IDR Working Group
Internet-Draft
Intended status: Standards Track
Expires: Jun 3, 2015

J. Uttaro
AT&T

P. Francois
IMDEA Networks

K. Patel
Cisco Systems

P. Mohapatra
Cumulus Networks

J. Haas
Juniper Networks

A. Simpson
R. Fragassi
Alcatel-Lucent

Dec 3, 2014

Best Practices for Advertisement of Multiple Paths in IBGP
draft-ietf-idr-add-paths-guidelines-07.txt

Status of this Memo

Copyright Notice

Abstract

Add-Paths is a BGP enhancement that allows a BGP router to advertise multiple distinct paths for the same prefix/NLRI. This provides a number of potential benefits, including reduced routing churn, faster convergence and better loadsharing.

This document provides recommendations to implementers of Add-Paths so that network operators have the tools needed to address their specific applications and to manage the scalability impact of Add-Paths. A router implementing Add-Paths may learn many paths for a prefix and must decide which of these to advertise to peers. This document analyses different algorithms for making this selection and provides recommendations based on the target application.

Table of Contents

1. Introduction

The BGP Add-Paths capability enhances current BGP implementations by allowing a BGP router to exchange with its BGP peers more than one path for the same destination/NLRI. The base BGP standard [RFC 4271] does not provide for such a capability. If a BGP router learns multiple paths for the same NLRI (from multiple peers), it selects only one as its best path and advertises the best path to its peers. The primary goal of Add-Paths is to increase the visibility of paths within an iBGP system. This has the effect of improving robustness in case of failure, reducing the number of BGP messages exchanged during such an event, and offering the potential for faster re-convergence. Through careful selection of the paths to be advertised, Add-Paths can also prevent routing oscillations.

The purpose of this document is to provide the necessary recommendations to the implementers of Add-Paths so that network operators have the tools needed to address their specific applications and to manage the scalability impact of Add-Paths while maintaining routing consistency. A router implementing Add-Paths may learn many paths for a prefix and must decide which of these to advertise to peers. This document analyses different algorithms for making this selection and provides recommendations based on the target application.

2. Terminology

In this document the following terms are used:

Add-Paths peer: refers a peer with which the local system has agreed to receive and/or send NLRI with path identifiers

Primary path: A path toward a prefix that is considered a best path by the BGP decision process [RFC 4271] and actively used for forwarding traffic to that prefix. A router may have multiple primary paths for a prefix if it implements multipath.

Diverse path: A BGP path associated with a different BGP next-hop and BGP router than some other set of paths. The BGP router associated with a path is inferred from the ORIGINATOR_ID attribute or, if there is none, the BGP Identifier of the peer that advertised the path.

Backup path: A diverse path with respect to the primary paths toward a prefix. The backup path can be used to forward traffic to the destination if the primary paths fail.

Optimal backup path: The backup path that will be selected as the new best path for a prefix when all primary paths are removed/withdrawn.

AS-Wide preferred paths: All paths that are considered as best when applying rules of the BGP decision process up to the IGP tie-break.

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].

3. Add-Paths Applications

[draft-pmohapat] presents the applications that would benefit from multiple paths advertisement in iBGP. They are summarized in the following subsections.

3.1. Fast Connectivity Restoration

With the dissemination of backup paths, fast connectivity restoration and convergence can be achieved. If a router has a backup path, it can directly select that path as best upon failure of the primary path. This minimizes packet loss in the dataplane. Sending multiple paths in iBGP allows routers to receive backup paths when path visibility is not sufficient with classical BGP. This is especially useful when Route Reflection is used.

Consider a network such as the one depicted in Figure 1 and suppose that none of the routers support Add-Paths. AS1 receives from AS3 2 paths (A and B) to a particular destination XYZ. Suppose path A is preferred over path B due to path A having a lower MED (multi-exit discriminator).

