高信頼同報バルク転送機構

山内長承[†,☆] 城下輝治[††]
佐野哲央[††] 高橋修[††]

コンピュータネットワークを介して同一の情報を多数の受信者に効率良く配布する機構として，高信頼同報バルク転送機構 RMTP（Reliable Multicast Transport Protocol）を設計し，評価する．本論文では，適用例から要求条件を整理した結果，数千から数万端末への数 10 MB のバルクデータの一斉同報配布，TCP と同程度の転送信頼性，送信サーバによる到着確認，情報秘匿や送受信者認証などの要求条件を得た．それを満たす同報再送機構を設計し，応答の送信サーバへの集中（応答爆発）を抑制するための Nack コーディングとバックオフによる応答遅延を検討する．RMTP の転送性能を評価するため，RMTP の再送方式に関する城下のモデルを用いてパケット損失率から再送回数，再送されるパケット数や端末数を推定し，この値を元にバックオフ時間の設定と転送性能の評価手順を示す．たとえばパケット損失率が 1%の場合，5000 端末に対する 2 MB の転送について，5 回の再送で収束し，そのときの各再送段階における転送パケット数，応答パケット数を推定できる．この値はバックオフ設計の基本データとなる．さらに，スター状とトリー状の 2 つのネットワーク構成例について応答爆発を推定しバックオフを含む総転送時間を解析した結果，スター状構成では並列の 1 対 1 通信に比べて RMTP では転送時間が余分な再送とバックオフにより 2.6 倍に長くなる一方，トリー状構成では転送時間が 1/200 となり，その中でバックオフによる性能劣化は 14%となるなどの推定ができ，再送やバックオフの設計に役立つことが分かった．

A Reliable Multicast Bulk Transfer Mechanism

NAGATSUGU YAMANOUCHI,[†,☆] TERUJI SHIROSHITA,[††] TETSUO SANO[††]
and OSAMU TAKAHASHI[††]

This paper proposes and evaluates a design of a reliable multicast bulk transfer mechanism, RMTP (Reliable Multicast Transport Protocol), as a tool for distributing identical information to a large number of receivers effectively. It first enumerates the requirements obtained from analyzing potential application cases with the results that a multicast distribution of dozens of megabyte data to thousands of receivers, transfer reliability compatible to TCP, and data arrival confirmation at the transmission server. We design a multicast retransmission mechanism, and explore methods of reducing the effect of Ack Implosion, namely, Nack coding and the backoff mechanism for response transmission. In order to evaluate RMTP performance with response backoff, we estimated the number of retransmission cycles, transmitted packets and terminals that need retransmission, by using Shiroshita's model. An example with 1% packet loss rate shows that an RMTP transmission of 2 MB data to 5000 terminals converges after five retransmissions, with the figures of each retransmission size. Then, we estimate the total transmission time with backoff delay in two network configurations, star and tree, and obtained the results that in the tree network RMTP reduces the transmission time to 1/200 against point-to-point transmissions, while backoff delay suffers 14% of the RMTP performance. These performance estimation models allow us to design retransmission, performance and backoff time.

1. はじめに

1.1 高信頼同報型バルク転送の必要性

コンピュータネットワークの普及にともない情報の受け手の数が増えかつ裾野が拡大するにつれて，1 対 1 の情報のやりとりだけではなく，同一の情報を多くの受け手が参照する利用形態が増えている．これはインターネットの世界で起こっているだけではなく，企

† 日本 IBM 東京基礎研究所
IBM Research, Tokyo Research Laboratory
☆ 現在，東京都立大学工学部客員研究教授
Presently with Department of Electronics and Information Engineering, Tokyo Metropolitan University
†† NTT 情報通信研究所
NTT Information and Communication Systems Laboratories

業内でもパソコン端末が従業員1人1人に配布されるにつれて，一定のオーソリティを持つ製作者が準備した信用できるデータを多くの受信者が参照する放送型の情報伝達が増えている．

ところが，コンピュータネットワークにおける情報伝達は1対1の通信を中心に発達してきたため，今までは1対多の放送型の伝達を，同じデータに対する，多数の受け手からの独立した1対1アクセスによって実現してきた．たとえば，インターネットで多用されているWebは，送り手のサーバ上に情報を用意し，それを多数の受け手がブラウザにより読み出すことにより伝達する機構であり，またグループウェアで伝達手段として用いられる共有データベースも同様に，送り手がサーバ上に用意した情報を多数の受け手が読み出す機構である．このような受け手ごとのアクセスは，受け手が自分の都合のよい時刻に読み出せる利点がある反面，受け手の数が大きくなると，サーバやネットワークにアクセスが集中し負荷が大きくなり応答性が悪くなる．さらに，このアクセスのトラフィックは同じ情報を繰り返し運ぶために生じているので，著しく無駄である．

多数の受け手に対して同じ情報を配布する手段として，放送型の伝達が有効である．たとえば，伝達媒体が電波や共有媒体の場合，情報を1回送出すればすべての受信者に到達できる．TCP/IPでは放送型の転送として，IP通信の枠組みを同報に拡張したIPマルチキャスト[1)]を利用することができる．

放送や同報の機構では，1対1通信と異なり，情報は送り放しのことが多い．途中経路の雑音や輻輳などにより情報が脱落しても，特別な処理は行わない．音声や動画の場合一瞬乱れる程度の被害で済むが，文字や数値情報の脱落や誤りは影響が大きい．したがって，1対1通信で達成されていると同様の転送信頼性が，1対多通信においても望まれる．

さらに，情報の正しい到着によって課金が成立するサービス，たとえば電子新聞やソフトウェア配送などの場合，データが受信端末に正しく到着したかどうかを送信サーバ側で正しく把握する必要がある．その種の信頼性も，上記の転送信頼性と並んで要求されることがある．本稿では，これらの信頼性の要求を整理し，それに対応する高信頼同報型バルク転送の仕組みを検討し，性能評価を行う．

