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

2008－DSM－48 (1)
2008－QAI－26 (1)
2008／3／6

高品質インターネットから品質指向のインターネットへの飛翔

中川晋一†‡ 三角真† 知念賢一‡ 宮地利幸† 宇多仁‡ 篠田陽一†‡

† 情報通信研究機構 〒184-8795 東京都小金井市貫井北町 4-2-1
‡ 北陸先端科学技術大学院大学 〒923-1211 石川県能美市旭台 1-1
E-mail: †{snakagaw, misumi, miyachi}@nict.go.jp ‡{k-chinen, zin, shinoda }@jaist.ac.jp

あらまし：Layer2 スイッチ等によって敷設された物理線を複数の Layer2 に分割し，一本の物理線を多層レイヤ 2 ネットワークとして運用することは 実験用テストベッド等で一般的になっている．これらの合理性と利便性について JGN と APII テストベッドの例を検討することにより説明する．技術的に困難な点，特にユーザにとっての用件について,これらテストベッドを用いて,実験を行った経験から述べる．この種の根とワークの問題点である物理的トポロジとユーザの使用する論理層トポロジの違いに起因する問題点に対し， Layer2 ネットワークのボトルネックリンク稼動状態推定法について報告する．
キーワード：高品質インターネット，QoS，テストベッド，JGN，ATM，ポリシー指向ルーティング

From Quality Assumed Internet Toward the Internet as the Quality Oriented Communication Infrastructure

Shin-ichi Nakagawa[1)2)3)] Makoto Misumi[1)] Ken-ichi Chinen[2)]
Toshiyuki Miyachi[1)] Zin Uta[2)] and Yoichi Shinoda[1)2)]

1) National Institute of Information and Telecommunication Technology, Japan
2) Japan Advanced Institute of Science and Technology
E-mail: 1){snakagaw, misumi, miyachi}@nict.go.jp, 2){k-chinen, zin, shinoda}@jaist.ac.jp,

Abstract: Layer2 tagging based multilayered network is becoming popular and useful solution for the congestions of requests and policies especially on the exchange points of the Internet and Experimental Testbeds. Firstly, the flexibility and usefulness of that solution will described from the result of negotiation for the initial JGN, APII project and other testbeds in this country. Secondary, its issues on the technical points are indicated from our own experiences at the real experiments using that environment. From that, the discrepancy of logical IP routing structure and the physical configuration of real network problem are concerned. We will propose a simple method for estimation of the Layer2 data transport quality form the user layer as the solution for the quality oriented Internet.
Key words: Layer2, testbed, JGN, ATM, Policy based routing, multilayered network

1. はじめに

ルータで構成されていた広域ネットワークが，Layer2 スイッチで構成される場合が増加している．1998 年から運用されてきた研究開発用ギガビットネットワーク(JGN)[1]は，1 世代目には ATM が，2 世代目(JGN2)ではギガビットイーサネットを基本とした共通の物理線を複数の Layer2 ネットワークとして運用する多層ネットワークである．

これらは Fig.1 に示すように，VC や VLAN で構成する Layer2 をユーザの要求毎に設定する．そのため，Fig.2 に示すように，実際の物理網 Layer1 と，それぞれのユーザが運用する Layer2 ネットワークのトポロジが必ずしも一致しない．以下，ATM や Tagged Switch (Layer2 スイッチ[2])によって構成され Layer1 ネットワークを複数ネットワークとするネットワークのことを，多層レイヤ 2 ネットワークと呼ぶ．

