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

田邊 昇[†1] 濱田芳博[†2] 山本淳二[†3]
今城英樹[†4] 中條拓伯[†2]
工藤知宏[†5,☆] 天野英晴[†6]

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．DIMMnet-1 は AOTF(Atomic On-The-Fly) という低遅延通信機構と BOTF(Block On-The-Fly) という高バンド幅通信機構を装備している．現在，Marini LSI の初期バージョンによって作成された電気リンク版および光リンク版 DIMMnet-1 は Pentium3 および Pentium4 ベースのパソコンの 100 MHz で駆動される DIMM スロット上で動作している．本報告では DIMMnet-1 プロトタイプの実機上での AOTF を用いた通信性能の評価結果を示す．

AOTF: A Low Latency Communication Mechanism of DIMMnet-1 Network Interface Plugged into a DIMM Slot

Noboru Tanabe,[†1] Yoshihiro Hamada,[†2] Junji Yamamoto,[†3]
Hideki Imashiro,[†4] Hironori Nakajo,[†2] Tomohiro Kudoh[†5,☆]
and Hideharu Amano[†6]

A high performance network interface architecture for PC clusters called DIMMnet-I that can be directly plugged into DIMM slot of PCs is presented. By using both a low latency AOTF (Atomic On-The-Fly) sending and a high bandwidth BOTF (Block On-The-Fly) sending, it can overcome the overhead caused by standard I/O like the PCI bus. Now, two types DIMMnet-1 prototype boards (providing optical and electrical network interface) consisting with a network interface controller chip Martini are available. They can be plugged into 100 MHz DIMM slot of PCs with Pentium 3 and Pentium 4. Experimental evaluation results of communication performance with the AOTF sending on a real system are shown.

1. はじめに

近年，高性能 PC を多数用いて並列処理を行ういわゆる PC クラスタが注目されている．高性能な PC クラスタ用に Myrinet[1)]，PCI-SCI[2)~4)]，MEMORY CHANNEL2[5),6)]等の高速ネットワークインタフェース（NIC）が各種開発されており，これらはいずれも PCI バスに接続される．光インタコネクションの持つ大きなバンド幅を有効に活用するには従来の PCI バスでは遅延時間もさることながらバンド幅の不足が深刻である．

一方，Infiniband[7)]が次世代のサーバ向け入出力の規格として提案され，製品が開発されつつある．しかし，最も価格性能比においてメリットのあるエンドユーザ用の量産 PC に，Infiniband が普及するかどうかは，Intel の撤退という状況を鑑みると望みが薄くなってきている．GigaE PM2[8)]を用いる等してすべてをコモディティ部品で構築するシステムよりも十分優れた性能を実現しつつ，価格性能比を最大にする PC クラスタを構築するためには，Infiniband 等とは

†1 株式会社東芝，研究開発センター
Corporate Research and Development Center, Toshiba
†2 東京農工大学
Tokyo University of Agriculture and Technology
†3 株式会社日立製作所
Hitachi Ltd.
†4 株式会社日立インフォメーションテクノロジー
Hitachi Information Technology
†5 新情報処理開発機構
Real World Computing Partnership
†6 慶應義塾大学
Keio University
☆ 現在，産業技術総合研究所
Presently with National Institute of Advanced Industrial Science and Technology

別のアプローチも検討に値する．

このような背景から我々は，従来のように PCI バス等の入出力バスではなく，メモリスロットに搭載されるタイプの NIC を検討してきた．このようなクラスの NIC を MEMOnet [9] と名付けた．MEMOnet は安価な PC 上で，PCI バスのバンド幅の限界を大幅に超越した NIC を実現可能とするのみならず，遅延時間においても優れた特性を示すと思われる．我々は MEMOnet のプロトタイプとして DIMM スロットに搭載される DIMMnet-1 [10),11)] を開発した．

この DIMMnet-1 や，同一の Martini LSI [12)] を用いた PCI 版 NIC である RHiNET2/NI [13)] には，AOTF および BOTF というプロテクションを確保しつつ低遅延な通信を実現する通信機構が搭載されている．これらは，1990 年頃に東芝で開発された高並列計算機 Prodigy [14)] の S-BUS 版ホストインタフェースに適用されている 2 ポートメモリへの書き込みをベースにした低遅延高バンド幅通信技術 [15)] や，RWCP 超並列東芝研究室で設計された超並列計算機 TS/1 の分散共有メモリアクセス機構である CTLB という通信制御情報の再利用機構 [16)] を，PC クラスタ用 NIC 向けに改良を施したものである．

低遅延通信を実現する他のアプローチとしては，1993 年頃から発表されている SHRIMP における VMMC [17)] や，1992 年頃から超並列計算機 JUMP-1 の通信機構として提唱された MBP [18)] がある．MBP は，多機能なメモリベースト通信を実現することが特徴とされている．この「CPU の MMU を介したメモリアクセスにより通信を起動することで低遅延通信とプロテクション維持を両立する方式」は，Prodigy の S-BUS 版ホストインタフェースにおいて MBP の提案に先立って実現され，その流れを汲む DIMMnet-1 の AOTF や BOTF にも，その特徴は受け継がれた．

一方，DIMMnet-1 ではメモリベースト通信という MBP と共通のアプローチをとりつつも，DIMM という大半のパソコンで利用可能な高性能なインタフェースを初めて NIC に採用した．さらに，MBP の思想とは逆に，高周波動作するホスト CPU からオフロードする機能を十分に絞り，送信側 CPU から受信側 CPU に至る経路全体にわたって通常動作時には単純なハードのみで処理されるよう注意して，ASIC 上のプロセッサには頼らない実現を徹底した．こうして，DIMMnet-1 では大幅に改善された低遅延通信と，凄まじい高速化をとげるパソコンの高い性能の有効利用を実現している．

本論文では，試作された DIMMnet-1 プロトタイプについて紹介し，そのアーキテクチャを解説する．その実機上で測定された AOTF を用いた細粒度通信性能として 4 バイトのラウンドトリップタイムやバリア同期および大域加算に関して報告する．最後に，その他の代表的な低遅延 NIC との違いについて明らかする．

2. DIMMnet-1 プロトタイプ

我々は MEMOnet や AOTF 等の種々のアーキテクチャの有効性を実証すべく，DIMMnet のプロトタイプ DIMMnet-1 を開発した．本章ではその概要を述べる．

2.1 DIMMnet-1 の概要

DIMMnet-1 は，PC66，PC100 または PC133 仕様の DIMM スロットに装着するネットワークインタフェースである．DIMMnet-1 の主な仕様を**表 1** に，その基本構造を**図 1** に示す．後述する Martini LSI は低遅延の FET バススイッチにより 2 バンクの SO-DIMM（ノート型 PC で用いられる汎用部品）を切り替えつつ，リンクインタフェースとデータの送受信をする．DIMM スロットの信号をじかに入力する DIMM 型 NIC 制御ポートを有する．メモリバス側のインタフェースは日本電子機械工業会規格の「プロセッサ搭載メモリ・モジュール（PEMM）動作仕様標準」[20)] に準拠した．PEMM 規格準拠のチップセットやマザーボードは現状では存在しないので，PEMM 準拠モード以外にも，PEMM で追加された 2 つの信号（バンクメモリへのアクセスを待たせる信号と割込み信号）

表 1 DIMMnet-1 の主な仕様
Table 1 Basic specifications of DIMMnet-1.

