6man
Internet-Draft
Updates: 6437 (if approved)
Intended status: Standards Track
Expires: September 17, 2021

C. Filsfils, Ed.
A. Abdelsalam, Ed.
Cisco Systems, Inc.
S. Zadok
Broadcom
X. Xu
Capitalonline
W. Cheng
China Mobile
D. Voyer
Bell Canada
P. Camarillo
Cisco Systems, Inc.

March 16, 2021

Structured Flow Label
draft-filsfils-6man-structured-flow-label-00

Abstract

This document defines the IPv6 Structured Flow Label. The seamless nature of the change to [RFC6437] is demonstrated. Benefits of the solution are explained. Use-cases are illustrated.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction 2
2. Structured Flow Label Format 3
3. Recommended Design 4
4. Seamless Migration from RFC6437 5
5. Benefits 6
6. IPv6 End-to-End Absolute Loss Measurements 6
7. Programmed Sampling of packets 7
8. Postcard-based Telemetry using packet Marking (PBT-M) 8
9. Acknowledgements 9
10. References 9
   10.1. Normative References 9
   10.2. Informative References 10
Appendix A. Entropy 10
Authors' Addresses 11

1. Introduction

The IPv6 header [RFC8200] contains a 20-bit field called "Flow Label" (FL) where the left-most bit is number 19 and the right-most bit is number0.

```
                  1                   0
 9 8 7 6 5 4 3 2 1 0 9 8 7 6 5 4 3 2 1 0
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|             Flow Label                |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 1: IPv6 Flow Label

The FL usage is specified in [RFC6437], briefly for entropy purpose.

Instead of using FL as a single 20-bit entropy structure, this document updates [RFC6437] and defines the 20-bit FL field as a structure of two fields:

- o FLC: 4-bit per-packet control bits for generic application marking (e.g., group-based policy)

o FLE: 16-bit per-flow entropy (equivalent to [RFC6437] definition)

This document shows that updating [RFC6437] is a seamless migration operation, simply because a majority of chipsets deployed in the Internet and private domains do not use FL as documented in [RFC6437]: they use a subset of the 20 bits of the FL as entropy, i.e. as documented in this document.

This document further shows that even if a chipset were to use the full FL as a 20-bit entropy input for ECMP hash, then the change proposed in this document would not cause any significant backward incompatibility.

The seamless nature of the change being explained, the document then explains the benefits of the Structured Flow Label definition. Three use-cases are provided. Several more are expected in the future in separate documents.

2. Structured Flow Label Format

We define the Structured Flow Label as shown in Figure 2

```
                   1                   0
 9 8 7 6 5 4 3 2 1 0 9 8 7 6 5 4 3 2 1 0
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  FLC  |           FLE                 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 2: Structured Flow Label Format

Where:

o FLC: 4-bit "[19, 16]": per-packet Control not included in ECMP hash

o FLE: 16-bit "[15, 0]": per-flow Entropy included in ECMP hash

FLE is defined as per [RFC6437]: i.e. [RFC6437] is strictly preserved, the only change is that it defines the usage of the 16 low-order bits "[15, 0]" instead of the full 20-bit of the Flow Label.

FLC is defined as follows: the 4-bit FLC field in the IPv6 header is used by the network for group-based policy marking. The value of the FLC bits in a received packet or fragment might be different from the value sent by the packet's source. FLC is not included in the ECMP hash computation. The definition of FLC is modeled on the definition of the "Traffic Class" [RFC8200].

In the same way that "Traffic Class" is not an input for ECMP hash, FLC is not an input for ECMP hash.

3. Recommended Design

This section provides design recommendation of how customer packets are being forwarded within an operator network.

All customer packets are typically encapsulated at the edge of the operator network as illustrated in Figure 3.

```
               +------------------------------------------+
               |              Operator IPv6 network       |
               |                                          |
   +---+     +-+--+         +-------+      +------+     +-+--+     +---+
   | A |<>-<>| R1 |<>-----<>|  R2  |<>---<>|  R3 |<>--<>| R4 |<>-<>| Z |
   +---+     +-+--+         +-------+      +------+     +-+--+     +---+
               | +--------+    +--------+    +--------+   |
               | | IP6(R1,|    | IP6(R1,|    | IP6(R1,|   |
               | |     R4,|    |     R4,|    |     R4,|   |
               | |     FL)|    |     FL)|    |     FL)|   |
   +--------+  | +--------+    +--------+    +--------+   |  +--------+
   |Pkt(A,Z)|  | |Pkt(A,Z)|    |Pkt(A,Z)|    |Pkt(A,Z)|   |  |Pkt(A,Z)|
   +--------+  | +--------+    +--------+    +--------+   |  +--------+
               |                                          |
               +------------------------------------------+
```

Figure 3: Packet forwarding within operator IPv6 network

When a customer packet is received at the edge router (R1) of operator IPv6 network, the packet is encapsulated using an outer IPv6 header. The outer IPv6 header defines the source edge router that encapsulates the packet (R1) and the destination edge router (R4) which has to decapsulate the packet before being forwarded towards its final destination.

