計算機アーキテクチャ　144−9
（２００１．７．２５）

RHiNET/MEMOnetネットワークインタフェース用 コントローラチップ Martiniの予備評価

渡邊 幸之介 [†1]　山本 淳二 [†2]　土屋 潤一郎 [†1]　田邊 昇 [†2]　西 宏章 [†2]
今城 英樹 [†3]　寺川 博昭 [†3]　上嶋 利明 [†3]
工藤 知宏 [†2]　天野 英晴 [†1]

[†1]慶應義塾大学　[†2]新情報処理開発機構　[†3]日立インフォメーションテクノロジー

RHiNETは分散配置されたPC/WSを相互接続してクラスタを構築するネットワークのプロトタイプである。RHiNETのネットワークインタフェースには、PCIバスに接続するRHiNET/NIと、メモリスロットを利用するMEMOnetがある。MartiniはRHiNET/NIとMEMOnetの両方をサポートする多機能ネットワークインタフェースコントローラであり、低レイテンシ高バンド幅な通信のためのハードウェア機構を備える。本論文ではMartiniがハードウェアでサポートする機能の予備評価について述べる。単純なリモートライト/リード機構は、2Gbps近いバンド幅の転送を実現する。また、PIO通信機構であるBOTF機構やAOTF 機構は、少量のデータを極めて低いレイテンシで転送する。

Preliminary Evaluation of Martini: the Network Interface Controller Chip for RHiNET/MEMOnet

Konosuke Watanabe [†1]　Junji Yamamoto [†2]　Jun-ichiro Tsuchiya [†1]　Noboru Tanabe [†2]
Hiroaki Nishi [†2]　Hideki Imashiro [†3]　Hiroaki Terakawa [†3]　Toshiaki Uejima [†3]
Tomohiro Kudoh [†2]　Hideharu Amano [†1]

[†1]Keio University　[†2]Real World Computing Partnership　[†3]Hitachi Information Technology

RHiNET is a network which enables efficient parallel processing by connecting PCs distributed in one or more floors. There are two types of the network interface, RHiNET/NI, using PCI bus, and MEMOnet, using memory slot as a connection port for the network interface card. Martini is a network interface controller chip which supports both of them. Its hardware provides a low latency and high bandwidth communication. This paper describes the preliminary evaluation of Martini's hardware supported functions. By executing common remote read/write functions by a hardwired logic, 2 Gbps bandwidth is achieved. Furthermore, short packet transfer using PIO functions, BOTF and AOTF, demonstrates the smaller latency than the traditional network interfaces.

1. はじめに

近年著しい性能向上を遂げているパーソナルコンピュータ (PC)やワークステーション (WS)は価格対性能比に優れた計算資源である。そこで、これらPC/WSを相互接続し、並列処理を行わせることで安価に大型計算機に匹敵する処理能力を実現する、クラスタコンピューティングが注目されている。

一般的なPC/WSクラスタは、集中配置したPC/WSをSAN (System Area Network)と呼ばれる高速な結合網で相互接続して構築する。これに対し、我々はオフィス等に分散配置されたPC/WS を相互接続することでクラスタを構成し、潜在する余剰計算力を活用するアプローチを提案している。分散配置されたPC/WSを相互接続してクラスタを構成するには、SAN並の低いレイテンシと高いバンド幅を持ち、かつLAN並のリンク長とトポロジの自由度を持つネットワークが必要となる。我々はこのようなネットワークのクラスをLASN (Local Area System Network) と呼び、研究・開発を行っている。

現在、LASNのプロトタイプとしてRHiNETを開発中である。RHiNETでは低レイテンシ高バンド幅の通信が要求されるため、我々は専用のネットワークインタフェースコントローラ “Martini” を開発している。

本稿では、Martiniがハードウェアでサポートする通信機構について述べ、その性能の予備評価について述べる。

2. RHiNET

LASNのプロトタイプであるRHiNETは、ネットワークインタフェースとネットワークスイッチ、およびホストとスイッチ間を接続する光インタコネクショ

ンで構成される。

現在開発中の RHiNET-2 では RHiNET-2/SW[5]、RHiNET-3/SW[2]、および NEC 製の Optical Interconnection IP (OIP) 用スイッチの OIP-SW が利用可能であり、それぞれ 1 ポート当たり 8Gbps、10Gbps、2Gbps のバンド幅を持つ。

