Network Working Group Z. Du
Internet-Draft P. Liu
Intended status: Informational China Mobile
Expires: August 26, 2021 February 22, 2021

Micro-burst Decreasing in Layer3 Network for Low-Latency Traffic
draft-du-detnet-layer3-low-latency-02

Abstract

This document introduces the problem of micro-bursts in layer3 network, and proposed a 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].

Table of Contents

1. Introduction

The DetNet architecture in RFC 8655 [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 mentioned in RFC 8578 [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], which explores the deterministic forwarding over a large-scale network.

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

2. Gaps for Large-scale Layer 3 Deterministic Network

According to RFC 8655 [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 large 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.

3. Rethinking the Problem in IP Forwarding

As a comparison, the current IP forwarding mechanism is considered to be a good example fulfilling the requirements of simplicity and scalability. However, 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 is unacceptable in the deterministic forwarding. Therefore, they have different design principles in a low layer.

The current forwarding mechanism in an IP router, which is based on statistical multiplexing, cannot provide the deterministic service because of various reasons. Even be given a high priority, a deterministic packet can experience a long congestion delay or be lost in a relatively light-loaded network, which is caused by micro-burst in the network.

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 (at the granularity of millisecond) or even packet loss.

The following paragraphs introduce the causes of the micro-burst.

Firstly, IP traffic has a 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, and 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

This document proposes a 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 do not distinguish different critical flows in the intermediate nodes on the forwarding plane.

4. Method to Decrease Micro-bursts

The method needs the cooperation of the edge nodes and the forwarding/core nodes in an IP network.

4.1. Working Flow of the Method

Generally, the method contains two steps:

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

Step2: per interface schedule in the core node. 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 the receiving rate (maybe an algorithm is needed here). If all work good, the buffer will be maintained with a proper depth.

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

4.2. 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-busrt 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 IEEE 802.1Qav, or the shaping methods in IEEE 802.1Qcr. 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.

4.3. 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 in the forwarding node. It is to say that, in the forwarding node, the critical flows should be aggregated and handled together.

This document suggests that the forwarding node can deploy a specific queue at each outgoing interface. 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.

The 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 uncertainties of micro-bursts.

As no per-flow state is maintained in the forwarding node, the sending rate of the shaper is hard to decide. In this document, the sending rate is suggested to be generated referring to the incoming rate of the queue. The purpose is to maintain a proper buffer depth for the queue.

5. Analysis of the Proposed Method

The method proposed does not need synchronization, just as the asynchronous mechanisms studied in IEEE 802.1 Qcr. Furthermore, the method has a larger aggregation granularity, which can fulfill the requirements of simplicity and scalability. However, it has a larger uncertainty in the forwarding than the TSN mechanisms, which needs to be further studied.

6. IANA Considerations

TBD.

7. Security Considerations

TBD.

8. Acknowledgements

TBD.

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>.

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

[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.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.

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