ATMスイッチ内の輻輳による TCPデータ転送性能低下の解析

中西 基起* Ishtiaq Ahmed* 石橋 勇人† 岡部 寿男* 金澤 正憲‡
*京都大学大学院情報学研究科
†大阪市立大学学術情報総合センター ‡京都大学大型計算機センター

概要

ATM（非同期転送モード：Asynchronous Transfer Mode）は、LANおよび基幹ネットワークで利用されている高速かつ広帯域な通信方式である。文字、音声、画像および映像などのマルチメディア通信が可能であると同時に、各通信に必要なサービス品質（Quality of Service）を保証させるための手段を備えている。一方、現在のLAN環境において主要となる通信プロトコルはTCP/IPである。よってATM上でTCP/IPを実装する必要があるが、両者が前提とするデータリンク層の性質が異なるなどの理由によりその実現には多くの問題があることが指摘されている。特にATMスイッチ内に高い負荷が存在する環境では、セルロスに起因するIPパケットの再送などによりTCPデータの転送性能が大幅に低下することが報告されている。本論文では、そのような高負荷時におけるTCPデータの転送性能を測定し、ATMレベルでの転送との関連を解析した。実験としてATMスイッチ内に高負荷な環境を設定し、TCPデータの転送性能の変化や極端な低下（deadlock）を観測し、ATM上でのTCP/IP実装における問題点を明らかにしている。また、ATMスイッチのセルバッファサイズや負荷となるトラフィックの帯域などがTCPデータ転送性能に及ぼす影響について解析した。

TCP throughput anomaly on a congested ATM link

Motoki Nakanishi*, Ishtiaq Ahmed*, Hayato Ishibashi†, Yasuo Okabe*and Masanori Kanazawa‡
*Graduate School of Informatics, Kyoto University
†Media Center, Osaka City University ‡Data Processing Center, Kyoto University

Abstract

ATM (Asynchronous Transfer Mode) is one of the key technologies for high-performance networks due to its QoS (Quality of Service). Many campus networks in Japan are using ATM as their backbone network. TCP/IP is most widely used protocol in the Internet. Using TCP on IP over ATM has several implementation problems and suffers throughput deadlock due to congestion. In this paper, we have analyzed the traffic dynamics of TCP on IP over congested ATM link. We investigated the reasons for TCP's throughput deadlock and suggested solution to improve the throughput. Then we analyzed the influence of some other parameters causing deadlock situations.

1 はじめに

高速かつ信頼性の高いネットワーク技術のひとつにATM（Asynchronous Transfer Mode）がある。ATMでは４つのサービスカテゴリが定義されており、そのため利用者の要求するサービス品質（QoS：Quality of Servive）に柔軟に対応でき、かつ有効に帯域を利用できる。

一方、InternetやLAN（Local Area Network）で主に利用されている通信プロトコルはTCP/IP

である。本来、TCP/IP は Ethernet などの比較的低速なネットワーク技術に対応して設計されており、かつ TCP/IP と ATM では前提とするデータリンク層の性質が異なっているため、ATM ネットワーク上で TCP/IP を利用する場合、様々な問題が発生することが指摘されている。例えば IP over ATM のデフォルト MTU サイズが Ethernet より数倍大きいことが TCP トラフィックのスループットを悪化させる場合がある [1]。また、ATM ではデータを 48[bytes] のペイロードに格納して転送するため、ひとつの IP パケットが多数のセルに分割されることになる。そのため、ATM レベルでのひとつのセル損失（セルロス）が上位のプロトコルレベルでは大幅なスループットの低下を招く原因となる [2]。

これまでの研究では、優先度が等しい複数の TCP トラフィックが輻輳した場合の振る舞いが扱われてきた。しかし、ATM が想定するマルチメディア環境では、動画像伝送などのストリームデータが広い帯域を占有する状況で、通常の TCP コネクションがどのような影響を受けるかも重要である。そこで本研究では、ATM アナライザが発生するトラフィックデータを優先度の高い CBR に割り当て、優先度の低い UBR 上の TCP トラフィックがどのような影響を受けるかの性能評価を実測により行った。その結果、ATM スイッチ内で輻輳が発生している状況では、TCP のデータ転送性能が急激に低下する場合（deadlock）があることを確認した。また送受信端末の Socket バッファサイズ、ATM スイッチ内のセルバッファサイズ、MTU サイズ、CBR トラフィックの帯域に注目し、これらの要因と deadlock との関連性を解析した。

