経路情報爆発問題と経路フィルタリングの話

ーIRS14 発表資料ー

2007／10／11

株式会社NTTデータ
吉野　誠吾

はじめに

この数ヶ月の間に経路情報の増大と経路フィルタリングに関わる様々なトピックスについて多くの方に教えていただいたので、共有させていただき今後の議論のベースになれば・・と思いまとめたものです。

不足、間違いなどはご指摘いただきたく、update した上で公開させて頂きたいと考えます。

コンテンツ

1. /8 より短い prefix
2. RIR 割振サイズとフィルタ
3. ripe-399：RIPE Routing Working Group Recommendations on Route Aggregation
4. ルータのリソース問題再び
5. Prefix フィルタ、AS(-PATH) フィルタ

1．/8 より短い prefix（その1）：事象

6 月に GSR のラインカード（E3）で、CPU HOG メッセージが出力された。
例：
SLOT 1: %SYS-3-CPUHOG: Task ran for 4024 msec (690/0), process = CEF IPC・・・・・

どうも、ここに書かれていることに該当したらしい。
http://www.cisco.com/japanese/warp/public/3/jp/service/tac/63/cpuhog-j.shtml#cpuhog-cef

ちょうど、こんな経路が飛んでいた。
*>i8.0.0.0/5
*>i16.0.0.0/4
*>i64.0.0.0/2
*>i128.0.0.0/1

1．/8 より短い prefix（その1）：説明

前頁の URL より引用。太字に変更は独断。

「CEF LC バックグラウンドプロセスが原因で発生する CPUHOG」
Cisco 12000 シリーズ インターネット ルータでは、パケット交換で使用するように、ラインカードごとに Forwarding Information Base（FIB: 転送情報ベース）が保持されています。**FIB ツリー構造上、短いサブネット マスク（/1 と /4 の間）でルーティングを変更すると**、次のようなメッセージがコンソール ログに表示されることがあります。
SLOT 1: %SYS-3-CPUHOG: Task ran for 4024 msec (690/0), process = CEF IPC Background, PC = 400B8908. -Traceback= 400B8910 408FF588 408FF6F4 408FFE8C 400A404C 400A4038

Cisco IOS ソフトウェアのプロセスの実行が **2000 ミリ秒（2 秒）を超えると**、CPUHOG メッセージが出力されます。非常に短いサブネット マスクに対して Cisco Express Forwarding（CEF）更新を実行すると、処理に必要な時間が 2000 ミリ秒を超えることがあり、こうしたメッセージがトリガされます。「CEF IPC バックグラウンド」プロセスは、転送ツリーに対するプレフィックスの追加と削除を制御する親プロセスです。
また **CPU が一定時間以上ロック ダウンされると、ファブリック Ping 障害によりラインカードがクラッシュするか、IPC 通信の切断タイムアウトにより FIB がディセーブルになる可能性があります。** こうした問題をトラブルシューティングする必要がある場合は、『Cisco 12000 シリーズ インターネット ルータにおけるファブリック Ping のタイムアウトおよび障害のトラブルシューティング』で重要な情報が見つかることがあります。
一般的に、/7 より短いマスクに関するルーティング更新は誤っているか、悪意があるかのいずれかです。こうした更新が処理されたり、伝搬しないように、すべてのお客様が**適切なルート フィルタリングを設定することをシスコは推奨**します。

1．/8 より短い prefix（その１）：コメント

- すぐに止まってしまう、というほどの危険性かは不明だが、負荷が上昇するのは確か。合わせ技で止まってしまう可能性はあるので、CEF を利用している場合、/1 – /4 は filter した方がよい。
- 馬渡さんフィルタも任意から必須に昇格させませんか？

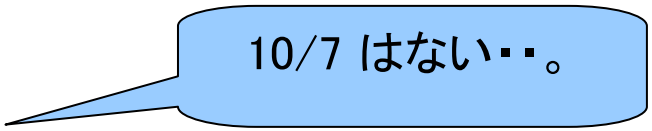
当日コメントあり：

資料末尾参照願います

1．/8 より短い prefix（その2）：事象

8/23。スペインの ISP より、インターネットに広告されている経路をカバーする /7 の経路が広告された。

2.0.0.0/7
4.0.0.0/7
6.0.0.0/7
8.0.0.0/7
12.0.0.0/7
（省略）
202.0.0.0/7
204.0.0.0/7
206.0.0.0/7
208.0.0.0/7
210.0.0.0/7
212.0.0.0/7
214.0.0.0/7
216.0.0.0/7
218.0.0.0/7
220.0.0.0/7
222.0.0.0/7



