IDR Working Group   W,. Hao
Internet-Draft   D. Eastlake
Intended status: Standards Track   S. Hares
Expires: September 21, 2016   Huawei Technologies
                                                         B. Pithawala
                                                           IP Infusion
                                                           M. Durrani
                                                         Cisco Systems
                                                                Y. Li
                                                  Huawei Technologies
                                                       March 20, 2016

Distribution of TRILL Link-State using BGP
draft-ietf-idr-ls-trill-01.txt

Abstract

   This draft describes a TRILL link state and MAC address reachability
   information distribution mechanism using a BGP LS extension.
   External components such as an SDN Controller can use the information
   for topology visibility, troubleshooting, network automation, etc.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

   BGP has been extended to distribute IGP link-state and traffic
   engineering information to some external components [RFC7752] such as
   the PCE and ALTO servers. The information can be used by these
   external components to compute a MPLS-TE path across IGP areas,
   visualize and abstract network topology, and the like.

   TRILL (Transparent Interconnection of Lots of Links) protocol
   [RFC6325] provides a solution for least cost transparent routing in
   multi-hop networks with arbitrary topologies and link technologies,
   using [IS-IS] [RFC7176] link-state routing and a hop count. TRILL
   switches are sometimes called RBridges (Routing Bridges).

   The TRILL protocol has been deployed in many data center networks.
   Data center automation is a vital step to increase the speed and
   agility of business. An SDN controller as an external component
   normally can be used to provide centralized control and automation
   for the data center network. Making a holistic view of whole network
   topology available to the SDN controller is an important part for
   data center network automation and troubleshooting.

```
                        +-------------+
                        |    SDN      |
                --------| Controller  |--------
                |       +-------------+        |
                |                              |
              + +                            + +
               +           +----------+        +
                           |          |
            +--------+     |IP Network|     +--------+
            |        |     |          |     |        |
  +---+ +---+|        |  +----+  +----+ |        |+---+ +---+
  |ES1|-|RB1|-| Area 1 |-|BRB1|  |BRB2|-| Area 2 |-|RB2|-|ES2|
  +---+ +---+|        |  +----+  +----+ |        |+---+ +---+
            |        |     |          |     |        |
            +--------+     +----------+     +--------+

     |<----TRILL ------>|<IP tunnel>|<-----TRILL ----->|
```

                     Figure 1: TRILL interconnection

   In Data Center interconnection scenario illustrated in figure 1, a
   single SDN Controller or network management system (NMS) can be used
   for end-to-end network management. End-to-end topology visibility on
   the SDN controller or NMS is very useful for whole network automation
   and troubleshooting. BGP LS can be used by the external SDN
   controller to collect multiple TRILL domain's link-state.

   If ESADI (End Station Address Distribution Information) protocol
   [RFC7357] is used for control plane MAC learning in each data center,
   BGP LS also can be used for MAC address reachability information
   synchronization across multiple TRILL domains. End-to-end unicast
   forwarding paths can be calculated based on the synchronized
   information.

   This document describes the detailed BGP LS extension mechanisms for
   TRILL link state and MAC address reachability information
   distribution.

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 [RFC2119]

   BGP - Border Gateway Protocol

   BGP-LS - BGP Link-State

Data label - VLAN or FGL (Fine Grained Label) [RFC7172]

IS - Intermediate System (for this document, all relevant
intermediate systems are RBridges).

NLRI - Network Layer Reachability Information

SDN - Software Defined Networking

RBridge - A device implementing the TRILL protocol

TRILL - Transparent Interconnection of Lots of Links

3. Carrying Trill Link-State Information in BGP

In [RFC7752], four NLRI types are defined as follows: Node NLRI, Link
NLRI, IPv4 Topology Prefix NLRI and IPv6 Topology Prefix NLRI. For
TRILL link-state distribution, the Node NLRI and Link NLRI are
extended to carry layer 3 gateway role and link MTU information.
TRILL specific attributes are carried using opaque Node Attribute
TLVs, such as nickname, distribution tree number and identifiers,
interested VLANs/Fine Grained Label, and multicast group address, and
etc.

To differentiate TRILL protocol from layer 3 IGP protocol, a new
TRILL Protocol-ID is defined.

```
      +------------+--------------------------------+
      | Protocol-ID | NLRI information source protocol |
      +------------+--------------------------------+
      | 1          | IS-IS Level 1                  |
      | 2          | IS-IS Level 2                  |
      | 3          | OSPFv2                         |
      | 4          | Direct                         |
      | 5          | Static configuration           |
      | 6          | OSPFv3                         |
      | TBD        | TRILL                          |
      +------------+--------------------------------+
```
             Table 1: Protocol Identifiers

ESADI (End Station Address Distribution Information) protocol
[RFC7357] is a per data label control plane MAC learning solution.
MAC address reachability information is carried in ESADI packets.
Compared with data plane MAC learning solution, ESADI protocol has
security and fast update advantage that are pointed out in [RFC7357].

