(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6464911号
(P6464911)
(45)発行日　平成31年2月6日(2019.2.6)　(24)登録日　平成31年1月18日(2019.1.18)

(51)Int.Cl.　　F I
H04L 12/807 (2013.01)　　H04L 12/807
H04L 29/08 (2006.01)　　H04L 13/00　307C

請求項の数 10　(全 32 頁)

(21)出願番号　特願2015-93799(P2015-93799)
(22)出願日　平成27年5月1日(2015.5.1)
(65)公開番号　特開2016-213581(P2016-213581A)
(43)公開日　平成28年12月15日(2016.12.15)
審査請求日　平成30年2月6日(2018.2.6)

(73)特許権者　000005223
富士通株式会社
神奈川県川崎市中原区上小田中４丁目１番１号
(74)代理人　100089118
弁理士　酒井　宏明
(72)発明者　李　忠翰
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内

審査官　森田　充功

最終頁に続く

(54)【発明の名称】情報処理システム、情報処理システムの制御方法及び受信装置

(57)【特許請求の範囲】
【請求項１】
パケットを送信する送信装置と、複数のスイッチ装置と、前記送信装置が送信したパケットを前記複数のスイッチ装置を介して受信する複数の受信装置と、前記送信装置と前記複数の受信装置を制御する制御装置とを有する情報処理システムにおいて、
各受信装置は、
前記送信装置が有する輻輳ウィンドウのサイズの増加を監視する監視部と、
前記送信装置に対する応答パケットを前記送信装置に送信してから、前記応答パケットに対する次のデータを前記送信装置から受信するまでの時間である往復遅延時間を測定する測定部と、
前記監視部が、前記輻輳ウィンドウのサイズの増加が定常状態になったことを監視した場合、パケットを受信するフローのうちボリュームが所定サイズ以上の対象フローの最小スループットを計算する計算部と、
過去に応答パケットに設定した受信ウィンドウのサイズである設定済ウィンドウサイズと、前記最小スループットから計算した受信ウィンドウのサイズである計算ウィンドウサイズとに基づいて、設定用の受信ウィンドウサイズである設定ウィンドウサイズを生成する生成部と、
前記設定ウィンドウサイズが設定された応答パケットを前記送信装置に送信する設定部とを有し、
前記制御装置は、

前記複数の受信装置から受信した平均往復遅延時間に基づいて、前記制御装置からの設定用の受信ウィンドウサイズである制御設定ウィンドウサイズを算出する算出部と、
前記制御設定ウィンドウサイズが設定された制御パケットを前記複数の受信装置の各々に送信する通信部と
を有することを特徴とする情報処理システム。
【請求項２】
前記生成部は、対象フローのスループット及び前記設定済ウィンドウサイズから計算される値と前記設定済ウィンドウサイズとの比率に基づいて、前記設定ウィンドウサイズを生成することを特徴とする請求項１記載の情報処理システム。
【請求項３】
前記複数のスイッチ装置には、
前記送信装置に接続する第１のスイッチ装置と、
前記受信装置に接続する第２のスイッチ装置と、
前記第１のスイッチ装置と前記第２のスイッチ装置とに接続する第３のスイッチ装置と、
前記第１のスイッチ装置と前記第２のスイッチ装置とに接続する第４のスイッチ装置とが含まれることを特徴とする請求項１又は２記載の情報処理システム。
【請求項４】
前記算出部は、前記対象フローが通過するリンクの使用可能バンド幅のうちの最小値に基づいて前記制御設定ウィンドウサイズを算出することを特徴とする請求項１、２又は３記載の情報処理システム。
【請求項５】
前記算出部は、前記使用可能バンド幅が前記最小値であるリンクにおいて前記対象フローにより使用されているバンド幅の総和を前記最小値に加えた値に前記平均往復遅延時間を乗じて前記対象フローの数で割った値を前記制御設定ウィンドウサイズとして算出することを特徴とする請求項４記載の情報処理システム。
【請求項６】
送信装置及び受信装置が仮想スイッチを用いて通信する場合には、前記監視部、前記測定部、前記計算部、前記生成部及び前記設定部は、前記仮想スイッチに含まれることを特徴とする請求項１～５のいずれか１つ記載の情報処理システム。
【請求項７】
前記生成部は、前記最小スループットとパケットのサイズのうち大きな方を前記計算ウィンドウサイズとすることを特徴とする請求項１～６のいずれか１つ記載の情報処理システム。
【請求項８】
前記生成部は、前記対象フローが複数あり、前記比率が０である場合に、前記比率を強制的に０と１の間の数にすることを特徴とする請求項２～７のいずれか１つ記載の情報処理システム。
【請求項９】
パケットを送信する送信装置と、複数のスイッチ装置と、前記送信装置が送信したパケットを前記複数のスイッチ装置を介して受信する複数の受信装置と、前記送信装置と前記複数の受信装置を制御する制御装置とを有する情報処理システムの制御方法において、
各受信装置が、
前記送信装置が有する輻輳ウィンドウのサイズの増加を監視し、
前記送信装置に対する応答パケットを前記送信装置に送信してから、前記応答パケットに対する次のデータを前記送信装置から受信するまでの時間である往復遅延時間を測定し、
前記輻輳ウィンドウのサイズの増加が定常状態になった場合、パケットを受信するフローのうちボリュームが所定サイズ以上の対象フローの最小スループットを計算し、
過去に応答パケットに設定した受信ウィンドウのサイズである設定済ウィンドウサイズ

と、前記最小スループットから計算した受信ウィンドウのサイズである計算ウィンドウサイズとに基づいて、設定用の受信ウィンドウサイズである設定ウィンドウサイズを生成し、
前記設定ウィンドウサイズが設定された応答パケットを前記送信装置に送信し、
前記制御装置が、
前記複数の受信装置から受信した平均往復遅延時間に基づいて、前記制御装置からの設定用の受信ウィンドウサイズである制御設定ウィンドウサイズを算出し、
前記制御設定ウィンドウサイズが設定された制御パケットを前記複数の受信装置の各々に送信する
ことを特徴とする情報処理システムの制御方法。
【請求項１０】
パケットを送信する送信装置と複数のスイッチ装置と前記送信装置を制御する制御装置と他の受信装置とともに情報処理システムを構築し、前記送信装置が送信したパケットを前記複数のスイッチ装置を介して受信し、かつ、前記制御装置により制御される受信装置において、
前記送信装置が有する輻輳ウィンドウのサイズの増加を監視する監視部と、
前記送信装置に対する応答パケットを前記送信装置に送信してから、前記応答パケットに対する次のデータを前記送信装置から受信するまでの時間である往復遅延時間を測定する測定部と、
前記監視部が、前記輻輳ウィンドウのサイズの増加が定常状態になったことを監視した場合、パケットを受信するフローのうちボリュームが所定サイズ以上の対象フローの最小スループットを計算する計算部と、
過去に応答パケットに設定した受信ウィンドウのサイズである設定済ウィンドウサイズと、前記最小スループットから計算した受信ウィンドウのサイズである計算ウィンドウサイズとに基づいて、設定用の受信ウィンドウサイズである設定ウィンドウサイズを生成する生成部と、
前記往復遅延時間の平均値を前記制御装置に送信する送信部と、
前記制御装置が前記送信部が送信した往復遅延時間及び他の受信装置が送信した往復遅延時間に基づいて設定用に算出した受信ウィンドウサイズである制御設定ウィンドウサイズを受信する受信部と、
前記制御設定ウィンドウサイズ又は前記設定ウィンドウサイズが設定された応答パケットを前記送信装置に送信する設定部と
を有することを特徴とする受信装置。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、情報処理システム、情報処理システムの制御方法及び受信装置に関する。
【背景技術】
【０００２】
近年、複数の情報処理装置を複数のスイッチ装置で接続して構築される情報処理システムがデータセンター等で利用されている。図３４は、データセンターにおける情報処理システムの一例を示す図である。図３４において、S_0～S_nは送信側エンドノードを示し、R_0～R_nは受信側エンドノードを示す。ここで、エンドノードは情報処理装置であり、送信側エンドノードはパケットを送信する情報処理装置であり、受信側エンドノードは送信側エンドノードが送信したパケットを受信する情報処理装置である。
【０００３】
ＳＷ＃００～ＳＷ＃０５は、ネットワークを構築するスイッチ装置である。ＳＷ＃００～ＳＷ＃０２は、エンドノードに接続されるスイッチ装置であり、リーフスイッチと呼ばれる。一方、ＳＷ＃０３～ＳＷ＃０５は、リーフスイッチに接続されるスイッチ装置であり、スパインスイッチと呼ばれる。送信側エンドノードが送信したパケットは、リーフス

イッチ及びスパインスイッチを経由して受信側エンドノードに送信される。
【０００４】
図３４に示す情報処理システムでは、輻輳によるパケットの再送等を防ぐために、レート制御（rate control）が行われる。図３５は、レート制御を説明するための図である。レート制御には、受信側エンドノードによるフロー制御（flow control）と送信側エンドノードによる輻輳制御（congestion control）がある。
【０００５】
フロー制御では、受信側エンドノードで受け取ることのできるパケットの数を示すＲＷＩＮ（Receive WINdow）を受信側エンドノードがフロー毎にＡＣＫで明示的に送信側エンドノードに通知する。図３５では、フロー＃１については、受信バッファ８９に２パケット格納されているので、ＲＷＩＮは３パケットである。また、フロー＃２については、受信バッファ８９に１パケット格納されているので、ＲＷＩＮは４パケットである。なお、図３５では、パケットは「ｐｔｋ」と表される。
【０００６】
ネットワークのスループットあるいはＲＴＴ（Round Trip Time：伝送遅延時間）からフロー毎に理論的な最大ＲＷＩＮを計算することが可能である。なお、ここでは、ＲＷＩＮをパケット数で示したが、パケット数にパケットサイズを乗じることでデータ量が得られることから、ＲＷＩＮはバイト数等のデータ量で表される場合もある。
【０００７】
輻輳制御は、送信側エンドノードが輻輳などネットワークの状態をパケットロスやタイムアウトから推定して受信側エンドノードに送るパケットの数を示す輻輳ウィンドウ（congestion window）をフロー毎に制御する。受信側エンドノードは、輻輳ウィンドウが知らせられないので、単位時間当たりのパケット数などで推定する必要がある。
【０００８】
図３５では、送信側エンドノードは、フロー＃１については、輻輳ウィンドウに基づいて２個のパケットを送信し、フロー＃２については、輻輳ウィンドウに基づいて３個のパケットを送信する。
【０００９】
ＴＣＰ（Transmission Control Protocol）を用いた通信では、フロー制御と輻輳制御が同時に働いて送信側エンドノードで送信するパケットの数はＲＷＩＮと輻輳ウィンドウのうちの小さい方のウィンドウで決まる。一般的には、ＲＷＩＮ≧輻輳ウィンドウであるが、もし、ＲＷＩＮ＜輻輳ウィンドウであると、受信側エンドノードによる明示的なレート制御が可能となる。
【００１０】
なお、第１の通信システムを用いた通信中の通信品質の低下に応じて、より低速な第２の通信システムにハンドダウンした場合に、通信相手に通知する受信ウィンドウサイズを縮小することで、スループットを向上する従来技術がある。
【００１１】
また、第１の伝送区間から伝送遅延がより大きい第２の伝送区間にパケットを送信する場合に、第２の伝送区間の往復遅延時間に基づいて最大送信可能データ量を増加又は減少させることで、第２の伝送区間の輻輳による伝送効率の低下を防止する従来技術がある。
【００１２】
また、往復遅延時間、パス最大転送単位、回線速度等のパラメータを収集し、パラメータに基づいてファイル転送の最適ウィンドウサイズを計算することで、スループットを向上する従来技術がある。
【００１３】
また、時間を２つのスロットに分けて、第１スロットではスループットを推定し、推定したスループットと受信側だけで制限を受ける場合の期待スループットとの比率に基づいて第２スロットでフロー毎にレート調整を行う従来技術がある。
【先行技術文献】

【特許文献】
【００１４】
【特許文献１】特開２０１１－１７６５４０号公報
【特許文献２】特開２００３－３２２９５号公報
【特許文献３】特開２００１－１９５３２６号公報
【非特許文献】
【００１５】
【非特許文献１】Haitao Wu, Zhenqian Feng, Chuanxiong Guo, Yongguang Zhang, "ICTCP:Incast Congestion Control for TCP in Data Center Networks," ACM CoNEXT 2010, November 30－December 3 2010, Philadelphia, USA.
【発明の概要】
【発明が解決しようとする課題】
【００１６】
しかしながら、図３４に示した情報処理システムにおいてインキャスト（incast）が検出されると、受信側エンドノード間でエレファント（elephant）フロー間の競合が発生するという問題がある。ここで、インキャストとは、例えば、データセンターにおいて同期する複数の送信側エンドノードが並列に受信側エンドノードにデータを送信した場合に発生する輻輳である。
【００１７】
また、エレファントフローとは、スループットに敏感でライフタイムが長くボリュームが大きいフローである。ライフタイムが短くボリュームが小さいフローはマイス（mice）フローと呼ばれる。マイスフローは、例えば、ボリュームが１Ｍバイト以下、ライフタイムが１０秒以下と定義される。データセンターでは、ボリュームについては９０％がエレファントフローで１０％がマイスフローであり、フローの数については１０％がエレファントフローで９０％がマイスフローである。
【００１８】
図３４では、ｎ＋１台の送信側エンドノードが接続するＳＷ＃００の送信側でインキャストが発生し、ｎ＋１台の受信側エンドノードが接続するＳＷ＃０２の受信側でエレファントフロー間の競合が発生している。
【００１９】
本発明は、１つの側面では、受信側エンドノード間でエレファントフロー間の競合の発生を防ぎ、エレファントフローを効率的に処理することを目的とする。
【課題を解決するための手段】
【００２０】
本願の開示する情報処理システムは、１つの態様において、パケットを送信する送信装置と、複数のスイッチ装置と、前記送信装置が送信したパケットを前記複数のスイッチ装置を介して受信する複数の受信装置とを有する。また、本願の開示する情報処理システムは、前記送信装置と前記複数の受信装置を制御する制御装置とを有する。各受信装置は、監視部と、測定部と、計算部と、生成部と、設定部とを有する。前記監視部は、前記送信装置が有する輻輳ウィンドウのサイズの増加を監視する。前記測定部は、前記送信装置に対する応答パケットを前記送信装置に送信してから、前記応答パケットに対する次のデータを前記送信装置から受信するまでの時間である往復遅延時間を測定する。前記計算部は、前記監視部が、前記輻輳ウィンドウのサイズの増加が定常状態になったことを監視した場合、パケットを受信するフローのうちボリュームが所定サイズ以上の対象フローの最小スループットを計算する。前記生成部は、過去に応答パケットに設定した受信ウィンドウのサイズである設定済ウィンドウサイズと、前記最小スループットから計算した受信ウィンドウのサイズである計算ウィンドウサイズとに基づいて、設定ウィンドウサイズを生成する。前記設定部は、前記設定ウィンドウサイズが設定された応答パケットを前記送信装置に送信する。また、前記制御装置は、算出部と通信部とを有する。前記算出部は、前記複数の受信装置から受信した平均往復遅延時間に基づいて、前記制御装置からの設定用の

