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

2006－HPC－107 (52)
2006／8／2

ハードウェア・エンジンを用いた10GbE上のTCP通信解析

吉野剛史[†] 玉造潤史[††] 稲上克史[†]
菅原豊[†] 稲葉真理[†] 平木敬[†]

10ギガビットイーサネットと，そのWAN PHY技術により広帯域・広域パケット交換ネットワークが一般化してきている．しかし，こうしたネットワークをEnd-to-end通信で利用するには多くの障害がある．インターネット上での転送制御の標準であるTCPはアルゴリズムの設計上，そのままでは大きな遅延下で性能低下を起こす．多数のスイッチが引き起こすパケット転送のバースト化はパケットの喪失や通信ホストの負荷上昇に繋がる．こうした問題を検出して原因と箇所を明らかにするためにはトラフィックの詳細な解析が必要となる．

我々が開発したネットワーク解析エンジンTAPEE (Traffic Analysis Precise Enhancement Engine) は，PCとの協調により，10ギガビットのワイヤーレートでのパケットヘッダ記録を可能にする．ヘッダと一緒に記録された高解像度タイムスタンプを含めての詳細なTCP通信解析により，広帯域・高遅延ネットワークに適したソフトウェア・ハードウェアの研究開発をサポートすることができる．

本稿では，TAPEEの詳細を述べ，その性能評価とともに実際の問題検出を行い，有効性を示す．

Analysis of TCP Communications over 10GbE Using Hardware Engine

Takeshi YOSHINO,[†] Junji TAMATSUKURI,[††]
Katsushi INAGAMI,[†] Yutaka SUGAWARA,[†] Mary INABA[†]
and Kei HIRAKI[†]

10 Gigabit Ethernet and its WAN PHY technology made high-bandwidth and wide-area packet-exchange networks popular. But there are many difficulties necessary to be solved for using these networks in end-to-end communications. TCP/IP is standard mechanism for transmission control on the Internet. But on large-delay network its performance comes down because of its design. Many switches on routes make burst transmission, leading to packet losses and pressure on hosts. It is necessary to analyze traffic in detail for detecting and clearing up problems and places.

We made hardware engine for analysis of 10GbE. Cooperating with PC servers it enables recording traffic at 10 gigabit wire rate. By analysis of TCP communication using recorded headers and high resolution timestamps we can support research and development of software and hardware suitable for network with high-bandwidth and large-delay.

In this paper, we describe detailed structure of TAPEE, and evaluate its performance. Finally we practice problem detection to demonstrate utility of TAPEE.

1. はじめに

End-to-end通信における10ギガビットイーサネット(10GbE)の利用が広まり始めている．また，10GbEとともに登場したWAN PHYは，SONET/SDHによる既存の広域ネットワークを流用しての広域パケット交換ネットワークの構築を可能にした．このような長距離，広帯域のネットワークはLong Fat-pipe Network (LFN)と呼ばれている．

しかし，エンドノードから10ギガビットという帯域幅を有効に活用するには多くの障害がある．インターネットにおけるデータ転送の標準プロトコルはTCP/IPであるが，広帯域かつ高遅延の環境下では十分な性能を発揮できない．数百ミリ秒の遅延は必要とするバッファ量の増加や，輻輳制御アルゴリズムが到達確認パケットに基いていることに起因する性能低下をもたらす．また，経路上のスイッチによってパケット間隔が変化するためにおきるバースト的な転送は，パケットの喪失や受信ホストへの負荷の上昇に繋がる．

このように，10GbEの有効利用は単にネットワークインタフェースを置き換えるだけでは達成できない．ソフトウェアや転送制御アルゴリズムの改良とチューニング，ハードウェアとソフトウェアの協調が必要になり，そのためにはネットワークトラフィックの詳細な解析が重要である．問題が発生しているならば発生箇所を確認して使用している機器やソフトを見直さなければいけなく，

† 東京大学大学院情報理工学系研究科
Graduate School of Information Science and Technology, University of Tokyo
†† 東京大学大学院理学系研究科
Graduate School of Science, University of Tokyo

