社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2004－ARC－159 (27)
2004／7／31

メモリスロットインタフェースの得失

田邊 昇[†1] 箱崎博孝[†2] 土肥康孝[†2]
中條拓伯[†3] 天野英晴[†4]

2004年6月にi915/i925チップセットによりPCI Expressが市場に本格投入され，11年間PCの標準インタフェースのデファクトであり続けたPCIから，徐々にその地位が世代交代されようとしている．本報告では，PCI Express元年にあたり，メモリスロット装着型ネットワークインタフェースであるDIMMnet-1およびDIMMnet-2によって開拓されてきたメモリスロットインタフェースの得失とその意義について，ネットワークインタフェース・高機能メモリモジュール・リコンフィギャラブルシステムの三つの応用面から考察する．特に，高機能メモリモジュールにおける意義についてはNAS CGベンチマークを用いて得た評価より，PCI Expressに対して十分に大きな優位性を確認した．

The Pros and Cons of Using Memory Slot Interface

NOBORU TANABE,[†1] HIROTAKA HAKOZAKI,[†2] YASUNORI DOHI,[†2] HIRONORI NAKAJO[†3]
and HIDEHARU AMANO [†4]

PCI Express debuted in the market by i915/i925 chip set in June 2004. The de fact standard of the standard interface of PC is going to change generation gradually from PCI to PCI Express. This report considers the pros and cons and the meanings of the memory slot host interface prototyped with DIMMnet-1 which is a network interface plugged into a memory slot, and DIMMnet-2 in the PCI Express first year from three application sides of a network interface, highly efficient memory module and reconfigurable system. Especially about the meaning in a highly efficient memory module, big superiority to PCI Express was observed by the evaluation with the NAS CG benchmark.

1. はじめに

1993年にIntel社のパーソナルコンピュータ(PC)用チップセット420TX・430LXに導入されて以来，PCIバス[1]はオフィスや家庭で用いられるComercial off-the-shelf(COTS)なPCのインタフェース(NIC)のデファクトスタンダードの地位を現在まで続けてきた．日進月歩で進歩するPCの分野において，この11年間という期間は極めて異例とも言える長さである．

ムーアの法則で進歩するCPUやメモリの性能向上に対して，PCIバスから次の標準への移行はなかなか進まず，ビデオカードや一部のNICにおいてPCIバスの能力の不足が懸念されてきた．そこでビデオカード向けには専用のインタフェースであるAGPが導入され，サーバー向けの高バンド幅NICやストレージインタフェースのためには，64bitPCIやPCI-Xといったインタフェースが策定・採用されてきた．しかし，コストが重視されるCOTSのPCにこれらは導入されなかった．

我々は，上記のような状況を鑑み，CPU性能と通信性能がバランスした高性能なPCクラスタ用NICの開発に際し，1999年にMEMOnet[5]というメモリスロットインタフェースをメモリ以外に使用するコンセプトを提唱した．さらにPC133仕様のSDR型DIMMスロットに装着可能なDIMMnet-1プロトタイプを作成して，その有効性を示してきた．

一方，ようやく2004年6月に，PCIバスの後継としてPCI ExpressがIntel社のi915・i925チップセットによってCOTSなPC市場に持ち込まれた．時を同じくしてメモリスロットインタフェースについても現在主流のDDRだけでなく，DDR2も前記チップセットから使える状況となった．この10年に1度という頻度で起こるPCの標準インタフェースの交代時期にあたり，メモリスロットインタフェースの意義について改めて考察する必要があると考えている．

本報告では，まずPC向け標準インタフェースを概観し，その問題点を述べる．次にメモリスロットインタフェースと，そのメモリ以外への応用に際しての問題点と利点を述べる．そのプロトタイプとしてDIMMnet-1およびFPGA版DIMMnet-2について紹介する．DIMMnet-1およびFPGA版DIMMnet-2のようなハードウェアの利用を念頭に，NIC，高機能メモリモジュール，リコンフィギャラブルコンピューティングの三つの応用面から見たメモリスロットインタフェースの得失について考察する．

2. PC向け標準インタフェースの現状とその問題点

本章では，PCI Express元年とも言える現段階で，低価格な高性能PCクラスタを構築していくことを念頭に，PC向け標準インタフェースの現状とその問題点について述べる．

2.1 PCI

1993年にIntel社のパーソナルコンピュータ(PC)用チップセット420TX・430LXに導入されて以来，PCIバスはPCのインタフェースのデファクトスタンダードの地位を現在まで続けてきた．当時としては32bit幅33MHzで133MB/sというバンド幅は高いものであった．

しかし，後述するAGPに分離されたビデオカードではもちろんのこと，最近ではギガビットイーサのカード1枚でもそのバンド幅は溢れてしまう状況にある．一部のサーバー機ではビット幅や周波数がそれぞれ2倍に改善されたPCIバスを用いてきた．しかし，それらは低価格なCOTSのPCには導入されては来なかった．

ビデオカードやNICのようなPCIバスでは足りない分野におけるその地位はPCI Expressに移行されようとしている．現在でも多くの拡張用カードはこのPCIバスで十分なケースが多く，それらは今後もしばらくはPCIバスを利用していくものと考えられる．

2.2 PCI-X

PCI-Xはサーバー機においてネットワークやストレージコントローラなどの用途に向け，PCIバスの周波数やビット幅を向上させてバンド幅を改善したインタフェースである．1GB/sのバンド幅が基本であるが，最近2GB/sの仕様もできた．

Myrinetや Infiniband仕様のPCクラスタ向けNICも，PCI-Xを採用したものが商品化されている．ただし，PCI-X関連の製品はサーバー機向けのニッチな市場向けであるため，サーバー機本体も含め高価であり，一般ユーザ向けのCOTS

