メモリベース概念に基づく次世代ネットワーク構築方式の研究開発

Research and development of the next generation network architecture based on a memory-based communication model

松本　尚 [1)]　　平木 敬 [2)]
Takashi MATSUMOTO　　Kei HIRAKI

[1)] 東京大学 大学院理学系研究科 情報科学専攻 (〒113-0033 文京区本郷 7–3–1
E-mail:**tm@is.s.u-tokyo.ac.jp**)

[2)] 東京大学 大学院理学系研究科 情報科学専攻 (〒113-0033 文京区本郷 7–3–1
E-mail:**hiraki@is.s.u-tokyo.ac.jp**)

ABSTRACT. Recently network bandwidth that is available at home or office increases rapidly. However, processing overheads of communication are too large to exploit the huge bandwidth efficiently. In this paper, we propose an innovative network-interface architecture "Memory-Based Processor II (MBP2)" that enables application programs to execute send/receive without system-calls under proper protection and virtualization. The MBP2 adopts memory-based architecture where direct communication between application spaces is performed without kernel services. We also introduce a built-in microprocessor "Casablanca" that can cooperate efficiently with DMA and Interrupt mechanisms.

1 背　景

近年、通信技術が発展して家庭や企業において使用可能なネットワークの転送能力が大幅に増大している。LAN として 100Mbit/sec のイーサネットは今やコモディティとして非常に安価に入手可能であり、ギガビットイーサネットも普及の兆しを見せている。公衆データ回線網も IT 革命の推進のために、各家庭へのファイバー敷設や高速無線通信によって、大きく能力が増強されつつある。一方、データ通信の利用形態も多様化して、文字情報だけではなく音声や動画情報が使用され始めている。しかし、現在広く使用されているインターネット上の通信方式（TCP/IP プロトコル）の実装は処理オーバヘッドが大きいため、将来より大きなバンド幅のネットワークが利用可能になっても、このオーバヘッドによりユーザがバンド幅を活用できない可能性がある。現時点（2000 年 5 月）においてもサーバマシン上で ATM や Ethernet のギガビットクラスの通信網を TCP/IP プロトコルでスループット最大まで使用するためには、最新鋭マイクロプロセッサを複数搭載した SMP 構成の計算機が通信処理のためだけに必要とされることが報告されている。また、複数台のマシンを集めてクラスタコンピューティングによってシステムの性能向上を図る場合に、このオーバヘッドが処理全体のボトルネックとなり性能向上効果が得られない。さらに、この通信オーバヘッドの問題は、セキュリティを保つための暗号プロトコルを通信に使用する場合には、暗号化／復号化の処理コストが非常に大きいため、より深刻な問題になる。

本研究開発では、メモリベース通信の概念を導入してより処理コストの低い通信ネットワークインタフェースアーキテクチャを設計し、このアーキテクチャに基づくネットワークカードのプロトタイプおよび通信用システム LSI を開発する。

2 目　的

背景で述べた通信の処理オーバヘッドを大幅に削減する新しいネットワークインタフェースアーキテクチャを研究開発するのが本研究開発の目的である。提案するネットワークインタフェースは情報家電からサーバマシンまで適用可能な低コストかつ高性能な方式であり、究極の目的は、汎用品（コモディティ）として広く製品に採用され、ネットワークインタフェースのデファクトスタンダードとなることである。汎用品を目指すために、通信プロトコルとして広く使われている TCP/IP および UDP/IP をサポートする。暗号プロトコルに関しては現在広まりつつあり、ネットワークインタフェースにおいて独立に処理可能な IPSec をサポートする。そして、クラスタ計算用およびソフトウェア分散共有メモリ実現のための通信方式として、*SSS–CORE* [1] [2]で性能と汎用性が実証されている MBCF [3] [4]を実装する。

技術的には低コストで処理オーバヘッドの極めて少ないネットワークインタフェースの方式を確立することが、本研究開発の目標である。このために、本研究開発

† 本研究開発の一部は情報処理振興事業協会「次世代応用基盤技術開発事業」の一課題として行われたものである。本稿は同開発事業論文集 (2000 年 6 月) に掲載された内容を 2000 年 9 月現在の状況に一部更新したものである。

では以下に挙げる革新的な技術を採用している。

- メモリベース通信 [5]の考え方
- 保護と仮想化が可能なユーザレベル I/O アクセス方式 [6]
- メモリコピー不要の DMA による機能結合
- 低オーバヘッドのオリジナルプロトコルスタック
- DMA と高速協調可能な組み込みマイクロプロセッサ
- 割り込み反応オーバヘッドがない組み込みマイクロプロセッサ [7]
- ネットワークインタフェース内に組み込まれたハードウェア暗号処理回路

これらの項目の詳細については技術的先進性の節で述べる。

3 研究開発の概要

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)**」[8]に基づいて、アーキテクチャ実証および評価用の MBP2 プロトタイプ (MBP2P) と MBP2 プロトタイプの大部分を 1 チップ LSI 化した MBP2 チップ (MBP2C) を研究開発した。