R1 also sets the Flow Label (FL) of the outer IPv6 header which is computed based on the hash of the customer packet. Every subsequent router (R2 and R3) within the operator network forwards the packet based on the information of the outer IPv6 header.

For example, ECMP hash calculation on routers R2 and R3 is performed using the outer IPv6 header information (R1, R4, FL).

4. Seamless Migration from RFC6437

Cisco and Broadcom report that the norm for their products:

- do not set entropy in the 4 most-specific bits of the FL field
- do not use the 4 most-specific bits of the FL as input for ECMP hash

The authors believe that the same is likely for other vendors and are gathering data for future versions of this document.

Even if a chipset were to use the 4 most-specific bits of the FL field as input for ECMP hash while the 4-most specific bits of the FL field were used by other nodes in the network as FLC semantics (hence, per-packet marking, potentially not per-flow constant), still the impact would not be significant. Let us take an example to illustrate this.

Let us assume that:

- Flow Z is to be routed across an operator network.
- Using the Structured FL format, all the packets of Z have an FLE value of 1010 1111 0100 0101.
- The operator leverages the FLC to mark the packets of Z into two subsets S1 and S2.
- S1 has FLC value of 0000.
- S2 has FLC value of 0010.

We can have the following two scenarios:

Scenario-1: Routers compliant to this document

These routers will only use FLE for ECMP decision and hence all packets of flow F will be routed on the same path.

Scenario-2: Routers implementing RFC6437

These routers will use the full 20-bit (FLC+FLE) for ECMP decision. This could (but not always) lead to having S1 packets taking a different path from the ones of S2.

However, the scenario-2 is unlikely as per the chipset implementation reported in this doc.

5. Benefits

o Seamless migration from RFC6437

o FLE of 16 bits is excellent to drive ECMP hash. [RFC8085] stated that 14 bits are sufficient Appendix A

o FLC of 4 bits provides up to 200 to 400% improvement packet marking capability for operator controlled use-cases

  * Without FLC, operators can only mark 6 bits of the IPv6 header (Traffic Class)

  * Many deployments consume 4 to 5 of these bits, leaving only 1 or 2 available

  * 4 new bits is a significant richness offered to operators to mark packets

o Several use-cases will illustrate the usage of these FLC bits:

  * IPv6 End-to-End absolute loss measurement

  * Programmed sampling of packets

  * Postcard-based Telemetry using packet Marking (PBT-M)

6. IPv6 End-to-End Absolute Loss Measurements

This section describes the usage of FLC bits to enable packet loss measurements [RFC8321] for IPv6 networks. We re-use the same reference topology from RFC8321 for our illustration (Figure 4).

```
    +--------+        +-------+        +--------+        +--------+
--<>|   R1   |<>----<>|  R2   |<>----<>|   R3   |<>----<>|   R4   |<>---
    +--------+        +-------+        +--------+        +--------+
            .                                                    .
            .              End-to-End Loss                       .
            <---------------------------------------------------->
```

Figure 4: End-to-End Absolute Loss Measurement

In order for an operator to enable End-to-End packet loss measurements between routers R1 and R4 for a given flow:

o The operator allocates one bit (C-bit) out of the FLC field to be used for packet coloring.

- o The operator configures R1 to use the C-bit to color packets of the flow of interest.

- o The operator configures R1 and R4 to match the C-bit and perform packet counting.

- o The operator configures R4 to clear the C-bit before forwarding the packet.

- o An SDN controller is used to collect the counters from R1 and R4 to perform End-to-End packet loss measurements.

The flow selection, flow identification, counters update, counters collection and counters correlation considerations are out of the scope of this doc. They can be realized using the techniques described in [RFC8321].

7. Programmed Sampling of packets

An operator can detect End-to-End packet loss by deploying the solution described in Section 6}.

In some cases, the operator needs to identify the node(s) or the link(s) where the packet loss happens. In order to so, the operator would need to collect packet loss measurement from each hop on the packet path. Figure 4 shows the combination of End-to-End and per-hop measurements.

An operator can detect End-to-End packet loss by deploying the solution described in Section 6}.

In some cases, the operator needs to identify the node(s) or the link(s) where the packet loss happens. In order to so, the operator would need to collect packet loss measurement from each hop on the packet path. Figure 5 shows the combination of End-to-End and per-hop measurements.

```
       +------+        +--------+        +--------+         +--------+
    --<>|  R1  |<>----<>|   R2  |<>----<>|   R3  |<>-----<>|   R4  |<>---
       +------+        +--------+        +--------+         +--------+
              .        .       .                 .         .
              .        .       .                 .         .
              .        <------->                 <--------->
              .        Node Loss                  Link Loss
              .                                            .
              .               End-to-End Loss              .
              <-------------------------------------------->
```

Figure 5: End-to-End and Per-Hop Loss Measurements

If the operator detects End-to-End packet loss, it will do the following:

- o The operator allocates another bit (H-bit) out of the FLC field to trigger per-hop measurements.
- o The operator configures R1 to enable H-bit for sample of the flow that experience End-to-End packet loss.
- o The operator configures each router on the packet path (R2 and R3 in Figure 5) to match the H-bit and perform packet counting
- o An SDN controller is used to collect the counters, perform End-to-End and per-hop loss measurements, and identify the node(s) or link(s) where the packet loss happens.