ネットワークインタフェースには、PCI バスに接続されるタイプの RHiNET/NI と、メモリスロットに接続されるタイプの MEMOnet[6] の 2 種類が存在する。PCI バスに装着する RHiNET/NI は、ノードとなる PC/WS を選ばないという特徴を持ち、一方メモリスロットを利用する MEMOnet は、より高いバンド幅と極めて低いレイテンシでの通信が可能という特徴を持つ。

RHiNET のネットワークインタフェースは、現在、コントローラに CPLD を用いた RHiNET-2/SW 用のものが完成している。しかし、CPLD を用いたコントローラでは、RHiNET-2/SW の性能を十分に活かすことができず、更には RHiNET-3/SW に要求される、より低いレイテンシと高いバンド幅の通信を実現するのが困難である。

そこで、我々は、RHiNET-2,3/SW の要求に対して十分な処理能力を提供するネットワークインタフェースとして、RHiNET/NI である RHiNET-2/NI と、MEMOnet のプロトタイプである DIMMnet-1 を開発している。これらネットワークインタフェースでは、より低レイテンシで高バンド幅なネットワークを実現すべく、コントローラに専用 ASIC “Martini”[1] を用いる。

3. Martini

3.1 Martini の概要

Martini は 0.14μm プロセスの ASIC であり、コアプロセッサ、内蔵メモリ、強力なハードウェア転送機構、ホストとネットワーク双方に対する多様なインタフェースを備えたシステム LSI である。単純なリモートメモリのライト機構 (PUSH プリミティブ) とリード機構 (PULL プリミティブ) のみをハードウェアで高速処理し、それ以外の複雑な機構や生起率の低いイベントに対してはコアプロセッサに割込みをかけることで処理を代行させる。この機構により、基本性能を落とすこと無く柔軟な処理を可能とする。

PUSH/PULL プリミティブはメモリ間のコピーを伴わないゼロコピー通信であり、ユーザレベルでこれを実現するために、Martini は内部に独自の TLB を持ち、アドレス変換を行う。また、PUSH/PULL プリミティブではデータ転送に DMA を用いているが、DMA 転送は少量のデータ転送時にはオーバヘッドが大きい。そこで Martini は少量のデータをより低いレイテンシで転送するために PIO 機構として Block On-The-Fly (BOTF) 機構と Atomic On-The-Fly (AOTF)[4] 機構を備える。

プリミティブ自体は、window と呼ばれる Martini 上のメモリ領域に対して必要な情報を書き込むことで起動する。

3.2 Martini のハードウエア構成

Martini のハードウェアは、細かなモジュール単位でパイプライン化されており、各モジュールはコアプロセッサが詳細に制御可能である。ハードウェアはインタフェース部とコア部に大きく分かれる。Martini のブロック図を図 1 に示す。

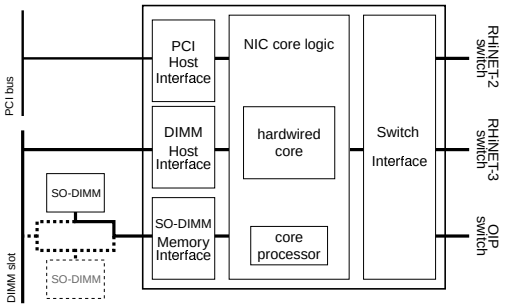

図 1　Martini のブロック図

3.2.1 インタフェース部

Martini は外部インタフェースとして、ホストインタフェース、DIMM メモリインタフェース、外部ネットワークインタフェースを備える。

ホストインタフェースは、ホストとなる PC/WS との間のインタフェースである。RHiNET-2/NI で用いるための PCI インタフェース (PCII) と、DIMMnet で用いるための DIMM ホストインタフェースを備える。

DIMM メモリインタフェースは、NIC 上にワークエリア等の目的で SO-DIMM を接続する際に利用する。

外部ネットワークインタフェースは、NIC が接続されるスイッチへのインタフェースであり、SWIF と呼ばれる。接続するネットワークに応じてプロトコルや転送周波数、bit 数等の変換を行う。SWIF は、RHiNET-2/SW で用いる 8Gbps のインタコネクション、RHiNET-3/SW で用いる 10Gbps インタコネクション、および OIP-SW で用いる 2.5Gbps のインタコネクションに対応する。

