計算機アーキテクチャ 145－8
（2001. 11. 29）

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

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

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という低遅延通信機構とBOTF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100MHz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．

Performance evaluation of communication on DIMMnet-1 network interface plugged into a DIMM slot

NOBORU TANABE,[†1,☆] YOSHIHIRO HAMADA,[†2] HITOSHI SUDA,[†2] JUNJI YAMAMOTO,[†1]
HIDEKI IMASHIRO,[†3] HIRONORI NAKAJO,[†2] TOMOHIRO KUDOH[†1] and HIDEHARU AMANO [†4]

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 plagged into 100MHz 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なういわゆる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)3)]，MEMORY CHANNEL2[4)5)]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バスではバンド幅および遅延ともに力不足である．

一方，Infiniband[6]が次世代のサーバー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ー用の量産PCに，Infiniband が普及するかどうか不透明である．GigaE PM2[7]を用いるなどして全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[8]と名付けた．MEMOnetは安価なPC上で，PCIバスのバンド幅や遅延時間の限界を超越したNICを実現可能とすると思われる．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[9)10)]を開発した．

†1 新情報処理開発機構
Real World Computing Partnership
☆ 現在，(株)東芝，研究開発センター
Presently with Corporate Research and Developpment Center, Toshiba
†2 東京農工大学
Tokyo University of Agriculture and Technology
†3 (株)日立インフォメーションテクノロジー
Hitachi Information Technology
†4 慶應義塾大学
Keio University

このDIMMnet-1や，同一のMartini LSI[11]を用いたPCI版NICであるRHiNET2/NI[12]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[13]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[14]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[15]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[16]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[17]がある．MBPは，多機能なメモリー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ではメモリーベースト通信というMBPと共通のアプローチを取りつつも，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に至る経路全体に渡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，凄まじい高速化を遂げるパソコンの高い性能の有効利用を実現している．

本報告では，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 制御ボートを有する．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」[19] 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）が無くても動作するモードの二つのモードを有する．

表 1 DIMMnet-1 の主な仕様

ホストとのインタフェース	DIMM および PEMM
NIC メモリ（共有バンク）	PC133,SO-DIMM2 枚
搭載可能 SO-DIMM 容量	64MB～1GB
低遅延共有メモリ (LLCM) 容量	128KB(オンチップ)
命令 SRAM 容量	128KB(オンチップ)
データ SRAM 容量	128KB(オンチップ)
オンチップ CPU	R3000 風 32bitRISC
通信リンクバンド幅	o2：各方向 8Gbps（全二重） o3:各方向 10Gbps（全二重） e(OIP):各方向 2.5Gbps（全二重） e(RN2):各方向 8Gbps（全二重）
NIC メモリバンド幅	1024MB/s(ホスト側) 1024MB/s(network 側)
最短送信時 NIC 遅延時間	105ns(DIMM～リンク)
最短受信時 NIC 遅延時間	90ns(リンク～LLCM)
NIC-LSI のテクノロジ	0.14 μm CMOS

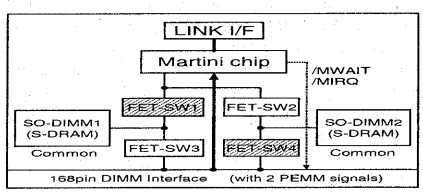

図 1 DIMMnet-1 の基本構造

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 同士が接続可能である．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 66MHz および 100MHz 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 の電気版の二種類が開発され，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．

表 2 DIMMnet-1 に接続可能なスイッチの仕様

スイッチ	RHiNET-2[20]	RHiNET-3[21]	OIP-SW[22]
光 port	8(or 0)	8	15
電気 port	0(or 8)	0	1
I/O ピン	800Mbps×9	1250Mbps×8	250Mbps×9
バンド幅	8Gbps	10Gbps	2.5Gbps
距離 (光)	100m	1km	100m
距離 (電気)	5m	-	5m
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

