計算機アーキテクチャ　152－11
（２００３．３．１０）

メモリスロット装着型ネットワークインタフェースDIMMnet-2の構想

田邊　昇[†] 濱田芳博[††] 三橋彰浩[††]
中條拓伯[††] 天野英晴[†††]

我々はDIMMスロット搭載型ネットワークインタフェースDIMMnet-1を開発した．DIMMスロットという誰も手を付けてこなかった場所にネットワークインタフェースを装着するというリスキーな開発であったために，開発を通して種々の問題点が明らかになった．本報告では，DIMMnet-1プロトタイプの試作経験により判明した問題点と，限られた設計リソースの制約から実装を見送った機能について述べ，これらを踏まえた次期バージョン開発に向けた改良方針に関して考察する．本報告はMEMOnetのネットワークインタフェースとしての発展の方向性とともに，キャッシュアーキテクチャを補完する高機能メモリとしての新機軸を提案している．

Concept of DIMMnet-2 network interface plugged into a DIMM slot

Noboru Tanabe,[†] Yoshihiro Hamada,[††] Akihiro Mitsuhashi,[††]
Hironori Nakajo[††] and Hideharu Amano [†††]

DIMMnet-1 prototype network interface plugged into a DIMM slot is developed. Because of risky development to plug a NIC into an unexplored place such as a DIMM slot, many problems become obvious. In this report, problems appearing on the prototyping are shown. In addition, functions passed up implementations due to lack of design resources are presented. We considered the concept how to enhance DIMMnet. We propose not only the direction for enhanced MEMOnet as a network interface, but also the new innovation for enhanced MEMOnet as a memory system to supplement cache memory architecture.

1. はじめに

パーソナルコンピュータ (PC) の価格性能比のめざましい進歩を背景に，PCクラスタが注目されている．PCを構成するCPUの処理性能や主記憶の容量およびバンド幅はMooreの法則に則って進歩をしている．

これに対し，コモディティPCにおけるI/Oバスやネットワークインタフェース (NIC) に対するニーズや進歩はMooreの法則からは程遠く，PCIバス[1)]が既に11年間もの間，デファクトスタンダードの地位を継続してきた．

NICについても一般ユーザーの当面の利用には100base-Tでほとんど十分であるため，コモディティPCにおけるPCIバス駆逐の原動力にはなっていない．

Intel はInfiniband[2)]の製造から撤退しており，コモディティPCにおけるポストPCIはPCI Express[1)]となることが予想されている．ただし，この領域のPCにおけるPCI Expressのニーズは前述のとおり高いものではないため，2004年頃に2.5Gbpsからスタートし，図1に示すようにMooreの法則からはかけ離れたスケジュールで徐々に進歩することが予想される．

このため，4X(10Gbps) 以上のInfinibandを用いることが好ましい領域のハイパフォーマンスコンピューティング用途にはPCI-X[1)]スロットを有する量産効果が薄く高価なサーバー型PCを用いざるを得ない．12X(30Gbps) のNICには次世代のPCI-X2.0[1)]スロットが必要で，さらに高価なサーバーが必要になってくる．

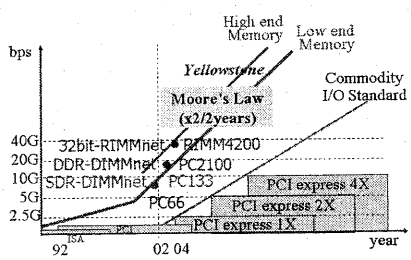

図1　標準I/OとMEMOnetのロードマップ
Fig. 1　Road map of I/O standard and MEMOnet

これに対し，筆者らはメモリスロットに装着されるNICをMEMOnet[3)]と名付け，DIMMスロットに装着されるNICのプロトタイプであるDIMMnet-1を試作してきた．MEMOnetはコモディティPCにおけるMoore則にそった進歩がなされるメモリスロットに装着されるために，低コストで高性能なPCの他の部位とのバランスの取れた性能向上を図ることが可能である．

本報告では，DIMMnet-1での経験をもとに，その改良に関して考察する．第二章ではDIMMnet-1の問題点を挙げ，第三章ではDIMMnet-1に計画はしていたが実際には実装されなかった機能について紹介し，第四章ではDIMMnet-2にお

† (株) 東芝，研究開発センター
Corporate Research and Developpment Center, Toshiba
†† 東京農工大学
Tokyo University of Agriculture and Technology
††† 慶應義塾大学
Keio University

ける改良方針について述べ，最後にまとめる．

2. DIMMnet-1 における問題点

2.1 メモリバスの変化の激しさへの不適応

同期 DRAM が出現して以降のメモリバスの進歩は凄まじい勢いであり，DIMMnet-1 の構想を練っていた PC100 が出始めた頃とは隔世の感がある．CPU の性能向上とともにメモリバスへの性能向上も要求されるためであり，この流れは今後とも続くことが予想される．

DIMMnet-1 は PC133 規格の DIMM スロット用に設計されたものの，現在では PC133 規格の量産型マザーボードを探すのが困難な状況になってきており，設計すべきメモリスロット仕様は極力先物である必要がある．新しい有望なメモリスロット規格が出現すれば，設計途中でも仕様変更に耐えうる柔軟性が必要になってくる．