MBP2 プロトタイプはマルチチップ構成の通信ネットワークカードであり、短期間に開発を終えるために、組み込みマイクロプロセッサおよびメディアアクセスコントローラには市販品を使用し、全体のコントローラは高速大容量のフィールドプログラマブルゲートアレイ（FPGA）である Xilinx 社の XCV1000-6FG680 を採用した。メモリデバイスには制御の容易なシンクロナス SRAM を同様に短期開発のために採用している。大容量 FPGA と言っても最先端のカスタム LSI と比べると容量が大幅に少ないため、MBP2 プロトタイプではハードウェア暗号処理回路の実装は諦め、暗号処理はカード上のマイクロプロセッサのファームウェアで実現する。MBP2 プロトタイプに関しては、*SSS–CORE* オペレーティングシステム上で開発を行うため、SBus 版と PCI バス版の二種類の実装が存在する。*SSS–CORE* オペレーティングシステムを開発用オペレーティングシステムに採用したのは、*SSS–CORE* は MBP2 の考案者である松本が設計開発したオペレーティングシステムであり、我々にとってデバイスドライバの開発コストが圧倒的に低いからである。

MBP2 チップは MBP2 プロトタイプの構成要素のうちカード内蔵メモリと 1GHz 以上の動作周波数が要求される部分を除いた全要素と暗号処理ハードウェアを 1 チップに集積した通信用システム LSI である。LSI を新規開発するため、組み込みマイクロプロセッサを新規開発する。MBP2 プロトタイプの市販組み込み用マイクロプロセッサと基本的に同じ命令体系を使用するが、拡張命令やキャッシュ方式や割り込み処理等は新規機能を盛り込み MBP2 上の処理を大幅に高速化する。メディアアクセスコントローラ（MAC）も新規開発することにより、市販 MAC の問題点を解決すると共に、10BASE,100BASE,1000BASE の 3 スピードを切替可能な通信用 LSI:MBP2C を作成する。低コスト通信インタフェース実現のために、MBP2 チップは外部接続されるメモリデバイスにビット当たり単価が圧倒的に安いシンクロナス DRAM を採用している。MBP2 チップは多くのシステムが採用している PCI バスに直結できる拡張バスインタフェースを内蔵している。MBP2C は 0.25μm テクノロジーのゲートアレイで開発される。

4 参加企業および機関

本研究開発には以下の 5 つの企業および機関が参加している。

- 東京大学大学院理学系研究科
- 日本アイ・ビー・エム株式会社 東京基礎研究所 インフォメーション・テクノロジー・ソリューション株式会社
- 三精システム株式会社
- 株式会社アックス
- 株式会社アーツテック

なお、インフォメーション・テクノロジー・ソリューション株式会社は日本アイ・ビー・エム株式会社の関連会社である。各参加機関のプロジェクト内における役割分担を表 1に示す。なお、MBP2P は本原稿作成時点で完動テスト済みであるが、MBP2C に関してはチップ製造が完了しておらず、論理設計完了の段階である。

