計算機アーキテクチャ 149－17
（２００２．８．２３）

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

田邊 昇[†1] 濱田芳博[†2] 三橋彰浩[†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 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 によって作成された光リンク版 DIMMnet-1 は 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 AOTF 送信機構と OTF 受信機構によるバリア同期や大域加算の遅延時間と，BOTF 送信機構や RDMA 送信機構を用いたバンド幅の DIMMnet-1 実機上での測定結果を報告する．

Performance Evaluation of Bandwidth and Global Operations on DIMMnet-1 Prototype

Noboru Tanabe,[†1] Yoshihiro Hamada,[†2] Akihiro Mitsuhashi,[†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-1 that can be directly plugged into DIMM slot of PCs is presented. It has a low latency AOTF (Atomic On-The-Fly) sending mechanism and a high bandwidth BOTF (Block On-The-Fly) sending mechanism. Now, DIMMnet-1 prototype boards providing optical network interface consisting with a network interface controller chip Martini is available. They can be plagged into 100MHz DIMM slot of PCs. Experimental evaluation of barrier synchronization and global sum latency with AOTF sending and OTF receiver on a real system are shown. The bandwidth with BOTF sending and RDMA on it is also presented.

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]，SCI-PCI[2] ，QsNET[3)4)5)] 等の高速ネットワークインタフェース (NIC) が各種開発されている．さらに 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，最近では 1X(2.5Gbps)[7)8)] や 4X(10Gbps)[9)10)] の製品が開発されてきた．しかし，これらは一部のサーバーマシンにしか搭載されていない 64bit66HMzPCI バスや PCI-X バス [11)] に接続されるため，最も価格性能比面で有利なエンドユーザ用の量産 PC において，本来の性能を発揮できる環境が提供されているとは言い難い．

CPU と入出力バスの進歩のスピードの差を考慮すれば，CPU と入出力バスの能力の差が今後開いていくことも予想される．ゆえに，処理能力と通信能力のバランスが要求される並列処理応用においてはとりわけ，CPU とともに性能向上の見込める通信手段の確保が望まれる．

光インターコネクションの持つ大きなバンド幅を鑑みる時，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，入出力バスに搭載される NIC とは別のアプローチも検討に値する．

このような問題意識にたち我々は，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ネットワークインタフェースを検討してきた．このようなクラスのネットワークインタフェースを MEMOnet[12)] 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ネットワークインタフェース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 を開発した．

メモリスロットは CPU との間でバンド幅的にも遅延時間的にも入出力バスより緊密に結合される．このような位置的なメリットを，成長著しい高性能 CPU と組み合わせて PC クラスタを構築する場合に最大限引き出すために，我々は Atomic on-the-fly(AOTF) および Block on-the-fly(BOTF) という送信機構および On-the-fly(OTF) 受信機構を提案してきた．DIMMnet-1 にはこれらが実装されている．

一方，現在の主流である通信方式は SCI-PCI[2] , QsNET[3)4)5)], Infiniband[6]，RHiNET[13)] に実装されている Remote DMA (RDMA) である．RDMA のハード化を中核アーキテクチャとしていた RHiNET のコントローラ ASIC 上に同居する形で実装された経緯から，DIMMnet-1 には RDMA も合わせて実装されている．よって，DIMMnet-1 上での各種の通信機構の性能比較は，今後の主流となるべき通信方式を考える上で意義がある．

本報告は以下のように構成される．第 2 章でアーキテクチャの概要を紹介する．第 3 章で DIMMnet-1 プロトタイプの概要を紹介する．第 4 章で AOTF 送信機構と OTF 受信機構によるバリア同期や大域加算の遅延時間と，BOTF 送信機構や RDMA 送信機構を用いたバンド幅の DIMMnet-1 実機上での測定結果を報告する．第 5 章で関連研究について述べ，第 6 章でまとめる．

†1 (株) 東芝, 研究開発センター
Corporate Research and Developpment Center, Toshiba
†2 東京農工大学
Tokyo University of Agriculture and Technology
†3 (株) 日立製作所
Hitachi Ltd.
†4 (株) 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
☆ 現在，産業技術総合研究所
Presently with National Institute of Advanced Industrial Science and Technology
†6 慶應義塾大学
Keio University

2. アーキテクチャ

我々は PC クラスタのための低遅延高バンド幅な NIC を構築するための種々のアーキテクチャを提案している．本章では DIMMnet-1 プロトタイプに実装されている主なアーキテクチャの概要を紹介する．

