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

2006－ＱＡＩ－19（４）
2006／5／24

Real Long Fat Network における TCP/IPv6 の通信性能評価

玉造 潤史† 吉野 剛史‡ 稲上 克史‡
菅原 豊‡ 稲葉 真理‡ 平木 敬‡

†東京大学大学院理学系研究科 〒113-0033 東京都文京区本郷 7-3-1
‡東京大学大学院情報理工学系研究科 〒113-0033 東京都文京区本郷 7-3-1
E-mail: {junji, ysn, inagami, sugawara, mary, hiraki}@is.s.u-tokyo.ac.jp

あらまし 海外回線を用いた実回線高帯域高遅延ネットワーク(Real Long Fat Network)におけるソフトウエアIPv6の性能を測定し，通信時の状態をワイヤーレートのIPパケットヘッダを採取可能なLFNロガーにより観測した．その振る舞いを擬似ネットワーク環境との比較により解析する．

キーワード TCP/IP 通信， IPv6， Long Fat Network

Performance evaluation of TCP/IPv6 on Real Long Fat Network

Junji TAMATSUKURI† Takeshi YOSHINO‡ Katsushi INAGAMI‡
Yutaka SUGAWARA‡ Mary INABA‡ and Kei HIRAKI‡

† Graduate school of Science, University of Tokyo 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033 Japan
‡ Graduate school of Information Science and Technology, University of Tokyo 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033 Japan
E-mail: {junji, ysn, inagami, sugawara, mary, hiraki}@is.s.u-tokyo.ac.jp

Abstract We measured Software TCP/IPv6 performance on Real Long Fat Network by oversea circuits, and observed its behavior by wire-rate IP header capture (LFN Logger). We compared the communication behaviors on Real network communications with Pseudo Long Fat Network by network emulator.

Keyword TCP/IP, IPv6, Long Fat Network

1. はじめに

Grid による科学技術計算やデータ複製のため，世界規模での離れた 2 地点を結ぶ高遅延広帯域のネットワーク Long Fat Network(LFN)の利用が増加している．LFN は多くの場合 SONET/SDH OC-192 による長距離回線を用いており，SONET を Ethernet として利用可能とする WAN PHY 技術も一般化してきた．Grid 構成要素であるホスト PC の 10GbE(10 ギガビットイーサネット) ネットワークインターフェース(NIC)も一般化した．現在の PC の多くが持っている I/O バスである PCI-X1.0 バスは帯域が 8Gbps であるため，単一のネットワーク接続が利用可能な帯域が制限されている．しかし，擬似的に構成された LFN 上でホスト PC は I/O バスの限界近い 7 Gbps 以上の TCP/IP 通信が可能であることはすでに示されている[1][2].

LFN 上においての TCP 通信は，大きな遅延のため大きな輻輳ウインドウを必要とすること，高速のパケット生成，パケットコントロールを行わなければならないこと，NIC から多くの割り込みが掛かること，などホストの負荷が大きい．しかし，データ転送においてデータの到着性保証と経路上での輻輳制御を行う TCP 通信が利用できることは重要である．

通常 TCP 通信によるホスト負荷を減少させるために 10GbE NIC はセグメント処理のオフロード(TSO: TCP Segment Offloading) や TCP スタック全体のオフロード(chelsio の TOE: TCP Offload Engine)をサポートしている．特に TOE は IPv4 では利用可能であるが，IPv6 では OS の TCP スタックによるソフトウエア処理によらねばならず LFN での活用は困難である．

我々はこれまでこれらの問題に対して擬似ネットワークを用いた環境でソフトウエア TCP/IPv6 の性能向上を行ってきた．本論文ではその結果に基づき実回線による Real LFN でのソフトウエア TCP/IPv6 による通信実験を行い，その性能と擬似環境との比較を示す．

実際のネットワーク回線では擬似ネットワーク環境と異なりネットワークを構成する物理回線，ルータなどの振る舞いが影響する．これらの実回線上での影響を測定するために 10Gbps レートでの高速なパケットログを収集する装置 Tapee (Traffic Analysis Precise Enhancement Engine)を開発し，計測に用いた．

