TSVWG
Internet-Draft
Intended status: Informational
Expires: April 18, 2013

L. Zhu
Huawei Technologies
H. Zhang
X. Gong
BUPT

October 15, 2012

# Tunnel Congestion Exposure
## draft-zhang-tsvwg-tunnel-congestion-exposure-00

Abstract

At present, tunneling technology has been widely applied in VPN, mobile communication network, IPv6 over IPv4, Mobile IP, multi-point delivery, and other fields. In the E2E link, there may already have been an effective congestion control mechanism, but we SHOULD also do traffic management in the tunnel to improve the performance of the entire network. Because of the particularity of the scenario of the tunnel, the existing E2E traffic management mechanism cannot be directly be deployed (e.g. VPN, IPv6 over IPv4 etc). In these cases, this document focuses on how to expose the congestion while the feedback mechanism is left for later study. This document describes the problem of identifying congestion in a tunnel segment of an end-to-end flow. A basic tunnel congestion exposure model is then described, followed by three example scenarios which use the basic model to derive tunnel congestion. Finally, a general solution that can be applied to IP-in-IP tunnels is described.

Table of Contents

## 1. Introduction

This document mainly describes four issues.

Firstly, this document describes the problem concerning congestion in the tunnel. At present, tunneling technology has been widely applied in, VPN, mobile communication network, IPv6 over IPv4, Mobile IP, multi-point delivery, and other fields. In this case, relying on end-to-end congestion management alone to deal with the congestion problem in the tunnel brings many drawbacks and seriously affects the performance of the entire network.

Secondly, this document proposes a basic tunnel congestion exposure model. In the model, the ingress and the egress of the tunnel are two dominant nodes which have the ability to handle admission, flow control and policy control.

In the basic model, in order to achieve a real-time understanding of the congestion status of the tunnel, the amount of tunnel congestions is fed back to the ingress of the tunnel using the tunnel encapsulation protocol signals. It's helpful for the ingress to achieve a real-time congestion status of the entire tunnel, and it also provides the possibility of using policy or flow control mechanisms to further reduce congestion in the local tunnel portion.

Thirdly, this document introduces three example scenarios where the proposed basic model can be applied.

Finally, this document presents a general solution. The general solution includes a statement of requirements and general procedures.

### 1.1. Document Overview

The main section in this document is the basic model described in chapter 4. A tunnel congestion exposure model is presented in the chapter. The model is relatively simple, but it can be used to sufficiently expose the tunnel congestion. Chapter 3 gives the general process of tunnel transmission and presents two major problems related to the tunnel congestion. Three scenarios are given in Chapter 5 that use the basic model. Chapter 6 introduces a general solution. Chapter 7 states a further study plan.

## 2. Conventions and Terminology

## 2.1. Conventions

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].

Table 1 recaps the names of the ECN codepoints [RFC3168].

| Binary codepoint | Codepoint name | Meaning |
|---|---|---|
| 00 | Not-ECT | Not ECN-capable transport |
| 01 | ECT(1) | ECN-capable transport |
| 10 | ECT(0) | ECN-capable transport |
| 11 | CE | Congestion experienced |

Table 1: Recap of Codepoints of the ECN Field [RFC3168]

## 2.2. Terminology

Further terminology used within this document:

- Tunnel: A tunnel is just a special type of connection across a network.
- Encapsulation: The process of adding control information as it passes through the layered model.
- Encapsulator: The tunnel endpoint function that adds an outer IP header to tunnel a packet, the encapsulator is considered as the "ingress" of the tunnel.
- Decapsulator: The tunnel endpoint function that removes an outer IP header from a tunnelled packet, the decapsulator is considered as the "egress" of the tunnel.
- Outer header: The header added to encapsulate a tunneled packet.
- Inner header: The header encapsulated by the outer header.
- E2E: End to End
- VPN: Virtual Private Network is a technology for using the Internet or another intermediate network to connect computers to isolated remote computer networks that would otherwise be inaccessible.

