散計算システム。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、アドレス変換バッファ（TLB）を有するプロセッサ、物理メモリ、及び、当該物理メモリに直接アクセス可能なネットワークインターフェースコントローラ（NIC）を備えた複数の計算機を、複数のデータリンクを介して相互接続した並列分散計算システムに関するものである。

【背景技術】

【０００２】

本願発明者は、非特許文献１に示すように、メモリベース通信同期方式に基づく通信同期機構のうちのメモリベース通信ファシリティ（MBCF；Memory Based Communication Facility）の開発を進めている。このMBCFは、特殊な通信同期ハードウェアを一切必要とせず、一般のネットワークインタフェースカード（NIC）を使用して、ソフトウェアのみで遠隔メモリ操作による高速高機能通信同期を実現する機構である。

【先行技術文献】

【非特許文献】

【０００３】

【文献】Matsumoto, T.: A Study on Memory-Based Communications and Synchronization in Distributed-Memory Systems. Dissertation Thesis, Graduate School of Science, Univ. of Tokyo (February 2001).

【発明の概要】

【発明が解決しようとする課題】

【０００４】

具体的に上記のMBCFは、アドレス変換バッファ（TLB）を有するプロセッサ、物理メ

モリ、及び、当該物理メモリに直接アクセス可能なネットワークインターフェースコントローラ（NIC）を備えた計算機を用いて構成される。例えば、送信元計算機のプロセス（以下、送信側プロセス）が、送信先計算機のプロセスを規定する操作対象プロセス（以下、受信側プロセス）の識別子と、当該受信側プロセスのメモリ領域を規定する操作対象アドレスと、書き込まれるデータサイズと、データ列とを含む操作要求パケットを送信する。そして、送信先計算機が、送信側プロセスが送信した操作要求パケットを受信し、受信側プロセス及び操作対象アドレスにより規定されるメモリ領域にデータ列を格納する。

【０００５】

ところで、本願発明者は、複数の計算機を、複数のデータリンクを介して相互接続した（リンクアグリゲーション結合された）並列分散計算システムを構成することを考えている。このとき、各ノード（計算機）には複数のデータリンクが存在することが想定されるが、複数のデータリンクが存在するがゆえに、複数のリンクにパケットを負荷分散することにより、遠隔メモリ操作要求パケットの到着順序が乱れてメモリコンシステンシが破れてしまうという問題が考えられる。

【０００６】

そこで本発明は、上記問題点を解決すべくなされたものであり、アドレス変換バッファ（TLB）を有するプロセッサ、物理メモリ、及び、当該物理メモリに直接アクセス可能なネットワークインターフェースコントローラ（NIC）を備えた複数の計算機を、複数のデータリンクを介して相互接続した並列分散計算システムにおいて、パケットの到着順序の乱れが存在する状況においてメモリコンシステンシを保持することを主たる課題とするものである。

【課題を解決するための手段】

【０００７】

すなわち本発明に係る並列分散計算システムは、アドレス変換バッファ（TLB）を有するプロセッサ、物理メモリ、及び、当該物理メモリに直接アクセス可能なネットワークインターフェースコントローラ（NIC）を備えた複数の計算機を、複数のデータリンクを介して相互接続した並列分散計算システムであって、送信元計算機のプロセス（以下、送信側プロセス）が、送信先計算機のプロセスを規定する操作対象プロセス（以下、受信側プロセス）の識別子と、当該受信側プロセスのメモリ領域を規定する操作対象アドレスと、書き込まれるデータサイズと、データ列とを含む操作要求パケットを送信し、前記送信先計算機が、前記操作要求パケットを受信し、前記受信側プロセス及び前記操作対象アドレスにより規定されるメモリ領域に前記データ列を格納するものであり、前記送信側プロセスは、前記受信側プロセス毎に前記操作要求パケットの送信数をカウントし、前記操作要求パケットに当該操作要求パケットのカウントアップ前又は後のカウント数を添付する送信数カウント部を有し、前記受信側プロセスは、前記送信側プロセス毎に前記操作要求パケットの受信数をカウントする受信数カウント部を有し、前記送信先計算機は、前記受信数カウント部のカウント数と、前記操作要求パケットに添付されたカウント数とが連続している場合に、前記操作要求パケットの処理を行い、前記操作要求パケットに添付されたカウント数によって前記受信数カウント部のカウント数を更新することを特徴とする。

【０００８】

このような並列分散計算システムであれば、複数の計算機を複数のデータリンクを介して相互接続し、複数の計算機が相互にMBCFで通信同期を行う並列分散計算システムにおいて、送信側プロセスが、操作要求パケットに当該操作要求パケットのカウントアップ前又は後のカウント数を添付する送信数カウント部を有し、受信側プロセスが、送信側プロセス毎に操作要求パケットの受信数をカウントする受信数カウント部を有しており、送信先計算機が、受信数カウント部のカウント数と、操作要求パケットに添付されたカウント数とが連続している場合に、操作要求パケットの処理を行い、操作要求パケットに添付されたカウント数によって前記受信数カウント部のカウント数を更新するので、操作要求パケットの到着順序の乱れを必要に応じて解消することができる。

【０００９】

未着の操作要求パケットが存在する場合の具体的な処理態様としては、前記送信先計算機は、前記受信側プロセスへの到着順序が乱れて未着の操作要求パケットが存在する場合には、当該未着の操作要求パケットの前までしか前記受信数カウント部によるカウント数の更新を行わずに、前記未着の操作要求パケット以降の操作要求パケットの到着に関しては、操作要求パケットの処理を行い、前記受信数カウント部とは別に操作要求パケットに添付されたカウント数を既着カウント記録部に記録する。