思うように性能が向上しなければ通信の振る舞いを詳しく調べて把握しなければならない．

ネットワークトラフィックの解析はPCとtcpdumpやEthereal といったソフトウェアだけを用いて行うことが出来る．しかし，こうした手法ではソフトウェア処理がボトルネックとなり，ワイヤーレート近くまで帯域を利用するマルチストリーム転送や，スイッチによってバースト化された高速トラフィックを解析することはできない．また，割り込みの取り扱いやコンテクストスイッチなどカーネルの振る舞いに影響されるため，正確な時刻データが得られない．メディアからNICへの入力時点でのパケットの振る舞いを測定することもできない．

以上の理由から，PC単体による解析では有効なデバッギングを行えない．ハードウェアで作られた専用解析機を用いるか，もしPCで解析するならば，転送処理を高速化，低負荷化するオフローディング技術と同様に，PCの解析能力を増強するハードウェアが必要になる．

本研究では，こうした要求に見合った解析手段提供のため，FPGAを用いた10GbE解析ハードウェアエンジンTAPEE（Traffic Analysis Precise Enhancement Engine）を開発した．この装置はPCとの協調により高速トラフィックのヘッダをハードウェアのタイムスタンプとあわせて長時間にわたり記録することができる．また，TCPのストリームを分析して通信解析を行うためのソフトウェアを整備し，統計・可視化を行った．

本稿は以降，次のように構成されている．2章でLFNにおけるTCP通信の現行課題について論じた上で，3章から6章で開発プラットフォームのTGNLE-1と，解析エンジンTAPEEの詳細について述べる．7章でパケット生成装置を使ったキャプチャ能力評価実験，8章でTAPEEを用いてのTCP通信の精細な解析と問題検出の2つを行いTAPEEの有効性を実証する．9章で他研究との関連，違いを述べ，10章でまとめる．

2. 広域・広帯域ネットワーク上のTCP通信

大陸間などの長距離通信においては，光の伝送遅延だけでも100msを超える状況となるため，遅延を考慮して通信を行う必要がある．TCPはその転送制御アルゴリズムがRound Trip Time（RTT）に影響を受ける設計となっており，遅延の大きい通信ではパフォーマンスの減少がおこる．1つはCongestion Windowのスケールアップであり，RTTが大きい経路ではスケールアップの速度が減少してしまうため，コネクション確立後の立ち上がりやパケットロス発生からの回復において帯域利用効率が悪化する．もう1つは大きなRTTのもとではIn-flightデータ量がWindow Size/RTTによって制限されることによる性能低下である．Windowサイズを大きくすることで対処できるが，Windowサイズを大きくすると数百MBのバッファメモリを抱え，ホストに大きな負荷がかかる．どちらの問題に対しても多くの改良研究がなされているが，確実なものは存在せず，パラメータを適切に選び，チューニングをかけることで初めて性能を発揮する．TAPEEでパケットのヘッダと時刻を記録してトレースすれば，WindowとIn-flightデータの変化を確認でき，再送との対応を見ることもできる．Windowの急激な上昇によるロス発生や負荷軽減のためのアルゴリズムが正しく働いているか調べ，TCPスタックや転送ソフトウェアのチューニングに反映させることができる．

また，長距離の経路は必然的に多くのスイッチを経由して構成され，各スイッチがパケットの間隔を変化させる可能性がある．SONET/SDHを用いたWAN PHY経路ならばSONETパケットへの変換でパケット間隔の変化も起こりうる．ルーティングを行う機器とL2スイッチングだけ行う機器でのパケット処理の違いも考えられる．パケットのバースト化はボトルネック経路でのパケット喪失に繋がり，TCP/IPでは特にパケット長の短いAckの間隔の減少でホストやNICに局所的に大きな負荷がかかる．このため，バーストを吸収する仕組みを備えたNICを用いてCoalesingをしたり，オペレーティングシステムとNICの協調を行い，場合によってはスイッチの見直しをする必要がある．バースト化の原因は様々なので，解析装置を使ってマイクロ秒単位でパケットを記録して回線構成や測定位置ごとに比較することが重要になる．