AS1 uses a route reflector RR1 to reduce the scale of its IBGP mesh. If the routers in AS1 are not configured for best-external then RR1 knows about only path A during steady state because router B suppresses/withdraws its advertisement of path (B) to RR1. If the routers in AS1 do support best-external then RR1 may have both paths in its Adj-RIB-IN, but regardless of the best-external configuration RR1 can only advertise its best path A to its peers, including router D.

```
========        =====================
=  +---+        +---+           +---+
=  |RTR|________|RTR|           |RTR|
=  | E |        | A |           | C |
=  +---+Path A->+---+    AS1    +---+
=      =        =    \         /    =
=      =        =     \       /     =
=      =        =      \     /      =
=      =        =       \   /       =
= AS3  =        =       +---+       =
=      =        =       |RR |       =
=      =        =       | 1 |       =
=      =        =       +---+       =
=      =        =       /   \       =
=      =        =      /     \      =
=      =        =     /       \     =
=      =        =    /         \    =
=  +---+Path B->+---+           +---+
=  |RTR|  ______|RTR|           |RTR|
=  | F |        | B |           | D |
=  +---+        +---+           +---+
========        =====================
```

Figure 1: Example Topology

Under these circumstances consider the steps required to restore traffic from router D to destination XYZ when the link between Router A and Router E fails. (Assume that router A set next-hop to self when advertising path A and that router B is not configured for best-external).

1. Router A sends a BGP UPDATE message withdrawing its advertisement of path (A).

2. RR1 receives the withdrawal, and propagates it to its other client peers, routers B, C and D.

3. When router B receives the withdrawal of path (A) it reruns its decision process and selects path (B) as its new best path. Router B advertises path (B) to RR1.

4. RR1 reruns its decision process and selects path (B) as its new best path. RR1 advertises path (B) to client peers A, C and D.

5. Router D reruns its decisions process, determines path (B) to be the best path, and updates its forwarding table. After this step traffic from router D to destination XYZ is restored (the traffic path has changed from A to B).

With the use of Add-Paths, the convergence time for the above path failure example can be reduced considerably. The main reason for the improvement is that Add-Paths allows router D to be aware of more than one path to destination XYZ prior to the failure of the best path (A). In steady-state (with no failures) router B decides, as before, that path (A) is its best path but because of its Add-Paths (or best-external) configuration it also advertises path (B) to RR1. Using Add-Paths RR1 can advertise both learned paths to its IBGP peers, including router D. Now consider again the scenario where the link between Router A and Router E fails. In this case, with Add-Paths, fewer steps are required to achieve re-convergence:

1. Router A sends a BGP UPDATE message withdrawing its advertisement of path (A).

2. RR1 receives the withdrawal, and propagates it to its other client peers, routers B, C and D.

3. Router D receives the withdrawal, reruns the decision process and updates the forwarding entry for destination XYZ.

3.2. Load Balancing

Increased path diversity allows routers to install several paths in their forwarding tables in order to load balance traffic across those paths.

3.3. Churn Reduction

When Add-Paths is used in an AS, the availability of additional backup paths means failures can be recovered locally with much less path exploration in iBGP and therefore less updates disseminated in eBGP. When the preferred backup path is the post-convergence path, churn is minimized.

3.4. Suppression of MED-Related Persistent Route Oscillation

As described in [oscillation], Add-Paths is a valuable tool in helping to stop persistent route oscillations caused by comparison of paths based on MED in topologies where route reflectors or the confederation structure hide some paths. With the appropriate path selection algorithm Add-Paths stops these route oscillations because the same set of paths are consistently advertised by the route reflector or the confederation border router and the routers receiving this set of paths make stable routing decisions about the best path.

4. Implementation Guidelines

This section discusses recommendations for the implementation of Add-Paths. The following topics are addressed:

- Considerations related to Add-Paths capability negotiation
- Receiving BGP routes from Add-Paths peers
- Advertising BGP routes to Add-Paths peers. This section discusses various path selection algorithms, which are the procedures available to an Add-Paths speaker for deciding which set of paths to advertise to an Add-Paths peer for particular prefixes.

4.1. Capability Advertisement

```
+---+           +---+
|RTR|___________|RTR|
| A |  <-BGP->  | B |
+---+           +---+
```

Figure 2: BGP Peering Example

In Figure 2, in order for a router A to receive multiple paths per NLRI from peer B, for a particular address family (AFI=x, SAFI=y), the BGP capabilities advertisements during session setup must indicate that peer B wants to send multiple paths for AFI=x, SAFI=y and that router A is willing to receive multiple paths for AFI=x, SAFI=y. Similarly, in order for router A to send multiple paths per NLRI to peer B, for a particular address family (AFI=x, SAFI=y), the BGP capabilities advertisements must indicate that router A wants to send multiple paths for AFI=x, SAFI=y and peer B is willing to receive multiple paths for AFI=x, SAFI=y. Refer to [Add-Paths] for details of the Add-Paths capabilities advertisement.

The capabilities of the local router MUST be configurable per peer and per address family, and SHOULD support the ability to configure send-only operation or receive-only operation. The default mode of operation is to both send and receive.