2.1 MEMOnet

従来のように PCI バス [11] 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 MEMOnet[12] 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，CPU の進歩に歩調を合わせて発展可能な NIC を実現可能とする．

ある時点で PCI-X や PCI express[11] などが PC 上で高いバンド幅を提供できて，PC クラスタ用 NIC のための標準 I/O 規格の制約が一時的に緩く見えることはあるかもしれない．

しかし，I/O バスに主記憶と同等のバンド幅が意味を持つ用途は PC クラスタ用 NIC 以外には稀であるため，実際の PC に搭載される I/O バスにはメモリバスのバンド幅が実装されることは，ほとんど無いものと思われる．さらに I/O バスの成長スピードは CPU の成長スピードより遅く，これは将来のさらなる状況悪化を意味する．

これに対し MEMOnet は，安価な PC に後付け可能な NIC のためにメモリスロットを用いる．このため MEMOnet は，CPU の高速化に追随した FSB のバンド幅に匹敵するバンド幅を，提供する可能性を常に維持できる．

例えば，現時点では Pentium4☆の FSB は 4.2GB/s であり，これに追随できる 32bitRIMM[14] も開発されている．よって将来計画されている Infiniband[6] の最大規格 12X(30Gbps) を超えるバンド幅を，MEMOnet は現時点の技術の組合せにより汎用 PC に導くことが可能と考えられる．

また, Yellowstone 技術 [15] により，主記憶バンド幅は 100GB/s までが近い将来得られる見込みができている．ゆえに，近未来に普及するであろう標準 I/O 規格の性能を大幅に凌駕する未来像を，MEMOnet 上には描くことができる．

2.2 AOTF 送信機構

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 1** に示す．

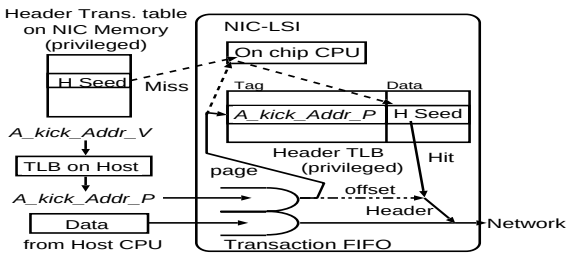

図 1 Atomic オンザフライ (AOTF) 送信

ここで，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 2** に DIMMnet-1 における HTLB の構成を示す．

パケットは起動に用いたアドレス（AOTF キックアドレス）

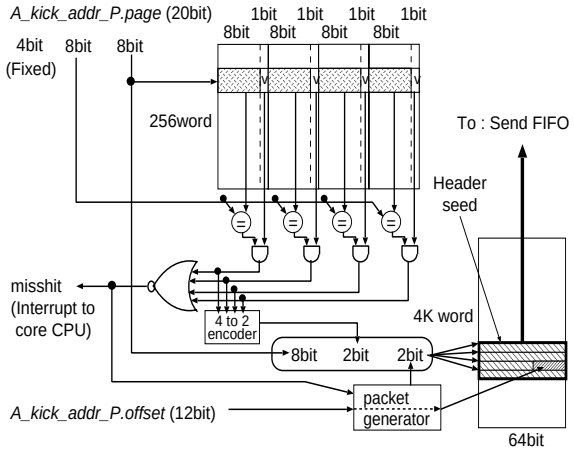

図 2 ヘッダ TLB(HTLB) の構造

の下位 bit 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．

2.3 OTF 受信機構

OTF 受信機構とは，アドレス変換や DMA コントローラの起動をすること無しに，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．

DIMMnet-1 では 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ヘッドを削除することが可能である．AOTF 送信に限って立てることができるヘッダー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込みバッファに押し込んでいく．書込みバッファは Martini 上のオンチップメモリである低遅延共有メモリ (LLCM) に，書き込めるタイミングで書き込む．

このように DIMMnet-1 では送信側の AOTF と受信側の OTFR が共同して極めて低遅延な通信を実現している．

2.4 BOTF 送信機構

Block On the fly(BOTF) 送信は，プロテクション刻印ウィンドーメモリにユーザモードでコピーされる複数ダブルワードにわたる一連の書き込みデータにプロテクション情報を付加して，ネットワークに送信する低遅延で高バンド幅な通信である．AOTF 送信と異なり送信データが 8 バイトを超えても構わないが，物理アドレスでのリモートアドレス指定はできない．

