複数パスを用いたトラヒックの動的振り分け方式の研究

2005MT013 福村 卓哉 2005MT083 小川 功 2005MT112 高橋 幸裕

指導教員 奥村 康行

1. はじめに

近年，ネットワーク化が進み，大容量データの FTP 転送，映像配信やテレビ会議を始めとする広帯域転送を必要とするアプリケーションが登場し，トラヒック量が増加してきている．トラヒックの増加とともに，データ転送の同時性，即時性が強く求められる傾向も強まっている．

そこで私たちは複数の物理的な回線を仮想的に束ね，あたかも 1 本の回線であるかのように扱う技術であるリンクアグリゲーションに注目した．つまり，1Gbps の回線 5 本を 5Gbps の回線 1 本とみなす．

リンクアグリゲーションにおいて，コネクションがリンクを選択する方法が重要となるが，先行研究により動的 L2 振り分け方式が提案されている．IEEE802.3ad で定義されているリンクアグリゲーションでは，コネクションのリンク選択方法に関しては言及しておらず，従来のリンク選択方式では高優先トラヒックの保護が難しいことから動的 L2 振り分け方式が提案された．この動的 L2 振り分け方式では，それぞれのリンクの帯域幅を一定として振り分けを行っている．

本研究では，先行研究で提案された振り分けアルゴリズムを改良し，帯域の異なるリンクを束ねて 1 本のリンクとして扱うための振り分けアルゴリズムを新たに提案する．また，コンピュータシミュレータとして実際のネットワークにより近い環境でシミュレーションができる Network Simulator ver.2 (NS-2)[1]を用いて，動的 L2 振り分け方式と，新たに提案する振り分けアルゴリズムにおけるパラメータの検証を行う．

2. 動的 L2 振り分け方式

2.1 リンクアグリゲーション

IEEE802.3ad で定義されている技術で，2 つのノード間に張られた複数のリンクを束ね，仮想的に一本のリンクとみなす技術である(図 1)．これはリンクの広帯域化と冗長化を実現する技術で，ある 1 本の回線が障害によりダウンした際などに，束ねられている残りのリンクでデータの送受信を続行できるというメリットを持つ．この技術の利用に際しては，使用する複数のリンクが全て同じ帯域であるという前提がある．

図1 リンクアグリゲーション

2.2 動的振り分け方式の主要構成要素

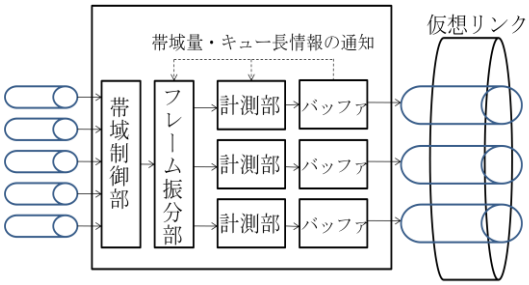

図2 動的L2振り分け方式を実現する回路モデル

動的 L2 振り分け方式の主要部は，帯域制御部，フレーム振り分け部，計測部及びバッファである．帯域制御部は各コネクションの優先度等の設定に基づいて集線し，複数のリンクを束ねた仮想リンクの帯域でフレームを読み出す．フレーム振分部はコネクションの優先度とリンクの使用状況に応じて，各リンクへ振り分ける．計測部とバッファは，各リンクの使用帯域，コネクション数及びキュー長を監視し，リンク変更に必要な情報をフレーム振分部に通知する．

2.3 要求条件

ノード間のリンクにおいてトラヒックの増加に対応し，高品質のサービスを提供するためには 4 つの必要要求条件があげられる．

(a) 既存設備の利用
(b) 帯域使用効率の向上
(c) 高優先トラヒックの保護
(d) TCP コネクションの公平化（フェアネス）

2.4 既存のリンク選択方式

リンクアグリゲーションにおいて，各コネクションが使用するリンクを複数のリンクの中から選択する方法として，リンク固定方式とリンク変動方式の2つの選択方式がある．

リンク固定方式（図3）は使用するリンクを固定し，常に同じリンクを使用する方式である．

リンク変動方式（図 4）はコネクションが使用するリンクを固定することなく，使用するリンクを動的に選択し，リンクを変更する．また複数リンクを並列に使用する方式である．

