ハイパフォーマンス
コンピューティング 87−19
（２００１．７．２６）

メモリスロットに搭載されるNICのクラスタ性能予測

濱田 芳博 † 中條拓伯 †
田邊 昇 †† 工藤知宏 †††

PC メモリスロットへ搭載し，低レイテンシかつ高バンド幅な通信を可能にするネットワークインタフェース (NIC) として新情報処理開発機構並列分散システムアーキテクチャ研究室で開発が行われている．本報告書では DIMMnet-1 について，ホスト CPU-NIC 間のバンド幅という観点から性能予測を行う．比較対象は PCI バス上へ搭載される Myrinet 社の NIC(33/66MHz) であり，これらの NIC は同等の送受信機構と物理接続を持つと仮定した．

Performance Estimation of a Cluster with a NIC Plugged into a Memory Slot

YOSHIHIRO HAMADA ,† HIRONORI NAKAJO ,† NOBORU TANABE ††
and TOMOHIRO KUDOH†††

Real World Computing Partnership has been developing a new network interface called DIMMnet-1 that is plugged into a memory bus in a PC in order to establish a low latency and high bandwidth network for cluster computing. In this paper performance estimation is discribed bandwidth of a data path between CPU and NIC. A target for comparision is Myrinet of Myricom. We assume that both NIC's holds the same transfer mechanisms and physical connection, and discuss advantage and disadvantage of memory bus and I/O bus.

1. はじめに

低価格な PC を高速なネットワークで接続し並列処理を行う PC クラスタが，今後の並列処理システムとして注目されている．しかしながら，PC 上の CPU 速度の向上に対して，ネットワークインタフェース (NIC) の速度や，それを接続するバスインタフェースの性能および接続方法について考慮しなければ，システム全体で十分な性能を発揮することは困難である．

クラスタコンピューティングのための代表的な NIC として Myricom 社の Myrinet[1] があげられる．しかしながら，Myrinet は PCI バスを通じて接続され，そのレイテンシやバンド幅が問題となると考えられる．それに対し，新情報処理開発機構で開発されている DIMMnet-1[2] はメモリバスを通じて接続され，前述の問題を解決することを目的としている．

クラスタコンピューティングにおいて，データ転送にかかる時間を短縮するため，転送処理をポインタの受渡しで行うゼロコピー通信が有効である．この方式を Myrinet へ適用する際には，メモリ上へ送受信領域を作成し，メモリ⇔PCI 間の DMA 転送によって，データ転送を CPU の計算処理と並行して実行する．DIMMnet-1 においても，その内部動作はゼロコピー通信と同様であるが，NIC がメモリバスへ接続されることによるホスト CPU ⇔ NIC 間のバンド幅の拡大や，細粒度のデータ転送，およびホスト CPU の NIC に対するポーリングの低レイテンシ化が行える．

しかしながら，DIMMnet-1 の送受信領域はホスト CPU と NIC により非キャッシュ属性としてアクセスされるが，Myrinet に関しては DMA がホスト CPU のキャッシュ一貫性を保つため，送受信領域をキャッシュ属性のまま利用可能である．非キャッシュ領域へのアクセスは，キャッシュ領域へのアクセスに対しバースト転送が行えないという点と，データの再利用ができないという点で，アプリケーションの実行時間が増大する要因となる．

この回避策として，DIMMnet-1 の設計においては，送信領域に使用されるメモリ領域に対し，Intel P6 プロセッサに搭載されている WriteCombining を使用するとしている[2]．WriteCombining は，非キャシュ属性へのメモリアクセスをバースト単位にまとめあげる機能である．これにより DIMMnet-1 の送受信領域に対して，CPU からの書き込みはキャッシュ属性の場合と同程度に行えるが，読み込みに関してはこの機能

† 東京農工大学
Tokyo University of Agliculture and Technology
†† （株）東芝 研究開発センター
TOSHIBA Corporate Research & Development Center
††† 新情報処理開発機構
Real World Computing Partnership

を利用したとしても，バンド幅は非キャッシュ属性のものとほとんど変らないことが判っている．これにより連続ワードデータの通信においては，DIMMnet-1がメモリバスへ接続される効果が相殺される可能性が考えられる．