ほとんどパケットそのものに近い状態でホスト CPU から NIC のハードウェアに渡されるので，NIC の回路が簡素で済み，かつ少ないクロック数でネットワークにパケットを出力できる．BOTF 送信におけるパケット生成メカニズムを図 3 に示す．

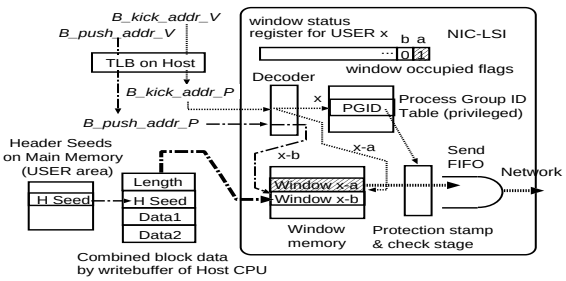

図 3 Block オンザフライ (BOTF) 送信

2.5 RDMA 送受信機構

DIMMnet-1 においても RHiNET や Infiniband，QsNET などと同様に Remote DMA(RDMA) の機能が使える．これらの RDMA 間には実装上の若干の差はあるものの本質的な部分での

☆ Pentium は Intel Corporation の登録商標です．本書に記載の商品の名称は，それぞれ各社が商標および登録商標として使用している場合があります．

違いはない．RDMA は AOTF や BOTF による通信よりもオーバヘッドが大きいが，非常に長いパケットを送る場合は BOTF を上回るバンド幅を実現できる．

RHiNET 等においては，PCI バスを経由して RDMA がホストの主記憶をアクセスする．これに対し，DIMMnet-1 では RDMA は NIC 基板上のバンク構成になった DIMM や通信制御 ASIC 内部の低遅延共有メモリ LLCM にアクセスする．DIMMnet-1 上の DIMM にホストから書き込んだ送信データをネットワークに送信する時には，RDMA による送信を起動する前に，送信データを保持している DIMM のバンクをホスト側から Martini 側に切り替える必要があるので，オーバヘッドが大きい．

一方，現在の DIMMnet-1 の実装では RHiNET 用に最適化された受信部を流用しているため，受信側におけるパケット毎に必要なインターバルが大きく，BOTF による短いパケットに伴うバンド幅が十分に出ない．このため，このような DIMMnet-1 に最適化されていない実装においては，パケット長に関する制約が緩い RDMA は，大量データを送信する際に高バンド幅を実現する手段として意味がある．

3. 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．

3.1 ス イッチ

DIMMnet-1 に接続可能なスイッチの仕様を以下に示す．本報告に記載の実験で用いられているのは光版の RHiNET-2/SW[16] である．なお，RHiNET-2/SW は条件つきながらマルチキャスト機能を持っており，後述する実験における同期完了通知パケットのような小さなパケットなら，1 個のパケットを送るのと同等の遅延時間 (240ns) で複数のポートに同一パケットをマルチキャストすることが可能である．

表 1 DIMMnet-1 に接続可能なスイッチの仕様

スイッチ	RHiNET-2[16]	RHiNET-3[17]	OIP-SW[18]
光 port	8(or 2)	8	15
電気 port	0(or 6)	0	1
I/O ピン	800Mbps×10	1250Mbps×8	250Mbps×8
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

3.2 通信制御 ASIC

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 上の制御チップである．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，ロックなどの複雑な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ことを意図して設計された．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いる．両者の違いを表に示す．特に第二版の Martini は 64bit66MHzPCI バスを用いる RHiNET-2/NI 向けに最適化されているため，DIMMnet-1 として用いる場合はほとんど全ての測定項目で第一版の方が性能的には高くなっている．

本報告における実験では主に第二版の Martini LSI を用いているが，一部の実験は第一版の Martini LSI を用いている．

表 2 二つの Martini LSI の違い

	第一版	第二版
電源電圧	>2.5V	2.5V
DIMM(Hz) : コア (Hz)	1 : 1	2 : 1
DIMM 最大周波数	100MHz	133MHz
AOTF&OTFR 最大周波数	100MHz	66MHz
コア最大周波数	66MHz	66MHz
DIMM → Window のビット幅	64bit	128bit
論理的なボトルネック	なし	コア部
最大バンド幅	528MB/s	528MB/s
状態読み取り並列度	2 windows	64 windows
リンク最大周波数	250MHz	400MHz
RHiNET2/SW との接続	不可	可
2 ノード間直接接続	可	不可
バグの数	多	少
歩留まり	悪	良

