Internet Engineering Task Force X. Wei
INTERNET-DRAFT Huawei Technologies
Intended Status: Informational L.Zhu
Expires: March 19, 2016 Huawei Technologies
L.Deng
China Mobile
September 16, 2015

Tunnel Congestion Feedback
draft-ietf-tsvwg-tunnel-congestion-feedback-00

Abstract

This document describes a mechanism to calculate congestion of a tunnel segment based on RFC 6040 recommendations, and a feedback protocol by which to send the measured congestion of the tunnel from egress to ingress . A basic model for measuring tunnel congestion and feedback is described, and a protocol for carrying the feedback data is outlined.

Table of Contents

1. Introduction 3
2. Conventions 3
3. Congestion Information Feedback Models 4
3.1 Direct Model 4
3.2 Centralized Model 4
4. Congestion Level Measurement 5
5. Congestion Information Delivery 7
5.1 IPFIX Extentions 7
5.1.1 ce-cePacketTotalCount 7
5.1.2 ect0-nectPacketTotalCount 8
5.1.3 ect1-nectPacketTotalCount 8
5.1.4 ce-nectPacketTotalCount 8
5.1.5 ce-ect0PacketTotalCount 9
5.1.6 ce-ect1PacketTotalCount 9
5.1.7 ect0-ect0PacketTotalCount 9
5.1.8 ect1-ect1PacketTotalCount 10
6. Congestion Management 10
7. Security 11
8. IANA Considerations 11
9. References 11
9.1 Normative References 11
9.2 Informative References 12
10. Acknowledgements 12
Authors' Addresses 12

1. Introduction

In IP network, persistent congestion (or named congestion collapse) lowers transport throughput, leading to waste of network resource. Appropriate congestion control mechanisms are therefore critical to prevent the network from falling into the persistent congestion state. Currently, transport protocols such as TCP, SCTP, DCCP, have their built-in congestion control mechanisms, and even for certain single transport protocol like TCP there can be a couple of different congestion control mechanisms to choose from. All these congestion control mechanisms are implemented on host side, and there are reasons that only host side congestion control is not sufficient for the whole network to keep away from persistent congestion. For example, (1) some protocol's congestion control scheme may have internal design flaws; (2) improper software implementation of protocol; (3) some transport protocols do not even provide congestion control at all.

In order to have a better control on network congestion status, it's necessary for the network side to do certain kind of traffic control. For example, ConEx [ConEx] provides a method for network operator to learn about traffic's congestion contribution information, and then congestion management action can be taken based on this information.

Tunnels are widely deployed in various networks including public Internet, datacenter network, and enterprise network etc. A tunnel consists of ingress, an egress and a set of interior routers. For the tunnel scenario, a tunnel-based mechanism which is different from ConEx is introduced for network traffic control to keep the network from persistent congestion. Here, tunnel ingress will implement congestion management function to control the traffic entering the tunnel.

In order to perform congestion management at ingress, the ingress must first obtain the inner tunnel congestion level information. Yet the ingress cannot use the locally visible traffic rates, because it would require additional knowledge of downstream capacity and topology, as well as cross traffic that does not pass through this ingress.

This document provides a mechanism of feeding back inner tunnel congestion level to the ingress. Using this mechanism the egress can feed the tunnel congestion level information it collects back to the ingress. After receiving this information the ingress will be able to perform congestion management according to network management policy.

2. 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 RFC 2119 [RFC2119]

3. Congestion Information Feedback Models

According to specific network deployment, there are two kinds of feedback model: direct model and centralized model.

3.1 Direct Model

```
                        Feedback
          |-----------------------------------------|
          |                                         |
          |                                         |
          |                                         V
      +----------+         tunnel            +-----------+
      |Egress    |========================== |Inress     |
      |(Exporter)|                           |(Collector)|
      +----------+                           +-----------+
```

(a) Direct Feedback Model.

Direct model means egress feeds information directly to ingress. In this model, egress collects network congestion level information and feedback the information to the ingress for congestion management. The ingress here will act as both the decision point that decides how to do congestion management and the action point that implements congestion management decision.

