異種混合ネットワークにおける自律型フロー分散制御方式

林 秀樹[†] 岩田 誠[†,††] 島村和典[†,††]

昨今のインターネットの急激な進展にともない，一部のネットワークノードの高速化が，反して，ネットワーク全体の性能を劣化させるという矛盾した現象が起こる問題が生じている．この原因は，高速ノードから高速バーストトラフィックが転送されると，従来の低速ノードではそのフローを受けきれずに，パケットロスが生じたり，あるいは，ノード自体がオーバーロードしてフリーズしたりして，ネットワーク全体の性能や信頼性を損なってしまうことにある．本論文では，ネットワークの拡充時の柔軟性やロバスト性を高めるために，高速ノードのトラフィックフローをそれと比較して低速な既設ノードで受信する場合でも，パケットロスや処理遅延を低減でき，同時に，複数経路に高速トラフィックを負荷分散する方式を提案する．さらに，本フロー分散制御方式が簡易な仕組みで実現できることを述べ，実用的なノード群を想定した典型的なネットワーク構成において低遅延で高いスループットを発揮できる方式であることをシミュレーションにより示す．

Autonomous Distributed Flow Control Scheme in Heterogeneous Networks

Hideki Hayashi,[†] Makoto Iwata[†,††] and Kazunori Shimamura[†,††]

With drastic growth of the Internet, there interconnect various performance routers or switches (nodes) in current networks. Furthermore, novel high-end nodes are continuously developed by introducing advanced communication technologies and these nodes are increasingly employed for augmenting current networks. However, burst traffic sent from a higher-performance node causes loss of packets at an existing lower-performance node. This implies a dilemma which a newly introduced higher-performance node might not upgrade but degrade total network performance. This paper proposes an autonomous distributed flow control scheme for an augmented network employing a higher-performance node. The scheme reduces both packet loss and latency by distributing the increased traffic to relay networks composed of the existing lower-performance nodes. Simulation results on a typical mixed node network show that the proposed scheme upgrades total network performance at equivalent level to the higher-performance node.

1. はじめに

近年，波長多重・光増幅技術の発展によって，信号の誤りや損失，遅延がきわめて少ない高品質な高速光伝送が可能になり，あらゆる形式の情報を自由に疎通できる情報通信ネットワークの実現が可能になりつつある．

しかし，昨今，このような高速光リンクを収容可能な高速ルータやスイッチ（ノード）を新規にネットワークに導入しても，その性能を十分に発揮できない問題が顕在化している．これは，新設ノードと既設ノードの性能差のため，パケットロスや処理遅延の変動が生じ，これらの既設ノードがネットワークの隘路になるからである．具体的には，TCP/IP の場合，既設ノードの能力を超える広告ウインドウによってフローが制御されると，送信側でのパケット廃棄の検出を通してのみ，経路上のボトルネックを間接的に検出することになる．このため，新旧ノード間の性能差が大きい場合，パケット廃棄検出，再送，ウインドウサイズの縮小を何度も繰り返し，これがフローの遅延や揺らぎを増加させる原因となる．また，フロー制御を行わない UDP/IP の場合，パケットが大量に低速ノードに流入し，最悪の場合，低速ノードがフリーズする可能性もあり，ネットワークの信頼性を大きく損ねてしまう．

このため，通常，トラヒックの増加傾向やその分布傾向を予測して，ネットワーク上に隘路が生じないように，複数の新規高速ノードが設備される．しかし，

† 通信・放送機構
Telecommunications Advancement Organization of Japan
†† 高知工科大学
Kochi University of Technology

インターネット接続事業者内では計画的にインフラ整備を行えても，インターネットの性格上，複数の事業者を含む統合的なネットワークに対して，計画的に新設ノードが設備されるとは限らない．また，トラフィックの分布状況は日々変動するため，これらの変動に応じて設備拡張を緻密に計画することも困難である．このため，現行のインターネットでは，ノード間の性能差を直接的に解消することが原理的に不可能であると考えざるをえない．よって，このような性能差によるパケットロスや遅延変動を可能な限り低減するフロー制御方式が必要になる．

さらに，仮に新旧ノード間で適切にフロー制御できたとしても，既設ノードがネットワーク内で隘路にならないように，トラヒックを負荷分散することが必須になる．フローを複数経路に負荷分散して送出し，経路あたりの負荷を下げる方式も多数提案されている[1),2)]．しかし，パケット単位あるいはパケットを細分化したセル単位で負荷分散する方式はスケジューリングやパケット順序補正を要するため，高速な専用ハードウェアが必要になり非効率的である．簡易な実装を目的として，フロー単位で分散する経路を選択する方式[3)]も提案されているが，フローの総量が大きいと特定のノードが高負荷になる場合があり，細粒度での負荷分散が難しい．

そこで本論文では，より柔軟で信頼性の高いネットワークを構築するために，新旧ノード間の性能差にともなうパケットロスや遅延変動を低減できるフロー制御を行うと同時に，フローをパケット単位で複数経路に自律的に負荷分散することによって，ネットワーク内のスループットを向上できる自律型フロー分散制御方式を提案する．

2. 複数経路フロー分散・集約方式

2.1 要　　件

高速ノード☆を新規に導入した場合に，ボトルネックが生じるネットワーク構成の典型として，**図 1** に示すような，高速ノード N_h 間にそれよりも劣る性能の複数の既設中継ノード（ないしは既設中継ネットワーク）N_e が接続される例を考える．高速ノード間に単一経路しか設定されない場合には，TCP/UDP パケットいずれの場合でも，前述したように，パケット廃棄や処理遅延の変動が増大し，高速ノードの転送性能を活用できない．たとえ，OSPF[4)] 等の経路選択アルゴリズムや IPv6 の経路制御ヘッダを用いて，複数経路のルーティングが選択されても，現状のノード装置には動的に変動するトラフィックに対して効果的に負荷分散する機能が実装されていないため，ネットワーク全体のスループット向上は困難である．

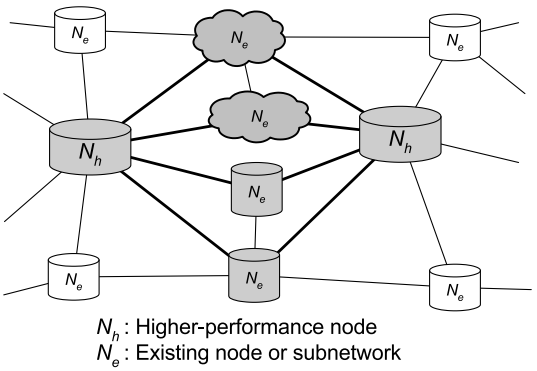

図 1 トラヒック負荷分散を必要とするネットワーク構成例
Fig. 1 An example network requiring traffic distribution.

