Vol.2010-IOT-11 No.5
2010/10/15

データセンタ間10Gbps回線におけるTCP中継器の実証実験

長谷川洋平† 地引昌弘†

近年，データセンタ間でのデータ共有など，長距離の大容量データ転送の要求が高まっている．一方，長距離通信ではTCPが原因となりスループットが制限されることが問題となる．このため，本研究ではTCP中継器を利用した長距離 TCP/IP 通信の高速化に取り組んでいる．本稿では，実際に敷設された約2600kmの10Gbps海底ケーブルで接続されたデータセンタにそれぞれ TCP 中継器を配置し，その効果を評価した結果を報告する．実験では，TCP中継器により，10Gbps回線を用いたファイルダウンロード時間を1/6に短縮するなどの効果を確認した．また，データセンタ内のネットワークにパケットロスがある場合には，最大で 70 倍ものスループット向上効果が得られた．

Evaluation of TCP Boosters on 10Gbps link between Data-Centers

Yohei Hasegawa† and Masahiro Jibiki†

TCP's performance is critical for data transfer throughput via a long distance network. We have proposed the TCP Booster (TCPB), which splits TCP connection in networks to enhance TCP throughput. With the TCPBs, TCP data transfer will be optimized for both congested terrestrial network and long distance submarine network. In this paper, we report field test results of the TCPB. In the field test, the TCPBs were placed in two data centers which were connected via 2600km 10Gbps network. With the TCPBs, file download via 10Gbps link was six times faster than the original TCP's throughput without the TCPBs. When packet loss occurred in the data-centers, the TCPBs improved TCP throughput by up to 70 times.

1. はじめに

近年，10Gbps回線の普及が進みつつあり，とうとう端末においても10Gbpsのイーサネットが利用可能となった．このように端末から長距離回線までの全てのネットワークで 10Gbps 通信環境が整いつつあり，長距離の大容量ファイル高速転送の実現が期待されている．

一方，インターネットで用いられる TCP/IP は長距離通信においてスループットを発揮しにくいことが知られている．TCPのおよそのスループットは端末間の往復遅延（RTT）に反比例し，また，同時にパケットロス率の平方根にも反比例し低下する[1]．このため，大きなRTTがあり，かつパケットロスが発生する場合は，特にスループットが低下してしまう．

この問題を軽減させるため，いくつもの高速 TCP が提案されてきた[2,3,4,5]．これら高速 TCP は，TCP の送信レート制御を変更することで，通常の TCP よりも高いスループットの達成を目指すものである．しかし，これら高速 TCP では比較的高いスループットが得られるものの，一般的にアグレッシブな制御を実現しているため，パケットロス率の悪化を招く可能性もある．また，他の TCP トラヒックと競合した際などに，他 TCP の性能を阻害する可能性もある．これら多様な要求を満たすような TCP を実現することは困難である．

このため，本研究では，ネットワーク内に配置した TCP 中継器にて TCP を終端し中継することで，特性の異なるネットワーク区間ごとにそれぞれ適した TCP コネクションを利用する通信方式を提案している．文献[9]では，実験室内において3320kmの長距離ケーブルを使用し，パケットロスが発生するネットワーク区間と，パケットロスが発生しない長距離ケーブル区間とのTCPを分割することで，スループットを向上できることを示した．

本稿では，前回の室内実験に引き続き，データセンタ間を結ぶ商用回線環境において実施した実験結果について報告する．実験では，TCP 中継器を約2600kmの海底ケーブル区間の両端に配置し，これを経由する2つのデータセンタ間の通信性能を計測した．

以降，本稿は次のように構成される．2章では長距離 TCP/IP 通信における性能問題と関連研究を説明する．3章では，提案している TCP 中継器について説明する．4章では約 2600km の商用通信回線を利用した TCP 中継器の性能評価結果を報告する．5章では本稿を総括する．

† 日本電気株式会社
NEC Corporation

Vol.2010-IOT-11 No.5
2010/10/15

2. 長距離 TCP/IP 通信における性能問題

ネットワークの遅延とパケットロスに対して TCP のスループットが大幅に低下することは広く知られており， TCP のスループットを向上させる方法については長年にわたる研究が行われてきている．TCP のおよそのスループットは端末間の往復遅延（RTT）に反比例し，また，同時にパケットロス率の平方根にも反比例し低下する．スループット B は簡単には次式のように表される[1].