3.2 Centralized Model

```
    Feedback    +-----------+
      --------->|Controller |#####################
      |         |(Collector)|                   #
      |         +-----------+                   #
      |                                         #
   +----------+         tunnel           +-----V-+
   |Egress    | ===========================|Ingress|
   |(Exporter)|                          +-------+
   +----------+
```

(b) Centralized Feedback Model

In the centralized model, the ingress only takes the role of action point, and it implements traffic control decision from another entity named "controller". Here, after egress has collected network

congestion level information, it feeds back the information to a controller instead of the ingress. Then the controller makes congestion management decision and sends the decision to the ingress to implement.

4. Congestion Level Measurement

This section describes how to measure congestion level in a tunnel.

There may be different approaches to packet loss detection for different tunneling protocol scenarios. For instance, if there is a sequence field in the tunneling protocol header, it will be easy for egress to detect packet loss through the gaps in sequence number space. Another approach is to compare the number of packets entering ingress and the number of packets arriving at egress over the same span of packets. This document will focus on the latter one which is a more general approach.

If the routers support Explicit Congestion Notification (ECN), after router's queue length is over a predefined threshold, the routers will marks the ECN-capable packets as Congestion Experienced (CE) or drop not-ECT packets with the probability proportional to queue length; if the queue overflows all packets will be dropped. If the routers do not support ECN, after router's queue length is over a predefined threshold, the routers will drop both the ECN-capable packets and the not-ECT packets with the probability proportional to the queue length. It's assumed all routers in the tunnel support ECN.

Faked ECN-capable transport (ECT) is used at ingress to defer packet loss to egress. The basic idea of faked ECT is that, when encapsulating packets, ingress first marks tunnel outer header according to RFC6040, and then remarks outer header of Not-ECT packet as ECT, there will be three kinds of combination of outer header ECN field and inner header ECN field: CE|CE, ECT|N-ECT, ECT|ECT (in the form of outer ECN| inner ECN).

In case all interior routers support ECN, the network congestion level could be indicated through the ratio of CE-marked packet and the ratio of packet drop, the relationship between these two kinds of indicator is complementary. If the congestion level in tunnel is not high enough, the packets would be marked as CE instead of being dropped, and then it is easy to calculate congestion level according to the ratio of CE-marked packets. If the congestion level is so high that ECT packet will be dropped, then the packet loss ratio could be calculated by comparing total packets entering ingress and total packets arriving at egress over the same span of packets, if packet loss is detected, it could be assumed that severe congestion has occurred in the tunnel. Because loss is only ever a sign of serious

congestion, so it doesn't need to measure loss ratio accurately.

The basic procedure of congestion level measurement is as follows:

```
           +-------+                 +------+
           |Ingress|                 |Egress|
           +-------+                 +------+
               |                         |
       +----------------+                |
       |cumulative count|                |
       +----------------+                |
               |                         |
               | <node id-i, ECN counts> |
               |------------------------>|
               |<node id-e, ECN counts>  |
               |<------------------------|
               |                         |
               |                         |
```

(a) Direct model feedback procedure

```
        +----------+    +-------+                 +------+
        |Controller|    |Ingress|                 |Egress|
        +----------+    +-------+                 +------+
             |              |                         |
             |      +----------------+                |
             |      |cumulative count|                |
             |      +----------------+                |
             |              |                         |
             |              | <node id-i, ECN counts> |
             |              |------------------------>|
             |              |                         |
             |                                        |
             |                                        |
             |       <node id-i, ECN counts>          |
             |       <node id-e, ECN counts>          |
             |<---------------------------------------|
             |                                        |
             |                                        |
             |                                        |
```

(b) Centralized model feedback procedure

Ingress encapsulates packets and marks outer header according to faked ECT as described above. Ingress cumulatively counts packets for three types of ECN combination (CE|CE, ECT|N-ECT, ECT|ECT) and then

the ingress regularly sends cumulative packet counts message of each type of ECN combination to the egress. When each message arrives, the egress cumulatively counts packets coming from the ingress and adds its own packet counts of each type of ECN combination (CE|CE, ECT|N-ECT, CE|N-ECT, CE|ECT, ECT|ECT) to the message and either returns the whole message to the ingress, or to a central controller.