†1 (株) 東芝，研究開発センター
Corporate Research and Developpment Center, Toshiba
†2 横浜国立大学
Yokohama National University
†3 東京農工大学
Tokyo University of Agriculture and Technology
†4 慶應義塾大学
Keio University

とは言い難い価格設定がなされるのが常である．よって，価格性能比の高い PC クラスタへの用途には，全体としてシステムのコストが高くついてしまう．

一方，Apple 社の Power Mac G5 の一部の機種では PCI-X を装備したものが販売されているので，PCI-X は高価という傾向は若干緩和する方向に向かっていると言える．しかし，Apple 社の PC 自体が Intel アーキテクチャの PC に比べてニッチである上，サーバー向けストレージ等が必要なユーザは，その中でもさらにごく一部となるため，低価格化は限定的なものと考えられる．

2.3 PCI Express

本年 6 月に，PCI バスの後継インタフェースとして PCI Express が Intel 社の i915・i925 チップセットによって PC 市場に持ち込まれた．論理的には PCI 互換であるので PCI 向けに記述されていたソフトウェアの書き換えが不要である．ハード的には 2.5Gbps の全二重シリアル通信路で構成され，8B10B 変換後のユーザーバンド幅は 250MB/s である．これを X1 として，これらを 2,4,8,16,32 レーン束ね合わせたものとして X2,X4,X8,X16,X32 が定義されている．

i915・i925 チップセットにおける PCI Express はサウスブリッジ側に一般拡張カード用の X1 が 4 本と，ノースブリッジ側にビデオカード向けの X16 が 1 本搭載されている．後者は PCI Express for Graphics と呼ばれており，仕様上はそこにビデオカード以外のカードを装着すると X1 のポートとして動作するようになっている．

一般ユーザにとってはビデオカード以外に X1 の PCI Express で当面足りなくなるという用途は見出しにくいのが現状であり，ビデオカード以外でも使えるようにしておいて，冷却問題などの種々の問題が起きて，PCI Express への移行という大目的に支障をきたすという可能性を封じておくのは半ば必然の戦略と言える．

このため，例えば PCI Express 8X 仕様の InfinibandNIC をここに装着しても，動いたとしても X1 モードで動作してしまうため，PCI-X バージョンのカードと比べてバンド幅が片方向で 1/4，双方向でも 1/2 になってしまい，メリットがない．

一方，サーバー機向けのチップセット E7525 ではビデオカード以外のスロットでも X8 をサポートしているものがあるため，PCI Express 8X 仕様の InfinibandNIC などの高性能 NIC は PCI-X の時と同様，少なくとも当面はサーバー機でのみ利用可能ということになる．

PCI Express はバンド幅的には PCI や PCI-X と比べて X16 や将来製品が出てくるであろう X32 まで考えれば相当高速化がなされている．しかし，トランザクション層，リンク層，物理層として以下のような回路を通過する．

表 1 PCI-Express における送受信側の回路

送信側	受信側
Header 生成	デシリアライザ
送信バッファ制御	エラスティックバッファ
FlowControl Credit 検査	10b/8b エンコード
VirtualChannel 調停	PIPE インターフェイス
TLP フレーミング	デスクランブリング
ECRC 生成	レーン to レーンデスキュー
LCRC 生成	受信デフレーミング制御
送信パケット調停	シーケンスナンバー検査
送信フレーミング制御	LCRC 検査
スクランブリング	パケット認識
PIPE インターフェイス	ECRC 検査
8b/10b エンコード	受信バッファ制御
シリアライザ	アプリケーションインターフェイス

これらのハードウェアオーバーヘッドが大きく，例えば PCI Express 上のレジスタを 1 回リードするにも往復でこれらの回路ブロックを通過することになり，エラスティックバッファが 125MHz7 クロック程度である他は 1〜2 サイクル程度の段になっており，500〜600ns 程度はかかってしまうことが予想される．

PCI Express を採用すれば長いデータのアクセスに関してはバンド幅が向上する分だけ転送時間が減るため全体として遅延時間が短くなる．しかし，短いデータについては転送時間は少なく PCI Express を採用しても改善は見込めないため，短いデータの読み出しが多用されるような状況下では効率が悪くなってしまう点が問題である．

PCI Express X1(250MB/s 全二重) のバンド幅でも足りないという一般ユーザの用途が新たに登場してこない限り，再び 10 年というスパンで PCI Express X1 が COTS の PC 向け標準インタフェースとして君臨し，進歩を止めてしまう可能性も低くないと考えられる．

3. メモリモジュールの現状と近未来

本章では，メモリスロットインタフェースを用いた NIC により低価格な高性能 PC クラスタを構築していくことを念頭に，現段階で PC に用いられているメモリモジュールの現状と，一部，近未来に用いられる可能性の高いメモリモジュールについて述べる．

3.1 SDR DIMM

SDR 型 SDRAM をベースにした DIMM は PC66, PC100, PC133 の規格のものがあり，1990 年代末から数年の間，PC のメモリモジュールとして最も多く使われてきたメモリモジュールである．64bit 幅のデータバスを用いてノースブリッジからの 66MHz〜133MHz のクロック信号に同期して 1 サイクルにつき 64bit のデータを転送するため，PC133 では 1GB/s のバンド幅を有する．

後述する DIMMnet-1 プロトタイプはこの SDR 型 DIMM の PC100,PC133 インタフェースで DIMM スロットに装着される．

組込み製品などでは今でも使用されているが，DDR 型の DIMM の台頭により，現時点では SDR 型 DIMM を使用できる COTS のマザーボードは市場から消滅しつつある．

3.2 RIMM

RIMM は Direct Rambus 型 DRAM を搭載したメモリモジュールで，i820・i840・i850 チップセットにおいて用いられた．SiS 社からは SiS R659 チップセットでは 4 本の PC1200 仕様の Direct Rambus チャネルを有し，合計で 9.6GB/s ものメモリバンド幅を実現している．

