idr
Internet-Draft
Expires: August 28, 2008

B. Dickson
Afilias Canada, Inc

February 25, 2008

Enhanced BGP Capabilities for Exchanging Second-best and Back-up Paths
draft-dickson-idr-second-best-backup-01

Status of this Memo

Copyright Notice

Abstract

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way to exchange prefix information, to permit multiple copies of a prefix with different paths to be announced and withdrawn.

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

Additional prefix instances have new optional BGP attributes, to control path selection.

Withdrawl of prefixes will require new attributes to disambiguate prefix instances.

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", persistant oscillations [1], and BGP "wedgies" [2].

Standardization of MRAI timers helps this, as well as RFC 5004 [4].

These RFCs identify the above issues as needing further work.

1.1. Localized Information

The problems listed above occur as a result of additional information not being available (either on a transient basis, or permanently.)

In the case of "path hunting", the information needed for achieving a stable final state is eventually received, but until it is, sub-optimal forwarding will occur, and possibly even transient routing loops.

The "problem" mechanisms involved are:

- the suppression of announcement of "second-best" paths, because of IBGP-received "best" paths;
- the suppression by route-reflectors, of IBGP non-best paths (i.e. those normally seen directly by IBGP peers)
- the suppression of announcement of "second-best" paths, because of EBGP-received "best" paths.
- the lack of explicit global mechanism for expressing de-prefering announcements via "back-up" providers.

When a prefix+path received is better than the local "best", the new "best" is normally sent.

However, once a new "best" is received, the side-effect is to force the speaker to WITHDRAW the previous best path within the same "regime" (IBGP mesh or EBGP peers).

When we consider the extra (e.g. suppressed) information, with special rules on what to send and how to treat it, the specified problems may go away, or be reduced in scope, duration, or likelihood.

1.2. The Withdrawl Problem

When a prefix (plus path) is withdrawn, the desired stable state is for the next-best path for that prefix (if one exists) to be chosen at each BGP speaker per its local policy.

If that second-best path is already on hand, the delay and intermediate states can be reduced or entirely avoided. This is especially true for both intra-AS and inter-AS "path hunting".

To avoid inconsistent behavior, routing loops, and routing-information loops, the second-best path received from a neighbor, should never be selected as a best path locally.

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.

1.3. The Uniqueness Problem

Currently, for each prefix, only one path for that prefix is ever announced from one peer to another (except in the instance of Route Reflectors). Because of this property, uniqueness, a withdrawl 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 was sent, then any attempt to withdraw a prefix+path would require that the specific path for the prefix being withdrawn be supplied in the withdrawl update message.

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

2. Proposed Changes

2.1. New Negotiated Option: USE_SECOND_BEST_AND_BACKUP

This is a new BGP Capabilities value, which can be optionally included in the capabilities negotiation. The specific value is a code-point to be assigned by IANA.

When negotiated:

- Update messages MUST be in the new format
- Updates without any of the new optional attributes are considered BEST
- For each prefix, at most one of each type (BEST, SECOND_BEST, BACKUP_ONLY, BACKUP_ONLY_SECOND_BEST) may be sent

2.2. New Optional Path Attribute: SECOND_BEST

This is a new BGP Path Attribute type. It MAY be used only if the USE_SECOND_BEST_AND_BACKUP capability has been negotiated. The type value is a new code point to be assigned by IANA.

This is an Optional, Non-Transitive, Non-Extended, Non-Partial attribute. All the "attr flag bits" (from BGP [3]) are zero. The length is 1, and the value is 1.

2.3. New Optional Path Attribute: BACKUP_ONLY

This is a new BGP Path Attribute type. The type value is a new code point to be assigned by IANA. This is an Optional, Transitive, Non-Extended, Non-Partial attribute, with the "attr flag bits" (from BGP [3]) set to appropriate values. The length is 1, and the value is 1.

2.4. New Optional Path Attribute: BACKUP_ONLY_SECOND_BEST

This is a new BGP Path Attribute type. It MAY be used only if the USE_SECOND_BEST_AND_BACKUP capability has been negotiated. The type value is a new code point to be assigned by IANA.