2.2 高周波スロットへの対応

DIMMnet-1 は PC133 規格のメモリスロットに装着することを意図して設計されたが，実際に安定動作したのは 100MHz に過ぎなかった．現在の PC の主流は DDR 型のメモリに移行しており，周波数も年々上がっており，メモリスロットに本来想定されていない容量性負荷を追加してしまう現在の DIMMnet-1 のバンク切り替え用 FET スイッチを用いた構造は，これらの高周波化には向いていない．

2.3 マザーボード上のデータ線のねじれ

DIMMnet-1 は最初の実験ではまともには稼動しなかった．その原因は，使用したマザーボード上のノースブリッジとメモリスロット間のデータ線の配線が，対応する番号どおりに接続されていないということにあった．そのため，これまでの全ての実験は上記のデータ線ねじれを予めソフトウェアで逆にねじってから DIMMnet-1 に書き込むという操作を行って執り行われた．これは性能低下の一因となった．

2.4 チップセット毎のアドレス重畳仕様の違い

設計当初は同じメモリモジュールへのアクセスを行う各社のノースブリッジのアドレス重畳仕様は同一かと考えていたが，実際には各社各様で，同じメーカーでもチップセットが異なると仕様が異なることもあるということが判った．Intel に関しては仕様書が web 上に公開されていたが，それ以外のメーカーでは取り寄せなければ入手できなかったり，事情を説明して必要な情報だけでも提供していただけるように依頼しても黙殺されることもあった．快く提供いただいた VIA 社のチップセットが Intel との係争の対象になってしまい，マザーボードが市場にあまり出てこないという状況も実際に経験した．このように，事前に仕様が明らかになっていないと対応できない現在の DIMMnet-1 の設計では，チップセットメーカーの商業上の戦略に左右される危険性があったり，新しいチップセットでは使えなくなったりするという問題があった．

2.5 主記憶として使えるメモリの減少

アプリケーションの中には少しでも多くの主記憶があることが HDD へのアクセスを回避して現実的な時間で処理するためには不可欠である場合がある．そのような場合，限られた本数のメモリスロットの一つが DIMMnet-1 に占有されてしまうことは好ましくない．

2.6 バンク切り替えの煩雑さ

DIMMnet-1 上の 2 枚の SO-DIMM から構成されるバンクメモリは，ホスト側とネットワーク側に独立したバンド幅を提供できる点では大きなメリットがある．しかし，受信データをホストからアクセスしたり，送信データをホストから書いた後に実際に送信する際にはバンクを切り替えなければならない．このため，利用形態に制約が発生し，例えばメッセージ交換におけるバッファにホストが頻繁にアクセスする必要がある場合にはバンク切り替えの存在が支障になる．

一方，Martini チップ上の 2 ポートメモリである LLCM(Low Latency Common Memory) は経験上，大変使いやすく，LLCM の容量が少ないことを除けば，バンク切り替えをせずにホストからアクセスできるという LLCM の性質は使い勝手の面から重要であることが判った．

2.7 PEMM 規格立ち消えによる割込み信号の消滅

設計当初，PEMM(Processor Enhanced Memory Module) 規格が公開され，今後のノースブリッジには割込み信号が設置されるかに思われたが，実際には PEMM 規格をサポートするノースブリッジは出現しなかった．現状の DIMMnet-1 単独ではホストからのポーリング以外でのイベントの検出はできない．DIMMnet-1 のようにオンチッププロセッサが低速な場合や，将来オンチッププロセッサを持たない構成にした場合は，ホストへの割り込みがサポートされていた方が好ましい．

2.8 受信側 RDMA におけるバブル

DIMMnet-1 プロトタイプの実装が，これに先立ち推進されていた RHiNET プロジェクトに相乗りする形で実施されたという経緯から，RDMA によって長いメッセージにおいて高いバンド幅を出すことが第一義に設計された回路ブロックを，AOTF や BOTF といった細粒度通信機構を使った通信においても，やむを得ず使わなければならなかった．

AOTF に関しては Mini-OTF 受信部という近道的専用受信部を導入して救うことができたが，BOTF については専用受信部を断念した．その結果，本来 BOTF は高バンド幅を実現する上でも高い可能性を持っていたにも関わらず，実際のプロトタイプ上では受信側 RDMA 部が発生するバブルによってバンド幅が大きく低下する．

2.9 全体として周波数が低い

DIMMnet-1 を構成する専用 LSI である Martini は，開発コスト低減のために予め拡散層メモリを作りこんだウェファを RHiNET スイッチとも共用しており，チップの両端にあるメモリにアクセスしなければならないような回路ブロックが存在するなど，先端的なプロセスを用いていた割には全体としては周波数が上げられなかった．PCI バスを利用する RHiNET と DIMM インタフェースを利用する DIMMnet-1 の同一チップ上での共存によってもチップ面積を小さくすることはできない上に，配線領域の余裕が少ない状況であったことも周波数や歩留まりが上がらない原因でもあった．