受信ウィンドウサイズである制御設定ウィンドウサイズを算出する。前記通信部は、前記制御設定ウィンドウサイズが設定された制御パケットを前記複数の受信装置の各々に送信する。

【発明の効果】

【００２１】

１実施態様によれば、エレファントフローを効率的に処理することができる。

【図面の簡単な説明】

【００２２】

【図１】図１は、実施例１に係る情報処理システムの構成を示す図である。

【図２】図２は、ＲＷＩＮ制御部の実現方法を説明するための図である。

【図３】図３は、通信階層におけるＲＷＩＮ制御部の位置づけを示す図である。

【図４】図４は、ＲＷＩＮ制御部の構成を示す図である。

【図５】図５は、ウィンドウテーブルがエレファントフロー毎に記憶する項目の一例を示す図である。

【図６】図６は、５タプル及びＤＳＣＰの格納場所を示す図である。

【図７】図７は、ＲＴＴの測定範囲を示す図である。

【図８】図８は、輻輳回避状態を説明するための図である。

【図９】図９は、ＲＴＴテーブルがＲＴＴ毎に記憶する項目の一例を示す図である。

【図１０】図１０は、スループット測定スロット及びＲＷＩＮ制御スロットを説明するための図である。

【図１１】図１１は、ＲＷＩＮテーブルがエレファントフロー毎に記憶する項目の一例を示す図である。

【図１２】図１２は、ｓｅｔＲＷＩＮの設定場所を示す図である。

【図１３】図１３は、輻輳ウィンドウの監視処理のフローを示すフローチャートである。

【図１４】図１４は、ＲＴＴの測定処理のフローを示すフローチャートである。

【図１５】図１５は、ＲＷＩＮの設定処理のフローを示すフローチャートである。

【図１６】図１６は、調整レート算出例を示す図である。

【図１７】図１７は、スループット測定スロットの途中にフローが出現した場合の調整レート算出例を示す図である。

【図１８】図１８は、前提条件及び事前準備を説明するための図である。

【図１９】図１９は、レート制御のシーケンスを示す第１の図である。

【図２０】図２０は、レート制御のシーケンスを示す第２の図である。

【図２１】図２１は、実施例１に係るレート制御の効果を示す図である。

【図２２】図２２は、実施例２に係る情報処理システムの構成を示す図である。

【図２３】図２３は、ＣｏＳを説明するための図である。

【図２４】図２４は、実施例２に係るレート制御を説明するための図である。

【図２５】図２５は、ＲＷＩＮ送信用のＰＤＵのフォーマットを示す図である。

【図２６】図２６は、平均ＲＴＴ送信用のＰＤＵのフォーマットを示す図である。

【図２７】図２７は、コントローラＲＷＩＮ制御部の構成を示す図である。

【図２８】図２８は、ＣＴＲＬ＿ＲＷＩＮテーブルがエレファントフロー毎に記憶する項目の一例を示す図である。

【図２９】図２９は、利用可能バンド幅を説明するための図である。

【図３０】図３０は、マトリクスの一例を示す図である。

【図３１】図３１は、コントローラによる処理のフローを示すフローチャートである。

【図３２】図３２は、レート制御のシーケンスを示す図である。

【図３３】図３３は、実施例１及び２に係るＲＷＩＮ制御プログラムを実行するコンピュータの構成を示す図である。

【図３４】図３４は、データセンターにおける情報処理システムの一例を示す図である。

【図３５】図３５は、レート制御を説明するための図である。

【発明を実施するための形態】

【００２３】
以下に、本願の開示する情報処理システム、情報処理システムの制御方法及び受信装置の２つの実施例を図面に基づいて詳細に説明する。実施例１では、１つの受信側エンドノードでのレート制御を説明し、実施例２では、複数の受信側エンドノード間でのレート制御を説明する。なお、これらの実施例は開示の技術を限定するものではない。
【実施例１】
【００２４】
まず、実施例１に係る情報処理システムの構成について説明する。図１は、実施例１に係る情報処理システムの構成を示す図である。図１に示すように、情報処理システム１は、S_0～S_nで表されるｎ＋１台の送信側エンドノード２と、R_0～R_nで表されるｎ＋１台の受信側エンドノード３と、ＳＷ＃００～ＳＷ＃０５で表される６台のスイッチ装置４とを有する。なお、送信側エンドノード２と受信側エンドノード３の数は異なってもよい。また、スイッチ装置４の数は、より多くでも少なくてもよい。
【００２５】
送信側エンドノード２は、情報処理装置であり、スイッチ装置４を介してパケットを受信側エンドノード３に送信する。受信側エンドノード３は、情報処理装置であり、送信側エンドノード２から送信されたパケットをスイッチ装置４を介して受信する。スイッチ装置４は、パケットを中継する装置である。
【００２６】
ＳＷ＃００～ＳＷ＃０２は、リーフスイッチであり、ＳＷ＃０３～ＳＷ＃０５はスパインスイッチである。ＳＷ＃００はS_0～S_nに接続され、ＳＷ＃０２はR_0～R_nに接続される。ＳＷ＃００～ＳＷ＃０２は、それぞれＳＷ＃０３～ＳＷ＃０５に接続される。
【００２７】
受信側エンドノード３は、ＲＷＩＮ制御部５を有する。ＲＷＩＮ制御部５は、各エレファントフローのスループットを定期的に測ってＲＷＩＮを制御してレートを調整する。図２は、ＲＷＩＮ制御部５の実現方法を説明するための図である。図２では、１台の受信側エンドノード３が１台のサーバとして動作する場合と、１台の受信側エンドノード３がｍ台の仮想サーバとして動作する場合のＲＷＩＮ制御部５の実現方法が示されている。なお、図２では、ｍ＝４である。
【００２８】
１台の受信側エンドノード３が１台のサーバとして動作する場合には、ＲＷＩＮ制御部５はＯＳ８（Operating System）の一部としてソフトウェアにより実現される。ＯＳ８は、ＮＩＣ（Network Interface Card）７を用いて送信側エンドノード２等の他の装置と通信を行う。また、送信側エンドノード２等の他の装置で実行されるＡＰＰ（APPlication：アプリケーション）と通信を行うＡＰＰ６がＯＳ８の制御の下で実行される。
【００２９】
１台の受信側エンドノード３がｍ台の仮想サーバとして動作する場合には、ｍ個のＯＳ８と、仮想サーバの通信を制御する仮想スイッチ９が受信側エンドノード３で動作し、ＲＷＩＮ制御部５は仮想スイッチ９の一部としてソフトウェアにより実現される。仮想スイッチ９は、ＮＩＣ７を用いて送信側エンドノード２等の他の装置と通信を行う。
【００３０】
図３は、通信階層におけるＲＷＩＮ制御部５の位置づけを示す図である。図３に示すように、ＲＷＩＮ制御部５は、ＴＣＰ、ＵＤＰ（User Datagram Protocol）が属するトランスポート層（Ｌ４）の上位層において動作する。
【００３１】
次に、ＲＷＩＮ制御部５の構成について説明する。図４は、ＲＷＩＮ制御部５の構成を示す図である。図４に示すように、ＲＷＩＮ制御部５は、ウィンドウテーブル５１と、ウィンドウ監視部５２と、ＲＴＴテーブル５３と、ＲＴＴ測定部５４と、ＲＴＴタイマ５５と、ＡｖｇＲＴＴ計算部５６と、ＮＴタイマ５７と、スループット測定部５８とを有する。また、ＲＷＩＮ制御部５は、スループット計算部５９と、最小スループット計算部６０

と、ＲＷＩＮテーブル６１と、ＲＷＩＮ計算部６２と、ＲＷＩＮ比較部６３と、ＲＷＩＮタイマ６４と、ｓｅｔＲＷＩＮ計算部６５と、ＲＷＩＮ設定部６６とを有する。また、ＲＷＩＮ制御部５は、タイムスロット設定部６７とコントローラ通信部６８とを有する。

【００３２】

ウィンドウテーブル５１は、輻輳ウィンドウの監視に用いられる情報を記憶する。図５は、ウィンドウテーブル５１がエレファントフロー毎に記憶する項目の一例を示す図である。図５に示すように、ウィンドウテーブル５１は、「インデックス」、「５タプル」、「前パケット数」、「現パケット数」、「ＲＴＴ」及び「初期ＡＣＫ」をエレファントフロー毎に記憶する。

【００３３】

「インデックス」は、エレファントフローのインデックスである。「５タプル」は、フローを特定するための情報であり、送信元ＩＰ（Internet Protocol）、送信先ＩＰ、送信元ポート、送信先ポート及びプロトコルである。送信元ＩＰは、パケットの送信元のＩＰアドレスである。送信先ＩＰは、パケットの送信先のＩＰアドレスである。送信元ポートは、パケットの送信元のポート番号である。送信先ポートは、パケットの送信先のポート番号である。プロトコルは、フローの通信プロトコルである。

【００３４】

図６は、５タプル及びＤＳＣＰ（Differentiated Services Code Point）の格納場所を示す図である。図６に示すように、送信元ポート及び送信先ポートはＴＣＰヘッダに含まれる。また、送信元ＩＰ、送信先ＩＰ、プロトコル及びＤＳＣＰはＩＰヘッダに含まれる。なお、ＤＳＣＰは、フローがエレファントフローであるか否かを示す情報の格納場所として使用される。

【００３５】

「前パケット数」は、前回のデータ転送で受信されたパケットの数である。「現パケット数」は、現在のデータ転送で受信されたパケットの数である。「ＲＴＴ」は、ＡＣＫが送信されてから次のデータが受信されるまでの時間である。図７は、ＲＴＴの測定範囲を示す図である。図７に示すように、ＲＴＴは、ｉ番目のデータに対してＡＣＫが送信された時刻から（ｉ＋１）番目のデータが受信された時刻までの時間として測定される。「初期ＡＣＫ」は、ＲＴＴを図るための最初のＡＣＫの時刻である。

【００３６】

図４に戻って、ウィンドウ監視部５２は、エレファントフローの輻輳ウィンドウが輻輳回避状態（Congestion Avoidance）になるまでをウィンドウテーブル５１を用いて監視する。ここで、輻輳回避状態とは、輻輳ウィンドウのサイズが前回から例えば１だけ増加するように緩やかに増加する定常状態である。

【００３７】

図８は、輻輳回避状態を説明するための図である。図８において、横軸はデータの転送回数を示すＲＴＴインデックスであり、縦軸は輻輳ウィンドウをパケット数で示す。図８に示すように、データ転送の初期は、輻輳が発生しないように輻輳ウィンドウが例えば指数関数的に激しく増加する。そして、輻輳ウィンドウがある程度の大きさになると増加が緩やかになる。このような輻輳ウィンドウの増加が緩やかになった定常状態が輻輳回避状態である。受信側エンドノード３は、各フローのスループットを輻輳回避状態で測定する。

【００３８】

ＲＴＴテーブル５３は、ＲＴＴの測定に用いられる情報を記憶する。ＲＷＩＮ制御部５は、ＲＴＴテーブル５３をエレファントフロー毎に有する。図９は、ＲＴＴテーブル５３がＲＴＴ毎に記憶する項目の一例を示す図である。図９に示すように、ＲＴＴテーブル５３は、「インデックス」、「ＲＴＴ」、「ＡＣＫタイム」、「ＡＣＫ番号」及び「失敗数」をＲＴＴ毎に記憶する。

【００３９】

「インデックス」は、ＲＴＴのインデックスである。「ＲＴＴ」は、ＡＣＫが送信され

てから次のデータが受信されるまでの時間である。「ＡＣＫタイム」は、ＡＣＫが送信された時刻である。「ＡＣＫ番号」は、ＡＣＫのシーケンス番号にＭＳＳ（Maximum Segment Size）が加えられた値である。ＭＳＳは、例えば１５００である。「失敗数」は、ＡＣＫ番号とデータのシーケンス番号であるデータ番号との比較で一致しなかった数である。

【００４０】

ＲＴＴ測定部５４は、ＡＣＫが送信されてから次のデータが受信されるまでの時間をＲＴＴテーブル５３を用いてＲＴＴとして測定する。ＲＴＴタイマ５５は、ＲＴＴの平均値を計算するスロットを測るタイマである。ここで、スロットとは一定の大きさの時間区間である。ＲＴＴ測定部５４は、ＲＴＴタイマ５５が０でない間はＲＴＴを測定する。ＡｖｇＲＴＴ計算部５６は、ＲＴＴタイマ５５が０になるとＲＴＴテーブル５３を用いて各エレファントフローのＲＴＴの平均値を計算する。

【００４１】

ＮＴタイマ５７は、スループットを測定するスロットを測るタイマである。スループット測定部５８は、輻輳回避状態になったエレファントフローのスループットをスループット測定スロットで測定する。ここで、スループット測定スロットとは、スループットを測定する時間帯である。ＲＷＩＮ制御部５は、時間をスループット測定スロットとＲＷＩＮ制御スロットに２分する。

【００４２】

図１０は、スループット測定スロット及びＲＷＩＮ制御スロットを説明するための図である。図１０において、横軸のＴｉｍｅは時間を表す。図１０に示すように、時間は一定の時間間隔のタイムスロットに分割され、タイムスロットは交互にスループット測定スロットとＲＷＩＮ制御スロットとなる。すなわち、スループット測定スロットがｅｎａｂｌｅされている場合にはＲＷＩＮ制御スロットがｄｉｓａｂｌｅされており、スループット測定スロットがｄｉｓａｂｌｅされている場合にはＲＷＩＮ制御スロットがｅｎａｂｌｅされている。例えば、タイムスロット（０）はスループット測定スロットであり、タイムスロット（１）はＲＷＩＮ制御スロットであり、タイムスロット（２）はスループット測定スロットであり、タイムスロット（３）はＲＷＩＮ制御スロットである。

