Network Working Group
Internet Draft
Updates: 4271 (if approved)
Intended Status: Standards Track
Expiration Date: March 9, 2011

E. Chen
P. Mohapatra
K. Patel
Cisco Systems
September 8, 2010

Revised Error Handling for BGP Updates from External Neighbors
draft-chen-ebgp-error-handling-00.txt

Status of this Memo

Copyright Notice

Abstract

In this document we partially revise the error handling of an UPDATE message from an external BGP neighbor. The essence of the revision is to avoid resetting an external BGP session by using the "treat-as-withdraw" approach when the whole NLRI field of a malformed UPDATE message can be parsed.

1. Introduction

The base BGP specification [RFC4271] requires that a BGP session be reset when an UPDATE message containing a malformed attribute is received. This behavior is undesirable in the case of optional transitive attributes as has been discussed and revised in [OPT-TRANS].

However, there are other situations where the behavior is also undesirable, but are outside the scope of [OPT-TRANS]. For example, there have been a few occurrences in the field where the AS-PATH attribute is malformed for a small number of routes. Resetting the BGP session would impact all the other valid routes in these cases.

Our goal is to minimize the scope of the network that is affected by a malformed UPDATE message, and also to limit the impact to only the routes involved. The constrain is that the protocol correctness must not be violated.

In this document we partially revise the error handling of an UPDATE message from an external BGP neighbor. The essence of the revision is to avoid resetting an external BGP session by using the "treat-as-withdraw" approach specified in [OPT-TRANS] when the whole NLRI field of a malformed UPDATE message can be parsed.

1.1. Specification of Requirements

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].

2. Revision to Base Specification

The revised error handling specified in this section is applicable only for processing an UPDATE message from an external BGP neighbor.

The error handling of the following case described in Section 6.3 of [RFC4271] remains unchanged:

> If the Withdrawn Routes Length or Total Attribute Length is too large (i.e., if Withdrawn Routes Length + Total Attribute Length + 23 exceeds the message Length), then the Error Subcode MUST be set to Malformed Attribute List.

The error handling of all other cases described in Section 6.3 of [RFC4271] that specify a sesion reset is conditionally revised as follows.

If a path attribute in an UPDATE message from an external BGP neighbor is determined to be malformed, the message containing that attribute SHOULD be treated as though all contained routes had been withdrawn ("treat-as-withdraw") when the whole NLRI field in the message can be parsed.

One exception is that the "attribute discard" approach [OPT-TRANS] SHOULD be used to handle a malformed optional transitive attribute for which the "attribute discard" approach is specified.

A BGP speaker MUST provide debugging facilities to permit issues caused by malformed UPDATE messages to be diagnosed. At a minimum, such facilities SHOULD include logging an error when such an attribute is detected. The malformed UPDATE message SHOULD be analyzed, and the root cause SHOULD be investigated.

3. Parsing of NLRI Fields

As described in [OPT-TRANS], we observe that in order to use the "treat-as-withdraw" approach for a malformed UPDATE, the NLRI field and/or MP_REACH and MP_UNREACH [RFC4760] attributes need to be successfully parsed. If this were not possible, the UPDATE would necessarily be malformed in some other way beyond the scope of this document and therefore, the procedures of [RFC4271] would continue to apply.

To facilitate the determination of the NLRI field in an UPDATE with malformed attributes, we strongly RECOMMEND that the MP_REACH or MP_UNREACH attribute (if present) be encoded as the very first path attribute in an UPDATE.

Traditionally the NLRIs for the IPv4 unicast address family are carried immediately following all the attributes in an UPDATE [RFC4271]. When such an UPDATE is received, we observe that the NLRI field can be determined using the "Message Length" and the "Total Attribute Length" (when they are consistent) carried in the message instead of relying on the length of individual attributes in the message.

Furthermore, it is observed that the NLRIs for the IPv4 unicast address family can also be carried in the MP_REACH attribute of an UPDATE when the IPV4 unicast address family capability is shared (i.e., both advertised and received) over a BGP session. For the same sake of better debugging and fault handling, we also RECOMMEND that the MP_REACH attribute be used and be placed as the very first path attribute in an UPDATE in this case.

4. Discussion

As discussed in [OPT-TRANS], the approach of "treat-as-withdraw" is not always safe to use. In the case of internal BGP sessions, the resolution of recursive nexthops can result in forwarding loops and blakholes when the BGP speakers inside a network have inconsistent routing information.

Depending on the network topology, the routing table, routes involved, and whether "tunnels" are used inside a network, the approach of "treat-as-withdraw" may work for internal BGP sessions only in some specific cases. Thus it may be deployed for internal BGP sessions only as a temporary measure to stop continuous session flaps due to malformed UPDATE messages. Such deployment must be carefully evaluated on a case-by-case basis.

5. IANA Considerations

This document makes no request of IANA.

6. Security Considerations

TBD

7. Acknowledgments

We would like to thank Robert Raszuk, Naiming Shen and Tony Li for their review and discussions.

8. References

8.1. Normative References

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

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

8.2. Informative References

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

[OPT-TRANS] Scudder, J. and E. Chen, "Error Handling for Optional Transitive BGP Attributes", Work in Progress, March 2010.

9. Authors' Addresses

Enke Chen
Cisco Systems, Inc.
170 W. Tasman Dr.
San Jose, CA 95134

EMail: enkechen@cisco.com

Pradosh Mohapatra
Cisco Systems, Inc.
170 W. Tasman Dr.
San Jose, CA 95134

EMail: pmohapat@cisco.com

Keyur Patel
Cisco Systems, Inc.
170 W. Tasman Dr.
San Jose, CA 95134

EMail: keyupate@cisco.com

Network Working Group
Internet-Draft
Intended status: Informational
Expires: April 28, 2011

Pierre Francois
Universite catholique de Louvain
Bruno Decraene
France Telecom
Cristel Pelsser
Internet Initiative Japan
Keyur Patel
Clarence Filsfils
Cisco Systems
October 25, 2010

Graceful BGP session shutdown
draft-ietf-grow-bgp-gshut-02

Abstract

This draft describes operational procedures aimed at reducing the amount of traffic lost during planned maintenances of routers, involving the shutdown of BGP peering sessions.

Table of Contents

1. Introduction

Routing changes in BGP can be caused by planned, manual, maintenance operations. This document discusses operational procedures to be applied in order to reduce or eliminate losses of packets during the maintenance. These losses come from the transient lack of reachability during the BGP convergence following the shutdown of an eBGP peering session between two Autonomous System Border Routers (ASBR).

This document presents procedures for the cases where the forwarding plane is impacted by the maintenance, hence when the use of Graceful Restart does not apply.

The procedures described in this document can be applied to reduce or avoid packet loss for outbound and inbound traffic flows initially forwarded along the peering link to be shut down. These procedures allow routers to keep using old paths until alternate ones are learned, ensuring that routers always have a valid route available during the convergence process.

The goal of the document is to meet the requirements described in [REQS] at best, without changing the BGP protocol or BGP implementations.

Still, it explains why reserving a community value for the purpose of BGP session graceful shutdown would reduce the management overhead bound with the solution. It would also allow vendors to provide an automatic graceful shutdown mechanism that does not require any router reconfiguration at maintenance time.

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 RFC 2119 [RFC2119].

2. Terminology

g-shut initiator : a router on which the session shutdown is performed for the maintenance.

g-shut neighbor : a router that peers with the g-shut initiator via (one of) the session(s) to be shut down.

Note that for the link-up case, we will refer to these nodes as g-no-shut initiator, and g-no-shut neighbor.

Initiator AS : the Autonomous System of the g-shut initiator.

Neighbor AS : the Autonomous System of the g-shut neighbor.

Affected path / Nominal / pre-convergence path : a BGP path via the peering link(s) undergoing the maintenance. This path will no longer exist after the shutdown.

Affected prefix : a prefix initially reached via an affected path.

Affected router : a router having an affected prefix.

Backup / alternate / post-convergence path : a path towards an affected prefix that will be selected as the best path by an affected router, when the link is shut down and the BGP convergence is completed.

Transient alternate path : a path towards an affected prefix that may be transiently selected as best by an affected router during the convergence process but that is not a post-convergence path.

Loss of Connectivity (LoC) : the state when a router has no path towards an affected prefix.

3. Packet loss upon manual eBGP session shutdown

Packets can be lost during a manual shutdown of an eBGP session for two reasons.

First, routers involved in the convergence process can transiently lack of paths towards an affected prefix, and drop traffic destined to this prefix. This is because alternate paths can be hidden by nodes of an AS. This happens when the paths are not selected as best by the ASBR that receive them on an eBGP session, or by Route Reflectors that do not propagate them further in the iBGP topology because they do not select them as best.

Second, within the AS, the FIB of routers can be transiently inconsistent during the BGP convergence and packets towards affected prefixes can loop and be dropped. Note that these loops only happen when ASBR-to-ASBR encapsulation is not used within the AS.

This document only addresses the first reason.

4. Practices to avoid packet losses

This section describes means for an ISP to reduce the transient loss of packets upon a manual shutdown of a BGP session.

4.1. Improving availability of alternate paths

All solutions that increase the availability of alternate BGP paths at routers performing packet lookups in BGP tables [BestExternal] [AddPath] help in reducing the LoC bound with manual shutdown of eBGP sessions.

One of such solutions increasing diversity in such a way that, at any single step of the convergence process following the eBGP session shutdown, a BGP router does not receive a message withdrawing the only path it currently knows for a given NLRI, allows for a simplified g-shut procedure.

