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

2003－DSM－29 (6)
2003／4／25

複数種NICを利用した広帯域LANのための拡張ARPとその実装

東京電機大学 理工学部 情報システム工学科
梅島 慎吾　桧垣 博章

我々は、スイッチングハブで構成されるイーサネットLANにおいて、動的に定められる通信相手との間に複数の通信路を設けることによって広帯域通信を実現する手法を提案している。ここでは、ARPを拡張することによって、送信先コンピュータに装着された複数のNICのMACアドレスを送信元のコンピュータが得ることを可能としている。本論文では、これらのコンピュータに装着されたNICの種類が単一ではない場合に、NICの組合せとそれらへの送信パケットの配分比を動的に決定することが可能となる機構を実現する方法を提案し、LINUXへの実装方法について述べるとともに、その性能評価について報告する。

Extended ARP for High Performance LAN Communication with Multiple Kinds of NICs

Shingo Umeshima　Hiroaki Higaki
{shin5, hig}@higlab.k.dendai.ac.jp
Department of Computers and Systems Engineering
Tokyo Denki University

The authors have been proposed a method to achieve higher bandwidth in an Ethernet LAN with switching hubs, not repeater hubs. In order to support communication where transmitter and receiver computers are dynamically determined according to requirement in application programs, ARP (address resolution protocol) has been extended. Here, the transmitter computer achieves all MAC addresses of multiple NICs in the receiver computer. However, it is assumed that bandwidths of the NICs are the same. In order to achieve higher bandwidth even in the case that multiple kinds of NICs are used, this paper proposes a novel extension of ARP.

1　背景と目的

近年、コンピュータネットワークの発達にともない、医療現場における高精細画像のようなマルチメディアデータの配送を可能とするLANの構築が求められている。現在のLAN構築にはイーサネット[2]が広く利用されている。イーサネットには100Base-TXや1000Base-Tといった広帯域仕様が定められているが、このようなマルチメディアデータの配送がともなうアプリケーションをサポートするには必ずしも十分ではなく、光ネットワークによる広帯域化を低コストで実現するには、時間を要することが見込まれる。これまでにも、限られた帯域幅を持つ通信メディアを複数用いることによる広帯域通信の実現手法が提案されてきた。しかし、バックボーンのルータ間を複数のワイヤで接続し、パケット群を分割して配送する技術のように、その多くは固定の通信相手との間に複数の通信路を設けるものである。複数の通信路を用いて、LANに接続されたコンピュータ間での広帯域幅の通信を実現するためには、各コンピュータに複数のイーサネットNICを装着し、アプリケーションの通信要求に従って、時々刻々変化する様々な通信相手との間に複数の通信路を用意することが必要である。イーサネットLANでは、スイッチングハブの導入により、従来のCSMA/CDに

代わって、コンピュータとスイッチングハブの全二重通信が可能となり、衝突と競合による実効帯域幅の低下を避けることが可能となっている。そこで、図1のコンピュータ $A-C$ 間のような接続を動的に設ける手法が求められる。

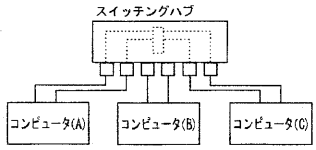

図 1: 複数 NIC を用いた広帯域通信機構

Linux へ実装されている bonding device [5] を用いることで、複数 NIC が装着されたコンピュータにおいて、パケット群をこれらの NIC に分割した送信が可能であるが、送信先 NIC が同一であるために、広帯域幅を得ることはできない。これに対して、我々は、送信元が送信先に装着された複数の NIC の MAC アドレスを獲得できる拡張 ARP [6] を提案した。ただし、提案した拡張 ARP では NIC の種類が異なる場合を想定していない。そこで、本論文では、送信先 NIC の種類を送信元に伝えられるように拡張したプロトコルを設計し、Linux への実装を行う。

2 拡張 ARP プロトコル

前章で述べた複数経路によるイーサネット LAN 内通信を実現するために導入される機構は、以下の条件を満たすことが求められる。

(1) 既存のネットワークアプリケーションへの変更が不要である。このためには、MAC アドレスの異なる複数の NIC に、同一の IP アドレスを与えることが求められる。

(2) すべてのコンピュータに提案機構が導入されていることを前提としない。提案機構が導入されているコンピュータと導入されていないコンピュータとが混在していても、TCP/IP による通信が正しく行なわれるものとする。