$$B \approx \min\left(\frac{W}{d}, \frac{C}{d\sqrt{p}}\right)$$

ただし，W は TCP の最大ウィンドウサイズ，d は RTT，p はパケットロス率，C は定数である．

このスループットモデルの概形を図 1 に示す．(W=512Kbyte, $C=\sqrt{3}/2$) このように，遅延とパケットロスが組み合わされた状況では，TCP のスループットはまったく発揮されない．近年の長距離回線では，FEC(前方誤り訂正符号)によってビットエラーが隠蔽されるため，パケットロスが発生しない長距離伝送を可能としている，しかし，この長距離リンクに接続されるネットワークでは一般に端末からのトラヒックが集約される過程でパケットロスが発生してしまい，結果として，端末間経路は，遅延とパケットロスの両方があるものとなってしまう．

この問題を軽減させるため，いくつもの高速 TCP が提案されてきた．例えば，High Speed TCP[2]，FAST TCP[3]，CUBIC TCP[4]，COMPOUND TCP[5]などがよく知られている．これら高速 TCP は，TCP の送信レート制御を変更することで，通常の TCP よりも高いスループットを達成することを目指すものである．

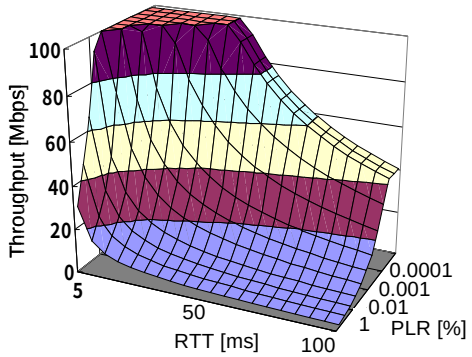


図 1 TCP のスループットと RTT，パケットロス率の関係

しかし，これら高速 TCP では比較的高いスループットが得られるものの，一般的にアグレッシブな制御を実現しているため，他の TCP トラヒックと競合した際などに，パケットロス率が上昇してしまい，他 TCP の性能を阻害する可能性もある．また，逆に，他 TCP トラヒックとの競合を考慮すると，理想的なスループットを発揮できない場合もある．このように多様な環境に適応し高い性能を発揮する TCP を実現することは難しい．

一方，ネットワーク内にて TCP を終端し中継することでスループットを向上させる方式なども研究されている[6,7,8]．もともと TCP 中継器は無線リンクなどビットエラーが発生するリンクが原因となるスループット低下を防ぐために用いられていたものである，これに対し，文献[6]の研究は，TCP の制御方法を変更することに加え，ネットワーク内にて TCP のフィードバックループを分割することで，パケット再送による性能低下も抑え，より一層高いスループットの発揮を目指したものである．我々は，TCP 中継器を用いることで，TCP のスループットを向上できることと，TCP 中継器が 10Gbps のスループットを達成できることを室内実験にて示した[9]．しかしながら，実際に敷設された長距離回線を用いて TCP 中継器の 10Gbps を実証した例はまだ報告されていない．

3. 長距離回線のための TCP 中継器

TCP 中継器は，端末間の TCP コネクションをネットワーク内で分割し，データを転送していく．例えば，2 つの TCP 中継器を用いることで，端末間のデータ転送は TCP1，TCP2，TCP3 の 3 つの TCP コネクションを介して行われることになる．図 3 には TCP 中継器を用いたパケットトランザクションの例を示す．この例では，コネクション開設を端末間で行った後，データ転送の際に TCP 中継処理を実施する例である．TCP 中継器は端末からは透過的に動作することができ，あたかも端末間で通信しているように利用できる．

本章では，これまで述べた問題を解決するため，長距離伝送回線をはさみ TCP 中継器を利用することを提案する．端末間の TCP1～3 はそれぞれ次のような部分ネットワーク区間を経由する．

TCP1，TCP3：パケットロスはあるものの遅延が小さい集線ネットワーク
TCP2：遅延はあるもののパケットロスが発生しない長距離伝送ネットワーク

これによって，TCP1～3 が経験する RTT もしくはパケットロス率のどちらかが小さくなることになる．図 1 を参照すると，それぞれのスループットが指数的に向上することがわかる． これによって，端末間の通信スループットを大幅に向上させることができる．

