中粒度メモリベース通信を支援する Memory-Based Processor II

松本 尚　野村真義
國澤亮太　平木 敬

中粒度のメモリベース通信を効率良く実現するネットワークインタフェースアーキテクチャ: Memory-Based Processor II (MBP2) を提案する。MBP2 は将来の汎用通信部品となる条件を満たすために汎用ネットワークインタフェースに対して上位コンパチビリティを持ち、ハードウェア量は同程度である。

Memory-Based Processor II
— A commodity supporting middle-grained memory-based communication —

Takashi MATSUMOTO , Masayoshi NOMURA ,
Ryota KUNISAWA and Kei HIRAKI

We propose a novel network interface architecture: "Memory-Based Processor II (MBP2)" which supports efficient middle-grained memory-based communications, legendary TCP/IP and UDP/IP. Although the hardware cost of the MBP2 is almost as same as that of conventional Network Interface Cards (NICs), in memory-based communications the MBP2 system is much superior to the NIC systems.

1. はじめに

筆者らはすべての通信同期を論理的なメモリアドレスを対象として実行するメモリベース通信という方式を 1992 年より提唱している [1]。この方式では通信同期の保護をメモリの保護に置き換えることができるとともに、メッセージパッシング型の通信よりもコピー回数やキュー操作回数を削減することができる。

WS クラスタや PC クラスタもしくはそれに近い形態を採る並列計算システムでは、専用のネットワークインタフェースを搭載するものが多い。しかし、近年のネットワークの進歩は目覚しく、WS/PC の汎用 LAN インタフェースにおいて通信速度が 1Gbit/s に達するものが現れている。そして、メモリベース通信を汎用ネットワークインタフェースに適用することによりユーザレベルの高速通信同期が実現可能であり、専用ネットワークが存在しなくても分散アプリケーションのみならず並列アプリケーションをも効率良く実行可能である。

本稿では、高スループットかつ低オーバヘッドのメモリベース通信を実現可能な汎用ネットワークインタフェースアーキテクチャを提案する。

2. 初代 MBP の概要と見直すべき点

Memory-Based Processor (**MBP**) [1] はメモリ管理ユニットの機能を遠隔メモリアクセスに拡張することによって、保護され仮想化されたユーザレベル高速通信同期を可能にするハードウェアサポート機構として世界初の物であった。MBP はローカルメモリバス上でメインプロセッサ (PE) に付加される形態でノードに組み込まれ、ノードを跨る共有メモリ空間を実現する。MBP は PE のメモリアクセスをスヌープして PE が遠隔メモリアクセスの必要な領域をアクセスすると、自動的にノード間の通信同期を行って遠隔メモリアクセスを実現する。つまり、PE は通常のメモリへの load/store によって、遠隔メモリアクセスが可能である。また、高効率な分散共有メモリ実現のために、MBP には複数のキャッシュコンシステンシプロトコルおよびメモリベース同期機能 [1] が実装されている☆。なお、MBP が実現する機能 (遠隔メモリ書き込み、遠隔メモリ読み出し、遠隔メモリ不可分操作、マルチキャスト、Ack コンバイニング、メモリベース同期機能、遠隔メモリキャッシュ機能等) を総称してメモリベース通信と呼ぶ。

前述のように MBP は PE の load/store をナイーブに遠隔メモリアクセスに拡張するために、細粒度の遠隔メモリアクセスを行う。しかし、通信効率を考えると、PE のメモリアクセス単位の粒度しかない細粒度通信は通信ヘッダの方が転送されるデータよりも量が大きく効率が悪い。近年通信速度の改善が目覚しいシリアルラインを用いた通信網では、パケット同期のためにさらに情報が付加されることが多いため、実質的な転送効率はさらに悪くなる。これらの通信用ヘッダは 20 ~ 50 バイト

