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

2005－DPS－122（20）
2005－CSEC－28（20）
2005／3／22

複数種NICによる広帯域通信のためのTCP再送抑制手法

東京電機大学 理工学部 情報システム工学科
加藤 剛史　桧垣 博章
E-mail: {kato, hig}@higlab.k.dendai.ac.jp

イーサネットで構築されたLANにおいて、マルチメディアデータの配送に広帯域幅を要求するネットワークアプリケーションが存在する。そのようなネットワークアプリケーションに十分な帯域幅を提供する手法のひとつとして、複数の通信路を利用する方法が考えられる。我々はネットワークアプリケーションへの変更を行わずに複数のNICを用いた通信を実現するため、ひとつのIPアドレスに複数のMACアドレスを対応付けることが可能な拡張ARPプロトコルを設計した。また、NICの種類に応じて各送信元NICに適切な送信先NICを対応付け、それぞれに適切な比率でIPデータグラムを配分する手法を導入した。これによってIPおよびUDPによる広帯域通信を実現することが可能となったが、配送順序保存を保証するTCPにおいては、通信路ごとの帯域幅が異なる場合に性能が劣化する。本論文では、送信元コンピュータにおいてIPデータグラム群の送信順序を制御することによって、通信路ごとの帯域幅の違いを吸収しTCPにおいても複数の通信路を用いてネットワークアプリケーションに広帯域幅を提供する手法を提案する。

Reduction of TCP Retransmission for Higher Throughput with Multiple-Route Transmission in Ethernet LAN

Takeshi Kato and Hiroaki Higaki
Department of Computers and Systems Engineering
Tokyo Denki University
E-mail: {kato, hig}@higlab.k.dendai.ac.jp

For supporting network applications which require higher throughput, multiple-route message transmission has been proposed. In an Ethernet LAN, dynamically determined pair of computers communicate. For realizing multiple-route message transmission in an Ethernet LAN, the authors have proposed an extended ARP by which a sender computer gets multiple MAC addresses of NICs of a receiver computer to which one IP address is assigned. In UDP/IP communication, higher throughput is achieved by using the extended ARP even though bandwidths of the NICs are different. In TCP/IP comunication, it is difficult to achieve higher throughput due to packet retransmission caused by wrong ordered reception if bandwidths of the NICs are different. In this paper, we introduce buffering of packets into the sender computer for waiting transmission of packets through a pair of wider bandwidth NICs. By using this method, 26.8% less packets are retransmitted and 260% more throughput is achieved.

1 背景と目的

現在、LAN の構築にはイーサネットが広く利用されており、その帯域幅は 10Mbps から 1Gbps へと順次拡大されてきた。しかし、医療現場において高精細画像を扱う医用画像情報システムのような大容量マルチメディアデータ配送を必要とするネットワークアプリケーションに対しては、十分な帯域幅を提供することはできない。このようなアプリケーションに十分な帯域幅を提供するために、光ファイバー等の新しいメディアが開発されているが、これらによる広帯域通信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。そこで、十分な帯域幅を低コストで提供する手法として、複数の通信路を用いて配送を行うことによる広帯域 LAN の実現が考えられる。限られた帯域幅を持つ通信路を複数用いることによって、送信元コンピュータと送信先コンピュータとの間に広帯域通信を実現する方法は、これまでにも様々な研究開発がなされ、広く用いられてきた。しかし、これらの方法は固定のコンピュータ対に対してのみ適用することが可能である。LAN に接続されたコンピュータが P2P 型のネットワークアプリケーションを実行する環境においては、各コンピュータに複数のイーサネット NIC を装着し、アプリケーションからの通信要求に従って時々刻々と変化する様々な通信相手との間に複数の通信路を用意しなければならない。

現在、イーサネット LAN ではリピータハブに代わり、衝突と競合を避けることが可能なスイッチングハブが広く利用されている。これを利用して、Linux に実装されている bonding device ではパケット群を複数の NIC に分散して送信することを実現している [2]。しかし、送信元が得られる送信先 MAC アドレスは 1 つであるため、送信先の 1 つの NIC にパケットが集中することとなり広帯域幅が得られない。そこで、複数の NIC を装