## 3. Problem Statement

Tunneling technology is one way to transmit data between different networks by using the Internet infrastructure. The tunnel can transmit the frames or packets of different protocols as its payload. The frames or packets are re-encapsulated using new headers of the tunneling protocol. The new header provides the routing information so that the encapsulated payload data can be transmitted through the public Internet. As soon as being transmitted to the endpoint of network (egress), the data will be de-encapsulated and forwarded to its final destination.

In the IP network, there are three typical encapsulation protocols: IP Encapsulation within IP [RFC2003]; Minimal Encapsulation within IP [RFC2004]; Generic Routing Encapsulation (GRE) [RFC1701].

As an example, the format of GRE encapsulation is shown in Figure 1.

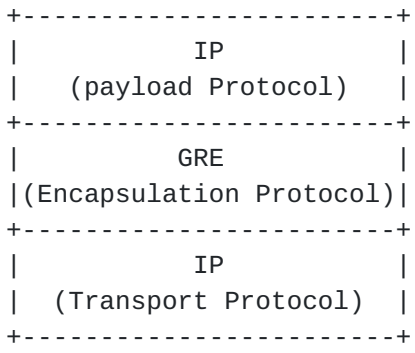


Figure 1: GRE encapsulation format

At present, tunneling technology is widely used in computer networks. It MAY be used to transport IPv4 in IPv6 networks and vice versa during the two protocols coexistence period. In Mobile IP, by introducing tunnels, a mobile node can communicate with its correspondent node using a fixed address (home address) at any time and place. In addition, tunneling technology has also played an important role in virtual private network (VPN), multi-point delivery, and mobile communication networks.

We perceive two problems related to congestion and tunneling technology in IP and mobile communication networks.

Consider the following situation: data packets need to be transmitted end-to-end over an ECN-enabled transport and part of this includes a tunnel segment.

The ingress, the egress and the intermediate nodes of the tunnel are ECN-enabled. At the egress of the tunnel, the data packets are de-capsulated, peeled off the transport protocol in outer header, and then the inner packets are forwarded to the destination endpoint.

In this case, it is useful to specify the behaviors of the encapsulation and de-capsulation; otherwise the following problems may occur:

- The packet congestion inside the tunnel cannot be indicated;
- E2E path congestion information may be lost.

For these two problems, RFC3168, RFC4301 and RFC6040 have presented some descriptions and specifications to all IP in IP tunnels.

The core idea is:

- Set the ECN bits as ECT(0)/ECT(1) at the ingress in order to inform the interior router to perform mark operations and indicate congestion events inside the tunnel;
- Perform "copy-to-copy" operations at the ingress and egress of the tunnel, so that the congestion information of the arriving packet or inside the tunnel can be forwarded to the final endpoint instead of being lost.

In the mobile communication network, end-user IP traffic maybe forwarded over multiple tunnel segments before it reaches the router that handles subscriber policy, charging etc. Congestion over these tunnel segments contribute to the overall congestion experienced E2E.

In the tunnel protocol deployment scenarios, we may use other methods, for example, out of-band signaling, to report the congestion information to NE (Network Element) which has the mechanisms to manage congestion based on per user policies or other flow control mechanisms.

However, tunnel congestion information should not be overlooked. In the E2E link, there may already have been an effective congestion control mechanism, but we SHOULD also do traffic management in the tunnel to improve the performance of the entire network. Because of the particularity of the scenario of the tunnel, the existing E2E traffic management mechanism cannot be directly be deployed (e.g. VPN, IPv6 over IPv4 etc). In these cases, this document focuses on how to expose the congestion while the feedback mechanism is left for later study.

The following assumptions are made in this document:

- Firstly, both payload and transport protocols are IP in this document, i.e., we consider only IP-in-IP tunnels;
- Secondly, the intermediate nodes within the tunnels, for example, the routers are ECN-enabled;
- Finally, no changes are made to the ECN mechanism.