† 東京大学 大学院理学系研究科 情報科学専攻
Department of Information Science, University of Tokyo

☆ 一部の機能は MBP に組み込まれたプログラマブルなプロトコル管理エンジンによって実現される

程度であるから、通信内容のデータサイズは数百バイト以上の粒度があることが望ましい。

この事実にも関わらず、MBP 考案時の 1991 年当時は高効率な共有メモリを実現するためには、細粒度の通信が必要であると考えていた。その主な理由として以下の二点が挙げられる。第一には、共有メモリモデルに従って記述された SPLASH [2] のようなアプリケーションプログラムにおいては、細粒度の遠隔メモリアクセスが多用されている。細粒度通信が MBP に採用されたもう一つの理由は、PE と MBP 間のインタフェースのオーバヘッド (遠隔メモリアクセスを開始 / 終了するときの PE 側のオーバヘッド) を極力抑えるためである。PE の load/store をそのまま遠隔メモリアクセスに拡張できれば余分な命令を PE で実行する必要がなく、非常に効率が高いと考えていた。これらの理由から MBP を使用するシステムは細粒度通信を基本とする設計がなされ、MBP は高効率な細粒度通信が可能な専用ネットワークと組み合わせて使うことが前提とされた。

けれども最近になって、ソフトウェア分散共有メモリ技術と最適化コンパイラ技術の進歩により、SPLASH 等に含まれる遠隔メモリアクセスをより大きな粒度の通信に最適化によって変換することが可能なことが判明し、そのための変換技術が開発された [3] [4] 。また、実用上重要な分散アプリケーションは、オーバヘッドの大きな通信手段を前提にして明示的に通信が記述されており、細粒度の通信を行うことはない。粒度の大きな通信が可能であり、通信頻度が少なくできるのであれば、ソフトウェア処理によって遠隔メモリアクセス時に多少のオーバヘッドが加わっても構わないはずである。それどころか、粒度の大きな通信の起動の方がオーバヘッドが小さくなる可能性を秘めている。なぜなら、細粒度の通信を多頻度で PE 外部に伝える方式では、データのバースト転送による恩恵を受けることができない。最近のプロセッサでは一回の外部メモリアクセスの間に数百の命令が実行可能であるため、内部キャッシュを活用してある程度粒度を大きくしてから、外部に転送した方がコスト的に有利な場合さえある。これらの事由から、細粒度通信と PE の単発 load/store による遠隔メモリアクセスの実現に拘る必然性がないことが明らかになった。

3. 中粒度メモリベース通信を実現する MBCF

Memory-Based Communication Facilities (**MBCF**) [5] [6] は汎用ネットワークインタフェースカード (NIC)[☆] を使ってシステムソフトウェアによって中粒度のメモリベース通信を可能にしたものである。機能的には MBP が提供するものと同じであるが、一回の通信同期に関与するデータサイズが可変であり、粒度の大きな通信同期が可能である。つまり、MBP のように PE のレジスタを介してメモリベース通信を行うのではなく、ローカルメモリの任意の部分を介してメモリベース通信を実現する。汎用 NIC を使いつつ保護と仮想化を実現するため、MBCF はカーネルレベルのプログラムで実現されており、メモリベース通信の要求側の処理はシステムコールによって起動される。このシステムコールによるオーバヘッドを小さく抑えるために、メモリベース通信起動専用のシステムコールを用意して余分な処理を極力排している。メモリベース通信の操作対象ノードでは、受信割り込みルーチン内において対象となるユーザアドレス空間が直接操作される。この時に、PE のメモリ管理ユニットをメモリベース通信のアドレス変換と保護に流用することで処理の高速化を図っている。MBCF によるメモリベース通信の結果は各ノードのユーザ空間のメモリ上に反映され、結果を受け取るためのシステムコールは基本的に必要とされない[☆☆]。すでに、Ethernet および Fast Ethernet 上の実装が済んでおり、ソフトウェアによる実装を活かしてパケット到着保証や順序保証のプロトコルがメモリベース通信同期機能とともに実装されている。なお、分散共有メモリ用のキャッシュの実現方式は MBP とは異なり、MBCF ではキャッシュプロトコルがユーザレベルのキャッシュエミュレーションと明示的な MBCF の使用によって実現され [5] [4]、通信最適化 [3] によって一回の通信粒度を大きくすることにより効率を向上させている。

