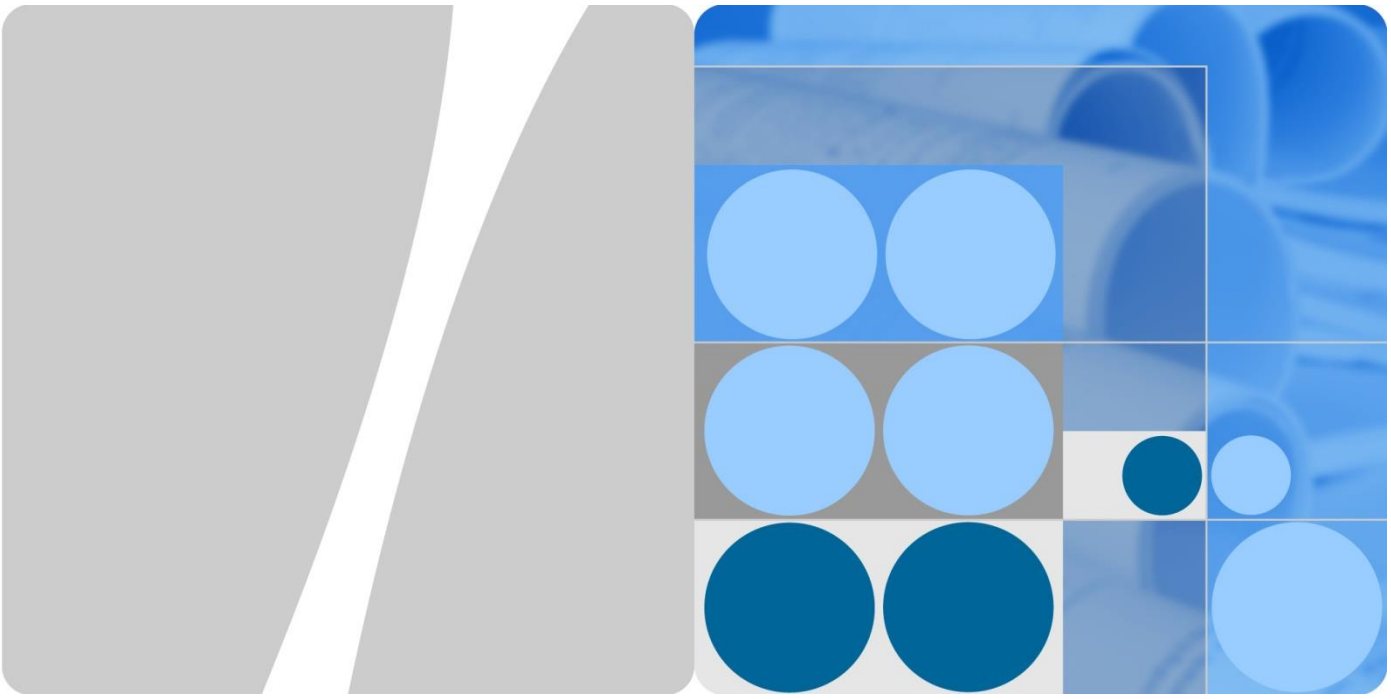


White Paper on Buffering Requirements for Data Center Switches

Issue 01

Date 2018-01-30

HUAWEI TECHNOLOGIES CO., LTD.

Trademarks and Permissions


and other Huawei trademarks are trademarks of Huawei Technologies Co., Ltd.

All other trademarks and trade names mentioned in this document are the property of their respective holders.

Notice

The purchased products, services and features are stipulated by the contract made between Huawei and the customer. All or part of the products, services and features described in this document may not be within the purchase scope or the usage scope. Unless otherwise specified in the contract, all statements, information, and recommendations in this document are provided "AS IS" without warranties, guarantees or representations of any kind, either express or implied.

The information in this document is subject to change without notice. Every effort has been made in the preparation of this document to ensure accuracy of the contents, but all statements, information, and recommendations in this document do not constitute a warranty of any kind, express or implied.

Huawei Technologies Co., Ltd.

Address: Huawei Industrial Base
Bantian, Longgang
Shenzhen 518129
People's Republic of China

Website: http://www.huawei.com

Email: support@huawei.com

Change History

Date	Version	Change Description	Author
2017-09-21	V1.0	This is the first draft.	Zheng Xiaolong 00413398
2017-09-26	V1.1	The application of classical theories and the setting of the ECN threshold are added.	Zheng Xiaolong 00413398

Contents

About This Document

This document answers two basic questions about the buffer of data center (DC) switches: buffer size required by a switch and the Explicit Congestion Notification (ECN) threshold of the buffer.

Based on a detailed analysis of the classical buffer theories and core viewpoints of the buffer in the industry, combined with simulation analysis, this document explores the relationship between the buffer, packet loss, throughput, and flow completion time (FCT) in a data center, and finally provides the guiding conclusion on buffer and threshold settings.

Keywords: DC, buffer, packet loss, ECN, FCT

NOTE

All simulation tests in this document are based on the NS3 simulation platform and the TCP congestion avoidance algorithm is New Reno.

1 Overview

1.1 Network Performance Requirements of a DC

In recent years, DCs have become important infrastructure for carrying rapidly developing applications and services such as Big Data, cloud computing, social networking, and Internet of Things (IoT).

The advances of informatization and the rapid development of mobile Internet industries, especially the explosive growth in industries such as video, live broadcast over the Internet, and game, pose higher requirements on users' access experience. The wide application of cloud computing technologies drives the dramatic increase in data storage scale, computing capability, and network traffic. In addition, the development of IoT, smart city, and artificial intelligence poses more requirements for DCs.

To address growing network requirements, the network performance in a DC must meet the following conditions:

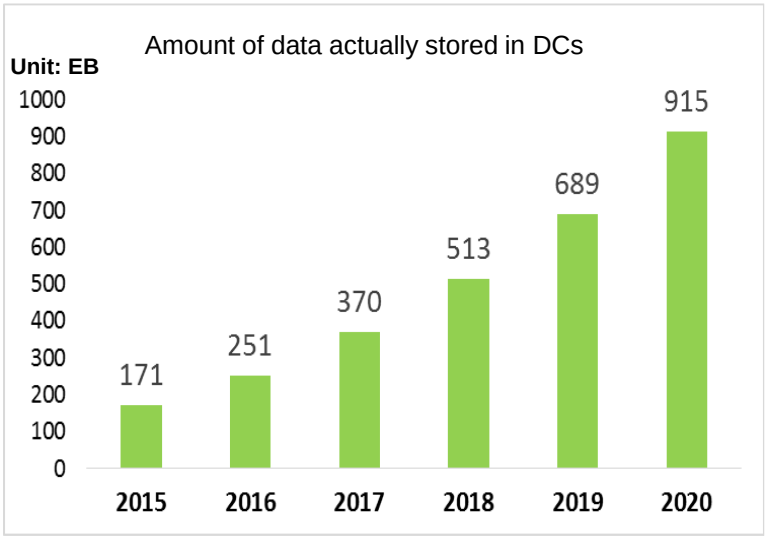
- Low latency

  The emergence and development of technologies such as deep learning and distributed computing drive fast growth in delay-sensitive services such as artificial intelligence and high-performance computing. With rapid development of computer hardware, network has replaced computing capability to become the new bottleneck of these applications, and low latency has become a key indicator that affects the computing performance of clusters. Delay-sensitive applications have even higher requirements on DC network latency. An E2E latency of 5 to 10 microseconds in a DC has become the target of mainstream vendors.

- High bandwidth and high throughput

  In the data era, a massive amount of data is generated in DCs each year, and the amount keeps increasing, as shown in Figure 1-1. The popularization of data-based applications, such as image recognition, promotes explosive growth in network data. As a result, low bandwidth cannot meet the requirement of applications that need a high transmission rate. In some application scenarios, low bandwidth has even become the bottleneck of user experience. High bandwidth and high throughput are critical to improving the performance of applications that involve transmission of large amount of data. To meet the requirements of these applications, enterprises such as Baidu, Tencent, and Alibaba have all had 100GE networks deployed in their DCs, and Alibaba even plans to deploy 400GE networks in 2020.

Figure 1-1 Amount of data actually stored in DCs



Data source: www.idcquan.com

- Extremely low packet loss rate

  If packet loss occurs, packets need to be retransmitted, which may cause retransmission timeout, resulting in bandwidth waste. Currently, many DCs provide lossless networks by enabling Priority-based Flow Control (PFC).

1.2 Confusion About Buffer for OTT Vendors in China

To meet the higher requirements for DC network performance, in addition to the optimization of upper layer applications, OTT vendors in China want to optimize the DC network architecture and switches to improve network performance. The buffer of DC switches becomes a major concern.