## 4. Basic Model

According to the general tunnel transmission process, this section introduces the abstract of the tunnel transmission and outlines a tunnel congestion exposure model as shown in Figure 2:

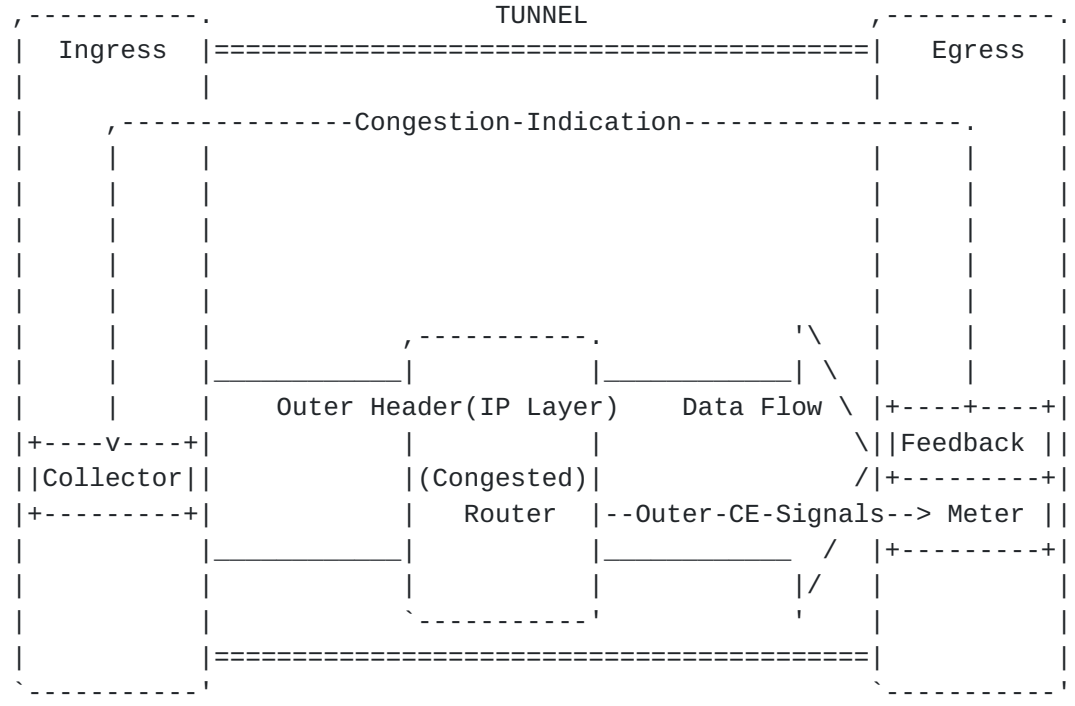


Figure 2: Tunnel Congestion Exposure Basic Model

This basic model MAY contain the following components: Ingress, Egress, Collector, Feedback and Meter.

By and large, the ingress and egress of the tunnel are gateway devices. In terms of the egress, it can calculate the amount of congestion and feed back the congestion information to the ingress. The collector is able to receive the congestion information which is

fed back from the egress. It MAY also have admission control and flow control functions.

General practice can be as follows:

At the egress, a module named meter can be added. The module records the outer header CE codepoint packets reaches the egress independently, and MAY need to estimate the congestion level inside the tunnel. In addition, a congestion information feedback module, called feedback, is also needed to be added. The feedback module is used to control the congestion information feedback. The feedback of the congestion information can be done via the extension of the encapsulation protocol of the tunnel.

A collector module for receiving the feedback congestion information from the egress SHOULD be added. The collector can be distributed in the ingress or other network elements that have the capability of handling the congestion (such as the PDN-GW in the mobile communication network described in section 5.2). Furthermore, some modules related with the congestion policy process can be added. However, no descriptions concerning this aspect are given in this document for the time being.

In this model, the tunnel is an IP-in-IP tunnel. Both the entry and egress of the tunnel are ECN-enabled and the intermediate routers in the tunnel path are also ECN-enabled.

