6T-1

I-BGP におけるルートリフレクターのスケーラビリティ評価に関する研究

長橋 賢吾 [1], **江崎 浩** [1]

東京大学大学院情報理工学系研究科 [1]

1 研究の目的と背景

I-BGP は, ドメイン間経路制御プロトコルである BGP[1] をドメイン内経路制御に応用した経路制御プロトコルである. 通常の I-BGP は, 経路のループを防ぐために近隣ルータから学習した経路情報を他の近隣ルータに転送しない. ゆえに, あるドメインにおいて I-BGP で経路制御をおこなう場合, すべての近隣ルータはフルメッシュでピアを確立しなくてはならない. フルメッシュに必要なピアの数は, $\frac{n(n-1)}{2}$ であり, ルータ数 (n) が増加するほど, その管理コストは増加する. ピア数増加を抑制するために,I-BGP では, ルートリフレクタ [2] を導入している. ルートリフレクタ (RR) とは, フルメッシュのかわりにルートリフレクタが，ルートリフレクタクライアントに対して経路情報を供給する. すなわちクライアントである I-BGP ルータは, ルートリフレクタのみピアを確立すれば, そのドメイン内における経路情報を取得することができる.

しかしながら, ルートリフレクタとピアを確立しているクライアント数が増えれば増えるほど, ルートリフレクタへの負荷は増大する. そこで，本論文では, この I-BGP におけるルートリフクレタの負荷に関して, 以下を数量的に調査解析することを目的とする.

1. 収容可能なクライアント数の評価
2. ルートリフレクタへの負荷に関する各自パラメータの重要度の評価

2 測定パラメータおよび測定

調査解析にあたって，I-BGP における以下のトポロジーを利用する.

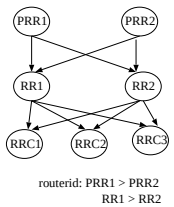

図 1: ルートリフレクタモデル

上記のトポロジーに以下のパラメータを投入する.

経路情報 インターネットの全経路 (full route) に耐えることが前提であり，routeviews[3] によれば，146,955 経路，その path attribute は，48,985 であり，これをパラメータとして利用する.

ルートリフレクタクライアント収容数 ルートリフレクタに収容するクライアントを 1，30，60，170 台と変化させ，その挙動を把握する.

ルータの実装 PC 上のルーティングソフトウェア zebra[4](OS FreeBSD4.10 メモリ 512MB)，および Cisco 社ルータの Cisco7200(IOS12.2(24a)，メモリ 256MB) の 2 種類のルータ実装において評価を実施する.

zebra，Cisco ともに 1 台，30 台とクライアント数を変化させても，ルートリフレクタは BGP Update を処理し，正常に収束に向かう．しかし，クライアント数を 60 台にした場合，Cisco は，図 2 のように 265.8sec で収束しているのに対して，zebra は，図 3 のように 400sec 経過しても TCP シーケンスは 217 と伸びない．その原因として,zebra のメモリ不足がある.zebra は各 peer ごとに 146,000 経路分の RIB を保持し, それを各 peer に対して同時に送信する. このとき, ルートリフレクタは自身のメモリ領域を確保できずに, 結果的に BGP パケットを送信することができなくなり, ゆえに, シーケンス番号は増加することなく, 一向に収束しない. つまり, zebra のような PC ベースのルーティングソフトウェアにおける限界性能は,CPU,BUS,NIC ではなく, メインメモリの容量に大きく依存すると言うことができる. 本測定の場合において, zebra のメインメモリは 512MB であり，そのときの zebra のクライアント数 60 が限界であったといえる.

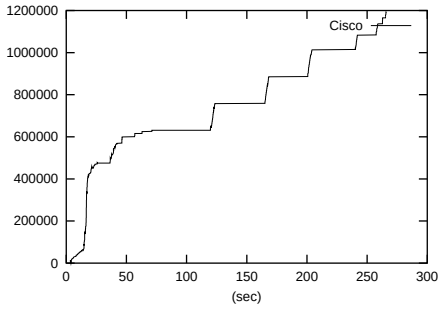

図 2: クライアント数 60 台での Cisco の TCP シーケンス

次に，Cisco ルータに 170 台のクライアントを接続した. 図 4 は, その際の,TCP シーケンスをあらわしている．この図では,170 台接続した場合,1104.6sec ですべての BGP

A Study of Scalability Measurement of Route Reflector

*Kengo NAGAHASHI, Graduate School of Information Science and Technology, The University of Tokyo.

†Hiroshi ESAKI, Graduate School of Information Science and Technology, The University of Tokyo.