このようなネットワークでは，経路上の中継ノード N_e のパケット転送能力，経路の伝送遅延は一般に異なっており，かつ，動的に変動する．このような複数の経路上にフローを効果的に負荷分散するために，本論文では以下の点に着目した．

- 既存の負荷分散方式[1)~3)] のような各 N_e への負荷の均一化ではなく，各 N_e に少しでも過度な負荷を与えないようにして，各 N_e の余剰処理能力を適応的に有効活用してフローを分配する．ただし，分配する経路に高速リンクで接続する N_h を含む場合，負荷分散せず当該 N_h のみに中継する．
- パケット単位またはパケットを細分化したセル単位で細粒度の負荷分散制御を行い，総合的なスループットを向上する．
- 複数の高速ノード N_h が同一の中継ノード N_e を負荷分散経路として共有する場合でも，適応的に負荷分散が可能なこと．
- エンドホスト間でパケットの順序逆転を完全に解消するのは一般に困難な問題とされている[5)] ため，集約側で順序補正を行う．ただし，複数経路に分散されたパケットを集約側（受信側）高速ノードで受信した時点でパケットの順序がなるべく正しくなる方式を導入して，受信側でのパケット並べ替えにともなう遅延時間を短縮する．
- 負荷分散・集約処理に要する遅延時間（中継経路

☆ ノードにおけるパケット転送処理は通常の場合パケットヘッダのみが処理対象となる．このためノードの性能は，単位時間あたり転送可能なパケット数 PPS（packets per second）で規定される．一般的には，最小サイズの IP パケットの転送能力が，そのノードが収容可能な伝送リンク速度の総和以下の場合も多く見受けられる．

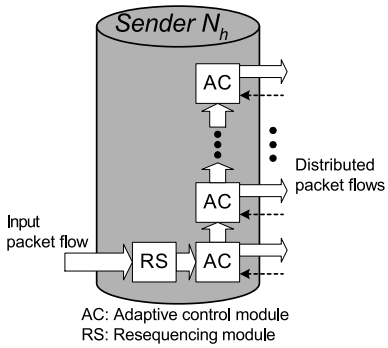

図 2 高速送信側での自律型フロー分散方式
Fig. 2 Autonomous flow distribution scheme.

での転送遅延を除外した時間）を経路に依存せずに均一にし，パケット順序逆転を低減する．

2.2 フロー分散方式

本論文では，**図 2** に示すように，負荷分散先の選択に要する処理の並列化のために，当該経路に送出可能かを適応的に判断する制御部 AC（Adaptive flow control module）を各経路ごとに用意して，以下のようにパイプライン並列に実現可能な方式を採用した．

(1) **送信番号付与部** RS：高速ノードから送出されるパケットはまず，受信側での順序保証のために，送信 N_h ID および送信番号を付与される．送信番号の記録は，IPv4 パケットではヘッダオプションの未使用部（予約クラス）の利用，また IPv6 パケットでは拡張ヘッダの宛先オプションの活用が考えられる．

(2) **適応的フロー制御部** AC：送信 N_h ID と番号が付与されたパケットは最初に，最下部の AC 部に入力され，当該ポートに送出可能かどうかを接続先 N_e の負荷状況に基づいて判断される．個別リンクのフロー制御については次章に詳述する．

(2-1) 送出可能と判断した場合，当該ポートよりパケットが送出される．

(2-2) 当該ポートへは送出すべきでない（接続先 N_e が過負荷である）と判断した場合，パケットは隣接する上部の AC 部へ送られ，再度そのポートへ送出可能かどうか判断される．同時に，後続のパケットは並行して，上記 (2-1) の操作対象となる．

(2-3) 各パケットは，送出可能なポートに到達するまで上部の AC 部へ送られる．

この一連の動作の結果，低速ノードの転送能力に制限されることなく，高速ノードの送出フローは並列に分散されて転送される．すなわち，本方式によって，低速ノードの転送能力に依存せず，代替経路がある限りは高速にフローを転送できる．

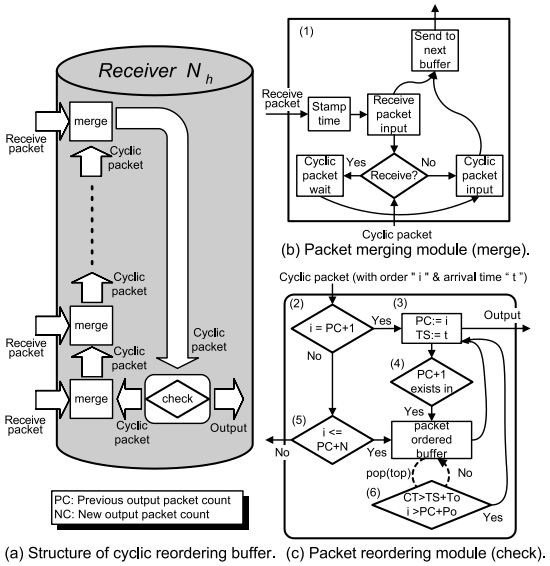

図 3 パケット順序補正が可能な環状型受信バッファ機構
Fig. 3 Cyclic reordering buffer mechanism.

2.3 フロー集約方式

複数の経路を経由して到着したパケットを元のフローに集約する際，到着間隔の揺らぎや到着順序逆転のリスクがある．これらの問題は不均質なネットワーク上で，しかも自律的にパケット単位で負荷分散する場合，可能性を零にはできない．このリスクを可能な限り少なくするには，負荷分散・集約処理に要する遅延時間 T_{da}（中継経路での転送遅延を除外した時間）が経路に依存せず等しいことが望ましい．これは，中継経路の転送遅延時間が一般には既知ではなく，経路ごとに T_{da} を差別化しても一般的な性能向上には寄与しないためである．

送信側高速ノードでは基本的に下部のポートから順にパケットを送出していくため，受信側ノードにおいても下部のポートから順にパケットを受信して整列する方式を採用し，T_{da} を均一にしている．しかし，必ずしも下部のポートから送出順にパケットを受信できるとは限らない．このため，受信バッファを環状型に接続して，複数経路から到着した受信パケットを集約すると同時に，順序逆転したパケットのみ再び巡回して再整列する，**図 3** に示す環状型受信バッファ機構を導入した．

この受信バッファの出力部に順序補正用の検査機構を設け，順序カウンタに従って次に送出すべきパケットか判定する．順序逆転し，先に出力部に到着したパケットは再度受信バッファを巡回する．ただし，出力すべきパケット順が僅差のものは，わずかなタイミング

で順序逆転したことも考えられ，引き続き出力すべきパケットが到着する可能性が高い．このようなパケットを再度巡回させると，フローを一巡分遅延させることになる．したがって，出力すべきパケットの順序に近い順序逆転パケットは巡回させず，出力部のバッファにおいて待機させ，無用な遅延を避けている．この小容量の出力待合せ用バッファを組み合わせれば，効率的なパケットの整列とこれにともなう遅延時間の短縮が同時に実現される．

