Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 11, 2011

P. Mohapatra
R. Fernando
C. Filsfils
R. Raszuk
Cisco Systems

March 10, 2011

Fast Connectivity Restoration Using BGP Add-path
draft-pmohapat-idr-fast-conn-restore-01

Abstract

A BGP route defines an association of an address prefix with an "exit point" from the current Autonomous System (AS). If the exit point becomes unreachable due to a failure, the route becomes invalid. This usually triggers an exchange of BGP control messages after which a new BGP route for the given prefix is installed. However, connectivity can be restored more quickly if the router maintains precomputed BGP backup routes. It can then switch to a backup route immediately upon learning that an exit point is unreachable, without needing to wait for the BGP control messages exchange. This document specifies the procedures to be used by BGP to maintain and distribute the precomputed backup routes. Maintaining these additional routes is also useful in promoting load balancing, performing maintenance without causing traffic loss, and in reducing churn in the BGP control plane.

Table of Contents

1. Introduction

Within an autonomous system, the availability of multiple routes to a given destination, where each of the routes has a different "exit point" from the local AS provides the following benefits:

- Fault tolerance: Knowledge of multiple "exit points" leads to reduction in restoration time after failure. For instance, a border router on receiving multiple paths to the same destination could decide to precompute a backup path and have it ready so that when the primary path becomes invalid, it could use the backup to quickly restore connectivity. Currently the restoration time is dependent on BGP protocol re-convergence that includes a set of withdraw and advertisement messages in the network before a new best path can be learnt.

- Load balancing: The availability of multiple paths to reach the same destination enables load balancing of traffic provided the paths for the given destination satisfy certain constraints.

- Churn reduction: The advertisement of multiple routes, in certain scenarios (Section 8.3.2), could lead to less churn in the network upon a failure, since the presence of multiple paths helps contain the failure to the local AS where the failure occurs.

- Graceful maintenance: The availability of alternate exit points allows one to bring down a router for maintenance without causing significant traffic loss.

Unfortunately, the border routers in an AS do not receive multiple paths for all prefixes. The reason is three-fold:

- The current BGP specification [RFC4271] specifies routers to advertise only the best path for a destination to speakers. The availability of multiple paths requires simultaneous distribution of multiple routes for a given prefix by a BGP speaker. We refer to this property of the network as "path diversity".

- When a router selects an IBGP learnt path as best, it does not announce any path for that prefix to IBGP though it may have EBGP learnt paths available. This loss of information leads to added churn and increases convergence time if the preferred path goes away. A mechanism to advertise the best-external path to IBGP is proposed in [I-D.ietf-idr-best-external].

- Most service providers deploy one of the scaling techniques like route reflectors [RFC4456] or confederations [RFC5065] inside the AS and avoid iBGP full mesh. Thus even when multiple paths exist,

the aggregation points (route reflectors or confederation border routers) advertise only the best path (as per the BGP base protocol).

As an effect of this behavior, the ingress border routers to an AS do not receive additional paths necessary to provide the benefits cited above: e.g. perform a local recovery during network failures or achieve load balancing in steady state across multiple exit points.

The mechanism to extend BGP to allow a given BGP speaker to advertise multiple paths simultaneously for a destination is defined in [I-D.ietf-idr-add-paths]. The current draft describes the use of this generic technique and certain additional procedures and implementation guidelines to enable the above applications.

More specifically, this document describes extensions to BGP decision process to select backup paths in a manner that ensures the important property of consistent route selection within an AS. It also introduces a new BGP attribute, Edge_Discriminator, that border routers should use to advertise multiple EBGP learnt paths for a given destination. To aid with better description of the applications, the draft illustrates certain use case scenarios for each.

One implication of multiple path advertisement is the associated cost, namely the performance overhead of processing and memory overhead of storing additional paths. It is anticipated that the benefits listed above outweigh the cost in most scenarios. Be that as it may, it is also expected that there will be configuration knobs provided to limit the number of additional paths propagated within an AS.

1.1. Requirements Language

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. Basic Idea

This document proposes two main additions to the BGP procedures:

1. The decision process is modified to determine backup paths along with the best path selection when multiple paths for a destination are available.

