NVO3 Working Group
INTERNET-DRAFT
Intended Status: Informational
Expires: March 23, 2017
H. Chen
Huawei Technologies
September 19, 2016

Using IPID for Performance Monitoring in VxLAN Network
draft-chen-nvo3-ipid-pm-01

Abstract

IP Identification(IPID)is a field in IP header primarily used to uniquely identify the group of fragments of a single IP packet. The value of IPID field in a packet from a specific traffic flow or source IP address keeps increasing until wrapped-around.

This document specifies a method by carefully examining IPID value to monitor the performance of VxLAN network. In this memo packet loss measurement is mainly considered. This method requires no extra hardware support, which means it is compatible with most of the deployed routers or switches. Such a mechanism is applicable to IPv4 network and potential useful in overlay network with different data encapsulation.

Copyright and License Notice

Table of Contents

1. Introduction

Performance Monitoring(PM) is a crucial part of network OAM, which mainly includes the packet loss and delay measurement. PM methods are usually classified into two categories: active(involving the addition of test traffic) or passive(no interference with normal traffic). Both of active and passive methods have their own strengths. Active method needs injecting test traffic from one measurement point to the other point, which can not be guaranteed to experience the same path with the data traffic where Equal Cost Multiple Paths(ECMP) exists. However, in overlay network(e.g. VxLAN) ECMP is common, which means passive method is more appropriate.

IP Identification(IPID) is a field in IP header, which can be used to implement the passive PM method. The example IPv4 header is shown in Figure 1. IPID is primarily used for uniquely identifying the group of fragments of a single IP packet. The value of IPID field in a packet from a specific traffic flow or source IP address keeps increasing until wrapped-around.

```
    0                   1                   2                   3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |Version|  IHL  |Type of Service|          Total Length         |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |         Identification        |Flags|      Fragment Offset    |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |  Time to Live |    Protocol   |         Header Checksum       |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                       Source Address                          |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                    Destination Address                        |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                    Options                    |    Padding    |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 1: Example IPv4 Header

IPID is required to be unique within the maximum lifetime for all packets with a given source address/destination address/protocol tuple. Hence, each packet in a specific flow has a unique IPID. Packets within a flow continuously increase the IPID value till it reaches the maximum value. Then it wraps around and increases again.

An example Controller-based VxLAN network can be shown as Figure 2. There is a controller connects to NVE A and NVE B. Assume there is a flow transmitted from VM1 to VM3(VM1->NVE A->SW M->SW N->NVE B->VM3),

it is necessary to implement the packet loss measurement at NVE A and NVE B.

This document specifies a method by carefully examining IPID value to monitor the performance of Controller-based VXLAN network. In this memo packet loss measurement is mainly considered. The Controller will specify which flow to be monitored. Before start monitoring, it will send the flow information to the specific NVEs. During the monitoring period, the Controller will collect statistical information from the specific NVEs in order to measure t packet loss and delay value.