以下の章は次のように構成される。第 2 章では実験の環境と方法を述べる。第 3 章では実験により deadlock が発生することを示し、Tcpdump を使用してその原因を解析する。第 4 章では deadlock になる条件を実験により導く。第 5 章では結論を述べる。

表 1: 実験で使用した機器

CPU	Pentium II 266MHz
端末 OS	FreeBSD 3.2-RELEASE
ATM NIC	Adaptec ANA-5940
ATM ドライバ	FreeBSD 3.2-RELEASE 付属
ATM スイッチ	富士通 EA1550
ATM アナライザ	Hewlett-Packard HP E5200A

2 実験環境と方法

2.1 実験環境

実験で使用した機器を表 1に示す。本実験で使用した ATM スイッチは出力バッファ型であり、各ポート毎、各サービスカテゴリ毎にセルバッファを割り当てられる。ATM アナライザは、内蔵したトラフィックジェネレータにより様々なタイプのトラフィックを発生させる機能を持つ。

これらの機器により構成した実験環境を図 1に示す。2 台の端末を ATM スイッチに接続し、固定型バーチャルコネクション（PVC）を張っている。端末と異なるポートに ATM アナライザを接続し、同様に PVC を張る。これら 2 つの PVC の出力ポートを同一ポートに設定することで 2 つのトラフィックを衝突させ、輻輳を発生させる。

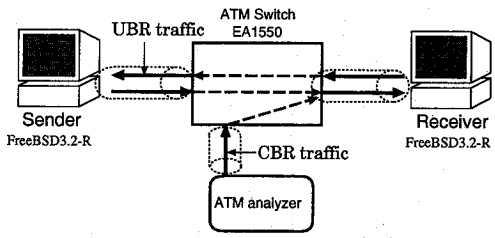

図 1: 実験環境

2.2 実験方法

実験には Netperf[3] を用いた。Netperf は TCP などのトラフィックを発生させ、そのときネットワークを流れたデータ量を測定することでスループットを計測する機能を持つ。

実験方法は以下の通りである。まず、ATM ア

ナライザのトラフィックジェネレータからCBRトラフィック*を発生させ、常に一定の帯域を占有している状況を設定する。次に、NetperfからTCPのUBRトラフィックを10秒間発生させ、そのスループットを測定する。このときCBRトラフィックとUBRトラフィックは同じ出力側ポートで合流することになる。よって、これら2つのトラフィックの合計帯域がATMの有効帯域を超過すれば、ATMスイッチ内は輻輳状態となりセルロスが発生する。このときUBRトラフィック上のTCP/IPがどのような影響を受けるかを観察する。

なお、実験結果におけるUBRトラフィックのスループットはNetperfによる測定を5回行った平均値を挙げている。また、Netperfでは測定時のパラメータとして送信端末のメッセージサイズを64K[bytes]に固定して測定を行った。これらの実験によって、セルバッファサイズ、Socketバッファサイズ†、CBRトラフィックの帯域、MTUサイズがUBRトラフィックのスループットに及ぼす影響を解析する。

3 TCPトラフィックのdeadlock

3.1 セルバッファサイズとスループット

まずCBRトラフィックの帯域を60M[bps]に設定し、UBRトラフィックのスループットを測定した。測定ではMTUサイズを512, 1500, 4352, 6500, 9180 [bytes]と変化させた。セルバッファサイズは1K[cells]であり、Socketバッファサイズは64K[bytes]である。測定結果を図2に示す。

セルバッファサイズが150K[bytes] (3K[cells])以上ではUBRトラフィックはMTUサイズに関わらずほぼ有効帯域に相当するスループットを得ている。一方、セルバッファサイズが150K[bytes] (3K[cells])以下の領域では急激にスループットが低下し、50K[bytes] (1K[cells])以下になると