2. In addition to using these backup paths for fast connectivity restoration locally, BGP speakers also advertise these paths to IBGP to increase the overall path diversity.

As alluded to in Section 1, BGP speakers that are the aggregation points (router reflectors or confederation border routers) need to announce backup paths to increase the path diversity at the ingress routers of an IBGP network (see Figure 2). It may also be useful, in certain cases, for the border routers to advertise multiple paths received via EBGP for a destination when it is redundantly connected and is transparently passing the NEXT_HOP field unchanged instead of setting it to self (see Figure 4). To this end, the draft defines a new attribute, Edge_Discriminator, that the border routers should advertise to ensure path selection consistency.

The following sections elaborate on these points.

3. Design Considerations

3.1. Ensuring Loop-Free Path Selection in an AS

It is critical that BGP speakers within an AS have an eventual consistent routing view of destinations and do not make conflicting decisions regarding best path selection that would otherwise cause forwarding loops. The current BGP protocol ensures this property by defining a decision process that takes the attributes of paths as input and determines a degree of preference of the paths by applying a constant function. A consistent view of attributes is disseminated through IBGP. Thus each BGP speaker within the AS determines the same degree of preference of the paths after applying the constant function independently. (The one exception is where IGP metric plays the tie breaking role. In this case, different routers may choose different next hops that are closer to them; but loop freedom is guaranteed.).

When the above mechanism is extended to select backup paths for the applications cited in this document, it is equally important to maintain the same consistency property for the backup paths, i.e. there should be no loops created when routers use the backup path in forwarding. The rest of the document goes into the details of this for various scenarios.

3.1.1. Border routers announcing single path

In scenarios where all border routers advertise a single external path (their best path or best-external path) into IBGP, a consistent

routing view of best path and backup paths can be created across the AS with the current BGP selection rules.

3.1.2. Border routers announcing multiple paths

There are scenarios where border routers need to advertise the best and backup EBGP learnt paths with NEXT_HOP unchanged to IBGP. If the border router sets next hop to self, the paths become indistinguishable and hence advertisement of only the best path is sufficient. An example scenario is depicted in Figure 4.

By using the add-path ([I-D.ietf-idr-add-paths]) extensions, the border routers could advertise multiple such EBGP-learnt paths. But doing so can potentially create an inconsistency between the paths that the sending and receiving routers select for forwarding. In other words, the routers in the IBGP mesh can make independent and separate decisions on the route selection since some of the values that play a role in the tie breaking steps of the decision process at the sender are not available to the rest of the BGP speakers of the AS. These are mainly (1) the interior cost, i.e. the metric to reach the external next hop, (2) BGP identifier of the peer, (3) the peer IP address. Due to this reduction in information, there can be inconsistency in the routing view within an AS.

Additionally, [RFC5004] proposes an extension to avoid best path transitions at the border router between external paths based on a temporal order of receiving the paths. This can also create an inconsistency across the BGP speakers in the path selection.

This document proposes two modifications to ensure consistency:

a. Border routers SHOULD not apply the modification to the selection rules as proposed in [RFC5004] to avoid best path transitions for parallel EBGP connection scenario where the border router wishes to transitively transmit the NEXT_HOP value unchanged.

b. To overcome the "information reduction" problem described above, the document specifies an attribute called "Edge_Discriminator attribute" that encodes the properties of each path advertised that would otherwise not be included using the normal attributes in a BGP UPDATE message (see Section 4).

3.1.3. Confederations

When an AS employs confederations and the confederation border routers advertise multiple paths, there is no way to distinguish the originator (the actual egress border router originating the prefix to the AS). To ensure consistent path selection, the confederation

border routers should create the ORIGINATOR_ID attribute as described in [RFC4456] that carries the BGP identifier of the originator of the route to the local AS.

3.2. Keeping Path Attributes Independent of Decision Process

In addition to providing consistency in path selection, the solution should satisfy the following important property: the attributes associated with a particular path should be invariant when a different path is advertised or withdrawn. Other things being equal, it is best to avoid the potential churn introduced by the feedback loops that would occur if path attributes were changed at the sender as a result of running the decision process. Thus we do not use any attributes with semantics like "this is my second best path", "this is my third best path", etc. This requirement precludes use of marking or other means of indicating path ordering from sender's perspective since a change in the ordering requires re-advertising most of the paths.