Vol.2010-IOT-11 No.5
2010/10/15

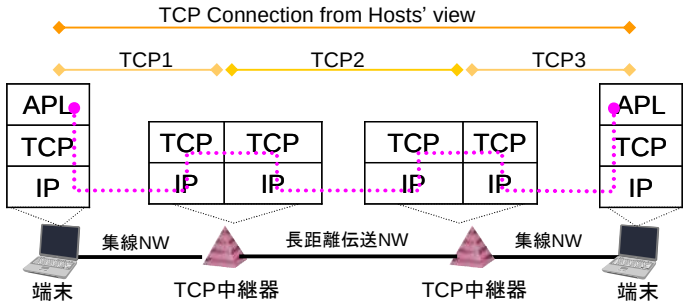


図 2　TCP 中継器を利用した通信の概要

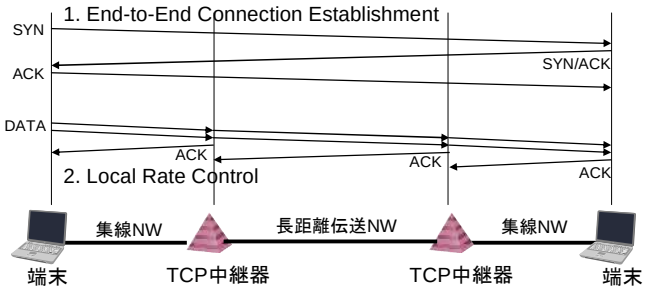


図 3　TCP 中継器によるパケットトランザクション

4.　性能評価

ここでは，約 2600km の商用回線を用いた環境における TCP 中継器の実証実験について報告する．まず，4.1 節にて使用した実験環境について説明する．次に，提案する TCP 中継器を導入することで一般的な端末の TCP/IP 通信スループットがどれほど向上するのか，通信性能を評価する．4.2 節では端末が 10Gbps 回線にて接続された場合の結果示し，4.3 節では端末が 1Gbps 回線にて接続された場合の結果を示す．

4.1　評価環境

今回の実験では，北米の通信キャリアが所有する 2 つのデータセンタとそれを結ぶ約 2600km の回線を利用した． 2 つのデータセンタを結ぶ回線の概要を図 4 に示す．端末と TCP 中継器が配置された 2 つのデータセンタをそれぞれデータセンタ 1，データセンタ 2 と呼ぶ．データセンタ 1 からの回線は約 80km の陸上 WDM 回線を経て陸揚局(Landing Station)に到達し，ここで海底ケーブル終端装置(Submarine Line Terminal Module)に接続される．陸揚局の SLTM からは約 2500km の海底ケーブルでデータセンタ 2 内の SLTM に接続される．データセンタ間の通信回線は計 2600km になり，パケットレベルのトラヒック合流が無いため，パケットロスは発生しない． WDM 装置は NEC DW4280 [10] を，SLTM は NEC T640M LTE SLTM-XG[11]が導入されている．

データセンタ内の回線接続の概要を図 5 に示す．それぞれのデータセンタにはクライアント／サーバとなる端末を 2 台ずつ配置した．クライアント 1 とサーバ 1 はギガビットイーサネット(GbE)でネットワークに接続され，クライアント 2 とサーバ 2 は 10 ギガビットイーサネット(10GbE)で接続される．クライアント／サーバと TCP 中継器はルータを介して接続される．このルータのスイッチファブリックには本実験以外のトラヒックも収容されている．また，さらに多様なネットワークのパケットロスや遅延を再現するため，クライアント 2 とサーバ 2 をネットワークエミュレータとしても使用した．

TCP 中継器は，長距離リンク区間を挟むように配置され，TCP 中継器間ではパケットレベルのトラヒックの合流が無いためパケットロスは発生しない．一方，端末からの TCP 中継器までの間では，ルータでのパケットロス，キューイング遅延が発生する可能性がある． また，実験ではネットワークエミュレータによっても遅延とパケットロスを発生させた評価も行った．