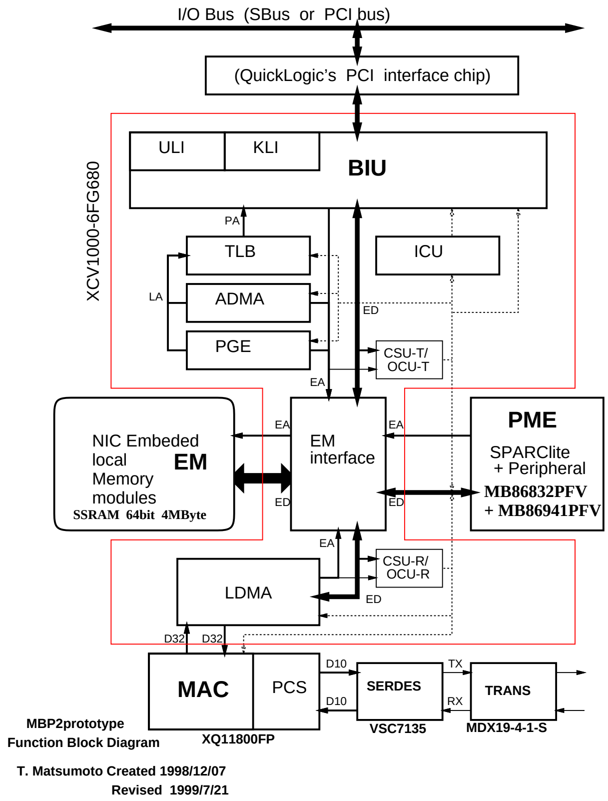

図 1　**MBP2** プロトタイプの構成

表1　MBP2 研究開発プロジェクトの担当

研究開発項目	担当企業／機関	担当者
アーキテクチャ設計	東京大学	松本尚
MBP2 詳細機能設計	東京大学	松本尚
組み込みプロセッサ詳細機能設計	東京大学	田中清史
MBP2P 論理回路設計	三精システム	及部晴康
MBP2P デバッグ＆テスト	東京大学 三精システム	松本尚 及部晴康
MBP2P 基板作製	三精システム	及部晴康、他
MBP2P 用部品調達	東京大学 三精システム	松本尚 及部晴康、他
MBP2C 論理回路設計 （全体）	日本 IBM/ITS	名村健、他
MBP2C 論理回路設計 （MAC）	日本 IBM	小林芳直
MBP2C 論理回路設計 （暗号処理）	日本 IBM	宗藤誠治
MBP2C 論理回路設計 （組み込みプロセッサ）	東京大学	田中清史
MBP2 ファームウェア開発	東京大学	松本尚
TCP/IP プロトコルスタック開発	アックス	竹岡尚三
IPSec エミュレータ開発	アックス	青笹茂
IPSec 関連技術調査	アーツテック	信国陽二郎
MBP2 用デバイスドライバ開発	東京大学	松本尚
プロジェクトアドミニストレーション	日本 IBM 東京大学 アーツテック	清水茂則、大矢幸雄、新島秀人 平木敬 信国陽二郎
MBP2P ＆ソフトウェア開発進行	東京大学	松本尚

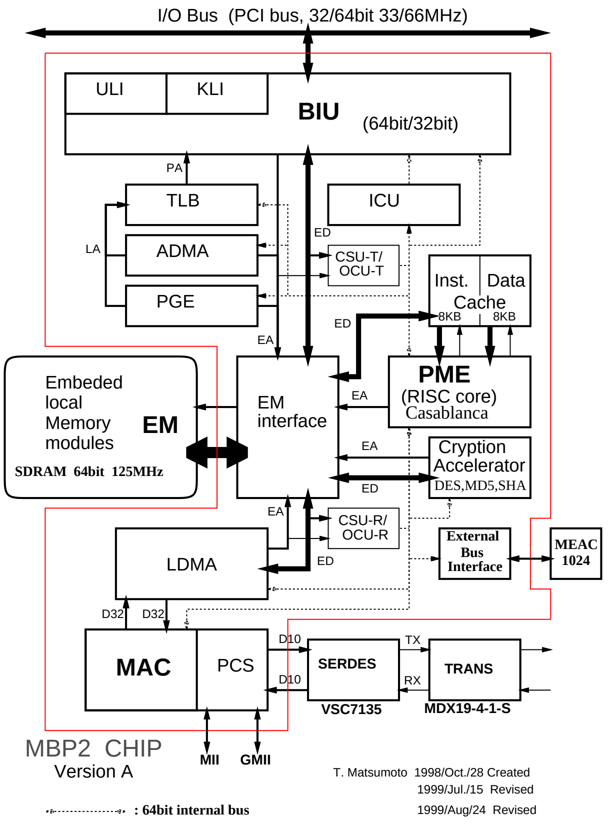

図2　MBP2 チップの構成

5　MBP2 のシステム構成

MBP2 アーキテクチャの説明をする前に各ブロックの紹介を兼ねてハードウェアのシステム構成を簡単に示す。図1に MBP2 プロトタイプの構成を図2には MBP2 チップの構成を示す。

- カード内蔵高速メモリ (**EM**)
  MBP2 ネットワークカード内蔵の高速メモリで、I/O バス、MAC、プロトコル管理エンジン (PME) および Cryption Accelerator（MBP2 チップの場合）の三（四）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（～四）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（64bit 幅以上）を採用する。メモリチップとして MBP2 プロトタイプでは設計期間の短縮のため高速 Synchronous SRAM（SSRAM）を採用し、MBP2 チップでは外づけ部品のコストを抑えるため Synchronous DRAM（SDRAM）を採用する。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。
- 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 型組み込みマイクロプロセッサ。メモリ管理機構や浮動小数点演算機構は持たない（要するに RISC コア）。MBP2 では SPARC 系の RISC コアを採用する。この RISC コア上のファームウェアによって、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 プロトコル等が実現される。メモリベース通信機能の対象ノード側での操作は 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 OCU は PME の高速化機構

として使用される。
MBP2 プロトタイプには PME として既存の富士通社製 SPARClite(MB86832) を使用し、超高速割り込み処理の実現と権利関係の完全なクリアのためにMBP2 チップには新規開発された SPARC V8 ユーザ命令準拠[1]の Casablanca（開発コード名）[7]を使用する。
MB86832 も Casablanca も命令とデータ各 8Kbyte / 2way のキャッシュを持っており、キャッシュラインサイズは 32byte である。 Casablanca は明示的なキャッシュラインの外部取り込み命令とダーティなキャッシュラインの書き戻し命令を用意することにより、データキャッシュを使用したまま EM 上で他の DMA ブロックと通信同期が可能である。
little endian のホストプロセッサによって書かれたパケットのヘッダ情報を効率良く扱うために、SPARC V9 仕様に見られる little endian 用の load /store ハードウェア機構を Casablanca に追加する。