RIMM には 16bit 幅のものと 32bit 幅のものの 2 系統があり，1 枚のメモリモジュールに対して現状で最も高いバンド幅を供給できるのが 32bit 幅の RIMM である．

DIMM と異なり RIMM では基板の配線の実装に関する規定があるためにデータ線のねじれという問題は無い．また，RAC と呼ばれる Rambus チャネル用の IP は RDRAM 側とコントローラ側では同一で，動作モードの設定により送信側にも受信側にもなり得る．よって，ホストのチップセット側から見て RIMM に見えるような ASIC を作成することは技術的には可能である．

ただし，メモリモジュールの価格や，Intel 社の Rambus からの撤退や，後述する DDR 型 DIMM の台頭により，市場的には RIMM を採用する PC は消滅に向かっていると考えられる．

3.3 DDR DIMM

DDR 型 SDRAM をベースにした DIMM は現時点で PC のメモリモジュールとして最も多く使われているメモリモジュールである．最新の i915・i925 チップセットでもマザーボードによっては使用することができる．64bit 幅のデータバスを用いてノースブリッジからの 66MHz〜200MHz のクロック信号に同期して 1 サイクルにつき 64bit のデータを立ち上がりと立下りの 2 回転送するため，200MHz の PC3200 では 3.2GB/s のバンド幅を有する．

後述する FPGA 版 DIMMnet-2 プロトタイプはこの DDR 型 DIMM の PC1600(DDR200) インタフェースで DIMM スロットに装着される．

3.4 DDR2 DIMM

DDR2 型 SDRAM をベースにした DIMM は現時点では DDR2 400,DDR2 533 の規格のものが存在し，i915・i925 チップセットなど，現時点で最新の PC で採用されているメモリモジュールとして登場したばかりのメモリモジュールである．DDR400 と DDR2 400 はバンド幅的にはどちらも 3.2GB/s であり，現状では価格に 4 倍程度の開きがあり，CPU の FSB バンド幅がボトルネックとなるため DDR2 を採用してもモジュール価格の高さの割には処理性能が向上しないため，当面あと 1 年程度は主流にはならないと考えられる．

しかし，DDR が DDR400 で打ち止めであるのに対し，DDR2 は DRAM 内部でのコア周波数が DDR の半分で動いている点や，ストローブ信号の差動信号である点などの理由により，さらなる高速化が確実視されている．消費電力も DDR2

の方が少なくて済むので，望ましい．さらに CPU の FSB バンド幅がさらに高速化してバンド幅的に DDR のデュアルチャネルでは足りなくなってくると，DDR2 への移行は促進される．よって年単位のスパンで考えれば DDR2 が次の主流になることはほぼ確実と考えられる．

以上の観点から，1〜2 年後の完成を目安に ASIC 化した DIMMnet-2 を作成することを想定すると，DDR2 ベースのインタフェースで作ることが望ましいと考えられる．

3.5 FB-DIMM

近未来に用いられる可能性の高いメモリモジュールとしては FB-DIMM[3] がある．Intel 社は 2005 年からサーバー市場向けに FB-DIMM が出荷されることを発表 [4] している．この FB-DIMM は後述する DIMMnet-2 のごとく，メモリバスに対して分岐を無くして負荷を 1 つしかみせないようにすることで高速化をはかる技術を用いた新型メモリモジュールであり，AMB(Advanced Memory Buffer) と称する LSI が数珠つなぎに接続され，DRAM チップは AMB に接続される．

FB-DIMM がサーバーのみに利用されるのであれば，その上では PCI Express X8 などを NIC 用に使うことができると考えられ，FB-DIMM 向けの DIMMnet を作る意義は高くない．しかし，この技術が COTS 側に転用されてくる場合は，FB-DIMM 向けの DIMMnet は効果的と考えられる．

4. メモリスロットインタフェースの利点

本章では，メモリスロットインタフェースの利点について，バンド幅，遅延時間，コストの 3 つの観点から述べる．

4.1 バンド幅

4.1.1 能力の高さ

メモリバスのバンド幅は 1 本あたり DDR400 で 3.2GB/s，DDR2 533 で 4.3GB/s あり，PCI-X よりも数倍高速であり，PCI Express X8 の双方向バンド幅 4GB/s に匹敵する NIC に使うには十分高い水準にある．PCI Express X8 は前述の通りサーバー機でしか NIC には使えないのに対し，DDR400 や DDR2 533 は安価な COTS の PC 上でこのバンド幅を NIC 等のために使用することができる．

4.1.2 進歩の継続性

PCI バスは 11 年の長きにわたり，少なくとも COTS の PC 上では性能が進歩しなかった．そのような実績や，COTS の PC の利用形態におけるニーズを鑑みると，その後継である PCI Express についても COTS の PC 上では性能が進歩しない可能性が低くない．これに対して，メモリスロットのバンド幅は CPU の能力や，それを支える FSB のバンド幅とともにムーアの法則に則った形で継続的に進歩してきたし，今後もその傾向は継続されることは確実である．

本年は PCI Express 元年にあたるため，短期的な視野でメモリスロットインタフェースを PCI Express との比較の上で評価すると，その意義が見えにくくなるが，5 年 10 年というスパンで両者の進歩の状況を比較していけば，CPU 能力とのバランスが約束されているメモリスロットインタフェースの意義は，年々高まってくると考えられる．

4.2 遅延時間

Pentium3(core850MHz-FSB100MHz-メモリ PC100) という環境での uncached 属性の主記憶領域への 8 バイトリード時の遅延の測定値は 173ns であった．これに対し，PCI Express では 500ns 程度が予測されており，メモリバスへのアクセスの方が約 3 倍程度遅延時間が少ないと考えられる．