MBCF の Ethernet 系ネットワークへの実装 (以下 MBCF/Ethernet と呼ぶ) を通して、専用システムコールのオーバヘッドはそれほど大きくないことが判明した。具体的には、SuperSPARC 85MHz において 4μsec 弱であり、最新プロセッサでは 1μsec を切ることも可能であろう。このコストは PE にとって外部メモリへのアクセス数回分に過ぎず、粒度を大きくして遠隔メモリアクセス (メモリベース通信) の頻度が少なくなっていれば、実行の大きな障害になるとは考えにくい。MBCF/Ethernet では汎用の Ethernet 用 MAC(Media Access Controller) を使用している。我々が使用した MAC は主記憶上のヘッダ部を含むパケットの最終イメージを DMA によって転送する形式になっている。このため、保護とデッドロック回避の観点からデータを送信ノードで 1 回コピーする必要が生じている。保護はパケットヘッダ部をユーザに勝手に生成させないためであり、デッドロック回避はパケット棄却によるデッドロック回避手法を許すためである。MBCF では複数のメモリベースコマンドを 1 パケットで送信可能にしているため、複数のデータ領域をマージして送る必要がある。この場合でも、これらの領域をポインタを使ってシステムコールに引き渡すことにより、最小の 1 コピーで複数コマンドのマージされたパケットを生成することができる。しかし、このコピーはキャッシュミスを起こす可能性の高いメモリ転送であるためコストがかなり大きく、例えば 1024 バイトのデータサイズにおけるメモリベース通信起動時のコピーコストは前出の条件

[☆] 付加的な通信用インタフェースカードに話を限定する意図はなく、マザーボードに組み込まれた通信インタフェースも便宜上 NIC と呼ぶ。

[☆☆] Memory-Based FIFO 等の機能において、不可分性を保証するために受信側にシステムコールが必要となる場合は存在する

下で約 12μsec である。

Ethernet の CSMA/CD のようにハードウェアレベルにおいてパケットの到着を保証しないネットワークインタフェースに用いられる到着保証ならびに順序保証プロトコルはかなり複雑である。しかし、パケットが消失せずに正常に届いている範囲 (つまり通常の範囲) では、到着保証順序保証のプロトコルのオーバヘッドは大きくない。具体的には、 SuperSPARC 85MHz 上の MBCF/Ethernet において、受信時に 2μsec 弱であり、送信時はほとんど無視できるレベルである。

4. Memory-Based Processor II

4.1 必 要 性

同一ハードウェア環境における MBCF と他の方式との比較によって、メモリベース通信の既存プロトコルに対する優位性は定性的に示されている [6]。

メモリベース通信の実現に際して、初代 MBP のように細粒度通信にこだわり、メインメモリに大量の高速タグメモリを付加するような実装は高価であり普及する見込みがない。実際、 MBP はフラグシップの超並列計算機の構成要素として考案されたものである。さらに、ユーザレベルのキャッシュエミュレーションと通信最適化技術の進歩によって、細粒度通信を行う分散共有メモリの必要性自体に疑問が投げ掛けられているため、初代 MBP のような実装形態は現時点では考えにくい。

それでは、 MBCF のように汎用 NIC を流用した PE 側のソフトウェアのみによる実装で十分なのであろうか。汎用 NIC の DMA はユーザ空間の任意のアドレス領域に対してデータ転送を行う能力が通常はない。また、仮にあったとしてもメモリベース通信のようなプロトコルを採用していないので、パケット内に示されたユーザ空間とアドレスを DMA の対象領域として使用することはできない。このため、 MBCF 方式の実装では PE が操作対象ノード内での最終的なデータ転送を行わざるを得ない。初代 MBP 流の実装と比べると、粒度が大きくなっているので頻度は少ないが、 1 回のデータ転送量が大きくなっているので、必ずしもデータ転送の負荷が十分に小さくなっているとは言えない。このため、 PE へのデータ転送のためのオーバヘッドはアプリケーションによっては無視できないものとなる。また、メモリベース通信では時分割の関係で操作対象タスクが実行されていないノードに対しても、いつでも通信や同期操作を行うことができる。この利点を享受して MBCF 流のメモリベース通信を多用すると、本来は他のタスクが使うべき PE 時間の多くをメモリベース通信実現のために使用することになり不公平になりかねない。これらのことから MBCF において、 PE が行っている処理 (主にデータ転送) を可能な限り PE 以外のハードウェアで実現する実装方式が望ましい。このメモリベース通信実装方式を持つ NIC を Memory-Based Processor II (**MBP2**) と呼ぶ。