本稿では擬似ネットワークと Real LFN での TCP/IPv6 の性能を示すとともに，Tapee による測定の詳細な結果と振る舞いの比較を示す．

2. 遅延発生装置による擬似 LFN

Real LFN での性能測定を行う前に，ネットワークエミュレータを用いて実際の回線と同じ遅延を挿入し擬似的に再現するネットワーク環境を構築した．この擬似 LFN 環境で遅延を変化させながらの通信性能を計測した．

2.1. 実験環境

使用したネットワークエミュレータは Anue 社製 H シリーズで，RTT 800ms までの遅延をワイヤーレートで実現できる．実験に用いたホストの仕様は表 1 の通りである．

CPU	Dual AMD opteron 250 (送信側), Dual AMD opteron 248(受信側)
Memory	2GB(Single Memory Bus)
MotherBoard	Rioworks HDAMA rev D
NIC	Chelsio N210
OS	Linux-2.14.7 original TCP stack

表 1:実験ホスト仕様

ネットワークドライバはカーネルに含まれて配布されているものを用いた．TCP の Congestion Algorithm は BIC-TCP を用い，通信アプリケーションは iperf version 2.0.2 を用いた．

これらの機材を NTT コミュニケーションズ大手町ビル内にある T-LEX に設置し，擬似環境と実回線との比較を行いながら実験できる環境を構築した．(図 1)

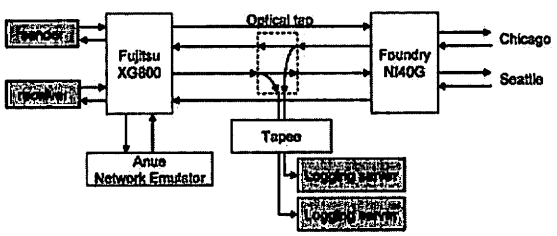

図 1:実験環境構成図(大手町 T-LEX)

エッジスイッチとして富士通の XG800 を用いており T-LEX のコアスイッチ NI40G との間に光タップを挿入し，計測するパケットを採取している．

2.2. Software TCP/IP 通信のチューニング

RTT の大きな高速 TCP/IP 通信を実現する上で必要なチューニング要素は(1)送受信に用いるウインドウバッファ，(2)iperf のアプリケーションバッファ,(3)TCP/IP スタックおよび NIC ドライバのパラメータである．(1)は理論的に次式 "ウインドウバッファ = 通信速度(Gbps) × RTT" によって求められる理論値を元に設定する．(2)はアプリケーションの性能に影響し OS のスケジュールとアプリケーションの動作によって決まる．送信側は 64kB, 受信側は 1024kB に設定した．これらの値は通信結果のバイナリサーチから最適値を求めたものである。(3)は現在の linux カーネルのドライバでは割り込み処理を高負荷な状況では Polling 処理に切り替える NAPI と Offload 機能として TSO が利用可能である．しかし，これら機能が不安定なため，NAPI は使わず送受信とも固定値の割り込み時間(データ送信側 40μs,データ受信側 100μs)になるよう受信処理の Coalescing を設定した．TSO の振る舞いを改善するために Outstanding な TSO Frame は 1 つだけに設定している．また，Delayed Ack はデフォルト値のまま用い，Ack パケット数はデータパケット数の半分になっている．

NIC からの割り込みは，プロセッサを固定して 1 台にかかるように設定した．また、アプリケーションも割り込みがかかる 1 台のプロセッサだけで処理し，TCP スタックも同一のプロセッサで動作する状況になっている．

2.3. 擬似 LFN での性能

擬似 LFN で RTT を変化させながら測定した性能を表 2 に示す．Packet サイズは 9,198Byte(IPv6 のデータ payload サイズ 9,138Byte)である．表中の w は iperf で設定する Window Buffer の最大サイズであり，送受信に同量のバッファを割り当てるため TCP スタックは最大で w 値の 2 倍のメモリを必要とする．性能はピークに達したときの安定性能であり，Scaling は TCP のウインドウサイズのスケーリングが終わるまでの時間である．

