2005－MPS－55（9）
2005／6／28

DIMMnet 通信インタフェース用パケット伝送レイヤ

濱田芳博[†1]　北村聡[†2]　西宏章[†2]
田邊昇[†3]　天野英晴[†2]　中條拓伯[†4]

PC クラスタ用のネットワークインタフェース (NIC) として開発された DIMMnet-1 は PC メモリバスへ直接搭載し，広帯域，低遅延な通信性能を目指したデバイスである．メモリバスは帯域拡張が PC の中で優先される箇所であるため，これを IO バスとして使用する本 NIC はその時点で利用可能な広帯域な通信リンクを使用して通信容量を拡張することが可能となる．しかし本 NIC が採用していた通信リンクは並列同期伝送を使用しており帯域の拡張性に欠く．この問題を解決するために高速シリアル伝送を通信リンクに採用した 2 種類の NIC（DIMMnet-1/bDais，DIMMnet-2）を開発し，現在 DIMMnet ネットワークプロトコルの拡張を行っている．これは大別して 2 つの部分より構成している．一つは通信プリミティブを含む CoreLogic であり，一つはスイッチを介してパケット伝送を行なう InfiniSWIF である．本論文では DIMMnet ネットワークプロトコルへ reliability と scalability を与える Concurrent Datagram 通信サービスの提案を行い，InfiniSWIF の設計と評価について示す．

A packet forwarding Layer for DIMMnet and its Hardware Implementationt

Yoshihiro Hamada ,[†1] Akira Kitamura,[†2] Hiroaki Nishi ,[†2]
Noboru Tanabe ,[†3] Hideharu Amano [†2] and Hironori Nakajo[†1]

DIMMnet-1 is a high-performance network interface for a PC cluster system in which bottleneck in accessing network is relaxed by using a PC memory bus instead of an I/O bus. Moreover, Though a PC memory bus has rather broad bandwidth earliest in a PC, the DIMMnet-1 can reorganize an NIC according to grawing bandwidth of a communication link. However, the NIC which uses a parallel transmission lacks a scalability of a bandwidth of a communication link. To overcome this problem, in developing the DIMMnet-1/bDais and the DIMMnet-2 as a new NIC, we employ a high speed serial transmission. Therefore we have to revise the DIMMnet network protocol. Now we have been developing it by two parts. The One is the CoreLogic which includes communication primitives. In this paper, we also describe the InfiniSWIF that includes functions of packet forwarding. And we propose the Concurrent Datagram communication service which is implemented in the InfiniSWIF and gives reliability and scalability for the DIMMnet network protocol.

1. はじめに

近年のネットワークインタフェース (NIC) では，PC における CPU の動作速度の向上や通信に利用可能な物理リンクの速度向上により，I/O バスのボトルネックが問題になっている．これを解消するために PCI バス規格を拡張した PCI-X，PCI-Express が規格化され，10Gbps の帯域を有するネットワークインタフェース (NIC) の製品化[8)9)] に利用されている．これに対し DIMMnet-1[1)] は早くからこの問題に対処するために試作が行われた NIC であり，PC メモリバスへ直接接続することにより広帯域化と低遅延化を行っている．

DIMMnet-1 の通信リンクは並列同期伝送を用いたオリジナルの設計である．しかしこの方式では搬送クロックに対する複数データ線のスキューにより，伝送レートの高クロック化あるいはデータ幅の拡大が困難になるため，通信リンクの帯域拡張性が乏しくなる．これは DIMMnet-1 のコンセプトの 1 つである，メモリバスが PC の中で優先して帯域拡張される部分という特徴を利用した帯域拡張性のあるネットワークプロトコルの実現に沿わない．このため新たな通信リンクとして，10GbE[8)] や InfiniBand[9)] 等の広帯域な NIC で採用されている高速シリアル伝送を採用し，通信プロトコルの拡張を行うことにした．この方式では伝送シンボルは一組の差動線で送受信されるため，伝送レートを上げやすくさらに複数の差動線を束ねて帯域拡張を行えるといった特徴を持つため，通信リンクの帯域拡張性に富むと言える．しかし DIMMnet 通信プロトコルの拡張においてスイッチを含めて新たなネットワークシステムを全て試作することは困難であるため，先に述べた高速シリアル伝送を通信リンク持つ InfiniBand 通信プロトコルを流用することにした．これにより InfiniBand スイッチを用いた DIMMnet

