マルチメディア通信と分散処理 74－18
グループウェア 15－18
（1996. 1. 25）

ATM高速通信ボードの構成と評価

丸山 充
NTTソフトウェア研究所

オンボード上で多重のTCP/IPセッションを高速に処理可能なATM通信ボードのプロトタイプを開発した．本稿では当プロトタイプボードのハードウェアアーキテクチャの紹介，および処理性能の評価結果を示す．ヘッダ／コンテンツ分離型プロトコルデータユニット管理方式，Virtual Circuit対応on-the-flyチェックサム算出機構の提案技術とウィンドウスケーリング，パイプライン再スケジューリング，チェイン型DMAの各技術を採用した当プロトタイプボードのデータ転送性能は，ボード対向時に133.4Mbpsを達成した．

Design and Evaluation of ATM High-speed Communication Board

Mitsuru MARUYAMA
NTT Software Laboratories

Using TCP/IP for the high-layer protocol and ATM for the low-layer protocol, the author has developed a prototype communications board that shifts the most of the protocol processing overhead to hardware. This paper proposes two new solutions: "PDU management using header contents parallel processing" and "on-the-fly checksum calculation for ATM virtual circuits". Using these techniques, single-connection TCP/IP protocol processing performance of 133.4 Mbps was achieved.

1 はじめに

近年，ネットワークの高速化と転送される情報のマルチメディア化により，各ノードの通信処理部には高速なプロトコル処理性能が要求されている．またマルチメディアサーバでは，動画／音声などの実時間メディアを複数のクライアント端末に同時サービスする事が要求されるため，高速性とともにプロトコルの多重処理性能も重要となっている．

Internet上で，広く使われているTCP/IPプロトコル (Transmission Control Protocol / Internet Protocol) を通信処理部で高速に処理するためには，(1) メモリ管理の高速化およびメモリコピー量の削減，(2)TCPチェックサム算出のオーバヘッドの削減，(3) プロセス切替えのオーバヘッドの削減を行なわなければいけないことが，明らかになっている [1, 2, 3, 4]．本問題を解決し，FDDI (Fiber Distributed Data Interface) を下位レイヤに用いTCP/IPのスループットの向上を図った例が，複数報告されている [1, 5]．

このたび，我々は下位レイヤにATM (Asynchronus Transfer Mode) を適用し，オンボード上で多重のTCP/IPセッションを高速に処理可能なATM通信ボードのプロトタイプを開発した．本稿では，高速プロトコル処理のためのハードウェア新技術として新たに開発した，ヘッダ／コンテンツ分離型プロトコルデータユニット管理方式，VC (Virtual Circuit) 対応on-the-flyチェックサム算出回路を提案する．またこれらの新規技術とウィンドウスケーリング，パイプライン再スケジューリング，チェイン型DMAの各技術を採用したプロトタイプボードを開発し，そのデータ転送性能を測定することで，提案技術の有効性を検証する．

2 ATM高速通信ボード

プロトタイプボードの構成を図1に，仕様を表1に示す．図1に示すようにプロトタイプボードは，汎用のCPUボードであるCPU部に新規ハードウェア要素であるNIC部が接続された構成をとる．プロトタイプボード上では，図2に示すプロトコル群をオンボードで処理を行なう．IPパケットをATMネットワーク上で転送する方式は各種存在するが，本プロトタイプボードでは現時点での相互接続性に優れているRFC1483 [6] 準拠のClassical IP方式を採用した．また信号方式は，PVC (Permanent Virtual Circuit) を実装している．ATMレイヤ部分に関しては，汎用SAR (Segmentail and Reassembly) LSIおよびTC (Transmission Convergence) LSIで処理を行