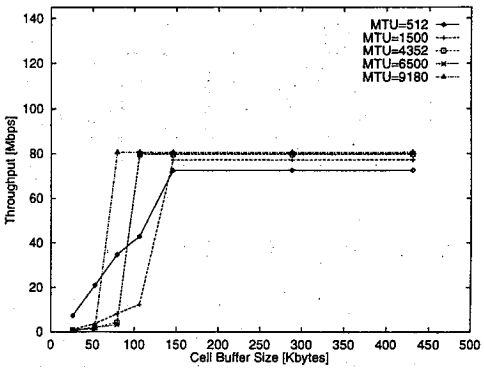

図2: セルバッファサイズとスループット

MTUサイズが512[bytes]の場合を除いてUBRトラフィックは有効帯域の2%以下のスループットしか得られていない。このようなほとんどスループットを得られていない状態をdeadlockと呼ぶことにする。

Floydらは、複数のTCPコネクションによるUBRトラフィックが輻輳する状況におけるセルバッファサイズとスループットの関係を、シミュレーションにより求めている[2]。Floydらの結果と比較すると、

- セルバッファサイズが50K[bytes] (1K[cells])未満の場合、複数のTCPコネクションによる輻輳では性能低下は10〜30%程度にとどまっているのに対し、TCPのトラフィックがCBRトラフィックにより妨害される状況では、ほとんどスループットが得られないdeadlockが発生する。

- 複数のTCPコネクションによる輻輳ではMTUサイズが大きいほど性能低下が大きいのに対し、TCPのトラフィックがCBRトラフィックにより妨害される状況では、MTUサイズが9180[bytes]よりも1500[bytes]の場合に性能低下が顕著である。

という際立った違いがあることがわかる。

*本論文では、CBRサービスでのトラフィックをCBRトラフィックと呼び、UBRサービスでのトラフィックをUBRトラフィックと呼んでいる。

†ここでいうセルバッファサイズはATMスイッチ内のバッファサイズを意味し、Socketバッファサイズは Netperfで指定する端末のSocketバッファサイズを意味する。

表 2: 各測定のパラメータと測定結果

CBR[Mbps]	60	
MTU[bytes]	1500	
Socket size[Kbytes]	25	38
Throughput	Good	Deadlock
Pattern	(a)	(b)

3.2 deadlock 発生のメカニズム

同じ MTU サイズにおいて、deadlock に陥るときと陥らないときでの TCP セグメントの流れ方の違いを分析した。Netperf により UBR トラフィックを発生させている状況下で、Tcpdump[4] を使用して端末間のセグメントの流れを観察し、その振る舞いを解析した。測定した 2 つのパターンを表 2に示す。

これら 2 パターンにおいて、送信端末からのセグメント送出時間とシーケンス番号との関係を示したのが図 3 である。パターン (a) のグラフと比較すると、パターン (b) のグラフは最終的に到達したシーケンス番号が小さい。また、「一定時間シーケンス番号が進まない状態（タイムラグ）」が度々発生しており、スループット低下を招いている。このような状態に陥る原因を明らかにするために、パターン (b) の Tcpdump データファイルを詳細に分析した。その結果から次のことがいえる。

- TCP において再送が起こる条件は、送信端末で重複確認応答（duplicate ACK）が 3 回受信されるか、タイムアウトが発生するかである [5]。問題となるタイムラグでは後者が発生している。
- タイムアウトを待たなければ再送できない状態になるタイミングでは、送信端末からのセグメントと受信端末からの ACK がネットワーク上から全て消滅し、送受信端末ともに何も送信しない状態になっている。

deadlock が発生するパターン (b) と発生しないパターン (a) の違いとしてひとつの注目すべき点は「最後に送信したセグメントのシーケンス番号と、その直後に受信する ACK のシーケンス番号との差（シーケンス番号の隔差）」である。パターン (a) と (b) を比較すると、前者は 1 回 ACK を受信する間に 2 回セグメントを送信しているが、後者は 1 回 ACK を受信する間に 3 回セグメントを送信するときがある。両者の ACK は常に 2 つのセグメントに対して確認応答をしている。よって 2 回セグメントを送信する場合は「シーケンス番号の隔差」は変化しないが、3 回セグメントを送信する場合に隔差（＝送信したが未確認のデータ量）が拡大する。パターン (a) ではこの「隔差」は一定値以上にならないが、パターン (b) では徐々に増加し、ある値を超過したときセグメント・ロスが発生する。