Increasing diversity with [AddPath] might lead to the respect of this property, depending on the path propagation decision process that add-path compliant routers would use.

Using advertise-best-external [BestExternal] on ASBRs and RRs helps in avoiding lack of alternate paths in route reflectors upon a convergence. Hence it reduces the LoC duration for the outbound traffic of the ISP upon an eBGP Session shutdown by reducing the iBGP path hunting.

Note that the LoC for the inbound traffic of the maintained router, induced by a lack of alternate path propagation within the iBGP topology of a neighboring AS is not under the control of the operator performing the maintenance. The procedure described in Section 4.2.2 should thus be applied upon the maintenance, even if the procedure described in Section 4.2.1 is not applied.

4.2. Graceful shutdown procedures for eBGP sessions

This section aims at describing a procedure to be applied to reduce the LoC with readily available BGP features, and without assuming a particular iBGP design in the Initiator and Neighbor ASes.

4.2.1. Outbound traffic

This section discusses a mean to render the affected paths less desirable by the BGP decision process of affected routers, still allowing these to be used during the convergence, while alternate paths are propagated to the affected routers.

A decrease of the local-pref value of the affected paths can be issued in order to render the affected paths less preferable, at the highest possible level of the BGP Decision Process.

This operation can be performed by reconfiguring the out-filters

associated with the iBGP sessions established by the g-shut initiator.

The modification of the filters MUST supplant any other rule affecting the local-pref value of the old paths.

Compared to using an in-filter of the eBGP session to be shut down, the modification of the out-filters will not let the g-shut initiator switch to another path, as the input to the BGP decision process of that router does not change. As a consequence, the g-shut initiator will not modify the state of its dataplane, and will not withdraw the affected paths over its iBGP sessions when it receives alternate paths. It will however modify the local-pref of the affected paths so that upstream routers will switch to alternate ones.

When the actual shutdown of the session is performed, the g-shut initiator will itself switch to the alternate paths.

In cases some BGP speakers in the AS override the local-pref attribute of paths received over iBGP sessions, the procedure described above will not work. In such cases, the recommended procedure is to tag the paths sent over the iBGP sessions of the g-shut initiator with an AS specific community. This AS specific community should lead to the setting of the lowest local-pref value. To be effective, the configuration related to this community MUST supplant or be applied after the already configured local-pref overriding.

An operator may decide to follow a simplified procedure and directly apply an in-filter reducing the local preference of the paths received over the eBGP session being brought down. While this procedure will be effective in many cases, corner cases as described in Appendix A.1 may happen, which may lead to some LoC for some affected destinations. The use of this simplified procedure does not lead to LoC when used in conjunction with [BestExternal].

4.2.2. Inbound traffic

The solution described for the outbound traffic can be applied at the neighbor AS. This can be done either "manually" or by using a community value dedicated to this task.

4.2.2.1. Phone call

The operator performing the maintenance of the eBGP session can contact the operator at the other side of the peering link, and let him apply the procedure described above for its own outbound traffic.

4.2.2.2. Community tagging

A community value (referred to as GSHUT community in this document) can be agreed upon by neighboring ASes and used to trigger the g-shut behavior at the g-shut neighbor.

4.2.2.2.1. Pre-Configuration

A g-shut neighbor is pre-configured to set a low local-pref value for the paths received over eBGP sessions which are tagged with the GSHUT community.

This rule must supplant any other rule affecting the local-pref value of the paths.

This local-pref reconfiguration SHOULD be performed at the out-filters of the iBGP sessions of the g-shut neighbor. That is, the g-shut neighbor does not take into account this low local-pref in its own BGP best path selection. As described in Section 4.2.1 this approach avoids sending withdraw messages that can lead to LoC in some cases.

4.2.2.2.2. Operational action upon maintenance

Upon the manual shutdown, the output filter associated with the maintained eBGP session will be modified on the g-shut initiator so as to tag all the paths advertised over the session with the GSHUT community.

4.2.2.2.3. Transitivity of the community

If the GSHUT community is an extended community, it SHOULD be chosen non-transitive.

If a regular community is used, this community SHOULD be removed from the path when the path is propagated over eBGP sessions.

Not propagating the community further in the Internet reduces the amount of BGP churn and avoids rerouting in distant ASes that would also recognize this community value. In other words, from a routing stability perspective, it helps concealing the convergence at the maintenance location. From a policy perspective, it prevents malignant ASes from using the community over paths propagated through intermediate ASes that do not support the feature, in order to perform inbound traffic engineering at the first AS recognizing the community.

ASes which support the g-shut procedure SHOULD remove the community

value(s) that they use for g-shut from the paths received from neighboring ASes that do not support the procedure or to whom the service is not provided.

There are cases where an interdomain exploration is to be performed to recover the reachability, e.g., in the case of a shutdown in confederations where the alternate paths will be found in another AS of the confederation. In such scenarios, the community value SHOULD be allowed to transit through the confederation but SHOULD be removed from the paths advertised outside of the confederation.

When the local-pref value of a path is conserved upon its propagation from one AS of the confederation to the other, there is no need to have the GSHUT community be propagated throughout that confederation.

4.2.2.2.4. Easing the configuration for G-SHUT

From a configuration burden viewpoint, it is much easier to use a single dedicated value for the GSHUT community.

First, on the g-shut initiator, an operator would have a single configuration rule to be applied at the maintenance time, which would not depend on the identity of its peer. This would make the maintenance operations less error prone.

Second, on the g-shut neighbor, a simple filter related to g-shut can be applied to all iBGP sessions. Additionnaly, this filter does not need to be updated each time neighboring ASes are added or removed.

The FCFS community value 0xFFFF0000 has been reserved for this purpose [BGPWKC].

4.2.3. Summary of operations

This section summarizes the configurations and actions to be performed to support the g-shut procedure for eBGP peering links.

4.2.3.1. Pre-configuration

On each ASBR supporting the g-shut procedure, set-up an out-filter applied on all iBGP sessions of the ASBR, that :

- sets the local-pref of the paths tagged with the g-shut community to a low value
- removes the g-shut community from the paths.
- optionally, adds an AS specific g-shut community on these paths to indicate that these are to be withdrawn soon. If some ingress ASBRs reset the local preference attribute, this AS specific g-shut community will be used to override other local

preference changes.

4.2.3.2. Operations at maintenance time

On the g-shut initiator :

- Apply an out-filter on the maintained eBGP session to tag the paths propagated over the session with the g-shut community.
- Apply an in-filter on the maintained eBGP session to tag the paths received over the session with the g-shut community.
- Wait for convergence to happen.
- Perform a BGP session shutdown.

4.2.4. BGP implementation support for G-Shut

A BGP router implementation MAY provide features aimed at automating the application of the graceful shutdown procedures described above.

Upon a session shutdown specified as to be graceful by the operator, a BGP implementation supporting a g-shut feature would

1. Update all the paths propagated over the corresponding eBGP session, tagging the GSHUT community to them. Any subsequent update sent to the session being gracefully shut down would be tagged with the GSHUT community.
2. Lower the local preference value of the paths received over the eBGP session being shut down, upon their propagation over iBGP sessions. Optionally, also tag these paths with an AS specific g-shut community. Note that alternatively, the local preference of the paths received over the eBGP session can be lowered on the g-shut initiator itself, instead of only when propagating over its iBGP sessions. This simplified behavior can lead to some LoC, as described in Appendix A.1, if not used in conjunction with [BestExternal].
3. Optionally shut down the session after a configured time.
4. Prevent the GSHUT community from being inherited by a path that would aggregate some paths tagged with the GSHUT community. This behavior avoids the GSHUT procedure to be applied to the aggregate upon the graceful shutdown of one of its covered prefixes.

4.3. Graceful shutdown procedures for iBGP sessions

If the iBGP topology is viable after the maintenance of the session, i.e, if all BGP speakers of the AS have an iBGP signaling path for all prefixes advertised on this g-shut iBGP session, then the shutdown of an iBGP session does not lead to transient unreachability.

However, in the case of a shutdown of a router, a reconfiguration of the out-filters of the g-shut initiator MAY be performed to set a low local-pref value for the paths originated by the g-shut initiator (e.g, BGP aggregates redistributed from other protocols, including static routes).

This behavior is equivalent to the recommended behavior for paths "redistributed" from eBGP sessions to iBGP sessions in the case of the shutdown of an ASBR.

5. Forwarding modes and forwarding loops

If the AS applying the solution does not rely on encapsulation to forward packets from the Ingress Border Router to the Egress Border Router, then transient forwarding loops and consequent packet losses can occur during the convergence process, even if the procedure described above is applied. Hence if zero LoC is required, encapsulation is required between ASBRs of the AS.

Using the out-filter reconfiguration avoids the forwarding loops between the g-shut initiator and its directly connected upstream neighboring routers. Indeed, when this reconfiguration is applied, the g-shut initiator keeps using its own external path and lets the upstream routers converge to the alternate ones. During this phase, no forwarding loops can occur between the g-shut initiator and its upstream neighbors as the g-shut initiator keeps using the affected paths via its eBGP peering links. When all the upstream routers have switched to alternate paths, the transition performed by the g-shut initiator when the session is actually shut down, will be loopfree. Transient forwarding loops between other routers will not be avoided with this procedure.

6. Dealing with Internet policies

A side gain of the maintenance solution is that it can also reduce the churn implied by a shutdown of an eBGP session.

For this, it is recommended to apply the filters modifying the local-pref value of the paths to values strictly lower but as close as possible to the local-pref values of the post-convergence paths.