3.3 ネットワークインタフェース

DIMMnet-1 は，PC66，PC100 または PC133 仕様の DIMM スロットに装着する NIC である．DIMMnet-1 の主な仕様を**表 3** に示す．なお，ここで示されているのは目標値ではなく，今回の実験で用いられている第二版の Martini を用いた光版 RHiNET2/SW 用プロトタイプの実機の値である．その基本構造を**図 4** 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．

表 3 DIMMnet-1 プロトタイプの主な仕様

ホストとのインタフェース	DIMM および PEMM
NIC メモリ	PC133,SO-DIMM2 枚
搭載 SO-DIMM 容量	256MB
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	各方向 8Gbps（全二重）
ボトルネック部バンド幅	各方向 400MB/s（全二重）
NIC メモリバンド幅	800MB/s(ホスト側) 800MB/s(network 側)

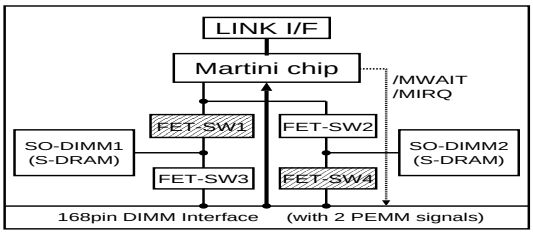

図 4 DIMMnet-1 の基本構造

4. 性 能 評 価

4.1 評 価 環 境

本章で述べる実験の評価環境を**表 4** に示す．

表 4 評価環境

環境	A	B	C	D	E
CPU	Pentium 3		Pentium 4		
Core	850MHz		1.5GHz		1.8GHz
FSB	100MHz				
チップセット	VIA 社 Pro133A		VIA 社 P4X266		
NIC	光版 DIMMnet-1				
Martini	ver.1	ver.2	ver.1	ver.2	
スイッチ	直結	RHiNET2/SW	直結	RHiNET2/SW	
OS	Linux(Kernel 2.4.2)				

4.2 バリア同期

DIMMnet-1 では，送信に AOTF 送信機構を用いてリモートアドレスに物理アドレス指定の 1 バイト書込みを行うパケットを送信し，受信に OTF 受信機構を用いて Martini のオンチップメモリである LLCM に書込み，8 バイトをホストからポーリングするという経路が最も低遅延である．そこで，バリア同期の実装の際にもその通信経路を用いる．

以上の手法を用いた，光版 RHiNET2 スイッチによって接続された光版 DIMMnet-1(第二版の Martini を使用) を用いた 7 ノード構成の PC クラスタ上のバリア同期と，Myrinet による PC クラスタ上でのバリア同期の所要時間の実測値を**図 5** に示す．測定には表 4 の環境 E を用いた．

このうち，一番上の Myrinet(M2M-PCI32C) については SCore5.0 の MPI によるバリア同期関数の測定結果である．真中の実線は同期完了通知のために通常の一対一通信を用いたものであり，一番下の破線は RHiNET2/SW のマルチキャスト機能を用いた場合のものである．

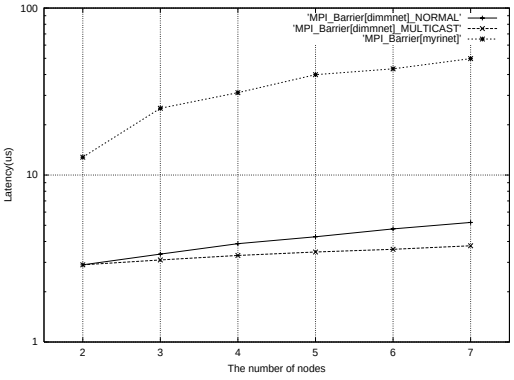

図 5 バリア同期遅延時間の比較

これより判るように，DIMMnet-1 では Myrinet によるよりも一桁程度高速にバリア同期がとれる．特にマルチキャストを用いる場合はマルチキャストを使わない場合に比べて台数が増えた場合の所要時間の増加が少ない．

その理由は，マルチキャストを用いない場合は同期参加ノードが一つ増えると root ノードに出入りするスイッチで，1 パケット処理するのにかかる遅延 (240ns) の往復分だけ増加するが，マルチキャストを用いる場合は，片方向分は一定なので，台数による遅延時間の増え方が半減するためである．

なお，本実験で用いた通信経路は大半が 50MHz で動作しているが，第一版の Martini では 100MHz でも動作していたため，もし第一版の Martini を RHiNET2/SW で接続する環境を構築できればさらに高速であると思われる．