†1 東京農工大学 工学研究科 電子情報工学専攻
Department of Electrical and Computer Engineering,Graduate school of technology,Tokyo University of Agriculture and Technology
†2 慶應義塾大学理工学部情報工学科
Department of Information and Computer Science,Faculty of Science and Technology,Keio University
†3 （株）東芝　研究開発センター
TOSHIBA Corporate Research & Development Center
†4 東京農工大学工学部情報コミュニケーション工学科
Department of Computer, Information and Communication Sciences,Faculty of Technology,Tokyo University of Agriculture and Technology

による PC クラスタを構成可能となる．

通信プロトコル拡張においては PC クラスタ用とするため低遅延な通信性能を実現したい．この用途として InfiniBand で規定されるトランスポート層の通信プリミティブ・サービスは通信遅延に問題を持つ．これは InfiniBand が通信帯域の有効利用に重きをおくため，多数の通信要求を NIC 内へキューイングし順次実行するよう構成されているためである．さらにこれを PC クラスタへ適用するには Scalability に問題がある．

この解決のため我々は InfiniBand で規定される通信サービスの改良を行い，独自の通信プリミティブの適用を行った．これまでに高速シリアル伝送を通信リンクへ用いた DIMMnet-1/bDais[5] と DIMMnet-2[4] の 2 種類の NIC を作成した．これらは FPGA と InfiniBand X4 コネクタ (10Gbps) を備え InfiniBand スイッチにより接続される．これらの通信プロトコルは CoreLogic レイヤとパケット伝送レイヤ (InfiniSWIF) の 2 つに分けて開発されており，本稿では InfiniSWIF についての設計と評価について述べる．

2. 従来技術と問題点

2.1 InfiniBand

InfiniBand は InfiniBand trade association[10] で規定される高速シリアル伝送を通信リンクに用いた通信規格[10] であり，2.5〜120Gbps の通信帯域が規定される．通信プリミティブには Send，RDMA Write，RDMA Read，Atomic Operation，Resync Operation が定義される．これらはメッセージの伝送機構である通信サービスを選択して使用可能である．通信サービスには 4 種類あり，パケット到着を保証する ReliableConnection(RC)，ReliableDatagram(RD) とこれを行わない，UnreliableConnection(UC)，UnreliableDatagram(UD) がある．

これらの通信プリミティブ・サービスは Queue Pair(QP) と呼ばれる Send Queue(SQ) と Receive Queue(RQ) からなる通信端点を介して行われる．SQ は送信要求を処理し，RQ は受信要求を処理する．それぞれにおいて通信要求はメッセージの格納先を示したディスクリプタを書き込むことで行う．**図 1** 中 (A) は RC 通信サービスの接続状況を示す．QP は接続毎に作成されるため，仕様において Scalability が必要な場合，RD 通信サービスの利用を促している．RD 通信サービスの接続状況は図中 (B) へ示すようであり，QP 間に接続ノードに毎に作成する End-to-End Context(EE) と呼ばれる再送制御に関した情報を挟みこむ．これにより 1 つの QP で多数の接続先との通信が可能になる．

2.2 問 題 点

PC クラスタを構成する場合において Scalability の確保は重要なポイントであると考える．しかし，InfiniBand アーキテクチャにおいてこれを実現するための RD 通信サービスは，図 1 中 (B) へ示すように，構造上 1 つのメッセージに対する通信が完結しない間は次のメッセージをオーバラップして送信することができないため通信帯域の利用効率低下は明らかである．このため MPI 等並列処理のためのプログラミングに使用されている通信サービスは RC が主である[13][14]．