For example, if an eBGP link is shut down between a provider and one of its customers, and another link with this customer remains active, then the value of the local-pref of the old paths SHOULD be decreased to the smallest possible value of the ’customer’ local_pref range, minus 1. Thus, routers will not transiently switch to paths received

from shared-cost peers or providers, which could lead to the propagation of withdraw messages over eBGP sessions with shared-cost peers and providers.

Proceeding like this reduces both BGP churn and traffic shifting as routers will less likely switch to transient paths.

In the above example, it also prevents transient unreachabilities in the neighboring AS that are due to the sending of "abrupt" withdraw messages to shared-cost peers and providers.

7. Link Up cases

We identify two potential causes for transient packet losses upon an eBGP link up event. The first one is local to the g-no-shut initiator, the second one is due to the BGP convergence following the injection of new best paths within the iBGP topology.

7.1. Unreachability local to the ASBR

An ASBR that selects as best a path received over a newly brought up eBGP session may transiently drop traffic. This can typically happen when the nexthop attribute differs from the IP address of the eBGP peer, and the receiving ASBR has not yet resolved the MAC address associated with the IP address of that "third party" nexthop.

A BGP speaker implementation could avoid such losses by ensuring that "third party" nexthops are resolved before installing paths using these in the RIB.

If the link up event corresponds to an eBGP session that is being manually brought up, over an already up multi-access link, then the operator can ping third party nexthops that are expected to be used before actually bringing the session up, or ping directed broadcast the subnet IP address of the link. By proceeding like this, the MAC addresses associated with these third party nexthops will be resolved by the g-no-shut initiator.

7.2. iBGP convergence

Similar corner cases as described in Appendix A.1 for the link down case, can occur during an eBGP link up event.

A typical example for such transient unreachability for a given prefix is the following :

1. A Route Reflector, RR1, is initially advertising the current best path to the members of its iBGP RR full-mesh. It propagated that path within its RR full-mesh. Another route reflector of the full-mesh, RR2, knows only that path towards the prefix.
2. A third Route Reflector of the RR full-mesh, RR3 receives a new best path orginated by the "g-no-shut" initiator, being one of its RR clients. RR3 selects it as best, and propagates an UPDATE within its RR full-mesh, i.e., to RR1 and RR2.
3. RR1 receives that path, reruns its decision process, and picks this new path as best. As a result, RR1 withdraws its previously announced best-path on the iBGP sessions of its RR full-mesh.
4. If, for any reason, RR3 processes the withdraw generated in step 3, before processing the update generated in step 2, RR3 transiently suffers from unreachability for the affected prefix.

The use of [BestExternal] among the RR of the iBGP full-mesh can solve these corner cases by ensuring that within an AS, the advertisement of a new route is not translated into the withdraw of a former route.

Indeed, "best-external" ensures that an ASBR does not withdraw a previously advertised (eBGP) path when it receives an additional, preferred path over an iBGP session. Also, "best-intra-cluster" ensures that a RR does not withdraw a previously advertised (iBGP) path to its non clients (e.g. other RRs in a mesh of RR) when it receives a new, preferred path over an iBGP session.

8. IANA considerations

Applying the g-shut procedure is rendered much easier with a reserved g-shut community value. The community value 0xFFFF0000 has been reserved from the FCFS community pool for this purpose.

9. Security Considerations

By providing the g-shut service to a neighboring AS, an ISP provides means to this neighbor to lower the local-pref value assigned to the paths received from this neighbor.

The neighbor could abuse the technique and do inbound traffic engineering by declaring some prefixes as undergoing a maintenance so as to switch traffic to another peering link.

If this behavior is not tolerated by the ISP, it SHOULD monitor the

use of the g-shut community by this neighbor.

ASes which support the g-shut procedure SHOULD remove the community value(s) that they use for g-shut from the paths received from neighboring ASes that do not support the procedure or to whom the service is not provided. Doing so prevents malignant ASes from using the community through intermediate ASes that do not support the feature, in order to perform inbound traffic engineering.

10. Acknowledgments

The authors wish to thank Olivier Bonaventure and Pradosh Mohapatra for their useful comments on this work.

11. References

[AddPath] D. Walton, A. Retana, and E. Chen, "Advertisement of Multiple Paths in BGP", draft-walton-bgp-add-paths-06.txt (work in progress).

[BestExternal] Marques, P., Fernando, R., Chen, E., and P. Mohapatra, "Advertisement of the best-external route to IBGP", draft-ietf-idr-best-external-00.txt, May 2009.

[REQS] Decraene, B., Francois, P., Pelsser, C., Ahmad, Z., Armengol, A., and T. Takeda, "Requirements for the graceful shutdown of BGP sessions", draft-ietf-grow-bgp-graceful-shutdown-requirements-06.txt, October 2010.

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, February 2006.

[Clarification4360] Decraene, B., Vanbever, L., and P. Francois, "RFC 4360 Clarification Request", draft-decraene-idr-rfc4360-clarification-00, October 2009.

[BGPWKC] "http://www.iana.org/assignments/bgp-well-known-communities".

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

Appendix A. Alternative techniques with limited applicability

A few alternative techniques have been considered to provide g-shut capabilities but have been rejected due to their limited applicability. This section describe them for possible reference.

A.1. In-filter reconfiguration

An In-filter reconfiguration on the eBGP session undergoing the maintenance could be performed instead of out-filter reconfigurations on the iBGP sessions of the g-shut initiator.

Upon the application of the maintenance procedure, if the g-shut initiator has an alternate path in its Adj-Rib-In, it will switch to it directly.

If this new path was advertised by an eBGP neighbor of the g-shut initiator, the g-shut initiator will send a BGP Path Update message advertising the new path over its iBGP and eBGP sessions.

If this new path was received over an iBGP session, the g-shut initiator will select that path and withdraw the previously advertised path over its non-client iBGP sessions. There can be iBGP topologies where the iBGP peers of the g-shut initiator do not know an alternate path, and hence may drop traffic.

Also, applying an In-filter reconfiguration on the eBGP session undergoing the maintenance may lead to transient LoC, in full-mesh iBGP topologies if

a. An ASBR of the initiator AS, ASBR1 did not initially select its own external path as best, and

b. An ASBR of the initiator AS, ASBR2 advertises a new path along its iBGP sessions upon the reception of ASBR1’s update following the in-filter reconfiguration on the g-shut initiator, and

c. ASBR1 receives the update message, runs its Decision Process and hence withdraws its external path after having selected ASBR2’s path as best, and

d. An impacted router of the AS processes the withdraw of ASBR1 before processing the update from ASBR2.

Applying a reconfiguration of the out-filters prevents such transient unreachabilities.

Indeed, when the g-shut initiator propagates an update of the old path first, the withdraw from ASBR2 does not trigger unreachability in other nodes, as the old path is still available. Indeed, even though it receives alternate paths, the g-shut initiator keeps using its old path as best as the in-filter of the maintained eBGP session has not been modified yet.

Applying the out-filter reconfiguration also prevents packet loops between the g-shut initiator and its direct neighbors when encapsulation is not used between the ASBRs of the AS.

Note that applying this simplified procedure in conjunction with [BestExternal] does not lead to LoC.

A.2. Multi Exit Discriminator tweaking

The MED attribute of the paths to be avoided can be increased so as to force the routers in the neighboring AS to select other paths.

The solution only works if the alternate paths are as good as the initial ones with respect to the Local-Pref value and the AS Path Length value. In the other cases, increasing the MED value will not have an impact on the decision process of the routers in the neighboring AS.

A.3. IGP distance Poisoning

The distance to the BGP nexthop corresponding to the maintained session can be increased in the IGP so that the old paths will be less preferred during the application of the IGP distance tie-break rule. However, this solution only works for the paths whose alternates are as good as the old paths with respect to their Local-Pref value, their AS Path length, and their MED value.

Also, this poisoning cannot be applied when nexthop self is used as there is no nexthop specific to the maintained session to poison in the IGP.

Authors' Addresses

Pierre Francois
Universite catholique de Louvain
Place Ste Barbe, 2
Louvain-la-Neuve 1348
BE

Email: pierre.francois@uclouvain.be
URI: http://inl.info.ucl.ac.be/pfr

Bruno Decraene
France Telecom
38-40 rue du General Leclerc
92794 Issi Moulineaux cedex 9
FR

Email: bruno.decraene@orange-ftgroup.com

Cristel Pelsser
Internet Initiative Japan
Jinbocho Mitsui Bldg.
1-105 Kanda Jinbo-cho
Tokyo 101-0051
JP

Email: pelsser.cristel@iij.ad.jp

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

Email: keyupate@cisco.com

Clarence Filsfils
Cisco Systems
De kleetlaan 6a
Diegem 1831
BE

Email: cfilsfil@cisco.com

Global Routing Operations Working Group
Internet-Draft
Intended status: Informational
Expires: March 12, 2011

T. Manderson
ICANN
September 8, 2010

MRT BGP routing information export format with geo-location extensions
draft-ietf-grow-geomrt-00.txt

Abstract

This document extends the Border Gateway Protocol (BGP) MRT export format for routing information to include terrestrial coordinates.

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 March 12, 2011.

Copyright Notice

Table of Contents

1. 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].

2. Introduction

Research is underway that analyzes the network behavior of routing protocol transactions from routing information base snapshots in relation to geographical coordinates. Specifically the BGP routing protocol is the subject of study and the analysis has been significantly aided by the availability and extension of the "MRT format" [I-D.ietf-grow-mrt] originally defined in the MRT Programmer's Guide [MRT PROG GUIDE].