この環状型受信バッファ機構の動作を図 3 (b)，(c) のフローチャートに従って以下に説明する．

(1) 環状型受信バッファ機構に接続する複数経路から負荷分散されたパケットが到着する．このとき，順序補正待ちのため巡回するパケットと到着パケットが同一合流部（merge）に同時合流する場合，到着パケットを優先させ，巡回パケットを待たせる．また，各パケットには受信時に時刻印 t を付与する．

(2) パケットが巡回して検査部（check）に到着すると，その送信 N_h ID と送信番号 i を，同一の送信 N_h ID から受信したパケットの中で直前に出力されたパケットの送信番号 PC と比較し，PC より 1 だけ大きい場合，すなわち連番であればそのパケットを出力可能と判断する．以降の手順では，送信 N_h ID ごとに各変数を管理する．

(3) パケット出力時に PC の値を i に更新する．さらに，最新出力パケットの受信時刻 TS を t に更新する．TS は手順 (6) のタイムアウト処理で利用する．

(4) さらに検査部はバッファ内に保留していたパケットの送信番号を検索し，新しい PC より 1 だけ大きい番号を持つパケットがバッファ内に存在するか検査する．もし存在すれば，引き続いてそのパケットも出力するために，手順 (3)，(4) を繰り返し，該当するパケットが存在する限りバッファを検索する．

(5) 検査部に到着したパケットの送信番号 i が PC と比較して 2 以上大きい場合，バッファに保留する範囲（バッファレンジ：N）の僅差の順序逆転であるかどうかを判断し，該当するものはバッファに保留し，そうでないものは再度環状バッファを巡回させる．

(6) 分散経路上でのパケットロスおよび受信 N_h を迂回するパケットに対応するためにタイムアウト処理を行う．すなわち，パケットバッファ中で最小の送信番号を持つパケットを周期的に参照し，CT $>$ TS $+$ To，かつ，i $<$ PC $+$ Po であれば，タイムアウトと判定し，そのパケットおよび後続の連番パケットを出力するために，手順 (3)，(4) を繰り返す．ここで，CT は現在時刻，To はタイムアウト時間，Po はタイムアウト時に強制的に出力すべきパケットの送信番号の差である．

以上のフロー分散・集約方式によって，遅延の大きな経路から小さな経路へとスケジューリングすることで，スケジューリング遅延（空いている AC 部に到着するまでの遅延）と経路間の差を相殺することが可能になる．同時に，遅延の小さな経路を通過したパケットが先に受信側に到着する可能性が高くなり，順序補正に要する遅延時間も短縮できる．したがって，多様な性能のノードが混在するネットワークにおけるパケット単位の負荷分散方式の課題である，パケットの到着間隔の揺らぎや順序逆転の問題が軽減される．

さらに，フロー分散機構ならびに集約機構ともにパイプライン並列に実現可能な方式を採用したため，各 AC 部，合流部，検査部は接続リンク速度で動作すれば十分である．従来の負荷分散方式[1),2)] では，送信側ではすべての経路の負荷状況を集約してスケジューリングし，さらに，受信側でもすべての経路からパケットを受けて順序補正する必要があったため，経路数に比例した負荷状況集約速度，順序補正速度が要求される．これに対して，本方式の要求性能は経路数に依存しないため，比較的低速なハードウェアで容易に実現でき，さらに，経路数の増加にも容易に対応できるスケーラブルな方式であるといえる．

3. 個別リンクのフロー制御方式

前章に述べたフロー分散・集約方式を効果的に実現するには，送信側の AC 部で実現すべき個別リンクのフロー制御を，負荷分散を前提とした方式にする必要がある．隣接ノードの情報をフィードバックしてリンクごとのフローを制御する方式として，昨今，AQM（Active Queue Management）が注目されている[6)]．AQM は，Drop-tail 方式を活用し，隣接する送信先の輻輳状況の程度をフィードバックしてフローを制御する方式である．フィードバック情報として，残存キュー量，到着レート，接続経路の容量等を用い，これらをもとに様々な閾値関数を計算する方式が提案されている．しかし，これらはいずれも，単一リンクないしはそれを含む経路のみを対象として，スループット向上，遅延時間短縮，およびパケット廃棄率低減を目指しており，複数経路による負荷分散を前提にはしていない．

負荷分散を前提にした場合，隣接する送信先の入力キューにパケットを滞留させるよりも，分散先が存在する限りは，そのパケットを他の経路に分散するほうが総合的な性能が向上する．したがって，本論文では，個別リンクのフロー制御方式として，低速の中継ノー

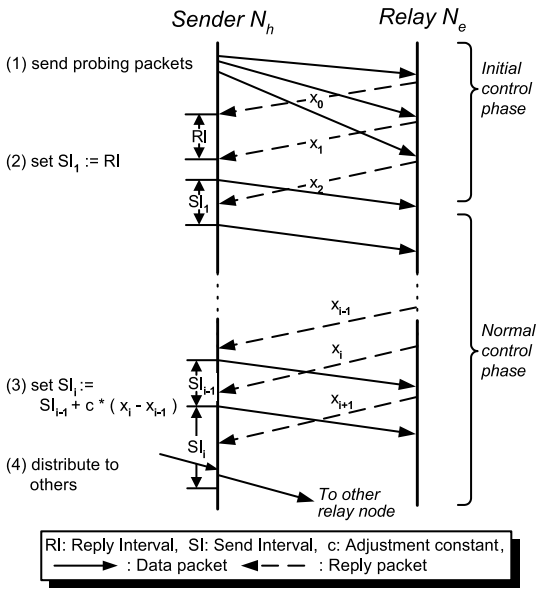

図 4 個別リンクのフロー制御信号シーケンス
Fig. 4 A sequence of link-by-link flow control signals.