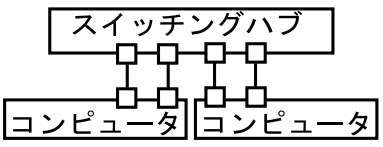

図 1: 複数 NIC を用いた広帯域通信

着したコンピュータを図 1 に示すように接続し、複数の NIC から複数の NIC へとフレーム群を配送することによって、広帯域通信を実現することができる [4]。この機構を、現在運用されている LAN へ導入するためには以下の条件を満たす必要がある。

[要求条件]

1. 既存のネットワークアプリケーションへの変更は不要である。そのためには、MAC アドレスの異なる複数の NIC に同一の IP アドレスを割り当てる必要がある。
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われるものとする。□

これらの条件を満足する手法として拡張 ARP が提案されている [6]。拡張 ARP を用いることによって、送信元コンピュータは送信先コンピュータの IP アドレスに対応する複数の MAC アドレスを得ることができる。そこで、送信元コンピュータの各 NIC に送信先コンピュータのいずれかの NIC をフレームの送信先として対応付けることによって、複数の NIC を装着した送信元コンピュータが複数の NIC を用いてフレーム群を受信することが可能になる。また、帯域幅の異なる複数種の NIC が送信元コンピュータ、送信先コンピュータに装着されている場合には、可能な範囲で帯域幅の等しい NIC を対応付けることとし [3]、さらに各 NIC に対するフレームの配分率を動的に調節することによって、送信元コンピュータから送信先コンピュータへの帯域幅を拡大することを実現している [5]。ところが、これまでに提案された手法はパケットの配送順序保存を保障しない UDP による通信を前提としている。しかし、送信元コンピュータと送信先コンピュータとの間に複数の通信路が存在する場合、帯域幅や伝送遅延といった各通信路の特性の違いにより送信元コンピュータが送信した順に送信先コンピュータが受信するとは限らない。そのため、高信頼な通信を提供するために再送機構とフロー制御機構を導入した TCP においては、通信路ごとの帯域幅が異なる場合に性能が劣化する。本論文では、送信元コンピュータにおいて IP データグラム群の送信順序を制御することによって、TCP の再送機構が機能することを抑制し、TCP においても複数の通信路を用いてネットワークアプリケーションに広帯域幅を提供する手法を提案する。

2 従来手法

拡張 ARP では、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 [1] を拡張して、送信元コンピュータが送信先コンピュータと、複数の NIC の MAC アドレスとその種類情報を交換することを可能としている。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 プログラムが拡張 ARP メッセージを受信しても、拡張部は無視される。拡張 ARP プログラムが従来の 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プロトコルアドレスと、拡張部の送信元プロトコルアドレスが一致しないことにより拡張 ARP メッセージでないことを検出する。送信元コンピュータ

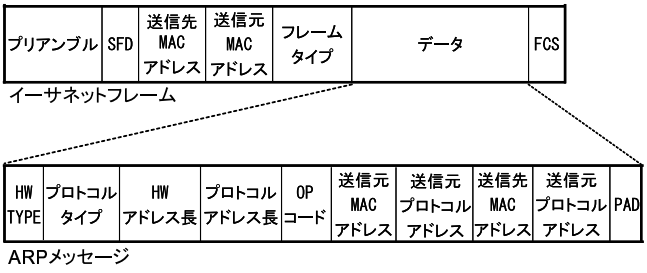

図 2: ARP メッセージ

図 3: 拡張 ARP メッセージの拡張部

は、送信先コンピュータの複数の NIC の MAC アドレスを取得することで、自身の複数の NIC から送信先コンピュータの複数の NIC へ配送することを実現する。また、NIC の種類情報を取得することによって広帯域の NIC 同士を優先的に割り当て、さらに各 NIC に対するフレームの配分率を動的に調節することによって、送信元コンピュータから送信先コンピュータへの帯域幅を拡大している。しかし、送信元コンピュータと送信先コンピュータとの間に複数の通信路が存在する場合、帯域幅や伝送遅延といった各通信路の特性を考慮せずにフレームを配分すると、たとえ各通信路が FIFO の特性を持っていて配送順序の保存を保障しているとしても、送信元コンピュータが送信した順に送信先コンピュータが受信するとは限らない。