本論文においては，DIMMnet-1 と Myrinet のような PCI バスに接続される NIC とをバンド幅などの観点から比較し，DIMMnet-1 が持つ高速なデータ通信の効果を生かした状態で，その有効性について検討を行う．

2. メモリバス接続と I/O バス接続 (PCI) との比較

2.1 PC 内部におけるデータ転送経路の相違

PC 内部における，DIMMnet-1 と Myrinet のデータ転送経路の相違は，ホスト CPU への接続が前者はメモリバスを使用し，後者は PCI バスを使用するという点である．

ここで PCI バスの DMA 転送に費される時間を調べるために Myrinet の PCI バス DMA 転送時バンド幅の測定を行う．測定環境については表 1 に示す．各種転送ブロックサイズによる DMA 転送時間の測定は，ホスト CPU 上のプログラムから Myrinet 上のプログラムへ DMA 開始を示し，これを受けた Myrinet 上のプログラムが DMA 起動～終了確認を行い，その間の時間をロジックアナライザで計測することにより行った．

測定結果として，図 1，2 に PCI 66MHz/33MHz のメモリ⇔ NIC 間のスループットを示す．図 1 にはメモリ→ NIC の DMA で，転送データがキャッシュにヒットしている場合と，ミスヒットしている場合のデータが示してある．これより，メモリ→ NIC 間の DMA 転送において，66MHz の PCI バスでは，転送データがキャッシュに存在する場合，WriteBack の影響を受け，CPU キャッシュサイズ内 (256K バイト) において WriteBack が発生しない場合と比較して，約 2 割程度のバンド幅低下が測定された．33MHz の PCI バスでは、バンド幅低下が発生していないことが判る．

表 1 PCI DMA 測定環境

	PCI66MHz	PCI33MHz
Myrinet	M3S-PCI64B-4 (64bit 66MHz)	M2M-PCI32C (32bit 33MHz)
lanai-gcc version	2.95.2..1.3	
CPU/FSB	PentiumIII 1GHz/FSB133MHz	
MEMORY	PC133 Registerd/ECC 付	PC133 Unbuffered
CHIPSET	ServerSet LE	Apollo Pro133A
LinuxKernel	2.4.2	
compiler	egcs-2.91.66	

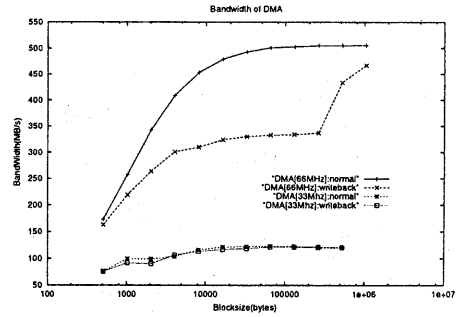

図 1 DMA Memory → NIC スループット

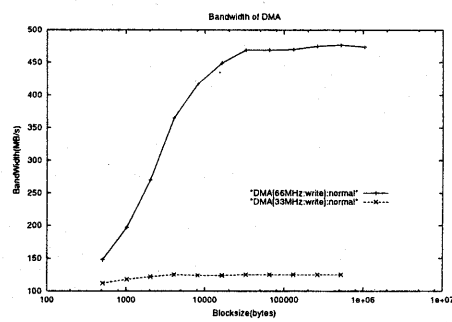

図 2 DMA NIC → Memory スループット

2.2 アプリケーションから見える送受信領域の相違

メモリ上のデータ送受信領域にアクセスするという点においては，アプリケーションから見える両者の NIC の違いは無いように思えるが，前述のように DIMMnet-1 のメモリ領域は非キャッシュ属性として扱わなければならず，その領域へのアクセスは，パーシャルでの操作しか行えないため，多大なアクセス遅延が生じる．これを回避するための WriteCombining の読み込みと書き込みに対する効果を調べるために，レジスタ⇔ Write-Back(WB)/Uncachable(UC)/WriteCombining(WC) 属性領域のバンド幅測定を行った．また測定中のメモリバスの動作を調べるために，PentiumIII のパフォーマンスカウンタを用い，バーストアクセスとパーシャルアクセスの発生回数を測定した．