ドにパケットをできる限りキューイングしないように，送出レートを制御する方式を導入した．

本方式では，中継ノードからのリプライ情報（リプライ間隔やキューの情報）によって，送信側の高速ノードのフロー送出を自律的に抑制する．これによって，送出側の高速ノードのバーストフローを中継ノードの転送能力のレベルに速やかに抑える．このリプライパケットには，ICMP パケット[7] の Information Reply（type 16：未使用）を利用する等，既存のプロトコルの活用が想定可能である．

本方式では，リプライ情報から求められるキュー量の増減から，接続先ノードの転送能力を推定して，これにパケットの送出間隔 SI を調整している．しかし，リプライパケットを受信できない，初期のパケット送出の場合には，転送能力が未知であるため，中継ノードのキューが溢れる恐れがある．この問題を解消するために，送信ノードと中継ノード間で送受する信号シーケンスを**図 4** のように規定している．

初期制御：まず最初に，リンク先の中継ノードの転送能力を以下のように推定して，SI の初期値を設定する．

(1) 送出側の高速ノードは続けて M 個（$M \geq 2$）のパケットを同じポートへ送出し，2 つ以上のリプライパケットを受ける．

(2) 各リプライパケットの到着間隔 RI を中継ノードの転送能力を反映したフィードバック情報であると仮定し，該当ポートへのパケット送出間隔 SI の初期値として設定する．SI の初期値を設定するまで，リプライパケットを待つ．

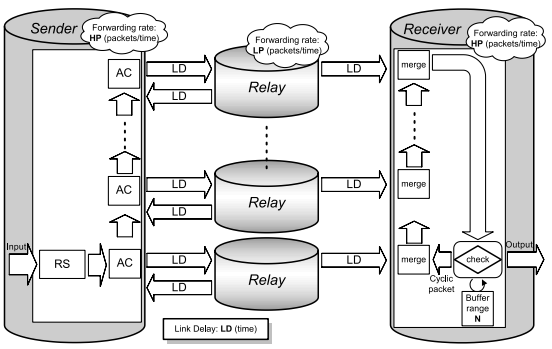

図 5 基本評価モデル
Fig. 5 Basic simulation model.

通常制御：

(3) 中継ノードからのリプライパケットが保持するキュー情報（バッファされているパケット数 x_i）をもとに SI を更新する．すなわち，中継ノードのバッファ内のパケット数の増減に応じて SI も，次式を用いて増減させる．

$$SI_i := SI_{i-1} + c * (x_i - x_{i-1})$$

ただし，c は高速ノードの転送性能に合わせた重み係数である．

(4) SI が初期化されれば，以前にパケットを送出した時間から現在までの時間経過 PI を監視する．次の送出パケットは，$PI \geq SI$ であれば送出される．そうでなければ，別の負荷分散先がある限り，次の AC 部へ転送される．

結果として，SI は中継ノードの転送能力に漸近し，自律的に中継ノードの転送能力に応じたフローレートでパケットの送出を続けることが可能となる．この最初に 2 つ以上のパケットを送出する方法は，Packet-pair による帯域推定の方式[8] を参考にした．Packet-pair 方式は，TCP/IP のコネクションエンドからの Ack パケットの受信間隔情報から，事前にコネクションリンクのボトルネック帯域を推定し，過度のパケット送出を抑制する方式である．本論文では，この手法を応用し，初期制御時にリプライパケットの到着間隔から，中継ノードの転送能力を推定している．

4. 評　　価

本章では，本論文で提案した自律型フロー分散制御方式の有効性をシミュレーションにより評価し，その結果について考察を行う．**図 5** は提案方式に関するシミュレーションの基本モデルを示している．この評価では，20 Bytes の IPv4 パケットの連続フローを想定したトラフィックを用い，定常状態に遷移するまで投

入する．単位時間は，高速ノード（Sender および Receiver）の転送能力にスケールさせ，高速ノードがこのパケットを連続転送可能な時間間隔（最小転送間隔）とし，これを 1 (time)☆とする．高速ノードの AC 部間パケット移動時間と環状型受信バッファ機構内のモジュール間パケット移動時間もここで定義した 1 (time) とする☆☆．この単位時間を用いて高速ノードの転送能力は，HP = 1 (packet/time) と表される．また，中継ノード（Relay）の転送能力を LP (packet/time) とすると，中継ノードの最小転送間隔は，1/LP (time/packet) と表される．以降の評価では，この 1/LP を用いて中継ノードの転送能力を表す．さらに高速ノードと中継ノード間のリンクにともなう遅延を LD (time) としている．本評価では，フロー分散から，フロー集約までの方式評価を行うものとし，高速ノードに内在する遅延は対象外とする．

まず本提案方式が複数経路上の中継ノードの転送能力に応じて，効果的にフローの負荷分散が行えることを確認するために，複数経路上のノードのパケットの転送能力のみを対象とした特性評価を行う．その効果を確認したうえで，現行のノード装置に定常的に加わる負荷による遅延，およびその遅延変動とパケット損失を想定し，実際のネットワークへの適用効果を評価する．

4.1 特性評価

本提案方式に対する特性評価として，高速ノードからの送出フローを複数種の中継ノードに分散させ，受信側の高速ノードで受信して出力されるまでの，ネットワークスループットとネットワークレイテンシの評価を行った．ここで，ネットワークスループットとは，送出側の高速ノードから，受信側の高速ノードへのフローのスループットを指す．ネットワークレイテンシとは同じ高速ノード間のフローのレイテンシを指し，具体的にはフロー分散方式の RS 部の入力から，フロー集約方式の環状型受信バッファ機構の出力までのレイテンシを指す．この評価では，中継ノードの転送能力のみ対象とし，局内接続を想定した最小限のリンク遅延のみ付加している．

特性評価では，高速ノードのフローをもれなく分散して中継できる最小限の並列中継ノード数を用い，ネットワークスループットとネットワークレイテンシを測定している．いずれの測定結果でもネットワークスループットは高速ノードの転送能力（最大スループット）に達することを確認した．よって評価結果は同種/異種ノードの並列構成パターン別の，ネットワークレイテンシ，および最大スループット到達時間に着目して特性評価を行っている．なお，以降の特性評価に用いたリンク遅延はすべて LD=3 (time) としている．

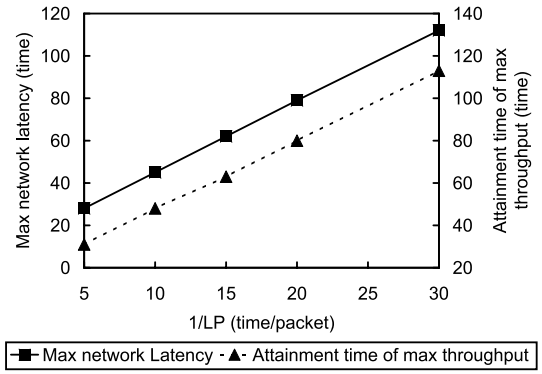

図 6 同種ノードで構成された中継経路の最小転送間隔 1/LP による最大ネットワークレイテンシと最大スループット到達時間の特性

Fig. 6 Max network latency and attainment time of max throughput characteristics for the variety of minimum forwarding interval 1/LP of homogeneous multiple pathways.

