多段接続された計算機間の遅延を抑制するパケットスケジューリング方式の設計と評価

岩嵜正明[†,††] 竹内理[†] 中野隆裕[†]
中原雅彦[†] 谷口秀夫[††]

従来，我々は，高精度周期スケジューリング機能を実現する OS カーネル Tactix を基盤に，実時間通信に不可欠な帯域保証技術を開発してきた．しかしながら，ビデオ会議等への応用で課題となるルータ多段接続時の遅延時間に関しては考慮されていなかった．本論文では，周期送信機能を持ったパケットスケジューラ内部での遅延発生メカニズムを詳細に分析し，帯域保証とともに遅延時間の低減を可能とする改善方式を提案する．また，この改善方式よるルータ多段接続時の遅延時間を実測し，背景トラフィックが混在する多段接続 Ethernet 環境において，帯域保証リアルタイムストリームに対して，パケットロスト率 0 を維持しつつ，ルータ 1 段あたりのパケット転送の最大遅延時間を，送信駆動周期の 2 倍以内に抑制できることを確認した．

Design and Evaluation of a Packet Scheduler to Reduce the Communication Delays over Cascaded Networks

Masaaki Iwasaki,[†,††] Tadashi Takeuchi,[†] Takahiro Nakano,[†]
Masahiko Nakahara[†] and Hideo Taniguchi[††]

In this paper, we propose a real-time packet scheduling method that enables the low latency packet forwarding for the real-time bi-directional continuous media communication on a LAN including cascaded IP routers. The experimental implementation of the packet scheduler can reduce the maximum latency in the single router to be less than twice the period of the isochronous transfer without any packet loss for a real-time stream. The packet scheduler can also reduce the maximum latency of cascaded routers to be less than the period of the isochronous transfer multiplied by the number of hops plus one with heavy background traffic.

1. はじめに

近年の IP ネットワーク技術やビデオ CODEC 技術の進歩により，これらを応用した IP 電話やビデオチャット，あるいは，Web カメラによる低コストなセキュリティ監視や多地点間ビデオ会議等，企業内における連続メディア通信の利用が急速に拡大している．

一方，ワームをはじめとするセキュリティ上の脅威に対処するため，企業内ネットワークにおいては，従来のように L2 スイッチを使用したフラットなセグメント上に多数のコンピュータを接続する構成は見直されている．具体的には，適切なサブネット分割を施し，各企業のセキュリティポリシに従ってサブネット間の通信をフィルタリングする，あるいは，プロトコル種別に応じてトラフィックを抑制するといった対策が進められている．

こういった状況下において，従業員が数百人以上の規模の組織では，基幹ネットワークを構成する高性能 IP スイッチ以外に，各フロア内にいくつも設置される低コストな IP スイッチにも，L3 ルーティング，パケットフィルタリング，QoS 保証等の高度な機能が求められるようになってきている．

本論文では，比較的低価格な L3-IP スイッチ（以下，ルータと記す）にこれらの高度な機能を実装することを前提に，ソフトウェアベースのパケットスケジューラを対象として，対話型の連続メディア通信に必要な遅延時間保証の実現について述べる．企業内ネットワークを流れるトラフィックは，リアルタイム性を要求されるストリーム型の連続メディア通信と，共有ファイルサーバへのアクセス等による突発的なデータ通信とが混在する．この突発的なデータ通信によって，連続

† 日立製作所システム開発研究所
Systems Development Laboratory, Hitachi Ltd.
†† 岡山大学大学院自然科学研究科
Graduate School of Natural Science and Technology, Okayama University

メディア通信の帯域が圧迫されないことに加えて，双方向の対話型通信に必要な遅延時間の保証を目指す．特に本論文では，ルータが多段接続された環境において，ソフトウェアによるパケットスケジューリング処理オーバヘッドが遅延時間に与える影響を検討する．

従来，我々は，高精度な周期スケジューリング機能[1),2)]を実現する Tactix オペレーティングシステムを開発し，Ethernet 上でも高品質なビデオストリーミングが可能なことを実証してきた[3)]．しかしながら，遅延時間に関しては，これまで十分な検討評価を行っていない．本論文では，帯域保証重視で設計してきた Tactix のパケットスケジューラを遅延時間保証の視点で見直し，ルータを多段接続した環境において，遅延時間を実測評価した結果について述べる．

2. 遅延メカニズムの分析

CSMA/CD 方式の Ethernet を介して多段接続したルータ群における遅延の原因は 2 つに大別できる．第 1 は，セグメント内の混雑による MAC 層での再送回数の増加，第 2 は，ルータ内部のパケットスケジューリング処理オーバヘッドである．

ここでは，まず，2.1 節で TTCP/ITM（Total Traffic Control Protocol/Isochronous Transfer Mode）適用時の MAC 層再送機能への影響を詳細に分析し，さらに，2.2 節で MAC 層再送による遅延時間分布が，パケットスケジューリングに及ぼす影響を明確にする．

2.1 MAC 層再送機能への TTCP/ITM の効果

これまでの研究[1),3)]で述べたように，TTCP/ITM は，約 10 ミリ秒程度の駆動周期ごと☆に各ノードから送信されるデータ量を制限し，駆動周期程度の短い時間間隔でセグメント内の総トラフィックを抑制☆☆する．これにより，送信時の衝突発生確率を低減させると同時に，衝突発生時には，MAC 層の再送機能を利用して，少ないリトライ回数で送信が成功する確率を向上させる．

図 1 は，この効果を確認するために，Half Duplex モードで動作する Shared Hub に 4 台の TTCP/ITM 機構を組み込んだホストを接続し，各ノードから 10 Mbps（合計で 40 Mbps）のレートでいっせいに送信を行い，100 Mbps Ether チップ内に実装された衝突カウンタを利用し，送信衝突回数（再送回数）分布をモニタリングした結果である．同図の横軸は再送回数，縦軸は発生頻度（パケット数/秒）を示す．

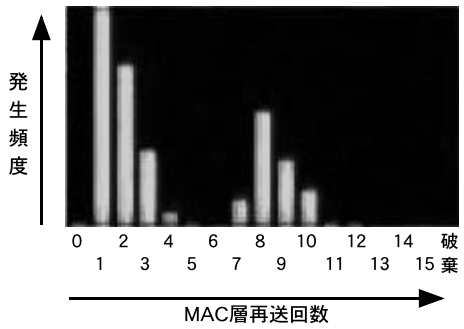

図 1 MAC 層における再送回数の分布
Fig. 1 Retransmit pattern in MAC layer.

TTCP/ITM を適用した場合，周期ごとのセグメント内の総トラフィックを 50 Mbps 程度以下に抑制すると，MAC 層での 10 回を超える再送はほとんど発生しなくなり，送信パケット破棄もなくなる．ただし，この場合，図 1 から分かるように，再送回数の分布に 2 つのピークが出現する．第 1 のピークは再送回数が 1 回から 5 回の範囲であり，第 2 のピークは 7 回から 10 回の範囲である．100 Mbps Ethernet の場合，MAC 層の再送間隔 T は次式で規定される．

$$T = 5.12\,\mu\text{sec} \times \text{random}(0, 2^{**}k)$$

ただし，$k = \min(10, \text{再送回数})$

よって，この式から，第 1 のピークでは累積遅延時間の最大値は 0.3 ミリ秒以内に収まるが，第 2 のピークでは 10 ミリ秒前後となる．

図 2 と**図 3** は，図 1 と同様な測定環境において，セグメント内の遅延時間の分布を，それぞれ TTCP/ITM 無効時と TTCP/ITM 有効時とで実測比較した結果である．図 2 の TTCP/ITM 無効時は，遅延時間分布の裾が大きく広がっており，通常の Ethernet では実時間転送が困難であることを裏付けている．一方，図 3 の TTCP/ITM 有効時は，遅延時間の分布が 10 ミリ秒前後より小さい範囲に収まっている．

