2004－ARC－159　(25)
2004／7／31

bDais:DIMMnet-1/InfiniBand 間ルータの評価

濱田芳博[†1]　荒木健志[†1]　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．DIMMnet-1 が利用可能なネットワークには，RHiNET 用に作成された SW2 スイッチによるものがある．しかしこれは，LocalAreaSystemNetwork(LASN) と呼ばれるネットワーククラスを対象にしており，SystemAreaNetwork(SAN) よりも広い範囲のインターコネクションを想定しているため，高精度でありコストが高い．これに対し DIMMnet-1 が想定するネットワーククラスは SAN であり，また安価なシステムの構築を目的としている．このため，InfiniBand スイッチによる商用ネットワークを用いた DIMMnet-1 クラスタを構成し評価を行うことになった．現在までに InfiniBand スイッチと DIMMnet-1 間のルータボードとして FPGA を搭載した bDais ルータの作成を行った．本報告では bDais と InfiniBand スイッチによる DIMMnet-1 クラスタの構成を説明し，遅延について評価を行う．結果より 8bytes のデータ転送の遅延は，DIMMnet-1/bDais/InfiniBand スイッチの組み合わせにおいて，InfiniBand HCA/InfiniBand スイッチの組み合わせの約 57%である．

bDais : evaluating a router between the DIMMnet-1 and the InfiniBand

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

The DIMMnet-1, which is plugged into a PC's Memory bus, is a network interface for a high performance cluster. It is connected by SW2-switchs. But SW2's network class is a Local Area System Network(LASN) that is larger than a System Area Network(SAN).It is a high precision and a high cost network.On the other hand , the DIMMnet-1 is a network interface for a SAN, and aims to construct PC clusters cheaply.So we need to make a router board to be able to construct the DIMMnet-1's PC clusters on a comoditty switch network such as the InfiniBand.That is called bDais.Now it is able to use.In this paper , we describe about the bDais and show a evaluating result about the latency. The latency of the combination of the DIMMnet-1 , the bDais and the InfiniBand switch is smaller than the combination of the InfiniBand HCA , the InfiniBand switch.It can reduce the latency by about 57%.

1. はじめに

近年の PC における CPU の動作速度の向上により，I/O バスのボトルネックが問題になりこれを解消するために PCI バス規格を拡張した PCI-X，PCI-Express が規格化されている．また近年は 10Gbps の帯域を有するネットワークインタフェース (NIC) が製品化[3)4)5)] されており，これらは PCI-X 規格を採用し広帯域化を行っている．これに対し，DIMMnet-1[7)] 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 が利用可能なネットワークには，RHiNET 用[8)] に作成された SW2[9)] スイッチによるものがある．このネットワークは光ケーブルにより接続され，エラーレートが小さい特殊な E/O O/E コンバータを用い，データリンク層上すなわち接続機器間で ForwardingErrorCorrect(FEC) を行うことにより信頼性を確保している．しかしこの方式では，伝送経路上で発生するエラービット数がエラー検出・訂正用コードの訂正能力以下である必要がある．また伝送方式には搬送クロックを用いたパラレル伝送を用いるが，データビットのずれ（スキュー）の訂正方式の実装はビット単位に行う必要があり実現が難しい．またスキューを許容内に抑えるためにはドライバ，レシーバ，アートワーク，ケーブル，コネクタ中でのスキューの合計を用いて設計を行う必要がある．これらより SW2 によるネットワークは高精度であり，コストが高くなることが判る．これは RHiNET が LocalAreaSystemNetwork(LASN)[9)] と呼ばれるネットワーククラスを対象にしており，SystemAreaNetwork(SAN) よりも広い範囲のインターコネクションを想定しているためである．

しかし DIMMnet-1 が想定するネットワーククラスは SAN であり，低コストなシステムであるため，LASN の使用は不適当である．このため安価な SAN を構成することを目的として，商用スイッチによるネットワークを用いた DIMMnet-1 クラスタを構成し評価を行うことになった．また，現在ホストインタフェースに DDR-SDRAM を使用した DIMMnet-2[6)] の開発が同時に行われいるが，この NIC におけるネットワーククラスも同様である．このため，使用する商用スイッチの選択においては DIMMnet-1/2 において必要な帯域を有し，設計仕様がオープンであることを求めた．これより，InfiniBand 規格のスイッチの使用を決定した．この規格は InfiniBand Trade Assosiation[2)] により定義され，仕様書も入手可能である．帯域としては 2.5, 10，30Gbps が利用可能である．これは SystemAreaNetwork や StrageAreaNetwork 向けのネットワークシステムである．PC クラスタでの利用においては，MPI[10)11)] や

