GROW Working Group
Internet-Draft
Intended status: Informational
Expires: July 8, 2011

R. Raszuk, Ed.
R. Fernando
K. Patel
Cisco Systems
D. McPherson
Verisign
K. Kumaki
KDDI Corporation

January 4, 2011

Distribution of diverse BGP paths.
draft-ietf-grow-diverse-bgp-path-dist-03

Abstract

The BGP4 protocol specifies the selection and propagation of a single best path for each prefix. As defined today BGP has no mechanisms to distribute paths other then best path between its speakers. This behaviour results in number of disadvantages for new applications and services.

This document presents an alternative mechanism for solving the problem based on the concept of parallel route reflector planes. Such planes can be build in parallel or they can co-exit on the current route reflection platforms. Document also compares existing solutions and proposed ideas that enable distribution of more paths than just the best path.

This proposal does not specify any changes to the BGP protocol definition. It does not require upgrades to provider edge or core routers nor does it need network wide upgrades. The authors believe that the GROW WG would be the best place for this work.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on July 5, 2011.

Copyright Notice

Table of Contents

1. Introduction

Current BGP4 [RFC4271] protocol specification allows for the selection and propagation of only one best path for each prefix. The BGP protocol as defined today has no mechanism to distribute other then best path between its speakers. This behaviour results in a number of problems in the deployment of new applications and services.

This document presents an alternative mechanism for solving the problem based on the concept of parallel route reflector planes. It also compares existing solutions and proposed ideas that enable distribution of more paths than just the best path. The parallel route reflector planes solution brings very significant benefits at a negligible capex and opex deployment price as compared to the alternative techniques and is being considered by a number of network operators for deployment in their networks.

This proposal does not specify any changes to the BGP protocol definition. It does not require upgrades to provider edge or core routers nor does it need network wide upgrades. The only upgrade required is the new functionality on the new or current route reflectors. The authors believe that the GROW WG would be the best place for this work.

2. History

The need to disseminate more paths than just the best path is primarily driven by three requirements. First is the problem of BGP oscillations [I-D.ietf-idr-route-oscillation]. The second is the desire for reduction of time of reachability restoration in the event of network or network element's failure. Third requirement is to enhance BGP load balancing capabilities. Those reasons have lead to the proposal of BGP add-paths [I-D.ietf-idr-add-paths].

2.1. BGP Add-Paths Proposal

As it has been proven that distribution of only the best path of a route is not sufficient to meet the needs of continuously growing number of services carried over BGP the add-paths proposal was submitted in 2002 to enable BGP to distribute more then one path. This is achieved by including as a part of the NLRI an additional four octet value called the Path Identifier.

The implication of this change on a BGP implementation is that it must now maintain per path, instead of per prefix, peer advertisement state to track which of the peers each path was advertised to. This

new requirement has its own memory and processing cost. Suffice to say that by the end of 2009 none of the commercial BGP implementation could claim to support the new add-path behaviour in production code, in part because of this resource overhead.

An important observation is that distribution of more than one best path by Autonomous System Border Routers (ASBRs) with multiple EBGP peers attached to it where no "next hop self" is set may result in bestpath selection inconsistency within the autonomous system. Therefore it is also required to attach in the form of a new attribute the possible tie breakers and propagate those within the domain. The example of such attribute for the purpose of fast connectivity restoration to address that very case of ASBR injecting multiple external paths into the IBGP mesh has been presented and discussed in Fast Connectivity Restoration Using BGP Add-paths [I-D.ietf-idr-add-paths] document. Based on the additionally propagated information also best path selection is recommended to be modified to make sure that best and backup path selection within the domain stays consistent. More discussion on this particular point will be contained in the deployment considerations section below. In the proposed solution in this document we observe that in order to address most of the applications just use of best external advertisement is required. For ASBRs which are peering to multiple upstream ASs setting "next hop self" is recommended.

The add paths protocol extensions have to be implemented by all the routers within an AS in order for the system to work correctly. It remains quite a research topic to analyze benefits or risk associated with partial add-paths deployments. The risk becomes even greater in networks not using some form of edge to edge encapsulation.

The required code modifications include enhancements such as the Fast Connectivity Restoration Using BGP Add-path [I-D.pmohapat-idr-fast-conn-restore]. The deployment of such technology in an entire service provider network requires software and perhaps sometimes in the cases of End-of-Engineering or End-of-Life equipment even hardware upgrades. Such operation may or may not be economically feasible. Even if add-path functionality was available today on all commercial routing equipment and across all vendors, experience indicates that to achieve 100% deployment coverage within any medium or large global network may easily take years.

While it needs to be clearly acknowledged that the add-path mechanism provides the most general way to address the problem of distributing many paths between BGP speakers, this document provides a much easier to deploy solution that requires no modification to the BGP protocol where only a few additional paths may be required. The alternative

method presented is capable of addressing critical service provider requirements for disseminating more than a single path across an AS with a significantly lower deployment cost.

3. Goals

The proposal described in this document is not intended to compete with add-paths. Instead if deployed it is to be used as a very easy method to accommodate the majority of applications which may require presence of alternative BGP exit points.

It is presented to network operators as a possible choice and provides those operators who need additional paths today an alternative from the need to transition to a full mesh.

It is intended as a way to buy more time allowing for a smoother and gradual migration where router upgrades will be required for perhaps different reasons. It will also allow the time required where standard RP/RE memory size can easily accommodate the associated overhead with other techniques without any compromises.

4. Multi plane route reflection

The idea contained in the proposal assumes the use of route reflection within the network. Other techniques as described in the following sections already provide means for distribution of alternate paths today.

Let’s observe today’s picture of simple route reflected domain:

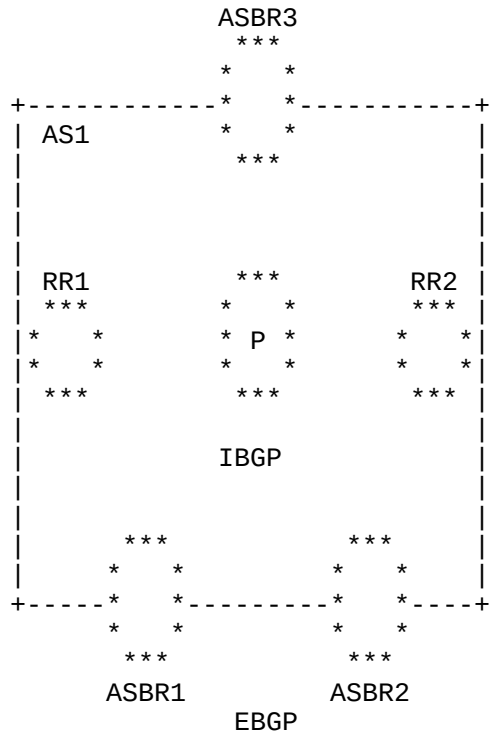


Figure1: Simple route reflection

Figure 1 shows an AS that is connected via EBGP peering at ASBR1 and ASBR2 to an upstream AS or set of ASes. For a given destination "D" ASBR1 and ASBR2 will each have an external path P1 and P2 respectively. The AS network uses two route reflectors RR1 and RR2 for redundancy reasons. The route reflectors propagate the single BGP best path for each route to all clients. All ASBRs are clients of RR1 and RR2.

Below are the possible cases of the path information that ASBR3 may receive from route reflectors RR1 and RR2:

1. When best path tie breaker is the IGP distance: When paths P1 and P2 are considered to be equally good best path candidates the selection will depend on the distance of the path next-hops from the route reflector making the decision. Depending on the positioning of the route reflectors in the IGP topology they may choose the same best path or a different one. In such a case

ASBR3 may receive either the same path or different paths from each of the route reflectors.

2. When best path tie breaker is Multi-Exit-Discriminator or Local Preference: In this case only one path from preferred exit point ASBR will be available to RRs since the other peering ASBR will consider the IBGP path as best and will not announce (or if already announced will withdraw) its own external path. The exception here is the use of BGP Best-External proposal which will allow stated ASBR to still propagate to the RRs its own external path. Unfortunately RRs will not be able to distribute it any further to other clients as only the overall best path will be reflected.

The proposed solution is based on the use of additional route reflectors or new functionality enabled on the existing route reflectors that instead of distributing the best path for each route will distribute an alternative path other then best. The best path (main) reflector plane distributes the best path for each route as it does today. The second plane distributes the second best path for each route and so on. Distribution of N paths for each route can be achieved by using N reflector planes.

Each plane of route reflectors is a logical entity and may or may not be co-located with the existing best path route reflectors. Adding a route reflector plane to a network may be as easy as enabling a logical router partition, new BGP process or just a new configuration knob on an existing route reflector and configuring an additional IBGP session from the current clients if required. There are no code changes required on the route reflector clients for this mechanism to work. It is easy to observe that the installation of one or more additional route reflector control planes is much cheaper and an easier than the need of upgrading 100s of route reflector clients in the entire network to support different protocol encoding.

Diverse path route reflectors need the new ability to calculate and propagate the Nth best path instead of the overall best path. An implementation is encouraged to enable this new functionality on a per neighbor basis.

While this is an implementation detail, the code to calculate Nth best path is also required by other BGP solutions. For example in the application of fast connectivity restoration BGP must calculate a backup path for installation into the RIB and FIB ahead of the actual failure.

To address the problem of external paths not being available to route reflectors due to local preference or MED factors it is recommended

that ASBRs enable the best-external functionality in order to always inject their external paths to the route reflectors.

4.1. Co-located best and backup path RRs

To simplify the description let's assume that we only use two route reflector planes (N=2). When co-located the additional 2nd best path reflectors are connected to the network at the same points from the perspective of the IGP as the existing best path RRs. Let's also assume that best-external is enabled on all ASBRs.

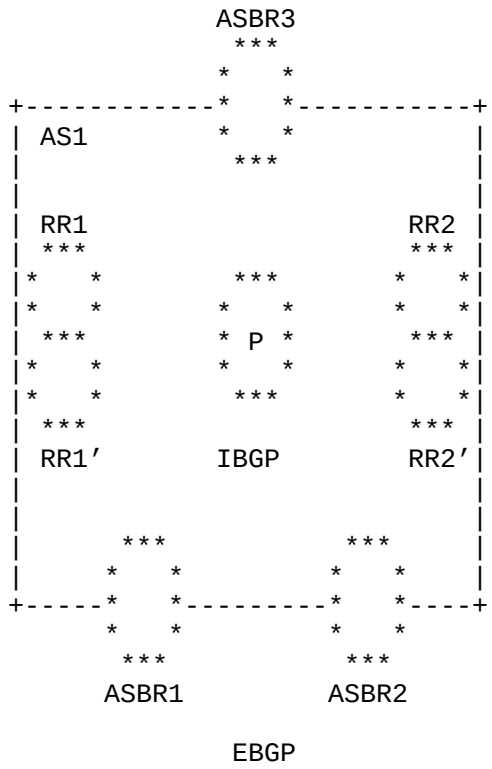


Figure2: Co-located 2nd best RR plane

The following is a list of configuration changes required to enable the 2nd best path route reflector plane:

1. Unless same RR1/RR2 platform is being used adding RR1' and RR2' either as logical or physical new control plane RRs in the same IGP points as RR1 and RR2 respectively.

2. Enabling best-external on ASBRs

3. Enabling RR1' and RR2' for 2nd plane route reflection. Alternatively instructing existing RR1 and RR2 to calculate also 2nd best path.

4. Unless one of the existing RRs is turned to advertise only diverse path to it's current clients configuring new ASBRs-RR' IBGP sessions

The expected behaviour is that under any BGP condition the ASBR3 and P routers will receive both paths P1 and P2 for destination D. The availability of both paths will allow them to implement a number of new services as listed in the applications section below.

As an alternative to fully meshing all RRs and RRs' an operator who has a large number of reflectors deployed today may choose to peer newly introduced RRs' to a hierarchical RR' which would be an IBGP interconnect point within the 2nd plane as well as between planes.