１．/8 より短い prefix（その２）：説明

short prefix なのでトラフィックには影響なし・・だった。
Aggregate されていたので、設定ミスか？
うちの uplink は AS2914（ntt.net）さんと AS10026 さんだが、この経路は AS10026 さんからは届いておらず、（良い悪いではなく）フィルタリングポリシーの違いが見て取れる。
ちなみに、IRR には登録されていない経路だった。
uRPF キラー。

1．/8 より短い prefix（その2）：コメント

- uRPF 使っている場合、もしくは、パケットを吸い出されるのが嫌であれば、/8 より短い経路はフィルタした方がよい
- IRR に登録されていなくても飛んでくる経路はある・・と再認識

2. RIR 割振サイズとフィルタ

ちょうど nanog で以下の URL、更新しようかっていうメールが。
「ISP Filter Policies」
http://www.nanog.org/filter.html
この URL には Verio（現 ntt.net）のポリシーへのリンクがある。

ここで再認識したのは、Tier-1 の peer では、/24 or shorter であれば、スルー＊1、＊2の事が多いらしく、IRR への登録有無とは関係ない、ということ。

＊1：経路数の多さを考えればやむ得ない
＊2：＋max prefix の制限を行っているところもあり

２．RIR 割振サイズとフィルタ

ここから話は脱線し、

A:「/24 までは受け取るのって、〇業要望？＊１」
B:「あれこれあって今は /24。」
B:「/21 とかだと経路数はもっと少なく（数万）抑えられるんだけど・・。」
A:「フィルタポリシーによって、経路数って随分違うんですね！？」
C:「RIRの割振サイズより細かいのを落とすとかなり経路減る・・」

ここで再び nanog にタイムリーな thread が・・。

＊１:トラフィックはより細かい経路に引き込まれる。/22 より /24 の方がトラフィックを得やすい⇒課金量が増える。

2. RIR 割振サイズとフィルタ

9/8 「Route table growth and hardware limits...talk to the filter」

RIR が公表している最小割振サイズの情報などを元に filter すると、約 23 万経路の現状で、14-15 万経路に収められる。ハードウェアリソースの限界に近づいている場合の延命策で使えないか・・というような内容。

2. RIR 割振サイズとフィルタ

RIR が公表している最小割振サイズ。例えば APNIC だとこんな感じ。

http://www.apnic.net/db/min-alloc.html

58/8 /21 portable allocations
59/8 /21 portable allocations
60/8 /21 portable allocations
61/8 /21 portable allocations
116/8 /21 portable allocations
（省略）
202/8 /24 portable assignments
203/8 /24 portable assignments
210/8 /21 portable allocations
211/8 /21 portable allocations
218/8 /24 portable allocations
219/8 /21 portable allocations
220/8 /21 portable allocations
221/8 /21 portable allocations
222/8 /21 portable allocations

JPNIC さんからの割振が /22 だったこともある（ whois.jp で確認。JPNIC さんには未確認）。「Stict Mode Ingress Prefix Filter Template 」では /22 で filter している。
この違いがなぜ生じているかは未調査・・。

2. RIR 割振サイズとフィルタ

以下のドキュメントも紹介されている。

「Strict Mode Ingress Prefix Filter Template 」
ftp://ftp-eng.cisco.com/cons/isp/security/Ingress-Prefix-Filter-Templates/T-ip-prefix-filter-ingress-strict-check-v18.txt

ちなみに、フィルタは RIR が公表している情報と違う(2007/10/3 現在)。
ちなみに、/1 - /4 も filter されるよう書かれている。
ちなみに、126/8 は filter されてしまいそう・・。126.*/16 は OK。

2. RIR 割振サイズとフィルタ

RIR の最小割振サイズで filter すると、一部 reachability が無くなるので注意（要らないとあきらめるか default を向けるか、が必要）。

このあたりをまとめた文章として ripe-399 を教えていただいた。

3. ripe-399

「RIPE Routing Working Group Recommendations on Route Aggregation」

経路の aggregation および deaggregation について、現状や推奨をまとめた文章。

3. What is Aggregation?

連続した IP アドレスを 1 つのアドレスブロックにまとめてルーティングすること。

4. The Internet Routing Table

4.1 What is Deaggregation?

aggregate せずに、より細かい経路(longer prefix)を広告すること。

4.2 General Deaggregation

Routing system security への対処や、未使用領域に飛んでくるトラフィックを減らしたい、などの commercial reason。(Routing system security って(続く)