測定環境は表 1 に示した PCI33MHz のものを使用した，測定結果を表 2，3，4，5 に示す．各々の表中の値は，キャッシュ属性が WB-WC-UC での測定値として示してある．また SIMD 命令を用いた測定ではデータ転送用としてレジスタを 2 つ使用し，WB への書き込みにおいては転送命令として一時のものと非一時のものを使用する．一時の命令は書き込みデータをキャッシュを経由させて行い，非一時のものは直接メモリへデータ書き込みを行うものである．表中で括弧で示してあるものが非一時の値である．

表 2 レジスタ→メモリ書き込みバンド幅 (通常命令使用)

	バンド幅	BURST	PARTIAL
単位	MB/s	回	回
512(bytes)	336-336-75	0-16-0	0-0-128
65536(bytes)	358-351-76	0-2048-0	0-0-16384
1(Mbytes)	173-351-76	0-32768-0	0-0-262144

表 3 レジスタ→メモリ書き込みバンド幅 (SIMD 命令使用)

	バンド幅	BURST	PARTIAL
単位	MB/s	回	回
512(bytes)	1145(986)-986-166	0-16-0	0-0-64
65536(bytes)	1429(1053)-1053-170	0-2048-0	0-0-8192
1(Mbytes)	171(1043)-1043-170	24576-32768-0	0-0-131072

表 4 メモリ→レジスタ読み込みバンド幅 (通常命令使用)

	バンド幅	BURST	PARTIAL
単位	MB/s	回	回
512(bytes)	221-45-35	16-0-0	0-128-128
65536(bytes)	220-45-35	2048-0-0	0-16384-16384
1(Mbytes)	220-45-35	32768-0-0	0-262144-262144

表 5 メモリ→レジスタ読み込みバンド幅 (SIMD 命令使用)

	バンド幅	BURST	PARTIAL
単位	MB/s	回	回
512(bytes)	435-142-74	16-0-0	0-64-64
65536(bytes)	466-148-74	2048-0-0	0-8192-8192
1(Mbytes)	466-148-74	32768-0-0	0-131072-131072

WC 領域への書き込みにおいては，通常命令，SIMD 命令いずれを用いても 32 バイトのバースト転送を発生させていることが判る．一方で UC 領域への書き込みは，通常命令を用いた場合は 4 バイトのパーシャルアクセスであり，SIMD 命令を用いた場合は 8 バイトのパーシャルアクセスを発生させていることが判る．これより，WC 領域へのアクセスは，仕様通りバースト転送を発生させいることが判る．WB 領域への書き込みにおいては，書き込みサイズがキャッシュ容量内におさまる 65536 バイトまではメモリへの転送が発生していない．これは書き込みデータがキャッシュ内に滞っているということである．キャッシュ容量を越える書き込においてはメモリバスへのアクセスが発生するが，この時 WB への書き込みバンド幅は WC の場合よりも低くなる．この差が顕著に現れるのは，SIMD 命令において一時の書き込み命令を利用した場合であり，書き込みサイズがキャッシュ容量に収まる内はメモリバス速度を超えたバンド幅が測定されている．しかしキャシュ容量を越えるとメモリバスへのアクセスが始まり，WC のバンド幅よりも低く，通常命令のバンド幅と変わりなくなる．また SIMD 命令において非一時の書きこみ命令を用いると，WC 領域への書き込みと同様のアクセスが行えることが判る．WC，UC 領域からの読み込みに対しては通常命令を使用した場合は 4 バイトのパーシャルアクセスであり，SIMD 命令を使用した場合は 8 バイトパーシャルアクセスが行われていることが判る．WB 領域からの読み込みは全て 32 バイトのバーストアクセスで行われ，WB 領域からの読み込みも同様に全て 32 バイトのバーストアクセスで行われている．

以上より，非キャッシュ属性領域へ WC を適用すると，書き込み操作に対しバーストアクセスを使用することができるため，連続ワードのアクセスが効率良く行えることが判る．しかし読み込み関しては UC よりも若干効率は良くなるものの，全てのアクセスがパーシャルでしか行われていず，バンド幅は WB の 4 ～ 5 分の 1 程度である．