1.2 高信頼同報型バルク転送の範囲と応用

本論文で対象とする高信頼同報型バルクデータ転送サービスの範囲を，具体的なサービスの応用例をあげることによって検討する．

1.2.1 有料情報の配送の例

情報を配送して課金するサービスで，電子新聞，株価速報などがあげられる．インターネット上ではすでにアクセス制限をしたWebや電子メールによる情報配送サービスが行われている．情報に課金できるためには，

(1) 転送の確実性，実用になる程度に確実にデータが配送されること．
(2) 受信の確認，情報が受信者の手元へ届いたことを送信者が確認できること．さらに，情報不達時には課金免除等のためその旨を送信サーバが知ることができること．
(3) 情報の秘匿・認証．公平な課金のため，料金を払っていない人が受信できないよう秘匿すること，また正当な情報であるかどうか送り手を認証できること，

などが満たされることが望ましい．

情報はテキストのほか写真・図版の類を含むが，日刊紙朝刊の例では圧縮の後2〜5 MBになる．受信者数は，全国紙では8百万部から1千万部を超えるものがある．

1.2.2 企業における業務情報の末端までの配布の例

全国に展開する多数の営業店の端末に対して，最新の商品情報・価格や，契約条件・販促プログラム等の営業情報を配布する．現在は，中央のサーバに置いた情報を，端末からアクセスする方法が多く使われている．Webの機構やグループウェアの共有DBを使う構成など，いろいろな実装が行われているが，始業時など多数の端末からサーバのデータをアクセスするためサーバやネットワークは十分な容量が必要となる．これに対して，高信頼同報バルク転送により各端末に必要なデータを効率良く同報配布する実装が考えられる．

このシステムの要件としては，転送の確実性，センタでの受信の確認，情報の秘匿・認証が必要となる．業務の円滑な遂行のためには，センタでの受信の確認に基づく代替転送手段の起動や，情報の認証による発信者の確認が必要となる．

端末数は全国展開する企業では5千から1万端末程度になる例がある．情報量は，商品情報などに静止画や将来的には動画を含めたデータを含める要求もあり，この場合1件につき圧縮した静止画で200 KBから1 MB程度，動画では数10 MBに及ぶ．配布の頻度は最大でも1日1回程度と考えられる．

1.2.3 企業におけるソフトウェア配布と管理の例

PCやワークステーションを多数運用している企業では，センタ主導でのソフトウェアの管理・更新を望む

場合が多い．更新用のデータを端末に配布し，端末にてインストール・カスタマイズ作業を実行し，更新に成功すればライセンス情報をセンタに集約して管理する，といった一連の作業が必要となる．このときデータ配送システムは，転送の確実性，センタでの受信の確認，情報の秘匿・認証が必要となる．秘匿はライセンス管理をより堅固なものとし，認証は予期しないまたは悪意のソフトウェア更新を防ぐ効果が期待できる．

端末数は，数千台に上る例がある．配布されるソフトウェアの量は，ワークステーションのソフトウェアやデータの配布の場合，200 MB 程度の例がある．この場合ネットワークの制約から，何日かに分割して配布するなどの工夫が行われる．配布の頻度は，安定時には数カ月に1回程度であるが，システム更改直後には1日に何回も転送することがある．

1.3 高信頼同報型バルク転送サービスの要件

このようなサービスを実現するための高信頼同報型バルク転送機構は，次の要件を満たす必要がある．

(1) 多数の受信端末に対して，かなりの大きさの同一情報を効率良く転送する．
受信端末数は，企業内の情報やソフトウェアの配送では，数千～数万端末にのぼる．任意の端末数に適用できることが望まれるが，企業内の利用では大半がこの範囲に収まると思われる．後述する応答爆発の問題から，集中応答を持つ方式は端末数に限界があるが，数千から数万の範囲に対応できれば実用になる．また，1千万部を配布する全国紙の電子新聞化は，2ないし3段の多段接続による対応を考えることができる．情報量は上記の例では現に 10 MB 程度，将来的には回線容量の増加にともない数 100 MB が必要とされる．
効率の良い転送，短時間での転送は，同報通信の性質を生かすことによって実現する．多数の端末からのアクセスを直列に実行すれば時間がかかり，並列に実行すればネットワークやサーバの負荷が非常に大きくなる．同報すれば，基本的には1端末に送信すると同じ時間，同じネットワーク負荷で多端末に転送できるので，効率良く短時間で転送できる．

(2) TCP などの1対1通信に並ぶ程度の転送信頼性がある．

(3) 送信者が受信対象者や転送の成否を管理・確認できる．
情報の課金やソフトウェアライセンスの集中管理などのため，センタサーバが受信者の制限と確認，データの到達の確認をすることができる．対象受信者以外への情報秘匿が必要であると同時に受信者側ではデータの出処を確認するため送信者を認証できる必要がある．

(4) バルク転送を対象とする．
転送するデータはあらかじめ準備されているものとし，始めと終わりが知られているとする．また，転送に実時間性を要求しない．ファイル全体が到着した時点でデータの統一性が保証されればよい．

以上のようなサービス要件から，次のような技術要件が整理できる．

(1) 多端末に対する同報転送
(2) 信頼性
  (a) TCP 程度のデータ転送の信頼性．
  (b) 送信者による受信者の管理．
  (c) 送信者による転送の確認．
(3) 転送情報の秘匿と受信者の認証．
(4) 送信者の認証．

2章では上記の機能要件を念頭において，マルチキャストを用いたバルクデータの同報転送機構 RMTP（Reliable Multicast Transport Protocol）を設計し，3章で性能を評価する．