- Local Direct Memory Access Unit (**LDMA**)
  EM と MAC の間においてパケットの DMA 転送を行う。各方向 1 本の計 2 本の DMA チャネルを持つ。 MAC とのハンドシェークを担う。
- Advanced Direct Memory Access Unit (**ADMA**)
  MBP2 の EM と通信カードが装着されたホストコンピュータのメインメモリとの間のデータ転送を行う 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。 MBP2C の ADMA は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。
- 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。 MBP2 の内部ブロックで MAC と並ぶ複雑な制御論理を必要とし、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。今回の MBP2 プロトタイプおよび MBP2 チップでは hardwired ロジックによって実装する。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および EM からメインメモリへのデータ転送にも使用される。
  MBP2C の ADMA と同様に、 MBP2C の PGE は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。実装上は、PGE はメインメモリ内構造体のトラバース機能付き ADMA ユニットであり、データパスは ADMA と共用される。
- 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 ミス時は、 PME が EM 内の MBP2 用のページテーブルからエントリを取り出して更新する。 ADMA および PGE には TLB をバイパスしてメインメモリをアクセスする手段も提供される。ただし、このバイパス機能はユーザレベルでは使用されないように保護されている。
- 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 および TCP 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。
- I/O バスインタフェースユニット (**BIU**)
  拡張 I/O バス（PCI バス,SBus）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 はホストコンピュータ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 経由でホストプロセッサは EM にアクセスできる。
- Cryption Accelerator Unit (**CAU**)
  通信速度（Gigabit）に準ずる速度で暗号化通信をサポートするための機能ブロックである。 EM 上のデータ領域に対して DES の encode/decode、または各種ハッシュ関数をユニット内の DMA 機能によって適用する。適用結果は EM 上の領域に DMA で格納される。 1 チップ版の MBP2C にのみ実装される。
- MEAC1024
  主に鍵交換アルゴリズムおよび公開鍵暗号アルゴリズムの高速実行の目的のため、剰余乗算や剰余冪乗算を高速に行う IBM 社製の MEAC1024 を MBP2 は外付けする。
- PE 割り込み制御ユニット (**ICU**)
  I/O バスを通してホストのメインプロセッサと MBP2 の PME 間で相互に割り込みを発生させるためのユニット。最大 64bit 幅のメッセージを伴うことができる。
- 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ホストプロセッサ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。
- ユーザレベルインタフェースユニット (**ULI**)

1 レジスタウィンドウがなく、割り込み用レジスタセットが複数用意されている点が V8 とは異なり、割り込みによるコンテクスト切替が超高速である。

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。ユーザレベルによる直接アクセスにも関わらず、アクセスしたプロセスが同定できる。科学技術振興事業団から特許出願された松本尚の発明 [6]に基づいている。

- 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。
- Serializer Deserializer Unit (SERDES)
  PCS からの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 プロトタイプ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。

6 技術的先進性

本節では MBP2 の技術的特長に関して説明する。

6.1 メモリベース通信に基づくアーキテクチャ

従来の通信は送信と受信のペアによって行われるメッセージパッシング型が多かった。通信相手のアドレスを指定して遠隔読み書きを行うものも存在したが、それらは保護や仮想化の機能を伴わないものであったため、汎用システムや複数プロセスが通信機能を同時使用する複雑なシステムでは使用できなかった。アドレス変換機能を通信ハードウェアに導入することにより保護と仮想化を伴ったユーザプロセス間の直接遠隔メモリアクセスが実現でき、これがアプリケーションプロセスから見れば理想的なオーバヘッドのない通信システムであることを 1992 年に松本が示した [5]。 1992 年に提案した初代 Memory-Based Processor (MBP) はプロセッサのローカルバスに接続され、細粒度通信を行う機構であった。しかし、その後の研究で細粒度通信が必要不可欠と思われた分散共有メモリを実現するための通信も、最適化コンパイラがあれば多少粒度が大きくても性能に影響がでないことが判明した [9] [10]。逆に、細粒度通信では通信のためのヘッダ等のオーバヘッドが大きく、ネットワーク上での効率が悪い。また、コンピュータシステム内でもバースト転送の恩恵に浴することが難しい。通信粒度を見直して汎用品として使用可能なことを考慮して考え直されたアーキテクチャが、 MBP2 である [8]。具体的には、 MBP は専用ネットワークを必要とするが、 MBP2 は 1000BASE-SX 等のイーサネットをネットワークとして仮定しており、ネットワークは通常のイーサネット用スイッチやハブが使用できる。 MBP は特定プロセッサのローカルバスに合わせて設計されるが、 MBP2 は拡張 I/O バスに挿入されるため適用可能な範囲が広い。