4.2. Receiving Multiple Paths

Currently, per standard BGP behavior, if a BGP router receives an advertisement of an NLRI and path from a specific peer and that peer subsequently advertises the same NLRI with different path information (e.g. a different NEXT_HOP and/or different path attributes) the new path effectively overwrites the existing path.

When Add-Paths has been negotiated with the peer, the newly advertised path should be stored in the RIB-IN along with all of the paths previously advertised (and not withdrawn) by the peer.

When an Add-Paths speaker has negotiated to receive multiple paths for (AFIx, SAFIy) from a peer all advertisements and withdrawals of NLRI within that address family from that peer MUST include a path identifier, as described in [Add-Paths]. The path identifiers have no significance to the receiving peer. If the combination of NLRI and path identifier in an advertisement from a peer is unique (does not match an existing route in the RIB-IN from that peer) then the route is added to the RIB-IN. If the combination of NLRI and path identifier in a received advertisement is the same as an existing route in the RIB-IN from the peer then the new route replaces the existing one. If the combination of NLRI and path identifier in a received withdrawal matches an existing route in the RIB-IN from the peer then that route shall be removed from the RIB-IN.

A BGP UPDATE message from an Add-Paths peer may advertise and withdraw more than one NLRI belonging to one or more address families. In this case Add-Paths may be supported for some of the address families and not others. In this situation the receiving BGP router should not expect that all of the path identifiers in the UPDATE message will be the same.

4.3. Advertising Multiple Paths

[Add-Paths] specifies how to encode the advertisement of multiple paths towards the same NLRI over an iBGP session, but provides no details about which set of multiple paths should be advertised. In this section, four path selection algorithms are described and compared with each other. These 4 algorithms are considered to be the most useful across the widest range of deployment scenarios. The list of possible path selection algorithms is much larger and for the interested reader Appendix A provides information about other path selection modes that were considered in historical versions of this document.

In comparing any two path selection algorithms the following factors should be taken into account:

Control Plane Load: When a router receives multiples paths for a prefix from an iBGP client it has to store more paths in its Adj-Rib-Ins.

Control Plane Stress: Coping with multiple iBGP paths has two implications on the computation that a router has to handle. First, it has to compute the paths to send to its peers, i.e. more than the best path. Second, it also has to handle the potential churn related to the exchange of those multiple paths.

MED/IGP oscillations: BGP sometimes suffers from routing oscillations when the physical topology differs from the logical topology, or when the MED attribute is used. This is due to the limited path visibility when a single path is advertised and Route Reflection is used. Increasing the path visibility by advertising multiple paths can help solve this issue.

Path optimality: When a single path is advertised, border routers do not always receive the optimal path. As an example, Route Reflectors typically send a single path chosen based on their own IGP tie-breaking procedure (although modifications to this are proposed in [BGP-ORR]). Increasing path visibility would also help routers to learn the path that is best suited for them w.r.t. the IGP tie-breaking.

Backup path optimality: Multiple paths advertisement gives routers the opportunity to have a backup path. However, some backup paths are better than others. Indeed, when a link failure occurs, if a router already knows its post-convergence path, the BGP re-convergence is straightforward and traffic is less impacted by the transient use of non-best forwarding paths.

Convergence time: Advertising multiple paths in iBGP has an impact on the convergence time of the BGP system. More paths need to be exchanged, but on the other hand, the routing information is propagated faster. With an increased path visibility, there is less path exploration during the convergence. Also, with the availability of backup paths, convergence time in case of failure is also reduced.

Target application: Depending on the application type, the number of paths to advertise for a prefix will vary. For example, for fast connectivity restoration, it may be sufficient to advertise only 2 paths to a peer so that it will have the best path and the optimal backup path. For load balancing purposes, it may be desirable to advertise more paths, but inclusion of the optimal backup path in the set may be less critical. For route oscillation elimination, it is required to advertise all group-best paths for a prefix.

4.3.1. Path Selection Modes

The following subsections describe the 4 main path selection modes considered in this draft. Each mode is considered either MANDATORY or OPTIONAL. A MANDATORY mode MUST be supported by any implementation that claims compliance with this document. An OPTIONAL made may be supported by some but not all implementations.

The path selection mode and any parameters applicable to the mode MUST be configurable per AFI/SAFI and per peer and SHOULD be configurable per prefix. To illustrate the value of this flexibility, consider a prefix P that belongs to an address family F requiring path IDs to be included with every NLRI (e.g. due to the Add-Paths capability negotiation with the peer). If P is one of a number of prefixes that would not benefit from the advertisement of multiple paths then it is perfectly valid to send only the best path.

