Network Working Group Z. Du
Internet-Draft P. Liu
Intended status: Informational China Mobile
Expires: January 8, 2023 July 7, 2022

Micro-burst Decreasing in Layer3 Network for Low-Latency Traffic
draft-du-detnet-layer3-low-latency-05

Abstract

It is complex to support deterministic forwarding in a large scale network because there is too much dynamic traffic in the network and the data model becomes hard to predict after traffic aggregation on the intermediate nodes. This document introduces the problem of micro-bursts in the layer3 network, and analyses the method to decrease the micro-bursts in layer3 network for low-latency traffic.

Requirements Language

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].

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

The DetNet architecture [RFC8655] is supposed to work in campus-wide networks and private WANs, including the large-scale ISP network scenario, such as the 5G bearing network, as mentioned in [RFC8578]. It is essential for the large-scale ISP network to be able to provide the low-latency service. The low-latency requirement exists in both L2 and L3 networks, and in both small and large networks.

However, as talked in [I-D.qiang-detnet-large-scale-detnet], deploying deterministic services in a large-scale network brings a lot of new challenges. A novel method called LDN (Large-scale Deterministic Network) is introduced in [I-D.qiang-detnet-large-scale-detnet] and [I-D.dang-queuing-with-multiple-cyclic-buffers], which explore the deterministic forwarding over a large-scale network.

This document also explores the deterministic service in the large-scale layer 3 network, and analyses the method based on micro-burst decreasing, which can benefit the forwarding of low-latency traffic in the large-scale network.

2. Gaps for Large-scale Layer 3 Deterministic Network

In this document, the large-scale network means that there are many dynamic flows in the network, but it is hard to do per-flow shaping on the intermediate nodes because they have high pressure on forwarding on the data plane.

According to [RFC8655], DetNet operates at the IP layer and delivers service over lower-layer technologies such as MPLS and IEEE 802.1 Time-Sensitive Networking [TSN]. However, the TSN mechanisms are designed for L2 network originally, and cannot be directly used in the large-scale layer 3 network because of various reasons. Some of them are described as below.

Some TSN mechanisms need synchronization of the network equipments, which is easier in a small network, but hard in a large network. It brings in some complex maintenance jobs across a long distance that are not needed before.

Some TSN mechanisms need a per-flow state in the forwarding plane, which is un-scalable. Aggregation methods need to be considered.

Some TSN mechanisms need a constant and forecastable traffic characteristics, which is more complicated in a large network which includes much more flows joining in or leaving randomly and the traffic characteristics are more dynamic.

The main aspects of the problems are the simplicity and the scalability. The former can ensure that the mechanism is easy to deploy, and the second can ensure that the mechanism is able to bear a large number of deterministic services. An analysis job about the requirements of the large scaled DetNet network is being done in [I-D.liu-detnet-large-scale-requirements].

3. Micro-burst Problem in IP Forwarding

The current IP forwarding mechanism is considered to be a good example fulfilling the requirements of simplicity and scalability. However, the traditional IP network is based on statistical multiplexing, and can only provide Best Effort service, short of SLA guaranteed mechanisms.

When we rethink the problem in the current IP forwarding mechanism, we can find that in the current IP network, a long delay in queuing, or some packet losses due to burst are acceptable; however, it may be unacceptable in the deterministic forwarding. Therefore, they have different design principles in the low layer.

The current forwarding mechanism in an IP router, which is based on statistical multiplexing, can not provide the deterministic service because of various reasons. Even be given a high priority, a critical packet can experience a long congestion delay or be lost in a relatively light-loaded network, which is caused by micro-bursts in the network. The "critical packet" here means that the packet is a DetNet packet, and is sensitive to the latency.

Micro-burst is a special case of network congestion, which typically lasts a short period, at the granularity of millisecond. In a micro-burst, a lot of data are received on the interface suddenly, and the temporary bandwidth requirement would be tens of or hundreds of the average bandwidth requirement, or even exceed the interface bandwidth.

In most cases, the buffer on the equipment can handle the micro-bursts. However, in some corner cases, micro-bursts bring in a long delay (for example, at the granularity of millisecond) or even packet loss.

We introduce the main causes of the micro-burst in the following paragraphs.

Firstly, IP traffic has an instinct of burstiness no matter in the macro or micro aspect, i.e., it does not have a constant traffic model even after aggregations.

Secondly, IP network has a flexible topology, where the incoming traffic may exceed the bandwidth of the outgoing interface. For example, an interface with a large bandwidth may need to send traffic to an interface with a smaller bandwidth, or multiple flows from several incoming interfaces may need to occupy the same outgoing interface.