もう一つ大きな MBP と MBP2 の違いに以下のものがある。 MBP は既存通信プロトコルや既存暗号通信プロトコルをサポートしないが、 MBP2 は TCP/IP や IPSec 等の既存通信および暗号方式をサポートし、 MBCF（メモリベース通信プロトコル）以外のプロトコルの実現においてもメモリベース通信の考え方が適用されて性能を向上させている。メッセージパッシング型の従来通信方式の実装では、プロセス間保護や同期のためにカーネル空間内に通信パケットのコピーを送信側、受信側において作る必要があった。よって、アプリケーションプログラムは送信や受信の度にシステムコールを行って、データをユーザ空間からカーネル空間にコピーする必要があった。システムコールによってカーネル空間に切替えるオーバヘッドばかりではなく、このコピーはホストプロセッサがアドレス保護下で行う必要があるため、通信処理オーバヘッドが増大する。これに対して MBP2 では、ユーザ空間内のデータを PGE の DMA でパケットとしてまとめて送信し、受信時は PGE の DMA で直接ユーザ空間内のバッファに転送することができる。このため、送信時はシステムコールもコピーも省略でき、受信時はコピーは必要であるがシステムコールなしにユーザ空間内だけで処理ができる。

6.2 保護と仮想化が可能なユーザレベル I/O アクセス方式

MBP2 では、保護や仮想化のレベルを落すことなしに、ユーザアプリケーションがシステムコールを行わずに通信が可能である。受信側はメモリベース通信同様に直接ユーザアドレス空間内の受信バッファでパケットを受け取ることによりシステムコールの必要性はなくなる。送信側に関しては、送信パケットのためのパラメータを揃えたことを NIC に知らせないと送信ができない。送信を行うプロセスがシステムに一つである場合はこの通知を行うレジスタを送信を行うプロセス空間内にマップすればよい。しかし、複数プロセスが送信を行う汎用システムでは、どのプロセスが送信要求を行っているか識別できる必要がある。悪意のあるプロセスが虚偽の送信要求通知を行って送信能力を下げようとしていることまで防止しようとすると、送信要求通知自身をどのプロセスが行ったか識別できる必要がある。この機能を低コストで可能にした機構が MBP2 の ULI である。送信要求通知用のレジスタの実体は一個であるが、物理アドレスのデコードを一部省略することで、このレジスタはページ単位のエイリアスを複数持っている。そして、このレジスタがアクセスされた際にはどのエイリアスからアクセスされたかも記録される構造になっている。この機構を利用して、送信を行うプロセス毎に異なったエイリアスアドレスを割り当てれば、アクセスしたプロセスが識別可能となる。他のプロセス用のエイリアスは自分の空間にマップされていないのでアクセスできないので、他プロセスへの「なりすまし」はできない。無駄な要求を多発するプロセスを認識することも可能なため、悪意を持ったプロセスの優先度を下げたり、極端な場合には実行を停止させることも可能である。このメモリエイリアスを利用した保護されたユーザレベル I/O アクセス方式は科学技術振興事業団から特許出願されている。

6.3 低オーバヘッドのオリジナルプロトコルスタック

従来一般に使用されてきた TCP/IP のプロトコルスタックは通信プロトコルごとに通信パケットのコピーを行う性能よりもプログラムの書き易さを重視したものであった。最近は極力コピーを行わないプロトコルスタックが Linux 等において開発されているが、ハードウェアの制約により送信時 1 回、受信時 2 回のコピーが発生していた。これに対して MBP2C では、DMA エンジンと MAC を最適化することにより TCP/IP においてもマイクロプロセッサ（PME またはホストコンピュータ）によるデータコピーを完全に排除可能なアーキテクチャになっている。このアーキテクチャを活かすために、PME が実行するプロトコルスタックを新規に開発する。ノーコピー版の MBCF 用プロトコルスタックは完成しており、PME はヘッダ生成、順序保証、到着保証、ヘッダ解釈の処理のみを実行し、データコピーは一切行わない。

6.4 メモリコピー不要の **DMA** による機能結合

MBP2P は MBP2C のサブセットであるため、本小節では MBP2C を例に取って説明する。送信時には、まず PGE がホストコンピュータのメインメモリからデータを DMA によって EM に転送し、送信パケットの原型を作る。次に、PME がヘッダ部の情報と EM に予め保持しているユーザプロセスの情報から送信に使うヘッダを作成する。この時暗号化が必要かどうかもチェックされ、暗号化が必要であれば、CAU を起動して DMA で暗号化されたパケットを EM 上に作成する。最終的にネットワークに出力するパケットイメージが完成したら LDMA を起動して MAC にパケットを送信し、外部に発信する。

受信時は、送信時の逆の経路をたどり、MAC が受信したパケットは LDMA により EM 上のバッファに受け取られ、PME がヘッダ部を解析する。PME が暗号の復号作業が必要と判断すると CAU を起動して、復号化された通信データを獲得する。受信パケットが MBCF プロトコルのパケットの場合は、ヘッダに指定されたタスク（プロセス）の指定された論理アドレスに対して指定された操作（遠隔書き込み、遠隔読み出し等）を ADMA によって施す。TCP/IP や UDP/IP のパケットの場合は、コネクションや受信ポートが指定する空間のバッファ領域に PGE を起動して DMA 転送する。