This is an Optional, Non-Transitive, Non-Extended, Non-Partial attribute. All the "attr flag bits" (from BGP [3]) are zero. The length is 1, and the value is 1.

2.5. New Update Format

Update messages are identical to existing format, with the exception of the new Withdrawl format, and the new optional Path Attributes (SECOND_BEST ,BACKUP_ONLY, and.BACKUP_ONLY_SECOND_BEST). If BGP capability USE_SECOND_BEST_AND_BACKUP has been negotiated, any Update MAY have a Path Attribute(s) which include SECOND_BEST, BACKUP_ONLY, and/or BACKUP_ONLY_SECOND_BEST. More than one instance of a given prefix, with distinct values of Path Attributes, MAY be sent between BGP speakers.

At most four instances may be sent, specifically one of each combination of with/without SECOND_BEST and BACKUP_ONLY and BACKUP_ONLY_SECOND_BEST: One with neither, one with SECOND_BEST only, one with just BACKUP_ONLY, and one with BACKUP_ONLY_SECOND_BEST. both SECOND_BEST and BACKUP_ONLY.

Two prefix paths are considered identical if they differ only in the presence or absence of any of the new attributes. An Update which contains a path which differs by either or both of these, will result in the path information for the prefix being modified.

2.6. New Withdraw Format

Since it is no longer possible to identify which instance of an prefix is affected by an update containing a withdrawl, a new format for Withdrawls is needed. For simplicity of implementations, this consists of four Withdrawl sections, one for each of the types (BEST, SECOND_BEST, BACKUP_ONLY, BACKUP_ONLY_SECOND_BEST). They occur in REVERSE order, to simplify state transitions if/when a "BEST" path is withdrawn. Each Withdrawl section has the same format as the original Withdrawl section.

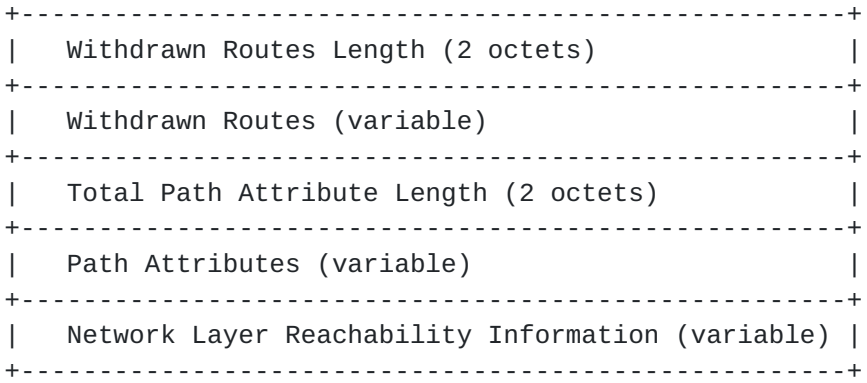


Figure 1

Withdrawn Routes Length: This 2-octets unsigned integer indicates the total length of the Withdrawn Routes field in octets. Its value allows the length of the Network Layer Reachability Information field to be determined, as specified below.

A value of 0 indicates that no routes are being withdrawn from service, and that the WITHDRAWN ROUTES field is not present in this UPDATE message.

Withdrawn Routes Field: This field now consists of four sub-fields and their respective lengths. The value for Withdrawn Routes Length above, must be the sum of the four lengths, plus 8 (the sum of the lengths of the Subfield Lengths).

The format and sequence of the subfields is as follows:

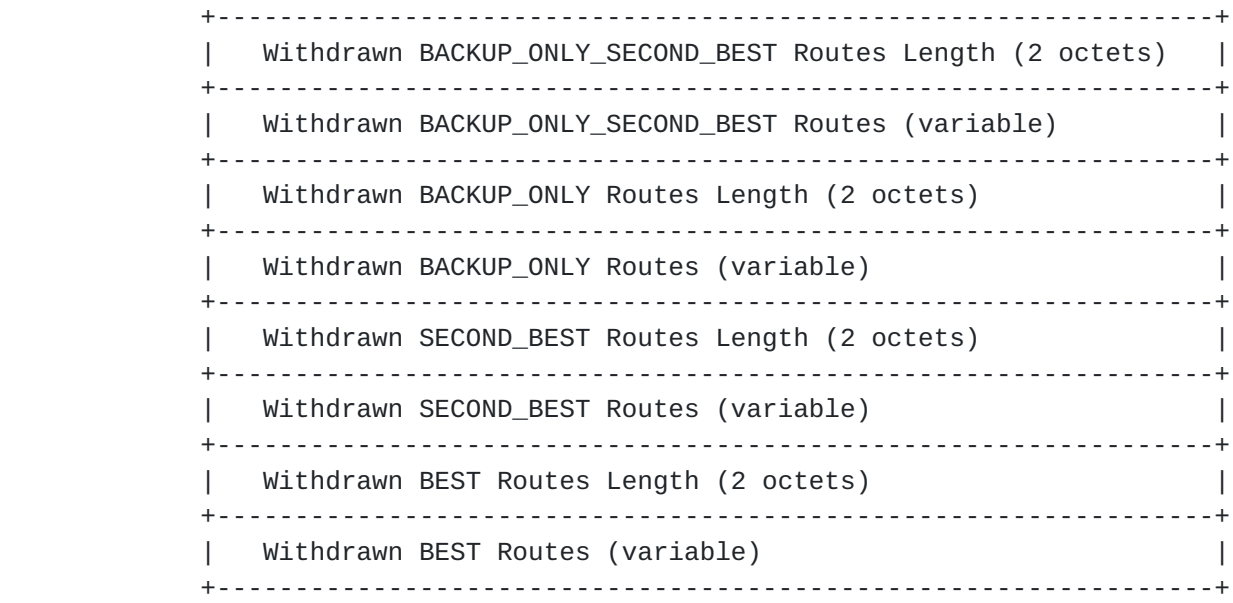
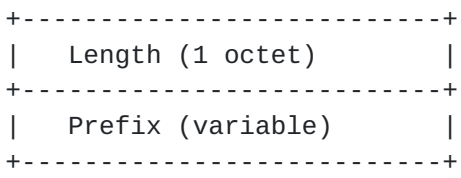


Figure 2

Withdrawn Routes Subfield Lengths These 2-octets unsigned integers indicates the total length of their respective Withdrawn Routes subfields in octets.

Withdrawn Routes Subfields: Each of these is a variable-length field that contains a list of IP address prefixes for the routes that are being withdrawn from service. Each IP address prefix is encoded as a 2-tuple of the form <length, prefix>, whose fields are described below:


Length (1 octet)
Prefix (variable)

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 [3]) are changed as follows:

- The following rule is placed before step (a): If paths with and without BACKUP_ONLY (or BACKUP_ONLY_SECOND_BEST) are both available, those with BACKUP_ONLY/BACKUP_ONLY_SECOND_BEST are eliminated

- The following rule is a modification to step (c): Step (c) is first performed INCLUDING paths with SECOND_BEST. If, at the end of the first attempt at step (c), only paths with SECOND_BEST remain, re-run step (c), this time EXCLUDING the paths with SECOND_BEST. After this modified version of step (c), the remaining paths MUST NOT have the SECOND_BEST attribute. In other words, Step (c) MUST remove any SECOND_BEST paths.

- The remainder of the usual BGP path selection rules are applied as normal

- If the final path selected has the BACKUP_ONLY/ BACKUP_ONLY_SECOND_BEST attribute, the attribute BACKUP_ONLY MUST be set.

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

- The already-selected "best" path is removed from the set of paths to compare

- The same rules are applied as for the "best" path

- The selected path is advertised with the attribute SECOND_BEST applied

- If the selected path had the BACKUP_ONLY attribute, the attribute BACKUP_ONLY_SECOND_BEST must be set.

The prefix instances for consideration of second-best path are the REMAINDER of non-SECOND_BEST instances, and the SECOND_BEST instance received on the in-RIB from which the best path was selected (if one exists). Only one SECOND_BEST instance received may be considered for the local (and out-RIB) SECOND_BEST path.