ホストインタフェース	SDR 型 DIMM および PEMM
共有バンクメモリ	PC133, SO-DIMM 2 枚
搭載 SO-DIMM 容量	64 MB〜1 GB
低遅延共有メモリ容量	128 KB（オンチップ）
命令 SRAM 容量	128 KB（オンチップ）
データ SRAM 容量	128 KB（オンチップ）
オンチップ CPU	R3000 風 32 bitRISC
通信リンクバンド幅	o2: 各方向 8 Gbps o3: 各方向 10 Gbps e(OIP): 各方向 2.5 Gbps e(RN2): 各方向 8 Gbps
バンクメモリバンド幅	1024 MB/s（ホスト側） 1024 MB/s（network 側）
最短送信時 NIC 遅延	105 ns（DIMM〜リンク）
最短受信時 NIC 遅延	90 ns（リンク〜LLCM）
NIC-LSI のテクノロジ	0.14 μm CMOS
対応するチップセット	Pro133, Pro266（Pentium3） P4X266, P4M266（Pentium4） KT133（Athlon, AthlonXP）

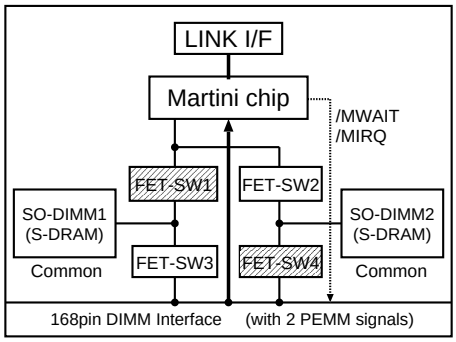


図 1 DIMMnet-1 の基本構造
Fig. 1 Basic structure of DIMMnet-1.

表 2 DIMMnet-1 に接続可能なスイッチの仕様
Table 2 Specification of switches for DIMMnet-1.

スイッチ	RHiNET2 [21]	RHiNET3 [22]	OIP-SW [23]
光 port	8 (or 2)	8	15
電気 port	0 (or 6)	0	1
I/O ピン	800 Mbps × 10	1.25 Gbps ×8	250 Mbps ×9
バンド幅	8 Gbps	10 Gbps	2.5 Gbps
距離（光）	100 m	1 km	100 m
距離（電気）	5 m	-	5 m
再送制御	N/A	OK	N/A
Table routing	OK	OK	N/A
Source routing	N/A	OK	OK
開発元	RWCP & 日立	RWCP & 日立	NEC & RWCP

がなくても動作するモードの 2 つのモードを有する．

2.2 DIMMnet-1 とスイッチの種類

DIMMnet-1 は**表 2** に示される 4 種類のスイッチおよび DIMMnet-1 どうしが接続可能である．DIMMnet-1 には電気版のスイッチに合わせたコネクタを搭載する基板（DIMMnet-1/e），光版 RHiNET2/SW に合わせたインタフェースを搭載する基板（DIMMnet-1/o2），光版 RHiNET3/SW に合わせたインタフェースを搭載する基板（DIMMnet-1/o3）の 3 種類の基板タイプがあり，現時点では DIMMnet-1/e（**図 2**）と DIMMnet-1/O2（**図 3**）が完成している．現在のところ，DIMM 上の周波数が 66 MHz および 100 MHz での動作が確認されている．

電気版のインタフェースを備えるスイッチとしては RWCP 光 NEC 研究室が開発した OIP（Optical IP）を用いた OIP スイッチと，RHiNET2/SW の電気版の 2 種類が開発され，現時点ではこれらはともに調整中である．DIMMnet-1 は OIP スイッチが持つ 1 つの電気ポートや電気版の RHiNET2/SW と LVDS レベルの電気信号を用いたケーブル接続により接続可能である．

図 2 DIMMnet-1/e
Fig. 2 DIMMnet-1/e.

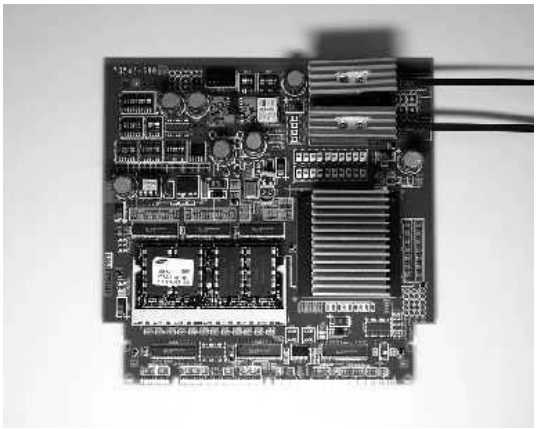


図 3 DIMMnet-1/o2
Fig. 3 DIMMnet-1/o2.

現時点では，光版のスイッチは RHiNET2/SW が完成しており，RHiNET3/SW は調整中である．

2.3 Martini LSI

Martini LSI は，PCI バスベースの RHiNET-2/NI と DIMM スロットベースの DIMMnet-1 の機能を 1 チップで実現する NIC 制御チップである．低遅延と高バンド幅が要求される単純なデータ転送はハードウェアのみによりサポートし，ロックや同期通信等の機能はチップ内に実装されたコアプロセッサにより実現する．モジュール単位のパイプライン化と代行機能により，コアプロセッサは，ハードウェアの一部を動作させながら，処理に介入することが可能であり，柔軟なソフトウェア/ハードウェア処理分担が可能となっている．

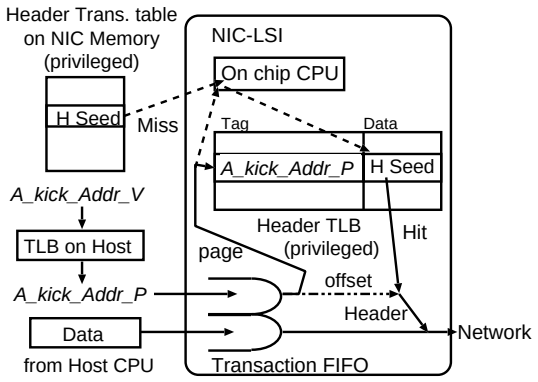


図 4 Atomic オンザフライ（AOTF）送信
Fig. 4 Atomic On-the-fly sending.

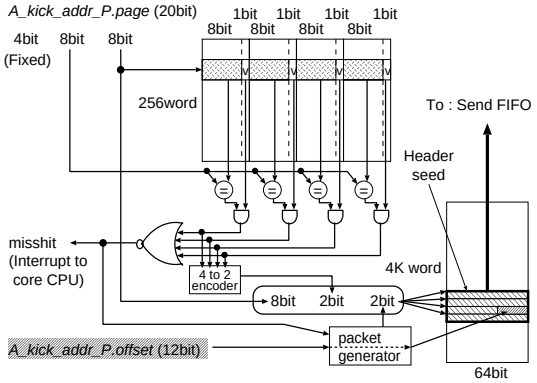


図 5 ヘッダ TLB（HTLB）の構造
Fig. 5 Structure of a header TLB.