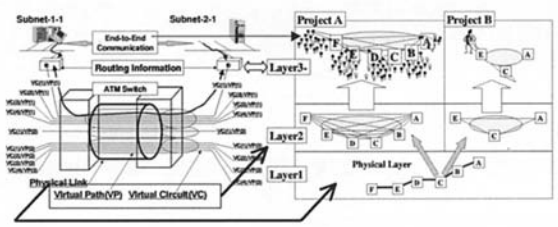


Fig.1 : Open Testbed Network for Layer2 based policy

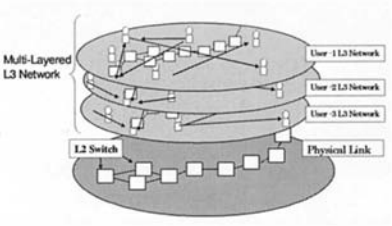


Fig. 2 Overview of Multi-Layered L2

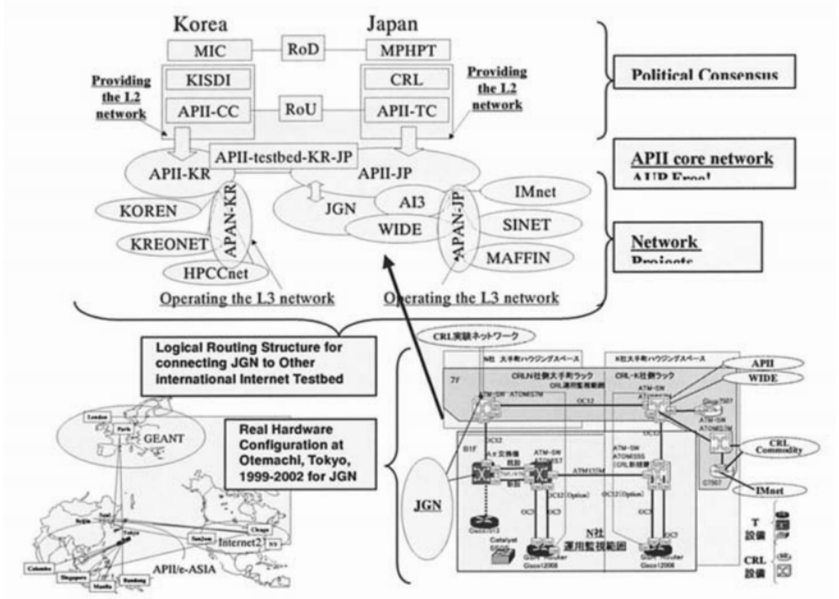


Fig. 3: Overview of Political Structure and Network Configuration of Internet Testbeds at Otemachi,

創成期のインターネットは，物理網とルータによって構成される論理網が単一であり，ユーザはルータに実装される smnp 等により，End to End の経路と途中経路の回線状態を traceroute[3]等の手法で知ることができた．多層レイヤ２ネットワークは，2 つのルータの間に多数のレイヤ２スイッチが介在するため，物理網と論理網が異なる．そのため従来の品質の推定が論理的に一致しないことも経験される．本研究では，この種のネットワークでの利便性，問題点に関して検討した結果とその対策について報告する．

2. 多層レイヤ２ネットワークの実例と運用

多層レイヤ２ネットワークの実例として，1999 年から 2001 年にかけて構築した JGN 創成期の各種テストベッド間接続の例を用いて説明し，特徴について説明する．

2.1. 相互接続点の多層レイヤ2ネットワークの実例

多層レイヤ２ネットワークによって構成されるネットワークでは，Fig.2 に示したように，物理的に広域であっても論理的には一つの LAN として運用することが可能である．その為，Fig.3 に示すように，各国ならびに各プロジェクトの Policy がそれぞれの Fund Owner の要求事項により異なる国際インターネットテストベッドの相互接続や，JGN と既存網の接続等では複雑なレイヤ２構造が必要になった[4].

本図は 1999 年から構築し 2002 年頃まで運用された APII(Asia Pacific Information Infrastructure) プロジェクト[5]の国際・国内接続の概念図である．政府間で交わされる RoD (Record of the Discussion)によって大枠の Policy のフレームワークが規定され，各国内のプロジェクトが接続先に対して覚書を研究者間で取り交わす MoU (Memory of the Understanding)により Layer3 接続を行って成立する．

これら実際の Policy の調整を行うのは研究者だが，物理網を提供する通信キャリアの管理する範囲を決めてネットワークの安定的な稼動を期待する．しかし，約款によって運営されている電話局舎内での作業であること，それまで公社側から民間企業への回線敷設はあっても民間キャリアからのファイバーの持ち込みが原則として許されなかったことなど，歴史的な困難を伴った．結果，Fig3 右下図のように研究用運用範囲とキャリアとしての運用範囲を区切り，それぞれが別々の VC (virtual citcuit)を運用すること，運用用の回線と実験用の回線を分離して 300m離れた２つの局舎を接続すること（右下図上側の OC12; 620Mbps ラインは研究，下の 135Mbps ラインはキャリア側）で責任の切り分けを明確化した．