【００１０】

未着の操作要求パケットが到着した後の具体的な処理態様としては、前記送信先計算機は、未着だった操作要求パケットの到着によって前記既着カウント記録部を用いて前記受信側プロセスへの到着記録の記録内容が前記受信数カウント部から連続した場合に、その連続している範囲内で前記受信数カウント部を更新し、前記既着カウント記録部から更新済みの記録を削除する。

【００１１】

操作要求パケットの処理順序を保障するためには、前記送信側プロセスは、自分が送ろうとする操作要求パケットがそれに先行する操作要求パケットの処理が済むまで操作されることを禁止する意図を伝えることとし、確実に操作が終了していて欲しい先行する操作要求パケットの送信数カウントを複数リンク同期情報として前記操作要求パケットに付加する。

【００１２】

操作要求パケットの処理順序を保障するための具体的な操作要求パケットの処理態様としては、前記送信先計算機は、前記複数リンク同期情報の付加された操作要求パケットを受信した場合に、前記受信側プロセスの前記受信数カウント部を参照して、前記複数リンク同期情報に記述されたカウント数が前記受信数カウント部のカウント数以下であれば、操作要求パケットの処理を実行し、前記受信数カウント部及び前記既着カウント記録部の更新作業を行う。

【００１３】

操作要求パケットの処理順序を保障するための具体的な操作要求パケットの処理態様としては、前記送信先計算機は、前記複数リンク同期情報の付加された操作要求パケットを受信した場合に、前記受信側プロセスの前記受信数カウント部を参照して、前記複数リンク同期情報に記述されたカウント数が前記受信数カウント部のカウント数より大きければ、前記受信数カウント部のカウント数が増えて、前記複数リンク同期情報に記述されたカウント数以上になるまで、操作要求パケットの処理を遅延し、前記受信数カウント部及び前記既着カウント記録部の更新作業も遅延させる。

【発明の効果】

【００１４】

このように構成した本発明によれば、複数の計算機を複数のデータリンクを介して相互接続し、複数の計算機が相互にＭＢＣＦで通信同期を行う並列分散計算システムにおいて、パケットの到着順序の乱れが存在する状況においてメモリコンシステンシを保持することが低コストで達成できる。

【図面の簡単な説明】

【００１５】

【図１】本発明の一実施形態における並列分散計算システムの全体構成を模式的に示す図である。

【図２】同実施形態の送信元計算機と送信先計算機の物理構成を示す模式図である。

【図３】同実施形態のＭＢＣＦ_ＷＲＩＴＥの使用環境を示す模式図である。

【図４】同実施形態の操作要求側におけるパケット生成過程を含む、操作要求パケット送信の動作手順を示す模式図である。

【図５】同実施形態の受信側ノードに操作要求パケットが到着した状況を示す模式図である。

【図６】同実施形態の受信側ノードにおけるパケット受信割込みルーチン内のＭＢＣＦ関連

処理を示す模式図である。
【図７】操作要求パケットの到着順序が乱れた状態を示す模式図である。
【図８】同実施形態の計算機の機能ブロック図である。
【図９】ノードNa、Nb、Ncにおいて、ノードNaの送信数カウント部及び受信数カウント部を代表させて示す模式図である。
【図１０】同実施形態の同期情報を挿入した遠隔メモリ操作の使用例１である。
【図１１】同実施形態の同期情報を挿入した遠隔メモリ操作の使用例２である。
【符号の説明】
【００１６】
１００・・・並列分散計算システム
２　・・・計算機
２Ｘ　・・・送信元計算機
２Ｙ　・・・送信先計算機
２１　・・・プロセッサ
２２　・・・物理メモリ
２３　・・・ネットワークインターフェースコントローラ（NIC）
３　・・・データリンク
２０１・・・送信数カウント部
２０２・・・受信数カウント部
２０３・・・既着カウント記録部
【発明を実施するための形態】
【００１７】
以下、本発明の一実施形態に係る並列分散計算システム１００について、図面を参照して説明する。
【００１８】
本実施形態の並列分散計算システム１００は、図１に示すように、複数の計算機２を複数のデータリンク３を介して相互接続したものである。複数の計算機２を複数のデータリンク３を介して相互接続することによりリンクアグリゲーション結合された並列分散計算システムが構成される。
【００１９】
各計算機２は、図２に示すように、アドレス変換バッファ（TLB；Translation Look-aside Buffer）を有するプロセッサ２１、物理メモリ２２、及び、当該物理メモリ２２に直接アクセス可能なネットワークインターフェースコントローラ（NIC；Network Interface Card）２３を備えたものである。
【００２０】
そして、並列分散計算システム１００は、特殊な通信同期ハードウェアを一切必要とせず、一般のネットワークインタフェースカード（NIC）２３を使用して、ソフトウェアのみで遠隔メモリ操作による高速高機能通信同期を実現するメモリベース通信ファシリティ（MBCF；Memory Based Communication Facility）を構築するものである。具体的に並列分散計算システム１００は、各計算機２のカーネル空間内に格納されたオペレーティングシステム（OS）によって、メモリベース通信ファシリティ（MBCF）を構築する。
【００２１】
並列分散計算システム１００は、例えば、以下に示す遠隔メモリ書き込みを行うwriteコマンド（MBCF_WRITE）や遠隔メモリ読み出しを行うreadコマンド（MBCF_READ）などの各種の操作コマンドのバリエーションを有する。
【００２２】
例えば、送信元計算機２（２Ｘ）のプロセス（以下、送信側プロセス）が、送信先計算機２（２Ｙ）のプロセスを規定する操作対象プロセス（以下、受信側プロセス）の識別子と、当該受信側プロセスのメモリ領域を規定する操作対象アドレスと、書き込まれるデータサイズと、データ列とを含む操作要求パケットを送信し、送信先計算機２が、操作要求