3. ripe-399

要は、自 AS が広告する経路より細かい（more specific）経路を他 AS で広告されるとトラフィックを奪われるという、所謂経路ハイジャックへの対処。）
　その他の deaggregate する理由として、CIDR Report の Top10 に載りたかったから・・。

4.3 iBGP and eBGP

　iBGP に顧客経路を入れてある場合で、eBGP に誤って経路を流してしまっている、というケース

4.4 Deaggregation to aid Multihoming

　トラフィックエンジニアリング目的。/24 がよく使われる。

4.5 Legacy Assignments

　CIDR 導入前の所謂歴史的 PI のアドレス。最近はこの領域も aggregate されていて、特に他の領域より経路数が多いということはない。

5. Impacts of the Routing Table Size

5.1 Router Memory （後ほど）

5.2 Router Processing Power （省略）

5.3 Routing Convergence （省略）

3. ripe-399

5.4 Network Performance

iBGP で顧客の経路を流している場合、顧客のリンクが flap した時に上がってきてもルーティング上は元に戻るのに時間がかかる・・という意味。

6. Solutions

6.1 The CIDR Report

http://www.cidr-report.org/

（TOP10 の他、自 AS がどんな経路を投げているか、どう aggregate することが可能か、などを表示できる。）

（こんなのもあります http://thyme.apnic.net/）

6.2 Filtering

RIR の最小割振サイズより細かい経路をフィルタリングする。どの程度のネットワークがこのフィルタを実装しているのかは不明。最小割振サイズで割振を受けているネットワークは、トラフィックエンジニアリング目的などで経路を分割することが制限される。

6.3 The "CIDR Police"

経路を監視して "こう aggregate したら・・" と助言する活動で昔存在した。

3. ripe-399

6.4 BGP Features

6.4.1 The NO_EXPORT BGP Community

隣接 AS 内でのみ扱い可能で、その先の AS には出て行かない。トラフィックエンジニアリング目的の場合には要検討。

6.4.2 The NOPEER BGP Community

RFC3765。(その名の通り:-)。実装はまだないようである。)

6.4.3 The "AS_PATHLIMIT:" Attribute

(まだ draft)AS PATH 中に現れる AS 数の最大値を定義するもので、受信側でこの attribute より多い AS 数を検知したらこの経路を処理しない。

6.4.4 Provider-Specific Communities

顧客に対して NO_EXPORT より細かく community による経路制御をできるようにすること。(すべての peer に 1 つ AS prepend とか、北米だけとか、広告しないとか)

3. ripe-399

7. Recommendations

7.1 Initial Allocations

RIR から割振を受けたサイズで広告して。

当日コメントあり:

資料末尾参照願います

7.2 Subsequent Allocations

所謂おかわりに対して、連続のアドレスブロックとなるよう RIR は配慮して。プロバイダも受け取ったら aggregate して。

7.3 Multihoming

分割する場合にも、割振サイズの経路も広告した方がよい。そうでないとバックアップがなくなる。分割は慎重に。経路情報へのインパクトを少なくしつつ、トラフィックエンジニアリングの効果を得るための tutorial が多数あるので、参考文献を参照して。

