Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: 22 January 2026

K. Patel
Arrcus, Inc
A. Vyavaharkar
Cisco Systems
N. Fazlollahi
Unaffiliated
A. Przygienda
Juniper Networks
A. Ananthamurthy
Cisco Systems
21 July 2025

Extension to BGP's Route Refresh Message
draft-idr-bgp-route-refresh-options-06

Abstract

[RFC2918] defines a route refresh capability to be exchanged between BGP speakers. BGP speakers that support this capability are advertising that they can resend the entire BGP Adj-RIB-Out on receipt of a refresh request. By supporting this capability, BGP speakers are more flexible in applying any inbound routing policy changes as they no longer have to store received routes in their unchanged form or reset the session when an inbound routing policy change occurs. The route refresh capability is advertised per AFI, SAFI combination.

There are newer AFI, SAFI types that have been introduced to BGP that support a variety of route types (e.g. IPv4/MVPN, L2VPN/EVPN). Currently, there is no way to request a subset of routes in a Route Refresh message for a given AFI, SAFI. This draft defines route refresh capability extensions that help BGP speakers to request a subset of routes for a given address family. This is expected to reduce the amount of update traffic being generated by route refresh requests as well as lessen the burden on the router servicing such requests.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Copyright Notice

Table of Contents

1. Introduction

[RFC2918] defines a route refresh capability to be exchanged between BGP speakers. BGP speakers that support this capability are advertising that they can resend the entire BGP Adj-RIB-Out on receipt of a refresh request. By supporting this capability, BGP speakers are more flexible in applying inbound routing policy changes as they no longer have to store copies of received routes in their unchanged form or reset the session when an inbound routing policy change occurs. The route refresh capability is advertised per AFI, SAFI combination.

Route refresh allows routers to dynamically request a full Adj-RIB-Out update from their peers when there's an inbound routing policy change. This is useful because routers that mutually support this capability no longer have to flap the peering session or store an extra copy of received routes in their original form. This helps by reducing memory requirements as well as eliminating the unnecessary churn caused by session flaps. [RFC2918] does not define a way for routers to request a subset of the Adj-RIB-Out for a given AFI, SAFI.

This draft defines new extensions to route refresh that will allow requesting routers to ask for a subset of the Adj-RIB-Out for a given AFI, SAFI combination. For example, routers could ask for specific route types from those address families that support multiple route types or, they could ask for a specific prefix.

As part of the new extensions, this draft combines elements of [RFC7313] and [RFC5291] and adds a new set of options to the route refresh message that will specify filters that can be applied to limit the scope of the refresh being requested. The new option format will apply to all new option types that may be defined moving forward.

1.1. Use Case Examples

The authors acknowledge that while the extensions being proposed in this draft could potentially be addressed by Route Target Constrain described in [RFC4684] by using route targets to identify desired subset of routes, this proposal includes address families where RT Constrain extension is not supported and avoids the necessity to assign and manage the route targets per desired set of routes. The approach in this draft is intended to be a single-hop refresh only, i.e., propagation of the refreshes in a way similar to RT Constrain routes is NOT intended.

Several possible use cases are discernible today:

* The capacity to refresh routes of a certain type within an address family is needed, e.g., auto discovery routes within the EVPN AF [RFC7432].

* In VPN scenarios where RT Constrain is not supported or configured, RDs can be used.

* In BGP LS [RFC7752] cases a speaker may choose to hold only a subset of routes and depending on configuration request a subset of routes. This document could provide further filters to support those use cases.

* On changes in inbound policy, when previously configured filters have been removed, only the according subset of routes may be requested.

2. 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].

3. Route Refresh Options Capability

A BGP speaker will use the BGP Capabilities Advertisement [RFC5492] to advertise the Route Refresh Options Capability to its peers. This new capability will be advertised using the Capability code [TBD] with a capability length of 0.

By advertising the Route Refresh Options Capability to a peer, a BGP speaker indicates that it is capable of receiving and processing the route refresh options described below. This new capability can be advertised along with the Enhanced Route Refresh Capability described in [RFC7313]. However, if the Route Refresh Options Capability has been negotiated by both sides of the BGP session, then it will override the Enhanced Route Refresh Capability.