4.3.1.1. Advertise N Paths

With the 'Advertise N Paths' mode (Add-N for short) a router advertises up to N paths per prefix towards an Add-Paths peer. The computational cost of this mode is the selection of the N paths. There must be a ranking of the paths in order to ensure consistency in the set of paths advertised to different Add-Paths peers. The recommended way for a router to consistently select N paths is to run its decision process N times and consider at each iteration only the paths that meet all of the following criteria:

(a) not selected during a previous iteration

(b)_diverse with respect to previously selected paths (see section 2 for the definition of a diverse path)

(c) not rejected by route filters or split horizon advertisement rules

The memory cost of this path selection mode is bounded: a router receives a maximum of N paths for each prefix from each peer. With N equal to 2, all routers know at least two paths and can provide local recovery in case of failure. If multipath routing is to be deployed in the AS, N can be increased to provide more alternate paths to the routers.

Path optimality and backup path optimality are not guaranteed, i.e. it is possible that the optimal path of a router (w.r.t. IGP tie-breaking) is not contained in the set of paths advertised by its Route Reflector. However, as the number of paths that it receives is higher than without Add-Paths, it is possible that the chosen nexthop is closer to the router in terms of IGP cost than the nexthop that would have been chosen without Add-Paths.

This solution helps to reduce routing oscillations, but not in all cases. Indeed, path visibility is still constrained by the maximum number of paths, and configurations with routing oscillations still exist.

This path selection mode is MANDATORY. The default value of N MUST be 2. The value of N MUST be configurable and MAY be upper bounded by an implementation.

The default value of 2 ensures the availability of a backup path (if 2 or more paths have been received) while maintaining minimum impact to memory and churn. If Add-N with N equal to 2 is insufficient to meet another objective (e.g. loadsharing or MED/IGP oscillation) there is always a large enough value of N that can selected, if N is configurable, to meet that objective.

4.3.1.2. Advertise All Paths

A simple rule for advertising multiple paths in iBGP is to advertise to iBGP peers all received paths minus those blocked by export filters or applicable split horizon rules. This solution is easy to implement, but the counterpart is that all those paths need to be stored by all routers that receive them, which can be quite expensive. If a path to a prefix P is advertised to N border routers, with a Full Mesh of iBGP sessions, all routers have N paths in their Adj-RIB-Ins. If Route Reflection is used and each client is connected to 2 Route Reflectors, it may learn up to 2*N paths.

This solution gives a perfect path visibility to all routers, thus limiting churn and losses of connectivity in case of failure. Indeed, this allows routers to select their optimal primary path, and to switch on their optimal backup path in case of failure.

However, as more paths are exchanged, the number of BGP messages disseminated during the initial iBGP convergence can be high, and convergence may be slower.

Routing oscillations are prevented with this rule, because a router won’t need to withdraw a previously advertised path when its best path changes.

This path selection mode is OPTIONAL.

4.3.1.3. Advertise All AS-Wide Best Paths

Another choice is to consider the set of paths with the same AS-wide preference [Basu-ibgp-osc], i.e. the paths that all routers would select based on the rules of the decision process that are not router-dependent (i.e. Local-preference, ASPath length and MED rules). Thus, for a given router, those paths only differ by the IGP cost to the nexthop or by the tie-breaking rules. The paths actually advertised to a peer are the set of AS-wide best paths minus those blocked by export filters or applicable split horizon rules.

The computational cost is reduced, as a router only has to send the paths remaining before applying the IGP tie-breaking rule. However, it is difficult to predict how many paths will be stored, as it depends on the number of eBGP sessions on which this prefix is advertised with the best AS-wide preference.

With this rule, the routing system is optimal: all routers can choose their best path (or best paths if multipath is used) based on their router-specific preferences, i.e. the IGP cost to the nexthop. Hot potato routing is respected. Also, MED oscillations are prevented, because the path visibility among the AS-wide preferred paths is total.

The existence of a backup path is not guaranteed. If only one path with the AS-wide best attributes exists, there is no backup path disseminated. However, if such a path exists, it is optimal as it has the same AS-wide preference as the primary

This path selection mode is OPTIONAL.