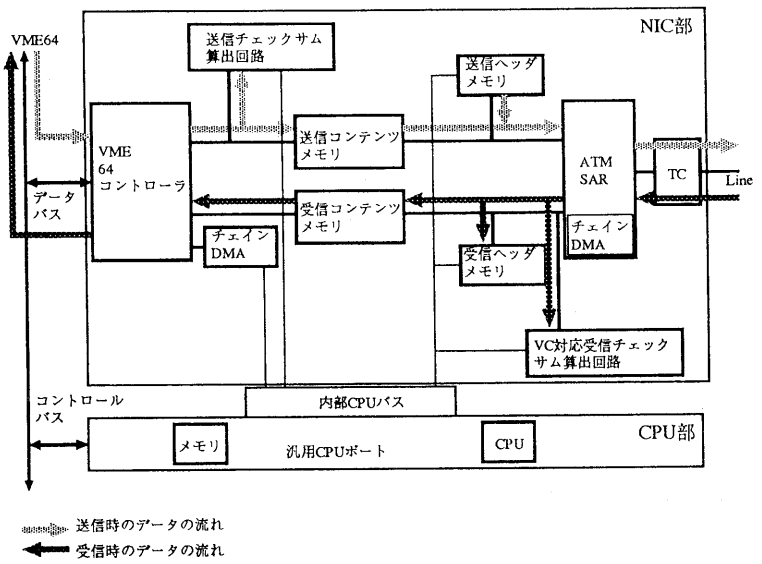
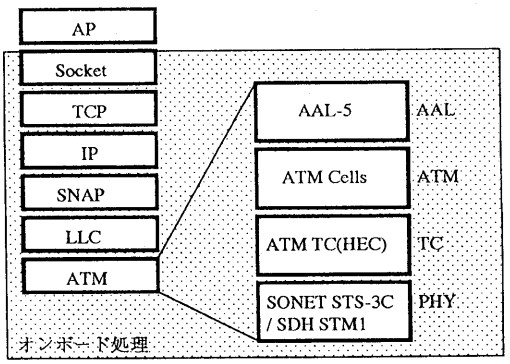


図 1　ATM 高速通信ボードの構成

表 1　ATM 高速通信ボードの仕様

CPU 部機能	仕様
CPU 性能	64 SPECint92
メモリ	8-MB RAM
I/O	Ethernet
OS	汎用 Real-time OS
コントロールバス	VME

NIC 部機能	仕様
ATM コントローラ	汎用 AAL-5 SAR LSI
ATM-TC インタフェース	UTOPIA 準拠
TC コントローラ	汎用 TC LSI
物理インタフェース	SONET/SDH 156 Mbps, 1ch
送信バッファメモリ	8-MB DRAM
受信バッファメモリ	8-MB DRAM
データバス	VME64
送受信チェックサム	1 の補数加算
ボードサイズ	VME ダブルハイト (6U)



SNAP: sub network access protocol
LLC: logical link control
AAL: ATM adaptation layer
TC: transmission convergence
PHY: physical layer

図 2　プロトコルの階層構造

なう．また他ボードとのインタフェースは，物理的にはVMEバスを用い，論理的にはUNIX OSのソケットインタフェースを採用している．TCP/IP プロトコルは，様々な方法でインプリメントがされているが，本稿では，4.3 BSD[7] を基本にし，一部4.4 BSD-Lite[8] の拡張を行なっている．

2.1　ATM 高速通信ボードの送信動作

送信時のデータの流れは，図1の上部に示すものである．本ボードは，外部のホストボードからのコマンドをCPU 部のコントロールバスを介して，受け取る事で動作を開始する．また動作の結果ステータスは，レスポンスとしてホストボードにコントロールバスを介して伝えられる．コマンド／レスポンスの送受にはホストボードとプロトタイプボードの両方がアクセス可能な共有メモリを使用する．本システムでは，CPU 部のメインメモリをVMEバス側からもアクセス可能な構成にして，共有メモリとして用いる．

データ送信時には，ホストボードから送信指示コ

マンドが発行される．送信指示コマンドのパラメータには，宛先IPアドレス，TCPポート番号に加えて，送信データを格納しているVMEバス上のアドレスも含まれる．プロトタイプボードでは，CPU部でこの情報を解釈し，該当のVMEアドレスから送信データをNIC部の送信コンテンツメモリにDMA転送する．同時に，転送と同じサイクルで同一データを送信チェックサム算出回路に転送し，データ全てのチェックサム算出が行なわれる．またCPU部では，IPアドレス，TCPポート番号などの情報を元に，TCP, IP, SNAP, LLCのヘッダ処理を行ない，NIC部でハードウェア算出されたチェックサム値にヘッダ部チェックサム値を補正した後，このデータを送信ヘッダメモリに書き込む．その後SARに転送起動を行なうと，SARは送信ヘッダメモリと送信コンテンツメモリ内のデータを連結して転送を行なう．SARでは該当のデータをAAL-5の形式とみなしてセルに分解し，TCでフレーミング処理後，回線に送出する．