パケットを受信し、受信側プロセス及び操作対象アドレスにより規定されるメモリ領域にデータ列を格納する（MBCF_WRITE）。

【００２３】

また、送信側プロセスが、送信先計算機２（２Ｙ）のプロセスを規定する操作対象プロセス（以下、受信側プロセス）の識別子と、当該受信側プロセスのメモリ領域を規定する操作対象アドレスと、読み出すデータサイズと、送信側プロセスのデータ格納領域アドレスとを含む操作要求パケットを送信し、送信先計算機２（２Ｙ）が、操作要求パケットを受信し、受信側プロセス及び操作対象アドレスにより規定されるメモリ領域からデータ列を読み出して、送信側プロセスのデータ格納領域に返送する（MBCF_READ）。

【００２４】

ここで、MBCF_WRITEの手順について、図３～図６を参照して詳述する。なお、その他のコマンドについても基本となる手順は共通している。

【００２５】

図３はMBCF_WRITEの使用環境を示しており、Pnode1が操作要求パケットの送信元ノード（送信元計算機）であり、Pnode２が受信側ノード（送信先計算機）である。送信元ノードのプロセッサは、自分のメモリのNIC DMA領域内に配送情報の書かれたヘッダとペイロードを含むパケットイメージを作成する。NIC DMA領域メモリには送信又は受信のためにNICが直接アクセスすることが可能である。送信用のパケットイメージの生成が終わると、プロセッサはNICにDMA読み出しによる送信動作を開始させるための指示（kick動作）を行う。受信側ノード（Pnode2）は、自分のメモリのNIC DMA領域内に到達するパケットのためのリングバッファを持っている。受信側ノードのNICは、パケットが自分のノード宛てのパケットであるか判断して（通常はMACアドレスで判断する）、自分宛てのパケットに関しては、リングバッファにそのコピーを生成する。その後、NICは、受信側ノードのプロセッサにパケットの到着を知らせるためにハードウェア割り込みを発生させる。

【００２６】

次に、図４を参照して、操作要求側におけるパケット生成過程を含む、操作要求パケット送信の動作手順を示す。

【００２７】

送信側プロセス（要求側タスク）では、受信側プロセス（要求先タスク）の識別子[Ltask1]、受信側プロセスの操作対象メモリアドレス[Laddr1]、受信側プロセスのメモリ空間操作のためのアクセスキー[AccessKey]、MBCFのコマンド種別[MBCF_WRITE]、遠隔書き込みを行うデータサイズ[n]、書き込むデータが格納された領域の先頭へのポインタ[Laddr0]から成るパラメータを用意する。そして、これらのパラメータを伴ってMBCF要求送信用システムコールを呼び出す。システムコールを受けてOSは、送信側プロセスのタスク表を参照して、受信側プロセスを示す論理タスクIDを物理タスクID[(Pnode2, Ptask5)]に変換する。物理タスクIDは物理的なノードIDであるPnode2を含んでいるため、この情報から受信側ノードへの経路情報（配送先情報）が設定できる。使用するネットワークがEthernetであれば、MACアドレスが配送先情報として使用される。この配送先情報によってNICは操作要求パケットを受信側ノードまで配送することが可能になる。そして、OSはNICに操作要求パケットを送信させる。

【００２８】

次に、図５及び図６を参照して、操作要求先における操作要求パケットの受信手順を示す。

図５は、受信側ノードに操作要求パケットが到着した状況を示す。操作要求パケットは受信側ノード[Pnode2]までネットワークによって運ばれてくる。パケット到着時に、受信側ノードのNICはDMAによって操作要求パケットのデータイメージをリングバッファにコピーし、その後に受信側ノードのプロセッサにパケットが到着したことを知らせるために、割込み信号を発生させる。

【００２９】

図６に受信側ノードにおけるパケット受信割込みルーチン内のMBCF関連処理を示す。NICからの受信割込み発生により、受信側ノードのプロセッサはパケット受信ルーチンに制御が切り替わり、NICによって要求される低レベルのパケット受信手順を最初に実行する。パケット受信ルーチンでは、まず物理タスクID（図中のPtask5）から受信側プロセスを特定する。具体的には、そのプロセスのプロセス構造体へのポインタを得る。次に、操作要求パケット内のAccessKeyが受信側プロセスのものと一致しているかどうかチェックして、一致している場合のみ受信側プロセスのメモリ空間内のメモリ操作を許可する。一致した場合は、メモリ空間のコンテクストを受信側プロセスのものに切り替えて、操作要求パケットで運ばれてきたnバイトのデータを操作対象論理アドレス（Laddr1）から、特権レベルとしてではなくユーザ権限のstore命令によって書き込む。そして、メモリ空間のコンテクストを割込み発生時のものに戻しておく。ここまでで割込みルーチン内の処理が終了する。

【００３０】

そして、本実施形態の並列分散計算システム１００は、複数のデータリンク３を有することから、図７に示すように、操作要求パケットの到着順序が乱れてしまうという問題が考えられる。このため、本実施形態の並列分散計算システム１００は、受信側プロセスへの操作要求パケットの到達順序をユーザの意図する範囲内において保障する機能を有している。