3.2.2 コントロール部

コントロール部はMartiniの制御を行う部分であり、ハードワイヤード処理部とコアプロセッサで構成される。

コアプロセッサはR3000と命令互換の32bit RISCであり、例外処理やPUSH/PULLプリミティブ以外の通信機構のネットワークインタフェース単体での実行に利用される。ハードワイヤード処理部と並列に動作することが可能な上、ハードワイヤード処理部の一部モジュールをステートレベルまで詳細に制御することができる。また、メモリとして256kbyteのオンチップSRAMを持つ。

ハードワイヤード処理部は、送信部、受信部、DMA 転送部、およびAOTF送信制御部で構成され、PUSH/PULLプリミティブの処理やBOTF/AOTFの処理、DMA転送処理等をサポートする。送信部はパケットの送信処理を、受信部はパケットの受信処理を行い、両者は並行して処理可能である。また、両者とも細かくモジュール化されており、パイプライン的に動作するため、高い処理性能を発揮する。DMA転送部は、送信部や受信部からの要求に基づき、PCII、DIMMメモリインタフェース、SWIF、コアプロセッサ用オンチップSRAMの間での任意の組合せでのDMA転送を制御する。AOTF 送信制御部は、極めて低いレイテンシで転送を行うAOTF機構のために他の転送機構とは独立した機構として用意されている。

3.3 Martiniの通信機構

3.3.1 PUSHプリミティブ

PUSHプリミティブは、自プロセスのメモリ領域を別ホストのプロセスのメモリ領域に転送する、リモートライト機構である。

PUSHプリミティブが起動すると、まずPUSHパケット発行側でMartini 内のTLBを参照し、転送するメモリ領域の仮想アドレスを物理アドレスに変換する。次にDMA要求を発行して、物理アドレスの領域からネットワークに対してデータをDMA転送する。

PUSHパケット受信側では、Martini内のTLBを参照して、まず転送先の領域の仮想アドレスを求め、次にそのアドレスを物理アドレスに変換し、DMA転送によってデータを書き込む。

DMA終了後、指定があれば、PUSHパケット受信側のネットワークインタフェースによってPUSHプリミティブ完了を示すパケットがPUSH パケット発行側のホストへ転送され、PUSH プリミティブを起動したプロセスのメモリ領域の指定アドレスにフラグがセットされる。これによりPUSHプリミティブ完了をプロセスが検知する。

3.3.2 PULLプリミティブ

PULLプリミティブは、別ホストのプロセスのメモリ領域を自プロセスのメモリ領域に転送する、リモートリード機構である。

PULLプリミティブが起動すると、PULLパケット発行側で受信領域や要求データ等の情報を含んだパケットが生成され、ネットワークへ送出される。

PULLパケット受信側では、Martini内のTLB参照により転送領域の仮想アドレスを取得し、さらに物理アドレスに変換する。次にDMA要求を発行し、このアドレスから、データをPULLEDパケットとしてネットワークへDMA転送する。

PULLパケット発行側では、PULLEDパケットに書かれた受信領域を物理アドレスに変換し、そこへDMA 転送でPULLEDパケットのボディを書き込む。DMA転送完了後は、PULLプリミティブを起動したプロセスのメモリ領域の指定アドレスに、PULLプリミティブ完了を示すフラグがセットされる。これによりPULLプリミティブの完了をプロセスが検知する。

3.3.3 Block On-The-Fly (BOTF) 機構

BOTFは、ホスト、ないしコアプロセッサがネットワークインタフェースに対して送信パケットを直接書き込むことでパケットを発行するPIO通信機構である。少量のデータ転送における、低レイテンシな通信をサポートする。

BOTFでは、windowに対してフリット単位で送出パケットを直接セットし、BOTF起動用の領域に書き込むことで、先に書き込んだ内容がパケットとして送出される。

3.3.4 Atomic On-The-Fly (AOTF) 機構

AOTFは、ホスト、もしくはコアプロセッサからの単一のメモリ書き込みでパケットを発行するPIO通信機構である。転送できるデータは1フリットという制限があるが、BOTFよりも更に低いレイテンシでの通信をサポートする。

AOTFでは、予めヘッダバッファと呼ばれる領域にパケットヘッダを格納しておき、AOTF起動領域に対して値を書き込むことで、書き込みアドレスからヘッダバッファのアドレスが導出され、ヘッダが生成される。これに、AOTF起動領域に書き込んだデータがボディとして付加されて、ネットワークへ送出される。