4.2 基本設計方針

前小節で述べたように、 MBP2 は MBCF プロトコルの PE 負荷を軽減するための機能を持った NIC である。つまり、 MBP2 は MBCF のカーネルプログラム部分 (送信システムコールと受信割り込みルーチンとパケット到着保証用タイマ割り込みルーチン) を MBP2 内のハードウェア機構と内蔵エンジンのプログラム (ファームウェア) に置き換えて高速化ならびに低オーバヘッド化したものと捉えることができる。さらに、 MBP2 が将来のパーソナルコンピュータやワークステーションの通信用 commodity となりうるように、基本設計方針として以下の項目を掲げる。

(1) 汎用 I/O バス (PCI,SBus 等) 上の I/O デバイスとしての実装
(2) 中粒度 (可変粒度) のメモリベース通信を実現
(3) Gigabit Ethernet 等の汎用通信規格 (Ethernet 系) を採用
(4) TCP/IP, UDP/IP をサポート
(5) 汎用 NIC と同程度のハードウェアコスト
(6) 汎用 NIC としても高性能高付加価値

いくつかの項目について説明を補足する。項目 (1) については、 MBP2 と PE の連係方式に制約を与える可能性があるが、通信用 commodity を目指すためにはやむを得ないと判断した。項目 (3) は、 LAN 用 commodity としての普及度と中粒度程度のデータが 1 パケットで転送可能なことから Ethernet 系の採用を決めた。項目 (5) は、既存の Gigabit Ethernet 用 NIC 自体が大容量の内蔵高速メモリ、 DMA 機構、プログラマブルエンジンを内蔵しているものが多い。これらと MBP2 の間の基本的差異は NIC 自体がパケット内の情報を使って、メモリ保護の下でユーザ空間への DMA 転送が可能かどうかである。項目 (6) は、 TCP/IP や UDP/IP の PE の処理オーバヘッドを従来よりも軽減することや、通信の実時間暗号化 / 解読をサポートすることなどを意味する。

4.3 ハードウェア構成

MAC の外部インタフェース方式によって、 MBP2 のハードウェア構成は影響を受ける。 MAC が入出力別系統のデータパスを持ち、 MAC 内蔵の DMA 機構がない場合を例に取って、 MBP2 のハードウェア構成例を図 1に示す。図内の各サブブロックの名称と機能は以下の通りである。

- 内部高速メモリ (**EM**)
  MBP2 内部の高速メモリで、 I/O バス、 MAC、プロトコル管理エンジン (PME) の三系統から独立にアクセスされる。 MBP2 内のデータ転送のバッファとしての役目を担うため、達成目標の通信スループットの二倍と PME 用のバンド幅を足し合わせたメモリアクセスバンド幅が必要とされる。この高バンド幅を達成するために、 EM には広データバス幅を採用するか、多バンク構成を採用することが望ましい。図内の EA は EM 用のアドレスバスを示し、 ED は EM 用のデータバスを示す。
- Advanced Direct Memory Access Unit (**ADMA**)
  PE のメインメモリと MBP2 の EM の間のデータ