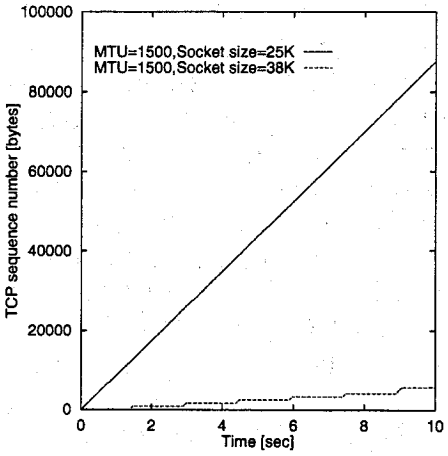

図 3: MTU=1500[bytes] のときのセグメント送出時間とシーケンス番号

また「送信したが未確認のデータ量」が送信端末の輻輳ウインドウに占める割合に注目すると、UBR トラフィックのスループットが良好な状態において、パターン (a) では「送信したが未確認のデータ量」は輻輳ウインドウの右端に到達している（図 4)。すなわちパターン (a) は ACK の受信によってウインドウがスライドしないと新たなセグメントが送信できない状態であるため、常に 2 つしかセグメントを送信しない。しかしパターン (b) では、まだウインドウに空きがあるためその後も「送信したが未確認のデータ量」が占める割合が増加し、ある値を超過するとセグメント・ロスが発生する。

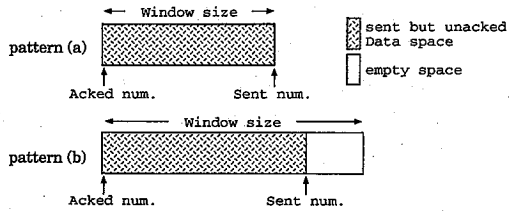

図 4: ウインドウ・サイズと送信したが未確認のデータ量

以上をまとめると、Socket バッファサイズが十分小さいときは輻輳ウインドウを全て使用してセグメントを送信しても輻輳が発生しないが、Socket バッファサイズが大きいと輻輳ウインドウを全て使用するまでに輻輳が発生し、セグメント・ロスが発生する。

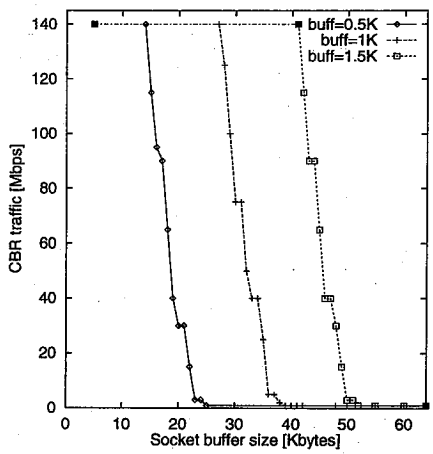

図 5: Socket バッファサイズと deadlock になる CBR トラフィック（セルバッファサイズ=0.5K, 1K, 1.5K[cells]）

4 deadlock 発生の条件

4.1 セルバッファサイズによる deadlock に陥る条件の変化

3.1節図 2で示したように、セルバッファサイズが十分大きいと deadlock は発生しない。そこで deadlock に陥るときのセルバッファサイズを調べるために、Socket バッファサイズを一定値に固定し、UBR トラフィックのスループットが deadlock に陥るまで CBR トラフィックの帯域を増加させて、deadlock を引き起こす最小の CBR トラフィックの帯域を求めた。MTU サイズを 1500[bytes] に、セルバッファサイズを 0.5K, 1K, 1.5K[cells] に設定した。Socket バッファサイズは 6K~64K[bytes] の範囲で変化させている。結果を図 5に示す。グラフは、UBR トラフィックのスループットが deadlock に陥ったときの CBR トラフィックの帯域を示している。但し、CBR=140M[bps] となっている部分（グラフの点線部分）では、CBR トラフィックの帯域が 140M[bps] でも deadlock が発生しないことを意味している。