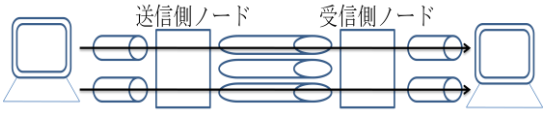

図3　リンク固定方式

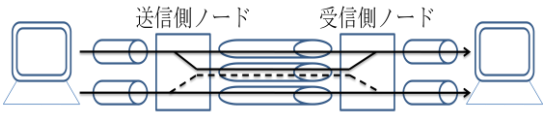

図4　リンク変動方式

しかし，これら2つのリンク選択方式は，すべてのトラヒックにたいして均一に制御しようとするため，既存設備の利用，帯域使用効率の向上，高優先トラヒックの保護，TCPコネクションの公平化という4つの要求条件を同時に実現することはできない．以下にその理由を示す．

2.4.1　リンク固定方式

送信側のみで制御し，既存設備が利用できるが，使用帯域に偏りが生じた場合，リンク全体の帯域使用効率が低下する．そのため，高優先トラヒックの保護を実現できるが，帯域使用効率の向上を実現できない．

2.4.2　リンク変動方式

リンク変動方式は，同時に複数リンクを並列して使用することで帯域使用効率が向上する．リンク変更を行う際にフレームの損失が生じる可能性あり，スループットが低下し，遅延が増大する．そこでこれらに対する制御として無制御，送受信側制御，受信側制御，送信側制御の 4 つがあげられる．

①無制御 ： フレームの損失を許容する方法．既存設備が利用できるが，スループットが低下し，高優先トラヒックも保護できない．

②送受信側連携制御 ： 送受信側で連携する方法．連携してリンク変更をするため，制御遅延が発生し，高優先トラヒックを保護できない．

③受信側制御 ： 受信側で制御を行う．高優先トラヒックを保護できるが，既存設備を利用できない．

④送信側制御 ： 送信側で制御を行う．既存設備を利用できるが，受信側で制御を行わないため高優先トラヒックを保護できない．

2.5　先行研究のリンク変更アルゴリズム

高優先トラヒックはリンクの状態に関わらずリンクを固定する．低優先トラヒックは一部のリンクにトラヒックが集中した場合，バッファオーバーフローが発生するためリンク変更を行う．リンク変更を行う際，変更元リンク，リンク変更コネクション，変更先リンクを決定する必要がある．変更先リンクの決定方法は，使用帯域の最も少ないリンクが変更先のリンクの候補になる．動的 L2 振り分.け方式では，フレーム順序逆転を許容する．フレーム順序逆転によるスループットの低下を最小限に抑え帯域使用効率を向上させるため，過剰なリンク変更は回避する．

閾値を 2 つ設けて th1 と th2 とする．リンクのバッファのキュー長がこれらの閾値を越えたときリンク変更を行う．また，リンク変更によるスループットの低下の影響を公平にするために，コネクション間で変更を行う回数を均等にする．そのため，リンク変更頻度の低いコネクションを選択する．キュー長が th2 を越えた時には，強制的にリンク変更を行う．

閾値 th1 が低い時，変更先リンクの輻輳を早期に防ぐことが可能で遅延を抑えることができる．一方閾値 th1 が高いと過剰なリンク変更を避けることができる．
これらの方法が先行研究[2]で研究された．

3.　複数パスによる動的振り分け方式

束ねるリンクの個々の帯域幅を変更するため，リンクアグリゲーションの枠を越え，複数パスへのコネクションの振り分けという形にネットワーク構成を変更し，その上で新たに複数パスによる動的振り分け方式のアルゴリズムを適用する事を提案する．

3.1　ネットワークの構成

本研究ではリンクが同帯域であったのを，帯域幅を変えて異なる帯域幅を含む動的振り分け方式を提案する．先行研究のアルゴリズムでシミュレーションを行うと，帯域幅の大きいリンクにも同量のコネクションしか振り分けられない．そこで帯域幅の異なるリンクに適切なコネクション数を振り分けるようなアルゴリズムを提案する．

3.2　計測用 TCP コネクション

ここであげるアルゴリズムは，リンクに含まれる TCP コネクションの 1 本を，図 5 で示すように計測用 TCP コネクションに変更する．他のデータ転送 TCP コネクションは転送をランダムに開始･終了するが，計測用 TCP コネクションはシミュレーション開始からシミュレーション終了まで張っておき，データ転送には使用せず計測のためだけに使用し，リンク変更を行わないよう高優先トラヒック扱いとする．