その結果，第一版の Martini は大変歩留まりが低い上，AOTF から Mini-OTF 受信部を経由して LLCM に至る経路のみ 100MHz で動作したものの，その他の部分は 66MHz でしか動作できなかった．その結果，第二版の Martini では DIMM インタフェース部のみがその他の部分の倍の周波数で動作するように設計変更がなされた結果，DIMMnet-1 のコア部は 100MHz の半分の 50MHz でしか動作できない状況になり，当初の予定の 40%弱のバンド幅しか得られないものとなってしまった．

2.10 オンチップ CPU の遅さ

Martini のオンチップ CPU は verilog で記述して論理合成によって生成した簡素な CPU である．ホストの CPU のように内部並列性をひきだすような機構もない上，周波数は 1/50 程度であり性能的には二桁程度の差がある．

機能的には TLB のメインテナンスだけでなく，一部の回路ブロックの処理を代行してハード実装されていない処理をファームウェアで実現できる．しかし，実際に使ってみると処理能力が足りず，オンチップ CPU を使った場合，大幅な性能低下を引き起こす．

2.11 リードバンド幅の低下

ネットワークからのメッセージを受信する領域は，キャッ

シュしていると何も手を打たない場合はキャッシュ上のデータと，受信データの不整合が発生する．PCI バスでは主記憶への PCI バスからの書き込みの際に，CPU に対してキャッシュ無効化がハード的になされるので問題ない．一方，メモリスロットにはその機能がない．

そこで設計当初は DIMMnet-1 上のメモリは全て uncached 属性にしようと考えたが，バンド幅が劇的に低下してしまった．書き込みに関しては write combining 属性に設定することでバンド幅がキャッシュ領域並みに向上させることができたが，リードバンド幅に関しては write combining 属性は無力であった．

そこで，Pentium3 に関してはキャッシュ属性にして SSE 命令セットの PREFETCHNTA 命令によってキャッシュを汚さないようにバーストリードし，Pentium4 に関してはライン単位での無効化を行う命令を用いることでリードバンド幅の高速化を図っている．

2.12 i845 の常時ブロックアクセス問題

Intel i845 チップセットに関しては第二版 Martini においてはアドレス重畳規則を埋め込んであったにも関わらず動作させることができなかった．その原因はこのノースブリッジがメモリバスについては非キャッシュ領域に対しても常に 64 バイトのブロックアクセスをしてしまうためであった．

この問題に関しては，このような特殊なアクセス形態を設定により抑制できるような配慮をチップセットメーカーに期待する．

2.13 汎用スイッチへの接続性

DIMMnet-1 が接続できるスイッチは現状では RHiNET 用に研究開発された専用スイッチしかない．光モジュール自身が研究対象である試作品であったため，高価であるだけでなく，増産が事実上不可能であった．

一方，近年では同等クラスのバンド幅を有する市販のスイッチが存在するため，これらを流用可能であることが普及に際しては重要である．

2.14 アトミックオペレーションがない

設計当初，アトミックオペレーションが必要になるような操作はオンチップ CPU 上でエミュレーションすることにしていた．しかし，実際にはオンチップ CPU を使用すると性能低下が激しくなるために，簡素でもよいから何らかのアトミックオペレーションがハードで実装されていることが重要であった．もし，アトミックオペレーションがハードで実装されていれば，オンチップ CPU を使わずにソフトウェア分散共有メモリで用いられるロック等を高速に実装できたと思われる．

2.15 専用の支持構造が必要

MEMOnet は規格外の形状の基板をメモリスロットに装着することになるので，メモリスロットのコネクタにより機械的に安定するかどうかの配慮が必要であり，DIMMnet-1 の場合は Martini チップが冷却フィンの装着を必要としたり，基板からケーブルを引き出したりする関係上，使用する筐体と使用するマザーボードを決めて，それら専用の支持構造を作成した．ただし，実用化にはより汎用な支持構造が必要である．

2.16 ラックマウント筐体に入らない

1U のラックマウント型 PC を構成要素に PC クラスタを構成することが流行した時期があったが，その際に DIMMnet-1 の基板は大きさ的に 1U サイズには収まらないという問題があった．DIMMnet-1 はプロトタイプという位置づけから，多数の設定用 DIP スイッチを搭載しており，電源電圧変換モジュールも搭載しているため基板サイズがかなり大きくなってしまっている．LSI パッケージ・SO-DIMM コネクタ・リンク用コネクタ（または光モジュール）の大きさもそれなりにあるので，通常の DIMM サイズには実装が困難であり，DIP スイッチや電圧変換モジュールを取り外すだけでは 1U ラックマウント筐体に納めることは困難である．1U 筐体に PCI カードを納める際に使われるライザーカードのような構造を持ち込むことも考えられるが，DIMMnet-1 の場合は既にタイミング的に PC133 の制限を越える状態にあったので，ライザーカードを挟むとさらにメモリバスの周波数を落とさざるを得なくなり困難であった．

