

(19) 日本国特許庁(JP)

(12) 公開特許公報(A)

(11) 特許出願公開番号

特開2005-295581

(P2005-295581A)

(43) 公開日 平成17年10月20日(2005. 10. 20)

(51) Int.Cl. ⁷	F I	テーマコード (参考)
H04L 12/56	H04L 12/56 200C	5K030
H04L 12/66	H04L 12/66 A	5K034
H04L 13/08	H04L 13/08	5K201
H04M 3/00	H04M 3/00 B	

審査請求 有 請求項の数 18 O L (全 14 頁)

(21) 出願番号	特願2005-135986 (P2005-135986)	(71) 出願人	596092698
(22) 出願日	平成17年5月9日 (2005. 5. 9)		ルーセント テクノロジーズ インコーポレーテッド
(62) 分割の表示	特願平10-27551の分割		アメリカ合衆国, 07974-0636
原出願日	平成10年2月9日 (1998. 2. 9)		ニュージャージー, マレイ ヒル, マウンテン アヴェニュー 600
(31) 優先権主張番号	60/037843	(74) 代理人	100064447
(32) 優先日	平成9年2月7日 (1997. 2. 7)		弁理士 岡部 正夫
(33) 優先権主張国	米国 (US)	(74) 代理人	100085176
(31) 優先権主張番号	08/961122		弁理士 加藤 伸晃
(32) 優先日	平成9年10月30日 (1997. 10. 30)	(74) 代理人	100106703
(33) 優先権主張国	米国 (US)		弁理士 産形 和央
		(74) 代理人	100096943
			弁理士 臼井 伸一

最終頁に続く

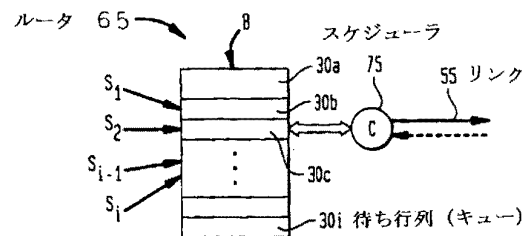
(54) 【発明の名称】 TCP接続の性能改善方法

(57) 【要約】

【課題】 フィードバック制御TCPネットワークの性能を改善する。

【解決手段】 フロー毎にキューイングする装置であり、複数の待ち行列に区分けされた所定サイズのバッファと、所定速度に従って各バッファからパケットを除去し、ネットワークを介して前記パケットを伝送するスケジューラと、受信パケットを受信することができ、待ち行列が使用可能である場合、パケットを待ち行列に入力することができるバッファ内の待ち行列の使用可能性を決定するためのコントロールデバイスとからなり、コントロールデバイスは更に、最長待ち行列ファーストスキームに従って待ち行列を選択し、待ち行列が使用可能でない場合、前記受信パケットの入力を可能にするために前記選択された待ち行列からパケットをドロップし、これにより、リンクによる公平性を高め、パケットスループットを高める。

【選択図】 図2



【特許請求の範囲】

【請求項 1】

T C P 接続の性能改善方法であって、

(i) 所定サイズのバッファを複数個の待ち行列に区分けするステップであって、各待ち行列は T C P 接続に対応して情報パケットを受け入れ、一時的に格納するための占有割当て b_i が割り当てられるとともに、各バッファからパケットを除去し、前記パケットを前記対応の T C P 接続を介して伝送するためのスケジューラにより処理されるステップ、

前記対応の T C P 接続に向けられたパケットが到着すると、

(ii) 該パケットを受け取るための前記 T C P 接続に対応する待ち行列の使用可能性を判定し、該待ち行列が使用可能である場合、前記パケットを該待ち行列に入力するステップ、及び

(iii) 該バッファが満杯の場合、前記パケットの入力を可能にするため、前記複数の待ち行列の中から 1 つの待ち行列を選択し、該選択された待ち行列からパケットを廃棄するステップ

からなることを特徴とする方法。

【請求項 2】

パケット到着時又は離脱時に、各待ち行列の現在の長さ q_i をトラッキングするステップを更に有する請求項 1 の方法。

【請求項 3】

トラッキングステップは、パケットが前記待ち行列に入力されるたびに、前記待ち行列に付随するカウンタを増分するか、又は、パケットが前記待ち行列から解放されるときに、前記カウンタを減分するステップを有する請求項 2 の方法。

【請求項 4】

前記スケジューラは所定の速度で前記各バッファからパケットを除去することができる請求項 1 の方法。

【請求項 5】

前記バッファ待ち行列が満杯の場合、前記満杯待ち行列に宛てられたパケットが到着すると、満杯ではない 1 つ以上の待ち行列からバッファスペースを一時的に借りるステップを有し、前記満杯待ち行列の前記現在の長さはその名目的な占有割当て b_i を越える請求項 2 の方法。