2. 高信頼同報バルク転送プロトコル RMTP の設計

2.1 IP マルチキャストの利用

多端末に対する同報型通信手段として，広く普及している IP ネットワーク上での稼働を想定し，IP マルチキャスト転送[1]を用いることとする．IP マルチキャスト転送は，IP パケット転送網上での同報機能を実現するもので，上位層に対する論理的なインタフェースは次のように整理できる．

(1) 受信者は特定の端末集合である．その集合は宛先グループとして定義され，1つのグループに1つの IP アドレスを割り当てる．

(2) 任意の送信者がグループに対して送信することができる．送信者は通常制限されない．

(3) 受信者は自分からグループに参加する．グループ参加のための制御プロトコル IGMP（Internet Group Management Protocol）を用いて参加要求 Join を発信し，直近のルータのマルチキャストパケット転送経路を設定させる．

(4) IP マルチキャスト機能としては，IP ネットワーク層のサービスまでを提供する．つまり，パケットの損失や順序反転が起こりうる．1対1の通

信で得られるTCPの整順・再送は，提供されていない．本論文で検討する機能の大半は，1対多通信のためのトランスポート層機能と位置付けられる．

IP層と下位層での同報機能の実装としては，次のことが行われる．

(1) パケットは送信者から複数の受信者へトリー状の経路を経て転送される．IPルータがトリーのノードとなり分岐がある場合複数の宛先にパケットをコピーして伝搬する．

(2) ルータは自分が当該の転送トリーのノードであること，およびマルチキャストパケットをコピー・送出するべき出力線を知らなければならない．このマルチキャスト経路情報は，マルチキャスト経路制御プロトコルによってルータ間を次々に伝搬される．

(3) 媒体がEthernetなど放送可能な場合，その媒体へ接続する中継ルータは同報または放送モードで1コピーを出力する．

高信頼同報バルク転送をこのようなIPマルチキャスト機構を用いて実現する場合の，個々の要件の実現方法について検討する．

2.2 転送信頼性と再送制御

データ転送の信頼性は，TCPと同様にエンド間で再送制御を行って実現する．ルータのホップごとに信頼性制御を行うことは，現在のIP転送機構の中では難しい．また前方誤り訂正（FEC）を用いてエンド間での信頼性を確保する方法は，冗長度をあらかじめ決めてしまうと訂正可能な誤りの個数が限られてしまうので，パケット損失が動的に変化する場合は訂正できない場合が起こる．したがって，エンド間での再送により転送信頼性を確保する方式をとる．

2.2.1 再送手順

1対1通信では主として再送方式とフロー制御が合わせて作られていて，大きく分類するとStart-Stop方式，Go-Back-N方式，Selective-Repeat方式が使われてきた．

Start-Stopはパケットを1つずつ送信，Ack返送にて確認，必要なら再送，という手順を踏み，1つのパケットを完全に転送した後，次のパケットを転送する．この方法は，マルチキャスト環境でもそのまま適用できるが，転送効率が悪く，特にマルチキャストではすべての端末で受信確認できるまで次に進めないため効率が低下する．

Go-Back-N（GBN）はTCPで使われている方法で，送信端は先送りできる範囲をウィンドウとして管理する．受信端はパケット受信ごとにAckを用いて到着を通知し，パケットが失われたときは，そのパケット以降すでに先送りしたすべてのパケットを再送する（**図1**）．マルチキャスト環境では，パケット受信ごとに各端末がいっせいに返すAck/Nackが送信サーバに集中し，サーバのソフトウェアやサーバ周辺のネットワークに対して大きな負荷となる（応答爆発，2.3節に詳述）．またパケット損失時，先送りしたパケットをすべて再送する場合に，すべての経路でこのパケットを転送することになるので，その転送量が非常に大きくなる．

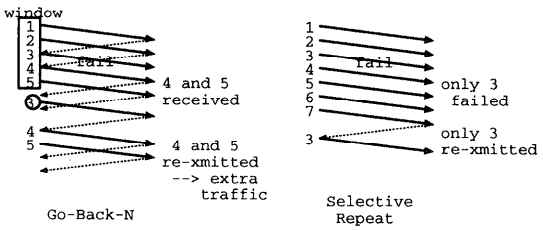


図1 GBNおよびSR方式における再送パケット
Fig. 1 Retransmitted packets in GBN and SR.

Selective-Repeat（SR）は，失われたパケットだけを選択的に再送する方式で，最近いろいろなプロトコルで採用されている．受信端末から，受信できなかったパケットの番号を送信端に通知し，再送する．特に同期性，順序性を要求されないバルク転送においては，データをまとめて転送した後で失われたパケットだけを選択的に再送する方法は，無駄な再送を防ぎ転送効率を向上させる．マルチキャスト環境でのバルク転送では，Ack/Nackの頻度を下げる効果があり応答爆発を軽減できる．またGBNで見られた先送りパケットの重複再送がない．したがって，マルチキャストによるバルク転送では，GBN方式に比べてSR方式が適している．

2.2.2 マルチキャスト再送とユニキャスト再送

多端末配送では，再送の場合においても同一データを複数端末に送信することが多いので，同報を用いることによって転送の効率を向上できる．しかし，必要とする端末のみにマルチキャストするように経路を変更するためには時間を要し，しかもその時間はネットワークの構成や経路に依存し一定ではないため，頻繁にグループを変更することは望ましくない．

再送の方法としては，**図2**のように端末2と4が再送の対象であったとして，1）端末1，2，3，4のすべてに再送する，2）端末2と4にマルチキャストで再送する，3）端末2と4にユニキャストで再送する，という選択があり，

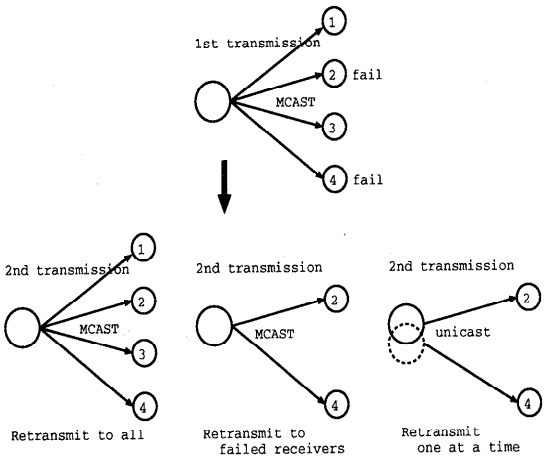


図2 マルチキャストを用いた再送方法
Fig. 2 Multicast Retransmission.