2.2 ATM高速通信ボードの受信動作

データ受信時には，SARが自局宛てのセルを受信するごとに，AAL-5のフォーマットに組み立てて転送を行なう．この時にSARは，ヘッダ部を受信ヘッダメモリに，コンテンツ部を受信コンテンツメモリに分離しながらデータ転送を行なう．またSARが各バッファに書き込むサイクルに同期して，受信チェックサム算出回路もバス上のデータをモニタし，VCごとのチェックサム値を算出する．受信ヘッダメモリにデータが転送されると同時に，CPU部によるヘッダ解析がはじまり，ヘッダが正常なことの確認，どのポート宛てのデータかの判別処理が行なわれる．受信データが全て受信コンテンツメモリに格納された後に，該当するVCの受信チェックサムの算出結果を読みだし，受信ヘッダ部の中の値と検証後，NIC部のVME64バスを介して外部のメモリボードにDMA転送を行なう．

3 プロトコル処理の高速化技術

本節ではプロトコル処理の高速化の技術として，ヘッダ／コンテンツ分離型プロトコルデータユニット管理方式，VC対応on-the-flyチェックサム算出機構の新規技術を提案し，ウィンドウスケーリング，パイプライン再スケジューリング，チェイン型DMAの従来提案技術についても述べる．

3.1 ヘッダ／コンテンツ分離型プロトコルデータユニット管理方式

通常TCPのメモリ管理には，mbufというリスト構造のデータを用いる．mbufデータ構造は，可変長のデータバッファを割り当てる際には，リストのチェインに接続するだけで済むために便利であるが，プロトコル内のデータの分割，結合の際には，データを改めてコピーしなおす必要があるために，バッファの割り当て管理，開放に時間がかかる問題点がある．また通常のワークステーションなどでは，プロトコル処理によるメモリ転送が頻繁におきる．データ送信時には，ユーザの連続したメモリ空間からカーネルの空間のチェインしたmbufデータ構造にCPUがバイト単位のコピーを行なう，逆にデータ受信時には，カーネル空間からユーザ空間へデータのバイト単位のコピーをCPUが行なうために時間がかかってしまう[5]．

我々のプロトタイプボードでは，これらの問題を避けるために，送受信に用いるバッファは，ヘッダ部およびコンテンツ部にハード的に分離している．ヘッダ部はSARとMPU部からの同時アクセスが可能であり，コンテンツ部はSARとVMEバスからのアクセスが可能な構成である．このように，ヘッダ部とコンテンツ部のアクセスパスを完全に分離することにより，CPUとDMAのメモリアクセス競合を避けることができる．また，コンテンツ部は図3に示すように，一定サイズのページに分割されており，それぞれのページはオフセットアドレスと有効長で管理する．送受信に用いられるバッファは，図3に示すように1つのヘッダ部を示すポインタと複数のコンテンツ部のページを示すポインタの組から構成されるクラスタで管理される．この構成のクラスタをあらかじめ複数用意しておき，メモリアロケートの場合には，これをサイクリックに割り当て使用する．またメモリ解放の場合は，クラスタ単位でのメモリ解放を行なうことで，mbufのメモリ管理のような複雑な操作を不必要にしている．

またデータ転送時には，上記クラスタ単位の転送パターンで，VME-送受バッファ間，SAR-送受バッファ間のDMA起動を行なうことで，CPUによるコピー操作を削減している．

3.2 VC対応on-the-flyチェックサム算出機構

TCPチェックサムは，16bit単位に1の補数加算を行なうもので，通常は，ヘッダ部とコンテンツ部を含む全データをCPUがすべて読み出して計算するためオーバヘッドになる．

on-the-flyチェックサム算出は，TCPのチェックサム算出をデータ転送中に行ない，オーバヘッドを削減する手法である．FDDIやEthernetのように1フレーム単位で受信を行なうものであれば，そのフレーム全体をメモリに転送するサイクルと同時にチェックサムを算出することが可能であった．しかしATMの場合には，細かいセル単位で受信される上に，異なるVCに属するセルがランダムに到着する可能性があり，該当のセルのデータ転送が，どのTCPセッションに属するものかを判別しなければいけない問題点がある．