3.2. Second Best - Basic Method

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

When the current "best" path is withdrawn, the withdrawl MAY be propogated without having to perform a full BGP table 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 withdrawl will propogate through the entire AS. This will either have no effect, or with a change in "best" without requiring non-local information to choose the new "best" path.

3.3. Second Best - Route Reflector

The "best" and "second best" are reflected. The same mechanism is used for determining both best and second-best per prefix. Updates must be reflected whenever the choice of either or both of the "best" or "second best" change. Withdrawls may be propogated immediately.

3.4. Second Best - Inter-AS Hybrid Method

When a withdrawl of the current best path is received from a peer doing USE_SECOND_BEST_AND_BACKUP, and the rules for sending updates require that an update for this prefix be sent to a peer who does not support USE_SECOND BEST_AND_BACKUP, 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.

When the selection of best path results in the selection of a path with BACKUP_ONLY, the path is sent as the best path. This is the only time where a BACKUP_ONLY path is sent as BEST, without preserving the BACKUP_ONLY attribute.

3.5. Backup Only - Basic Method

The main reason for establishing the BACKUP_ONLY attribute is to permit the global implementation of actual "backup only" announcements. It is not to facilitate change of policies, or to circumvent local policies, instead it is to make possible the implementation of policies where those have been negotiated by two or more parties.

Currently, there are several documented scenarios in the "Wedgies" RFC [2] where the mutually desired policy is either unable to be implemented, or does not deterministically reach the desired state.

Use of the BACKUP_ONLY attribute on announcements sent to a backup provider, permit these problems to be resolved.

The same prefix is announced to both the primary and backup provider. When announced to the primary provider, the BACKUP_ONLY attribute is NOT set. When announced to the backup provider, the BACKUP_ONLY attribute IS set.

The propogation of the BACKUP_ONLY instance will be limited by the availability of multiple paths and the use of SECOND_BEST peerings.

In Figure 10 (of Appendix C), the BACKUP_ONLY instance will be seen by the backup provider, and be passed with both SECOND_BEST and BACKUP_ONLY to the backup provider's transit provider. The latter will prefer any other instace without BACKUP_ONLY, even if it has applied a LOCAL_PREFERENCE to the received prefix instance. Should the other instance be withdrawn, the BACKUP_ONLY will be selected and subsequently propogated. The withdrawl will also eventually result in an Update with the BACKUP_ONLY attribute but WITHOUT the SECOND_BEST attribute (since the prefix will now only be reachable via the backup provider.)

3.6. Backup Only - Route Reflector

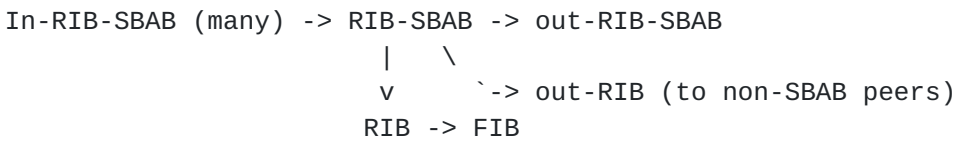
Route Reflectors operate the same as always. The BACKUP_ONLY attribute MUST be preserved during reflection. Thus, if "Second Best" is in operation, then the BACKUP_ONLY attribute of both best and second-best MUST be preserved on both instances. And, if "Second Best" is not in use, then the selected "best" prefix, if it has BACKUP_ONLY set, must be reflected with BACKUP_ONLY as well.

3.7. IBGP vs EBGP

The same rules apply for EBGP->EBGP, EBGP->IBGP, IBGP->EBGP, and IBGP->IBGP. If a particular peering has had USE_SECOND_BEST_AND_BACKUP negotiated, then any update for a particular prefix that results in new selection of either or both of best and second-best, the new selections (and possible withdrawl of old selections) is sent to the appropriate peers. Additionally, updates which have BACKUP_ONLY MAY be sent.

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.