(1) すべてに再送する場合経路設定は不要だが，必要のないリンクに再送トラフィックが発生する点は望ましくない．
(2) 2，4にだけマルチキャストすると，無駄トラフィックは発生しないが，経路を設定し直すために時間を要する．
(3) ユニキャストでは無駄トラフィックは発生しないが，宛先の受信端末数が多い場合には，マルチキャストによる効率良い転送ができない．

という問題がある．

RMTPでは，1）を原則とするが，なるべく2）に近づくように経路の刈り込みを行う方式とした．すなわち，すべてのデータを正常に受信し終えた端末がマルチキャストグループから離脱するようにし，その部分への無駄な転送を防ぐ一方，端末を単調減少させることによって経路設定を待たないで次の再送を始めることができる．

2.3 応答爆発問題

応答爆発（Ack Implosion）は，1対多通信において受信端から送信端へフィードバックを送る場合，多数の応答パケットが送信端に集中するため，送信サーバおよび周辺のネットワークに大きな負荷がかかる問題で，応答のパケットが失われるうえ，他のトラフィックにも影響を与える．

応答爆発に対して，a）近傍の受信端末間で不足データを再送して補い，Nackを送信サーバまで返さないことにより，端末数のスケーラビリティを確保する，b）複数の端末から送られるAck/Nackをマルチキャスト転送トリーの各分岐ノードで1つに併合する，c）スケーラビリティは追求せず実用範囲での規模を実現する，の3つの立場が考えられる．たとえばAT&T RMTP[2]やSRM[3]はa）に属し，StarBurst MFTP[4]やNTT-IBM RMTP[5]はc）に属する．またb）はPGM-RTP[7]で提案されている．一般に，a）は近傍グループの啓性や代表端末選択の機構が複雑なほか，送信サーバによる受信状況の確認がしにくいなど集中管理が難しくなる．また，b）は途中のノードにAck/Nackを併合する機構が必要なため，既存のインターネット上ではルータを改変するか併合サーバを多数設置する必要がある．それらに対して，応答爆発の存在下で実用となる数の端末が収容できるのであれば，制御はc）の方が簡単であるし，ネットワークの改変も必要がない．StarBurst MFTPやNTT-IBM RMTPは，5千から1万端末程度を対象として設計されているため，企業内の利用では大半の要求を満たすほか，消費者向け大規模配布においても2段ないし3段の多段結合を用いて対応できる．

応答爆発の軽減・抑制手段としては，応答パケットのパケット長と，一時に到着するパケット数を減らす工夫が考えられる．

2.3.1 Nackのコーディングの工夫

Nackパケットは未受信パケットのシーケンス番号を返すので，パケット長が大きくなる．コーディングを工夫してなるべく短くすることにより，爆発を軽減することができる．具体的には，シーケンス番号を列記する方法のほか，列記に加えて区間表示の混在を許す方式（**図3**），ビットマップで表現する方式（StarBurst）などが考えられる．区間表示は，パケット損失が連続して起こる場合特に有効である．

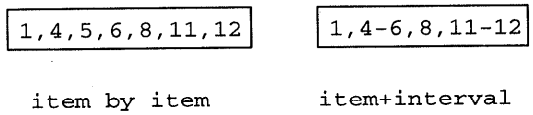


図3 Nackコーディング方式
Fig. 3 Nack coding schemes.

次に簡単な比較を掲げる．Ack/Nackの対象となるパケット総数をNとすると，ビットマップ表示ではN/8バイトが必要になる．他方，番号の表示では，たとえば現行のRMTPの仕様では，1つのシーケンス番号の表示は16ビット，列記の場合は区切り符号は用いず続けて書き，区間表示は2つのシーケンス番号の間に特別な16ビットパターンを挿入する．これを用いた場合，パケット誤り率（PER）をPとすると，列記の場合N*P*2バイトを要する．列記とビットマップが均衡するのはN/8=2NPの点であり，このときのPはP=1/16=17%となる．PERがこれより

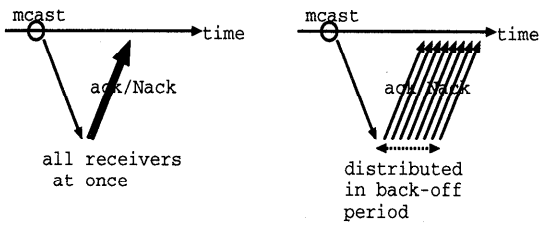


図4 バックオフ機構
Fig. 4 Back off time mechanism.