そこで、送信先 IP アドレスから送信先 MAC アドレスを得るための ARP [4] を拡張し、送信元が送信先の複数の NIC の MAC アドレスを得ることを可能とした [6]。これは、図2に示す ARP メッセージのパディング部以降に図3に示す拡張部を追加する新しいメッセージフォーマットを定義することにより実現されている。

従来の ARP プログラムは、拡張部のある ARP メッセージを受信しても、拡張部を無視するだけである。また、拡張した ARP プログラムが、従来の ARP プログラムの送信した ARP メッセージを受信した場合は、ARP メッセージに含まれる送信元 IP アドレスと、拡

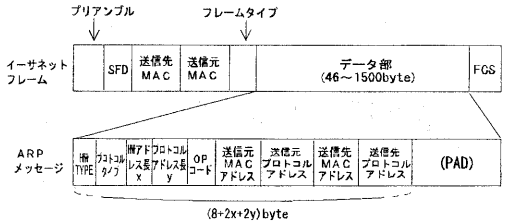

図 2: ARP フォーマット

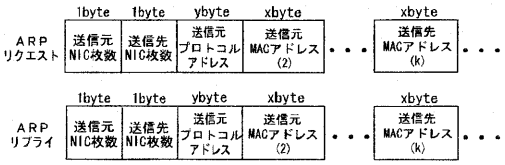

図 3: 拡張 ARP フォーマット (拡張部)

張部の送信元 IP アドレスフィールドに相当するビット列とが一致しないことにより、拡張部を含まないことが検出できる。拡張部に含まれる送信元 IP アドレスは、ARP リクエストのパディングをそのままコピーして ARP リプライを送信する FreeBSD の実装に対しても有効に作用する [6]。ところが、このプロトコルでは、送信先各 NIC の種類 (ここでは帯域幅のみを考える) を得ることができず、送信先の NIC の種類を考慮していない。

表 1: NIC の組合わせと帯域幅

NIC の組み合わせ			帯域幅 [Mbps]
1000Base-T	⇒	1000Base-T	387.89
100Base-TX	⇒	100Base-TX	95.72
1000Base-T	⇒	100Base-TX	95.77
100Base-TX	⇒	1000Base-T	95.73

表 1 は、1000Base-T と 100Base-TX の NIC の組合わせによる実効帯域幅の測定結果を示したものである。測定には、Pentium III 1GHz、192MByte メモリのパーソナルコンピュータを用いた。また、1000Base-T の NIC として、CAMEO 社製 TP83820GB-PCI32 を、100Base-TX の NIC として、Compaq 社製 NC3163 を用いた。表 1 から、送信元 NIC と送信先 NIC が 1000Base-T どうしの組合せであれば広帯域通信を実現することができるが、1000Base-T と 100Base-TX の組合せである場合には、100Base-TX どうしの組合せと同程度の帯域幅しか得られないことが分かる。したがって、異なる帯域幅の NIC が装着されている場合には、送信元と送信先の NIC の組合せ方が重要であるといえる。

ここで、1000Base-T と 100Base-TX の NIC をそれぞれ 1 枚ずつ装着した 2 台のコンピュータ間で 1000Base-T どうし、100Base-TX どうしが通信するように、使用するコンピュータのカーネルに静的に対応付けを行う処理を追加し、コンピュータ間の実効帯域を測定した。測定には先に使用したコンピュータと NIC を用いた。測定結果を表 2 に示す。

表 2: 同種の NIC どうしの通信の実効帯域幅

手法	帯域幅 [Mbps]
従来手法	191.87
同種どうし通信させる手法	191.88

表 2 から、適切な組合せを行った場合でも、実効帯域幅は従来手法と変わらず、1 対の 1000Base-T の NIC を使用した場合より狭帯域となることが分かる。従来手法 [6] では、送信制御機構にすべての NIC を平等に使用するラウンドロビンを採用している。そのため、1000Base-T と 100Base-TX の NIC の双方に同じ割合で送信パケットを割り当てることが、このような結果となる原因であると考えられる。すなわち、広帯域通信を実現するためには、1000Base-T の NIC へより多くの送信パケットを配分することが必要であることが分かる。そこで、送信先コンピュータに装着された複数 NIC の MAC アドレスだけでなく、各 NIC の種類も送信元コンピュータが獲得可能な ARP プロトコルを設計する。