【請求項 6】

待ち行列の使用可能性を決定する前記ステップは、所定サイズの前記バッファが満杯であるか否かを決定するステップを有する請求項 5 の方法。

【請求項 7】

前記選択された待ち行列から解放されるべき前記パケットは、前記待ち行列に内在する最も古いパケットである請求項 2 の方法。

【請求項 8】

待ち行列を選択する前記ステップは、

(a) 各待ち行列について現在の割当て b_i を確立するステップと、

(b) 前記各待ち行列の現在の待ち行列長さの値 q_i を取得するステップと、

(c) 現在の待ち行列長さの値 q_i と、各待ち行列の割当てられたバッファ占有率 b_i との間の差を計算するステップと、

(d) 計算された差で最も大きな値を有する前記待ち行列を選択するステップとを有する請求項 2 の方法。

【請求項 9】

待ち行列を選択する前記ステップは、

(a) 各待ち行列について現在の割当て b_i を確立するステップと、

(b) 現在の待ち行列長さの値 q_i が各待ち行列の割当てられたバッファ占有割当て b_i を越える一連の待ち行列を計算するステップと、

(c) 前記一連の待ち行列からランダムに待ち行列を選択するステップとを有する請求項

10

20

30

40

50

2の方法。

【請求項10】

前記スケジューラは異なる所定速度に従って各バッファからパケットを除去し、待ち行列を選択する前記ステップは、

(a)そのそれぞれの待ち行列長さ q_i とその所定のサービス速度から各待ち行列に格納されたパケットにより経験されるキューイング遅延を計算するステップと、

(b)最も長いキューイング遅延を経験する待ち行列を選択するステップとを有する請求項2の方法。

【請求項11】

前記選択ステップは、

(a)1つ以上のピンを確立するステップと、ここで各ピンは所定の長さの1つ以上の待ち行列に結合される、

(b)最長長さを有する1つ以上の待ち行列に結合されたピンから待ち行列をランダムに選び取るステップを更に有する請求項2の方法。

【請求項12】

複数のソースからの情報パケットをTCP/IPネットワーク内の単一の通信リンクへ通信するルータであって、

(a)複数個の待ち行列に区分けされた所定サイズのバッファであって、各待ち行列は、情報パケットを受信し一時的に格納するための占有割当て b_i が割り当てられ、各待ち行列はTCP接続に対応しているバッファ、

(b)各バッファからパケットを除去し、また、前記TCP接続を介して前記パケットを伝送するためのスケジューラ、及び

(c)前記バッファ内の待ち行列の使用可能性を判定するコントロール手段であって、パケットが向けられた前記TCP接続に対応する待ち行列が使用可能である場合、該待ち行列に該受信パケットを入力し、前記バッファが満杯の場合、該受信パケットの入力を可能にするため、該複数の待ち行列の中から1つの待ち行列を選択するとともに前記スケジューラが該選択された待ち行列からパケットを廃棄できるようにするためのコントロール手段

からなることを特徴とするルータ。

【請求項13】

パケットが入力されるたびに、又は前記待ち行列から解放されるたびに、前記待ち行列の現在の長さ q_i をトラッキングするための、前記各待ち行列に結合された手段を更に有する請求項12のルータ。

【請求項14】

前記トラッキング手段は、前記待ち行列に結合されたカウンタデバイスを有し、前記カウンタデバイスは、パケットが前記待ち行列に入力されるたびに増分し、パケットが前記待ち行列から解放されるときは減分する請求項13のルータ。

【請求項15】

前記コントロール手段は、

(a)前記各待ち行列の現在の待ち行列長さの値 q_i を取得する手段と、

(b)現在の待ち行列長さの値 q_i と、各待ち行列について割当てられたバッファ占有割当て b_i との間の差を計算する手段とを有し、計算された差で最大の値を有する前記待ち行列を選択する請求項13のルータ。

【請求項16】

前記コントロール手段は、

(a)現在の待ち行列長さの値 q_i が各待ち行列について割当てられたバッファ占有割当て b_i を越える一連の待ち行列を計算する手段と、

(b)前記一連の待ち行列からランダムに待ち行列を選択する手段とを更に有する請求項13のルータ。

【請求項17】

10

20

30

40

50

前記選択された待ち行列から解放されるべき前記パケットは、前記待ち行列に内在する最も古いパケットである請求項 12 のルータ。

【請求項 18】

前記スケジューラは異なる所定速度に従って各バッファからパケットを除去し、前記コントロール手段は、

(a) そのそれぞれの待ち行列長さ q_i とその所定のサービス速度から各待ち行列に格納されたパケットにより経験されるキューイング遅延を計算する手段と、

(b) 最も長いキューイング遅延を経験する待ち行列を選択する手段とを更に有する請求項 13 のルータ。

【発明の詳細な説明】

10

【技術分野】

【0001】

本発明は TCP プロトコルを処理する通信ネットワークに関する。更に詳細には、本発明はフィードバック制御ネットワークトラヒック用の接続ごとのキューイング (queuing) 及びバッファ管理スキームに関する。

【背景技術】

【0002】

データパケットの高信頼トランスポートを実行するためのトランスポートプロトコル "TCP" 処理系は周知である。図 1 は、TCP/IP ネットワーク (例えば、インターネット、イントラネット) における従来技術による常用のルータ装置の概要図である。

20

【0003】

図 1 において、ローカル・エリア・ネットワーク (LAN) 20 はルータ 50 を介して広域ネットワーク (WAN) 60 に接続されている。一般的に、ルータ 50 のアーキテクチャは、複数のデータストリーム (例えば、LAN 20 に接続されたソースデータ端末からのもの) が単一のバッファにより受信され、かつ、リンク 55 を介してパケットを送信するサーバにより処理されるようなものである。

【0004】

データパケットがその宛先 (例えば、広域ネットワーク 60 に配置されたデータ端末) に到着すると、パケットの受信機は、受信された各データパケットについて、暗黙的に又は明快に、肯定応答 "ACK" パケットを発生する。ソース端末からのデータ送信は、先行データ送信からの受信肯定応答パケットの速度により制御される。