One of the deployment model of this scenario can be achieved by simple upgrade of the existing route reflectors without the need to deploy any new logical or physical platforms. Such upgrade would allow route reflectors to service both upgraded to add-paths peers as well as those peers which can not be immediately upgraded while in the same time allowing to distribute more then single best path. The obvious protocol benefit of using existing RRs to distribute towards their clients best and diverse bgp paths over different IBGP session is the automatic assurance that such client would always get different paths with their next hop being different.

The way to accomplish this would be to create a separate IBGP session for each N-th BGP path. Such session should be preferably terminated at a different loopback address of the route reflector. At the BGP OPEN stage of each such session a different bgp_router_id may be used. Correspondingly route reflector should also allow its clients to use the same bgp_router_id on each such session.

4.2. Randomly located best and backup path RRs

Now let's consider a deployment case where an operator wishes to enable a 2nd RR' plane using only a single additional router in a different network location to his current route reflectors. This model would be of particular use in networks where some form of end-to-end encapsulation (IP or MPLS) is enabled between provider edge routers.

Note that this model of operation assumes that the present best path

route reflectors are only control plane devices. If the route reflector is in the data forwarding path then the implementation must be able to clearly separate the Nth best-path selection from the selection of the paths to be used for data forwarding. The basic premise of this mode of deployment assumes that all reflector planes have the same information to choose from which includes the same set of BGP paths. It also requires the ability to ignore the step of comparison of the IGP metric to reach the bgp next hop during best-path calculation.

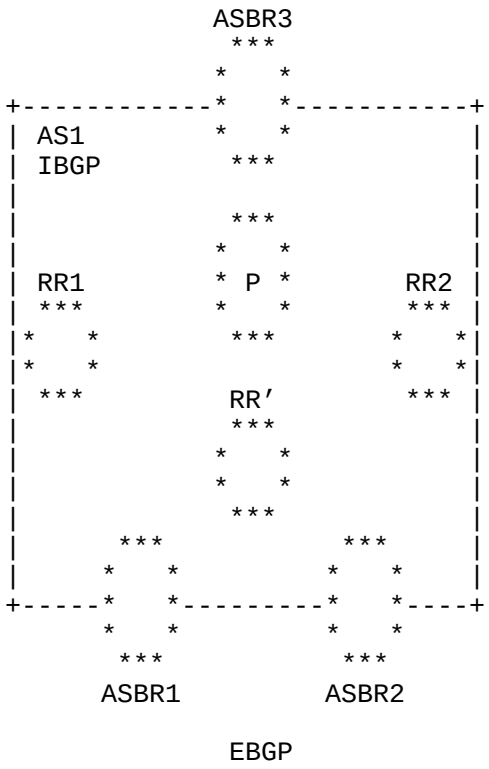


Figure3: Experimental deployment of 2nd best RR

The following is a list of configuration changes required to enable the 2nd best path route reflector RR' as a single platform or to enable one of the existing control plane RRs for diverse-path functionality:

1. If needed adding RR' logical or physical as new route reflector anywhere in the network

2. Enabling best-external on ASBRs

3. Disabling IGP metric check in BGP best path on all route reflectors.

4. Enabling RR' or any of the existing RR for 2nd plane path calculation

5. If required fully meshing newly added RRs' with the all other reflectors in both planes. That condition does not apply if the newly added RR'(s) already have peering to all ASBRs/PEs.

6. Unless one of the existing RRs is turned to advertise only diverse path to it's current clients configuring new ASBRs-RR' IBGP sessions

In this scenario the operator has the flexibility to introduce the new additional route reflector functionality on any existing or new hardware in the network. Any of the existing routers that are not already members of the best path route reflector plane can be easily configured to serve the 2nd plane either via using a logical / virtual router partition or by having their bgp implementation compliant to this specification.

Even if the IGP metric is not taken into consideration when comparing paths during the bestpath calculation, an implementation still has to consider paths with unreachable nexthops as invalid. It is worth pointing out that some implementations today already allow for configuration which results in no IGP metric comparison during the best path calculation.

The additional planes of route reflectors do not need to be fully redundant as the primary one does. If we are preparing for a single network failure event, a failure of a non backed up N-th best-path route reflector would not result in an connectivity outage of the actual data plane. The reason is that this would at most affect the presence of a backup path (not an active one) on same parts of the network. If the operator chooses to build the N-th best path plane redundantly by installing not one, but two or more route reflectors serving each additional plane the additional robustness will be achieved.

As a result of this solution ASBR3 and other ASBRs peering to RR' will be receiving the 2nd best path.

Similarly to section 4.1 as an alternative to fully meshing all RRs & RRs' an operator who may have a large number of reflectors already deployed today may choose to peer newly introduced RRs' to a hierarchical RR' which would be an IBGP interconnect point between planes.

4.3. Multi plane route servers for Internet Exchanges

Another group of devices where the proposed multi-plane architecture may be of particular applicability are EBGP route servers used at many of internet exchange points.

In such cases 100s of ISPs are interconnected on a common LAN. Instead of having 100s of direct EBGP sessions on each exchange client, a single peering is created to the transparent route server. The route server can only propagate a single best path. Mandating the upgrade for 100s of different service providers in order to implement add-path may be much more difficult as compared to asking them for provisioning one new EBGP session to an Nth best-path route server plane. That will allow to distribute more then single best BGP path from a given route server to such IX peer.

The solution proposed in this document fits very well with the requirement of having broader EBGP path diversity among the members of any Internet Exchange Point.

5. Discussion on current models of IBGP route distribution

In today's networks BGP4 operates as specified in [RFC4271]

There are a number of technology choices for intra-AS BGP route distribution:

1. Full mesh
2. Confederations
3. Route reflectors

5.1. Full Mesh

A full mesh, the most basic iBGP architecture, exists when all the BGP speaking routers within the AS peer directly with all other BGP speaking routers within the AS, irrespective of where a given router resides within the AS (e.g., P router, PE router, etc..).

While this is the simplest intra-domain path distribution method,

historically there have been a number of challenges in realizing such an IBGP full mesh in a large scale network. While some of these challenges are no longer applicable today some may still apply, to include the following:

1. Number of TCP sessions: The number of IBGP sessions on a single router in a full mesh topology of a large scale service provider can easily reach 100s. While on hardware and software used in the late 70s, 80s and 90s such numbers could be of concern, today customer requirements for the number of BGP sessions per box are reaching 1000s. This is already an order of magnitude more then the potential number of IBGP sessions. Advancement in hardware and software used in production routers mean that running a full mesh of IBGP sessions should not be dismissed due to the resulting number of TCP sessions alone.

2. Provisioning: When operating and troubleshooting large networks one of the top-most requirements is to keep the design as simple as possible. When the autonomous systems network is composed of hundreds of nodes it becomes very difficult to manually provision a full mesh of IBGP sessions. Adding or removing a router requires reconfiguration of all the other routers in the AS. While this is a real concern today there is already work in progress in the IETF to define IBGP peering automation through an IBGP Auto Discovery [I-D.raszuk-idr-ibgp-auto-mesh] mechanism.

3. Number of paths: Another concern when deploying a full IBGP mesh is the number of BGP paths for each route that have to be stored at every node. This number is very tightly related to the number of external peerings of an AS, the use of local preference or multi-exit-discriminator techniques and the presence of best-external [I-D.ietf-idr-best-external] advertisement configuration. If we make a rough assumption that the BGP4 path data structure consumes about 80-100 bytes the resulting control plane memory requirement for 500,000 IPv4 routes with one additional external path is 38-48 MB while for 1 million IPv4 routes it grows linearly to 76-95 MB. It is not possible to reach a general conclusion if this condition is negligible or if it is a show stopper for a full mesh deployment without direct reference to a given network.

To summarize, a full mesh IBGP peering can offer natural dissemination of multiple external paths among BGP speakers. When realized with the help of IBGP Auto Discovery peering automation this seems like a viable deployment especially in medium and small scale networks.

5.2. Confederations

For the purpose of this document let's observe that confederations [RFC5065] can be viewed as a hierarchical full mesh model.

Within each sub-AS BGP speakers are fully meshed and as discussed in section 2.1 all full mesh characteristics (number of TCP sessions, provisioning and potential concern over number of paths still apply in the sub-AS scale).

In addition to the direct peering of all BGP speakers within each sub-AS, all sub-AS border routers must also be fully meshed with each other. Sub-AS border routers configured with best-external functionality can inject additional exit paths within a sub-AS.

To summarize, it is technically sound to use confederations with the combination of best-external to achieve distribution of more than a single best path per route in a large autonomous systems.

In topologies where route reflectors are deployed within the confederation sub-ASes the technique describe here does apply.

5.3. Route reflectors

The main motivation behind the use of route reflectors [RFC4456] is the avoidance of the full mesh session management problem described above. Route reflectors, for good or for bad, are the most common solution today for interconnecting BGP speakers within an internal routing domain.

Route reflector peerings follow the advertisement rules defined by the BGP4 protocol. As a result only a single best path per prefix is sent to client BGP peers. That is the main reason why many current networks are exposed to a phenomenon called BGP path starvation which essentially results in inability to deliver a number of applications discussed later.

The route reflection equivalent when interconnecting BGP speakers between domains is popularly called the Route Server and is globally deployed today in many internet exchange points.

6. Deployment considerations

The diverse BGP path dissemination proposal allows the distribution of more paths than just the best-path to route reflector or route server clients of today's BGP4 implementations.

From the client's point of view receiving additional paths via separate IBGP sessions terminated at the new router reflector plane is functionally equivalent to constructing a full mesh peering without the problems that such a full mesh would come with set of problems as discussed in earlier section.

By precisely defining the number of reflector planes, network operators have full control over the number of redundant paths in the network. This number can be defined to address the needs of the service(s) being deployed.

The Nth plane route reflectors should be acting as control plane network entities. While they can be provisioned on the current production routers selected Nth best BGP paths should not be used directly in the date plane with the exception of such paths being BGP multipath eligible and such functionality is enabled. On RRs being in the data plane unless multipath is enabled 2nd best path is expected to be a backup path and should be installed as such into local RIB/FIB.

The proposed architecture deployed along with the BGP best-external functionality covers all three cases where the classic BGP route reflection paradigm would fail to distribute alternate exit points paths.

1. ASBRs advertising their single best external paths with no local-preference or multi-exit-discriminator present.

2. ASBRs advertising their single best external paths with local-preference or multi-exit-discriminator present and with BGP best-external functionality enabled.

3. ASBRs with multiple external paths.

Let's discuss the 3rd above case in more detail. This describes the scenario of a single ASBR connected to multiple EBGP peers. In practice this peering scenario is quite common. It is mostly due to the geographic location of EBGP peers and the diversity of those peers (for example peering to multiple tier 1 ISPs etc...). It is not designed for failure recovery scenarios as single failure of the ASBR would simultaneously result in loss of connectivity to all of the peers. In most medium and large geographically distributed networks there is always another ASBR or multiple ASBRs providing peering backups, typically in other geographically diverse locations in the network.

When an operator uses ASBRs with multiple peerings setting next hop self will effectively allow to locally repair the atomic failure of

any external peer without any compromise to the data plane. The most common reason for not setting next hop self is traditionally the associated drawback of loosing ability to signal the external failures of peering ASBRs or links to those ASBRs by fast IGP flooding. Such potential drawback can be easily avoided by using different peering address from the address used for next hop mapping as well as removing such next hop from IGP at the last possible BGP path failure.

Herein one may correctly observe that in the case of setting next hop self on an ASBR, attributes of other external paths such ASBR is peering with may be different from the attributes of its best external path. Therefore, not injecting all of those external paths with their corresponding attribute can not be compared to equivalent paths for the same prefix coming from different ASBRs.

While such observation in principle is correct one should put things in perspective of the overall goal which is to provide data plane connectivity upon a single failure with minimal interruption/packet loss. During such transient conditions, using even potentially suboptimal exit points is reasonable, so long as forwarding information loops are not introduced. In the mean time BGP control plane will on its own re-advertise newly elected best external path, route reflector planes will calculate their Nth best paths and propagate to its clients. The result is that after seconds even if potential sub-optimality were encountered it will be quickly and naturally healed.