For an RBridge that is announcing participation in ESADI, the RBridge
can distribute MAC address reachability information to external

components using BGP. A new NLRI type of ''MAC Reachability NLRI''
is requested for the MAC address reachability distribution.

```
        +------+---------------------------+
        | Type | NLRI Type                 |
        +------+---------------------------+
        | 1    | Node NLRI                 |
        | 2    | Link NLRI                 |
        | 3    | IPv4 Topology Prefix NLRI |
        | 4    | IPv6 Topology Prefix NLRI |
        | TBD  | MAC Reachability NLRI     |
        +------+---------------------------+
              Table 2: NLRI Types
```

The MAC Reachability NLRI uses the format as shown in the following
figure.
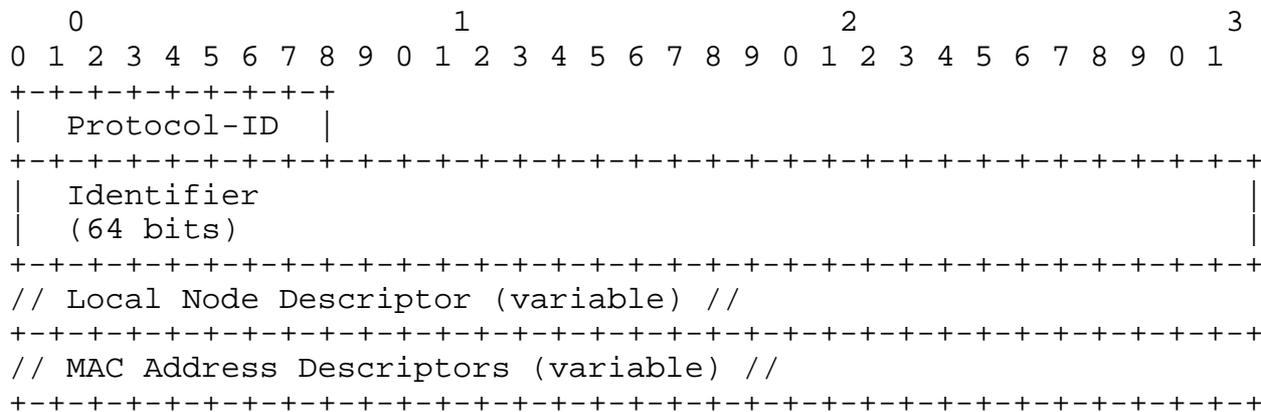
```
    0                   1                   2                   3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
   +-+-+-+-+-+-+-+-+
   |  Protocol-ID  |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |  Identifier                                                   |
   |  (64 bits)                                                    |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   // Local Node Descriptor (variable) //
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   // MAC Address Descriptors (variable) //
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
              Figure 2: The MAC Reachability NLRI format
```

3.1. Node Descriptors

   The Node Descriptor Sub-TLV types include Autonomous System and BGP-
   LS Identifier, iS-IS Area-ID and IGP Router-ID. TRILL uses a fixed
   zero Area Address as specified in [RFC6325], Section 4.2.3. This is
   encoded in a 4-byte Area Address TLV (TLV #1) as follows:
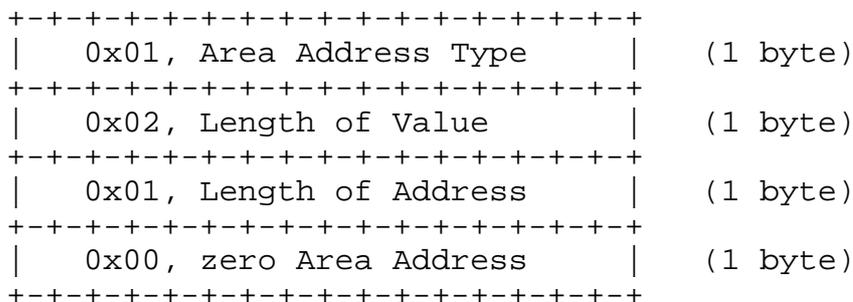
```
        +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
        |   0x01, Area Address Type     |   (1 byte)
        +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
        |   0x02, Length of Value       |   (1 byte)
        +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
        |   0x01, Length of Address     |   (1 byte)
        +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
        |   0x00, zero Area Address     |   (1 byte)
        +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
          Figure 3: Area Address TLV
```