小さいと列記が，大きいとビットマップが有利となる．また，区間表示については，仮に誤りパケットが5つずつ連番の塊（バースト誤り）になっている場合，誤り区間の発生期待値はN*P/5，1つの誤り区間の記述は始めのシーケンス番号，区間表示符号，終わりのシーケンス番号がともに16ビットであるので合計6バイトとなる．このときの均衡点はN/8=(6*N*P/5)になり，P=5/48=5%となる☆．

2.3.2 バックオフによる同時到着応答数の削減

受信端で応答の送出タイミングを分散させる（バックオフ．**図4**参照）．応答を送出するタイミングを端末ごとに異なる時間で遅らせれば，送信サーバ側での集中が緩和される．送信サーバの単位時間に受信できる応答の上限を元に，応答を遅らせる時間幅を長くする．過度に長くすると，転送効率が低下する．

受信端では受信端末の総数を知ることができないので，送信サーバから端末総数ないしは応答遅延時間を通知する必要がある．通知の方法は

- 送信サーバが端末ごとに遅らせる時間を個別に指定する．この方法は端末数分のユニキャストを繰り返す必要がある．
- 応答遅延時間の最大幅を通知し，端末側でその期間中から乱数を用いて時間を選ぶ．送信サーバからの通知は同一値をマルチキャストすることができるが，応答遅延がどのように分布するかを送信サーバが制御できない．

RMTPでは後者を用いている．

2.3.3 Nackの中間返送による同時到着数の削減

すべての転送終了後にまとめてNackを送るのではなく，転送の途中で区切ってNackを返す．1回のNackのデータ長が短くなると同時に，応答タイミングを受信端末ごとに変えることにより，集中を避けることができる．

2.4 コネクション管理

高信頼同報通信では，特定多数の受信者に情報を配信するので，受信端の認証，転送結果の送信端での確認などのため，エンド間でコネクションを管理する必要がある．

受信端末の認証は，課金をともなうサービスやライセンス付きソフトウェアの配布など，配布対象を限定する場合に必要となる．

コネクションは，送信端がコネクション確立要求メッセージを，受信対象となるマルチキャストグループアドレスに対して送信し，受信端が返答メッセージを返すことで確立する．このとき，マルチキャスト経路が確立できていない受信端は，確立要求メッセージを受信することができないので，コネクションの対象から外される．受信端は返答メッセージに載せて受信端の認証情報を返すことにより，送信サーバは正当な受信者のみを転送対象とする☆☆．

認証はコネクション確立時に行い，拒否された端末はコネクションを確立できない．認証結果は1つのデータの送信に対して有効とし，機構上は1回のコネクション中は有効とする．1回のコネクション中は途中で新たな受信端末が加わることはないので，たとえば定期的な再認証などを考える必要はない．

認証の方式は固定パスワード，ワンタイムパスワード，チャレンジ応答などが考えられるが，固定パスワード方式は繰り返して利用する場合に安全性が確保できない問題があり，ワンタイムパスワードは両端で同期したパスワードを生成することが難しく同期がずれることがありうる．ここではチャレンジ応答方式を用いる．具体的には，送信サーバが受信端末に対してコネクション確立要求メッセージをマルチキャストで送る際にチャレンジ情報を載せて送り，受信端末は確立応答メッセージをユニキャストで返す際にチャレンジ応答をあわせて返答する．チャレンジ応答は，送られてきたチャレンジ情報と，各端末が送信サーバと共有している端末ごとに異なるパスワードを元に，MD5アルゴリズムによって生成している．送信サーバは端末から送られてきた応答を認証し，異なっていればその受信端末とのコネクションを拒否する．

なお，送信者の認証，転送情報の秘匿には送信データ全体に対する暗号化が有効であるが，この処理は上位アプリケーション内で実現するものとし，RMTP

☆ ちなみに，この区間表示を用いた場合，最悪のパターンはOXXXOXXXOXXX（PER = 75%，表示長は6*N/4 = 1.5*N）となり，またパターンOXXOXXOXXOXXのときはPER = 67%，表示は区間表示にはならず列記となり，4*N/3 = 1.3*Nとなる．

☆☆ この受信者の認証は，IPマルチキャスト経路制御を制約するものではないので，不正な受信者がルーティングを設定し，パケットを取り込むことができる．

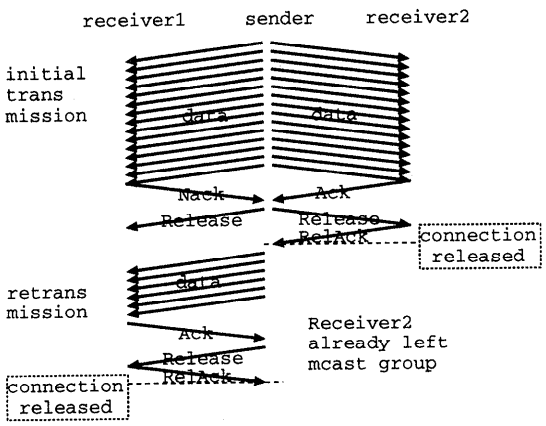


図5 コネクション解放のタイミング
Fig. 5 Connection close timing.

では除外している．

返答メッセージは短いが全端末が応答するため，応答爆発を起こす可能性があり，バックオフの対象とする．確立要求のパラメータとして送信されるコネクション初期化情報の一部に，バックオフ時間を含める．

コネクションの切断・解放は，全端末が全データを正しく受信し終えるまで待ってからいっせいに切断するのではなく，図5のように各端末が全データを受信し終えたらその時点で切断する．この方式は，すべてのデータが揃った端末はこの転送作業からなるべく早く解放できるほか，切断と同時にIPマルチキャストグループを離れることにより，再送マルチキャストパケットの伝搬範囲を小さくできる利点がある．

2.5 転送の流れ

RMTPの全体の転送の流れを図6に示す．

プロトコルヘッダの形式やステートマシンの定義は，RMTPプロトコル仕様書[5),6)]に詳述されている．

送信サーバでは，送信データのほか，同報配布の対象となる受信者のリストが準備されているものとする．受信端末では同報配布サービスの送信サーバのアドレス，受信ポート番号のほか，同報配布サービスのグループアドレスが既知であり，受信端末起動時にIGMPプロトコルを用いてそのグループにJoinしておく．

コネクション設定

初めにコネクションを設定する．その手順は，

(1) 送信サーバからコネクション確立要求をマルチキャストにて送信する．コネクション確立要求には，転送するデータの大きさ，その他のオプションに必要なパラメータを添付する．