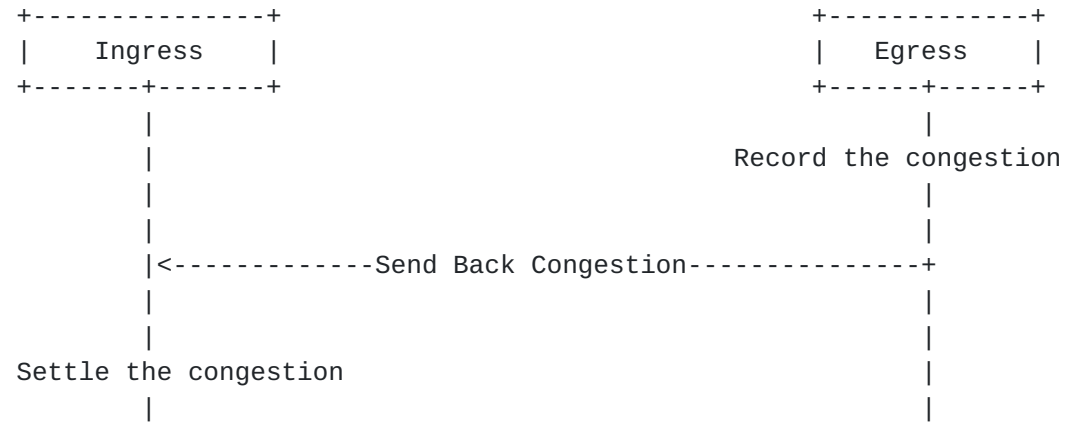


Figure 3: the general process of the congestion feedback

Figure 3 demonstrates the general process of the congestion information feedback.

The general method to feed back the congestion information is to do

some extensions to the encapsulation protocol messages. The details of the feedback process may differ in terms of the encapsulation protocols of the tunnel, but the general process is as shown in Figure 3.

## 5. Example Scenarios

This chapter presents three scenarios based on the basic model described in the previous chapter. This chapter focuses on two aspects. On the one hand, it describes the process of tunnel congestion exposure for each scenario, and on the other hand it stresses the significance of congestion exposure.