After communication with OTT vendors in China, we find that OTT vendors have two questions about the buffer:

- What is the size of the buffer required by a switch? The buffer size has always been of great concern. The large buffer can absorb burst traffic, reduce lost packets, and increase the queuing latency. The small buffer ensures a low queuing latency but cannot absorb burst traffic, which affects the bandwidth usage of links. OTT vendors urgently want to know the buffer size of a device that can meet service requirements in the target scenario.
- How to set the ECN threshold after ECN is enabled? ECN enables the device to feed back congestion information as early as possible to avoid long queuing latency and reduce lost packets. However, the setting of ECN threshold has a great impact on network performance. If the threshold is too low, the link will be underflow. If the threshold is too high, the latency increases, and the advantage of congestion feedback enabled by ECN cannot be displayed. Therefore, OTT vendors urgently need the guidance on the ECN threshold setting.

1.3 Objectives

In this white paper, we analyze the viewpoints and conclusions about buffer of DC switches in the industry. Based on the understanding of buffer and combined with theoretical and simulation analysis, we attempt to answer the two basic questions about the buffer: buffer size required by a switch and the ECN threshold of the buffer.

Currently, the mainstream scenarios on the DC network include TCP congestion control network based on tail drop, TCP congestion control network based on ECN, and TCP congestion control network enabled with differentiated scheduling of elephant and mice flows. Therefore, this white paper analyzes buffering requirements of switches in three Data Center Network (DCN) scenarios and is expected to provide guidance on the design and usage of buffer.

2 Classical Theories of Buffering Requirement

2.1 1BDP Theory

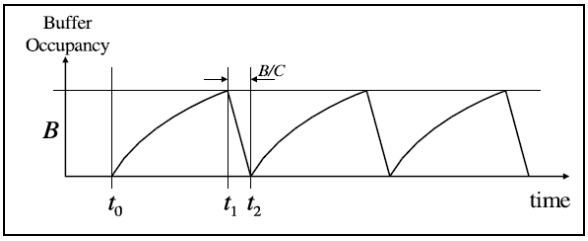
The earliest classical theory about the buffer size of a network device is the bandwidth-delay product (BDP) theory proposed by Villamizar and Song in 1994, which is also called rule of thumb. BDP is calculated as follows:

BDP = C x RTT

C indicates the bandwidth of the bottleneck link, and Round Trip Time (RTT) indicates the round-trip transmission latency of the link.

Assumptions of 1BDP theory are as follows: (1) The transport layer protocol is the TCP based on packet loss, and the packet loss rate is reduced by half. (2) Only one elephant flow exists on the link. (3) The network has only one hop. Figure 2-1 shows the buffer occupancy of the TCP flow. (4) The buffering requirement aims to ensure that the link bandwidth is fully occupied.

Figure 2-1 Buffer occupancy of a TCP flow



At t_1, the TCP sender detects packet loss. In this case, the TCP congestion window (CWND) decreases from Wmax to Wmax/2, and the inflight of the TCP flow is equal to Wmax. Since the inflight data is in the transmission channel or buffer queue, the inflight is calculated as follows:

Inflight = BDP + B

B indicates the buffer size, and Wmax is calculated as follows:

Wmax = BDP + B

At t_2, the buffer queue is empty, the inflight is equal to or less than CWND, and the sender sends data with the CWND of Wmax/2. In this case, there is no queuing latency for data packets, and the RTT of the data packet is equal to the round-trip transmission latency of the link, so the sending rate of the TCP flow is calculated as follows:

Sending rate = Wmax/2/RTT

To ensure that the link bandwidth is fully occupied, the sending rate of the TCP flow should be equal to the bandwidth of the bottleneck link C. So C is calculated as follows:

C = Wmax/2/RTT

And Wmax is calculated as follows:

Wmax = 2C x RTT

In conclusion, Wmax can be calculated as follows:

Wmax = BDP + B = 2C x RTT

So B can be calculated as follows:

B = BDP = C x RTT

In other words, **in the single-flow scenario, 1BDP is enough to ensure the full occupancy of link bandwidth**.

In the network topology of more than one hop, when the buffer of the current device is empty at t2, the RTT of the flow includes the buffer queuing latency of other devices except that of the current device. Therefore, both the queuing latency and transmission latency need to be considered to ensure that the link is underflow.

2.2 Nick Mckeown Theory

On the basis of the 1BDP theory, the Stanford professor Nick Mckeown considers multi-flow synchronization and asynchronization and further extends theories of buffering requirement.

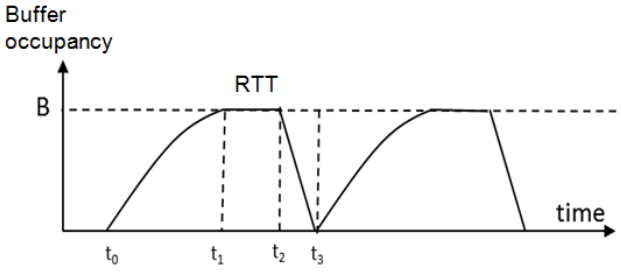
1. Multi-flow synchronization. Flow synchronization refers to the in-phase synchronization of CWND curves of different flows, that is, CWNDs increase and decrease simultaneously. Generally, multi-flow synchronization is increased when multiple flows are sent simultaneously, packets of multiple flows are lost simultaneously on a switch, and RRTs of multiple flows are similar. For multiple flows transmitted on the same bottleneck link, the integrated CWND curve is the integration of multiple CWND curves, which is similar to the CWND curve of a single flow. Therefore, the derivative logic of the 1BDP theory is applied to obtain the following conclusion: **In a multi-flow synchronization scenario, 1BDP is enough to ensure the full occupancy of link bandwidth**.
2. Multi-flow asynchronization. **When flows are not synchronized**, the integrated CWND curve of multiple flows will be staggered. If the number of flows is large, **assuming that CWNDs of multiple flows are independent identically distributed**, the integrated CWND follows the normal distribution according to the central limit theorem. In this scenario, there is $\sqrt{n}$ relationship between the integrated CWND of multiple flows and CWND of a single flow, where **n** is the number of elephant flows. Therefore, in a multi-flow asynchronization scenario, to ensure that the link is underflow, the buffer required by a switch is BDP/ $\sqrt{n}$.

2.3 Applicability of Classical Theories

The classical theories are proposed for buffering requirement of backbone routers on the WAN. Compared with the WAN, the data center has fewer hops, shorter latency, and different service traffic models. Are the classical theories applicable on the DCN?

Is the BDP theory still applicable in the multi-flow scenario? Considering that buffer overflows and packet loss occurs, the sender needs three redundant acknowledgments (ACKs) to reduce the CWND by half. The buffer occupancy curve of multiple flows is shown in Figure 2-2. If there are **n** flows in the DCN, the CWND of flow i is W_i, and the RTT of flow i is RTT_i.

Figure 2-2 Buffer occupancy of multiple TCP flows



At t_1, the buffer overflows and packet loss occurs. However, the source cannot immediately detect packet loss. After receiving an ACK, the source continues to increase the CWND and sends data packets. As a result, the buffer is occupied. At t_2 after 3RTT, each flow detects packet loss after receiving three redundant ACKs. In this case, CWNDs of multiple flows decrease from W_i to $W_i/2$. Before the CWND decreases at t_2, the sum of CWND of each flow is $\sum_n W_i$, which is the total number of inflight data packets. These packets are in the transmission channel or buffer queue, so $\sum_n W_i$ is calculated as follows:

$$\sum_n W_i = \text{BDP} + \text{B}$$

During t_2 to t_3, since the CWND of each flow is smaller than the inflight, each flow stops sending packets, and the buffer queue is empty.

At t_3, each flow has received sufficient ACKs so that the inflight is less than or equal to $\sum_n W_i/2$. Therefore, each flow starts to transmit data. To ensure that the link is underflow, the sum of the sending rates of flows should be the bandwidth of the bottleneck link C. Therefore, $\sum_n \frac{W_i/2}{RTT_i} = C$. Assuming that RTT_{max} is equal to max{RTT_i}, C is calculated as follows:

$$C = \sum_n \frac{W_i/2}{RTT_i} \geq \sum_n \frac{W_i/2}{RTT_{\max}}$$

So $\sum_n W_i$ is calculated as follows:

$$\sum_n W_i \leq 2C \times RTT_{\max}$$

The following formula is used to substitute the above formula:

$\sum_n W_i$ = BDP + B

And the conclusion is obtained as follows:

$$B \le C \times RTT_{\max} = BDP_{\max}$$

Therefore, 1BDP is enough to ensure the full occupancy of link bandwidth. That is, the 1BDP theory is still applicable in the data center.