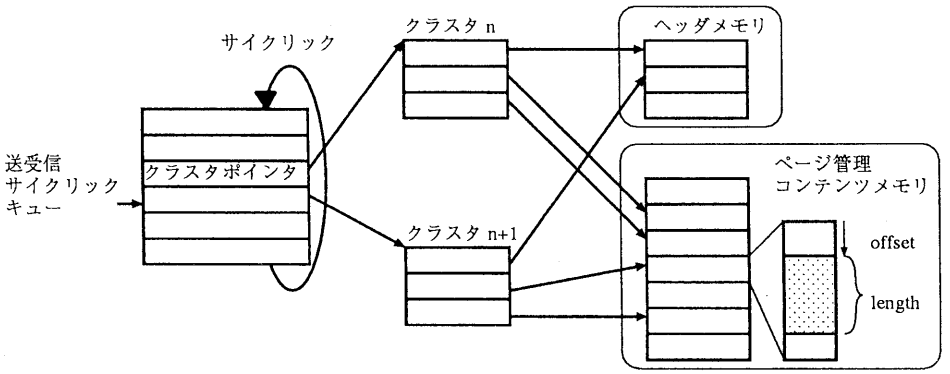


図 3　ヘッダ／コンテンツ分離型プロトコルデータユニット管理方式

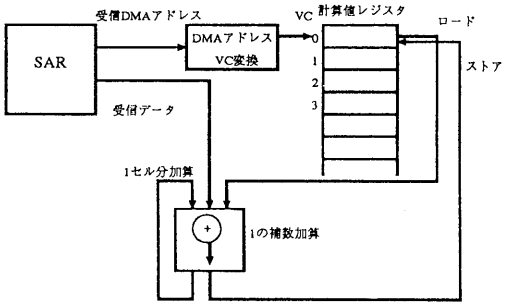


図 4　VC 対応 on-the-fly チェックサム算出機構

図 4 に示す VC 対応 on-the-fly チェックサム算出機構では，SAR からメモリに転送するときのアドレスが，VC ごとに異なるように設定しておく．実際の DMA データ転送中には，転送するアドレスを比較することにより，どの VC に属するデータを転送しているかを判別する．VC が決定すると，VC ごとに持つ計算値レジスタから途中計算値をロードし，データ転送サイクルで 1 の補数加算を行ない，転送終了後にその加算結果値を計算レジスタにストアする．これにより，複数の VC に属するセルをランダムに受信する場合でも，それぞれのセッションごとのチェックサム値を算出することが可能になる．

なお送信時には，VME バスから送信データバッファ内のページ単位で転送が行なわれるために，従来と同様に DMA サイクル中に該当のチェックサムを算出している．

3.3　その他の高速化技術

ここでは，従来提案しているプロトコル処理の高速化技術の中で，本プロトタイプボードで採用しているものを示す [1, 9]．具体的には，単体性能の向上のためにウィンドウスケーリングを多重性能の向上のために，パイプライン再スケジューリングおよびチェイン型 DMA をそれぞれ示す．

(1)　ウィンドウスケーリング

従来，TCP のウィンドウサイズは 16bit であったために，連続して 64KB までしか送出することができなかった．今回は，1 対向時のデータ転送の高速化を図るために RFC1323 [10] 準拠のウィンドウスケーリングの機構を実装することで，広いウィンドウサイズの転送が可能になった．

(2)　パイプライン再スケジューリング

多重処理性能の向上のためには効率良いパイプライン制御が必須である．送信パイプラインを乱す影響として，ACK などの非同期な受信動作があげられる．

ACK パケットなど非同期に受信されるパケットは，割り込みにより OS に通知し，受信処理の起動を行なってしまう．TCP では ACK 受信により，相手端末の正常な受信を確認し，送信メモリの開放処理が行なわれる．ACK 受信が非同期に起きると，その度にカーネル内でコンテキストスイッチが起き，送信処理が中断されるために送信パイプラインの動作を乱してしまう．このため，ACK およびそれに基づく処理をパイプライン動作上問題がないところまで遅延させることで，効率のよい送信パイプライン制御が可能となる．

(3)　チェイン型 DMA

効率良いパイプライン動作のためには，オーバヘッドの少ない DMA 起動が必要となる．本プロトタイプボードでは，SAR, バッファメモリ間および VME, バッファメモリ間の 2 つの DMA が存在する．それぞれの DMA は，コントローラに与えるコマンド群をメモリ上にチェインする事ができ，DMA コントローラは，そのコマンドをメモリ上から自律的に

表 2　評価に用いたパケットのオーバヘッド

オーバヘッド要因	実効転送レート
回線速度 SONET/OC-3C	155.520 Mbps
SONET のオーバヘッド	149.760 Mbps
セルヘッダのオーバヘッド 48/53	135.632 Mbps
パケット化オーバヘッド 8192/8256 *1	134.581 Mbps