### 5.1. VPN Scenario

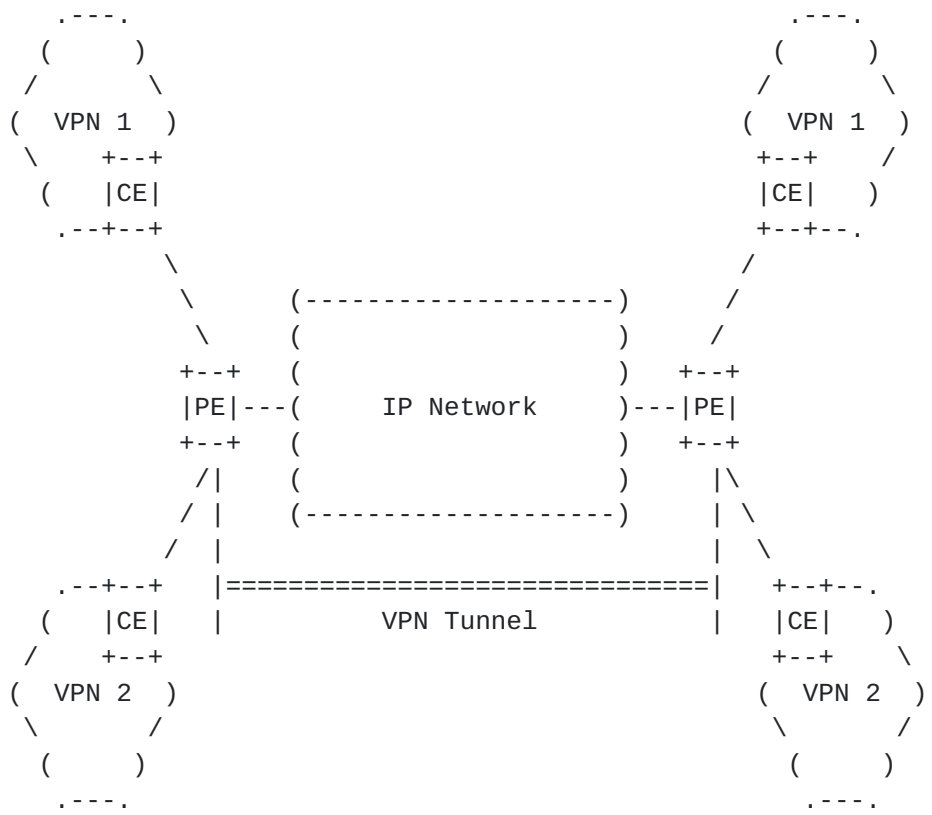


Figure 4: VPN Scenario

Figure 4 is a simplified VPN multi-instance model. A VPN tunnel is established between two PEs. Before creating the VPN tunnel, the routing in the network between two PEs has been configured (e.g., in large networks, the BGP routing protocol is generally used). CEs are

connected to the networks where the users locate. Both ends of the CE devices are located respectively in the VPN instance 1 (VPN 1) and VPN instance 2 (VPN 2).

CE and PE are defined as follows:

- CE (Customer Edge): User network edge devices, which have interfaces directly to the SP (Service Provider) networks. A CE can be a switch and can also be a host. Usually, the CE cannot sense the presence of the VPN and it does not need to support the VPN features.

- PE (Provider Edge): IP network edge devices, which are connected to CEs directly. In the VPN network, all processing of VPN instances is on the PEs.

In this scenario, the VPN Tunnel is the object drawing our attention. The two PEs are the powerful nodes that acting as the entry and egress of the tunnel, receiving and sending back the congestion information. The PEs can also do more operations of the traffic management. The process is almost the same as that of the basic model.

### 5.2. Mobile Scenario

Figure 5 is the scenario in the EPS (Evolved Packet System), where two PMIP tunnels exist, i.e., PMIP tunnel between eNodeB and the S-GW, and between the S-GW and the PDN-GW. In this scenario, we can just expose the congestion information of the backhaul tunnel segment (Note: It is an assumption that core network has extremely low possibility of congestion). With the extensive use of mobile communication network and variable traffic rate per user, backhaul network congestion problems get more unpredictable. If the congestion information in the backhaul network can be exposed, the PDN-GW MAY reuse the exposed congestion information to do flow control, which is helpful to improve the performance of the entire network and enhance the user experience.

There are significant differences between this scenario and the VPN scenario.

- First of all, the congestion information is reported to PDN-GW rather than the ingress.

- Secondly, no congestion feedback exists in the backhaul network in both the uplink and downlink.

- In addition, the object which is used to reporting the tunnel congestion information (PDN-GW) and the module used to record the congestion information are not in the same tunnel section.

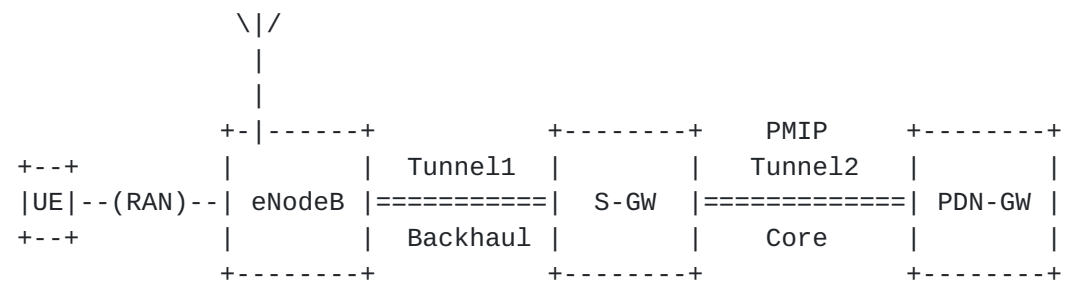


Figure 5: Mobile Scenario

In this scenario, the processing in the uplink and downlink SHOULD be distinguished. The general process is as follows:

- The egress records the congestion information according to the description in the basic model.

