Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: May 1, 2021

H. Bidgoli, Ed.
Nokia
V. Voyer
Bell Canada
A. Stone
Nokia
P. Parekh
Cisco System
S. Krier
A. Venkateswaran
Cisco System, Inc.
October 28, 2020

Advertising p2mp policies in BGP
draft-hb-idr-sr-p2mp-policy-01

Abstract

SR P2MP policies are set of policies that enable architecture for P2MP service delivery.

A P2MP policy consists of candidate paths that connects the Root of the Tree to a set of Leaves. The P2MP policy is composed of replication segments. A replication segment is a forwarding instruction for a candidate path which is downloaded to the Root, transit nodes and the leaves.

This document specifies a new BGP SAFI with a new NLRI in order to advertise P2MP policy from a controller to a set of nodes.

This document introduces two new route types within this NLRI, one for P2MP policy and its candidate paths that need to be programmed on the Root node and another for the replication segment and forwarding instructions that needs to be programmed on the Root, and optionally on Transit and Leaf nodes.

It should be noted that this document does not specify how the Root and the Leaves are discovered on the controller, it only describes how the P2MP Policy and Replication Segments are programmed from the controller to the nodes.

Status of This Memo

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

Copyright Notice

Table of Contents

1. Introduction

The draft [draft-ietf-pim-sr-p2mp-policy] defines a variant of the SR Policy [draft-ietf-spring-segment-routing-policy] for constructing a P2MP segment to support multicast service delivery.

A Point-to-Multipoint (P2MP) Policy contains a set of candidate paths and identifies a Root node and a set of Leaf nodes in a Segment Routing Domain. The draft also defines a Replication segment, which corresponds to the state of a P2MP segment on a particular node. The Replication segment is the forwarding instruction for a P2MP LSP at the Root, Transit and Leaf nodes.

For a P2MP segment, a controller may be used to compute a tree from a Root node to a set of Leaf nodes, optionally via a set of replication nodes. A packet is replicated at the root node and optionally on Replication nodes towards each Leaf node.

We define two types of a P2MP segment: Spray and Replication.

A Point-to-Multipoint service delivery could be via Ingress Replication (aka Spray in some SR context), i.e., the root unicasts individual copies of traffic to each leaf. The corresponding P2MP segment consists of replication segments only for the root and the leaves.

A Point-to-Multipoint service delivery could also be via Downstream Replication (aka TreeSID in some SR context), i.e., the root and some downstream replication nodes replicate the traffic along the way as it traverses closer to the leaves.

It should be noted that two replication nodes can be connected directly, or they can be connected via unicast SR segment or a segment list.

The leaves and the root of a p2mp policy can be discovered via the multicast protocols or procedures like NG-MVPN [RFC6513] or manually configured on the PCC (CLI) or the PCE.

Base on the discovered root and leaves the controller builds a P2MP policy and advertise it to the head-end router (i.e. the root of the P2MP Tree). The advertisement uses BGP extensions defined in this document. In addition, the controller builds the replication segments on each segment of the tree, Root, Transit and Leaf nodes and downloads the forwarding instructions to the nodes via BGP extensions defined in this document.

As it was mentioned a SR p2mp policy is a variant of the SR policy and as such it reuses the concept of a candidate path. This draft reuses some of the concepts and TLVs mentioned in [draft-ietf-idr-segment-routing-te-policy]

A candidate path with in the P2MP policy can contain multiple path-instances. A path-instance can be viewed as a P2MP LSP. For candidate path global optimization purposes two or more path-instances can be used to execute make before break procedures.

Each path-instance is a P2MP LSP as such each path-instance needs a set of replication segments to construct its forwarding instructions.

2. Conventions used in this document

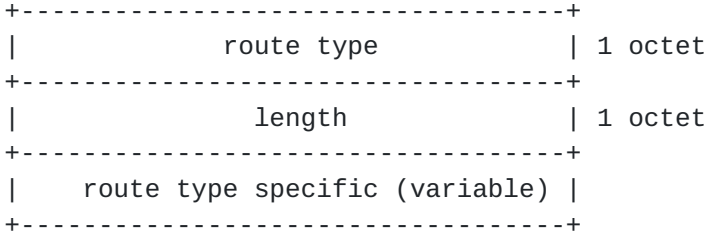
The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

3. P2MP Policy and Replication Segment Encoding

3.1. P2MP Policy SAFI and NLRI

This document defines a new BGP NLRI, called the P2MP-POLICY NLRI.