3 提案プロトコル

前章で述べた拡張 ARP と同様に、送信元と送信先の間では、1 対の ARP リクエスト、ARP リプライが送受信され、1 つの ARP メッセージに複数 NIC の情報を格納する。図 4 に、提案プロトコルにおける ARP リクエスト、ARP リプライの拡張部のフォーマットを示す。各 NIC について、ハードウェアアドレス長の

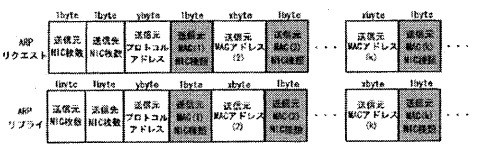

図 4: 提案 ARP フォーマット (拡張部)

ハードウェアアドレスフィールド (図 4 では MAC アドレスフィールド) とともに、1 バイトのハードウェア種別フィールド (図 4 では NIC 種類フィールド) を用意する。ただし、従来 ARP との互換性のため、1 枚目の NIC(従来 ARP フォーマット部に情報が格納されている NIC) のハードウェア種別フィールドは、拡張部に設ける。ARP リプライの受信により、送信先の複数 NIC の MAC アドレスとハードウェア種別を得た送信元は、以下の 2 つを決定する。

(1) 送信元の各 NIC が、送信先のいずれの NIC を宛先として送信するか、その組合せを決定する。
(2) 各 NIC に対して、パケット群をどのような割合で分割するか、その配分率を決定する。

提案プロトコルを Linux オペレーティングシステムのカーネルに実装した。ただし、NIC は、100Base-TX と 1000Base-T の 2 種類だけであると仮定した。組合せは、以下のように決定する。

(a) NIC の組合せは 1 対 1 とする。もし、送信元と送信先の NIC の枚数が同一であるならば、すべての NIC を用いて通信するが、枚数が異なるならば、少ない方の枚数のみを用いることとする。
(b) 帯域幅が同一であるものを優先的に組合わせ、残ったものについて、異なる帯域幅のものを組合わせる。

4 Linux カーネルへの実装

前章で述べた提案プロトコルを Linux カーネル (カーネルバージョン 2.2.17-0vl10) に実装する。実装するにあたり以下の機構を新たに実現する。

- ひとつの IP アドレスを複数の MAC アドレスと各 MAC アドレスに対応するハードウェア種別を対応付けることが可能な拡張 ARP プロトコル。
- 送信元、送信先の環境に応じた通信 NIC の対応付けとパケットの送信
- IP 層からの送信要求を受けたデータリンク層において、送信すべき IP データグラムを通信路の本数や種類により決定する配分率に従い、複数の NIC のデバイスドライバに振り分ける制御機構 (スケジューラ)。なお、受信に関しては、IP が配送順序を保障しないプロトコルであることから、特別な機構を導入する必要がない。

4.1 拡張 ARP プロトコル

実装には Linux カーネルの arp.c、dev.c、neighbour.c、neighbour.h、netdevice.h に変更を行なった。コンピュータに装着されている NIC に関する情報が格納されている構造体である device 構造体のメンバに NIC の帯域幅情報を格納する dev_speed を追加した。dev_speed のは提案 ARP メッセージフォーマットの拡張部の NIC 格納する以降に記述する ARP リクエストおよび ARP リプライの送受信、送信すべき IP データグラムの各 NIC への配分処理に使用される。ARP リクエストおよび ARP リプライを送信する関数 `arp_send` に、以下の処理を追加した。

- 従来の ARP メッセージフォーマットに加え、図4の拡張部分を持つ ARP メッセージを送信する。拡張部分には、送信元コンピュータと送信先コンピュータに装着された NIC の数、送信元コンピュータの IP アドレス、送信元コンピュータと送信先コンピュータに装着された NIC の MAC アドレスとそのハードウェア種別のフィールドが含まれる。

