TRILL Working Group   W. Hao
INTERNET-DRAFT   Y. Li
Intended Status: Standard Track   Huawei Technologies
                                                          M. Durrani
                                                               Cisco
                                                            S. Gupta
                                                          IP Infusion
                                                               A. Qu
                                                            MediaTec
                                                               T. Han
                                              Huawei Technologies
Expires: September 2016   March 02, 2016

Centralized Replication for BUM traffic in active-active edge connection
draft-ietf-trill-centralized-replication-04.txt

Abstract

   In TRILL active-active access scenario, RPF check failure issue may
   occur when pseudo-nickname mechanism in [RFC7781] is used. This
   draft describes a solution to resolve the RPF check failure issue
   through centralized replication. All ingress RBridges send BUM
   (Broadcast, Unknown unicast and Mutlicast) traffic to a centralized
   node with unicast TRILL encapsulation. When the centralized node
   receives the BUM traffic, it decapsulates the packets and forwards
   them to all destination RBridges using a distribution tree
   established as per TRILL base protocol[RFC6325]. To avoid RPF check
   failure on a RBridge sitting between the ingress RBridge and the
   centralized replication node, some change of RPF calculation
   algorithm is required. RPF calculation on each RBridge should use
   the centralized node as ingress RB instead of the real ingress
   RBridge which is denoted as RBv in [RFC7781] to perform the
   calculation.

Status of this Memo

Internet-Drafts are working documents of the Internet Engineering
Task Force (IETF), its areas, and its working groups. Note that
other groups may also distribute working documents as Internet-
Drafts.

Internet-Drafts are draft documents valid for a maximum of six
months   and may be updated, replaced, or obsoleted by other
documents at any time. It is inappropriate to use Internet-Drafts
as reference material or to cite them other than as "work in
progress."

The list of current Internet-Drafts can be accessed at
http://www.ietf.org/1id-abstracts.html

The list of Internet-Draft Shadow Directories can be accessed at
http://www.ietf.org/shadow.html.

Copyright Notice

Table of Contents

1. Introduction

The IETF TRILL (Transparent Interconnection of Lots of Links)
[RFC6325] protocol provides loop free and per hop based multipath
data forwarding with minimum configuration. TRILL uses IS-IS
[RFC6165] [RFC6326bis] as its control plane routing protocol and
defines a TRILL specific header for user data.

Classic Ethernet (CE) devices typically are multi-homed to multiple
edge RBridges which form an edge group. All of the uplinks of CE are
bundled as a Multi-Chassis Link Aggregation (MC-LAG). An active-
active flow-based load sharing mechanism is normally implemented to
achieve better load balancing and high reliability. A CE device can
be a layer 3 end system by itself or a bridge switch through which
layer 3 end systems access to TRILL campus.

In active-active access scenario, pseudo-nickname solution in
[RFC7781] can be used to avoid MAC flip-flop on remote RBs. The
basic idea is to use a virtual RBridge RBv with a single pseudo-
nickname to represent an edge group that MC-LAG connects to. Any
member RBridge of that edge group should use this pseudo-nickname
rather than its own nickname as ingress nickname when it injects
TRILL data frames to TRILL campus. The use of the nickname solves
the address flip flop issue by making the MAC address learnt by the
remote RBridge bound to pseudo-nickname. However, it introduces
another issue of incorrect packet drop which will be described as
follows. When a pseudo-nickname is used by an edge RBridge as the
ingress nickname to forward BUM traffic, any RBridges (RBn) sitting
between the ingress RB and the distribution tree root will treat the
traffic as it is ingressed from the virtual RBridge RBv. If the same
distribution tree is used by different edge RBridges of the same RBv,
the traffic may arrive at RBn from different ports. Then the RPF
check fails, and some of the traffic receiving from unexpected ports
will be dropped by RBn.

This document proposes a centralized replication solution for broadcast, unknown unicast and multicast(BUM) traffic forwarding to resolve the issue of incorrect packet drop caused by RPF check failure in virtual RBridge case. The basic idea is that all ingress RBridges send BUM traffic to a centralized node which is recommended to be a distribution tree root using unicast TRILL encapsulation. When the centralized node receives the packets, it decapsulates and forwards them to all destination RBridges using a distribution tree established as per TRILL base protocol.