†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

DSM[12] などの幾つかの実装例が報告されており良好な結果を示している．

現在まで InfiniBand スイッチによる DIMMnet-1 クラスタを構成するため，FPGA を搭載した bDais ルータボード[13] の開発を行った．DIMMnet-2 においてはルータボードと同種の FPGA を実装しており，直接 InfiniBand へ接続可能である．本報告においては 8bytes データ転送時のレイテンシでは DIMMnet-1(電気版) / bDais / InfiniBand スイッチの接続では InfiniBand HCA / InfiniBnad スイッチの接続に対し 57%の時間で転送が可能であることが判った．またこの結果より，DIMMnet-2 においては InfiniBand スイッチへ直接接続されるため，DIMMnet-1 / bDais の組み合わせや InfiniBand HCA よりも低遅延になると考えられる．

本稿は 2，3，4 章で DIMMnet-1 と InfiniBand の概要を示し，両 NIC の比較を行う．次に 5 章において bDais ルータボードを示し，これを用いた DIMMnet-1/bDais/InfiniBand のクラスタ構成を示す．6，7 章において bDais の設計要求と実装を示し，8 章において 8bytes データ転送時の遅延について評価を行い，10 章へまとめる．

2. DIMMnet-1

通信 LSI として Martini[8] を搭載した**図 2** へ示す NIC であり，PC100 へ準拠したメモリバスへ接続される．通信リンクに E/O O/E コンバータを搭載した光リンクを持つ光版 (**図 1**) と，LVDS12 対を入出力に持つ電気版 (図 2) の 2 種類が存在する．光版においては通信リンクは 8Gbps であり，電気版は 2.5Gbps である．外部スイッチとの通信プロトコルは両者共に SW2[9] を使用している．また，本稿における bDais ルータは電気版と接続される．

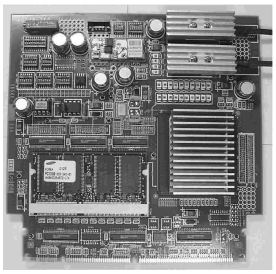

図 1 DIMMnet-1：光版

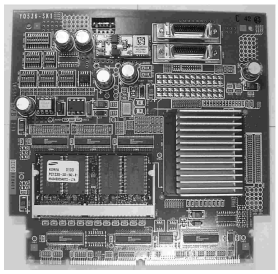

図 2 DIMMnet-1：電気版

2.1 Martini

RHiNET/DIMMnet-1 用として開発された通信 LSI である．外部スイッチとの接続プロトコルに SW2 を使用した場合，データリンク層において ECC による ForwardingErrorCorrect(FEC) が行われる．これは Martini が LocalAreaSystemNetwork(LASN) と呼ぶ SAN よりも広いネットワーククラスで並列計算を行うため，パケット再送に伴う輻輳の排除，上位層でのパケット配送保証に伴うオーバヘッドの削減を行うためである．このため，通信プリミティブが実装されるトランスポート層においてはエラーパケットの検出，再送機構等を実装せずに済む．トランスポート層へは**表 1** へ示す 3 種類の通信プリミティブがハードワイアードにより実装されている．それぞれにおいて RemoteDMA による PUSH が可能である．通信を行うユーザプロセスは通信端点として NIC 上の Window 領域，あるいは AOTF キック領域を自身のアドレス空間へマッピングする．RDMA と BOTF のプリミティブ起動には Window 領域が用いられる．RDMA では 1 枚の Window を使用し，BOTF では複数の Window を交互に使用して通信の起動が行われる．

2.2 ユーザインタフェース

ドライバ及び通信ライブラリは Linux kernel 2.4.2 用に作成している．NIC は複数あるメモリスロットの内 PC 物理アドレス空間の低い方へ割り付けられる 2 番目のスロットへ搭載される．また，配線長の問題により 3 番目以降のスロットへ本 NIC を搭載して使用することは難しい．このため，NIC のアドレス空間のみを OS 起動時に Reserve するために Kernel ソースを若干変更している．先に述べた Window や AOTF キック領域は，ライブラリとドライバを介してユーザプロセスへマッピングされる．

表 1 通信プリミティブ

	転送サイズ	送信側
AOTF	1〜8bytes	送信データを AOTF キック領域へ書き込むことにより通信の起動を行う．パケットヘッダは Martini 内部へ AOTF キックアドレスを基にキャッシュされており，書き込みデータと共にパケットが送出される．
BOTF	8〜472bytes	パケットヘッダと送信データを Window 領域へ書き込むことにより通信起動を行う．
RDMA	8〜	送信側と受信側での DMA 情報を書き込むことにより通信起動を行う．