4. Martiniの予備評価

4.1 評価環境

評価はx86 Linux上でCadence社のVerilogシミュレータ NC Verilog Simulator v3.20を用いて行い、PCIバスのシミュレーションモデルにSynopsys社のSmart Modelを利用した。

また、シミュレーション上でのホストの動作は、C++で記述した。これには、Verilogシミュレータ側からホスト上のプログラムを起動し、シミュレータとプログラムの間に通信路を形成して相互にやり取りを行うことで、ホスト上のプログラムをシミュレーションに組み込む、独自開発のライブラリを用いている。

4.2 評価条件

シミュレーションの評価条件を以下に示す。

- ネットワークスイッチ: RHiNET-2/SW
- ノード–スイッチ間の伝送遅延: 100ns
- PCIバス: 64bit/33MHz
- Martini動作周波数: 66MHz

上記環境で2台のホスト間でのデータ転送を行い、Martiniの通信性能を評価した。

4.3 PUSH/PULLプリミティブの評価

4.3.1 PUSHプリミティブの処理時間の内訳

あるホストでPUSHプリミティブを起動し別ホストの1024Byteのデータを転送した際の、PUSHパケットの発行側における処理時間の内訳を図2に、PUSHパケットの受信側における処理時間の内訳を図3に示す。

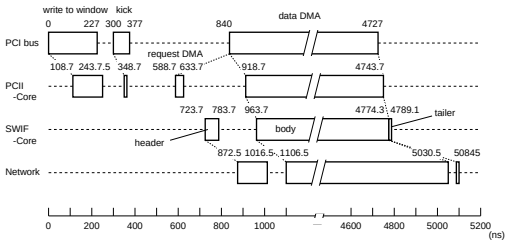

図2 PUSHパケット発行側の処理時間内訳

発行側では、windowに必要な情報を書き込み、その後kickアドレスに対して書き込むことでプリミティブが起動される。起動後は、TLBの参照によるアドレス変換等の処理が行われた後、PCIIからSWIFへのDMA要求が発行される。これと並行してパケットヘッダが生成され、DMAの要求が受け付けられるとヘッダはネットワークへ送出される。要求が発行されてからデータがネットワークに送出され始めるまでに1106.5ns を要している。

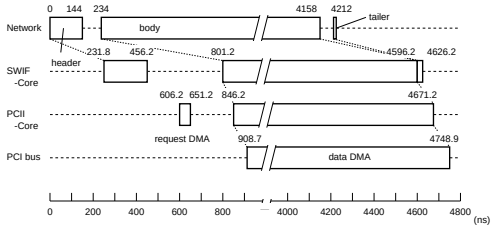

図3 PUSHパケット受信側の処理時間内訳

受信側では、まずパケットのヘッダが受信され、それを元にDMA要求が発行されてパケットのボディがホストのメモリへPCI経由でDMA転送される。ネットワークよりパケットヘッダが到着してから、有効なデータがPCIバス経由でDMA転送されるまでに908.7nsを要している。

これらより、PUSHプリミティブを起動してからリモートにデータが書かれ始めるまでのレイテンシは、ネットワークによる遅延を除くと約2.0μsとなる。ネットワークによる遅延は、RHiNET-2/SW内部の遅延が約300ns程度であり、伝送路による遅延は1mあたり5nsなので、20mの光インタコネクションでスイッチに結合した場合、500ns程度と見積られる。よって、ネットワークによる遅延も含めたPUSHプリミティブ発行からリモートでのデータ書き込み開始までのレイテンシは約2.5μsとなる。

4.3.2 PULLプリミティブの処理時間の内訳

あるホストでPULLプリミティブを起動し、別ホストから1024byteのデータを転送した際の、PULLパケットの発行側における処理時間の内訳を図4と図6に、PULLパケットの受信側における処理時間の内訳を図5に、それぞれ示す。

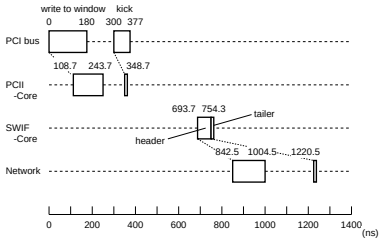

図4 PULLパケット発行側の送信処理時間内訳

発行側では、PULLプリミティブに必要な情報をwindowに書き込み、kickアドレスに対してデータを書くことで、PULLプリミティブが起動される。PULLパケットはボディを持たないので、TLBの参照によるアドレス変換等の処理が行われた後、ヘッダとテ