RC では Scalability の問題に加え通信遅延に問題があると考える．もちろん PCI バスの DMA 起動のオーバヘッドは大きいと考えるが，これ以外にも通信端点となる QP は仕様上 2^{24} 個作成可能であるため，これらの内から通信を行う QP を選択しこれを開始するまでの遅延が大きいと考える．

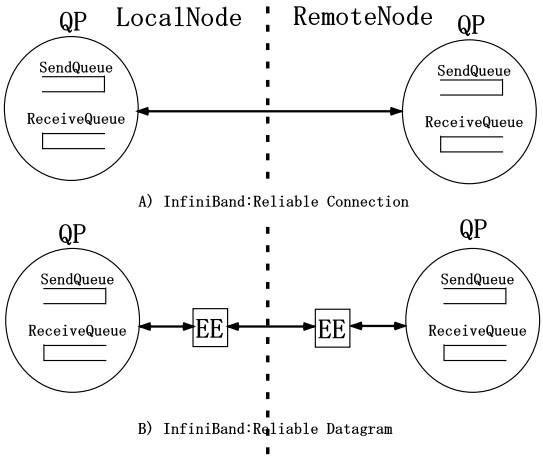

図 1 通信サービス

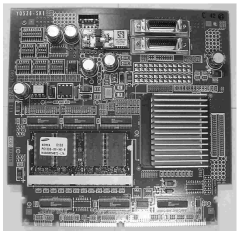

図 2 DIMMnet-1

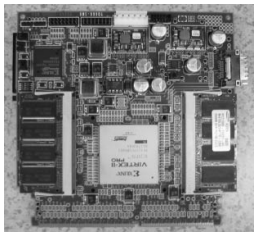

図 3 DIMMnet-2

3. DIMMnet 通信インタフェース

3.1 DIMMnet-1/bDais

本 NIC は，DIMMnet-1NIC と bDais ルータの 2 つのボードにより構成される．これらのボード間は PC 筐体内において 12 対の差動ケーブルによって接続され，DIMMnet-1 のパケットは bDais により InfiniBand サブネットワークを用いて中継される．

図 2 は DIMMnet-1 であり，PC100 のメモリスロットへ搭載される．ASIC によりネットワーク制御チップとして作成された MartiniLSI[2] を搭載し片側 2.5Gbps の通信容量を持つ全 2 重の物理リンクを備える．

図 3 は bDais ルータボードであり FPGA を中心に，DIMMnet-1 と InfiniBand ネットワークへの通信ポートにより構成される．ここで使用される FPGA は Virtex-II Pro[11] であり，最大 3.125Gbps まで実現可能な RocketIO(MGT) と呼ばれる高速シリアル伝送用トランシーバと，PowerPC プロセッサのコアを備える．本実装において MGT は InfiniBand に対する通信ポートの構成に使用し，PowerPC は Subnet Management Agent (SMA) を実装し使用した．

本 NIC において InfiniSWIF は bDais 上 FPGA へ実装される．CoreLogic レイヤは MartiniLSI へ実装されているものを指す．

3.2 DIMMnet-2

図 3 へ示す DIMMnet-2 は，PC1600 のメモリスロットへ搭載される．本ボードは bDais と同種の FPGA と X4 コネクタ (10Gbps) を 1 つ備え InfiniBand ネットワークへ直接接続される．本 NIC において InfiniSWIF と CoreLogic レイヤは同一の FPGA 上へ実装される．また本 NIC における CoreLogic レイヤは DIMMnet-1 とは異なるものであり，様々な点において改良されている．