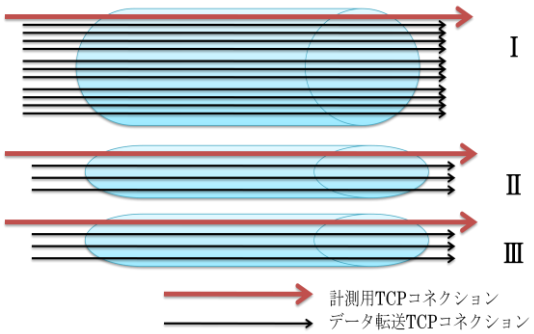

図5　計測用TCPコネクション

この新たに提案するアルゴリズムでは，以下の手順で変更先リンクを選別する．

ⅰ） 計測用コネクションのスループットを計測する．
ⅱ） スループットが最大のコネクションを選出する．
ⅲ） 選出したコネクションの所属リンクを確認する．
ⅳ） そのリンクを変更先リンクとする．

TCP ではフローの数に関係なく帯域を最大限に利用しようとするため（RTT，Window による制限があるため，最大限までは利用することができないが，より多くの帯域幅を使用しようとする），あるリンクにおいて，流れるフローが増えればコネクション一本一本のスループットは低下し，フローが減ればコネクション一本一本のスループットは上昇する．したがって，リンク帯域に余力があればあるほど，コネクション一本毎のスループットが大きい値を示すことになる．そのため，計測用 TCP コネクションからスループットが最大のものを選出し変更先リンクとする．

先行研究のアルゴリズムではリンクの帯域幅は全て同じ帯域に固定せざるを得ないが，新たに提案するアルゴリズムでは，各リンクの帯域幅にかかわらず（全て異なるリンクを使用したとしても）コネクションは平等に振り分けられる．また，この新たに提案するアルゴリズムは UDP にも適応できる。

3.3　シミュレーションのパラメータ

以下のようにパラメータを設定し，先行研究のアルゴリズムと新たに提案するアルゴリズムを用い，実験を行う．

リンク本数を 3 本とし，その帯域幅をそれぞれ Link1 では 1Gbps,Link2 では 100Mbps，Link3 では 10Mbps に設定し，フロー数を 70 とする．その他のリンク遅延，キュー長閾値 th1，th2，リンク変更周期に関しては，すべて先行研究と同じ値を用いて，遅延 16ms，キュー長閾値 th1 を 15frame，th2 を 100frame，リンク変更周期を 5ms のままで，シミュレーション時間は全て 600 秒とする．

3.4　リンク毎のスループットの比較

この節では 3.3 節で示したパラメータを用いて先行研究のアルゴリズムと新たに提案するアルゴリズムを使用した場合のスループットの時間遷移を示す．

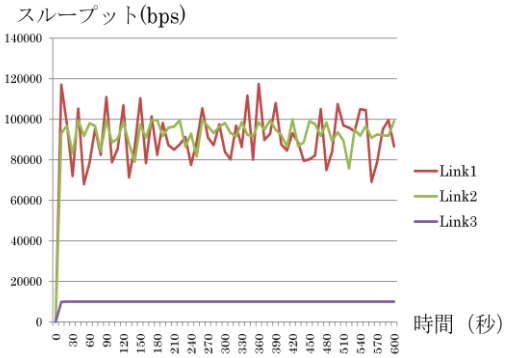

図6 先行研究のアルゴリズムを用いた場合のスループット特性

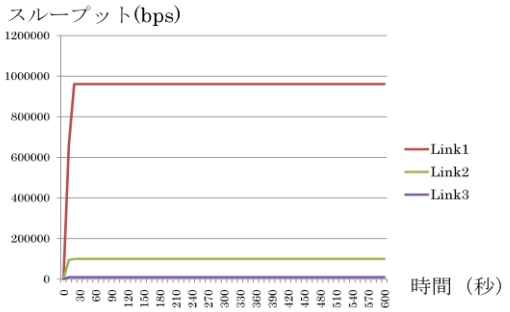

図7 新たに提案するアルゴリズムを用いた場合のスループット特性

図 6 は先行研究のアルゴリズムを適用した場合を示す．この実験結果では Link3 は限界までリンクが使用されている．しかし，Link 1 は帯域幅が 1Gbps あり、余力があるにも関わらず 100Mbps 程度しか使われておらず Link2 と同じような値を示している．

先行研究のアルゴリズムでは、スループットの最も小さいリンクを選出し、そのリンクを変更先リンクとする性質により，帯域幅 1Gbps の Link 1 と帯域幅 100Mbps の Link2 は変更先リンクの候補から外れることになる．