2.3 データ送受信における実行時間比較

WC 領域からの連続ワードの読み出しバンド幅が低いという点が，アプリケーションが DIMMnet-1 を使用する上でどの程度影響するかを検討する．ここでは NIC に DIMMnet-1 あるいは，Myrinet を使用する．両方の NIC の送信受信機構及び，接続は図 3 のように想定し，この環境において送信側プログラムバッファから受信側プログラムバッファへデータ移動にかかる時間を Myrinet に対する DIMMnet-1 の比率として求める．図において A～I の値は各々の伝送経路のデータ転送レートであり，GAP1，GAP2，GAP3 は DMA の動作に必要な時間である．便宜上 Myrinet 内の DMA は一つで表し，DIMMnet-1 も同一のものを使用するとする．データの送受信量は 512 バイト，65536 バイト，1M バイトで行い，GAP1 と GAP2 と GAP3 の和を GAP として，0～1ms の変数とする．DIMMnet-1 における受信側 NIC の DMA の転送レート F は H が PCI 66MHz/64bit であれば 2 倍，PCI33MHz/32bit であれば 4 倍であるとする．また E は B，D，F，H よりも十分に大きいとする．両者の NIC における送信処理と受信処理にかかる時間については，DIMMnet-1 は NIC 上の受信完了状態のため NIC 内の状態表示レジスタ (UC 属性) をメモリバス経由で 1 ワード分読むとし，この時間は 117ns である．Myrinet においては，送信処理において DMA を起動するために，PCI バス経由で 1 ワードのデータ書きこみを行い，この処理に必要な時間は 33MHz で 140ns 66MHz で 115ns である．また受信完了は NIC が 4 ワードの DMA 転送をメモリ上の特定の位置に行い，この領域より 1 ワードの読み出しを行い完了状態を検査する．DMA 転送に必要な時間は，33MHz で 0.70us，66MHz で 0.87us でありメモリ上の完了状態を読み出す時間は 42ns である．

D，H の転送レートは図 1 と図 2 より得た．またメモリ→メモリ間のバンド幅について測定した結果を表 6 に示す．測定環境は表 1 へ示した PCI33MHz のものを用いた．表 6 においては転送元のデータがメモリに存在する場合でありキャッシュ属性を WB，WC，

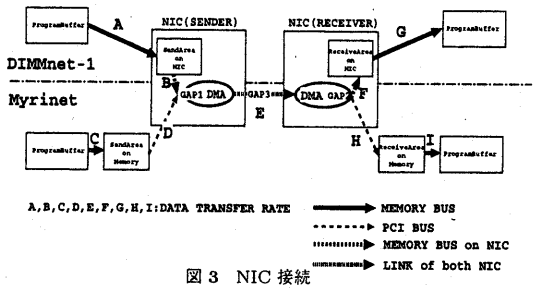

図 3 NIC 接続

UC とし，転送先のキャッシュ属性は WB とする．つまり受信領域からプログラム内バッファへデータを読み込む時のバンド幅である．また送信側のプログラムが転送データを送信領域へ書き込む時のバンド幅には表 2，3 のデータを利用した．

表 6 メモリ→メモリバンド幅

	WB		WC	
	NORM	SIMD	NORM	SIMD
512bytes	165MB/s	238MB/s	40MB/s	73MB/s
65536bytes	170MB/s	238MB/s	41MB/s	74MB/s
1Mbytes	103MB/s	98MB/s	34MB/s	49MB/s

図 4 へ PCI 33MHz Myrinet とのデータ送受信にかかる実行時間の比較結果を示し，図 5 と図 6 へ PCI 66MHz Myrinet との比較結果を示す．これらの図において，GAP を示している．縦軸は Myrinet におけるデータ送受信の実行時間に対する，DIMMnet-1 による実行時間の短縮度である．PCI33MHz との比較では SIMD 命令を用いた 512 バイトの転送時に，GAP が 1us より低い範囲で実行時間短縮が現れている．PCI66MHz との比較では，いずれにおいても実行時間短縮は現れていない．また，PCI において Write-Back が発生する場合には実行時間の増大が，65536，1M バイトの転送時に減少している．以上の結果より DIMMnet-1 の受信領域に対する WC 属性の影響より，これを利用するアプリケーションの速度低下が現れていることが判る．