3. DIMMnet-1 で実装を見送った機能

3.1 送信ログと End to End 再送

設計当初，DIMMnet-1 には主記憶上にパケットイメージが存在するとは限らない Programed I/O により生成されたパケットの受信側での CRC エラー発生時などにおける再送を送信側 NIC で行えるようにするため，送信パケットのログ保存を指定できるパケットフォーマットを定義していた．さらに，シーケンス番号も付加できるようなパケットフォーマットを定義していた．しかし，RHiNET2 スイッチではリンクの信頼性が高く，RHiNET3 スイッチではリンクレベルの再送を行えるようになっていたので，End to End の再送を行うためのこの機能は省略され，実装されなかった．

3.2 受信側でのステータス書き込み指定

設計当初，DIMMnet-1 には送信データの到着を受信側で所定のアドレスへのポーリングにより知ることを可能にするために，受信側のステータス格納アドレスを指定できるパケットフォーマットを定義していた．しかし，これは実際には省略され，実装されなかった．このため，送信側はメッセージ本体の他に別途メッセージ送信が完了したことを示すフラグを転送する必要が出てしまい，ソフトウェアオーバーヘッド増加の原因となった．

3.3 リモート間接書き込み

リモート間接書込みは DIMMnet-1 の設計初期段階から企画されていた．[3] リモート間接書込みとは「送信側がポインタと書込みデータを含むパケットを送信することで，そのポインタで指し示される受信側にあるアドレス情報に基づき遠隔書込みが行われる通信方式」と定義している．図 **2** にリモート間接書込みのコンセプトを示す．

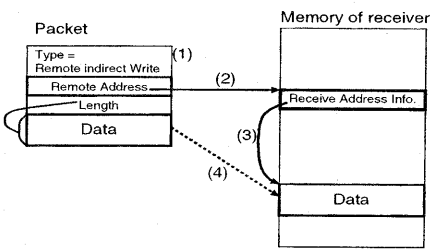

図 **2** リモート間接書込み
Fig. 2 Remote Indirect Write

DIMMnet-1 では本機能はハードでは実装されず，オンチッププロセッサによるエミュレーションではバンド幅の低下が著しかった．[24]

3.4 メッセージ交換支援

リモート間接書き込みの変形として，書き込んだ後のリングバッファアドレス更新を追加することにより，リモート FIFO 書き込みを企画していた．[3] この FIFO を大量に SO-DIMM 領域に形成することでメッセージ交換の高速化を図ろうとしていた．しかし，この機能も DIMMnet-1 においては実装が見送られた．

3.5 定形的データ長 PUSH の On-The-Fly 受信

データバスサイズ (1~8 バイト) やキャッシュラインサイズ (典型的には 32 バイトまたは 64 バイト) のデータをリモートライトするパケットに関しては，DMA コントローラを設定するまでも無く，メモリインタフェースに定形的データ長の書込みであることを伝えて専用のステートマシンで即座に書き込みを開始させることで受信側の遅延時間を節約可能にする On-The-Fly 受信部を設置する案があった．しかし，DIMMnet-1 には AOTF 送信部が発生した 1~8 バイトの書き込みのみサポートする Mini-OTF 受信部のみが実装され，それ以上のサイズに関しては実装が見送られた．その結果，BOTF が発生する短いパケットは全て RDMA 受信部を経由し，バブル発生によるバンド幅低下を免れなかった．

4. DIMMnet-2 における改善対策案

4.1 非分岐型インタフェースによる DDR 化

DIMMnet-2 では，メモリスロットには 1 個の LSI が接続される形態を取ることによって分岐を抑え，本来想定されていない容量性負荷の増加を防ぐことで対応できるメモリバスの高周波化を実現する．

メモリバスからみて LSI の向こう側に SO-DIMM が設置される．SO-DIMM ポートはデュアルチャネルとし，128bit 幅の 1 ポートメモリを擬似的に 64bit 幅の 2 ポートメモリに見せかけることで，ホストからのアクセスとネットワークからのアクセスのバンド幅を確保する．このようにすることで RDMA においても送受信同時実行時にバンド幅半減ということはなくなる．

ホスト側のメモリバスは最初のプロトタイプでは DDR-DRAM に準拠とするが，デファクトスタンダードの激しい動きを注視しつつ，ある時点で実現可能な最新のスタンダードに準拠するものとする．よって，できるだけホスト側のメモリバスに対して柔軟性のある方式を採用する．

4.2 プリフェッチ機能付き window メモリ

上記の方針を採用する結果，メモリバスからみて LSI の向こう側に SO-DIMM が設置されることから，BIOS におけるタイミング設定による調整の範囲を超えるおそれがある．

そこで，LSI 内にホストから BIOS 設定範囲内にアクセスできるバッファを設け，SO-DIMM 上のデータをホストが読み出す場合は，まず，バッファに対するプリフェッチ要求を出した後，バッファに要求したデータが準備されるタイミング以降に実際のデータ読み出しを行う．