このように TTCP/ITM は，通常の方式による Ethernet への送信に比較すると，きわめて高品質な遅延時間の低減が実現できる☆☆☆．しかし，その遅延時間分布には約 10 ミリ秒を超える広がりが残存している．

2.2 MAC 層遅延分布のスケジューラへの影響

Tactix パケットスケジューラ内部での遅延発生メカニズムには，前述の MAC 層再送による最大 10 ミリ秒前後の遅延時間分布の広がりと，周期駆動スレッド

☆ この周期は CSMA/CD 方式の 100 Mbps Ethernet における送信衝突時の再送バックオフタイムに依存している．

☆☆ 具体的には，ストリーミング開始時に予約した帯域幅に相当するパケット数以内に送信量を抑制する．

☆☆☆ 総トラフィックが物理帯域の 50%の場合，10 回連続して衝突が発生する確率は 0.5 の 10 乗（約 0.1%）以下となる．

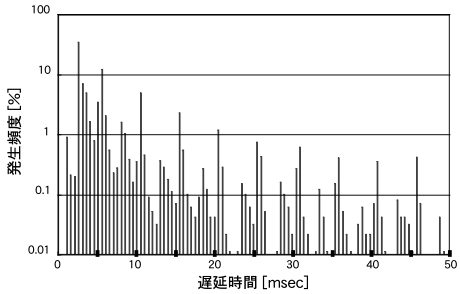

図 2 TTCP/ITM 無効時の遅延時間分布
Fig. 2 Delay pattern without ITM.

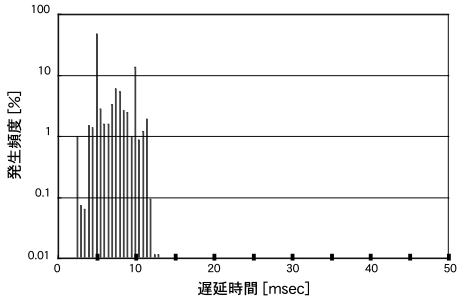

図 3 TTCP/ITM 有効時の遅延時間分布
Fig. 3 Delay pattern with ITM.

の存在とが関与している．以下，まず，2.2.1 項でパケットスケジューラの構造について概説し，2.2.2 項でこの遅延発生メカニズムについて述べる．

2.2.1 Tactix パケットスケジューラの構造

ここでは，2.2.2 項の議論の前提となる Tactix パケットスケジューラの構成要素とその役割について述べる．本パケットスケジューラは，**図 4** に示すように，処理機能を SLIH（Second Level Interrupt Handler）スレッド，IP スレッド，RT-IP スレッド☆，ITM スレッドの 4 種類のスレッドに分割した構造としている．これにより，各スレッド内の処理を単純化し，同時に，スレッド内での複数パケットの一括化処理によるオーバヘッドの低減を図っている．

また，これらのスレッド間の制御フローは，図 4 の矢印で示すように閉ループを形成し，受信から送信に至る全過程において，バッファをキュー間のつなぎ替えのみで受け渡し，CPU によるバッファ内容のコピーを完全に排除している．さらに，受信側でのバッファ資源の割当てと送信側での解放を均衡させることで，高速ネットワークで発生しがちな受信バッファ枯渇を回避している．

図 4 に示すパケットスケジューラの構成要素，および，構成要素間の制御フローの概略を以下に説明する．

(1) **デバイス**：Ethernet デバイスのハードウェアである．ルータの場合，少なくとも 2 つのデバイスが存在する．

(2) **FLIH**（First Level Interrupt Handler）：デバイスからの受信完了や送信完了の割込みによって起動される割込みハンドラである．割込み元デバイスを特定し，SLIH ベクタを介して適切な SLIH スレッドの起動をアイソクロナススケジューラに要求する．

(3) **アイソクロナススケジューラ**：高精度なスレッドの周期駆動機能を有し，スレッド群をスケジュールする．割込み発生後は，FLIH からの要求に従って，適切な SLIH スレッドをスケジュールする．

(4) **SLIH スレッド**：デバイスを制御し，送受信処理を実行するイベント駆動型のドライバであり，各デバイスごとに SLIH スレッド☆☆の実体が 1 つ存在する．各 SLIH スレッドは送信処理部と受信処理部から構成し，さらに，送信処理部，受信処理部ともに，デバイスへコマンドを送るダウンコール部と，デバイスからの割込みで起動されるアップコール部から構成する．

(5) **IP スレッド**：IP スレッドはイベント駆動型スレッドで，SLIH スレッドからの NRT 受信キューへのパケット到着をブロックして待ち，NRT 受信キューに混在して到着した NRT パケットを宛先別にルーティングし，適切な NRT 送信キューにキューイングする．NRT 受信キュー，NRT 送信キューともに FIFO キューである．

(6) **RT-IP スレッド**：RT-IP スレッドは周期駆動スレッドで，SLIH スレッドからの RT 受信キューへのパケット到着を駆動周期ごとにチェックし，RT 受信キューに混在して到着した RT パケットを宛先別にルーティングし，さらに，ストリームごとに各 RT 送信キューへ分離キューイングする．RT 受信キュー，ストリームごとの各 RT 送信キューはすべて FIFO キューである．

(7) **ITM スレッド**：ITM スレッドは周期駆動スレッドで，IP スレッドや RT-IP スレッドから，

☆ 以下，リアルタイムを RT，非リアルタイムを NRT と略す．

☆☆ Tactix では SLIH は独自のコンテクストを持った完全なイベント駆動型スレッドとして実装している．ただし，その実行優先順位は通常のイベント駆動スレッドや周期駆動スレッドよりも高い．

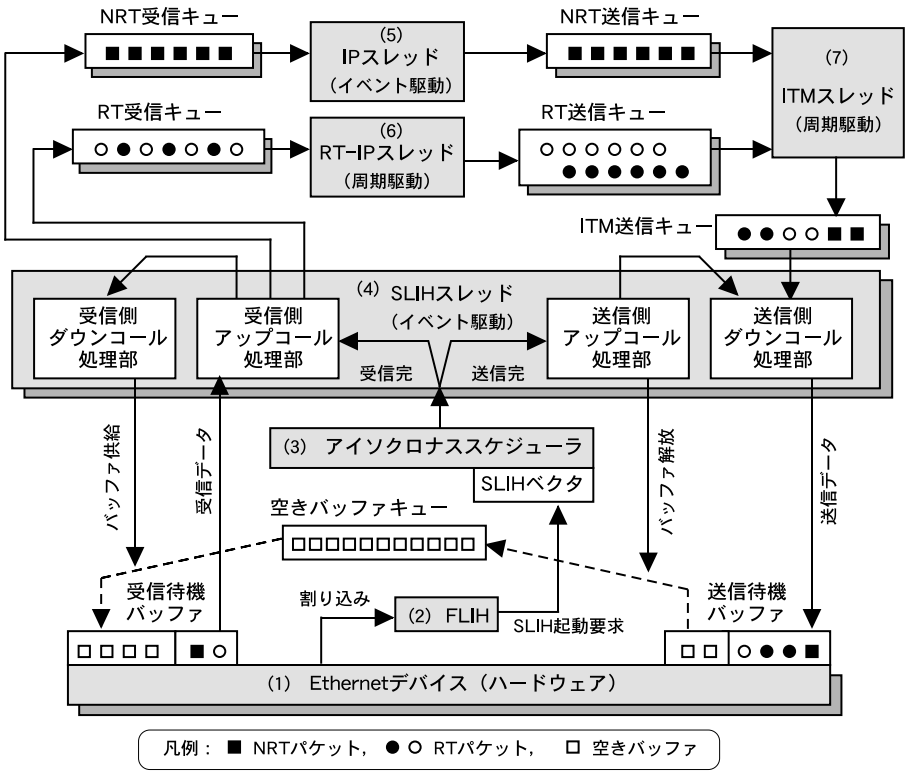

図 4 パケットスケジューラの構造
Fig. 4 Packet scheduler architecture.

