2007－DSM－44 (6)
2007／3／9

BGPの信頼性の向上を目指した新しい peering 手法について

有賀征爾　鈴木昭徳†

現在，バックボーンネットワークの経路制御方式として BGP が標準的に利用されている．しかし，BGP では経路制御メッセージの配送とデータトラフィックの配送を同一回線で行なっているため，データトラフィックが輻輳した際，経路制御メッセージの損失による大規模な通信途絶や通信不安定などの通信に著しい影響が発生する場合がある．そこで，経路制御メッセージを別回線で送受信することにより，輻輳時の経路制御の信頼性を向上する手法について，その実現性と問題点の検討を行なった．

New BGP peering method for increasing routing stability

Seiji Ariga　Akinori Suzuki†

Currently, BGP is a standard routing protocol operated in Internet backbone network. BGP is exchanging routing information and data traffic over the same physical channel. Thus, when there is a congestion over this channel, routing information exchange will be disrupted and it will cause routing instability and data loss. To resolve this issue, we've tried to separate routing channnel and data traffic channel. We will discuss about advantages and disadvantages of this method.

1. はじめに *

今日、インターネットはサービス・アプリケーションを動かすインフラとしての重要性をますます高めている。特に音声伝送(VoIP)や映像放送(ライブ放送、オンデマンド放送)などは、アクセスラインの広帯域化にともなって広く行なわれるようになった。これら実時間性が重要なアプリケーション(リアルタイムアプリケーション)にとってネットワークの安定性・疎通性は非常に重要である。そこで通信路を高速に切り替えるために BFD（Bi-directional Forwarding Detection）[1] などのプロトコルなども策定されている．一方，BFD などにより疎通性に障害が起こったことが検知されトラフィックが迂回される場合，大量のトラフィックが迂回してしまう．特にインターネットトラフィックの主要な交換ポイントとなっている IX (Internet Exchange) では数十 Gbps のトラフィックが交換されており，そのトラフィックが一度に移動してしまうと，IX につながっているルータ，およびそのルータにつながる AS（Autonomous System）内部のルータ・ネットワークに多大な影響を与えることとなる．

つまり，インターネットはこれまで必ずしも配送網として(リアルタイムの)安定性・疎通性を保証してこなかったが，これはある局所的な疎通性の問題を広範囲に伝播させないために，障害に対し必ずしもリアルタイムに対応をしないという設計上の選択でもあった．たとえばインターネットのバックボーンでは現在 BGP (Border Gateway Protocol) [2] を使って経路交換を行なっているが、主要なルータを既定値で使っている場合、BGP は通信の断検知に90秒から180秒かかる。通信断の原因には、機器の故障、回線の切断などがあるが、昨今では DDoS (Distributed Denical of Service)攻撃によるものも増えている。そこで本稿では、DDoS などの過大なトラフィックが IX へ流入することによって起こる輻輳が BGP の通信断などを引き起こすことによって、ネットワークに与える影響を検討する。次いで、次いで、そのような環境化においても安定的にデータ配送を継続するためのネットワークデザインを検討する。

2. IX における問題

現在，IX ではルータは BGP を用いて計制御情報の交換を行なっている．BGP では定期的（既定値では 30 秒から 60 秒に一回）に keepalive メッセージを送受信することで，通信相手への到達性を検証している．しかしトラフィックが一箇所に集中した場合，輻輳によりこの keepalive メッセージが IX のスイッチ内で失わことがある．その結果，通信経路に物理的な障害が無いにも関わらず，ルータとしては BGP の peering 相手との疎通性が失われたと判断し，トラフィックの迂回を始める．これ

*† NTT コミュニケーションズ株式会社
NTT Communications Corporation

は輻輳回避の手法としては有効であるが，一定時間の通信断をともなってしまう．トラフィックが迂回されるには，通信断の検知（複数回の keepalive の喪失により検知）時間と経路情報の再計算，その収束時間がかかる．これは一般に IX で使われている機器の場合，検知に 180 秒，収束時間に 30 秒程度かかる [3] ことを意味する．つまり，IX 内で輻輳が起こった場合おおよそ 3 分半程度通信が途絶してしまう．

しかし，トラフィックが迂回されることにより IX 内での輻輳が解決され，ふたたび BGP の keepalive が行なえるようになる．すると，BGP の peering が回復し再びトラフィックが戻ってくる．このようにして IX にトラフィックが集中した場合，人手による操作が行なわれない限り，先に述べたような大量のトラフィックが移動し続けてしまう場合がある．