まず，全経路の中継ノードの転送能力が同一である複数経路に，フローを分散させた場合の評価を行った．**図 6** は，1/LP がそれぞれ 5，10，15，20，30 (time/packet) の中継ノードを想定し，これらより同一のノードで並列構成する，5 種の複数経路で評価した最大ネットワークレイテンシと最大スループット到達時間の特性である．中継ノードの最小転送間隔に応じて，両特性とも線形増加を示し，この方式が分散構成した中継ノードの 1/LP に応じたスケーラビリティを持つことを示している．

次に，各経路の中継ノードの転送能力が異なる複数経路に，フローを分散させた場合のネットワークレイテンシと最大スループット到達時間の特性を評価した．1/LP が，5，10，15 (time/packet) の 3 種の異種中継ノードを想定し，各経路にいずれかの中継ノードを用いた複数経路を考える．ただし，下部の AC 部に接続する経路から上部の AC 部に接続する経路の方向に，3 種の中継ノードをローテーション順に，各経路へ割り当てる．ローテーションの方法は，順列 ${}_3P_3$=6 より 6 通り想定し，6 種の複数経路を考える．

図 7 は，これら 6 種の複数経路で最大ネットワークレイテンシと最大スループット到達時間の特性評価

☆ 高速ノードが 20 Bytes のパケットを 10 Gbits/s で連続投入できる場合，最小転送間隔は 16 ns となる．

☆☆ パケット転送に関わるノード内部の一部の処理であるため 1 (time) より短縮できると考えられるが，方式のオーバヘッドの最悪条件として 1 (time) を定義する．

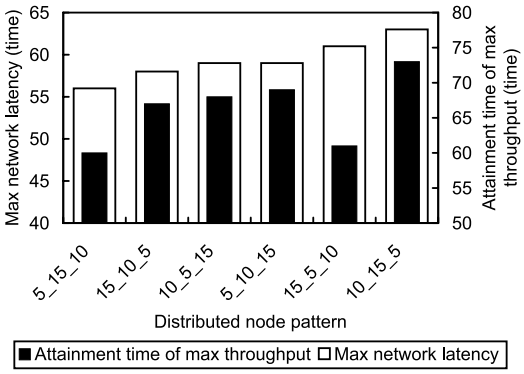

図 7 異種ノードで構成された中継経路へのフロー分散順序による最大ネットワークレイテンシと最大スループット到達時間の特性
Fig. 7 Max network latency and attainment time of max throughput characteristics for the order of flow distributed to heterogeneous multiple pathways.

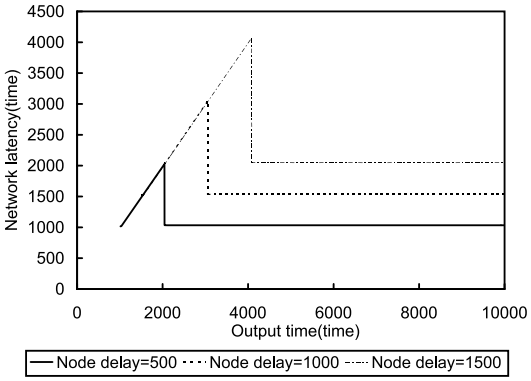

図 8 同種ノードで構成された中継経路へフロー分散したネットワークレイテンシ特性
Fig. 8 Network latency characteristics for the flow distributed to homogeneous multiple pathways.

を行ったものである．この図 7 から，中継ノードの性能が低い経路順にフローが分散するパターン，すなわち 1/LP が … → 15 → 10 → 5 → … の順の複数経路構成の方が，… → 5 → 10 → 15 → … の構成より，最大スループット到達時間が早く，ネットワークレイテンシが小さい傾向が見られた．この理由は，前者の構成では下部 AC 部からの送出レートより，上部 AC 部からの送出レートが速い場合が多く，パケットが下部 AC 部から送出されず，上部 AC 部に転送された場合でも，その上部 AC 部への転送遅延が受信ノード到着時には解消されるためである．逆順になる後者の複数経路構成では，上部 AC 部に転送されたパケットは下部 AC 部から送出されるより遅いレートで送出される場合が多く，このため受信ノードへのパケットの到着間隔が広がる傾向にある．さらに … → 15 → 5 → … のような転送能力の逆転差の大きい経路間で分散送出した連続パケットは，順序逆転することが多く，この順序補正の待合せ遅延の累積が，ネットワークレイテンシを大きくし，最大スループットの到達時間を遅くすると考えられる．15 → 5 → 10 → … の複数経路構成の最大スループット到達時間特性は，上述の傾向と異なる．この理由は，15 → 5 →の転送能力逆転差が最下部の経路にあるため，早期に順序逆転による遅延が確定し，その後，遅延変動が少なくなるためと考えられる．

以上のことから，あらかじめ中継ノードの性能差が既知である場合，本提案方式では，性能の低い順にフローを分散させる方が，フロー分散効率を高めることが分かる．

4.2 典型的なネットワークへの適用例の評価

実際の典型的なネットワーク構成を想定して，本提案方式の実用可能性についても予備的評価を行った．

この評価では，図 5 の基本評価モデルにおいて，単位時間を基準に，送出ノード－中継ノード間，および中継ノード－高速ノード間のリンク遅延をそれぞれ LD=250 (time) とした．また，転送能力の異なる中継ノードを 3 種類想定し，それぞれ 1/LP を，3，5，15 (time/packet) とし，さらにこの中継ノードにそれぞれ 500，1000，1500 (time) の処理遅延（ノード遅延）を加味した．これらのパラメータは現行の数種のギガビットルータ[9)~11)] を参考にして設定している．

また，本評価では実際のノードで生じる遅延変動とパケット損失を想定し，これらの評価条件も設定した．遅延変動とパケット損失は中継ノードで生じるものとし，遅延変動については，1/LP を平均値とする指数分布に従って転送間隔が変動するものとした．さらにその転送間隔の処理タイミングに合わせて，処理遅延もそれぞれ 500，1000，1500 (time) を平均値とする指数分布に従って変動するものとした．パケット損失はランダムに発生するものとし，損失率を 5%とした．

これらの評価条件のもとで，複数経路を 9 経路に固定し，上記 3 種のノードを用いた同種/異種ノードの並列構成によるネットワークレイテンシを測定した．

図 8 は，想定した 3 種の中継ノードを用いて，同一の中継ノードで並列構成する 3 種の複数経路にフロー分散させたときの，ネットワークレイテンシ特性である．いずれの複数経路構成の場合も，ノード遅延にリンク遅延合計 500 (time) を加えた遅延，1000，1500，2000 (time) にネットワークレイテンシが収束している．この結果より，定常状態に遷移した，通常

制御状態では本提案方式の遅延の影響は見られず，効率良く分散されていることが分かる．初期の大きなレイテンシは，初期制御時の SI 設定用のリプライパケット待ち時間のためである．たとえば，ノード遅延が 500 (time) の中継ノードへの初期制御の場合，リプライパケットが到着するまでの待ち時間（往復リンク遅延＋中継ノードの遅延=1000 (time)）に，経路上でのレイテンシ（リンク遅延合計＋中継ノードの遅延=1000 (time)）を加えた遅延が初期制御時の最大遅延となっているのが分かる．