*1: LLC/SNAP, IP, TCP 各ヘッダ 8, 20, 20 バイト, ペイロード 8192 バイト, AAL-5 トレイラ 8 バイト, セル分割フラグメント 8 バイト で 8256 バイト (172 セル分) として計算

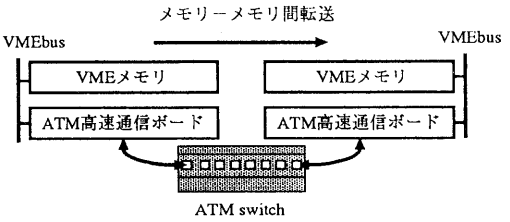


図 5　試験環境

獲得して，対象となるメモリ領域に対する DMA 動作を連続して行なうようになっている．この機能により，DMA 起動に基づくオーバヘッドを削減することが可能となる．

4　評価

本章では，プロトタイプボードの転送性能の評価を 1 対向送受信および多重送受信性能に分けて示す．

4.1　評価環境

本ボードの評価で扱っている TCP パケットのペイロードは，8192 バイトとした．この条件では，表 2に示すように各オーバヘッドを取り除くと，限界転送レートは 134.581Mbps になる．表で明らかなように，一番オーバヘッドになっている部分が，セルヘッダのオーバヘッドである．

試験環境は，図 5に示すように，本プロトタイプボードを 2 枚 ATM Switch を介して接続を行ないメモリ - メモリ間で TCP による通信性能の測定を 1 対向および多重セッションで行なった．スループットの測定は，プロトタイプボード単体の転送性能を評価するために，単体でデータの送受が可能なベンチマークプログラムをプロトタイプボード上に作成し，転送時間はプロトタイプボード上のタイマを用いて計測した．また，測定中は ATM アナライザを接続して，総計セル数と時間を計測して，検証を行なった．

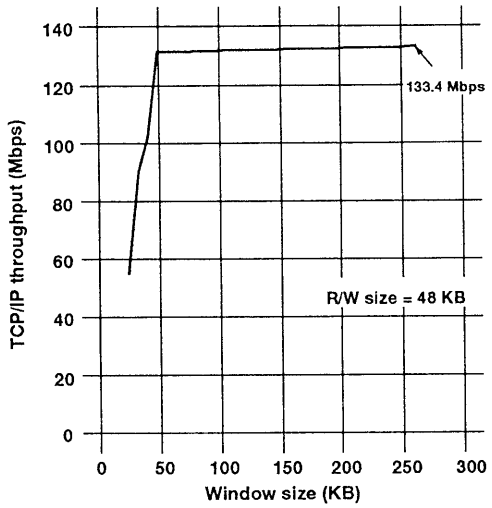


図 6　1 対向送受信性能

4.2　1 対向送受信性能

図 6が，本プロトタイプボードの 1 対向送受信性能である．横軸が，ウィンドウサイズ，縦軸がスループットを示している．グラフにより明らかなように最高スループットは，ウィンドウサイズが 256KB の時で，133.4Mbps であり，限界転送レートの 99% を達成していることが分かる．これにより，プロトコル処理性能の高速化の各機構は十分目的を達していることがいえる．

またウィンドウサイズが 48KB 以上では，ほぼ最高スループットと同じ値を示していることが分かり，あまりウィンドウスケーリングの効果は発揮されていない事も明らかになった．

4.3　多重送受信性能

図 7が，ウィンドウサイズごとの多重処理性能を示したものである．横軸に多重数，縦軸がスループットを示している．縦軸のスループットは，各セッションのスループットを加算したものを示している．

この結果，ウィンドウサイズが 16KB の時は多重時のピーク値 k が 131.2Mbps に達し，ウィンドウサイズ 16KB の場合には，60 多重時でも 1% 程度しか性能が下がらないことが分かる．ウィンドウサイズ 24KB では 60 多重時に 4.3% まで性能が下がり，ウィンドウサイズ 48KB では，50 多重時に 6.4% まで性能が下がる．またウィンドウサイズ 256KB では，15 多重を限界にデータが取れなかった．ウィンドウサイズを増やして，多重処理を行なった時に性能が低下する原因は，送受コンテンツメモリのページ数が不足してしまい，それ以上の要求があると，バッファ解放待ちになってしまう事による．またウィンドウサイズを 8KB に落すと 最大スループットが 10.6Mbps