This memo documents an extension to the "MRT format" [I-D.ietf-grow-mrt] and introduces an additional definition of a MRT Type field and related Subtype fields.

3. Geo-location aware MRT Routing Information Type

The following additional Type is defined for the TABLE_DUMP_v2+GEO format, which extends the TABLE_DUMP_V2 type.

[TYPE NUMBER] TABLE_DUMP_v2+GEO

The TYPE NUMBER, an FCFS entry from the future IANA MRT registry is yet to be assigned.

4. TABLE_DUMP_v2+GEO Type

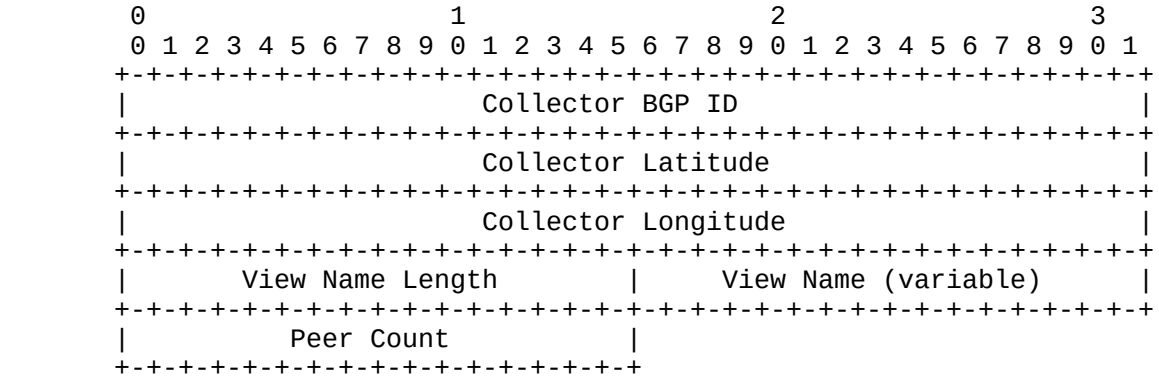
The TABLE_DUMP_v2+GEO Type updates the TABLE_DUMP_v2 Type to include Geo-location information in the form of WGS84 [WGS 84] formatted coordinates. The following subtypes as used with the TABLE_DUMP_V2 Type, are used in the TABLE_DUMP_v2+GEO Type and their formats have been augmented to include the WGS84 coordinates.

1 PEER_INDEX_TABLE
2 RIB_IPV4_UNICAST
3 RIB_IPV4_MULTICAST
4 RIB_IPV6_UNICAST
5 RIB_IPV6_MULTICAST
6 RIB_GENERIC

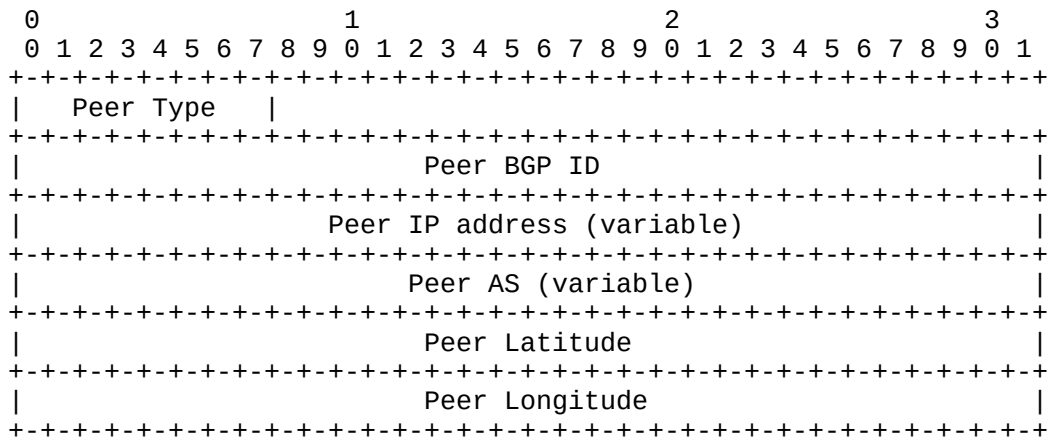
The extended PEER_INDEX_TABLE MRT record provides the BGP ID of the collector, the latitude and longitude in WGS84 [WGS 84] format, an optional view name, and a list of indexed peers.

The format and function of the Collector BGP ID, the View Name Length and View Name, Peer Count are as defined by the TABLE_DUMP_V2 MRT format [I-D.ietf-grow-mrt].

The Collector Latitude and Collector Longitude are the geographical coordinates of the collector in WGS84 [WGS 84] datum decimal degrees format stored as a single precision float in the 32 bits allocated to the Latitude and Longitude.



The format of the peer entries is shown below. The PEER_INDEX_TABLE record contains Peer Count peer entries.



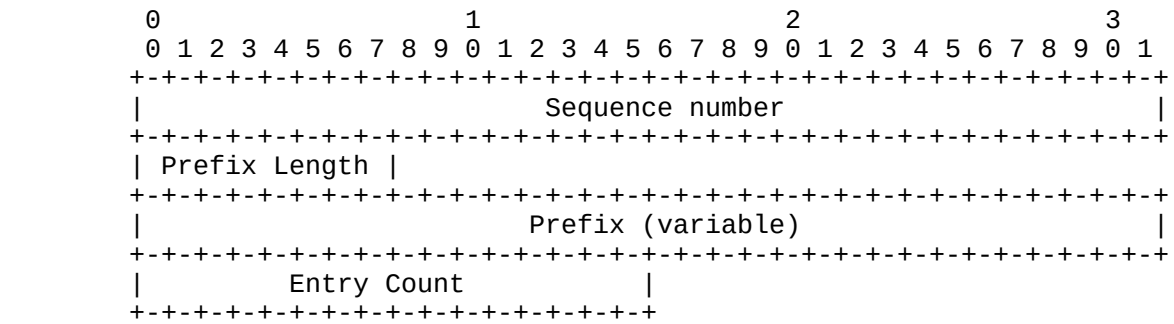
The Peer Type, Peer BGP ID, Peer IP, Peer AS, Peer Latitude, and Peer Longitude fields are repeated as indicated by the Peer Count field. The position of the Peer in the PEER_INDEX_TABLE is used as an index in the subsequent TABLE_DUMP_V2+GEO MRT records. The index number begins with 0.

The Peer Latitude and Peer Longitude are the geographical coordinates of the collector in WGS84 [WGS 84] datum decimal degrees format stored as a single precision float in the 32 bits allocated to the Latitude and Longitude.

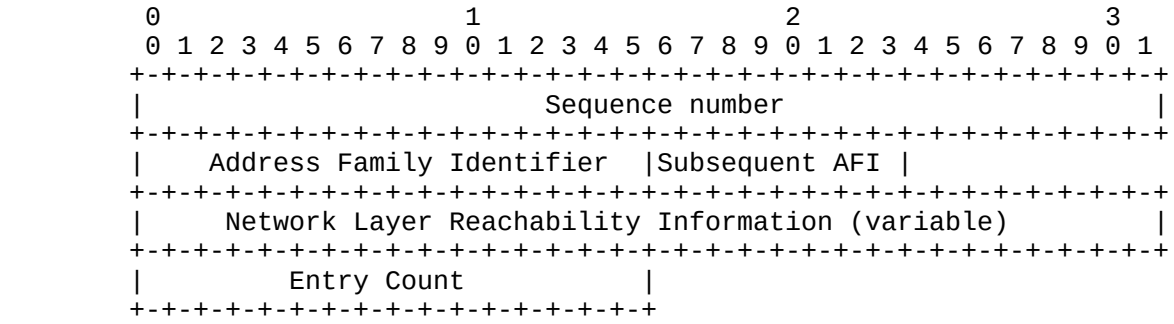
The Peer Type field remains as defined in the TABLE_DUMP_V2 MRT format [I-D.ietf-grow-mrt].

The records which follow the PEER_INDEX_TABLE record constitute the RIB entries and their formats remain unchanged from TABLE_DUMP_V2+ GEO.

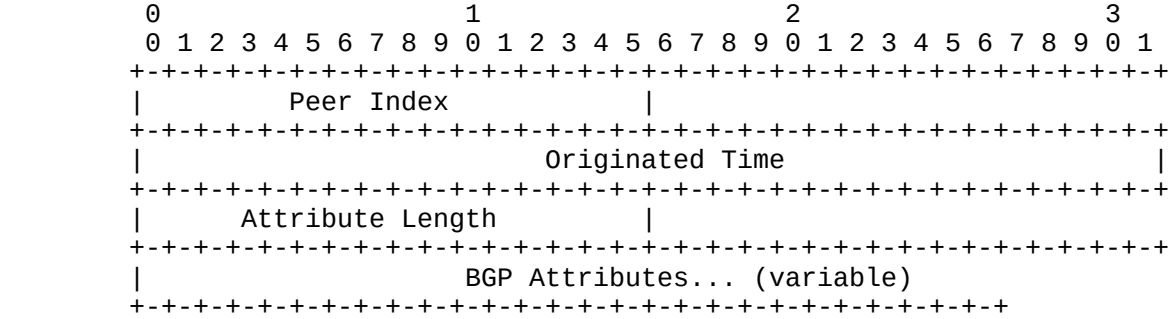
That is the common format for the RIB_IPV4_UNICAST, RIB_IPV4_MULTICAST, RIB_IPV6_UNICAST, and RIB_IPV6_MULTICAST remains as defined for TABLE_DUMP_V2 and the header is shown below for informational purposes only.