トランスポートプロトコルに TCP を用いるネットワークアプリケーションに対して、複数の通信路による広帯域通信を用いることを考える。TCP はアプリケーションに対してセグメントの紛失、重複がなく、配送順序保存を保障するプロトコルである。これらの性質は、下位層のプロトコルが紛失や重複、配送順序の入れ替えなどが発生したとしても、TCP の機能によってアプリケーションに対して保障することができる。したがって、複数の通信路による通信をそのまま用いることが可能である。しかし、実際には以下の理由により広帯域通信が提供されない。

まず、下位層で配送順序が保障されないことにより、送信先コンピュータで受信された TCP セグメント群が長時間バッファリングされ、帯域幅が低下する。図 4 に示すように、狭帯域幅の通信路を用いて配送される TCP セグメント T_i を含むフレームが先に送信を開始されたにも関わらず、以降に送信を開始された TCP セグメント群 $T_{i+1}, \cdots, T_j (j > i)$ を含むフレームの方が先に送信先コンピュータに受信されることが考えられる。このとき、TCP では $T_{i+1}, \cdots, T_j$ を T_i が受信されるまでバッファに保存し、アプリケーションへの配送は遅延される。

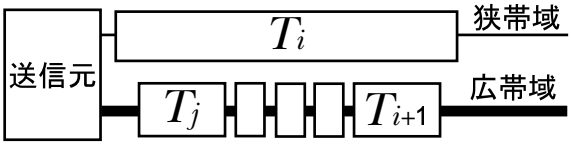

図 4: 到着順序の変動

つぎに、先着した TCP セグメント群に対する受信確認によって TCP の再送機構が機能して帯域幅が低下する。TCP では、累積的受信確認あるいは選択的受信確認を用いて、紛失したフレームに含まれる TCP セグメントを再送信している。ここでは、$T_{i+1}, \cdots, T_j$ の各 TCP セグメントの到着に対して、送信先コンピュータが受信確認の TCP セグメントを送信するが、このいずれにも T_i が未受信である情報が含まれている。これらを受信した送信元コンピュータは、閾値 k 回 (多くの実装が $k = 3$ としている) 以上の未受信情報を得ると、T_i を再送信する。すなわち、$j - i > k$ を満たすならば、T_i が紛失していない場合でも再送信される。

さらに、TCP セグメントの再送信によって、TCP のフロー制御におけるウインドウサイズが縮小され、配送中の状態にある TCP セグメント群に含まれるデータ量が低く制限され、帯域幅の拡大が実現されない。

3 提案手法

上記の問題を解決するために、アプリケーションから送信要求されたデータを含む TCP セグメント群の送信順序を、これらが送信される NIC の帯域幅に応じて変更することによって、送信先コンピュータにおける受信順序が送信元コンピュータのアプリケーションが配送要求した順序と等しい、あるいは大きく異ならないようにする。図 5 に示すように、送信元コンピュータ

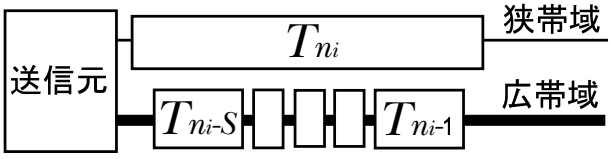

図 5: 送信順序の変更

のアプリケーションが配送要求した TCP セグメント群 $T_i, \cdots, T_{i+S}$ のうち T_{i+S} を狭帯域の NIC で先に送信し、他の TCP セグメントを広帯域 NIC から送信する。これによって、先に送信を開始された T_{i+S} が受信される以前に、この TCP セグメントと同時あるいはこれ以降に送信を開始された TCP セグメント群 $T_i, \cdots, T_{i+S-1}$

の方が先に受信されることになり、広帯域通信が実現できる。なお、定数 S をスキップ数と呼ぶ。

以上の手法は、次のフレームの送信に使われる NIC の種類によって送信すべきセグメントを決定する必要があるため、次のフレームの送信に使われる NIC の種類を参照する必要がある。そこで、この機構をデータリンク層のパケット配分機構の直前に行われるように実装するものとする。Linux の実装では、データリンク層にセグメントごとに分割された後のパケットがセグメントの順番で渡される。そこで、パケットの並び替えを行うために、TCP コネクションごとに長さ S のバッファを使用する。以下に送信順序制御機構の実装について説明する。