3. 10GbE Testbed TGNLE-1

解析エンジンを実装するプラットフォームハードウェアにはTGNLE-1[3]を用いた．TGNLE-1はFPGAを用いた再構成可能な10GbEパケットプロセッサである．TGNLE-1は300pin MSAとIntel IXF18104 10 Gigabit Ethernet LAN PHYチップを用いた2つの10GBASE-LRインタフェースを備え，それぞれの受信インタフェースが他方の送信インタフェースとFPGAを介してつながっている．パケット処理を行うFPGAにはXilinx社のVirtex-II Pro XC2VP50を用いている．同じくXilinx社のSPI-4.2のSink/Source IP CoreによりFPGAとIXF18104との間でイーサネットフレームの受け渡しを行う．受信したパケットはFPGAにユーザが実装した回路によって自由に処理した上で再送出することが出来る．

また，TGNLE-1はUSBのインタフェースを備えており，PCからTGNLE-1内のレジスタの読み書きが行える．各FPGAにはバッファ用として2GBのDDR266 SDRAMが接続されており，擬似遅延機能やペーシング機能で用いる．本研究ではDRAMを用いた機能は使用

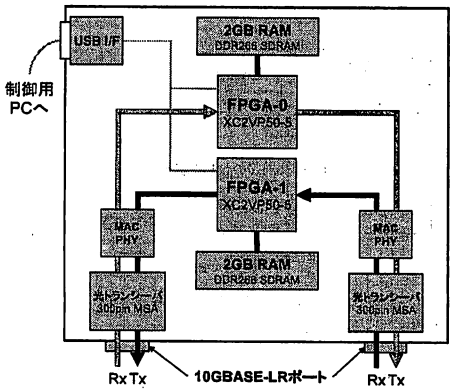

図 1 TGNLE-1 の構成図

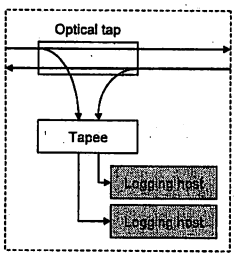

図 2 TAPEE の構成図

していない．

FPGA 内の基本構造は受信バッファ，FIFO，および送信バッファからなっている．受信バッファブロックは受信パケットに付加情報を付与して FIFO へ格納する形に変換を行う．送信バッファブロックは付加情報を取り出し，適切なタイミングで送信する．これらの基本ブロックに必要な変更を加えることで，解析のための各機能を実装した．本研究で行った実装は TAPEE を実現するパケットヘッダ記録サポート回路と，パケット生成回路である．

4. TAPEE

TAPEE は PC によるパケット記録をサポートする装置であり，専用の回路をロードした TGNLE-1 と記録用 PC，および光 Tap により構成される．光 Tap は入力された光を 2 つのファイバに向けて元の 50%の強さで分割するものを使い，一方を本来の送信先である通信ホストへ，もう一方を TAPEE へ送るよう接続した．

TAPEE は受信パケットからヘッダを含む数十オクテットを切り出し，ソフトウェア的処理に影響されない正確なタイムスタンプを打つ．さらに，取り出したデータを複数個まとめて 1 つのパケットに再構成して記録ホストへ送出する．こうして，ディスクへの記録と割り込み

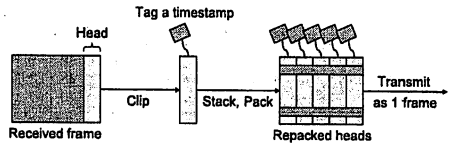

図 3 TAPEE の概念図

表 1 記録ホストの仕様

機種	IBM eServer x345
CPU	Intel Xeon 2.4GHz Dual
メモリ	2GB
NIC	Chelsio T110 Chelsio TOE Network Driver 2.1.4
OS	Linux kernel 2.6.12.6 Red Hat Enterprise Linux WS 4