Similarly the the RIB_GENERIC format is unchanged and is shown here:



The RIB entries that follow the RIB entry headers are also unchanged from MRT [I-D.ietf-grow-mrt]:



5. Implementation Note

In implementation of the formats above where a Collector has an assigned Latitude and Longitude but a Peer does not. It is currently recommended that the Collector's coordinates are replicated in the Peer's Latitude and Longitude. The inquiring researcher can then make the decision on the interpretation of the routes 'as seen' at those coordinates, or disregard any geographical information for the peer based on the comparison of the Collector and Peer coordinates.

The TABLE_DUMP_v2+GEO format MUST not be used if the Collector's Latitude and Longitude have not been defined.

6. Acknowledgements

Thanks to Andrew Clark, Ernest Foo, Dave Meyer, Larry Bluck, and Jeffrey Haas for reviewing this document.

This document describes a small portion of the research towards the author's PhD.

7. IANA Considerations

This section requests the Internet Assigned Numbers Authority (IANA) register the Type code values (as FCFS as defined in the "MRT format" [I-D.ietf-grow-mrt] and Subtype code values related to the TABLE_DUMP_v2+GEO type as an entry in the MRT namespaces, in accordance with BCP 26, RFC 5226 [RFC5226].

8. Security Considerations

This extension to the "MRT format" [I-D.ietf-grow-mrt] defines fields that are of a descriptive nature and provide information that is useful in the analysis of routing systems. As such, the author believes that they do not constitute an additional security risk.

9. References

9.1. Normative References

[I-D.ietf-grow-mrt]
Blunk, L., Karir, M., and C. Labovitz, "MRT routing information export format", draft-ietf-grow-mrt-11 (work in progress), March 2010.

[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.

9.2. Informative References

[MRT PROG GUIDE]
Labovitz, C., "MRT Programmer's Guide", November 1999, <http://www.merit.edu/networkresearch/mrtprogrammer.pdf>.

[WGS 84] Geodesy and Geophysics Department, DoD., "World Geodetic System 1984", January 2000, <http://earth-info.nga.mil/GandG/publications/tr8350.2/wgs84fin.pdf>.

Author's Address

Terry Manderson
ICANN

Email: terry.manderson@icann.org

GROW Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 29, 2011

E. Jasinska
Limelight Networks
N. Hilliard
INEX
R. Raszuk
Cisco Systems
N. Bakker
AMS-IX B.V.

October 26, 2010

Internet Exchange Route Server
draft-jasinska-ix-bgp-route-server-01

Abstract

The growing popularity of Internet exchange points (IXPs) brings a new set of requirements to interconnect participating networks. While bilateral exterior BGP sessions between exchange participants were previously the most common means of exchanging reachability information, the overhead associated with dense interconnection has caused substantial operational scaling problems for Internet exchange point participants.

This document outlines a specification for multilateral interconnections at IXPs. Multilateral interconnection is a method of exchanging routing information between three or more BGP speakers using a single intermediate broker system, referred to as a route server. Route servers are typically used on shared access media networks such as Internet exchange points (IXPs), to facilitate simplified interconnection between multiple Internet routers on such a network.

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), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt.

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html.

This Internet-Draft will expire on April 29, 2011.

Copyright Notice

Table of Contents

1. Introduction to Multilateral Interconnection

Internet exchange points (IXPs) provide IP data interconnection facilities for their participants, typically using shared Layer-2 networking media such as Ethernet. The Border Gateway Protocol (BGP) [RFC4271], an inter-Autonomous System routing protocol, is commonly used to facilitate exchange of network reachability information over such media.

In the case of bilateral interconnection between two exchange participant routers, each router must be configured with a BGP session to the other. At IXPs with many participants who wish to implement dense interconnection, this requirement can lead both to large router configurations and high administrative overhead. Given the growth in the number of participants at many IXPs, it has become operationally troublesome to implement densely meshed interconnections at these IXPs.

Multilateral interconnection is a method of interconnecting BGP speaking routers using a third party brokering system, commonly referred to as a route server and typically managed by the IXP operator. Each of the multilateral interconnection participants (usually referred to as route server clients) announces network reachability information to the route server using exterior BGP, and the route server in turn forwards this information to each other route server client connected to it, according to its configuration. Although a route server uses BGP to exchange reachability information with each of its clients, it does not forward traffic itself and is therefore not a router.

A route server can be viewed as similar in function to an [RFC4456] route reflector, except that it operates using EBGP instead of iBGP. Certain adaptions to [RFC4271] are required, to enable an EBGP router to operate as a route server, which are outlined in Section 4 of this document. Operational considerations to be taken into account in a route server deployment are subject of Section 5.

The term "route server" is often in a different context used to describe a BGP node whose purpose is to accept BGP feeds from multiple clients for the purpose of operational analysis and troubleshooting. A system of this form may alternatively be known as a "route collector" or a "route-views server". This document uses the term "route server" exclusively to describe multilateral peering brokerage systems.

1.1. Specification of Requirements

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

2. Bilateral Interconnection

Bilateral interconnection is a method of interconnecting routers using individual BGP sessions between each participant router on an IXP in order to exchange reachability information. While interconnection policies vary from participant to participant, most IXPs have significant numbers of participants who see value in interconnecting with as many other exchange participants as possible. In order for an IXP participant to implement a dense interconnection policy, it is necessary for the participant to liaise with each of their intended interconnection partners and if this partner agrees to interconnect, then both participants' routers must be configured with a BGP session to exchange network reachability information. If each exchange participant interconnects with each other participant, a full mesh of BGP sessions is needed, as detailed in Figure 1.

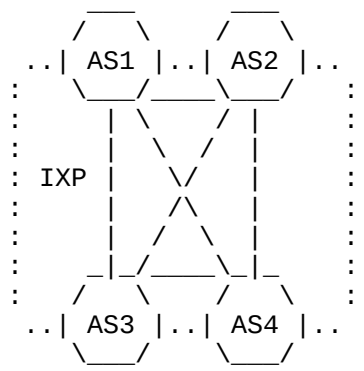


Figure 1: Full-Mesh Interconnection at an IXP

Figure 1 depicts an IXP platform with four connected routers, administered by four separate exchange participants, each of them with a locally unique autonomous system number: AS1, AS2, AS3 and AS4. Each of these four participants wishes to exchange traffic with all other participants; this is accomplished by configuring a full mesh of BGP sessions on each router connected to the exchange, resulting in 6 BGP sessions across the IXP fabric.

The number of BGP sessions at an exchange has an upper bound of n*(n-1)/2, where n is the number of routers at the exchange. As many exchanges have relatively large numbers of participating networks, the quadratic scaling requirements of dense interconnection tend to cause operational and administrative overhead at large IXPs. Consequently, new participants to an IXP require significant initial resourcing in order to gain value from their IXP connection, while existing exchange participants need to commit ongoing resources in order to benefit from interconnecting with these new participants.

3. Multilateral Interconnection

Multilateral interconnection is implemented using a route server configured to use BGP to distribute network layer reachability information (NLRI) among all client routers. The route server preserves the BGP NEXT_HOP attribute from all received NLRI UPDATE messages, and passes these messages with unchanged NEXT_HOP to its route server clients, according to its configured routing policy. Using this method of exchanging NLRI messages, an IXP participant router can receive an aggregated list of prefixes from all other route server clients using a single BGP session to the route server instead of depending on BGP sessions with each other router at the exchange. This reduces the overall number of BGP sessions at an Internet exchange from n*(n-1)/2 to n, where n is the number of routers at the exchange.

In practical terms, this allows dense interconnection between IXP participants with low administrative overhead and significantly simpler and smaller router configurations. In particular, new IXP participants benefit from immediate and extensive interconnection, while existing route server participants receive reachability information from these new participants without necessarily having to adapt their configurations.

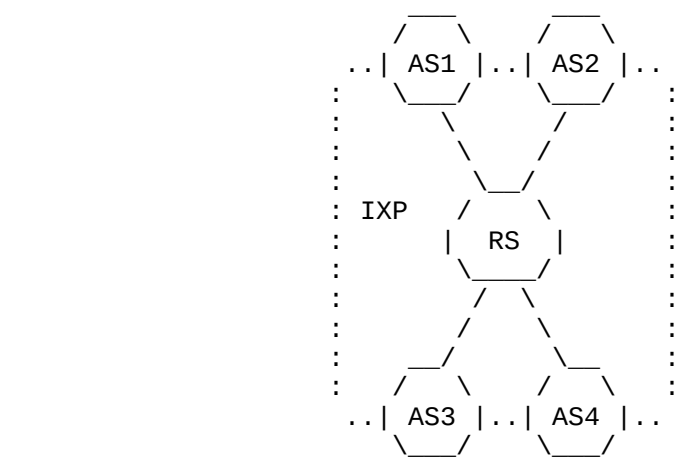


Figure 2: IXP-based Interconnection with Route Server

As illustrated in Figure 2, each router on the IXP fabric requires only a single BGP session to the route server, from which it can receive reachability information for all other routers on the IXP which also connect to the route server.

4. Technical Considerations for Route Server Implementations

4.1. Client UPDATE Messages

A route server MUST accept all UPDATE messages received from each of its clients for inclusion in its Adj-RIB-In. These UPDATE messages MAY by omitted from the route server's Loc-RIB or Loc-RIBs, due to filters configured for the purposes of implementing routing policy. The route server SHOULD perform one or more BGP Decision Processes to select routes for subsequent advertisement to its clients, taking into account possible configuration to provide multiple NLRI paths to a particular client as described in Section 4.3.2.2 or multiple Loc-RIBs as described in Section 4.3.2.1. The route server SHOULD forward UPDATE messages where appropriate from its Loc-RIB or Loc-RIBs to its clients.

