idr
Internet-Draft
Expires: January 14, 2009

B. Dickson
Afilias Canada, Inc

July 13, 2008

Enhanced BGP Capabilities for Exchanging Additional Nth-Best Paths
draft-dickson-idr-well-ordered-nth-best-01

Status of this Memo

Copyright Notice

Abstract

This Internet Draft describes an enhanced way to exchange prefix information, so as to permit multiple copies of a prefix, with different paths, to be announced and withdrawn.

This negotiated capability facilitates faster local (inter-AS) and global (intra-AS) convergence, reduces path-hunting, improves route-reflector behaviour, including eliminating persistent oscillations.

Additional prefix instances have a new wire format for updates and

withdrawals, to control path selection.

Benefits are seen both when deployed intra-AS, and on inter-AS peering.

Author's Note

This Internet Draft is intended to result in this draft or a related draft(s) being placed on the Standards Track for idr.

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 [5].

Intended Status: Proposed Standard.

Table of Contents

1. Background

Even when all the best current practises are observed, operational problems may be experienced when running a BGP network.

These include slow convergence due to "path-hunting" and persistant oscillations [1].

Standardization of MRAI timers helps path-hunting, and oscillations can be worked around with RFC 5004 [3].

However, both of these RFCs identify the above issues as needing further work.

1.1. The Best Path Chaining and the Best Path Tree

In a stable system of BGP speakers, for every given prefix, the selected best paths should form a spanning tree. At each node, the best path selected points further up the tree. The root of the tree is the destination, i.e. the originator of the prefix. The path from any leaf to the root forms a "chain" of best paths.

There are any number of ways that path attributes may be modified over time, at arbitrary places in this tree. When this happens, individual segments of the tree may conceptually "stretch" or "shrink". These changes may have no effect on the overall set of choices of best path, or they may cause a cascade effect "below" that point in the tree, with nodes migrating to new locations in a new version of the tree.

However, each node makes its choice of best path locally, and every time a node changes its selection of best path, that change is visible to its peers, and may in turn affect their own choice of best path. This propogation of changes is not instantaneous, and owing to the non-tree-like nature of the actual connectivity between nodes, can and does result in race conditions.

Depending on connectivity, peering policy, and initial conditions, the behavior may border on that of systems best describe through chaos theory. The time to reach a stable state, while generally bounded, is often far from fast, not necessarily predictable, and not necessarily consistent.

1.2. The Withdrawal Problem

Under normal circumstances, a change in attributes for a prefix will "flow" along the tree of best paths, without disrupting the structure of the tree itself signficantly. Even when a node selects a new best

path (and thus re-attaches itself to the tree in a new location), it typically will continue to pass the new attributes along the branch of the tree for which it is the root.

However, under certain circumstances, its choice of new best path, requires it to WITHDRAW the prefix from those peers, and effectively sever the branch. It is in the after-effects of this truncation that much of the path-hunting behavior gets triggered.

When a withdrawal effectively severs a branch of the tree, all the nodes on the tree will need to find new paths to the root. The problem is, that it takes some time for them to learn this fact.

In the mean time, the nodes in the severed branch may continue to use, and propogate, paths that are technically infeasible.

The idea is to fast-track the flooding of the infeasibility of paths throughout all parts of the tree below a given link, so as to minimize the use of infeasible paths.

1.3. The Uniqueness Property

Currently, for each prefix, only one path for that prefix is ever announced from one peer to another (ignoring Route Reflectors). Because of this property, uniqueness, a withdrawal on a prefix does not require path information. This also means that a change of best path is accomplished via an update for a prefix with the new path information.

If, however, more than one path for a given prefix were sent, then any attempt to withdraw a prefix+path would require some mechanism to distinguish between prefix instances.

In an environment where multiple path announcments per prefix are possible, but only one "best" path per prefix is maintained, then two steps would be involved in changing the "best" path. In no particular order, that would be the withdrawal of the old prefix+ path, and the announcement of the new prefix+path.

2. Proposed Changes

What is being proposed is, maintaining a set of "N best" for each prefix, and sending ALL of these rather than just the "best" path.