これら回線では，運用と運営の部分と実験の部分を分離することは物理網では不可能だが，Layer2 の多層化によって WIDE，APAN 等それぞれのネットワークの AUP のマッチしたプロジェクトが申し合わせて Layer2 接続を行い，Layer3 での相互接続を実現した．Link Owner は，Layer2 の相互接続を Carry する責任と帯域の管理を負うが，Layer3 の Policy には関与しない．このような相互接続するペアに対して Layer3 の AUP(Acceptable Use Polocy)を Free（制限なし）という運用（Layer2 AUP Free Policy）によって，実質的に Policy を切り分けて運営することが可能となった．このようなネットワークの多層化によって，AUP の合意できたテストベッド間での相互接続（IP レイヤ）が可能とな

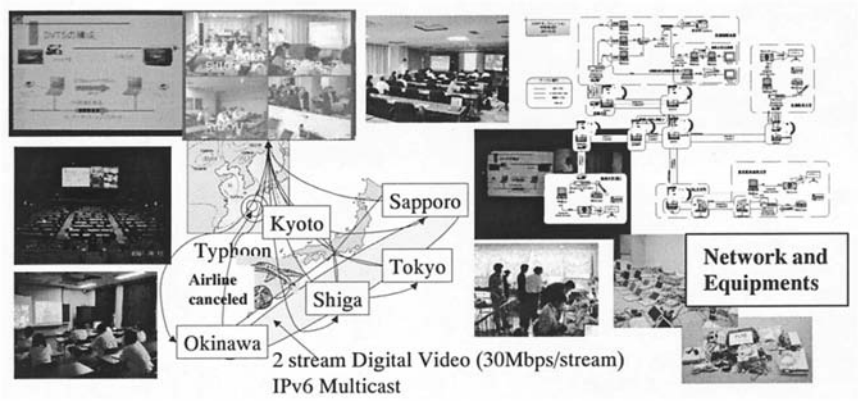


Fig. 4. DDW-Japan Experiment Overview at 2001

り，各プロジェクト担当者は年度報告をリンクオーナに対して行うという，成果指向の運営が可能となった．

2.2. AUP-Free 多層レイヤ2ネットワークの実際

このような Layer2 AUP-Free の運営形態は，回線を共有する場合に実際的かつ効率的である．しかし，一方でリンクオーナでさえ実際にネットワーク上で何が行われているかを把握することができなくなるなどの弊害もある．別のテストベッドでは帯域割り当てのポリシーを帯域割り当て委員会(Board of Resource Allocation)に集中して運用しているものもあったが，実験者が事前にこの委員会に一つ一つの実験に関して申請を行い，検討されてからしか行えない等の非効率が生じることも問題であった[6].

これらのことから JGN や APII テストベッドでは，帯域割り当ての原則を UBR(Unspecified Bit Rate)とするなど工夫が行われた．ユーザに対してレイヤ 2 を提供し，最大帯域を提示してベストエフォット型サービスと称するわが国独特のサービス形態はこの頃から商用サービスとして運用が開始され現在に至っている．