(2) 受信端末はコネクション確立要求を受信して，

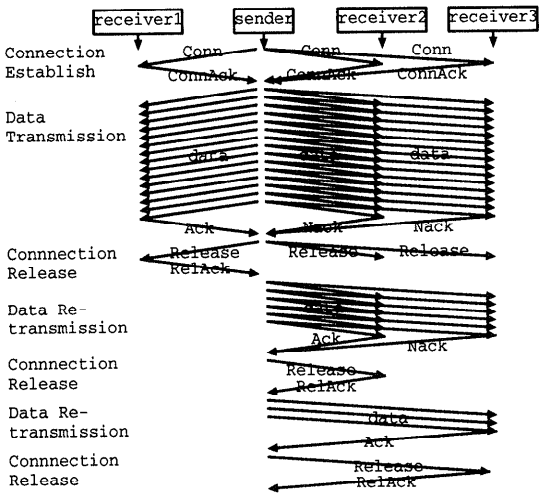


図6 RMTPの転送プロセス
Fig. 6 Overall RMTP transmission process.

返答を返す．この返答には送信サーバが受信端末を認証するために必要なデータが含まれる．

(3) 送信サーバは認証できた受信端末を，同報配布対象として登録する．この時点で同報グループの経路に組み込まれていない受信端末は，以後の転送の対象とならない．

データ転送

(1) データを転送する．
送信サーバは転送対象のデータをパケットに分割して，マルチキャストにて次々に送信する．パケットにシーケンス番号を付し，欠落に備える．連続した転送の最大数はNackの最大長に制約される．

(2) Ack/Nackを返す．
各端末は，コネクション時に通知された総データ長を見て，パケットの欠落を判定する．すべてのパケットが受信できていればAckを返し，コネクション切断へ移行する．欠落パケットがある場合は，不足しているパケットをNackを用いて送信サーバに通知する．

(3) 送信サーバは，すべての受信端末からAckまたはNackが届いた後，図7のような表を用いて端末の受信成否を集計し，未受信端末があるパケットは再送する．AckもNackも返さない端末は，タイムアウトにより問合せを行って受信状況を集める．

コネクション解放

(1) 1回のデータ転送または再送フェーズが終了するごとに，全データを正しく受信し終えた端末

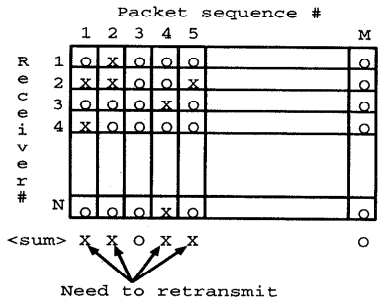


図7 受信成否の表
Fig. 7 Success-failure table.

を解放する．送信サーバはAckを返した受信端末があれば，それらの端末のためにコネクション解放メッセージをマルチキャストにて送る．

(2) 全データを正しく受信しAckを返した受信端末は，コネクション解放メッセージを受信すると，コネクション解放に対する確認返答を返し，使っていた資源を解放する．

(3) 送信サーバは受信端末からの解放確認を受けて解放処理を行い，次の再送サイクルに入る．

再送サイクルは，再送要求のあるパケットのみを対象にしてマルチキャストによるデータ転送とコネクション解放の手順を繰り返す．このときのマルチキャストグループは，コネクション解放により離脱した端末を除く全端末とする．個々のパケットについて未受信端末の集合は異なるが，グループの組直しは時間を要するのでパケットごとに行わず，未受信パケットが1つでもある端末すべてを含むグループに対してマルチキャスト再送する．このようにして，再送対象となるパケット数も未受信データのある端末数も減少しながら再送が進行し，最後にすべての端末がすべてのデータを受信し，コネクションを解放した時点で，この転送が終了する．

3. 性能解析

ここではRMTPの基本手順を用いた場合の転送時間について分析する．転送時間の要素として，データ転送時間，再送回数と各再送における転送データ量，応答爆発に対するバックオフの遅延があげられる．

3.1 データ転送時間

マルチキャストの場合，IPレベルの転送時間は端末数に依存せずユニキャストと同じ転送時間と考えられる．ルータ内の処理時間はユニキャストより若干増えるが，全体から見ると無視できるほど小さい．

他のトラフィックがない環境では，ボトルネックとなる回線の回線速度がデータ転送時間を支配する．他のトラフィックと混在する場合については，別途輻輳制御の問題として検討する[8),9)]．

3.2 再送回数と各再送における転送データ量

パケット誤りによる再送回数と各再送における転送データ量の見積りモデルは，城下が文献10)，12)に示しており，下記の漸化式によって求められる．

このモデルでは次の点を前提条件としている．まず，パケット誤りはすべてのパケットに対して同一の確率eで発生するとする．つまりトリー状の経路では上流で発生した損失は下流に伝搬するが，このモデルではその効果は無視し，各受信端末は送信サーバと独立した回線で接続され個々の回線のパケット誤り率をすべてeとする．本モデルでは再送の機構から考えてパケット誤りとパケット損失は区別しない．

各再送段階では，パケットごとに集計して1つでも未受信の端末があれば，そのパケットは再送の対象とする．各端末では送られてくるパケットにより自らの未受信パケットが埋まれば次回からそのパケットを再送要求しない．したがって，未受信パケットの埋まり方に関してはどの端末のどのパケットが失われるかによらず，全体の未受信延べパケット数 S_k は1回の再送につきeの割合で減少していく．すなわち $S_k = e * S_{k-1} = e^k * S_0$ となる．

文献10)，12)と同様，第k回目の再送の対象となる，1つ以上未受信パケットのある端末数を N_k，1つ以上未受信端末のあるパケット数を M_k とすると，これらについて，漸化式

$$N_k = \{1-(1-e)^{S_{k-1}/N_{k-1}}\}N_{k-1}$$
$$M_k = \{1-(1-e)^{S_{k-1}/M_{k-1}}\}M_{k-1}$$