3. 受信領域からの読み出しバンド幅改善

前章より DIMMnet-1 においては連続ワードのデータ送受信を行う場合．WC 領域からの読み出しの遅さが原因し，NIC の接続位置をメモリバスへ変更した効果が消えてしまうことが判る．

WC 領域からの連続ワードの読み込みが遅いのは，メモリアクセスがパーシャルで行われるからである．これを改善するために，DIMMnet-1 の送受信領域をキャッシュ属性として用いることを検討する．

3.1 ホスト CPU のキャッシュフラッシュ

キャッシュ一貫性が行われないデバイスに対して

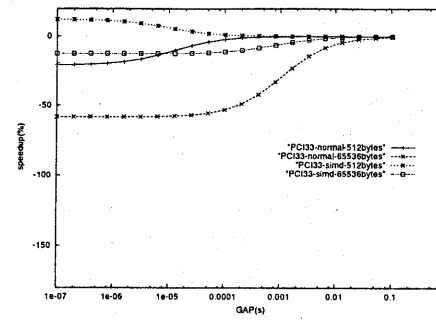

図 4 PCI33Myrinet に対するデータ転送実行時間短縮度 (WriteBack なし)

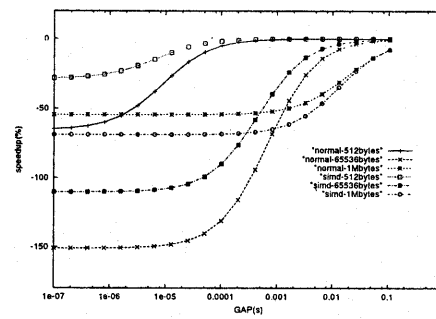

図 5 PCI66Myrinet に対するデータ転送実行時間短縮度 (WriteBack なし)

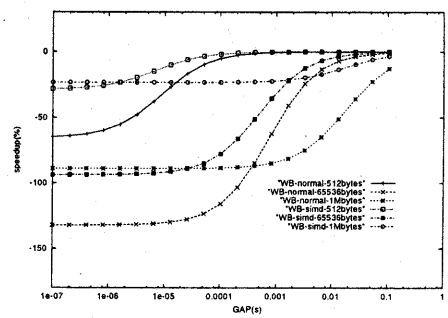

図 6 PCI66Myrinet に対するデータ転送実行時間短縮度 (WritwBack あり)

キャッシュ属性を適用するためには，メモリアクセスの際にキャッシュの状態をホスト CPU 側で更新すれば良い．PentiumIII においてこのために使用できる命令として，CPU 内のキャッシュをフラッシュする WBINVD がある．しかしこの命令はキャッシュ内の全てのデータを WriteBack または Invalidate するため，キャッシュの格納状況によってその実行に要する時間が異なる．これを調べるため，測定を行う前にメモリよりデータを読み込みかつ更新を行った後，WBINVD

を実行し処理時間の測定を行った．ここで，WBINVD は特権命令なのでカーネル内で動作する必要があり，ドライバへの IOCTL 呼び出しとして実行したので，実行時間はこの呼び出しにかかる時間も含んでいる．また測定中に発生する Writeback の回数をパフォーマンスカウンタで取得した．この結果を図 7 に示す．

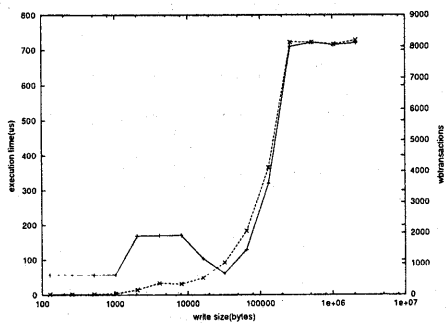

図 7　WBINVD の実行実行時間

結果より PentiumIII においてキャッシュフラッシュにかかる時間は，WriteBack の発生する回数に相関しており，57〜720us の間で変化している．

ここでキャッシュフラッシュを用い，DIMMnet-1 の受信領域へキャッシュ属性を適用した状態での前章での受信プログラムの実行時間比較を行う．キャッシュフラッシュの実行は受信領域よりのデータを読み出す前に一回行うとする．比較対象は PCI 66MHz Myrinet で WriteBack なしの状態に限定する．この結果について図 8 へキャッシュフラッシュにかかる時間が最小である 57us のものと，図 9 へキャッシュフラッシュにかかる時間が最大である 720us のものを示す．