【００３１】

ユーザの意図する範囲内のみの保障というのは、具体的には、受信側プロセスのメモリ上にA, B, C, D, Fという変数があり、送信側プロセスがそのA, B, C, DにそれぞれMBCF_WRITEで遠隔書き込みを行い、最後に、それらの書き込みが済んだことを示す値（例えば１）をFにMBCF_WRITEで書き込むことにすると、A, B, C, DのMBCF_WRITEの間の順序保障は不要であるが、それら４つのMBCF_WRITEとFに対するMBCF_WRITEの間の順序保障は不可欠である。逆に、全部のMBCF_WRITE間で順序保障を行うことは過剰の保障であり、余分なオーバヘッドを発生させる可能性がある。よって、本実施形態では、Dの書き込みまでのMBCF操作要求パケットがすべて処理された後に、Fの書き込みを行うという同期情報を、FのMBCF_WRITE要求パケットにのみ付加する。

【００３２】

操作要求パケットの到達順序をユーザの意図する範囲内において保障する機能は、具体的には、図８に示すように、送信元計算機２（２Ｘ）の各送信側プロセスは、受信側プロセス毎に操作要求パケットの送信数をカウントし、操作要求パケットに当該操作要求パケットのカウントアップ前又は後のカウント数を添付する送信数カウント部２０１を有している。また、送信先計算機２（２Ｙ）の各受信側プロセスは、送信側プロセス毎に操作要求パケットの受信数をカウントする受信数カウント部２０２を有している。

【００３３】

そして、送信先計算機２（２Ｙ）は、受信数カウント部２０２のカウント数と、操作要求パケットに添付されたカウント数とが連続している場合に、操作要求パケットの処理を行い、操作要求パケットに添付されたカウント数によって受信数カウント部２０２のカウント数を更新する。

【００３４】

ここで、送信先計算機２（２Ｙ）は、受信側プロセスへの到着順序が乱れて未着の操作要求パケットが存在する場合には、当該未着の操作要求パケットの前までしか受信数カウント部２０２によるカウント数の更新を行わない。また、未着の操作要求パケット以降の操作要求パケットの到着に関しては、以下に述べる複数リンク同期情報が付加されていなければ、操作要求パケットの処理を行い、受信数カウント部２０２とは別に操作要求パケットに添付されたカウント数を既着カウント記録部２０３に記録する。

【００３５】

そして、送信先計算機２（２Ｙ）は、未着だった操作要求パケットの到着によって既着カウント記録部２０３を用いて受信側プロセスへの到着記録の記録内容が受信数カウント

部２０２から連続した場合に、その連続している範囲内で受信数カウント部２０２を更新し、既着カウント記録部２０３から更新済みの記録を削除する。

【００３６】

このとき、送信側プロセスは、自分が送ろうとする操作要求パケットがそれに先行する操作要求パケットの処理が済むまで操作されることを禁止したい場合、つまり到着順序をユーザ（送信側プロセス）が保障したい場合は、確実に操作が終了していて欲しい操作要求パケットの送信数カウントを複数リンク同期情報として操作要求パケットに付加する。

【００３７】

送信先計算機２（２Ｙ）は、複数リンク同期情報の付加された操作要求パケットを受信した場合に、受信側プロセスの受信数カウント部２０２を参照して、複数リンク同期情報に記述されたカウント数が受信数カウント部２０２のカウント数以下であれば、操作要求パケットの処理を実行し、受信数カウント部２０２及び既着カウント記録部２０３の更新作業を行う。

【００３８】

また、送信先計算機２（２Ｙ）は、複数リンク同期情報の付加された操作要求パケットを受信した場合に、受信側プロセスの受信数カウント部２０２を参照して、複数リンク同期情報に記述されたカウント数が受信数カウント部２０２のカウント数より大きければ、受信数カウント部２０２のカウント数が増えて、複数リンク同期情報に記述されたカウント数以上になるまで、操作要求パケットの処理を遅延し、受信数カウント部２０２及び既着カウント記録部２０３の更新作業も遅延させる。

【００３９】

次に、受信数カウント部２０２のカウント数の更新方法について、既着カウント記録部２０３にビットベクタを用いた一例を説明する。

ビットベクタを使ってパケット番号に抜けが無い状態でカウンタをカウントアップが可能になったら、カウンタとビットベクタを更新する。具体的な更新方法は以下である。

【００４０】

ビットベクタ長8bitで101番のパケットが欠落して、102番、103番が到着している場合は、カウンタの値は100で、ビットベクタは二進数で00000110となる。つまり、最下位より(102-100)、(103-100)番目のbitが1にセットされる。

【００４１】

ここに105番目のパケットが到着すると、カウンタの値は100のままで、ビットベクタは二進数で00010110となる。つまり、(105-100)番目のbitが1にセットされる。

【００４２】

この状態で、101番目のパケットが到着すると、(101-100)番目のbitが1にセットされ、ビットベクタは二進数で00010111となる。一番右端のbitに1がセットされているので、このことはカウンタの値に連続したパケットの到着を意味し、一番右端のbitがゼロになるまでこのビットベクタを右シフトして、そのシフト数をカウンタに加える。

【００４３】

つまり、シフト数は3になるので、カウンタの値は103で、ビットベクタは二進数で00000010となる。ここに104番目のパケットが到着すると、(104-103)番目のbitが1にセットされ、ビットベクタは二進数で00000011となり、先ほどと同じく、1が続く2をカウンタに加えてカウンタの値は105で、ビットベクタは二進数で00000000となる。ビットベクタがオールゼロの状態は、パケットに抜けが無くパケットがカウンタの値まで到着している状態である。