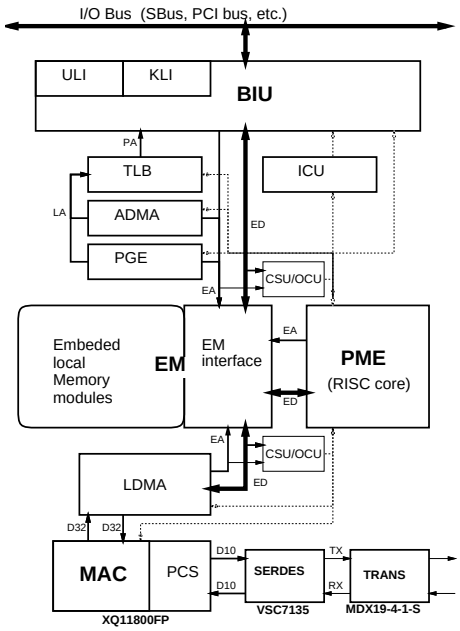

図1　MBP2 の構成

転送を行う 1 チャネルの DMA ユニット。メインメモリ側のアドレス指定にはコンテクスト ID と論理アドレスの組 (図中 LA) を使用し、 MBP2 の内部 TLB でメインメモリの物理アドレス (PA) に変換してアクセスする。アドレス境界やデータ境界がアラインされていない領域も過不足なく DMA 転送する。

- Media Access Controller (**MAC**)
  Ethernet プロトコルを実現するコントローラ。
- Physical Coding Sublayer (**PCS**)
  オートネゴシエーション、全二重通信のフロー制御のための一時休止、 8B/10B のエンコードとデコードを実現する。
- Local Direct Memory Access Unit (**LDMA**)
  内部高速メモリと MAC の間においてパケットの DMA 転送を行う。
- Translation Look-aside Buffer (**TLB**)
  ADMA および PGE が I/O バスを介してメインメモリをアクセスする際に参照して物理アドレスを得る。タスク ID (コンテクスト ID) のフィールドを各エントリに持ち、複数のアドレス空間が混在可能。 TLB ミス時の処理は PME に行わせる。
- プロトコル管理エンジン (**PME**)
  EM の一部に内蔵されたプログラムを EM 上のデータ領域を使って実行する RISC プロセッサエンジン。メモリ管理機構や浮動小数点演算機構はない (要するに RISC コア)。 EM に高速 SRAM を採用するため、 PME 用のキャッシュメモリが別途必要となることはない。機能的には、 MAC のコントロール、到着保証および順序保証プロトコル、 EM 内の通信用バッファ管理、 TLB 用ページ管理、 IP プロトコル、 TCP プロトコルを実現する。操作対象側のメモリベース通信機能は PME が ADMA を適宜制御することによって実現される。 CSU および POU は PME の高速化機構として使用される。
- パケット生成エンジン (**PGE**)
  メインメモリ内の送信用データ構造体をトラバースしてパケットを EM 内に生成する。送信時のポインタによるデレファレンス段数が通常の Ethernet 用の構造体よりも深くなるため、プログラマブルエンジンによる実現と専用 hardwired ロジックによる実現の両者が考えられる。通常の Ethernet 用の NIC が提供するメインメモリ内の受信用データ構造体のトラバースにも使用される。送受信共にサイズの大きなデータ転送は PGE が ADMA を起動して実現される。
- Serializer Deserializer Unit (SERDES)
  パラレル入出力をシリアル入出力に変換する。
- Gigabit Ethernet Transceiver (TRANS)
  MBP2 の初期バージョンにはマルチモードファイバー用の 850nm レーザー送受信モジュールが使用される。
- PE 割り込み制御ユニット (**ICU**)
  I/O バスを通して PE に対して割り込みを発生させる。 PME および PGE が使用する。
- I/O バスインタフェースユニット (**BIU**)
  汎用 I/O バスと MBP2 を接続するためのインタフェース。 ADMA および PGE がメインメモリをアクセスするため、バスマスタ機能が必要である。また、 EM は PE のメインメモリ空間内にマップすることが可能であり、 BIU 経由で PE は EM にアクセスできる。
- チェックサム生成ユニット (CSU) および
  パケット順序チェックユニット (OCU)
  CSU は TCP のためのデータチェックサムを計算するユニット。 OCU は MBCF/Ethernet の到着保証および順序保証プロトコルを高速化するためのユニット。共に、実際には送信用と受信用の二組が存在する。パケット受信時は LDMA が MAC から EM へパケット転送中に計算され、パケット送信時は PGE/ADMA がメインメモリから EM へデータ転送中に計算される。 PME によって参照される。
- ユーザレベルインタフェースユニット (**ULI**)
  ユーザレベルインタフェースは PGE を起動するのに使用される。複数の PE を持つシステムでは PE ごとにエントリポイントがあり、 PE の MMU でユーザプログラムは一つのエントリポイントしかアクセスできないように保護されている。
- カーネルレベルインタフェース (**KLI**)
  カーネルプログラムだけが EM に直接アクセスできるように PE のページをマップして、 PME とカーネルは EM を介してインタフェースする。ただし、カーネル側から PME に割り込み実行を要求できるインタフェースも用意する。