以上よりキャッシュフラッシュが 57us である場合，転送量が 65536 バイトと 1M バイトであれば，Myrinet に対し DIMMnet-1 を利用する受信側アプリケーションの実行時間は短くなる．しかし 512 バイトでは実行時間の増加がみられ，またキャッシュフラッシの所要時間が 57us → 720us へ増加した場合はさらに増加することが判る．キャッシュフラッシュによるバンド幅確保の操作は，レイテンシを許容できない範囲で犠牲にしてしまうと言える．Pentium 4 においては，キャッシュフラッシュをキャッシュラインサイズで行う CLFLUSH 命令が実装されている．これを利用すれば，WBINVD のようにレイテンシを犠牲にすることはないので，受信領域へのキャッシュ属性適用を効率良く行える．つまり，DIMMnet-1 における受信領域からの読み込みバンド幅低下の問題は，PentiumIII 固有の問題と言える．

3.2　非一時プリフェッチの利用

[2] で示された DIMMnet-1 予測と同様に，非一時プリフェッチの利用の検討を行う．これは PentiumIII へ

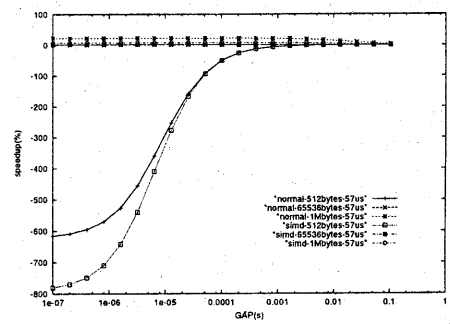

図 8　PCI66Myrinet に対する受信プログラム実行時間短縮度 (WriteBack なし，キャッシュフラッシュ57us)

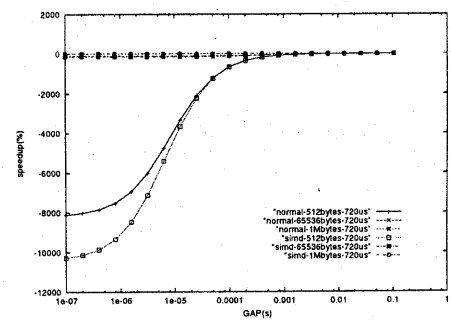

図 9　PCI66Myrinet に対するデータ転送実行時間短縮度 (WritwBack なし，キャッシュフラッシュ720us)

実装されている PrefetchNTA という命令でキャッシュ属性領域からの読み込みを，キャッシュへの影響を最小限に抑えて実行するものである．この命令の利用により受信領域からの読み込みにおいて，読み込みデータがキャッシュに保存されない状態で行えれば，受信領域にキャッシュ属性を適用可能である．つまり読み込み操作においてバースト操作が使用可能になる．

PrefetchNTA の動作を確認するために，メモリからの読み出しの実験を行った．キャッシュにヒットしていないメモリ領域から普通にデータを読み出す場合と，PrefetchNTA を経由してデータを読み込む場合でそれぞれの操作をした後，普通にデータを読み込む．この時，パフォーマンスカウンタにより，メモリバスに対するバーストリードの発生回数を調べた．結果は表 7 のようになった．これより PrefetchNTA による読み込み後ではデータがバスを移動していることが判る．また，普通読み出し後ではデータがキャッシュ上に存在するため，バス上をデータが通過していないことが判る．つまり，PrefetchNTA の利用により受信領域をキャッシュ属性として用いるられる可能性があると言える．

次に，これを用い DIMMnet-1 の送受信領域へキャッシュ属性を適用した状態で，前章での受信プログラム

の実行時間比較を行う．PrefetchNTA を用いたメモリ→メモリ間のバンド幅を表 8 に示し，受信側プログラムがデータをバッファへコピーする場合に用いるとする．比較対象は PCI 66MHz Myrinet で WriteBack なしの状態に限定する．