Nick's theory must meet the following two basic assumptions: multi-flow asynchronization and independent CWNDs. If the CWNDs of multiple flows on the network do not meet the synchronization and independence requirements, the $\sqrt{n}$ relationship between the integrated CWND of multiple flows and CWND of a single flow will become invalid. On the DCN, especially the DCN that focuses on applications such as Big Data and image recognition, incast patterns are very common in the partion and aggregation applications such as the MapReduce and parameter servers. In an incast traffic pattern, if multiple flows are sent simultaneously or almost at the same time, CWNDs of multiple flows start to increase. Since the data center has fewer hops and RTTs of multiple flows are close, packet loss occurs when the buffer overflows, so CWNDs decrease simultaneously, resulting in the improvement of multi-flow synchronization. Therefore, **CWNDs of multiple flows on the DCN are synchronized in many application scenarios**. Generally, congestion occurs on the DCN in the incast traffic pattern or scenario when load balancing fails. In this case, multiple flows share the same bottleneck link, and CWNDs of multiple flows are limited by the same physical bandwidth. So **the CWNDs of multiple flows on the DCN are not independent of each other when congestion occurs**.

In conclusion, in most multi-flow scenarios, the assumptions of multi-flow asynchronization and independent CWNDs cannot be met. So **Nick's theory is not applicable in the DC**.

To verify the applicability of classical theories, we simulate the link throughput of 120 synchronized elephant flows when the buffer size is 1BDP and BDP/$\sqrt{n}$ in Figure 2-3. Figure 2-4 and Figure 2-5 show the results.

Figure 2-3 Test topology

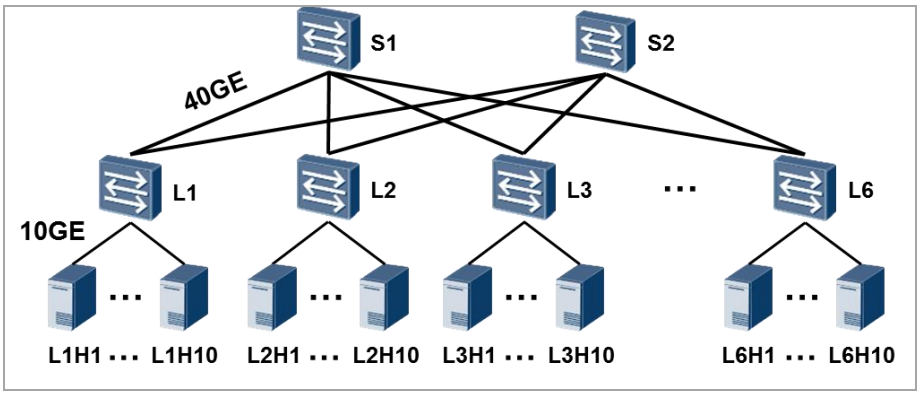


Figure 2-4 Link throughput when the buffer is equal to 1BDP

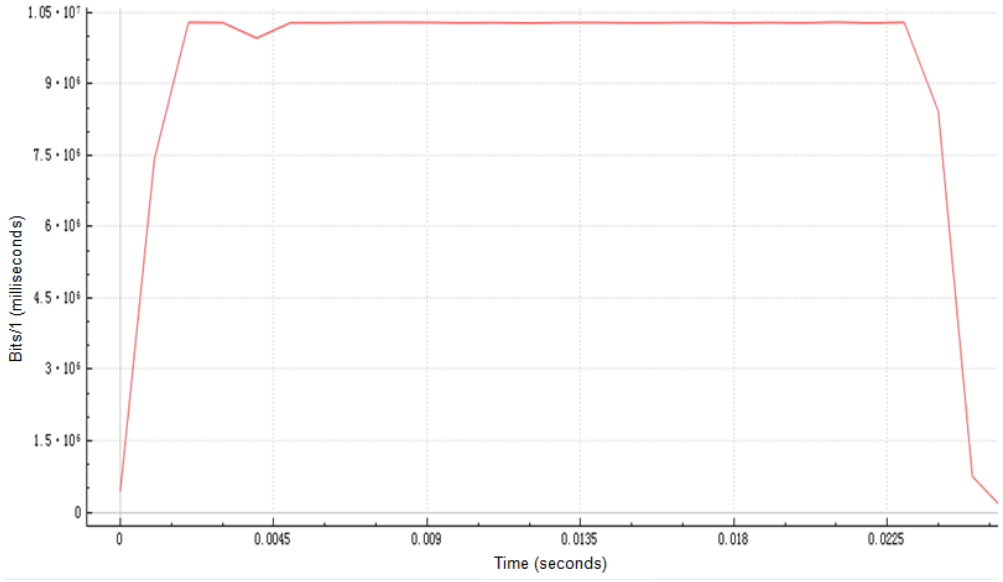
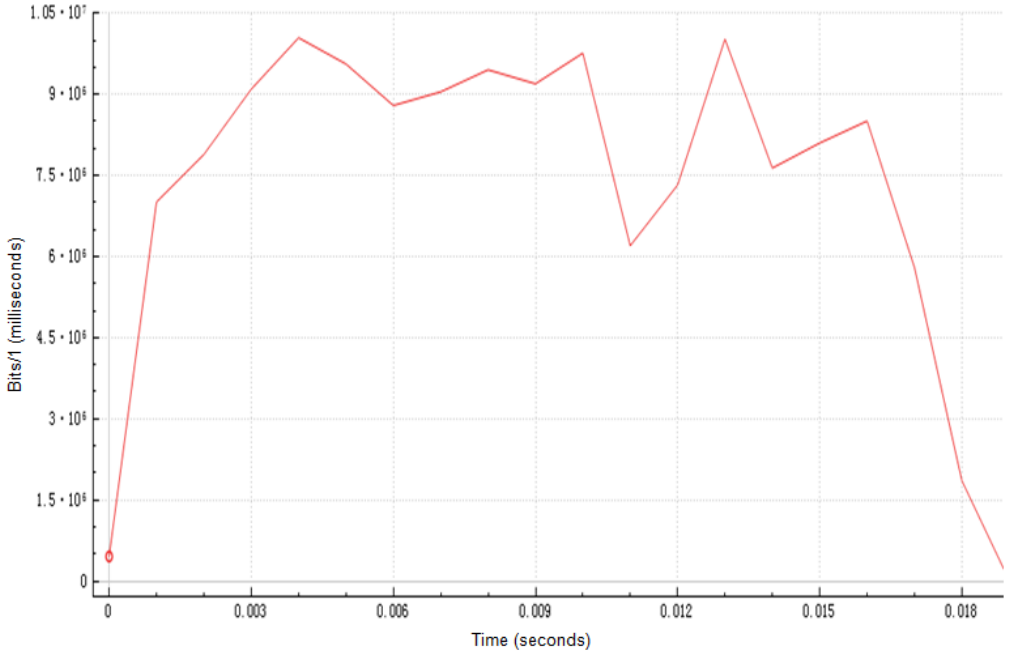


Figure 2-5 Link throughput when the buffer is equal to BDP/$\sqrt{n}$



The results show that when the buffer is equal to 1BDP, the link bandwidth can be fully occupied. When the buffer is equal to BDP/$\sqrt{n}$, the link bandwidth is underflow. Therefore, in the incast scenario of the DC, Nick's theory is not applicable.

3 Buffering Requirement Based on Tail Drop

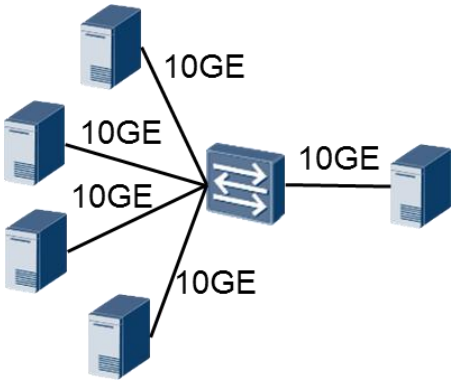
Currently, mainstream DCNs in China use the TCP based on packet loss, and most switches in the DC are configured with the buffer queue management mechanism based on tail drop. Therefore, TCP congestion control based on tail drop is a common scenario on the live network. Broadcom and Arista released the white paper about buffering requirements in this scenario in 2015 and 2016 respectively, and put forward the conclusion of large buffer on switches.

Based on simulation analysis and combined with white papers of Broadcom and Arista, this chapter analyzes the core problems related to buffering requirements in the TCP scenario based on tail drop, including the impact of packet loss, functions of large buffer, large buffer required in the DC, buffer size, and changes of buffering requirements after bandwidth upgrade.