7.4 BGP Enhancements

（BGP の community でうまく伝播範囲を制御して）

7.5 Proxy Aggregation

注意を要する。blackhole になる可能性。同じプロバイダにマルチホームしているなど他に影響がないと判断できる場合にのみ使用したほうがよい。

3. ripe-399

7.6 IP version 6

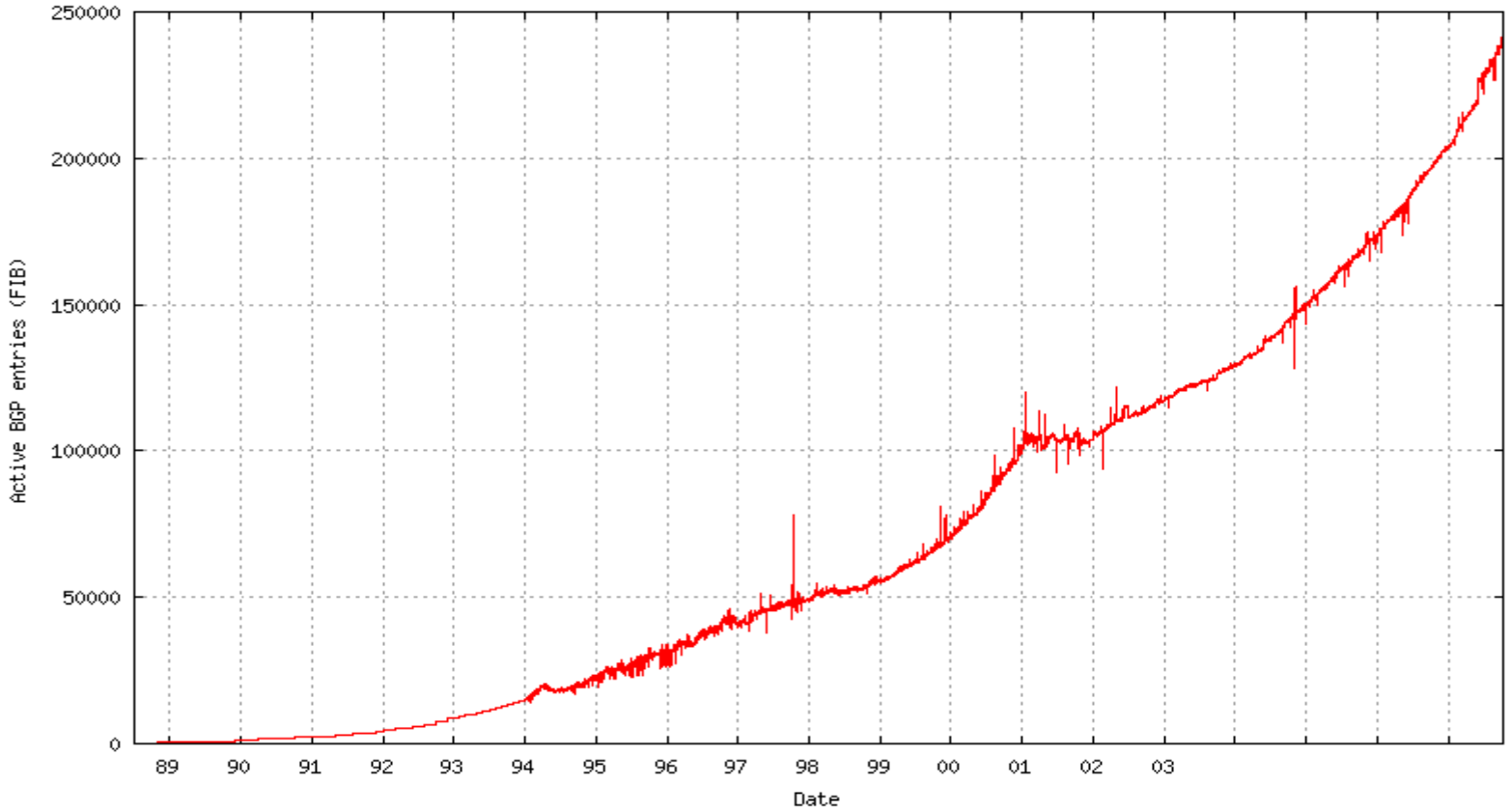
v6 でも同じ。

（メモ）

multiple origin については触れられていない。

4. ルータのリソース問題再び

2007/10/11 時点で、239623経路見えている。
ここ5年間ほどは exponential な伸び。年率 16% 前後・・4年強で倍。



http://bgp.potaroo.net/as2.0/bgp-active.html

４．ルータのリソース問題再び

IRS11（2006/12/11）の「20 万経路に挑戦」や「CEF苦労話」関連。

基本的にはメモリリソースの増強（ルータの更改）で対処する問題ではあるが、いくつかの理由により時間稼ぎとしての対処策は整理しておいた方がよいと考えた。

・何らかの理由で装置購入の許可が得られない
・更改する装置がバグ等でトラぶった場合に、逃げ道を失う
・IPv4 アドレス枯渇時期に経路数が急増する可能性も0ではない

４．ルータのリソース問題再び

メモリが不足すると何が起きるか！？

・一部経路への到達性がなくなる。
・BGP session が flap する。
・FIB が disable になる
・OSPF neighbor が切れて二度とつながらなくなる
・telnet/ssh できなくなる（注：最近の IOS は最低限のメモリは空ける）