【００４４】

また、以下の方法であっても良い。

今、カウンタの値は105で、ビットベクタは二進数で00000000で106番のパケットが到着した（順序の乱れがない）とすると、一旦、ビットベクタの(106-105)番目のbitが1にセットされる。このとき、カウンタの値は105で、ビットベクタは二進数で00000001となり、ビットベクタの一番右端のbitに1がセットされているので、1bit分右シフトして

、カウンタの値は+1で106となり、ビットベクタは二進数で00000000となる。
【００４５】
この方法では、順序の乱れがないことを確認する手順を入れずに、抜けがある場合の手順と同一で処理することが可能となる。
【００４６】
なお、上記では、8bitのビットベクタを用いているが、データリンクが８本までのリンクアグリゲーションを仮定すると、ビットベクタは32bitぐらいあれば、実用上問題はない。万が一、ビットベクタのビット幅を超える順序の混乱が生じた場合は、ビットベクタのビット幅を超える番号を持つパケットは受け取りを拒否して再送対象にする。
【００４７】
次に、図９に示すように、３つの計算機２において、それぞれをノードNa、Nb、Ncとし、
ノードNaに、プロセスTa1、Ta2、Ta3、
ノードNbに、プロセスTb1、Tb2、Tb3、
ノードNcに、プロセスTc1、Tc2、Tc3、
の各３つのプロセス（タスクもいう。）が存在して、相互にMBCFで通信同期を行っている場合を考える。なお、送信側プロセスは、MBCFを使って同一計算機の中の自らを含む他プロセスへのメモリ操作要求を遠隔メモリ操作の一環として行うことができる。
【００４８】
この場合、プロセスTa1には、送信数カウント部２０１（送信数カウンタ）が、プロセスTa1、Ta2、Ta3、Tb1、Tb2、Tb3、Tc1、Tc2、Tc3に対応して、送信数カウンタXa1_Ta1、Xa2_Ta1、Xa3_Ta1、Xb1_Ta1、Xb2_Ta1、Xb3_Ta1、Xc1_Ta1、Xc2_Ta1、Xc3_Ta1の９個が存在する。ここでは、自分自身や同じノードのプロセスに対しても順序制御すると仮定している。なお、これらではパケット到着順序の乱れが想定されないので順序制御を省略可能であるが、通信相手の受信側プロセスが同一計算機内に存在するかどうかをユーザやプログラマが意識しなければならないことは、プログラム作成の困難度を増す。このため、受信側プロセスがどの計算機に存在するかに関わらず、同一仕様の送信数カウントと複数リンク同期情報を含むMBCF操作要求パケットを使用するようにシステムを構成する。
【００４９】
そして、例えば、プロセスTa1がプロセスTb2に向かって操作要求パケットを発行した場合には、送信数カウンタXb2_Ta1がカウントアップされる。同じように、他のプロセスにも９個ずつ送信数カウンタが存在する。
【００５０】
また、プロセスTa1には、受信数カウント部２０２（受信数カウンタ）が、プロセスTa1、Ta2、Ta3、Tb1、Tb2、Tb3、Tc1、Tc2、Tc3に対応して、受信数カウンタRa1_Ta1、Ra2_Ta1、Ra3_Ta1、Rb1_Ta1、Rb2_Ta1、Rb3_Ta1、Rc1_Ta1、Rc2_Ta1、Rc3_Ta1の９個が存在する。
【００５１】
そして、例えば、プロセスTa1宛の操作要求パケットをノードNaがプロセスTc2から受信し、その操作要求パケット内のカウント値が（Rc2_Ta1+１）であれば、メモリ操作後に受信数カウンタRc2_Ta1をカウントアップする。操作要求パケット内のカウント値が（Rc2_Ta1+１）より大きければ、メモリ操作後に受信数カウンタRc2_Ta1に対応した既着カウント記録部２０３（ビットベクタを用いて構成）に記録して、追い越された操作要求パケットのメモリ操作終了後まで受信数カウンタRc2_Ta1のカウントアップを遅延させる。
【００５２】
プロセスTc2からのプロセスTa1への操作要求パケットに複数リンク同期情報（同期オプション）が付加されている場合は、受信数カウンタRc2_Ta1の値が同期オプションで記述されたカウント値以上であった場合は、同期オプションなしの操作要求パケットと同じ