それぞれの送信キューにキューイングされたパケットを受け取る．RT パケットについては各ストリームごとに予約されている帯域に相当する駆動周期あたりのパケット数を ITM 送信キューに移す．NRT パケットについては NRT トラフィック全体に許可されている帯域に相当する駆動周期あたりのパケット数を ITM 送信キューに移す．

以上のように，Tactix パケットスケジューラ内部では，受信パケット群を RT パケットと NRT パケットに振り分けて各キューへデマルチプレクスする処理，さらに，RT パケットの場合には，ストリームごとの各キューへデマルチプレクスする処理等が行われる．すなわち，複数の周期駆動スレッドやイベント駆動の割込みハンドラが，それぞれの時間的な制約条件を満足しつつ，各キュー☆を介して適切に同期しながら，これらの処理が実行される．なお，これらのキューへのアクセスに対するスレッド間排他制御は，細粒度プリエンプト制御[5] により，アトミックな操作中はプリエンプトを禁止することで，リアルタイム性能を損なわないように配慮している．

2.2.2 スケジューラ内部での遅延拡大

TTCP/ITM を使用する場合，2.1 節に述べたように MAC 層での再送回数の分布に 2 つのピークが発生し，セグメント内の遅延時間分布には数ミリ秒から 10 ミリ秒前後の広がりが生じる．このため，Tactix パケットスケジューラでは，周期駆動スレッドが，遅延の小さいパケット群を N 周期目に受信できた場合，比較的遅延の大きいパケット群を受信するタイミングは，次の駆動周期（N + 1 周期目）となる可能性が高くなる．

この様子を**図 5** に示す．同図 (a) は，受信と周期駆動のタイミングを示している．この図の最下段の濃灰色部分は，周期駆動スレッドに割り当てられたタイムスロットを示している．RT-IP スレッドや ITM スレッドはこれらの予約されたタイムスロット以外では CPU 実行権を得ることができない．このため，これらのスレッドが N 周期目において処理可能なパケット群は，同図の最上段に示す薄灰色の時間領域内で受信が完了していなければならない．同図の中段に示す薄灰色の時間領域に受信したパケット群は，N + 1 周期

☆ NRT 受信キュー，RT 受信キュー，NRT 送信キュー，RT 送信キュー，ITM 送信キューは，各々，デバイスごとに存在する．

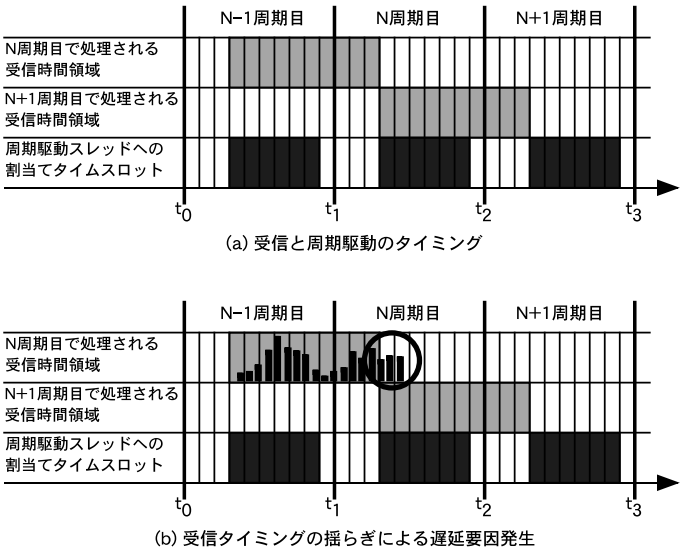

図 5 MAC 層再送遅延時間分布とスケジューリング遅延
Fig. 5 Scheduling delay caused by retransmission.

目以降で処理されることになる．

図 5 (b) は，図 3 の遅延時間分布を重ねて表示している．本来，この遅延時間分布がこの図の最上段の時間領域（駆動周期 = 10 ミリ秒）内に収まることが望ましい．しかし，この図の丸印で示すように，ある確率（セグメント内のトラフィックによって数%から数十%）でパケット到着が遅延し，これらの受信遅延したパケット群は，N + 1 周期目にルーティング処理され，次段への送信は駆動周期分遅延してしまう．

以下，**図 6** に示す構成例を用いて，このパケット受信タイミングの分散が，スケジューラ内部の遅延や次段へのパケット送信に与える影響を説明する．図 6 のシステムでは，左端の NRT 送信ノードと RT 送信ノードがそれぞれパケット群を送出し，これらのパケット群はルータノードを経由して入力側セグメントから出力側セグメントへ転送される．同図の Ethernet スイッチは 100 Mbps Half Duplex の Shared Hub モードで動作し，各セグメント内では CSMA/CD 方式によるパケット衝突を低減させるために TTCP/ITM による送信制御が施されていることを前提とする．

図 7 は，3 駆動周期にわたって，ルータノードにおけるパケットの受信→周期駆動スレッドの実行→次段への送信のタイミングを表している．パケット群 1 とパケット群 3 は，NRT 送信ノードから N − 1 周期目に一括送信されたが，MAC 層再送によって受信タイミングが分散化している．パケット群 5 とパケット群 7 についても同様である．また，RT 送信ノードから一括送信されたパケット群 2 とパケット群 4，および，

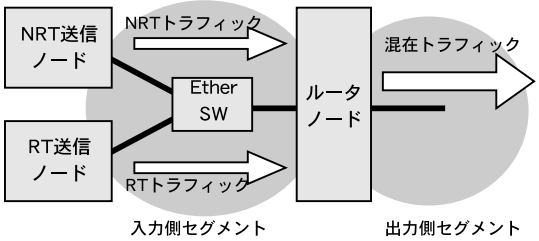

図 6 システム構成例
Fig. 6 An example configuration.

パケット群 6 とパケット群 8 についても同様である☆．

N − 1 周期目において，パケット群 1 とパケット群 2 は，周期駆動スレッドの実行以前に受信が完了しており，これらのパケット群はこの周期内で次段への送信が開始される．一方，パケット群 3 とパケット群 4 については，周期駆動スレッドの実行開始までに受信が完了していないため，パケット群 5 やパケット群 6 と一緒に N 周期目にルーティング処理され，次段へ送信される．この図 7 に示す状況では，各周期における受信パケットの総数は，当該周期内に送信可能なパケットの総数（TTCP/ITM で許可されているストリームごとの帯域に相当する）以下に収まっており，スケジューラ内部での遅延は最小（ほぼ駆動周期と同程度）に収まっている．

一方，**図 8** は，MAC 層での再送増加により，N − 1

☆ 説明を簡略化するため，パケット群が受信単位であるように記述しているが，実際には，パケット群 1 とパケット群 2 に含まれるパケットは混在した順序で受信される．

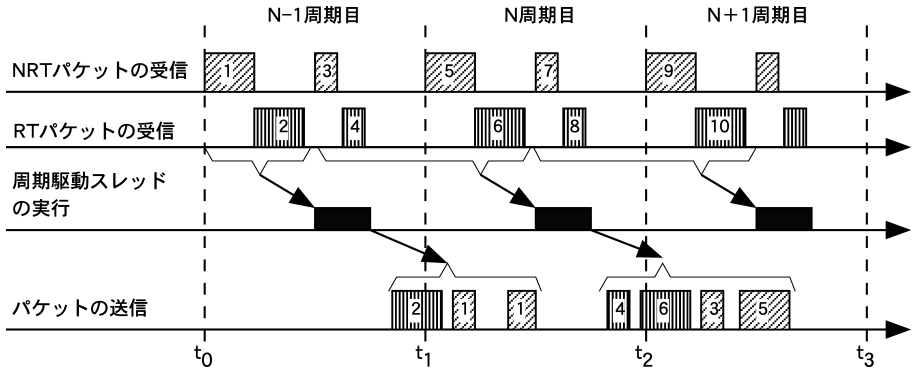

図 7 遅延が駆動周期以内に収まる場合
Fig. 7 Minimum delay case.