その結果変更先リンクはスループットの小さい帯域幅 10Mbps の Link3 となり，結果的に Link1，Link2 に振り分けられるコネクションが減る．

図 7 は新たに提案するアルゴリズムを適用した場合のスループット特性を示す．これをみると全てのリンクでほぼ最大限まで帯域が利用されていることが分かる．これは新たに提案するアルゴリズムでは，計測用 TCP コネクションからスループットが最大のものを選出し変更先リンクとする性質によりリンク変更の選択が行われたためである．また，それぞれのリンクのスループットの変動がほとんどないという特徴もみられる．

3.5 フレームロス

この節では 3.3 節で示したパラメータを用いて，先行研究のアルゴリズムと新たに提案するアルゴリズムを使用した場合のフレームロス数を示しそれらについて考察する．

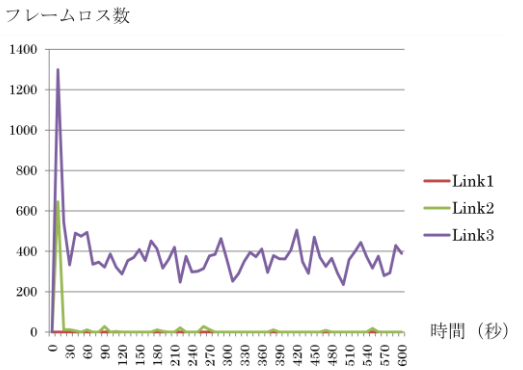

図8 先行研究のアルゴリズムを適用した場合のフレームロス数

図 8 は先行研究のアルゴリズムを適用した場合のフレームロス数を示す．Link 1 でのフレームロスが皆無であるが他の 2 つのリンクではフレームロスが増大している．これは Link 1 に事実上余力があるにも関わらずフローが他のリンクへと振り分けられ，結果としてバッファオーバーフローが発生したためと考えられる．新たに提案するアルゴリズムを適用した場合のフレームロス数は，全てのリンクでフレームロス数は 0 であった．この理由は各リンクへ公平に振り分けがなされたためと考えられる．

3.6 先行研究とのアルゴリズムの比較

この節では 3.4 節，3.5 節の結果をもとに，先行研究のアルゴリズムと新たに提案するアルゴリズムの比較を行う．

先行研究のアルゴリズムでは，スループットが安定せず余力のあるリンクへの振り分けが行われないため，Link1 は帯域幅が 1Gbps あるにもかかわらず、100Mbps 程度しか使われておらず，Link 2 と Link 3 のフレームロス数が多くなっている．

それに対して，新たに提案するアルゴリズムでは，図 7 を見てわかるように全てのリンクにおいて物理帯域をほとんど使用している。これは新たに提案するアルゴリズムが全てのリンクに対して公平な振り分けをしたことにより，スループットが安定し，フレームロス数も 0 となった．このことから異なる帯域を束ねる時では新たに提案するアルゴリズムが有効なことが分かる．

4. コネクションの公平性

この章では新たに提案するアルゴリズムを用いた場合の計測用 TCP コネクションのみを抜き出したスループットの時間遷移を示す．各リンクにおいて計測用 TCP コネクションのスループットが全てのリンクでおよそ 5Mbps の値で一定であることからコネクションの公平性が保たれているといえる．

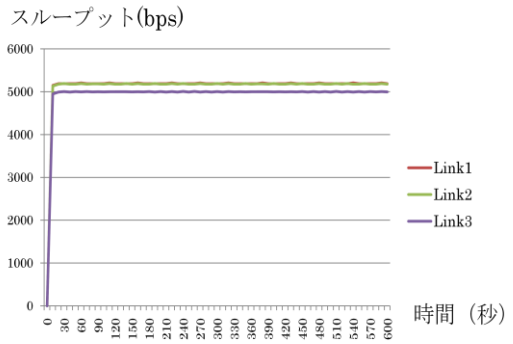

図9 新たに提案するアルゴリズムを用いた場合の計測用TCPコネクションのスループット特性

5. 今後の課題

本研究では振り分けのためのアルゴリズムの提案にとどまっており，各種パラメータの最適値を求めることができていない．リンク本数を変えた場合のキュー長閾値 th1，th2，リンク変更周期を含めた場合のパラメータの最適値の導出を行うため，他のネットワーク構成について研究する必要がある．

6. 参考文献

[1] 銭飛，“NS2 によるネットワークシミュレーション”，森北出版 (2006 年).

[2] 秦野 智也，笠原 康信，吉原 慎一，岩田 敏行，前田 洋一，“帯域使用効率向上させるコネクション優先度に基づく動的 L2 振り分け方式”，信学技報，CS2005-13，2005 年 8 月．