プリフェッチの要求は，プリフェッチ要求制御語をユーザー空間にマップされた LSI 内レジスタに書き込むことにより行う．

図 3 にプリフェッチ機能付き Window メモリの構成を示す．バッファは DIMMnet-1 同様に Window 化され，アクセス権限のある Window のみがユーザー空間にマップされる．各 Window は 512 バイトの大きさに対して 1 個の 64bit のフラグレジスタが付いており，リセット後にフラグのあるビットに対応するダブルワードに書き込みや読み出しがあるとそのフラグが反転する．ホストはこの 64bit のフラグレジスタをリードすることで，どこまで必要なデータが準備できているかを知ることができる．

なお，DIMMnet-2 においてはプリフェッチ window もリモートライトの書き込み対象となる．これは，超並列計算機 TS/1 におけるリモートのベクトルレジスタ間のチェイニングを行うプロセッサ間チェイニング [27] の変形である．

4.3 バリエーション付きベクトルロードストア

プリフェッチのバリエーションとしては連続ブロックアクセス要求の他に，等間隔 (ブロックストライド) アクセス要求や，マスクベクトルアクセス要求，簡易リストベクトルアクセス要求 (数個の不規則不連続データへのコンパクション型プリフェッチ要求) を出せるようにする．

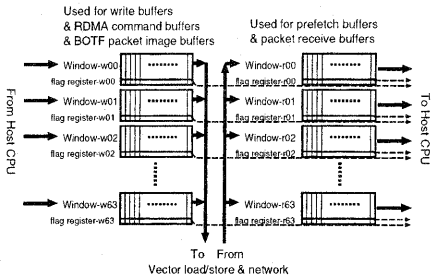

図 3 プリフェッチ機能付き Window メモリ
Fig. 3 Window memory with prefetch

ストアについても同様なバリエーションをサポートする．さらに，書き込み先は SO-DIMM だけでなくリモートメモリも指定可能とする．つまり，これは DIMMnet-1 における BOTF の変形であり，バッファ上にデータがセットされた状態で送信制御語を書き込む (キックする) ことにより，ストアバリエーション付きのパケットを生成し，リモートノードでそのバリエーションに応じた展開アクセスが行われる．

これらにより，不連続アクセスに伴うホスト CPU のキャッシュやメモリバスの効率的利用を可能にする．

4.4 スロット限界容量以上の多バンクメモリ実装

DIMMnet-2 では EMS のように，ある瞬間に見える窓は狭いが，バンクを切り替えながら窓を動かすことで広大な空間を一つのプロセスに見せる．

メモリスロット上の物理アドレスがどのバンクアドレス拡張レジスタでアドレス拡張されて SO-DIMM をアクセスするかを，カレントプロセス ID(CPID) レジスタから知るというハード機構を実装する．自分のプロセス ID(PID) に対応するバンク拡張アドレスレジスタはユーザモードで書き換え可能な場所にマップされ，ここを書き換えることで別のバンクが見えるようになる．

プロセススイッチのたびに，カレントプロセスが NIC にアクセスすることを予約したプロセスかどうかを検査し，そうならば今から次のプロセススイッチまでアクセスを行うプロセスに対応する PID は何番だということをカーネルが NIC に設定する．

4.5 プラグアンドプレイねじれ解消

プリフェッチの導入により緩和されたタイミング制約を背景に，DIMMnet-2 ではデータ線のねじれ解消論理を LSI 内に実装する．さらに初期化時にデータ線の対応を CPU 側に知らせるためのステートマシンを実装し，どのラインが何ビット目に接続されているかを CPU 側の初期化プログラムから検出可能にする．検出された接続パターンからねじれ解消のための設定値を算出し，これを LSI にロードする．

アドレスについてはアクセス時に CPU からメモリスロットに出力されたアドレスをデータとして CPU 側に返納させることを可能とする初期化回路により，アドレスの重畳規則を CPU 側の初期化プログラムから検出可能にする．検出された重畳規則から物理アドレス再構成のための設定値を算出し，これを LSI にロードする．

4.6 バーストアクセス保持型 AOTF

DIMMnet-2 における AOTF 送信部では，AOTF キックアドレスに該当するアドレスへのバーストアクセスを検出し，これらを一つのヘッダーと連結させることでリモート物理アドレスによるリモートブロック書き込みを実現する．

4.7 定形的データ長 PUSH の On-The-Fly 受信

改良型 AOTF が発生するデータバスサイズ (1~8 バイト) やキャッシュラインサイズ (32 バイトおよび 64 バイト) のデータをリモートライトする PUSH パケットに関しては，ヘッダー中にフラグが立ち，受信側でそのフラグを検出すると物理アドレスに基づく On-The-Fly 受信が行われる．

BOTF が発生するパケットについても物理アドレス指定を可能にし，受信側で PID が登録されていることを並列比較器によりチェックし，アドレス上位が PID に対応するもので上書きされた後に上記と同様の On-The-Fly 受信が行われる．

4.8 Infiniband 用スイッチの利用

DIMMnet-2 では物理層は 4X タイプの Infiniband に合わせ，リンク層，すなわち Infiniband のスイッチが解釈できるヘッダー (LRH = Local Route Header) だけ Infiniband に合わせて設定することでコストパフォーマンスが高い市販の 4X タイプの Infiniband スイッチを利用する．