このように MBP2 では機能ブロックが EM を介したメモリ結合になっており、データに対する処理や移動が必要な機能ブロックにはすべて DMA 能力がある。このため、PME もホストコンピュータもデータコピーやデータ全体をスキャンするような処理を行う必要がない。

6.5 **DMA** と高速協調可能な組み込みマイクロプロセッサ

PGE, ADMA, LDMA, CAU の各 DMA 機能は DMA 要求を EM 上にリングバッファであるディスクリプタリングに複数登録してバッファリングできるようになっている。DMA の終了や失敗といった情報はディスクリプタリング上の該当エントリが上書きされることで PME に通知される。また、送受信パケットのヘッダ部はヘッダを生成・解釈するために PME が読み出す必要がある。つまり、ディスクリプタリングやパケットのバッファ領域は DMA と PME の共有メモリとなっており、両者が更新する。しかし、組み込み用の安価なマイクロプロセッサにはデータキャッシュの外部スヌープ機能がない。このため、これらの領域の更新を PME が間違いなく認識するためにはノンキャッシャブル領域として指定する必要がある。現実に、MBP2P の SPARClite では、これらの領域をノンキャッシャブルにすることで、データの無矛盾性を保持している。けれども、ノンキャッシャブルに指定することは EM アクセスコストを大幅に上昇させる。暗号化を伴わない TCP/IP パケットでもヘッダは 54 バイト以上の大きさがある。これをノンキャッシャブルで 4byte ずつ読み出すと、PME は 14 回の EM アクセスを行うことになる。それに対して、キャッシュが使用可能であれば、32byte のキャッシュライン単位で 2 回読み込めば PME はヘッダ解析処理が可能となる。1 回の EM アクセスは MBP2C では 11 サイクル以上かかるため、この回数の少なさは非常に重要である。MBP2C の PME である Casablanca には明示的なキャッシュライン fill 命令と明示的な dirty キャッシュライン writeback 命令が存在する。これらの命令を使うとキャッシュライン単位で EM とデータのやり取りが可能であり、コンシステンシを守ったまま性能を大幅に向上できる。もちろん、これらの命令を使用するタイミングは注意深く決定されなければならないため、ファームウェアの最適化作業は仕事量が増大する。

6.6 割り込み反応オーバヘッドがない組み込みマイクロプロセッサ

各ブロックの DMA が処理を終了する度に何らかの処理を PME が行う必要がある。たとえ、パケット処理としては最後の DMA であったとしても、DMA のためのディスクリプタやバッファを回収する必要がある。この DMA の終了検知や ULI からの送信要求を割り込みベースで受け取るとすると、たとえパケットのデータ長が 1Kbyte あったとしても 1Gbit/s のフルスピード時には 8μsec に受信で 2 回、送信で 3 回、全二重なら 5 回の割り込みを処理する必要がある。暗号用の CAU の DMA を使用するとさらに処理を行う割り込みの回数が増加し（CAU は wirespeed 近くで動作する予定）、TLB ミスによる割り込みもさらに加わる。通常の組み込みマイクロプロセッサでは割り込み処理のためのコンテクスト切替に数 μsec かかってしまうため、この頻度の割り込みをとても処理できない。MBP2P の SPARClite は市販品であり、プロトタイプ用であるため、割り込み処理の高速化は不可能であるが、MBP2C の Casablanca はオリジナル CPU であり、究極の割り込み処理高速化手法を採用している。Casablanca は RISC コア内に複数コンテクスト分のレジスタセットを持っており、それぞれのセットが通常走行用、内部割り込み用、外部割

り込み用といった用に用途が決まっており、割り込みが発生するとパイプラインストールすら起こすことなしにレジスタセットが切り替わり、処理が継続される。新たな割り込み用レジスタセットにおいて、多重割り込みを禁止して処理を行う場合には、割り込みによるコンテクスト切替のオーバヘッドはゼロになる。よって、Casablanca ではポーリングよりも割り込みベースで処理を行った方が、無駄なポーリングアクセスがない分高速に処理が行える。また、ポーリングよりも優先度を考慮したプログラムが作り易い。

6.7 NIC 内に組み込まれた暗号処理回路

ここ数ヶ月の間に半導体大手メーカが通信処理用の LSI をアナウンスし、多くの物は暗号処理回路を内蔵もしくは外づけできる構造になっている。しかし、MBP2 を最初に提案した 1998 年の時点では、メインプロセッサの能力増強用の暗号処理専用カードは存在したが、ネットワークカード内に暗号処理回路を持つ物はなかった。現時点でもギッガビットネットワークのコントローラーとそのスピードに対応した暗号処理回路を持つネットワークインタフェースはアナウンスされていない。