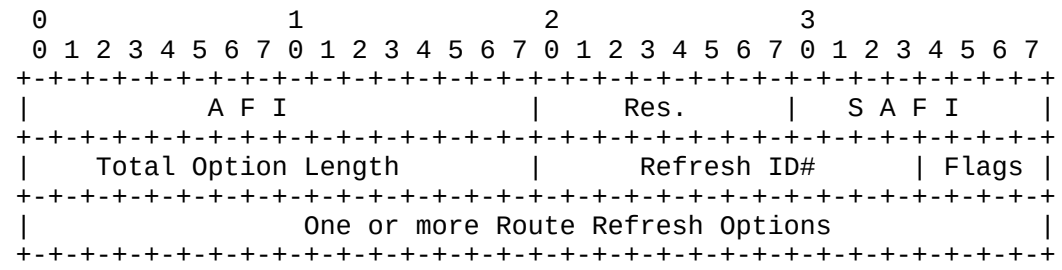
4. Route Refresh Sub-Types

[RFC7313] defines route refresh BGP message sub-types that utilize the "Reserved" field of the Route Refresh message originally defined in [RFC2918]. Currently, there are three sub-types defined and this draft proposes three additional sub-types which will be used to indicate a Route Refresh message that includes options before any ORF field of the Route Refresh message as well as BoRR and EoRR Route Refresh messages with options.

```
      0 - Normal route refresh request [RFC2918]
          with/without Outbound Route Filtering (ORF) [RFC5291]
      1 - Demarcation of the beginning of a route refresh
          (BoRR) operation
      2 - Demarcation of the ending of a route refresh
          (EoRR) operation
    + 3 - Route Refresh request with options and optional
          ORF [RFC5291]
    + 4 - BoRR with options
    + 5 - EoRR with options
    255 - Reserved
```

When the Route Refresh Options Capability has been negotiated by both sides of a BGP session, both peers MUST use message types 3, 4 and 5. The requesting speaker MUST use the refresh ID for all refresh requests including those without any options, i.e., requests for the full BGP Adj-RIB-Out.

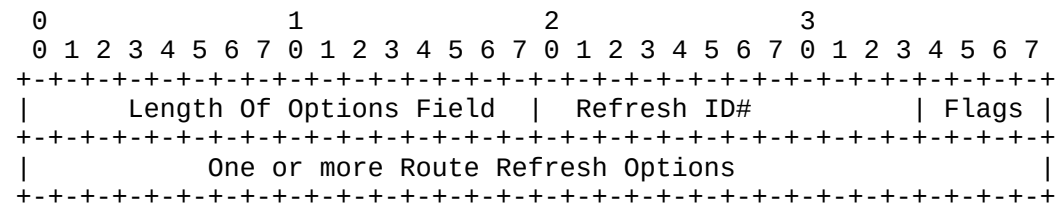
The Route Refresh Request Message with options will now be formatted as shown below



5. Route Refresh Option format

[RFC2918] defines the route refresh BGP message that includes only the AFI, SAFI of the routes being requested. This draft proposes extending the basic message by including options that will indicate to the remote BGP speaker that a subset of the entire Adj-RIB-Out is being requested. The remote BGP speaker will select routes that match the specified options and the flag settings.

As described in the previous section, the options will be added to the Route Refresh message before the ORF field of the message. Outbound Route Filtering is described in [RFC5291]. The options will assume the following format



6. Route Refresh Option Length

The Option Length field will occupy the two octets immediately following the Route Refresh message containing the AFI, SAFI and sub-type. The purpose of this field is to allow the BGP speaker to calculate the length of any attached ORF fields by subtracting the Option Length from the Route Refresh message length.