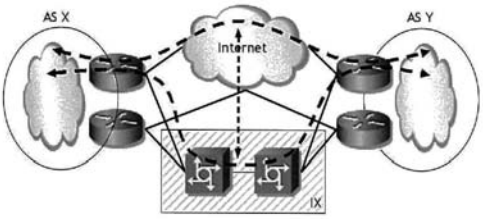

図 1　IX を利用した AS 間の接続

図 1 のように通常 IX で peering を行なっている AS 間は別の回線でも接続されているが，効率的にトラフィックを交換するために IX 経由での交換を優先している場合が多い．したがって，IX 経由での疎通性が回復するとトラフィックは IX を経由しようとする．

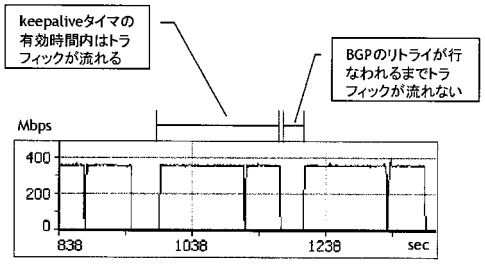

図2　実験環境における AS X－AS Y 間の IX 経由のトラフィック

図 2 は実験環境で図 1 の構成を模した上で IX に過大なトラフィックを流し，AS X と AS Y が IX 経由で交換するトラフィックをグラフ化したものである．定期的（BGP の keepalive タイマが切れるタイミング）にトラフィックが落ち込んでいることが分かる．図 1 のような環境では，通常 IX を経由して交換されている膨大なトラフィックはインターネットを経由することとなり，経路上の複数のネットワークに問題を及ぼす可能性がある．

このように輻輳検知のために BGP の peering が切断される，設計通りにトラフィックを迂回させることが，かえってネットワーク全体に問題を引き起こす場合がある．

この問題を解決するために，次に新しい peering を手法を検討する．

3.　IX における新しい peering 手法の検討

先に述べた問題は，BGP が経路制御メッセージの交換とデータトラフィックの交換を同一の物理リンクを用いて行なっており，データトラフィックの輻輳によって経路制御情報の交換ができなくなるため発生する．そこで本稿では経路制御メッセージの配送網とデータトラフィックの配送網を分離することを提案する．

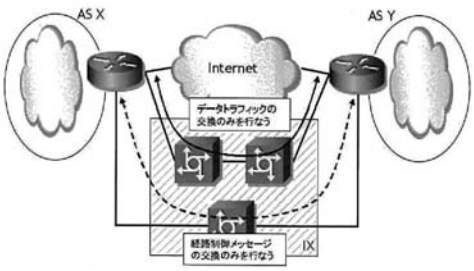

図 3　経路制御メッセージの配送網とデータトラフィックの配送網の分離

図 3 のようにデータトラフィックに用いられるスイッチ群とは別に，経路制御メッセージの配送網用にスイッチを用意する．IX に接続する各ルータからは，データトラフィックの配送に用いられるインタフェースとは別に，経路制御メッセージ配送用のスイッチに接続を行ない，そこで他の AS と BGP の peering を確立する．これにより、データトラフィックに影響されずに安定して経路制御メッセージを交換することができるようになる．経路制御メッセージによるトラフィックは，データトラフィックに比して非常に小さいため，経路制御メッセージ配送用のスイッチは，既存の IX において利用されているスイッチに比べて小容量のもので十分である．

3.1　本手法における障害検知

既存の peering 手法では装置故障，回線故障，輻輳のいずれの障害も peering の切断というかたちで検知が可能であった．しかし本手法ではデータトラフィック配送網が経路制御メッセージ配送網とは分離されているため，データトラフィック配送網に障害があった場合でも，経

路制御メッセージの配送には影響がなく，障害のある回線にデータトラフィックを流し続けてしまう．2. で示した問題は輻輳のみによってもトラフィックが移動してしまう，といったものであったが，本手法では回線障害があった場合でも当該回線にトラフィックを流し続けようとしてしまう．そこで，データトラフィック配送網経由でも peering を確立する．ただし，BGP の local-preference の値を下げるなどで経路制御メッセージ配送網経由の peering の方が優先されるよう設定する．これにより peering の切断という形でデータトラフィック配送網の障害を検知することが可能となる．ただし，あくまで障害検知のための peering であるため切断が起こってもデータトラフィックの配送には影響はない．

一方，当該 peering は経路制御メッセージ配送網経由の peering のバックアップとして機能することができる．本手法では既存の手法とは逆に，データトラフィック配送網に障害が起こっていなくても経路制御メッセージ配送網に障害が起こることで経路情報の交換ができなくなり，結果的にデータトラフィック配送網が使われなくなってしまう．そこで，データトラフィック配送網の peering をバックアップとすることで，経路制御メッセージ配送網に障害が起こってもロス無しでデータトラフィックの交換を継続することができる．