4. Edge_Discriminator attribute

Edge_Discriminator attribute is an optional non-transitive attribute that is composed of a set of Type-Length-Value (TLVs) encodings. The type code of the attribute is to be assigned by IANA. Each TLV contains an attribute of the path from the border router that is not otherwise sent as part of the UPDATE message. The TLV is structured as follows:

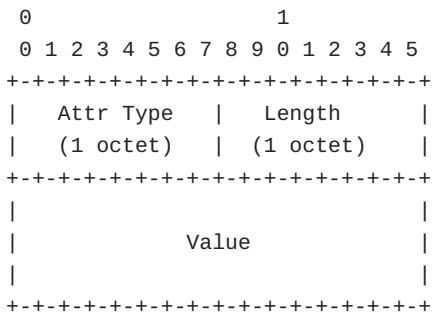

Figure 1: Edge_Discriminator attribute format

a. Attr Type (1 octet): It identifies the type of the attribute that is encoded by the border router. Unknown types are to be ignored and skipped upon receipt. This document defines the following types:

* Interior Cost: Attr Type = 1

* peer BGP Identifier: Attr Type = 2

* IPv4 Peer Address: Attr Type = 3

* IPv6 Peer Address: Attr Type = 4

b. Length (1 octet): the total number of octets of the Value field.

c. Value (variable): The value field encodes the attribute of the corresponding type. For "Interior Cost" type, it encodes the four octet metric value. For "BGP Identifier" type, it encodes the four-octet router identifier of the neighbor for the path. For "IPv4 Peering Address" type, the 4 byte BGP IPv4 peering address is encoded. For "IPv6 Peering Address" type, the 16 byte BGP IPv6 peering address is encoded.

A brief description of how a BGP speaker constructs the attribute is provided in Section 6.

