GROW Working Group   R. Raszuk, Ed.
Internet-Draft   NTT MCL
Intended status: Informational   R. Fernando
Expires: January 2, 2013   K. Patel
                                                       Cisco Systems
                                                        D. McPherson
                                                            Verisign
                                                           K. Kumaki
                                                     KDDI Corporation
                                                           July 2012

                    Distribution of diverse BGP paths.
                  draft-ietf-grow-diverse-bgp-path-dist-08

Abstract

   The BGP4 protocol specifies the selection and propagation of a single
   best path for each prefix. As defined and widely deployed today BGP
   has no mechanisms to distribute alternate paths which are not
   considered best path between its speakers. This behaviour results in
   number of disadvantages for new applications and services.

   The main objective of this document is to observe that by simply
   adding new session between route reflector and it's client Nth best
   path can be distributed. Document also compares existing solutions
   and proposed ideas that enable distribution of more paths than just
   the best path.

   This proposal does not specify any changes to the BGP protocol
   definition. It does not require software upgrade of provider edge
   routers acting as route reflector clients.

Status of this Memo

This Internet-Draft will expire on January 2, 2013.

Copyright Notice

Table of Contents

1. Introduction

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

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

This proposal does not specify any changes to the BGP protocol
definition. It does not require upgrades to provider edge or core
routers nor does it need network wide upgrades. The only upgrade
required is the new functionality on the new or current route
reflectors.

2. History

The need to disseminate more paths than just the best path is
primarily driven by three requirements. First is the problem of BGP
oscillations [RFC3345]. The second is the desire for faster
reachability restoration in the event of network link or network
element's failure. Third requirement is to enhance BGP load
balancing capabilities. Those reasons have lead to the proposal of
BGP add-paths [I-D.ietf-idr-add-paths].

2.1. BGP Add-Paths Proposal

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

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

This new requirement comes with its own memory and processing cost.

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

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

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

While it needs to be clearly acknowledged that the add-path mechanism
provides the most general way to address the problem of distributing
many paths between BGP speakers, this document provides a much easier
to deploy solution that requires no modification to the BGP protocol
where only a few additional paths may be required. The alternative
method presented is capable of addressing critical service provider
requirements for disseminating more than a single path across an AS

with a significantly lower deployment cost. That in the light of
number of general network scaling concerns documented in RFC4984
[RFC4271] "Report from the IAB Workshop on Routing and Addressing"
may provide a significant advantage.

3. Goals

The proposal described in this document is not intended to compete
with add-paths. It provides an interim solution until the
standardization and implementation of add-paths and until support for
that function can be deployed across the network.

It is presented to network operators as a possible choice and
provides those operators who need additional paths today an
alternative from the need to transition to a full mesh. The Nth best
path describes a set of N paths with different BGP next hops with no
implication of ordering or preference among said N paths.

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

4. Multi plane route reflection

The idea contained in the proposal assumes the use of route
reflection within the network.

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

```
                            ASBR3
                             ***

                            *     *
          +------------*    *-----------+
          | AS1           *     *       |
          |                 ***         |
          |                             |
          |                             |
          |                             |
          | RR1            ***       RR2 |
          | ***           *     *    *** |
          |*     *        * P *       *     *|
          |*     *        *     *       *     *|
          |  ***           ***           ***  |
          |                             |
          |               IBGP          |
          |                             |
          |                             |
          |      ***              ***    |
          |     *     *          *     *  |
          +-----*    *---------*    *----+
                *     *          *     *
                 ***              ***
              ASBR1            ASBR2
                       EBGP
```
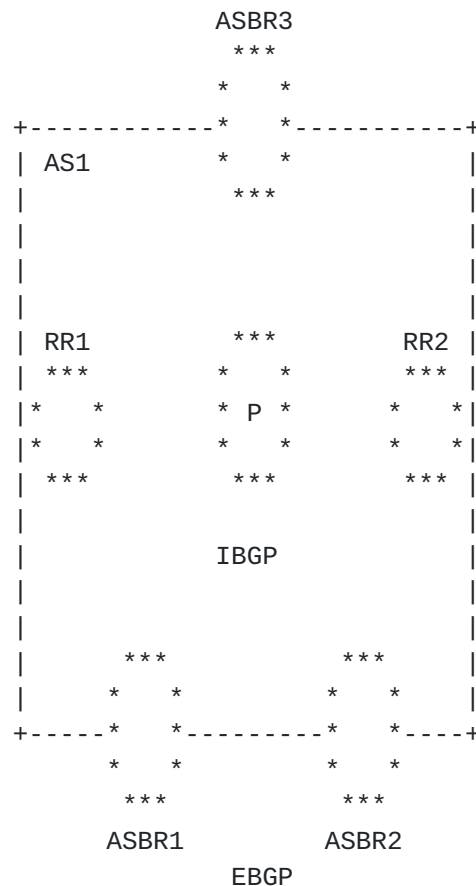
Figure1: Simple route reflection