7. Summary of benefits

The diverse BGP path dissemination proposal provides the following benefits when compared to the alternatives:

1. No modifications to BGP4 protocol.

2. No requirement for upgrades to edge and core routers. Backward compatible with the existing BGP deployments.

3. Can be easily enabled by introduction of a new route reflector, route server plane dedicated to the selection and distribution of Nth best-path or just by new configuration of the upgraded current route reflector(s).

4. Does not require major modification to BGP implementations in the entire network which will result in an unnecessary increase of memory and CPU consumption due to the shift from today’s per prefix to a per path advertisement state tracking.

5. Can be safely deployed gradually on a RR cluster basis.

6. The proposed solution is equally applicable to any BGP address family as described in Multiprotocol Extensions for BGP-4 RFC4760 [RFC4760]. In particular it can be used "as is" without any modifications to both IPv4 and IPv6 address families.

8. Applications

This section lists the most common applications which require presence of redundant BGP paths:

1. Fast connectivity restoration where backup paths with alternate exit points would be pre-installed as well as pre-resolved in the FIB of routers. That would allow for a local action upon reception of a critical event notification of network / node failure. This failure recovery mechaism based on the presence of backup paths is also suitable for gracefully addressing scheduled maintenane requirements as described in [I-D.decraene-bgp-graceful-shutdown-requirements].

2. Multi-path load balancing for both IBGP and EBGP.

3. BGP control plane churn reduction both intra-domain and inter-domain.

An important point to observe is that all of the above intra-domain applications based on the use of reflector planes but are also applicable in the inter-domain Internet exchange point examples. As discussed in section 4.3 an internet exchange can conceptually deploy shadow route server planes each responsible for distribution of an Nth best path to its EBGP peers. In practice it may just equal to new short configuration and establishment of new BGP sessions to IX peers.

9. Security considerations

The new mechanism for diverse BGP path dissemination proposed in this document does not introduce any new security concerns as compared to base BGP4 specification [RFC4271].

10. IANA Considerations

The new mechanism for diverse BGP path dissemination does not require any new allocations from IANA.

11. Contributors

The following people contributed significantly to the content of the document:

Selma Yilmaz
Cisco Systems
170 West Tasman Drive
San Jose, CA 95134
US
Email: seyilmaz@cisco.com

Satish Mynam
Cisco Systems
170 West Tasman Drive
San Jose, CA 95134
US
Email: mynam@cisco.com

Isidor Kouvelas
Cisco Systems
170 West Tasman Drive
San Jose, CA 95134
US
Email: kouvelas@cisco.com

12. Acknowledgments

The authors would like to thank Bruno Decraene, Bart Peirens, Eric Rosen, Jim Uttaro, Renwei Li and George Wes for their valuable input.

The authors would also like to express special thank you to number of operators who helped to optimize the provided solution to be as close as possible to their daily operational practices. Especially many thx goes to Ted Seely, Shan Amante, Benson Schliesser and Seiichi Kawamura.

13. References

13.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.

[RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008.

13.2. Informative References

[I-D.decraene-bgp-graceful-shutdown-requirements]
Decraene, B., Francois, P., pelsser, c., Ahmad, Z., and A. Armengol, "Requirements for the graceful shutdown of BGP sessions", draft-decraene-bgp-graceful-shutdown-requirements-01 (work in progress), March 2009.

[I-D.ietf-idr-add-paths]
Walton, D., Retana, A., Chen, E., and J. Scudder, "Advertisement of Multiple Paths in BGP", draft-ietf-idr-add-paths-04 (work in progress), August 2010.

[I-D.ietf-idr-best-external]
Marques, P., Fernando, R., Chen, E., and P. Mohapatra, "Advertisement of the best external route in BGP", draft-ietf-idr-best-external-02 (work in progress), August 2010.

[I-D.ietf-idr-route-oscillation]
McPherson, D., "BGP Persistent Route Oscillation Condition", draft-ietf-idr-route-oscillation-01 (work in progress), February 2002.

[I-D.pmohapat-idr-fast-conn-restore]
Mohapatra, P., Fernando, R., Filsfils, C., and R. Raszuk, "Fast Connectivity Restoration Using BGP Add-path", draft-pmohapat-idr-fast-conn-restore-00 (work in progress), September 2008.

[I-D.raszuk-idr-ibgp-auto-mesh]
Raszuk, R., "IBGP Auto Mesh", draft-raszuk-idr-ibgp-auto-mesh-00 (work in progress), June 2003.

[RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, April 2006.

[RFC5065] Traina, P., McPherson, D., and J. Scudder, "Autonomous System Confederations for BGP", RFC 5065, August 2007.

Authors' Addresses

Robert Raszuk (editor)
Cisco Systems
170 West Tasman Drive
San Jose, CA 95134
US

Email: raszuk@cisco.com

Rex Fernando
Cisco Systems
170 West Tasman Drive
San Jose, CA 95134
US

Email: rex@cisco.com

Keyur Patel
Cisco Systems
170 West Tasman Drive
San Jose, CA 95134
US

Email: keyupate@cisco.com

Danny McPherson
Verisign
21345 Ridgetop Circle
Dulles, VA 20166
US

Email: dmcpherson@verisign.com

Kenji Kumaki
KDDI Corporation
Garden Air Tower
Iidabashi, Chiyoda-ku, Tokyo 102-8460
Japan

Email: ke-kumaki@kddi.com

Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: August 24, 2011

R. Shakir
C&W

February 20, 2011

Operational Requirements for Enhanced Error Handling Behaviour in BGP-4
draft-shakir-idr-ops-reqs-for-bgp-error-handling-01

Abstract

BGP-4 is utilised as a key intra- and inter-Autonomous System routing protocol in modern IP networks. The failure modes as defined by the original protocol standards are based on a number of assumptions around the impact of session failure. Numerous incidents both in the global Internet routing table and within Service Provider networks have been caused by strict handling of a single invalid UPDATE message causing large-scale failures in one or more Autonomous Systems.

This memo describes the current use of BGP-4 within Service Provider networks, and outlines a set of requirements for further work to enhance the mechanisms available to a BGP-4 implementation when erroneous data is detected. Whilst this document does not provide specification of any standard, it is intended as an overview of a set of enhancements to BGP-4 to improve the protocol’s robustness to suit its current deployment.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on August 24, 2011.

Copyright Notice

Table of Contents

1. Introduction

Where BGP-4 [RFC4271] is deployed in the Internet and Service Provider networks, numerous incidents have been recorded due to the manner in which [RFC4271] specifies errors in routing information should be handled. Whilst the behaviour defined in the existing standards retains utility, the deployments of the protocol have changed within modern networks, resulting in significantly different demands for protocol robustness. Whilst a number of Internet Drafts have been written to begin to enhance the behaviour of BGP-4 in terms of the handling of erroneous messages, this draft intends to define a set of requirements for ongoing work. These requirements are considered from the perspective of a Network Operator, and hence this draft does not intend to define the protocol mechanisms by which such error handling behaviour is to be implemented.

1.1. Role of BGP-4 in Service Provider Networks

BGP was designed as an inter-Autonomous System (AS) routing protocol and hence many of the error handling mechanisms within the protocol specification are designed to be conducive to this role. In general, this consideration as an inter-AS routing propagation mechanism results in the view that a BGP session propagates a relatively small amount of network-layer reachability information (NLRI) between two ASes. In this case, it is the expectation of session resilience for those adjacencies that are key to routing continuity (for example, it is expected that two networks peering via BGP would connect multiple times in order to safeguard equipment or protocol failure). In addition, there is some expectation of multiple paths to a particular NLRI being available - it would be expected that a network can fall back to utilising alternate, less direct, paths where a failure of a more direct path occurs.

Traditional network architectures would deploy an Interior Gateway Protocol (IGP) to carry infrastructure and customer prefixes, with an Exterior Gateway Protocol (EGP) such as BGP being utilised to propagate these prefixes to other Autonomous Systems. However, with the growth of IP-based services, this is no longer considered best practice. In order to ensure that convergence is within acceptable time bounds, the amount of routing information carried within the IGP is significantly reduced - and tends to be only infrastructure prefixes. iBGP is then utilised to propagate both customer, and external prefixes within an AS. As such, BGP has become an IGP, with traditional IGPs acting as a means by which to propagate the routing information which is required to establish a BGP session, and reach the egress node within the local routing domain. This change in role presents different requirements for the robustness of BGP as a routing protocol - with the expectation of similar level of

robustness to that of an IGP being set.

Along with this change in role, the nature of the IP routing information that is carried has changed. BGP has become a ubiquitous means by which service information can be propagated between devices. For instance, BGP is utilised to carry routing information for IP/MPLS VPN services as described in [RFC4364]. Since there is an existing deployment of the protocol between PE devices in numerous networks, it has been adapted to propagate this routing information, as its use limits number of routing protocols required on each device. This additional information being propagated represents a large change in requirement for the error handling of the protocol - where session failure occurs, it is likely a complete service outage for at least a subset of a network's customers is experienced where an erroneous packet may have occurred within a different sub-topology or even service (a different address family for example). For this reason, there is a significant demand to avoid service affecting failures that may be triggered by routing information within a single sub-topology or service.

Both within Internet and multi-service routing architectures, a number of BGP sessions propagate a large proportion of the required routing information for network operation. For Internet routing, these are typically BGP sessions which propagate the global routing table to an AS - failure of these sessions may have a large impact on network service, based on a single erroneous update. In an multi-service environment, typical deployments utilise a small number of core-facing BGP sessions, typically towards route reflector devices. Failure of these sessions may also result in a large impact to network operation. Clearly, the avoidance of conditions requiring these sessions to fail is of great utility to any network operator, and provides further motivation for the revision of the existing behaviour.

Whilst the behaviour in [RFC4271] is suited to ensuring that BGP messages with erroneous routing information in are limited in scope (by means of session reset), with the above considerations, it is clear that this mechanism is not suited to all deployments. It should, however, be noted that the change in scope affects the handling only of errors occurring after BGP session establishment. There is no current operational requirement to amend the means by which error handling in session establishment, or liveliness detection, are performed.

1.2. Overview of Operator Requirements for BGP-4 Error Handling

It is the intention of this document to define a set of criteria for the manner in which a revised error handling mechanism in BGP-4 is

required to conform. The motivation for the definition of these requirements can be summarised based on certain behaviour currently present in the protocol that is not deemed acceptable within current operational deployments, or where there is a short-fall in the tool set available to an operator. These key requirements can be summarised as follows:

- It is unacceptable within modern deployments of the BGP-4 protocol that a single erroneous UPDATE packet affects prefixes that it does not carry. This requirement therefore requires some modification to the means by which erroneous UPDATE packets are handled, and reacted to - with a particular focus on avoiding the use of the NOTIFICATION message.

- It is recognised that some error conditions may occur within the BGP-4 protocol may not always be handled gracefully, and may result in conditions whereby an implementation cannot recover. In these (and similar) cases, it is unacceptable for an operator that this reset of the BGP-4 session results in interruption to forwarding packets (by means of withdrawing prefixes installed by BGP-4 into a device's RIB, and subsequently FIB). To this end, there is a requirement to define a session reset mechanism which provides session re-initialisation in a non-destructive manner.

- Further to the requirements to provide a more robust protocol, the current visibility into error conditions within the BGP-4 protocol is extremely limited - where further modifications to this behaviour are to be made, complexity is likely to be added. Thus, to ensure that BGP-4 is manageable, there are requirements for mechanisms by which the protocol can be examined and monitored.

This document describes each of these requirements in further depth, along with an overview of means by which they are expected to be achieved. In addition, the mechanism by which the enhancements meeting these requirements are to interact is discussed.