同期フラグのポーリングなど，少ない量のデータに対してリードを頻繁に行なわねばならない用途ではこの差は大きく，処理時間全体にも響いてくるケースがありうる．

4.3 コスト

10Gbps クラス以上の高バンド幅な NIC を利用することを考える場合，標準インタフェースを採用するならば PCI-X2.0(双方向で 2GB/s) または PCI Express X8(双方向で 4GB/s) が必要になってくる．しかし，これらは高価なサーバー機でないと利用することができない．

PCI-X1.0(双方向で 1GB/s) であれば，Power MAC G5 の一部の機種には装備されているのでこれらを用いた PC クラスタを構築することもできないこともないが，送受信が同時に起こる場合はバンド幅が足りないし，Intel ベースの COTS な PC と比べて PC 本体が 3 倍程度のコストとなることを覚悟する必要がある．

これに対して，メモリスロットインタフェースによれば，DDR 型 DIMM で DDR200 ベースでも 2.1GB/s，最新の DDR2-533 ベースでは 4.3GB/s のバンド幅を安価な COTS の PC 上で提供できるため，10Gbps クラス以上の高バンド幅な NIC を安価な PC 上で利用することが可能となる．

10Gbps クラス以上の高バンド幅な NIC を利用する PC クラスタ用ノードの想定仕様とその予想コストを**表 2** に示す．

表 2 10Gbps クラスの NIC を利用する PC クラスタ用ノードの想定仕様とその予想コスト

	MAC G5	IA server	COTS PC
本体コスト	30〜40 万円	30 万円〜	10 万円
10G NIC 種類	IB	IB	DIMMnet-2
通信リンク仕様	IB 4X	IB 4X	IB 4X
通信バンド幅	2GB/s	2GB/s	2GB/s
ホスト I/F 種類	PCI-X1.0	PCI Exp.X8	DDR2-533
I/F バンド幅	1GB/s	4GB/s	4.3GB/s

5. メモリスロットインタフェースの問題点

本章では，DIMMnet-1 プロトタイプの試作の経験から抽出された問題点について述べる．これらのうち (1)〜(3) は DIMMnet-2 によって克服されるが，(4)(5) は本質的に完全な克服は不能であり，それを受け入れた上で問題が大きくならないような対応が DIMMnet-2 においては講じられる．(6) は性能と引き換えに本質的に受け止めなければならない問題である．

5.1 データ配線のねじれ

マザーボード上のデータ線配線に関する規定が無いため，マザーボードによってはデータ線がチップセット上のビット番号と DIMM 上のビット番号で不整合な配線（ねじれた配線）になっていることが大半であった．中にはバイト単位でねじれた上に 1 バイト内でもねじれているものも存在した．このねじれは単純なメモリとして使う分にはこれでも問題にならないが，NIC 等の別の目的に使う場合にはねじれを解消するための手段が必要になる．

DIMMnet-2 においては FPGA 版では FPGA の配線の可変性を利用してねじれ解消を行なうが，ASIC 版においてはホスト側から 1bit ずつ bit を変化させたデータを繰り返し書き込み，その位置情報を 64bit バスを 1bit として使うことでホストに送り返すことにより，ねじれ解消情報を生成し，ロジックアナライザ等を使用しなくともねじれが解消できるような仕組みを導入することで解決する．

前述の RIMM や FB-DIMM に対応した MEMOnet を想定した場合は，このようなデータ線ねじれの問題は無いと考えられる．

5.2 アドレスマルチプレクス規則が不統一

チップセット毎にアドレスのマルチプレクス規則が異なる．このため，DIMM に 2 回にわけて入力されたアドレス信号から，ホスト CPU が出力したアドレスを復元するための論理回路に柔軟性があることが望ましい．仕様を公開していないチップセットメーカーもあるので，何らかの方法でその規則を抽出する機構を有することが望ましい．

5.3 分岐配線による動作マージン減少

PC133 程度の周波数でも動作マージンが少なくなるため，バンク切り替えスイッチなど，メモリバス上に分岐配線を作るような構造を取ることは，より高い周波数領域を目指す場合は好ましくない．

DIMMnet-2 においてはバンク切り替えスイッチの機能は FPGA または ASIC の論理として実装し，DIMM スロット側にはその FPGA または ASIC 一つが負荷として接続されるように構成することで，DDR 型の DIMM に対応する．

このため，少なくとも全ての DRAM 領域へのリードアクセスに関しては，DIMMnet-1 のように通常の DRAM と何ら変わらないアクセスをすることはできなくなっている．常にプリフェッチコマンドでプリフェッチ Window という高速メモリに読みたい領域をプリフェッチしておき，プリフェッチ完了後にプリフェッチ Window をバーストリードするという方式を取る．

前述の FB-DIMM もこの考え方と類似しており，ASIC で分岐を断ち切り，DRAM を接続した ASIC 間を高速シリアル

リンクで接続している．その場合，メモリスロット自体に分岐がなくなっているので，DIMMnet-2 と同様な実装方式で対応可能と考える．

5.4 最大容量で限定されたアドレス空間サイズ

1 本の DIMM に到達するアドレス信号は DIMM の最大容量で限定される．よって CPU が 64bit 化されてより広大な空間を指し示せるアドレスが CPU から出力されたとしても，1 本の DIMM に到達するアドレス信号でリモートノードの全メモリをそこにマップすることは事実上不可能である．よって DIMM 上に全てのリモートメモリをアクセスできる機能を実現するには，リモートアドレスはデータとしてメモリモジュール側に伝達する手段が必要である．

このため，DIMMnet-1 および DIMMnet-2 では，リモートノードのメモリの一部を AOTF 送信機構のヘッダー TLB というハードウェアに対応関係を設定することでローカルアドレス空間内に一部のリモートメモリをマップする．さらに BOTF 送信機構や RDMA 送信機構に対してはリモートアドレスをデータとして Window メモリ経由で伝達することで，リモートメモリへのアクセスを可能にしている．