TCP セグメント T を含むパケットが送信順序制御機構に渡されると、つぎのフレームの送信に使われる NIC の種類によって 2 通りに分岐する。つぎのフレームが狭帯域の NIC で送信される場合、S 個のパケットの送信を保留するためにバッファに S 個のパケットが格納されていなければ、T をバッファの最後に追加する。バッファに S 個のパケットが保留されている時は T を狭帯域の NIC で送信する。そのつぎの送信に使われる NIC が広帯域 NIC であれば、バッファに保留されたパケットを送信しても良いが、狭帯域の NIC で複数のフレームが連続して送信される場合もあるため、ここではつぎの送信に関する処理を行わない。つぎのフレームが広帯域の NIC で送信される場合は、いずれかのパケットが送信される。このとき、バッファに保留されたパケットがある場合は、保留されているパケットを先に送信する。この時点で、パケットの保留が行われていないという状態は、コネクション確立時から 1000Base-T の NIC が選択されていた場合と、後述するタイムアウトによる送信が発生した場合に起こりうるため、保留されたパケットの有無を確認する必要がある。バッファに保留されているパケットを送信する場合は、バッファの先頭のパケットから順に全てのパケットを広帯域の NIC で送信する。すなわち、つぎの送信に使われる NIC が狭帯域の NIC になった場合は、バッファに保留されたパケットが残っていても、T をバッファの最後に追加し、つぎのパケットを待つ。この判断は、広帯域 NIC の連続送信回数が S より小さい場合に必要である。たとえば、TCP セグメント群 $T_1, T_2, \cdots, T_8$ を送信する時に広帯域 NIC の連続送信回数が 3、$S = 4$ のときは図 6 の順で送信されるが、広帯域 NIC で T_1, T_2, T_3 の 3 つを送信した後、長さ 4 のバッファには T_4 が残されている。また、広帯域 NIC の連続送信回数が S より大きい場合は、バッファに S 個のパケットが保留される前に狭帯域 NIC での送信が発生する。たとえば、TCP セグメント群 $T_1, T_2, \cdots, T_8$ を送信する時に広帯域 NIC の連続送信回数が 3、$S = 2$ のときは図 7 の順で送信され、広帯域 NIC の連続送信

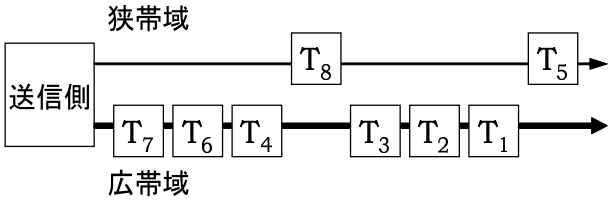

図 6: 広帯域 NIC の連続送信回数が S より小さい場合

回数分のパケットがバッファに保留される前に、狭帯域 NIC での送信が行われる。パケットが保留されていな

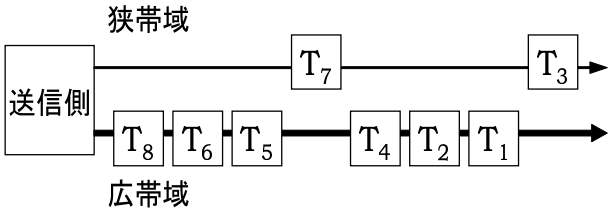

図 7: 広帯域 NIC の連続送信回数が S より大きい場合

かった、もしくは保留されたパケットをすべて送信した後、広帯域 NIC での送信が続けられる場合は T を送信し、つぎのフレームが狭帯域の NIC で送信される場合は T をバッファに追加して、つぎのパケットを待つ。以上の送信順序制御機構のフローチャートを図 8 に示す。