暗号化(セキュリティ機能)によって守りたいのは外部との通信であって、コンピュータ内部のデータではない。コンピュータに違法侵入されなければコンピュータ内部のデータを暗号化する必要性は薄い。暗号化と復号化がネットワークインタフェースカード内で処理されると、そのためのデータ転送のためにホストコンピュータのメインプロセッサや I/O バスを使用する必要がなくなり、ホストコンピュータの使用効率を改善することができる。特に、ギガビットネットワークでは wirespeed で 125M x 4(全二重、読み書き) = 500Mbyte/s のデータトラフィックが暗号処理のみに必要になる。このトラフィックを PCI バス上で実現することはほとんど不可能である。

7 MBP2 プロトタイプの現状

MBP2 チップはチップ製造が完了していないため、実証試験が行えていない。本節では、機能的には当初目標通りに完成している MBP2 プロトタイプの現状について報告する。ただし、チューンナップ作業が現在も進行中であるため、スペックは変更される可能性が高い。ファームウェアとデバイスドライバを含めて一通り完成しているのは SBus 版 MBP2 プロトタイプである。PCI バス版はカードはほぼ完成しているが、ホストコンピュータ側のデバイスドライバが未完成であるため簡単なテストしかなされていない。今後、本節で述べる MBP2 プロトタイプは SBus 版のことである。

図 3に MBP2 プロトタイプ基板の外観を示す。

7.1 動作環境

MBP2 プロトタイプの動作環境は表 2の通りである。
図 4に MBP2 プロトタイプの動作環境の外観を示す。

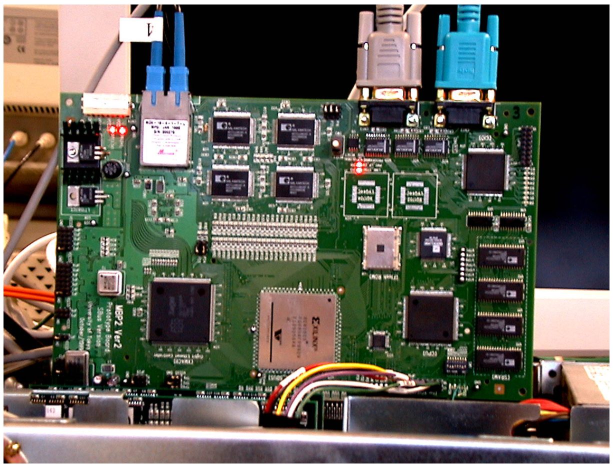

図 3 MBP2 プロトタイプ基板の外観

表 2 MBP2 プロトタイプの動作環境

ホストコンピュータ	Sun SPARCstation 20
オペレーティングシステム	*SSS–CORE* Ver.1.2-p
SPARClite モニタ端末	RS232 (9600bps)
FPGA 設計ダウンロード	Xilinx JTAG ケーブル
イーサネット入出力	1000BASE-SX (850nm)

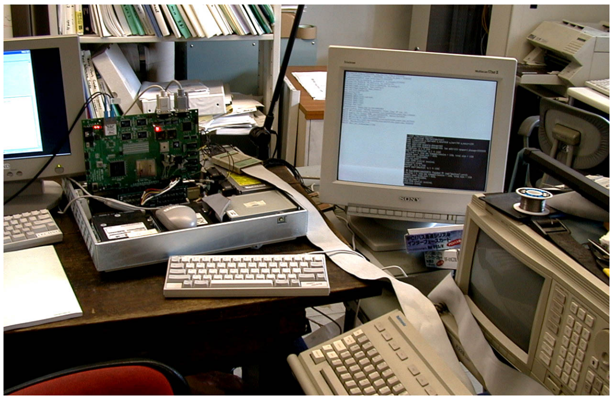

図 4 MBP2 プロトタイプの動作環境

7.2 ハードウェア基本スペック

MBP2 プロトタイプの現状(2000 年 9 月)の基本スペックを表 3に示す。

表 3 MBP2 プロトタイプの HW 基本スペック

SPARClite 動作周波数	100MHz
SPARClite 外部バス周波数	20MHz
FPGA 使用容量	約 60 万ゲート
FPGA 動作周波数	40MHz
SBus 動作周波数	25MHz
SBus 最大転送幅	4byte
SBus 最大バースト転送サイズ	32byte

7.3 ファームウェアスペック

MBP2 プロトタイプのファームウェアのスペックを表 4にしめす。 MBP2P のファームウェアはポーリングベースで書かれている。これは SPARClite では割り込みよりもポーリングの方が高速プログラムが開発可能なためである。また、 MBP2C の Casablanca と同じソフトウェアが使用できるメリットもある。 TCP/IP や UDP/IP もノーコピーのプロトコルスタックが MBP2 アーキテクチャ上は可能であるが、開発日程の関係で現状ではデータコピーを行っている。