Thirdly, the IP node has been designed to send traffic as quickly as possible, and it is not aware whether the downstream node’s buffer can handle the traffic. For example, Figure 1 below shows the problem of the current IP scheduling mechanism. Before the scheduling in an IP network, the packets are well paced, but after the scheduling, the packets will be gathered even the total traffic rate is unchanged. When an IP outgoing interface receives multiple critical flows from several incoming interfaces, the situation becomes worse. However, an IP router will try to send them as soon as possible, so occasionally, in some later hops, micro-bursts will emerge.

```
          _     _     _     _     _     _     _     _     _     _     _
         | |   | |   | |   | |   | |   | |   | |   | |   | |   | |   | |
      ---------------------------------------------------------------------
                          Before scheduling in an IP network

          _  _  _  _  _  _                 _  _  _  _  _
         | || || || || || |               | || || || || |
      ---------------------------------------------------------------------
                          After scheduling in an IP network
```

Figure 1: Change of the traffic characteristics in an IP network

4. Analysis of the Method to Decrease Micro-bursts

This document analyses the method to support the low latency traffic bearing in an IP network, such as the 5G bearing network, by avoiding micro-bursts in the network as much as possible. The principle in this method is to forward critical and BE traffic separately, and does not distinguish different critical flows on the forwarding plane on the intermediate nodes.

As talked before, the target method should be scalable and easy to deploy. As the intermediate nodes have high pressure on forwarding packets, the target method should not bring in too much complex process on the data plane. Several requirements are listed as follows.

The first is that the DetNet traffic should support aggregation. The intermediate nodes should not do per-flow process on the date plane.

The second is that separation process of the control plane and data plane on the intermediate nodes. The status of the aggregated DetNet traffic on the control plane may change frequently in the large-scale network. We should not assume that the control plane on an intermediate node can interact with the data plane frequently, for example, to change a shaper parameter frequently. On the data plane, some self-decision process should be supported.

5. An Example of Method to Decrease Micro-bursts

In this section, we describes an example of method fulfilling the requirements mentioned in the last section. It is a traffic forwarding method in the DetNet network, which can decrease micro-bursts for the critical traffic. It needs the cooperation of the edge nodes and the forwarding/intermediate nodes in an IP network.

5.1. Working Flow of the Method

Generally, the method contains two steps:

Step1: per flow schedule on the edge node. The purpose is to make sure that each critical traffic has a constant traffic model.

Step2: per interface schedule on the intermediate nodes. Traffic are aggregated to ensure the scalability, and the pacing also makes sure that they do not gather. The purpose is to make the critical traffic be forwarded as the shape when outgoing the edge, not as quickly as possible. We assume that the sending rate of the buffer for the critical traffic is the same as or similar to the receiving rate (how to achieve this is out of scope of this document). If all work well, the buffer will be maintained with a proper depth.

Other requirements include an RSVP-TE liked mechanism with a good scalability, which should be used to make sure the bandwidth is not exceeded on the interface.

5.2. Process of Edge Node

The edge node of the IP network can recognize each critical flows just as in the TSN network, and then give them individually a good shaping. In fact, in TSN mechanisms, no micro-burst will emerge for critical traffic, and each TSN mechanism is proved to be effective under certain conditions.

This document suggests the edge node to shape the critical traffic by using the CBS method in [IEEE802.1Qav], or the shaping methods in [IEEE802.1Qcr]. Generally, the shaping methods can generate a paced traffic for each critical flow.

The parameters of the shaper, such as the sending rate, can be configured for each flow by some means.

5.3. Process of Forwarding Node

For the forwarding node, it is uneasy to recognize each critical flow because of the high pressure of forwarding a large amount of packets. It is suggested that no per-flow state is maintained on the forwarding node. It is to say that, on the forwarding node, the critical flows should be aggregated and handled together.

We do not distinguish each critical flow on the forwarding node, but all the packets of critical flows should have a common identification to be recognisable, which also stands for that the packet is time sensitive. The forwarding node can obtain the identification in the critical packet, and accordingly forward it to a specific queue.

This document suggests that the forwarding node can deploy a specific queue on each outgoing interface to buffer the time sensitive packets waiting to be sent. When receiving a packet of critical traffic, the forwarding node will forward it to the specific queue on the outgoing interface according to its destination address and its identification. The queue will buffer all critical traffic that need to go out through that interface, and will pace them by using methods mentioned in the last section.

A shaping method in TSN is used here instead of the original forwarding method in an IP router, which can make the critical traffic be forwarded orderly instead of as soon as possible. Therefore, micro-bursts can be decreased in the network.

If all the forwarding nodes can do their jobs properly, i.e., they can well pace the critical traffic, no or rare micro-bursts for the critical traffic would take place. In this way, the critical traffic will have a relatively low latency in the IP network with less uncertainty of micro-bursts.

As no per-flow state is maintained on the forwarding node, the sending rate of the shaper is hard to decide. As said in the last session, the sending rate is suggested to be adjusted referring to the incoming rate of the queue. In other words, before sending the critical packet, we should shape the specific queue by using a shaper parameter based on the computing result of the incoming rate of the queue. The purpose is to maintain a proper buffer depth for the queue.