2. Avoiding use of NOTIFICATION

The error handling behaviour defined in RFC4271 is problematic due to the limited options that are available to an implementation. When an erroneous BGP message is received, at the current time, the implementation must either ignore the error, or send a NOTIFICATION message, after which it is mandatory to terminate the BGP session. It is apparent that this requirement is at odds with that of protocol robustness.

There is significant complexity to this requirement. The mechanism defined in [I-D.chen-ebgp-error-handling] describes a means by which no NOTIFICATION message is generated for all cases whereby NLRI can be extracted from an UPDATE. The NLRI contained within the erroneous UPDATE message is considered as though the remote BGP speaker has provided an UPDATE marking it as withdrawn. This results in a limit in the propagation of the invalid routing information, whilst also ensuring that no traffic is forwarded via a previously-known path that may no longer be valid. This mechanism is referred to as "treat-as-withdraw".

Whilst this behaviour results in avoiding a NOTIFICATION message, keeping other routing information advertised by the remote BGP speaker within the RIB, it may result in unreachability for a sub-set of the NLRI advertised by the remote speaker. Two cases should be considered - that where the entry for a prefix in the Adj-RIB-In of the neighbour propagating an erroneous packet is utilised, and that where the prefix installed in the device’s RIB is learnt from another BGP speaker. In the former case, should the identified NLRI not be treated as withdrawn, the original NLRI is utilised within the global RIB. However, this information is potentially now invalid (i.e. it no longer provides a valid forwarding path), whilst an alternate (valid) path may exist in another Adj-RIB-In. By continuing to utilise the NLRI for which the UPDATE was considered invalid, traffic may be forwarded via an invalid path, resulting in routing loops, or black-holing. In the second case, no impact to the forwarding of traffic, or global RIB, is incurred, yet where treat-as-withdraw is implemented, possibly stale routing information is purged from the Adj-RIB-In of the neighbour propagating errors.

Whilst mechanisms such as "treat-as-withdraw" are currently documented, the proposals are limited in their scope - particularly in terms of restrictions to implementation only on eBGP sessions. This limitation is made based on the view that the BGP RIB must be consistent across an autonomous system. By implementing treat-as-withdraw for a iBGP session, one or more routers within the Autonomous System may not have reachability to a prefix, and hence blackholing of traffic, or routing loops, may occur. It should,

however, be considered if this view is valid, in light of the manner in which BGP is utilised within operator networks. Inconsistency in a RIB based on a single UPDATE being treated as withdrawn may cause a inconsistency in a single sub-topology (e.g. Layer 3 VPN service), or a service not operating completely (in the case of an UPDATE carrying service membership information). Where a NOTIFICATION and teardown is utilised this is destructive to all sub-topologies in all address family identifiers (AFIs) carried by the session in question. Even where mechanisms such as multi-session BGP are utilised, a whole AFI is affected by such a NOTIFICATION message. In terms of routing operation, it is therefore far less costly to endure a situation where a limited sub-set of routing information within an AS is invalid, than to consider all routing information as invalid based on a single trigger.

It is considered that, if extended to cover iBGP, the mechanisms described in [I-D.chen-ebgp-error-handling] and [I-D.ietf-idr-optional-transitive] provide a means to avoid the transmission of a NOTIFICATION to a remote BGP speaker based on a single erroneous message, where at all possible, and hence meet this requirement. The failure cases whereby NLRI cannot be extracted from the UPDATE message represent a case whereby the receiving system cannot handle the error gracefully based on this mechanism.

3. Recovering RIB Consistency

The recommendations described in Section 2 may result in the RIB for a topology within an AS being inconsistent across the AS' internal routers. Alternatively, where such mechanisms are deployed at an AS boundary, interconnects between two ASes may be inconsistent with each other. There are therefore risks of traffic blackholing, due to missing routing information, or forwarding loops. Whilst this is deemed an acceptable compromise in the short term, clearly, it is suboptimal. Therefore, a requirement exists to provide mechanisms by which a BGP speaker is able to recover the consistency of the Adj-RIB-In for a particular neighbour.

It is envisaged that during such routing inconsistencies, the local BGP speaker is aware that some routing information was not able to be processed - due to the fact that an UPDATE message was not parsed correctly. If the 'treat-as-withdraw' mechanism described within Section 2 is utilised, it is also possible for the local BGP speaker to have determined the set of NLRI for which an erroneous UPDATE message was received. In this scenario, by utilising targeted mechanisms to re-request the specific NLRI that was unreachable, this routing information can be re-transmitted from the remote BGP speaker. Such a request requires extension to the existing BGP-4 protocol, in terms of specific UPDATE generation filters with a transient lifetime. It is envisaged that the work within [I-D.zeng-one-time-prefix-orf] provides a mechanism allowing targeted elements of the Adj-RIB-In for a BGP neighbour to be recovered.

In addition to such cases where specific routing information is known to be erroneous, the more general case where either a large amount of the Adj-RIB-In is contained in UPDATE messages subject to treat-as-withdraw, or the specific prefixes are unknown to the local BGP speaker must be considered. In this case, there is a requirement for a BGP speaker to re-request the entire RIB advertised by a remote neighbour. In this case, where such re-advertisement is required, it is envisaged that a ROUTE-REFRESH as per the description in [RFC2918] is utilised. [I-D.keyur-bgp-enhanced-route-refresh] provides a means by which the ROUTE-REFRESH mechanism can be extended in order to meet this requirement.

It is of particular note for both means of recovering RIB consistency described that these are effective only when considering transitive errors within an implementation - for instance, should an RFC interpretation error within an implementation be present, regardless of the number of times a specific UPDATE is generated, it is likely that this error condition will persist. For this reason, there is an requirement to consider the means by which such consistency recovery mechanisms are utilised. It is not advisable that a transitive

filter and advertisement mechanism is triggered by all error handling events due to the load this is likely to place on the neighbour receiving such a request. Where this BGP speaker is a relatively centralised device - a route reflector (as described by [RFC4456]) for example - the act of generation of UPDATE messages with such frequency is likely to cause disproportionate load. It is therefore an operational requirement of such mechanisms that means of request dampening be required by any such extension.

4. Reducing the Impact of Session Reset

Even where protocol enhancements allow errors in the BGP-4 protocol to cease to trigger NOTIFICATION messages, and hence reset a BGP session, it is clear that some error conditions may not be exited. In particular, errors due to existing state, or memory structures, associated with a specific BGP session will not be handled. It is therefore important to consider how these error conditions are currently handled by the protocol. It should be noted that the following discussion and analysis considers only those NOTIFICATION messages generated in response to errors in UPDATE messages (as defined by Section 6.3 in [RFC4271]).

The existing NOTIFICATION behaviour triggers a reset of all elements of the BGP-4 session, as described in Section 6 of [RFC4271]. It is expected that session teardown requires an implementation to re-initialise all structures and state required for session maintenance. Clearly, there is some utility to this requirement, as error conditions in BGP are, in general, exited from. However, this definition is responsible for the forwarding outages within networks utilising BGP for route propagation when each error is experienced. The requirement described in Section 2 is intended to reduce the cases whereby a NOTIFICATION is required, however, any mechanism implemented as a response to this requirement by definition cannot provide a session reset to the extent of that achieved by the current behaviour.

In order to address this, there is a requirement for a means by which a BGP speaker can signal that an unhandled error condition in an UPDATE message occurred - requiring a session reset - yet also continue to utilise the paths advertised by the neighbour that are currently in use within the RIB. In this case, the Adj-RIB-In received from the neighbour is not considered invalid, despite a NOTIFICATION, and session reset, being required. This set of requirements is akin to those answered by the BGP Graceful Restart mechanism described in [RFC4724]. Since the operational requirement in this case is to provide a means to achieve a complete session restart without disrupting the forwarding path of those prefixes in use within a BGP speaker’s RIB, it is expected that utilising a procedure similar to the Graceful Restart mechanism meets the error handling requirement. By responding to an error condition (repeated or otherwise) with a message indicating that an error that cannot be handled has occurred, forcing session reset, whilst retaining forwarding information within the RIB allows forwarding to all prefixes within a system’s RIB to continue, whilst the session restarts. By placing a time bound on the restart lifetime, should an error condition not be transient - for example, should an error have occurred with the BGP process, rather than a specific of the BGP

session - the remote BGP speaker is still detected as an invalid device for forwarding.

It should, however, be noted that a protocol enhancement meeting this requirement is not able to solve all error conditions - however, a complete restart of the BGP and TCP session between two BGP speakers implements an identical recovery mechanism to that which is achieved by the existing behaviour. Where an error condition such as memory or configuration corruption has occurred in a BGP implementation, it is expected that a mechanism meeting this requirement continues to detect this, by means of a bound on time for session restart to occur. Whilst there may be some consideration that packets continue to be forwarded through a device which can be in an failure mode of this nature for a longer period, due to this requirement, the architecture of modern IP routers should be considered. A divided forwarding and control plane is common in many devices, as well as process separation for software-based devices - corruption of a specific protocol daemon does not necessarily imply forwarding is affected. Indeed, where forwarding behaviour of a device is affected, it is envisaged that a failure detection mechanism (be it Bidirectional Forwarding Detection, or indeed BGP KEEPALIVE packets) will detect such a failure in almost all cases, with the symptomatic behaviour of such a failure being an invalid UPDATE message in very few other cases.

5. Operational Toolset for Monitoring BGP

A significant complexity that is introduced through the requirements defined in this document is that of monitoring BGP session status for an operator. Although the existing error handling behaviour causes a disproportionate failure, session failure is extremely visible to most operational personnel within a Network Operator due to both existing definitions of SNMP trap mechanisms for BGP, along with the forwarding impact typically caused by such a failure. By introducing mechanisms by which errors of this nature are not as visible, this is no longer the case. There is a requirement that where subsets of the RIB on a device are no longer reachable from a BGP speaker, or indeed an AS, that some mechanism to determine the cause is available to an operator. Whilst, to some extent, this can be solved by mandating a sub-requirement of each of the aforementioned requirements that a BGP speaker must log where such errors occur, and are hence handled, this does not solve all cases. In order to clarify this requirement, the example of the transmission of an erroneous Optional Transitive attribute can be considered. Since, by definition, there is no requirement for all BGP speakers to parse such an attribute, a receiving router may treat NLRI as withdrawn based on an erroneous attribute not examined by its neighbour. In this case, the upstream device or network, propagating the UPDATE, has no visibility of this error. Operationally, however, it is of interest to the upstream router operator that such invalid information was propagated.

The requirement for logging of error conditions in transmitted BGP messages, which are visible to only the receiver, cannot be achieved by any existing BGP message, or capability. It is envisaged that each erroneous event should be transmitted to the remote peer - including the information as to the set of NLRI that were considered invalid. Whilst with some mechanisms this is achieved by default (for example, One-Time Prefix ORF [I-D.zeng-one-time-prefix-orf] (Outbound Route Filtering) will transmit the set of prefixes that are required), the operator requirement is to know which prefixes may have been unreachable in all cases. It is envisaged that an extension to meet this requirement will allow for such information to be transmitted between peers, and hence logged. Such a mechanism may provide further utility as a either a diagnostic, or logging toolset.

It should be noted that numerous work items within the IETF exist at the time of writing that begin to solve this requirement. Within the IDR working group both [I-D.raszuk-bgp-diagnostic-message] and [I-D.ietf-idr-advisory] provide mechanisms by which such information can be propagated in-band to an existing BGP session. Transmitting such diagnostic information in-band is considered the optimal means by which to propagate details of errors present in UPDATE messages, due to the fact that no additional protocols (and hence security and

trust concerns) must be configured between two Autonomous Systems (where the errors occur at an AS boundary), and the load on each BGP speaker is increased only due to an additional capability, rather than an additional code base, and protocol. Clearly, any mechanism implemented in-band to a BGP session is required to be relatively lightweight, since the information provided over the session is an enhancement to the operational visibility of the protocol, and should not disrupt core protocol operations. Other, out-of-band, mechanisms - such as that proposed in [I-D.ietf-grow-bmp] are likely to provide mechanisms by which further insight into BGP operation can be achieved. The fact that such a protocol is implemented independently of the BGP protocol results in further flexibility to provide detailed protocol data, without introducing further complexity to the BGP protocol itself.