The counting of packets can be at the granularity of the all traffic from the ingress to the egress to learn about the overall congestion status of the path between the ingress and the egress. The counting can also be at the granularity of individual customer's traffic or a specific set of flows to learn about their congestion contribution.

5. Congestion Information Delivery

As described above, the tunnel ingress needs to convey message of cumulative packet counts of each type of ECN combination to tunnel egress, and the tunnel egress also needs to feed the message of cumulative packet counts of each type of ECN combination to the ingress or central collector. This section describes how the messages could be conveyed.

The message can travel along the same path with network data traffic, referred as in band signal; or go through a different path from network data traffic, referred as out of band signal. Because out of band scheme needs additional separate path which might limit its actual deployment, the in band scheme will be discussed here.

Because the message is transmitted in band, so the message packet may get lost in case of network congestion. To cope with the situation that the message packet gets lost, the packet counts values are sent as cumulative counters. Then if a message is lost the next message will recover the missing information.

IPFIX [RFC7011] is selected as a choice of candidate protocol. IPFIX is preferred to use SCTP as transport. SCTP allows partially reliable delivery [RFC3758], which ensures the feedback message will not be blocked in case of packet loss due to network congestion.

When sending message from ingress to egress, the ingress acts as IPFIX exporter and egress acts as IPFIX collector; when sending message from egress to ingress or controller, the egress acts as IPFIX exporter and ingress or controller acts as IPFIX collector.

5.1 IPFIX Extentions

5.1.1 ce-cePacketTotalCount

Description: The total number of incoming packets with CE|CE ECN marking combination for this Flow at the Observation Point since the Metering Process (re-)initialization for this Observation Point.

Abstract Data Type: unsigned64

Data Type Semantics: totalCounter

ElementId: TBD1

Statues: current

Units: packets

5.1.2 ect0-nectPacketTotalCount

Description: The total number of incoming packets with ECT(0)|N-ECT ECN marking combination for this Flow at the Observation Point since the Metering Process (re-)initialization for this Observation Point.

Abstract Data Type: unsigned64

Data Type Semantics: totalCounter

ElementId: TBD2

Statues: current

Units: packets

5.1.3 ect1-nectPacketTotalCount

Description: The total number of incoming packets with ECT(1)|N-ECT ECN marking combination for this Flow at the Observation Point since the Metering Process (re-)initialization for this Observation Point.

Abstract Data Type: unsigned64

Data Type Semantics: totalCounter

ElementId: TBD3

Statues: current

Units: packets

5.1.4 ce-nectPacketTotalCount

Description: The total number of incoming packets with CE|N-ECT ECN marking combination for this Flow at the Observation Point since the Metering Process (re-)initialization for this Observation Point.

Abstract Data Type: unsigned64

Data Type Semantics: totalCounter

ElementId: TBD4

Statues: current

Units: packets

5.1.5 ce-ect0PacketTotalCount

Description: The total number of incoming packets with CE|ECT(0) ECN marking combination for this Flow at the Observation Point since the Metering Process (re-)initialization for this Observation Point.

Abstract Data Type: unsigned64

Data Type Semantics: totalCounter

ElementId: TBD5

Statues: current

Units: packets

5.1.6 ce-ect1PacketTotalCount

Description: The total number of incoming packets with CE|ECT(1) ECN marking combination for this Flow at the Observation Point since the Metering Process (re-)initialization for this Observation Point.

Abstract Data Type: unsigned64

Data Type Semantics: totalCounter

ElementId: TBD6

Statues: current

Units: packets

5.1.7 ect0-ect0PacketTotalCount

Description: The total number of incoming packets with ECT(0)|ECT(0)

ECN marking combination for this Flow at the Observation Point since the Metering Process (re-)initialization for this Observation Point.