4.3 大域加算

NIC に搭載される CPU はホストの CPU と比較して低周波数であるとともに，通常，回路規模削減のために浮動小数演算器は入っていない．このため，バリア同期のような整数系の処理を NIC 上で行うという実現例 [19] はあっても，浮動小数系の大域演算を NIC 上で高速に行うことは困難である．

上述のバリア同期は，同期専用ハードウェアによるものでもなく，NIC 上の低速な CPU によるソフトウェアによるものでもなく，ホストの高速な CPU によるソフトウェアによって実装されている．にも関わらず，非常に高速に実行されている．このことは，同期のような整数系の処理のみならず，大域加算や最大値検索のような浮動小数演算系の大域演算に対しても，同じような手法で特別な専用回路を設けることなく高速に実行できることを意味している．

DIMMnet-1 による大域演算の高速性を確認するために，大域加算を各種のデータ型において実行した．本実験は表 4 の環境 A,B,C,D での測定である．大域演算を行うデータ型を変更した時のデータ送信受信に使用する命令を表 5 へ示す．また大域加算遅延の測定結果を表 6 へ示す．第一版の Martini を用いた場合は 2 μ秒を切る遅延時間で 2 ノードに分散する浮動小数の総和が双方のノードに通知されていることが判る．一方，低遅延で有名な SCI-PCI や QsNET の場合は，この時間では単方向のリモート書込みすら終わらない．

表 5 各実験使用送信受信命令

	送信	受信
unsigned	通常命令 (mov)/unsigned	通常命令 (mov)/unsigned
unsigned64	MMX 命令 (movq)	通常命令 (mov)/unsigned
ull	MMX 命令 (movq)	通常命令 (mov)/ull
float	浮動小数命令 (flds/fstps)	浮動小数命令 (flds/fstps)
float64	MMX 命令 (movq)	浮動小数命令 (flds/fstps)
double	MMX 命令 (movq)	浮動小数命令 (fldl/fstpl)

表 6 光版 DIMMnet-1 の大域加算遅延

環境	A	B	C	D
CPU(Core/FSB)	P3(850M/100M)		P4(1.5G/400M)	
Martini	Ver.1	Ver.2	Ver.1	Ver.2
DIMM 部	100MHz	100MHz	100MHz	100MHz
Core 部	100MHz	50MHz	100MHz	50MHz
スイッチ	無	有	無	有
unsigned	1859ns	2741ns	-	-
unsigned64	1864ns	2744ns	2101ns	3001ns
ull	2106ns	3022ns	2651ns	3501ns
float	1854ns	2739ns	-	-
float64	1863ns	2755ns	2165ns	2971ns
double	1861ns	2757ns	2163ns	2939ns

4.4 バンド幅

本節では，BOTF 送信機構や RDMA 送信機構を用いたバンド幅の DIMMnet-1 実機上での測定結果を報告し，今後の改善の指針や，両者の特質について考察する．

4.4.1 RDMA によるバンド幅

表 4 の環境 B での DIMMnet-1 による RDMA を用いた場合の通信バンド幅の実測値を**図 6** に示す．

図 6 中 (A) はスイッチを介した 2 台での測定であり，(B) はスイッチを介した 1 台での測定，すなわち送受信同時実行時の片方向分のバンド幅である．(A) より RDMA による継続バンド幅において 330MB/s が得られていることが判った．しかし (B) のように一台の NIC 内部で送信と受信が並行して動作した場合，SO-DIMM が送信／受信 DMA の競合資源となりピーク時の 0.6 倍にバンド幅が低下することが判った．

4.4.2 BOTF によるバンド幅

BOTF によるバンド幅測定は，送信データ量を 464(bytes) 毎に分割し，2 枚の Window メモリにより交互にパケットを発生

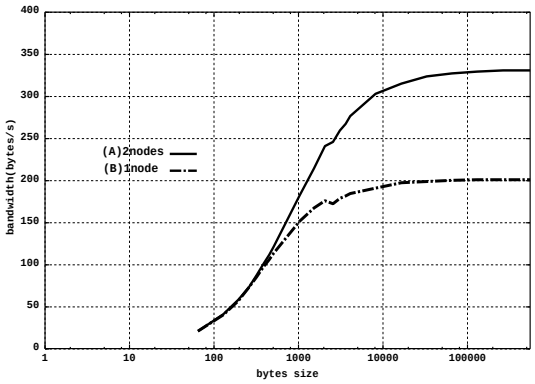

図 6 光版 DIMMnet-1(2nd) の RDMA 通信バンド幅

させることにより行った．ここで，Window メモリのキャッシュ属性は Write Combining とし，データ受信領域には LLCM を用いた．また送信データは CPU キャッシュ上にあるものとした．