で記録ホストにかかる負荷を軽減し，高速なトラフィックを高精細な時刻データを用いて解析することを可能にする．

TCP の通信解析にはヘッダ情報さえ取得できれば十分であり，ヘッダを解析すれば Ack や Seq の遷移，転送レートをはじめとして，In-flight データの量，再送や duplicate ack の発生，パケットのバースト等が確認できる．

記録ホストには現在 IBM eServer x345 を利用しており，仕様は表 1 の通りである．x345 は Chelsio T110 を搭載し，TAPEE から受け取った再パックパケットのダンプイメージをディスクに記録する．

記録した TCP ヘッダは PC で読み取り TCP の振る舞いを見る場合はストリームごとにわけて時刻，Ack，Seq 等の遷移を確認，Gnuplot を用いた可視化を行えるようソフトウェアを整備した．遅延や時間間隔などは PC で集計を行ったうえでヒストグラムなど見やすい形に可視化される．

TAPEE は測定パケットから抽出した情報を直接 10GbE でホストに渡すため，測定機器特有の機器や仕様に沿う必要は無く，PC 上でユーザがヘッダ情報を自由に処理して分析することができるという柔軟性を持っている．

5. パケット生成装置

今回 TAPEE の性能評価には TAPEE と同じく TGNLE-1 上に実装したパケット生成装置を用いた．これは PC から設定した長さのダミーパケットを自分で FIFO に挿入し続け，それを送信するものである．送出レートはクロック（133MHz）レベルで調整できる．パケット生成装置を用いてネットワーク機器をプッシュすること転送限界や高負荷下で起きる異常を検出することができる．ダミーパケットにはシーケンスがうってあり，評価実験では TAPEE でキャプチャしたデータにあるこ

表 2 FPGA の消費リソース

機能	スライス	DCM	BlockRAM (RAMB16s)
パススルー	7364 (31%)	5 (62%)	30 (12%)
パケット生成	6638 (28%)	5 (62%)	28 (12%)
再パック	7545 (31%)	5 (62%)	30 (12%)

のシーケンスを確認することで喪失の有無を調べる.

6. FPGA の消費資源

使用した Xilinx Virtex-II Pro XC2VP50 は 23616 スライスで構成されており，133MHz のクロックで動作させている．各機能の回路を実装する際に消費したスライス数は表 2 の通りである．使用した CAD は Xilinx ISE 6.3.03i，記述言語は Verilog HDL である．

7. 解析能力の検証

ワイヤーレートの測定能力を検証するため，パケット生成機能によってシーケンス付きトラフィックを生成しドロップせずにキャプチャできるかどうかを確認した．実験環境の構成は以下の通りである．パケット生成回路をロードしてある TGNLE-1 の 2 つの 10GbE インタフェースを XG800 に接続し，それぞれを実験用ホスト opteron5 と opteron7 と同じ VLAN に載せてある．パケット生成 TGNLE-1 と XG800 の経路上に光タップを挿入して信号をコピーし，TAPEE に送ってキャプチャする．実験では TAPEE の能力を確認すると共に，単に 10GbE をそなえたマシンでキャプチャする場合の性能も調べ，比較した．

まず MTU が 9000 octet のジャンボフレームトラフィックに対しての実験を行った．MTU が 9000 のとき，MAC 宛先アドレスと EtherType フィールドおよび FCS のあわせて 18 octet を加えるとパケット長は 9018 octet であり，Inter-frame gap と Preamble を含めるて 1 パケットあたり 9038 オクテット分の信号を占有する．よって，ワイヤーレートで送信した場合 LAN PHY での毎秒パケット数は 10Gbps / 9038 = 約 138kpps となる．先頭 80 octet を最大 20 個再パックするように設定し，実際に 20 分間のキャプチャを行った結果，全くドロップすることなく約 1 億個のパケットのヘッダ情報を記録することが出来た．

MTU が 1500 octet のスタンダードフレームの場合はハードディスクへの記録処理が追いつかず，連続記録時間はライトバッファとして利用可能なメモリ量によって制限された．現在の構成では約 20 秒が上限となった．実際，MTU に 1500 octet を用いる複数のホストが 10Gbps の帯域を埋め尽くす場合，10Gbps / 1538 octet = 約 813kpps でパケットが飛ぶことになり，この場合，TAPEE が出力する再パックデータのレートは約