また、ARP リクエストおよび ARP リプライを受信する関数 arp_rcv には、以下の処理を追加した。

- ARP リクエストの拡張部分からも、MAC アドレスとそのハードウェア種別を取り出し ARP キャッシュへ保存する。

4.2 拡張 ARP キャッシュ

Linux では ARP テーブルの各エントリはひとつの IP アドレスに対応しており、その情報は neighbour 構造体に格納され、管理されている。従来の ARP では 1 エントリにひとつの MAC アドレスが対応付けられている。また論文 [6] で作成された ARP キャッシュでは neighbour 構造体を図 5 のようにメンバ ha を複数の MAC アドレスが格納されている配列の先頭アドレスへのポインタとし、NIC の数を格納する neighbour 構造体のメンバ nic_counts を追加することでひとつの IP アドレスを複数の MAC アドレスの対応付けを行っている。本論文では、MAC アドレスを格納する配列 hw_addr とその MAC アドレスを持つ NIC の帯域幅情報を格納する kind をメンバとする struct neigh_ha を定義し、図 5 のようにメンバ ha をこの構造体配列の先頭アドレスのポインタとすることで、図 6 のようにひとつの IP アドレスを複数の MAC アドレスと各 MAC アドレスに対応するハードウェア種別への対応付けを行っている。さらに、送信元、送信先 NIC の組み合わせを行った後、1000Base-T どうしの組み合わせの数、100Base-TX どうしの組み合わせの数を格納する giga_link,fast_link、自身の NIC と送信先の NIC の対応付けをさせた情報を格納する neigh_table、異なる組み合わせの対応付けを行った場合のフラグを格納する change_speed、送信先が本手法を導入しているかどうかを示すフラグを格納する ex_arp_flag の追加を行った。また、ひとつの neighbour 構造体に複数の MAC アドレスを対応付けることによって neighbour 構造体のメンバ primary_key に格納される IP アドレスをキーとして neighbour 構造体を探す関数 __neigh_lookup と neighbour 構造体に追加したこれらの情報を用いることで、前章で述べた、ARP リプライを受信する関数 arp_rcv で送信元 NIC と送信先 NIC の対応付けをすることができる。

4.3 送信制御機構 (スケジューラ)

IP 層からの送信要求に対してはデータリンク層 dev.c でどの NIC を用いて送信するか決定され、送

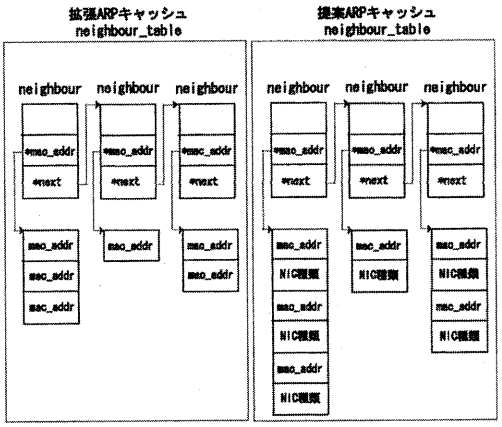

図 5: ARP テーブルの構造

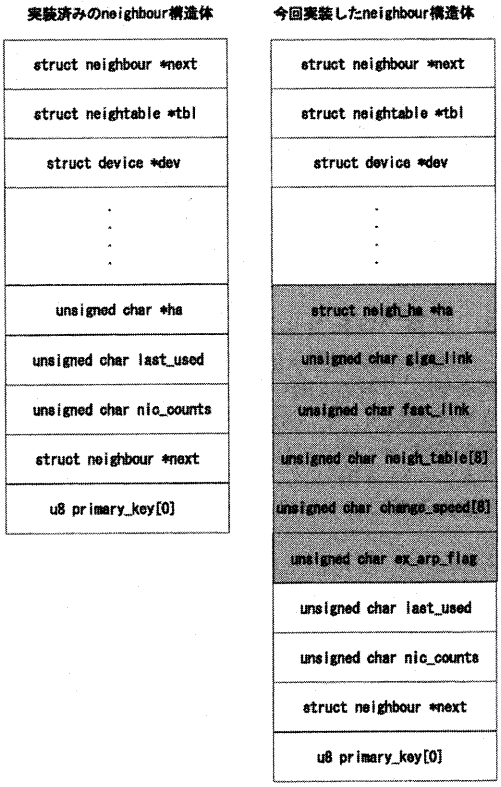

図 6: neighbour 構造体

信に用いる NIC の device 構造体へのポインタを獲得することでパケットを NIC へ割り当てている。論文 [6] で作成された送信制御機構ではパケットの NIC への割

り当てにすべてのNICを平等に使用するラウンドロビンを採用しているため、本論文で提案している広帯域のNICへより多くのパケットを送信させる手法を実現することはできない。そこでneighbour構造体のメンバgiga_link,fast_linkの情報から、あらかじめ定義されている配分率を選択し、その配分率とdevice構造体のメンバdev_speedをもとに、各NICへのパケットの割り当てを制御、またARPリプライの受信時に対応付けを行った送信NICへ送信を行うことで、より広帯域な通信を実現することができる。