3. 多層レイヤ２ネットワークの問題点

ここでは，多層レイヤ 2 ネットワークでの実験を行った結果から技術課題について説明する．

3.1. 多層レイヤ2ネットワーク実験の実例

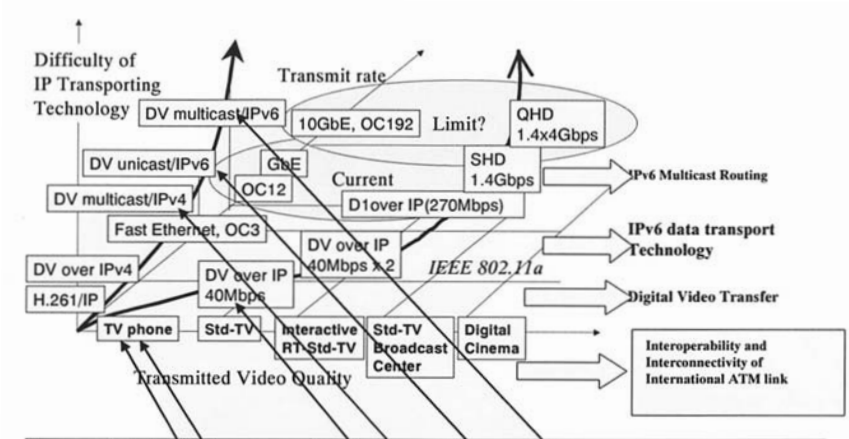
多層レイヤ２ネットワークである JGN を使って多地点双方向会議を行った実験の概要を Fig 4 に示す[7]. この実験は，デジタルビデオ（D1 に対して約５分の１圧縮のフレーム間圧縮のない伝送方式，帯域約 30Mbps を消費する）を５箇所から集めて合成し，IPv6 Multicast で 5 箇所にリアルタイムで伝送するものであった．図の左側が会議の模様，右上がネットワークトポロジ，写真は事前準備の模様である．それぞれの会場への回線は地方のキャリアのサービスクラスを CBR（この場合伝送帯域が最大 130Mbps を越える事はないので OC3）で用意し，JGN を UBR で使用する．各会場からの打ち上げのデータは大手町に集め，４会場分を１画像に合成し，マルチキャストルータを用いてパケットコピーする．

それぞれの会場のオペレータはこれら障害を常に取り除く作業を行う必要があった．末端のネットワークから Ping などによって得られる情報は，Layer3 ネットワークでのパケットロスと RTT の延長であり，設置した L3 のノードへの到達性の情報である．本例のように，UBR で提供されたネットワークを用いて，実験を実際に行う場合，画像を伝送するためには回線の設定だけではなく様々な問題が生じた．要点を以下に列挙する．

ATM 多層レイヤ２ネットワーク実験の障害原因

1. 足回り回線の調達困難
2. 回線の物理的切断
3. Layer2 ネットワークの設定ミス
4. Layer2 ネットワーク伝送装置の故障
5. Layer3 ルータの故障
6. ルータの能力不足
7. ルーティングに関する人的ミス
8. ATM セル損失
9. パケットロス
10. Shaping parameter 調整でのパケットロス
11. 各会場の画像系障害
12. 音声回り込みによるハウリング
13. 端末 PC の故障
14. 電源遮断
15. その他の人的ミス

実際には多数の Layer2 ネットワークスイッチが介在していることから，どの Layer2 のセグメントが障害になっても Layer3 にとってはパケットロスの増加としか感知できない．本実験のように 1 画像 30Mbps 超の動画像の送受には数万パケット/秒以上の伝送が行われ



Year	Month	Title	Place at Japan	Sites	Category	Bandwidth
1998	Jun.	APEC-TELMIN3	Tokyo	Singapore	P-P	8M
1998	Nov.	IST'98	Tokyo	Madrid	P-G(Unicast)	2M
1999	Sep.	JP-CA-DVTS/V4	Tokyo	Otawa	P-P	40M
1999	Dec.	IST'99	Tokyo	Wiena	P-G(Unicast)	2M
2000	July	I-Grid, DV-land	Yokohama, Kyushu, Okinawa, Hokkaido,	CERN	IPv4 Multicast	20Mbps(In Japan, 135Mbps)
2001	Jan.	JP-Eu DV/IPv6	Tokyo, Beppu	London, Madriid	IPv6 Multicast	40Mbps
2001	Apr.	AIC25, IPv4/6/4	Tokyo	Shanghai	IPv4/V6 tunnel	2Mbps
2001	Jul.	CRL-UCL DV/6	Kobe	London	IPv6 P-P	20Mbps(In Japan, 135Mbps)
2001	Nov.	JGN-v6-AI3	Okinawa	Thai	IPv4/Satellite	512Kbps

Fig.5: History Map of Experiment for Broadband Multimedia Transport Using JGN and APII Testbed Until 2001. : 3D Mapping with Quality, Transmit Rate and Difficulties for implementation.

るため，たとえ 0.5%のパケットロスであったとしても，特別なバッファを使わずに UDP で伝送する DVTS 等のシステムでは，即刻フレームロスや音声の途絶を招く．実験を成功させるためには，時々刻々生じるこれら障害について，それぞれのオペレータが情報を共有し対処する必要があった．これら問題を try and error で一つ一つ解決して行った実験と目的とした技術をの連関関係を Fig.5 に示す[8][9][10].