7. Route Refresh ID

The Refresh ID field will occupy twelve bits following the Route Refresh Options Length. It is infeasible to use a wide number like a 64-bit unsigned integer since this number must be stored per route entry associated with any peer supporting this feature, at least during the time any refresh is pending. It is a value assigned by the requesting BGP speaker. It MUST be a strictly monotonically increasing number per peer AFI and SAFI using sequence number arithmetic based on two-complements given in Appendix A. It is comparable to the calculations standardized in [RFC1982] but fixes several of its anomalies. The purpose of this field is to allow the requesting BGP speaker to correlate concurrent, overlapping refresh requests and ultimately delete correct stale routes. The Refresh ID MUST be reflected in the BoRR and EoRR messages sent by the BGP speaker servicing the refresh request.

A Refresh ID value MUST NOT be reused until an EoRR with this ID has been received by the requesting speaker or the last resort time has expired. The behavior is unspecified otherwise. More specifically, defining the interval [LID, HID] by the values

LID = MAX(lowest requested Refresh ID# without BoRR,
lowest received BoRR without EoRR)

and

HID = highest requested Refresh ID#

the requesting speaker MUST only use values V where V >: LID and V >: HID as defined by the relation given in Appendix A. Beside that, HID =>: LID MUST hold by the same algebra.

If no such number V exists, LID must catch up to HID, i.e. no further requests can be issued. To use a 3 bit example in Appendix A, if LID was 1 and HID was 4, we cannot progress to unsigned 5 since 1 ? 5. When LID progresses to unsigned 2 however, we have 5 >: 2 and 5 >: 4 and we can choose a V.

Value of 0 MUST NEVER be used as Refresh ID and is considered an "invalid" ID.

The sending speaker MUST NOT reorder the BoRR messages on sending in case it received multiple requests, i.e., the BoRRs MUST follow in the same sequence as the requested Route Refresh IDs.

8. Route Refresh Option Flags

This draft defines several route refresh option flags:

* ’O’-bit specifies whether the receiving BGP speaker MUST logically OR the attached options or logically AND them (in case of the bit being clear). When the flag is clear, the router on the receiving end SHOULD logically AND the options and only refresh routes that match all received options. If the option flag is set, the router SHOULD select routes that match using a logical OR of the options. In any case the set of routes sent between the according BoRR and EoRR MUST contain at least the logically requested set.

* ’C’ bit indicates that the receiving BGP speaker MUST clear immediately all the received Route Refresh Requests with Options, either pending or being processed. EoRRs MUST NOT be sent. The Refresh ID# on the request MUST be set as the (in unsigned terms) next possible number L for which LID >: L and HID >: L per Appendix A or in other words we "wrap around the sequence number space" on reset. The C flag MUST NOT be set on BoRR or EoRR messages and CAN be used only with refresh requests.

* by ’S’ bit indicate a refresh is being spontaneously originated by the BGP speaker which received requests and has them pending. The receiving BGP speaker MUST immediately clear all their pending Route Refresh requests with the sending peer. The Refresh ID# on the request MUST be set as the the largest unsigned number L for which LID >: L and HID >: L. When this flag is set, the receiving BGP speaker MUST use this sequence number for its next request. To use example from Appendix A, if the peer received LID 4 and HID 5 (i.e. it didn’t send BoRR for 4 yet but received request for 5 already) it will reset the sequence number to 1 by those rules. Now, if there is a request with 6 in flight, it will be seen as 1 >: 6 when arriving.

The precise format is indicated below

```
 0 1 2 3 4 5 6 7
+-+-+-+-+-+-+-+-+
|  .... |C|O|S|R|
+-+-+-+-+-+-+-+-+
```

C Clear pending requests and reset Refresh ID# space.