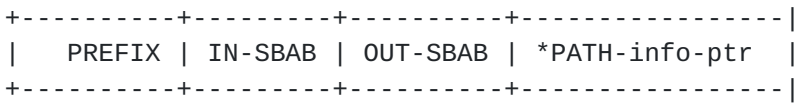


Figure 4

Where IN-SBAB and OUT-SBAB are 4-bit fields indicating what the SECOND_BEST_AND_BACKUP (SBAB) are attributes (BEST, SECOND_BEST, BACKUP_ONLY, SECOND_BEST_BACKUP_ONLY). IN-SBAB are the attributes received from a peer, and for ONLY those prefixes selected for inclusion into the RIB-SBAB, what the corresponding attributes are.

For example, if all external peers have NOT negotiated SBAB, those prefixes would have SBAB binary values of 1000. Each In-RIB-SBAB would have at most one instance. And for each prefix, at most one In-RIB-SBAB would be selected as best, and have its corresponding OUT-SBAB set to binary value 1000.

This forward-chaining allows for processing of SBAB updates to determine whether withdrawals need to be flooded to peers, and if so, what SBAB attribute to apply to the withdrawals that are flooded. 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-SBAB to RIB is "select prefixes marked 'best'".
- The process for the step RIB-SBAB to out-RIB is also "select prefixes marked 'best'".

- The process for the step RIB-SBAB to out-RIB-SBAB is the same as ordinary RIB to out-RIB, except for preservation of SBAB attributes (if any).

5. Security Considerations

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

6. IANA Considerations

IANA will need to assign new code points for BGP Capabilities for USE_SECOND_BEST_AND_BACKUP. IANA will need to assign new code points for BGP Attribute Types for SECOND_BEST, BACKUP_ONLY and BACKUP_ONLY_SECOND_BEST.

7. Acknowledgements

The author wishes to acknowledge the helpful guidance of Joe Abley, Tony Li, and Yakhov Rehkter. 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] Griffin, T. and G. Huston, "BGP Wedgies", RFC 4264, November 2005.

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

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

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 5 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 6.

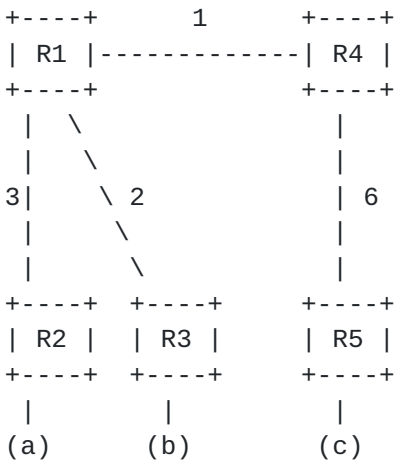


Figure 5

Path	AS_PATH	MED
a	1 3	10
b	2 3	1
c	2 3	0

Figure 6

With the addition of "second best", 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 the second_best in the best path calculation, the persistent oscillation problem is resolved.

Appendix C. BGP Wedgie Examples

The following examples from RFC 4264 [2] show the effects of the proposed changes, in resolving "wedgie" issues.

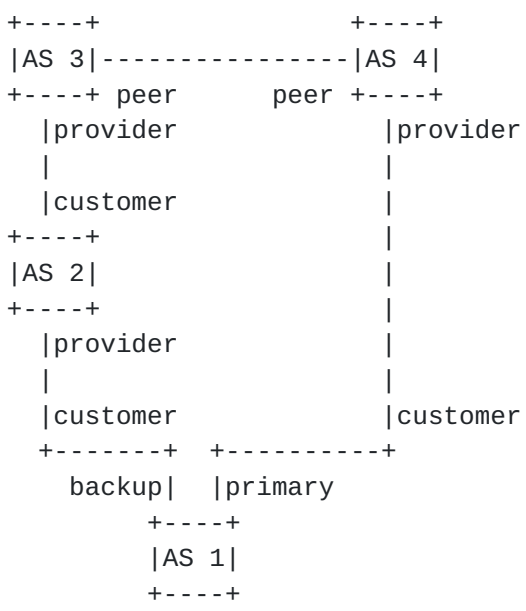


Figure 10

In Figure 10 above, the announcement via the backup link is sent with BACKUP_ONLY.