RTT (ms)	送信側 w (MB)	受信側 w (MB)	理論値 w (MB)	性能 (Gbps)	Scaling (sec)
0	1	1	0	7.13	0
50	40	50	42.5	7.12	4
100	80	100	85	7.12	8
200	160	200	170	7.12	17
300	240	300	255	7.12	33
400	320	400	340	7.12	40
500	400	500	425	7.12	50

表 2:擬似 LFN 環境における TCP/IPv6 性能

この結果では，RTT=500msまでの性能はリニアに送受信に用いるWindow Bufferサイズを変更することで性能向上が見られる．また，ピークに達したときの性能もRTTによらず同じ性能を得ることができることがわかる．Scaling時間は以前のTCP/IPスタックよりもDelayed Ackを正しく使用しているためRTTが大きな場合に多くの時間を要するようになっている．BI-TCPでは，スケーリング時間はかかるが安定的にピーク性能に到達し，持続することができる．

3. 世界一周回線による Real LFN 実験

実際の長距離回線を用いた Real LFN における TCP/IP 通信は擬似 LFN 環境よりも難しいと考えられる．回線のもつ特性や経路上のネットワーク機器の影響を大きく受け，パケットロスやパケット到着のジッタが発生するためである．Real LFN での性能と振る舞いを調べるため擬似環境における測定結果を元にして実際回線での通信実験を行い詳細なデータ収集を行った．

3.1. 実験環境

今回実験に用いたのは，JGN2 (Tokyo – Chicago), Surfnet (Chicago – Amsterdam), CA*net (Amsterdam – Seattle), IEEAF (Seattle – Tokyo) の回線である．CA*net は L1 の light path として提供されており，残りの回線は OC-192 であり，接続は全て 10GbE WAN PHY によるものである．これら用いて，起点終点が IEEAF 回線の接続ポイントである T-LEX になるように接続し，L3 ルータの設置ポイント間の距離合計 32,000km, RTT が約 500ms の世界一周経路の Real LFN を構築した．

経路上の各回線接続ポイントには Hitachi GS4000 2 台，Foundry NI40G 1 台が設置されルーティングを行い，ほかに 2 台の Force10 E1200 が L2 接続に用いられている．詳細な構成図を図 2 に示す．

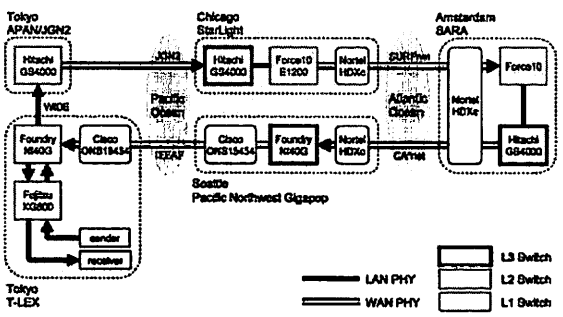

図 2: 海外回線構成図

通信に用いたホストおよび TCP スタックの設定は擬似環境 RTT=500ms の場合と同一であり，スイッチの設定を変えるだけで全く同じ動作をする状況で測定を行った．

3.2. LFN でのパケットログ採取

Real LFNの振る舞いを高精度に測定するため、10GbEヘッダキャプチャリングを行うための装置 Tapee (Traffic Analysis Precise Enhancement Engine)を開発し用いた． TapeeはFPGAによる再構成可能なネットワーク機器TGNLE-1[3]を用いて作られている．記録対象の通信路から光タップで信号を採取する．受信パケットフレームからヘッダを含む数十オクテットを取り出し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と送出する．こうして，tcpdump など通常のパケットキャプチャシステムよりもディスクへの記録とNICからの割り込みで記録ホストにかかる負荷を軽減し，高速トラフィックの高精細な時刻データを用いた解析を可能にする．