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].The acronyms and terminology in [RFC6325] is used herein with the following additions:

BUM - Broadcast, Unknown unicast and Multicast

CE - As in [CMT], Classic Ethernet device (end station or bridge).

   The device can be either physical or virtual equipment.

3. Centralized Replication Solution Overview

When an edge RB receives BUM traffic from a CE device, it uses unicast TRILL encapsulation instead of multicast encapsulation to send the packets to a centralized node. The centralized node is recommended to be a distribution tree root.

The TRILL header of the unicast TRILL encapsulation contains an "ingress RBridge nickname" field and an "egress RBridge nickname" field. If ingress RB receives the BUM packet from a port which is in an MC-LAG, it should set the ingress RBridge nickname to be the pseudo-nickname rather than its own nickname to avoid MAC flip-flop on remote RBs as per [RFC7781]. The egress RBridge nickname is set to the special nickname of the centralized node which is used to differentiate the centralized replication purpose unicast TRILL encapsulation from a normal unicast TRILL encapsulation. The special nickname is called R-nickname.

When the centralized node receives a unicast TRILL encapsulated packet with R-nickname as egress nickname, it decapsulates the packet. Then the centralized node replicates and forwards the BUM

packet to all destination RBs using one of the distribution trees
established as per TRILL base protocol. It is recommended to use the
distribution tree whose tree root is the centralized node itself.
When the centralized node forwards the BUM traffic, ingress nickname
remains same as that in frame it received to ensure that the MAC
address learnt by all egress RBridges bound to pseudo-nickname.

When the replicated packet is forwarded on each RBridge along the
distribution tree starting from the centralized node, RPF check will
be performed as per [RFC6325]. For any RBridge sitting between the
ingress RBridge and the centralized replication node, the incoming
port of such BUM packet should be the centralized node facing port
as the multicast traffic always comes from the centralized node in
this solution. However the RPF port as the result of distribution
tree calculation as per [RFC6325] will be the real ingress RB facing
port as it uses virtual RBridge as the ingress RB, so RPF check will
fail. To solve this problem, some change of RPF calculation
algorithm is required. RPF calculation on each RBridge should use
the centralized node as ingress RB instead of the real ingress
virtual RBridge to perform the calculation. As a result, RPF check
will point to the centralized node facing port on the RBridge for
multi-destination traffic. It prevents the incorrect frame drop by
RPF check.

To differentiate the centralized purpose unicast TRILL encapsulation
from normal unicast TRILL encapsulation on a centralized node, R-
nickname is introduced for centralized replication. Only when the
centralized node receives unicast TRILL encapsulation traffic with
egress nickname equivalent to R-nickname, it decapsulates the packet
and then forwards the packet to all destination RBs through a
distribution tree by re-encapsulation as per aforementioned. The
centralized nodes should announce its R-nickname to all TRILL campus
through TRILL LSP extension.

4. Frame duplication from remote RB

Frame duplication may occur when a remote host sends multi-
destination frame to a local CE which has an active-active
connection to the TRILL campus. To avoid local CE receiving multiple
copies from a remote RBridge, the designated forwarder (DF)
mechanism should be supported for egress direction multicast traffic.

DF election mechanism allows only one port of one RB in an MC-LAG to
forward multicast traffic from TRILL campus to local access side for
each VLAN. The basic idea of DF is to elect one RBridge per VLAN
from an edge group to be responsible for egressing the multicast

traffic. [RFC7781] describes the detailed DF election mechanism
among member RBridges involving in a MC-LAG.

If DF election mechanism is used for frame duplication prevention,
access ports on an RB are categorized as three types: non mc-lag,
mc-lag DF port and mc-lag non-DF port. The last two types can be
called mc-lag port. For each of the mc-lag port, there is a pseudo-
nickname associated. If consistent nickname allocation per edge
group RBridges is used, it is possible that same pseudo-nickname
associated to more than one port on a single RB. A typical scenario
is that CE1 is connected to RB1 & RB2 by mc-lag1 while CE2 is
connected to RB1 & RB2 by mc-lag 2. In order to conserve the number
of pseudo-nickname used, member ports for both mc-lag1 and mc-lag2
on RB1 & RB2 are all associated to pseudo-nickname pn1.