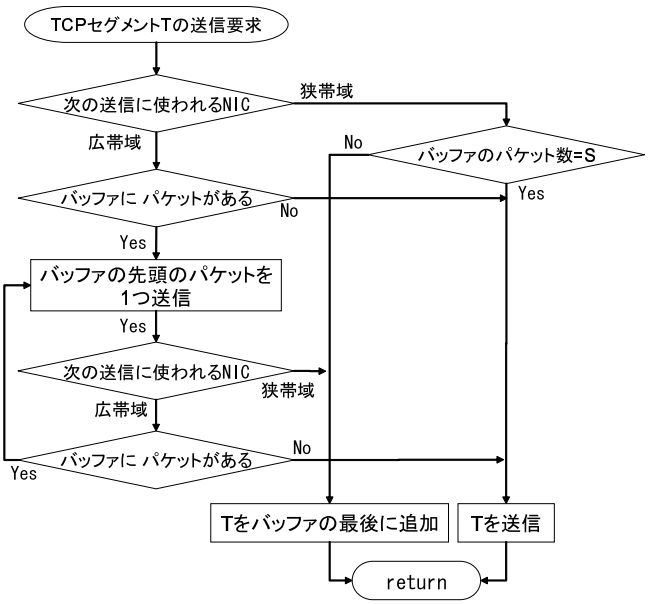

図 8: 送信順序制御機構

この機構を Linux カーネルに実装する上で、以下の変更を行う必要がある。

3.1 タイムアウトによる送信

アプリケーションから送信要求されたデータは、TCPによってセグメントごとに分割され、ソケットの情報とともに下位層に渡される。そのため、送信順序制御機構では、TCPのコネクションを識別することはできるが、1回の送信要求で生成されるTCPセグメントの数を把握することができない。また、送信順序制御機構が動作するのはTCPセグメントの送信要求が発生したときのみであるため、TCPの処理がセグメント群をすべて下位層に渡した後も、送信順序制御機構が送信を保留したフレームは送信されないままバッファに残ってしまう場合がある。この場合、TCPのタイムアウトによって、送信されなかったフレームに含まれるセグメントの再送信が発生するか、アプリケーションからの次の送信要求が発生するなどして、バッファにS個のフレームがたまるまでTCPセグメントの送信が保留される。これらのフレーム群はバッファにS個のフレームがためられた後のTCPセグメントの送信と共に送信される。そのため、最終的には通信が成立することになるが、再送信の発生によって解決した場合はウインドウサイズが大幅に縮小される。

今回の実装では、この問題を解決するためにタイマを利用し、タイムアウト時間を設けている。バッファにフレームが追加される度にタイムアウト時間を更新し、タイムアウトが発生するとバッファに残されたフレームをすべて送信することによって、バッファ内のフレームに含まれるセグメントの再送信を防いでいる。また、タイムアウトが発生した時のバッファ内にあるフレーム数が$m(m \leq S)$のとき、以下の規則で送信することにより、セグメントの到達順序が大きく変わることを防いでいる。

- つぎのフレームが、広帯域のNICで送信されるときは、バッファの先頭のフレームを送信する。
- つぎのn個のフレームが、狭帯域のNICで送信されるときは、バッファの後ろからn番目のフレームを送信する。

タイムアウトによってバッファから送信されたセグメントの受信確認が、TCPの再送信までのタイムアウト時間に間に合わないと再送信が発生するため、送信順序制御のタイムアウト時間は、受信確認がTCPの再送信までのタイムアウト時間に間に合う長さにする必要がある。また、次のTCPセグメントが送信される前に送信順序制御のタイムアウトが発生すると、送信順序制御機構が機能しないため、送信順序制御のタイムアウト時間はTCPのオーバーヘッドによるTCPセグメントの送信間隔より十分長くする必要がある。今回の実装では送信順序制御のタイムアウト時間を10msとした。

3.2 ウインドウサイズの拡大

TCPではフロー制御のためにセグメントの送信量をウインドウサイズで制限している。コネクション確立時はウインドウサイズが最小で、セグメントの到着を確認するごとに徐々に拡大され、再送信が発生すると縮小される。ところが、ウインドウサイズがSより小さい場合、TCPではウインドウサイズの分だけセグメントを送信しているにもかかわらず、そのセグメント群を含むパケットがすべて送信順序制御機構のバッファに納められてしまう。この問題は、コネクション確立時のウインドウサイズが小さいときに起こりやすく、さらに再送信の発生によってウインドウサイズが縮小され、帯域幅が一向に広がらない状態に陥りやすい。

このような状態を防ぐため、ウインドウサイズが$S+1$未満とならないようにしなければならない。今回の実装では、ウインドウサイズの初期値の計算、およびウインドウサイズを縮小するときの再計算を行う関数で、ウインドウサイズが$S+1$未満にならないよう変更した。