```
                          ***************************
                          *     +--------------+    *
                          *     |  Controller  |    *
                          *     +-|---------|--+    *
                          *    /  |         |   \   *
                          *   /   |         |    \  *
          +---------+     *  /    |         |     \ *     +---------+
          |+---+    |     * /     |         |      \*     |    +---+|
          ||VM1|    |   +--/+   +-|-+     +-|-+   +-\-+   |    |VM3||
          |+---+    +---+NVE+---+SW +-----+SW +---+NVE+---+    +---+|
          |    +---+|   +-A-+   +-M-+     +-N-+   +-B-+   |+---+    |
          |    |VM2||     *                         *     ||VM4|    |
          |    +---+|     *      VxLAN Overlay      *     |+---+    |
          +---------+     *         Network         *     +---------+
             Tenant       *                         *       Tenant
             System       *                         *       System
                          ***************************
```

Figure 2: Example Controller-based VxLAN Network

This method requires no extra hardware support, which means it is compatible with most of the deployed routers or switches. Such a mechanism is applicable to IPv4 network and potential useful in overlay network with different data encapsulation.

2. Terminology

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].

This document makes use of the following terms, additional terms are defined in [RFC7348]

- o ECMP - Equal Cost Multiple Paths
- o IPID - IP Identification
- o MSB - Most Significant Bit
- o OG - Observation Group
- o PM - Performance Monitoring

3. IPID Overview

This document mainly considers the IPID in IPv4 header. As defined in[RFC791], IPID field holds 16 bits. It is used together with the source and destination address, and the protocol fields, to identify datagram fragments for reassembly.

There used to be some experimental works using IPID field for other purposes, such as for adding packet-tracing information to help trace packets with spoofed source addresses[Savage_2000]. However, [RFC6864] prohibits these kind of uses. It claims that the IPv4 ID field MUST NOT be used for purposes other than fragmentation and reassembly. Besides, [Chen_2004] describes that the 16-bit IPID field carries a copy of the current value of a counter in a host's IP stack. Current versions of Windows implement this counter as a global counter. That is, IPID value is continuously increasing per source IP address. On the contrary, current versions of Linux implement this counter as a per-flow counter. That is, IPID value is continuously increasing in a per flow fashion. The authors also did extensive experiment to prove the incremental feature of IPID value. To sum up, IPID field can only be set by the Tenant-system and used as a sequence number of packets flow.

Observing IPID's incremental feature, it is possible to take one bit in IPID field as the Criterion bit(C bit), to divide one packets flow into several Observation Groups(OGs). By collecting the observed packet number and starting time of each OG from the relevant NVEs, the controller is able to measure packet loss and delay of each flow.

The VxLAN encapsulation [RFC7348] includes an outer IP header and an inner IP header, both of which have its own IPID field - i.e., the outer IPID and the inner IPID respectively. Because it's the inner header that reflects the real flow info, this memo only use the inner IPID for performance monitoring.

Theoretically, each bit of IPID field can be used as the C bit. But selecting the Criterion bit is a little bit tricky, because high-

order bit varies slowly while low-order bit varies quickly. The selection of C bit have to take the flow rate into consideration. To illustrate, as Figure 3 shows, if taking IPID's most significant bit(MSB) as the C bit, then each OG contains up to 2^15 = 32,768 packets. In the real deployment in data center network, most of the user traffic is usually lower than the rate of 1G bps. In this case, IPID will wrap-around in approximate 0.8s. When user traffic is up to 10G bps, the IPID will wrap-around more quickly, may be less than 80ms.

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

Figure 3: Example Criterion Bit

Figure 4 is a simple example to illustrate how the C bit is used to divide the packets flow into sequential OGs. Assuming the first packet observed holds the IPID value 0x00FC(bit 8 = 0). The first 4 packets hold the same C bit(C = 0) while the last 4 packets hold the same C bit(C = 1).