Martini LSI は 3 つのバージョンが開発されているが，最初の 2 つのバージョンは論理的なデバッグが不十分である．特に最初のバージョンは遅延チューニングが不十分かつ，レイアウト上の問題もあり電源電圧を規定より落とさなければ使えないため，予定された周波数での動作ができない．本論文中の実験で用いられたのは最初のバージョンの Martini LSI である．

なお，規定の電源電圧においては最初のバージョンの Martini を用いた場合でも，DIMMnet-1 へのホストからのアクセスは正常に動作した．よって，DIMMnet の基本的なコンセプトが PC133 上で実現可能であることは，プロトタイプの試作により実証されたといえる．

3. Atomic オンザフライ（AOTF）送信

Atomic On-the-fly（AOTF）送信は，後述するヘッダ TLB（HTLB）を用いることにより，メモリバス上の 1 回の書き込みアクセスによって起動される低オーバヘッドな送信アーキテクチャである．送信すべきデータがレジスタ上に存在すれば，CPU がレジスタ上にあるデータをユーザモードのまま所定の仮想アドレスに書き込むというわずか 1 命令を実行するだけでパケット送信を起動できる．AOTF 送信におけるパケット生成メカニズムを**図 4** に示す．

なお，AOTF 送信機能は Martini LSI に搭載されており，DIMMnet-1 のみならず，Martini LSI を用いた PCI バスベースの NIC である RHiNET-2/NI でも利用可能である．

3.1 ヘッダ TLB

AOTF 送信はヘッダ TLB（HTLB）により実現される．HTLB は AOTF 送信起動のために割り当てられたアドレス（AOTF キックアドレス）へのアクセスからパケットヘッダを連想するハードウェアである．**図 5** に DIMMnet-1 における HTLB の構成を示す．HTLB はヘッダシードとアドレスの対応関係を保持する．ヘッダシードとは送信すべきパケットのヘッダから，リモートアドレス部の下位が削除されたものである．これが登録されるヘッダ変換テーブルや，そのキャッシュである HTLB はユーザモードからは直接は触れることのできない場所に配置される．

DIMMnet-1 の HTLB は 4 ウェイセットアソシアティブ構成で 9 ビット幅 1024 エントリのタグ部を有する．ヘッダシードは 64 ビット幅 4096 語構成のオンチップメモリに記憶される．DIMMnet-1 では HTLB とヘッダ変換テーブルの管理はホスト CPU および Martini LSI 上のコア CPU の双方から行うことが可能である．

次に，HTLB を用いた AOTF 送信の動作を図を用いて順を追って説明する．

(1) AOTF キックアドレスの仮想アドレス（図 4 の A_kick_addr_V）は CPU 内部の TLB により物理アドレス（図 4 の A_kick_addr_P）に変換され，チップセットに渡される．なお，DIMMnet の場合は上記のアドレスは DIMM スロット上では通常 ROW アドレスと COLUMN アドレスにマルチプレクスされて現れるので，チップセットに応じて元の A_kick_addr_P に復元される．

(2) A_kick_addr_P はデータとともにトランザクション FIFO に格納される．なお，DIMMnet-1 の場合は 1〜8 バイトのデータ送信をサポートするために，ここでバイトイネーブル信号から生成したデータ長もあわせて格納している．

(3) トランザクション FIFO から取り出した A_kick_addr_P の上位アドレス（図 5 におけ

る A_kick_addr_P.page）から HTLB はヘッダシードを連想する．

(4) AOTF キックアドレスの下位 bit（図 5 における A_kick_addr_P.offset で DIMMnet-1 の場合 12 bit）がヘッダシードのリモートアドレスフィールドに上書きされてヘッダが完成する．なお，DIMMnet-1 の場合は 1〜8 バイトのデータ送信をサポートするために，ここでトランザクション FIFO から取り出したデータ長もヘッダシードのデータ長フィールドに上書きする．

(5) 起動時に書き込まれたデータをヘッダに添付することでパケットが生成される．

3.2 ヘッダの再利用性

DIMMnet-1 では AOTF 送信のほかに BOTF 送信[26]が利用可能であり，どちらもヘッダの再利用性がある．しかし，これらの 2 つの送信機構のヘッダの再利用性には差がある．まず，ヘッダ再利用に際して，AOTF になくて BOTF にある送信側のオーバヘッドとしては，先行するパケットが送り終わる前に Window メモリを上書きすると先行パケットが正しく送れないので，再利用しようとする Window メモリの上書き可否ステータスチェックが必要である．これは非キャッシュ領域へのリードになるのでその所要クロック数は比較的大きいものである．

これに対し，AOTF はトランザクション FIFO のクレジット（ユーザに与えられたアクセス回数）を使い切るまでは，前回いつ利用したヘッダを再利用するのかによらず，送る前のステータスチェックが不要である．よって，再利用する場合の送信側の遅延時間は，AOTF の方が BOTF よりその分短くなる．

さらに再利用できるヘッダの数は，BOTF では 1 ユーザに与えられる Window メモリの数で抑えられるので，DIMMnet-1 の場合は NIC を 1 ユーザで占有したとしても最大 64 にすぎない．ところが AOTF の場合は HTLB のミスヒットが起きない HTLB のエントリ数だけでも 1,024 あり，かなり大規模なクラスタでも十分な再利用性が確保できる．

3.3 AOTF にともなう CPU の TLB ミス

AOTF 送信では AOTF キックアドレスの仮想アドレスから物理アドレスへの変換に際してホスト CPU の TLB エントリを消費する．つまり AOTF によりアクセスされるリモートのページは CPU の TLB エントリ 1 つに対応しており，そのアクセスパターンによっては CPU の TLB ミスを増加させるケースを想定できる．あまり多くの PE にまたがる広い領域をあえて AOTF のみで飛び飛びにアクセスするような使い方である．PE 数が多いシステムで長いメッセージを全対全通信するような場合が最悪の状況となるが，DIMMnet-1 上では，CPU 内の TLB エントリを 1 つしか消費せず，長いのメッセージの送信にも適した BOTF を用いることで回避可能である．

しかし，通常，1 つの PE が送信相手とする PE 数はさほど多くなく，後述する実験のような AOTF が想定している典型的な利用状況では，あまり問題にならないケースが多いと考えられる．

3.4 リモートアドレスの物理アドレス表現

ヘッダ変換テーブルやそのキャッシュである HTLB はユーザモードからは直接は触れることのできない場所に配置されるので，AOTF 送信ではプロテクションをつかさどるプロセスグループ ID（PGID）やリモートアドレスの上位をユーザが勝手に書き替えたりできない．よって，リモート DMA（RDMA）送信[25]や BOTF 送信[26]と異なり，AOTF 送信はリモートアドレスを仮想アドレスだけでなく物理アドレスでも登録することができる．

4. AOTF 送信を支援する受信機構

4.1 OTF（On-The-Fly）受信機構

OTF 受信機構とは，BOTF や RDMA による送信の場合は必ず必要になるアドレス変換や DMA コントローラの起動をすることなしに，パケットヘッダの情報から所定の長さのデータ部を直接メモリに書き込む機構である．

AOTF 送信に限っては，上述のとおりリモートアドレスを物理アドレスでも登録することができるため，受信時のリモートにおけるアドレス変換のオーバヘッドを削除することが可能である．

DIMMnet-1 の OTF 受信部である Mini OTF 受信部では AOTF 送信に限って立てることができるヘッダ中のフラグを受信部が判定し，アドレス部と 1〜8 バイトのデータ部を書き込みバッファに押し込んでいく．書き込みバッファは後述する低遅延共有メモリ（LLCM：Low latency common memory）に，書き込めるタイミングで書き込む．

4.2 低遅延共有メモリ（LLCM）

低遅延共有メモリ（LLCM）とは，主に AOTF によりリモートノードから書き込まれたデータをホスト CPU から低遅延でポーリングするために用いられるオンチップのマルチポートメモリのことである．LLCM はオンチップであるゆえに小容量であるが，マルチポートであるゆえに，DIMMnet-1 の SO-DIMM による疑似的な 2 ポートメモリとは異なり，バンク切

替えを行うことなしに，パケットの送受信とホストからのアクセスを同時に行うことが可能である．

PCI バス型の NIC では，PCI バス上の資源をホストからポーリングしてしまうとそれだけで PCI バスのバンド幅を使いきってしまい，パケットの受信そのものを妨げてしまう．このため，受信用のバンド幅確保のために遅延時間を犠牲にして主記憶上の cacheable な領域までポーリングすべきデータをあらためて NIC から DMA 転送をかけ，DMA 転送時のキャッシュ無効化にともなうリフィル時の値の変化をホストから検出する．同一の Martin チップを用いる PCI 型 NIC である RHiNET-2/NI も LLCM は搭載しているが，PCI ベースゆえに上記のような動作をせざるをえない．