3.1.1. IGP Router-ID

   Similar to layer 3 IS-IS, TRILL protocol uses 7-octet "IS-IS ID" as
   the identity of an RBridge or a pseudonode, IGP Router ID sub-TLV in
   Node Descriptor TLVs contains the 7-octet "IS-IS ID". In TRILL
   network, each RBridge has a unique 48-bit (6-octet) IS-IS System ID.
   This ID may be derived from any of the RBridge's unique MAC addresses
   or configured. A pseudonode is assigned a 7-octet ID by the DRB
   (Designated RBridge) that created it, the DRB is similar to the
   "Designated Intermediate System" (DIS) corresponding to a LAN.

3.2. MAC Address Descriptors

   The ''MAC Address Descriptor'' field is a set of Type/Length/Value
   (TLV) triplets. ''MAC Address Descriptor'' TLVs uniquely identify an
   MAC address reachable by a Node. The following attributes TLVs are
   defined:

   +--------------+----------------------+----------+-----------------+
   | TLV Code | Description | Length | Value defined |
   | Point | | | in: |
   +--------------+----------------------+----------+-----------------+
   | 1 | MAC-Reachability | variable | section 3.2.1 |
   +--------------+----------------------+----------+-----------------+
              Table 3: MAC Address Descriptor TLVs

3.2.1. MAC-Reachability TLV
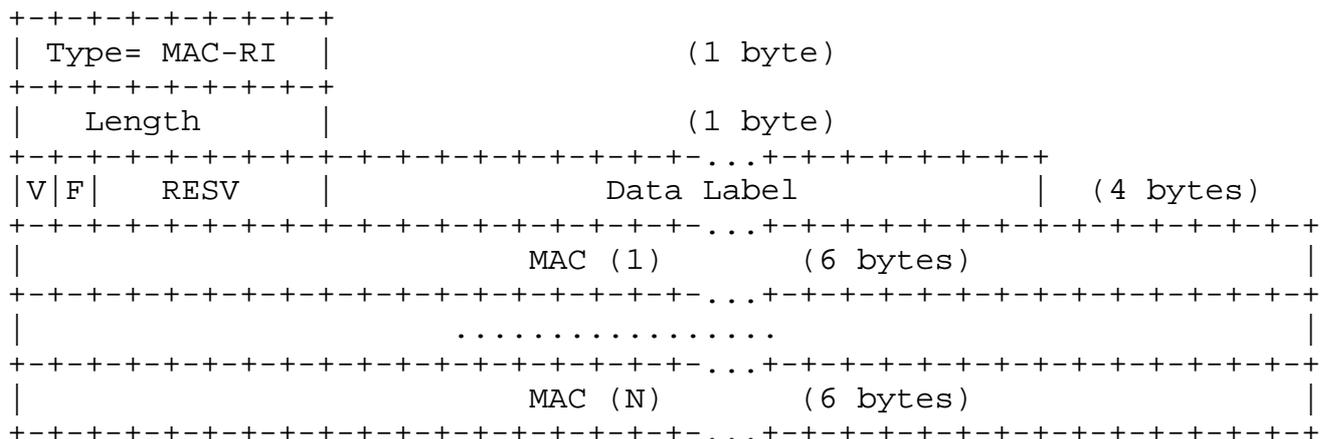
      +-+-+-+-+-+-+-+-+
      | Type= MAC-RI | (1 byte)
      +-+-+-+-+-+-+-+-+
      | Length | (1 byte)
      +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-...-+-+-+-+-+-+-+
      |V|F| RESV | Data Label | (4 bytes)
      +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-...-+-+-+-+-+-+-+-+-+-+-+-+-+
      | MAC (1)   (6 bytes) |
      +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-...-+-+-+-+-+-+-+-+-+-+-+-+-+
      | .... |
      +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-...-+-+-+-+-+-+-+-+-+-+-+-+-+
      | MAC (N)   (6 bytes) |
      +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-...-+-+-+-+-+-+-+-+-+-+-+-+-+
   Figure 4: MAC-Reachability TLV format

   Length is 4 plus a multiple of 6.

The bits of 'V' and 'F' are used to identify Data Label type and are
defined as follows:

```
+----------+-------------------------+
| Bit      | Description             |
+----------+-------------------------+
| 'V'      | VLAN                    |
| 'F'      | Fine Grained Label      |
+----------+-------------------------+
```
        Table 4: Data Label Type Bits Definitions

Notes: If BGP LS is used for NVO3 network MAC address distribution
between external SDN Controller and NVE, Data Label can be used to
represent 24 bits VN ID.