4.9 スイッチインタフェースの PCI 基板化

小規模クラスタ向けの直接接続だけでなく，フォールトトレランスの観点からも通信リンクを複数有することが重要である．そのためには，通信リンクは安価かつコンパクトに実装できなければならない．

DIMMnet-2 においては Infiniband の規格には 12X タイプという 4X タイプを 3 本束ねた仕様が定義されていることも考慮し，4X タイプの Infiniband のリンクを 3 ポート外部に引き出せる PCI ボードを併用する．

4.10 PCI 基板によるホスト割込み

PEMM 規格のチップセットは今後も出荷されることはなく，さらに ASIC によるオンチップ CPU の能力はホスト CPU に比べて圧倒的に低くなるという見地に立ち，DIMMnet-2 では併用する PCI 基板には，NIC-LCI からホストに対する割込み信号線を導き，NIC-LSI における予期せぬイベントに対するホストでの対応やハードで実現されていない複雑な機能のホストによるエミュレーションを円滑に進める．この PCI 基板の存在は DIMMnet-2 のプラグアンドプレイ化にも寄与する．

4.11 ハードワイヤード TLB リフィル

バンド幅が大きいということは大量のデータが一気に流れ込んでくるということは十分ありうることであり，その場合 TLB のミスヒットが頻発する状態に陥るので対策を打っておくことが重要である．

ソフトで対応した場合はミス時にマイクロ秒オーダーがかかるが，ハードで数十ナノ秒オーダーで済むため，大量データ受信やミスヒットしやすいアクセスパターンだった場合のバンド幅は向上すると考えられる．

4.12 アトミックオペレーション

DIMMnet-2 ではアトミックオペレーションをハードワイヤードで実現する．キューベースのロックよりも競合が増えたときに効率的なロックである HBO Lock[28] の実装には Compare & Swap アトミックオペレーションのみが必要であるので，そのハードサポートを考慮する．

4.13 ハードワイヤードリモート間接アクセス

DIMMnet-2 ではハードワイヤードなリモート間接アクセス機構を実現する．これにより受信側の状態をチェックせずに送信を開始する one sided なメッセージ交換を実現し，メッセージ交換における通信遅延の短縮化を試みる．

4.14 オンチッププロセッサ廃止

ASIC によるオンチップ CPU の能力はホスト CPU に比べて圧倒的に低く，TLB のリフィルやアトミックオペレーションもハードワイヤードで行われ，その他のイベントにはホストに割り込みをかけられるようにするので，DIMMnet-2 ではオンチップ CPU は廃止する．これにより ASIC のダイサイズの縮小，動作周波数の向上，設計コストの削減を図る．

4.15 End to End 再送

DIMMnet-2 では市販のスイッチに接続可能とするために，スイッチにおけるリンクレベルの再送を行うことができない．そこで，DIMMnet-2 では End-to-end の NIC 間でエラー発生時の再送を行う．順序管理用のシーケンス番号はハード的にヘッダーに組み込まれる．

AOTF や BOTF といった Programed I/O により生成されたパケットの受信側での CRC エラー発生時などにおける再送を送信側 NIC で行えるようにするため，再送制御が必要なパケットまたはそれを再生可能な情報を送信側に保持する．

AOTF に関してはヘッダーにフラグが立っているパケットのログをとっておく方式の他に，ホストからの AOTF 起動時のトランザクションを保持しているトランザクション FIFO のエントリ解放時期を ACK 受信まで延期する方法も考えられるので，どちらにするかは今後検討する．

BOTF に関してはホストから書き込まれるパケットイメージを保持する window メモリの解放時期を ACK 受信まで延期してしまうと，その間メモリのプリフェッチ等が阻害されるおそれがあるため，ヘッダーにフラグが立っているパケットのログをとっておく方式をとる．ACK によって対応するパケットのログは解放される．ログは一旦オンチップのバッファに保持し，溢れた場合に SO-DIMM 側に吐き出されるが，その前に ACK が戻ってきた場合は SO-DIMM へのアクセスは発生しない．

RDMA に関しては，SO-DIMM 上にパケットイメージが残っているのでそれを再送に用いる．

4.16 可変なプロテクションスタンプ位置

BOTF において刻印されるプロテクション情報の位置を可変にする．これにより，スイッチの種類が変わった場合や，Infiniband でもスイッチだけでなくルーターが混在するシステムへの対応をする場合にも対応可能になる．

4.17 後続パケットの第 3 層以上ヘッダー省略

最大パケット長を超えるメッセージはパケットに分割されるが，DIMMnet-2 では AOTF や BOTF が発生する細粒度パケットによるメッセージ送信の高バンド幅化を促進するために，後続パケットのネットワーク層以上のヘッダーは省略形を用いる．

Infiniband のリンク層ヘッダーはノードの ID として約 5 万ノードが指定できるので，PC クラスタとしては十分なノード数をアドレスできると考え，ネットワーク層ヘッダーの IP 部 (SGID,DGID)32 バイト分は DIMMnet-2 では省略する．