5.5 主記憶として利用可能なスロットの減少

メモリスロットインタフェースを採用すると，そのデバイスによりスロットが消費されるので，結果として主記憶として利用可能なスロット数が減少する．通常の仮想記憶においてページングされる主記憶がその PC の全メモリスロットを消費して実現される容量がないとスラッシングにより事実上動かない応用に対しては不利になる．

DIMMnet-2 においては，この問題に対しては，DIMMnet-2 上にメモリスロットの容量限界以上のメモリを搭載可能にし，ページフォルトハンドラにおいて通常は HDD に行くべきところを，可能な状態であれば裏のメモリバンクとの間でスワップイン・スワップアウトを行なわせることで通常の主記憶として使える領域が少なくなることによる悪影響を軽減可能とする予定である．この手法はページフォルトの頻度は上がってしまうので多少の性能低下は発生するが，HDD との間でのスラッシング状態にはならないので，劇的な性能低下は起こらないようにできるものである．

5.6 寿命の短さ

DIMMnet-1 は SDR 型 SDRAM ベースの DIMM インタフェースをホストインタフェースとするが，そろそろそのような仕様の主記憶を有する COTS な PC は入手が困難になってきている．2000 年近辺での最新のメモリ規格である PC133 の COTS な PC での利用は 3 年ほどで終息に向かった．このように，PCI バスのようにその寿命が 10 年以上も続く標準インタフェースに比べ，メモリスロットの寿命は短いといわざるを得ない．よって，ホストインタフェースの部分だけでもそのようなサイクルで更新していかねばならない．3 年もすればゴミ同然になってしまうと言われる COTS な PC の世界では，この寿命の短さはムーアの法則に従って成長する CPU 性能とのバランスと引き換えに甘んじて受け止める必要がある．

インタフェースの寿命が短いがゆえに，極力先物の仕様を採用するか，あるいは極力開発期間を短く抑えないと，でき上がった頃には使えるマザーボードが入手困難になるという状況になりうるため，注意が必要である．

6. プロトタイプ

我々は，CPU 性能と通信性能がバランスした高性能な PC クラスタ用 NIC の開発に際し，1999 年に MEMOnet というメモリスロットインタフェースをメモリ以外に使用するコンセプトを提唱し，PC133 仕様の SDR 型 DIMM スロットに装着可能な DIMMnet-1 プロトタイプを作成して，その有効性を示してきた．その改良版として DDR 型 DIMM スロットに装着可能な DIMMnet-2 プロトタイプを作成中である．本章ではこれらの二つのプロトタイプの概要を述べる．

6.1 DMMnet-1 プロトタイプ

DIMMnet-1 プロトタイプは PC133 仕様の SDR 型 DIMM スロットに装着可能な PC クラスタ向け NIC の最初のプロトタイプである．コントローラ LSI としては RWCP にて開発した Martini という ASIC を用いており，この Martini は PCI バージョンの NIC である RHiNET とも共通に使われている．

その構成を図 1 に示す．このように，バンク切り替え用の FET スイッチが DIMM スロットにスタブ（分岐）を増やす方向で追加されるために，電気的なマージンを食いつぶし，SO-DIMM のソケットを用いた実機上では周波数的には 100MHz では安定動作するものの，133MHz での安定動作するまでの調整はできなかった．ソケットを用いない実機では 133MHz で動作したが，微妙な配線長の変化やコネクタの反射などの影響の変化によってギリギリの状態で動いているものと思われる．

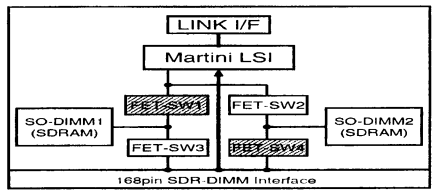

図 1 DIMMnet-1 基板の構成

通信機構としてはリモートメモリの一部をローカルな仮想記憶空間にマップして 1〜8 バイトの細粒度なリモートアクセスを高速化する AOTF(Atomic On-The-Fly) 送信機構，キャッシュラインサイズ程度の長さの 1〜512 バイトの短めのリモートアクセスを高速化する BOTF(Block On-The-Fly) 送信機構，より長いデータ長のリモートアクセスを高速化する RDMA(Remote DMA) 送受信機構，AOTF で送信されたリモートアクセスを高速に受信する Mini OTF 受信機構，Mini OTF 受信機構から書き込まれるリモートアクセスを低遅延でホストに伝達することを目的とした低遅延マルチポートメモリである LLCM(Low Latency Common Memory) を有する．

内部の動作周波数が予定の半分で実装された実機上でも，ホストからの書込み→ AOTF 送信機構→ Mini OTF 受信機構 → LLCM →ホストからの読み出し，という経路での通信は片道 1 マイクロ秒を切っており，この低遅延通信を用いて高速なバリア同期などの大域演算が実装された．

6.2 FPGA 版 DIMMnet-2 プロトタイプ

FPGA 版 DIMMnet-2 プロトタイプは 2004 年 3 月に基板が完成し，現在 FPGA 内部論理を作成中のプロトタイプで，動作周波数的に ASIC よりは不利な状況にある FPGA ベースでの論理検証を主目的に DDR200 スロットでの動作を目標に設計を進めている．

その構成を図 2 に示す．DDR での動作をさせるため，DIMM スロットからのスタブ (分岐) を作らないように構成されている．その結果，現状では DDR200 に設定されたホスト側から DIMMnet-2 基板上の SO-DIMM への読み書きができる状態になっており，分岐配線回避の方向での DDR 化というコンセプトが成立することを実験確認した．