A new SAFI is defined: the SR P2MP Policy SAFI, (Codepoint tbd assigned by IANA). The following is the format of the P2MP-POLICY NLRI:



- The Route type field defines the encoding of the rest of the P2MP-POLICY NLRI.
- The length field indicates the length in octets of the route type.
- This document defines the following route types:
  - 2MP Policy route
  - Non-Shared Replication Segment
  - Shared Replication Segment

The NLRI containing the SR P2MP Policy is carried in a BGP UPDATE message [RFC4271] using BGP multiprotocol extensions [RFC4760] with an AFI of 1 or 2 (IPv4 or IPv6) and with a SAFI of "TBD" (assigned by IANA from the "Subsequent Address Family Identifiers (SAFI) Parameters" registry).

All other recommendations of [draft-ietf-idr-segment-routing-te-policy] section SR Policy SAFI and NLRI, should be taken into account for P2MP policy.

3.1.1. P2MP Policy Route

- Root-ID: IPv4/IPv6 address of the head-end (root) of the p2mp tree
- Tree-ID: a unique 4 octets identifier of the p2mp tree on the head- end (root)router.
- Distinguisher: 4-octet value uniquely identifying the policy in the context of <Tree-ID, Originating Router's IP> tuple. The distinguisher has no semantic value and is solely used by the SR P2MP Policy originator to make unique (from an NLRI perspective) multiple occurrences of the same SR P2MP Policy.

3.1.2. Non-shared Tree Replication segment Route

A non-shared tree is used when the label field of the PMSI Tunnel Attribute (PTA) is set to 0 as per [draft-parekh-bess-mvpn-sr-p2mp]. In short this route type is used when there is no upstream assigned label in the PTA and aggregate of MVPNs into one P-Tunnel is not desired.

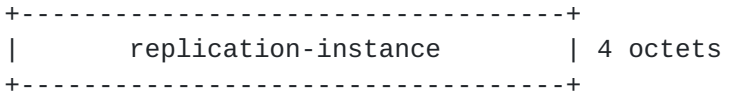
- Root-ID: IPv4/IPv6 address of the head-end (root) of the p2mp tree

- Tree-ID: a unique 4 octets identifier of the p2mp tree on the head- end (Root)router

- path-instance-id, identifies the path-instance with in the p2mp-policy. Each candidate path can have one, two or more path-instance. Path-instance is used for global optimization of the candidate path via make before break procedures.

3.1.3. Shared Tree Replication Segment Route

A shared tree is used when the label field of the PTA is NOT set to Zero. This route type is used when there is an upstream assigned label in the PTA and aggregate of MVPNs into one P-Tunnel is desired.



- replication-instance: is a unique identifier of the replication segment on a specific node. Each node can assign its own replication- id for a replication segment.

3.2. Tunnel Encapsulation Attribute

The content of this new NLRI is encoded in the tunnel Encapsulation Attribute originally defined in [ietf-idr-tunnel-encaps] using two new Tunnel-Type TLV (codepoint is TBD, assigned by IANA from the "BGP Tunnel Encapsulation Attribute Tunnel Types" registry) one for P2MP Policy and another for Replication segment.

3.2.1. SR P2MP policy encoding

```
SR P2MP Policy SAFI NLRI: <route-type p2mp-policy>
Attributes:
   Tunnel Encaps Attribute (23)
      Tunnel Type: (TBD, P2MP-Policy)
          Preference
          Policy Name
          leaf-list (optional)
               remote-end point
               remote-end point
               ...
          path-instance
              active-instance-id
              instance-id
              instance-id
              ...
```

- SR P2MP-POLICY NLRI and P2MP Policy route type.

- Tunnel Encapsulation Attribute is defined in [ietf-idr-tunnel-encaps]

- Tunnel-Type is set to P2MP-Policy Tunnel-Type TBD (assigned by IANA from the "BGP Tunnel Encapsulation Attribute Tunnel Types" registry).