3.2 既存の手法との相互運用性

本手法では，既存の peering に新たに経路制御メッセージの交換用の peering を追加することによって経路制御メッセージの配送の信頼性を高めている．つまり既存の peering はそのままに，必要に応じて peering を追加することで本手法を実現することができるため，既存の手法との相互運用性に問題はない．

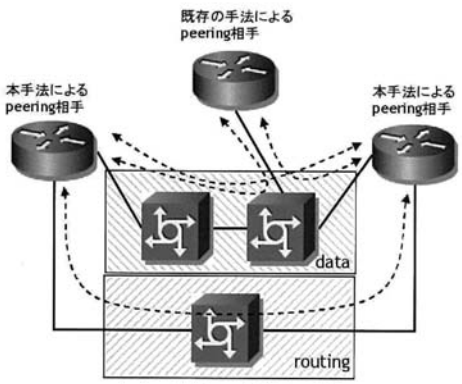

図4 既存の手法との共存

これによって，すべての peering を一度に移行することなく，徐々に本手法を導入できる．

4. 実験環境における本手法の検証

図5の環境において本手法の検証を行なった．図中の回線はすべて Gigabit Ethernet となっている．

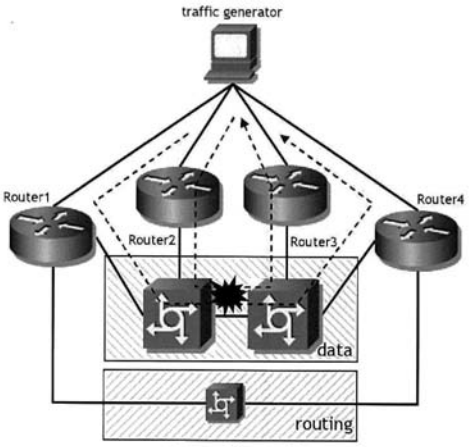

図5 検証環境

トラフィックジェネレータからルータ 1，ルータ 2 へそれぞれ 800Mbps のトラフィックを印加し，このトラフィックはそれぞれルータ 4，ルータ 3 を経由してトラフィックジェネレータへと戻る．これによって，Gigabit Ethernet で接続されている IX のスイッチ間の容量を超えるトラフィックが流れ，輻輳が発生している状況での通過トラフィックの状況を計測する．

ルータ 1，ルータ 4 は「data トラフィック用スイッチ」経由と「routing トラフィック用スイッチ」経由の両方で peering を行なっている．一方ルータ 2 とルータ 3 は static ルーティングによりそれぞれトラフィックを相手に向ける．（通常の IX では BGP の peering を行なうが，本検証ではルーティングの影響を排除するため static ルーティングで設定している）

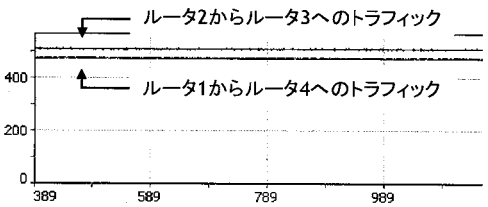

図6 ルータ間のスループット

図6はトラフィックジェネレータからルータ 1 とルータ 2 へ印加したトラフィックが，ルータ 2，ルータ 3 を経由してどれだけトラフィックジェネレータで受信され

たかを示すグラフである．図2のグラフとは異なり，スイッチ間で輻輳が起こっているにも関わらずトラフィックの低下は発生していない．それぞれスイッチ間の物理帯域である1Gbpsの半分，500Mbps程度のトラフィックが流れている．（物理帯域 1Gbps であるため輻輳が起こり，当然印加したトラフィック 800Mbps すべては流れない）このとき，「dataトラフィック用スイッチ」を経由して行なっているpeeringは，輻輳により複数回切断されていた．なおルータ2からルータ3へのトラフィックは先に述べた通りstaticルーティングで設定されているため，輻輳による経路の変更は発生しない．

5. まとめ

本稿では既存の実装を用いて，既存の IX のスイッチとは異なるスイッチで peering を行なうことによって経路制御メッセージの配送網とデータトラフィックの配送網を分離した．網を分離することによって経路制御メッセージの配送がデータトラフィックの輻輳の影響を受けなくなり，データトラフィックの安定的な送受信を実現することができた．

参考文献

1) draft-ietf-bfd-base-05「Bidirectional Forwarding Detection」D. Katz, D. Ward June, 2006
2) RFC4271「A Border Gateway Protocol 4 (BGP-4)」Y. Rekhter, Ed., T. Li, Ed., S. Hares, Ed.. January, 2006
3) JANOG17「ネットワークにおける高速切り替え方式の検討」鈴木 昭徳, January, 2006