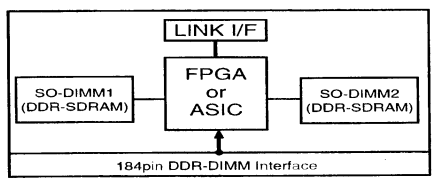

図 2 DIMMnet-2 基板の構成

通信機構は概ね DIMMnet-1 のものを踏襲するが，Martini 内部での TLB のリフィルをハード化したり，リモート間接書き込みのハード化をするなど，従来ファームウェアで実装されていたため大変遅かった重要な動作に関して，低速な内蔵 CPU の助けを極力借りない方針で，種々の高速化に関する最適化を施している．

なお，DIMMnet-2 では PC クラスタ向けの NIC としての利用のみならず，プリフェッチ機能を有するメモリモジュールとしても機能するように設計されている．そのプリフェッチ機能には連続アクセスのみならず，等間隔アクセスやリストベクトルアクセス（間接参照）もサポートされており，ホスト CPU がとっているキャッシュアーキテクチャでは非常に効率が悪い種類のアプリケーションに対しても，キャッシュが有効に働くようにすることができる．

DIMMnet-2 のプリフェッチコマンドはプリフェッチ Window

と称する一種のベクトルレジスタに対してベクトルロードを実行するようにできている．ホスト CPU 側からはプリフェッチが完了したことをフラグを検査することで知ることができる．このフラグのポーリングがメモリスロットインタフェースを採用している DIMMnet-2 においては PCI Express などの標準インタフェースと比べて高速に実行できる点が重要である．

7. 応用からみたメモリスロットインタフェース

本章では FPGA 版 DIMMnet-2 のようなハードウェアの利用を念頭に，NIC，高機能メモリモジュール，リコンフィギャラブルコンピューティングの三つの応用面から見たメモリスロットインタフェースの得失について考察する．

7.1 NIC 応用

NIC 応用の観点からは，DIMMnet-1 プロトタイプ開発時は主に高いバンド幅が安価な COTS な PC 上で実現でき，10Gbps の送受信を同時に行ってもワイヤスピードで送受信ができる点がメモリスロットインタフェースの際立った利点であった．

一方，4X タイプの Infiniband も 10Gbps の通信リンクを有するが，そのホストインタフェースは PCI-X1.0 で実装されていたために単方向通信時はホストインタフェースがネックにならず 900MB/s 程度のバンド幅が出るが，送受信が同時に起こる双方向通信実行時には 500MB/s 弱までバンド幅が低下してしまうことが報告 [12] されている．POWER MAC G5 の一部の機種が PCI-X1.0 を装備しているため，今ではその上で Infiniband の NIC が動作可能だが，上記同様のバンド幅不足となることは確実である．

さらに Infiniband は現在の主流の 4X(10Gbps) に加え，最近ではスイッチ側は 12X(30Gbps) のポートを有するスイッチ製品が出てきており，スイッチ間は 12X で接続可能だが，NIC 側は PCI-X では大幅に能力不足のため事実上対応できない．DIMMnet-2 であれば DDR DIMM インタフェースを採用しているので，DDR400(3.2GB/s) や DDR2-533(4.2GB/s) に対応すれば 12X の Infiniband スイッチとの接続に対しても概ねワイヤスピードでの送受信が可能となる．ただし，このクラスの NIC にはサーバー機に限定されるものの PCI Express X16(片方向 4GB/s) によっても対応は可能と考えられる．

しかし，PC クラスタの全体処理性能を律速してしまいがちなバリア同期や総和などの大域通信においては，細粒度通信となるため PCI Express の通信バンド幅高速化による遅延時間短縮効果はほとんど期待できず，シリアル通信に伴うオーバーヘッド増加分が直接性能低下に響いてしまう．

一方，メモリスロットインタフェースでは DIMMnet-1 で確立された，ホストからの書込み → AOTF → Mini OTF 受信機構 → LLCM → ホストからの読み出し，という経路での高速細粒度通信に適しており，高速な大域通信が実現できる．

7.2 高機能メモリモジュール応用

DIMMnet-2 では PC クラスタ向けの NIC としての利用のみならず，プリフェッチ機能を有するメモリモジュールとしても機能するように設計されている．そのプリフェッチ機能には連続アクセスのみならず，等間隔アクセスやリストベクトルアクセス（間接参照）もサポートされており，ホスト CPU がとっているキャッシュアーキテクチャでは非常に効率が悪い種類のアプリケーションに対しても，キャッシュが有効に働くようにすることができる．

DIMMnet-2 のプリフェッチコマンドはプリフェッチ Window と称する一種のベクトルレジスタに対してベクトルロードを実行するようにできている．ホスト CPU 側からはプリフェッチが完了したことをフラグを検査することで知ることができるが，このフラグポーリングはデータ長が少ないリードの連続となるため，リード遅延が性能に大きく影響を及ぼす．その影響を評価するために，プリフェッチ機能付きメモリモジュールを用いた NAS CG ベンチマークの実行に際して，ポーリング遅延を変化させた時にどの程度の処理速度向上率への影響が出るかを，文献 [11] に記載の評価プログラムに対して，空ループにより人工的に作り出した 100ns 単位の遅延時間をポーリング部分で挿入することで調査した．なお，NAS CG ベンチマークは RWCP 製 C 言語版を改造したものを用いた．Class C の実験については Linux のメモリ割り当ての制約上動作不能であるため，一部の巨大な static 配列を，calloc 関数によって main() の最初の部分で 1 回確保することで動作させた．測定環境を**表 3** に示し，測定結果を**図 3** に示す．

前述のようにメモリバスへのアクセス遅延は 100ns 台であり，PCI Express へのアクセス遅延は 500ns 以上であると推定されている．PCI Express のバンド幅が主記憶バンド幅と同一だったとしても，図 3 に示されるように遅延が 100ns 台の DIMMnet 方式に比べ，500ns 以上の PCI Express はバンド幅がたとえ主記憶と同等であったとしても明らかな性能低下を起こすことが判る．