3. InfiniBand

3.1 HCA

InfiniBand において NIC は HostChannelAdapter(HCA) と呼ばれる．通信ポートには X1，X4，X12 規格があり，それぞれ 2.5Gbps の上下 2 組のリンクを 1 組，4 組，12 組持ち，2.5Gbps，10Gbps，30Gbps の帯域を 1 物理チャネルとして使用可能．リンク間のフロー制御にはクレジットベースが用いられる．仮想チャネルは VirtualLane(VL) と呼ばれ最大 16 本まで 1 つの物理チャネルに割付けられる．現在 InfiniBand HCA は Mellanox[4] や Voltaire[5] から販売されており，本稿での 4.4 へ述べる通信性能評価には Voltaire の HCA400 を使用している．これは，PCI-X を PC へのインタフェースとして用い，通信リンクには X4 規格を用いている．

3.2 QueuePair

HCA における通信は QueuePair（QP）を基に実行される．**図 3** へ示す様に，これは Send Queue(SQ) と Receive Queue(RQ) の 2 つからなる．ユーザプロセスは WorkQueueElement（WQE）と呼ばれる通信要求を SQ か RQ へ書き込むことにより発行する．WQE により実行可能な通信プリミティブは，通信終了時は，HCA により CompletionQueueElement（CQE）が QP へ対応づけられた CompletionQueue(CQ) へ格納されるため，ユーザプロセスはこれを読み出すことにより，通信終了状態を知ることができる．

QueuePair と VL のマッピングは ServiceLevel(SL) を通じて行われ，これにより QualityOfService(QOS) の実装を示唆している．現状ではこれはサブネットワーク間をパケットルーティングされる際に VL の再割り当てを行うために使用されている．

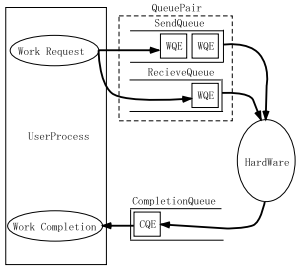

図 3 QueuePair による通信要求

3.3 通信サービス

HCA のトランスポート層において以下へ示す通信サービスが提供される．通信プリミティブはこれらを使用して実装される．

- **ReliableConnection（RC）**
  送信元と送信先の QP を一対一で対応付け通信を行う．パケットの到達保証と到達順序保証が行われる
- **UnreliableConnection（UC）**
  送信元と送信先の QP を一対一で対応付け通信を行う．受信パケットの正当性が検査され，不当なパケットは破棄される．再送制御は行われない

表 2 通信プリミティブ

通信プリミティブ	RC	RD	UC	UD	内容
Send	○	○	○	○	リモート側で Receive が実行されていれば, ローカル側メモリ領域からデータを読み出しリモートへ送信する．この際送信バッファを複数まとめるための Gather が指定可能.
Receive	○	○	○	○	リモート側よりの Send 要求を受け付け，受信データをローカル側のバッファ領域へ格納する．この際受信データを複数の受信バッファへ分割するための Scatter が指定可能.
RDMA - WRITE	○	○	○	×	ローカル側のメモリ領域のデータをリモート側のメモリ領域へ RemoteDMA により書き込みを行う.
RDMA - READ	○	○	×	×	リモート側のメモリ領域をローカル側のメモリ領域へ RemoteDMA により読み込みを行う.
ATOMIC	○	○	×	×	Compare & Swapp や fetch & add をリモートメモリに行う.

注）表中で RC，RD，UC，UD **の略語は各プリミティブにおいてトランスポート層で使用可能な通信サービス** (3.3 **参照**) **を示す．**

- **ReliableDatagram（RD）**
  End-to-EndContexts（EEC）を用い，送信側の QP を多数の受信側 QP と対応付けることが可能．パケットの到達保証と到達順序保証が行われる．
- **UnreliableDatagram（UD）**
  送信側の QP を多数の受信側 QP と対応付けることができる．受信パケットの正当性が検査され，不当なパケットは破棄される．再送制御は行われない．

3.4 Reliable サービス

通信サービスにおける Reliable サービスは，ノード間でのパケット送受信の保証はを行う．これは QueuePair 間で，パケットシーケンス番号を用いた Ack/Nack プロトコルにより実現される．

3.5 ユーザインタフェース

InfiniBand においてユーザインタフェースは Verb 層として定義される．ここでは各種通信プリミティブを使用するための API が定義される．

4. DIMMnet-1/InfiniBand 比較

4.1 仮想チャネル

DIMMnet-1 において仮想チャネルは 4 本備える．SW2 側の仮想チャネルは 16 本あり，NIC 上の設定値によりこれらは対応づけられる．これは並列プロセスグループの通信を区別するために使用される．InfiniBand における仮想チャネルは VL と呼ばれ 16 本まで実装可能である．VL は QueuePair と ServiceLevel(SL) により対応づけられる．通信の区別は QueuePair 毎に可能であり，これは 2^{24} 個まで作成可能である．InfiniBand は DIMMnet-1 よりもハードウェアレベルでより細かく通信を区別している．