30

【0005】

従って、周知のように、ソースからのデータパケット送信の制御は、Richard W. Stevens, "TCP/IP Illustrated", Vols. 1-3, Addison Wesley 1994 に概説されるように、TCP - Reno と TCP - Tahoe の、適応型スライディングウィンドウスキームの 2 つの主要変形に従い、受信機からのデータ ACK パケットのフィードバックにより決定される。

【0006】

TCP は長ラウンドトリップ時間を伴う接続に対して固有の不公平性 (unfairness) を示すことが確認された。すなわち、短ラウンドトリップ時間接続は、長ラウンドトリップ時間接続よりも大きな帯域幅を得てしまう。なぜなら、短ラウンドトリップ時間接続の場合の TCP ウィンドウは、長ラウンドトリップ時間接続のウィンドウよりも速く成長してしまうからである。これは、短ラウンドトリップ時間接続では接続肯定応答 (ACK) がより速く受信されるためである。この固有の不公平性を解消する効果的な解決策は未だ発見されていない。

40

【0007】

TCP に固有の別の問題点は、ウィンドウ同期 (window synchronization) 現象である。例えば、2 つの等しいラウンドトリップ時間接続がリンクにあるとすれば、そのリンクは飽和してしまい、バッファに蓄積が始まる。或る時点で、そのリンクに接続されたバッファはオーバーフローしてしまう。そして、これらの接続はおおよそ同時にパケット損失を

50

経験する。結果的に、各接続はそのウインドウを事実上同時に縮小し、ネットワークスループット速度も同時に低下してしまい、その結果、ネットワークリンクは十分に活用されない。

【 0 0 0 8 】

T C P 接続ネットワーク性能を低下させるその他の問題点は、位相効果、ボトルネックリンクによる低リバースパスの物理的な帯域幅非対称性、バーストラヒックの生成、分離性の欠如、より侵略的なトランスポートプロトコル又は悪意的ユーザからの保護性の欠如などである。

【 0 0 0 9 】

ネットワーク性能を改善するために実行された或る解決法は、ランダム・アーリー・デ
イテクション (" R E D ") スキームである。このスキームは、バッファが飽和される前
に、パケットをドロップするように機能する。バッファはモニターされ、平均バッファ
占有率が特定の閾値を超えたら、パケットは、例えば、平均バッファ占有率 (移動時間平
均) の関数である特定の確率に従ってランダムにドロップされ、結果的に、T C P 接続は
そのウインドウをランダムにカットする。

【 0 0 1 0 】

これは、ウインドウを非同期化するのに特に有効である。特に、F I F O - R E D スキ
ームでは、リンクを多数使用する接続は、一層多くのパケットをドロップし易い。従っ
て、結果的に、その対応ウインドウサイズもカットされる。これは、このような高速度接続
のバイアスを低下させる傾向がある。

【 0 0 1 1 】

米国特許出願第 0 8 / 8 5 8 3 1 0 号明細書に開示された別のスキームは、リバースパ
スが混雑した時に、公平性とスループットを高める効果を有する A C K パケットに関する
フロントストラテジーからのドロップを実行する。

【 0 0 1 2 】

非 T C P 開ループトラヒック (例えば、いわゆる " リーキー・バケット (leaky-bucket)
" 条件付タイプのストリーム) の場合、等しい又は異なる重みに従い、各入力接続間で帯
域幅が共用されるように実行されるスケジューリングスキームに従って、接続リンクの性
能を最大化するための公平待ち行列スケジューリングスキームが実行されている。

【 0 0 1 3 】

最近の研究努力は、エンド・ツー・エンド遅延のバウンド (bound) 及び分離 (isolation)
を保証する手段として、公平待ち行列スキームを使用することに集中的に向けられている
。これに関連する研究は基本的に、非フィードバック制御リーキーバケット管理トラヒッ
クに関するものであり、重み付公平キューイングスキームによる交換機及びルータが発達
した。

【 0 0 1 4 】

例えば、A.K. Parekh and R.G. Gallager, "A Generalized Processor Sharing Apprao
ch to Flow Control in the Single Node Case", Proc. of INFOCOM '92, pp. 915-924,
May 1992及びD. Stiliadis and A. Varma, "Design and Analysis of Frame-based Queui
ng: A New Traffic Scheduling Algorithm for Packet-Switched Networks", Proc of AC
M SIGMETRICS '96, pp. 104-115, May 1996参照。

【 0 0 1 5 】

交換機における公平キューイング (F Q) の関心は、様々なバッファ構成 (例えば、物
理的又は論理的に分離されたバッファ、共用バッファなど) のようなバッファ構成) を処理
するためのスケジューラの使用に集中している。

【 発明の開示 】

【 発明が解決しようとする課題 】

【 0 0 1 6 】

トラヒックを制御し、フィードバック制御 T C P ネットワークの性能を改善するための
公平キューイングスキームの実現が強く望まれている。

10

20

30

40

50

【 0 0 1 7 】

更に、TCP接続に関する相当のスループットを可能にするパケットドロップ機構を実行する公平キューイングスキームの実現が強く望まれている。

【課題を解決するための手段】

【 0 0 1 8 】

前記課題は、本発明による、フィードバック制御TCPネットワークにおける公平キューイングスキームと共に併用されるフロー/接続毎の共用バッファ管理スキームにより解決される。