- AS 4 sends the "best" (the direct link to AS 1) to AS 3.
- AS 2 sends its "best", which is the BACKUP_ONLY path from AS 1, to AS 3, also with BACKUP_ONLY (since it is a transitive attribute).
- AS 3 and AS 4 exchange their respective "best" paths.
- AS 3 prefers the path "4 1" over "2 1" because "2 1" is BACKUP_ONLY.
- AS 3 sends a revised BACKUP_ONLY update to AS 4 as SECOND_BEST.
- AS 3 sends the new "best" to AS 2.
- AS 2 sends a revised BACKUP_ONLY update to AS 3 as SECOND_BEST.

This state will be reached regardless of sequence of disconnects and reconnects.

Link failures will also result in propogation of withdrawls of "best"

and the SECOND_BEST promotions will result in immediate correct behavior.

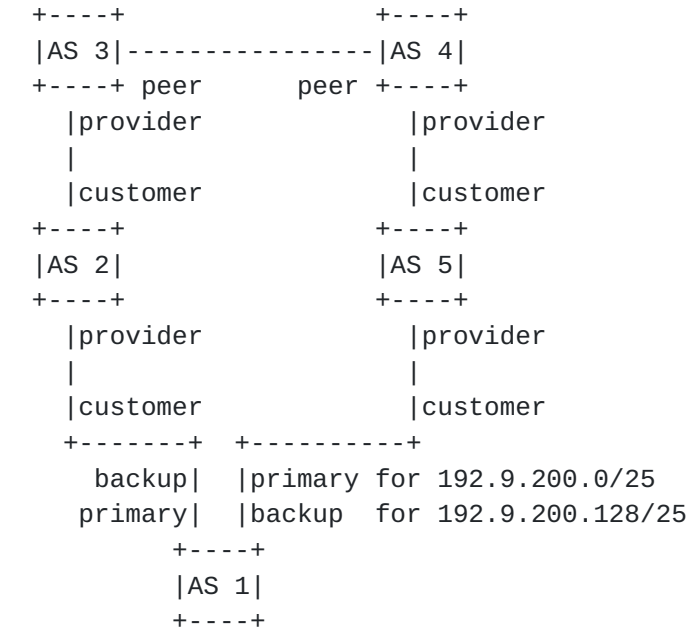


Figure 11

In Figure 11 above, the announcements via the backup links will work the same as in Example 1.

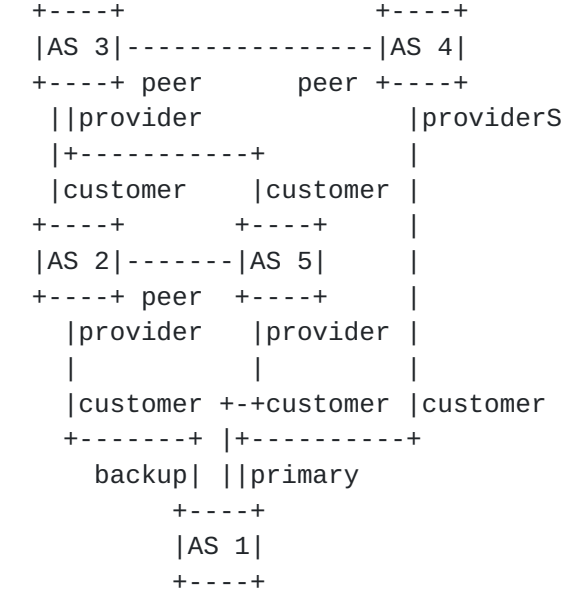


Figure 12

In Figure 12 above, the announcements via both backup links will

result in:

- AS 2 selecting its best path via "3 4 1" (the only path it hears from AS 3)
- AS 2 hearing two paths from AS 5:
  - its "second best" path "5 3 4 1"
  - another path marked SECOND_BEST and BACKUP_ONLY with path "5 1"
- AS 2 hearing a BACKUP_ONLY directly from AS 1

Any announcement that AS 3 hears from AS 2 or AS 5 will always be marked BACKUP_ONLY. Thus, any combination of break/restore on any links in any order, will always result in the desired state being reached.

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

Acknowledgment

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