これに対し DIMMnet-1 上の LLCM へのポーリングは LLCM や SO-DIMM への受信をまったく妨げないので，バンド幅を犠牲にしたり，cacheable な領域に DMA される値へのポーリングを用いることによる遅延時間の増加を発生させたりすることなく，ホストからのポーリングを実行することが可能である．

ホストからのポーリングに適すると思われる LLCM 上のデータの例としては，以下に示すものがあげられる．

- ACK 付きの通信において送信側の LLCM に書かれるステータス
- ACK 付きの通信において受信側の LLCM 書かれるステータス☆
- AOTF や BOTF によってメッセージ本体を書き終わった後で，AOTF によって受信側の LLCM に書き込まれる受信完了フラグ
- 消費者側から AOTF によって生産者側の LLCM に書き込まれる消費者側リングバッファの先頭位置
- 生産者側から AOTF によって消費者側の LLCM に書き込まれる消費者側リングバッファの末尾位置
- バリア同期のために AOTF で受信側の LLCM に書き込まれる，フェーズを示すカウント値
- 大域演算のために AOTF で受信側の LLCM に書き込まれるデータ
- ソフトウェア分散共有メモリの実装における ACK
- マルチグレインタスク間の同期フラグまたはデータ

LLCM への AOTF によるアクセスを用いた処理の例として，消費者のリングバッファと生産者が 1 対 1 に対応しているケースのデータ受渡しのプロトコルを以下に示す．

生産者側

(1) 生産者側のメモリ上にある消費者側バッファの末尾位置を確認．
(2) 生産者側の LLCM 上にある消費者側バッファの先頭位置を確認．
(3) 上記に基づき空きがない場合は (2) へループ（ポーリング）．
(4) 空きがある場合は消費者のバッファの末尾位置に AOTF または BOTF でデータを書き込む．
(5) 生産者側のメモリ上にある消費者側バッファの末尾位置を更新し，AOTF により消費者側の LLCM 上にある消費者側バッファの末尾位置を更新．
(6) (1) へループ．

消費者側

(1) 消費者側のメモリ上にある消費者側バッファの先頭位置を確認．
(2) 消費者側の LLCM 上にある消費者側バッファの末尾位置を確認．
(3) 上記に基づきデータがない場合は (2) へループ（ポーリング）．
(4) 上記に基づきデータがある場合は消費者側バッファの先頭位置からデータを取り出す．
(5) 消費者側のメモリ上にある消費者側バッファの先頭位置を更新し，AOTF により生産者側の LLCM 上にある生産者側バッファの先頭位置を更新．
(6) (1) へループ．

5. 性能評価

本章では，AOTF 送信の低遅延性を DIMMnet-1 プロトタイプの実機を用いて評価する．DIMMnet-1 においては，AOTF 送信部→ Mini-OTF 受信部→ LLCM →ホストによる読み出しという経路で 1〜8 バイトをリモートライトするのが最も高速なホストへのデータの伝達方法である．よって，本章の実験においてはすべてこの通信経路を用いており，AOTF が最も効果的な状況での性能を評価する．

5.1 測定環境

以下の実験において用いた測定環境を**表 3** に示す．

DIMM や FSB がともに 133 MHz となる本来の設計値にはなっていないので，予定より低い性能が観測されるはずである．また，今回の実験で用いている通信リンクの動作モードでは，128 バイト以下のパケットにはパディングデータが付加される仕様になってい

☆ 今回の実装では省略された機能なので現時点では利用はできないが，仕様上は受信側ステータス格納アドレス付きのパケットフォーマットも定義されている．

表 3 DIMMnet-1 設定/測定環境
Table 3 Setting of DIMMnet-1 and experimental environment.

測定環境	A	B	C	D	E	F
基板種別	電気版 (e)		光版 (o2)			電気版 (e)
Link モード	OIP		RHiNET2			OIP
Link(MHz)	125		250			125
Link(MB/s)	250		500			250
CPU	Pentium3					Pentium4
コア (MHz)	850 MHz					1.5GHz
FSB(MHz)	100					400
DIMM(MHz)	66	100	66	100		100
MEMORY	256 MB(PC133)					
CHIPSET	VIA Pro133A					P4X266
LinuxKernel	2.4.2					
Compiler	egcs-2.91.66					

表 4 uncacheable 領域への 8 バイトアクセス時 CPU タイム
Table 4 CPU times for 8 bytes access to uncacheable area.

CPU	P3-850 MHz	P3-850 MHz	P4-1.5GHz	
FSB	100 MHz	100 MHz	400 MHz	
DIMM	66 MHz	100 MHz	100 MHz	
MMX	on	on	on	off
write	53 ns	53 ns	59 ns	54 ns
read	204 ns	173 ns	276 ns	469 ns

たり（光版：RHiNET2 スイッチモード時），DIMM 周波数より後述する SWIF 部の周波数が低い利用状況では頻繁にバブルが挿入される（62.5 MHz 以上の DIMM 上で電気版：OIP スイッチモードを使用時）．このため，遅延時間的には最適の状態にはなっていない．

今回の実験に用いた測定環境における uncacheable 領域へのアクセス時 CPU タイムの測定結果を**表 4** に示す．read 時には CPU タイムにチップセット遅延の往復分が折り込まれるが，write 時のチップセット遅延はプログラムでは正確には測定できない．その値はおおむね read と write の差の半分（60 ns）以下と考えられる．

5.2 ラウンドトリップ時間

DIMMnet-1 における AOTF 送信を用いた LLCM への通信によるラウンドトリップ時間（RTT）とその内訳を測定する．

5.2.1 ラウンドトリップ時間測定法

DIMMnet-1 においては，AOTF 送信部→ Mini-OTF 受信部→ LLCM（Maritini 内部の低遅延共有メモリ）→ホストによる読み出しという経路で 1〜8 バイトをリモートライトするのが最も高速なホストへのデータの伝達方法である．今回の測定では，この経路で 4 バイトを送信し，ホストにより LLCM をポーリングして値の変化を検知し，変化があった場合にそのデータを最初にリモートライトをかけてきたノードの LLCM にリモートライトして送り返す．時間測定は CPU 内の内部クロックに同期したカウンタを読むことにより行った．なお，カウンタを読む関数の実行時間自体は今回の測定環境では Pentium3 で 38 ns，Pentium4 で 53 ns かかる．ただし，コンテクストスイッチによる遅延増加はけた違いに多くなるので，多数回測定した際にけた違いに遅くなるものはコンテクストスイッチの影響を受けたと判断し，除外した．

5.2.2 周辺回路遅延測定法

Verilog による機能シミュレーション上では，NIC が搭載されるメモリスロット上に最初の信号が発生してから 14 クロック（133 MHz 動作時に 115 ns）で通信リンクインタフェースへの出力が始まる．しかし，DIMMnet-1 を用いた実際の測定環境では，異種クロックドメイン間同期化回路，シリアライザ・デシリアライザ，光インタフェースやケーブル等，上記の機能シミュレーションでは組み込まれていない遅延要因がいくつか存在する．

一方，Martini LSI にはデバッグ用に，SWIF という低速クロックドメインに属して光インタフェースに導かれる高速クロックドメインへの橋渡しをする回路ブロック内で自己ループをさせる機能を持っている．これによって高速系およびケーブルを使ったループによる遅延時間と，SWIF 間直結自己ループによる遅延時間を測定することにより，SWIF より外部の回路の遅延時間を測定できる．

5.2.3 測 定 結 果

DIMMnet-1 における AOTF 送信を用いた LLCM への通信による対向通信時ラウンドトリップ時間，高速系およびケーブルを使ったループ（外部ループ）による遅延時間と，SWIF 内直結自己ループ（内部ループ）による遅延時間，それらの差から得られた SWIF より外部の回路の遅延時間の測定結果を**表 5** に示す．

5.2.4 考 察

Verilog による AOTF 通信のシミュレーションにおける Martini の DIMM に同期動作する部分（送信側，受信側）および SWIF（送信側，受信側）の遅延を**表 6** に示す．本表における一番右の列が論理的な最短所用クロック数および本来の動作周波数（DIMM：133 MHz，SWIF：100 MHz）で動作した場合の遅延時間である．

上記の Verilog によるシミュレーションによる Martini 内部回路の遅延の合計は，内部ループラウンドトリップ時間より小さい．その差分は，Martini から CPU 側の外部で消費される時間である．その内訳は

表 5 AOTF 通信によるラウンドトリップ時間
Table 5 Roundtrip latency by AOTF sending.