- Preference, leaf-list, remote-end point, Policy Name, path-instance, instance are defined in this document.

- Additional sub-TLVs may be defined in the future.

3.2.2. replication segment encoding

```
replication segment SAFI NLRI: <route-type non-sahred/shared
                                tree replication-segment>
     Attributes:
         Tunnel Encaps Attribute (23)
            Tunnel Type: (TBD Replication-Segment)
                replication-sid (equivalent to binding Sid)
                downstream-nodes
                    segment-list
                        segment
                        segment
                        ...
                    segment-list
                        segment
                        segment
                        ...
                    ...
```

- SR P2MP-POLICY NLRI and non-shared tree Replication segment route type or shared tree Replication segment route type.

- Tunnel Encapsulation Attribute is defined in [ietf-idr-tunnel-encaps].

- Tunnel-Type is set to Replication Segment Tunnel Type, TBD (assigned by IANA from the "BGP Tunnel Encapsulation Attribute Tunnel Types" registry).

- tree-identifier, replication-sid (binding sid), down-stream-nodes, segemtn-list and segment-list are defined in this document.

- Additional sub-TLVs may be defined in the future.

3.3. P2MP Policy Sub-TLVs

EACH P2MP policy NLRI represents a candidate path for a P2MP policy. A P2MP policy can have multiple candidate paths and would need multiple P2MP policy NRLI to download all the candidate paths.

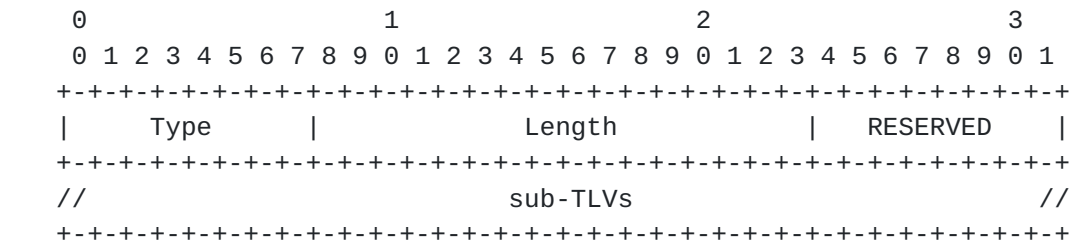
3.3.1. preference Sub-TLV

As defined in preference Sub-TLV section in [draft-ietf-idr-segment-routing-te-policy]

3.3.2. leaf-list Sub-TLV

The leaf list sub-tlv identifies a set of leaves for the tree. Each leaf is a remote endpoint as defined in [ietf-idr-tunnel-encaps] The leaf-list sub-tlv is optional. The PCE can choose to download the

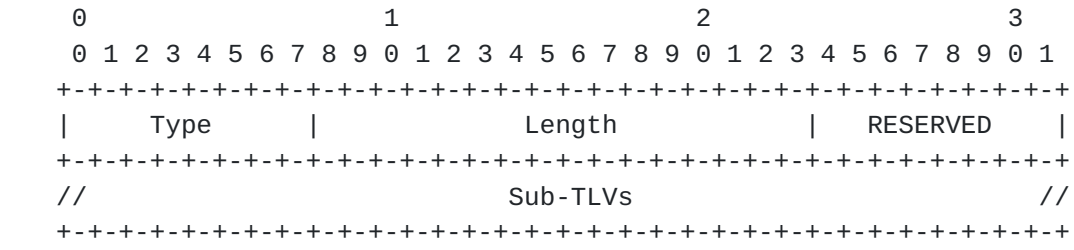
leaf list every time it is configured or learns a new leaf. If the PCE chooses to download this optional sub-tlv it should download the entire set of the end-points every time the endpoint list has been modified. The leaf list has informational value but is optional since it is not required for the root to operate. However, it must be noted that in some cases the end-points list can become very large with 100s of leaves.