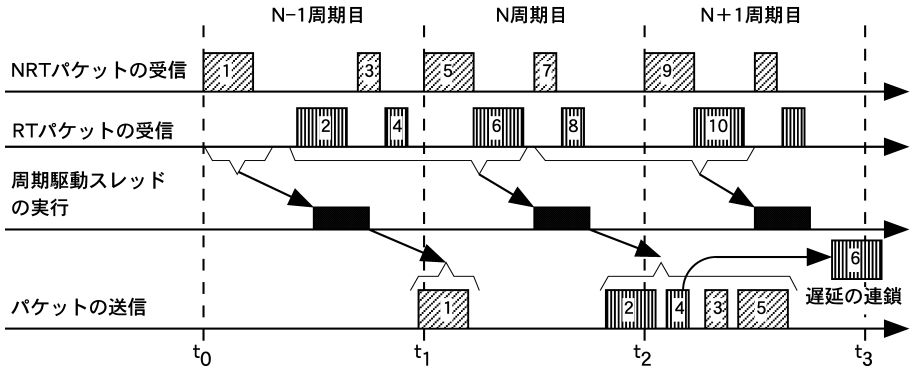

図 8 駆動周期を超える遅延が発生する場合
Fig. 8 Increasing delay case.

周期目のパケット群 2 の受信が，周期駆動スレッドの実行開始までに間に合わなかった場合を示す．この場合，パケット群 2 はパケット群 3〜6 とともに N 周期目でルーティング処理され，ITM スレッドによる送信処理に渡される．ところが，パケット群 2，4，6 のパケット総数は，当該周期内に送信可能な RT パケットの総数を超過している．このため，パケット群 6 の送信は N + 1 周期目へ遅延させられる．帯域保証を実現する TTCP/ITM 機構によって，各ストリームの駆動周期ごとのパケット送信量が厳密に制限されているため，このようにスケジューラ内部での遅延が拡大する．

さらに，いったん生じた転送遅延の拡大幅は連鎖的に（玉突き的に）後続の周期に伝播する．このため，図 8 の N − 1 周期目のような受信遅延が頻発すると，遅延が累積し，受信ノードで観測される再生レートが送信ノードでの再生レートに比べて低下してしまう．この結果，たとえば，送信側で再生時間 60 秒の映像が，受信側での再生には 70 秒を要するといった問題が発生する．

従来，Tactix では，この転送レート低下の問題を回避するために，RT 帯域割当て時にある程度の余裕を持たせるという「経験的な手法」を採用していた．たとえば，転送データの再生レートが 1.0 Mbps の場合，ルータの帯域割当て時に，あらかじめ余裕を見込んで 1.2 Mbps 割り当てておけば，遅延したパケット群を次の周期で送信できる☆．

しかしながら，この「経験的な手法」には，余裕分の帯域が必要となるために，割当て可能なストリーム数が減少する，あるいは，どの程度の余裕帯域を確保すべきかが定量的に予測できないといった欠点があった．

3. 遅延低減方式

本章では，トラフィックの動的な変動やルータ段数が予測できない現実的な環境において，前述の「経験的な手法」に代わる，確実で効率的な遅延低減方式を

☆ 送信ノードからの送信レートは 1.0 Mbps であるから，実際にルータを通過する転送データの平均レートは 1.0 Mbps であり，パケットの遅延が発生しなければ，余裕帯域は使用されずに無駄になるのみである．

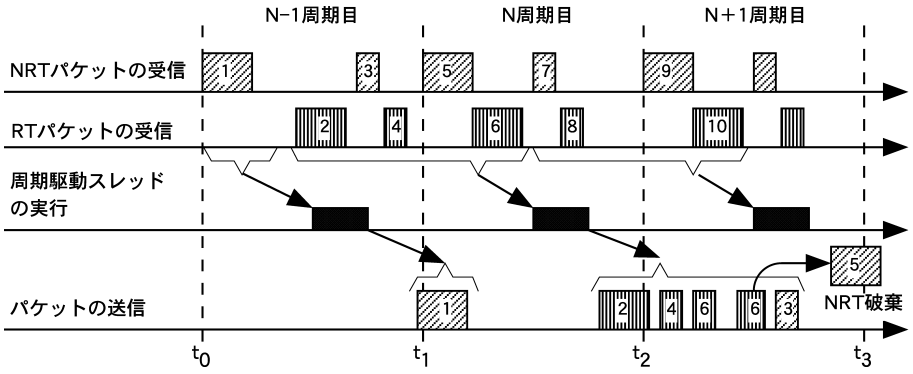

図 9 周回遅れ RT パケットの優先送信
Fig. 9 Delayed packet first.

提案する．具体的には，3.1 節で，2.2.2 項に述べた 2 つの課題，すなわち，

(1) 遅延分布の広がりに起因するスケジューラ内部での遅延拡大
(2) 送信量抑制による遅延の玉突き的な伝播

を解決する方式について述べる．また，3.2 節では，割込み処理と周期駆動スレッドとの同期を実現するイベントボックスについて述べる．

3.1 周回遅れ RT パケットの優先送信

図 8 に示す問題を解決するには，RT パケット群 2，4，6 のすべてを N 周期目に送信できればよい．しかしながら，パケット群 2〜6 のすべてを N 周期目に送信しようとすると，TTCP/ITM によって許容された帯域を超過してしまう．この問題を解決するには，当該 RT ストリームへの帯域割当てを一時的に増加させることが必要である．

これまでに開発したアイソクロナススケジューリングや TTCP/ITM 等の QoS 保証技術は，「リソース使用量の予測が可能な RT 処理に対して優先的にリソースを割り当て，残ったリソースを NRT 処理に割り当てる」という設計方針に基づき，RT 処理の開始を契機として，その完了までの期間にわたり，静的にリソースを割り当てている．一方，周回遅れの RT パケットを優先送信するには，設計方針は共通であるが，受信遅延の発生を契機として，1 駆動周期内に期間を限定する動的なリソース割当てを実現しなければならない．

しかしながら，TTCP/ITM では，各 RT ストリームに割り当てた帯域は，セグメント内の帯域管理サーバ☆によって集中管理されている[3]．また，ルータを多段接続した環境では，RTIPSIG[4] によって，複数のセグメントの各帯域管理サーバにまたがって，固定値の帯域が割り当てられている．このため，一時的に帯域を増加させようとすると，帯域管理サーバや経路上のルータ群との制御通信が必要となり，駆動周期内でこれらの制御通信を完了することができないという問題が生じる．

そこで，余分に RT パケットを送出するために必要な帯域を，自ノードに割り当てられた NRT 帯域を削減することで補う．すなわち，**図 9** に示すように，周回遅れの RT パケット群 2 を含めて送出すべきすべての RT パケット群を N 周期目に送出し，この過剰な RT パケット送出によって不足する帯域を，同じ N 周期目に送出予定であった NRT パケット送出用の帯域を削減することで補う．具体的には，ITM スレッドが各 RT 送信キューからパケットを取り出す際，割当て帯域の超過分を算出し，この超過分に相当するパケット数を NRT 帯域の割当て分から差し引く．図 9 の場合，NRT パケット群 5 を破棄することで，不足する RT 帯域を補っている．破棄された NRT パケット群は，後に TCP 等により再送される．

この周回遅れの RT パケットを優先的に送出する方式によれば，パケットスケジューラ内部における RT パケットの遅延は，図 7 に示すケースと同等な最小値（ほぼ駆動周期と同程度）に収めることができる．同時に，この方式では，周回遅れの影響が玉突き的に後続の周期に伝播するという問題も解消できる．

なお，この方式が適用できるためには，各ノードに十分な NRT 帯域が割り当てられていることが前提となる．また，各ノードに割り当てられた NRT 帯域に余裕がある場合（実際の使用量が少ない場合）には，上述の NRT パケットの破棄も発生せず，周回遅れの RT パケットを優先送信することの実質的なデメリッ