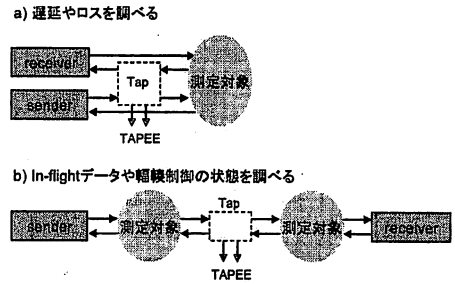

図 4 測定の形態

78MB/sec となる．今回使用したホストの構成はこのようなレートでの長時間記録には性能不足である．解決策としては，RAID 等による 2 次記憶の性能増強や複数ホストを用いたロードバランシングが考えられる．

一方，TAPEE を用いずに XG800 に接続されたホスト単体で先頭 80 octet の記録を行った場合，MTU9000 でも約 8%のパケットを喪失し，記録できなかった．

以上のように，TAPEE との協調の結果，ジャンボフレーム（MTU = 9000）による場合はハードディスクの許す限り，Standard パケットでもメモリの許す限りワイヤーレートでの記録が行えることが示された．

8. 実際の問題検出

本装置は既に 1) で用いられており，実 LFN と擬似 LFN のミクロな振る舞いの違いを明らかにすることを可能にした．本稿でも，いくつかの状況下で TAPEE を用いた測定を行い，問題検出を行った．

TCP 通信に対する，測定の形態は図 4 に示すように 2 つある．目的に応じて測定箇所を選び，記録後複数のログをあわせて解析する．

8.1 ミクロなレート遷移

現在，我々は PCI-X 2.0 バスが利用可能な IBM eServer x260 を 2 台用いて，Chelsio T310 を使った転送実験を行っている．PCI-X 1.0 では最大周波数のものでもバス転送速度がボトルネックになるが，PCI-X 2.0 を用いることで 10Gbps のデータを受け渡すのに十分な速度になる．x260 は OS カーネルが Linux Kernel 2.6.12，32GB のメモリを持ち，CPU は Intel Xeon 3.66GHz 4 個である．

この x260 対向環境での Iperf による TCP 転送のミクロな振る舞いを TAPEE で測定した．TCP の輻輳制御方式は BIC-TCP を用いている．実験は 2 種類の構成で行った．1 つ目の実験は IPv4 で，受信側で TCP Offload Engine（TOE）を，送信側で TCP Segmentation Offload（TSO）を有効にして転送を行ったものであり，結果を示したグラフは図 5 である．2 つ目の実験

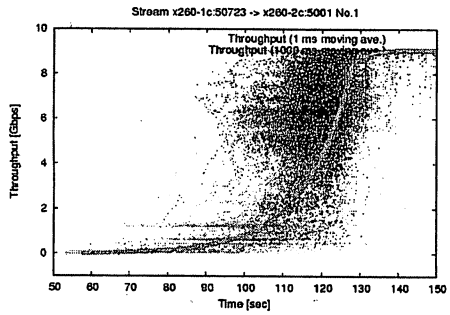

図 5　T310 IPv4 実験 転送レートの 1 ミリ秒と 1 秒での移動平均

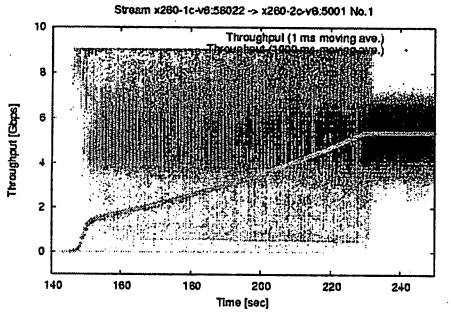

図 6　T310 IPv6 実験 転送レートの 1 ミリ秒と 1 秒での移動平均

はIPv6で行ったもので，結果を示したグラフは図 6 である．使用しているNICはIPv6でTOEやTSOが使えないため，ソフトウェアがTCPの処理を行っている．結果のグラフはそれぞれ転送レートの移動平均をプロットしたもので，1ミリ秒の平均が濃い色の散布図，1秒の平均が薄い色の曲線である．