Although it is claimed that the proposed method is simpler than the TSN mechanisms, forwarding/intermediate nodes also need to be updated. The detailed realization of the method on the intermediate nodes is out of scope of this document.

5.4. Analysis of the Proposed Method

The method proposed does not need synchronization, just as the asynchronous mechanisms studied in [IEEE802.1Qcr]. Furthermore, the method has a larger aggregation granularity, which can fulfill the requirements of simplicity and scalability as much as possible. However, in theory, it has a larger uncertainty on the forwarding than the zero congestion loss target in the TSN mechanisms.

We compare three mechanisms in the following paragraphs. The first is the priority based light-load mechanism, i.e., the traditional method. The second is the TSN mechanism, such as CQF. The third is the proposed mechanism.

In the first mechanism, we only give a high priority to the critical traffic, and thus the scalability of the deterministic system is good. However, the uncertainty on the forwarding plane perhaps can not fulfill the requirements in the industry network where SLA requirements are very essential. Perhaps, it is only able to work well when a small amount of critical traffic exist in the network.

If we use the scheduling method in the TSN, such as CQF. Its uncertainty is very low, but its scalability is not very good as said in Section 2. It should be noted that in a large deterministic system, the ISP normally will not guarantee the user 100 percent reliability, instead of which it perhaps is a value very close to.

The proposed method has a better scalability than the TSN mechanisms, and a better reliability than the priority based method. If we assume that different services need different deterministic levels, this method may be helpful for the service that does not need a very high deterministic level. For example, the method can be used in the consumption Internet, in which the deterministic service needs a relatively lower deterministic level than the industry Internet.

6. IANA Considerations

This document has no IANA actions.

7. Security Considerations

Detailed security considerations can refer to [I-D.ietf-detnet-bounded-latency] and [I-D.ietf-detnet-security].

8. Acknowledgements

Thanks for the valuable comments from Janos Farkas, Lou Berger, and David Black.

9. References

9.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC8655] Finn, N., Thubert, P., Varga, B., and J. Farkas, "Deterministic Networking Architecture", RFC 8655, DOI 10.17487/RFC8655, October 2019, <https://www.rfc-editor.org/info/rfc8655>.

9.2. Informative References

[I-D.dang-queuing-with-multiple-cyclic-buffers]
Liu, B. and J. Dang, "A Queuing Mechanism with Multiple Cyclic Buffers", draft-dang-queuing-with-multiple-cyclic-buffers-00 (work in progress), February 2021.

[I-D.ietf-detnet-bounded-latency]
Finn, N., Boudec, J. L., Mohammadpour, E., Zhang, J., and B. Varga, "DetNet Bounded Latency", draft-ietf-detnet-bounded-latency-10 (work in progress), April 2022.

[I-D.ietf-detnet-security]
Grossman, E., Mizrahi, T., and A. J. Hacker, "Deterministic Networking (DetNet) Security Considerations", draft-ietf-detnet-security-16 (work in progress), March 2021.

[I-D.liu-detnet-large-scale-requirements]
Liu, P., Li, Y., Eckert, T., Xiong, Q., and J. Ryoo, "Requirements for Large-Scale Deterministic Networks", draft-liu-detnet-large-scale-requirements-02 (work in progress), April 2022.

[I-D.qiang-detnet-large-scale-detnet]
Qiang, L., Geng, X., Liu, B., Eckert, T., Geng, L., and G. Li, "Large-Scale Deterministic IP Network", draft-qiang-detnet-large-scale-detnet-05 (work in progress), September 2019.

[IEEE802.1Qav]
IEEE 802.1, "IEEE 802.1Qav-2009 - IEEE Standard for Local and metropolitan area networks-- Virtual Bridged Local Area Networks Amendment 12: Forwarding and Queuing Enhancements for Time-Sensitive Streams", 2009, <https://standards.ieee.org/standard/802_1Qav-2009.html>.

[IEEE802.1Qcr]
IEEE 802.1, "IEEE 802.1Qcr-2020 - IEEE Standard for Local and Metropolitan Area Networks--Bridges and Bridged Networks - Amendment 34: Asynchronous Traffic Shaping", 2020, <https://standards.ieee.org/standard/802_1Qcr-2020.html>.

[RFC8578] Grossman, E., Ed., "Deterministic Networking Use Cases", RFC 8578, DOI 10.17487/RFC8578, May 2019, <https://www.rfc-editor.org/info/rfc8578>.

[TSN] IEEE 802.1, "Time-Sensitive Networking (TSN) Task Group", 2012, <https://1.ieee802.org/tsn/>.

Authors’ Addresses

Zongpeng Du
China Mobile
No.32 XuanWuMen West Street
Beijing 100053
China

Email: duzongpeng@foxmail.com

Peng Liu
China Mobile
No.32 XuanWuMen West Street
Beijing 100053
China

Email: liupengyjy@chinamobile.com