測定環境	A	B	C	D	E	F
CPU	Pentium3-850 MHz				Pentium4-1.5GHz	
リンク	電気	電気	光	光	光	電気
SWIF	62.5 MHz	62.5 MHz	100 MHz	100 MHz	100 MHz	62.5 MHz
RTT 実測値（対向）	2,340 ns	1,940 ns	2,251 ns	1,840 ns	2,005 ns	2,122 ns
RTT 実測値（外部ループ）	1,026 ns	851 ns	1,091 ns	907 ns	882 ns	922 ns
RTT 実測値（内部ループ）	918 ns	705 ns	946 ns	756 ns	748 ns	796 ns
SWIF 外遅延	108 ns	146 ns	145 ns	151 ns	134 ns	126 ns

表 6 Verilog レベルで把握できている遅延時間
Table 6 Known latency by Verilog simulator.

測定環境	A	B, F	C	D, E	設計値（サイクル数）
DIMM	66 MHz	100 MHz	66 MHz	100 MHz	133 MHz
SWIF	62.5 MHz	62.5 MHz	100 MHz	100 MHz	100 MHz
ホストからの書き込み	45 ns	30 ns	45 ns	30 ns	21.5 ns (3)
トランザクションキュー処理	30 ns	20 ns	30 ns	20 ns	15 ns (2)
ヘッダ TLB 参照	45 ns	30 ns	45 ns	30 ns	21.5 ns (3)
転送サイズ判定	15 ns	10 ns	15 ns	10 ns	7.5 ns (1)
送信バッファハンドシェイク	15 ns	10 ns	15 ns	10 ns	7.5 ns (1)
送信側 SWIF での遅延	64 ns	64 ns	40 ns	40 ns	40 ns (4)
受信側 SWIF での遅延	144 ns	144 ns	70 ns	70 ns	70 ns (7)
LLCM への書込み	15 ns	10 ns	15 ns	10 ns	7.5 ns (1)
合計	343 ns	318 ns	245 ns	220 ns	192.5 ns

今回のピンポン通信による測定用ソフトウェア自体のオーバヘッドと，CPU が書き込み命令を実行してからチップセットのノースブリッジを経由して Martini LSI に至るまでの遅延（write 時 CPU タイムとチップセット遅延の合計），CPU が読み出し命令を実行してからチップセットのノースブリッジを経由して Martini LSI 内部の LLCM から読み出されるまでの遅延（read 時 CPU タイム），実際の受信から受信確認のポーリングまでのずれ（平均値はポーリング間隔の半分）からなると考えられる．

850 MHz の Pentium3 上で FSB100 MHz，DIMM 100 MHz，SWIF100 MHz の光版 DIMMnet-1 を内部ループバックおよび外部ループバックさせた場合の経路ごとの遅延時間内訳は**図 6** に示すようになる．

5.3 バリア同期時間

AOTF による LLCM へのリモートライトとホストからのポーリングを用いたバリア同期時間の測定を行う．

5.3.1 バリア同期の実現法

DIMMnet-1 においては，AOTF 送信部→ Mini-OTF 受信部→ LLCM（Maritini 内部の低遅延共有メモリ）→ホストによる読み出しという経路で 1〜8 バイトを送信するのが最も高速なホストへのデータの伝達方法である．Martini LSI に内蔵されるコア PU で LLCM にリモートライトされたデータをポーリング

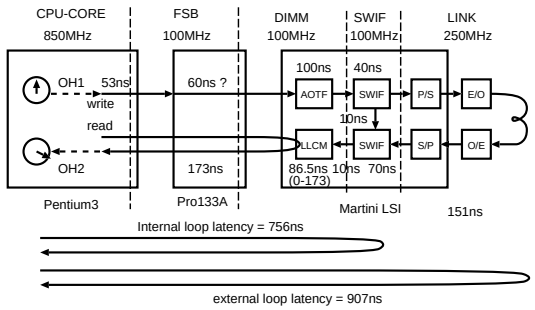


図 6 ループバック時の経路ごとの遅延時間内訳
Fig. 6 Map of latency for loopback test.

する方法も考えられる．しかし，今回の測定では**図 7** に示すように，上記の経路を用いてホストが介在する手法[24]によってバリア同期を実現した．その手順を以下に示す．

(1) 子ノードが AOTF で木構造の親にあたるノード側の LLCM 上にある 1 バイトのフラグをフェーズを示すカウント値で更新．
(2) 親ノードでは親ノード側の LLCM 上にある 8 バイトのフラグをリード．
(3) 上記 8 バイト中の同期に関連するバイト位置をマスクにより切り出す．
(4) 上記に基づき同期に関連するすべてのデータが更新されていない場合は (2) へループ（ポーリング）．

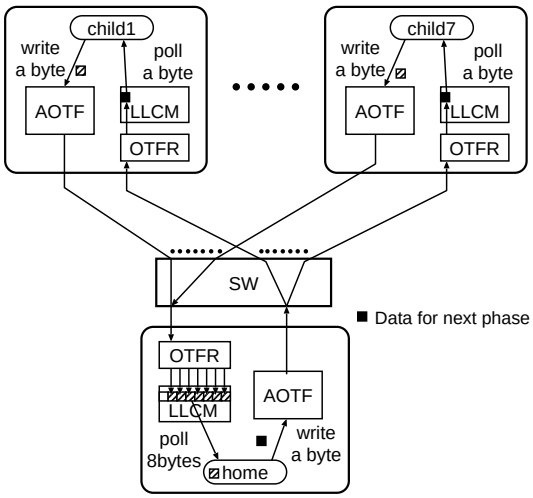


図 7 AOTF を用いた 8 ノードまでのバリア同期
Fig. 7 Barrier synchronization for 8 nodes with AOTF sending.