6. Operational Complexities Introduced by Altering RFC4271

The existing NOTIFICATION and subsequent teardown of a BGP session upon encountering an error has the advantage that a consistent approach to error handling is required of all implementations of the BGP-4 protocol. This is of operational advantage, as it provides a clear expectation of the behaviour of the protocol. The requirements defined herein add further complexity to the error-handling within BGP, and hence are liable to compromise the existing deterministic protocol behaviour. It is therefore deemed that there is a further requirement to provide a clear method by which an erroneous UPDATE should be reacted to, in order that all protocol implementations provide a consistent means by which recovery is achieved. A further complexity is introduced due to the disparate nature of the work items altering the BGP error handling behaviour - since all items are likely to be implemented as a BGP capability [RFC5492], situations are likely to occur between devices (especially those with different BGP implementations), where some of the mechanisms referenced are unsupported. This adds further barriers to a standard definition of the BGP-4 error handling behaviour.

In general, the approach considered ideal upon encountering an erroneous UPDATE message can be divided into two cases - those where the NLRI can be determined from the message, and those where it cannot be. The latter case is the simpler of the two. In this case, there is a requirement for the implementation to reset the BGP session, utilising the reduced-impact approach, described in Section 4. In the case where the remote BGP speaker is in a transient error condition related to specific peer data structures, or state, a single instance of this behaviour is likely to exit the error condition. In the case of implementation errors, it is possible that the BGP session in question may enter a continuous loop of being reset, with a partial RIB being held by one or more of the BGP speakers due to an non-deterministic order of UPDATE propagation. It is therefore a requirement that within this reduced-impact procedure any subsequent UPDATE messages that would result in further session resets are ignored. Whilst this results in a condition where an undetermined amount of the RIB is inconsistent, partial reachability is maintained. In this case, the operational toolsets discussed in Section 5 is likely to provide mechanisms by which this condition can be brought to the attention of the relevant operators. This requirement to accept a partial RIB, which results in potential invalid traffic forwarding is a direct result of the deployments of BGP-4, as described in Section 1.1.

The case where NLRI can be determined from an erroneous UPDATE provides further complexities. In this case, a BGP speaker is aware of the sub-set of the RIB which have been identified as being

contained within invalid UPDATE messages. This allows a local BGP speaker to re-request single prefixes, utilising a mechanism such as "one-time prefix ORF". However, a similar result is achieved by re-requesting the entire RIB - albeit with greater resource requirements. It is therefore expected that the process of recovery utilises a staged set of mechanisms to attempt to restore consistency of the RIB:

1. Where available, a mechanism capable of requesting only the NLRI determined to have been contained within a invalid UPDATE should be utilised. However, since it is possible that such an error condition can be transient in nature, it is likely that more than one request is to be transmitted (assuming the first does not return a valid UPDATE message). In order to allow a deterministic process, there is a requirement for a limit on the number of specific requests transmitted to be defined.

2. Where a specific refresh mechanism is not available, a peer should re-request the entire RIB. Again, there is a requirement to limit the number of complete RIB requests that should be sent via an implementation, in order to provide a bound both on the expected level of load a device may experience, and on the time for which the RIB may be inconsistent.

3. Finally, a session reset should be performed, as per the reduced-impact NOTIFICATION requirement defined in Section 4. At this point, a similar challenge to that discussed above exists, should the error condition persist. In this case, as defined above, there is a requirement to ignore those UPDATE messages that continue to be erroneous.

It is envisaged that where limits are required, these will be defined on a per memo-basis, or within a further revision of the requirements described herein.

Whilst the approach described above provides a standard means by which error recovery may be handled on a per UPDATE basis, further complexities are raised where multiple errors occur. Clearly, following this procedure causes control-plane load on both the BGP speakers - for this reason, consideration of how repeated use of the mechanisms discussed in this document is required. It is notable that errors may not occur with UPDATE messages relating to only a single NLRI, independent errors in multiple NLRIs may be experienced. For this reason, it is required that an implementation rate limits the number of error handling events sourced towards a particular neighbour. It is expected that such rate limiting, or event suppression is achieved on a per-session basis, where state information is already held, rather than on a per-prefix basis as it

is envisaged that such behaviour presents significant scaling problems, and introduces further state requirements for an implementation of the protocol. It is recommended that where a flag indicative of erroneous behaviour is implemented, the state of such a value is maintained independently of session establishment.

7. IANA Considerations

This memo includes no request to IANA.

8. Security Considerations

The requirements outlined in this document provide mechanisms by which erroneous BGP messages may be responded to with limited impact to forwarding operation. This is of benefit to the security of a BGP speaker in general. Where UPDATE messages may have been propagated by a single malicious Autonomous System or router within a network (or the Internet default free zone - DFZ), which are then propagated to all devices within the same routing domain, all other NLRI available over the same session become unreachable. This mechanism may provide means by which an Autonomous System can be isolated from required routing domains (such as the Internet), should the relevant UPDATE messages be propagated via specific paths. By reducing the impact of such failures, it is envisaged that this possibility may be constrained to a specific set of NLRI, or a specific topology.

Some mechanisms meeting the requirements specified in this document, particularly those within Section 5 may provide further security concerns, however, it is envisaged that these are addressed in per-enhancement memos.

9. Acknowledgements

The author would like to thank Rob Evans, David Freedman, Tom Hodgson, Sven Huster, Jonathan Newton, Neil McRae, Thomas Mangin, Tom Scholl and Ilya Varlashkin for their review and valuable feedback.

10. References

10.1. Normative References

[I-D.chen-ebgp-error-handling]
Chen, E., Mohapatra, P., and K. Patel, "Revised Error Handling for BGP Updates from External Neighbors", draft-chen-ebgp-error-handling-00 (work in progress), September 2010.

[I-D.ietf-grow-bmp]
Scudder, J., Fernando, R., and S. Stuart, "BGP Monitoring Protocol", draft-ietf-grow-bmp-05 (work in progress), December 2010.

[I-D.ietf-idr-advisory]
Scholl, T., Scudder, J., Steenbergen, R., and D. Freedman, "BGP Advisory Message", draft-ietf-idr-advisory-00 (work in progress), October 2009.

[I-D.ietf-idr-optional-transitive]
Scudder, J. and E. Chen, "Error Handling for Optional Transitive BGP Attributes", draft-ietf-idr-optional-transitive-03 (work in progress), September 2010.

[I-D.keyur-bgp-enhanced-route-refresh]
Patel, K., Chen, E., and B. Venkatachalapathy, "Enhanced Route Refresh Capability for BGP-4", draft-keyur-bgp-enhanced-route-refresh-01 (work in progress), October 2010.

[I-D.raszuk-bgp-diagnostic-message]
Raszuk, R., Chen, E., and B. Decraene, "BGP Diagnostic Message", draft-raszuk-bgp-diagnostic-message-00 (work in progress), October 2010.

[I-D.zeng-one-time-prefix-orf]
Zeng, Q. and J. Dong, "One-time Address-Prefix Based Outbound Route Filter for BGP-4", draft-zeng-one-time-prefix-orf-01 (work in progress), October 2010.

[RFC2918] Chen, E., "Route Refresh Capability for BGP-4", RFC 2918, September 2000.

[RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.

[RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, April 2006.

[RFC4724] Sangli, S., Chen, E., Fernando, R., Scudder, J., and Y. Rekhter, "Graceful Restart Mechanism for BGP", RFC 4724, January 2007.

[RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, February 2009.

10.2. Informational References

[RFC5881] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD) for IPv4 and IPv6 (Single Hop)", RFC 5881, June 2010.

Author's Address

Rob Shakir
Cable&Wireless Worldwide

Email: rob.shakir@cw.com

Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: September 15, 2011

S. Tsuchiya, Ed.
Cisco Systems
S. Kawamura
NEC BIGLOBE, Ltd.
R. Bush
C. Pelsser
Internet Initiative Japan, Inc.

March 14, 2011

Route Flap Damping Deployment Status Survey
draft-shishio-grow-isp-rfd-implement-survey-01

Abstract

BGP Route Flap Damping [RFC2439] is a mechanism that targets route stability. It penalyzes routes that flap with the aim of reducing CPU load on the routers.

But it has side-effects. Thus, in 2006, RIPE recommended not to use Route Flap Damping (see RIPE-378).

Now, some researchers propose to turn RFD, with less aggressive parameters, back on [draft-ymbk-rfd-usable].

This document describes results of a survey conducted amoung service provider on their use of BGP Route Flap Damping.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on September 15, 2011.

Copyright Notice

Table of Contents

1. Survey Purpose

RIPE published some recommendations such as RIPE-178 [RIPE-178],RIPE-210 [RIPE-210],RIPE-229 [RIPE-229] and RIPE-378 [RIPE-378].

The purpose of this survey is to understand the current usage and requirements of Route Flap Damping [RFC2439] among service providers.

2. Survey’s target and period

2.1. For Japan

Target: Japan Network Operator Group janog@janog.gr.jp

Period: Jan 28,2011 - Feb 12,2011

2.2. All

Target: All operators that will answer the survey following the publication of this document.

Period:Mar 7,2011 - May 25,2011

Please open the following url and answer the questionaire.

https://www.surveymonkey.com/s/rfd-survey

3. Survey Results

3.1. Q1.Do you use Route Flap Damping ?

3.1.1. Japan

Yes: 5

No: 13

1 respondant skipped this question

3.1.2. All

No results yet!

3.2. Q2.If you select No on Q1,why?

3.2.1. Japan

Do not have the need: 3

Did not know about the feature: 2

No benefits expected: 3

Customers would complain:1

Because I read RIPE-378 [RIPE-378]:2

Other: 3

3.2.2. All

No results yet!

3.3. Q3.If you select Yes on Q1,what parameter do you use?

3.3.1. Japan

Default parameters: 3

RIPE-178 [RIPE-178]: 0

RIPE-210 [RIPE-210]: 0

RIPE-229 [RIPE-229]: 0

Other: 3

1 person answered Q3, even if he selected "No" on Q1.

3.3.2. All

No results yet!

3.4. Q4.Do you know Randy Bush et. al's report ''Route Flap Damping Considered Usable?''

3.4.1. Japan

Yes: 12

No: 7

One person skipped Q1, but answered Q4.

3.4.2. All

No results yet!

3.5. Q5.IOS’s max-penalty is currently limited to 20K. Do you need this limitation to be relaxed to over 50K?

3.5.1. Japan

Yes: 10

No: 9

3.5.2. All

No results yet!

3.6. Q6.If you have any comments, please fill this box.

Free format

3.6.1. Japan

-Our peer seems to have damping enabled, and our prefix gets damped sometimes.

-We do not enable damping because we think that customers want a non-damped route.

-From the perspective of a downstream ISP, if our upstream told us that an outage occurred because a route was damped, I may call and ask "is it written in the agreement that you will do this?"

-We use damping pretty heavily

-I had RFD turned on until this morning when I discovered our router has CSCtd26215 issues. I would like to turn on a "useful" RFD.

3.6.2. All

No results yet!

4. Summary of data

From the survey we see that there are many service providers with RFD

disabled. The reason varies among providers, but it is clear that there are those who wish that RFD was made useful. [draft-ymbk-rfd-usable] describes how to improve RFD with minor changes to some parameters. From the comments in the survey, the most significant fear of enabling RFD is its impact on customers.

5. Acknowledgements

We thank the 19 respondant to this survey.

6. IANA Considerations

This document has no actions for IANA.

7. Security Considerations

This document has no security considerations.

8. References

8.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC2439] Villamizar, C., Chandra, R., and R. Govindan, "BGP Route Flap Damping", RFC 2439, November 1998.