5. Local forwarding behavior on ingress RBridge

When an ingress RBridge(RB1) receives BUM traffic from a local
active-active accessing CE(CE1) device, the traffic will be injected
to TRILL campus through TRILL encapsulation, and it will be
replicated and forwarded to all destination RBs which include
ingress RB itself along a TRILL distribution tree, the traffic will
also return to the ingress RBridge. To avoid the traffic looping
back to original sender CE, ingress nickname of pseudo-nickname can
be used for traffic filtering.

If there are two CEs of CE1 and CE2 connecting to the ingress RB1
associated with same pseudo-nickname, CE1 needs to locally
replicates and forwards to CE2, because another copy of the BUM
traffic between CE1 and CE2 through TRILL campus will be blocked by
the traffic filtering.

If CE1 and CE2 are not associated with same pseudo-nickname, the
copy of the BUM traffic between CE1 and CE2 through TRILL campus
won't be blocked by the traffic filtering. To avoid duplicated
traffic on receiver CE, the local replicated BUM traffic between
these two CEs on ingress RB1 should be blocked.

In summary, to ensure correct BUM traffic forwarding behavior for
each CE, local replication behavior on ingress RBridge should be
carefully designed as follows:

1. Local replication to the ports associated with the same pseudo-
nickname as that associated to the incoming port.

2. Do not replicate to mc-lag port associated with different pseudo-
nickname.

3. Do not replicate to non mc-lag ports.

The above local forwarding behavior on the ingress RB of RB1 can be called centralized local forwarding behavior A.

If ingress RB of RB1 itself is the centralized node, BUM traffic injected to TRILL campus won't loop back to RB1. In this case, the local forwarding behavior is called centralized local forwarding behavior B. The local replication behavior on RB1 is as follows:

1. Local replication to the ports associated with the same pseudo-nickname as that associated to the incoming port.

2. Local replication to the mc-lag DF port associated with different pseudo-nickname. Do not replicate to mc-lag non-DF port associated with different pseudo-nickname.

3. Local replication to non mc-lag ports.

[6]. Loop prevention among RBridges in a edge group

If a CE sends a broadcast, unknown unicast, or multicast (BUM) packet through DF port to a ingress RB, it will forward that packet to all or subset of the other RBs that only have non-DF ports for that MC-LAG. Because BUM traffic forwarding to non-DF port isn't allowed, in this case the frame won't loop back to the CE.

If a CE sends a BUM packet through non-DF port to a ingress RB, say RB1, then RB1 will forward that packet to other RBridges that have DF port for that MC-LAG. In this case the frame will loop back to the CE and traffic split-horizon filtering mechanism should be used to avoid looping back among RBridges in a edge group.

Split-horizon mechanism relies on ingress nickname to check if a packet's egress port belongs to a same MC-LAG with the packet's incoming port to TRILL campus.

When the ingress RBridge receives BUM traffic from an active-active accessing CE device, the traffic will be injected to TRILL campus through TRILL encapsulation, and it will be replicated and forwarded to all destination RBs which include ingress RB itself through TRILL distribution tree. If same pseudo-nickname is used for two active-active access CEs as ingress nickname, egress RB can use the nickname to filter traffic forwarding to all local CE. In this case, the traffic between these two CEs goes through local RB and another copy of the traffic from TRILL campus is filtered. If different ingress nickname is used for two connecting CE devices, the access

ports connecting to these two CEs should be isolated with each other. The BUM traffic between these two CEs should go through TRILL campus, otherwise the destination CE connected to same RB with the sender CE will receive two copies of the traffic.