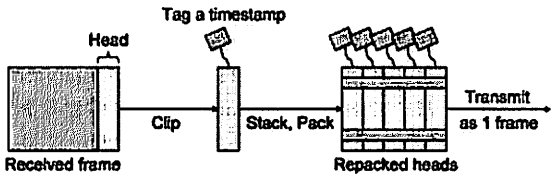

図 3: Tapee 概念図

3.3. Real LFN での性能

実回線上での TCP/IPv6 の通信の状況を図 4 に示す．

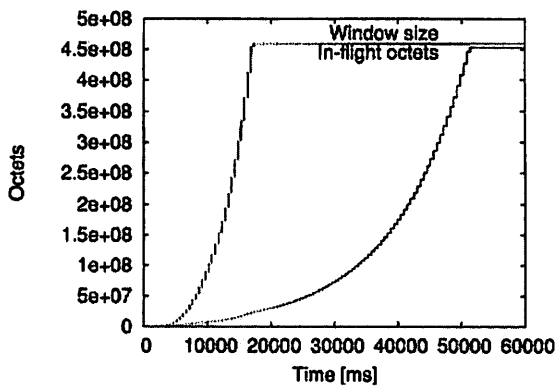

図 4:実回線での性能とウインドウコントロール

ピーク時の通信速度は，7.12Gbps であり，性能的には擬似環境での性能と同じであった．また，ウインドウのスケーリングに必要な時間も変わらず，ホストのマクロな振る舞いは同一であった．つまり，ホストアプリケーションからの見え方は同じになっている．この結果は十分にチューニングされたホストであれば Real LFN 上でも擬似環境と同一の性能を得ることができることを示している．

4. 擬似 LFN と Real LFN の比較

Tapee によって計測した擬似 LFN 環境でのパケットヘッダーログと Real LFN 環境でのヘッダーログを解析しこれらの違いを比較することでそれぞれのパケットの振る舞いを解析した.

特に, Real LFN での TCP パケットの振る舞いが 10Gbps レベルのネットワーク回線と機器を通った場合にどのような振る舞いを示すかを知ることは実際にネットワークを活用する場合に重要である.

Tapee には 2 個の 10Gbps インターフェースが装備されているため, 送信側, 受信側ホストのパケットを次の 2 通りに分類し解析を行った.

1. Sender から送出され LFN を越え, Receiver に到着する Data パケット
2. Receiver から送出され LFN を越え, Sender に到着する Ack パケット

これらを, Real LFN と擬似 LFN の RTT=500ms で TCP/IPv6 通信した場合のパケットで解析した.

4.1. Real LFN での Data パケットの振る舞い

まず, はじめに示すのは Real LFN での Data パケットの振る舞いである. この Data パケットの到着間隔をプロットしたものが図 5 である. RTT=500ms では TCP の輻輳ウインドウのスケーリングは約 50 秒(50,000ms) 掛かっている. 約 20 秒後に送出速度が約 600Mbps に達するまでは非常に緩やかにパケット間隔が減少し, その後は CPU と NIC への負荷が上昇するため若干のばらつきが発生している. しかし, 約 50 秒後に限界性能 7.12Gbps に到達するとほぼパケット間隔は収束する.

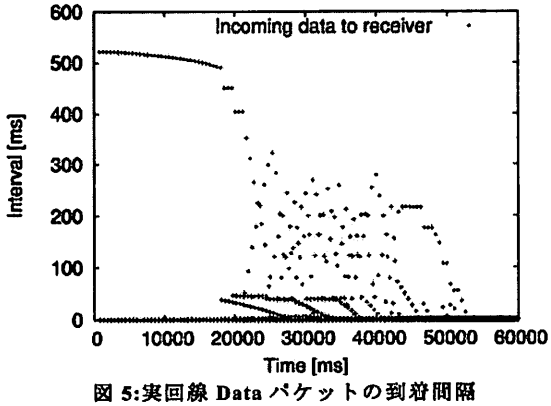

図 5:実回線 Data パケットの到着間隔

Jumbo Frame を使用しているためパケットサイズは 9,208Byte であり 7.12Gbps で通信した場合パケットの間隔は約 10 μ s となる. この状況を確認するために図 5 を 0.4ms までの部分で拡大したものが図 6 である.