グラフから、つぎのことがわかる。

- Socket バッファサイズが十分小さければ CBR トラフィックが大きくても deadlock は生じない。
- deadlock が生じ得る最小の Socket バッファサイズはセルバッファサイズに比例する。
- その値より Socket バッファサイズを増やすと、deadlock を生じさせるのに必要な CBR トラフィックの帯域は線形に減少し、その傾きはセルバッファサイズには依存しない。
- Socket バッファサイズがある値以上では、ごくわずかの CBR トラフィックでも deadlock を引き起こす。

4.2 MTU サイズによる deadlock に陥る条件の変化

セルバッファサイズを 1K[cells] として、MTU サイズが 1500, 9180[bytes] の場合の Socket バッファサイズと deadlock を引き起こす最小の CBR トラフィックの帯域との関係を求めた。Socket バッファサイズは 6K~64K[bytes] の範囲で変化させている。測定結果を図 6に示す。

グラフから、次のことがわかる。

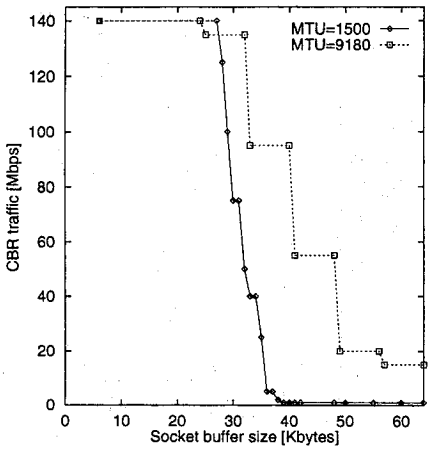

図 6: Socket バッファサイズと deadlock になる CBR トラフィック（MTU=1500, 9180[bytes]）

- deadlock が生じ得る最小の Socket バッファサイズは MTU サイズにほとんどよらない。
- その値より Socket バッファサイズを増やすと、deadlock を生じさせるのに必要な CBR トラフィックの帯域は線形に減少するが、その傾きは MTU サイズ（正確には MSS）にほぼ反比例している。

なお、MTU サイズが 9180[bytes] の場合にグラフは階段状になっているのは、Socket バッファサイズを細かく変化させても TCP が MSS に基づいてウインドウサイズを丸めるという TCP の仕様によるものと考えられる。

5 結論

本稿では、優先度の異なる 2 つのトラフィックを ATM スイッチで合流させることで輻輳を発生させ、その状況での TCP のデータ転送性能を観察し、優先度の低いトラフィックのスループットが deadlock に陥る原因を解析した。また、TCP トラフィックのスループットを決定する要因として、送受信端末の Socket バッファサイズ、スイッチ内のセルバッファサイズ、MTU サイズ、妨害トラフィックを挙げ、これらが UBR トラフィックのスループットに及ぼす影響について解析した。また、各要因と UBR トラフィックのスループットとの関連性について明らかにした。

今後はこれらの解析結果を元に、セルロスの発生する条件をモデリングしたい。

謝辞 日頃からご討論頂く京都大学大型計算機センターの諸氏に深謝します。

参考文献

[1] K. Moldeklev, P. Gunningberg: How a Large ATM MTU Causes Deadlocks in TCP Data Transfers, IEEE/ATM Transactions on Networking, Vol. 3, No. 4, Aug. 1995.

[2] A. Romanow, S. Floyd: Dynamics of TCP Traffic over ATM Networks, IEEE Journal on Selected Areas in Communications, Vol. 13, No. 4, May 1995.

[3] R.Jones: Netperf, A benchmark for measuring network performance, Hewlett-Packard Co., 1993.

[4] Tcpdump 3.4, Lawrence Berkeley National Laboratory, Network Research Group

[5] G.R.Wright, W.R.Stevens: TCP/IP Illustrated, Vol.1 "The Protocols", Addison-Wesley, 1994.