第二版の Martini を用いた光版の DIMMnet-1 は表 4 の環境 B で測定し，第一版の Martini を用いた光版の DIMMnet-1 は表 4 の環境 A において通信リンクを自己ループさせ，DIMM の周波数を 66MHz に落した状態で測定した．すなわち結果は送受信同時実行時の片方向分のバンド幅である．DIMMnet-1 による BOTF を用いた場合の通信バンド幅の実測値を**図 7** に示す．

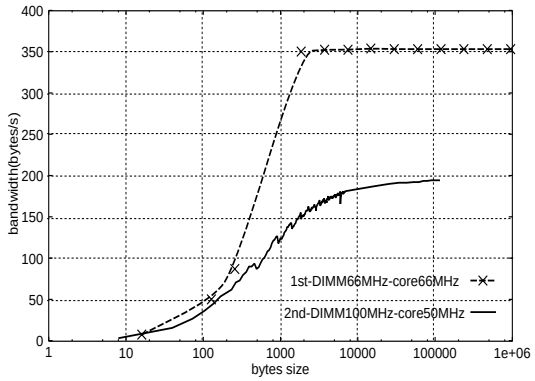

図 7 光版 DIMMnet-1(1st,2nd) の BOTF 通信バンド幅

66MHZ の DIMM スロット上で動作している第一版の Martini を用いた場合では BOTF によって単方向で 350MB/s，双方向で 700MB/s を超えるバンド幅が確認された．

一方，第二版の Martini を用いた場合の結果はスイッチを介した 2 台での測定であるが，1 台の測定でも RDMA を用いた時のようにバンド幅が低下することはなかった．しかし，BOTF の継続バンド幅は 190MB/s であり RDMA と比較して約 40%のバンド幅低下が発生していた．

4.5 バンド幅に関する考察

第一版の場合は，66MHz の DIMM スロット上での動作であるにもかかわらず，内部で分周されていないため，第二版よりも高速に動作している．低周波数で動作している PCI バス関係等を削除して，ダイザイズを減らし，133MHz 化することで，さらに上記の 2 倍程度の性能は達成できると考えられる．

第二版の場合は，133MHz ではなく 100MHz の DIMM スロット上で動作させ，しかも内部では 2 分周された 50MHz で動いているものであり，当初の設計値とはかけ離れたところで動作しているため絶対性能が低くなっている．

第二版を用いた同一の環境での比較で，RDMA に対して BOTF はバンド幅が低かった原因は，今回の実装では受信側を RHiNET の RDMA 通信用に最適化された Martini の RDMA 受信部を流用したことにある．ヘッダ解析から DMA 開始までの処理と，DMA によるメモリ書込みがパイプライン化されておらず，パケット間インターバルが 41 サイクルと大きい．

RDMA の方はパケット長を十分に長く設定することができるので，パケット長を十分大きくすればパケット間インターバルの弊害は無視できるようになる．これに対し，BOTF の方は Window メモリのサイズの限界により送信データ長を本プロトタイプの場合は 464 バイトに区切ってパケット化する必要がある．464 バイトのデータ書込み時間の 58 サイクルに対して，パケット間インターバルが 41 サイクルというのは大変大きく，これが約 40%の性能低下をもたらす主要な原因となっている．

上記の問題は，受信部において先行するパケットのデータ部の DMA によるメモリ書込み (58 サイクル) と，後続するパケットのヘッダ解析から DMA 開始までの処理 (パケット間インターバル 41 サイクルに相当) を並列実行できるようにすることで解決できるものと考えられる．

よって，RDMA と BOTF による主記憶上にメッセージが準備できた状態でのバンド幅の差は，有効データに対するパケットヘッダの占める比率の差と，オーバヘッドの差のトレードオフの関係に帰着されるため，RDMA の優位性は将来的には小さくなると考えられる．

今回の測定はメモリ上に送信メッセージが準備できている状態からのバンド幅であり，実際に並列処理で用いる場合は RDMA ではメモリ上に送信メッセージを作り上げる作業が発生する．その上でヘッダー情報とほぼ同等な情報量を持つコマンドを Window メモリに書き込む必要があり，それらの時間は BOTF 送信時のパケットイメージを Window メモリに書き込む時間とほとんど差がない．このため，実際に並列処理を行う際の通信への CPU 負荷は RDMA と BOTF で大差は無い．