4.2 通信サービス比較

DIMMnet-1 はデータリンク層で EEC による FEC を行い，パケット伝送の保証を行なう．これに対し InfiniBand ではトランスポート層で Reliable と Unreliable サービスが定義されている．パケット伝送は Reliable サービスにおいて，QueuePair 間で Ack/Nack プロトコルによるパケット再送制御を用いて保証される．また送信側と受信側における QueuePair の対応付けにより，connection 型と datagram 型に分けられる．InfiniBand において DIMMnet-1 と同様の通信サービスは，ReliableDatagram(RD) である．

4.3 通信プリミティブ比較

DIMMnet-1 が備える通信プリミティブは RDMA による PUSH を実行する．複数あるプリミティブの相違は送信側においてユーザプロセスからの通信要求時にパケットデータを PIO による即値として与えるか，DMA のためのディスクリプタとして与えるかであり，簡潔に実装されている．これに対し，InfiniBand では RDMA による WRITE 以外にも READ や Send/Receive，AtomicOperation といった豊富なプリミティブを備え，これらと先の通信サービスを組み合わせることが可能であり，DIMMnet-1 に対し複雑な実装といえる．

4.4 通信性能比較

両 NIC の通信性能の比較を行う．評価環境は**表 3** へ示す．それぞれの測定環境において NIC を搭載した PC を 2 台用意し，ケーブルで直結した対抗通信の環境で評価を行った．比較は 8〜4096bytes での遅延時間と，継続帯域幅が得られるまでのスループットで行う．DIMMnet-1 については電気版と光版について，AOTF，BOTF，RDMA による結果を用いる．また，4.3 へ述べた様に DIMMnet-1 の通信プリミティブは InfiniBand における ReliableDatagram サービスを用いた RDMA-WRITE と同等であると考える．このため，InfiniBand 側はこの方式を用いたプログラムで比較を行うべきである．しかし，本稿においてこのサービスによる通信を行なうプログラムを用意できなかった．代わりに Reliable Connection による RDMA-WRITE を用いたプログラムを作成し比較を行った．1 対 1 の環境で通信性能を測定する場合において両者の差は少ないと考える．また参考として Unreliabe Connection による RDMA-WRITE との比較を行った．

表 3 評価環境

	DIMMnet		InfiniBand
NIC	電気版	光版	Voltaire HCA 400/PCI-X
CPU	Pentium3 1.25 GHz	Pentium3 850 MHz	Pentium4 3.06 GHz
ChipSet	VIA Apollo Pro 133T	VIA Apollo Pro 133A	ServerWorks GC-SL
OS	Linux Kernel 2.4.2		Linux Kernel 2.4.18-10
compiler	gcc 2.91.66		gcc 2.96-110
driver	Ver2.6		ib-host-hpc-1.2.0_63-1

4.4.1 8bytes 遅延

表 4 へ 8bytes データ送信時の遅延の比較結果を示す．DIMMnet においてはこのサイズのデータ転送は AOTF が最も速いためこれを用いた．また，InfiniBand においては RC は ReliableConnection を用いた RDMA-WRITE の結果であり，UC は UnriliableConnection を用いた結果である．結果より InfiniBand の遅延は DIMMnet と比較して遅く，電気版との比較で RC が 3.94 倍，UC が 3.88 倍となる．また InfiniBand において UC が RC と比較し若干速いが，両者間の差は小さい．

表 4 8bytes 送信時遅延比較

	DIMMnet-1		InfiniBand	
	電気版	光版	RC	UC
8bytes 遅延（us）	1.388	1.27	5.47	5.39

4.4.2 遅　　延

図 4 へ 8〜4096bytes データ転送時の遅延の比較結果を示す．DIMMnet-1 においてこのサイズのデータ転送には BOTF が速いためこれを用いた．InfiniBand においては先程と同様に RC，UC の RDMA-Write である．結果より DIMMnet-1 は，512bytes 付近より低いサイズのデータ転送の遅延は InfiniBand より 2〜3us 小さい．しかし，512bytes 以上のデータ転送で遅延が急激に増加し，InfiniBand の遅延より大きくなる．4096bytes 時点で電気版の遅延は InfiniBand の 5 倍であり，光版は 3 倍に達する．

4.4.3 帯　　域

図 5 へ 1024〜80744bytes データ転送時の遅延の帯域の比較結果を示す．DIMMnet-1 においてこのサイズのデータ転送では RDMA が最も高い帯域となるためこれを用いた．InfiniBand においては先程と同様に RC の RDMA-Write である．結果より，InfiniBand の帯域は DIMMnet-1 に対して高く 64Kbytes 付近で約 870MB/s で継続した帯域を示す．X4 規格においては 10Gbps の帯域を有するが，8B/10B により符号化されるためこれは最大限の帯域を得ていることになる．これに対し電気光版は 6Kbytes 付近で 140MB/s の継続帯域を示し，同様に光版は 240MB/s を示す．