【 0 0 1 9 】

本発明によれば、(1)長ラウンドトリップ時間を伴う接続に対するTCPの固有の不公平性を軽減する、(2)異なるTCPバージョンを用いる接続がボトルネックリンクを共用する場合、分離性を与える、(3)一層侵略的なトラフィックソース、不品行なユーザ又はリバースパス輻輳の場合のその他のTCP接続からの保護性を与える、(4)両方向トラフィックの存在下におけるACK圧縮作用を軽減する、(5)ACK損失(トラフィックのバーストを起こす)を経験するユーザに顕著な悪影響を及ぼすその他の接続から保護する、及び(6)全体的なリンク利用率を低下させることなく、“欲張り”接続を伴うボトルネックを共用する双方向接続に対する低待ち時間を与える、などのようなTCPの種々の目標を達成することができる。

【 0 0 2 0 】

更に詳細には、共用バッファアーキテクチャは、各接続のための帯域幅予約保証 r_i により実行される。速度 r_i のそのバッファを完全に使用する第1の接続と、十分に利用されない(無駄使いされる)、速度 r_2 の第2の接続があり、第1の接続が更に一層の帯域幅を必要とする場合、共用バッファスキームでは、 r_2 接続からバッファリング(帯域幅)を借りることもできる。

【 0 0 2 1 】

第2の接続がそのバッファスペースを矯正しなければならない場合、別の利用バッファからのデータを押し出し、着信データパケットのための場所を空けなければならない。

【 0 0 2 2 】

接続毎の待ち行列(キュー)管理スキームは、共用バッファアーキテクチャにおけるパケットドロップメカニズム(例えば、最長待ち行列ファースト(“LQF”))をサポートし、FIFO-REDバッファ管理スキームよりも優れたTCP性能を生じる。本明細書では、公平性は、全リンク容量に対する個々のスループットの割合の平均値に対する標準偏差の比により計算する。

【発明を実施するための最良の形態】

【 0 0 2 3 】

図2は、様々なソース $s_1, \dots, s_2$ から発信するパケットトラフィックを処理するTCPネットワーク接続のルータ65のための接続毎の待ち行列アーキテクチャを示すブロック図である。

【 0 0 2 4 】

ネットワーク接続要素は、汎用の共用バッファメモリBを有する。このメモリBは、速度Cでリンク55上のデータパケットを処理するスケジューラ75と単一(ボトルネック)のリンク55との接続のために、“i”個の複数の待ち行列30a, ..., iに分割されている。各バッファ接続iは、接続iの保証バッファサイズである名目バッファ割当て b_i を有する。“フロー毎のキューイング”又は“接続毎のキューイング”として知られているこのアーキテクチャでは、待ち行列に入れられるべき到着パケットのきめ細かい動的分類が必要である。

【 0 0 2 5 】

“ソフト状態”アプローチは、潜在的に非常に多数のおそらく活性な接続(この場合、大多数の接続は実際には活性ではないであろう)へ導く接続状態を維持する。ネットワークノード内の関連状態はその他のガーベジ・コレクション(ごみ集め)の手段によりタイ

10

20

30

40

50

ムアウト（時間切れ）を起こさせるか、矯正させるか又は遅延させるために待機される。従って、スケーラビリティ(scalability)及び共用は、接続毎のサーバーのための基本的な要件である。必要な全ての操作は、0 (1) 複雑さにより実行可能であり、リソース（バッファ又は帯域幅）は所定の接続には静的に割当てられない。

【 0 0 2 6 】

操作中、スケジューラ 7 5 は、各待ち行列について同等であるか又は所定の重みに応じた、 r_i に等しい速度で各個別待ち行列 i を処理する。速度 r_i で広い帯域幅を使用する特定の待ち行列（接続）（例えば、待ち行列 3 0 a）は、その他の待ち行列よりも長い待ち行列を有する傾向がある。全ての待ち行列接続が、それらの各帯域幅割当を完全に利用している場合、待ち行列 3 0 a はおそらくバッファオーバーフローを経験するであろう。

10

【 0 0 2 7 】

或る割当てメモリ（例えば、3 0 c）が完全に利用されていない場合、本発明の公平キューイング（FQ）スキームは、待ち行列 3 0 b に到着するデータパケットのために、必要に応じて、十分に利用されていない待ち行列からバッファスペースを利用するか又は借りることができる。従って、バッファ待ち行列 3 0 a の予約割当て b_i を越えることができる。

【 0 0 2 8 】

第 1 の高速バッファ待ち行列 3 0 a のためのパケットを受信する第 2 のバッファが満杯になると、別の十分に利用されていないバッファ（例えば、待ち行列 3 0 i）は、待ち行列 3 0 a に宛てられる新たなデータパケットにバッファスペースを貸す。

20

【 0 0 2 9 】

同時に 2 つ以上の待ち行列がバッファオーバーフローを経験することもでき、そのため、十分に利用されていないバッファスペースから借りることもできる。従って、2 つ以上の待ち行列がその予約割当て b_i を越えることもできる。本発明は、TCP 接続ネットワークのフォワード及びリバース接続の両方に適用可能であり、また、データ及びACKトラヒックフローの両方に同等に適用可能である。

【 0 0 3 0 】

接続 i が $(b_i + 1)$ 個以上のバッファを必要とする場合、総占有率が B 未満であるとすれば、使用可能なプールからスペースが割当てられる。

【 0 0 3 1 】

図 3 (a) は、本発明の接続毎の待ち行列スキームを実行するための一般的な流れ図である。ステップ 2 0 1 に示されるように、接続 i について宛てられたパケットが到着すると、最初のステップ 2 0 3 で、バッファ B の現に残っている使用可能なスペースを有するカウンタをチェックし、新たに到着したパケットを確実に収容するための十分なメモリスペースがバッファ内に残存しているか否か決定する。