The SDN controller aspects, flow sampling, flow selection, flow identification, counters update, counters collection and counters correlation considerations are out of the scope of this doc. Some of these considerations can be realized using the techniques described in [RFC8321].

8. Postcard-based Telemetry using packet Marking (PBT-M)

This section describes the usage of FLC bits to enable packet marking for PBT-M [I-D.song-ippm-postcard-based-telemetry].

PBT-M enables each router along the packet path exports its telemetry data to the telemetry collector in the form of postcards as illustrated in Figure 6. In PBT-M a single bit is needed to mark the packet which then matched by each node to trigger telemetry export from intermediate routers.

```
                    +-----------------------+
       +------------>|   Telemetry Collector |<---------------+
       |             +-----------------------+                |
       |                 ^                 ^                  |
       | Postcard        | Postcard        | Postcard         | Postcard
       |                 |                 |                  |
     +------+        +------+          +------+          +------+
  --<>|  R1  |<>----<>|  R2  |<>-----<>|  R3  |<>-----<>|  R4  |<>---
     +------+        +------+          +------+          +------+
```

Figure 6: Postcard-Based Telemetry using packet Marking (PBT-M)

An operator that wants to deploy PBT-M, can do the following:

- o Allocates one bit (T-bit) out of the FLC field to be used for PBT packet marking.
- o Configures R1 to enable T-bit for sample of the flow of interest
- o Configures each router to match the T-bit and perform packet counting and send a postcard with its telemetry data to the Telemetry collector.
- o An SDN controller is used to the collected postcards and analyze them.

The SDN controller aspects, flow sampling, flow selection, flow identification, postcard generation, postcard collection and postcard correlation and postcard processing considerations are out of the scope of this doc. Some of these considerations are defined in [I-D.song-ippm-postcard-based-telemetry].

9. Acknowledgements

TBD

10. References

10.1. Normative References

[RFC6437] Amante, S., Carpenter, B., Jiang, S., and J. Rajahalme, "IPv6 Flow Label Specification", RFC 6437, DOI 10.17487/RFC6437, November 2011, <https://www.rfc-editor.org/info/rfc6437>.

10.2. Informative References

[I-D.song-ippm-postcard-based-telemetry]
Song, H., Zhou, T., Li, Z., Mirsky, G., Shin, J., and K. Lee, "Postcard-based On-Path Flow Data Telemetry using Packet Marking", draft-song-ippm-postcard-based-telemetry-08 (work in progress), October 2020.

[RFC8085] Eggert, L., Fairhurst, G., and G. Shepherd, "UDP Usage Guidelines", BCP 145, RFC 8085, DOI 10.17487/RFC8085, March 2017, <https://www.rfc-editor.org/info/rfc8085>.

[RFC8200] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", STD 86, RFC 8200, DOI 10.17487/RFC8200, July 2017, <https://www.rfc-editor.org/info/rfc8200>.

[RFC8321] Fioccola, G., Ed., Capello, A., Cociglio, M., Castaldelli, L., Chen, M., Zheng, L., Mirsky, G., and T. Mizrahi, "Alternate-Marking Method for Passive and Hybrid Performance Monitoring", RFC 8321, DOI 10.17487/RFC8321, January 2018, <https://www.rfc-editor.org/info/rfc8321>.

Appendix A. Entropy

Section 5.1.1 of [RFC8085] discusses the usage of UDP for Source Port Entropy and states that 14 bits of Entropy are sufficient for most ECMP applications:

- In IPv6 UDP tunnel, the BCP suggests the usage of UDP source port for ECMP hash calculation.

- A sending tunnel endpoint selects a source port value in the UDP header that is computed from the inner packet information.

- To provide sufficient entropy, the sending tunnel endpoint maps the encapsulated traffic to one of a range of UDP source values.

- The value SHOULD be within the ephemeral port range, i.e., 49152 to 65535, where the high order two bits of the port are set to one.

- The available source port entropy of 14 bits (using the ephemeral port range) plus the outer IP addresses seems sufficient for entropy for most ECMP applications.

Authors' Addresses

Clarence Filsfils (editor)
Cisco Systems, Inc.
Belgium

Email: cf@cisco.com

Ahmed Abdelsalam (editor)
Cisco Systems, Inc.
Italy

Email: ahabdels@cisco.com

Shay Zadok
Broadcom
Israel

Email: shay.zadok@broadcom.com

Xiaohu Xu
Capitalonline
China

Email: Xiaohu.xu@capitalonline.net

Weiqiang Cheng
China Mobile
China

Email: chengweiqiang@chinamobile.com

Daniel Voyer
Bell Canada
Canada

Email: daniel.voyer@bell.ca

Pablo Camarillo Garvia
Cisco Systems, Inc.
Spain

Email: pcamaril@cisco.com