処理を行う。
【００５３】
一方で、受信数カウンタRc2_Ta1の値が同期オプションに記述されたカウント値より小さい場合は、受信数カウンタRc2_Ta1の値が同期オプションに記述されたカウント値以上になるまでメモリ操作や既着カウント記録部（ビットベクタ）への登録も含めて、実行を遅延させる。
【００５４】
＜２．本実施形態の効果＞
このように構成した本実施形態の並列分散計算システム１００によれば、複数の計算機２を複数のデータリンク３を介して相互接続し、複数の計算機２が相互にＭＢＣＦで通信同期を行うものにおいて、送信側プロセスが、操作要求パケットに当該操作要求パケットのカウントアップ前又は後のカウント数を添付する送信数カウント部２０１を有し、受信側プロセスが、送信側プロセス毎に操作要求パケットの受信数をカウントする受信数カウント部２０２を有しており、送信先計算機２（２Ｙ）が、受信数カウント部２０２のカウント数と、操作要求パケットに添付されたカウント数とが連続している場合に、操作要求パケットの処理を行い、操作要求パケットに添付されたカウント数によって受信数カウント部２０２のカウント数を更新するので、操作要求パケットの到着順序の乱れが発生する環境下において、ユーザの意図する範囲内でメモリ操作順序を保障する、つまり、メモリコンシステンシを維持することができる。
【００５５】
従来、分散配置されたメモリに対して共有メモリアクセスを許す分散共有メモリ型並列計算機において、メモリの操作順序を保障するために、すべての遠隔メモリ操作に対して確認応答（Acknowledgement, 以下ではAck）が要求元に返送され、順序を保障しないとならない場合は、それまでの遠隔メモリ操作すべてに対するAckが戻って来たことを確認後に、順序を保障すべき遠隔メモリ操作の要求を行っていた。この順序保障の仕組みはメモリバリアと呼ばれている。それに対して、本発明においては、順序保障に対して遠隔メモリ操作のAckを必要とせず、直近に要求した遠隔メモリ操作に続けて、順序を保障すべき遠隔メモリ操作の要求を行うことが可能である。このことは先行する遠隔メモリ操作のAckが返ることを待つ必要がないことを意味し、大幅な性能向上を可能にする。また、送信側プロセスは自分が保持する送信数カウント部２０１の値を使って複数リンク同期情報を付加することができ、送信先計算機２（２Ｙ）は同じ計算機内の受信側プロセスに属する受信数カウント部２０２の値を使って、同期成立・非成立の判定が可能である。つまり、同期を行うために必要な情報を、送信側プロセスも送信先計算機２（２Ｙ）も極めて低コストで入手することが可能である。
【００５６】
本発明を使うことによって、Ackを待つ必要がないこと以上に同期によるオーバヘッドコストを削減できる可能性がある。その可能性について、具体例によって説明する。受信側プロセスに、A1, B1, C1というメモリ領域があり、それらに値の格納を終えたことを示すF1というフラグ変数がある。同様に、同じ受信側プロセスにA2, B2, C2, A3, B3, C3というメモリ領域があり、A2, B2, C2に値の格納を終えたことを示すF2というフラグ変数、A3, B3, C3に値の格納を終えたことを示すF3というフラグ変数がある。すべてのメモリ領域とフラグへの値の格納をMBCF_WRITEによって実現するものとする。A1, B1, C1とF1、A2, B2, C2とF2、A3, B3, C3とF3の遠隔書き込みの間にはそれぞれ守るべき順序関係があるが、AnとBnとCnの間やFm, Fn （ただし、m ≠ n）の間には順序関係はない。これを単純に図１０のように、MBCF_WRITE要求パケットを発行すると、F1, F2, F3への書き込みは、それぞれ直前のC1, C2, C3への書き込みが行われた後であることを保障せねばならない。複数のリンクが存在する下では、要求を出した順番で届くことは保障されないため、送信先計算機２（２Ｙ）において、操作要求パケット処理を同期のために遅延させなければならない可能性がかなりある。これに対して、図１１のように、無関係なMBCF_WRITEを間に挿む形で、順序を保障すべきメモリ領域への書き込みとフラグ変

数への書き込みの距離を離すことにより、付加するリンク同期情報のカウント値が、直前に発行したMBCF要求カウントの送信数カウント部２０１の値ではなく、３パケット前の送信数カウント部２０１の値で済むことになる。このことは送信先計算機２（２Ｙ）への要求パケットの到着は直前の２つのパケットより早くても構わないことを意味し、図１１のケースは図１０のケースよりも明らかに、フラグ変数への書き込みが同期待ちによって、送信先計算機２（２Ｙ）において遅延される可能性が低い。つまり、本発明を使用し、同期の必要な操作要求と不要な操作要求、ならびに順序関係がない操作要求の発行順序を最適化することにより、同期のオーバヘッドコストをより削減できる可能性がある。

【００５７】

その他、本発明は前記実施形態に限られず、その趣旨を逸脱しない範囲で種々の変形が可能であるのは言うまでもない。

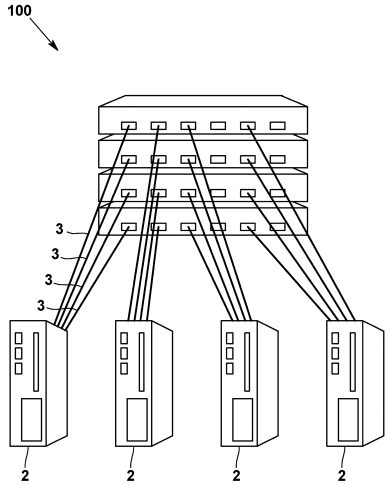
【産業上の利用可能性】

【００５８】

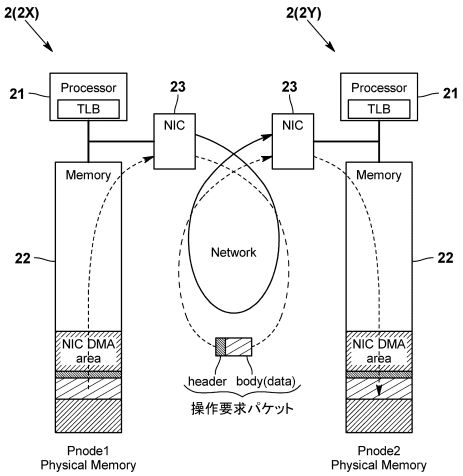
本発明は、複数の計算機を複数のデータリンクを介して相互接続し、複数の計算機が相互にMBCFで通信同期を行う並列分散計算システムにおいて、パケットの到着順序の乱れが存在する状況においてメモリコンシステンシを保持することを低コストで達成することができる。