一方，BOTF の場合はこの作り上げる作業の中で送信が行われているのに対し，RDMA の場合は全く別のオーバヘッドとして，今回測定されたバンド幅に基づいて消費される転送時間に上乗せされる．よって 並列処理に用いる際の実質的なバンド幅は今回測定された数字以上に BOTF の方が高いものと思われる．

ただし，受信したデータを CPU によって何ら加工することなく store & forward させる場合に限っては，RDMA における CPU 負荷は BOTF における CPU 負荷を大幅に下回る．よってファイルサーバやルータ用途に DIMMnet-1 を用いる場合は RDMA が有効であり，受信データに関して CPU 処理することが本質的である並列処理用途では BOTF が有効であるといえる．

5. 関 連 研 究

5.1 QsNET，RHiNET

Quadrics 社の QsNET[3)4)] の NIC である Elan は COMPAQ の AlphaServer SC シリーズ [5)] の結合網として採用されている．これは「RDMA を基本的にはハードで実行する PCI バス版の NIC」という意味で，本来の RHiNET[13)](DIMMnet-1 向けに後から追加した AOTF 送信部，BOTF 送信部，OTF 受信部を除いた構成) に非常に近い．

しかし，QsNET や RHiNET は PCI バスがネックとなり，送受信を同時に行うとバンド幅が低下する．特に QsNET は AlphaServer の主記憶間で送受信すると片方向あたり 80MB/s までバンド幅が低下してしまう．ServerWorksHE ベースの Pentium3 環境では単方向が 317MB/s に対し，双方向は片方向あたり 160MB/s に低下する．[3)]RHiNET も上記のケースや，DIMMnet-1 の RDMA を用いた双方向通信の場合の性能低下と同様に，半減に近い性能低下が発生するのは避けられない．一方，DIMMnet-1 の BOTF を用いた双方向通信の場合は性能低下がない．

また，これらの PCI バス型 NIC では，今回のバリア同期や大域演算の実験のようなホストからのポーリングを NIC 上のメモリに対して行うことは，ポーリングが PCI バス上で本来の処理に伴うデータの受信そのものを阻害してしまうため好ましくな

い．よって，ポーリングすべきデータは主記憶上まで DMA されなければならず，その分遅延は増加する．他の通信が行われていない状況での QsNET の同期は 4 ノードで 5〜10 μ秒程度であるが，その背景で通信が行われていると 8〜200 μ秒程度まで増加してしまう．[3] 一方，DIMMnet-1 では LLCM へのホストからのポーリングが DIMM 側に受信されるであろうデータの受信を阻害することはない．

5.2 Infiniband

PC クラスタに使用可能な Infiniband[6] の HCA(Host Channel Adapter) の実装には，64bit66MHzPCI による 1X タイプのもの (Essential 社の IB-Now HCA[8]，JNI 社の InfiniStar IBP-1X02[7]) や，PCI-X による 4X タイプのもの (Mellanox 社の MTEK23108[9]，IBM 社の HCA[10]) がアナウンスされている．

しかし，64bit66MHzPCI や PCI-X[11] は，少なくとも現時点では PC に標準的に装備されているものではなく，一部の高価なサーバーにのみ搭載される．また，I/O バスにおける遅延の問題もこの種の実装においては残存する．

将来的には HCA が I/O バス上ではなく，専用のチップセット上に実装されるようになるかもしれない．しかし，それらはクラスタの構成要素をサーバー製品とすることを強いるため，全体として高価なシステムになると考えられる．

一方，DIMMnet-1 は安価な PC にも標準的に搭載されているメモリスロットに搭載されるために，全体として安価なソリューションを提供できる．将来的にも CPU や FSB の速度向上に追随した性能向上が見込める．

また，通信方式という観点から Infiniband と DIMMnet-1 と比較すれば，DIMMnet-1 は RDMA という Infiniband と同様の通信方式を有するとともに，AOTF や BOTF という Programmed I/O ベースの低遅延通信方式も有し，大域処理等のより粒度の細かい並列処理にも対応できる．

6. ま と め

本報告では，AOTF 送信と OTF 受信によるバリア同期や大域加算の遅延時間と，BOTF 送信や RDMA 送信を用いたバンド幅の DIMMnet-1 実機上での測定結果を述べた．

今回の実験で用いられた DIMMnet-1 プロトタイプは 133MHz ではなく 100MHz の DIMM スロットで動作させ，さらに内部が二分周されているバージョンという不完全な状態での測定である．にもかかわらず，バリア同期に関しては Myrinet 上の Score/MPI との比較を行い，約 10 倍の高速実行が達成されることが示された．