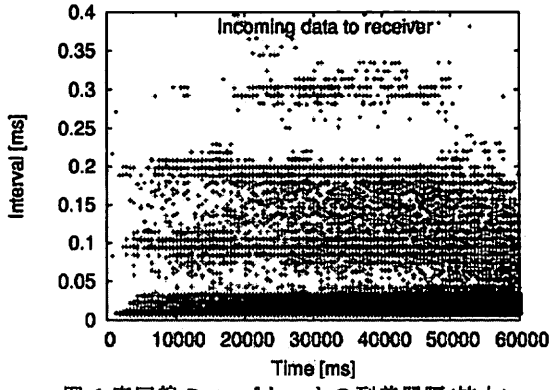

図 6:実回線 Data パケットの到着間隔(拡大)

図 6 では, ほぼ 10 μ s を中心としたパケット分布の非常に濃い部分と 0.1ms,0.2ms と 0.1ms ごとのパケットの分布がある部分がある. 前者は理論値どおりのパケット送出間隔であり, 後者は OS の 1TICK(1,000Hz) ごとのタイムスライスが出ていると考えられる.

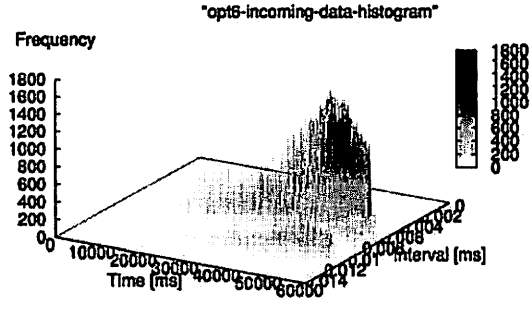

図 7:実回線 Data パケットの到着間隔
(図 6 のヒストグラム)

さらに, 図 6 のパケットの分布密度を見るためにヒストグラムとしたものが図 7 である. 全期間を通してパケットがタイムスライス単位で分散し続けていることが分かる. これは, RTT=500ms の場合, 送信側の CPU の Usage は, ほぼ 100%に達する. そのため送出時から送出するパケット間隔のばらつきがあり, 受信側ホストに到着してもその送出側ホストの振る舞いがそのまま現れている. つまり, パケットは受信側には送信時のパケット間隔が変わらずに届いている.

4.2. Ackパケットの振る舞い

TCP 通信において Ack パケットの振る舞いは重要である. 送出パケットに対応する Ack パケットを受信す

ることにより送出したデータをバッファ上から開放でき，輻輳ウインドウを大きくすることができる．しかし，送信側ホストは送出するTCPパケットの生成処理のため負荷が高く，小さなパケットが多数到着するAckパケット処理は困難である．また，データパケットが輻輳のためパケットロスした場合は経路上の帯域が足りないためウインドウが小さくなることは必要であるが，同じppsで小さなパケットが返されるAckが大きなppsのためにロスすることは本来避けなければならないことである．このため，Delayed Ackによりパケット数を削減してパケットロスの安定した配送を行っている．この状況が図8である．

Delayed Ackによりパケット数は半分であり，パケットを生成する受信側ホストは送信側ホストよりも負荷が低く約60%のロードである．そのため，パケットの散らばりがデータパケットよりも小さく安定して送られていることがわかる．

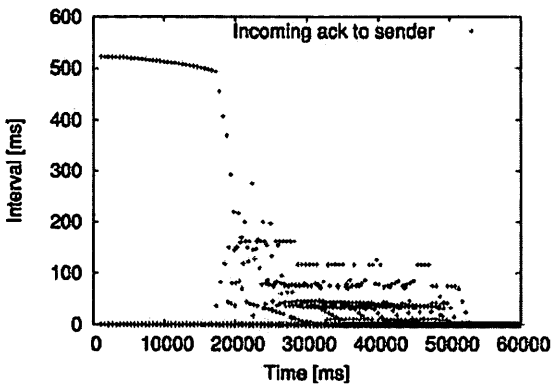

図 8:実回線 Ackパケットの到着間隔

図9はデータパケットと同様にパケットの集中している0.4msまでの部分を拡大したものである．この分散には2種類の間隔が含まれている．

ひとつはデータパケットと同様にOSのタイムスライスによるもの，もうひとつは格子的に示される点である．このような格子のような間隔はデータパケットでもわずかではあるが見ることができる．