30

【 0 0 3 2 】

新たに到着したパケットを確実に収容するための十分なメモリスペースがバッファ内に残存している場合、ステップ 2 0 8 で処理が継続され、到着パケットが属する接続 i を識別し、そして、ステップ 2 1 1 で、このパケットを、この接続に対応する待ち行列に格納する。このスキームに潜在的に含まれることは、到着パケットが適正に分類され、そして、待ち行列のうちの一つに割当てられることである。

40

【 0 0 3 3 】

ステップ 2 0 3 において、新たに到着したパケットを確実に収容するための十分なメモリスペースがバッファ内に残存していないと決定される場合（例えば、現在の占有率 b_i が b_i 未満である待ち行列 3 0 j がバッファを必要とする場合）、ステップ 2 2 5 において追出しスキームが実行され、到着パケットのために場所を空ける。特に、呼出された TCP プロトコルの実行に応じて、追出しが行われる待ち行列を選択するための 2 つの方法が使用できる。

【 0 0 3 4 】

特に、図 4 (b) に示されるように、第 1 の実施態様における追出しメカニズムは、L

50

Q F スキームであり、別の待ち行列から予約された最大量のメモリを借りる待ち行列、すなわち、 $(q_i - b_i)$ が全ての接続を通して最大であるような接続 i を選択する。例えば、ステップ 250 に示されるように、バッファ内の各待ち行列の現在のバッファ割当てに関する決定が行われる。

【0035】

次いで、ステップ 260 で、各待ち行列の現在の待ち行列長さ q_i を取得し、ステップ 270 で、各待ち行列の $(q_i - b_i)$ 差に関する計算を行う。最後に、ステップ 275 において、最大の $(q_i - b_i)$ 差を有する待ち行列を選択する。従って、その予約割当て b_i からの最大偏差は最長待ち行列であり、このため、図 3 (a) のステップ 226 で示されるように、到着パケットと先入れ相関関係を有する待ち行列からパケットがドロップされる。 10

【0036】

当業者ならば前記の最長待ち行列を最初に追出すスキーム (L Q F スキーム) を実行するためのその他のアルゴリズムを案出することもできるので、本発明は図 4 (b) に示された方法論に限定されない。例えば、最長遅延ファースト ("L D F") ドロップینگメカニズムは、割当て処理速度 r_i が全て等しい場合、L Q F スキームと同等に実行できる。なぜなら、待ち行列が同じ速度で処理される場合、遅延は各接続について同一だからである。同様に、処理速度が均一でない場合、待ち行列長さが同一であっても、遅延は異なる。従って、L Q F スキームは L D F の特例である。

【0037】

第 2 の最長待ち行列よりもかなり長い待ち行列を有する多数の接続を有するシステムで実行する場合、最長待ち行列ファースト (L Q F) は過度のバースト損失をもたらすことがある。すなわち、2 個以上の非常に接近したパケットは、その割当てを越える待ち行列から連続的にドロップされる。例えば、T C P - R e n o 型処理系はバースト損失の存在下で不良動作を行うことが知られているので、T C P - R e n o アーキテクチャを実行する接続の性能は悪影響を受ける。従って、前記 L Q F におけるバースト損失量を低下するために、スキームは、それぞれの割当てを越えるバッファからランダムに抽出するランダムジェネレータを使用するように変更されている。 20

【0038】

特に、第 2 の実施態様では、各バックログ接続 (backlogged connection: 予備接続) i は、 $b_i = B / n$ (ここで、 n はバックログ接続の個数である) で示される名目バッファ割当てを有する。図 5 (c) に示されるように、ステップ 255 で、各待ち行列のメモリ割当て b_i を取得する。次いで、ステップ 265 で、バックログ接続を例えば、2 個のサブセットにグループ分けする。一方は、 b_i よりも大きな占有率 b_i を有するグループであり、他方は、 $q_i - b_i$ の関係を有するグループである。 30

【0039】

割当てより上 (すなわち、 $q_i > b_i$) の待ち行列群から、ステップ 285 で示されるように、1 つの待ち行列がランダムに選択され、そして、ステップ 226 で示されるように、1 つのパケットが先頭からドロップされる。

【0040】

異なる接続に関するバッファ占有率を平坦化し、そして、システムに過負荷をかける接続から最適に保護するための試みにおいて、L. Georgiadis, I. Cidon, R. Guerin, and A. Khamisy, "Optimal Buffer Sharing," IEEE J. Select. Areas Commun., vol. 13, pp. 1229-1240, Sept. 1995 に記載されるような開ループトラヒックについて実行されるスキームと同様な方法で、特定の追出し (pushout) スキームは最長待ち行列の先頭からパケットをドロップする。 40

【0041】

図 6 に示されるように、ハードウェアの観点から、カウンタ 90 a, ..., 90 i は、それぞれの付随待ち行列の占有率 b_i を保持するようにプログラムされたコントロールプロセッサ 92 と共に、各待ち行列 30 a, ..., 30 i に関連するように図示されて 50

いる。従って、図4(b)及び図5(c)のステップ260及び265において、現在の待ち行列長さは、例えば、ポーリングにより、又は、図6のテーブル99のようなレジスタテーブル(ここで、レジスタは最長待ち行列長さを示す)から位置指定することにより取得される。

【0042】

別のカウンタ95は、総バッファ占有率Bのトラックを保持するために図示されている。従って、パケットが到着すると、プロセッサ92は、カウンタ95で使用可能な総メモリを決定するために、ステップ203(図3(a)参照)におけるチェックを行う。