が成り立つ．

これに，端末数 $N_0 = 5000$，データ数 $M_0 = 2000$ パケット（データ長2MBを長さ1KBのパケットに分割），パケット誤り率を $e = 10^{-6}$，10^{-4}，10^{-2} を与えると，(S_k, M_k, N_k) の値は，

	e=10E-6	e=10E-4	e=10E-2
k=0	10M, 2K, 5K	10M, 2K, 5K	10M, 2K, 5K
k=1	10, 10, 10	1K,787,906	100K,2K, 5K
k=2		0.1,0.1,0.1	1K,787,906
k=3			10, 10, 10
k=4			0.1,0.1,0.1

のように計算できる．空欄は値が0.01以下である．これによると，パケット誤り率1%のときは，5回の再送で5000端末に2000パケット（2MB）を送れることが分かる．またそのときの総転送パケット数は ΣM_k=4799個，サーバに到着する応答の総パケット

数は ΣN_k=10918 個になることが分かる．また，各再送段階での応答パケット数 N_k は，次節の応答爆発の評価に用いる．

また，端末数を一定（5000 端末）とし，いくつかのデータ長（パケット数）について再送回数を求めると，

PER packets	10E-1	10E-2	10E-3	10E-4	10E-5	10E-6
1000	9	5	3	3	2	2
2000	10	5	4	3	2	2
5000	10	5	4	3	2	2
10000	10	5	4	3	2	2
15000	11	6	4	3	3	2
20000	11	6	4	3	3	2

のようになり，再送回数はデータ長に関してあまり敏感でないことが分かる．

3.3 バックオフによる応答爆発の抑制と性能劣化

2.3.2 項に述べたように，応答爆発を抑制するため，受信端末側で応答送出を遅らせて到着を分散させるバックオフ機構を採用する．このとき，応答を待つために転送プロセスの進行が遅れ，性能は低下する．バックオフ時間は，応答の総量がバックオフ時間内に平均に分散して発生したときに，応答流入量がサーバ・ネットワークの処理容量を越えないように設定するのがよい．

前節で求めた5000端末，2000パケット（2 MB），パケット損失率 10^{-2} の例について，Nack データ量を推定する．第0回目の転送を行った後，延べ $S_1 = 100$ K パケットの再送要求があり，Nack データのうちの損失パケットの列記表示のために 200 KB を要する．また，Nack データのヘッダ部分について求めると，1回目の再送では N_1=5000 端末すべてから Nack が送られ，端末あたりの平均の再送要求数は 100 K 個/5000 端末=20 個で列記表示をすると 40 B になるから端末あたり1パケットに収まる．したがって，Nack パケットの総数は端末数と同じ 5000 個と推定され，ヘッダ（IP，UDP，RMTP を含む．物理層のヘッダはここの計算では除外する）のために 40 B × 5000 = 200 KB を要する．合計すると 400 KB のデータがサーバへ流入する．なお，再送は第1回目が最も多く，第2回再送への要求を同様に求めると約 38 KB となり，ほとんど無視できる．

次に，応答の集中に対するネットワーク構造の影響の評価方法[11]について，2つの極端な場合を例にして検討する．ネットワークが中継ノードのないスター状の構成で送信サーバから各端末への回線容量が 128 Kbps の場合，伝搬遅れは全端末に共通になり，応答はいっせいに返される．送信サーバの応答処理能力を毎秒 1000 パケット（1つの応答の処理に 1 mS を要する）とすれば，5000 端末からの応答を 5 秒間に分散するバックオフ設定により溢れを防ぐことができる．このバックオフ時間をコネクション設定応答およびデータ転送応答に用いるとすると，全転送時間は，コネクション設定時間（転送 7 mS，バックオフ 5 秒），データ転送（2 MB データの転送時間 125 秒，バックオフ 5 秒），再送（1，2，3，4 回目の合計で，データ転送 175 秒，バックオフ 20 秒）を合わせて，およそ 330 秒となる．その他，細かいオーバヘッドは含めていない．1 対 1 通信を用いた場合と比較すると，たとえば，TCP のコネクションを 5000 端末に対して同時並列に開くことができたとすると 126 秒（パケット誤り率 1%を仮定）となり，RMTP の場合 2.6 倍の時間を要している．これは個別に再送すると各端末に対して 10 パケットを並列に再送すればよいのに対して，マルチキャストで再送すると端末間で OR されて 2799 パケット再送する時間とバックオフの時間がかかっているためである．このうちバックオフ時間は 30 秒，9%に相当する．

なお，総転送時間にはパケット損失の判定に要するタイムアウト時間を考慮する必要があるが，ここにあげた例の場合，バックオフ時間に比較して無視できるほど小さいことが分かる．

ネットワークが平衡した 8 段の 3 分木の構成で，端末は末節にのみ収容，各リンク回線容量が 128 Kbps の場合では，受信端がいっせいに応答を返すと，送信サーバに直結される 3 本のリンクの転送能力 376 Kbps（48 KB/s，600 パケット/秒に相当）が転送の隘路となり，溢れを起こす．ネットワークの応答転送能力が送信サーバの応答処理能力を下回るので，ネットワークに合わせてバックオフ時間を 8.3 秒とすると，全転送時間は約 350 秒となる．このネットワークでは 1 対 1 通信を用いるとサーバ直下の 3 本のリンクが隘路となって並列度が上げられず，376 Kbps で転送する結果 70000 秒（約 20 時間）を要するので，RMTP は転送時間を 1/200 に短縮することになる．RMTP の転送時間中バックオフ時間の占める割合は 50 秒，14%になるが，第 2 回目以降の再送ではバックオフが必要なくなるため取り除くと全転送時間 325 秒のうちバックオフは 8%となる．