表 3 評価環境

機種名	Dell プレシジョン 360
CPU	Pentium4
FSB 周波数	800MHz
コア周波数	2.4GHz
L1 キャッシュ容量	8KB
L2 キャッシュ容量	512KB
L1 キャッシュラインサイズ	64B
L2 キャッシュラインサイズ	128B
メモリ種類	PC3200 (DDR SDRAM)
メモリバス本数 (総バンド幅)	2 (6.4GB/s)
メモリ容量	4GB
OS	Linux 2.4.20-8
コンパイラ	gcc 3.2.2
最適化オプション	-O3

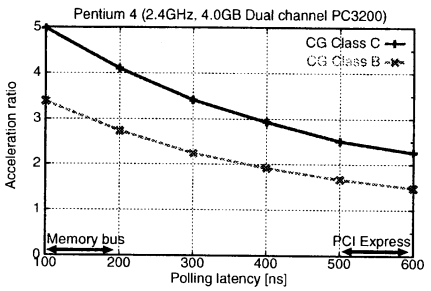

図 3 プリフェッチ機能付きメモリモジュールによる NAS CG における処理速度向上率のポーリング遅延による変化

CG ベンチマークはメモリバンド幅が極めて敏感に効くアプリケーションであるので，主記憶バンド幅の 63%の 4GB/s のリードバンド幅を有する X16 仕様の PCI Express Graphics 側ではなく，主記憶バンド幅の 3.9%の 250MB/s のリードバンド幅しか持たない X1 側にこのプリフェッチ機構を実装してしまうと，さらに一桁の性能低下を起こし，通常の DIMM を使ってキャッシュミスヒットを多発させつつ実行する処理速度 (ClassC で 55.66MFLOPS) より遅くなることが予想される．

このように，高機能メモリモジュールへの応用を考えた場合には，ホストインタフェースは PCI Express では不十分であり，低遅延性と高バンド幅の両方を兼ね備えるメモリスロットインタフェースとする意義は大きいと言える．

7.3 リコンフィギャラブル応用

DIMMnet-2/NI の試作ボードは,NIC 用 ASIC の開発用プロトタイプとしての役割を果たすだけでなく，新しいリコンフィギャラブルシステムとして優れた特徴を持っている．

ホスト PC に大規模な FPGA とメモリを搭載したボードを接続し，ホスト上での処理の一部を高速化するアクセラレータとしてのリコンフィギャラブルシステムの利用が進んでいる．これらのリコンフィギャラブルなアクセラレータは以下の応用分野で利用が盛んである．

- 単純な数値演算ではなく，ビット処理，比較，多数のメモリアクセスを含む．
- 処理アルゴリズムが対象によって柔軟性が必要になる．場合によっては，処理用のハードウェア構成自体を対象に適合するように変更することが性能向上に有用である．
- ニッチ市場であり，専用 ASIC を開発するほど多数用いられない．
- 粗粒度にジョブを分割することが難しく，PC クラスタなどで並列処理を行うと，通信処理がボトルネックになる可能性がある．

これらに当てはまる，ボリュームレンダリングなどのグラフィックス処理 [13]，グラフ同定問題等の計算困難な問題 [14]，細胞シ

ミュレーション[18]などのバイオインフォマティクス分野の大規模演算等ではホストCPUの数十倍から数百倍の性能向上が報告されている．さらに，最近は，FPGA自体の性能向上に加えて，IPとして組み込まれた演算器を用いることで，大規模な数値演算に対する応用[16]が活発化している．

これらのリコンフィギャラブルアクセラレータの多くは，PCIバスでホストCPUに接続されており，ホストCPUとの協調処理を行うと，PCIバスの転送遅延と容量ボトルネックが問題となる場合がある．例えばPCIバスに接続するReCSiPボード上でホストとの協調作業で細胞シミュレーションを実行する場合，通常の実装方法ではPCIバスの転送容量がボトルネックとなり，処理性能が向上しないことが報告されている[18]．

このボトルネックを解消するために，メモリバスを利用する方法がいくつか試みられている．Pilchard FPGA co-processor platform[15]は，PC100, PC133のDIMMに接続され，小規模なVirtexを用い，ボード上にアプリケーション用のメモリは搭載していない．Nuron社からアナウンスされたAcBは，同様にメモリバスに接続され，FPGAおよび64MByteのメモリを搭載しているが，プロトタイプのみで終わっている☆．スイスの連邦研究所で開発されているTKDM[17]は，これらの先行研究を踏まえたシステムで，2種類のFPGAと大量のメモリを搭載している点に特徴がある．

しかし，これらのシステムは，全てがSDR-SDRAM対応であり，現在のPCに接続することが困難である．また，NICを装備していない．DINNnet-2/NIは，DDR-SDMAMに対応することで，これらのシステムの2倍の転送容量を実現すると共に汎用PCへの接続の道を開いている．PCI-Expressなどの普及により，将来，転送容量の点ではメモリバスと同等のI/Oバスの利用が可能になっても，レイテンシィの点ではメモリバスが有利である．現在，PCIバスに接続するリコンフィギャラブルアクセラレータは，PCIバスのレイテンシィを考慮して，大きなレイテンシィがあまり性能に影響しない実装法を取っている場合が多い．しかし，メモリバスへの接続により，ホストPCとリコンフィギャラブルシステムがより密接に協調して処理を行うことが可能となる．すなわち，DIMMnet-2/NIは，アクセラレータとしてのリコンフィギャラブルシステムの応用方式を広げる可能性を持つ．