☆ 通常は，各セグメントに接続するルータの 1 つが帯域管理サーバを兼ねる．

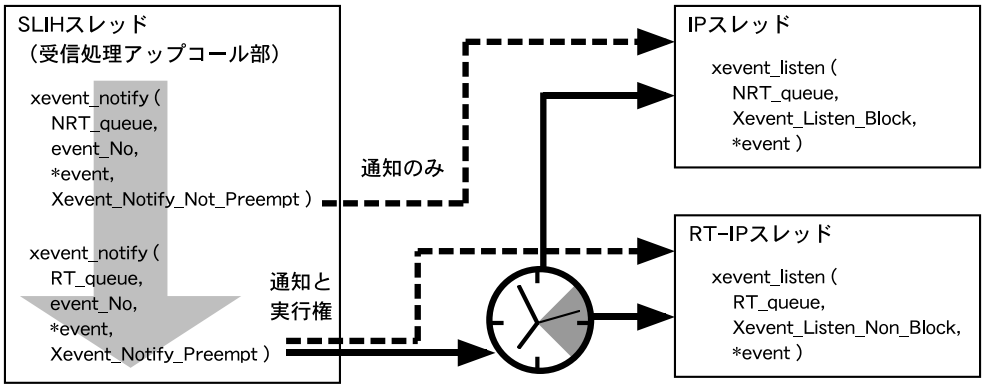

図 10 イベントボックスによるスレッド間同期
Fig. 10 Synchronization using event box.

トは小さいと考えられる．

3.2 割込みイベントと周期駆動の同期

TTCP/ITM を実装したパケットスケジューラ内には，2.2.1 項に述べたように，正確な周期送信を行うための周期駆動スレッドと，Ethernet からの受信割込みに高速応答するためのイベント駆動スレッドが併存する．このため，駆動周期ごとに事前予約したタイムスロット（1 タイムスロット = 1 ミリ秒）精度で厳密に実行される周期駆動スレッドと，駆動周期（= 10 ミリ秒）に近い揺らぎを持つパケット受信イベントの同期を実現するには，以下の制約条件を克服する必要がある．

(1) 周期駆動スレッドは suspend できない：パケットスケジューラの内部処理において，他スレッドからのイベント発生の通知を待つために，周期駆動スレッド（RT-IP スレッドや ITM スレッド）が自身の実行をブロックすると，デッドラインミスの発生に直結し，駆動周期以上の遅延が発生する．周期駆動スレッドは，周期的に CPU 実行権を得た時点で，イベント発生を能動的に確認するノンブロック型で待ち合わせなければならない．

(2) 周期駆動スレッドへの hand-off はできない：受信完了割込みを契機に起動される SLIH スレッドの受信側アップコール処理において，SLIH スレッドからの受信イベント発生の通知は，RT-IP スレッドと IP スレッドの両方になされなければならない．ただし，被通知側スレッドが周期駆動スレッドの場合，イベント発生時に被通知側スレッドが実行可能とは限らないため，RT-IP スレッドと IP スレッドのどちらが高優先順位かは時刻に依存する．このため，通知側から被通知側への CPU 実行権の単純なハンドオフ制御は行えない．

これらの制約条件を満たしつつ，割込みイベントと周期駆動スレッドとの同期問題を解決するために，**図 10** に示すスレッド間同期機能を提供するイベントボックス（E-box）を設計し，スレッド間でのキューを介したバッファ受け渡しに適用する．

(1) E-box を利用してイベント発生を受理するスレッドは，関数 xevent-listen 発行時に Xevent-Listen-Block（ブロック型インタフェース），または，Xevent-Listen-Non-Block（ノンブロック型インタフェース）を選択してイベント発生を待ち合わせることができる．ITM スレッドや RT-IP スレッドは，各実行周期ごとに能動的にイベント発生をチェックすることで，自身の実行ブロックによりデッドラインミスが発生するといった問題を回避できる．

(2) E-box を利用してイベント発生を通知するスレッドは，関数 xevent-notify 発行時に Xevent-Notify-Preempt（通知後に CPU 走行権を他スレッドに明け渡す）か，Xevent-Notify-not-Preempt（通知後も自身が CPU 走行権を継続使用する）かを選択できる．

SLIH は各 Ethernet デバイスごとにスレッドが存在し，個々の SLIH が起動される契機はパケット群の受信時刻に依存し，さらに，個々の SLIH の受信割込み後の走行時間は受信したパケット数に依存する．このため，SLIH がすべての受信パケットを図 4 の NRT 受信キューや RT 受信キューにつなぎ終えた時点で，周期駆動の RT-IP スレッドが実行可能な時刻かどうかは保証されない．よって，SLIH は，このキューイング完了を IP スレッドと RT-IP スレッドの両方に通知してから，自身を割込み待ち状態に遷移させて，CPU

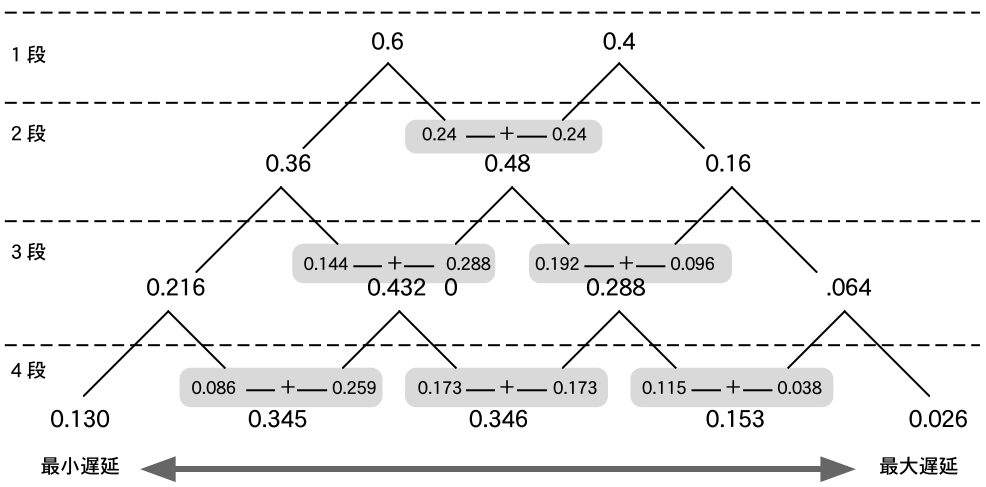

図 11 多段接続時の遅延時間分布予測
Fig. 11 Delay prediction of cascaded routers.

実行権を解放する必要がある☆．すなわち，2 つのスレッドのいずれか一方に最初にキューイング完了を通知した時点では，まだ，2 番目のスレッドへの通知が完了していない．もし，最初の通知を終えた時点で，SLIH が CPU 走行権を解放すると，2 番目の通知は，次に SLIH が起動される契機（すなわち，次の受信割込み発生）まで遅延する．関数 xevent-notify の仕様は，SLIH が最初の通知を終えた時点では CPU 実行権を保持したままとし，2 番目の通知を終える時点で CPU 実行権を解放することを可能としており，この問題を解決する．

4. 遅延低減の効果

本章では，3 章に提案した遅延低減方式の効果を評価する．まず，4.1 節では，遅延発生率を定義し，これを用いたルータ多段接続時の遅延時間予測モデルについて述べる．4.2 節では，ルータ 1 段の遅延発生率を実測した結果について述べる．4.3 節では．Tactix ルータを多段接続したシステムの遅延時間を実測し，前述の予測モデルと比較する．

なお，以下の各節に述べる実測は，2 章に述べた CSMA/CD 方式による遅延時間分布の広がりに起因するルータ内遅延の増大，および，これに対する遅延低減方式の有効性を確認することを目的としている．このため，実測に用いた各システムでは，すべての送受信ノードやルータ間を Half Duplex モードの Shared Hub で接続している．

4.1 優先送信方式の遅延分布予測

各ルータにおいて，受信した RT パケットが駆動周期内に送信される確率を $(1-\beta)$ とすると，次の駆動周期において送信される確率は β となる．以下，β の値を遅延発生率と呼ぶ．簡単のため，経路上の全ルータの β が同一であると仮定すると，N 段のルータ群を最小遅延時間で通過できる確率は $(1-\beta)$ の N 乗，最大遅延時間を要する確率は β の N 乗となる．