4.4 通信バッファリング方式

CRC チェックや TCP/IP のデータチェックサムをハードウェアサポートするために、市販の高性能 MAC チップは store&forward（SF）方式を採用している。しかし、低レイテンシを実現するためには cut through（CT）方式の方が望ましい。TCP/IP のようにパケットのヘッダ部に後続データに関するチェックサムフィールドを設けなければ、少なくとも送信時にはメモリベース通信に対して CT 方式が適用可能である。MBP2 の初期バージョンでは市販の MAC チップを使うため CT 方式を適用可能かどうかは MAC チップの機能次第である。しかし、将来的にはメモリベース通信用パケット送信は CT 方式で行えるバージョンの MBP2 を開発する予定である。

汎用 I/O バス経由の DMA によるデータ転送の最悪条件を考慮すると、CT 方式ではバスからのデータ供給が間に合わなく可能性があるシステムが多い。それにも関わらず、CT 方式の MBP2 では楽観的にパケットデータの DMA によるデータフェッチ終了前にネットワークへの送信を開始する予定である。DMA が間に合わない場合はパケットの CRC を故意に不正な値にする機能を MAC 周辺に実装してパケットを棄却する。

4.5 到着保証の必要性と再送用コピー

Ethernet 方式を MBP2 のネットワークに採用した場合でも、full duplex かつフロー制御 (IEEE802.3x) を行えば、通信媒体上でのパケット消失の危険性をなくすことができる。しかし、前述のように、上位レベルの到着保証プロトコルはデッドロック回避の役目を兼ねており、他のデッドロック回避メカニズムを採用しない限り到着保証プロトコルを廃止できない。また、Gigabit Ethernet 以外ではパケット消失の可能性がある CSMA/CD 方式が大量に使われており、将来的に CSMA/CD 方式自体が衰退するか否かは不明である。さらに、MBCF のソフトウェア実装経験から判断して、到着保証プロトコルの平常時の時間コストは送信側のコピー作成時間を除けば大きくない。送信側コピーは PE のメモリから MBP2 への DMA 転送中に並行して MBP2 内メモリに書き出してコピーを作れば、余分な時間コストは発生しない。このため、MBP2 では MBCF/Ethernet と同じ到着保証および順序保証プロトコルを採用する。

4.6 ユーザインタフェース

メモリベース通信の結果ならびにステータス等はユーザが要求時に指定したユーザ空間内のメモリに反映されるため、ユーザインタフェースとして規定すべきものはメモリベース通信を要求するインタフェースのみである。一つのユーザ空間（タスク）ごとにメモリベース通信要求用のリングバッファが MBP2 に登録されている。送信要求したタスクを認識するために、コンテクスト切り替え時にカーネルがタスク ID（コンテクスト ID）を MBP2 のカレントタスク ID レジスタにセットする。マルチプロセッサへの対応はユーザが送信を起動するエントリアドレス（ULI）とカレントタスク ID レジスタをプロセッサ台数分用意する。このタスク ID を使って MBP2(PME および PGE) はリングバッファの情報やタスクに関する各種情報を獲得できる。