図 10 に実験結果を示す．これより，いずれの場合においても，Myrinet よりも短い時間でプログラムの実行を終えていることが判る．中でも SIMD 命令を用いたデータ転送量 512 バイトのものは GAP が 1us～1ns の間で 35～40%の速度向上を示している．また通常命令を利用したとしても，同様の範囲内で 28～32%の速度向上を示している．この結果においても，DIMMnet-1 の通信で転送データ量が低い場合に良好な結果を示しており，Myrinet のような PCI バスに接続される NIC に対し，パフォーマンスの差を出すにはこの点を利用したアプリケーションを作成または作成可能にする必要があると考える．

表 7　バーストリード発生回数

	PrefetchNTA 後	普通読み出し後
512bytes	16	0
65536bytes	2048	0

表 8　メモリ→メモリバンド幅

	WB	
	NORM	SIMD
512bytes	172MB/s	270MB/s
65536bytes	232MB/s	529MB/s
1Mbytes	146MB/s	155MB/s

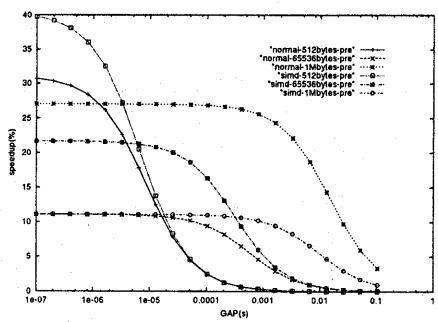

図 10　PCI66Myrinet に対するデータ転送実行時間短縮度
(WriteBack なし，PrefetchNTA 使用)

ここまでの DIMMnet-1 と Myrinet の比較では，DMA 動作のための遅延であるギャップが同一であると仮定して行ってきたが，実際の Myrinet と DIMMnet-1 においては DMA の数が異なる．Myrinet はホストメモリ⇔NIC 内メモリの DMA と，NIC 内メモリ⇔パケットインタフェースの DMA 2 つを持つため，データ送受信においては 4 つの DMA による 5 つのギャップが存在する．DIMMnet-1 においては DMA は 2 つで実装されているため，3 つのギャップが存在する．この差を考慮し，各々の DMA 間のギャップが同値であるとして，本節で行った実験を行った．結果を参考データとして図 11 に示す．図において横軸は 1 つのギャップの値である．各々の値において DIMMnet-1 は 3 倍 Myrinet は 5 倍のギャップを持つとした．

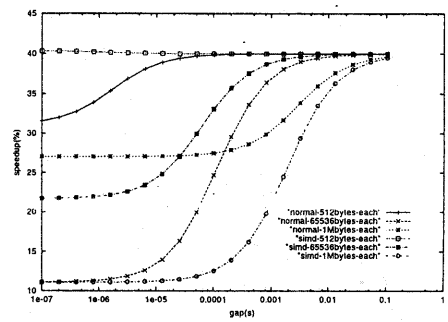

図 11　PCI66Myrinet に対するデータ転送実行時間短縮度
(WriteBack なし，PrefetchNTA 使用，ギャップ数考慮)

4. おわりに

DIMMnet-1 の受信領域に WC 属性を用いた場合，PentiumIII においては連続ワードのデータ転送においてパーシャルアクセスによるバンド幅の低さにより，Myrinet の様な PCI へ接続する NIC の方が高速なデータ通信が行えると考えたが，この場合であってもデータ転送量が 512bytes 程度で SIMD 命令を用いた状態であれば，PCI 33MHz Myrinet に対して数%の速度向上が予測される．また受信領域にキャッシュ属性を用いれば，転送量が 512bytes，SIMD 命令を用いた状態でギャップが 100ns～1us の間において，PCI 66MHz Myrinet と比較して 35～40%の速度向上が予測される．これより DIMMnet-1 を利用するアプリケーションが高速な送信を利用するには，小さな粒度でのデータ通信を行い，SIMD 命令の使用及び受信領域へのキャッシュ属性の適用が必要であると考える．

また本論文内で行った評価は，NIC 内の DMA が DIMMnet-1 と Myrinet で同一の数で行っており，実際には図 11 の様に，ギャップ変化においても常に速度向上が見られると考える．

参考文献

1) http://www.myri.com
2) Tanabe,Yamamoto,Kudoh. MEMnet : Network Interface attached on memory slot.In IPSJ SIGNotes 99-ARC-134(SWoPP'99),pp73-78,Japan,Augast 1999