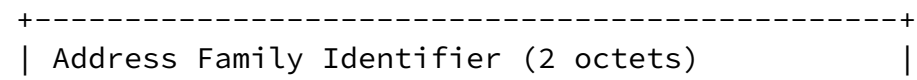
When any of the "N best" becomes infeasible, a withdrawal is sent. If a withdrawal is received, it receives special fast-track handling, taking advantage of the "N best" information. If any of the N best

is affected by the withdrawal, it is immediately flooded to peers without doing a prefix BGP path comparison (since those results have already been pre-calculated).

The supposition is that pruning all infeasible branches, while maintaining information on N best paths, allows for fast removal of all best paths which are dependent on infeasible paths, and fast reconvergence with pre-computed alternate paths. It is expected that the N-best mechanism should act as a stop-gap until, but not actually replace, full prefix path comparisons to generate a new set of "N best" paths.

2.1. USE_N_BEST Capability

The USE_N_BEST Capability is a new BGP capability [RFC2842]. The Capability Code for this capability is specified in the IANA Considerations section of this document. The Capability Length field of this capability is variable. The Capability Value field consists of zero or more of the tuples <AFI, SAFI> as follows:



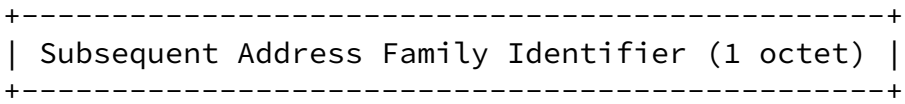


Figure 1

The meaning and use of the fields are as follows:

Address Family Identifier (AFI): This field carries the identity of the Network Layer protocol for which the BGP speaker intends to advertise multiple paths. Presently defined values for this field are specified in [IANA-AFI].

Subsequent Address Family Identifier (SAFI): This field provides additional information about the type of the Network Layer Reachability Information carried in the attribute. Presently defined values for this field are specified in [IANA-SAFI].

When advertising the USE_N_BEST Capability to a peer, a BGP speaker conveys to the peer that the speaker is capable of receiving multiple paths as well as the single path from the peer for address families that the speaker supports. When a tuple <AFI, SAFI> is included in the capability, it indicates that the BGP speaker intends to advertise multiple paths for the <AFI, SAFI>. If the USE_N_BEST Capability is also received from the peer, the speaker would then follow the procedures for advertising "Best N" paths to the peer for the specified <AFI, SAFI>.

When advertising "Best N" paths:

- Update messages MUST be in the new format [4], and ADD_PATH_ORDERED must also be advertised
- For each prefix, at most one of each ordinal value, 1 through N, may be sent
- The sender is responsible for selecting its own path ordinals
- The sender is responsible for maintaining the sequence order per prefix
- As a result of withdrawals, the sequence sent might not start at 1, and might be sparse

3. Modifications to BGP Behavior

3.1. Changes to Path Selection Rules

The path selection rules for BGP (section 9.1.2.2 of BGP4 [2]) are changed as follows:

- The following rule is a modification to step (c).

  It MAY only be needed when the node is acting as a Route Reflector. If a node is NOT a Route Reflector, a simplified modification (remove any paths NOT marked BEST) MAY be used. (This modification exists to resolve the Persistent Oscillation problem only.)

  The modification to step (c) is:

  Step (c) is first performed INCLUDING paths NOT marked as BEST.

  If, at the end of the first attempt at step (c), no paths marked BEST remain, re-run step (c), this time EXCLUDING all paths NOT marked BEST.

  After this modified version of step (c), it should be observed (and asserted) that only paths marked BEST must remain.

  In other words, Step (c) MUST remove any non-BEST paths.
- The remainder of the usual BGP path selection rules are applied as normal

The path selection rules for "Nth Best" path are as follows:

- The already-selected (N-1) best paths are removed from the set of paths to compare
- The same rules are applied as for the "best" path (including the modification to step (c), above)
- The selected path is advertised (to any peers with whom Nth-best has been negotiated), with the ordinal value of N applied

The prefix instances for consideration of Nth-best path are the REMAINDER of non-yet-selected instances. NB: Only the best (lowest

received ordinal), not-yet-selected instance of any IN-RIB may be selected for the local (and out-RIB) Nth-best path.