5. おわりに

本報告では，DIMM スロット装着型ネットワークインタフェースの DIMMnet-1 プロトタイプの試作経験により判明した問題点を述べた．これらはリスキーな開発を実際に手がけてみないとわからなかったことが多い．さらに，限られた設計リソースの制約から実装を見送った機能について述べた．実用化に際してはこれらの問題点の解決や未実装項目の実装が重要であり，DIMMnet-2 における改良方針に関して考察した．ネットワークインタフェースとしての発展の方向性とともに，キャッシュアーキテクチャを補完する高機能メモリとしての新機軸を打ち出している．

なお，DIMMnet-1 は RWCP のもとで研究開発されたが，2002 年度から 5ヵ年の計画で総務省の研究資金サポートのもとで，その改良版の MEMOnet が開発されることになっている．

今後は，本報告で述べられた DIMMnet-2 における改良方針をベースに，より詳細な検討と設計を進め，FGPA ベースの簡易版プロトタイプを作成した後に，ASIC 化した改良版の

MEMOnet が開発される予定である．

さらに，DIMMnet-1 および DIMMnet-2 を用いたソフトウェア分散共有メモリやコンパイラの研究も併せて行われることが予定されている．

謝辞

本研究は総務省戦略的情報通信研究開発制度の一環として行われたものである．DIMMnet-1 に関しては新情報処理開発機構が推進してきた RWC (Real World Computing) プロジェクトの並列分散コンピューティング技術研究の一環として研究開発されたものである．

産業技術総合研究所の工藤氏，(株) 日立製作所の山本氏，西氏，慶應義塾大学の渡辺氏，元・慶應義塾大学の土屋氏，元・東京農工大学の須田氏，(株) 日立 IT の今城氏，上嶋氏，金野氏，寺川氏，慶光院氏，岩田氏，山本氏，柏原氏，大杉氏をはじめ Martini LSI および DIMMnet-1 の開発に携わった全ての方々と，DIMMnet-2 の開発に関する議論にご参加いただいている和歌山大学の国枝教授，上原講師，齋藤助手，慶應義塾大学の大塚氏，伊豆氏，北村氏に感謝いたします．

