Tunnel congestion Feedback

(draft-wei-tunnel-congestion-feedback-00)

IETF 87 Berlin, Germany

Problem Statement

- Nature of end user flows, length of sessions can lead to significant variability in aggregate bandwidth demands and latency in managed networks such as mobile backhaul.

- IP-in-IP tunnels used to carry end user flows in a backhaul network can use RFC 6040, Appendix C to calculate congestion experienced in the tunnel.

- However, there is no standard feedback mechanism for congestion information from Egress to Ingress router.

This draft provides a general model and feedback mechanism.

Congestion Feedback Model

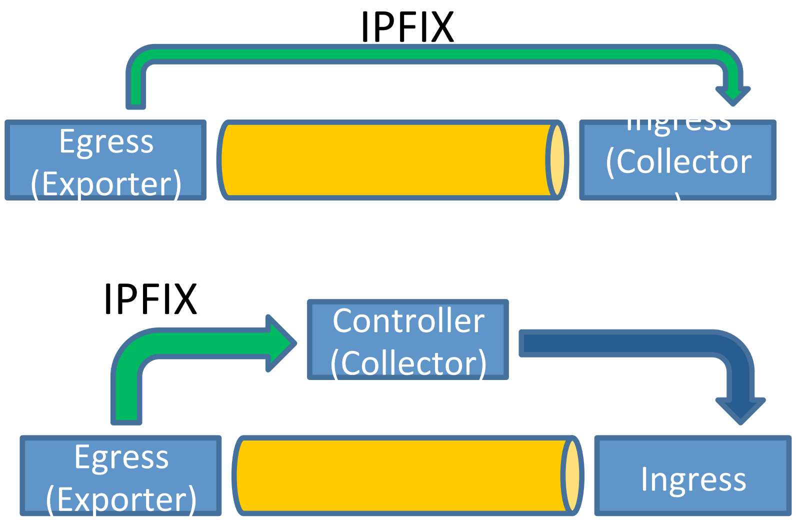
```
,-----------.             IP-in-IP TUNNEL            ,-----------.
|  Ingress  |========================================|  Egress   |
|           |                                        |           |
|     ,-----------Congestion-Feedback-Signals--------------.     |
|     |     |                                        |     |     |
|     |     |                                        |     |     |
|     |     |                                        |     |     |
|     |     |                                        |     |     |
|     |     |                                        |     |     |
|     |     |            ,-----------.          '\   |     |     |
|     |     |____________|           |__________| \  |     |     |
|+----v----+|    Outer Header(IP Layer)  Data Flow \ |+----+----+|
||Collector||            |           |              \||Feedback ||
|+---------+|            |(Congested)|              /|+---------+|
|| Manager ||            |   Router  |Outer-CE-Signals--> Meter ||
|+---------+|____________|           |________    /  |+---------+|
|           |            |           |          |/   |           |
|           |            `-----------'          '    |           |
|           |========================================|           |
`-----------'                                        `-----------'
```

How to calculate congestion in tunnel

- Uses RFC6040 Appendix C.
- Egress calculates congestion in a statistical way. It simply calculates the proportion of packets not marked in the inner header that have a CE marking in the outer header.

- Egress uses moving average (MA) algorithm to calculates congestion, and take the result as current congestion of the tunnel.

IPFIX for Feedback



The information conveyed in IPFIX is in the form of Information Element (IE).
A lot of IEs have been defined in RFC5102, but no congestion related IE has been defined.

A new IE indicating congestion level needs to be defined.
Congestion level may be presented in the form of :

(1) congestion volume [for the statistical method this one is improper.];
(2) percentage.

IPFIX is primarily used to convey Information about IP flows passing through a network element (extend to convey tunnel congestion):

- TCP based (congestion friendly)
- Contains basic mechanisms (periodic, triggered requests etc.)
- Extensible / flexible information export model

Next steps

Comments from group on this approach.

- model for feedback based on RFC 6040, Appendix C.
- Is IPFIX an acceptable method for tunnel congestion feedback?