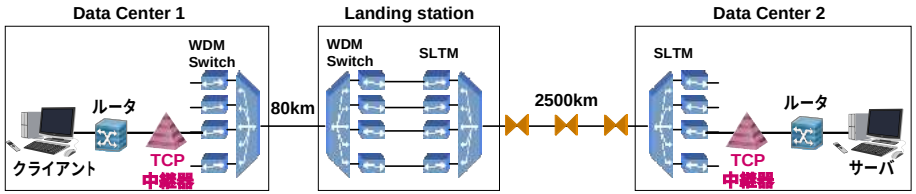


図 4　データセンタ間長距離回線の概要

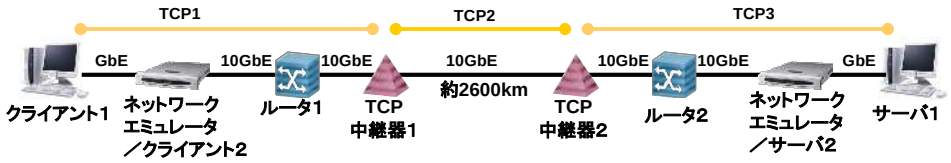


図 5　データセンタ内の回線接続

TCP 中継装置とネットワークエミュレータには NEC の Express サーバを用いた. TCP 中継器のスペック概要を表 1 に示す．また，クライアント 1 とサーバ 1 のスペックを表 2 に示す.

クライアントとサーバ間の通信性能を計測するためには，次の 2 種類のトラヒックを使用し，受信側でスループットを計測した.

1．HTTP によるファイルダウンロード.
サーバには Apache version 2.0.63, クライアントには wget version 1.11.4 を使用し，ファイルダウンロードでの性能を計測した.

2．Iperf 2.0.4 によるトラヒックを流した場合.
クライアントからサーバへ TCP トラヒックを 60 秒間発生させ，受信側でスループットを計測した.

表 1　TCP 中継器とネットワークエミュレータのスペック

Machine	NEC Express server 5800/R120a-2
CPU	Intel XEON 5570 2.93 GHz ×2
RAM	12GB (DDR3-1066, 1GB×12)
NIC	NEC 10GBASE-SR接続ボード N8103-123A (Chelsio S310E-SR)

表 2　クライアント 1 とサーバ 1 のスペック

Machine	N/A
OS	Windows XP (RWIN=64K) Linux (RWIN = 256KB)
CPU	Intel Core2Duo(2.66GHz)
RAM	4GB
NIC	1000BASE-T

4.2　長距離 10Gbps 通信の検証

ここでは, 10Gbps インタフェースを持つクライアント 2 とサーバ 2 の間の通信性能を計測した結果を示す. 図 6 には, クライアント 2 がサーバ 2 の WEB サーバから DVD1 枚分のデータに相当する 4.7GB のファイルをダウンロードするのに要した時間について，TCP 中継器を使用した場合と使用しない場合についてそれぞれ評価した．なお，端末の TCP バッファサイズはそれぞれ使用した Linux ディストリビューションのデフォルト設定の 256KB である.

結果，TCP 中継器を使用しなかった場合に 91 秒を要したダウンロード時間が，TCP 中継器を使用した場合は 16 秒にまで短縮された．このときのスループットは約 400Mbps から 2.4Gbps に上昇した．このように，TCP 中継器により約 6 倍の性能向上が得られた．このほか，実験では，TCP コネクションの本数を増やした場合に TCP 中継器が合計 10Gbps のスループットを達成できることも確認した.

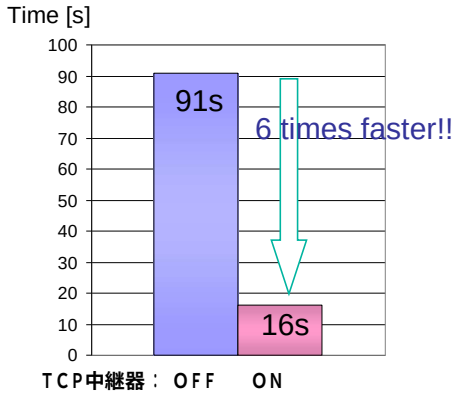


図 6　DVD1 枚分(4.7GB)のファイルダウンロード時間

4.3　ユーザ端末とデータセンタ

続いて，クライアント 1 で動作する OS に Windows XP，Linux をそれぞれ使用し，TCP 中継器を使用した場合と, 使用しない場合について計 4 通りの設定で iperf の計測をした結果を図 7 に示す.

Linux をクライアントとした場合，TCP 中継器が無いと 390Mbps しかスループットが発揮されなかったのに対し, TCP 中継器を利用すると 937Mbps のスループットが計測された．端末の回線インタフェースは 1Gbps のイーサネットであり，ヘッダ長のオーバヘッドを考慮した最大スループットは 940Mbps ほどである．実験では，これにほぼ等しい性能を発揮できたことがわかる.