3.1 Impact of Packet Loss

7 For the TCP based on packet loss, packet loss is an important factor that affects the DCN performance. This section uses the topology shown in Figure 3-1 to perform simulation analysis on the impact of packet loss.

Figure 3-1 Topology of traffic simulation on a ToR switch



3.1.1 Impact of Packet Loss on Bandwidth Usage

A direct perception of the impact on bandwidth usage is that if the packet loss rate is high, the bandwidth usage is low. In fact, packet loss has different impacts on bandwidth usage in different scenarios.

- Impact of packet loss on bandwidth usage in the case of all elephant flows under different loads. In simulation analysis, different flows are configured with different loads, and different packet loss rates and throughput loss are obtained on interfaces of different switches. As shown in Figure 3-2, the throughput loss is calculated as follows:

  Throughput loss = (1 – actual bandwidth usage/load) x 100%

  The following figure shows that when the network works at 50% load, the packet loss rate is 0.2% and the throughput loss reaches 35%, and **the bandwidth usage of large flows is greatly affected**. Since the stable CWND of an elephant flow is large, the CWND decreases by half after packet loss occurs, resulting in a large bandwidth loss.

Figure 3-2 Packet loss rate and throughput loss of elephant flows under different loads

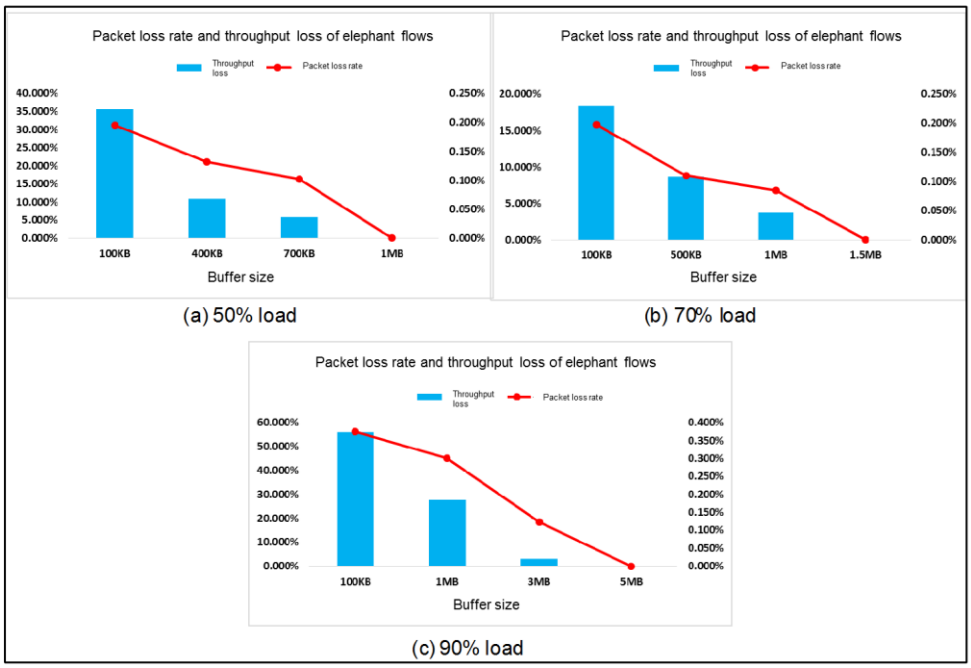
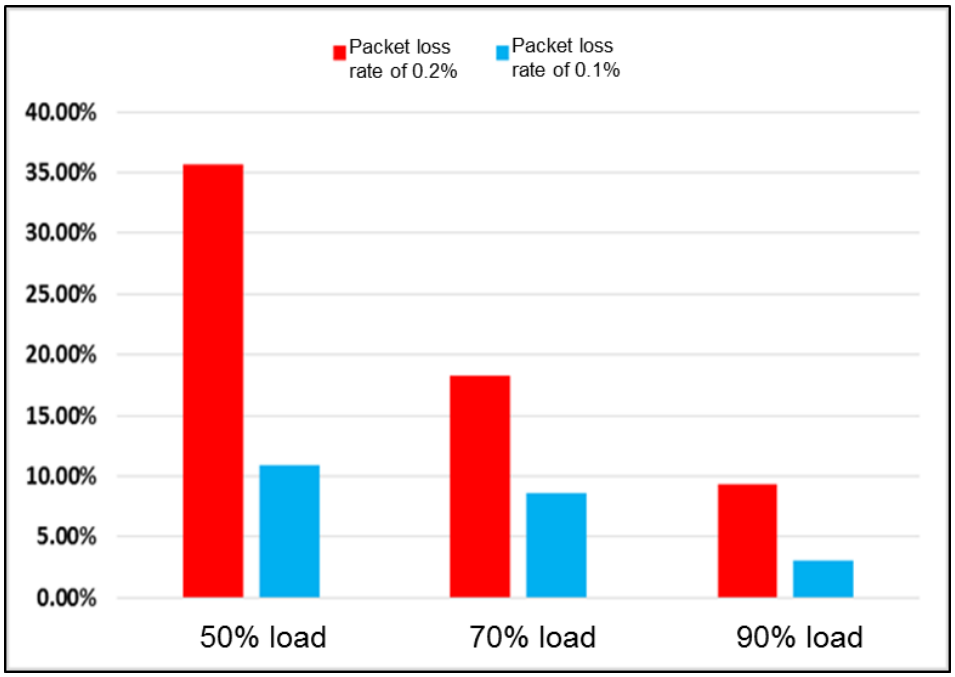


Figure 3-3 shows the relationship between the packet loss rate and bandwidth loss under different loads (the number of flows).

Figure 3-3 Relationship between the packet loss rate and throughput loss of elephant flows under different loads



As shown in the above figure, if the packet loss rate decreases, the bandwidth loss decreases significantly. **When the packet loss rate is the same, if the load is heavy (the number of elephant flows is large), the bandwidth loss is little.** The reason is that a larger number of elephant flows result in the smaller stable CWND of each flow. If packet loss occurs and CWND decreases by half, the bandwidth loss is little.

- Impact of packet loss on bandwidth usage in the case of all mice flows under different loads. In simulation analysis, each host randomly sends 2930, 4100, and 5400 mice flows with 5 KB to 100 KB within 0s to 1s when the network works at 50%, 70%, and 90% load respectively. Figure 3-4 shows the packet loss and bandwidth loss of mice flows under different loads. Figure 3-5 shows link throughput curves of mice flows with different packet loss rates.

  Figure 3-4 shows that when the network works at 90% load, the packet loss rate is 0.016% and the throughput loss reaches 51.8%, that is, **a small packet loss rate seriously deteriorates the average bandwidth usage of mice flows**. Since there are a few data packets, if packet loss occurs, there is a high probability that retransmission cannot be triggered due to the lack of three redundant ACKs. As a result, RTO occurs and the link is idle, deteriorating the average bandwidth usage seriously.

Figure 3-4 Packet loss rate and throughput loss of mice flows under different loads