- Type: TBD
- Length: the total length (not including the Type and Length fields) of the sub-TLVs encoded within the leaf-list sub-TLV.
- RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt.
- sub-TLVs: One or more remote endpoint sub-TLVs. Note the remote endpoint object is defined in [ietf-idr-tunnel-encaps]

3.3.3. path-instance Sub-TLV

The path instance sub-tlv contains a set of instance-ids (P2MP LSPs). These LSPs can be used for MBB procedure under a candidate path. Each LSP Instance-id has a unique id (4 octets) with in the root and the P2MP policy. The PCE SHOULD always download all instance-ids to the node.



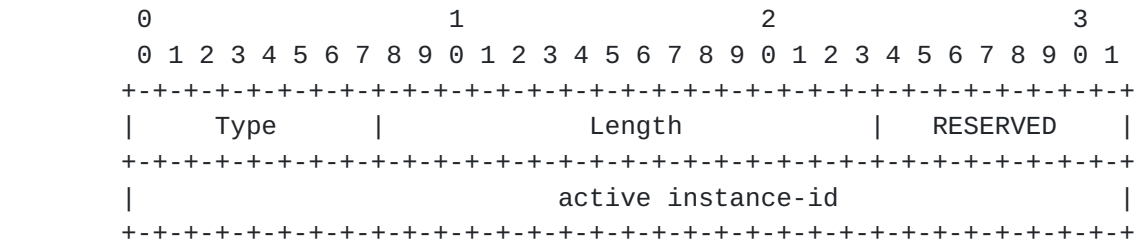
- Type: TBD
- Length: the total length (not including the Type and Length fields) of the sub-TLVs encoded within the Segment List sub-TLV.

- RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt

- sub-TLVs: * active instance-id * one or more instance-id

3.3.3.1. active instance-id Sub-TLV

An instance-id is equivalent to a P2MP LSP. The instance-id is unique in context of the <root node,p2mp policy> in other word is unique per <root node,tree-id>. The Active instance-id is used to identify the P2MP LSP which should be active amongst the collection of LSPs.



- Type: TBD.

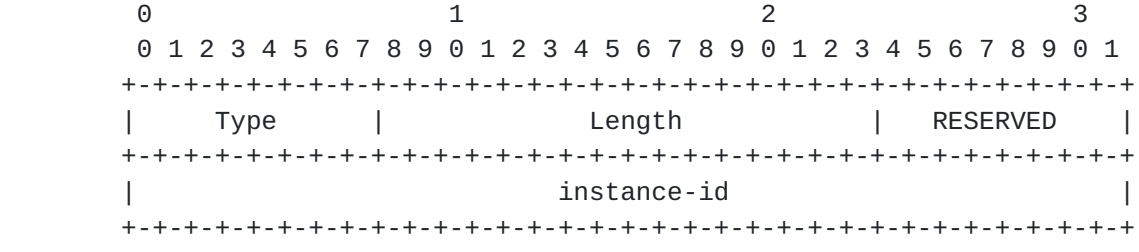
- Length: the total length (not including the Type and Length fields) of the sub-TLVs encoded within the Segment List sub-TLV.

- RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt.

- active instant-id: The identifier of the active instance-id

3.3.3.2. instance-id Sub-TLV

An instance-id is equivalent to a P2MP LSP. The instance-id is unique in context of the <root node, p2mp policy> in other word is unique per <root node,tree-id>



- Type: TBD

- Length: the total length (not including the Type and Length fields) of the sub-TLVs encoded within the Segment List sub-TLV.

- RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt.

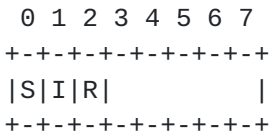
- instan-id: a 32 bit unique identifier. The instance-id is unique with in the context of the <root node, p2mp policy>

3.4. replication segment Sub-TLVs

3.4.1. Replication SID (Binding SID)

The Replication SID is form of a Binding SID as it is defined in [draft-ietf-idr-segment-routing-te-policy]. The definition of replication sid with in P2MP Policy is defined in [draft-ietf-spring-sr-replication-segment]. On the transit and leaf node the replication SID can be used to identify the replication segment and the forwarding information at the node. How ever on the head-end node (Root), the replication segment acts as a Binding SID to direct the traffic into the P2MP Tree. It should be noted that two replication SIDs can be directly connected or connected via a SR binding SID or node/adjacency SID.