イラのみで構成されるPULLパケットがネットワークに送出される。プリミティブを起動してからネットワークにパケットが送出されるまでには842.5nsを要している。

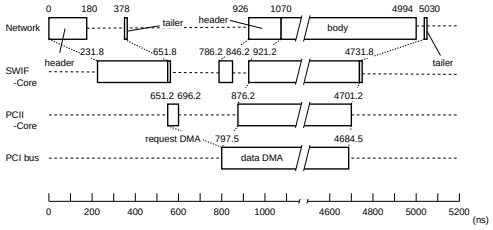

図5　PULLパケット受信側の処理時間内訳

受信側では、PULLパケットのヘッダが受信されると、ヘッダの情報を元にPCIIからSWIFへのDMA転送を要求する。それと並行してPULLパケットへの応答パケットであるPULLEDパケットのヘッダが生成され、ボディのDMA要求が受け付けられるとネットワークへ送出される。PULLパケットのヘッダ受信からPULLEDパケットのヘッダ送出開始までは926nsを要している。

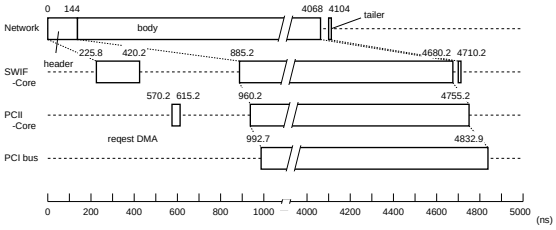

図6　PULLパケット発行側の受信処理時間内訳

発行側にてPULLEDパケットのヘッダが受信されると、PUSHパケット受信時とほぼ同様に、ヘッダの情報を元にDMA要求が発行され、ボディがホストのメモリへPCIバス経由でDMA転送される。ヘッダを受信してからボディがPCIバスへ送出され始めるまで、992.7nsを要している。

これらより、プリミティブの起動からリモートのデータがローカルメモリに書かれ始めるまでに要するレイテンシは、ネットワークの遅延を含めない場合約2.8μsとなり、PUSHと同様のネットワークの遅延を往復分含めると約3.8μsとなる。

4.3.3　バンド幅の比較

PUSHプリミティブとPULLプリミティブの転送容量を変更した際の転送時間を測定し、実効バンド幅を導出した。転送時間はプリミティブ起動から、プリミティブ完了のフラグがセットされるまでの時間とした。結果を図7に示す。

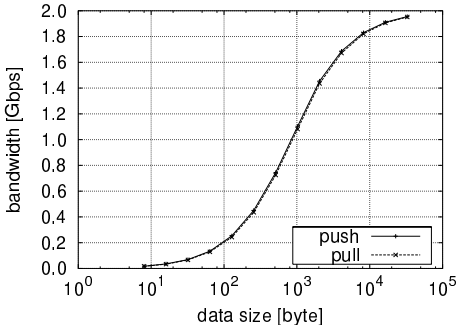

図7　PUSH/PULLプリミティブのバンド幅

転送容量が4kbyte以上では、バンド幅は1.6Gbps以上となり、転送サイズをより大きくするにつれ2Gbpsへ近付いている。64bit/33MHzのPCIバスの最大バンド幅が約2.1Gbpsであることから、PCIバスが通信のボトルネックとなっているものと考えられる。

4.4　BOTFの評価

4.4.1　BOTFの処理時間の内訳

BOTF機構を利用して32byteのボディを持つPUSHパケットを発行した際の処理時間の内訳を図8に示す。

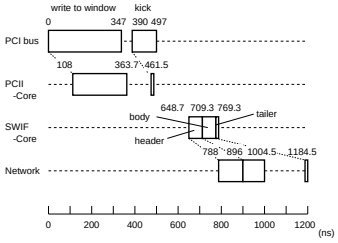

図8　BOTFの処理時間内訳

BOTFは、パケットとして送出するフリットとその個数をwindowに書き込み、kickアドレスに対して書き込むことでパケットが送出される。BOTFでは、転送するフリット数に応じて転送開始までの所要時間が異なるが、kickアドレスに対して書き込んでからネットワークにパケットヘッダが送出されるまでの所要時間は398.5nsである。

4.4.2　BOTFとPUSHプリミティブの比較

BOTFでPUSHパケットを発行した場合と、PUSHプリミティブを起動してPUSHパケットを発行した場合との間で、転送サイズを変更して比較を行った。結果を図9に示す。