他、予期せぬ事態。

４．ルータのリソース問題再び

注意しておくべきことは！？

・トラフィック増によりインタフェースをアップグレードするのと違って、一気に多数のルータの更改が必要となる。急に対処するのは困難。

・RP/RE だけではなく、LC/PFE も同じ問題がありえる。

・リソースの閾値が分かりにくい（余裕が多めに必要）

・そもそもどこを確認しておけばいいのか分かりにくい（DRAM の空きだけではなく TCAM も・・表示が分かりにくいとか）

４．ルータのリソース問題再び

身を守るには経路数を減らすしかないが、現状送り手側の抑制は難しい。
環境問題と同じで長期的な働きかけが必要。
では、受け手側でできることは？

- filter (with default)　　捨てる
- proxy aggregate　　まとめる

４．ルータのリソース問題再び

方法１：RIR の最小割振サイズより細かい経路をフィルタ
　割振サイズの経路をアナウンスしていないネットワークへの reachability がなくなるので、upstream の AS に default を向ける必要あり（要調整）
　default route を生成するルータが upstream との境界ルータでない場合（sinkhole の作り方によって）は、対応が難しい。

Transit AS 対応：
　顧客に捨てた経路が伝わらないので
・顧客にも default route を向けてもらう。
　もしくは、
・route-map で 1 つに aggregate した経路を載せてあげる。
　（Origin が自 AS に書き換わるので相当慎重に対応する必要がある）

４．ルータのリソース問題再び

方法2：捨ててもトラフィックに影響がない経路を選んでフィルタ
　①があれば、②と③は捨てても大丈夫・・！？恐らく①だけがなくなることはないと思われる・・。default を向ける必要はない。
（例）
　以下の②、③を捨てる。

192.168.0.0/16	１－①
192.168.0.0/24	１－②
192.168.2.0/24	１－③

４．ルータのリソース問題再び

方法１のTransit AS 対応：
　顧客に捨てた経路が伝わらないので

・顧客にも default route を向けてもらう。

　もしくは、

・route-map で 1 つに aggregate した経路を載せてあげる。
　（Origin が自 AS に書き換わるので相当慎重に対応する必要がある）

４．ルータのリソース問題再び

比較検討

　方法１は、Transit を提供する場合、顧客にも同様に default を向けてもらう必要がある。
　方法２は削減効果が小さい。
　aggregate する場合は要注意。

まとめ
　構成上可能であれば方法１、短期間の延命でよければ方法２がよいのでは？（異論ありだと思いますが）
　他の案があれば是非コメントください。

４．ルータのリソース問題再び

当日コメントあり：

資料末尾参照願います

（参考）

FIB Compression（注：実装はあるらしいが不明。今後主流となるかも不明）

RIB から FIB を生成する際に、不要な情報を削除する機能。

例えば、

192.168.0.0/16　AS300 AS200 AS100　①

192.168.0.0/24　AS300 AS200 AS100　②

192.168.2.0/24　AS300 AS200 AS100　③

の②、③はなくても動作するので FIB に載せない。

ただし、①しか FIB にない状態で、①が withdraw されたら・・、②、③を FIB に復活させる必要がある。

このためアルゴリズムが複雑になり、CPU リソースを多く必要とするという問題はある。

４．ルータのリソース問題再び

（参考）
「Router Scaling Trends」
http://www.ripe.net/ripe/meetings/ripe-54/presentations/Router_Scaling_Trends.pdf

経路数増大に関するルータへの要求に関していろいろまとまっている。
既存の技術で 1,000万経路対応ルータも作れる・・とのこと。

TCAM ではなくて RLDRAM と並列処理で・・とか。

当面、数年後経路数がある程度の伸びとなっても更改するルータは存在しそう。買えるかどうかは別問題として。

5. Prefix フィルタ、AS(-PATH) フィルタ

Ingress の経路フィルタを設定する目的は、

・隣接 AS のミスからの防御

・hijack からの防御

にあるが、日本でまだ主流？の AS(-PATH) フィルタ（以下、AS フィルタ）では、問題が多いのでは？

注：Prefix フィルタの元情報を得る方法として、

・メール等連絡する方法

・IRR を参照する方法

とあるが、基本的に IRR を参照するものとして以下説明。

5. Prefix フィルタ、AS(-PATH) フィルタ

効果の比較

想定ケース:
①:フルルートを誤ってアナウンス
②:IGP の経路を誤ってアナウンス
③:hijack

当日コメントあり:
資料末尾参照願います

ケース	Prefix フィルタ	AS フィルタ
①	○	△ （Origin が書き換えられるとだめ）
②	△ （フィルタの内容による）	×
③	△ （IRR の情報が偽られるとだめ）	×