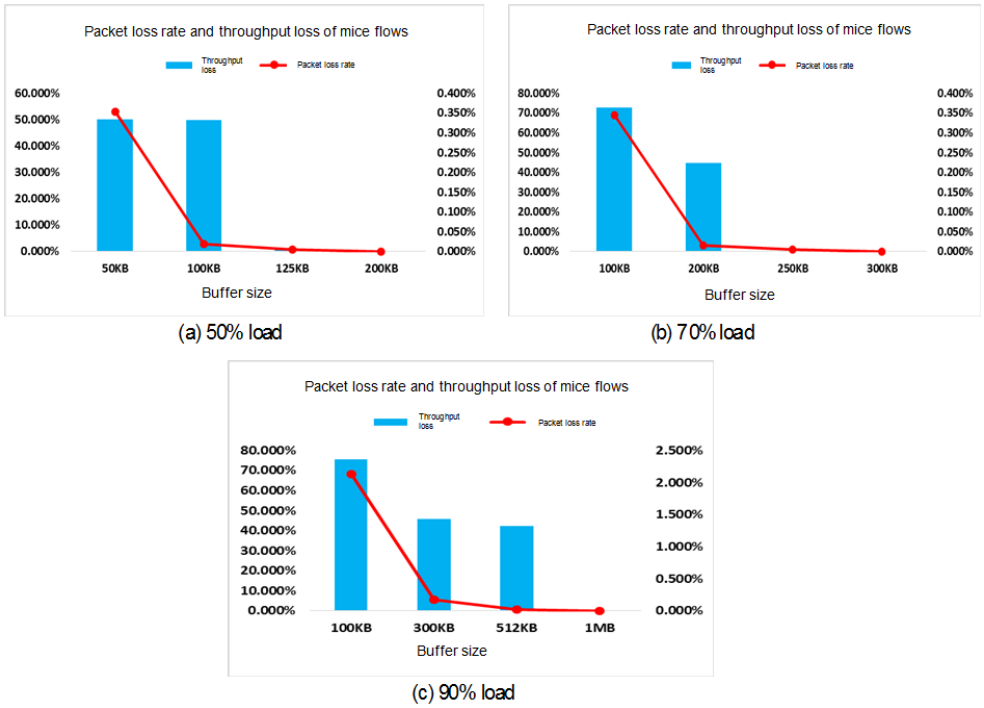


Figure 3-5 Link throughput of mice flows with different packet loss rates

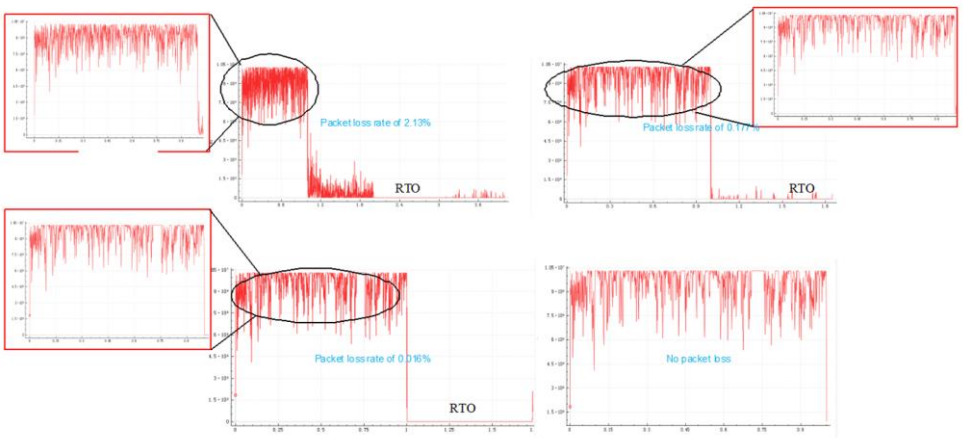


Figure 3-5 shows that the bandwidth loss is little within 1s. After 1s, the bandwidth usage decreases significantly. Since the CWND of each flow is very small, even if the packet loss or RTO occurs, the bandwidth can be quickly occupied by new flows within 1s, and the bandwidth loss is little. After 1s, packets are discarded or retransmitted. Since no data is to be sent, the link is idle, resulting in low bandwidth usage. Therefore, **when the link is in the non-idle period, packet loss has little impact on the bandwidth usage of mice flows. Packet loss reduces the average bandwidth usage of mice flows, which is caused by the unoccupied bandwidth due to RTO**.

3.1.2 Impact of Packet Loss on FCT

Flow Completion Time (FCT) refers to the duration from the time when the first data packet is sent to the time when the last data packet is received by the receiver. From the perspective of user experience, FCT is the time required for downloading a video, time spent on opening a web page, and time for submitting a purchase order to the server. In recent years, FCT or average FCT has attracted more and more attentions and becomes the mainstream optimization target of the network.

Based on the simulation experiment in section 3.1.1, we obtain the relationship between the packet loss rate and average FCT of elephant and mice flows under different loads, which are shown in Figure 3-6 and Figure 3-7.

Figure 3-6 Relationship between the packet loss rate and average FCT of elephant flows under different loads

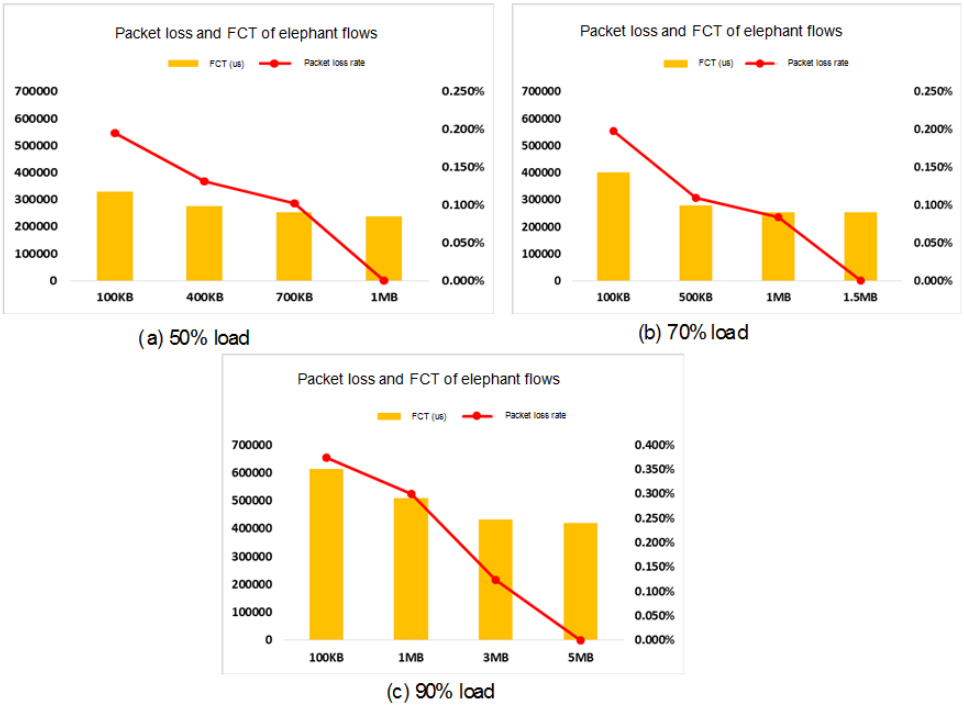
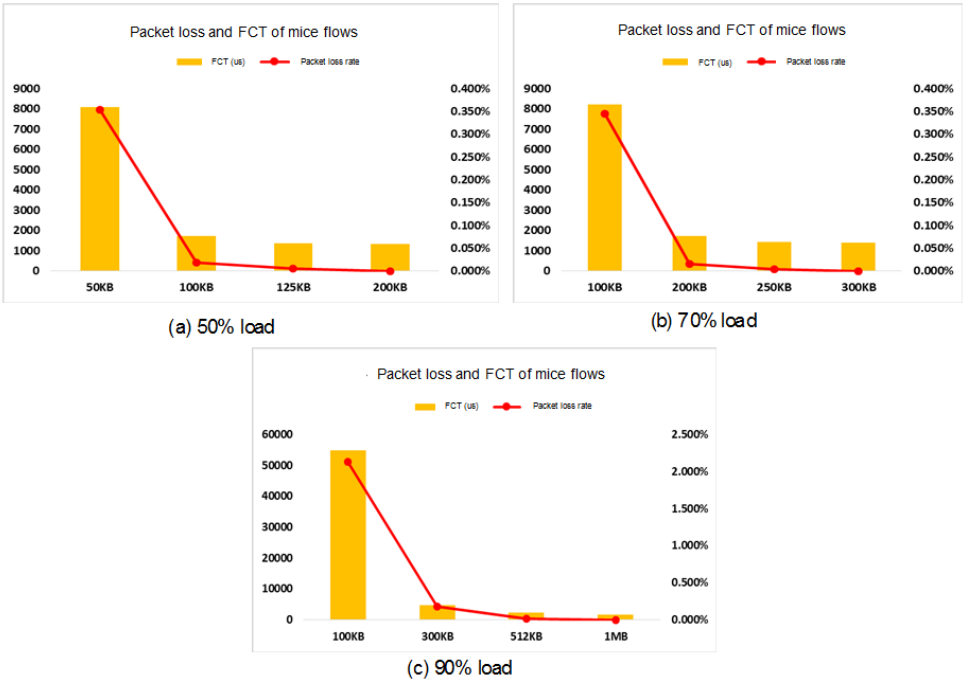


Figure 3-7 Relationship between the packet loss rate and average FCT of mice flows under different loads



The preceding figures show that **packet loss has a larger impact on the average FCT of mice flows than that on the average FCT of elephant flows**. Since the RTO is prone to occur due to packet loss of mice flows, the FCT of a single flow deteriorates from us to 200 ms, greatly affecting the average FCT. Elephant flows are sensitive to the throughput, so the throughput loss caused by packet loss in the simulation experiment does not exceed 50%. Therefore, the growth of the FCT does not increase to 2 times.

When mixed flows are transmitted, reducing the packet loss rate of a mice flow can reduce the RTO and significantly increase the average FCT. However, reducing the packet loss rate of an elephant flow can also increase the average FCT, but the increase is not larger than that of the mice flow.