Abbreviations used: RR - Route Reflector, P - Core router.

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

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

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

choose the same best path or a different one. In such a case
ASBR3 may receive either the same path or different paths from
each of the route reflectors.

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

There is no requirement of path ordering. The "Nth best path" really
describes set of N paths with different bgp next hops.

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

As diverse-path functionality may be enabled on a per peer basis one
of the deployment model can be realized to continue advertisement of
overall best path from both route reflectors while in addition new
session can be provisioned to get additional path. That will allow
the non interrupted use of best path even if one of the RRs goes down
provided that the overall best path is still a valid one.

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

Diverse path route reflectors need the new ability to calculate and
propagate the Nth best path instead of the overall best path. An

implementation is encouraged to enable this new functionality on a
per neighbor basis.

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

To address the problem of external paths not being available to route
reflectors due to local preference or MED factors it is recommended
that ASBRs enable the best-external functionality in order to always
inject their external paths to the route reflectors.

4.1. Co-located best and backup path RRs

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

```
                             ASBR3
                              ***
                             *    *
          +------------*    *-----------+
          | AS1           *    *        |
          |                ***          |
          |                             |
          | RR1                    RR2  |
          |  ***                  ***   |
          |*    *        ***      *    *|
          |*    *       *    *    *    *|
          |  ***       * P *      ***   |
          |*    *       *    *    *    *|
          |*    *        ***      *    *|
          |  ***                  ***   |
          | RR1'         IBGP      RR2' |
          |                             |
          |                             |
          |      ***           ***      |
          |    *    *        *    *     |
          +-----*    *---------*    *----+
               *    *          *    *
                ***            ***
              ASBR1          ASBR2


                             EBGP
```
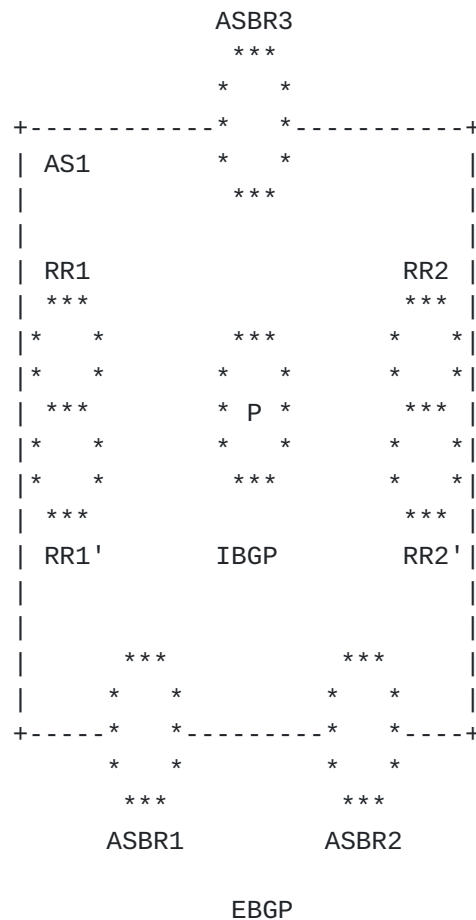
Figure2: Co-located 2nd best RR plane

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

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

   2. Enabling best-external on ASBRs

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

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

   The expected behaviour is that under any BGP condition the ASBR3 and
   P routers will receive both paths P1 and P2 for destination D. The

availability of both paths will allow them to implement a number of
new services as listed in the applications section below.