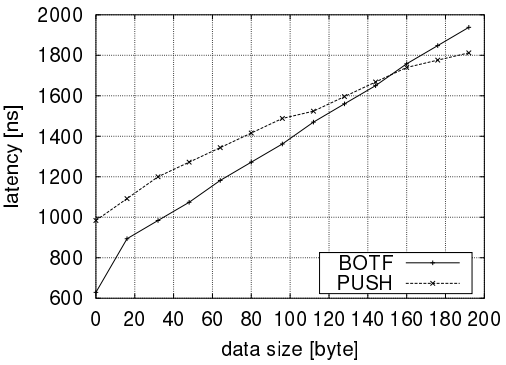

図9 BOTFとPUSHプリミティブのレイテンシの比較

PIO機構であるBOTFは、転送するボディのサイズが160byte未満の場合、PUSHプリミティブでDMAを用いてデータを送出するよりも低いレイテンシを実現しており、ソフトウェア処理するプリミティブに用いる短いパケットの転送等に向いていると言える。

4.5 AOTFの評価

4.5.1 AOTFの処理時間の内訳

AOTFを発行した際の処理時間の内訳を図10 に示す。

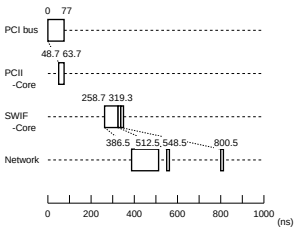

図10 AOTFの処理時間内訳

AOTFでは、PCIバスよりデータが書き込まれてからネットワークにヘッダが送出されるまで、386.5nsを要している。BOTFで8byteのデータを送出する場合、PCIバス経由のwindowへの書き込み開始からパケットヘッダの送出開始まで698.5nsを要することから、8byteのボディの転送についてはBOTFより更に低いレイテンシでの通信が可能である。

5. ま と め

本稿では、RHiNET-2/NIでの使用時における、Martiniがハードウェアサポートする通信機構の性能を、シミュレーションにて評価した。

Martiniは、PUSH/PULLプリミティブ、BOTF、AOTFのいずれも十分に小さいレイテンシを示している。特に1フリットの転送の場合はAOTFが、十数フリットまでの転送の場合はBOTFが、それぞれDMAを用いるPUSHプリミティブでの転送よりも低いレイテンシを実現しており、設計意図に十分適った性能を発揮している。

一方、PUSH/PULLプリミティブにおけるバンド幅に関しては、転送サイズを大きくしても2Gbps程度にしか到達せず、64bit/33MHzのPCIバスがボトルネックとなった。DIMMnet-1として利用する場合や、Martiniが64bit/66MHzのPCIバスに対応可能となった場合等は、より高いバンド幅を得られるものと予想される。

参 考 文 献

1) 山本 淳二, 渡辺 幸之介, 土屋 潤一郎, 今城 英樹, 西 宏章, 田辺 昇, 工藤 知宏, 天野 英晴. RHiNETの概要とMartiniの設計/実装 情報処理学会研究報告 2001-ARC-144, 2001.

2) 西 宏章, 上野 龍一郎, 多昌 廣治, 稲沢 悟, 西村 信治, 工藤 知宏, 天野 英晴. LASN用 10Gbps/port 8×8 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, pp.13-18, 2000.

3) 田邊 昇, 山本 淳二, 工藤 知宏. メモリスロット搭載型ネットワークインタフェースDIMMnet-1における細粒度通信機構. 情報処理学会研究報告 2000-ARC-137, pp.65-70, 2000.

4) Noboru Tanabe, Junji Yamamoto, Hiroaki Nishi, Tomohiro Kudoh. On-the-flu Sending : A Low Latency High Bandwidth Message Trasfer Mechanism. I-SPAN2000, pp.186-193, 2000.

5) 西 宏章, 多昌 廣治, 西村 信治, 山本 淳二, 工藤 知宏, 天野 英晴. LASN用 8Gbps/port 8×8 One-chip スイッチ: RHiNET-2/SW. 2000年記念並列処理シンポジウム (JSPP2000), pp.173-180, 2000.

6) 田邊 昇, 山本 淳二, 工藤 知宏. 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. 情報処理学会研究報告 99-ARC-134(SWoPP'99), pp.73-78, 1999.