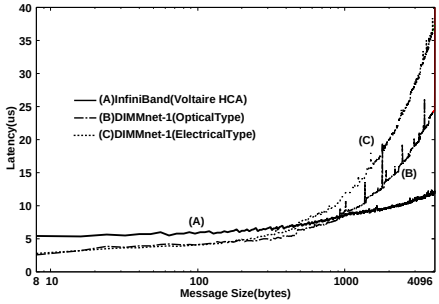

図 4 Latency

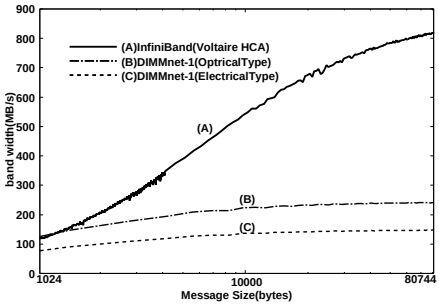

図 5 BandWidth

4.4.4 ま と め

InfiniBand で定義される通信プリミティブは高機能であり，これらが QueuePair により実装されることにより，多様な通信プリミティブを一枚の HCA 上で実行することが可能である．これに対し DIMMnet-1 が実装している通信プリミティブは簡潔なものである．この違いは通信性能評価に現れており，512bytes 以下での遅延は DIMMnet-1 が電気版，光版共に InfiniBand よりも小さい．DIMMnet-1/bDais や DIMMnet-2 では遅延の削減に重点をおいて実装を進めるべきである．帯域については，DIMMnet-2 において InfiniBand 並に改善されていくと考える．

5. bDais による DIMMnet-1 クラスタ構成

本章において DIMMnet-1，bDais，InfiniBand スイッチによる DIMMnet-1 クラスタ構成を示す．

5.1 bDais ルータボード

bDais ルータボードは，DIMMnet-1 のパケットを InfiniBand スイッチで構成されたネットワークへ通過させるためのエッジルータである．本ボードは**図 6** へ示すボードであり，FPGA を中心に DIMMnet-1 と InfiniBand スイッチへの通信ポートを持ち，16Mbit の SSRAM を備える．また，PCI バスへ対してのエッジを持つが，これは PC 筐体内でボードを支持するために使用するものであり，電源以外の信号線は配線していない．

使用 FPGA 特徴

FPGA には Xilinx 社 Virtex-IIpro(XC2VP20FF896) を用いている．本 FPGA は 3.125Gbps までの高速シリアル伝送を行える MGT トランシーバと，PowerPC プロセッサ (PPC405) をハードウェアマクロとして備え，内部メモリとして BlockRAM(BRAM) が 1,584Kbit 使用可能である．

5.2 InfiniBand スイッチ

InfiniBand によるネットワークの最小単位はサブネットでありスイッチにより構成される．このネットワークは，48K-1 のユニキャストアドレスと 16K-1 のマルチキャストアドレスを識別可能である．本稿においてスイッチは，Voltaire 社の ISR6000 (SW6IB4，PEM2IB4)[5] を使用する．本スイッチは X4 ポート (6.2 参照) を 8 個備え，各ポートは X1,X4 何れかの規格を選択的に使用可能である．また InfiniBand 規格[2] においてはスイッチで構成されるネットワークをサブネットワークと呼び SubnetManager(SM) と呼ばれるプロセスにより，サブネットワーク上のルーティングテーブルの管理を行う．本スイッチはこの SM を備え，これは VoltaireFoundationManager(VFM) と呼ばれる．

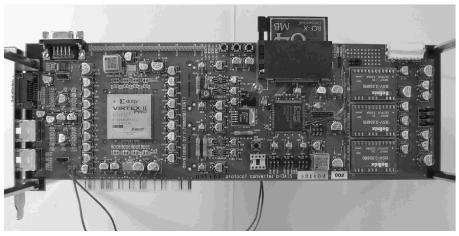

図 6 bDais ルータボード

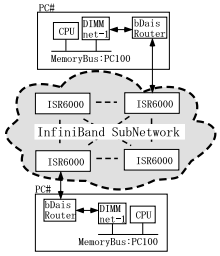

図 7 bDais/InfiniBand サブネットによる DIMMnet-1 クラスタ構成

図 8 DIMMnet-1-bDais-InfiniBand スイッチの接続

5.3 bDais による DIMMnet-1 クラスタ構成

DIMMnet-1 クラスタは**図 7** へ示す様に，ISR6000 の InfiniBand サブネットワークに bDais 経由で NIC を接続し構成する．**図 8** へ図中 PC#の接続状態を示す．