- In the uplink direction, the egress is S-GW. Therefore, the S-GW will report congestion information to the PDN-GW using the PMIP messages.

- In the down-link direction, the egress is the eNodeB. The eNodeB feeds back the congestion information to the S-GW first using the PMIP messages. Then the S-GW transfers the congestion information to the PDN-GW.

### 5.3. IPv6 over IPv4 Scenario

The tunnel used to connect IPv6 isolated islands in an IPv4 network is called IPv6 over IPv4 tunnel.

In the early stage of the transition from the IPv4 Internet to the IPv6 Internet, the IPv4 networks have been well deployed while the IPv6 networks are isolated network islands spread around the world. Using a dedicated line to connect these isolated IPv6 islands is not economical obviously. The usual practice is to use tunneling technology. By using the tunneling technology to create a tunnel in the IPv4 network, the IPv6 islands can be interconnected through the IPv4 networks. This is similar to the deployment of VPNs in the IP networks through the tunneling technology.

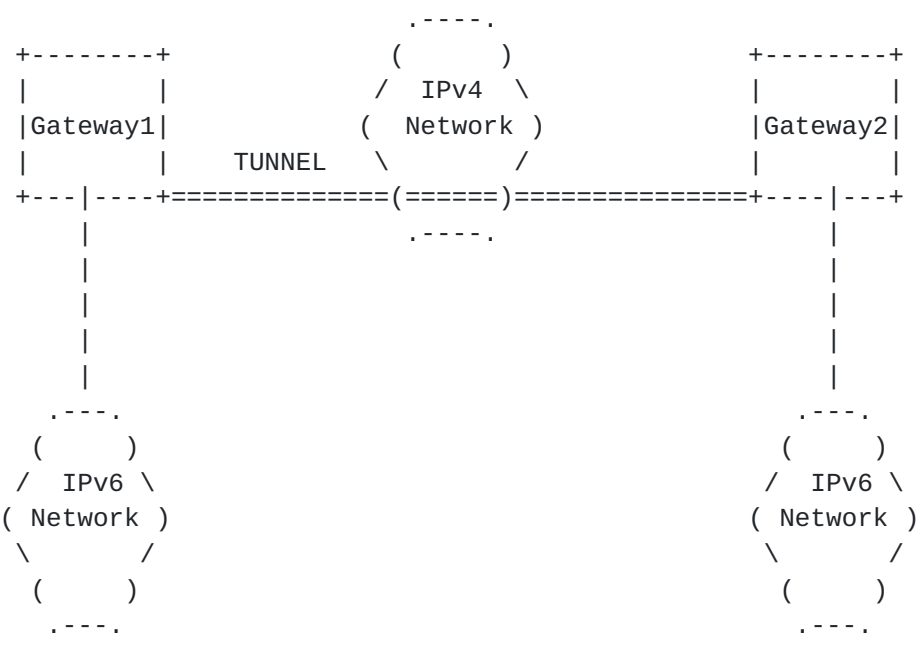


Figure 6: IPv6 over IPv4 Scenario

In this scenario, the two gateways, namely Gateway1 and Gateway2, perform admission control. They are the endpoints of the connection to the two IPv6 isolated islands and serve as the ingress and egress of the tunnel. It can be seen from above, the specific congestion exposure mechanisms are consistent with the basic model.

The congestion information is exposed at the ingress, we can use the congestion information to do a lot of significant things.

## 6. The General Solution

### 6.1. Statement of Requirements

- All tunnels are IP-in-IP type.
- Tunnel path supports the ECN mechanism--that is, the ingress, the egress and intermediate nodes are ECN enabled.
- The tunnel is configured to support feedback congestion information.
- A powerful device exists in the tunnel which is used for processing the congestion information.