【図面】

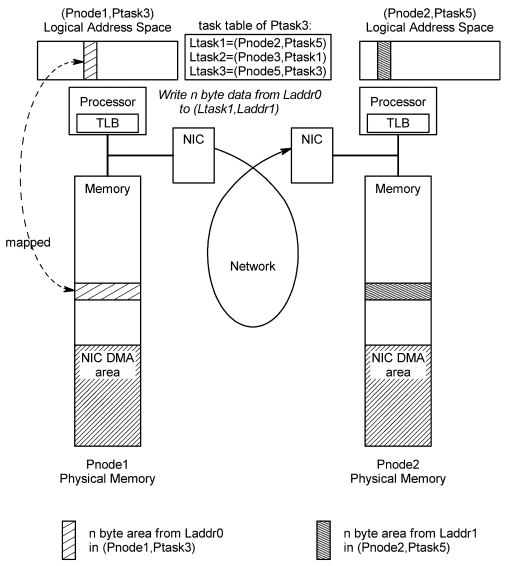
【図１】



【図２】

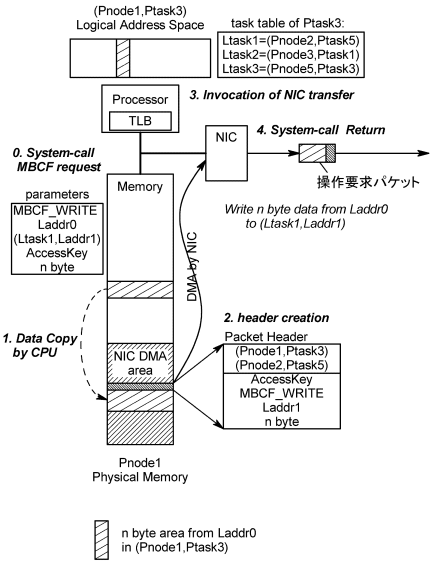


【図３】



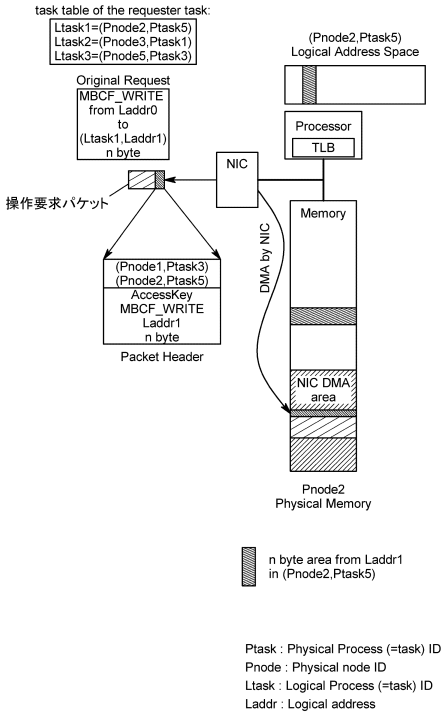
Ptask : Physical Process (=task) ID
Pnode : Physical node ID
Ltask : Logical Process (=task) ID
Laddr : Logical address

【図４】



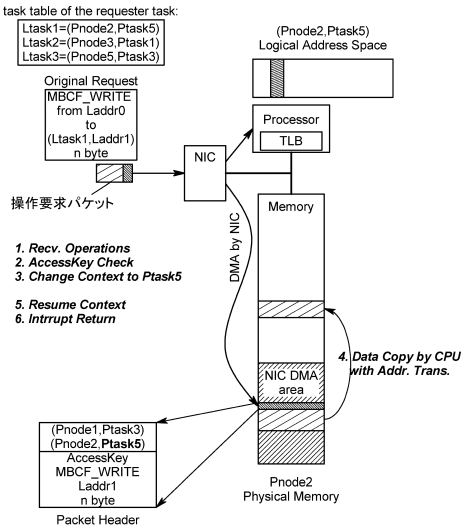
Ptask : Physical Process (=task) ID
Pnode : Physical node ID
Ltask : Logical Process (=task) ID
Laddr : Logical address

【図５】



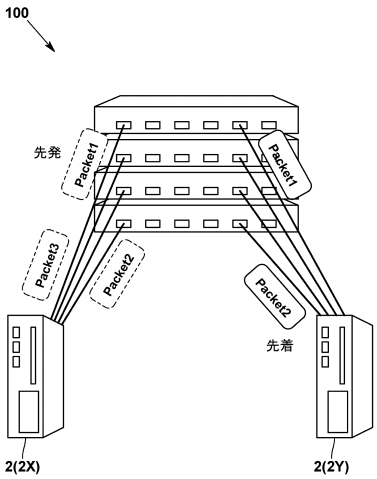
Ptask : Physical Process (=task) ID
Pnode : Physical node ID
Ltask : Logical Process (=task) ID
Laddr : Logical address