以上の例で分かるとおり，RMTP の同報・再送機構の性能評価には，再送回数や各再送時のデータ量のほかバックオフ時間が無視できない．バックオフ時間設定のためには応答データ量，応答端末数の推定が必要であり，ここに掲げた手順に従って求めることがで

きる．

送信サーバの応答受信能力が溢れを起こす場合については，受信バッファを 50 KB としたときのモデルと解析を文献 10) に示している．そこではバックオフ機構に加えて，Ack/Nack が失われ，POLL によって Ack/Nack を再送する場合の所要時間を求めている．

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 RMTP について，想定する応用例から要求を抽出し，その中で転送信頼性を満たすための再送機構の設計について詳述した．再送の導入にともなう応答爆発は，同報端末数に上限を与えてしまうが，企業内配布の実用規模である数千から数万端末を可能とするための爆発抑制手法を検討した．

全体の転送性能を評価するため，再送機構をモデル化し，再送回数，再送対象パケット数，再送を必要とする端末数を評価した．また応答爆発抑制のため，バックオフタイムによる応答遅延機構を用いるが，その遅延時間の所要値の推定法を示し，2 つのネットワーク構成について転送所要時間に対する影響を評価した．

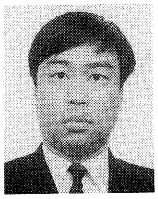
RMTP は NTT および IBM にて独立に試作実装され，評価実験されている．少数の実端末による実験のほか，端末エミュレータによる端末の評価実験が NTT によって行われている[10]．また RMTP の実用化が図られており，実働環境での評価が期待できる．

本稿で触れていない問題として，ネットワーク構成の応答爆発に対する影響の詳しい評価，サーバの応答処理能力の詳細な評価がある．これらはより正確な性能予測に必要となるパラメータである．より詳細な検討を進めており，前者は文献 11) に，後者は文献 12) に一部の結果を発表しているが，さらに詳細な検討を進めている．また，受信者の認証に付帯して鍵の配布など運用にともなう問題があるが，これらについては実用化を進める中で詳細を詰めている．高信頼マルチキャストにおけるフロー制御・輻輳制御の要求，手法，評価も，実用化するうえで重要なポイントであり，詳しい検討を進めている[8),9)]．これらについては稿を改めて発表する．

参考文献

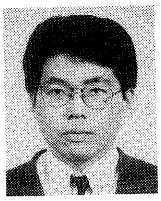
1) Deering, S.E.: Host Extensions for IP Multicasting, IETF RFC1112 (1989).
2) Lin, J.C. and Paul S.: RMTP: A Reliable Multicast Transport Protocol, *Proc. IEEE Infocom '96*, pp.1414–1424 (1996).
3) Floyd, S., Jacobson, V., Liu, C., McCanne, S. and Zhang, L.: A Reliable Multicast Framework for Light-weight Sessions and Application Level Framing, *Proc.ACM SIGCOMM '95*, pp.342–356 (1995).
4) 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> (1997).
5) Shiroshita, T., Sano, T., Takahashi, O. and Yamanouchi, N.: Reliable Multicast Transport Protocol, IETF Internet Draft <draft-shiroshita-rmtp-spec-01.txt> (1997).
6) Shiroshita, T., Sano, T., Takahashi, O. and Yamanouchi, N.: Reliable Multicast Transport Protocl Version 2. IETF Internet Draft <draft-shiroshita-rmtpv2-spec-00.txt> (1997).
7) Speakman, T., Farinacci, D., Lin, S. and Tweedly, A.: PGM Reliable Transport Protocol, IETF Internet Draft <draft-speakman-pgm-spec-01.txt> (1998).
8) Sano, T., Shiroshita, T., Takahashi, O. and Yamashita, M.: Monitor-based Flow Control for Reliable Multicast Protocol and Its Evaluation, *Proc. IEEE IPCCC '97*, pp.403–409 (1997).
9) 山内長承，佐野哲央，城下輝治，高橋 修：高信頼マルチキャストにおけるフロー・輻輳制御，インターネットコンファレンス'97論文集，pp.63–77 (1997).
10) 城下輝治ほか：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, pp.137–144 (1996).
11) 山内長承，城下輝治，佐野哲央，高橋 修：高信頼同報での再送機構の Ack Implosion の再評価，第 55 回情報処理学会全国大会論文集，6T-08 (1997).
12) 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* (1996).

(平成 9 年 10 月 15 日受付)
(平成 10 年 3 月 6 日採録)



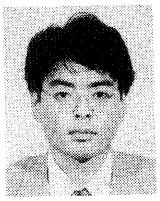
山内 長承（正会員）

昭和28年生．昭和50年東京大学工学部電子工学科卒業．昭和58年同大学院情報工学専門課程博士課程中退．昭和53～59年スタンフォード大学大学院在学．昭和59年日本アイビーエム（株）入社．現在，東京基礎研究所勤務．東京都立大学工学研究科に客員教授として出向．主としてOS，並列プログラムの検証，計算機ネットワークの応用の研究開発に従事．工学博士．ACM，IEEE，日本ソフトウェア科学会各会員．



城下 輝治

1982年京都大学工学部数理工学科卒業，1984年同大学院修士課程修了．同年日本電信電話公社入社．現在NTT情報通信研究所知的通信処理研究部主任研究員．主としてマルチメディア通信処理と通信プロトコルの研究開発に従事．電子情報通信学会，ACM各会員．



佐野 哲央

1992年大阪大学基礎工学部情報工学科卒業，1994年同大学院修士課程修了．同年NTTに入社．以来マルチキャスト通信プロトコルの研究開発に従事．現在，情報通信研究所知的通信処理研究部勤務．電子情報通信学会会員．

高橋 修（正会員）

1975年北海道大学大学院情報工学専攻修士課程修了．同年，日本電信電話公社入社．現在，NTT情報通信研究所知的通信処理研究部主幹研究員．主としてマルチメディア通信サービスとプロトコルの研究・開発に従事．