4.2. Attribute Transparency

As a route server primarily performs a brokering service, modification of attributes could cause route server clients to alter their BGP best-path selection process for received prefix reachability information, thereby changing the intended routing policies of exchange participants. Therefore, contrary to what is

specified in section 5. of [RFC4271], route servers SHOULD NOT update well-known BGP attributes received from route server clients before redistributing them to their other route server clients. Optional recognized and unrecognized BGP attributes, whether transitive or non-transitive, SHOULD NOT be updated by the route server and SHOULD be passed on to other route server clients.

4.2.1. NEXT_HOP Attribute

The NEXT_HOP, a well-known mandatory BGP attribute, defines the IP address of the router used as the next hop to the destinations listed in the Network Layer Reachability Information field of the UPDATE message. As the route server does not participate in the actual routing of traffic, the NEXT_HOP attribute MUST be passed unmodified to the route server clients, similar to the "third party" next hop feature described in section 5.1.3. of [RFC4271].

4.2.2. AS_PATH Attribute

AS_PATH is a well-known mandatory attribute which identifies the autonomous systems through which routing information carried in the UPDATE message has passed.

As a route server does not participate in the process of forwarding data between client routers, and because modification of the AS_PATH attribute could affect route server client best-path calculations, the route server SHOULD NOT prepend its own AS number to the AS_PATH segment nor modify the AS_PATH segment in any other way.

4.2.3. MULTI_EXIT_DISC Attribute

MULTI_EXIT_DISC is an optional non-transitive attribute intended to be used on external (inter-AS) links to discriminate among multiple exit or entry points to the same neighboring AS. If applied to an NLRI UPDATE sent to a route server, the attribute (contrary to section 5.1.4 of [RFC4271]) SHOULD be propagated to other route server clients and the route server SHOULD NOT modify the value of this attribute.

4.2.4. Communities Attributes

The BGP COMMUNITIES ([RFC1997]) and Extended Communities ([RFC4360]) attributes are attributes intended for labeling information carried in BGP UPDATE messages. Transitive as well as non-transitive Communities attributes applied to an NLRI UPDATE sent to a route server SHOULD NOT be modified, processed or removed. However, if such an attribute is intended for processing by the route server itself, it MAY be modified or removed.

4.3. Per-Client Prefix Filtering

4.3.1. Prefix Hiding on a Route Server

While IXP participants often use route servers with the intention of interconnecting with as many other route server participants as possible, there are several circumstances where control of prefix distribution on a per-client basis is important for ensuring that desired interconnection policies are met.

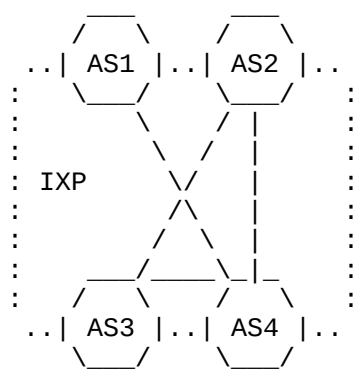


Figure 3: Filtered Interconnection at an IXP

Using the example in Figure 3, AS1 does not directly exchange prefix information with either AS2 or AS3 at the IXP, but only interconnects with AS4.

In the traditional bilateral interconnection model, prefix filtering to a third party exchange participant is accomplished either by not engaging in a bilateral interconnection with that participant or else by implementing outbound prefix filtering on the BGP session towards that participant. However, in a multilateral interconnection environment, only the route server can perform outbound prefix filtering in the direction of the route server client; route server clients depend on the route server to perform their filtering for them.

If the same prefix is sent to a route server from multiple route server clients with different BGP attributes, and traditional best-path route selection is performed on that list of prefixes, then the route server will select a single best-path prefix for propagation to all connected clients. If, however, the route server has been configured to filter the calculated best-path prefix from reaching a particular route server client, then that client will receive no

reachability information for that prefix from the route server, despite the fact that the route server has received alternative reachability information for that prefix from other route server clients. This phenomenon is referred to as "prefix hiding".

For example, in Figure 3, if the same prefix were sent to the route server via AS2 and AS4, and the route via AS2 was preferred according to BGP's traditional best-path selection, but AS2 was filtered by AS1, then AS1 would never receive this prefix, even though the route server had previously received a valid alternative path via AS4. This happens because the best-path selection is performed only once on the route server for all clients.

It should be noted that prefix hiding will only occur on route servers which employ per-client prefix filtering; if an IXP operator deploys a route server without prefix filtering, then prefix hiding does not occur, as all paths are considered equally valid from the point of view of the route server.

There are several techniques which may be employed to prevent the prefix hiding problem from occurring. Route server implementations SHOULD implement at least one method to prevent prefix hiding.

4.3.2. Mitigation Techniques

4.3.2.1. Multiple Route Server RIBs

The most portable means of preventing the route server prefix hiding problem is by using a route server BGP implementation which performs the per-client best-path calculation for each set of prefixes which results after the route server's client filtering policies have been taken into consideration. This can be implemented by using per-client Loc-RIBs, with prefix filtering implemented between the Adj-RIB-In and the per-client Loc-RIB. Implementations MAY optimize this by maintaining prefixes not subject to filtering policies in a global Loc-RIB, with per-client Loc-RIBs stored as deltas.

This problem mitigation technique is highly portable, as it makes no assumptions about the feature capabilities of the route server clients.

4.3.2.2. Advertising Multiple Paths

The prefix distribution model described above assumes standard BGP session encoding where the route server sends a single path to its client for any given prefix. This path is selected using the BGP path selection decision process described in [RFC4271]. If, however, it were possible for the route server to send more than a single path

to a route server client, then route server clients would no longer depend on receiving a single best path to a particular prefix; consequently, the prefix hiding problem described in Section 4.3.1 would disappear.

We present two methods which describe how such increased path diversity could be implemented.

4.3.2.2.1. Diverse BGP Path Approach

The Diverse BGP Path proposal as defined in [I-D.ietf-grow-diverse-bgp-path-dist] is a simple way to distribute multiple prefix paths from a route server to a route server client by using a separate BGP session from the route server to a client for each different path.

The number of paths which may be distributed to a client is constrained by the number of BGP sessions which the server and the client are willing to establish with each other. The distributed paths may be established from the global BGP Loc-RIB on the route server in addition to any per-client Loc-RIB. As there may be more potential paths to a given prefix than configured BGP sessions, this method is not guaranteed to eliminate the prefix hiding problem in all situations. Furthermore, this method may significantly increase the number of BGP sessions handled by the route server, which may negatively impact its performance.

4.3.2.2.2. BGP ADD-PATH Approach

The [I-D.ietf-idr-add-paths] Internet draft proposes a different approach to multiple path propagation, by allowing a BGP speaker to forward multiple paths for the same prefix on a single BGP session. As [RFC4271] specifies that a BGP listener must implement an implicit withdraw when it receives an UPDATE message for a prefix which already exists in its Adj-RIB-In, this approach requires explicit support for the feature both on the route server and on its clients.

If the ADD-PATH capability is negotiated bidirectionally between the route server and a route server client, and the route server client propagates multiple paths for the same prefix to the route server, then this could potentially cause the propagation of inactive, invalid or suboptimal paths to the route server, thereby causing loss of reachability to other route server clients. For this reason, ADD-PATH implementations on a route server SHOULD enforce send-only mode with the route server clients, which would result in negotiating receive-only mode from the client to the route server.

5. Operational Considerations for Route Server Installations

5.1. Route Server Scaling

While deployment of multiple Loc-RIBs on the route server presents a simple way to avoid the prefix hiding problem noted in Section 4.3.1, this approach requires significantly more computing resources on the route server than where a single Loc-RIB is deployed for all clients. As the [RFC4271] Decision Process must be applied to all Loc-RIBs deployed on the route server, both CPU and memory requirements on the host computer scale approximately according to O(P * N), where P is the total number of unique prefixes received by the route server and N is the number of route server clients which require a unique Loc-RIB. As this is a super-linear scaling relationship, large route servers may derive benefit from deploying per-client Loc-RIBs only where they are required.

Regardless of any Loc-RIB optimization implemented, the route server's control plane bandwidth requirements will scale according to O(P * N), where P is the total number of unique prefixes received by the route server and N is the total number of route server clients. In the case where P_avg (the arithmetic mean number of unique prefixes received per route server client) remains roughly constant even as the number of connected clients increases, this relationship can be rewritten as O((P_avg * N) * N) or O(N^2). This quadratic upper bound on the network traffic requirements indicates that the route server model will not scale to arbitrarily large sizes.

5.1.1. Tackling Scaling Issues

The network traffic scaling issue presents significant difficulties with no clear solution - ultimately, each client must receive a UPDATE for each unique prefix received by the route server. However, there are several potential methods for dealing with the CPU and memory resource requirements of route servers.

5.1.1.1. View Merging and Decomposition

View merging and decomposition, outlined in [RS-ARCH], describes a method of optimising memory and CPU requirements where multiple route server clients are subject to exactly the same routing policies. In this situation, the multiple Loc-RIB views required by each client are merged into a single view.

A variation of this approach may be implemented on route servers by ensuring that separate Loc-RIBs are only configured for route server clients with unique export peering policies.