3.3. BGP-LS attribute

3.3.1. Node Attribute TLVs

3.3.1.1. Node Flag Bits TLV

A new Node Flag bit is added as follows:
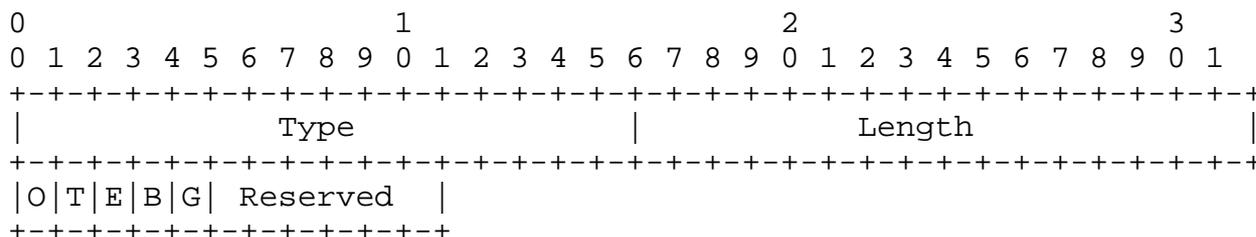
```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|              Type             |            Length             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|O|T|E|B|G| Reserved  |
+-+-+-+-+-+-+-+-+-+-+-+
```
           Figure 5: Node Flag Bits TLV format

The new bit and remaining reserved bits are defined as follows:

```
+----------+----------------------------+-----------+
| Bit      |         Description         | Reference |
+----------+----------------------------+-----------+
| 'G'      | Layer 3 Gateway Bit        | [RFC7176] |
| Reserved | Reserved for future use    |           |
+----------+----------------------------+-----------+
```
       Table 5: Node Flag Bits Definitions

3.3.1.2. Opaque Node Attribute TLV

The Opaque Node Attribute TLV is used as the envelope to
transparently carry TRILL specific information. In [RFC7176], there
are the following Sub-TLVs in the Router Capability and MT-
Capability TLVs and the Group Address (GADDR) TLV that need to be

carried. Future possible TRILL TLVs/Sub-TLVs extension also can be
carried using the Opaque Node Attribute TLV.

```
Descriptions           IS-IS TLV/Sub-TLV
---------------------------------------
TRILL-VER              22/13
NICKNAME               22/6
TREES                  22/7
TREE-RT-IDs            22/8
TREE-USE-IDs           22/9
INT-VLAN               22/10
VLAN-GROUP             22/14
INT-LABEL              22/15
RBCHANNELS             22/16
AFFINITY               22/17
LABEL-GROUP            22/18
GMAC-ADDR              142/1
GIP-ADDR               142/2
GIPV6-ADDR             142/3
GLMAC-ADDR             142/4
GLIP-ADDR              142/5
GLIPV6-ADDR            142/6
```

Table 6: TRILL TLVs/Sub-TLVs

3.3.2. Link Attribute TLVs

Link attribute TLVs are TLVs that may be encoded in the BGP-LS
attribute with a link NLRI. Besides the TLVs that has been defined
in [RFC7752] section 3.3.2 table 9, the following 'Link Attribute'
TLV is provided for TRILL.

TLV Code Point	Description	IS-IS TLV /Sub-TLV	Defined in:
TBD	Link MTU	22/28	[RFC7176]/2.4

Table 7: Link Attribute TLVs

4. Operational Considerations

This document does not require any MIB or Yang model to configure
operational parameters.

An implementation of this specification[idr-ls-trill], MUST do the
malformed attribute checks below, and if it detects a malformed
attribute, it should use the 'Attribute Discard' action per [RFC7606]
section 2.

An implementation MUST perform the following expanded [BGP-LS]
syntactic check for determining if the message is malformed:

o Does the sum of all TLVs found in the BGP LS attribute correspond
   to the BGP LS path attribute length ?

o Does the sum of all TLVs found in the BGP MP_REACH_NLRI attribute
   correspond to the BGP MP_REACH_NLRI length ?

o Does the sum of all TLVs found in the BGP MP_UNREACH_NLRI
   attribute correspond to the BGP MP_UNREACH_NLRI length ?

o Does the sum of all TLVs found in a Node-, Link, prefix (IPv4 or
   IPv6) NLRI attribute correspond to the Node-, Link- or Prefix
   Descriptors 'Total NLRI Length' field ?

o Does any fixed length TLV correspond to the TLV Length field in
   this document ?

o Does the sum of MAC reachability TLVs equal the length of the
   field?