3.2 Functions of Large Buffer

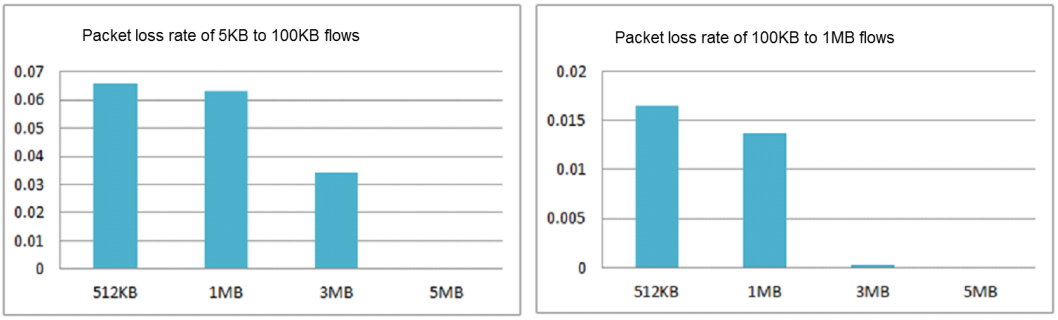
As shown in section 3.1, packet loss reduces the bandwidth usage and deteriorates the FCT. Therefore, a direct way to optimize the DCN performance is to use large buffer switches. The large buffer has the following functions:

3.2.1 Absorbing Burst Traffic, Reducing Packet Loss, and Guarantee the Throughput

The initial design of the buffer is to absorb burst traffic and guarantee the throughput. In the DC, a large amount of many-to-one traffic exists. **The large buffer can effectively absorb burst traffic and reduce packet loss and retransmission to guarantee the average link throughput**.

We simulate the topology shown in Figure 3-1 in the incast traffic model. Each host sends 60 flows and a total of 240 flows are transmitted on the network. In the traffic model, elephant and mice flows with the proportion of 2 to 8 are transmitted to obtain the packet loss rate under different buffer sizes in the following figure. Figure 3-8 shows that **larger buffer results in lower packet loss rate**.

Figure 3-8 Packet loss rates of elephant and mice flows under different buffer sizes



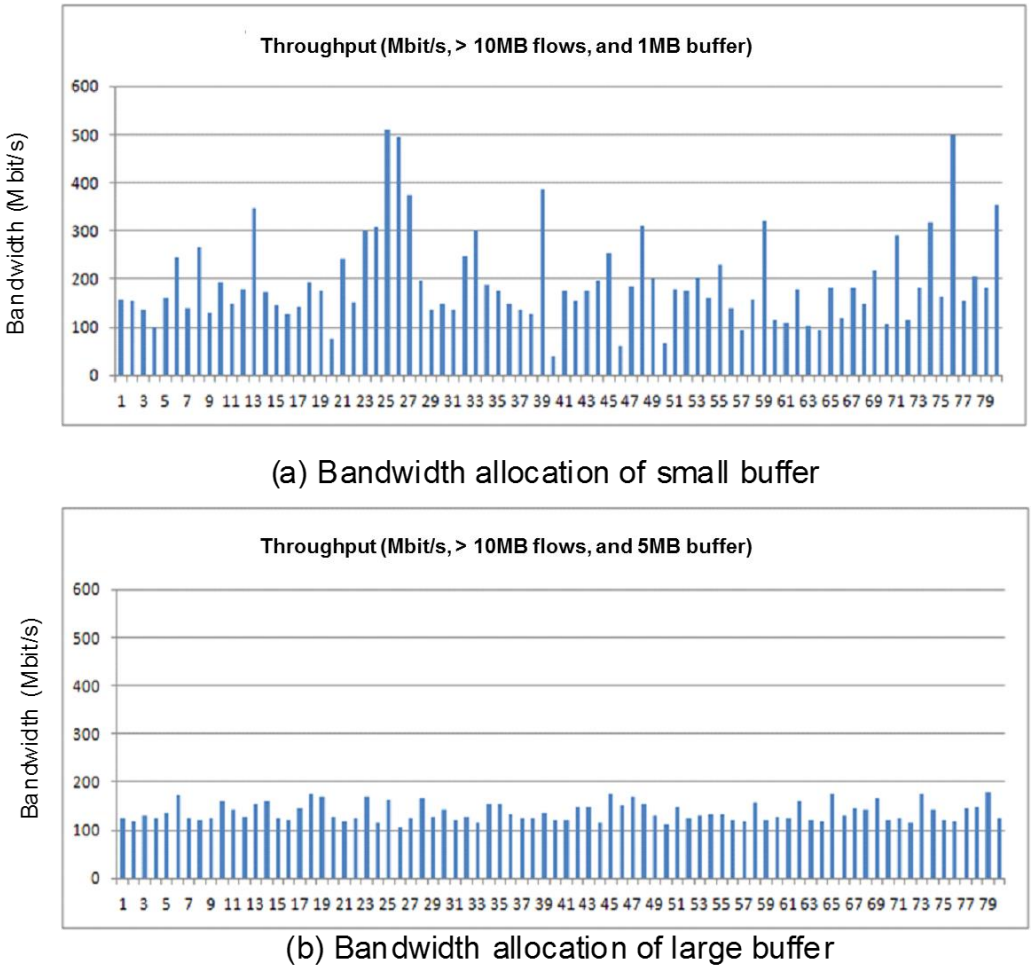
(a) Packet loss rate of mice flows under different buffer sizes (b) Packet loss rate of elephant flows under different buffer sizes

3.2.2 Allocating Bandwidth Evenly

After researching the fairness in bandwidth allocation of multiple flows in large and small buffers, we conclude that **bandwidth is allocated more evenly in large buffer**. The reason is that the small buffer causes unfairness in packet loss. Some unlucky packets may be lost and do not obtain the bandwidth, while the sending rate of some lucky packets increases steadily and the packets obtain high bandwidth, resulting in uneven bandwidth allocation for multiple flows.

We construct a similar scenario to simulate bandwidth allocation of multiple flows in large and small buffers, as shown in Figure 3-9. The result shows that **the bandwidth allocation in large buffer is fairer than that in small buffer**. Since a small number of data packets can be stored in small buffer, data packets of different flows are stored according to the sequence in which packets are received. As a result, the flow that occupies the buffer early has higher bandwidth, and packets received later are discarded, resulting in unfairness in packet loss and rate reduction. A large number of data packets can be stored in large buffer. When the buffer overflows, the number of data packets discarded in each flow is positively related to the CWND. As a result, more packets are discarded and the sending rate reduces when higher bandwidth is occupied. In this way, the sending rate of packets is fairer, which results in fairness in bandwidth allocation.

Figure 3-9 Bandwidth allocation of multiple flows

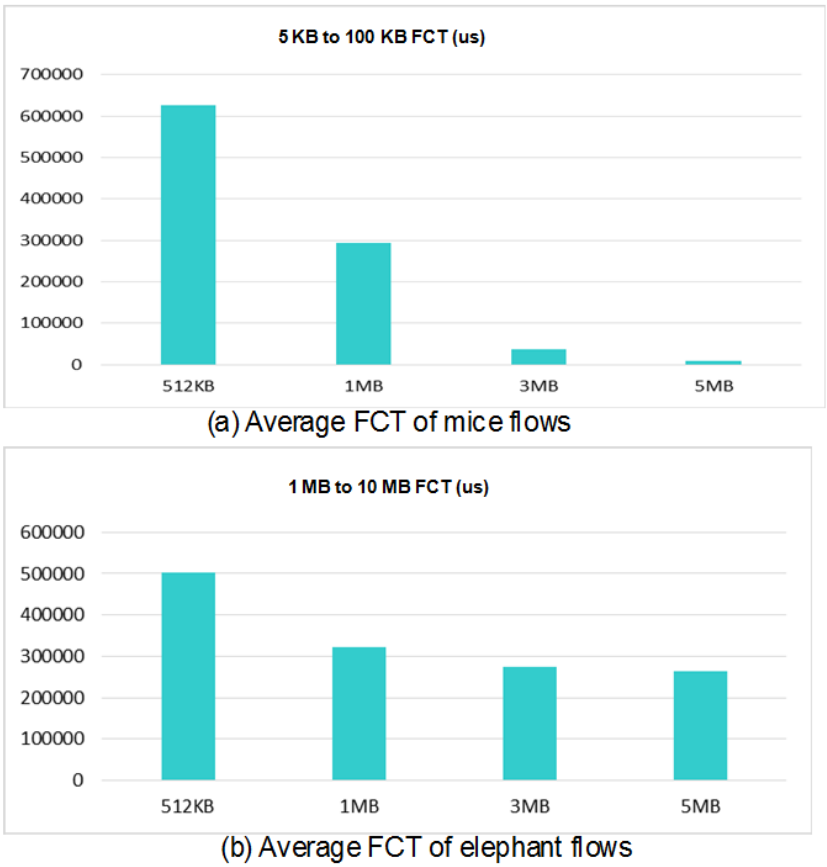


3.2.3 Optimizing FCT

After researching the change of average FCTs of large and small buffers under different loads, we conclude that the average FCT increases significantly with the increasing load in small buffer. If the load is heavier, the FCT is better optimized in large buffer. When the network works at 95% load, the FCT in large buffer is optimized by 50 times.