3.2. N Best - Basic Method

Once the capabality for doing so has been negotiated between a pair of BGP speakers, each sends the best N paths for each prefix. The path information will include the additional ordinal value on the

each Nth-best path.

When the current "best" path is withdrawn, the withdrawal MAY be propogated without having to perform a full BGP prefix path selection. The current "second best" path in the local-RIB is promoted to "best". This is because the alternate candidates have already been evaluated and "second-best" has already been selected.

Whenever an AS consists of a mesh of BGP speakers who have negotiated this capability, the withdrawal will propogate through the entire AS. This will either have no effect, or will cause a change in "best", which does not require non-local information in order to choose the new "best" path.

The second-best path from a neighbor MUST ONLY be considered as a candidate for best path, when the previous best path from that neighbor is withdrawn. When this occurs, the path in question is promoted to "best" status.

3.3. N Best - Route Reflector

The N best are all reflected. The same mechanism is used for determining the best N per prefix. Updates must be reflected whenever the choice of any of the best N change. Withdrawals may be propogated immediately.

3.4. N Best - Inter-AS Hybrid Method

When a withdrawal of the current best path is received from a peer doing USE_N_BEST, and the rules for sending updates require that an update for this prefix be sent to a peer who does not support USE_N_BEST, the current second-best instance of the prefix is sent to that peer in an Update. The neighbor does not need the withdrawal, since the new path replaces the old path.

3.5. IBGP vs EBGP

The same rules apply for EBGP->EBGP, EBGP->IBGP, IBGP->EBGP, and IBGP->IBGP. If a particular peering has had USE_N_BEST negotiated, then any update for a particular prefix that results in new selection

of any of the N best paths, the new selections (and possible

withdrawal of old selections) is sent to the appropriate peers.

4. Implementation Guidelines

In order to encourage effective implementation schemes, and to demonstrate some of the benefits of deployment, here are some suggestions for facilitating fast propogation of path changes, which are anticipated as improving behavior. This applies in particular to Path Hunting issues.

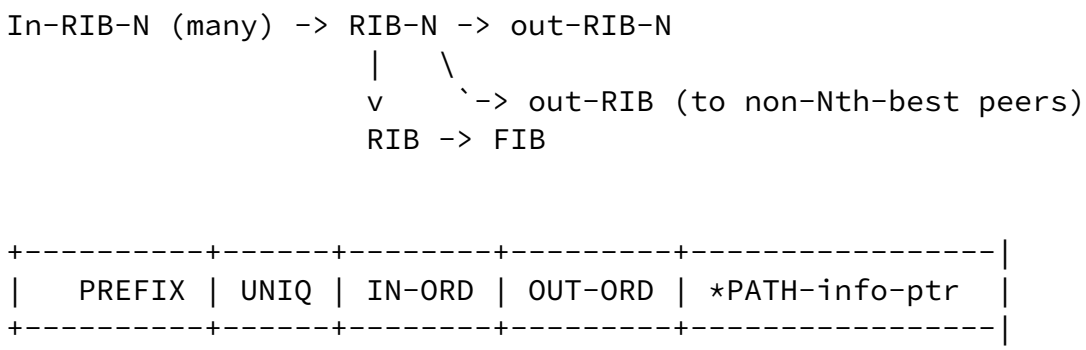


Figure 2

Where IN-ORD and OUT-ORD indicate the preference order (from BEST to Nth-BEST) of the sender, or ourselves, and UNIQ is chosen to uniquely identify the prefix; BGP Originator is used for UNIQ. IN-ORD are the values sent from a peer. OUT-ORD is non-zero for ONLY those prefixes selected for inclusion into the RIB-N.

For example, if all external peers have NOT negotiated Nth-Best, those prefixes would have an ordinal value of 1. Each In-RIB-N would have at most one instance. And for each prefix, at most one In-RIB-N would be selected as best, and have its corresponding OUT-ORD set to 1.