この格子のように現れる間隔は送信側のパケットの分散だけをみると現れてこない．つまり，このパケットの収束はReal LFNによって発生したものである．

この分散をヒストグラムにしたものが図10である．上記の分散は観測されているもののほとんどのパケットの間隔はないことが分かる．これはAckパケットのサイズが非常に小さいため，配送系路上でパケットのマージが行われパケットの間隔がなくなっていることを意味している．多くのパケットが間隔なしで届くようにReal LFN上で変化している．

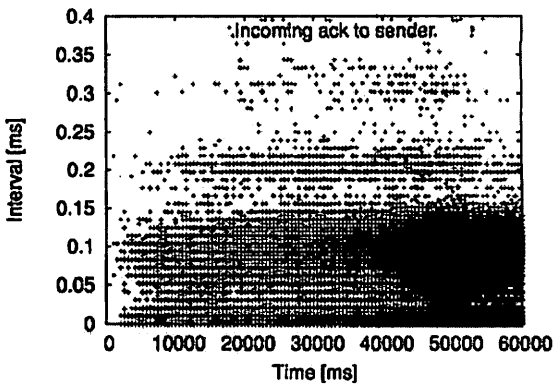

図 9:実回線 Ackパケットの到着間隔(拡大)

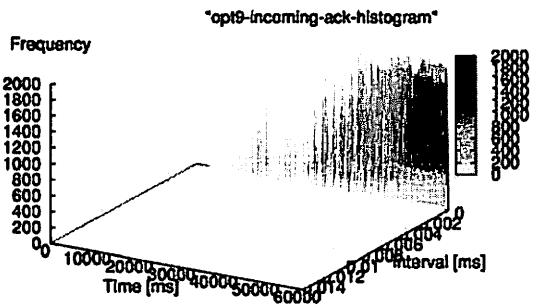

図 10:実回線 Ackパケットの到着間隔
(図9のヒストグラム)

この結果と対比するために，擬似環境でのAckパケットの振る舞いを示す．

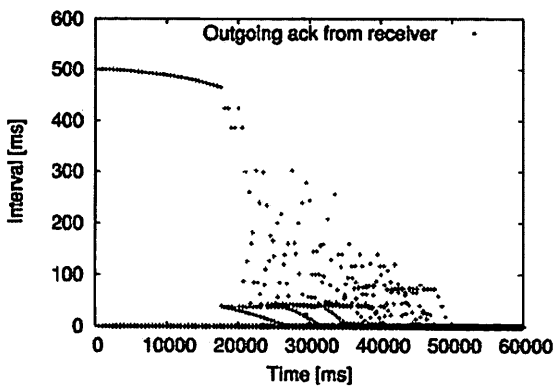

図 11:擬似環境 Ackパケットの到着間隔

図11は擬似LFN上でのAckパケットのマクロな振る舞いである．Real LFNでのパケットの到着と比較して大きな差はない．これを拡大したものが図12である．擬似環境でもタイムスライスによるパケットの分布が濃い部分が現れるが，実回線では現れていた格子状の

分布は存在していない.

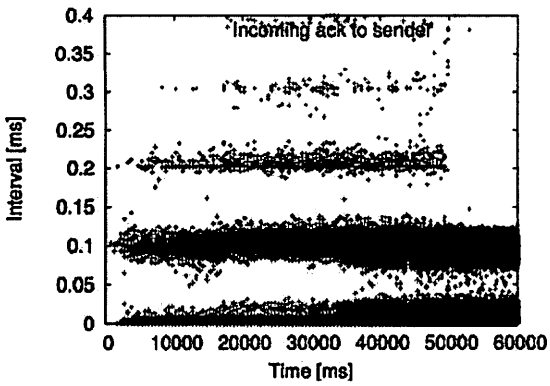

図 12:擬似環境 Ack パケットの到着間隔(拡大)