このプロトコルを実装した複数のコンピュータと従来のARPが実装されたLinux、FreeBSD、Windows2000、Solaris8、Solaris9がインストールされたコンピュータとをギガビットスイッチングハブに接続した結果、正しく動作することが確認された。すなわち、提案プロトコルを実装したLinuxコンピュータどうしでは、上記の組合せによる通信が行なわれ、Linuxコンピュータと他のオペレーティングシステムのコンピュータでは、1対のNICを用いた通信が行なわれる。

5 性能評価

前章で述べた手法をLinuxカーネル(カーネルバージョン2.2.17-0vl10)へ実装し、図7に示す環境で動作の確認を行った。コンピュータA,Cには100Base-TX(A1,C1)と1000Base-T(A2,C2)のNICを1枚ずつ装着した。そのIPアドレスとMACアドレスの対応表は表3に示す通りである。AからCへpingコマンド(ICMPエコー)を実行し、A,C間に流れるIPデータグラムをBでethereal [1]を用いて観測した結果を図8に示す。図8より提案手法の動作確認をすることができた。

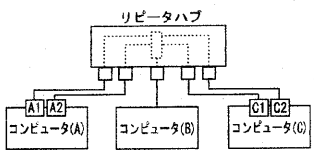

図 7: 実験環境

表 3: IPアドレスとMACアドレスの対応

A	192.168.100.100	A1	00:50:8B:EB:7A:F1
		A2	00:40:F4:2A:27:69
C	192.168.100.200	C1	00:50:8B:EB:7D:B7
		C2	00:40:F4:2A:29:53

次に、各NICへのパケットの分割における配分率については、配分率によって送信元と送信先の間で得られる全体の帯域幅をnetperf [3]測定した。測定には、Pentium III 1GHz、192MByteメモリのパーソナルコンピュータを用いた。また、1000Base-TのNICとして、CAMEO社製TP83820GB-PCI32を、100Base-TXのNICとして、Compaq社製NC3163を用いた。1000Base-Tと100Base-TXが送信元、送信先ともに各1枚ずつ装着されている場合の実験結果を図9に示す。ここでは、1000Base-Tと100Base-TXが送信元、送信先ともに各1枚ずつ装着されているとし、同一帯域幅の組み合わせ(A),と異なる帯域幅の組み合わせ(B)について測定を行った。

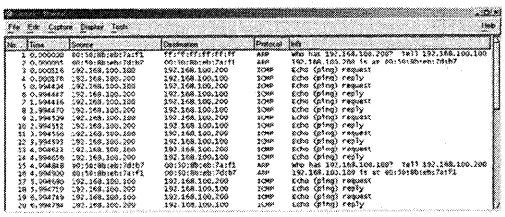

図 8: IPデータグラムの観測結果

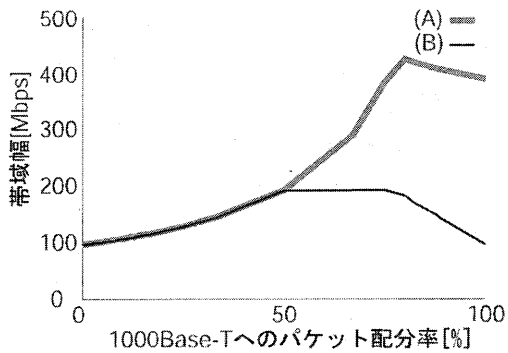

図 9: 配分率と帯域幅の関係

提案手法を導入していない拡張ARPでは、50%の可能性で異種のNICの組合せができることから、広帯域通信が実現できない可能性がある。しかし、提案手法を用いることより、確実に広帯域通信が提供されることが分かる。また、得られる帯域幅は各NICへのパケットの配分率によって大きく変化することが分かる。

次に通信に使用するNICの枚数を変えた場合の配分率の変化に対する帯域幅を測定した。測定には(A)の通信環境において、使用するNICに100Base-TXのNICを1枚追加した場合(C)、1000Base-TのNICを1枚追加した場合(D)、100Base-TX、1000Base-TのNICを1枚ずつ追加した場合(E)の3つの場合について行った。測定結果を図10に示す。次に(A)の通信において、NICの種類を変えた場合の配分率の変化に対する帯域幅を測定した。100Base-TXのNICを3com