4.3.1.4. Advertise ALL AS-Wide Best and Next-Best Paths (Double AS Wide)

This variant of "Advertise All AS Wide Best Paths" trades-off the number of paths being propagated within the iBGP system for post-convergence alternate paths availability and routing stability. A BGP speaker running this mode will select, as candidates for advertisement, its AS Wide Best paths, plus all the AS Wide Best paths obtained when removing the first ones from consideration. The paths actually advertised to a peer are the double-AS wide candidate paths minus those blocked by export filters or applicable split horizon rules.

Under this mode, a BGP speaker knows multiple AS-Wide best paths or the AS-Wide best path and all the second AS-Wide best paths, so that routing optimality and backup path availability are ensured. Note that the post-convergence paths will be known by each BGP node in an AS supporting this mode.

The computation complexity of this mode is relatively low as it requires the router to run the usual BGP Decision Process up to and including the MED rule. The set of paths remaining after that step form the AS-Wide best paths. Next, a best path selection algorithm is run up to and including the MED rule, based on the paths that are not in the set of AS-Wide best paths.

The number of paths for a prefix p, known by a given router of the AS, is the number of AS-Wide best and second AS-Wide best paths found at the Borders of the AS.

MED Oscillations are avoided by this mode, both for the primary and alternate paths being picked under this mode.

This path selection mode is OPTIONAL.

4.3.1.5. Advertise Used Multipaths

Many BGP implementations support BGP Multipath, allowing a BGP router to use multiple BGP next-hops for forwarding towards a prefix/NLRI when the corresponding paths are considered equally preferred. In cases where the deployment of Add-Paths is mostly aimed at providing multiple paths for load balancing with BGP Multipath, a natural approach for a BGP speaker supporting Add-paths is to advertise the paths that are selected by its BGP multipath selection algorithm.

BGP Multipath selection algorithms can vary depending on the implementation and configuration options. An Add-Paths mode based on BGP multipath is considered practical because it lets the BGP path propagation be aligned with the load balancing objectives expressed by the operator configuring BGP multipath.

In some deployment scenarios, it is likely that such a mode leads to the selection and advertisement of a large number of paths for some NLRI, and hence should be controlled as per the mechanism described in section 4.3.2. In case the number of multipaths exceeds the upper bound on the number of advertised paths the ones that should be advertised are those with the highest degree of preference by the BGP decision process. This can be achieved if the advertising router has strictly ordered all of its paths.

This path selection mode is OPTIONAL.

4.3.2. Derived Modes from Bounding the Number of Advertised Paths

For some of the modes discussed in section 4.3.1 the number of paths selected by the algorithm (M) is not predictable in advance, and depends on factors such as network topology. For such modes, implementations MAY support the ability to limit the number of advertised paths to some value N that is less than M.

It must be noted that the resulting derivative mode may no longer meet the properties stated in section 4.3.1 (which assumes N=M). This is particularly true for the MED oscillation avoidance property. The use of such bounds thus needs to be considered carefully in deployments where MED oscillation avoidance is a key goal of deploying Add-path. If fast recovery is the main objective then it is reasonable and sufficient to set N to 2. If the main goal is improved load-balancing then limiting N to number of ECMP paths supported by the forwarding planes of the receiving routers is also a reasonable practice.

4.3.3. Derived Modes from Adding N More Paths

Some modes discussed in section 4.3.1 may result in only one or a few selected paths, depending on network topology and/or router configuration, and this small number of paths may not meet minimum requirements for backup path or load balancing purposes. When using such modes implementations MAY support the ability to add N more paths to the set returned by the basic selection algorithm as described in section 4.3.1. The N more paths should be the N next-best paths, as determined by the BGP decision process.

It must be noted that the resulting derivative mode may no longer meet the properties stated in section 4.3.1 (which assumes N=0).

5. Deployment Considerations

This section proposes a potential strategy for introducing Add-Paths into an existing network and discusses considerations related to scalability, routing consistency and routing churn.

5.1. Introducing Add-Paths into an Existing Network

There are many possible ways that Add-Paths can be introduced into an existing deployed network. It is not a practical goal for this document to list all of these options and discuss the pros and cons for each one. It is however valuable to consider an example migration strategy that may be relatively common among layer 3 service providers that currently use route reflectors for scaling. This example migration strategy is attractive for several reasons:

1. It involves incremental steps that allow the impact of Add-Paths to be carefully evaluated before proceeding to the next step.

2. It recognizes the fact that many routers will require at least a software upgrade to support Add-Paths, and it will not be practical to upgrade all of these routers all at once.