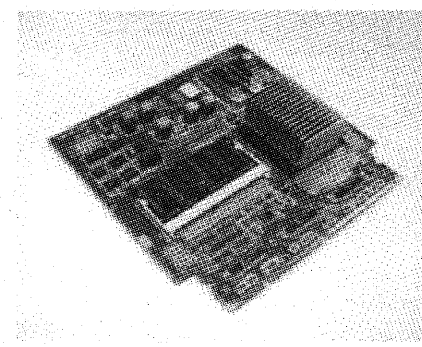

図 2 DIMMnet-1/e

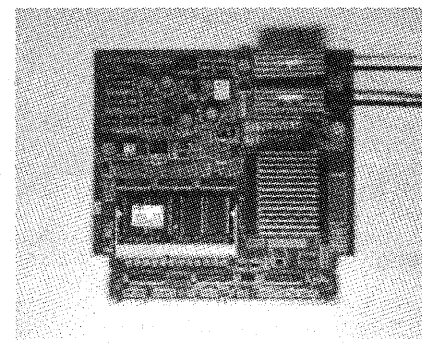

図 3 DIMMnet-1/o2

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，ロックやバリア，同期通信など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．

なお，Martini LSI は 2 つのバージョンが開発されており，最初のバージョンでは動作周波数チューニングやデバッグが不十分であったり，レイアウト上の問題があり電源電圧を規定より落とさないと使えないため予定された周波数での動作ができず，さらにいくつかの機能が省略されている．本報告における実験

で用いられているのは最初のバージョンの Martini LSI である．

なお，最初のバージョンにおいても，規定の電源電圧における実験では SO-DIMM へのアクセスや DIMM 型 NIC 制御ボートを介した内部資源へのアクセスは 133MHz での動作が確認されており，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現時点でも確認できている．

2.4 対応するチップセット

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．Martini LSI は特定のチップセットに対応でき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．第一バージョンの Martini が対応できることが判っているチップセットには Intel 社の BX,i810,i815，VIA technology 社の Pro133,Pro266,KT133,KT266,P4X266,P4M266 がある．第二バージョンの Martini はこれに加え，Intel 社の i845 も対応する．なお，上記には DDR-SDRAM に対応しているチップセットも含まれるが，Martini は SDR-SDRAM にしか対応していないので，それらのチップセットを用いている場合でも，SDR-SDRAM 型の DIMM スロットを持つマザーボードに限り使用が可能である．

2.5 判明した問題点

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 目のデータ線同士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ーヘッドのため，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，ソフト対応が必要のないデータ線のストレート接続がされたマザーボードも存在することが判っており，チップセットメーカーからマザーボードメーカー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．

3. Atomic オンザフライ (AOTF) 送信

Atomic On the fly(AOTF) 送信は，ヘッダー TLB(HTLB) を用いることにより，メモリバス上の一つの書き込みアクセストランザクションによって起動される低オーバー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ー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．

AOTF 送信機能は最初のバージョンの Martini LSI にも搭載されており，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．

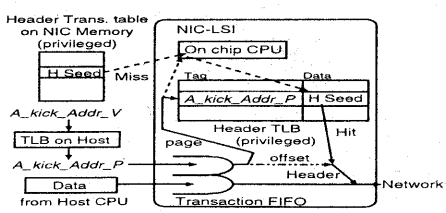

図 4 Atomic オンザフライ (AOTF) 送信

3.1 ヘッダ TLB

AOTF 送信はヘッダー TLB(HTLB) により実現される．HTLB は物理アドレスからヘッダーシードを連想し，パケットを生成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．

ヘッダーシードとは送信すべきパケットのヘッダーから，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ー変換テーブルや，そのキャッシュである HTLB はユーザー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 ではこの HTLB とヘッダー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．

パケットは起動に用いたアドレス（AOTF キックアドレス）の下位 bit（DIMMnet-1 の場合 12bit）のオフセットをヘッダーシードのリモートアドレスフィールドに上書きしてヘッダーを完成させ，起動時に書き込まれた 1〜8 バイトのデータを添付することで生成される．

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．ヘッダーシードは 64 ビット幅 4096 語構成のオンチップメモリに記憶される．

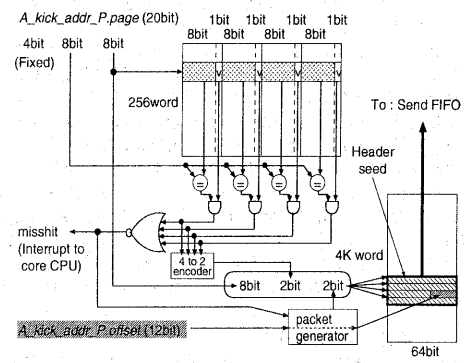

図 5 ヘッダ TLB(HTLB) の構造

3.2 リモートアドレスの物理アドレス表現

ヘッダー変換テーブルやそのキャッシュである HTLB はユーザー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ーが勝手に書き換えたりできない．よって，この AOTF 送信に限り，リモートアドレスを物理アドレスで登録することができ，受信時のリモートにおけるアドレス変換のオーバーヘッドを削除することが可能である．