8.2. Informative References

[I-D.ymbk-rfd-usable]
Pelsser, C., Bush, R., Patel, K., Mohapatra, P., and O. Maennel, "Making Route Flap Damping Usable", draft-ymbk-rfd-usable-00 (work in progress), March 2011.

[RIPE-178]
Barber, T., Doran, S., Panigl, C., and J. Schmitz, ""RIPE Routing-WG Recommendation for coor-dinated route-flap damping parameters"", Feb 1998, <ftp://ftp.ripe.net/ripe/docs/ripe-178.txt>.

[RIPE-210]
Barber, T., Doran, S., Karrenberg, D., Panigl, C., and J. Schmitz, ""RIPE Routing-WG Recommendation for coordinated

route-flap damping parameters"", May 2000, <ftp://ftp.ripe.net/ripe/docs/ripe-210.txt>.

[RIPE-229]
Panigl, C., Schmitz, J., Smith, P., and C. Vistoli, ""RIPE Routing-WG Recommendations for Coordinated Route-flap Damping Parameters"", Oct 2001, <ftp://ftp.ripe.net/ripe/docs/ripe-229.txt>.

[RIPE-378]
Smith, P. and C. Panigl, ""RIPE Routing Working Group Recommendations On Route-flap Damping"", May 2006, <http://www.ripe.net/ripe/docs/ripe-378>.

[Route Flap Damping Considered Usable?]
Pelsser, C., Maennel, O., Patel, K., and R. Bush, ""Route Flap Damping Considered Useable"", Nov 2011, <http://ripe61.ripe.net/presentations/222-101117.ripe-rfd.pdf>.

Appendix A. Additional Stuff

This becomes an Appendix.

Authors' Addresses

Shishio Tsuchiya (editor)
Cisco Systems
Shinjuku Mitsui Building, 2-1-1, Nishi-Shinjuku
Shinjuku-Ku, Tokyo 163-0409
Japan

Phone: +81 3 6434 6543
Email: shtsuchi@cisco.com

Seiichi Kawamura
NEC BIGLOBE, Ltd.
14-22, Shibaura 4-chome
Minatoku, Tokyo 108-8558
JAPAN

Phone: +81 3 3798 6085
Email: kawamucho@mesh.ad.jp

Randy Bush
Internet Initiative Japan, Inc.
5147 Crystal Springs
Bainbridge Island, Washington 98110
US

Phone: +1 206 780 0431 x1
Email: randy@psg.com

Cristel Pelsser
Internet Initiative Japan, Inc.
Jinbocho Mitsui Buiding, 1-105
Kanda-Jinbocho, Chiyoda-kun 101-0051
JP

Phone: +81 3 5205 6464
Email: cristel@iij.ad.jp

Network Working Group
Internet-Draft
Intended status: Informational
Expires: July 7, 2011

Z. Uzmi
A. Tariq
LUMS
P. Francis
MPI-SWS

January 03, 2011

FIB Aggregation with SMALTA
draft-uzmi-smalta-01.txt

Abstract

Concerns about the growth of the global routing table has led to proposals for short-term FIB-reduction and long-term RIB-reduction solutions. The simplest type of FIB-reduction solution is "FIB aggregation", whereby individual routers locally reduce the FIB size without any changes to external operation. The draft [I-D.zhang-fibaggregation] describes and analyzes several FIB aggregation schemes. This current draft describes and analyzes another point in the design space, called SMALTA. Compared to the approaches in [I-D.zhang-fibaggregation], SMALTA provides better aggregation and does not introduce non-routable entries in the FIB, but is also more complex.

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on July 7, 2011.

Table of Contents

1. Introduction

FIB Aggregation is an approach to shrinking the size of a router's FIB without requiring any changes to the external behavior of the router. While FIB Aggregation per se is not a new idea, interest has recently increased alongside the growing concern over global routing table growth. FIB Aggregation represents a short-term but simple fix that can extend the lifetime of existing routers as well as reduce FIB size requirements for routers in the future. The draft [I-D.zhang-fibaggregation] provides a very good overview of the problem space, and so is not repeated here.
[I-D.zhang-fibaggregation] also describes four FIB Aggregation algorithms, called Level 1 through Level 4, each adding complexity but providing better aggregation over its predecessor.

These 4 levels all have the characteristic that they require a complete crawl through the FIB to produce full aggregation. They can also be incrementally updated to respond to changes in routes, but these incremental updates don't maintain full aggregation. The third and fourth levels have the characteristic that they introduce non-routable space into the routing table. In other words, packets destined for prefixes that are not in the RIB, and would therefore otherwise be dropped, may nevertheless be forwarded towards the destination.

This draft introduces another FIB aggregation algorithm, SMALTA, that can be viewed as a 5th level in that it is yet more complex, and provides still better aggregation. Like the first four levels, SMALTA's incremental updates cause the FIB to deviate from its fully aggregated state and, therefore, also requires a periodic re-compute of the full FIB. Unlike the third and fourth levels, SMALTA introduces no non-routable space into the FIB. It is semantically identical to the non-aggregated FIB.

In a nutshell, what distinguishes SMALTA from the first four levels is the following. In the first four levels, aggregation is always done either by assigning a nexthop to an ancestor prefix, or by removing nexthop(s) of descendent prefix(es). In other words, if there is a prefix in the pre-aggregated tree, it will either remain unchanged or will be aggregated with a less specific prefix in the aggregated tree. SMALTA provides additional reduction in the FIB by allowing prefixes to be de-aggregated i.e. to be split into one or more specific prefixes.

2. Requirements notation

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT",

"SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

3. Design of SMALTA

SMALTA is composed of two algorithms: the one-shot aggregation and an incremental update algorithm. The implementations of these algorithms assume a binary tree structure, but is not limited to this and can be easily applied to other data structures. Furthermore, similar to the example set in [I-D.zhang-fibaggregation], the one-shot aggregation and update algorithms in SMALTA do not introduce any new data structure in the RIB or the FIB, using the existing structures that stores these tables.

SMALTA one-shot aggregation algorithm takes the current snapshot of the RIB and produces an aggregated version of it. This aggregated table is downloaded to the FIB, as a monolithic download, during router startup or after a BGP hard reset. This algorithm can also be invoked at any time to improve the amount of aggregation, for example (i) at regular intervals to ensure that the number of entries in the forwarding table remains small, (ii) when the FIB size exceeds a threshold, or (iii) on-demand when there occurs a significant routing change such as addition or deletion of a physical interface, changes to the IGP metrics or a BGP session restart.

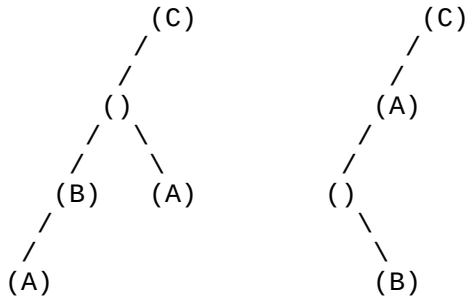
The update algorithm incrementally incorporates any subsequent BGP updates (withdraw or announce) into the aggregated table. Any resulting change in the aggregated table is then reflected in the FIB through an on-demand RIB to FIB download. The update algorithm is efficient for single updates, but slightly degrades the aggregated state of the table. The update algorithm is invoked whenever there is a change in the primary route of a prefix in the RIB. When that happens, SMALTA computes the necessary changes to the prefixes in the aggregated table. Note that some changes to the aggregated table result in no required changes to the FIB, and others may require multiple changes to the FIB. Section 4 gives results of measurements showing the average and worst-case number of FIB changes per BGP update.

In the following sections, we provide examples that give the basic idea of how the SMALTA algorithms work. Complete details are available in a technical report [TR.Uzmi-SMALTA]

3.1. Example 1: Simple SMALTA One-Shot Aggregation

Figure 1A represents an original (non-aggregated) table with four prefixes, using the convention introduced in

[I-D.zhang-fibaggregation].



ORIGINAL TABLE ONE-SHOT AGGREGATED TABLE

FIGURE 1A FIGURE 1B

Note that none of the Level 1 through Level 4 algorithms can further aggregate this table. By contrast, SMALTA would aggregate this table as shown in figure 1B.

With SMALTA, the prefix with next hop B in the original table is de-aggregated in the aggregated table. This de-aggregation creates an opportunity for the prefixes with next hop A to be aggregated. The resulting aggregated table is semantically equivalent to the original table.

SMALTA's one-shot aggregation is derived from ORTC [Paper.Draves-ORTC]. Like its precursor, one-shot runs three passes over a tree data structure:

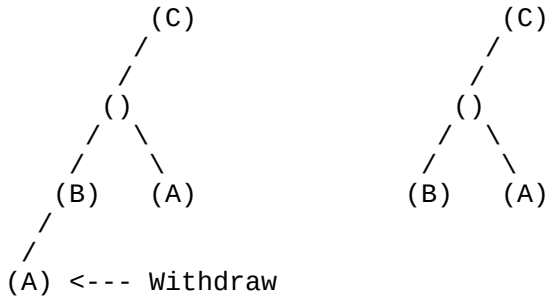
1. A normalization pass in which the prefixes are expanded such that every node in the binary tree has two or no children,
2. A post-order traversal up the tree, wherein each node is assigned a set of next hops, and
3. A pass where the algorithm assigns next hops to prefix nodes in the tree starting from the root and traversing through to the leaves, removing any unnecessary leaves.

Unlike ORTC, SMALTA one-shot is not provably optimal, though it is very close to optimal. This is because SMALTA places some minor constraints on the ORTC algorithm in order to allow the aggregated table and the original non-aggregated table to be implemented as a single data structure.

3.2. Example 2: SMALTA Update Algorithm for BGP WITHDRAW

To incorporate a withdraw for a prefix P1, the ownership of the IP space originally covered by P1 should be given up in favor of the next hop of its immediate ancestor prefix P in the original table. This is simply achieved by setting nexthop of P1 equal to that of P in the aggregated table. Next we restore all nearest descendent prefixes of P1 if their nexthop does not match the new nexthop of P1. Finally, we reclaim the space covered by the deaggregates of P1 by setting their nexthops equal to that of P.

Figure 2A shows the non-aggregated table before the withdraw, and indicates to which prefix the withdraw occurs. Figure 2B shows the non-aggregated table after the withdraw. Figure 2C shows the aggregated table that would result from applying the one-shot algorithm to the non-aggregated table of Figure 2A. Note that as a result of one-shot, there is in fact no nexthop in the entry to which the withdraw is being applied. Figure 2D shows the aggregated table after applying the update to the table of Figure 2C, using SMALTA update algorithm.

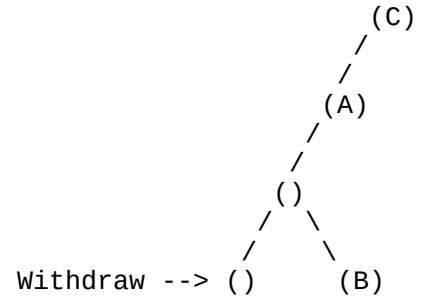


WITHDRAW IN ORIGINAL TABLE

FIGURE 2A

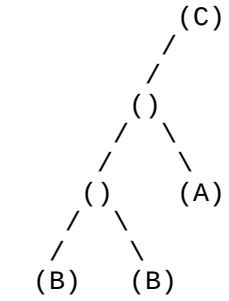
RESULT OF WITHDRAW ON ORIGINAL TABLE

FIGURE 2B



WITHDRAW IN AGGREGATED TABLE

FIGURE 2C



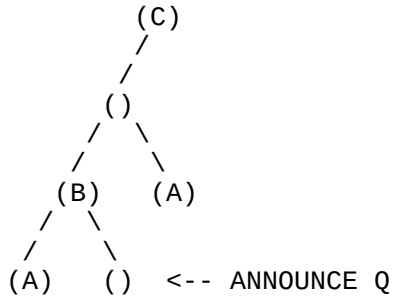
RESULT OF WITHDRAW
ON AGGREGATED TABLE

FIGURE 2D

3.3. Example 2: SMALTA Update Algorithm for BGP ANNOUNCE

In order to incorporate an announce for a prefix P1, the IP address space of P1 should be assigned the nexthop of P1 if that next hop does not match the next hop of P, the immediate ancestor of P1. Next, all nearest descendent prefixes of P1 that have a nexthop that differs from the announced nexthop are restored. Finally, the deaggregate prefixes of P that are specifics of P1 have their nexthops set to that of P1.