(5) 上記に基づき同期に関連するすべてのデータが更新された場合は，さらに上位の親がいる場合は (1) へループ．

(6) 最上位にある home ノードではフェーズを示すカウント値を進め，その値を AOTF によって同期に関連するすべての子ノードにマルチキャスト（スイッチのマルチキャスト機能を使うか，1 対 1 通信を繰り返し実行）．

(7) 子ノードでは子ノード側の LLCM 上にある 1 バイトのフラグをリード．

(8) 上記のフラグの値の変化がない場合は (6) へループ（ポーリング）．

なお，今回の実験では，スイッチが利用できなかったために，2 ノードでのバリア同期を対向通信環境によって実現した．また，MMX 命令を用いている場合ではデータ転送命令以外での MMX 命令を使用していない．

5.3.2 測 定 結 果

AOTF による LLCM へのリモートライトとホストからのポーリングを用いたバリア同期時間の測定の結果を**表 7** に示す．

5.3.3 考 察

今回測定されたバリア同期遅延時間は，同期専用ハードを追加することで実現されている SCC[28] の性能（1.6〜3.3 μs）に匹敵する性能を，不完全なチューニング状態にある DIMMnet-1 によりソフト的に実現できたことが示されている．マルチユーザ対応が困難な SCC に比べ，多くのユーザ数，同期グループ数に対応できる点でも本方式が優れている．

Pentium3 上では MMX 命令を用いた方が 8 バイトのリードを 1 回でできるために高速化している．これに対し Pentium4 上では表 4 に示されるようにリードそのものは MMX を使った方が高速であるにもかかわらず，emms 命令により MMX 命令使用を終了させた後に発生する原因不明のオーバヘッドが観測された．このため，MMX を使用した方がバリア同期時間も遅い．Pentium4 上での最適化における MMX 使用には注意を要すると思われる．

8 ノードまでのバリア同期は 1 回の 8 バイトリードによって判定できるので，今回の測定結果に対して，スイッチにおける 1 つの出力ポートへの 1〜7 個の 1 バイトのリモートライトパケットを出力する際の遅延時間（RHiNET2/SW の場合 1 個あたり約 240 ns）と，1 個のマルチキャストパケットのスイッチでの通過遅延時間（RHiNET2/SW の場合 1 個のパケットをスイッチがマルチキャストする機能がある）を加えたものが 8 ノードまでのバリア同期時間となると考えられる．8 ノードを超えるノード数 N の場合は 8 進木構造で対応することができ，その場合は上記に 8 進木の階層数 $log_8 N$ を乗じた時間でバリア同期がとれ，上記に $log_2 N$ を乗じた時間がかかる 2 ノードの同期を基本に Suffle Exchange 等で台数を増やす方式[27] よりも，台数が多くなっても遅延の増加は少ないと考えられる．

5.4 大域加算時間

AOTF による LLCM へのリモートライトとホストからのポーリングを用いた大域加算時間の測定を行う．

5.4.1 大域加算の実現法

前述のバリア同期の場合とほぼ同様に，各ノードに分散するデータの加算結果を全ノードに伝達する大域加算を実現することができる．バリア同期と異なり大域加算の場合は参加しているノードに対応するバイトのみを切り出すためのマスク操作が不要である．LLCM にリモートから書かれた所定の型のデータを自ノードのデータに加算して，他ノードの LLCM に結果を書き込み，受信側では LLCM に結果が書かれることをポーリングすることで終了を判定する．なお，MMX 命令を用いている場合ではデータ転送命令以外での MMX 命令を使用していない．

5.4.2 測 定 結 果

2 ノード対向環境における AOTF による LLCM へのリモートライトとホストからのポーリングを用いた大域加算時間の測定の結果を**表 8** に示す．

5.4.3 考 察

測定結果から分かるように，大域加算はおおむねバ

表 7 AOTF による 2 ノード対向環境でのバリア同期時間
Table 7 Barrier synchronization time for 2 nodes with AOTF sending.

測定環境	A	B	C	D	E	F
CPU	Pentium3-850 MHz				Pentium4-1.5GHz	
リンク	電気	電気	光	光	光	電気
バリア実測値（MMX あり）	2,375 ns	2,026 ns	2,255 ns	2,075 ns	2,616 ns	2,765 ns
バリア実測値（MMX なし）	2,569 ns	2,135 ns	2,435 ns	2,275 ns	2,283 ns	2,425 ns

表 8 AOTF による 2 ノード対向環境での大域加算時間
Table 8 Global sum operation time for 2 nodes with AOTF sending.

測定環境	A	B	C	D	E	F
大域加算実測値（MMX あり，unsigned）	2,379 ns	1,958 ns	2,288 ns	1,897 ns	2,101 ns	2,281 ns
大域加算実測値（MMX あり，ull）	2,798 ns	2,246 ns	2,677 ns	2,196 ns	2,651 ns	2,824 ns
大域加算実測値（MMX あり，float）	2,397 ns	1,975 ns	2,286 ns	1,903 ns	2,165 ns	2,303 ns
大域加算実測値（MMX あり，double）	2,377 ns	2,015 ns	2,286 ns	1,912 ns	2,163 ns	2,237 ns

リア同期と同等か，若干短時間で実行される．バリア同期におけるマスク操作や比較演算にかかる時間と，大域加算におけるデータの加算にかかる時間はおおむね同等で，1 バイトのリモートライトを行うのも，4〜8 バイトのリモートライトを行うのも，DIMMnet-1 ではあまり実行時間に差はないのでこのような結果が出たと思われる．

バリア同期の場合は浮動小数演算ではないために，NIC 上のプロセッサ等を用いた実装[29]も可能である．一方，大域加算の場合は浮動小数演算も高速に実行できる必要があるため，浮動小数演算が苦手な NIC 上のプロセッサ等を用いた実装は適さない．今回の実装はホスト上の CPU で実行させているので処理は加算だけでなく MPI で定義されているような種々の演算に対応は容易であり，それが 2μ 秒程度という短時間で実行できている意義は大きい．

6. 他の低遅延 NIC との違い

商用の NIC の中で低遅延なものの代表として，2μ 秒を切るリモートライト遅延時間を持つ Dolphin 社の PCI-SCI（D330）と COMPAQ 社の MEMORY CHANNEL-2 の 2 機種を取り上げ，DIMMnet-1 との違いを**表 9** に示す．

DIMMnet-1 では周波数の高さからくる高速化に加え，少ないクロック数でパケットにできるヘッダテンプレート（ヘッダシード）を連想する HTLB により，低遅延が実現されている．遅延やバンド幅といった基本的な性能指標や，性能に反映される周波数の高さや物量の豊富さの面だけでなく，プロテクションやマルチユーザの NIC 内滞在といった機能面でも DIMMnet-1 はこれらの製品を上回る特徴を備えている．

7. ま と め

試作された DIMMnet-1 プロトタイプについて紹介し，そのアーキテクチャを解説した．その実機上で測定された AOTF による細粒度通信性能に関して 8 バイトのラウンドトリップタイムやバリア同期および大域加算に関して報告した．レイアウトの不具合により規格外電源電圧で動作しているため不完全な状態ながら，きわめて優れた低遅延性を観測できていることが示された．また，その他の代表的な低遅延 NIC として PCI-SCI（D330）や MEMORY CHANNEL2 との違いについて明らかにした．

当初のターゲットであった PC クラスタとは異なるが，本プロトタイプで有効性が示された AOTF や BOTF は低遅延な通信機構として，並列計算機の専用ネットワークへのインタフェース部にも応用可能であると思われる．

今後は，ソフトウェア環境の整備を進め，アプリケーションによる評価を中心に，新バージョンの Martini LSI や RHiNET2/SW を用いた DIMMnet-1 の実機上での評価を進める予定である．

本論文の実験で示した範囲の使用形態では MPI の API でも十分に AOTF の低遅延性を利用することが可能である．DIMMnet-1 が提供している通信はリモートライトやリモートリードといった One sided 通信がベースとなっているので，MPI-2 や OpenMP との整合性も良いと思われる．ただし，今回の実装では SO-DIMM の領域はバンクを切り替えないとホスト CPU からはアクセスしにくいため，本論文で高速性を示したバリア同期等でバンク切替えのタイミングをつかみ，バンク切替えを適切に制御できる独自の API を併用することが必要と思われる．

表 9 代表的な低遅延 NIC と DIMMnet-1 の違い
Table 9 Difference between typical low latency NICs and DIMMnet-1.