5. Prefix フィルタ、AS(-PATH) フィルタ

その他比較

項目	Prefix フィルタ	AS フィルタ
更新の検知	○:自動化可能	×:メールでの連絡
config 生成	○:自動化可能	×:手作業が入る
config の量	×:比較的多い	○:比較的少ない
hijack 時の対応の容易さ *1	△:or-longer を開けておかないとフィルタにかかってしまう	○:変更不要

*1:hijack 時の対応とは、hijack 経路より細かい経路を投げ返すことでトラフィックを奪い返そうとすること。

5. Prefix フィルタ、AS(-PATH) フィルタ

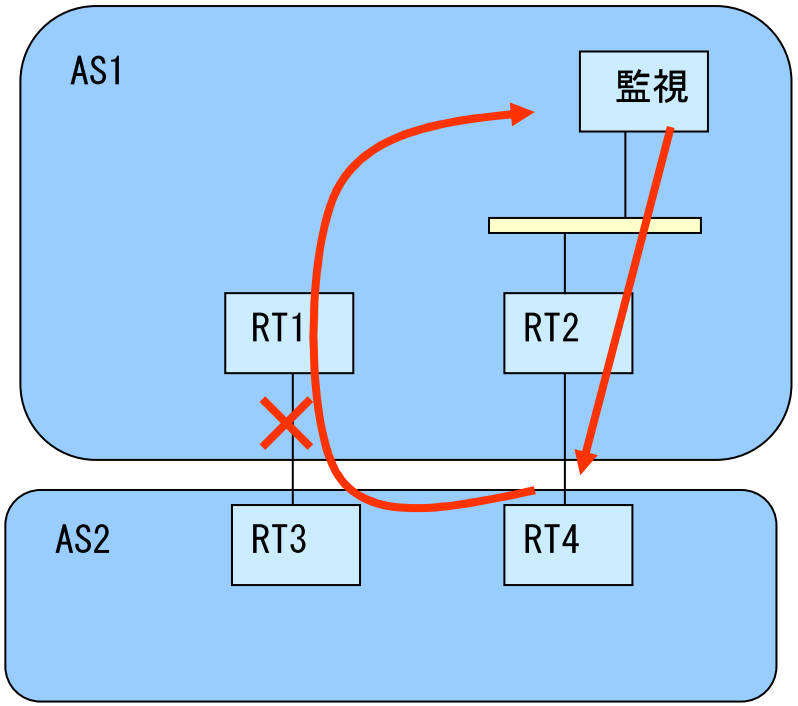
まとめ

・Prefix フィルタへ移行するというのはいかがでしょう？
・その際、or-longer を開けておくか、community をつければ accept する(*1)等、hijack 時の対応を容易にしてあげるというのはいかがでしょう？

*1：or-longer を開けると、IGP の経路を突っ込まれる問題に対して脆弱になるので。

（付録）NO_ADVERTISE community の使い方例

BGP の well known community に NO_ADVERTISE というものもある。
AS 間が複数の接続点で接続している状況で、監視のパケットの行き返りを同一回線に固定したい場合、監視装置のセグメントを含む specific な経路に NO_ADVERTISE をつけてアナウンスする方法が取れる。



AS2 から AS1 へ向かうルートは RT3->RT1 の経路がベストになっているとすると、例えば、RT1-RT3 間が IX などでリンクダウンを検出できない状態で故障が発生すると、RT4 に対する監視でアラームがあがることがある。

RT2 から RT4 に対して NO_ADVERTISE をつけて監視装置のセグメントをアナウンスすると RT4 からの返りだけ RT4->RT2 の経路となる。

いただいたコメント

・/1 – /4 は特定の実装依存の話なので、馬渡さんフィルタとはちょっと違う話ですよね。–> （こういう実装問題も含めて）/8 より短いものは必須でフィルタした方がよい、という意見に変更させてください。
・FIB Compression は違う AS からの広報の場合はどうなるの？–> それはまとめないと思います。
・FIB Compression は Stub の AS だと効果が大きいようです。
・連続したアドレスを後から割振られた場合に、aggregate すべき・・という点について、現実的には最初に割振られたサイズで、フィルタや管理情報など各種情報が紐づいてしまうので、難しい場合があります。
・Prefix フィルタについては、Origin が違っていてもスルーなので Origin もセットでチェックするのが望ましいと思います。そういう実装はまだないですが、試みとして Zebra でチェックする・・をやっていました。–> IRS 8 の発表ですね。