7. Centralized replication forwarding process

```
                         +-----------+
                         |   (RB5)   |
                         +-----------+
                               |
                         +-----------+
                         |   (RB4)   |
                         +-----------+


                      |      |     |
                   --------     |     --------
                   |            |           |
                +------+     +------+     +------+
                |(RB1) |     |(RB2) |     | (RB3)|
                +------+     +------+     +------+
                  *   |        *   |        * |  ^
                  *   |        *   |        * |   ^
                  *    ---------*-------------*--    ^
                  ***************************** |      ^
           MC-LAG1 *                    MC-LAG2 |       ^
              +------+                   +------+    +------+
              | CE1 |                    | CE2  |    | CE3  |
              +------+                   +------+    +------+
```
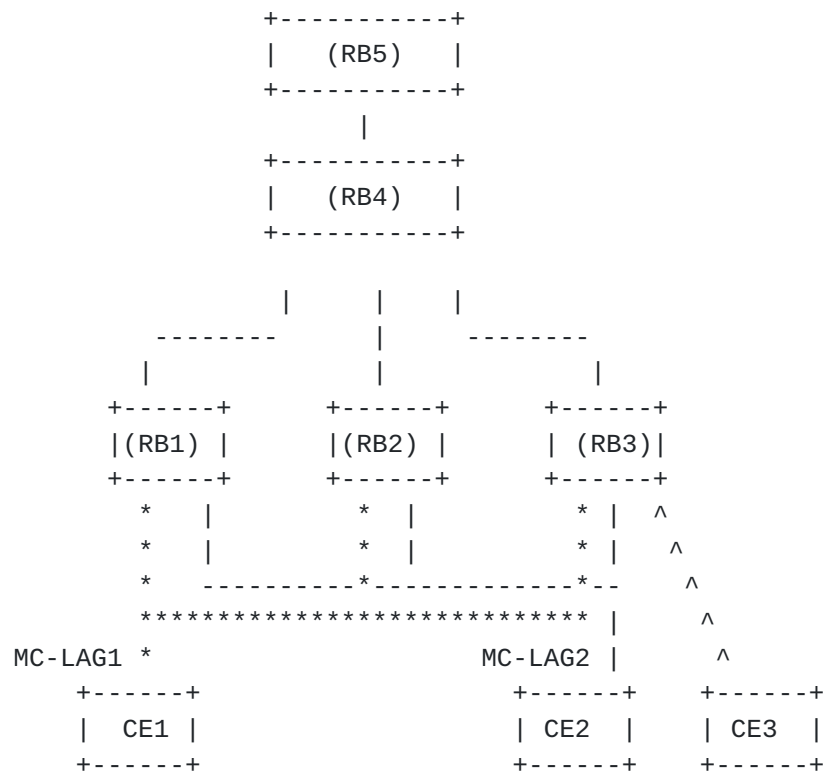
                 Figure 1 TRILL Active-active access

Assuming the centralized replication solution is used in the network of above figure 1, RB5 is the distribution tree root and centralized replication node, CE1 and CE2 are active-active accessed to RB1,RB2 and RB3 through MC-LAG1 and MC-LAG2 respectively, CE3 is single homed to RB3. The RBridge's own nickname of RB1 to RB5 are nick1 to nick5 respectively. RB1,RB2 and RB3 use same pseudo-nickname for MC-LAG1 and MC-LAG2, the pseudo-nickname is P-nick. The R-nickname on the centralized replication node of RB5 is S-nick.

The BUM traffic forwarding process from CE1 to CE2,CE3 is as follows:

1. CE1 sends BUM traffic to RB3.

2. RB3 replicates and sends the BUM traffic to CE2 locally. RB2 also
   sends the traffic to RB5 through unicast TRILL encapsulation.
   Ingress nickname is set as P-nick, egress nickname is set as S-
   nick.

3. RB5 decapsulates the unicast TRILL packet. Then it uses the
   distribution tree whose root is RB5 to forward the packet. The
   egress nickname in the trill header is the nick5. Ingress
   nickname is still P-nick.

4. RB4 receives multicast TRILL traffic from RB5. Traffic incoming
   port is the up port facing to distribution tree root, RPF check
   will be correct based on the changed RPF port calculation
   algorithm in this document. After RPF check is performed, it
   forwards the traffic to all other egress RBs(RB1,RB2 and RB3).

5. RB3 receives multicast TRILL traffic from RB4. It decapsulates
   the multicast TRILL packet. Because ingress nickname of P-nick is
   equivalent to the nickname of local MC-LAGs connecting CE1 and
   CE2, it doesn't forward the traffic to CE1 and CE2 to avoid
   duplicated frame. RB3 only forwards the packet to CE3.

6. RB1 and RB2 receive multicast TRILL traffic from RB4. The
   forwarding process is similar to the process on RB3, i.e, because
   ingress nickname of P-nick is equivalent to the nickname of local
   MC-LAGs connecting CE1 and CE2, they also don't forward the
   traffic to local CE1 and CE2.