Windows XP をクライアントとした場合，TCP 中継器を使用しない場合にはスループットが 18.3Mbps しか得られなかったのに対し，TCP 中継器を使用することで，806Mbps までスループットが向上した．これは 44 倍も高速化されたことになる.

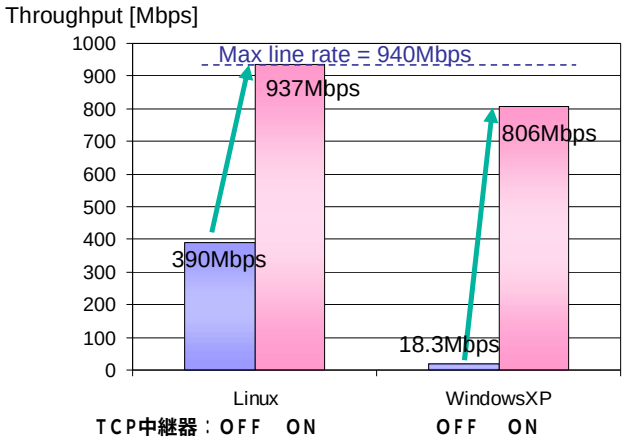


図 7　TCP 中継器の有無による DVD ダウンロード時間の違い

次にデータセンタ内におけるネットワークにも遅延とロスがある環境での TCP 中継器の効果を確認するため，ネットワークエミュレータでパケットロスと遅延を設定し iperf のスループットを計測した結果を示す．表 3 から表 6 には，クライアントに Linux もしくは WindowsXP を使用し， TCP 中継器を使用した場合と使用しない場合それぞれの結果を示す．また，これらをグラフ化したものを図 8，図 9 に示す．ネットワークエミュレータは,パケットロス率を 0%から 0.1%の間で設定し，遅延の設定を 0ms から 10ms の間で設定した．

Linux を用いた評価結果では，TCP 中継器を使用することで，TCP 中継器を使用しない場合と比較して，全ての評価結果で 2.4 倍以上のスループット改善が見られた．また，最大では約 70 倍のスループット改善が見られた(エミュレータ設定：パケットロス率＝0.1%，遅延 0ms の場合)．

WindowsXP を用いた評価結果でも，同様に TCP 中継器の効果が確認でき，全ての評価結果で 3.5 倍以上のスループット改善が見られ，最大では 62 倍のスループット改善が見られた(エミュレータ設定：パケットロス率＝0.1%，遅延 0ms の場合)．

以上の評価により，TCP 中継器の高い効果を確認することができた．ネットワークエミュレータを用いた評価により，データセンタ内などで遅延やパケットロスがある場合にも高いスループット改善効果が得られることが確認できた．特にデータセンタ内でパケットロスが発生した場合には，提案方式の効果が非常に高いことを確認した．

表 3　Linux クライアントのスループット(TCP 中継器無し)

		Delay [ms]				
		0	1	2	5	10
PLR [%]	0	390	345	305	233	21
	0.001	44.5	39.2	37.1	26.5	17.6
	0.01	29.7	22.7	17.8	13.5	12.7
	0.1	8.61	7.2	5.67	4.68	3.53

表 4　Linux クライアントのスループット(TCP 中継器あり)

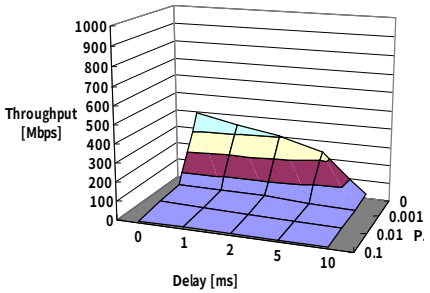
		Delay [ms]				
		0	1	2	5	10
PLR [%]	0	937	910	848	792	77
	0.001	936	509	311	142	71.7
	0.01	899	317	217	91.9	44.8
	0.1	606	124	73.5	32.1	14.5

表 5　WinXP クライアントのスループット(TCP 中継器無し)

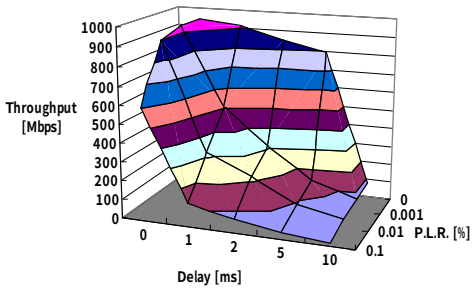
		Delay [ms]				
		0	1	2	5	10
PLR [%]	0	18.3	16.1	14.3	10.7	7.37
	0.001	17.3	15.1	14.1	10.2	7.23
	0.01	15.1	13.6	11.7	8.62	5.82
	0.1	7.62	6.85	6.07	4.81	2.73