6. 設 計 要 求

6.1 概 要

bDais ルータにおいて DIMMnet-1/InfiniBand 間のルーティングに必要な要件を OSI 参照モデルに沿って述べる．**図 9** へはルータ上へ実装するパケット処理レイヤを示す．以下においては DIMMnet-1 パケットについての処理を SW2 と示し，InfiniBand パケットについて Infini と示す．

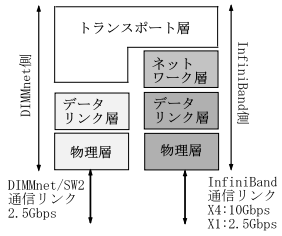

図 9 ルータレイヤ

6.2 物 理 層

SW2 LVDS12 対を入出力各々に持つ 10bit の同期伝送であり，伝送レートは 2.5Gbps である．

Infini X12 で差動信号 12 対，X4 で 4 対 X1 で 1 対を入出力各々に持つシリアル非同期伝送であり，伝送レートはそれぞれ 30Gbps，10Gbps，2.5Gbps である．本実装では X4，X1 規格を使用する．

6.3 データリンク層

- フロー制御
  SW2 80bit 単位で slack buffer 法により行う．
  Infini パケット単位でクレジットベース法により行う．
- エラー検出訂正
  SW2 フリット毎に ECC を付加し，機器間での FEC を行いエラーフリーのネットワークを実現している．

6.4 ネットワーク層

Infini DIMMnet-1 と InfiniBand 間のルーティング処理を行う．両者においてスイッチでのパケット中継はテー

ブルルーティングにより行なわれるので，これらを対応づけ相互に付け替えを行なう必要がある．またパケットフォーマットの差違は，InfiniBand のフォーマットで DIMMnet-1 パケットをカプセル化することにより吸収する．

6.5 トランスポート層

Infini CRC によるエラー検出のみでありパケットロスが発生する．このため SW2 と同等のネットワークを実現するため，ルータ間で通過パケットに対する再送制御を行い，これを保証する必要がある．

7. 実　　装

7.1 概　　要

図 10 へ実装中の FPGA 内部回路を示しこれを基に以下に説明を行う．

7.2 Infini モジュール

7.2.1 通信ポート I/O

MGT(伝送レート 2.5GHz) を使用して実装する．X1 規格を用いる場合は 1 つ，X4 には 4 つ使用する．各 MGT は 8B/10B 使用，送信 FIFO は 16bit で構成しているため，64bit データの処理周期は X1 規格使用時に 31.25MHz，X4 では 125MHz である．

7.2.2 データパケット処理

VL#を挟んで，6.4，6.5 節の処理を行う Hi と，6.3 節の処理を行う Lo により構成される．VL#はデータ送受信を行うための仮想チャネルであり，1 チャネルは 3 つの AsynchronousFIFO(AFIFO) により構成され，総計 620Kbit 分の BRAM を使用し実装している．また，現在 Hi における 6.5 節の再送制御処理は未実装である．

7.2.3 マネージメントパケット処理

Lo と PPC405 が VL15 と FIFO Cont. を通して接続される．VL15 は VL#と同様の構成の仮想チャネルであり，SM とのパケット送受信に用いルーティングに関するパラメータの受け渡しを行なう．本実装においてはこの処理は PPC405 上のプロセスにより行っている．現在はスイッチに搭載された VoltaireFoundationManager(VFM) による SM を使用した場合には bDais をスイッチに対しリンクアップすることができない．この代わりに Mellanox 社より提供される minism を使用しているが，これを用いた場合マルチキャストが不可能になる．現在 VFM 使用についてはデバッグを行っている．

7.3 SW2 モジュール

7.3.1 通信ポート I/O

伝送クロックに対し送信データを両エッジに割り付けるため，使用 FPGA の I/O 機能である DDR I/O を使用して実装した．伝送クロックは 125MHz であり，1 フリット (80bit) の処理周期は 31.25MHz である．

7.3.2 パケット処理

AFIFO を挟んで，6.3，6.5 節の処理を行う Hi と 6.2 節の処理を行う Lo により構成される．

7.4 ク ロック

使用クロックはグループ A/B/C/D/E へ分けられる．A は DIMMnet RX からの伝送クロックを信号源として使用する．B/C/D/E はボード上のオシレータが信号源である．A/B/D/E は通信ポート I/O のために 125MHz，C は SW2 の 1 フリット処理周期に合わせ 31.25MHz である．

8. 評　　価

8.1 概　　要

図 11 へ示す接続により，DIMMnet-1 / bDais / InfiniBand スイッチ (以降 DIMMnet と記す) と InfiniBand HCA / InfiniBand スイッチ (以降 InfiniBand と示す) の (A) 及び (B) の経路における 8bytes データの通過遅延について比較評価を行う．各々の環境は表 3 へ示す. スイッチには 5.2 で示した Voltaire6000 を使用し，SM には Mellanox 社より提供される minism を使用した．以下へ各環境における遅延測定方法と結果を示し最後にまとめる．