### 6.2. The General Procedure

- In case of congestion, network element (ingress and egress of the tunnel, intermediate routers etc.) performs marking operations on packets according to AQM algorithm.

- The egress of the tunnel calculates congestion information by recording the number of congestion and total package periodically.

- The egress of the tunnel feeds back congestion information to the functional traffic management entity (such as: the ingress of the tunnel). In the tunnel, if the transport protocol is UDP, the congestion information is fed back by extending signaling of the encapsulation protocol. In this step, the encapsulation signaling SHOULD be extended. For example, in the IPSec tunnel, the IPSec signaling messages should be extended which are used for sending back the congestion information to the collector module.

- Congestion information receiving object (functional entity executing flow management) disposes congestion information when the congestion threshold level is reached. Here, we can define the different congestion levels according to the actual situations or requirements. For the simplest condition, we can divide the congestion conditions into three types: the high congestion, moderate congestion, and low-grade congestion. This process involves things like collecting congestion information, making policy control based on the congestion information and etc.

## 7. Next Steps

At present, this document focuses on how to expose the tunnel congestion to the ingress of the tunnel which has flow control and policy control functions etc. In this document, a basic model for congestion exposure is proposed, a general solution is introduced, and several scenarios for applying the basic model are described. However, no excessive details are given.

In the following versions, more details and processes of the tunnel congestion exposure will be introduced. Which type of congestion information and how to use the information will also be discussed. In the near future, more than one document will be used to describe these practices.

## 8. Security Considerations

TBD

## 9. IANA Considerations

This document includes no request to IANA.

## 10. Acknowledgments

The authors would like to thank Wendong Wang, Li Yuhong and Xirong Que for their technical guidances towards to this draft.

The authors would like to thank their colleagues for their sincerely help and comments when drafting this document.

## 11. References

### 11.1. Normative Reference

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC3168] Ramakrishnan, K., Floyd, S., and D. Black, "The Addition of Explicit Congestion Notification (ECN) to IP", RFC 3168, September 2001.

[RFC4301] Kent, S. and K. Seo, "Security Architecture for the Internet Protocol", RFC 4301, December 2005.

[RFC6040] Briscoe, B., "Tunnelling of Explicit Congestion Notification", RFC 6040, November 2010.

### 11.2. Informative References

[3GPP.23.203] 3GPP, "Policy and charging control architecture", 3GPP TS 23.203 10.7.0 , June 2012.

[3GPP.23.402] 3GPP, "Architecture enhancements for non-3GPP accesses", 3GPP TS 23.402 V11.3.0 , June 2012.

[3GPP.29.274] 3GPP, "Technical Specification Group Core Network and Terminals; 3GPP Evolved Packet System (EPS); Evolved General Packet Radio Service (GPRS) Tunneling Protocol for Control plane (GTPv2-C)", 3GPP TS 29.274 V11.2.0 , March 2012.

[I-D.ietf-conex-abstract-mech]
Mathis, M. and B. Briscoe, "Congestion Exposure (ConEx) Concepts and Abstract Mechanism", draft-ietf-conex-abstract-mech-05 (work in progress), July 2012.

[RFC1701] Hanks, S., Li, T., Farinacci, D., and P. Traina, "Generic Routing Encapsulation (GRE)", RFC 1701, October 1994.

[RFC2003] Perkins, C., "IP Encapsulation within IP", RFC 2003, October 1996.

## Authors' Addresses

Lei Zhu
Huawei Technologies
Huawei Building, Q20 No.156 Beiqing Rd.Z-park
Haidian District, Beijing 100095
P. R. China

Email: lei.zhu@huawei.com

Huabing Zhang
Beijing University of Posts and Telecommunications
Xitucheng road 10
Haidian District, Beijing 100876
P. R. China

Email: zhanghb29@gmail.com

Xiangyang Gong
Beijing University of Posts and Telecommunications
Xitucheng road 10
Haidian District, Beijing 100876
P. R. China

Email: xygong@bupt.edu.cn