4. 性能評価

4.1 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．

表 3 DIMMnet-1 設定/測定環境

測定環境	(A)	(B)	(C)	(D)	(E)
DIMMnet-1 種別	電気版		光版		
リンク周波数 (MHz)	125		250		
リンクバンド幅 (MB/s)	250		500		
CPU	P3 850MHz				P4 1.5GHz
FSB(MHz)	100				400
DIMM 周波数 (MHz)	66	100	66	100	100
MEMORY	256MB(PC133)				
CHIPSET	VIA Pro133A				P4X266
LinuxKernel	2.4.2				
Compiler	egcs-2.91.66				

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．その値は概ね read と write の差の半分以下と考えられる．

DIMM や FSB がともに 133MHz 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．

表 4 uncacheable 領域への 8 バイトアクセス時 CPU タイム

CPU	P3-850MHz	P3-850MHz	P4-1.5GHz	
FSB	100MHz	100MHz	400MHz	
DIMM	66MHz	100MHz	100MHz	
MMX	on	on	on	off
write	71ns	71ns	53ns	53ns
read	184ns	141ns	269ns	472ns

4.2 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とその内訳を測定する．

4.2.1 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 38ns,Pentium4 で 53ns 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．

4.2.2 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 (133MHz 動作時に 115ns) 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．

4.2.3 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．

4.2.4 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．

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ーヘッドと，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．

850MHz の Pentium3 上で FSB100MHz，DIMM100MHz，SWIF100MHz 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．

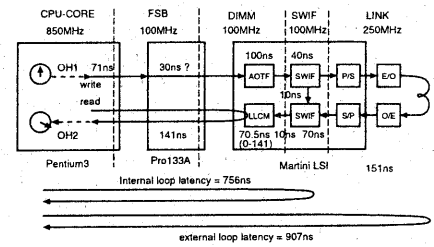

図 6 ループバック時の経路ごとの遅延時間内訳

上記の内訳から推定すると，最初にタイマーカウンタを読みに行ってから AOTF キックの write アクセス命令実行までのソフトオーバーヘッド (OH1) と，LLCM から読まれた値が CPU のレジスタに入ってから再びタイマーカウンタを読みに行くまでのソフトオーバーヘッド (OH2) の合計は 214ns 程度ということになる．

4.3 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．

4.3.1 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，ホストが LLCM 上の 8 バイトをポーリングして，バリア同期完了を意味する値になっていることをホスト上で判定し，同期完了検出時に同期に参加したノードの LLCM にインクリメントした同期変数 1 バイトを AOTF によりリモートライトすることによって同期完了を伝達する手法 [23] によってバリア同期を実現した．柔軟な同期参加ノード数に対応可能とするマスク処理も実装した上で測定している．

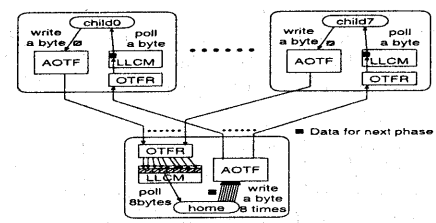

図 7 AOTF を用いた 8 ノードまでのバリア同期

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．

4.3.2 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．

表 5 AOTF 通信によるラウンドトリップ時間

測定環境	(A)	(B)	(C)	(D)	(E)
SWIF	62.5MHz	62.5MHz	100MHz	100MHz	100MHz
RTT 実測値 (対向)	2340ns	1940ns	2251ns	1840ns	2042ns
RTT 実測値 (外部ループ)	1026ns	851ns	1091ns	907ns	885ns
RTT 実測値 (内部ループ)	918ns	705ns	946ns	756ns	755ns
SWIF 外遅延	108ns	146ns	145ns	151ns	130ns

表 6 Verilog レベルで把握できている遅延時間

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

表 7 AOTF による 2 ノード対向環境でのバリア同期時間

測定環境	(A)	(B)	(C)	(D)	(E)
バリア実測値 (MMX あり)	2375ns	2026ns	2255ns	2075ns	2616ns
バリア実測値 (MMX なし)	2569ns	2135ns	2435ns	2275ns	2283ns