3.2. Multi-modality ネットワークの課題

本例のように MTU 1500bytes 以上の IP パケットを ATM で伝送する場合，IP の手順によるフラグメンテーションに加えて ATM セルに分割されることによる伝送効率の低下と，ネットワーク輻輳などによる ATM セルの損失は深刻な結果を招く．また，前述のようにそれぞれの物理線をルータによって接続され，物理網の障害を traceroute などで用いて知ることができたインターネットに比べ，それぞれのルータ間にユーザからは直接見えない多数のネットワークスイッチが介在するため，単純な Layer3 の障害検知方法ではネットワーク障害を知る事ができないことも問題である．Fig.6 は町沢らの行った[11]，JGN 東京大手町を基点として ATM 折り返しによって計測した伝播遅延を地理的距離の縮尺で表した図である．

計測の結果，一部のセグメントで地理的距離との乖離が見られた．その後，通信キャリアの協力により，０種網の情報開示が行われた．ネットワークの物理網の情報は通信キャリアにとってセキュリティの問題も大きかったため，ほとんど開示されることはないが協力が得られた．提供された回線は，以下の物理網を経

Fig. 6: Correlation between the result of Network ATM roundtrip time Measurement of JGN(one) backbone and Geographical Map.

由していることがわかった．回線キャリアＡは，東京－北陸間の直接の物理網を保持していなかった．そのため，キャリアＡは，大手町を基点として長野で別のキャリアＢの回線に乗り換え，キャリアＢ側の事情で千葉県内まで一度折り返した後，金沢のキャリアＢの局舎で再収容しキャリアＡの回線として到達した．そのため，東京－長野間を２回通ってから金沢に到達することになり，地理的距離感よりも遅延が大きくなった．このように、提供されている物理網の回線品質をユーザが計測できることは難しい．しかし，放送型のストリームデータ伝送などでは，あらかじめ伝送路の遅延特性を把握して Jitter を設定する必要があり下位レイヤの情報を取得することも重要と考えられた．

4. ユーザによるレイヤ２伝送品質推定の例

これらのことから，レイヤ２伝送品質推定をもく t 暦としてRDTSCを用いたLayer3でのパケット送受によるLayer2リンク稼動状態の推定方法を示す．実験の概念をFig.7に示す。

4.1. RDTSCを用いたレイヤ2稼働率推定法

Fig.7 のような条件では，一定のトラヒックを発生する TG から１つめのスイッチに対して一定間隔で等長のパケットが到着するが，スイッチ側の CSMA/CD 機構によって一様にブロックされる事により，スイッチにとってパケットはランダムに到着する．同様に mSN を発したパケットもスイッチに対してランダムに到着する．TG に比べて mSN から出すパケットは十分に頻度が小さく，十分に短いため，大数効果により TG-RNg 間のパケットのストア-フォワード枠に一様に混入し，２つのスイッチ間のボトルネックに対して TG-RNg の起こすブロッキングエラーとほぼ同じ確率でブロッキングエラーに遭遇する事が期待される．ブロッキングによる負荷遅延は表１より伝送速度が 100Mbps の Fast Ethernet の場合，64 バイトの１パケットあたりのスイッチ内滞留時間は 1/148800 秒=6.72×10^{-6} 秒である．CPU の制御周波数分の１秒での精度(3 GHz の場合，１RDTSC カウントは 0.3×10-9 秒)である．最近の PC であれば通常の精度として得ることが可能である．

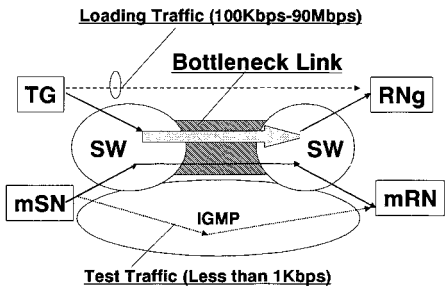


Fig. 7　Schema of Measurement for　Estimation of Layer2 Bottleneck capacity

表１イーサネットスイッチ交換性能の例

1000M⇔1000M	1,488,000pps
100M⇔100M	148,800pps
10M⇔10M	14,880pps

pps: packet per second

一定時間間隔で到着する計測用パケットが，スイッチに到着した時，無負荷時には一定時間 Ｔｒでフォワードされる筈が，ボトルネックリンクを定常的に埋める負荷トラヒックが存在している場合，確率ｐで付加遅延 Td だけ待たされたのちにフォワードされる．従って，ｎ番目に到着したパケットの伝送時間 Tn は次式で与えられる．

$$Tn = \begin{cases} Tr & (P(n) < (1-p)) \\ Tr + Td & (P(n) >= p) \end{cases} \quad (式1)$$