5. Calculation of Best and Backup Paths

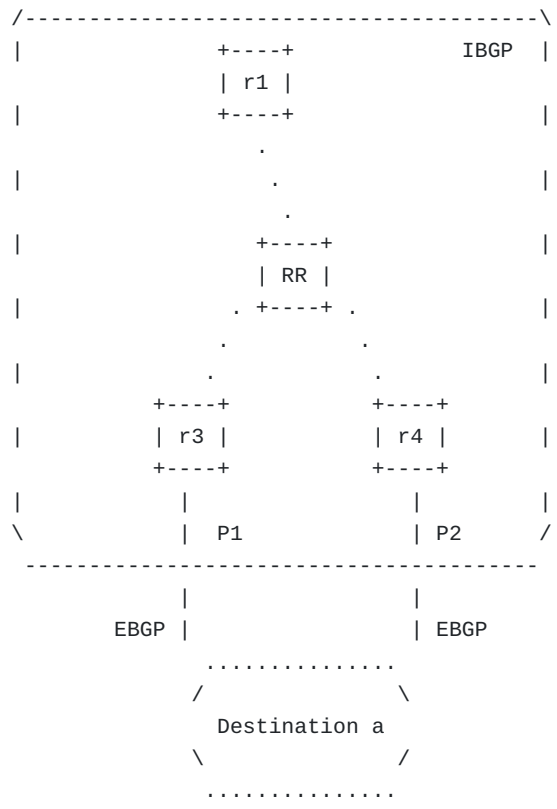

Figure 2: Basic RR topology

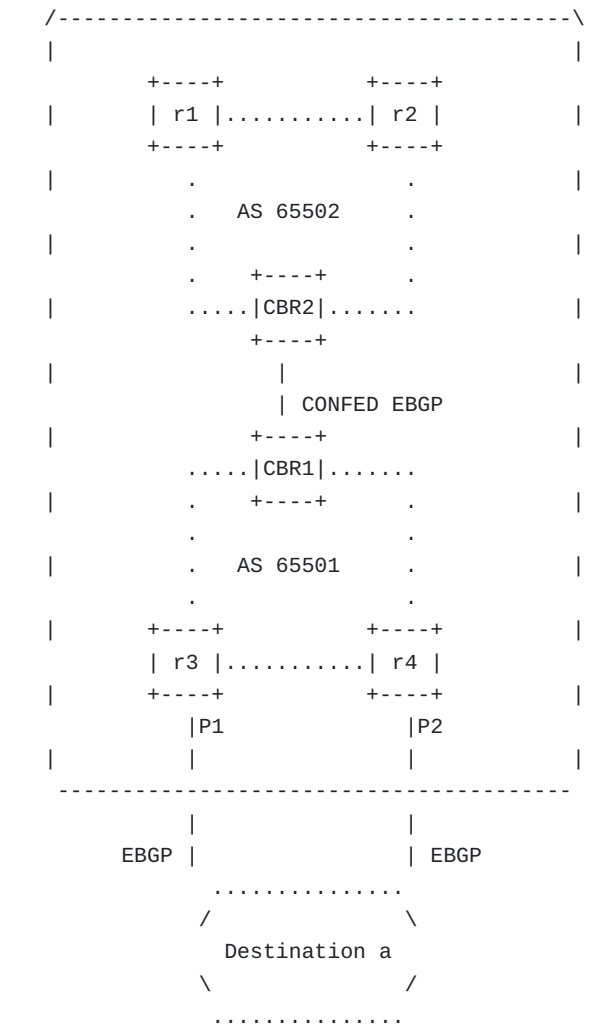

Figure 3: Confederation topology

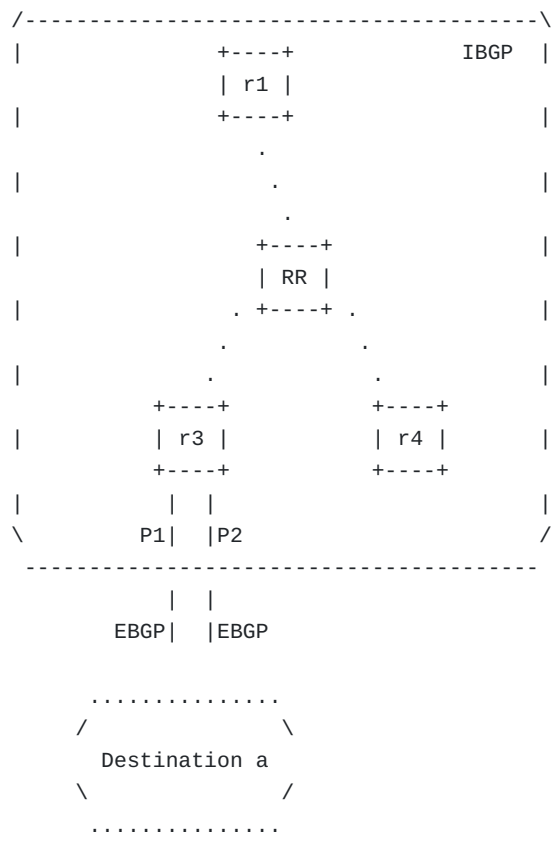

Figure 4: Border router with parallel eBGP links

The decision process as described in [RFC4271] is followed to determine the overall best path for a destination. In addition, the following rule SHOULD be inserted into the tie breaking rules of the BGP decision process after step f) (Sect. 9.1.2.2: [RFC4271]) and after the CLUSTER_LIST length check step (Sect. 9: [RFC4456]): a BGP speaker SHOULD apply the tie breaking steps (steps (e), (f), and (g) as defined in [RFC4271]) with the values encoded in the Edge_Discriminator attribute.

Note that the above step effectively compares multiple paths that are advertised by the same egress border router (since the BGP Identifier comparison step earlier would have eliminated paths from different egress border routers).

Consider the network in Figure 4. r3 learns two paths P1 and P2 for destination a and wishes to advertise both to the iBGP mesh with NEXT_HOP value unchanged. We need to ensure that both r3 and the other ingress routers in the network (r1, r4) make a consistent route

selection for the best and the backup paths for destination a. The current tie breaking rules [step f) comparison of router ID or ORIGINATOR_ID and step g) comparision of peering ID] are insufficient since at the ingress routers, both the paths will be received with same values for each of the above parameters. Hence an additional tie breaking rule comparing the original values that the border router itself used to tie break the paths is required.