This forward-chaining allows for expedited processing of updates. We can immediately determine whether any given withdrawals need to be flooded to peers, and if so, what ordinal to use on the forwarded update. This flooding MAY be performed in parallel to normal BGP table update processing.

For clarity, it should be pointed out that:

- The process for the step RIB-N to RIB is "select prefixes with OUT-ORD == 1".

- The process for the step RIB-N to out-RIB is also "select prefixes with OUT-ORD == 1".
- The process for the step RIB-N to out-RIB-N is the same as ordinary RIB to out-RIB, except for preservation of Ordinal values.

5. Security Considerations

No additional security considerations beyond those already present in BGP are introduced.

6. IANA Considerations

IANA will need to assign a new code point for BGP Capabilities for USE_N_BEST.

7. Acknowledgements

The author wishes to acknowledge the helpful guidance of Joe Abley, and Tony Li. The author thanks the following for feedback during the review and revision process: Joel M. Halpern, Tony Li. The author also wishes to acknowledge the insight gained from his Scottish Deerhound, Skylar, winning a Reserve Best-in-Show. (The selection method of "second best" comes from the Reserve system used at the group and best-in-show levels of dog shows).

8. References

8.1. Normative References

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

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

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

[4] Dickson, B., "Enhanced BGP Capabilities for Exchanging Second-Best Paths", July 2008.

8.2. Informative References

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

Appendix A. Path-Hunting Examples

(These will be included in a subsequent version of this ID.)

Appendix B. Persistent Oscillation Examples

Consider the example in Figure 3 where o R1, R2, R3, R4, and R5 belong to one AS. o R1 is a route reflector with R2 and R3 as its clients. o R4 is a route reflector with R5 as its client. o The IGP metrics are as listed. o External paths (a), (b), and (c) are as described in Figure 4.

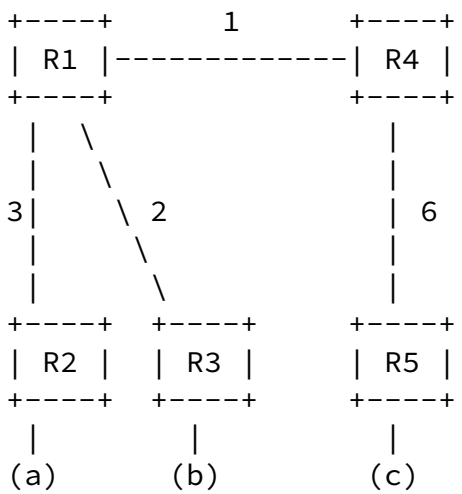


Figure 3

Path	AS_PATH	MED
a	1 3	10
b	2 3	1

```
 c       2 3     0
```

Figure 4

With the addition of "Nth Best", and locally limiting N to 2, we have:

R1 has the following:

```
Path    AS_PATH MED IGP-metric
 a       1 3    10   3 (received:best) (best)
 b       2 3     1   2 (received:best)
 c       2 3     0   7 (received:best) (second_best - not sent)
```

R4 has the following:

```
Path    AS_PATH MED IGP-metric
 a       1 3    10   4 (received:best) (best - not sent)
 c       2 3     0   6 (received: best) (second_best)
```

This results in R1 having:

```
Path    AS_PATH MED IGP-metric
 a       1 3    10   3 (received:best) (best)
 b       2 3     1   2 (received:best)
 c       2 3     0   7 (received:second_best) (second_best - not sent)
```

By including N best (for N=2) in the best path calculation, the persistent oscillation problem is resolved.

Author's Address

Brian Dickson
Afilias Canada, Inc
4141 Yonge St,
Suite 204
North York, ON M2P 2A8
Canada

Email: brian.peter.dickson@gmail.com
URI: www.afilias.info

Full Copyright Statement

Intellectual Property

on the procedures with respect to rights in RFC documents can be found in BCP 78 and BCP 79.

Copies of IPR disclosures made to the IETF Secretariat and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the IETF on-line IPR repository at http://www.ietf.org/ipr.

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to implement this standard. Please address the information to the IETF at ietf-ipr@ietf.org.

Acknowledgment

Funding for the RFC Editor function is provided by the IETF Administrative Support Activity (IASA).