8. BUM traffic load balancing among multiple centralized nodes

To support unicast TRILL encapsulation BUM traffic load balancing,
multiple centralized replication node can be deployed and the
traffic can be load balanced on these nodes in vlan-based.

Assuming there are k centralized nodes in TRILL campus, each
centralized node has different R-nickname, VLAN-based(or FGL-based)
loadbalancing algorithm used by ingress active-active access RBridge
is as follows:

   1. All R-nicknames are ordered and numbered from 0 to k-1 in
   ascending order.

   2. For VLAN ID m, choose the R-nickname whose number equals (m
   mod k) as egress nickname for BUM traffic unicast TRILL
   encapsulation.

For examples, there are 3 centralized nodes (CN) which has one R-nickname respectively, the CN nodes will be ordered based on the ordering algorithm for each R-nickname from CN0 to CN2. Assuming there are 5 VLANs from 1 to 5 spreading among edge RBridges, the traffic in VLAN 1 will go to CN1, VLAN 2 will go to CN2, and so on.

When an ingress RBridge participating active-active connection receives BUM traffic from local CE, the RB decides to send the traffic to which centralized node based on the VLAN-based loadbalancing algorithm, vlan-based loadbalancing for the BUM traffic can be achieved among multiple centralized nodes.

9. Co-existing with CMT solution

```
                    +------+     +------+
                    |(RB6) |     |(RB7) |
                    +------+     +------+
         ----------------|-----------|---------------------
            |         |         |         |            |
        +------+   +------+   +------+   +------+    +------+
        |(RB1) |   |(RB2) |   |(RB3) |   |(RB4) |    |(RB5) |
        +------+   +------+   +------+   +------+    +------+
            |         |           |         |           |
         ------------           ------------------------
              |                          |
           +------+                   +------+
           | CE1 |                    | CE2 |
           +------+                   +------+
```
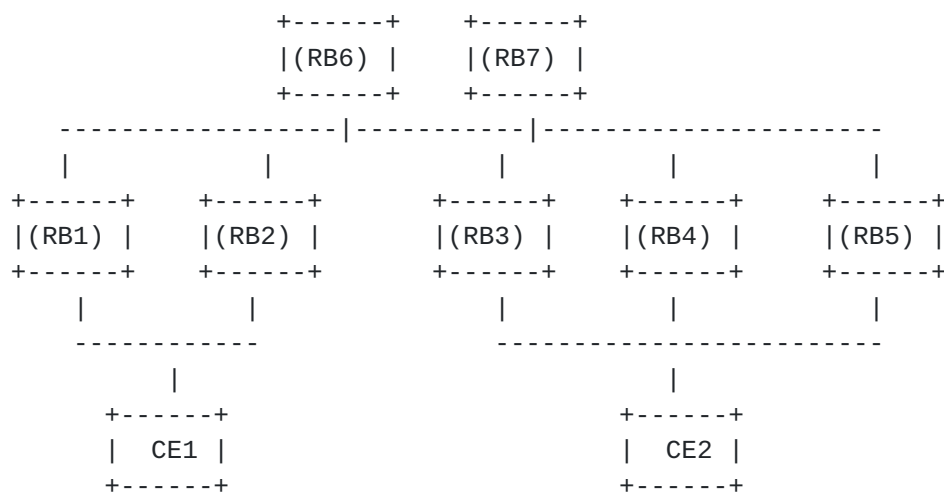Figure 2 CMT and centralized replication co-existing scenario

Both the centralized replication solution and CMT solution rely on pseudo-nickname concept to avoid MAC flip-flop on remote RBridges, these two solutions can co-exist in single TRILL network. Each solution can be selected by each edge group RBridges independently. As illustrated in figure 2, RB1 and RB2 use CMT for CE1's active-active access, RB3,RB4 and RB5 use the centralized replication for CE2's active-active access.

For the centralized replication solution, edge group RBridges should announce local pseudo-nickname using Nickname Flags APPsub-TLV with C-flag, the nickname with C-flag is called "C-nickname". A transit RBridge will perform the centralized replication specific RPF check algorithm if it receives TRILL encapsulation traffic with C-nickname as ingress nickname.