3. It reduces convergence time (in stages) with a relatively moderate increase in router memory and CPU demands.

The example migration strategy assumes a starting point of a deployed network with one or more RR clusters. None of the routers in the network support Add-Paths without an upgrade, but some do support best-external. Two of the clusters in this network are shown in Figure 3. In cluster 2, PE1, PE2, RRy and RRz are configured for best-external. This makes RRy and RRz aware of all external paths received by PEs in cluster 2 and ensures that RRy and RRz can advertise a path to the RRs in cluster 1 if it happens that the best overall route is learned from cluster 1. It doesn’t however allow other clusters to be aware of more than one path per prefix learned by cluster 2.

```
       ==========                ==================
       =        =                =                =
       =    +---+                +---+      +---+ =
       =    |RR |----------------|RR |  <-BE|   | =
       =    |a  |                |y  |------|PE1| =
       =    |   |                |   |      |   | =
       =    +---+                +---+      +---+ =
       =      | =                = |  \    /      =
       =      | =                = |   \  /       =
       =      | =                = |    \/        =
       =      | =                = |    /\        =
       =      | =                = |   /  \       =
       =      | =                = |  /    \      =
       =    +---+                +---+      +---+ =
       =    |RR |----------------|RR |------|   | =
       =    |b  |                |z  |  <-BE|PE2| =
       =    |   |                |   |      |   | =
       =    +---+                +---+      +---+ =
       =        =                =                =
       ==========                ==================
      RR Cluster 1                  RR Cluster 2
```

Figure 3: RR Cluster Before Add-Paths

The following sequence of steps occurs in the example migration strategy:

1. The route reflectors are upgraded in each cluster, one by one, to support Add-Paths. This allows the intra- and (eventually) inter-cluster RR-to-RR sessions to start using Add-Paths. All RRs are configured to use the Add-N, N=2 path selection algorithm. The effect of this step is to slightly reduce convergence time when the best and second-best paths for a prefix are learned by a single cluster (such as cluster 2 in Figure 3).

2. The clients are upgraded in each cluster, one by one, to support Add-Paths. On the RRs Add-Paths is configured to use the Add-N, N=2 path selection algorithm towards upgraded client peers. At this step clients are configured in the receive-only Add-Paths mode. This means that best-external continues to operate as before in the client-to-RR direction. The effect of this step is to ensure that all clients have two paths per prefix for ECMP or fast failover, assuming at least 2 paths are available.

3. The clients are re-configured to use Add-Paths in the transmit direction towards their RR peers. This causes Add-Paths to replace

the best-external behavior. The effect of this step is to free up CPU and memory resources related to the storage of paths that are third best or worse. If a cluster such as the one in Figure 3 had 50 clients, and 10 of these learned an external route for the same prefix, then the RRs in that cluster would need to store up to 12 paths for that prefix. This would be true even if the 2 best overall paths came from another cluster. Contrast this with the use of Add-Paths in the client-to-RR direction. For the same case the route reflectors need only store the 2 paths learned from non-client peers.

5.2. Scalability Considerations

In terms of scalability, we note that advertising multiple paths per prefix requires more memory and state than the current behavior of advertising the best path only. A BGP speaker that does not implement Add-Paths maintains send state information in its prefix data structure per neighbor as a way to determine that the prefix has been advertised to the neighbor. With Add-Paths, this information has to be replicated on a per path basis that needs to be advertised. Mathematically, if "send state" size per prefix is ’s’ bytes, number of neighbors is ’n’, and number of paths being advertised is ’p’, then the current memory requirement for BGP "send state" = n * s bytes; with Add-Paths, it becomes n * s * p bytes. In practice, this value may be reduced with implementation optimizations similar to attribute sharing. Receiving multiple paths per prefix also requires more memory and state since each path is a separate entry in the Adj-RIB-Ins.

5.3. Routing Consistency Considerations

As discussed in previous sections Add-Paths can help routers select more optimal paths and it can help deal with certain route oscillation conditions arising from incomplete knowledge of the available paths. But depending on the path selection algorithm and how it is used Add-Paths is not immune to its own cases of routing inconsistencies. If the BGP routers within an AS do not make consistent routing decisions about how to reach a particular destination, route oscillations may occur and these route oscillations may result in traffic loss.