8.2 DIMMnet

bDais 上では図 10 へ示す InfiniHI モジュールが未完のため，このモジュールから左側の SW2 側モジュールと InfiniBand 側モジュールに分けそれぞれの遅延を測定し，これらと InfiniHi モジュールの遅延の和を求めこれを 8bytes

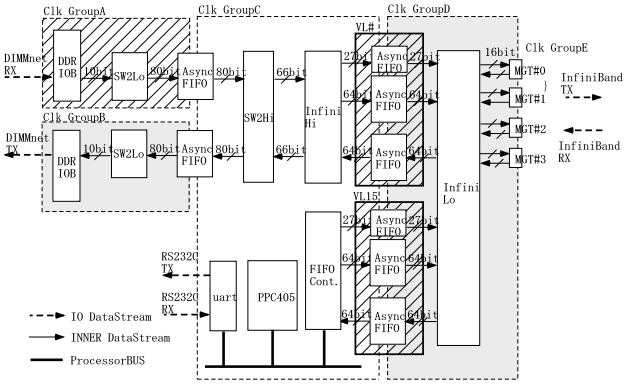

図 10 FPGA 回路構成

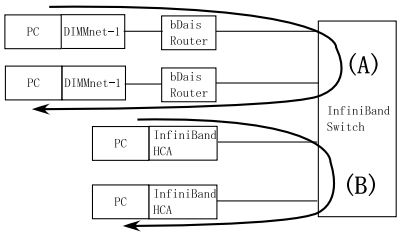

図 11 遅延測定環境

データの転送に要する遅延とした．InfiniHi モジュールにおける遅延は，InfiniBand パケットフォーマットでカプセル化した AOTF パケットを送信 FIFO へ書き込むまでの処理に要する時間とした．

8.2.1 SW2 側モジュール遅延測定

測定環境は**表 8** へ示す PC と 1 台の bDais を 50cm のケーブルで接続する．bDais で使用するコンフィグは図 10 へ示す SW2 HI モジュールでパケットを折り返す様にした．遅延は PC 上で AOTF により 8bytes データを送信し，自身で受信するまでに要する時間を測定し求めた．この結果 1783ns となった．また，DIMMnet-1 をケーブルによりループバック接続した場合の同様の測定値は 1388ns となった．これより，bDais における SW2 側モジュールの通過遅延は 580ns となる．この部分にかかる処理の内訳を**表 6** と**表 5** へ示す．

表 5 受信処理

処理部	処理内容	クロックドメイン	クロック数
SW2Lo	フレーム信号検出〜フレームデータ受信	GroupA	4
	IDLE フレーム判定，受信 FIFO へ書き込み		1
SW2Hi	FIFO 読み出し	GroupC	1
	SW2FLIT →共通 FLIT 変換 ECC チェック，書式変換		4
合計 (ns)			180(ns)

表 6 送信処理

処理部	処理内容	クロックドメイン	クロック数
SW2Hi	共通 FLIT → SW2FLIT 書式変換,ECC 生成	GroupC	4
	送信 FIFO 書き込み		1
SW2Lo	送信 FIFO 読み出し	GroupA	1
	フレームデータ送信		4
合計 (ns)			180(ns)

8.2.2 InfiniBand 側モジュール遅延測定

2 台の bDais をスイッチへ 5m のケーブルで接続する．bDais で使用するコンフィグは図 10 へ示す Infini 側のモジュールにおいて，PPC を VL15 以外に VL0 へも接続す

るものを作成した．遅延の測定は bDais をスイッチへリンクアップした後，PPC より bDais 間で VL0 使用して InfiniBand パケットフォーマットでカプセル化した AOTF パケットを転送することにより測定を行った．測定した時間は，送信側 bDais の送信 FIFO への WriteEnable が Assert されてからと受信側 bDais の受信 FIFO の Empty が Assert されるまでの時間であり，これらの間隔はそれぞれの信号をデバックピンへ出力し，ロジックアナライザにより測定した．この結果 785ns となった．また bDais 同士をケーブルで直接接続し同様の測定値は 400ns となった．これよりスイッチの通過遅延は 385ns であり，bDais の Infini 側モジュール通過遅延は 400ns である．この処理の内訳を **表 7**，**表 8** へ示す．

表 7 送信処理

処理部	処理内容	クロックドメイン	クロック数
InfiniLo	転送サイズ判断・パケットヘッダ作成・CRC 生成準備	GroupD	4
MGT	8B/10B Encode	GroupE	4
	TX FIFO 通過		4
	TX Serdes 通過		1.5
合計 (ns)			108(ns)