O Use logical OR of attached options

S Synchronize sequence numbers

R Reserved bit

9. Route Refresh Options

This draft introduces new options carried within the Route Refresh message as shown in the following figure

```
 0               1               2               3
 0 1 2 3 4 5 6 7 0 1 2 3 4 5 6 7 0 1 2 3 4 5 6 7 0 1 2 3 4 5 6 7
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   Type        |            Length             |     Value     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                  Value (cont’d).                              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

The option Type is a 1 octet field that uniquely identifies individual options. The Length is a 2 octet field that contains the length of the option Value field in octets. The option Value is a variable length field that is interpreted according to the value of the option Type field.

The following types are being defined in this draft and additional types can be defined subsequently as needed

+ 1 - Route Type
+ 2 - NLRI Prefix
+ 3 - Route Distinguisher Prefix

The Route Type option would specify a particular route type that is being requested. This option applies specifically to those AFI/SAFI combinations that support multiple route types, e.g. L2VPN/EVPN and MUST be otherwise ignored. The value field would be the route type specifying which route type was being requested. The length of the option depends on the AFI/SAFI.

The NLRI Prefix option would specify a request for all matching address prefixes with their lengths equal to or greater than the specified prefix per AFI/SAFI definitions. The value field would contain the address prefix according to the NLRI specification of the AFI/SAFI contained in the Route Refresh message. For those AFI/SAFI combinations that specify NLRIs containing a type and/or RD, the value field MUST exclude the type and RD and SHOULD only include any remaining NLRI fields. If the requesting speaker expects its peer to also match the type and/or RD, the speaker CAN include the type and RD prefix options accordingly. The length field would contain the length of the value field in bits.

The Route Distinguisher prefix option would specify an RD prefix that is being requested for AFs that support it. The receiving BGP speaker would then refresh all routes in the specified AFI/SAFI that matched the requested RDs. The Value field would contain the RD, its length and the mask length of the RD prefix. This option applies specifically to those AFI/SAFI combinations that support route distinguishers and MUST be otherwise ignored.

10. Operation

A BGP speaker that understands and supports Route Refresh Options SHOULD advertise the Route Refresh Options Capability in its Open message. The following procedures for route refresh are only applicable if the BGP speaker originating the route refresh has received the route refresh options capability and supports it.

When originating a Route Refresh message, a BGP speaker SHOULD use and set these options if it wants to restrict the scope of updates being refreshed. The specific options being sent will be set according to the operator’s command.

When a BGP speaker receives a route refresh message that includes any options, it MUST parse the options and strongly SHOULD use them to filter outgoing NLRIs when refreshing the Adj-RIB-Out to the requesting BGP speaker.

If a BGP speaker receives the route refresh message with the message subtype set to BoRR with options as described above, then it needs to process all the included options and MUST mark all matching routes as stale as described in [RFC7313].

If a BGP speaker receives the route refresh message with the message subtype set to EoRR with options as described above, then it needs to process all the included options and delete any remaining stale routes that match the options received with the EoRR as described in [RFC7313].

A BGP speaker responding to a route refresh request MUST set the message subtypes of the BoRR and EoRR messages so that each BoRR message has a matching EoRR message. This means a BoRR message without options SHOULD only be followed eventually by an EoRR message without options. Similarly, a BoRR message with options MUST eventually be followed by an EoRR message with the same options. If BoRR and EoRR message options do not match, the outcome is unpredictable as remaining staled routes pending a refresh may get inadvertently deleted. BGP speakers MUST NOT summarize EoRR messages by combining options in order to allow the requesting BGP speaker to uniquely identify the included sets of routes when concurrent refreshes are originated with overlapping sets of routes.

Observe that overlapping refreshes with different options are possible and in such case the according BoRR and EoRR messages are associated by using their Refresh ID#. The BGP speaker responding to the route refresh requests MAY perform the refreshes in parallel. In case of concurrent refreshes overlapping same routes, the responding speaker MUST ensure that the sent advertisements will result in deletion of the omitted routes at the time all EoRRs have been received by the remote speaker or it MUST explicitly advertise withdrawals to correct any anomalies.

The BGP speaker requesting a refresh from its peers SHOULD maintain a locally configurable upper bound on how long it will keep matching stale routes once a BoRR has been received. Each subsequent BoRR SHOULD reset this period so that any remaining stale routes are only flushed after the last BoRR has been received in case there are multiple back-to-back refreshes being sent out and the last matching EoRR is never received or arrives too late. This is an implementation specific detail.

A BGP speaker may spontaneously originate a refresh to one or more of its peers depending on operator intervention, or due to a policy or configuration change, etc. In such a case, the speaker MUST refresh the entire Adj-RIB-Out. The speaker MUST also send BoRR/EoRR with the options field with the ’S’ flag set and a sequence number which lies outside the range of the sequence numbers that are currently in use with the receiving BGP speaker.

11. Error Handling

The handling of malformed options MUST follow the procedures mentioned in [RFC7606]. This draft obsoletes some of the error handling procedures in [RFC7313] if the Route Refresh Options Capability is sent. In addition, this draft mandates the following behavior at the receiver of the route refresh request upon detection of:

Length errors - If the message length minus the fixed-size message header is less than 4, the procedure in [RFC7313] MUST be followed. Also, if the overall length of all the options or any individual option length exceeds the total number of remaining bytes, the same procedure MUST be followed.

Option type errors - Any unknown option type CAN be ignored for AND’ed options. In case of OR’ed options the receiving speaker MUST ignore all the options and de-facto treat it as a full AFI/SAFI Adj-RIB-Out refresh. Such event SHOULD be logged in either case to notify the operator.

Option value errors - Length errors which cannot be distinguished from value field errors at the receiver are treated the same as value errors. The receiver MUST send a NOTIFICATION message with the Error Code "ROUTE-REFRESH Message Error" and the subcode of Invalid Message Length to the peer. The Data field of the NOTIFICATION message MUST contain the complete ROUTE-REFRESH message.

BoRR with "unknown" or "invalid" Refresh ID# - The receiver MUST discard all pending requests and issue a Route Refresh Request with Options. The options MUST be empty and the clear flag MUST be set to resynchronize the RIBs. "Unknown" means here a BoRR which is not in the interval