To further analyze the difference of FCT optimization between elephant and mice flows in large buffer, we simulate and test FCTs of elephant and mice flows in different buffers respectively. Figure 3-10 shows the result that the **FCT of a mice flow in large buffer is improved by 60 times, and the FCT of an elephant flow is improved by 2 times**. Since mice flows are sensitive to packet loss, packet loss can be reduced in large buffer to avoid retransmission caused by RTO. As a result, FCTs of mice flows decrease from 200 ms to us, greatly improving FCTs. Elephant flows are sensitive to throughput, so lost packets decrease and throughput is guaranteed in large buffer, improving FCTs of elephant flows. However, the improvement of throughput for elephant flows in large buffer is limited, FCTs of elephant flows are improved slightly.

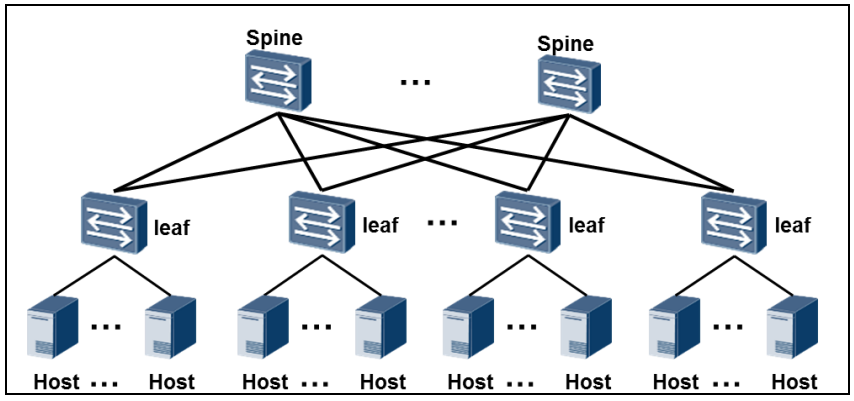
Figure 3-10 FCTs of elephant and mice flows in large and small buffers



3.3 Larger Buffer Required in the DC

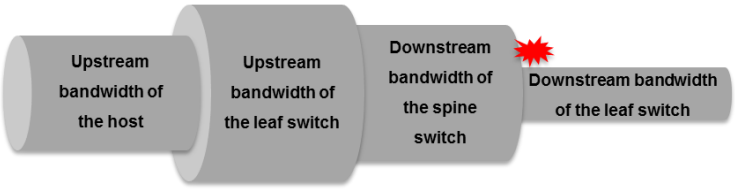
The large buffer is simple and effective for improving the DCN performance. In the DC, the Clos architecture is used as an example (as shown in Figure 3-11). Which one of the following spine and leaf switches requires large buffer?

Figure 3-11 Clos architecture



Assuming that the DC network is the Clos architecture, the bandwidth pipe model is used to describe the DCN congestion when inter-ToR traffic is transmitted in a DC. If many-to-one traffic is transmitted in the DC, the number of destination hosts is smaller than that of source hosts, and the DC oversubscription is less than 1, the bandwidth pipe model in this scenario is shown in Figure 3-12.

Figure 3-12 Bandwidth pipe model of inter-ToR traffic

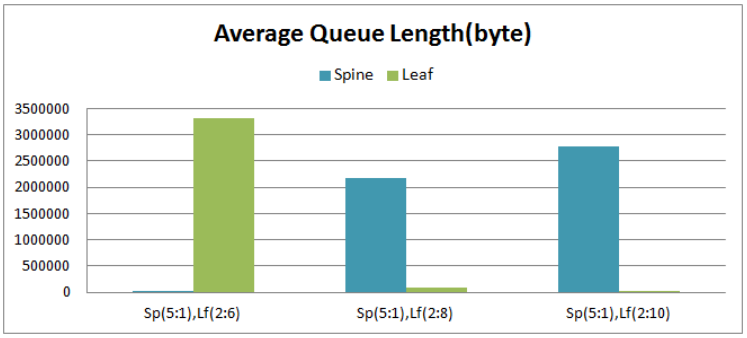


As shown in the preceding figure, the downstream bandwidth of the spine and leaf switches is two bandwidth bottlenecks. Based on the TCP, network congestion firstly occurs on the downstream bandwidth of a leaf switch that has the smallest bandwidth, and the leaf switch becomes the congestion point. In this case, the spine switch does not become the congestion point. If the traffic model of a service changes, the number of destination hosts increases, and the total downstream bandwidth of the leaf switch is greater than that of the spine switch. In this case, the downstream bandwidth of the spine switch becomes the bandwidth bottleneck and the spine switch becomes the congestion point. For intra-ToR traffic, the leaf switch must be the congestion point. Therefore, **either the downstream bandwidth of the spine switch or the downstream bandwidth of the leaf switch becomes the bandwidth bottleneck and congestion point, and large buffer is required on the congestion point**.

The congestion point in the DC is determined by the network topology and service traffic model. The network topology determines the number of spine switches, downstream bandwidth of spine switches, and downstream bandwidth of leaf switches. The traffic model determines the number of source and destination hosts and whether the intra- or inter-ToR traffic is transmitted. If the intra-ToR traffic is transmitted, a leaf switch is the congestion point that requires large buffer. If the inter-ToR traffic is transmitted, and the total number calculated by the number of spine switches x downstream bandwidth of a spine switch (total downstream bandwidth of spine switches) is greater than the total number calculated by the number of destination hosts x downstream bandwidth of a leaf switch (total downstream bandwidth of leaf switches), the downstream bandwidth of the leaf switch becomes the bandwidth bottleneck and the leaf switch becomes the congestion point that requires large buffer. Otherwise, the spine switch becomes the congestion point that requires large buffer.

In a scenario where two spine switches and six leaf switches are deployed, the downstream bandwidth of leaf switches is 10GE and the downstream bandwidth of spine switches is 40 GE. We simulate and test the congestion of spine and leaf switches on the network when different hosts are configured as destination hosts that are connected to a leaf switch. In Figure 3-13, Sp (5:1) indicates that all hosts connected to the first five leaf switches are source hosts, and Lf (2:x) indicates that **n** hosts connected to the sixth leaf switch are destination hosts. That is, the total downstream bandwidth of leaf switches is n x 10GE.

Figure 3-13 Average queue length of spine and leaf switches in different scenarios



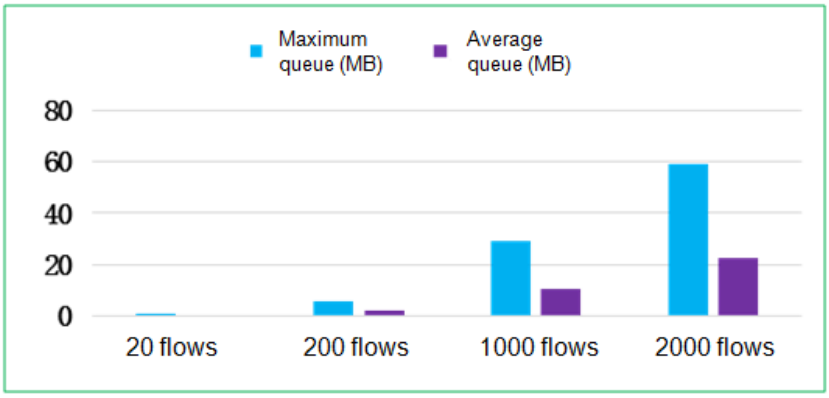
As shown in the preceding figure, when the number of destination hosts connected to a leaf switch is smaller than 8, that is, the total downstream bandwidth of leaf switches is smaller than the total downstream bandwidth of spine switches (80GE), the queue length of spine switches is almost 0 and the buffer occupancy is little, while the queue length of leaf switches is long and the buffer occupancy is high. When the number of destination hosts connected to a leaf switch is equal to or larger than 8, that is, the total downstream bandwidth of leaf switches is smaller than the total downstream bandwidth of spine switches (80GE), the queue length of spine switches is long and the buffer occupancy is high, while the queue length of leaf switches is almost 0 and the buffer occupancy is little. Therefore, **in a DC, if the intra-ToR traffic is mainly transmitted, the leaf switch is the major congestion point that requires large buffer. If the inter-ToR traffic is mainly transmitted, the downstream bandwidth of the leaf switch in the traffic model is the bandwidth bottleneck. For the network that the oversubscription is greater than 1 or has small number of destination hosts, the leaf switch is the major congestion point that requires large buffer. Otherwise, the spine switch is the major congestion point that requires large buffer**.