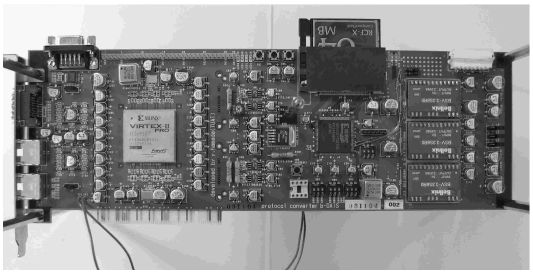

図 3 the bDais router board

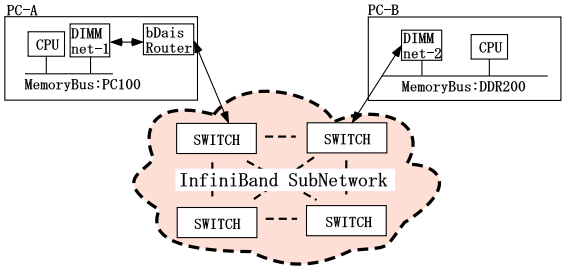

図 4 overview of DIMMnet networks

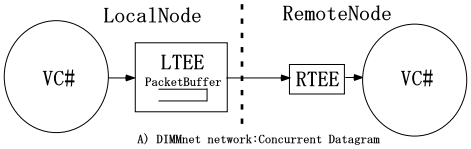

図 5 the concurrent datagram

4. 設　　計

4.1 DIMMnet ネットワーク

DIMMnet ネットワークは**図 4** へ示すように InfiniBand スイッチで構成されるサブネット単体へ，DIMMnet NICs が接続し PC クラスタを構成する．各 NIC のパケットは InfiniBand パケットの書式でカプセル化されてネットワーク通過する．信頼性確保はノード間で行い，この際の接続は**図 5** へ示すようになる．図において VC#は仮想チャネルを意味する．これは NIC 内に並列プロセスグループの通信を切り分ける目的で数個存在する．ここへは通信プリミティブによりパケットが直接入出力される．信頼性通信を用いる場合は，次節で提案する Concurrent Datagram(CD) 通信サービスを用いる．CD は 1 つの送信先へのパケット伝送：シングルキャストについて保証を行う．信頼性通信を行わない場合には，Unreliable Datagram(UD) を用いる．

4.2 提案方式：Concurrent Datagram(CD) 通信サービス

図 5 へ示す Concurrent Datagram(CD) 通信サービスは図 1B) へ示す RD サービスで問題となる通信帯域の有効利用を保ちつつ Scalability の確保を行う．これは RD サービスにおける EE Context に似た Local Temporary End-to-End Context(LTEE) 中に再送バッファを設けたため，LTEE 毎に信頼性に関する処理を行えるためである．しかし，LTEE と Remote Temporary End-to-End Context(RTEE) に注目すれば，この構造は 図 1A) へ示す RC に似ている．しかし CD において通信端点は VC#であり，RC での QP と異なり高々数個しか存在しない．つまり送信側において通信の実行順序は VC#へ入力されたパケット順に行えば良く，RC の様に多数の通信端点を扱うために生ずる遅延は抑えられると考える．

LTEE は CD において送信側に配置され，送信先へのパケットイメージを保持する再送用の PacketBuffer と，次節へ示す信頼性に関する処理を行う Automatic Repeat reQuest(ARQ) に関する情報を収めている．RTEE は受信側におかれる送信元への ARQ に関する情報が格納される．DIMMnet ネットワークでは単体のサブネットで用いるため，これらの情報はユニキャストとマルチキャストを含めて 65,534 個管理する必要があり，割合大きなメモリ領域が必要になり実装が困難になる．そこで 4.3 節で定義する Concurrent ARQ(C-ARQ) を用いて LTEE や RTEE をキャッシュ上へ配置し，実装に必要なメモリ容量の削減を行った．LTEE と RTEE に付けている Temporary は，これらが必要の無い場合には廃棄されることを意味している．