【００４３】

スループット計算部５９は、各エレファントフローのスループットを計算する。最小スループット計算部６０は、エレファントフローのスループットの中で最小スループットを計算する。

【００４４】

ＲＷＩＮテーブル６１は、ＲＷＩＮの計算に用いられる情報を記憶する。図１１は、ＲＷＩＮテーブル６１がエレファントフロー毎に記憶する項目の一例を示す図である。図１１に示すように、ＲＷＩＮテーブル６１は、「インデックス」、「バイト数」、「ＲＴＴ」、「スループット」及び「ｃａｌＲＷＩＮ」をエレファントフロー毎に記憶する。

【００４５】

「インデックス」は、エレファントフローのインデックスである。「バイト数」は、スループット測定スロットで測定されたバイト数である。「ＲＴＴ」は、ＡｖｇＲＴＴ計算部５６により計算された平均ＲＴＴである。「スループット」は、エレファントフローのスループットである。「ｃａｌＲＷＩＮ」は、最小スループットから計算されたＲＷＩＮである。

【００４６】

ＲＷＩＮ計算部６２は、最小スループットとＲＴＴからＲＷＩＮを計算する。ＲＷＩＮの単位はバイトである。また、ここで使用されるＲＴＴは、各エレファントフローのＲＴＴを平均した値である。

【００４７】

ＲＷＩＮ比較部６３は、ＲＷＩＮ計算部６２が計算したＲＷＩＮと１パケット分を比較して大きい方の値をｃａｌＲＷＩＮとする。ＲＷＩＮ計算部６２が計算したＲＷＩＮがパ

ケットのサイズより小さい場合には、ＲＷＩＮが１パケット分ない場合であるので、ＲＷＩＮ比較部６３は、１パケットの大きさをｃａｌＲＷＩＮとする。

【００４８】

ＲＷＩＮタイマ６４は、ＲＷＩＮ制御スロットを測るタイマである。ｓｅｔＲＷＩＮ計算部６５は、ＲＷＩＮ制御スロットでｃａｌＲＷＩＮとαから調整レートｓｅｔＲＷＩＮを計算する。具体的には、ｓｅｔＲＷＩＮ計算部６５は、ｓｅｔＲＷＩＮ＝ｃａｌＲＷＩＮ＋ｃａｌＲＷＩＮ×α／２によりｓｅｔＲＷＩＮを計算する。ここで、αは以下の式（１）で定義される。

【数１】

$$\alpha = \frac{\mathrm{MAX}_{i=0,I}\left|f_i - \mathrm{Rate}\right|}{\mathrm{Rate}} \qquad \cdots(1)$$

【００４９】

式（１）で、Ｒａｔｅは１つ前のＲＷＩＮ制御スロットで計算された調整レートであり、f_iはｉ番目のエレファントフローの現在のスループットである。$\mathrm{MAX}_{i=0,I}$は、０番目のエレファントフローからＩ番目のエレファントフローについて、｜f_i－Ｒａｔｅ｜の最大値をとる。Ｉ＋１がエレファントフローの数である。また、f_iには、最小スループットとして選択されたエレファントフローのスループットは含まれない。

【００５０】

式（１）に示すように、αは、「エレファントフローの現在のスループットと１つ前の調整レートの差の最大値」と「１つ前の調整レート」との比率である。ただし、０≦α≦１であり、αが１より大きい場合は１とする。αが１に近い場合は、１つ前の調整レートと現在のエレファントフローのスループットとの差が大きい場合であり、αが０に近い場合は、１つ前の調整レートと現在のエレファントフローのスループットとの差が小さい場合である。また、ｓｅｔＲＷＩＮ計算部６５は、複数のエレファントフローがあって、α＝０である場合には、例えば、強制的にα＝０．１とする。ｓｅｔＲＷＩＮ計算部６５は、強制的にαの値を０．１以外の値にしてもよい。

【００５１】

ＲＷＩＮ設定部６６は、ｓｅｔＲＷＩＮ計算部６５により計算されたｓｅｔＲＷＩＮをＡＣＫに設定する。図１２は、ｓｅｔＲＷＩＮの設定場所を示す図である。図１２に示すように、ｓｅｔＲＷＩＮは、ＴＣＰヘッダの中のＲＷＩＮフィールドに設定される。

【００５２】

タイムスロット設定部６７は、スループット測定スロット、ＲＷＩＮ制御スロット及びＲＴＴスロットの時間間隔をそれぞれＮＴタイマ５７、ＲＷＩＮタイマ６４及びＲＴＴタイマ５５に設定する。時間間隔は、例えば、１ｍｓ（ミリ秒）、１００ｍｓ、ＲＴＴ等である。スループット測定スロットとＲＷＩＮ制御スロットの時間間隔は異なってもよい。

【００５３】

コントローラ通信部６８は、ＲＴＴの平均値をコントローラに送信し、ＲＷＩＮの設定値を受信してＲＷＩＮ設定部６６に渡す。なお、コントローラについては、実施例２で説明する。

【００５４】

次に、輻輳ウィンドウの監視処理のフローについて説明する。図１３は、輻輳ウィンドウの監視処理のフローを示すフローチャートである。図１３に示すように、ウィンドウ監視部５２は、エレファントフローのＳＹＮパケットを受信し（ステップＳ１）、ＳＹＮパケットから５タプルを取り出す（ステップＳ２）。

【００５５】

そして、ウィンドウ監視部５２は、５タプルの情報を含むエレファントフローの情報をウィンドウテーブル５１に記録し（ステップＳ３）、ＡＣＫが送信された時刻をウィンドウテーブル５１に初期ＡＣＫとして記録する（ステップＳ４）。そして、ウィンドウ監視部５２は、パケットを受信したか否かを判定し（ステップＳ５）、受信していない場合に

は、パケットの受信を待つ。

【００５６】

一方、パケットを受信した場合には、ウィンドウ監視部５２は、受信したパケットの５タプルでウィンドウテーブル５１を検索してエレファントフローを特定する（ステップＳ６）。また、ウィンドウ監視部５２は、ＲＴＴフラグがｆａｌｓｅであるか否かを判定する（ステップＳ７）。ここで、ＲＴＴフラグは、ＲＴＴの測定が行われたか否かを示すフラグであり、ｆａｌｓｅは測定が行われていないことを示す。また、ステップＳ７における「＝＝」は、「等しい」を表す記号である。

【００５７】

そして、ＲＴＴフラグがｆａｌｓｅでない場合には、ＲＴＴが測定されているので、ウィンドウ監視部５２は、ステップＳ１０へ進む。一方、ＲＴＴフラグがｆａｌｓｅの場合には、ＲＴＴ測定部５４が、ＡＣＫ←→データ間のＲＴＴを測定し（ステップＳ８）、ＲＴＴフラグをｔｒｕｅに設定する（ステップＳ９）。

【００５８】

そして、ウィンドウ監視部５２は、ステップＳ６で特定したエレファントフローに関するウィンドウテーブル５１の情報に基づいて、パケット受信間隔がＲＴＴより小さいか否かを判定する（ステップＳ１０）。その結果、パケット受信間隔がＲＴＴより小さい場合には、データ受信中なので、ウィンドウ監視部５２は、ウィンドウテーブル５１の現パケット数に１を加え（ステップＳ１１）、ステップＳ５に戻る。

【００５９】

一方、パケット受信間隔がＲＴＴより小さくない場合には、ウィンドウ監視部５２は、ウィンドウテーブル５１の現パケット数が前パケット数より大きいか否かを判定する（ステップＳ１２）。その結果、現パケット数が前パケット数より大きくない場合には、輻輳回避状態ではないので、ウィンドウ監視部５２は、現パケット数を前パケット数とし（ステップＳ１３）、現パケット数を０とする（ステップＳ１４）。そして、ウィンドウ監視部５２は、ステップＳ５に戻る。

【００６０】

一方、現パケット数が前パケット数より大きい場合には、ウィンドウ監視部５２は、前パケット数から現パケット数への増加値が１であるか否かを判定し（ステップＳ１５）、１でない場合には、輻輳回避状態ではないので、ステップＳ１３へ移動する。一方、増加値が１である場合には、ウィンドウ監視部５２は、ウィンドウ数を１増加し（ステップＳ１６）、ウィンドウ数が２より大きいか否かを判定する（ステップＳ１７）。ここで、ウィンドウ数は、前パケット数から現パケット数への増加値が１である回数を表す。すなわち、ウィンドウ数は、ウィンドウの状態が輻輳回避になった回数を表す。

【００６１】

そして、ウィンドウ数が２より大きい場合には、ウィンドウ監視部５２は、輻輳回避状態になったと判定し、処理を終了する。一方、ウィンドウ数が２より大きくない場合には、輻輳回避状態になったとはまだ判定できないので、ウィンドウ監視部５２は、ステップＳ１３へ移動する。

【００６２】

このように、ウィンドウ監視部５２は、ウィンドウテーブル５１を用いて１回のデータ転送におけるパケット数を監視することで、エレファントフローが輻輳回避状態になったか否かを判定することができる。

【００６３】

次に、ＲＴＴの測定処理のフローについて説明する。図１４は、ＲＴＴの測定処理のフローを示すフローチャートである。図１４に示すように、ＡｖｇＲＴＴ計算部５６は、対象のエレファントフローについてウィンドウテーブル５１から５タプルを取得してＲＴＴテーブル５３を作成する（ステップＳ２１）。

【００６４】

そして、ＡｖｇＲＴＴ計算部５６は、ＲＴＴタイマ５５が０でないか否かを判定し（ス

テップＳ２２）、０でない場合には、ＲＴＴを測定する時間帯なので、パケットがＡＣＫであるか否かを判定する（ステップＳ２３）。その結果、パケットがＡＣＫでない場合には、ＡｖｇＲＴＴ計算部５６は、ＡＣＫが来るまで待つ。

【００６５】

一方、パケットがＡＣＫである場合には、ＡｖｇＲＴＴ計算部５６は、ＡＣＫの時刻をＲＴＴテーブル５３のＡＣＫタイムに記録し（ステップＳ２４）、ＡＣＫ番号にＭＳＳを加える（ステップＳ２５）。そして、ＡｖｇＲＴＴ計算部５６は、データパケットを受信したか否かを判定し（ステップＳ２６）、データパケットを受信しない場合には、受信するまで待つ。

【００６６】

一方、データパケットを受信した場合には、ＡｖｇＲＴＴ計算部５６は、ＡＣＫ番号とデータ番号が等しいか否かを判定する（ステップＳ２７）。その結果、等しくない場合には、ＡｖｇＲＴＴ計算部５６は、失敗数に１を加え（ステップＳ２８）、失敗数が３より大きいか否かを判定する（ステップＳ２９）。そして、ＡｖｇＲＴＴ計算部５６は、失敗数が３より大きくない場合には、ステップＳ２６に戻り、失敗数が３より大きい場合には、ステップＳ２２に戻る。

【００６７】

一方、ＡＣＫ番号とデータ番号が等しい場合には、ＲＴＴ測定部５４が、ＡＣＫ時刻とデータパケットを受信した時刻からＲＴＴを測定し（ステップＳ３０）、ＲＴＴをＲＴＴテーブル５３に記録する（ステップＳ３１）。そして、ＡｖｇＲＴＴ計算部５６は、ステップＳ２２に戻る。

【００６８】

一方、ＲＴＴタイマ５５が０である場合には、ＡｖｇＲＴＴ計算部５６は、ＲＴＴテーブル５３に記録された複数のＲＴＴから平均ＲＴＴを計算し（ステップＳ３２）、コントローラ通信部６８がコントローラに平均ＲＴＴをＢＰＤＵ（Bridge Protocol Data Unit）化して送信する（ステップＳ３３）。なお、コントローラについては、実施例２で説明する。

【００６９】

そして、ＡｖｇＲＴＴ計算部５６は、ＲＷＩＮテーブル６１に平均ＲＴＴを記録し（ステップＳ３４）、ＲＴＴテーブル５３をクリアし（ステップＳ３５）、ステップＳ２２に戻る。

【００７０】

このように、ＡｖｇＲＴＴ計算部５６がＲＴＴの平均値を算出してＲＷＩＮテーブル６１に記録することによって、スループット計算部５９はＲＴＴの平均値を用いてエレファントフローのスループットを算出することができる。

【００７１】

次に、ＲＷＩＮの設定処理のフローについて説明する。図１５は、ＲＷＩＮの設定処理のフローを示すフローチャートである。図１５に示すように、ＲＷＩＮ制御部５は、スロットフラグがｔｒｕｅであるか否かを判定する（ステップＳ４１）。ここで、スロットフラグは、スループット測定スロットであるかＲＷＩＮ制御スロットであるかを示すフラグであり、ｔｒｕｅである場合にはスループット測定スロットであることを示し、ｆａｌｓｅである場合にはＲＷＩＮ制御スロットであることを示す。

【００７２】

そして、スロットフラグがｔｒｕｅでない場合には、ＲＷＩＮ制御部５は、ＲＷＩＮタイマ６４が０でないか否かを判定し（ステップＳ４２）、０でない場合には、ＲＷＩＮフラグがｆａｌｓｅであるか否かを判定する（ステップＳ４３）。ここで、ＲＷＩＮフラグは、現在のＲＷＩＮ制御スロットでＲＷＩＮ制御を行ったか否かを示すフラグであり、ｔｒｕｅである場合にはＲＷＩＮ制御を行ったことを示し、ｆａｌｓｅである場合にはＲＷＩＮ制御を行っていないことを示す。そして、ＲＷＩＮフラグがｆａｌｓｅでない場合には、ＲＷＩＮ制御部５は、ステップＳ４２に戻る。