ここで，ｐ：ボトルネック帯域に対する負荷トラヒックの占有率である．本研究では，一定伝送時間 Tr で到着する計測用パケットが Td 遅れて到着する事象確率 P(n)を実測し，未知確率ｐを求める．これによって，ボトルネック稼働率を直接推定する．

本系における伝送時間 Tr_n は

$$Tr_n = RDr_n - RDs_n \quad \cdots(式2)$$

n は送出シーケンス番号，RDr，RDs は，送信時，受信時のそれぞれの RDTSC 値である．しかし，同種類のものでも搭載される水晶のばらつきによって個体差があることが知られており，単純に（式 2）から算出する事はできない．送受信ノード間での RDTSC の補正が必要である．以下，提案する補正方法ならびにボトルネック稼動状態の推定方式に関して述べる．

ボトルネックリンクに負荷がかかっておらず，送信側から送信されたパケットが受信ノードに付加遅延なく到着すると仮定する．時刻 t に送受信ノード間でシーケンス番号 n，n+1 の送受信があったとすると，送信間隔と受信間隔の比 Dr(t)は次式で与えられる．

$$Dr(n) = \frac{RDsn+1-RDsn}{RDrn+1-RDrn} \quad \cdots(式3)$$

これを計測し平均値を求め，Dr_{const} 値を求める．

$$Dr_{const} = average(N_0,\ Dr(n)) \quad \cdots(式4)$$

伝送時間 Tr_n は次式で与えられる．

$$Tr_n = RDr_n - Dr_{const} \cdot RDs_n \quad \cdots(式5)$$

また，同時に無負荷時の伝送時間 Tr_{n1} を次式で求める．

$$Tr_{nl} = average(N_0,\ Tr(n)) \quad \cdots(式\ 6)$$

ここで，ボトルネックを定常的に占有するトラヒックを与え，Tr_n を N1 個取得しボトルネックリンクの稼働率 P_b を算出する．

$$P_b = \frac{Freq\ (N1 \mid Tr_n > Tr_{nl})}{N1} \quad \cdots(式\ 7)$$

以上により，両端非同期の PC によるレイヤ 2 ネットワーク伝送路の稼動状態を推定する．以下に実験結果を示す．

4.2. 実験結果と考察

実験は Fig.7 に示した 2 台ストアアンドフォワード型の Layer2 スイッチを 100Mbps イーサネットで接続し，両端に Linux の稼動するクロック周波数約 3 GHz の PC を用いて行った．クロストラヒックを発生する TG から iperf を用いて UDP パケットを帯域幅 100Mbps に対して 10，50Mbps と増加させ，mSN から 1 秒間隔で 64 バイトのパケットを mRN に送信し，mRN での計測から伝送時間推定値 Trn を求めた．

1Mbps, 10Mbps, 50Mbps, 90Mbps をボトルネックに対して負荷したときの伝送時間 Tn (Observed Transport Time)の経時変化を Fig.8 に示す．実測された Tn は，mSN, mRN 各機材の特性によって，制御周波数が異なり，式 7 の補正に加えて初期値によるドリフトの補正も必要であった．以上より 100Mbps のボトルネックリンクに 1, 10, 50, 90%の負荷を与えた時，Tn が延長するし，延長頻度と比例することが示された(Fig.9)．比較のため，Ping を用いて同様の実験を行い，無負荷時(0.10ms)に比べて往復で 0.05ms (片道 25μ秒)の延長を認めたものの頻度を Conventional Method として Fig.9 に示した．回帰係数 R2 値は Conventional 0.909 に比べ，Proposed では 0.995 と改善した．本法は，計測精度など検討の余地を残すが，両端の PC での同期を必要とする他の方式に比べ，相対値のみでボトルネックの稼動率を推定でき有用と思われる．