NIC	Memory channel2 [5),6)]	PCI-SCI（D330）[2)~4)]	DIMMnet-1
リモートライト時間	1.76 μs [5)]	1.46 μs [3)]	270 ns☆
実測ラウンドトリップ時間	4.34 μs（8byte）[5)]	8.2 μs（0 B），12.0 μs（32 B）[4)]	1.84 μs（8byte）☆☆
実測バリア同期時間	不明	4.18 μs（2 ノード）[4)]	2.06 μs（2 ノード）☆☆
単方向通信継続バンド幅	100 MB/s [5)]	200 MB/s	1017 MB/s（BOTF）☆☆☆
双方向通信継続バンド幅	133 MB/s 以下	304 MB/s	2034 MB/s（BOTF）☆☆☆
ホスト I/F	PCI（32 bit，33 MHz）	PCI（64 bit，66 MHz）	SDR-DIMM（64 bit，133 MHz）
リンクバンド幅	133 MB/s×2	667 MB/s×2	1064 MB/s×2
送信手段	PIO のみ	PIO，RDMA	AOTF，BOTF，RDMA
パケットあたりペイロード長	4〜256 B（4 B 単位可変）	1 B，64 B，128 B 固定（63 B 以下の端数は 1 B 用パケットに分割）	AOTF: 1〜8 B（1 B 単位可変） BOTF: 1〜464 B （1 B 単位可変）
送信起動手法	store 命令	store 命令	store 命令
通信制御情報再利用手段	PCT（Page Control Table）	ATC（Address Translation Cache）と外部 SRAM 上の ATT（Address Translation Table）	HTLB（Header TLB）と外部 DRAM 上の Header 変換テーブル
再利用される情報	PCI-GLOBAL アドレス対応関係と属性フラグ	PCI-SCI アドレス対応関係と属性フラグ	汎用かつ短時間でパケット化可能なヘッダのテンプレート（32 B）
連想手法	直接アドレシング	ダイレクトマッピング	4way セットアソシアティブ
キューイングできる送信要求数	不明	32	AOTF: 2,048，BOTF: 64
NIC 内共存可能ユーザ数	不明	1（DMA 用の制御状態レジスタが多重化されていないため）	64
送受信両側の対応付け	送信前に両側の PCT を設定する必要あり	送信前に送信側の ATT におけるソースノード ID を受信側のテーブル（256 エントリ）上に設定する必要がある．	事前の一致は不要（受信側 TLB でミスヒットが起こればリフィルされる）
受信側でのプロテクション	アドレスに該当する PCT エントリの存在を検査	ソースノード ID を検査後，アドレスの上下限を検査	アドレス変換スキップフラグを検査後，必要に応じてプロセスグループ ID とプロセス ID と領域 ID とアドレスを TLB で検査

☆ 表 6 の 133 MHz 動作の Martini 内遅延に AOTF キック実行から DIMM 上に信号が現れるまでの時間（チップセットに依存するうえ，ソフト的には正確に測定できない部分であるため 10 クロックサイクルを仮定）を加算したものである．
☆☆ 通常より低い周波数（100 MHz）動作の DIMM 上で，かつ使用したマザーボード向けのデータ線ねじれ解消のソフトウェアオーバヘッドを含む値である．
☆☆☆ 133 MHz 動作 DIMM への CPU からのコピー動作の実測バンド幅からの 133 MHz 動作時の推定値である．

本論文で明らかになったように DIMMnet-1 はきわめて優れた低遅延性を有することから，リモートライトのみを行える Memory Channel 上にソフト的に作られるコヒーレントな細粒度分散共有メモリをベースにした shasta と同様なトランスレータが DIMMnet-1 においても有効と考えられ，今後開発される予定である．

謝辞 （株）日立製作所の西氏，東京農工大学の須田氏，三橋氏，慶應義塾大学の土屋氏，渡辺氏，（株）日立 IT の上嶋氏，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 の開発に携わったすべての方々に感謝いたします．なお，本研究は新情報処理開発機構が推進した 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行われたものである．

参 考 文 献

1) Myricom Corp. http://www.myri.com/
2) Dolphin Corp.: PCI-SCI Adapter Card D320/D321 Functional Overview Part No.D1950–10299 (1999.11).
3) Dolphin Corp.: PCI-64/66 - PCI-SCI Adapter Card for System Area Networks. http://www.dolphinics.com/products/hardware/pci64.html
4) Scali Computer Corp.: ScaBench — Scali's MPI Benchmark Suite. http://www.scali.com/performance/ssp212/scabench.html
5) Fillo and Gillett: Architecture and Imple-

mentation of MEMORY CHANNEL 2, *Digital Technical Journal*, Vol.9, No.1, (1997).
6) Compaq Corp.: MEMORY CHANNEL 技術概要, OpenVMS Cluster 構成ガイド, pp.333–347.
7) InfiniBand Trade Association, available from http://www.infinibandta.org/
8) 住元, 堀, 手塚, 原田, 高橋, 石川：GigaE PM II：Gigabit Ethernet による高速通信ライブラリの設計, 情報処理学会計算機アーキテクチャ研究会, Vol.99, No.67, pp.61–66 (1999. 8).
9) 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 (1999.8).
10) 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 (2000.3).
11) 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, Vol.2001, No.77, pp.99–104 (2001.7).
12) 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 (2000.11).
13) 山本, 渡邊, 土屋, 今城, 寺川, 西, 田邊, 工藤, 天野：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, Vol.2001, No.76, pp.37–42 (2001.7).
14) 田邊, 中村, 鈴岡, 小柳：並列 AI マシン Prodigy の試作と通信性能評価, 電子情報通信学会論文誌, Vol.J74-D-I, No.4, pp.264–272 (1991.4).
15) 田邊：マルチプロセッサシステム, 公開特許公報, 特願平 2-157491（出願 1990.6）, 特開平 4-48371（公開 1992.2）.
16) 鈴木, 田邊, 菅野, 小柳：超並列 Teraflops マシン TS1—分散共有メモリアーキテクチャ, 情報処理学会第 48 回全国大会, 4B-4 (1994).
17) Blumrich, Li, Alpert, Dubnicki, Felten and Sandberg: Virtual Memory Mapped Network Interface for the SHRIMP Multicomputer, *ISCA'94*, pp.142–153 (1994.4).
18) 松本, 平木：超並列計算機上の共有メモリアーキテクチャ, 電子情報通信学会コンピュータシステム研究会 CPSY92-26, pp.47–55 (1992).
19) 五島, 斎藤, 小西, 秤谷, 森, 富田：並列計算機 JUMP-1 の分散共有メモリ・システム, 情報処理学会論文誌, No.SIG8(HPS 2), pp.15–27 (2000.11).
20) 日本電子機械工業会：日本電子機械工業会規格：プロセッサ搭載メモリ・モジュール（PEMM）動作仕様標準, EIAJ ED-5514 (1998.7).
21) 西, 多昌, 西村, 山本, 工藤, 天野：LASN 用 8 Gbps/port 8x8 One-chip スイッチ：RHiNET-2/SW, 2000 年記念並列処理シンポジウム（JSPP2000）, pp.173–180 (2000.5).
22) 西, 上野, 多昌, 稲沢, 西村, 工藤, 天野：LASN 用 10 Gbps/port 8x8 ネットワークスイッチ：RHiNET-3/SW, 情報処理学会計算機アーキテクチャ研究会, Vol.2000, No.110, pp.13-18 (2000.11).
23) Yoshikawa and Matsuoka: Optical Interconnections for Parallel and Distributed Computing, *Proc. IEEE*, Vol.88, No.6, pp.849–855 (2000.6).
24) Tanabe, Hamada, Yamamoto, Kudoh, Imashiro, Nakajo and 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*) (2001.8).
25) 山本, 渡辺, 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 (2002.5).
26) 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 (2002).
27) 田中, 久保田, 佐藤, 関口：並列アルゴリズムにおける Collective 通信の性能比較, 情報処理学会研究報告, 96-HPC-62, pp.19–26 (1996.8).
28) 早川, 関口, 岩根：Beowulf クラスタにおける高精度実行時間測定の検討と評価, 情報処理学会 HPC 研究会, Vol.2001, No.77, pp.111–116 (2001.7).
29) Buntinas, D., et al.: Performance Benefits of NIC-Based Barrier on Myrinet/GM, *Proc. Workshop on Communication Architecture for Clusters* (*CAC*) *with IPDPS'01* (2001).
30) Scales, Gharachorloo and Thekkath: Shasta: A Low Overhead, Software-Only Approach for Supporting Fine-Grain Shared Memory, *ASPLOS'96* (1996.10).

