高信頼同報での再送機構の Ack Implosion の再評価

6T－8

山内長承† ・城下輝治‡ ・佐野哲央‡ ・塩川鎮雄‡

†日本 IBM 東京基礎研究所・東京都立大学工学部、‡NTT 情報通信研究所

1　高信頼同報バルク転送と RMTP

本稿では、IP マルチキャストを用いた多端末への高信頼バルクデータ同報転送プロトコル RMTP[1, 2, 3] における、Ack Implosion の程度とバックオフ時間の見積りを再評価する。

高信頼バルク同報は、インターネットにおける電子新聞の配送 ([1]) の他、非常に多数の端末への業務用情報の配信設定、非常に多数の WS へのソフトウェアの導入・更新などの応用において、転送時間の短縮、ネットワーク及び送信サーバーの負荷の軽減の効果が期待できる。

そのデータ転送信頼性を確保するためには、データの受信端にて到着を確認し、不足するものがあれば再送依頼を送信サーバーへ送る手順が必須となる。1 対多端末の通信では、受信端からの返答は端末数が多くなるにつれて量が大きくなり、送信サーバーに接続されるネットワークや送信サーバー自身の処理プログラムに対して大きな負荷となり、受信端末数の増加に応じてスケールすることが難しい。

この問題に対し、受信端末をグループ化してその中でお互いに再送する方式など、分散した再送機構の提案もあるが [5, 9]、送信サーバーがデータ配送をすべて管理監督することができなくなり、業務上必須のデータを配布するには適さない。それに対して、RMTP は Ack Implosion を軽減する工夫を施すことによって、実用上役に立つ数千程度の範囲での受信端末数を実現しようとするものである。

高信頼同報バルク転送の手順例として、RMTP での全体の転送手順を図 1 に示す。コネクション設定フェーズで、送信サーバーと個々の受信端末の間にコネクションを設定する。これは送信端からマルチキャストを用いてコネクション設定メッセージを送り、受信端がそれに応答する。この応答はすでに Ack Implosion の可能性があるが、Ack パケットが短いため程度は軽い。

次にデータ転送フェーズに入り、送信サーバーはマルチキャストで一連のデータを送信してしまう。この間に起こったパケット損失は受信側に記録され、送信終了後に Nack による再送要求を送り返す。すべてのデータを受信できた端末は Ack を返送する。Ack か Nack かを必ず返送することにより、送信端での受信状況の把握が確実なものとなる。この Ack/Nack は Implosion を起こす可能性がある。

すべてのデータを受信でき Ack を返した端末は、送信サーバーからのコネクション解放メッセージをトリガーにして、同報サービスから離れる。再送を必要とする受信端末は、1 回目のデータ転送と同様の手順で受信に失敗したデータを受信する。再送もマルチキャストを用い、コネクション解放せずに残っているすべての受信端末に対して（そのデータが必要であるなしにかかわらず）送られる。受信端末は必要なデータを受信し、もしすべてのデータが揃えばコネクション解放へ、まだ不足があれば再送要求の Nack を送信サーバーに返し、更に再送を受ける。この手順を、すべての端末がすべてのデータを受けとるまで、繰り返す。

2　Ack Implosion の軽減

1 回の Ack Implosion は、返答データ量と受信端末の数の積に比例する。この影響を軽減するために、

a) 再送要求 (Nack) のコーディングを工夫し、データ長を小さくする。単に再送を要求するパケットの番号を列挙するのでなく、StarBurst[6] ではビットマップ表現、RMTP[4] ではパケット損失のバースト性を考慮して区間表示を採用している。

b) 同時に Nack を送る端末数を減らす。たとえば、バックオフ機構により返答送出時刻をランダマイズする。各々の端末に異なるバックオフ時間を割当て、Nack 送出時にその時間を待ってから送信させる。これにより Nack の到着が分散され、サーバーの負荷集中が軽減する。