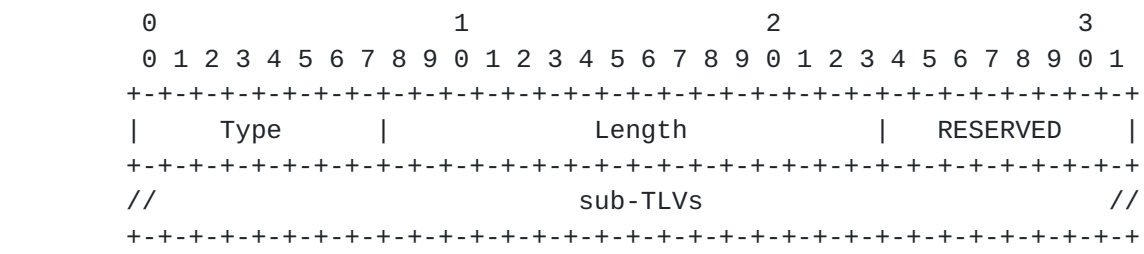
As it was mentioned earlier the sr-te-policy binding sid sub-tlv is used for replication sid. This draft defines a new flag for replication sid at transit and leaf node



R-FLAG: is Replication SID. Replication SID can be used to define the forwarding information of the transit or leaf nodes.

3.4.2. down stream nodes Sub-TLV

The down-stream nodes sub-tlv is the list of down stream nodes for this replication segment. Two replication segments can be directly connected or they can be connected via a sr segment-list. As such the down stream nodes sub-tlv is a list of segment-lists. Each segment- list connects two replication segments via a replication sid or a segment list.



- Type: TBD.
- Length: the total length (not including the Type and Length fields) of the sub-TLVs encoded within the down-stream nodes sub-TLV.
- RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt. o
- sub-TLVs: One or more segment list sub-TLVs.

3.4.3. segment list Sub-TLV

The segment list Sub-TLV is defined in [ietf-spring-segment-routing-policy]. It should be noted that P2MP policy the optional weight Sub-TLV is not used and can optionally be set to 1. The segment-list Sub-TLV contains zero or more segment Sub-TLVs.

3.4.4. segment Sub-TLV

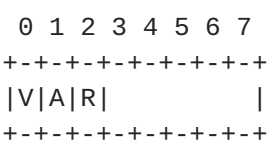
The segment sub-Tlv is identified in [draft-ietf-idr-segment-routing-te-policy]. As it was mentioned before two replication segments can be connected directly to each other or via a segment list. If they are connected directly to each other then the segment list can be constructed via:

- If the replication segment is steered via IPv4 or IPv6 nexthops or interface then the segment type E or G can be used with the new R flag set.
- If the replication segment is steered via a SR Unicast node or adjacency SID then segment type A can be used with the new R flag set. Unicast SR segment types can also be configured for steering.

If they are connected via SR domain then the segment list can contain multiple different types of SIDs, such as Node, Adjacency or Binding SIDe. In this case the replication sid is at the bottom of the stack and of type A with the R flag set. The SR node/adjacency or binding

sids steer the packet through a SR domain until it reaches another replication segment. where the bottom of the stack replication sid identifies the forwarding information on that replication segment.

A replication segment can use the same type of segment types defined in [draft-ietf-idr-segment-routing-te-policy]. To identify a replication segment explicitly a new flag is defined.



Where R-Flag is set for a segment Sub-TLV that identifies a Replication Segment. It should be noted that in a segment list only the last segment can have the R flag set. Multiple replication segments can not be stacked on top of each other. That said there can be special cases for Link Protection where a bypass tunnel is build via a shared replication segment. As an example when the PCE downloads a bypass tunnel for link protection that is only constructed via shared replication segments to protect a group of non-shared replication segments.

4. P2MP Policy Operation

Inline with [draft-ietf-idr-segment-routing-te-policy] the consumer of an P2MP Policy is not the BGP process. The BGP process is used for distributing the P2MP policy NLRI and its route-types but its installation and use is outside the scope of BGP. The detail for P2MP Policy can be found in [draft-ietf-pim-sr-p2mp-policy]

4.1. Configuration and advertisement of P2MP Policies

The controller usually is connected to the receivers via a route reflector. As such one or more route-target SHOULD be attached to the advertisement of P2MP Policy NLRI and its route-type. Each route target identifies one head-end (root nodes) for P2MP Policy route or one or more head-end, transit and leaf nodes for the Non- Shared/ Shared Tree Replication Segment route, for the advertised P2MP Policy.