また，N 段めのルータで駆動周期内に送信されたパケット群は，次段ルータにおいては，1 周期遅れの駆動周期で送信されるパケット群と同時に送信される．この関係を**図 11** に示す．図 11 は，β の値が 0.4 の場合を示している．1 段目のルータで 60%のパケットが駆動周期内に送信され，残り 40%のパケットは次の駆動周期に送信される．2 段目のルータでは，0.6 の自乗（36%）のパケットのみが最小遅延時間で送信される．0.4 の自乗（16%）のパケットは最大遅延時間で送信され，$0.6 \times 0.4 + 0.4 \times 0.6$（48%）のパケットがその中間の時間で送信される．図 11 に示すように，以下，ルータ段数が増加した場合も同様に予測できる．

遅延発生率 β の値は，ルータの CPU 性能，パケットスケジューラの実装方式，および，セグメント内の総トラフィック等の要因に影響される．このため，ルータを多段接続した系全体の遅延時間分布を予測するには，この β の値の実測が必要である．

4.2 遅延発生率の測定

ここでは前節に述べた遅延発生率 β を測定する．この測定では，**図 12** に示すシステムを使用し，RT-ping

☆ Ethernet デバイスからの割込みを受け付ける SLIH は，RT-IP スレッドや IP スレッドよりも高い優先順位を持っており，SLIH が CPU 実行権を解放しない限り，その他のスレッドは実行できない．

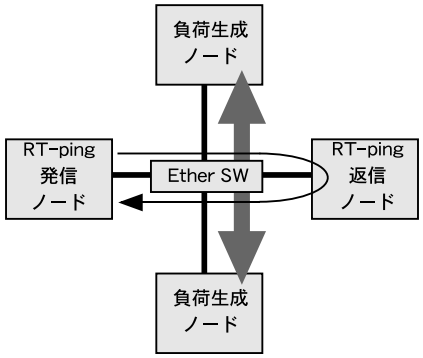

図 12 遅延発生率の実測システム
Fig. 12 Simple delay measurement system.

返信ノードを Tactix ルータとして機能させ，往復通信の応答時間（往復遅延時間）の分布を計測☆1し，その分布パターンからパケットスケジューラ内の遅延発生率 β を求める．

図 12 のシステムは，2 台の負荷生成ノード，RT-ping 発信ノード，RT-ping 返信ノードを，Ethernet スイッチで接続している．Ethernet スイッチは，100 Mbps Half Duplex の Shared Hub モードで動作させる．セグメント内での CSMA/CD 方式による送信遅延増大やパケット廃棄を避けるため，TTCP/ITM による帯域保証機能を全ノードに組み込んでいる．各ノードの CPU はクロック周波数 200 MHz または 300 MHz で，各ノード☆2とも十分な実メモリ（64 MB）を搭載している．

このシステム上で，まず，2 台の負荷生成ノード間で背景トラフィック（NRT トラフィック）を発生させる．この背景トラフィックをパラメータとして変化させながら，RT-ping 発信ノードと RT-ping 返信ノード間で RT パケットの応答時間を計測する．なお，RT-ping 発信ノードは，応答時間の計測に影響を与えないよう，パケット送受信時の時間測定誤差を最小化している．また，RT-ping 返信ノードは，受信した RT パケットを RT-IP スレッドおよび ITM スレッド経由で返信する内部動作となっており，ルータ動作とほぼ等しい．

この測定結果を**図 13** に示す．図 13 の各グラフは，パラメータとして背景トラフィックを変化させた場合の RT-ping の遅延時間☆3の分布を表している．各グラフの横軸は 0.5 ミリ秒きざみの遅延時間を，縦軸はその発生頻度（パケット数）を対数スケールで示している．これらのグラフから，遅延時間分布の第 1 のピークは駆動周期の約 1/2，すなわち，約 5 ミリ秒付近に出現すること，および，駆動周期の 10 ミリ秒付近に遅延時間分布の第 2 のピークが出現することが観測される．また，各ピークの高さは背景トラフィックの影響を受けて変動するが，ピークの位置は変化しないことが分かる．

第 1 のピークと第 2 のピークの高さを比較すると，背景トラフィックが 0 の場合，第 1 のピークに相当するパケット数が全体の 95%以上を占めており，遅延発生率 β は 5%未満である．しかし，背景トラフィックが 50 Mbps になると，第 1 のピークに相当するパケット数は全体の 70%程度に減少し，遅延発生率 β が約 30%となる．

4.3 多段接続時の遅延時間

本節では，多段接続したルータノード群を経由する場合の遅延時間を計測する．評価に用いるシステムは，**図 14** に示す構成を基本とし，1 台から 4 台までのルータノード，2 台の負荷生成ノード，RT-ping 発信ノード，RT-ping 返信ノードを，Ethernet スイッチで接続している．各 Ethernet スイッチは，100 Mbps Half Duplex の Shared Hub モードで動作させる．各 Ethernet セグメント内での CSMA/CD 方式による送信遅延増大やパケット廃棄を避けるため，TTCP/ITM による帯域保証機能を全ノードに組み込んでいる．各ルータノード☆4の CPU はクロック周波数 200 MHz，その他のノード☆5の CPU はクロック周波数 300 MHz で，各ノードとも十分な実メモリ（64 MB）を搭載している．

このシステム上で，まず，RTIP/RTIPSIG [4] を用い，2 台の負荷生成ノード間で，ルータノードを経由する RT トラフィック，および，NRT トラフィックを双方向に発生させる．RTIP/RTIPSIG と TTCP/ITM によって，すべての RT トラフィックの帯域を保証しつつ，NRT トラフィックを増減させる．RT トラフィックと NRT トラフィックの総量，あるいは，両トラフィックの比率をパラメータとして変化させながら，RT-ping 発信ノードと RT-ping 返信ノード間で RT パケットの遅延時間の分布を計測する．

図 15 のグラフ (a) からグラフ (e) は，各ルータ段数

☆1 Tactix ルータの入力側セグメントと出力側セグメントを同一セグメントとし，往復遅延時間を計測することで，ノード間クロック同期の問題を回避し，時間測定の精度を向上させている．
☆2 Pentium 200 MHz を搭載した日立製 FLORA 350 DM3，および，Pentium Pro 300 MHz を搭載した日立製 FLORA TS1．NIC はパケットの一括送受信機能やヘッダギャザ機能を備えた Intel 製 21140 チップを使用．
☆3 片道の遅延時間として，応答時間を 1/2 している．
☆4 Pentium 200 MHz を搭載した日立製 FLORA 350 DM3.
☆5 Pentium Pro 300 MHz を搭載した日立製 FLORA TS1. NIC はパケットの一括送受信機能やヘッダギャザ機能を備えた Intel 製 21140 チップを使用.