【００７３】
一方、ＲＷＩＮフラグがｆａｌｓｅである場合には、スループット計算部５９が各エレファントフローのスループットを計算し（ステップＳ４４）、ＲＷＩＮ計算部６２が最スループットからＲＷＩＮを計算する（ステップＳ４５）。そして、ＲＷＩＮ比較部６３がｃａｌＲＷＩＮを計算する。そして、ｓｅｔＲＷＩＮ計算部６５がｓｅｔＲＷＩＮを計算し（ステップＳ４６）、ＲＷＩＮ設定部６６が各エレファントフローに対してｓｅｔＲＷＩＮをＡＣＫに設定する（ステップＳ４７）。そして、ＲＷＩＮ制御部５は、ＲＷＩＮフラグをｔｒｕｅに設定し（ステップＳ４８）、ステップＳ４２に戻る。
【００７４】
また、ステップＳ４２においてＲＷＩＮタイマ６４が０である場合には、ＲＷＩＮ制御スロットからスループット測定スロットに変更するために、スロットフラグにｔｒｕｅを設定し（ステップＳ４９）、ステップＳ４１へ戻る。
【００７５】
また、ステップＳ４１においてスロットフラグがｔｒｕｅである場合には、ＲＷＩＮ制御部５は、ＮＴタイマ５７が０でないか否かを判定する（ステップＳ５０）。その結果、０でない場合には、スループット測定部５８がスループットを測定し（ステップＳ５１）、ＲＷＩＮ制御部５は、ステップＳ５０へ戻る。一方、０である場合には、スループット測定スロットからＲＷＩＮ制御スロットに変更するために、ＲＷＩＮ制御部５は、スロットフラグをｆａｌｓｅに設定し（ステップＳ５２）、ステップＳ４１に戻る。
【００７６】
このように、ＲＷＩＮ制御スロットにおいてＲＷＩＮ設定部６６が各エレファントフローに対してｓｅｔＲＷＩＮをＡＣＫに設定することで、ＲＷＩＮ制御部５は各エレファントフローに対してスループットが均等になるようにフロー制御を行うことができる。
【００７７】
次に、調整レート算出例について図１６及び図１７を用いて説明する。図１６は、調整レート算出例を示す図である。図１６において、Ｅ＃０～Ｅ＃２は、輻輳回避状態にあるエレファントフローである。図１６に示すように、タイムスロット（０）では２つのエレファントフローＥ＃０及びＥ＃１があり、タイムスロット（１）でエレファンントフローＥ＃２が出現したとする。
【００７８】
スループット測定スロットであるタイムスロット（０）でスループット測定部５８がスループットを測定し、ＲＷＩＮ制御スロットであるタイムスロット（１）でスループット計算部５９がスループットを計算する。ここでは、スループット計算部５９が計算したＥ＃０及びＥ＃１のスループットがそれぞれ１Ｇｂｐｓ（ギガビット／秒）及び１．２Ｇｂｐｓであったとする。
【００７９】
すると、最小スループットは１Ｇｂｐｓであり、前回の調整レートはないので、αをデフォルト値の０として、ｓｅｔＲＷＩＮ＝ｃａｌＲＷＩＮ＝１Ｇｂｐｓである。ここで、ＲＴＴ＝０．５ｍｓとすると、ｓｅｔＲＷＩＮ＝１Ｇｂｐｓ×０．５×10^{3}ｓ＝５００Ｋビット＝６２．５Ｋバイトとなる。なお、ＲＴＴの値は、Ｅ＃０とＥ＃１のＲＴＴの平均値である。
【００８０】
その後、スループット測定スロットであるタイムスロット（２）でスループット測定部５８がスループットを測定し、ＲＷＩＮ制御スロットであるタイムスロット（３）でスループット計算部５９がスループットを計算する。ここでは、スループット計算部５９が計算したＥ＃０、Ｅ＃１及びＥ＃２のスループットがそれぞれ１Ｇｂｐｓ、１．１Ｇｂｐｓ及び１．２Ｇｂｐｓであったとする。
【００８１】
すると、最小スループットは１Ｇｂｐｓであり、α＝０．２Ｇｂｐｓ／１Ｇｂｐｓ＝０．２であり、ｃａｌＲＷＩＮ＝１Ｇｂｐｓであるので、ｓｅｔＲＷＩＮ＝１Ｇｂｐｓ＋１

Ｇｂｐｓ×０．２／２＝１．１Ｇｂｐｓである。ここで、ＲＴＴ＝０．５ｍｓとすると、ｓｅｔＲＷＩＮ＝１．１Ｇｂｐｓ×０．５×10^3ｓ＝５５０Ｋビット＝６８．７５Ｋバイトとなる。なお、ＲＴＴの値は、Ｅ＃０～Ｅ＃２のＲＴＴの平均値である。このように、調整レートより測定スループットが大きいと、ＲＷＩＮ制御部５は、調整レートを上げる。

【００８２】

なお、タイムスロット（３）でスループット計算部５９が計算したＥ＃０、Ｅ＃１及びＥ＃２のスループットがそれぞれ１Ｇｂｐｓ、１．１Ｇｂｐｓ及び０．８Ｇｂｐｓであったとすると、最小スループットは０．８Ｇｂｐｓであり、α＝０．１Ｇｂｐｓ／１Ｇｂｐｓ＝０．１であり、ｃａｌＲＷＩＮ＝０．８Ｇｂｐｓであるので、ｓｅｔＲＷＩＮ＝０．８Ｇｂｐｓ＋０．８Ｇｂｐｓ×０．１／２＝０．８４Ｇｂｐｓである。ここで、ＲＴＴ＝０．５ｍｓとすると、ｓｅｔＲＷＩＮ＝０．８４Ｇｂｐｓ×０．５×10^3ｓ＝４２０Ｋビット＝５２．５Ｋバイトとなる。この場合、調整レートは変動する。

【００８３】

図１７は、スループット測定スロットの途中にフローが出現した場合の調整レート算出例を示す図である。図１７に示すように、タイムスロット（０）ではエレファントフローＥ＃０があり、タイムスロット（２）でエレファンントフローＥ＃１が出現したとする。

【００８４】

スループット測定スロットであるタイムスロット（０）でスループット測定部５８がスループットを測定し、ＲＷＩＮ制御スロットであるタイムスロット（１）でスループット計算部５９がスループットを計算する。ここでは、スループット計算部５９が計算したＥ＃０のスループットが１Ｇｂｐｓであったとする。

【００８５】

すると、最小スループットは１Ｇｂｐｓであり、前回の調整レートはないので、αをデフォルト値の０として、ｓｅｔＲＷＩＮ＝ｃａｌＲＷＩＮ＝１Ｇｂｐｓである。ここで、ＲＴＴ＝０．５ｍｓとすると、ｓｅｔＲＷＩＮ＝１Ｇｂｐｓ×０．５×10^3ｓ＝５００Ｋビット＝６２．５Ｋバイトとなる。

【００８６】

その後、スループット測定スロットであるタイムスロット（２）でスループット測定部５８がスループットを測定すると、途中でＥ＃１が出現する。そして、ＲＷＩＮ制御スロットであるタイムスロット（３）でスループット計算部５９がスループットを計算する。ここでは、スループット計算部５９が計算したＥ＃０及びＥ＃１のスループットがそれぞれ１Ｇｂｐｓ及び０．６Ｇｂｐｓであったとする。

【００８７】

すると、途中で出現したＥ＃１のスループットは使用せず、最小スループットは１Ｇｂｐｓであり、α＝０であるので、ｓｅｔＲＷＩＮ＝ｃａｌＲＷＩＮ＝１Ｇｂｐｓである。ＲＴＴ＝０．５ｍｓとすると、ｓｅｔＲＷＩＮ＝１Ｇｂｐｓ×０．５×10^3ｓ＝５００Ｋビット＝６２．５Ｋバイトとなる。そして、ＲＷＩＮ制御部５は、途中で出現したＥ＃１のＲＷＩＮも６２．５Ｋバイトに設定する。

【００８８】

そして、ＲＷＩＮ制御スロットであるタイムスロット（５）でスループット計算部５９がスループットを計算する。ここでは、スループット計算部５９が計算したＥ＃０及びＥ＃１のスループットがそれぞれ１Ｇｂｐｓ及び０．８Ｇｂｐｓであったとする。

【００８９】

すると、最小スループットは０．８Ｇｂｐｓであり、α＝０である。ここで、複数のエレファントフローがあり、α＝０であるので、ｓｅｔＲＷＩＮ計算部６５は、強制的にα＝０．１とする。ｓｅｔＲＷＩＮ計算部６５は、強制的にαの値を０．１以外の値にしてもよい。すると、ｓｅｔＲＷＩＮ＝０．８Ｇｂｐｓ＋０．８Ｇｂｐｓ×０．１／２＝０．８４Ｇｂｐｓである。ここで、ＲＴＴ＝０．５ｍｓとすると、ｓｅｔＲＷＩＮ＝０．８４Ｇｂｐｓ×０．５×10^3ｓ＝４２０Ｋビット＝５２．５Ｋバイトとなる。このように、

スループット測定スロットの途中にフローが出現した場合には、途中で出現したフローのスループットは、次の次のＲＷＩＮ制御スロットで使われる。

【００９０】

次に、レート制御のシーケンスについて図１８～図２０を用いて説明する。なお、図１９は図１６に示した例に対応し、図２０は図１７に示した例に対応する。図１８は、前提条件及び事前準備を説明するための図である。図１８において、実線の矢印はエレファントフローのデータを示し、破線の矢印はエレファントフローのＡＣＫを示し、点線の矢印はマイスフローを示す。

【００９１】

図１８に示すように、前提条件として、送信側エンドノードS_0～S_2は、エレファントフローのパケット及びマイスフローのパケットを受信側エンドノードR_0に送信する。そして、ＳＷ＃００でインキャストが発生し、ＳＷ＃０２でエレファントフロー間の競合が発生する。

【００９２】

また、ウィンドウ監視部５２は、エレファントフローの輻輳ウィンドウが輻輳回避状態になることを監視する（ステップｔ１）。そして、ＲＴＴ測定部５４は、ＡＣＫとデータの間のＲＴＴを測定する（ステップｔ２，ステップｔ３）。ＲＴＴの測定は、エレファントフローの通信が終わるまで実施される。

【００９３】

図１９は、レート制御のシーケンスを示す第１の図である。図１９は、ＲＷＩＮ制御スロットでエレファントフローＥ＃２が出現する場合を示す。なお、図１９及び図２０において、T_nはタイムスロット（ｎ）を表し、T_0、T_2及びT_4はスループット測定スロットを表し、T_1、T_3及びT_5はＲＷＩＮ制御スロットを表す。また、実線の矢印はデータパケットを示し、破線の矢印はＲＷＩＮ制御があるＡＣＫを示し、点線の矢印はＲＷＩＮ制御がないＡＣＫを示す。

【００９４】

図１９に示すように、まず、図１８に示した前提条件及び事前準備が整う（ステップｔ１１）。すなわち、インキャストが発生し、エレファントフローの輻輳ウィンドウが輻輳回避状態になる。そして、T_0において、スループット測定部５８がエレファントフローＥ＃０及びＥ＃１のスループットを測定する（ステップｔ１２）。

【００９５】

そして、T_1において、ｓｅｔＲＷＩＮ計算部６５がｓｅｔＲＷＩＮ（６２．５ＫＢ）を計算し、ＲＷＩＮ設定部６６がＥ＃０とＥ＃１のＡＣＫにｓｅｔＲＷＩＮを設定して送信する（ステップｔ１３）。ｓｅｔＲＷＩＮの計算には、Ｅ＃０及びＥ＃１のスループットとしてそれぞれ１Ｇｂｐｓ及び１．１Ｇｂｐｓが用いられる。また、T_1において、Ｅ＃２が輻輳回避状態になって出現する（ステップｔ１４）。ただし、Ｅ＃１のＡＣＫにはｓｅｔＲＷＩＮは設定されない。

【００９６】

そして、T_2において、スループット測定部５８がエレファントフローＥ＃０、Ｅ＃１及びＥ＃２のスループットを測定する（ステップｔ１５）。そして、T_3において、ｓｅｔＲＷＩＮ計算部６５がｓｅｔＲＷＩＮ（６８．７５ＫＢ）を計算し、ＲＷＩＮ設定部６６がＥ＃０とＥ＃１とＥ＃２のＡＣＫにｓｅｔＲＷＩＮを設定して送信する（ステップｔ１６）。ｓｅｔＲＷＩＮの計算には、Ｅ＃０、Ｅ＃１及びＥ＃２のスループットとしてそれぞれ１Ｇｂｐｓ、１．１Ｇｂｐｓ及び１．２Ｇｂｐｓが用いられる。

【００９７】

なお、ｓｅｔＲＷＩＮの計算に、Ｅ＃０、Ｅ＃１及びＥ＃２のスループットとしてそれぞれ１Ｇｂｐｓ、１．１Ｇｂｐｓ及び０．８Ｇｂｐｓが用いられた場合には、ｓｅｔＲＷＩＮの値は５２．５ＫＢとなる。

【００９８】

図２０は、レート制御のシーケンスを示す第２の図である。図２０は、スループットの

測定途中でエレファントフローが出現する場合を示す。図２０に示すように、まず、図１８に示した前提条件及び事前準備が整う（ステップｔ２１）。そして、T_0において、スループット測定部５８がエレファントフローＥ＃０のスループットを測定する（ステップｔ２２）。

【００９９】

そして、T_1において、ｓｅｔＲＷＩＮ計算部６５がｓｅｔＲＷＩＮ（６２．５ＫＢ）を計算し、ＲＷＩＮ設定部６６がＥ＃０のＡＣＫにｓｅｔＲＷＩＮを設定して送信する（ステップｔ２３）。ｓｅｔＲＷＩＮの計算には、Ｅ＃０のスループットとして１Ｇｂｐｓが用いられる。

【０１００】

そして、T_2において、スループット測定部５８がエレファントフローＥ＃０のスループットを測定する（ステップｔ２４）。また、T_2において、エレファントフローＥ＃１が輻輳回避状態になって出現する（ステップｔ２５）。そして、T_3において、ｓｅｔＲＷＩＮ計算部６５がｓｅｔＲＷＩＮ（６２．５ＫＢ）を計算し、ＲＷＩＮ設定部６６がＥ＃０のＡＣＫにｓｅｔＲＷＩＮを設定して送信する（ステップｔ２６）。ｓｅｔＲＷＩＮの計算には、Ｅ＃０のスループットとして１Ｇｂｐｓが用いられる。ただし、ｓｅｔＲＷＩＮの計算にはＥ＃１のスループットは用いられず、Ｅ＃１のＡＣＫにはｓｅｔＲＷＩＮは設定されない。