Backup Slides

How to calculate congestion in tunnel

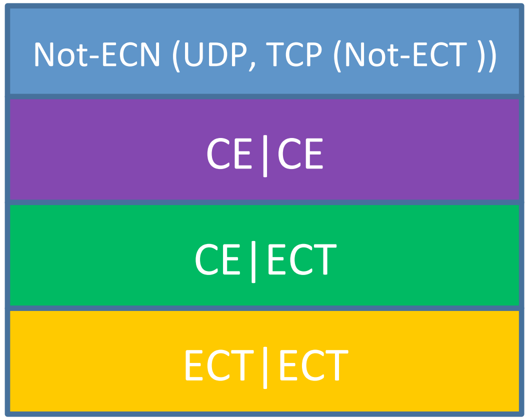
- The algorithm to calculate congestion.

The basic idea of calculating congestion statistically in tunnel is :
Calculating the congestion level of a subset of traffic flows in the tunnel, and take the result as congestion level of the whole tunnel.
Rationale: all the traffics are treated equally by router according to RED, and when congestion occurs in the router, the router randomly selects the packets to mark.

Here we take the traffic that is ECN capable and not congestion-marked before tunnel to calculate congestion.

When ingress is conformant to RFC6040, the packets collected by egress can be divided into to 4 categories, shown as the figure.
The tag before "|" stands for ECN field in outer header; and the tag after "|" stands for ECN field in inner header.

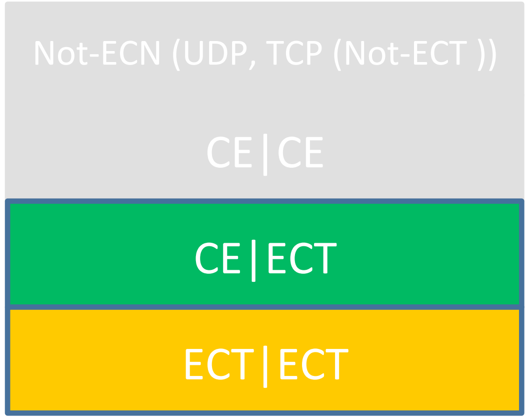
"Not-ECN" means the traffic that don't support ECN, such as UDP and Not-ECT marked TCP;
"CE|CE" means the ECN capable packets that have CE-marked before entering tunnel;
"CE|ECT" means ECN capable packets that CE-marked in tunnel;
"ECT|ECT" means ECN capable packets that have not congested in tunnel.



How to calculate congestion in tunnel

Assuming the quantity of CE|ECT packets is A, the quantity of ECT|ECT packets is B, then the congestion level (C) can be calculate as following:

$$C=A/(A+B)$$



Not-ECN (UDP, TCP (Not-ECT))

CE|CE

CE|ECT

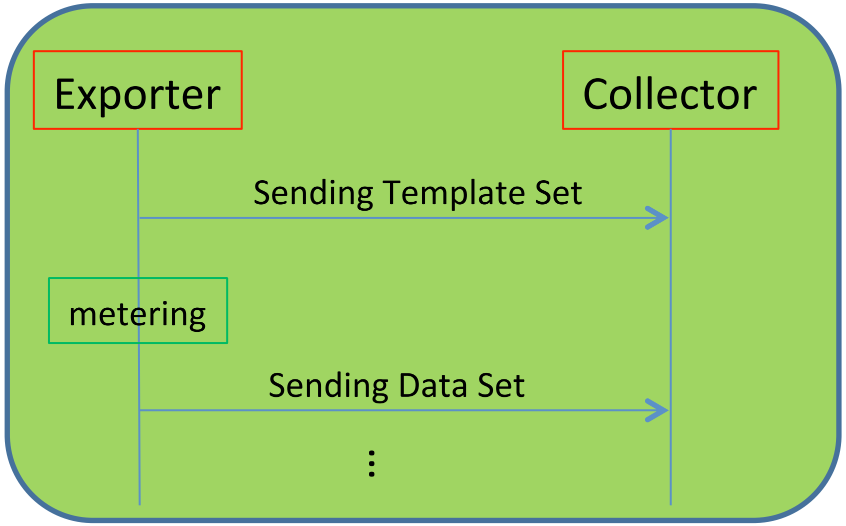
ECT|ECT

Here as an example, we take 100 packets to calculate the moving average. As analyzed above, we just need to take CE|ECT and ECT|ECT packets into consideration. Every time we calculate the congestion, we use the current packet (CE|ECT or ECT|ECT) and the last 99 packets (CE|ECT or ECT|ECT) to get the moving average result which stands for current congestion.