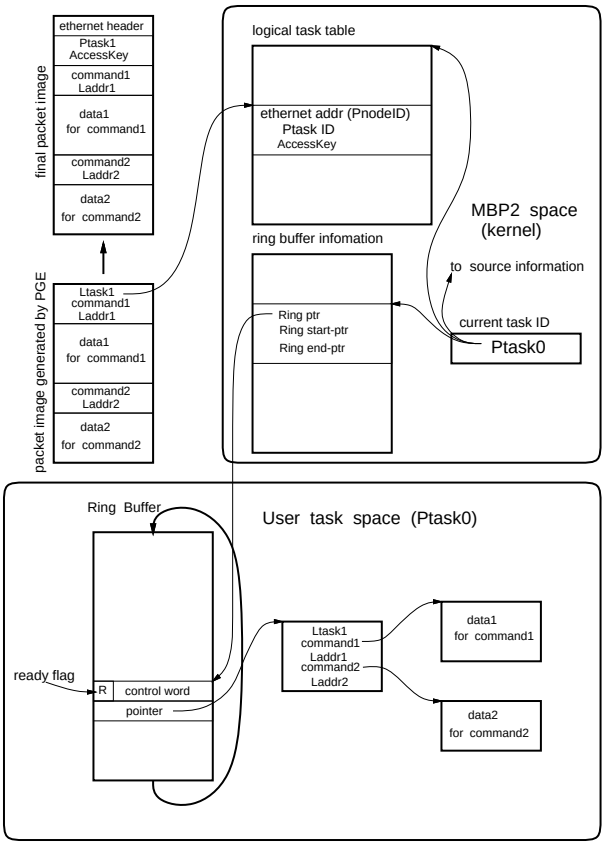

図2 送信要求ユーザインタフェースの構造

MBP2(PGE) は ULI を通して送信要求を受ける度に、送信要求したタスクのリングバッファをスキャンする（図 2参照）。PGE はセットされた ready フラグのあるエントリをリングバッファから順次読み出し、そのエントリのポインタが指すメモリベース通信要求構造体の中味を参照してパケット生成処理を行う。通信要求構造体はパケットに添付すべきデータ領域へのポインタを含む可能性があり、PGE は構造体をトラバースしながらパケットを生成する。PGE が生成したパケットは PME に渡され、論理タスク番号は Ethernet アドレス（物理ノード番号）、物理タスク番号といったグローバルな値に変換され、到着保証プロトコルのためのフィールドや要求タスクに関するフィールドが埋められて、パケットとして完成され、MAC を介して実際に送信される。PGE はパケット生成処理を終了したら、フラグをクリアしリングバッファへのポインタを次のエントリに移動させる。次のパケット生成が可能になったら、リングバッファを再び参照に行き、ready フラグセがットされたエントリが連続して存在し続ける限りパケット生成を続ける。クリアされた ready フラグを持ったエントリを読み出した場合は、次回のリングバッファのスキャン

をそのエントリから始める。

MBP2 では、パケット生成時にメモリ上のデータ領域から MBP2 への DMA 転送が PE の処理と並行して行われる。そこで、PE 自体がデータ転送してパケット作成を行っていた MBCF/Ethernet とは異なり、パケット作成において参照されるユーザデータ領域をパケット生成完了まで変更せずに保存する必要がある。この保存期間は要求送信用のリングバッファの ready フラグ (クリアされるまで領域を保存) によって、ユーザレベルで検知することができる。

5. 汎用 I/O バスによる制約事項

汎用 I/O バス上のデバイスとして実装される MBP2 はいくつかの制約を受ける可能性がある。ここではそれらの制約と解決方法について述べる。

5.1 ユーザタスクへのアップコール

Memory-Based Signal (MB-Signal) は操作対象タスク内のユーザルーチンの非同期呼び出しを伴うため、PE への割り込みが不可避である。しかし、パケットが運んだユーザデータのメモリへの転送を、MBP2 に行わせることにより PE の処理コストを削減することができる。つまり、MBP2 は PME によるパケットの順序保証やアクセスキーのチェック後、メモリベース通信のヘッダ部のみカーネルメモリに転送して PE に割り込みを掛け、パケット内のデータ部は PE の割り込みルーチンの指示に従って ADMA が直接転送する。