参 考 文 献

1) PCI-SIG, http://www.pcisig.com/
2) InfiniBand Trade Association, http://www.infinibandta.org/
3) 田邊, 山本, 工藤：メモリスロットに搭載されるネットワークインタフェース MEMnet, 情報処理学会計算機アーキテクチャ研究会, Vol. 99, No. 67, pp. 73-78, (Aug. 1999)
4) 田邊, 山本, 工藤：メモリスロット搭載型ネットワークインタフェース DIMMnet-1 における細粒度通信機構, 情報処理学会計算機アーキテクチャ研究会, Vol. 2000, No.23, pp. 65-70, (Mar. 2000)
5) Tanabe, Yamamoto, Nishi, Kudoh, Hamada, Nakajo, Amano : MEMOnet : Network interface plugged into a memory slot, *IEEE International Conference on Cluster Computing (CLUSTER2000)*, pp.17-26 (Nov. 2000)
6) Tanabe, Yamamoto, Nishi, Kudoh, Hamada, Nakajo, Amano : On-the-fly Sending : A Low Latency High Bandwidth Message Transfer Mechanism, *5th International Symposium on Parallel Architectures, Algorithms, and Networks (I-SPAN2000)*, pp.186-193 (Dec. 2000)
7) 山本, 田邊, 西, 土屋, 渡辺, 今城, 上嶋, 金野, 寺川, 慶光院, 工藤, 天野：高速性と柔軟性を併せ持つネットワークインタフェース用チップ: Martini, 情報処理学会計算機アーキテクチャ研究会, Vol.2000, No.110, pp.19-24 (Nov. 2000)
8) 田邊, 山本, 今城, 上嶋, 濱田, 中條, 工藤, 天野：DIMM スロット搭載型ネットワークインタフェース DIMMnet-1 の試作, 情報処理学会 HPC 研究会 (SWoPP2001), Vol.2001, No.77, pp.99-104 (Jul. 2001)
9) 濱田, 中條, 田邊, 工藤：メモリバスに接続される NIC による PC クラスタの性能予測. 情報処理学会計算機ハイパフォーマンスコンピューティング研究会 (SWoPP 2001) 2001-HPC-87, pp.105-110 (Jul. 2001)
10) 渡邊, 山本, 土屋, 田邊, 西, 今城, 寺川, 上嶋：RHiNET/MEMOnet ネットワークインタフェース用コントローラチップ Martini の予備評価, 情報処理学会計算機アーキテクチャ研究会 (SWoPP 2001) 2001-ARC-144, pp. 49-54 (Jul. 2001)
11) 山本, 土屋, 寺川, 田邊, 渡邊, 今城, 西, 工藤："RHiNET の概要と Martini の設計／実装", 情報処理学会計算機アーキテクチャ研究会 (SWoPP 2001) 2001-ARC-144, pp. 37-42 (Jul. 2001)
12) Tanabe, Hamada, Yamamoto, Kudoh, Imashiro, Nakajo, Amano : A prototype of high bandwidth low latency network interface plugged into a DIMM slot, International Conference on Advances in Infrastructure for Electronic Business, Science and Education on the Internet (SSGRR2001), http://www.ssgrr.it/en/ssgrr2001/papers/Noboru%20Tanabe.pdf (Aug. 2001)
13) 田邊, 濱田, 山本, 今城, 中條, 工藤, 天野：DIMM スロット搭載型ネットワークインタフェース DIMMnet-1 の通信性能評価, 情報処理学会計算機アーキテクチャ研究会 (DegisgnGaia2001) 2001-ARC-145, pp.51-56 (Nov. 2001)
14) Tanabe, Hamada, Suda, Nakajo, Yamamoto, Imashiro, Kudoh, Amano "DIMMnet-1 : A Low Latency High Bandwidth Network Interface for PC Cluster", 5th International Workshop on Innovative Architecture for Future Generation High-Performance Processors and Systems (IWIA2002), (Jan. 2002)
15) Tanabe, Yamamoto, Nishi, Kudoh, Hamada, Nakajo, Amano : Low Latency High Bandwidth Message Transfer Mechanisms for a network interface plugged into a memory slot, *Cluster Computing Journal*, Vol.5, No.1, pp.7-17 (Jan. 2002)
16) Watanabe, Yamamoto, Tsuchiya, Tanabe, Nishi, Kudoh, and Amano "Preliminary Evaluation of Martini: A Novel Network Interface Controller Chip for Cluster-based Parallel Processing", Twentieth IASTED International Conference Applied Informatics (AI 2002) (Feb. 2002)
17) 山本, 渡邊, 土屋, 原田, 今城, 寺川, 西, 田邊, 上嶋, 工藤, 天野："高性能計算をサポートするネットワークインタフェース用コントローラチップ Martini", 並列処理シンポジウム JSPP2002, pp.35-42 (May 2002)
18) 田邊, 山本, 濱田, 中條, 工藤, 天野：DIMM スロット搭載型ネットワークインタフェース DIMMnet-1 とその高バンド幅通信機構 BOTF, 情報処理学会論文誌, Vol.43, No.4, pp.866-878 (Apr. 2002)
19) 田邊, 濱田, 三橋, 山本, 今城, 中條, 工藤, 天野："DIMMnet-1 プロトタイプによるバンド幅と大域演算性能の評価", 情報処理学会計算機アーキテクチャ研究会 (SWoPP2002) 2002-ARC-149, pp.97-102 (Aug. 2002)
20) Tanabe, Hamada, Suda, Nakajo, Yamamoto, Imashiro, Kudoh, Amano : "Low Latency Communication on DIMMnet-1 Network Interface Plugged into a DIMM Slot", International Conference on Parallel Computing in Electrical Engineering (ParElec2002), pp.9-14 (Sep. 2002)
21) 濱田, 三橋, 田邊, 中條, 工藤："高速通信インタフェース DIMMnet-1 の通信バンド幅評価", 第 1 回 情報科学技術フォーラム (FIT2002) (Sep. 2002)
22) 三橋, 濱田, 田邊, 中條, 工藤："高速通信インタフェース DIMMnet-1 を備えた PC クラスタの評価", 第 1 回 情報科学技術フォーラム (FIT2002) (Sep. 2002)
23) 山本, 渡邊, 土屋, 原田, 今城, 寺川, 西, 田邊, 上嶋, 工藤, 天野："高性能計算をサポートするネットワークインタフェース用コントローラチップ Martini", 情報処理学会論文誌ハイパフォーマンスコンピューティングシステム, Vol.43, No.SIG6(HPS5), pp.122-133 (Sep. 2002)
24) 田邊, 濱田, 三橋, 山本, 今城, 中條, 工藤, 天野："DIMMnet-1 における Martini オンチッププロセッサによる通信の性能評価, 情報処理学会計算機アーキテクチャ研究会 (DegisgnGaia2002) 2002-ARC-150, pp.53-58 (Nov. 2002)
25) Tanabe, Hamada, Mitsuhashi, Nakajo, Yamamoto, Imashiro, Kudoh, Amano "Performance Evaluation of Bandwidth and Global Operations on DIMMnet-1 Prototype", 6th International Workshop on Innovative Architecture for Future Generation High-Performance Processors and Systems (IWIA2003), (Jan. 2003)
26) 田邊, 濱田, 山本, 今城, 中條, 工藤, 天野：DIMM スロット搭載型ネットワークインタフェース DIMMnet-1 とその低遅延通信機構 AOTF, 情報処理学会論文誌ハイパフォーマンスコンピューティングシステム, Vol.43, No.SIG(HPS6), pp.10-23 (Jan. 2003)
27) 田邊 "超並列テラフロップスマシン TS/1 における並列処理～プロセッサ間チェイニングとその応用～", 情報処理学会論文誌, Vol.36, No.3, pp.658-668 (Mar. 1995)
28) Radovic and Hagersten : "Hierarchical Back-Off Lock for Non-Uniform Communication Architectures", 9th International Symposium on High Performance Computer Architecture (HPCA-9), pp.241-252 (Feb. 2003)