Optimizing an Add-Paths deployment for scalability may run counter to routing consistency goals, and in these circumstances operators have to decide the correct tradeoff for their particular deployment. For example the Advertise All Paths mode, if applied to many prefixes, is far from ideal from a scalability perspective but it does guarantee routing consistency and correctness. A path selection mode that allows better control over scalability is the Advertise N paths mode, but this is susceptible to routing inconsistency. First, if the N paths do not include the best path from each neighbor AS group then route oscillation cannot be precluded. Second, if the advertising router (e.g. an RR) advertises N paths to peer_n and M paths to peer_m, and N < M, care must be exercised to ensure that all paths advertised to peer_n are included in the paths advertised to peer_m. This can be assured as long as the advertising router has strictly ordered all of its paths.

5.4. Consistency between Advertised Paths and Forwarding Paths

When using Add-Paths, routers may advertise paths that they have not selected as best, and that they are thus not using for traffic forwarding. This is generally not an issue if encapsulation is used in the AS as described in [RFC4364] and all forwarding decisions, including by the tunnel egress router, are based on label information - i.e. if only the ingress router performs an IP FIB lookup. In this situation the dataplane path followed by the packets is the one intended by the ingress router, and corresponds to the control plane path it selected.

On the other hand, if Add-Paths is used in a network without encapsulation, some scenarios can result in forwarding deflection or loops. Such forwarding anomalies already occur without Add-Paths, when the routers on the forwarding path do not have a synchronized view of the best path. They will deflect the traffic to their own local view of the best path, and, when multiple deflections occur, forwarding loops can occur. With Add-Paths, the issue can be exacerbated due to routers advertising non-best paths. As discussed above, encapsulation can help with this issue, but only to the extent that it allows downstream routers to forward without an IP FIB lookup.

A first example of such issue is when the Local-Pref of non-primary paths received over iBGP sessions is modified. The ingress router may thus select as best a path non-preferred by the egress, and the egress router will thus deflect the traffic.

Another example is when the best path is selected based on tie-breaking rule. When the ingress and the egress base their path selection on the router-id of the neighbor that advertised the path to them, the result may be different for each of them. This specific issue is described and solved in [draft-pmohapat].

In general, if the network forwards on a hop-by-hop basis and does not make use of encapsulation, it is necessary to advertise the best path. The second path that is advertised should be the second best path using one of the path selection modes described previously. Additional paths are discretionary with the presumption that they can be forwarded on a hop-by-hop basis.

Similarly, if the network uses encapsulation, the best path should be advertised for consistency, the second best path should be advertised for fast routing convergence. All further paths and their choice for selection are completely discretionary; the destination is presumed to be reachable via encapsulation.

5.5. Interactions with Route Filtering

As noted in the previous section, modification of advertised paths may lead to inconsistent route selection. This is true even when the Add-Paths feature is not in use. Similarly, the use of route filtering, when used carelessly for iBGP, may result in inconsistent route selection in an AS with the possibility of introducing forwarding loops.

The Add-Paths feature has additional considerations for route filtering since the receiver of multiple paths is unable to determine by inspection of the received NLRI which path corresponds to the sender’s active path for the prefix. The sender SHOULD send the best path when sending multiple paths for a destination. The receiver must take care when rejecting destinations to not discard the best path but permit alternate paths. A failure on either the part of the sender or receiver to distribute/receive the best path may result in inconsistent route selection.

An implementation MAY support the ability to suppress advertisement of all alternate paths when the export policy would otherwise suppress the best path.

5.6. Routing Churn

As noted in section 3.3 using Add-Paths between IBGP peers can help to reduce routing churn with EBGP peers. This benefit does however come at the cost of potentially increased churn between the IBGP Add-Paths peers. In a non Add-Paths deployment a change in the preference order of non-best paths requires no updates to be sent to peers. But when a router has Add-Paths peers changes in non-best path preference may no longer be invisible and increased route churn may be observable. Choosing the right path selection mode and parameters - for example not setting N unnecessarily large in the Add-N mode, is important to minimizing this additional churn.

6. Security Considerations

TBD

7. Acknowledgments

This document was prepared using 2-Word-v2.0.template.dot.

8. Contributors

Virginie Van den Schrieck
Email: v.vandenschrieck@gmail.com

Rohit Gupta
Apple Inc
Email: rxgupta@apple.com

9. IANA Considerations

TBD

10. References

10.1. Normative References

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

10.2. Informative References

[Add-Paths] Walton, D., Retana, A., Chen E., Scudder J., "Advertisement of Multiple Paths in BGP", draft-ietf-idr-add-paths-07, June 17, 2012.