Once the best path is chosen, eliminate that path and all paths that have the same BGP Identifier or NEXT_HOP as the choosen best path. Note that as specified in [RFC4456], if the path carries the ORIGINATOR_ID attribute, that should be treated as the BGP Identifier. Then rerun the best path procedure to choose the backup path. The Tie Breaking rules of the BGP decision process for second best path selection are also modified as described above.

This mechanism can be recursively used to calculate multiple backup paths if desired.

6. Advertising Multiple Paths

The technique outlined in [I-D.ietf-idr-add-paths] is used to advertise best and backup paths selected with the rules described in Section 5. For the purposes of the applications cited in this document, the "Path Identifier" is always treated as an opaque value with no semantics.

When an egress border router chooses to advertise multiple paths learnt via EBGP to IBGP, it SHOULD include the Edge_Discriminator attribute as defined in Section 4 for each of the paths. The attribute is constructed by encoding the following properties of the path in TLV format:

- The interior cost to reach the NEXT_HOP of the path, encoded with type 1.
- The BGP identifier of the EBGP peer from which it received the path, encoded with type 2.
- The peer address of the EBGP peer from which it received the path, encoded either with type 3 or 4.

7. Deployment Considerations

To ensure consistency in path selection process across all the routers in an AS, the deployment considerations from the individual

scaling technology employed in the network should be inherited/ applied. For example, as specified in [RFC4456], the intra-cluster IGP metric values should be better than the inter-cluster IGP metric values. Similar considerations as specified in [RFC5065] should be designed.

8. Applications

8.1. Fast Connectivity Restoration

Consider the network in Figure 2. All 4 routers indicated are part of a single AS. r3 and r4 are the border routers. Suppose r3 and r4 receive paths P1 and P2 for the same prefix. Also assume that P1 is the preferred exit.

There are two scenarios to consider:

- case 1: P1 is the preferred exit for all routers within the AS (including r4). In this case, if r4 follows [RFC4271], r4 withdraws P2 from the IBGP cloud.

- case 2: P2 is preferred exit by r4. In this case, if RR follows [RFC4271], RR gets both paths, chooses one and sends it to r1.

In both the cases above, 'r1' holds only a single path and only after a failure that makes P1 unavailable, it receives the alternate path (P2).

However, if both paths were available to 'r1' and all other border routers in the network, then they could precompute backup paths and keep them ready to restore connectivity upon being notified of a failure. The failure notification could be triggered due to a link failure between 'r3' and its EBGP neighbor. This failure could be propagated to other routers in r3's AS either via IGP or BGP, resulting in invalidating on all these routers their primary paths that were advertised by that neighbor to r3 (and that r3 subsequently re-advertised into IBGP). Once these paths are invalidated, all these routers could switch to the precomputed backup paths, without waiting for any additional BGP advertisements.

8.2. Load Balancing

In the above network, not only can the additional path be used as a standby best, but can also be used in steady state to load balance traffic across the two exit points.

8.3. Churn Reduction

There are two aspects to reducing churn - Inter-domain and Intra-domain.

8.3.1. Inter-domain Churn Reduction

Consider the network diagram in Figure 5.

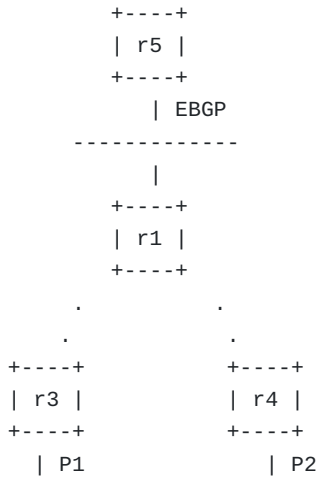

Figure 5

'r5' is an EBGP peer of 'r1'. Today, if path P1 goes away, due to the non-availability of other paths, 'r1' sends a withdraw to r5 thus triggering a churn in the Internet. This could be significant if there are multiple prefixes involved. On the other hand, if r1 had an alternate path (with identical attributes), then this churn could be entirely avoided by r1 performing a local repair.