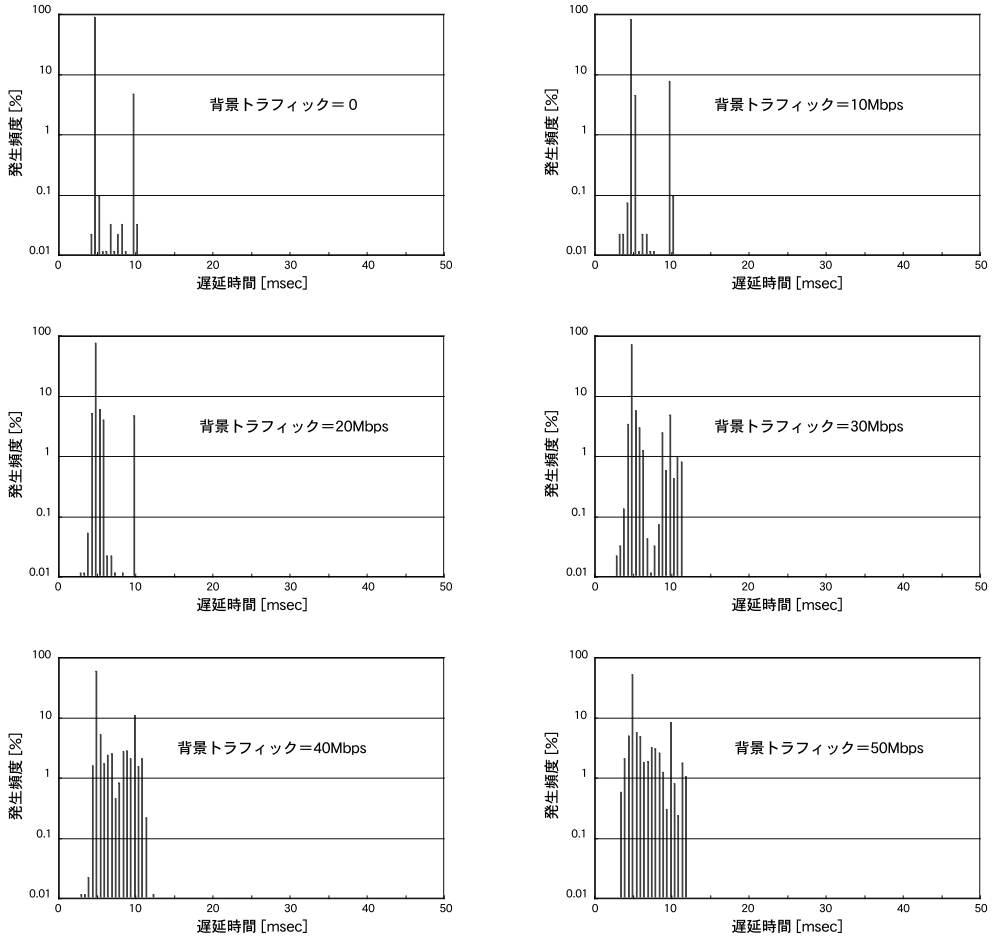

図 13 セグメント内の遅延分布の実測結果
Fig. 13 Delay ratio in a segment.

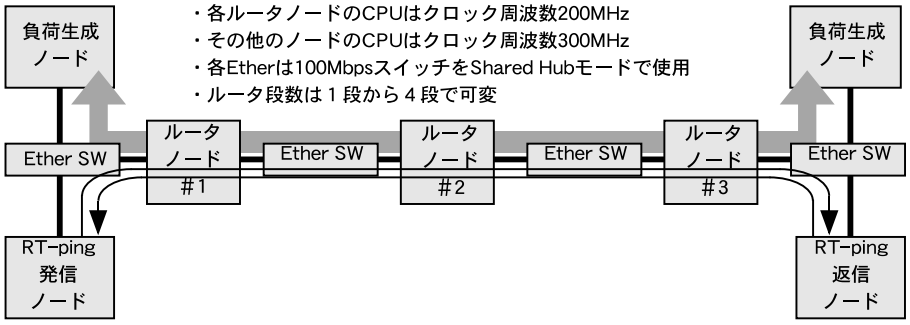

図 14 多段接続時の遅延時間実測システム
Fig. 14 Cascaded delay measurement system.

における往復転送の遅延時間分布の測定結果を示している．この測定結果は，負荷生成ノード間の NRT トラフィックが 20 Mbps，RT トラフィックが 20 Mbps，合計 40 Mbps の背景トラフィック☆を付加した場合のデータである．これらのグラフの横軸はミリ秒単位の遅延時間を，縦軸はその発生頻度をリニアスケールで示している．

グラフ (a) は，ルータを経由しないセグメント内の

☆ NRT，RT ともに，上りと下りで各 10 Mbps である．

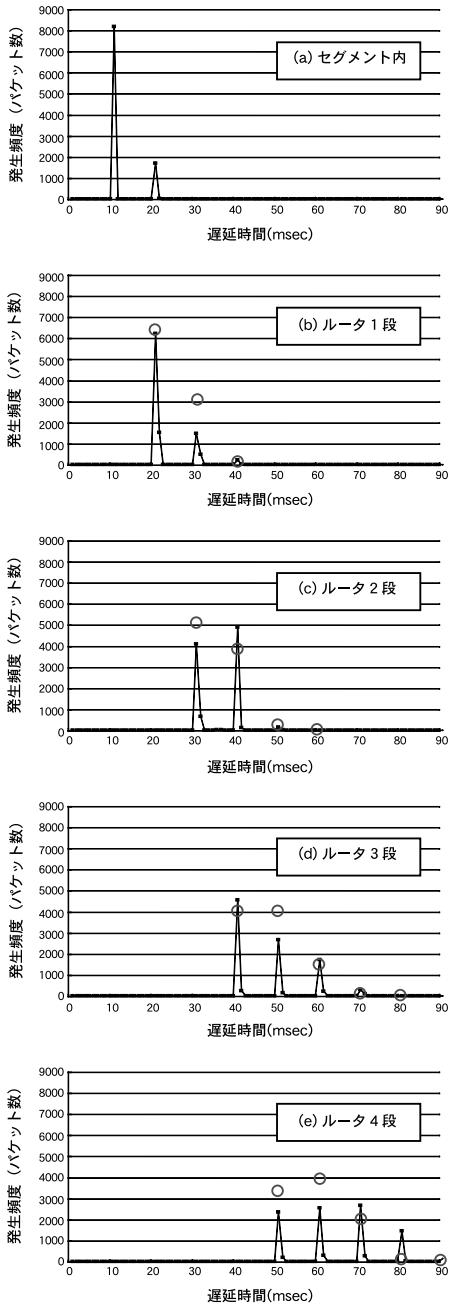

図 15 多段接続時の遅延時間分布
Fig. 15 Delay patterns of cascaded routers.

遅延時間分布☆を表すが，第 1 のピークが駆動周期の 10 ミリ秒の位置に，第 2 のピークが駆動周期の 2 倍の 20 ミリ秒の位置に出現している．両ピークの高さ（パケット数に比例）を比較すると，第 1 のピークが約 80%強の比率を占めており，遅延発生率 β は 20%以下である．

本測定のトラフィック条件（背景トラフィック = 40 Mbps）であれば，各ルータや Ethernet スイッチにおいてパケット破棄はほとんど発生せず，どのセグメントのトラフィックもほぼ同一で，各ルータの CPU 性能も同一であるから，β の値はほぼ等しいと考えられる．グラフ (b) からグラフ (e) に示す丸印は，すべてのルータにおける遅延発生率 β が 20%と仮定した場合の遅延分布の予測結果を表している．この予測結果と実測結果はおおむね一致していることが分かる．

上記の図 13 および図 15 の実測結果から，以下のことが確認できる．

(1) 図 13 から，40 Mbps の背景トラフィックでは，遅延発生率 β が 20%程度に達しているため，駆動周期内に前段ルータからの受信が完了せず，ルータ内で最大 1 周期遅れで転送されているパケットが存在する．図 15 の遅延時間分布の実測結果と $\beta = 20\%$時の予測結果の一致はこれを裏付けている．

(2) 図 15 のグラフの素データとなる RT-ping の応答遅延時間は，0.5 ミリ秒単位でパケット数の分布を採取しているが，このデータを見る限り，片道あたりで，駆動周期 × (ルータ段数 + 1) を超えて遅延している RT パケットは存在しない．

これらの結果から，1 周期分の遅延が発生しても，この遅延が後続周期に玉突き的に伝播していないことが確認できる．すなわち，遅延が発生した次の周期において，当初の予約帯域を超えて，遅延した RT パケット群を優先送信する機構が有効に機能していると判断できる．