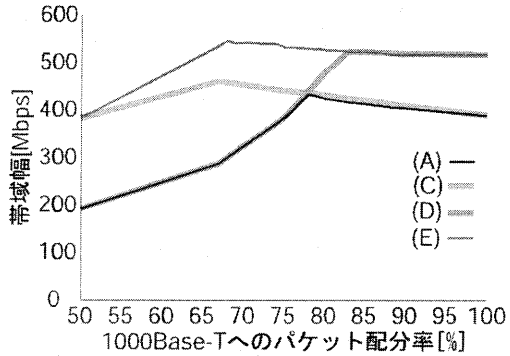

図 10: 使用する NIC の枚数を変更した場合

社製 3C905-TX に変更した場合 (F) の測定結果を図 11 に示し、1000Base-T の NIC を Intel 社製 PRO/1000 MT に変更した場合 (G) の測定結果を図 12 に示した。また変更した各 NIC1 枚どうしを対向させた通信の実行帯域幅を測定した結果を表 4 に示す。

表 4: 変更した NIC の実効帯域幅

NIC	帯域幅 [Mbps]
Intel PRO/1000 MT	827.28
3COM 3C905-TX	95.72

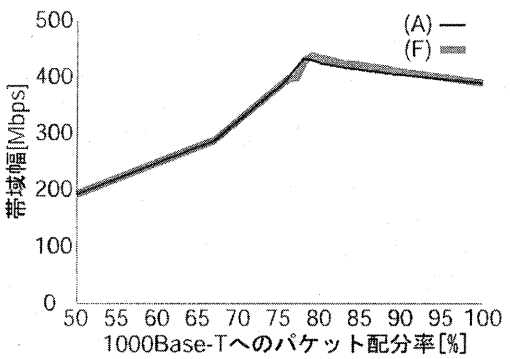

図 11: 100Base-TX を変更した場合

図 12 と表 4 から分かる通り (G) では通信の広帯域化ができていない。これは、複数枚の NIC を 32bitPCI バスに装着したため、バス帯域幅の限界を超えてしまったことが原因であると考えた。今回使用したコンピュータでは 32bitPCI バスと 64bitPCI バスが独立の構成となっている。そこで、1000Base-T の NIC を 64bitPCI バスへ装着した場合 (H) についても測定を行った。図 10, 図 11, 図 12 から、どの場合においても最も広い帯域幅が確保できる配分率は異なることが分かる。これらの結果から、最も広い帯域幅が確保できる最適な配分率は、NIC 製品の組合わせ、通信に使用する NIC の枚数や種類によっても異なることから、各 NIC の帯域幅の組合せから、デフォルトの配分率での通信を行ない、通信を継続しながら最適な配分率へと調整する機能を検討する必要がある。

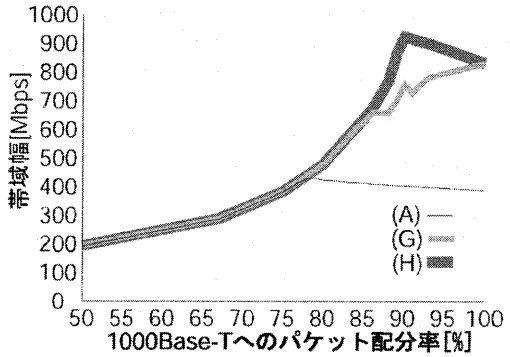

図 12: 1000Base-T を変更した場合

6 まとめと今後の課題

本論文では、帯域幅の異なる複数の NIC が装着されたコンピュータ間で広帯域幅の通信を実現するための拡張 ARP プロトコルとその Linux への実装について述べた。今後は、NIC の帯域幅の種類を限定せず、狭帯域幅の複数枚の NIC と広帯域幅の 1 枚の NIC とを対応付けることが可能な NIC 対の決定方法、パケット群の配分率を動的に最適化する手法について検討する。

参考文献

[1] Gerald, C., "ethereal," http://www.ethereal.com/.

[2] IEEE Project 802 CSMA/CD Working Group 802.3, "Operating Rules of IEEE Project 802 Working Group 802.3," http://grouper.ieee.org/groups/802/3/.

[3] Jones,R., "Netperf Homepage," http://www.netperf.org/netperf/NetperfPage.html

[4] Plummer, D.C., "An Ethernet Address Resolution Protocol," RFC 826 (1982).

[5] Thomas, D., "bonding device," tadavis@lbl.gov.

[6] 林, 梅島, 桧垣, "複数 NIC とスイッチングハブを用いた広帯域通信機構の LINUX への実装," 信学技報, Vol. 101, No. 639, pp. 33–38 (2002).