次の，**図 9**，および**図 10** は，前項の特性評価のように，3 種の中継ノードをローテーションして各経路に割り当てた複数経路構成の，ネットワークレイテンシ特性である．特性評価で得られた結果を確認するために，この評価では中継ノードの性能の低い順にローテーションして割り当てた複数経路と，逆順に割り当てた複数経路の 2 種の構成を用いる．パターン A は，性能の低いノード順（ノード遅延順で，1500 → 1000 → 500 → 1500 → ⋯ の順序）に構成した複数経路であり，パターン B はその逆順で構成した複数経路である．この結果は，前項の評価から見られた特性と同様，性能の低いノード順にフローを分散させた方が，ネットワークレイテンシの特性が良いことを示している．

パターン A のネットワークレイテンシでは，初期の特性に初期制御時のリプライパケット待ちの遅延の影響が見られる．図 8 と比較すると，この初期特性はノード遅延が 500 (time) の場合の初期特性と一致しているのが分かる．初期制御時に，SI の初期値を設定するため，リプライパケットを待ち合わせる場合，送信番号の小さなパケットほど上部の AC 部で待ち合わせることになる．パターン A の経路構成の場合，上部 AC 部の方が性能の高い中継ノードに接続される傾向にある．したがって，上部 AC 部から先に SI が設定され，下部で待ち合わせていたパケットも続いて同じ AC 部から送出されていく．このため，順序逆転が起こりにくく，最小のノード遅延の影響しか出ないことになる．初期特性以降のネットワークレイテンシは各経路の異なるノード遅延のため SI の設定時刻に差を生じ，早く SI が設定された経路からパケットが送出されるため，ネットワークレイテンシが階段状に上昇している．

パターン B の場合，最大ノード遅延 1500 (time) とリンク遅延合計 500 (time) に加えて，最大ノード遅延と最小ノード遅延の差（最大ノード遅延ギャップ）1000 (time) が順序補正のための待ち時間として累積されるため，ほぼ 3000 (time) にネットワークレイテンシが収束する．初期の大きなレイテンシは，図 8 のノード遅延が 1500 (time) の場合の特性と一致し，複数経路内で最大ノード遅延を持つ経路の初期制御時のリプライパケット待ち時間の影響であることが分かる．パターン B の場合，初期制御時，送信番号の小さいパケットが上部の性能の低い中継ノードに接続する AC 部で SI 設定待ちとなる傾向にある．このため送信番号の大きなパケットが先に送出され，順序補正の待合せが必要となり，最大ノード遅延の影響が出る．本評価から，異種ノードから構成される複数経路にフロー分散する場合，最悪の場合でも，最大ノード遅延とノード間の最大遅延ギャップを基準にレイテンシが想定可能であることが分かった．

図 11 は，上述のパターン A に対して，遅延変動かパケットロス，およびその両方の影響がある場合の

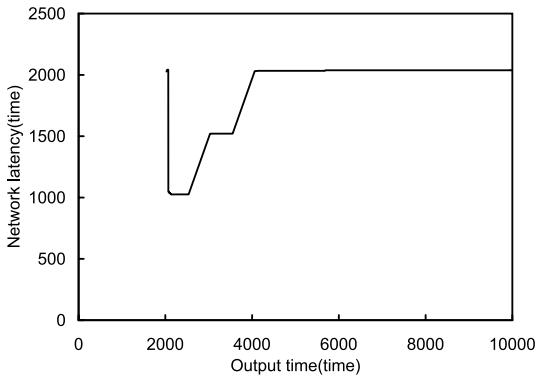

図 9 異種ノードで構成された中継経路へのフロー分散順序パターン A によるネットワークレイテンシ特性
Fig. 9 Network latency characteristics for the flow distributed by pattern A to heterogeneous multiple pathways.

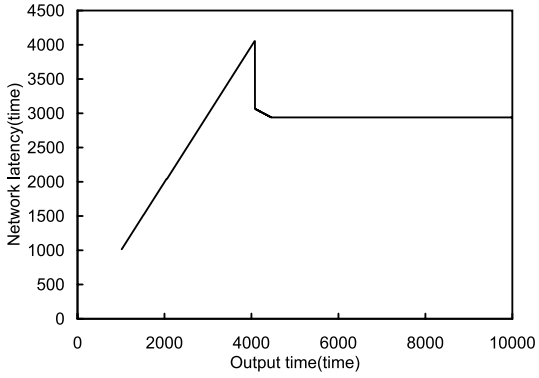

図 10 異種ノードで構成された中継経路へのフロー分散順序パターン B によるネットワークレイテンシ特性
Fig. 10 Network latency characteristics for the flow distributed by pattern B to heterogeneous multiple pathways.

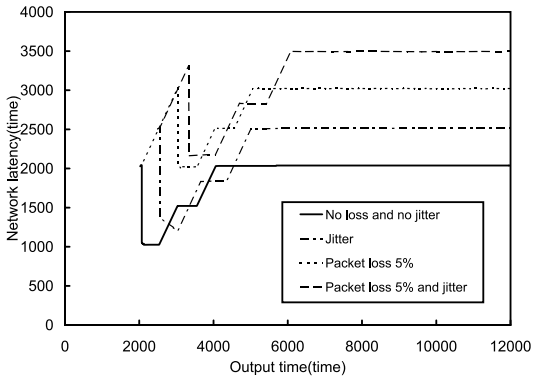

図 11 異種ノードで構成された中継経路へのフロー分散順序パターン A の遅延変動とパケットロスに対するネットワークレイテンシ特性

Fig. 11 Effect of delay jitter and packet loss on network latency characteristics for the flow distributed by pattern A to heterogeneous multiple pathways.

ネットワークレイテンシ特性である．遅延変動のみの場合，ほぼ 500 (time) だけのネットワークレイテンシ増加で収まっている．遅延変動によるレイテンシは，単一リンクのフローであれば，通常はその影響が積算されていくと考えられるが，本評価のように遅延変動の生じるノードの存在するリンクを並列して用いる場合，その影響は相殺され，遅延の増加が 500 (time) 程度で収束することが分かった．パケットロスのみの場合はほぼ 1000 (time) ネットワークレイテンシが増加している．これは，この評価モデルの受信側高速ノード順序補正待合せ時間を 1000 (time) としているためであり，パケットロスの発生にともなった，パケット待合せ時間がネットワークレイテンシに積算されたと考えられる．遅延変動とパケットロスの両方が起こる場合は，これらの積算遅延分である，ほぼ 1500 (time) だけネットワークレイテンシが増加することが分かる．いずれのケースも時間の経過とともにネットワークレイテンシが一定となり，特性が安定することが分かった．