さらに，DIMMnet-2/NIは，従来のメモリバス接続型リコンフィギャラブルシステムに比べて大容量のメモリを2スロット有し，大量のデータ記憶と同時アクセスが必要なアプリケーションに対応する．また，ボード間交信用にInfinibandインタフェースを持つ点も大きな特徴である．このNICを利用することで，ホストとの協調処理，転送に付随する処理，他のボードとの負荷分散を含めた新しいリコンフィギュラブルシステム実現の可能性を開くことができる．

8. ま と め

本報告では，PC向け標準インタフェースを概観し，その問題点を述べた．特に，PCI Expressはポスト PCIとして期待されているが遅延時間に関しては問題を有する．さらに，メモリスロットインタフェースの現状と近未来を概観し，そのメモリ以外への応用に際しての問題点と利点を述べた．DDRスロットでの動作を疑問視する見方もあったが，FPGA版DIMMnet-2プロトタイプ作成による実験確認により，その疑いは払拭されたものと考える．

さらにDIMMnet-1およびFPGA版DIMMnet-2のようなハードウェアの利用を念頭に，NIC，高機能メモリモジュール，リコンフィギャラブルシステムの三つの応用面から見たメモリスロットインタフェースの得失について考察した．これら全てに関してバンド幅の有効性を明らかにしたとともに，前者2つに関してはPCI Expressでは到達し得ない低遅延性の有効性についても明らかにした．

高機能メモリモジュール応用においては，定量的評価からPCI Expressではその遅延特性から主記憶同等のバンド幅(6.4GB/s)を実装できたとしても性能低下が大きく，低遅延性と高バンド幅の両方を兼ね備えるメモリスロットインタフェースとする意義は大変大きいことを確認した．

リコンフィギャラブルシステム応用に関しては低遅延性の有効性については現時点では明らかではないが，未開拓な分野であるため，今後の研究により有効性が発揮される分野が開拓される可能性を秘めている．

今後は，FPGA版DIMMnet-2の内部論理を完成させ，実機評価を進めるとともに，並列処理のためのソフトウェア環境や，プリフェッチ機構の利用を容易にするコンパイラ等のソフトウェア環境の整備を行なう予定である．

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．PCI Express for Graphics上のビデオカード以外の動作についてご教授いただきました Intel(株) の鈴木氏，PCI Expressの IPにおける処理内容の概要をご教授いただきました (株) 図研の中村氏，Rambus のRACについてご教授いただきました Rambus 社の太田氏に感謝いたします．DIMMnet-2の開発に関する議論にご参加いただいている横浜国立大学の安藤氏，立命館大学の国枝教授，和歌山大学の齋藤講師，平石氏，笠松氏，京都大学上原助教授，東京農工大学の並木助教授，浜田氏，荒木氏，森氏，慶應義塾大学の西助手，渡辺氏，大塚氏，北村氏に感謝いたします．DIMMnet-2基板作成をご担当いただいている日立ITの上嶋氏，岩田氏，今城氏に感謝いたします．

参 考 文 献

1) PCI-SIG, http://www.pcisig.com/
2) DDR2, http://www.memforum.org/memorybasics/ddr2/
3) FB-DIMM, http://www.memforum.org/memorybasics/fb_dimm/
4) Intel corp. "Intel Press Release (Feb.18, 2004)", http://www.intel.com/pressroom/archive/releases/20040218comp.htm
5) 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)
6) 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)
7) 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)
8) 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)
9) 田邊, 濱田, 中條, 天野："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", 情報処理学会計算機アーキテクチャ研究会, 2003-ARC-152, pp.61-66 (Mar. 2003)
10) 田邊, 土肥, 中條, 天野："プリフェッチ機能を有するメモリモジュール", 情報処理学会計算機アーキテクチャ研究会, 2003-ARC-154, pp.139-144 (Aug. 2003)
11) 田邊, 中武, 箱崎, 土肥, 中條, 天野："プリフェッチ機能付きメモリモジュールによる不連続アクセスの連続化", 情報処理学会計算機アーキテクチャ研究会, 2004-ARC-157, pp.139-144 (Mar. 2004)
12) J. Liu, B. Chandrasekaran, W. Yu, J. Wu, D. Buntinas, S. P. Kini, P. Wyckoff, and D. K. Panda："Micro-Benchmark Level Performance Comparison of High-Speed Cluster Interconnects", Hot Interconnect 11, (Aug. 2003)
13) 森, 出雲, 高山, 丸山, 津邑, 五島, 中島, 富田 大規模データの並列可視化を支援する FPGA 搭載 PCI ボード, 第1回リコンフィギャラブルシステム研究会予稿集 pp.15-20 (2003)
14) S.Ichikawa, S.Yamamoto, Data Dependent Circuit for Subgraph Isomorphism Problem IEICE Trans. on Inf & Syst. Vol.E86-D, No.5, pp.796–802 (2003)
15) P.Leong, M.Leong, O.Cheung, T.Tung, C.Kwok, M.Wong and K.Lee, Pilchard - a reconfigurable omputing platform with memory slot interface. In Proc. of IEEE Symp. on Field Programmable Custom Computing Machines(FCCM) (Apr. 2001)
16) 佐々木, 溝口, 長嶋 Car-Parrinello 計算向け三次元 FFT ロジックの開発 先進的計算基盤シンポジウム SACSIS2004 論文集, pp.407–414 (2004)
17) Plessl, C., Platzner, M., TKDM - a reconfigurable co-processor in a PC's memory slot, Field-Programmable Technology (FPT), 2003. Proceedings. 2003 IEEE International Conference on (Dec.2003)
18) Yasunori Osana, Tomonori Fukushima, Hideharu Amano, Implementation of ReCSiP: A ReConfigurable Cell SImulation Platform, 13th International Conference on Field Programmable Logic and Applications(FPL), 2003. Proceedings. (Sep,2003)

☆ Nuron社はIntelに吸収され現在参照可能な home pageが存在しない