8.3.2. Intra-Domain Churn Reduction

Since advertising multiple paths in general increases the path diversity at the border routers, some of the control plane churn in terms of a stream of advertisements, withdraws, and re-advertisements can be reduced, thus improving the stability of the network.

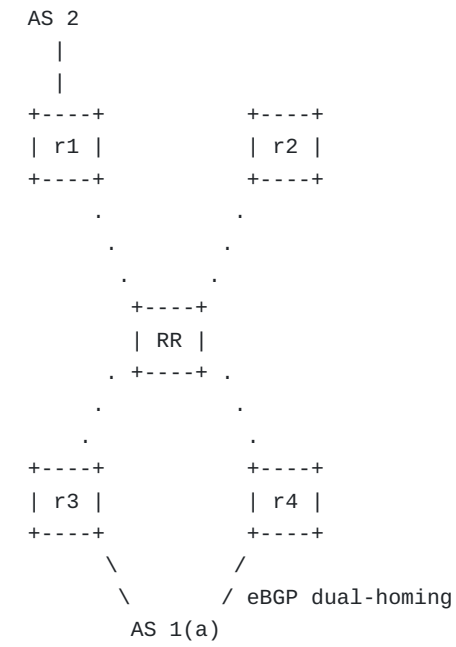

Figure 6

Assuming router r3's path is the best path in the AS, RR advertises the corresponding route information to the iBGP network. If r3 goes down (or the peering link [r3, AS1] fails and r3 didn't change the next hop to itself), the following will be sequence of updates from router r1 to AS 2:

- Initial update for all prefixes when r1 chooses best path,
- Withdraws for all prefixes when r1 detects failure,
- Re-advertisement of all prefixes when the RR chooses router r4's path as the new best path and advertises to r1.

With both the paths advertised and received on router r1, the sequence of updates reduces to:

- Initial update for all prefixes when r1 chooses best path,
- Re-advertisement of all prefixes when r1 detects failure and chooses router r4's path as the new best path

8.4. Graceful Maintenance

[I-D.ietf-grow-bgp-graceful-shutdown-requirements] defines requirements for graceful maintenance of routers in a service provider network. Current BGP operations treat this as a sudden link or node failure and try to reconverge that can take in the order of seconds or minutes.

With the procedures defined in this document, since alternate paths are available at the ingress routers, taking down egress routers from the network does not result in a network-wide reconvergence event.

9. Acknowledgements

The authors would like to thank Enke Chen for the many discussions resulting in this work. In addition, the authors would also like to acknowledge valuable review and suggestions from Eric Rosen, Yakov Rekhter, and John Scudder on this document.

10. IANA Considerations

This document defines a new BGP optional non-transitive attribute type, called Edge_Discriminator attribute. The attribute type is to be assigned by IANA.

This document introduces Attr TLVs within the above attribute. The type space for these should be set up by IANA as a registry of 1-octet attr types. These should be assigned on a first-come-first-serve basis.

This document defines the following attr types that should be assigned in the registry:

Attr	Type
Interior Cost	1
Peer BGP Identifier	2
IPv4 Peer Address	3
IPv6 Peer Address	4

11. Security Considerations

There are no additional security risks introduced by this design.

12. References

12.1. Normative References

[I-D.ietf-grow-bgp-graceful-shutdown-requirements]
Takeda, T., Decraene, B., Francois, P., pelsser, c., Ahmad, Z., and A. Armengol, "Requirements for the graceful shutdown of BGP sessions", draft-ietf-grow-bgp-graceful-shutdown-requirements-07 (work in progress), January 2011.

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

[I-D.ietf-idr-best-external]
Marques, P., Fernando, R., Chen, E., and P. Mohapatra, "Advertisement of the best external route in BGP", draft-ietf-idr-best-external-03 (work in progress), March 2011.

[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.

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

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

12.2. Informative References

[RFC5004] Chen, E. and S. Sangli, "Avoid BGP Best Path Transitions from One External to Another", RFC 5004, September 2007.

Authors' Addresses

Pradosh Mohapatra
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: pmohapat@cisco.com

Rex Fernando
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: rex@cisco.com

Clarence Filsfils
Cisco Systems
Brussels,
Belgium

Email: cfilsfil@cisco.com

Robert Raszuk
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: raszuk@cisco.com