表 8 受信処理

処理部	処理内容	クロックドメイン	クロック数
MGT	8B/10B Decode	GroupE	1
	RX FIFO 通過		18
	RX Serdes 通過		1.5
InfiniLo	受信サイズ・受信 VL 判断・CRC 検査準備	GroupD	4
合計 (ns)			196(ns)

8.3 InfiniBand

UnrelibleConenection(UC) における RDMA-WRITE を用いて 8bytes データ転送時の片道遅延を測定する．5m のケーブルでスイッチに接続した 2 台のノード間の RoundTripTime を PingPong プログラムにより測定しこの半分を片道遅延とした．測定に用いたプログラムは Verb 層の関数により記述している．結果，UC におけるスイッチを介した片道遅延は 5660ns となった．

8.4 ま と め

DIMMnet（DIMMnet-1(電気版)/bDais/InfiniBand スイッチ）における測定結果を**表 9** へまとめる．全ての遅延を合計すると 8bytes のデータ転送にかかる遅延は 3208ns となった．InfiniBand における UC の片道遅延は 5660ns であり，DIMMnet は約 57%で 8bytes のデータを転送可能である．ところで，InfiniBand HCA における遅延が以外に大きい．bDais における InfiniLo 部分は規格のデータリンク層の仕様に基づいて作成しているため，この部分で消費される時間は VoltaireHCA においても同様であると考える．この時間は 400ns であり，スイッチとケーブルの遅延を加えて 785ns である．8bytes のデータ転送にかかる遅延時間よりこの値を引くと約 5us となる．この時間の内訳については不明であるが，PCI バスや DMA 起動にかかる遅延にしては大きく，QueuePair に関係した処理が大きく関与していると考える．

表 9 DIMMnet:8Bytes データ通過遅延

処理内容	通過遅延 (ns)
SW2/DIMMnet-1 通過遅延	1783
Infini 通過遅延	785
InfiniHi(FIFO) 書き込み遅延	640
合計	3208

9. 関 連 研 究

DIMMnet-2 においては InfiniBand スイッチへ直接接続されるため，8 章における DIMMnet-1/bDais の組み合わせ，InfiniBand HCA より低遅延になる．DIMMnet-2 においてはトランスポート層の改良により 4.4 で示した InfiniBand HCA 同等の帯域を得られれば，通信性能としてはこれを超えるものになると考える．

10. お わ り に

本稿では bDais ルータボードを紹介し，これを用いた InfiniBand サブネットワークによる DIMMnet-1 クラスタの構成を示した．この構成における 8bytes データ転送時の遅延は Voltaire 社の InfiniBand HCA と比較して 57%の時間で実行可能であることが判った．ところで，InfiniBand HCA における遅延は以外に大きい．これは PCI バスや DMA 起動にかかる遅延にしては大きく，高機能な通信プリミティブを実現するための QueuePair に関係した処理が大きく関与していると考えている．

本稿での評価は bDais 回路において再送制御が未実装であったため，InfiniBand HCA との比較は Unreliable Connection を用いて行った．今後再送制御を実装していく必要があるが，先の低遅延性を損なわない様に簡潔に実装していく必要がある．

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．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 基板の調整を行なって頂いた，日立 IT 株式会社の岩田氏，松尾氏に感謝致します．

参 考 文 献

1) InfiniBand Trade Association.InfiniBand architecture Specification Release1.1,November 6 2002.
2) http://www.infinibandta.org/home
3) http://www.intel.com/network / connectivity / resources / doc_library / tech_specs / spec_pro10GbE_LR_server_adapter.htm
4) http://www.mellanox.com/products/hca.html
5) http://www.voltaire.com
6) 田邊，他，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，情報処理学会アーキテクチャ研究会，2003-ARC-152，Mar.2003.
7) N.Tanabe, et al "MEMOnet:Network interface plugged into a memory slot.", In CLUSTER2000
8) J.Yamamoto,et al,Martini:An ASIC of network interface for high speed network with flexibility.,Japan
9) 西，他 "LASN 用 8Gbps/port8x8One-chip スイッチ:RHiNET-2/SW",JSPP2000 pp173-180,(May 2000)
10) J.Liu, et al,MPI over InfiniBand: Early Experiences.Technical Report, OSU-CISRC-10/02-TR25, Computer and Information Science, the Ohio State University,January 2003.
11) J.Liu, et al,High Performance RDMA-Based MPI Implementation over InfiniBand.In 17th annual ACM International Conference on Supercomputing(ICS'03),June 2003.
12) R. Noronha, and D. K. Panda.,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.
13) 濱田，他， bDais:DIMMnet-1/InfiniBand 間ルータの開発，情報処理学会アーキテクチャ研究会，sacsis2004,May 2004.