Abstract Data Type: unsigned64

Data Type Semantics: totalCounter

ElementId: TBD7

Statues: current

Units: packets

5.1.8 ect1-ect1PacketTotalCount

Description: The total number of incoming packets with ECT(1)|ECT(1) ECN marking combination for this Flow at the Observation Point since the Metering Process (re-)initialization for this Observation Point.

Abstract Data Type: unsigned64

Data Type Semantics: totalCounter

ElementId: TBD8

Statues: current

Units: packets

6. Congestion Management

After tunnel ingress (or controller) receives congestion level information, then congestion management actions could be taken based on the information, e.g. if the congestion level is higher than a predefined threshold, then action could be taken to reduce the congestion level.

The design of network side congestion management SHOULD take host side e2e congestion control mechanism into consideration, which means the congestion management needs to avoid the impacts on e2e congestion control. For instance, congestion management action must be delayed by more than a worst-case global RTT, otherwise tunnel traffic management will not give normal e2e congestion control enough time to do its job, and the system could go unstable.

The detailed description of congestion management is out of scope of this document, as examples, congestion management such as circuit breaker [CB] and congestion policing [CP] could be applied. Circuit

breaker is an automatic mechanism to estimate congestion, and to terminate flow(s) when persistent congestion is detected to prevent network congestion collapse; Congestion policing is used in data center to limit the amount of congestion any tenant can cause according to the congestion information in the tunnels.

7. Security

This document describes the tunnel congestion calculation and feedback. For feeding back congestion, security mechanisms of IPFIX are expected to be sufficient. No additional security concerns are expected.

8. IANA Considerations

This document defines a set of new IPFIX Information Elements (IE). New registry for these IE identifiers is needed.

TBD1~TBD8.

9. References

9.1 Normative References

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

[RFC3168] Ramakrishnan, K., Floyd, S., and D. Black, "The Addition of Explicit Congestion Notification (ECN) to IP", RFC 3168, September 2001, <http://www.rfc-editor.org/info/rfc3168>.

[RFC3758] Stewart, R., Ramalho, M., Xie, Q., Tuexen, M., and P. Conrad, "Stream Control Transmission Protocol (SCTP) Partial Reliability Extension", RFC 3758, May 2004, <http://www.rfc-editor.org/info/rfc3758>.

[RFC6040] Briscoe, B., "Tunnelling of Explicit Congestion Notification", RFC 6040, November 2010, <http://www.rfc-editor.org/info/rfc6040>.

[RFC7011] Claise, B., Ed., Trammell, B., Ed., and P. Aitken, "Specification of the IP Flow Information Export (IPFIX) Protocol for the Exchange of Flow Information", STD 77, RFC 7011, September 2013, <http://www.rfc-editor.org/info/rfc7011>.

9.2 Informative References

[CONEX] Matt Mathis, Bob Briscoe. "Congestion Exposure (ConEx) Concepts, Abstract Mechanism and Requirements", draft-ietf-conex-abstract-mech-13, October 24, 2014

[CB] G. Fairhurst. "Network Transport Circuit Breakers", draft-ietf-tsvwg-circuit-breaker-01, April 02, 2015

[CP] Bob Briscoe, Murari Sridharan. "Network Performance Isolation in Data Centres using Congestion Policing", draft-briscoe-conex-data-centre-02, February 14, 2014

10. Acknowledgements

Thanks Bob Briscoe for his insightful suggestions on the basic mechanisms of congestion information collection and many other useful comments. Thanks David Black for his useful technical suggestions. Also, thanks Anthony Chan and John Kaippallimalil for their careful reviews.

Authors' Addresses

Xinpeng Wei
Beiqing Rd. Z-park No.156, Haidian District,
Beijing, 100095, P. R. China
E-mail: weixinpeng@huawei.com

Zhu Lei
Beiqing Rd. Z-park No.156, Haidian District,
Beijing, 100095, P. R. China
E-mail:lei.zhu@huawei.com

Lingli Deng
Beijing, 100095, P. R. China
E-mail: denglingli@gmail.com