【図６】

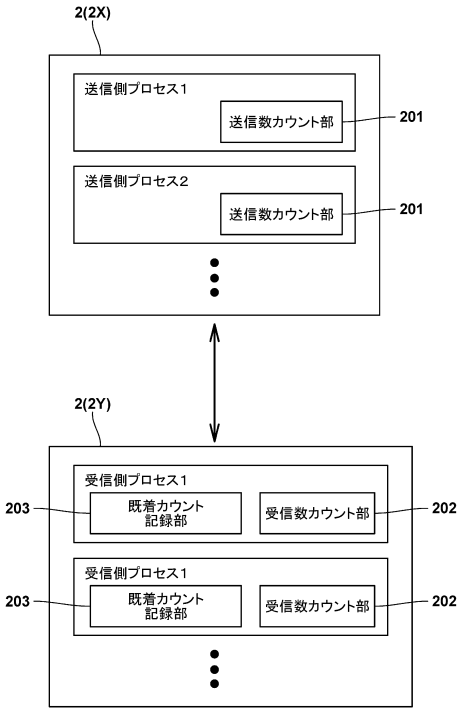


Ptask : Physical Process (=task) ID
Pnode : Physical node ID
Ltask : Logical Process (=task) ID
Laddr : Logical address

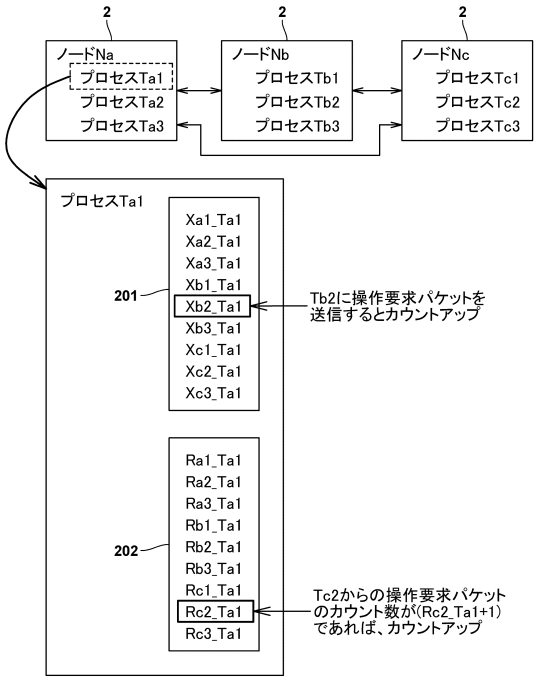
【図７】



【図８】



【図９】



【図１０】

送信数カウント	複数リンク同期情報	コマンド	対象アドレス	サイズ（バイト数）	データ列
1	———	MBCF_WRITE	A1のアドレス	N	a_1d_1,...
2	———	MBCF_WRITE	B1のアドレス	N	b_1d_1,...
3	———	MBCF_WRITE	C1のアドレス	N	c_1d_1,...
4	3	MBCF_WRITE	F1のアドレス	4	1
5	———	MBCF_WRITE	A2のアドレス	N	a_2d_1,...
6	———	MBCF_WRITE	B2のアドレス	N	b_2d_1,...
7	———	MBCF_WRITE	C2のアドレス	N	c_2d_1,...
8	7	MBCF_WRITE	F2のアドレス	4	1
9	———	MBCF_WRITE	A3のアドレス	N	a_3d_1,...
10	———	MBCF_WRITE	B3のアドレス	N	b_3d_1,...
11	———	MBCF_WRITE	C3のアドレス	N	c_3d_1,...
12	11	MBCF_WRITE	F3のアドレス	4	1

【図１１】

送信数カウント	複数リンク同期情報	コマンド	対象アドレス	サイズ（バイト数）	データ列
1	———	MBCF_WRITE	A1のアドレス	N	a_1d_1,...
2	———	MBCF_WRITE	A2のアドレス	N	a_2d_1,...
3	———	MBCF_WRITE	A3のアドレス	N	a_3d_1,...
4	———	MBCF_WRITE	B1のアドレス	N	b_1d_1,...
5	———	MBCF_WRITE	B2のアドレス	N	b_2d_1,...
6	———	MBCF_WRITE	B3のアドレス	N	b_3d_1,...
7	———	MBCF_WRITE	C1のアドレス	N	c_1d_1,...
8	———	MBCF_WRITE	C2のアドレス	N	c_2d_1,...
9	———	MBCF_WRITE	C3のアドレス	N	c_3d_1,...
10	7	MBCF_WRITE	F1のアドレス	4	1
11	8	MBCF_WRITE	F2のアドレス	4	1
12	9	MBCF_WRITE	F3のアドレス	4	1

フロントページの続き

(72)発明者　松本 尚
新潟県新潟市西区真砂一丁目１５番３６号 株式会社情報科学研究所内

審査官　三坂 敏夫

(56)参考文献　特開２０１４－１９１４９７（ＪＰ，Ａ）
特表２０１０－５１０５９０（ＪＰ，Ａ）
松本 尚，メモリーベース通信ファシリティの評価，電子情報通信学会技術研究報告，社団法人電子情報通信学会，2001年09月27日，Ｖｏｌ．１０１ Ｎｏ．３２９，第31頁－第40頁，ISSN:0913-5685
松本 尚 他，汎用超並列オペレーティングシステム：ＳＳＳ－ＣＯＲＥ，情報処理学会研究報告，社団法人情報処理学会，1996年08月27日，Ｖｏｌ．９６ Ｎｏ．７９，第115頁－第120頁，ISSN:0919-6072
丹羽 純平 他，汎用超並列オペレーティングシステムＳＳＳ－ＣＯＲＥ上の非対称分散共有メモリにおけるコンパイル技法，コンピュータソフトウェア，日本ソフトウェア科学会，1998年05月15日，ＶＯＬ．１５ ＮＯ．３，第242頁－第246頁，ISSN:0289-6540

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