[draft-pmohapat] Mohapatra, P., Fernando, R., Filsfils, C., and R. Raszuk, "Fast Connectivity Restoration Using BGP Add-path", draft-pmohapat-idr-fast-conn-restore-02.txt, Oct 3, 2011.

[oscillation] Walton, D., Retana, A., Chen, E., Scudder, J., "BGP Persistent Route Oscillation Solutions", draft-walton-bgp-route-oscillation-stop-06.txt, June 14, 2012.

[Basu-ibgp-osc] Basu, A., Ong, C., Rasala, A., Sheperd, B., and G. Wilfong, "Route oscillations in iBGP with Route Reflection", Sigcomm 2002.

[BGP-ORR] Raszuk, R., Cassar, C., Aman, E., Decraene, B., "BGP Optimal Route Reflection", draft-raszuk-bgp-optimal-route-reflection-01, March 11, 2011.

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

Appendix A. Other Path Selection Modes

A.1. Advertise Neighbor-AS Group Best Path

[walton-osc] proposes that a router groups its paths based on the neighbor AS from which it was learned, and to advertise the best path in each of those groups.

The control plane stress induced by this solution is the computation of the per-neighbor path group, and the application of the decision process to each of them. The Control-Plane load is bounded by the number of neighboring ASes advertising a prefix, which cannot be known a-priori.

Path optimality and backup path optimality are not guaranteed, as the paths advertised are not all the AS-wide preferred paths. Backup path availability is not guaranteed. Indeed, if only one AS advertises this prefix, even on multiple eBGP sessions, only one of the paths may be selected and advertised.

A.2. Best LocPref/Second LocPref

This selection method consists in grouping the paths by Local Preference. A router sends to its peers all paths with the highest Local Preference. If there is only a single path with the highest Local Preference, it also sends all paths with the second best Local Preference.

This method ensures that all routers know all paths with the best local preference. As local preference are often related to the type of peering of the peer the path comes from, this ensures that in case of failure, routers have a backup path of equivalent quality. This prevents for example that a router switches temporarily on a peer path while an alternate path from a customer is available but hidden at the border of the AS. Such a situation could result in a temporary withdrawal of the prefix on some eBGP sessions when the router selects the path via the peer.

The advertisement of the Second Local Preference occurs when there is no alternate path with the same quality as the best path. This way, fast convergence is still ensured. Backup path is optimal, as it has the second AS-Wide preference, which becomes the AS-wide best preference upon failure of the primary one.

Sending all the paths with a given Local Preference also has a positive impact on routing optimality. Indeed, this allows border routers to have an increased path visibility and to choose their best path based on their own criteria.

The computational cost of this solution is reduced when there are several paths with the best local preference. In this case, it is sufficient to stop the decision process after the first rule to have the set of paths to be advertised. When it is necessary to advertise the paths with second local-preference, the additional cost is to apply a second time the first rule of the decision process, which is still reasonable. The memory cost depends on the number of paths with the best local preference.

A.3. Advertise Paths at decisive step -1

When the goal is to provide fast recovery by advertising candidate post-reconvergence paths, one can choose to stop the decision process just before the step where only one path remains. If the decision process comes to IGP tie-break, all remaining paths are advertised. This way, routers advertise as many paths as possible with a quality as similar as possible.

This path selection is an intermediary solution between the two preceding ones. Here, instead of stopping the decision process at the local preference step or the IGP step, we stop it before the rule that removes the best potential backup paths. This way, we minimize the number of paths to advertise while guaranteeing the presence of a backup path. Primary and backup path optimality is ensured, as all paths with the same AS-wide preference as the best paths are included in the set of paths advertised.

Authors' Addresses

Jim Uttaro
AT&T
200 S. Laurel Avenue
Middletown, NJ 07748 USA
Email: uttaro@att.com

Pierre Francois
Institute IMDEA Networks
Avda. del Mar Mediterraneo, 22
Leganese 28918
ES
Email: pierre.francois@imdea.org

Pradosh Mohapatra
Cumulus Networks
pmohapat@cumulusnetworks.com

Roberto Fragassi
Alcatel-Lucent
600 Mountain Avenue
Murray Hill, New Jersey
Email: roberto.fragassi@alcatel-lucent.com

Adam Simpson
Alcatel-Lucent
600 March Road
Ottawa, Ontario K2K 2E6
Canada
Email: adam.simpson@alcatel-lucent.com

Keyur Patel
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134 USA
Email: keyupate@cisco.com

Jeffrey Haas
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
USA
Email: jhaas@juniper.net