[10]. Network Upgrade Analysis

Centralized nodes need software and hardware upgrade to support
centralized replication process, which stitches TRILL unicast
traffic decapsulation process and the process of normal TRILL
multicast traffic forwarding along distribution tree.

Active-active connection edge RBs need software and hardware upgrade
to support unicast TRILL encapsulation for BUM traffic, the process
is similar to normal head-end replication process.

Transit nodes need software upgrade to support RPF port calculation
algorithm change.

[11]. TRILL protocol extension

Two Flags of "R" and "C" in Nickname Flags APPsub-TLV [RFC7180bis]
are introduced, the nickname with "R" flag is called R-nickname, the
nickname with "C" flag is called C-nickname. R-nickname is a
specialized nickname attached on a centralized node to differentiate
unicast TRILL encapsulation BUM traffic from normal unicast TRILL
traffic. C-nickname is set on each edge group RBridge, C-nickname is
a specialized pseudo-nickname for transit RBridges to perform
different RPF check algorithm.

When active-active edge RBridges use centralized replication to
forward BUM traffic, the R-nickname is used as the egress nickname
and the C-nickname is used as ingress nickname in TRILL header for
unicast TRILL encapsulation of BUM traffic.

[11.1]. "R" and "C" Flag in Nickname Flags APPsub-TLV

```
      +--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
      |   Nickname                                    |
      +--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
      |IN|SE|R | C|    RESV                           |
      +--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
                  NICKFLAG RECORD
```

   o R. If R flag is one, it indicates that the advertising TRILL
switch is a centralized replication node, and the nickname is used
as egress nickname for edge group RBridges to inject traffic to
TRILL campus when the edge group RBridges use centralized
replication solution for active-active access. If flag is zero, that
nickname will not be used for that purpose.

o C. If C flag is one, it indicates that the TRILL traffic
with this nickname as ingress nickname requires special RPF check
algorithm. If flag is zero, that nickname will not be used for that
purpose.

12. Security Considerations

This draft does not introduce any extra security risks. For general
TRILL Security Considerations, see [RFC6325].

13. IANA Considerations

This document requires no IANA Actions. RFC Editor: Please remove
this section before publication.

14. References

14.1. Normative References

[1] [RFC6165] Banerjee, A. and D. Ward, "Extensions to IS-IS for
       Layer-2 Systems", RFC 6165, April 2011.

[2] [RFC6325] Perlman, R., et.al. "RBridge: Base Protocol
       Specification", RFC 6325, July 2011.

[3] [RFC6326bis] Eastlake, D., Banerjee, A., Dutt, D., Perlman, R.,
       and A. Ghanwani, "TRILL Use of IS-IS", draft-eastlake-isis-
       rfc6326bis, work in progress.

[4] [RFC7180bis] Eastlake, D., Zhang, M., Perlman, R., Banerjee, A.
       Ghanwani and Gupta.S, "TRILL: Clarifications, Corrections, and
       Updates", draft-ietf-trill-rfc7180bis-00, work in progress.

14.2. Informative References

[1] [RFC7781] Zhai,H., et.al., "RBridge: Pseduonode nickname",
       draft-hu-trill-pseudonode-nickname-07, Work in progress,
       November 2011.

[2] [TRILAA] Li,Y., et.al., " Problem Statement and Goals for
       Active-Active TRILL Edge", draft-ietf-trill-active-active-
       connection-prob-00, Work in progress, July 2013.

[3] [CMT] Senevirathne, T., Pathangi, J., and J. Hudson,
       "Coordinated Multicast Trees (CMT)for TRILL", draft-ietf-
       trill-cmt-00.txt Work in Progress, April 2012.

[15](). Acknowledgments

Authors' Addresses

Weiguo Hao
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China
Email: haoweiguo@huawei.com

Yizhou Li
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China
Email: liyizhou@huawei.com

Muhammad Durrani
Cisco
Email: mdurrani@cisco.com

Sujay Gupta
IP Infusion
RMZ Centennial
Mahadevapura Post
Bangalore - 560048
India
Email: sujay.gupta@ipinfusion.com

Andrew Qu
MediaTec
Email: laodulaodu@gmail.com

Tao Han
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China
Email: billow.han@huawei.com