NOTES:
(1) There only shows a simple method of moving average, some other method, such as weight moving average may also be used .
(2) The calculation is based on the assumption that all the packets are treated equally by routers, but the existence of DSCP may has some impacts for this.
(3) According to the calculation process the UDP traffic may don’t have too much impact.
(4) In 3GPP scenario, the congestion may be calculated by ECN field, e.g. when the congestion occurs in RAN.

The basic procedure of congestion feedback

Here an example is shown to illustrate how IPFIX can be used for congestion feedback, the information conveyed here may be incomplete. The exact information to be conveyed from exporter to collector needs further discussion.



(1) Sending Template Set

The exporter use Template Set to inform the collector how to interpret the IEs in the following Data Set.

Set ID=2	Length = octets
Template ID= 257	Field Count =
exporterIPv4Address = 130	Field Length = 4
collectorIPv4Address = 211	Field Length = 4
Congestion Level = TBD1	Field Length = 2
Enterprise Number = TBD2	

Exporter sends Data Set periodically or by trigger.

(2) Sending Data Set

The exporter meters the traffic and sends the congestion information to collector by Data Set.

Set ID = 257	Length = octets
192.0.2.12	
192.0.2.34	
15	

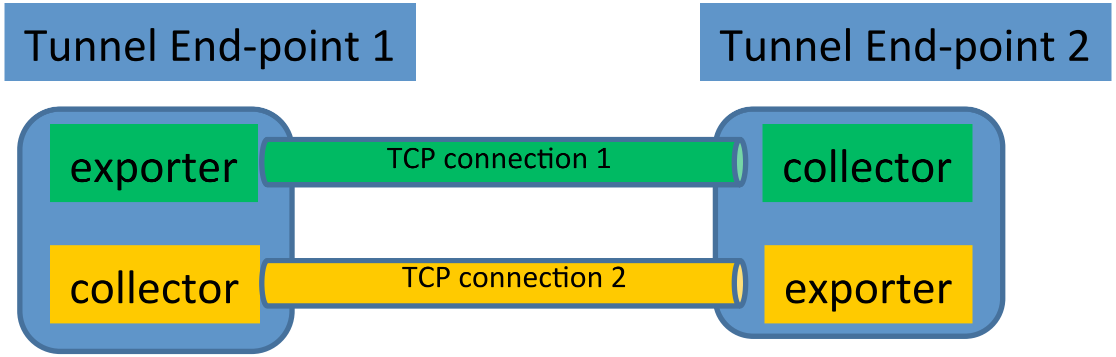
The basic procedure of congestion feedback

Congestion information to be reported:

Congestion volume	mandatory
Egress IP address	mandatory
Ingress IP address	mandatory
……	……

Evaluation of IPFIX

The message flow of IPFIX is unidirectional, which means the information is only from exporter to collector. But in the tunnel scenario, the ingress/egress can host both exporting process and collecting process, so for a pair of ingress and egress there should be two TCP connections to convey IPFIX message bidirectionally. *[NOTES: The need of two TCP connections between ingress and egress may be a shortcoming here. But I am wondering if information in two direction can be transported through one TCP connection.]*



The scenario that two TCP connections are established between two tunnel end-point, each connection for one direction. There will be an exporter and a collector process on each tunnel end-point.

Relationships between tunnel congestion control and e2e ECN control

- Tunnel congestion control is a kind of local congestion control, here we can take tunnel as a administration domain, and it only responds to the congestion happened in the tunnel.
- As all we know, there have been a end-to-end ECN mechanism for congestion control which responds to congestion along the whole path.
- The tunnel congestion control is consistent with e2e ECN control.
- The tunnel congestion feedback provides network administrator with network congestion level information that can be used as an input for network management.
- If the tunnel is congested it will be a waste of resource to allow new traffics enters, because they may eventually dropped in the tunnel, it’s better to have a control on these new traffics at ingress.