4.3.3 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[25] の性能 (1.6〜3.3μs) 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が，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，MMX 命令使用後に発生する謎のオーバー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 240ns) 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 ノードを越えるノード数 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[24] 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. 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，2 μ 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 の二機種を取り上げ，DIMMnet-1 との違いを表 8 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ーテンプレート (ヘッダーシード) 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．

6. 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に関して報告した．レイアウトの不具合による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，極めて高い性能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．

DIMMnet-1 ではメモリーベースト通信という MBP と共通のアプローチを取りつつも，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 に至る経路全体に渡り，肝心な部分のみのハード化の徹底を行った．こうして，DIMMnet-1 では大幅な低遅延通信と，凄まじい高速化を遂げるパソコンの高い性能の有効利用を実現している．

今後は，バンド幅の評価を中心に，第二版の Martini LSI や RHiNET2/SW を用いた DIMMnet-1 の実機上での評価と，ソフトウェア環境の整備を進める予定である．さらに，DIMMnet-1 における高速な細粒度通信性能を活かすと思われる Shasta[26] のようにある程度通信の粒度を細かいところに最適化したコンパイラの開発が望まれる．

謝辞 新情報処理開発機構の西氏，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，(株) 日立 DDC 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. available from http://www.myri.com/
2) Dolphin Corp. : PCI-SCI Adapter Card D320/D321 Functional Overview Part no.:D1950-10299, available from http://www.dolphinics.com/pdf_filer/PCI_SCI_Overview.pdf (1999.11)
3) Dolphin Corp. : SCI SISCI Performance Test Results, available from http://www.dolphinics.com/ scibenchmarks/testresults.html

表 8 代表的な低遅延 NIC と DIMMnet-1 の違い

NIC	MEMORY CHANNEL 2[4)5)]	PCI-SCI (D330)[2)3)]	DIMMnet-1
リモートライト時間	1.76 μs	1.46 μs	270ns
単方向通信継続バンド幅	100MB/s	200MB/s	1017MB/s(BOTF)
双方向通信継続バンド幅	133MB/s 以下	304MB/s	2034MB/s(BOTF)
ホスト I/F	PCI(32bit,33MHz)	PCI(64bit,66MHz)	SDR-DIMM(64bit,133MHz)
リンクバンド幅	133MB/s×2	667MB/s×2	1064MB/s×2
送信手段	PIO のみ	PIO, RDMA	AOTF, BOTF, RDMA
パケット当りペイロード長	4〜256B (4B 単位可変)	1B,64B,128B 固定 (63B 以下の端数は 1B 用パケットに分割)	AOTF:1〜8B (1B 単位可変) BOTF:1〜464B (1B 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ーのテンプレート (32B)
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:2048, 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, 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

4) Fillo and Gillett : Architecture and Implementation of MEMORY CHANNEL 2 , *Digital Technical Journal*, Vol.9(1) (1997)

5) Compaq Corp. : MEMORY CHANNEL 技術概要, OpenVMS Cluster 構成ガイド, pp.333-347, available from http://digital.compaq.co.jp/openvms/document/jv73/pdf/jopenvms_073_clus_conf.pdf

6) InfiniBand Trade Association, available from http://www.infinibandta.org/

7) 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)

8) 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)

9) 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)

10) 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, 情報処理学学会 HPC 研究会, Vol.2001, No.77, pp.99-104 (2001.7)

11) 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)

12) 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)

13) 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)

14) 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)

15) 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)

16) Blumrich, Li, Alpert, Dubnicki, Felten and Sandberg : Virtual Memory Mapped Network Interface for the SHRIMP Multicomputer, *ISCA'94*, pp.142-153 (1994.4)

17) 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)

18) 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)

19) 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)

20) 西, 多昌, 西村, 山本, 工藤, 天野 : LASN 用 8Gbps/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)

21) 西, 上野, 多昌, 稲沢, 西村, 工藤, 天野 : LASN 用 10Gbps/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)

22) Yoshikawa, Matsuoka : Optical Interconnections for Parallel and Distributed Computing, *Proceedings of the IEEE*, Vol. 88, No. 6, 2000 pp.849-855, (2000.6)

23) 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)*, available from http://www.ssgrr.it/en/ssgrr2001/papers/Noboru%20Tanabe.pdf (2001.8)

24) 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)

25) 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, 情報処理学学会 HPC 研究会, Vol.2001, No.77, pp.111-116 (2001.7)

26) Scales, Gharachorloo and Thekkath : Shasta: A Low Overhead, Software-Only Approach for Supporting Fine-Grain Shared Memory, *ASPLOS'96* (1996.10)