【0043】

最長待ち行列を決定するために、プロセッサ92はソート済み構造を実行することもできる。例えば、各キューに入れる、キューから外す又はドロップ操作の時点で、待ち行列の対応カウンタがそれに応じて増分又は減分された後、最長待ち行列構造を常に知っておくために、その待ち行列占有率値 q_i を現在の最長待ち行列占有率値と比較するように、ソート済み構造を実行する。

【0044】

これらフロー毎のスキームの変法は、指数関数的に増大するサイズを有する一連のピンを使用することにより効率的に実行することができる。待ち行列が変更される場合(例えば、キューに入れる、キューから外す又はドロップする場合)は常に、システムは最高占有ピンのトラックを保持しながら、適当なピン(例えば、現在のピンと同一のピン、一つ上のピン又は一つ下のピン)に移動させる。

【0045】

バッファが満杯の場合、最高占有ピンからの待ち行列が選択され、その最初のパケットがドロップされる。バッファがバイト単位で測定される場合、この操作は、可変パケットサイズのため、新たに到着したパケットを収容するために十分なスペースが解放されるまで、繰り返さなければならない。真のLQDを行うために、最高占有ピンにおける待ち行列をソート済みリストとして維持するか、又はいつも最長待ち行列をサーチしなければならない。

【0046】

FIFO-RED及びFQ-REDスキームと比較される本発明のバッファ管理スキームの改善された性能のシミュレーションについて説明する。図7は、このシミュレーションシステム119を示す概要図である。図7において、ソース101a, ..., 101fはルータ120への高速アクセスパスを有する。ルータ120は唯一のボトルネックであり、本発明の接続/フロー毎のバッファ管理スキームを実行するものである。

【0047】

比較は、TCP-Tahoe及びTCP-Renoソース、バースト及び欲張りソース(greedy source)、リバースパス輻輳を有する片方向及び両方向トラヒックソース、及び広範囲に異なるラウンドトリップ時間の混合を用いて行われる。アクセスパス遅延は、様々なラウンドトリップ時間のモデルを作るために広範囲にわたって設定され、宛先はパケット毎にACKと仮定した。

【0048】

片方向トラヒックの場合、ACKは非輻輳パスを介して送信される。両方向トラヒックの場合、リターンパスはルータを経由し、そのため、キューイング(待合わせ)遅延が生じることがある。特に、ルータがFIFOスケジューリングを使用する場合、ACK及びデータパケットは待ち行列内で混合される。公平キューイングの場合、ACKは別のフローとして処理される。非対称トラヒックの場合、リターンリンクの帯域幅は、宛先からルータまで低下される。そのため、相当なリバースパス輻輳とACK損失が生じる。

【0049】

シミュレーションシステム119におけるTCP-Tahoe及びTCP-Renoの処理系は、4.3-Tahoe BSD及び4.3-Reno BSDのTCPフローと輻輳制御挙動とをそれぞれモデル化する。REDモデルは有向パケットであり、最小閾値

としてバッファサイズの25%を使用し、最大閾値としてバッファサイズの75%を使用する。待ち行列重みは0.002である。

【0050】

図8(a)及び(b)は、従来技術によるFIFO-RED及びLQF法と比較されるような、利用率及び公平性のそれぞれの観点から、擬似TCPネットワークで公平待ち行列LQD及びRNDドロップ法を実行した場合の、改善された性能を示す特性図である。

【0051】

特に、図8(a)及び(b)は、図7の擬似ネットワークにおける10Mbps/100kbps容量(TCP-Tahoe及びTCP-Renoの20ms-160msラウンドトリップ時間)を有する非対称ボトルネックリンクによる20個のTCP接続のためのバッファサイズ(で表したリンク速度)の関数として、スループット(図8(a))及び公平性係数(図8(b))により評価される改善されたTCP性能を示す。

10

【0052】

図示されているように、ライン137a及び138aでそれぞれ示されるFQ(公平キューイング)-RED及びFIFO-REDの両戦略は、ライン139及び140で示されるFQ-LQF及びFQ-RNDドロップ法よりも非常に低いスループットを有する。なぜなら、再送パケットに対応するACKが、擬似非対称値が3(これは帯域幅非対称ではない)の場合の時間が66%も喪失されるからである。これは、TCPサイクルの少なくとも66%でタイムアウト(時間切れ)を生じ、スループットを著しく低下させる。

【0053】

フォワードパスにおける多数の損失及びフォワードパスにおける再送パケットの損失のために、その他のタイムアウトも起こる。一方、リバースパスにおける先頭からのドロップはこれらのタイムアウトを殆ど完全に除去する。タイムアウトは不経済なので、両方のREDスキームとも、FIFO-LQDを含むその他のスキームよりも著しく低いスループットを有する。

20

【0054】

更に、図8(b)に示されるように、FQ-RND及びFQ-LQDワークの両方とも非常に良好である。なぜなら、これらは、フロー毎の待ち行列の効果と、先頭からのドロップのタイムアウト除去効果とを併有するからである。FQ-LQDは、再送パケットのドロップに対する組込みバイアスを有するという別の効果も有する。これは、ソースが最初の重複ACKの受信による損失を検出したときに、パケットの送信を中止するからである。

30

【0055】

第3の重複ACKが受信された後にだけ、再送パケットが送信される。介入間隔中、ソースがTCPにより強制的に不活動化される場合、フローに対応する待ち行列は、少なくともその最小保証速度で排出される。従って、再送パケットが到着する場合、最長待ち行列であることは殆どない。従って、再送パケットのドロップングに対する固有バイアスである。このバイアスは非対称ネットワークに限定されないが、低速リバースチャネル膨張のために、非対称要因、第1の重複ACK受信と第3の重複ACK受信との間の間隔により、バイアスは非対称ネットワークにおいて高められる。