5.1.1.2. Destination Splitting

Destination splitting, also described in [RS-ARCH], describes a method for route server clients to connect to multiple route servers and to send non-overlapping sets of prefixes to each route server. As each route server computes the best path for its own set of prefixes, the quadratic scaling requirement operates on multiple smaller sets of prefixes. This reduces the overall computational and memory requirements for managing multiple Loc-RIBs and performing the best-path calculation on each. In order for this method to perform well, destination splitting would require significant co-ordination between the route server operator and each route server client. In practice, such levels of co-ordination are unlikely to work successfully, thereby diminishing the value of this approach.

5.1.1.3. NEXT_HOP Resolution

As route servers are usually deployed at IXPs which use flat layer 2 networks, recursive resolution of the NEXT_HOP attribute is generally not required, and can be replaced by a simple check to ensure that the NEXT_HOP value for each prefix is a network address on the IXP LAN’s IP address range.

5.2. NLRI Leakage Mitigation

NLRI leakage occurs when a BGP client unintentionally distributes NLRI UPDATE messages to one or more neighboring BGP routers. NLRI leakage of this form to a route server can cause connectivity problems at an IXP if each route server client is configured to accept all prefix UPDATE messages from the route server. It is therefore RECOMMENDED when deploying route servers that, due to the potential for collateral damage caused by NLRI leakage, route server operators deploy NLRI leakage mitigation measures in order to prevent unintentional prefix announcements or else limit the scale of any such leak. Although not foolproof, per-client inbound prefix limits can restrict the damage caused by prefix leakage in many cases. Per-client inbound prefix filtering on the route server is a more deterministic and usually more reliable means of preventing prefix leakage, but requires more administrative resources to maintain properly.

5.3. Route Server Redundancy

As the purpose of an IXP route server implementation is to provide a reliable reachability brokerage service, it is RECOMMENDED that exchange operators who implement route server systems provision multiple route servers on each shared Layer-2 domain. There is no requirement to use the same BGP implementation or operating system

for each route server on the IXP fabric; however, it is RECOMMENDED that where an operator provisions more than a single server on the same shared Layer-2 domain, each route server implementation be configured equivalently and in such a manner that the path reachability information from each system is identical.

5.4. AS_PATH Consistency Check

As per [RFC4271] every BGP speaker who advertises a route to another external BGP speaker prepends its own AS number as the last element of the AS_PATH sequence. Therefore the leftmost AS in an AS_PATH attribute is equal to the autonomous system number of the BGP speaker that sent an UPDATE message.

[RFC4271] suggests in section 6.3 that a BGP speaker MAY check the AS_PATH attribute of each UPDATE message received for consistency, if the leftmost AS in the AS_PATH is in fact the one of the sender.

Route servers do not modify the AS_PATH attribute (as described in Section 4.2.2), since they do not participate in the traffic exchange. Therefore a consistency check on the AS_PATH of an UPDATE received by a route server client would fail. It is therefore RECOMMENDED that route server clients disable the AS_PATH consistency check towards the route server.

5.5. Implementing Routing Policies

Prefix filtering is commonly implemented on route servers to provide prefix distribution control mechanisms for route server clients. There are a few commonly used strategies available.

5.5.1. Communities

Prefixes sent to the route server are tagged with certain COMMUNITIES attributes agreed upon beforehand between the operator and all participants. Based on the values, routes are propagated to all other participants, a subset of participants, or none. This allows for one-way filtering policies to be implemented on the route server; if a participant chooses not to exchange routes with a certain other participant, he will have to instruct the route server to not announce his own routes and filter incoming routes on his own router.

5.5.2. Internet Routing Registry

Filters configured on the route server can be constructed by querying an Internet Routing Registry database for RPSL [RFC2622] objects placed there by participating operators. Import and export statements for the route server’s ASN in an aut-num object define

their desired policy, from which the configured filters are derived.

6. Security Considerations

On route server installations which do not employ prefix-hiding mitigation techniques, the prefix hiding problem outlined in section Section 4.3.1 can be used in certain circumstances to proactively block third party prefix announcements from other route server clients.

7. IANA Considerations

The new set of mechanism for route servers does not require any new allocations from IANA.

8. Acknowledgments

The authors would like to thank Chris Hall, Ryan Bickhart and Steven Bakker for their valuable input.

In addition, the authors would like to acknowledge the developers of BIRD, OpenBGPD and Quagga, whose open source BGP implementations include route server capabilities which are compliant with this document.

9. References

9.1. Normative References

[I-D.ietf-grow-diverse-bgp-path-dist]
Raszuk, R., Fernando, R., Patel, K., McPherson, D., and K. Kumaki, "Distribution of diverse BGP paths.", draft-ietf-grow-diverse-bgp-path-dist-02 (work in progress), July 2010.

[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.

[RFC1997] Chandrasekeran, R., Traina, P., and T. Li, "BGP Communities Attribute", RFC 1997, August 1996.

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

[RFC2622] Alaettinoglu, C., Villamizar, C., Gerich, E., Kessens, D., Meyer, D., Bates, T., Karrenberg, D., and M. Terpstra, "Routing Policy Specification Language (RPSL)", RFC 2622, June 1999.

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

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, 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.

[RS-ARCH] Govindan, R., Alaettinoglu, C., Varadhan, K., and D. Estrin, "A Route Server Architecture for Inter-Domain Routing", 1995, <http://www.cs.usc.edu/research/95-603.ps.Z>.

9.2. Informative References

[RFC1863] Haskin, D., "A BGP/IDRP Route Server alternative to a full mesh routing", RFC 1863, October 1995.

[RFC3418] Presuhn, R., "Management Information Base (MIB) for the Simple Network Management Protocol (SNMP)", STD 62, RFC 3418, December 2002.

[RFC4223] Savola, P., "Reclassification of RFC 1863 to Historic", RFC 4223, October 2005.

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

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

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

Authors' Addresses

Elisa Jasinska
Limelight Networks
2220 W 14th St
Tempe, AZ 85281
US

Email: elisa@llnw.com

Nick Hilliard
INEX
4027 Kingswood Road
Dublin 24
IE

Email: nick@inex.ie

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

Email: raszuk@cisco.com

Niels Bakker
AMS-IX B.V.
Westeinde 12
Amsterdam, NH 1017 ZN
NL

Email: niels.bakker@ams-ix.net

INTERNET-DRAFT Danny McPherson
Ryan Donnelly
Frank Scalzo
VeriSign, Inc.
Expires: March 2011 September 10, 2010
Intended Status: Best Current Practice

Unique Per-Node Origin ASNs for Globally Anycasted Services
<draft-mcpherson-unique-origin-as-00.txt>

Status of this Memo

http://www.ietf.org/shadow.html

Abstract

This document makes recommendations regarding the use of unique origin ASNs for globally anycasted critical infrastructure services.

Table of Contents

1. Terminology

This document employs much of the following terminology, which was taken in full from Section 2 of [RFC 4786].

Anycast: the practice of making a particular Service Address available in multiple, discrete, autonomous locations, such that datagrams sent are routed to one of several available locations.

Anycast Node: an internally-connected collection of hosts and routers that together provide service for an anycast Service Address. An Anycast Node might be as simple as a single host participating in a routing system with adjacent routers, or it might include a number of hosts connected in some more elaborate fashion; in either case, to the routing system across which the service is being anycast, each Anycast Node presents a unique path to the Service Address. The entire anycast system for the service consists of two or more separate Anycast Nodes.

Catchment: in physical geography, an area drained by a river, also known as a drainage basin. By analogy, as used in this document, the topological region of a network within which packets directed at an Anycast Address are routed to one particular node.

Local-Scope Anycast: reachability information for the anycast Service Address is propagated through a routing system in such a way that a particular anycast node is only visible to a subset of the whole routing system.

Local Node: an Anycast Node providing service using a Local-Scope Anycast Address.

Global Node: an Anycast Node providing service using a Global-Scope Anycast Address.

Global-Scope Anycast: reachability information for the anycast Service Address is propagated through a routing system in such a way that a particular anycast node is potentially visible to the whole routing system.

Service Address: an IP address associated with a particular service (e.g., the destination address used by DNS resolvers to reach a particular authority server).

2. Introduction

IP anycasting [RFC 4786] has been deployed for an array of network services since the early 1990s. It provides a mechanism for a given network resource to be available in a more distributed manner, locally and/or globally, with a more robust and resilient footprint, commonly yielding better localization and absorption of systemic query loads, as well as better protections in the face of DDoS attacks, network partitions, and other similar incidents. A large part of the Internet root DNS infrastructure, as well as many other resources, has been anycasted for nearly a decade.

While the benefits realized by anycasting network services is proven, some issues do emerge with asserting routing system reachability for a common network identifier from multiple locations. Specifically, anycasting in BGP requires injection of reachability information in the routing system for a common IP address prefix from multiple locations. These anycasted prefixes and network services have traditionally employed a common origin autonomous system number (ASN) in order to preserve historically scarce 16-bit AS number space utilized by BGP for routing domain identifiers in the global routing system. Additionally, a common origin AS number was used in order to ease management overhead of resource operations associated with acquiring and maintaining multiple discrete AS numbers, as well as to avoid triggering various operations- oriented reporting functions aimed at identifying "inconsistent origin AS announcements" observed in the routing system. As a result, the representation of routing system path attributes associated with those service instances, and that anycasted prefix itself, typically bear no per-instance discriminators in the routing system (i.e., within the network control plane itself).