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

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

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

4.2. Randomly located best and backup path RRs

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

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

```
                               ASBR3
                                ***
                               *   *
              +------------*   *-----------+
              | AS1         *   *           |
              | IBGP         ***            |
              |                             |
              |             ***             |
              |            *   *            |
              | RR1       * P *        RR2 |
              | ***        *   *        ***  |
              |*   *        ***        *   *|
              |*   *                   *   *|
              | ***         RR'         ***  |
              |             ***             |
              |            *   *            |
              |            *   *            |
              |             ***             |
              |     ***              ***    |
              |    *   *            *   *    |
              +-----*   *----------*   *----+
                    *   *          *   *
                     ***            ***
                    ASBR1          ASBR2


                              EBGP
```
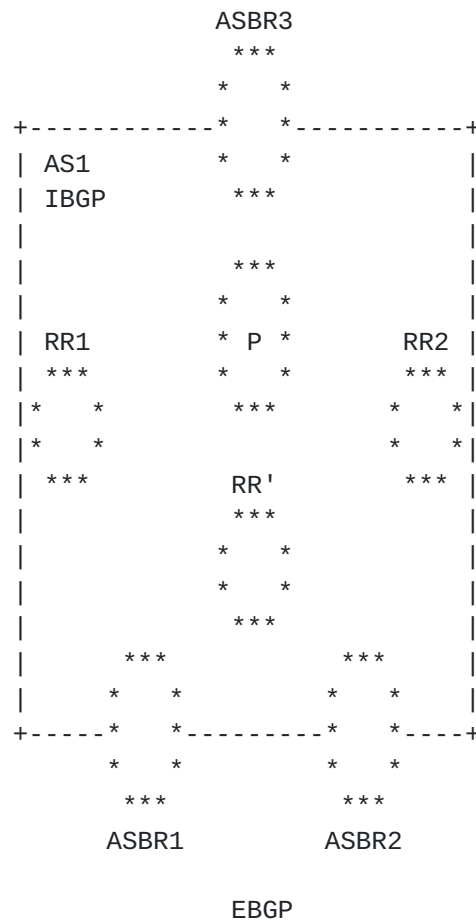
Figure3: Experimental deployment of 2nd best RR

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

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

2. Enabling best-external on ASBRs

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

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

5. If required fully meshing newly added RRs' with the all other
    reflectors in both planes. That condition does not apply if the

      newly added RR'(s) already have peering to all ASBRs/PEs.

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

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

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

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

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

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

   It is recommended that an implementation will advertise overall best
   path over Nth diverse-path session if there is no other BGP path with
   different next hop present. That is equivalent to today's case where
   client is connected to more than one RR.

4.3. Multi plane route servers for Internet Exchanges

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

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

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

5. Discussion on current models of IBGP route distribution

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

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

   1. Full mesh

   2. Confederations

   3. Route reflectors

5.1. Full Mesh

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

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

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

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

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

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

5.2. Confederations

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

Within each sub-AS BGP speakers are fully meshed and as discussed in

section 2.1 all full mesh characteristics (number of TCP sessions,
provisioning and potential concern over number of paths still apply
in the sub-AS scale).

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

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

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

5.3. Route reflectors

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

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

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

6. Deployment considerations

Distribution of diverse BGP paths proposal allows the dissemination
of more paths than just the best-path to route reflector or route
server clients of today's BGP4 implementations. As deployment
recommendation it needs to be mentioned that fast connectivity
restoration as well as majority of intra-domain BGP level load
balancing needs can be accommodated with only two paths (overall best
as well as second best). Therefor as deployment recommendation this
document suggests use of N=2 with diverse-path.