などのほか、なるべく Ack を送る頻度を少なくするよう、フロー制御は TCP のいわゆる Go-Back-N 方式ではなくレート制御を用いたり、送信途中の状態を問わないバルク転送の性質を利用して、Ack/Nack を返すのはすべての転送を行なった後に 1 回だけとするなどの工夫をしている。

最大バックオフ時間の設定については、バックオフ時間が長いと Nack を収集する時間が長くなり全体の転送性能が低下する。バックオフ時間が短いと、Nack の集中により Nack が失われる結果となり、その再送処理のために全体の転送性能が低下する。その観点から、Ack Implosion の適正な推定が必要になる。

3　Ack Implosion の推定

第 k 回再送における Ack/Nack の量は、再送パケット数の推定 [1, 3] から求められる。Ack を送る端末は $N_0 - N_k$ であり、Ack の IP/UDP ヘッダを含むデータ長 $28B^*(N_0 - N_k)$ が転送される。Nack については、再送を必要とするパケットの個数が全体で S_k 個、1 端末あたりの平均が S_k/N_k 個となり、Nack のデータ量としてはヘッダ共通部分が $28B^*(N_k)$、パケット指定部分が $2B^*S_k$ 個（バーストによる節約効果がないとして）となる。[1] の例にある 5000 端末、2000 パケット、ビット誤り率 10^{-6} の例では、第 1 回目の再送が $N_1 = 5000$, $S_1 = 100000$ であるから、Ack を送る端末はなく、Nack データの総量が 340KB、Nack の平均長が 68B となる。

Reevaluating Ack Implosion Effect of Retransmission in Reliable Multicast Transfer

Nagatsugu Yamanouchi†, Teruji Shiroshita‡, Tetsuo Sano‡, Shizuo Shiokawa‡

†IBM Research, Tokyo Research Laboratory & Dept. Elec. and Info. Eng., Tokyo Metropolitan University, ‡NTT Information and Communication Systems Laboratories

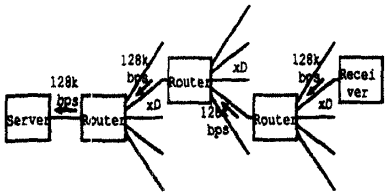

図 1: Multicast Tree.

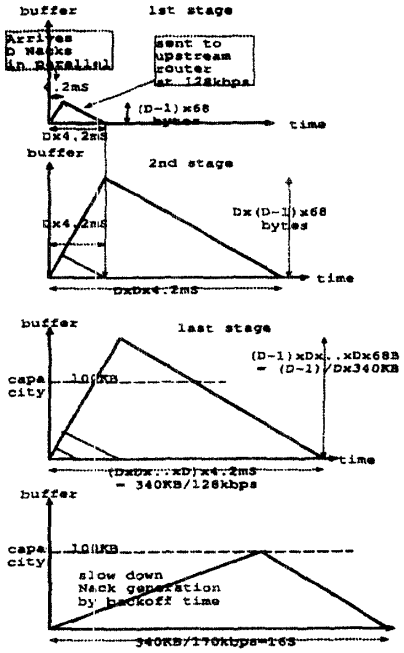

図 2: Flow balance of routers in each stage.

サーバーの処理能力は、TCP/IP の処理能力評価 [7] から推して、数十 Mbps 程度の回線からの入力ならば処理可能であると思われる[1]ので、ネットワークの飽和状況を解析する。接続のモデルとして一定分岐数 D のトリー（図 1）を考え、すべてのリンクが比較的低速な回線（たとえば 128Kbps）と仮定する。平衡トリーであればマルチキャストパケットの受信端への到着はほとんど同時になり、Ack 送出も同時となる。もっともリーフに近いルーターでは D本の入り線から同時に平均 68B の Nack が 128Kbps で到着し、1 本の出線から 128Kbps で出てゆく。ルーター内の滞留量は 4.2mS の期間 $(D-1)$*128Kbps で上昇し $(D-1)$*68B に達し、その後さらに $(D-1)$*4.2mS かかってすべて上位のルーターへ送出される。次のレベルのルーターでは D*4.2mS の期間で同時に D^2個の Nack が 128Kbps で到着し、滞留量が $D*(D-1)$*68B となり、その後 $D*(D-1)$*4.2mS かかってすべて上位ルーターへ送出される（図 2）。最上流のルーターでは 340KB の Nack が集まり、最大滞留量は $(D-1)/D$*340KB となる。ルーターのバッファ容量がこれより小さい場合はバッファ溢れにより Nack が失われるため Nack の再送を行なわなければならない。