付録：試作から判明した問題と改良方針

A.1 チップセット相性問題

DIMM スロットに対して供給されるアドレスは, 物理アドレスを ROW アドレスと COLUMN アドレスの 2 サイクルにマルチプレクスされて来るが, そのマルチプレクス規則がチップセットのノースブリッジに

よって異なる．よって Martini LSI はその仕様を入手することができた少数のチップセットに対応した配線を選択するロジックを備えており，これによって CPU が発生した物理アドレスを DIMM 上の信号から復元して用いている．しかし，チップセットのアドレスマルチプレクス仕様は必ずしも一般に公開されているとは限らないため，事前に用意した配線を選択する方式だと，今後，新たな DIMMnet 用 LSI を設計する場合にチップセット仕様がその時点であきらかにされていない場合は対応できない．この部分が後日リリースされてくる新しいチップセットを用いたマザーボードへの適用性を阻害している．

A.2 データライン相性問題

実際にマザーボードを入手し，動作を確認し始めた頃には，CPU からの書き込みデータが Martini LSI に正しく伝わらないという現象が観測された．それは，マザーボード上での DIMM スロットとノースブリッジの間のデータ線の配線に関する規定や規格が存在しないために，両者の n bit 目のデータ線どうしが必ずしもストレートに接続されていないために起こった現象であることが判明した．そのため，必要があればソフトウェアで事前にデータを各マザーボード対応した規則でねじってから DIMMnet-1 に書き込む必要があり，本報告で用いられた 2 種類の PC ではそのようなソフト的な対応をして動作させている．このソフトウェアオーバヘッドのため，マザーボードによっては DIMMnet-1 の実行性能は低下が発生する．なお，ソフト対応が必要ないデータ線がストレート接続されたマザーボードも存在することが分かっており，チップセットメーカからマザーボードメーカに出される実装ガイドラインの中で，データ線のストレート接続を推奨していただくことが今後望まれる．

A.3 ハード的改良の指針

パソコンの部品の進化のスピードは目覚ましく，その恩恵を享受するにはチップセットの変更や，マザーボードの変更にも柔軟に対応できるような作りにすることが望まれる．さらには，チップセットやマザーボードへの柔軟な適応だけでなく，スイッチについても市販のものを用いることができればより望ましい．

試作された Martini LSI は最も普及している Unbufferd 型 DIMM のみの対応だったために，タイミング的な余裕が厳しいことが予想された関係上，チップセットごとのアドレスマルチプレクス規則やマザーボード上でのデータ線ねじれへの柔軟な対応が可能な選択的配線の実装は採用しなかった．しかし，タイミング的余裕が増加する Registered 型に対応するか，Unbufferd 型向けでも高速な FPGA で実装できるのであれば，そのような可変構造が埋め込める可能性がある．

Rambus 型の場合は，アドレスのマルチプレクス規則がパケットのフォーマットとして規定されている．マザーボード上のデータ線ねじれの問題についても同様のことがいえ，論理的な意味での相性問題については Rambus 化は 1 つの有望な解になっていると思われる．

スイッチについては，現状の Martini LSI を用いた DIMMnet-1 ではリンクインタフェースが表 2 のものにしか対応していないために，同等クラスのバンド幅を持つものとして市販されつつある Infiniband や 10Gbit Ethernet 用スイッチ等には，現状では直接接続はできない．

本プロトタイプに実装された現状の AOTF 送信部では，ヘッダ長が 32 バイトまでという制約や，ヘッダへのリモートアドレス等をはめ込む位置が固定になっていることにともなう制約がある．BOTF 送信部もプロテクション刻印を行う位置が固定となっている．さらに誤り検査符号の仕様が Ethernet とは一致しない．しかし，これらの変更を行うことは軽微な修正で済み，原理的には困難ではない．外部の変換回路によって対応することも可能である．

特に，AOTF や BOTF といった通信機構の特徴は，パケットフォーマットについてはソフト的に書き替えられる柔軟性を有しているため，変更箇所は RDMA の場合よりも大幅に少なくて済む．

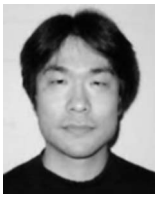
以上のように，外部の変換回路を作成するか，軽微な論理変更と対応した物理層回路を作成することにより，将来的には市販のスイッチと組み合わせて利用できる実装形態は実現可能と思われる．

ただし，AOTF の低遅延性が真に発揮できるのは，TCP 層のようなソフトウェアが必須とならない場合であり，パケットが破棄されてしまうようなスイッチには適さない．そのようなスイッチにも対応可能にするためには，今回試作した Martini LSI で省略された機能である NIC における end-to-end なハードウェア式再送機構を実装する必要がある．

(平成 14 年 6 月 7 日受付)
(平成 14 年 10 月 21 日採録)

田邊　昇（正会員）

1985 年横浜国立大学工学部卒業．1987 年横浜国立大学大学院工学研究科修了．同年（株）東芝に入社．1998 年より 2001 年まで新情報処理開発機構つくば研究センターに出向．並列処理，並列アーキテクチャに関する研究に従事．現在,（株）東芝・研究開発センター勤務．博士（工学）．電子情報通信学会会員．



濱田　芳博

2001 年東京農工大学工学部卒業．現在，東京農工大学大学院工学研究科（前期課程）在学中．電子情報工学専攻．

山本　淳二（正会員）

1991 年慶応義塾大学理工学部卒業．1997 年慶応義塾大学大学院理工学研究科博士課程単位取得退学．同年新情報処理開発機構入社．2002 年より（株）日立製作所・研究開発本部に勤務．並列処理・ネットワークに関する研究に従事．博士（工学）．

今城　英樹（正会員）

1989 年 釧路工業高等専門学校卒業．同年（株）日立コンピュータエレクトロニクス入社．以来，大形計算機のハードウェア開発に従事．現在,（株）日立インフォメーションテクノロジーにて各種 ASIC 開発のコンサルティング業務に従事．

中條　拓伯（正会員）

1961 年生まれ．1987 年神戸大学大学院工学研究科電子工学専攻修了．1989 年神戸大学工学部情報知能工学科助手を経て，現在，東京農工大学工学部情報コミュニケーション工学科助教授．1998 年より 1 年間イリノイ大学スーパコンピューティング研究開発センター（CSRD）にて客員助教授．プロセッサアーキテクチャ，分散共有メモリ，クラスタコンピューティングに関する研究に従事．電子情報通信学会，IEEE-CS 各会員．博士（工学）．

工藤　知宏（正会員）

1991 年慶應義塾大学大学院理工学研究科博士課程単位取得退学．東京工科大学講師，助教授を経て，1997 年より新情報処理開発機構並列分散システムアーキテクチャつくば研究室長，2002 年より産業技術総合研究所．博士（工学）．

天野　英晴（正会員）

1986 年慶應義塾大学大学院理工学研究科修了．工学博士．現在，同大学情報工学科教授．計算機アーキテクチャの研究に従事．