From the client's point of view receiving additional paths via

separate IBGP sessions terminated at the new router reflector plane
is functionally equivalent to constructing a full mesh peering
without the problems that such a full mesh would come with set of
problems as discussed in earlier section.

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

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

The use of terminology of "planes" in this document is more of a
conceptual nature. In practice all paths are still kept in the
single table where normal best path is calculated. That means that
tools like looking glass should not observe any changes nor impact
when diverse-path has been enabled.

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

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

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

3. ASBRs with multiple external paths.

This section focuses on discussion of the 3rd above case in more
detail. This describes the scenario of a single ASBR connected to
multiple EBGP peers. In practice this peering scenario is quite
common. It is mostly due to the geographic location of EBGP peers
and the diversity of those peers (for example peering to multiple
tier 1 ISPs etc...). It is not designed for failure recovery
scenarios as single failure of the ASBR would simultaneously result
in loss of connectivity to all of the peers. In most medium and
large geographically distributed networks there is always another

ASBR or multiple ASBRs providing peering backups, typically in other geographically diverse locations in the network.

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

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

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

7. Summary of benefits

Distribution of diverse BGP paths proposal provides the following benefits when compared to the alternatives:

1. No modifications to BGP4 protocol.

2. No requirement for upgrades to edge and core routers (as required in draft-ietf-idr-add-paths-07). Backward compatible with the existing BGP deployments.

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

current route reflector(s).

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

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

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

8. Applications

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

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

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

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

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

9. Security considerations

The new mechanism for diverse BGP path dissemination proposed in this
document does not introduce any new security concerns as compared to

base BGP4 specification [RFC4271] especially when compared against
full iBGP mesh topology.

In addition authors observe that all BGP security issues as described
in [RFC4272] do apply to the additional BGP session or sessions as
recommended by this specification. Therefor all recommended
mitigation techniques to BGP security are applicable here.

10. IANA Considerations

Following [RFC5226] authors declare that the new mechanism for
diverse BGP path dissemination does not require any new allocations
from IANA.

11. Contributors

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

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

Satish Mynam
Juniper Networks
1194 N. Mathilda Ave
Sunnyvale, CA 94089
US
Email: smynam@juniper.net

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

12. Acknowledgments

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

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

13. References

13.1. Normative References

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

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

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

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

13.2. Informative References

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

[I-D.ietf-idr-add-paths]
           Walton, D., Chen, E., Retana, A., and J. Scudder,
           "Advertisement of Multiple Paths in BGP",
           draft-ietf-idr-add-paths-07 (work in progress), June 2012.

[I-D.ietf-idr-best-external]
           Marques, P., Fernando, R., Chen, E., Mohapatra, P., and H.

            Gredler, "Advertisement of the best external route in
            BGP", draft-ietf-idr-best-external-05 (work in progress),
            January 2012.

   [I-D.pmohapat-idr-fast-conn-restore]
            Mohapatra, P., Fernando, R., Filsfils, C., and R. Raszuk,
            "Fast Connectivity Restoration Using BGP Add-path",
            draft-pmohapat-idr-fast-conn-restore-02 (work in
            progress), October 2011.

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

   [RFC3345] McPherson, D., Gill, V., Walton, D., and A. Retana,
            "Border Gateway Protocol (BGP) Persistent Route
            Oscillation Condition", RFC 3345, August 2002.

   [RFC4272] Murphy, S., "BGP Security Vulnerabilities Analysis",
            RFC 4272, January 2006.

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

Authors' Addresses

   Robert Raszuk (editor)
   NTT MCL
   101 S Ellsworth Avenue Suite 350
   San Mateo, CA 94401
   US

   Email: robert@raszuk.net

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

   Email: rex@cisco.com

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

Email: keyupate@cisco.com

Danny McPherson
Verisign
21345 Ridgetop Circle
Dulles, VA 20166
US

Email: dmcpherson@verisign.com

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

Email: ke-kumaki@kddi.com