```
 [ MAX(lowest requested Refresh ID# without BoRR,
       highest received BoRR+1 respecting sequence number arithmetic),
   highest requested Refresh ID# ]
```

EoRR with unknown Refresh ID# - Those SHOULD be ignored and a warning or error MUST be logged.

BoRR or EoRR with incorrect options - analogous to BoRR with unknown Refresh ID#.

EoRR with known Refresh ID# but without preceding BoRR - analogous to EoRR with unknown Refresh ID#. Observe that this can be caused by the peer expiring last resort timer and reusing the ID# for another request before the EoRR is received. This should be extremely unlikely given the size of the refresh ID space.

12. IANA Considerations

This draft defines a new route refresh options format for BGP Route Refresh messages.

This draft defines a new route refresh capability for BGP Route Refresh messages. We request IANA to record this capability to create a new registry under BGP Capability Codes as follows:

+74 Route Refresh Options Capability

This draft defines 3 new route refresh message subtypes for BGP Route Refresh messages. We request IANA to record these subtypes to create a new registry under BGP Route Refresh Subcodes as follows:

+ 3 - Route Refresh with options
+ 4 - BoRR with options
+ 5 - EoRR with options

13. Security Considerations

This extension to BGP does not change the underlying security issues inherent in the existing [RFC7313] and [RFC4271].

14. Acknowledgements

The authors would like to thank Anant Utgikar for initial discussions resulting in this work. John Scudder and Jeff Hass provided further comments.

15. References

15.1. Normative References

[RFC1982] Elz, R. and R. Bush, "Serial Number Arithmetic", RFC 1982, DOI 10.17487/RFC1982, August 1996, <https://www.rfc-editor.org/info/rfc1982>.

[RFC2918] Chen, E., "Route Refresh Capability for BGP-4", RFC 2918, DOI 10.17487/RFC2918, September 2000, <https://www.rfc-editor.org/info/rfc2918>.

[RFC4684] Marques, P., Bonica, R., Fang, L., Martini, L., Raszuk, R., Patel, K., and J. Guichard, "Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC 4684, DOI 10.17487/RFC4684, November 2006, <https://www.rfc-editor.org/info/rfc4684>.

[RFC5291] Chen, E. and Y. Rekhter, "Outbound Route Filtering Capability for BGP-4", RFC 5291, DOI 10.17487/RFC5291, August 2008, <https://www.rfc-editor.org/info/rfc5291>.

[RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, DOI 10.17487/RFC5492, February 2009, <https://www.rfc-editor.org/info/rfc5492>.

[RFC7313] Patel, K., Chen, E., and B. Venkatachalapathy, "Enhanced Route Refresh Capability for BGP-4", RFC 7313, DOI 10.17487/RFC7313, July 2014, <https://www.rfc-editor.org/info/rfc7313>.