図 12 は，本方式を適用した複数の送信側高速ノードが，一部の中継ノードを共有し，共通の受信側高速ノードへトラフィックを分散させて送出する場合の評価モデルである．この評価でも前述の評価と同様に，中継ノードは 1/LP が 3，5，15 (time/packet)，それぞれの処理遅延が 500，1000，1500 (time) のものを 3 種類を想定し，受信側高速ノードに接続する経路は 9 経路とする．この 9 経路への中継ノードの割り当て方法は，前述のパターン A と同様とする．送信側高速ノード数は 2 とし，それぞれ複数経路を 6 つ持つ．

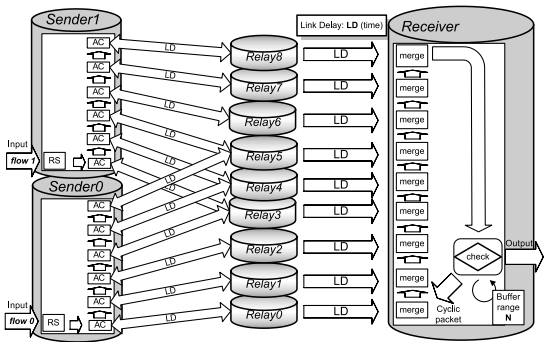

図 12 応用評価モデル

Fig. 12 Applied simulation model.

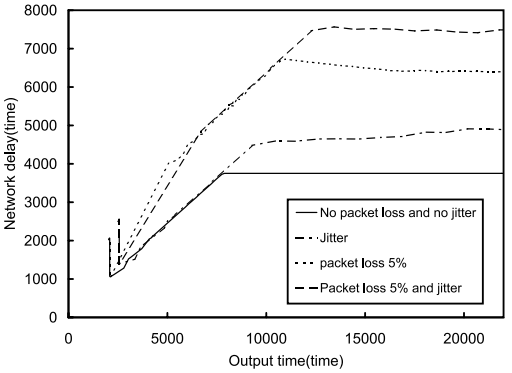

図 13 送信側高速ノードが 2 つの場合のネットワークレイテンシ特性（flow 0））

Fig. 13 Network latency characteristics for the flow 0 in the case of two sender N_h existing.

そのうち送信側高速ノード間で共有する経路数は 3 つとした．また，この評価で用いた順序補正待合せ時間は 3000 (time) とした．

図 13 は，図 12 の評価モデル構成における flow 0 に着目したネットワークレイテンシ特性である．単一の送信側高速ノードを適用した場合の特性である図 11 に比べて，ネットワークレイテンシの増加が見られる．たとえば，パケットロスなし，遅延変動なしの場合を図 13 と図 11 の特性から比較すると，図 13 のネットワークレイテンシは，図 11 の場合の 2 倍近い値を示している．これは，両者の特性評価が同じ中継ノード構成を用いており，図 11 の評価時に単一のフローのみで用いていた帯域を図 13 の評価時には 2 つのフローで共有しているためである．すなわち，フローあたりの可用帯域がほぼ半分であるため，ネットワークレイテンシが 2 倍になっている．遅延変動がある場合の特性も，使用できる帯域が狭いため，大きな遅延のあるリンクを自律的に避けることが難しくなり，この結果，図 11 の評価に比べてネットワークレイテンシ

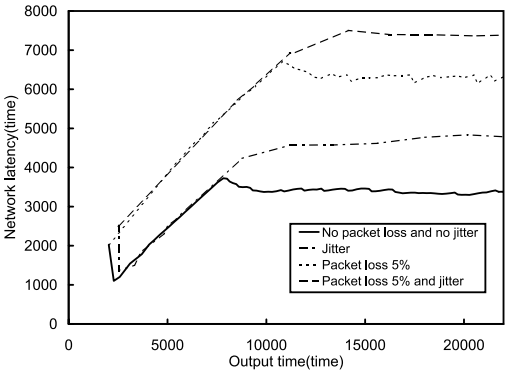

図 14 送信側高速ノードが 2 つの場合のネットワークレイテンシ特性（flow 1）
Fig. 14 Network latency characteristics for the flow 1 in the case of two sender N_h existing.

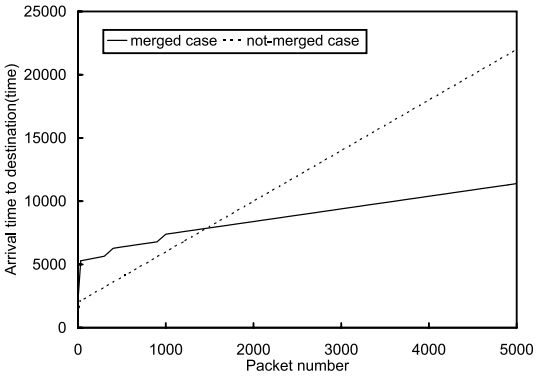

図 15 フロー到着時間比較
Fig. 15 A comparison of arrival time to destination.

が増加している．パケットロスがある場合は，この評価で適用した順序補正待合せ時間 3000 (time) がネットワークレイテンシに積算されていることが分かる．遅延変動とパケットロスのある場合は，これらの遅延積算分がネットワークレイテンシに反映されている．以上のことは，**図 14** の評価結果についても，同様なことがいえる．ただし，図 13 と図 14 を比較すると，遅延変動がない場合に，若干図 14 の方がネットワークレイテンシが少なく，最大値に達した時点で変動が見られる．この原因は，接続構成に起因すると考えられる．入力されたパケットはまず，送出側高速ノードの最下部の AC 部において，出力の可否を判定される．この評価構成では，flow 1 が入力する送出側高速ノード（Sender1）の下部の AC 部は，別の送出側高速ノード（Sender0）の上部 AC 部とも接続される中継ノードに接続している．そのため，出力可否の判定には，Sender0 のトラフィックの影響を受ける．この結果，入力されたパケットが下部の AC 部で出力できるか，上位 AC 部に送られるかが，他のフローの影響を受け変動しやすいのでネットワークレイテンシが若干不安定になると思われる．しかし，このような flow 0 と flow 1 の特性の差異は，遅延変動がある場合には見られない．つまり，遅延変動の影響は，このような接続構成に起因する特性よりも支配的であることが分かる．

これらのネットワークレイテンシ特性は，いずれのケースもほぼ一定値に安定し，その値は想定可能なものである．

次に，送出側高速ノードから複数の中継ノードに分散されたトラフィックが，中継ノードの設定誤り，または本方式の適用誤りによって，受信側高速ノードを経由せず，中継ノードの一部から直接宛先ノードに送られるトラフィックが発生する場合の性能評価を行う．