5.2 MBP2 と PE の排他制御

Memory-Based FIFO (MB-FIFO) は MB-Signal 同様にメモリベース通信の宛先となっている管理構造体とは異なる場所にデータバッファが存在している。MB-Signal とは異なり PE 上のタスクに対するアップコールの必要はないため、PME が構造体内のポインタを手繰って処理を完了させることが可能である。しかし、PME と PE の MB-FIFO 管理構造体へのアクセス競合が問題となる。I/O バス上のプロセッサである PME とメインプロセッサ (PE) の間には低コストの排他制御手段が通常提供されていない。このため、MB-Signal と同様に MB-FIFO の管理構造体の操作は割り込みによって PE が行い、データ転送を MBP2 の ADMA 機構が行う。なお、同様の理由から、メモリベース不可分操作も PE への割り込みによって不可分性を保証し、実際のデータ転送は MBP2 の ADMA によって実現される。ただし、データサイズが小さい場合は PE がデータ転送も行う。

5.3 MBP2 と PE のページ管理の整合性

PE (のあるタスク) と MBP2 (ADMA または PGE) の両者から同一論理アドレスでアクセス可能になっているメインメモリ領域のみが MBP2 による直接操作対象領域である。MBP2 の TLB ミス時に PME が PE のページテーブルを直接参照するのであれば、ページが無効化された場合に MBP2 の TLB の該当エントリをフラッシュする必要がある。高速化のために PME が EM 内に設けた MBP2 専用ページテーブルを参照するのであれば、MBP2 がアクセスする可能性のある PE ページの割り当ての変更に応じて、MBP2 の TLB フラッシュだけではなく MBP2 専用ページテーブルも適宜更新する必要がある。

5.4 MBP2 からアクセス可能なメインメモリの範囲

汎用 I/O バス上からメインメモリの全領域をアクセスできるように設定できないマシンが存在する☆。この制約のあるマシンに対しては、MBP2 から直接アクセスできないメモリ操作用に PE による割り込みでメモリベース機能を実現するコマンドを提供する。

メモリベース通信の要求用ユーザインタフェースに関しても、MBP2 がバスマスタになってメインメモリをアクセスするため、マシンによっては同様の問題が生じる可能性がある。MBP2 がアクセス不可能な範囲にあるデータをパケットに添付する必要がある時は、ユーザレベルでアクセス可能な領域にコピーして、コピー領域へのポインタを送信用構造体にセットする。

6. おわりに

ソフトウェアメモリベース通信のメインプロセッサへのオーバヘッドを大幅に削減する Memory-Based Processor II (MBP2) を提案した。通信用 commodity を目指すために、MBP2 は汎用ネットワークインタフェースの形態で実装される。

謝　辞

本研究の一部は新情報処理開発機構 (RWC) にサポートされている。

参考文献

1) 松本 尚, 平木 敬: 超並列計算機上の共有メモリアーキテクチャ. 信技報, CPSY 92-26, pp.47-55 (August 1992).
2) S. C. Woo, et al.: The SPLASH-2 Programs: Characterization and Methodological Considerations. In *Proc. of the 22nd ISCA*, pages 24–36, June 1995.
3) J.Niwa, et al.: Efficient Implementation of Software Release Consistency on Asymmetric Distributed Shared Memory. In *Proc. of the 1997 ISPAN*, pages 198–201, (December 1997).
4) T. Matsumoto, et al.: Compiler-Assisted Distributed Shared Memory Schemes Using Memory-Based Communication Facilities. In *Proc. of the 1998 PDPTA*, (July 1998).
5) 松本 他: メモリベース通信による非対称分散共有メモリ. 情処コンピュータシステムシンポジウム論文集, pp.37–44 (November 1996).
6) T. Matsumoto and K. Hiraki. MBCF:A Protected and Virtualized High-Speed User-Level Memory-Based Communication Facility. In *Proc. of the 1998 ICS*, (July 1998).

☆ ただし、高性能 WS や PC では I/O バスのブリッジ IC のレジスタを I/O バス側から動的に書き換えて、全メモリ領域へのアクセスを可能にしている物も多い。