Service level query capabilities may or may not provide a mechanism to identify which anycast node responded to a particular query, although this is likely both service (e.g., DNS or NTP) and implementation dependent. For example, NSD, Unbound, and BIND all provide ’hostname.bind or hostname.id’ [HNAME] query support that enables service-level identification of a given server. Tools such as traceroute are also used to determine which location a given query is being routed to, although it may not reveal local-scope anycast instances, or if there are multiple servers within a given anycast node, which of the servers responded to a given query, in particular when multiple servers within an anycast node are connected to a single IP router. When utilizing these service level capabilities, query responses are typically both deterministic and inherently topology-dependent, however, these service level identifiers at the data plane provide no control plane (routing system) uniqueness.

As more services are globally anycasted, and existing anycasted services realize wider deployment of anycast nodes for a given service address in order to accommodate growing system loads, the difficulty of providing safeguards and controls to better protect those resources expands. Intuitively, the more widely distributed a given anycasted service address is, the more difficult it becomes for network operators to detect operational and security issues that affect that service. Some examples of such security and operational issues include BGP route leaks affecting the anycasted service, rogue anycast nodes appearing for the service, or the emergence of other aberrant behavior in either the routing system, the forward query datapath, or query response datapath. Diagnosis of the routing system issues is complicated by the fact that no unique discriminators exist in the routing system to identify a given local or global anycast node. Furthermore, both datapath and routing system problem identification is compounded by the fact that either incident type can be topologically-dependent.

Additionally, while it goes without saying that anycasted services should always strive for exact synchronization across all instances of an anycasted service address, if local policies or data plane response manipulation techniques were to "influence" responses within a given region in such a way that those response are no longer authentic or that they diverge from what other nodes within an anycasted service were providing, then it should be an absolute necessity that those modified resources only be utilized by service consumers within that region or influencer’s jurisdiction.

Mechanisms should exist at both the network and service layer to make it abundantly apparent to operators and users alike whether any of the query responses are not authentic. For DNS, DNSSEC [RFC 4033] provides this capability at the service layer, assuming validation is being enforced by recursive name servers, and DNSSEC deployment at the root and top level domain (TLD) levels is well underway [DNSSEC-DEPLOY]. Furthermore, control plane discriminators should exist to enable operators to know toward which of a given set of instances a query is being directed, and to enable detection and alerting capabilities when this changes. Such discriminators may also be employed to enable anycast node preference or filtering keys, should local operational policy require it.

3. Recommendation for Unique Origin ASNs

In order to be able to better detect changes to routing information

associated with crtical anycasted resources, globally anycasted services with partitioned origin ASNs SHOULD utilize a unique origin ASN per node where possible.

Discrete origin ASNs per node provide a discriminator in the routing system that would enable detection of leaked or hijacked instances more quickly, and would also enable operators that so choose to proactively develop routing policies that express preferences or avoidance for a given node or set of nodes associated with an anycasted service. This is particularly useful when it is observed that local policy or known issues exist with the performance or authenticity of responses returned from a specific anycast node, or that enacted policies meant to affect service within a particular region are affecting users outside of that region as a result of a given anycast catchment expanding beyond its intended scope.

Furthermore, inconsistent origin AS announcements associated with anycasted services for critical infrastructure SHOULD NOT be deemed undesirable by routing system reporting functions, but should instead be embraced in order to better identify the connectedness and footprint of a given anycasted service.

While namespace conservation and reasonable use of AS number resources should always be a goal, the introduction of 32-bit ASNs significantly lessens concerns in this space. Globally anycasted resources, in particular those associated with critical infrastructure-enabling services such as root and TLD name servers, SHOULD warrant special consideration with regard to AS number allocation practices during policy development by the constituents of those responsible organizations (e.g., the Regional Internet Registries). Additionally, defining precisely what constitutes "critical infrastructure services" or "special consideration" (e.g., some small range of 32-bit AS numbers might be provided) is left to the constituents of those organizations. Additionally, critical infrastructure employment of 32-bit ASNs for new nodes might well help to foster adoption of native 32-bit ASN support by network operators.

One additional benefit of unique origin AS numbers per anycast node is that Resource PKI (RPKI) Secure Inter-domain Routing [SIDR] machinery, and in particular, that of Route Origin Authorizations (ROAs), and routing policies that may be derived based on those ROAs, can be employed with per anycast node resolution, rather than relying on a single ROA and common origin AS to cover all instantiations of an anycasted prefix (possibly hundreds) within the global routing system. For example, deployments that incorporate partitioned ASN anycast models that have a single ASN bound to all nodes but cross organizational or political boundaries, a situation may arise where

nobody would be deemed appropriate to hold the key for the ROA. Additionally, a globally anycasted service within a given IP prefix that shares a common ASN might be taken totally offline because of the revocation of a ROA for that origin.

4. Additional Recommendations for Globally Anycasted Services

Two additional recommendations for globally anycasted critical infrastructure services are related to publication of information associated with a given node's physical location, and which adjacent upstream ASNs an origin AS interconnects with. The former would allow operators to better define and optimize preferences associated with a given node to align with local policy and service optimizations. The latter would allow expression though policy such as Routing Policy Specification Language [RFC 4012] specified in Internet Routing Registries (IRRs) in a manner that illustrates a discrete set of upstream ASNs for each anycast node, rather than the current model where all upstream ASNs associated with a common origin AS may or may not be expressed. This information would provide an additional level of static AS path validation or monitoring and detection models by network operators, and perhaps explicit network layer source address validation in the datapath.

5. Security Considerations

The recommendations made in this memo aim to provide more flexibility for network operators hoping to better monitor and prevent issues related to globally anycasted critical infrastructure resources. Anycast itself provides considerable benefit in the face of certain attacks, yet if a given instance of a service can appear at many points in the routing system and legitimate instances are difficult to distinguish from malicious ones, then anycast expands the service's attack surface rather than reducing it.

The recommendations made in this document are expressed to assist with visibility and policy specification capabilities in order to improve the availability of critical Internet resources. Use cases where the recommendations outlined in this memo may have helped to more easily detect or scope the impact of a particular incident are illustrated in [RENESYS-BLOG].

Furthermore, while application layer protection mechanisms such as DNSSEC provide integrity and authentication, they often do so at the cost of introducing more failure conditions. For example, if a recursive name server is performing DNSSEC validator functions and receives a bogus response to a given query as a result of a man-in-the-middle (MITM) or injected spoofed response packet such as a cache poisoning attempt, the possibility might exist that that packet is processed by the server and results in some temporal or persistent DoS condition. The unique origin AS mechanism outlined in this document provides the capability for network operators to expressly avoid anycast node catchments known to regularly elicit bogus responses, while allowing the anycasted service address to remain available otherwise.

6. Deployment Considerations

Maintenance of unique ASNs for each node within an anycasted service may be challenging for some critical infrastructure service operators initially, but for globally anycasted resources there needs to be some type of discriminator in the control plane to enable detection, remediation, and optimally, preventative controls for dealing with routing system anomalies that are intensified by the application of anycast. Additionally, this technique sets the stage to employ RPKI-enabled machinery and more secure and explicit routing policies, which all operators should be considering.

The granularity of data publication related to anycast node location should be left to the devises of each services operator, but some reasonable level of detail to inform operators as outlined herein should be provided by each critical services operator.

Adjacent AS information for a given origin AS can be obtained through careful routing system analysis already and should present no new threat. However, network interconnection and peering policies may well present some challenges in this area. That said, interconnection with networks that provide critical infrastructure services should certainly be given due consideration as such by network operators when evaluating interconnection strategies.

Some root and TLD operators today identify erroneous anycast prefix announcements by detecting prefix announcements with an origin AS other than the common origin AS shared via all nodes. This detection model would need to be expanded to account for unique origin ASNs per node, and given that AS paths are trivial to manipulate, the above

technique would only assist in the event of unintentional configuration errors that reoriginate the route (e.g., it doesn't even detect leaks that preserve the initial path elements)..

Finally, anycast node presence at exchange points that employs route servers may make enumeration of adjacent ASNs for a given node challenging. While this is understood, service operators should make every effort to enumerate the set of adjacent ASNs associated with a given anycast node's origin AS.

7. Acknowledgements

Thanks to David Conrad, Steve Kent, Mark Kosters, Andrei Robachevsky, Paul Vixie, Brad Verd, Andrew Herrmann, and Randy Bush for early review and comments on this concept.

8. IANA Considerations

This document requires no direct IANA actions, although it does provide general guidance to number resource allocation and policy development organizations,

9. References

9.1. Normative References

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

[RFC 4786] Abley, J., and Lindqvist, K., "Operation of Anycast Services", RFC 4786, BCP 126, December 2006.

9.2. Informative References

[RFC 4012] Blunk, et al., "Routing Policy Specification Language next generation (RPSLng)", RFC 4012, March 2005.

[RFC 4033] Arends, et al., "DNS Security Introduction and Requirements", RFC 4033, March 2005.

[DNSSEC-DEPLOY] "Root DNSSEC", <http://www.root-dnssec.org/>

[HNAME] ISC, "Which F-root node am I using?" <http://www.isc.org/community/f-root/which_node>

[RENESYS-BLOG] Zmijewski, E., "Accidentally Importing Censorship", Renesys Blog, March 30, 2010. <http://www.renesys.com/blog/2010/03/fouling-the-global-nest.shtml>

[SIDR] Lepinski, M., Kent, S., "An Infrastructure to Support Secure Internet Routing", October 2009, Internet-Draft, "Work in Progress".

10. Authors' Addresses

Danny McPherson
Verisign, Inc.
Email: dmcpherson@verisign.com

Ryan Donnelly
Verisign, Inc.
Email: rdonnelly@verisign.com

Frank Scalzo
Verisign, Inc.
Email: fscalzo@verisign.com

Copyright Statement