If no route-target is attached to the NLRI, then it is assumed that the originator sends the P2MP Policy update directly to the intended receiver. In such case, the NO_ADVERTISE community MUST be attached to the P2MP Policy update.

4.2. Reception of an P2MP Policy NLRI

When a BGP speaker receives an P2MP Policy NLRI the following rules apply:

- The P2MP Policy update MUST have either the NO_ADVERTISE community or at least one route-target extended community in IPv4-address format. If a router supporting this document receives an P2MP Policy update with no route-target extended communities and no NO_ADVERTISE community, the update MUST NOT be processed. Furthermore, it SHOULD be considered to be malformed, and the "treat-as-withdraw" strategy of [RFC7606] is applied.

- If one or more route-targets are present, then at least one route-target MUST match one of the BGP Identifiers of the receiver in order for the update to be considered usable. The BGP Identifier is defined in [RFC4271] as a 4 octet IPv4 address. Therefore the route- target extended community MUST be of the same format.

- If one or more route-targets are present and no one matches any of the local BGP Identifiers, then, while the P2MP Policy NLRI is acceptable, it is not usable on the receiver node.

4.3. Global Optimization for P2MP LSPs

When a P2MP LSP needs to be optimized for any reason (i.e. it is taking on an FRR Path or new routers are added to the network) a global optimization is possible. Note that optimization works per candidate path. Each candidate path is capable of global optimization. To do so each candidate path contains two or more path- instances. Each path instance is a P2MP LSP, each P2MP LSP is identified via a path-instance-id (equivalent to an lsp-id [RFC3209]). After calculating an optimized P2MP LSP path the PCE will program the candidate path with a 2nd path instance and its set of replication segments for this path-instance on the root, transit and leaf nodes. After the optimized LSP replication segments are downloaded a MBB procedure is performed and the previous instance of the path instance is deleted and removed from head-end node and its corresponding replication segments from head-end, transit and leaves.

5. IANA Consideration

- A new SAFI is defined: the SR P2MP Policy SAFI, (Codepoint tbd assigned by IANA)

- 3 new Route type field defines the encoding of the rest of the P2MP- POLICY SAFI

  * P2MP Policy Route

  * Non-Shared Replication Segment

  * Shared Replication Segment

- Two new Tunnel type to be assigned by IANA

6. Security Considerations

TBD

7. Acknowledgments

8. References

8.1. Normative 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>.

8.2. Informative References

[draft-ietf-idr-segment-routing-te-policy]
.

[draft-ietf-pim-sr-p2mp-policy]
"D. Yoyer, C. Filsfils, R.Prekh, H.bidgoli, Z. Zhang, "draft-ietf-pim-sr-p2mp-policy"", October 2019.

[draft-ietf-spring-segment-routing-policy]
.

[draft-ietf-spring-sr-replication-segment]
"D. Yoyer, C. Filsfils, R.Prekh, H.bidgoli, Z. Zhang, "draft-ietf-pim-sr-p2mp-policy "", July 2020.

[draft-parekh-bess-mvpn-sr-p2mp]
.

[ietf-idr-tunnel-encaps]
.

[ietf-spring-segment-routing-policy]
.

[RFC4271] .

[RFC4760] .

[RFC6513] .

[RFC7606] .

Authors' Addresses

Hooman Bidgoli (editor)
Nokia
Ottawa
Canada

Email: hooman.bidgoli@nokia.com

Daniel Voyer
Bell Canada
Montreal
Canada

Email: daniel.yover@bell.ca

Andrew Stone
Nokia
Ottawa
Canada

Email: andrew.stone@nokia.com

Rishabh Parekh
Cisco System
San Jose
USA

Email: riparekh@cisco.com

Serge Krier
Cisco System, Inc.

Email: sekrier@cisco.com

Arvind Venkateswaran
Cisco System, Inc.
Ottawa
Canada

Email: arvvenka@cisco.com