DIMMnet-1 によればホストの高速 CPU を用いた高速な実装が可能なので，バリア同期のような整数演算系の大域演算のみならず，浮動小数演算の総和などの大域演算も高速化できる．今回の測定では 2 μ s という短時間で 2 ノードに分散する浮動小数の総和が双方のノードに通知された．一方，この時間では SCI-PCI や QsNET の場合は単方向のリモート書込みすら終わらない．

バンド幅については，並列処理にいて重要な双方向のバンド幅については RDMA ではメモリアクセスという本質的な資源競合のために性能が大きく劣化するのに対し，BOTF では全く性能低下がないことが確認された．

一方，単方向バンド幅に関しては，今回の実装では受信側を RDMA 通信用に最適化された Martini の RDMA 受信部を流用しているため，BOTF では RDMA による単方向バンド幅よりも 40%の性能低下が発生してしまった．小さなパケットでも高いバンド幅を出せるよう，受信部のパイプライン化といった BOTF 向けの受信部への最適化が必要である．

今後は，低レベルな API を用いたアプリケーションによる評価，メッセージ交換の実装，MPI を用いたアプリケーションによる評価，並列コンパイラへの適用を進める．ハード面では今回の評価結果を参考に，最新の高バンド幅メモリインタフェースに基づいた MEMOnet に特化した実装を検討する予定である．

謝辞 (株) 日立製作所の西氏，(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，(株)NEC ソリューションズの土屋氏，慶應義塾大学の渡辺氏，元・東京農工大の須田氏をはじめ 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. http://www.myri.com/
2) Dolphin Corp. “PCI SCI-64 Adapter Card” http://www.dolphinics.no/dics_pci-sci64.htm
3) F. Petrini, W. Feng, A. Hoisie, S. Coll, E. Frachtenberg “The Quadrics Network : High-Performance Clustering Technology”, IEEE Micro, pp.46-57 (2002)
4) F. Petrini, S. Coll, E. Frachtenberg, A. Hoisie “Hardware- and Software-Based Collective Communication on the Quadrics Network”, Proc. Int’l Symp. on Network Computing and Application 2001, pp.24-35 (2001)
5) COMPAQ “AlphaServer SC : Scalable Supercomputing”, No.135D-0900A-USEN (2000)
6) InfiniBand Trade Association, http://www.sysio.org/
7) JNI Corp. “InfiniStar IBP-1X02 PCI-to-InfiniBand HCA Module Dual 2.5 Gb IB Ports”, http://www.jni.com/Products/IB/PDFs/IBP.pdf
8) Essential Communications “IB-Now InfiniBand Adapter”, http://www.esscom.com/resources/IB-NowAdapter.pdf
9) Mellanox Technologies Inc. “InfiniHost MT23108” http://www.mellanox.co.il/products/shared/Infinihostglossy.pdf
10) IBM Corp. “IBM PCI-X to InfiniBand Host Channel Adapter Summary Datasheet - Version 03”, http://www-3.ibm.com/chips/techlib/techlib.nsf/techdocs/852569B20050FF7785256990006DB7D5/$file/hcasds03_pub.pdf
11) PCI-SIG, http://www.pcisig.com/
12) 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)
13) 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”, JSPP2002, pp.35-42 (2002)
14) Samsung Semiconductor ”High performance of 32bit RIMM 4200”, http://www.samsungelectronics.com/semiconductors/dram/rambus_dram/32bit_rimm.htm
15) Rambus ”Yellowstone Memory Signaling Technology”, http://www.rambus.com/technology/yellowstone_overview.html
16) 西村，工藤，西，山本，原澤，福田，敷地，坏 “RHiNET-2/SW: 並列光インタコネクションを搭載した並列計算機システム用高速ネットワークスイッチ，電子情報通信学会論文誌，VOL. J84-C NO.9，pp. 756-765，(2001)
17) 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”, 情報処理学会研究報告 2000-ARC-140-4，pp.13-18 (2000)
18) T. Yoshikawa, H. Matsuoka “Optical Interconnections for Parallel and Distributed Computing”, Proceedings of the IEEE, Vol. 88, No. 6, 2000 pp.849-855 (2000)
19) D.Buntinas et al. “Performance Benefits of NIC-Based Barrier on Myrinet/GM”, Proceedings of the Workshop on Communication Architecture for Clusters (CAC) with IPDPS’01 (2001)