表 6　WinXP クライアントのスループット(TCP 中継器あり)

		Delay [ms]				
		0	1	2	5	10
PLR [%]	0	806	201	115	49	26.1
	0.001	774	199	142	48.2	25.4
	0.01	736	182	104	46.9	24.7
	0.1	472	133	73.3	29.3	15.8

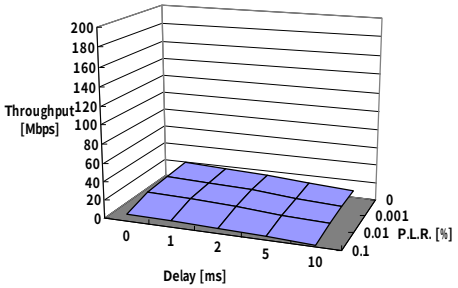


(a) TCP 中継器なし

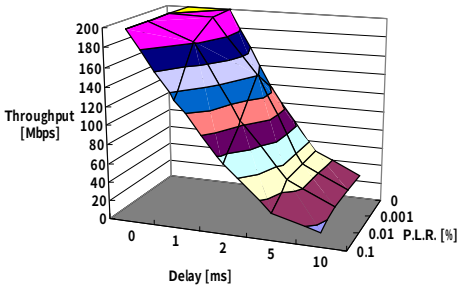


(b) TCP 中継器あり

図 8　Linux クライアントのスループット



(a) TCP 中継器なし



(b) TCP 中継器あり

図 9　Windows クライアントのスループット

5.　おわりに

本稿では，ネットワークを特性毎に分解するよう TCP 中継器を利用する提案方式を用いて，約 2600km の 10Gbps 回線で接続される 2 つのデータセンタ間での実証実験結果を報告した．実験では，一般的な端末間 TCP/IP 通信に提案する TCP 中継器を組み合わせることで 44 倍にスループットを向上させた．また，ネットワークエミュレータを用いてデータセンタ内のネットワークにパケットロスがあることを想定した実験では，最大で 70 倍ものスループット向上効果が得られた．

今後は TCP 中継器を，多様な環境に対応させるべく検討を進める予定である．例えば，TCP 中継器を通過するが，IP フォワードされる背景トラヒックへの影響などについても検討する．

参考文献

1)　J. Padhye, V. Firoiu, D. Towsley, J. Kurose, “Modeling TCP Throughput: A Simple Model and its Empirical Validation,” Proc. ACM SIGCOMM, SEP 1998.
2)　S. Floyd, “HighSpeed TCP for large congestion windows,” RFC3649, IETF, DEC 2003.
3)　C. Jin, D. Wai, and S. Low, “FAST TCP: Motivation, architecture, algorithms, performance,” Proc. IEEE INFOCOM, vol. 4, pp.2490-2501, MAR 2004.
4)　I. Rhee and L. Xu, “CUBIC: A new TCP-friendly high-speed TCP variant,” Proc.　PFLDNet, 2005.
5)　K. Tan, J. Song, Q. Zhang and M. Sridharan, “A Compound TCP Approach for High-speed and Long Distance Networks,” Proc. IEEE INFOCOM, APR 2006.
6)　T. Murase, H. Shimonishi, and Y. Hasegawa, “TCP overlay network architecture,” Proc. Comm. Conf IEICE’02, B-7-49, SEP, 2002.
7)　Y. Liu, Y. Gu, H. Zhang, W. Gong, and D. Towsley, “Application Level Relay for High-bandwidth Data Transport,” Proc. GridNets 2004, OCT 2004.
8)　長谷川洋平，村瀬勉, “TCP 中継ノードの高速プロトコル処理方式と性能評価,” 信学会ソサイエティ大会, SEP 2003.
9)　長谷川洋平，地引昌弘，“長距離 10Gbps 回線における TCP 中継器の評価，”情報処理学会，IOT 研究会，2010 年 3 月．
10)　“SpectralWave DW4200,” available at http://www.nec.co.jp/spectralwave/dw4200/.
11)　Y. Sato, T. Nakada, “ T640SW LTE Terminal Equipment for Optical Submarine Cable Systems,” NEC Technical Journal Vol.5, No.1, Feb. 2010.