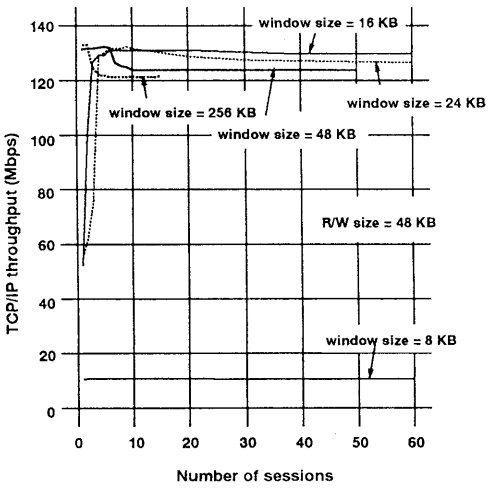


図 7　多重送受信性能

程度までしか出ない.

以上のことから，適切なウィンドウサイズにおいて多重動作を行なう限りでは，ほぼ理想どおりのパイプライン制御ができていることが明らかになった.

5　まとめ

高速性と多重性を兼ね備えた，ATM 通信ボードの処理技術を提案した．これらの技術を採用したプロトタイプボードの転送スループットを測定し，新技術の有効性を検証した．得られた主要な結果は以下のとおりである.

- 提案した技術の採用により，TCP/IP プロトコル処理性能は，ウィンドウサイズが 256KB の時で，133.4Mbps であり，限界性能の 99% を達成している.
- 多重性能はウィンドウサイズ 16-24KB 時に 60 多重時に 1-4.3% しか減少しない.
- 多重動作時のウィンドウサイズの拡大は，バッファメモリ不足を招き，逆にスループットを下げる原因になってしまう.

今後，本プロトタイプボードをマルチメディアサーバに組み込み，システム化を図るとともに，RTP (real-time transport protocol) などの新規プロトコルの実装，評価を行なうためのテストベッドとして使用していく予定である.

謝辞　本研究の機会を与えていただいた，NTT ヒューマンインタフェース研究所 メディア応用システム研究部の中野博隆部長ならびに 西村一敏グループリーダに感謝します．また本研究を最後まで支援してくださった，NTT ソフトウェア研究所 広域コンピューティング研究部の後藤滋樹部長ならびに高橋直久グループリーダに感謝いたします.

参考文献

[1] 丸山, 中野, 西村: “多重度の向上を目指したプロトコル処理技術の提案と評価”, 信学技報, IN93-113, pp.45-52, 1994.

[2] D. D. Clark, V. Jacobson, J. Romkey, and H. Salwen: “An Analysis of TCP Processing Overhead”, IEEE Communication Magazine, June, pp. 23-29, 1988.

[3] C. Papadopoulos and G. M. Parulkar: “Experimental Evaluation of SUNOS IPC and TCP/IP Protocol Implementation”, IEEE /ACM Transactions on Networking, Vol. 1, No. 2, pp. 199-216, 1993.

[4] P. Druschel, M. B. Abbott, M. A. Pagels, and L. L. Peterson: “Network Subsystem Design”, IEEE Network Magazine, Vol. 7, No. 4, pp. 8-17, 1993.

[5] C. Dalton, G. Watson, D. Banks, C. Calamvokis, A. Edwards, and J. Lumley: “Afterburner: A network-independent card provides architectural support for high-performance protocols”, IEEE Network Magazine, Vol. 7, No. 4, pp. 36-43, 1993.

[6] J. Heinanen: “Multiprotocol Encapsulation over ATM Adaptation Layer 5”, IETF RFC-1483, 07/20/1993.

[7] S. J. Leffler et al.: “The Design and Implementation of the 4.3BSD UNIX Operating System”, ISBN 0-201-06196-1, Addison-Wesley, 1989.

[8] Gary R. Wright and W. Richard Stevens: “TCP /IP Illustrated, Volume2: The Implementation”, ISBN 0-201-63354-X, Addison-Wesley, 1995.

[9] M. Maruyama, H. Sakamoto, Y. Ishibashi, and K. Nishimura: “High-speed hardware architecture for high-definition videotex system” , Journal of Electronic Imaging , Vol. 1, No. 4, pp. 349-357, 1992.

[10] D. Borman, R. Braden, V. Jacobson: “TCP Extensions for High Performance”, IETF RFC-1323, 05/13/1992.