なお，図 15 の測定条件においては，以下に述べるとおり，周回遅れの RT パケットを優先送信しても，NRT パケットの廃棄は発生しない可能性が高い．すなわち，物理帯域 100 Mbps から使用禁止帯域 30 Mbps を引いた 70 Mbps がセグメント全体の割当て可能帯域となる．この値から上り 10 Mbps と下り 10 Mbps の RT 帯域を引いた 50 Mbps が，セグメント全体で NRT 帯域として利用できる．よって，ルータ 2 台のみが接続するセグメントの場合，各ノードごとの NRT 送信許可帯域は 25 Mbps となる．また，ルータ 1 台と送受信ノード 2 台が接続したセグメントの場合，各ノードごとの NRT 送信許可帯域は約 16.7 Mbps となる．このため，NRT 背景トラフィックの上りと下り各 10 Mbps では，少なくとも約 6.7 Mbps の余裕（未使用 NRT 帯域）がある．

☆ 図 13 と違い，往復の遅延時間を表している．

5. 関連研究

パケットスケジューリングに関する最近の研究としては，文献 6)〜8) がある．

文献 6) と文献 7) は関連する研究であり，Ethernet 上でのリアルタイム通信を目的として，我々が提案した TTCP/ITM [1),3)] と同様に，各ノードが一定時間内に送出するデータ量をある上限値以内に抑制すれば，セグメント内のパケット衝突確率を低減させて高品質な通信が実現できることを述べている．文献 6) は，セグメント内の通信品質が改善できることを統計的手法で証明しており，また，文献 7) は，カーネルのスケジューリング機構についても論じているが，多段接続したルータ群を経由する場合の遅延時間保証に関してはいっさい論じていない．

文献 8) は，多重チャネル光ネットワーク上のパケットフローの QoS を保証するプリエンプティブなスケジューリングアルゴリズムについて論じている．高品質なパケットスケジューリングを目的として，カーネルのスケジューリングアルゴリズムを幅広く比較検討しているが，詳細な実装や遅延特性についてはまったく論じられていない．

6. まとめ

これまで，帯域保証重視で設計してきた Tactix のパケットスケジューラについて，その遅延発生メカニズムを詳細に分析した．その結果，MAC 層での再送回数増加により，受信タイミングに送信周期と同程度の揺らぎが生じ，これが引き金となってスケジューラ内部での遅延の拡大，および，玉突き的な遅延の後続周期への伝播が発生することを明らかにした．

さらに，これらの遅延を低減させる「周回遅れ RT パケットの優先送信」方式を提案し，ルータを多段接続した Ethernet 環境において，リアルタイムストリームの遅延時間を予測および実測評価した．この結果，数十 Mbps の背景トラフィックが混在する多段接続 Ethernet（100 Mbps Half Duplex モード）環境において，帯域を保証したリアルタイムストリームに対して，パケットロスト率 0 を維持しつつ，ルータ 1 段あたりのパケット転送の最大遅延時間を，送信周期の 2 倍以内に抑えていることを確認した．また，ルータ多段接続環境における系全体でのパケット転送の最大遅延時間は，ルータ段数に 1 を加え，これに送信周期を乗じた値以内に抑えられていることを確認した．

また，本方式の特徴として，ルータ単体におけるパケット転送の遅延発生率 β を測定することにより，ルータ多段接続時における系全体の遅延時間分布を予測できる．この β の値は，各セグメント内における総トラフィックの最大許容値を増減することで制御可能であり，対象アプリケーションが必要とするリアルタイム通信の品質に応じて，遅延時間分布を調整することが可能である．

なお，本論文では，TTCP/ITM による帯域保証技術が確立できている CSMA/CD 方式の Half Duplex Ethernet を対象に検討を行ったが，最近の有線 LAN の主流となっている Full Duplex の Switching Hub を用いた Ethernet，あるいは，CSMA/CA 方式を採用する無線 LAN 等は，それぞれ異なった特性を有しており，これらへの Tactix 技術の適用が今後の課題である．

参考文献

1) Iwasaki, M., Takeuchi, T., Nakahara, M. and Nakano, T.: Isochronous Scheduling and its Application to Traffic Control, *The 19th IEEE Real-Time Systems Symposium* (*RTSS'98*), Madrid, Spain (Dec. 1998).
2) 竹内 理，岩嵜正明，中原雅彦，中野隆裕：連続メディア処理向き OS の周期駆動保証機構の設計と実装，情報処理学会論文誌，Vol.40, No.3, pp.1204–1215 (1999).
3) 中野隆裕，岩嵜正明，中原雅彦，竹内 理：Ethernet 上で QoS を保証する通信方法の設計と実装，情報処理学会論文誌，Vol.41, No.2, pp.322–332 (2000).
4) 竹内 理，岩嵜正明，中原雅彦，中野隆裕：アイソクロナススケジューラを応用した QoS 保証型通信の設計と実装，情報処理学会論文誌，Vol.40, No.10, pp.3737–3751 (1999).
5) 中原雅彦，岩嵜正明，竹内 理，中野隆裕：連続メディア処理向けマイクロカーネルにおける内部排他制御方式，情報処理学会論文誌，Vol.40, No.6, pp.2635–2644 (1999).
6) Kweon, S.-K. and Shi, K.G.: Statistical Real-Time Communication over Ethernet, *IEEE Trans. Parallel and Distributed Systems*, Vol.14, No.3, pp.322–335 (2003).
7) Mehra, A., Indiresan, A. and Shin, K.G.: Structuring Communication Software for Quality of Service Guarantees, *IEEE Trans. Softw. Eng.*, Vol.23, No.10, pp.616–634 (1997).
8) Jackson, L.E. and Rouskas, G.N.: Deterministic Preemptive Scheduling of Real-Time Tasks, *IEEE Computer*, pp.72–79 (May 2002).

(平成 18 年 10 月 10 日受付)
(平成 19 年 3 月 18 日採録)

岩嵜　正明（正会員）

昭和 33 年生．昭和 56 年九州工業大学工学部電子工学科卒業．昭和 58 年九州大学大学院・総合理工学研究科情報システム学修士課程修了．同年（株）日立製作所中央研究所入所，平成 5 年同社システム開発研究所に異動．現在，同所主幹研究員．入社以来，並列推論マシン，メインフレームシステム，超並列スーパコンの OS 研究開発を経て，HiTactix の研究を開始，現在 Linux 関連の研究開発に従事．電子情報通信学会，IEEE 各会員．

竹内　　理（正会員）

昭和 44 年生．平成 4 年東京大学理学部情報科学科卒業．平成 6 年同大学大学院理学系研究科情報科学専攻修士課程修了．同年（株）日立製作所システム開発研究所入社．連続メディア処理向きマイクロカーネルの研究，特にリアルタイムスケジューリング方式，リアルタイム通信方式，異種 OS 共存技術，ストリーミングサービスアーキテクチャ，OS デバッグ方式の研究に従事．

中野　隆裕（正会員）

昭和 44 年生．平成 5 年電気通信大学電気通信学部情報工学科卒業．平成 7 年同大学大学院電気通信学研究科情報工学専攻修士課程修了．同年（株）日立製作所システム開発研究所入社．オペレーティングシステムの研究，特に連続メディア処理向きマイクロカーネルや，ストレージシステム向け組み込みカーネルに関する研究・開発に従事．

中原　雅彦（正会員）

昭和 40 年生．昭和 63 年東京農工大学工学部数理情報工学科卒業．平成 2 年同大学大学院工学研究科修士課程修了．同年（株）日立製作所システム開発研究所入社．入社以来，ワークステーションの性能評価，並列計算機用オペレーティングシステム，連続メディア処理向きマイクロカーネル等の研究・開発を経て，現在は携帯電話向け通信サーバ等の研究・開発に従事．

谷口　秀夫（正会員）

昭和 53 年九州大学工学部電子工学科卒業．昭和 55 年同大学大学院修士課程修了．同年日本電信電話公社電気通信研究所入所．昭和 62 年同所主任研究員．昭和 63 年 NTT データ通信（株）開発本部移籍．平成 4 年同本部主幹技師．平成 5 年九州大学工学部助教授．平成 15 年岡山大学工学部教授．博士（工学）．オペレーティングシステム，実時間処理，分散処理に興味を持つ．著書『オペレーティングシステム』（昭晃堂）等．電子情報通信学会，ACM 各会員．