In addition, the following checks need to be made for the fields
specific to the BGP LS for TRILL:

o PROTOCOL ID is TRILL,

o NLRI types are valid per table 2,

o MAC Reachability NLRI has correct format including:

   * Identifier (64 bits),

   * local node descriptor with AREA address TLV has the form found
      in figure 2,

o opaque TLV support the range of ISIS-TLV/SUB-TLV shown in table 3,
   and link TLVs support the range in figure 8.

5. Security Considerations

   Procedures and protocol extensions defined in this document do not
   affect the BGP security model. See [RFC6952] for details.

6. IANA Considerations

   This section complies with [RFC7153]. For all of the following
   assignments, [this document] is the reference.

   IANA is requested to requested to assign one Protocol-ID for "TRILL"
   from the BGP-LS registry of Protocol-IDs

   IANA is requested to assign one NLRI Type for "MAC Reachability" from
   the BGP-LS registry of NLRI Types.

   IANA is requested to assign one Node Flag bit for "Layer 3 Gateway"
   from the BGP-LS registry of BGP-LS Attribute TLVs.

   IANA is requested to assign one new TLV type for "Link MTU" from the
   BGP-LS registry of BGP-LS Attribute TLVs.

7. Acknowledgements

   Authors like to thank Andrew Qu, Jie Dong, Mingui Zhang, Qin Wu,
   Shunwan Zhuang, Zitao Wang, Lili Wang for their valuable inputs.

8. 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,
               <http://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,
               <http://www.rfc-editor.org/info/rfc4271>.

   [RFC6325]   Perlman, R., Eastlake 3rd, D., Dutt, D., Gai, S., and A.
               Ghanwani, "Routing Bridges (RBridges): Base Protocol
               Specification", RFC 6325, DOI 10.17487/RFC6325, July 2011,
               <http://www.rfc-editor.org/info/rfc6325>.

   [RFC6952] Jethanandani, M., Patel, K., and L. Zheng, "Analysis of
              BGP, LDP, PCEP, and MSDP Issues According to the Keying
              and Authentication for Routing Protocols (KARP) Design
              Guide", RFC 6952, DOI 10.17487/RFC6952, May 2013,
              <http://www.rfc-editor.org/info/rfc6952>.

   [RFC7153] Rosen, E. and Y. Rekhter, "IANA Registries for BGP
              Extended Communities", RFC 7153, DOI 10.17487/RFC7153,
              March 2014, <http://www.rfc-editor.org/info/rfc7153>.

   [RFC7172] Eastlake 3rd, D., Zhang, M., Agarwal, P., Perlman, R., and
              D. Dutt, "Transparent Interconnection of Lots of Links
              (TRILL): Fine-Grained Labeling", RFC 7172,
              DOI 10.17487/RFC7172, May 2014,
              <http://www.rfc-editor.org/info/rfc7172>.

   [RFC7176] Eastlake 3rd, D., Senevirathne, T., Ghanwani, A., Dutt,
              D., and A. Banerjee, "Transparent Interconnection of Lots
              of Links (TRILL) Use of IS-IS", RFC 7176,
              DOI 10.17487/RFC7176, May 2014,
              <http://www.rfc-editor.org/info/rfc7176>.

   [RFC7357] Zhai, H., Hu, F., Perlman, R., Eastlake 3rd, D., and O.
              Stokes, "Transparent Interconnection of Lots of Links
              (TRILL): End Station Address Distribution Information
              (ESADI) Protocol", RFC 7357, DOI 10.17487/RFC7357,
              September 2014, <http://www.rfc-editor.org/info/rfc7357>.

   [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,
              <http://www.rfc-editor.org/info/rfc7606>.

   [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,
              <http://www.rfc-editor.org/info/rfc7752>.

Authors' Addresses

Weiquo Hao
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China

Phone: +86-25-56623144
Email: haoweiguo@huawei.com

Donald E. Eastlake
Huawei Technologies
155 Beaver Street
Milford , MA 01757
USA

Phone: +1-508-333-2270
Email: d3e3e3@gmail.com

Susan Hares
Huawei Technologies
7453 Hickory
Saline , MI 48176
USA

Phone: +1-734-604-0332
Email: shares@ndzh.com

Sujay Gupta
IP Infusion

Email: sujay.gupta@ipinfusion.com

Muhammad Durrani
Cisco Systems

Phone: +1-408-527-6921
Email: mdurrani@cisco.com

Yizhou Li
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China

Email: liyizhou@huawei.com