Figures 3A and 3C give the non-aggregated and aggregated tables respectively before the announce, and show to which prefix the announce takes place (corresponding to Figures 2A and 2C). Figure 3B shows the non-aggregated table after the BGP update, and Figure 3D shows the aggregated table after the SMALTA update algorithm is applied.

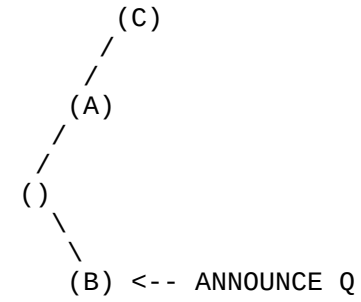


ANNOUNCE IN
ORIGINAL TABLE

FIGURE 3A

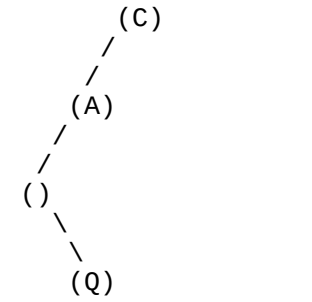
RESULT OF ANNOUNCE
ON ORIGINAL TABLE

FIGURE 3B



ANNOUNCE IN
AGGREGATED TABLE

FIGURE 3C



RESULT OF ANNOUNCE
ON AGGREGATED TABLE

FIGURE 3D

4. Analysis

4.1. SMALTA One Shot Aggregation

The extent of aggregation achieved by the SMALTA one-shot aggregation algorithm is given in Table 1. This table also lists the number of

entries in the aggregated table, obtained after applying Level 1 and Level 2 aggregation from [I-D.zhang-fibaggregation]. Level 3 and Level 4 aggregation results are not provided because these levels create extra routable space and hence can not be fairly compared with the SMALTA one-shot aggregation algorithm which does not create any extra routable space. (The authors have also produced a version of SMALTA that does create extra routable space, and provides substantially more aggregation than the version here. This algorithm may be described in a future version.)

It can be seen from Table 1 that SMALTA one-shot algorithm consistently provides better aggregation as compared to that provided by Level 1 and Level 2 algorithms.

In Table 1:

#(OT) = Number of entries in the original table

#(AT one-shot) = Number of entries in aggregated table after applying SMALTA one-shot algorithm. (% of #(OT) is also shown)

#Level 1 Aggregation = Number of entries in FIB after applying Level 1 Aggregation [I-D.zhang-fibaggregation].

#Level 2 Aggregation = Number of entries in FIB after applying Level 2 Aggregation [I-D.zhang-fibaggregation].

Year	#(OT)	#(AT) (one-shot)	#Level 1 Aggregation	#Level 2 Aggregation
2004	160818	54340 (33.7%)	103565 (64.39%)	69572 (43.2%)
2005	176474	55801 (31.6%)	109169 (61.86%)	73169 (41.4%)
2006	203082	75356 (37.1%)	139763 (68.8%)	96993 (47.7%)
2007	240162	83282 (34.6%)	161147 (67.0%)	110065 (45.8%)
2008	269532	84170 (31.2%)	178676 (66.2%)	117139 (43.4%)

TABLE 1: FIB Aggregation with SMALTA one-shot and comparison with Level 1 and Level 2 aggregation.

4.2. Update Processing in SMALTA

A single BGP update may change the primary route for a single prefix. This may, in turn, change the next hop for multiple prefixes (or for no prefix) in the aggregated table. Table 2 shows the number of prefixes in the aggregated table whose next hop is modified (either changed or removed) as a result of a single change in the primary route of a prefix. These modified entries are then added or deleted from the aggregated FIB through RIB to FIB download. As indicated in Table 2, the average number of such modified entries remains small over various years.

It may be noted that when aggregation is not used, a change in the primary route of a prefix is translated to precisely one RIB to FIB download. Thus, on average, the number of RIB to FIB downloads in SMALTA are comparable to the number of RIB to FIB downloads when no aggregation is used.

The table also shows the maximum number of entries modified in the aggregated table as a result of a single change in primary route for a prefix. This worst case usually occurs when a BGP update arrives for very short prefixes (/8 in most cases). This is expected because shorter prefixes represent a larger IP address space, with the possibility that a larger number of more specific prefixes will fall within that space.

Year	No. of updates	Avg #entries modified	Max. #entries modified	Prefix for Max. modified
2004	86904	1.0494	300	/8
2005	140920	1.0962	157	/8
2006	121651	0.9660	499	/8
2007	246634	1.3141	376	/11
2008	240802	0.4031	64	/8

TABLE 2: Number of entries modified when directly incorporating updates into aggregated table

4.3. Updates and Aggregated Table Size

SMALTA one-shot aggregation is performed only infrequently. In between two episodes of one-shot aggregation, any BGP updates are incorporated using the SMALTA update algorithm. When incorporating the BGP updates incrementally, it is important to track how the size of the aggregated table changes. Ideally, we would want to keep the FIB in its fully aggregated state (what would have resulted from one-shot aggregation).

We used the routeviews tables from 2004 to 2008, one table from each year, and applied a corresponding update trace to the aggregated table using SMALTA update algorithm. The resulting number of entries (in the aggregated table), as a percentage of the number of un-aggregated entries, are shown in Table 3.

Table 3 also indicates the number of entries in the aggregated table if SMALTA one-shot algorithm is applied after incorporating the updates in the original un-aggregated table.

Year	No. of updates	#entries in aggregated table (SMALTA update) (% of unaggregated)	#entries in aggregated table (SMALTA one-shot) (% of unaggregated)
2004	86904	35.50	34.38
2005	140920	36.69	33.08
2006	121651	45.51	39.9
2007	246634	60.18	40.25
2008	240802	32.99	31.72

TABLE 3: Number of Entries in the aggregated table after applying updates using (i) SMALTA one-shot, and (ii) SMALTA update

5. IANA Considerations

There are no IANA considerations.

6. Security Considerations

Because SMALTA does not change the external behavior of a router, there are no security considerations.

7. References

7.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

7.2. Informative References

[I-D.zhang-fibaggregation]
Zhang, B., Wang, L., Zhao, X., Liu, Y., and L. Zhang, "FIB Aggregation", draft-zhang-fibaggregation-02 (work in progress), October 2009.

[Paper.Draves-ORTC]
Draves, R. and J. Doe, "Optimal Routing Table Constructor", INFOCOM 1999, July 1999.

[TR.Uzmi-SMALTA]
Uzmi, Z., Jawad, S., Tariq, A., and P. Francis, "Prefix Aggregation with SMALTA", Technical Report URL http://www.mpi-sws.org/~zartash/TR-MPI-SMALTA.pdf, July 2010.

Authors' Addresses

Zartash Uzmi
Lahore University of Management Sciences
LUMS, D.H.A.
Lahore 54792
Pakistan

Phone: +92 42 35608202
Email: zartash@gmail.com

Ahsan Tariq
Lahore University of Management Sciences
LUMS, D.H.A.
Lahore 54792
Pakistan

Phone: +92 42 35608000
Email: ahsan.tariq11@gmail.com

Paul Francis
Max Planck Institute for Software Systems
Gottlieb-Daimler-Strasse
Kaiserslautern 67633
Germany

Phone: +49 631 930 39600
Email: francis@mpi-sws.org

Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: March 19, 2012

I. Varlashkin
Easynet Global Services
R. Raszuk
NTT MCL Inc.
September 16, 2011

Carrying next-hop cost information in BGP
draft-varlashkin-bgp-nh-cost-02

Abstract

This document describes new BGP SAFI to exchange cost information to next-hops for the purpose of calculating best path from a peer perspective rather than local BGP speaker own perspective.

Table of Contents

1. Motivation

In certain situation route-reflector clients may not get optimum path to certain destinations. ADDPATH solves this problem by letting route-reflector to advertise multiple paths for given prefix. If number of advertised paths sufficiently big, route-reflector clients can choose same route as they would in case of full-mesh. This approach however places additional burden on the control plane. Solutions proposed by [BGP-ORR] use different approach - instead of calculating best path from local speaker own perspective the calculations are done using cost from the client to the next-hops. Although they eliminate need for transmitting redundant routing information between peers, there are scenarios where cost to the next-hop cannot be obtained accurately using this methods. For example, if next-hop information itself has been learned via BGP then simple SPF run on link-state database won't be sufficient to obtain cost information. To address such scenarios this document proposes a solution where cost information to the next-hops is carried within BGP itself using dedicated SAFI.

2. NEXT-HOP INFORMATION BASE

To facilitate further description of the proposed solution we introduce new table for all known next hops and costs to it from various routers on the network.

Next-Hop Information Base (NHIB) stores cost to reach next-hop from arbitrary router on the network. This information is essential for choosing best path from a peer perspective rather than BGP-speaker own perspective. In canonical form NHIB entry is triplet (router, next-hop, cost), however this specification does not impose any restriction on how BGP implementations store that information internally. The cost in NHIB is does not have to be an IGP cost, but all costs in NHIB MUST be comparable with each other.

NHIB can be populated from various sources both static and dynamic. This document focuses on populating NHIB using BGP. However it is possible that protocols other than BGP could be also used to populate NHIB.

3. BGP BEST PATH SELECTION MODIFICATION

This section applies regardless of method used to populate NHIB.

When BGP speaker conforming to this specification selects routes to be advertised to a peer it SHOULD use cost information from NHIB

rather than its own IGP cost to the next-hop after step (d) of 9.1.2.2 in [RFC4271].

4. USING BGP TO POPULATE NHIB

This section describes extension to base BGP specification that allows BGP to be used for exchanging next-hop information between BGP speakers via new SAFI in order to populate NHIB. Although next-hops costs are exchanged via dedicated SAFI, this information is vital to best path selection process for other AFI/SAFI (e.g. IPv4 and IPv6 unicast). It's therefore recommended that next-hop cost information is exchanged before other AFI/SAFI.

4.1. NEXT-HOP SAFI

This document introduces Next-Hop SAFI (NH SAFI) with value to be assigned by IANA and purpose of exchanging information about cost to next-hops.

4.2. CAPABILITY ADVERTISEMENT

A BGP speaker willing to exchange next-hop information MUST advertise this in the OPEN message using BGP Capability Code 1 (Multiprotocol Extensions, see [RFC4760]) setting AFI appropriately to indicate IPv4 or IPv6 and SAFI to the value assigned by IANA for NH SAFI. Note that if BGP speaker whishes to exchange cost information for both IPv4 and IPv6, then it MUST advertise two capabilities: one NH SAFI for IPv4 and one NH SAFI for IPv6.

4.3. INFORMATION ENCODING

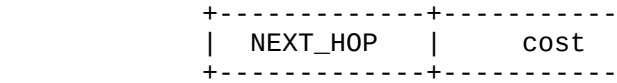
To request cost to a next-hop from peer or to inform peer about cost to a next-hop BGP attribute 14 is used as follow:

1. AFI is set to indicate IPv4 or IPv6 (whichever is appropriate)

2. SAFI is set to NH SAFI

3. Network Address of Next-Hop field is zeroed out

4. NLRI field is encoded as shown in the next figure



Where cost is 32-bit unsigned integer (value described below), and

NEXT_HOP is AFI-specific address of the next-hop cost to which is being communicated or requested. Size of NEXT_HOP field is inferred from total length of attribute 14.

To request cost to arbitrary next-hop from a peer, BGP speaker sets cost field to zero.

To inform peer about cost to a next-hop BGP speaker sets cost to actual cost value.

To inform peer that a next-hop is not reachable the cost is set to all-ones (0xFFFFFFFF).