どちらの転送レートも1秒平均で見ると輻輳ウィンドウの増加に従って緩やかに上昇しているが，1ミリ秒で見ると違いがある．IPv4実験では，低速転送時にパケット間隔きちんと確保されてバーストが抑えられているが，IPv6実験では，転送開始時点から9Gbpsと1Gbpsの間を振動している．これはTCPの転送制御がウィンドウ方式で行われているため，ウィンドウの空き部分のデータを一気に送信しようとする結果である．イーサネットは転送レートを制御する機能を持っていないため，TCPスタックから送信リクエストが来ると，このようにバースト的な転送と転送無しの状態を繰り返すことになる．最終的に輻輳ウィンドウが帯域とRTTの積と一致してからはバースト性がなくなっている．ただし，ここで言う帯域とはホストの転送レート限界値のことである．

パケットのバースト化問題は現在ペーシングなどの研究によって解決が試みられているものであるが，こうしてTAPEEを用いることで実際に可視化して確認することが出来る．

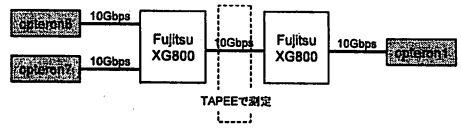

図 7　複数ストリーム実験の構成

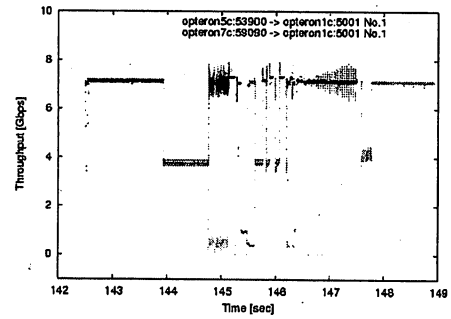

図 8　2 つのストリームが競合する様子

以上のように詳細な時刻データにもとづく分析を行うことで，構成の違いによる振る舞いの変化を明らかにすることが可能になる．

8.2　複数のストリームが競合する様子

送信ホストを2台用意し，1台の受信ホストとの間でIperfを行いTAPEEで測定した結果が図 8 である．実験環境を図 7 に示す．通信ホストopteron1,5,7に用いたPCはOpteron 248のデュアルでChelsio T110を搭載し，カーネルはLinux Kernel 2.6.12.6である．

2つのストリームは10Gbpsのインターフェースからそれぞれ送出され，スイッチから出る際にボトルネックである1つの10Gbps経路に合流する．測定はこのボトルネック経路に対して行い，通過するDataパケットとAckパケットを記録・解析した．

図8が2つのストリームが競合する様子を示したものである．それぞれのグラフが転送レートの1ミリ秒移動平均をあらわす．

図 9 がヘッダから分かるいくつかの情報を抽出し，転送レートと一緒に描画したものである．Ackの到着時点には高さ1の棒グラフがひかれており，それが重複Ackの場合は何回目の重複かを高さであらわす棒グラフを描いてある．シーケンスが前パケットのシーケンスとペイロード長の和より大きい場合，または小さい場合についてそれぞれy座標-0.25と-0.5の位置に点をうってある．

このグラフから減速が発生した箇所で重複Ackが発生し，再送が起こっていることがわかる．ヘッダの各フィールド値は全て別途記録されているので問題箇所を特定し

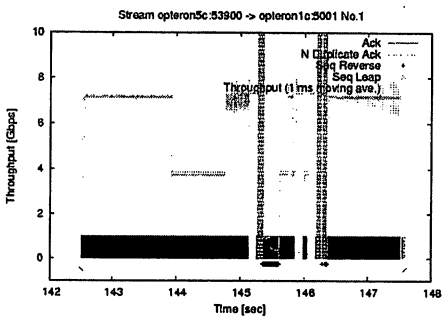

図 9 多重 Ack，シーケンス異常発生と減速の対応

た後は，より詳細な解析が行える．例えば，シーケンスとペイロード長を確認した結果，データパケットが失われていれば測定箇所より送信ホスト側でパケット喪失がおきていることがわかる．