また，今回は SOCKET Lib へのデータグラム書き込み直前に TSC を参照しネットワークインターフェイスをはさんだ伝送路の付加遅延の推定を試みたが，さらに，他のアプリケーションでは，バッファ書き込みの直前に TSC を読み出し，Application 間での伝送ゆらぎや伝送遅延も記録する事も可能と思われ，今後の検討課題とする．

5. まとめ：品質指向インターネットへの期待

「高品質インターネット研究会」発足当時の 1998 年から一般化してきた，多層レイヤ 2 ネットワークについて，ポリシー構造，ネットワーク構成の特徴，アプリケーション実験，ネットワークトポロジや障害の問題点について検討した．これらから End to End でのレイヤ 2 ネットワークのボトルネック稼動状態の推定方式を例示し，ユーザレベルでのレイヤ 2 ネットワーク伝送品質計測可能性を検討し有用性を示した．

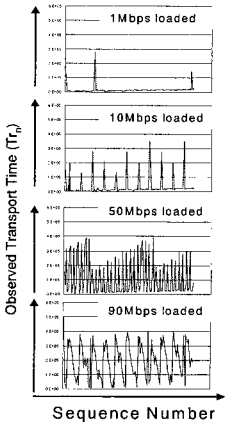


Fig.8 Trends of Observed Tr_n value with each loaded traffics at

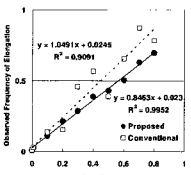


Fig.9 Comparison of Observed Frequency of Tn Elongation between conventional method and

謝辞

本研究を行うにあたり，御協力を頂いた情報通信研究機構 久保田文人センター長、町沢朗彦主任研究員，ならびに元同研究機構理事 塩見正博士，総務省 寺崎明氏，松井房樹氏，鈴木薫氏，雨宮明氏，島田淳一氏，他関係各位に深謝する．本研究は情報通信研究機構運営費交付金〈新世代ネットワーク研究センター〉,平成 17,18,19 年度厚生労働省がん研究助成金研究総合研究「がん情報ネットワークを利用した総合的がん対策支援の具体的方法に関する研究 」若尾班等の支援を得て行った．関係各位に謝意を表する。

参考文献

[1] T. Saito H. Esaki, Gigabit Network, IOS Press, ISSN 1348-513X, 2003
[2] Kennedy Clark, Cisco Lan Switching: Layer 2 Technologies (Ccie Professional Development),Cisco Systems ; ISBN: 1587052164 ,2005
[3] Information Network Div. HP Company, Netperf: A Network Performance Benchmark Rev.2, http://www.netperf.org, 1995
[4] S. Nakagawa, Harmonization of Acceptable Users Policies of the Networks IPSJ Symposium Series, Vol 99-7, pp 357-361, 1999
[5]S. Nakagawa and et.al.,Proposal of Next Generation Peta-bits Network Testbed, Proceedings of The International Symposium on Towards Peta-Bit Ultra-Networks , pp 25-33, 2003
[6] S. Nakagawa, A. Machizawa,et.al., Proposal of Policy Model for the Next Generation Internet Testbed, Journal of CRL, 48-2, pp 23-33, 2001
[7] 櫻田武嗣他，ATM NIC によるマルチキャストルーティングの限界, IEICE IN 研究会, 2002,(101-639),pp9-15
[8] S.Nakagawa, QoS Evaluation Method for Stream Data Transport with ICMP - An Experiment Networking and Telemedicine Demonstration at APEC TELMIN3 -, Proceedings of IEEE ICOIN-13,11C-1.1-1.6 1999
[9] 杉浦一徳他インターネットにおける TCP 協調型の DV 転送技術、信学技報、CS99-143, CQ99-66, pp19-24, 2000
[10] M. Katsumoto,et.al, Design of the VoD System for High-quality Video and audio with D1 over IP, Proceedings of IEEE International Conference on Networking (ICOIN-15), pp714-719, 2001
[11] 町澤朗彦 他, ATM による DV 伝送システムの遅延と品質評価について, DICOMO 論文集、2000
[12] 中川晋一 他，ネットワークにやさしい低頻度非同期ノード間通信によるレイヤ 2 ネットワークボトルネック稼動状態推定法,DEWS2006 Proceedings, 2b-o2, 200