図 12 をヒストグラムにしたものが図 13 で図 10 と比較することにより，実回線と擬似環境との比較ができる．擬似環境では，パケットは送出された間隔を維持したまま受信側に届くためこの分散は Ack パケットの送出時の分散そのものである．

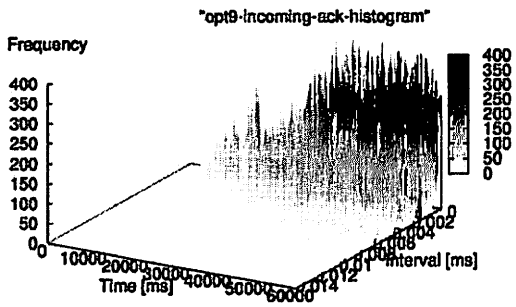

図 13:擬似環境 Ack パケットの到着間隔 (図 12 のヒストグラム)

まず，第 1 にマクロには安定して通信しているようであっても，ミクロなパケット送出間隔の分散があることは擬似環境のパケットヒストグラムが比較的大きな分散を持っていることから分かる．PCI-X バスがボトルネックとなる構成であるため，CPU が生成したパケットが NIC に渡される間隔のまま送出されている．

ところが，Ack のような小さなパケットは LFN を通過することでパケット間隔が消滅し，小さなパケットのバーストとして到着していることが分かった．今回のネットワーク構成では途中経路が WAN-phy であるため，Ethernet から SONET の Framing に変換される際に小さなパケットはマージされて配送され，それが受信端でのパケット間隔の無い到着となって現れていると考えられる．

5. まとめ

擬似 LFN と Real LFN それぞれの環境で TCP/IPv6 通信をした場合に同じ性能が得られることが示された．その性能はホスト PC の PCI-X 1.0 バスの限界に近い 7.12Gbps であり，RTT=500ms までの結果は変わらなかった．

しかし，ミクロなパケットの振る舞いは両環境では大きく異なり，特に Ack パケットの間隔は Real LFN を通過することでほぼ無くなることが分かった．

本研究の成果は TCP/IPv6 通信の性能向上に貢献しており，Internet2 が認定する Land Speed Record の IPv6 Single/Multi stream の記録を更新した[4],[5]．その記録は今回測定した経路とほぼ同一のものであり，転送性能は平均 6.96Gbps である．

現在ボトルネックがホスト I/O バスにあるが，PCI-X 2.0 が利用可能となりつつある．このようなホストでは CPU の処理能力が十分高ければボトルネックはネットワークとなり現在の状況とは異なる結果となることも考えられる．今後は，そのような環境での TCP 通信の性能向上を目指す．

謝　辞

本研究の実施に当たって東京大学情報基盤センタ加藤朗先生には多くの支援を頂いた．また,実験のため 10Gbps 回線を JGN2, Surfnet, IEEAF, Canarie から提供を受けた．

本研究は,文部科学技術省 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(2)15300014「アプリケーショントランスペアレントな大域データインテンシブ機構」,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で補助された．

文　献

[1] J.Tamatsukuri, K.Inagami, T.Yoshino, Y.Sugawara, M,Inaba and K,Hiraki, "Experimental Results of TCP/IP data transfer On 10Gbps IPv6 Network", 4th Workshop on Protocols for Fast Long-Distance Network (PFLDnet2006), Feb 2006. http://www.hpcc.jp/pfldnet2006/paper/s5_01.pdf,.

[2] 中村誠，玉造潤史，菅原豊，稲葉真理，平木敬,"擬似ネットワーク環境における TCP/IP の性能評価", 電子情報通信学会技術研究報告 IA2005-7,pp.1-8,Jun 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, pp.85-92, Aug 2005.

[4] Data Reservoir Project., ``Internet2 Land Speed Record in IPv6 single and multiple TCP stream",http://data-reservoir.adm.s.u-tokyo.ac.jp/lsr-20051114/index.html, Nov. 2005.

[5] Internet2 Land Speed Record, http://lsr.intenet2.edu