9. 関連研究

10GbE に対する商用のハードウェアアナライザは IXIA，SmartBits，Sniffer などをはじめ多数存在し，必要な処理の多くを専用ハードウェアが担うが，本装置はそれらと異なり，ヘッダデータをそのままユーザが PC で受け取り，自由に解析できるようにする PC 解析能力増幅装置という位置づけをしている．

本研究のほか，10GbE に対するプログラマブルなパケット処理装置の開発には GtrcNET-10[2]，XGE-ProtoDevel[4] などがある．前者は高精細なレート測定機能を備えているが，具体的な問題検出へ実用した研究ではなく，長時間にわたっての全パケットダンプを用いた詳細解析を可能にしたものでもない．後者は長時間の解析やエラーエミュレーション機能を備えているが，LFN での TCP 通信ではなく，並列計算向けのプロトコル開発を対象としたシステム整備を行っている．

10. おわりに

我々は 10GbE 上での TCP 通信解析に適したネットワーク解析エンジン TAPEE を開発した．広帯域・高遅延での TCP 通信は多くの問題点を抱えており，問題検出に重要な高精細かつワイヤーレートに対応できる解析手段を TAPEE は提供する．TAPEE は，ハードウェアでの前処理によって PC サーバ単体では解析し得ないトラフィックの精細な解析を可能にし，また，ヘッダデータを直接 PC サーバに落とし込むことで，ユーザの目的にあった柔軟な解析作業を可能にした．記録は長時間にわたって可能であり，問題の発生と限られた記録時間のタイミングをあわせる必要は無い．

評価実験ではパケット生成装置によるプッシュ実験で TAPEE の記録性能を確認し，実際に TCP 通信の測定を行い問題検出への有効性を実証した．ミクロな現象は netstat による統計や SNMP による荒い解析ではわからず，こうした解析によってはじめて明らかにすることが出来，有効なデバッグが可能になった．

謝辞 本研究の実施に当たって東京大学の加藤朗助教授，JGNII の方々，IEEAF の方々，Pacific Northwest Gigapop（米 Seattle），AlaxalA Networks の方々には多くの支援を頂いた．また，実験に用いた 10GbE の広域回線は JGNII，SURFnet，IEEAF，CANARIE から提供を受けている．本研究は文部科学省 科学技術振興調整費「重要課題解決型研究等の推進-分散共有型研究データ利用基盤の整備」，科学技術研究費基盤研究 B「メモリ階層を考慮した高速検索アルゴリズムとそのハードウェア化の研究」，および 21 世紀 COE「情報科学技術戦略コア-大域ディペンダブル情報基盤」により実施された．

参考文献

1) 玉造潤史，吉野剛史，稲上克史，　菅原豊，稲葉真理，　平木敬：Real Long Fat Network における TCP/IPv6 の通信性能評価，電子情報通信学会技術研究報告 IA2006-4，電子情報通信学会，pp.19–24 (2006).
2) 児玉祐悦，工藤知宏，清水敏行：10GbE 対応ネットワークテストベッド GtrcNET-10 の構成と評価，並列/分散/協調処理に関するサマー・ワークショップ (SWoPP2005)，情報処理学会 (2005).
3) 菅原豊，稲葉真理，　平木敬：細粒度パケット間隔制御の実装と評価，情報処理学会技術研究報告 OS-100，情報処理学会，pp.85–92 (2005).
4) 中島耕太，住元真司，　佐藤充，久門耕一，　石川裕：性能問題を解決する 10Gbps 級ネットワーク向け通信プロトコル開発支援ツールの提案，先進的計算基盤システムシンポジウム (SACSIS2005)，情報処理学会，pp.321–328 (2005).