表 4 **MBP2** プロトタイプのファームウェアスペック

MBCF プロトコル	ノーコピー版
TCP/IP プロトコル	送信: 3 コピー版 受信: 2 コピー版 ソフトウェアチェックサム
UDP/IP プロトコル	送信: 1 コピー版 受信: 2 コピー版 ノーチェックサム
IPSec ESP プロトコル	DES-CBC, SHA-1, MD5
IPSec AH プロトコル	SHA-1, MD5
ARP プロトコル	自ノードアドレスのリプライ
ICMP プロトコル	Echo のサポート
MBP2 モニタ機能	RS232 端末による
MAC モニタ機能	RS232 端末による
EM モニタ機能	RS232 端末による

7.4 デバイスドライバとライブラリ

MBP2 プロトタイプの現状のホスト側ソフトウェアの一覧を表 5に示す。表 5に示されたユーザアプリケーション用の 4 種類のインタフェースは、プロセスが通信開始時にシステムコールによって MBP2 に該当プロセスの情報を登録すると、後はシステムコールを使うことなく送受信 / 遠隔メモリアクセスが可能である。

表 5 **MBP2P** 用ホスト側システム **SW** の一覧

プログラム種類	備考
TCP/IP インタフェース	SSS–CORE 形式
UDP/IP インタフェース	SSS–CORE 形式
ソケットインタフェース	UNIX ライク
MBCF インタフェース	SSS–CORE 形式
各種パラメータ設定	特権システムコール
MAC 物理層制御	特権システムコール

8 おわりに

本テーマの公式の開始時点は 1999 年 4 月であるが、実質的には 1999 年 1 月ぐらいから研究開発作業をスタートしていた。基本的なアーキテクチャに対するアイデアはプロジェクトに応募した時点で確立していたが、2000 年 1 月の締切までに組み込みマイクロプロセッサを含む通信ハードウェアをゼロから設計製作し、その上のファームウェアとホストコンピュータのデバイスドライバを同じくゼロから開発するという仕事は非常に日程的に困難なものであった。時間的余裕があれば工夫する余地がある設計箇所や機能を十分に検討することができなかったことは残念である。単年度会計の枠にしばられない研究開発助成金の在り方が検討されることを望む。MAC と組み込みプロセッサに市販品を使用することで短期完成を目指した MBP2 プロトタイプは、逆にこれらの市販品の詳細情報が入手できないために、いくつものトラブルが発生し、非常に開発工数が大きくなった。MBP2 チップの生産はまだ済んでいないが、 MBP2 プロトタイプでは当初予定の機能を実現しており、アプリケーションプログラムがシステムコールを使うことなく各種通信がユーザレベルのみで実現されている。この MBP2 プロトタイプは MBP2 アーキテクチャの機能検証と MBP2 チップ用のテストベクタ生成に使用可能である。 MBP2 プロトタイプのファームウェアの一部機能はアーキテクチャを十分に活用できていないため、ファームウェアは徐々に最適化して高速化していく予定である。また、 PME が SPARClite から Casablanca に置き換わることで大幅な処理性能の高速化が達成できるはずであるため、 MBP2 チップの完成を急ぐ予定である。

参考文献

[1] 松本 尚, 平木 敬: 汎用並列オペレーティングシステム SSS–CORE の資源管理方式. 日本ソフトウェア科学会第 11 回大会論文集, pp.13–16 (October 1994).

[2] 松本 尚, 他: 汎用超並列オペレーティングシステムカーネル SSS-CORE. 第 17 回技術発表会論文集, 情報処理振興事業協会, pp.175-188 (October 1998).

[3] 松本 尚, 平木 敬: 汎用超並列オペレーティングシステム SSS–CORE のメモリベース通信機能. 第 53 回情報処理学会全国大会講演論文集 (1), pp.37–38 (September 1996).

[4] Matsumoto, T. and Hiraki, K.: MBCF:A Protected and Virtualized High-Speed User-Level Memory-Based Communication Facility. In *Proc. of the 1998 ACM Int. Conf. on Supercomputing*, pp.259–266 (July 1998).

[5] 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).

[6] 松本 尚: アクセス方法及びアクセス処理プログラムを記録した記録媒体. 科学技術振興事業団, 特願平 11-255272 (September 1999).

[7] 田中清史, 松本 尚, 平木敬 : Casablanca: 実時間処理 RISC コアの設計と実装. 計算機アーキテクチャ研究会報告, 情報処理学会, ARC-135, pp.51-56 (August 1999)

[8] 松本 尚, 他: 中粒度メモリベース通信を支援する Memory-Based Processor II, 計算機アーキテクチャ研究会報告, 情報処理学会.ARC-130-18, pp.103-108 (August 1998)

[9] 丹羽, 稲垣, 松本, 平木: 非対称分散共有メモリ上におけるコンパイル技法. 情報処理学会研究報告 97–HPC–67, 情報処理学会, Vol.97, No.75, pp.121–126 (August 1997).

[10] 丹羽 純平, 稲垣 達氏, 松本 尚, 平木 敬: 非対称分散共有メモリ上における最適化コンパイル技法の評価, 情報処理学会論文誌, Vol.39 No.6 pp.1729–1737 (June 1998).