```
Index    H               C                   L
                        +-+
  1      0 0 0 0 | 0 0 0|0|| 1 1 1 1 | 1 1 0 0 <-+
  2      0 0 0 0 | 0 0 0|0|| 1 1 1 1 | 1 1 0 1   | Group 1
  3      0 0 0 0 | 0 0 1|0|| 1 1 1 1 | 1 1 1 0   | (C = 0)
  4      0 0 0 0 | 0 0 1|0|| 1 1 1 1 | 1 1 1 1 <-+
  5      0 0 0 0 | 0 0 1|1|| 0 0 0 0 | 0 0 0 0 <-+
  6      0 0 0 0 | 0 0 1|1|| 0 0 0 0 | 0 0 0 1   | Group 2
  7      0 0 0 0 | 0 0 1|1|| 0 0 0 0 | 0 0 1 0   | (C = 1)
  8      0 0 0 0 | 0 0 1|1|| 0 0 0 0 | 0 0 1 1 <-+
 ...        ...         +-+     ...       ...      Group k
```

Figure 4: Example C bit based OG division

To illustrate, as shown in Figure 2 VM1 initiates a communication to VM3. The packets flow from VM1 to VM3 will go through NVE A/B and underlay switch M/N . The Controller will send a PM command to NVE A and NVE B simultaneously. The PM command specifies the following information:

1. which bit in IPID field will be taken as the C bit;

2. the basic flow information, including IP address of VM1 and VM3

and the the protocol type(e.g. TCP or UDP).

On receipt of this command, NVE A/B will count the transmitted /received packets respectively in each OGs. The OGs are divided based on the value of C bit. An integrated OG could be determined by two adjacent reversal of C bit. To illustrate, as shown in Figure 4, reversal from 0 to 1 could be seen as the start point of group 2 while reversal from 1 to 0 could be seen as the end point of group 2.

When NVE A and B start to count, firstly they have to determine the integrated OGs. Then NVE A and NVE B will report the counting results to the Controller.

The example counting results of NVE A is shown as below

Group index	C bit	pkt num
1	1	a
2	0	b
3	1	c
4	0	d

Table 1: Example counting results of NVE A

Each time an integrated OG is counted, NVE A will report the results to the Controller. The controller will record the time on receipt of the results as t_A.

The example counting results of NVE B is shown as below

Group index	C bit	pkt num
1	0	k'
2	1	a'
3	0	b'
4	1	c'

Table 2: Example counting results of NVE B

NVE B will report the counting results to the controller in the same way as NVE A. The controller will also record the time on receipt of the results as t_B.

In order to determine whether these two OGs are matched, the

Controller has to go through the following two step

1. compare the C bit value of these two OGs,

2. compare |t_A - t_B| with the value of T, where T is the time duration of one single OG. T is determined by the configuration of C bit and the flow rate.

For example, OG(1) in Table 1 has C = 1 while OG(1) in Table 2 has C = 0. These two OGs do not have the same C bit value, thus the Controller does not consider these two OGs are matched. On the other hand, OG(2) in Table 2 is the next immediate OG and has C = 1. These two OGs have the same C bit value, then the Controller will go to next step to compare |t_A - t_B| with T. If |t_A - t_B| < T, then the Controller considers these two OGs are matched. Otherwise, the Controller considers these two OGs are not matched and simply ignores them. For the case these two OGs are matched, packet number counted in these two OGs can be used to determine whether the packet loss take place between NVE A and NVE B.

4. Packet Loss Measurement

Packet loss measurement could be done by comparing the counted packet number between the matched OGs. In the example of Section 3, packet loss could be computed as follows:

Pkt_Loss = |a - a'| + |b - b'| + |c - c'|.

5. Security Considerations

Security considerations are not addressed in this document.

6. IANA Considerations

No IANA action is needed for this document.

7. References

7.1 Normative References

[RFC791] Postel, J., "Internet Protocol", September 1981.

7.2 Informative References

[Chen_2004] Chen, W., Huang, Y., Ribeiro, B., Suh, K., Zhang, H., Silva, E., Kurose, J. and D. Towsley, "Exploiting the IPID field to infer network path and end-system characteristics", 2004.

[RFC6864] Touch, J., "Updated Specification of the IPv4 ID Field", February 2013.

[RFC7348] Mahalingam, M., Dutt, D., Duda, K., Agarwal, P., Kreeger, L., Sridhar, T., Bursell, M. and C. Wright, "Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", August 2014.

[Savage_2000] Savage, S., Wetherall, D., Karlin, A. and T. Anderson, "Practical Network Support for IP Traceback", October 2000.

Authors' Addresses

Hao Chen
Huawei Technologies
101 Software Ave., Yuhuatai Dist.
Nanjing, Jiangsu 210012
China

Phone: +86-25-56628107
EMail: philips.chenhao@huawei.com