【０１０１】

そして、T_4において、スループット測定部５８がエレファントフローＥ＃０及びＥ＃１のスループットを測定する（ステップｔ２７）。そして、T_5において、ｓｅｔＲＷＩＮ計算部６５がｓｅｔＲＷＩＮ（５２．５ＫＢ）を計算し、ＲＷＩＮ設定部６６がＥ＃０とＥ＃１のＡＣＫにｓｅｔＲＷＩＮを設定して送信する（ステップｔ２８）。ｓｅｔＲＷＩＮの計算には、Ｅ＃０及びＥ＃１のスループットとしてそれぞれ１Ｇｂｐｓ及び０．８Ｇｂｐｓが用いられる。また、ｓｅｔＲＷＩＮを計算する際にα＝０となるが、複数のエレファントフローが存在するため、ｓｅｔＲＷＩＮ計算部６５は強制的にα＝０．１として、ｓｅｔＲＷＩＮを少し大きい数値に設定する。

【０１０２】

図２１は、実施例１に係るレート制御の効果を示す図である。図２１は、７本の２５．６ＭＢのエレファントフローと５０本の２５６ＫＢのマイスフローを用いて１０Ｇのリンクでインキャストを発生させ、各エレファントフローのスループットを従来技術Ａ～Ｄと実施例１に係るレート制御で比較したものである。縦軸はスループットであり、横軸はエレファントフローのインデックスを示すフローインデックスである。図２１に示すように、実施例１に係るレート制御は、従来技術と比較して、スループットが高い値で平均化されている。

【０１０３】

上述してきたように、実施例１では、最小スループット計算部６０が、輻輳回避状態にある複数のエレファントフローのスループットから最小のスループットを計算する。また、ＲＷＩＮ計算部６２が、最小のスループットに基づいてＲＷＩＮを計算し、ＲＷＩＮ比較部６３がｃａｌＲＷＩＮを計算する。そして、ｓｅｔＲＷＩＮ計算部６５は、前回の調整レートと現在のスループットの差の絶対値を最小スループット以外の全てのエレファントフローについて計算し、最も大きな値と前回の調整レートに基づいてαを計算する。そして、ｓｅｔＲＷＩＮ計算部６５は、ｃａｌＲＷＩＮとαに基づいてｓｅｔＲＷＩＮを計算し、ＲＷＩＮ設定部６６がｓｅｔＲＷＩＮを各エレファントフローのＡＣＫに設定する。したがって、ＲＷＩＮ制御部５は、競合するエレファントフロー間でスループットを平均化することができる。

【０１０４】

また、実施例１では、ＲＷＩＮ制御部５は、仮想スイッチ９が存在する場合には仮想スイッチの一部として実現され、仮想スイッチ９が存在しない場合にはＯＳ８の一部として実現される。したがって、受信側エンドノード３は、仮想スイッチ９が存在する仮想化シ

ステムであるか否かにかかわらず、競合するエレファントフロー間でスループットを平均化することができる。

【０１０５】

また、実施例１では、ＲＷＩＮ比較部６３は、ＲＷＩＮ計算部６２が計算したＲＷＩＮとパケットサイズとを比較し、大きい方をｃａｌＲＷＩＮとする。したがって、ＲＷＩＮ制御部５は、ｓｅｔＲＷＩＮの大きさを１パケット以上とすることができる。

【０１０６】

また、実施例１では、複数のエレファントフローが存在してα＝０になった場合に、ｓｅｔＲＷＩＮ計算部６５は強制的にα＝０．１とする。したがって、ＲＷＩＮ制御部５は、ｓｅｔＲＷＩＮを大きくすることができ、スループットを向上することができる。

【０１０７】

なお、実施例１では、αをｃａｌＲＷＩＮに乗じる際に、αを２で割ることとしたが、ネットワークの状態、特徴等を考慮して、αを２で割らない、αの平方根を用いる、αを他の自然数で割る、αを所定の指数で割ることとしてもよい。また、ｃａｌＲＷＩＮにαを乗じる代わりにｃａｌＲＷＩＮをαの平方根で割ってもよい。

【実施例２】

【０１０８】

ところで、上記実施例１では、１つの受信エンドノード３で複数の競合するエレファントフローのレートを制御する場合について説明したが、エレファントフローは複数の受信エンドノード３間で競合する場合もある。そこで、実施例２では、複数の受信エンドノード３間で複数のエレファントフローが競合する場合のレート制御について説明する。

【０１０９】

図２２は、実施例２に係る情報処理システムの構成を示す図である。なお、ここでは説明の便宜上、図１に示した装置と同様の役割を果たす装置については同一符号を付すこととしてその詳細な説明を省略する。図２２に示すように、情報処理システム１ａは、ｎ＋１台の送信側エンドノード２と、ｎ＋１台の受信側エンドノード３と、６台のスイッチ装置４と、コントローラ１０とを有する。

【０１１０】

スパインスイッチＳＷ＃０３～ＳＷ＃０５は、パケットのヘッダを見てエレファントフローのパケットであれば、ＳＹＮパケット、ＦＩＮパケット及びＲＳＴパケットをスヌープしてコントローラ１０に送信する。スパインスイッチＳＷ＃０３～ＳＷ＃０５は、エレファントフローであるか否かをパケットのヘッダのＤＳＣＰ、ＣｏＳ（Class of Service）等を用いて判定する。図２３は、ＣｏＳを説明するための図である。図２３に示すように、ＣｏＳは、パケットのヘッダのＶＬＡＮタグに含まれる。

【０１１１】

また、各受信エンドノード３は、エレファントフローのＲＴＴの平均値を算出すると、算出した平均ＲＴＴをコントローラ１０に送信する。

【０１１２】

コントローラ１０は、スパインスイッチＳＷ＃０３～ＳＷ＃０５から送信されたパケット及び各受信エンドノード３から送信された平均ＲＴＴに基づいて、複数の受信エンドノード３間で競合するエレファントフローのレート制御を行う。コントローラ１０はＲＷＩＮを計算して各受信側エンドノード３に送信し、各受信側エンドノード３はコントローラ１０から送信されたＲＷＩＮを用いてエレファントフローのレート制御を行う。

【０１１３】

図２４は、実施例２に係るレート制御を説明するための図である。図２４に示すように、コントローラ１０はコントローラＲＷＩＮ制御部５ａを有する。コントローラＲＷＩＮ制御部５ａはネットワーク１１からエレファントフローに関する情報を収集してＲＷＩＮを計算し、計算したＲＷＩＮを各受信側エンドノード３のＲＷＩＮ制御部５に送信し、ＲＷＩＮ制御部５がレート制御を行う。各受信側エンドノード３のＲＷＩＮ制御部５は、自身が計算したｓｅｔＲＷＩＮよりコントローラ１０から受信したＲＷＩＮを優先してレー

ト制御を行う。

【０１１４】

このように、コントローラ１０がＲＷＩＮを計算することによって、実施例２に係る情報処理システム１ａは受信側エンドノード３間で競合するエレファントフローのレート制御を公平に行うことができる。

【０１１５】

図２５は、ＲＷＩＮ送信用のＰＤＵのフォーマットを示す図である。図２５に示すように、ＲＷＩＮ送信用のＰＤＵでは、送信先アドレスとしてあらかじめ予約された「０１－８０－Ｃ２－００－００－ＸＸ」が用いられる。また、図２５に示す２バイトのＢＰＤＵにＲＷＩＮの値が設定される。

【０１１６】

図２６は、平均ＲＴＴ送信用のＰＤＵのフォーマットを示す図である。図２６に示すように、平均ＲＴＴ送信用のＰＤＵでは、送信先アドレスとしてあらかじめ予約された「０１－８０－Ｃ２－００－００－ＸＸ」が用いられる。また、図２６に示す３８バイトのＢＰＤＵに５タプルと平均ＲＴＴが設定される。

【０１１７】

図２７は、コントローラＲＷＩＮ制御部５ａの構成を示す図である。図２７に示すように、コントローラＲＷＩＮ制御部５ａは、５タプル検出部７１と、ＣＴＲＬ＿ＲＷＩＮテーブル７２と、パス計算部７３と、パス検索部７４と、ＲＴＴ測定部７５と、スループット計算部７６と、ＡＢＷ計算部７７とを有する。また、コントローラＲＷＩＮ制御部５ａは、ＲＷＩＮ計算部７８と、ＲＷＩＮ設定部７９と、通信部８０とを有する。

【０１１８】

５タプル検出部７１は、ＳＹＮパケットのヘッダから５タプルを取り出す。ＣＴＲＬ＿ＲＷＩＮテーブル７２は、ＲＷＩＮの算出に用いられる情報を記憶する。図２８は、ＣＴＲＬ＿ＲＷＩＮテーブル７２がエレファントフロー毎に記憶する項目の一例を示す図である。図２８に示すように、ＣＴＲＬ＿ＲＷＩＮテーブル７２は、「インデックス」、「５タプル」、「パス」、「ＲＴＴ」、「ＡＢＷ」、「スループット」、「ＲＷＩＮ」及び「トポロジー」をエレファントフロー毎に記憶する。

【０１１９】

「インデックス」は、エレファントフローを識別する識別子である。「５タプル」は、エレファントフローを特定するための５タプルである。「パス」は、エレファントフローが通る端から端までのパスである。「ＲＴＴ」は、ＡＣＫとデータ間の平均ＲＴＴである。

【０１２０】

「ＡＢＷ」は、パス上で最小の利用可能バンド幅である。図２９は、利用可能バンド幅を説明するための図である。図２９に示すように、利用可能バンド幅は、リンク容量から使用中の帯域幅を引いた帯域幅である。例えば、リンク容量が１０Ｇｂｐｓで使用中の帯域幅が５Ｇｂｐｓのとき、利用可能バンド幅は５Ｇｂｐｓである。

【０１２１】

「スループット」は、エレファントフローのスループットである。「ＲＷＩＮ」は、ＡＢＷとスループットで計算されたＲＷＩＮである。「トポロジー」は、スイッチ装置間のリンクの有無を示し、マトリクスで定義される。図３０は、マトリクスの一例を示す図である。図３０に示すように、スイッチ装置間にリンクがある場合には、マトリクスの対応する要素は「○」であり、スイッチ装置間にリンクがない場合には、マトリクスの対応する要素は「－」である。

【０１２２】

図２７に戻って、パス計算部７３は、５タプル、マトリクス等のネットワーク情報を用いてエレファントフローの端から端までのパスを計算し、ＣＴＲＬ＿ＲＷＩＮテーブル７２に記録する。パス検索部７４は、ＣＴＲＬ＿ＲＷＩＮテーブル７２を検索して同じリンクでエレファントフローが２以上であるかを判定する。

【０１２３】
ＲＴＴ測定部７５は、受信側エンドノード３が送信した平均ＲＴＴを通信部８０から受け取り、ＣＴＲＬ＿ＲＷＩＮテーブル７２に記録する。スループット計算部７６は、平均ＲＴＴとフロー数から各エレファントフローのスループットを計算する。
【０１２４】
ＡＢＷ計算部７７は、リンク数とリンク容量から各リンクの利用可能バンド幅を計算し、計算した利用可能バンド幅の最小値ＡＢＷをＣＴＲＬ＿ＲＷＩＮテーブル７２に記録する。具体的には、ＡＢＷ計算部７７は、以下の式（２）でＡＢＷを計算する。
【数２】

$$ABW = MIN_{l=0,L}\{Linkcap(l) - Linkused(l)\} \quad \cdots(2)$$

式（２）で、Ｌｉｎｋｃａｐはリンク容量であり、Ｌｉｎｋｕｓｅｄは使用中の帯域幅である。
【０１２５】
ＲＷＩＮ計算部７８は、ＡＢＷ、各エレファントフローのスループット、複数の受信側エンドノード３から受信した平均ＲＴＴの平均値に基づいてＲＷＩＮを計算する。具体的には、ＲＷＩＮ計算部７８は、以下の式（３）でＲＷＩＮを計算する。
【数３】

$$RWIN=\left\{\left(ABW+\sum_{i=0}^{I}BW(f_i)\right)Avg(RTT)\right\}/n \quad \cdots(3)$$

式（３）でＢＷ（f_i）はエレファントフローｉが使用するバンド幅（スループット）であり、Ａｖｇ（ＲＴＴ）は、複数の受信側エンドノード３から受信した平均ＲＴＴの平均値である。
【０１２６】
ＲＷＩＮ設定部７９は、ＲＷＩＮ計算部７８により計算されたＲＷＩＮを競合が起きている受信側エンドノード３に通信部８０を介して送信する。通信部８０は、受信側エンドノード３から送信される平均ＲＴＴの受信、受信側エンドノード３へのＲＷＩＮの送信、スイッチ装置４からのパケットの受信等の通信を行う。
【０１２７】
次に、コントローラ１０による処理のフローについて説明する。図３１は、コントローラ１０による処理のフローを示すフローチャートである。図３１に示すように、通信部８０は、パケットを受信し、受信したパケットがＳＹＮパケットであるか否かを判定する（ステップＳ６１）。
【０１２８】
その結果、ＳＹＮパケットでない場合には、通信部８０は、パケットがＦＩＮパケット又はＲＳＴパケットであるか否かを判定する（ステップＳ６２）。その結果、ＦＩＮパケットでもＲＳＴパケットでもない場合には、通信部８０は、パケットが平均ＲＴＴのＢＰＤＵかを判定する（ステップＳ６３）。その結果、パケットが平均ＲＴＴのＢＰＤＵでない場合には、コントローラ１０は、ステップＳ６１に戻る。
【０１２９】
一方、受信したパケットが平均ＲＴＴのＢＰＤＵである場合には、ＲＴＴ測定部７５は平均ＲＴＴをＣＴＲＬ＿ＲＷＩＮテーブル７２に記録し（ステップＳ６４）、ＲＴＴフラグをｔｒｕｅに設定する（ステップＳ６５）。また、ステップＳ６２において、受信したパケットがＦＩＮパケット又はＲＳＴパケットである場合には、コントローラ１０は、ステップＳ７０に進む。
【０１３０】
また、ステップＳ６１において、受信したパケットがＳＹＮパケットである場合には、５タプル検出部７１は、ＳＹＮパケットから５タプルを取り出す（ステップＳ６６）。そして、パス計算部７３は、トポロジーを読み込み（ステップＳ６７）、エレファントフロ

