idr
Internet-Draft
Expires: January 14, 2009

B. Dickson
Afilias Canada, Inc

July 13, 2008

Enhanced BGP Capabilities for Exchanging Second-Best Paths
draft-dickson-add-paths-ordered-01

Status of this Memo

Copyright Notice

Abstract

This Internet Draft describes an enhanced format for encoding prefix information, to permit multiple copies of a prefix with different paths to be announced and withdrawn.

Prefix instances using the new format include both unique identifiers, and ordinals to control path selection.

Withdrawal of prefixes requires a slight modification to disambiguate prefix instances.

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 [4].

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 the "best N" for each prefix, and sending all of these rather than just the "best" path.

The supposition is that pruning all infeasible branches, while maintaining information on the next N best paths, allows for fast removal of all (possibly 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 BGP path selection to generate a new set of "best N" paths.

2.1. How to Identify a Path

As defined in [RFC4271], a path refers to the information reported in the path attribute field of an UPDATE message. As the procedures specified in [RFC4271] allow only the advertisement of one path for a particular address prefix, a path for an address prefix from a BGP peer can be keyed on the address prefix.

In order for a BGP speaker to advertise multiple paths for the same address prefix, a new identifier (termed "Path Identifier" hereafter) needs to be introduced so that a particular path for an address prefix can be identified by the combination of the address prefix and the Path Identifier.

Depending on the application and the configuration of a particular peer, the Path Identifier for a path can be an AS number, or a BGP Identifier, or an opaque number, with which a path is associated by

the BGP speaker that advertises the path.

2.2. Extended NLRI Encodings

In order to carry the Path Identifier in an UPDATE message, the existing NLRI encodings specified in [RFC4271, RFC2858] are extended as the following:

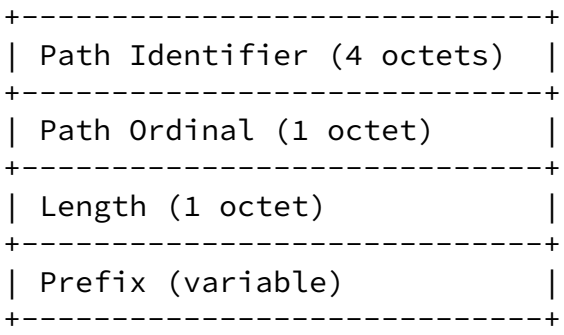

Figure 1

and the NLRI encoding specified in [RFC3107] is extended as the following:

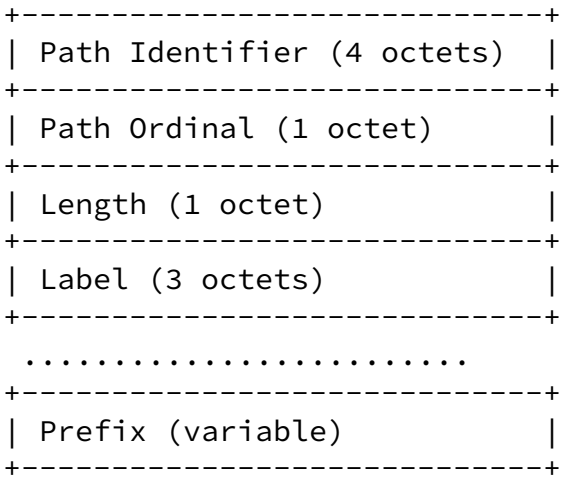

Figure 2

Update messages are otherwise identical to existing format. If BGP capability ADD_PATHS_ORDERED has been negotiated, every Update MUST have the New Update Format. More than one instance of a given prefix, with distinct values of Path Attributes, MAY be sent between BGP speakers.

At most N instances may be sent, where N is the value sent along with the ADD_PATHS_ORDERED capability.

Two prefix paths are considered identical if they differ only in the value of the ordinal. An Update which contains a path which differs from the previous path with that value of UNIQ (identifier), will result in the path information for the prefix and UNIQ being modified.

The Ordinal must be non-zero, but the rules governing values of Ordinal(s) used are specific to RFCs which refer to this document. For example, BGP Equal-Cost Multipath may allow two paths with the same Ordinal to be used. Similarly, BGP N-best Paths may require per-prefix Ordinals be unique.

2.3. ADD_PATH_ORDERED Capability

The ADD_PATH_ORDERED 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, MOV> as follows:

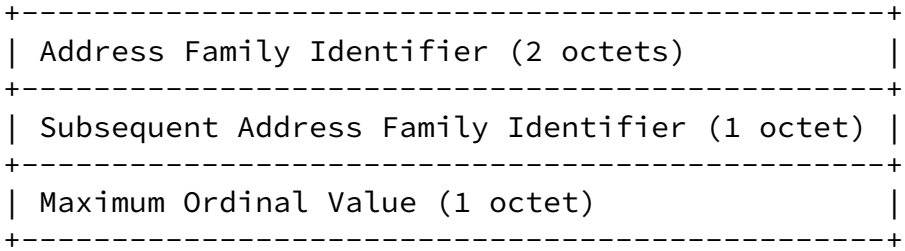

Figure 3

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].

Maximum Ordinal Value (MOV): This field specifies the maximum value the speaker will send in the Ordinal field of any Update. It does not mean that that the speaker will necessarily send any particular Ordinal value within that range, nor that more than one Ordinal value will be used. The value is an unsigned 8-bit value greater than zero.

When advertising the ADD_PATH_ORDERED 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, MOV> is included in the capability, it indicates that the BGP speaker intends to advertise multiple paths for the <AFI, SAFI, MOV>. If the ADD-PATH Capability is also received from the peer, the speaker would then follow the procedures for advertising multiple paths to the peer for the specified <AFI, SAFI, MOV>.

3. Security Considerations

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

4. IANA Considerations

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

5. Acknowledgements

The author wishes to acknowledge the helpful guidance of Joe Abley, Tony Li, and Yakhov Rehkter. The author thanks the following for feedback during the review and revision process: Joel M. Halpern, Tony Li. The author has based much of this document on an expired Internet Draft, "draft-walton-bgp-addp-paths-05", and has used substantial portions of that draft verbatim. The original authors of that draft were Daniel Walton, Alvaro Retana, and Enke Chen, of Cisco Systems.

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).

6. References

6.1. 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.

6.2. Informative References

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

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

ietf-ipr@ietf.org.

Acknowledgment

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