バッファが小さい場合、バックオフにより Nack の送出をずらして集中を避けることが考えられる。最上流ルータへの入り側フロー総計を I、出側フローを O (=128Kbps) とすると、バッファ内の最大滞留量は $(I-O)$*(最大になる時間) =$(I-O)$*(全データ量/I) となるので、これを仮に 100KB とすると、I=170Kbps となる。つまり、全 Nack が 340KB/170Kbps=16 秒に分散して到着すれば、最上流ルーターは 100KB のバッファでも溢れない。また、回線速度が下流から上流へ向かって大きくなれば滞留しにくくなり、極端な場合各ルーターの入り側の回線速度の D 倍の回線を出側に用意すれば、このルーターでは滞留しないことになるが、最上流では結局 128Kbps*端末個数分（この例では 5000 端末なので 640Mbps）の回線容量が必要になる。

[7] の条件に比べて否定的な要素は、RMTP の処理が TCP ほど効率が良くない可能性がある、RMTP の現在の実装がカーネル外であるのでモード切替やタスクスケジューリングのオーバーヘッドがある、などいろいろあるが、いずれも決定的に性能を劣化させる要因とは考えられず、回避の方法が考え得る。

4 まとめ

マルチキャストを用いた高信頼バルクデータ転送のもつ Ack Implosion 問題について、RMTP の転送方式を軸にしてそのデータ量の見積りや適正なバックオフ時間について検討した。静的な見積りは可能なことがわかったが、動的な環境変化、例えば他のトラフィック、雑音等の動的変化に対する最適化には程遠い。そのため現在は、RMTP は他のトラフィックの影響の少ないイントラネット環境を中心に試行しているが、より広範な環境で利用できるためには動的な設定を検討することが必要である。

参考文献

[1] 城下、高橋、佐野、山下、山内、串田: 高信頼マルチキャスト通信プロトコル (RMTP) の各種ネットワークへの適用性. 信学技法 SSE95-196/IN95-140. 1996 年 3 月

[2] 城下、高橋、佐野、山下、中村、山内、串田: インターネットに適用可能な高信頼一斉分配システム. 情処 AVM 研究会, AVM 11-2. 1995 年 12 月

[3] Shiroshita, T., Sano, T., Takahashi, O., Yamashita, M., Yamanouchi, N., and Kushida, T.: Performance evaluation of reliable multicast transport protocol for large-scale delivery. Proc. IFIP PfHSN. October 1996.

[4] Shiroshita, T., Sano, T., Takahashi, O., and Yamanouchi, N.: Reliable Multicast Transport Protocl. IETF Internet Draft <draft-shiroshita-rmtp-spec-00.txt>. March 1997.

[5] Lin, J. C. and Paul, S.: RMTP* A Reliable Multicast Transport Protocol. Proc. IEEE INFOCOM 96, pp. 1414-1424. April 1996.

[6] Miller, K., Robertson, K., Twwedly, A., and White, M.: StarBurst Multicast File Transfer Protocol (MFTP) Specification. IETF Internet Draft <draft-miller-mftp-spec-01.txt>. June 1997.

[7] Clark, D. D., Jacobson, V., Romkey, J., and Salwen, H.: An Analysis of TCP Processing Overhead. IEEE Communication Magazine. June 1989.

[8] Jacobson, V.: Congestion Avoidance and Control. Proc. ACM Sigcomm '88. August 1988.

[9] Floyd, S., Jacobson, V., Liu, C., McCanne, S., and Zhang, L.: A Reliable Multicast Framework for Light-weight Sessions and Application Level Framing. Proc. ACM Sigcomm '95. August 1995.