ーのパスを計算して（ステップＳ６８）、ＣＴＲＬ＿ＲＷＩＮテーブル７２にパスを保存する（ステップＳ６９）。

【０１３１】

そして、パス検索部７４は、ＣＴＲＬ＿ＲＷＩＮテーブル７２を検索し、同じパス上でエレファントフローの数が２より大きいか否かを判定し（ステップＳ７０）、２より大きくない場合には、コントローラ１０は、ステップＳ６１に戻る。一方、２より大きい場合には、コントローラ１０は、ＲＴＴフラグがｔｒｕｅであるか否かを判定し（ステップＳ７１）、ｔｒｕｅでない場合には、受信側エンドノード３から平均ＲＴＴを受信していないので、ステップＳ６１に戻る。

【０１３２】

一方、ＲＴＴフラグがｔｒｕｅである場合には、通信部８０は、パス上のスイッチ装置４からリンク数を収集する（ステップＳ７２）。そして、ＡＢW計算部７７は、各リンクの利用可能バンド幅を計算し（ステップＳ７３）、利用可能バンド幅の最小値ＡＢＷを計算する（ステップＳ７４）。

【０１３３】

そして、通信部８０は、最小値ＡＢＷを有するスイッチ装置４からエレファントフロー数を収集する（ステップＳ７５）。そして、スループット計算部７６は、Ａｖｇ（ＲＴＴ）とエレファントフロー数でスループットを計算する（ステップＳ７６）。そして、ＲＷＩＮ計算部７８は、ＡＢＷとスループットでＲＷＩＮを計算する（ステップＳ７７）。そして、ＲＷＩＮ設定部７９は、エレファントフロー間競合が生じた受信側エンドノード３に通信部８０を介してＲＷＩＮを設定する（ステップＳ７８）。

【０１３４】

このように、コントローラ１０がＲＷＩＮを計算し、計算したＲＷＩＮを受信側エンドノード３に送信することによって、エレファントフローが競合する受信側エンドノード３間でスループットを均等化することができる。

【０１３５】

次に、レート制御のシーケンスについて説明する。図３２は、レート制御のシーケンスを示す図である。図３２に示すように、まず、前提条件及び事前準備が整う（ステップｔ３１）。すなわち、インキャストが発生する。

【０１３６】

そして、S_0はR_0へエレファントフローＥ＃０のＳＹＮパケットを送信し、S_nはR_nへエレファントフローＥ＃１のＳＹＮパケットを送信する。そして、コントローラ１０の５タプル検出部７１はＳＹＮパケットから５タプルを取り出し、パス計算部７３はパスを計算する（ステップｔ３２）。また、R_0はS_0へエレファントフローＥ＃０のＡＣＫを送信し、R_nはS_nへエレファントフローＥ＃１のＡＣＫを送信する。

【０１３７】

そして、エレファントフローの輻輳ウィンドウが輻輳回避状態になる（ステップｔ３３）。そして、受信側エンドノード３のＲＴＴ測定部５４がＲＴＴを測定し、ＡｖｇＲＴＴ計算部５６が平均ＲＴＴを計算する（ステップｔ３４）。そして、R_0のコントローラ通信部６８が平均ＲＴＴをＢＰＤＵ化して送信し（ステップｔ３５）、R_nのコントローラ通信部６８が平均ＲＴＴをＢＰＤＵ化して送信する（ステップｔ３６）。

【０１３８】

そして、コントローラ１０の通信部８０はリンク数を収集して、ＡＢW計算部７７は最小値ＡＢＷを計算する（ステップｔ３７）。また、コントローラ１０の通信部８０はエレファントフロー数を収集して、スループット計算部７６はスループットを計算する（ステップｔ３８）。そして、コントローラ１０のＲＷＩＮ計算部７８はＲＷＩＮを計算し、通信部８０はＢＰＤＵ化したＲＷＩＮを各受信側エンドノード３に送信する（ステップｔ３９）。

【０１３９】

そして、コントローラ１０からのＲＷＩＮを各受信側エンドノード３のＲＷＩＮ設定部

６６がＡＣＫに設定する（ステップｔ４０）。そして、R_0はS_0へエレファントフローＥ＃０のＡＣＫを送信し、R_nはS_nへエレファントフローＥ＃１のＡＣＫを送信する。

【０１４０】

上述してきたように、実施例２では、スループット計算部７６が各エレファントフローのスループットを計算し、ＡＢＷ計算部７７が利用可能バンド幅の最小値ＡＢＷを計算する。そして、ＲＷＩＮ計算部７８が、スループットから計算される各エレファントフローの利用可能バンド幅とＡＢＷとＡｖｇ（ＲＴＴ）を用いてＲＷＩＮを計算し、ＲＷＩＮ設定部７９がＲＷＩＮを各受信側エンドノード３に送信する。そして、コントローラ１０からのＲＷＩＮを各受信側エンドノード３のＲＷＩＮ設定部６６がＡＣＫに設定する。したがって、コントローラＲＷＩＮ制御部５ａは受信側エンドノード３間でエレファントフロー間の競合の発生を防ぎ、情報処理システム１ａはエレファントフローを効率的に処理することができる。

【０１４１】

また、実施例１及び２では、ＲＷＩＮ制御部５及びコントローラＲＷＩＮ制御部５ａについて説明した。しかし、ＲＷＩＮ制御部５及びコントローラＲＷＩＮ制御部５ａが有する構成をソフトウェアによって実現することで、同様の機能を有するＲＷＩＮ制御プログラムを得ることができる。そこで、ＲＷＩＮ制御プログラムを実行するコンピュータについて説明する。

【０１４２】

図３３は、実施例１及び２に係るＲＷＩＮ制御プログラムを実行するコンピュータの構成を示す図である。図３３に示すように、コンピュータ９０は、メインメモリ９１と、ＣＰＵ（Central Processing Unit）９２と、ＬＡＮ（Local Area Network）インタフェース９３と、ＨＤＤ（Hard Disk Drive）９４とを有する。また、コンピュータ９０は、スーパーＩＯ（Input Output）９５と、ＤＶＩ（Digital Visual Interface）９６と、ＯＤＤ（Optical Disk Drive）９７とを有する。

【０１４３】

メインメモリ９１は、プログラムやプログラムの実行途中結果などを記憶するメモリである。ＣＰＵ９２は、メインメモリ９１からプログラムを読み出して実行する中央処理装置である。ＣＰＵ９２は、メモリコントローラを有するチップセットを含む。

【０１４４】

ＬＡＮインタフェース９３は、コンピュータ９０をＬＡＮ経由で他のコンピュータに接続するためのインタフェースである。ＨＤＤ９４は、プログラムやデータを格納するディスク装置であり、スーパーＩＯ９５は、マウスやキーボードなどの入力装置を接続するためのインタフェースである。ＤＶＩ９６は、液晶表示装置を接続するインタフェースであり、ＯＤＤ９７は、ＤＶＤの読み書きを行う装置である。

【０１４５】

ＬＡＮインタフェース９３は、ＰＣＩエクスプレス（ＰＣＩｅ）によりＣＰＵ９２に接続され、ＨＤＤ９４及びＯＤＤ９７は、ＳＡＴＡ（Serial Advanced Technology Attachment）によりＣＰＵ９２に接続される。スーパーＩＯ９５は、ＬＰＣ（Low Pin Count）によりＣＰＵ９２に接続される。

【０１４６】

そして、コンピュータ９０において実行されるＲＷＩＮ制御部プログラムは、ＤＶＤに記憶され、ＯＤＤ９７によってＤＶＤから読み出されてコンピュータ９０にインストールされる。あるいは、ＲＷＩＮ制御部プログラムは、ＬＡＮインタフェース９３を介して接続された他のコンピュータシステムのデータベースなどに記憶され、これらのデータベースから読み出されてコンピュータ９０にインストールされる。そして、インストールされたＲＷＩＮ制御部プログラムは、ＨＤＤ９４に記憶され、メインメモリ９１に読み出されてＣＰＵ９２によって実行される。

【０１４７】

なお、実施例１及び２では、インキャストが発生している場合について説明したが、本

発明はこれに限定されるものではなく、インキャストが発生していない場合のエレファントフロー間の競合にも同様に適用することができる。

【０１４８】

また、実施例１及び２では、エレファントフロー間で競合が発生している場合について説明したが、本発明はこれに限定されるものではなく、例えばマイスフローなど他の種類のフロー間で競合が発生している場合にも同様に適用することができる。

【０１４９】

また、実施例１及び２では、ネットワークがスパインスイッチとリーフスイッチで構成され、リーフスイッチの下にエンドノードが接続される場合について説明した。しかしながら、本発明はこれに限定されるものではなく、例えば、リーフスイッチの下にＴＯＲ（Top Of Rack）スイッチが接続し、ＴＯＲにエンドノードが接続される場合など他の構成のネットワークにも同様に適用することができる。

【符号の説明】

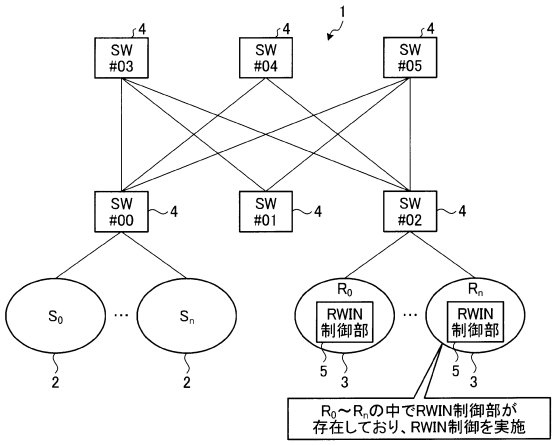
【０１５０】

１，１ａ　情報処理システム
２　送信側エンドノード
３　受信側エンドノード
４　スイッチ装置
５　ＲＷＩＮ制御部
５ａ　コントローラＲＷＩＮ制御部
６　ＡＰＰ
７　ＮＩＣ
８　ＯＳ
９　仮想スイッチ
１０　コントローラ
１１　ネットワーク
５１　ウィンドウテーブル
５２　ウィンドウ監視部
５３　ＲＴＴテーブル
５４　ＲＴＴ測定部
５５　ＲＴＴタイマ
５６　ＡｖｇＲＴＴ計算部
５７　ＮＴタイマ
５８　スループット測定部
５９　スループット計算部
６０　最小スループット計算部
６１　ＲＷＩＮテーブル
６２　ＲＷＩＮ計算部
６３　ＲＷＩＮ比較部
６４　ＲＷＩＮタイマ
６５　ｓｅｔＲＷＩＮ計算部
６６　ＲＷＩＮ設定部
６７　タイムスロット設定部
６８　コントローラ通信部
７１　５タプル検出部
７２　ＣＴＲＬ＿ＲＷＩＮテーブル
７３　パス計算部
７４　パス検索部
７５　ＲＴＴ測定部
７６　スループット計算部

７７　ＡＢＷ計算部
７８　ＲＷＩＮ計算部
７９　ＲＷＩＮ設定部
８０　通信部
８９　受信バッファ
９０　コンピュータ
９１　メインメモリ
９２　ＣＰＵ
９３　ＬＡＮインタフェース
９４　ＨＤＤ
９５　スーパーＩＯ
９６　ＤＶＩ
９７　ＯＤＤ

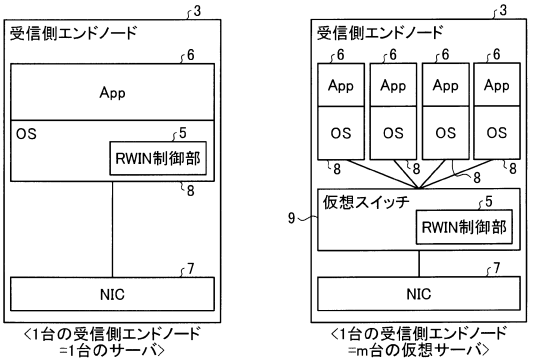
【図１】

実施例1に係る情報処理システムの構成を示す図



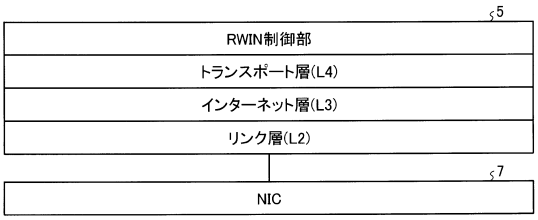
【図２】

RWIN制御部の実現方法を説明するための図



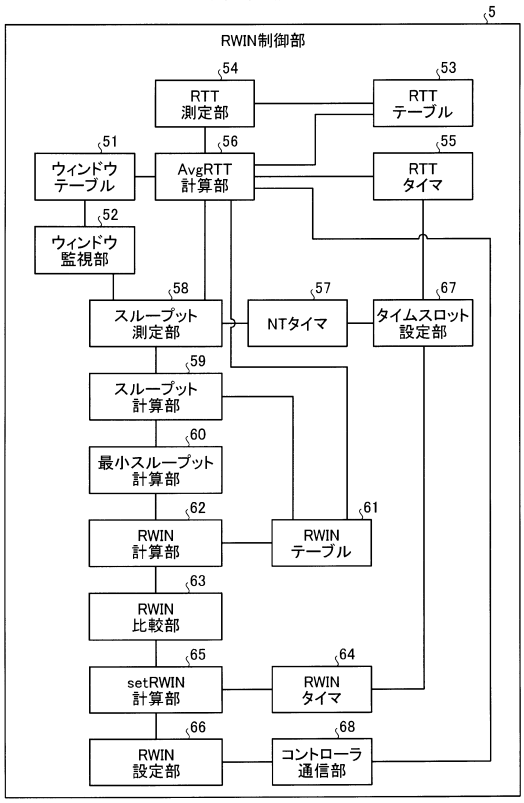
【図３】

通信階層におけるRWIN制御部の位置づけを示す図



【図４】

RWIN制御部の構成を示す図



【図５】

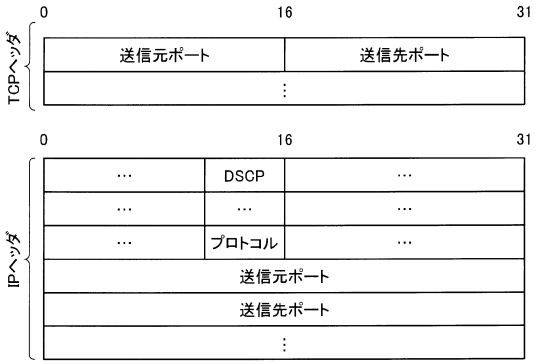
ウィンドウテーブルがエレファントフロー毎に記憶する項目の一例を示す図

インデックス	5タプル	前パケット数	現パケット数	RTT	初期ACK

インデックス:エレファントフローインデックス
5タプル:送信元IP、送信先IP、送信元ポート、送信先ポート、プロトコル
前パケット数:以前(i-1)に受信したパケットの数
現パケット数:現在(i)受信したパケットの数
RTT:ACK↔データ間のRTT
初期ACK:RTTを測るための最初のACKの時刻