4.3 Concurrent ARQ

Concurrent ARQ(C-ARQ) は Stop and Wait ARQ(SW-ARQ) と Go Back N-ARQ(GBN-ARQ) を組み合わせ，ノード毎の ARQ に関する情報である LTEE や RTEE をキャッシュ上へ配置し，これらの状態により使用する ARQ 方式を切り替える方式である．C-ARQ では通信開始時は送信側だけで ARQ に関する情報を保持すれば実装可能な SW-ARQ を使用する．しかし SW-ARQ では通信帯域の利用効率が悪いため受信側へ ARQ に関する情報を登録した後これを改善することが可能である GBN-ARQ へ切り替える．

5. 評　　価

InfiniSWIF を DIMMnet-1/bDais と DIMMnet-2 へ用いた場合の遅延とスループットについてシミュレーションにより評価を行う．通信サービスには CD と UD を用いる．比較には InfiniBand HCA を通信サービス RC で用いた．何れの環境も InfiniBand switch で接続する．各 NIC における InfiniSWIF 内部の動作クロックは**表 1** へ示す．

表 1 working clock of the InfiniSWIF

	DIMMnet-1/bDais	DIMMnet-2
NIC clk	31.25MHz	100MHz
TP clk	62.5MHz	100MHz
DL clk	125MHz	125MHz
MAC clk	125MHz	125MHz

5.1 遅 延 評 価

スイッチにより接続した各 NIC 中 InfiniSWIF における送受信ノード間でのパケット通過遅延について評価を行う．設計値よりこの通過遅延は式 1 のように表される．式中のパラメータを**表 2** へ示す．また式中「n」は送信する DIMMnet パケット長の総計を 8bytes 単位で表す．CD 通信サービスにおいては，送受信両 ARQ キャッシュに登録が終了した定常状態 (hit) と，これらが登録されていない初期状態 (misshit) に分けてパラメータを示している．DIMMnet パケットのオーバヘッドは DIMMnet-1/bDais と DIMMnet-2 で異なるが，ここではオーバヘッドの大きい DIMMnet-1/bDais の 24bytes で統一する．これより，8bytes を伝送する場合のパケット長は 32bytes である．この際 CD での通過遅延は DIMMnet-1/bDais で 1.22μs であり，DIMMnet-2 で 1.02μs である．InfiniBand HCA のソフトウェア分を含まない 8bytes データの通過遅延は 3.39μs である．InfiniSWIF は通信プリミティブの処理を含まないため，InfiniBand HCA と直接比較することはできないが，これよりも低遅延な通

表 2 InfiniSWIF 各部オーバヘッドクロック数

	UD	CD	
		hit	misshit
A	6	11	15
B	4	4	4
C	MACtx+MACrx+Switch Latency		
D	7	7	7
E	2	7	7

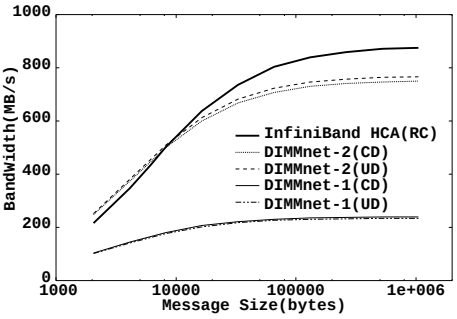

図 6 throuput of the InfiniSWIF

信プロトコルを実現できる可能性は示されたと考える．ところで UD に対する CD の遅延増加の割合は何れも 10%であった．また CD の misshit 時の hit 時に対する増加割合は 5%となった．

$$TL = \frac{n}{NICclk} + \frac{A+E}{TPclk} + \frac{B+D+n}{DLclk} + \frac{C}{MACclk} \quad (1)$$