4 評価

提案手法をLinuxコンピュータに実装し、その有効性を実験によって確認した。実験には、PentiumIII 1GHzのCPUと192MBのメモリを搭載したコンピュータ2台と、スイッチングハブLSW-GT-8Wを用いた。各PCには、1000Base-TのNICであるPRO/1000-MTと100Base-TXのNICである3C905-TXを1枚ずつ装着した。オペレーティングシステムはLinux(カーネルバージョン2.2.17)である。図9にスキップ数の変化に対する帯域幅と再送信発生率を示す。スキップ数が0の場合が従来手法である。スキップ数が8のとき帯域幅は最大である373Mbpsとなっており、従来手法の143Mbpsに対して260%の拡大が実現されている。この帯域幅は、同じ環境でUDPの通信を行ったときの帯域幅914Mbpsの40.8%である。この実験のNICへのパケット配分率は、UDPの配送において最大帯域幅を得た1000Base-Tに90%、100Base-TXに10%としている。また、再送信発生率は従来手法の9.79%に対して提案手法では7.17%となっており、26.8%の削減となっている。なお、このデータは再送を観測するソフトの負荷の影響を受けて帯域が低下しているため、ホストで再送状況を観測せずに実験したところ、パケット配分率を1000Base-Tに99.6%、100Base-Tに0.4%、スキップ数を18としたときに最大帯域幅714Mbpsとなり、1000Base-T単体のときの689Mbpsを超える結果となった。

一方、図10はウインドウサイズの時間変化を示したものである。図10より提案手法では従来手法の3倍のウインドウサイズを得ることができている。

以上により、送信順序制御機構を導入することによ

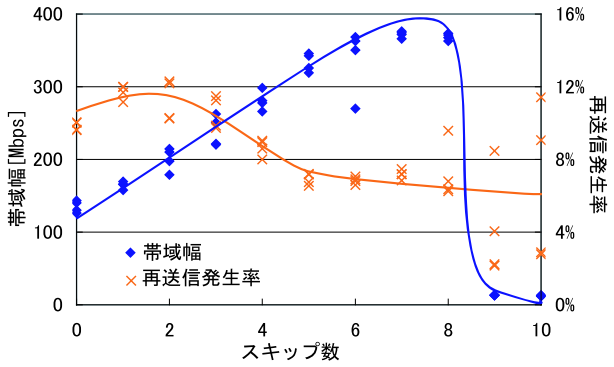

図 9: スキップ数の変化に対する帯域幅と再送信発生率

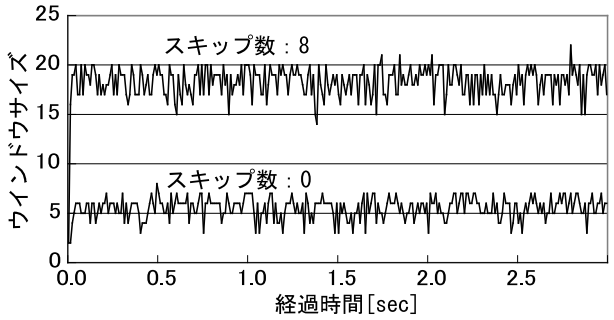

図 10: ウインドウサイズの時間変化

り TCP の通信に対しても広帯域通信を提供することができるといえる。

5 まとめ

本論分では、LAN に接続された複数の NIC を装着したコンピュータ間の通信を拡張 ARP に基づいて広帯域化する手法を TCP 通信に用いる場合に問題となる再送信要求の発生を抑制するために、狭帯域通信路で配送されるフレームの送信順序を早める手法を提案した。しかし、再送信の発生を完全には抑えられていないことから、スキップ数の最適値が逐次変動していると考えられるため、スキップ数の設定を動的に行うことが今後の課題となる。また、多様な NIC 数の組み合わせに対するスキップ数を定め、提案手法が一般に有効であることを示す。

参考文献

[1] David, C. P. , “An Ethernet Address Resolution Protocol,” RFC 826 (1982).

[2] Thomas, D., “bonding device,” tadavis@lbl.gov (1985).

[3] 加藤, 梅島, 森田, 桧垣, “複数種 NIC による広帯域 LAN のための ARP の拡張と実装,” 電子情報通信学会総合大会論文集, p. 351 (2003).

[4] 出口, 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構築と評価,” 信学技報, Vol. 100, No. 670, pp. 129–134 (2001).

[5] 中田, 杉木, 梅島, 桧垣, “複数種 NIC による広帯域通信のための配分率制御機構の実装,” 電子情報通信学会論文集, p. 327 (2004).

[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).