[RFC7606] Chen, E., Ed., Scudder, J., Ed., Mohapatra, P., and K. Patel, "Revised Error Handling for BGP UPDATE Messages", RFC 7606, DOI 10.17487/RFC7606, August 2015, <https://www.rfc-editor.org/info/rfc7606>.

15.2. Information References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <https://www.rfc-editor.org/info/rfc4271>.

[RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <https://www.rfc-editor.org/info/rfc7432>.

[RFC7752] Gredler, H., Ed., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and Traffic Engineering (TE) Information Using BGP", RFC 7752, DOI 10.17487/RFC7752, March 2016, <https://www.rfc-editor.org/info/rfc7752>.

[Wikipedia]
Wikipedia, "https://en.wikipedia.org/wiki/Serial_number_arithmetic", 2016.

Appendix A. Sequence Number Binary Arithmetic

The only reasonably reference to a cleaner than [RFC1982] sequence number solution is given in [Wikipedia]. It basically converts the problem into two complement's arithmetic. Assuming a straight two complement's substractions on the bit-width of the sequence number the according >: and =: relations are defined as:

U_1, U_2 are 12-bits aligned unsigned version number

D_f is (U_1 - U_2) interpreted as two complement signed 12-bits
D_b is (U_2 - U_1) interpreted as two complement signed 12-bits

U_1 >: U_2 IIF D_f > 0 AND D_b < 0
U_1 =: U_2 IIF D_f = 0

The >: relationsship is symmetric but not transitive. Observe that this leaves the case of the numbers having maximum two complement distance, e.g. (0 and 0x800) undefined in our 12-bits case since D_f and D_b are both -0x7ff.

A simple example of the relationship in case of 3-bit arithmetic follows as table indicating D_f/D_b values and then the relationship of U_1 to U_2:

U2 / U1	0	1	2	3	4	5	6	7
0	+/+	+/-	+/-	+/-	-/-	-/+	-/+	-/+
1	-/+	+/+	+/-	+/-	+/-	-/-	-/+	-/+
2	-/+	-/+	+/+	+/-	+/-	+/-	-/-	-/+
3	-/+	-/+	-/+	+/+	+/-	+/-	+/-	-/-
4	-/-	-/+	-/+	-/+	+/+	+/-	+/-	+/-
5	+/-	-/-	-/+	-/+	-/+	+/+	+/-	+/-
6	+/-	+/-	-/-	-/+	-/+	-/+	+/+	+/-
7	+/-	+/-	+/-	-/-	-/+	-/+	-/+	+/+

U2 / U1	0	1	2	3	4	5	6	7
0	=	>	>	>	?	<	<	<
1	<	=	>	>	>	?	<	<
2	<	<	=	>	>	>	?	<
3	<	<	<	=	>	>	>	?
4	?	<	<	<	=	>	>	>
5	>	?	<	<	<	=	>	>
6	>	>	?	<	<	<	=	>
7	>	>	>	?	<	<	<	=

Authors' Addresses

Keyur Patel
Arrcus, Inc
United States of America
Email: keyur@arrcus.com

Aamod Vyavaharkar
Cisco Systems
821 Alder Drive
Milpitas, CA 95035
United States of America
Email: avyavaha@cisco.com

Niloofar Fazlollahi
Unaffiliated
United States of America
Email: Niloofar_fazlollahi@yahoo.com

Tony Przygienda
Juniper Networks
1194 N. Mathilda Ave
Sunnyvale, CA 94089
United States of America
Email: prz@juniper.net

Krishnaswamy Ananthamurthy
Cisco Systems
821 Alder Drive
Milpitas, CA 95035
United States of America
Email: kriswamy@cisco.com