4.4. SESSION ESTABLISHMENT

BGP speakers willing to exchange next-hop information SHOULD NOT establish more then one session for given AFI and NH SAFI, even using different transport addresses. This can be ensured for example by checking peer's Router Id.

4.5. INFORMATION EXCHANGE

Typically NH SAFI sessions will be established between route-reflectors and its internal peers (both clients and non-clients). As soon as the NH SAFI session is ESTABLISHED requests for next-hop cost and information information about next-hop costs MAY be sent independently. That is, route-reflector MAY send multiple requests without waiting for response, and its peers MAY send cost information before or after receiving such request. On the other hand, Router Reflectors SHOULD request cost information from their internal peers as soon as possible (due to reasons stated in section "BGP best path selection modification"). BGP speaker does not need to track outstanding requests to the peer.

When a BGP speaker receives request for cost information it MUST reply with actual cost (not necessarily IGP cost, but whatever has been chosen to be carried in NH SAFI) to given next-hop or with cost set to all-ones indicating that next-hop is unreachable.

Note that BGP speaker MUST use longest match rather than exact match for the next-hop.

When a BGP speaker detects change in cost to previously advertised next-hop with delta equal or exceeding configured advertisement threshold, it SHOULD inform peer by advertising new cost or 0xFFFFFFFF.

When a BGP speaker discovers new next-hop among candidate routes it

SHOULD request cost information from the peer.

4.6. TERMINATION OF NH SAFI SESSION

When BGP speaker terminates (for whatever reason) NH SAFI session with a peer, it SHOULD remove all cost information received from that peer unless instructed by configuration to do otherwise.

4.7. GRACEFUL RESTART AND ROUTE REFRESH

NH SAFI sessions could use graceful restart and route refresh mechanisms in the same way as it’s used for IPv4 and IPv6 unicast.

5. Security considerations

No new security issues are introduced to the BGP protocol by this specification.

6. IANA Considerations

IANA is requested to allocate value for Next-Hop Subsequent Address Family Identifier.

7. References

7.1. Normative References

[RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.

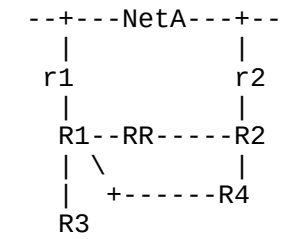
7.2. Informative References

[I-D.raszuk-bgp-optimal-route-reflection] Raszuk, R., Cassar, C., Aman, E., and B. Decraene, "BGP Optimal Route Reflection (BGP-ORR)", draft-raszuk-bgp-optimal-route-reflection-01 (work in progress), March 2011.

[RFC2918] Chen, E., "Route Refresh Capability for BGP-4", RFC 2918, September 2000.

Appendix A. USAGE SCENARIOS

A.1. Trivial case



In this scenario r1 and r3 along with NetA are part of AS1; and R1-R4 along with RR are in AS2.

If RR implements non-optimized route-reflection, then it will choose path to NetA via R1 and advertise it to both R3 and R4. Such choice is good from R3 perspective, but it results in suboptimal traffic flow from R4 to NetA.

Using NH SAFI the route-reflector will learn that cost from R4 to R1 is 8 whereas to R2 it's only 1. RR will announce NetA to R4 with next-hop set to R2, while its announce to R3 will still have R1 as next-hop. Both R3 and R4 now will send traffic to NetA via closest exit, achieving same behaviour as if full iBGP mesh would have been configured.

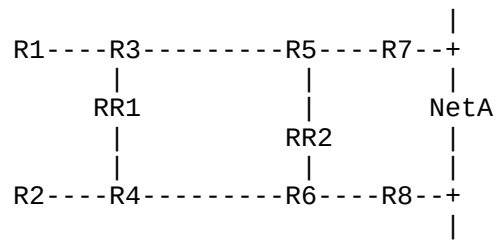
A.2. Non-IGP based cost

When it's desirable to direct traffic over an exit other than the one with smallest IGP cost, NH SAFI can be used to convey cost which is not based on IGP. For example, network operator may arrange exit points in order of administrative preference and configure routers to send this instead of IGP cost. Route reflector then will then calculate best path based on administrative preference rather than IGP metrics.

Network operators should excercise care to ensure that all routers up to and including exit point do not devert packets on to a different path, otherwise routing loops may occur. One way to achieve this is to have consistent administrative preference among all routers. Another option is to use a tunneling mechanism (e.g. MPLS-TE tunnel) between source and the exit point, provided that the router serving as exit point will send packets out of the network rather than diverting them to another exit point.

A.3. Multiple route-reflectors

This example demonstrates that NH SAFI peerings are necessary only between routers that already exchange other AFI/SAFI.



In the above network the routers R1-R4 are clients of RR1, and R5-R8 are clients of RR2. RR1 and RR2 also peer with each other and use ADDPATH.

RR2 learns about NetA from R7 and R8. Since it sends not just best-path but all prefixes to RR1, there is no need for RR2 to learn cost information from R1 and R2 towards R7 and R8. On the other hand RR1 does exchange NH SAFI information with R1 and R2 so that each of them can receive routes, which are best from their perspective.

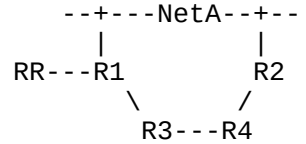
As addition to ADDPATH a mechanism could be devised that would allow RR2 to learn how many alternative routes does it need to send to RR1. For example, if NetA would also be connected to R9 (not shown) but all clients of RR1 prefer R7 as exit point and R9 as next-best, then there is no need for RR2 to send NetA routes with next-hop R8 to RR1.

Discussion: authors would like to solicit discussion whether there is sufficient interest in such mechanism.

A.4. Inter-AS MPLS VPN

Previous example could be transposed to Inter-AS MPLS VPN Option C scenario. In this case route reflectors RR1 and RR2 can be from different autonomous system. Essentially the behaviour of routers remains as already described.

A.5. Corner case



In the above network cost from R3 to R1 is 10, all other costs are 1. If RR advertises NetA to R3 based on cost information received from R3, but uses its own cost when advertising NetA to R4, there will be a loop formed. This is the reason why section "BGP best path selection modification" requires RR to have next-hop cost information for every next-hop and every peer.

Note that the problem is the same as if RR would not use extensions described in this document and R3 would peer directly with R1 and R2, while R4 would peer only with RR.

Authors' Addresses

Ilya Varlashkin
Easynet Global Services

Email: ilya.varlashkin@easynet.com

Robert Raszuk
NTT MCL Inc.
101 S Ellsworth Avenue Suite 350
San Mateo, CA 94401
US

Email: robert@raszuk.net

Network Working Group
Internet-Draft
Intended status: BCP
Expires: August 22, 2011

L. Vegoda
ICANN

February 18, 2011

Time to Remove Filters for Previously Unallocated IPv4 /8s
draft-vegoda-no-more-unallocated-slash8s-01

Abstract

It has been common for network administrators to filter IP traffic coming from unallocated IPv4 address space. Now that there are no longer any unallocated IPv4 /8s, this practise is more complicated, fragile and expensive. Network administrators are advised to remove filters based on the registration status of the address space.

This document explains why any remaining filters for unallocated IPv4 /8s should now be removed on border routers and documents those IPv4 unicast prefixes that should not be routed across the public Internet.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on August 22, 2011.

Copyright Notice

Table of Contents

1. Introduction

It has been common for network administrators to filter IP traffic coming from unallocated IPv4 address space. Now that there are no longer any unallocated IPv4 /8s, this practise is more complicated, fragile and expensive. Network administrators are advised to remove filters based on the registration status of the address space.

This document explains why any remaining filters for unallocated IPv4 /8s should now be removed on border routers and documents those IPv4 unicast prefixes that should not be routed across the public Internet.

2. Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14, RFC 2119 [RFC2119].

3. Traffic Filtering Options

3.1. No Longer Filtering Based on Address Registration Status

Network administrators who implemented filters for unallocated IPv4 /8s did so in the knowledge that those /8s were not a legitimate source of traffic on the Internet and that there was a small number of filters to implement. Now that there are no longer any unallocated unicast IPv4 /8s, there will be legitimate Internet traffic coming from all unicast /8s that are not reserved for special purposes in an RFC.

Removing ingress filters based on the registration status of the IPv4 address is a simple approach that will avoid blocking legitimate Internet traffic.

3.2. Continuing to Filter Traffic from Unallocated IPv4 Space

Some network administrators might want to continue filtering unallocated IPv4 addresses managed by the Regional Internet Registries (RIRs). This requires significantly more granular ingress filters and the highly dynamic nature of the RIRs' address pools means that filters need to be updated on a daily basis to avoid blocking legitimate incoming traffic.

4. Prefixes That Should Not be Routed Across the Internet

Network operators who only wish to filter traffic originating from addresses that should never be routed across the Internet can deploy a set of ingress filters designed to block traffic from address blocks reserved for special purposes. These are:

- 0.0.0.0/8 (Local identification) [RFC1122];
- 10.0.0.0/8 (Private use) [RFC1918];
- 127.0.0.0/8 (Loopback) [RFC1122];
- 169.254.0.0/16 (Link local) [RFC3927];
- 172.16.0.0/12 (Private use) [RFC1918];
- 192.0.2.0/24 (TEST-NET-1) [RFC5737];
- 192.168.0.0/16 (Private use) [RFC1918];
- 198.18.0.0/15 (Benchmark testing) [RFC2544];
- 198.51.100.0/24 (TEST-NET-2) [RFC5737];
- 203.0.113.0/24 (TEST-NET-3) [RFC5737];
- 224.0.0.0/4 (Multicast) [RFC5771]; and
- 240.0.0.0/4 (Future use) [RFC1112].

A full set of special use IPv4 addresses can be found in [RFC5735]. It includes prefixes that are intended for Internet use.

5. Security Considerations

The cessation of filters based on unallocated IPv4 /8 allocations is an evolutionary step towards reasonable security filters. While these filters are no longer necessary, and in fact harmful, this does not obviate the need to continue other security solutions. These other solutions are as necessary today as they ever were.

6. IANA Considerations

This document makes no request of IANA.

7. Normative References

[RFC1112] Deering, S., "Host extensions for IP multicasting", STD 5, RFC 1112, August 1989.

[RFC1122] Braden, R., "Requirements for Internet Hosts - Communication Layers", STD 3, RFC 1122, October 1989.

[RFC1918] Rekhter, Y., Moskowitz, R., Karrenberg, D., Groot, G., and E. Lear, "Address Allocation for Private Internets", BCP 5, RFC 1918, February 1996.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC2544] Bradner, S. and J. McQuaid, "Benchmarking Methodology for Network Interconnect Devices", RFC 2544, March 1999.

[RFC3927] Cheshire, S., Aboba, B., and E. Guttman, "Dynamic Configuration of IPv4 Link-Local Addresses", RFC 3927, May 2005.

[RFC5735] Cotton, M. and L. Vegoda, "Special Use IPv4 Addresses", BCP 153, RFC 5735, January 2010.

[RFC5737] Arkko, J., Cotton, M., and L. Vegoda, "IPv4 Address Blocks Reserved for Documentation", RFC 5737, January 2010.

[RFC5771] Cotton, M., Vegoda, L., and D. Meyer, "IANA Guidelines for IPv4 Multicast Address Assignments", BCP 51, RFC 5771, March 2010.

Appendix A. Acknowledgments

Thanks are owed to Kim Davies, Terry Manderson, Dave Piscitello and Joe Abley for helpful advice on how to focus this document. Thanks also go to Andy Davidson, Philip Smith and Rob Thomas for early reviews and suggestions for improvements to the text and Carlos Pignataro for his support and comments.

Author's Address

Leo Vegoda
Internet Corporation for Assigned Names and Numbers
4676 Admiralty Way, Suite 330
Marina del Rey, CA 90292
United States of America

Phone: +1-310-823-9358
Email: leo.vegoda@icann.org
URI: http://www.iana.org/