40

【0056】

再送パケットの損失は不経済なタイムアウトを起こすので、このバイアスは、図8(b)においてライン141で示されるように、FQ-LQDの性能を改善する。ライン142で示されるFQ-RNDも同様にこのバイアスを有するが、著しく低い。その理由は、FQ-LQDに対するものと若干類似している。

【0057】

第1の重複ACK受信と第3の重複ACK受信との間の間隔中に、フローの待ち行列は、(ソースは不活動化されているので)少なくともその保証速度と等しい速度で流出し、待ち行列占有率はこのフローの予約パラメータ以下にまで落ちるであろう。この場合、再送パケットが到着すると、集合バッファが満杯であったとしても、再送パケットは失わ

50

れない。これらの効果より、F Q - L Q D 及び F Q - R N D は最高の性能を有する。

【 0 0 5 8 】

[発明の効果]

以上説明したように、本発明によれば、(1)長ラウンドトリップ時間を伴う接続に対する T C P の固有の不公平性を軽減する、(2)異なる T C P バージョンを用いる接続がボトルネックリンクを共用する場合、分離性を与える、(3)一層侵略的なトラヒックソース、不品行なユーザ又はリバースパス輻輳の場合のその他の T C P 接続からの保護性を与える、(4)両方向トラヒックの存在下における A C K 圧縮作用を軽減する、(5) A C K 損失 (トラヒックのバーストを起こす) を経験するユーザに顕著な悪影響を及ぼすその他の接続から保護する、及び (6) 全体的なリンク利用率を低下させることなく、“ 欲張り ” 接続を伴うボトルネックを共用する双方向接続に対する低待ち時間を与える、などのような T C P の種々の目標を達成することができる。

10

【 図面の簡単な説明 】

【 0 0 5 9 】

【 図 1 】 従来技術による T C P ネットワーク接続を示すブロック図である。

【 図 2 】 多数の T C P 接続のための共用バッファアーキテクチャのブロック図である。

【 図 3 】 (a) はバッファ割当て及びパケットドロップスキームを行うために履行される方法を示す流れ図であり、図 3、図 4、図 5 で 1 つの流れ図を形成する。

【 図 4 】 (b) は L Q F パケットドロップ法を示す流れ図である。

【 図 5 】 (c) は R N D パケットドロップ法を示す流れ図である。

20

【 図 6 】 各バッファ待ち行列に関連するハードウェアを示すブロック図である。

【 図 7 】 公平キューイングを処理するルータを有する T C P / I P ネットワークのシミュレーションを示すブロック図である。

【 図 8 】 (a) は図 7 の擬似ネットワークにおけるボトルネックリンクによる 2 0 個の T C P 接続のためのバッファサイズ (で表したリンク速度) の関数として、スループットにより評価される改善された T C P 性能を示す特性図である。 (b) は図 7 の擬似ネットワークにおけるボトルネックリンクによる 2 0 個の T C P 接続のためのバッファサイズの関数として、公平性係数 (全リンク容量に対する個々のスループットの割合の平均値に対する標準偏差の比) により評価される改善された T C P 性能を示す特性図である。

【 符号の説明 】

30

【 0 0 6 0 】

3 0 待ち行列 (キュー)

5 5 ルータ

6 5 ルータ

7 5 スケジューラ

9 2 コントロールプロセッサ

9 5 カウンタ

9 9 レジスタテーブル

1 0 1 ソース

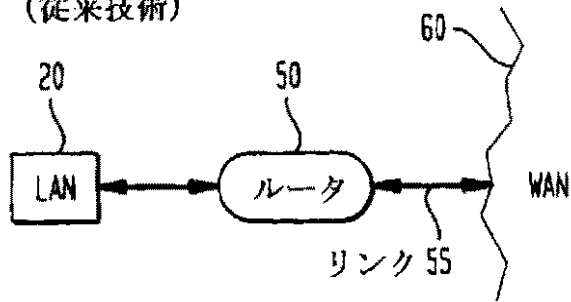
1 1 9 シミュレーションシステム

1 2 0 ルータ

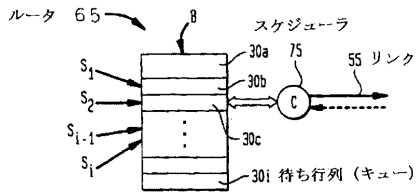
40

【図 1】

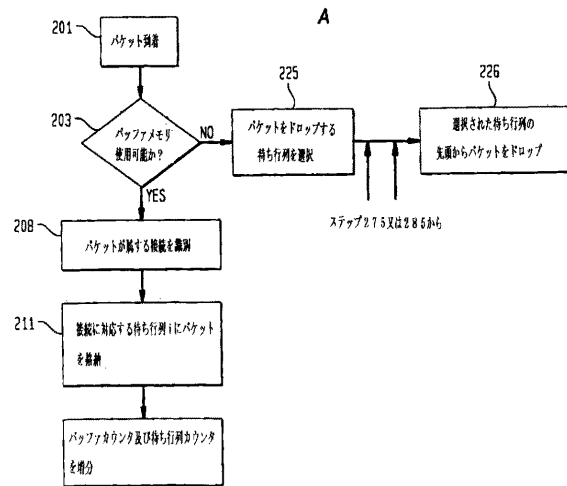
(従来技術)