5.2 スループット評価

InfiniSWIF のスループットは設計より式 2 で表される．式中の GAP はパケットを繰り返し伝送するのに必要なオーバヘッドであり，シミュレーションにより数えたところ UD において 7 クロックであり，CD においては 18 クロックであった．CD において GAP が増加する原因は，ACK 受信時に LARQ-Cache の内 LTEE の内容変更をパケット送信処理と並行に実行できないためである．**図 6** へ計算結果を示す．DIMMnet-1/bDais，DIMMnet-2 において継続帯域は InfiniBand HCA よりも低いが，これは各々の CoreLogic 帯域で律速されるためであり，DIMMnet-1/bDais では 250MB/s であり，DIMMnet-2 では 800MB/s である．ところで各々における UD と CD の継続帯域は，DIMMnet-1/bDais で 240MB/s，234MB/s. DIMMnet-2 で 769MB/s，749MB/s であった．CD の UD に対する低下割合は DIMMnet-1/bDais で 2.5%であり，DIMMnet-2 で 2.6%であった．

$$TP = TL + \frac{GAP \times PKTNUM}{NICclk} + \frac{N}{NICclk} \quad (2)$$

- PKTNUM :メッセージを構成するパケット数
- GAP :繰り返し伝送のオーバヘッド（クロック数）
- N :メッセージ長 (wword)

6. 結　　論

DIMMnet 通信プロトコルへ信頼性を与える Concurrent Datagram (CD) 通信サービスについて提案を行い，これをパケット伝送レイヤである InfiniSWIF へ実装した．シミュレーションによる評価において CD 用いても，これを用いない場合と比較して極端な通過遅延の増加，スループット低下を発生しないことが判った．また通過遅延の評価では InfiniBand のハードウェアでの遅延時間と比較して 2μs 以上小さく，低遅延な通信プロトコルを構成できる可能性が示されたと考える．

本稿においては，CD 通信サービスをシングルキャストにおいて適用した．今後はこれをマルチキャスト可能に改良しバリア同期の実装に適用したいと考える．

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．bDais 基板の作成は東京エレクトロンデバイス株式会社によって行なわれたものであり，同社設計開発センターの小田島氏，Eric 氏，成田氏，開発営業グループの菅原氏に感謝致します．また DIMMnet-1 基板の調整，DIMMnet-2 基板の製造を行なって頂いた日立 IT 株式会社の岩田氏，今城氏，松尾氏，上嶋氏に感謝致します．

参 考 文 献

1) N.Tanabe, et al "MEMOnet:Network interface plugged into a memory slot." IEEE International Conference on Cluster Computing (CLUSTER2000) 17-26,2000
2) 山本, 他 "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 ."
3) 西，他 "LASN 用 8Gbps/port8x8One-chip スイッチ:RHiNET-2/SW",JSPP2000 pp173-180(2000). IPSJ Computer Architecture No140,
4) 北村, 他, DIMMnet-2 ネットワークインタフェースボードの試作，ARC Vol.2004 No.80,151-156(2004).
5) 濱田，他， bDais:DIMMnet-1/InfiniBand 間ルータの開発，先進的計算機基盤シンポジウム SACSIS2004 ，Vol.2004,No.6 pp.133-134(2004).
6) InfiniBand Trade Association.InfiniBand architecture Specification Release1.2 October 2004. Final
7) D.Dunning et al, "The Virtual Interface Architecture" IEEE Micro,March/April 1998, pp.66-73
8) http://www.intel.co.jp/jp/network /connectivity/products /pro10GbE_LR_server_adapter.htm
9) http://www.mellanox.com/products/hca.html
10) http://www.infinibandta.org/home
11) http://www.xilinx.com/
12) XILINX "RocketIO^{TM}Tranceiver User Guide" UG024(V2.3)Feburuary 24,2004
13) J.Liu, et al,High Performance RDMA-Based MPI Implementation over InfiniBand.In 17th annual ACM International Conference on Supercomputing(ICS'03),June 2003.
14) R. Noronha, et al,Designing High Performance DSM Systems using InfiniBand: Opportunities, Challenges and Experiences. , Technical Report, OSU-CISRC-11/03-TR60, Computer and Information Science department, the Ohio State University, Oct. 2003.