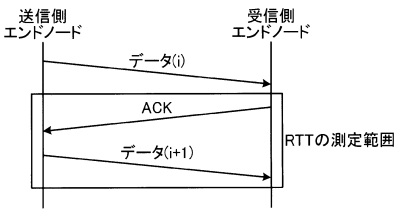
【図６】

5タプル及びDSCPの格納場所を示す図



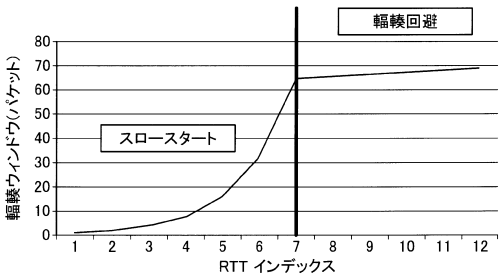
【図７】

RTTの測定範囲を示す図



【図８】

輻輳回避状態を説明するための図



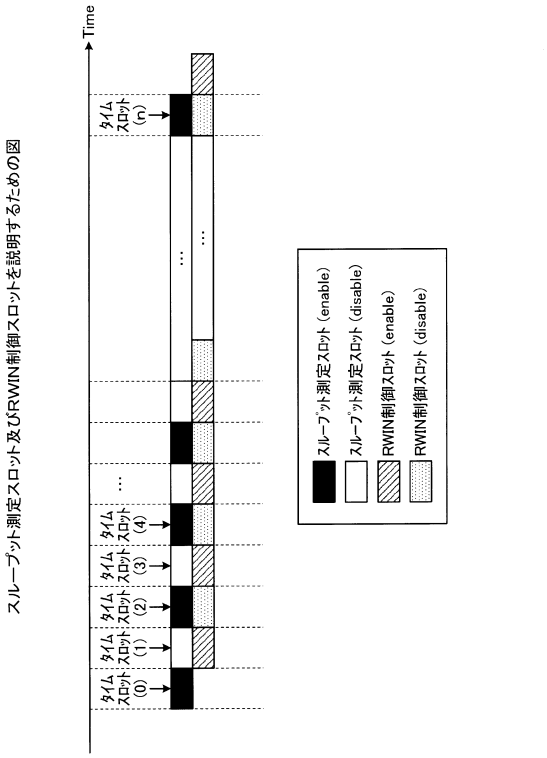
【図９】

RTTテーブルがRTT毎に記憶する項目の一例を示す図

インデックス	RTT	ACK タイム	ACK 番号	失敗数

インデックス:RTTインデックス
RTT:ACK↔データ間のRTT
ACKタイム:ACKが送信された時刻
ACK番号:ACKシーケンス番号+MSS(例えば1500)
失敗数:ACK番号とデータ番号を比較して一致しなかった数

【図１０】



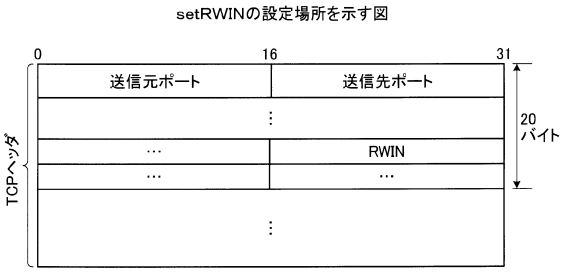
【図１１】

RWINテーブルがエレファントフロー毎に記憶する項目の一例を示す図

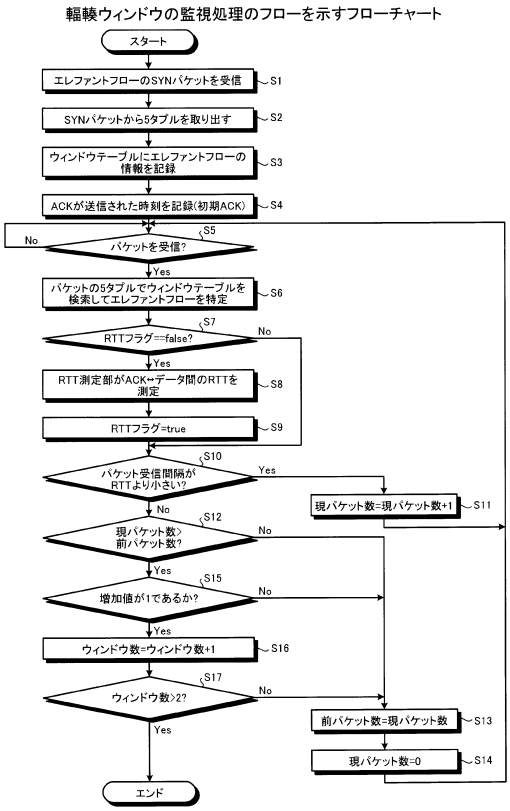
インデックス	バイト数	RTT	スループット	calRWIN

インデックス:エレファントフローインデックス
バイト数:スループット測定スロットで測定されたバイト数
RTT:ACK↔データ間の平均RTT
スループット:エレファントフローのスループット
calRWIN:最小スループットから計算されたRWIN