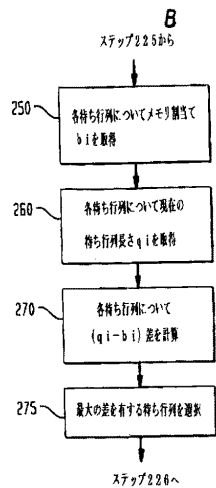
【図 2】



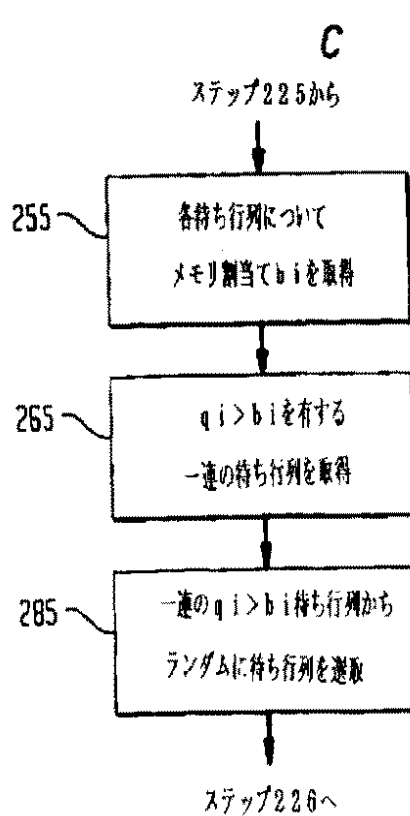
【図 3】



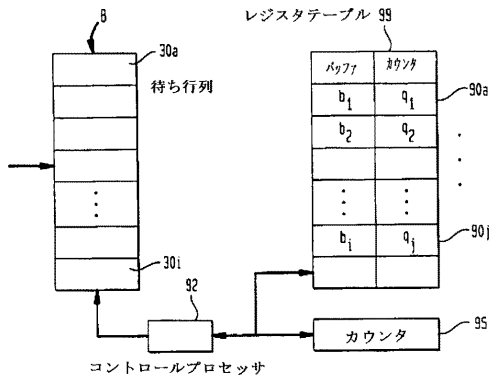
【図 4】



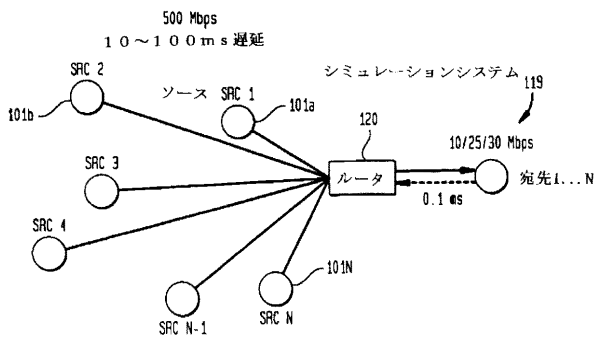
【図 5】



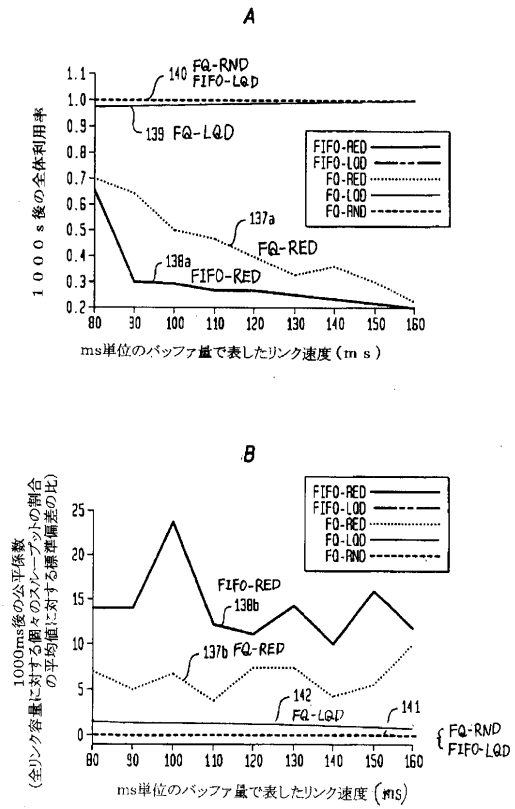
【図 6】



【図 7】



【図 8】



フロントページの続き

- (74)代理人 100101498
弁理士 越智 隆夫
- (74)代理人 100096688
弁理士 本宮 照久
- (74)代理人 100104352
弁理士 朝日 伸光
- (72)発明者 アビジット カマー チョーデュリー
アメリカ合衆国, 0 7 9 2 0 ニュージャージー, スコッチ プレインズ, ハンティントン ロード 77
- (72)発明者 ティー・ヴィー・ラクシュマン
アメリカ合衆国, 0 7 7 2 4 ニュージャージー, イートンタウン, ヴィクトリア ドライブ 118
- (72)発明者 ディミトリオス スティリアディス
アメリカ合衆国, 0 7 0 3 0 ニュージャージー, ホボケン, パーク アヴェニュー 106
- (72)発明者 バーナード シュター
アメリカ合衆国, 0 7 7 4 7 ニュージャージー, アバーディーン, アイダホ レイン 12
- Fターム(参考) 5K030 GA03 HA08 HD03 HD06 KA05 LB05 LC02
5K034 AA02 DD01 EE11 HH21
5K201 AA01 BB05 EA05 EC01 EE07