この場合，該当する中継ノードに到着したパケットは，受信側高速ノードに到達しないためパケットロスと見なされ，順序補正待合せ時間を経過してから，続きの番号を持つパケットが受信ノードから送出される．この順序補正待合せのための時間が，受信側高速ノードで待ち合わせたパケットの遅延時間となる．本方式を適用せず，当該中継ノードのみを用いて，宛先ノードにトラフィックを送出する場合の性能と，本方式を適用し，当該中継ノードのためパケットロスの誤認識が起こる状態で，宛先ノードへのトラフィックを分散・集約させた場合の性能を比較評価する．

この比較評価において，評価モデル図 5 の構成の中継ノードの数，および，転送間隔，処理遅延は上述のパターン A と同様とする．このうち最上部の中継ノード（転送間隔 3 (time)，処理遅延 500 (time)）は，送信側高速ノードから受信したトラフィックを受信側高速ノードに送出せず，直接宛先ノードに送出するものとする．宛先ノードから当該中継ノードまで，また受信側高速ノードまでの伝送遅延はそれぞれ 250 (time) とする．また，受信側高速ノードの順序補正待合せ時間は 1000 (time) とした．

この条件で総計 5,000 個のパケットが宛先ノードに届くまでの時間を比較評価したものが，**図 15** である．この結果より，本方式のパケットロスの誤認識による順序補正待合せ時間の遅延蓄積よりも，中継ノード単体でパケットを送出していくことによる遅延蓄積の方が大きく，このような場合でも本方式の適用効果があることが分かる．

以上の結果より，実ネットワークを想定した，大きな遅延時間差，遅延変動，またパケットロスを与えた

評価の結果からも，本提案方式の有用性と適用の優位性が示すことができたと考える．

5. ま と め

本論文では，ネットワークの拡充時の柔軟性やスケーラビリティを提供するために，高速ノードのトラフィックフローをそれと比較して低速な既設ノードで受信する場合でも，パケットロスや処理遅延を低減でき，同時に，複数経路に高速トラフィックを負荷分散してネットワーク内の多様なノードを活用できる方式を提案した．さらに，実用的なノード群を想定した典型的なネットワーク構成において低遅延で高いスループットを発揮できる方式であることをシミュレーションにより示した．

ネットワークの拡大や発展にともなう，新規ノードの導入や部分的な設備更改の際に，残された既存ノードが結果としてボトルネックとなる問題も，本提案方式の活用によって緩和されることが期待できる．低い性能のノードを活用しつつ，高速ノードの送出する高速フローが導通できるため，設備更改のどの段階においても，ネットワーク内のボトルネックを緩和し，全体の性能を引き上げることが可能となる．さらに本提案方式の特長である自律制御機構では，接続形態に応じた分散スケジューリングの設計が不要である．送出ノードは接続された中継ノードの性能を自律的に知ることにより，適切にフローの分散化を図れる．ネットワークの拡大や発展が継続し，接続形態が固定できない現在および将来のネットワークに対するフロー分散制御方式として，有効であると考える．

本論文では，ノード間のリンク故障等のトラブルによるパケットの欠落に対しては，アプリケーション側が対処するものとして評価した．しかし，フローの分散化に起因する潜在的な遅延や，再送時のフローの不安定化のリスクが懸念されることから，この課題について，本提案方式の中で対処することの有効性を検討中であり，今後の課題として残されている．

謝辞 本研究は，通信・放送機構ギガビットネットワーク研究開発プロジェクトにて行われた．関係諸氏に深く感謝いたします．また，本研究にあたり貴重なご意見ならびにご示唆をいただいた高知工科大学教授寺田浩詔先生に深く感謝いたします．

参 考 文 献

1) Adiseshu, H., Parulkar, G. and Varghese, G.: A Reliable and Scalable Striping Protocol, *Proc. ACM SIGCOMM*, pp.131–141 (1996).
2) Wang, J. and Nahrstedt, K.: Parallel IP Packet Forwarding for Tomorrow's IP Routers, *Proc. IEEE Workshop on High Performance Switching and Routing* (*HPSR'01*) (2001).
3) Jo, J.Y., Kim, Y., Chao, H.J. and Merat, F.: Internet Traffic Load Balancing using Dynamic Hashing with Flow Volume, *Proc. SPIE ITCom* (2002).
4) Moy, J.: OSPF Version 2, RFC2178 (1997).
5) Bennett, J.C.R., Partridge, C. and Shectman, N.: Packet Reordering is Not Pathological Network Behavior, *IEEE/ACM Trans. Networking*, Vol.7, No.6, pp.789–798 (1999).
6) Wydrowski, B. and Zukerman, M.: QoS in Best-Effort Networks, *IEEE Communication Magazine*, Vol.40 No.12, pp.44–49 (2002).
7) Postel, J.: Internet Control Message Protocol, RFC792 (1981).
8) Keshav, S.: A Control-Theoretic Approach to Flow Control, *Proc. ACM SIGCOMM*, pp.3–15 (1991).
9) http://www.juniper.co.jp/solutions/literature/test_report/lightreading03.html#01
10) http://www.hitachi.co.jp/Prod/comp/network/gr2000/
11) http://www.hitachi.co.jp/Prod/comp/network/switch/summit/

(平成 15 年 5 月 16 日受付)
(平成 15 年 12 月 2 日採録)

林 秀樹（正会員）

昭和 63 年大阪大学工学部通信工学科卒業．平成 2 年同大学院工学研究科博士前期課程修了．同年日本テレコム（株）入社．平成 14 年通信・放送機構出向，現在に至る．通信・放送機構高知通信トラヒックリサーチセンター研究員．QoS 技術に関する研究に従事．電子情報通信学会会員．

岩田　誠（正会員）

昭和 61 年大阪大学工学部電子工学科卒業．平成 3 年同大学院工学研究科博士後期課程単位取得退学．同年大阪大学工学部助手，平成 9 年高知工科大学情報システム工学科助教授，平成 13 年同教授，現在に至る．博士（工学）．平成 11 年より通信・放送機構高知通信トラヒックリサーチセンター研究フェロー，平成 14 年より東北大学電気通信研究所 IT21 センター客員助教授兼任．並列ネットワークプロセッサとそのソフトウェア環境の研究に従事．電子情報通信学会，IEEE 各会員．

島村　和典（正会員）

昭和 47 年大阪大学工学部電気工学科卒業．昭和 49 年同大学院工学研究科博士前期課程修了．同年日本電信電話公社（現日本電信電話株式会社）入社．画像通信研究部長，NTT Berhad Director General 等を歴任．平成 10 年より，高知工科大学情報システム工学科教授，現在に至る．博士（工学）．平成 11 年より，通信・放送機構ギガビットネットワーク研究開発プロジェクトサブリーダ兼任．次世代インターネット技術・ディジタル画像通信技術・フルメディア通信アプリケーションの研究に従事．DAVIC Specification 1.0 個人貢献業績賞，画像電子学会論文賞受賞．映像情報メディア学会，電子情報通信学会，画像電子学会，IEEE 各会員．