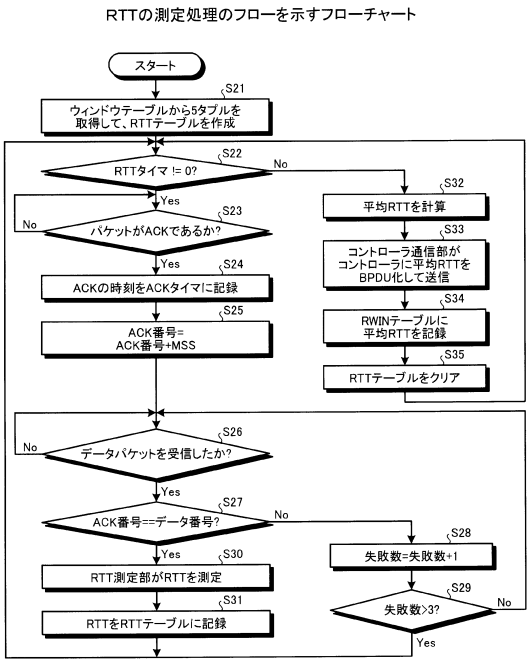
【図１２】



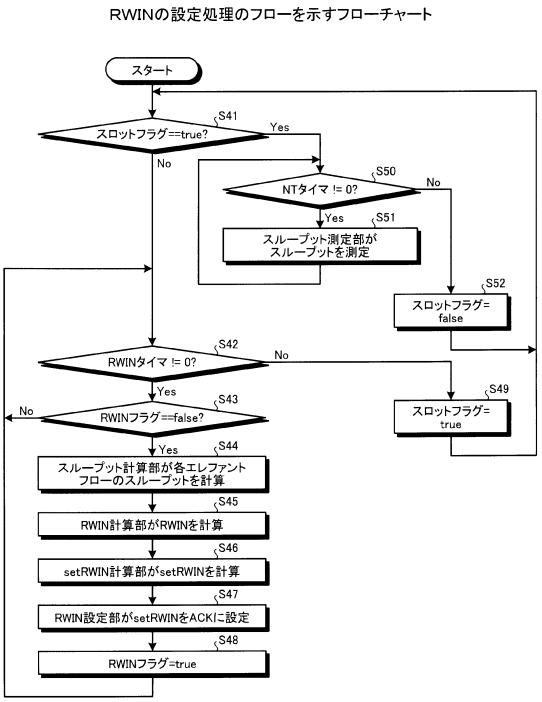
【図１３】



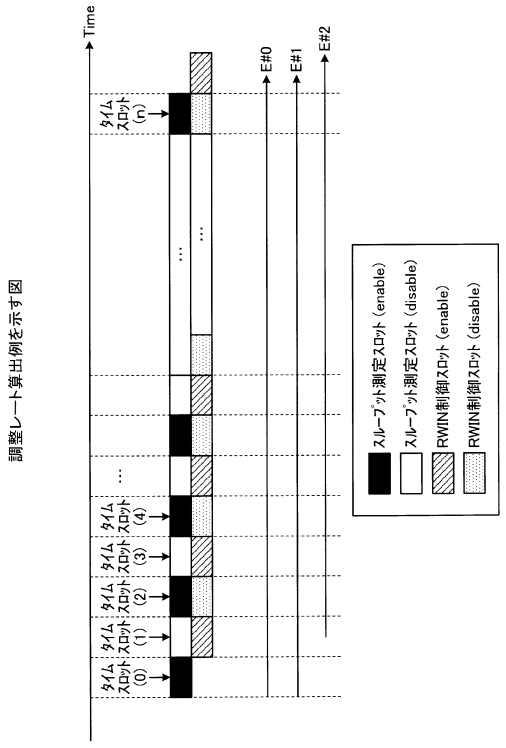
【図１４】



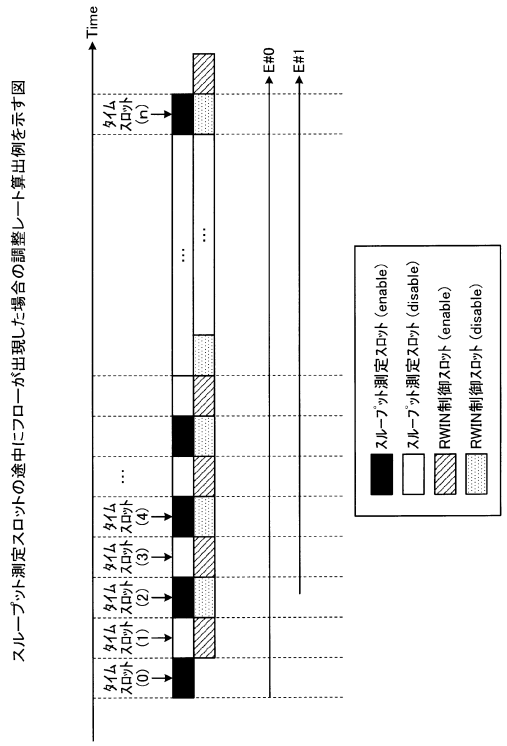
【図１５】



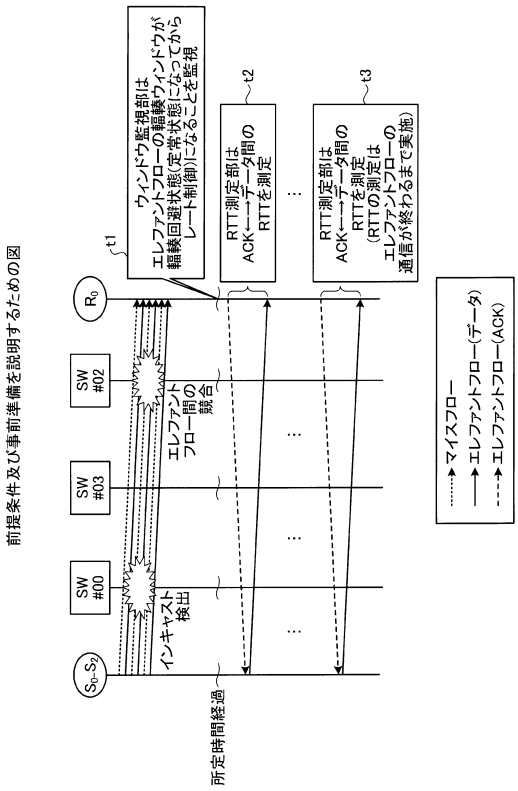
【図１６】



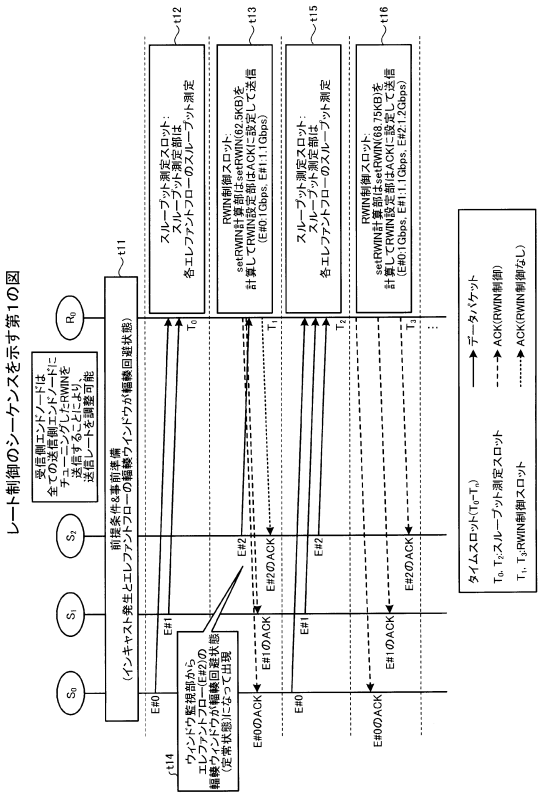
【図１７】



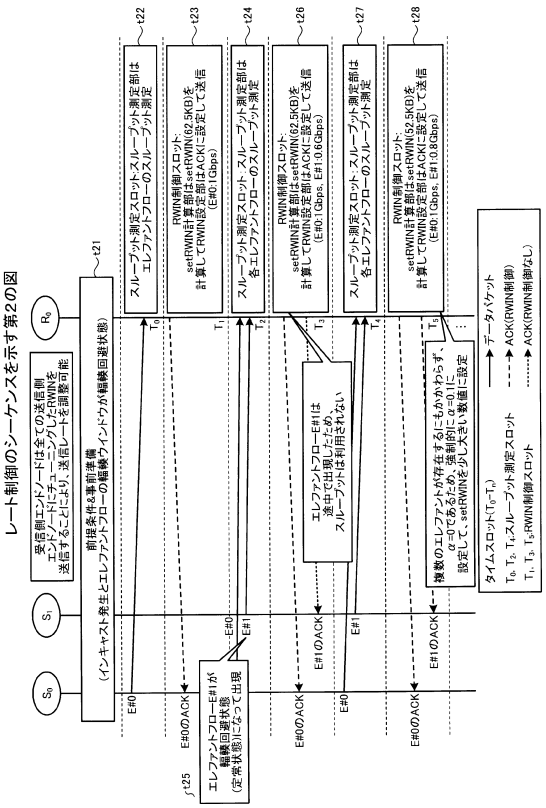
【図１８】



【図１９】

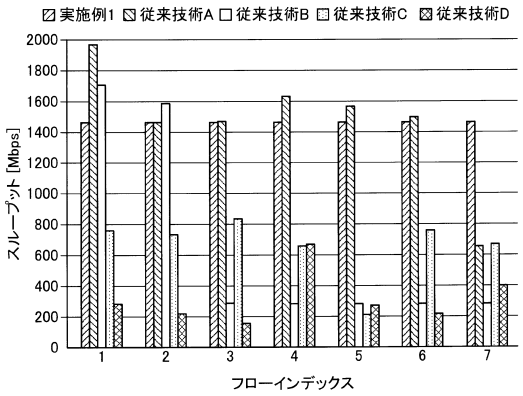


【図２０】



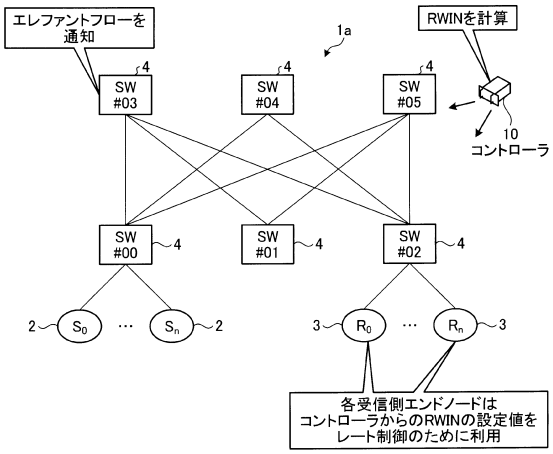
【図２１】

実施例1に係るレート制御の効果を示す図



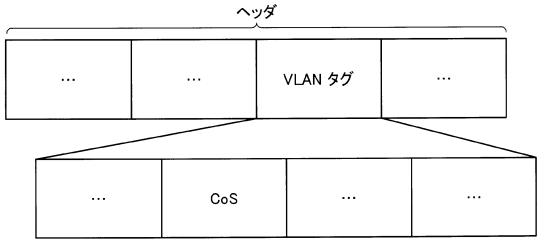
【図２２】

実施例2に係る情報処理システムの構成を示す図



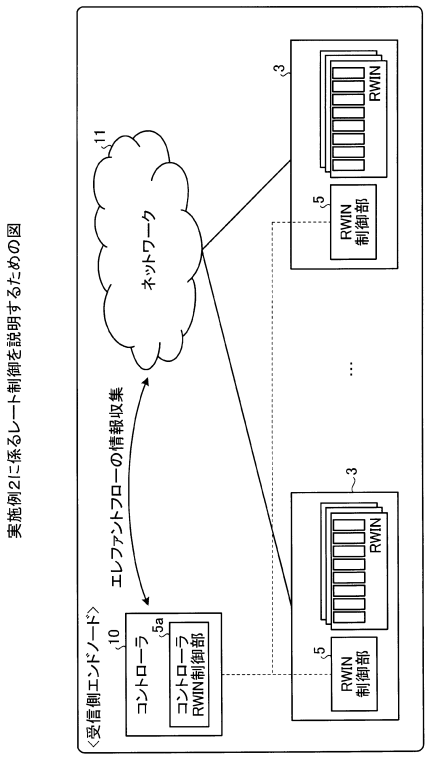
【図２３】

CoSを説明するための図



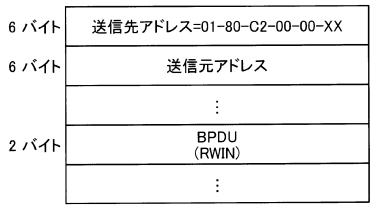
【図２４】

実施例2に係るレート制御を説明するための図



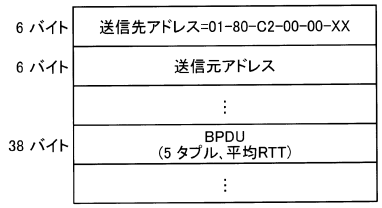
【図２５】

RWIN送信用のPDUのフォーマットを示す図



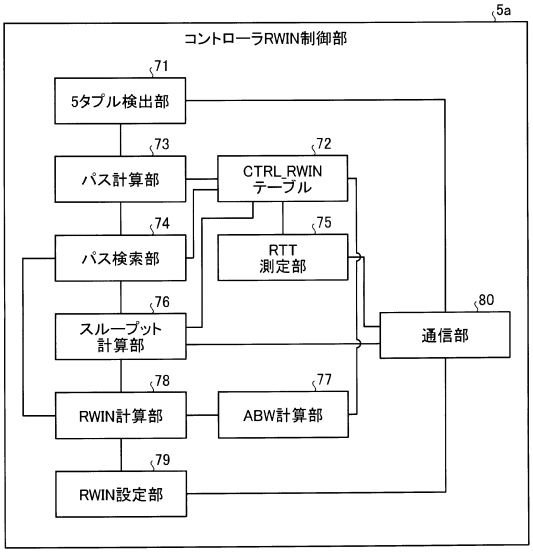
【図２６】

平均RTT送信用のPDUのフォーマットを示す図



【図２７】

コントローラRWIN制御部の構成を示す図



【図２８】

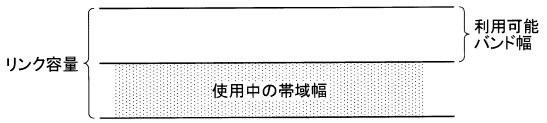
CTRL_RWINテーブルがエレファントフロー毎に記憶する項目の一例を示す図

インデックス	5タプル	パス	RTT	ABW	スループット	RWIN	トポロジー

インデックス:エレファントフロー識別子
5タプル:エレファントフローの5タプル
パス:エレファントフローが通る端から端までのパス
RTT:ACK↔データ間の平均RTT
ABW:パス上で最小の利用可能バンド幅
スループット:エレファントフローのスループット
RWIN:ABWとスループットで計算されたRWIN
トポロジー:マトリクスで定義

【図２９】

利用可能バンド幅を説明するための図



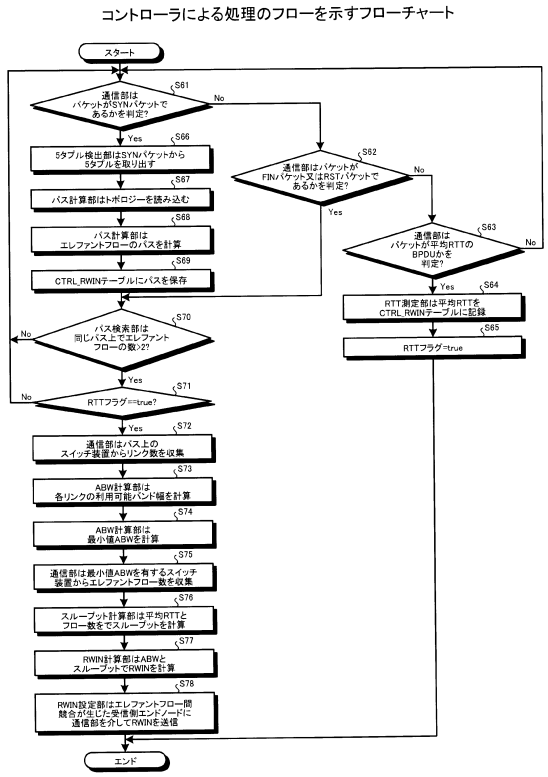
【図３０】

マトリクスの一例を示す図

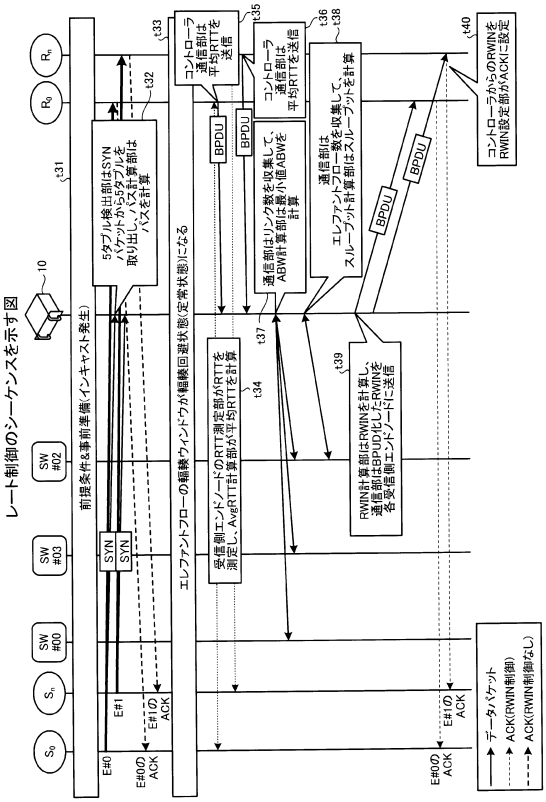
トポロジー	SW#00	…	SW#yy
SW#00	－	…	○
SW#01	○	…	○
⋮	⋮		⋮
SW#xx	○	…	－

－:スイッチ装置の間でリンクが存在しない
○:スイッチ装置の間でリンクが存在する

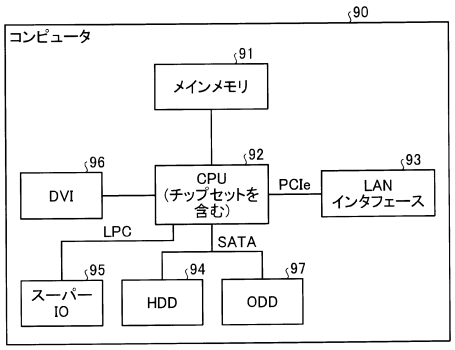
【図３１】



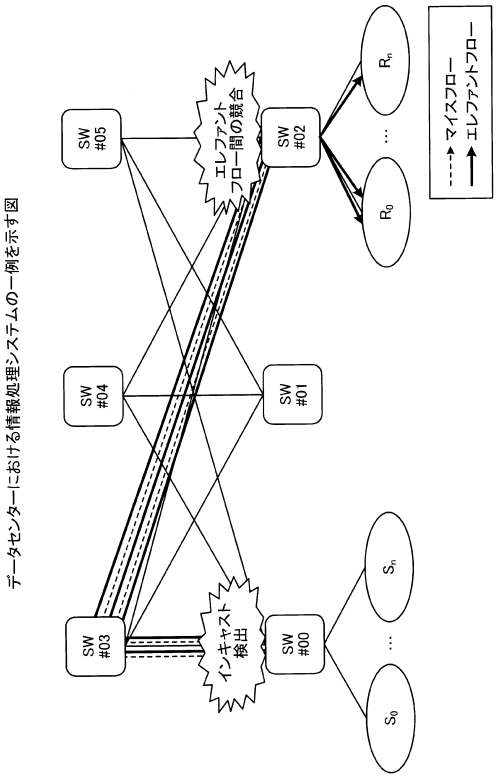
【図３２】



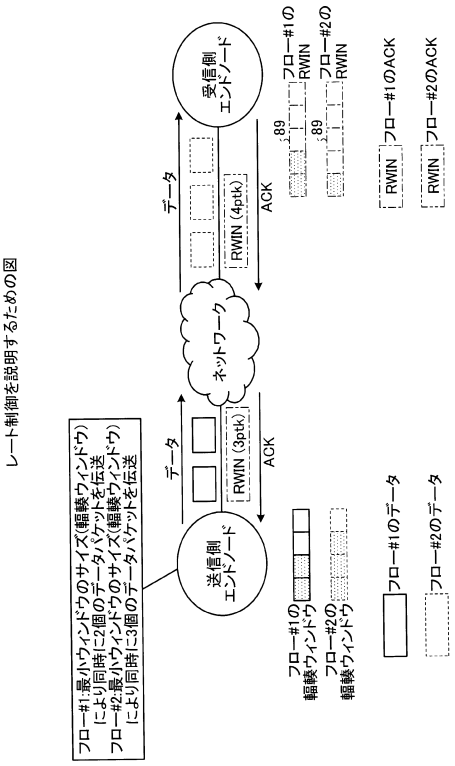
【図３３】



【図３４】



【図３５】


レート制御を説明するための図
フロー#1:最小ウィンドウのサイズ(輻輳ウィンドウ)により同時に2個のデータパケットを伝送
フロー#2:最小ウィンドウのサイズ(輻輳ウィンドウ)により同時に3個のデータパケットを伝送
送信側エンドノード
データ
RWIN (3ptk)
ACK
ネットワーク
データ
RWIN (4ptk)
ACK
受信側エンドノード
フロー#1の輻輳ウィンドウ
フロー#2の輻輳ウィンドウ
フロー#1のデータ
フロー#2のデータ
89
フロー#1のRWIN
89
フロー#2のRWIN
RWIN
フロー#1のACK
RWIN
フロー#2のACK

フロントページの続き

(56)参考文献 米国特許出願公開第２００７／００７６７２６（ＵＳ，Ａ１）
特開２００５－２１８０６９（ＪＰ，Ａ）
特開２０１１－０３５６０８（ＪＰ，Ａ）
特開２００９－２０６７３３（ＪＰ，Ａ）
吉水 卓 Suguru YOSHIMIZU, 高速ＴＣＰの公平性を改善する輻輳制御アルゴリズムの検討 Congestion Control Algorithms to Improve Fairness of High-Speed TCPs, 電子情報通信学会技術研究報告 Ｖｏｌ．１０９ Ｎｏ．３２７ IEICE Technical Report, 日本, 社団法人電子情報通信学会 The Institute of Electronics,Information and Communication Engineers, ２００９年１２月 3日, 第109巻, p.25-30

(58)調査した分野(Int.Cl., ＤＢ名)
Ｈ０４Ｌ １２／８０７
Ｈ０４Ｌ ２９／０８