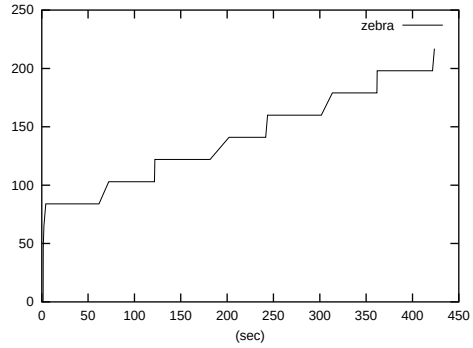

図 3: クライアント数 60 台での zebra の TCP シーケンス

Update が終了し, 収束している.

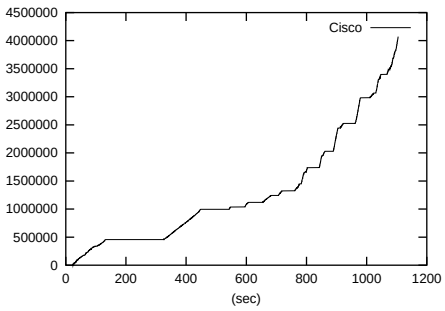

図 4: クライアント数 170 台での Cisco の TCP シーケンス

さらに (1)PRR_1 と RR_1 とのピアが切断,(2)PRR_2 経由で経路を受信する状態で再び PRR_1 とのピアを確立した. 図 5 は, 再び PRR_1 とのピアを確立したときの TCP シーケンスをあらわしている.

この図からは,1365sec 経過しても，前節の 60 台での zebra の測定結果と同様に本来であれば 1,000,000 バイト程度まで到達する TCP シーケンスが，わずか 500 バイト程度しか伸びていない．これは PRR_1 と RR_1 とのピアの再確立時の処理によるもので，PRR_1 と RR_1 は以下のように振舞う.

- PRR_1 と RR_1 および RR_2 のピアが再確立
- RR_1 および RR_2 は, ベストパスの配布先が変更されたので, その BGP Update をそれぞれの RRC に配布する
- 多量の RRC に BGP Update メッセージを送信するために，RR_1 の CPU のキューが溢れてしまう
- CPU のキューが溢れた結果，大量の BGP Update を送信できず，TCP 処理が進まない

3 測定に関する考察

第 1 章において，本論文では以下の目的を設定した.

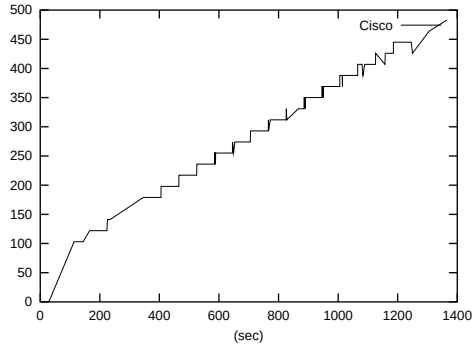

図 5: PRR_1 のピアを再確立した後の TCP シーケンス

1. 収容可能なクライアント数の評価
2. ルートリフレクタへの負荷に関する各自パラメータの重要度の評価

(1) に関しては，一般的にはルートリフレクタのハードウェア構成に依存するが，本測定の結果では，PC ルータ (zebra) の場合であれば，クライアント 60 台前後，Cisco ルータでは，クライアント 170 台ではいったん収束するがピア断やルータダウンなどで簡単に発散状態となることがわかった．(2) に関しては，本測定では，(a) クライアント数 (1,30,60,70 台)，(b) ルータ実装 (zebra,Cisco), (c) 経路情報をパラメータとして投入した．その結果，負荷にかかるパラメータとして重要なのは経路情報数 x クライアント数であり，この値が大きければ，大きいほどルートリフレクタにかかる負荷が大きいことがわかった. そして，この値が大きくなると，(a) ルータのメモリ不足, (b) ルータの出力キューが捌けない状態を引き起こし，結果的に収束しない状態に陥ることがわかった.

4 むすび

前章での考察を踏まえ，今後の課題としては以下が考えられる.

- 今回の収束ケースは, 限られたトポロジーであったので, 大規模なトポロジーでも今回のケースが再現できるかについての検証
- プライマリ・バックアップ型に依存しない冗長ルートリフレクタアーキテクチャーの提案

参考文献

[1] Y. Rekhter T. Li. *A Border Gateway Protocol 4 (BGP-4).* RFC 1771, Match 1995.

[2] R. Chandra T. Bates. *BGP Route Reflection An alternative to full mesh IBGP.* RFC 1997, June 1996.

[3] Oregon Exchange BGP Route Viewer. *Host:route-views.oregon-ix.net.*

[4] Zebra Project. *http://www.zebra.org/.*