3.4 Buffer Size

To check whether the buffer size required in the DC is sufficient, we need to clarify the criteria for the buffer. If the buffer is set to ensure that the link is underflow, 1BDP is enough according to the conclusion in chapter 3. Considering that the main objective of DCN optimization is FCT, FCT optimization is considered as the criterion to determine whether the buffer is enough.

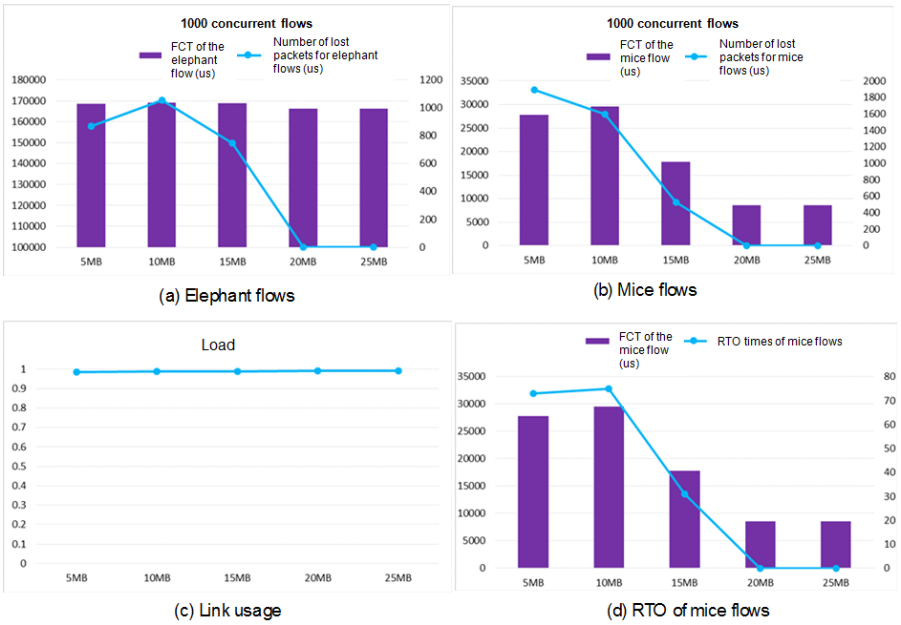
Based on the topology shown in Figure 3-14, we simulate the buffer occupancy under different concurrent flows. Figure 3-15 shows the result that more concurrent flows result in longer queue length, requiring larger buffer to reduce packet loss and optimize the FCT.

Figure 3-14 Maximum and average queue lengths of buffer under different concurrent flows



The industry points out that the maximum number of concurrent flows of a single host in a DC is 100 to 1000. So in this example, we use 1000 concurrent flows (proportion of the number of elephant flows to mice flows is 2:8) to simulate different buffer sizes and collect statistics about FCTs of elephant and mice flows, number of lost packets, link usage, and number of RTO times of mice flows. Figure 3-15 shows the result.

Figure 3-15 Performance of 1000 concurrent flows in different buffers

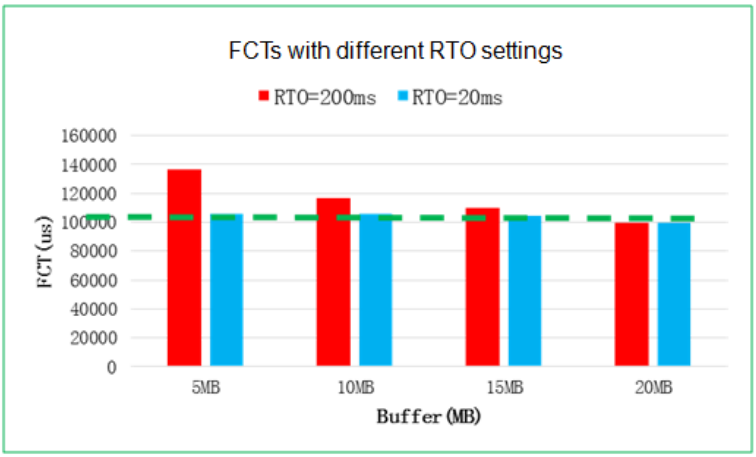


As shown in the preceding figure, when the buffer is larger than 1BDP, the link bandwidth can be fully occupied and the link is underflow. In this case, packet loss has little impact on the FCT of an elephant flow, and increasing the buffer size can slightly improve the FCT. The FCT of a mice flow is affected by the number of RTO times. Increasing the buffer size can significantly reduce the packet loss rate for mice flows and the number of RTO times. Therefore, the FCT of the mice flow is optimized. When there is no RTO for the mice flow, increasing the buffer size cannot optimize the FCT. In this scenario, the buffer of 20 MB for each 10GE port can meet the requirement of FCT optimization. **To determine whether the**

buffer is enough, we need to ensure that the link is underflow and there is no RTO for mice flows.

To further study the impact of RTO on FCT, we simulate and compare the network performance between the scenario when the RTO is 200 ms and the scenario when the RTO is 20 ms. Figure 3-16 shows the result.

Figure 3-16 FCTs with different RTO settings



As shown in the preceding figure, after the RTO is reduced from 200 ms to 20 ms, the buffering requirement is reduced from 20 MB to 5 MB. Since the RTO is reduced, the sender that sends mice flows can quickly detect the RTO and retransmit the lost data packets, so the FCT is optimized. That is, if a small RTO is set, the impact of RTO on the FCT is reduced and the performance does not deteriorate seriously. As a result, the buffering requirement is reduced.

In conclusion, the buffer size required by DC switches depends on the number of concurrent flows. The larger number of concurrent flows requires the larger buffer. The buffer should be large enough to ensure that the link is underflow and there is almost no RTO for mice flows. If there are 1000 concurrent flows per 10GE port, 20 MB buffer per port is enough.

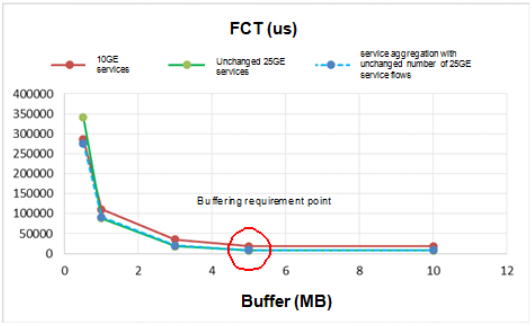
3.5 Changes of Buffering Requirement After Bandwidth Upgrade

The sharp increase of network traffic has higher requirements for the link bandwidth. Currently, many enterprises have started to upgrade the bandwidth and deploy 25GE, 40GE or even 100GE networks. Does the buffering requirement change after bandwidth upgrade?

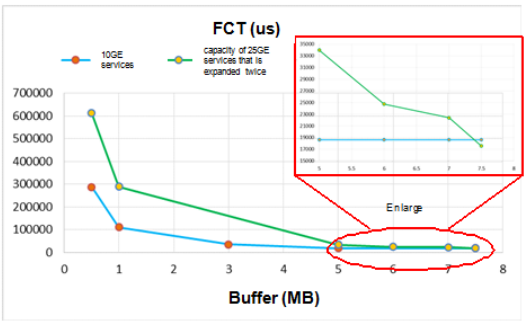
We use the topology of Figure 3-1 and compare the performance of networks using10GE and 25GE links. Assuming that each host sends 60 flows and a total of 240 flows are transmitted on the network using 10GE links, services may change after bandwidth upgrade. We simulate four situations based on the assumption: 10GE services, unchanged 25GE services (number of flows remains unchanged), service aggregation with unchanged number of 25GE service flows (total number of flows remains unchanged and the number of source hosts is reduced to 2), and capacity of 25GE services that is expanded twice (total number of flows increases to

120 flows per host). The following figure shows the network performance in different scenarios.

The result shows that when the link is upgraded from 10GE to 25GE, the buffering requirement of ensuring the FCT remains unchanged after services are not expanded, while the buffering requirement increases by 50% after services are expanded by twice. Therefore, **after bandwidth upgrade, the buffering requirement remains unchanged if services are not expanded. If services are expanded, the buffering requirement increases proportionally**.



(a) Non-increasing services



(b) Service capacity extension

3.6 Conclusion

This chapter describes the core problems about buffering requirements of DC switches in the TCP congestion control scenario based on tail drop.

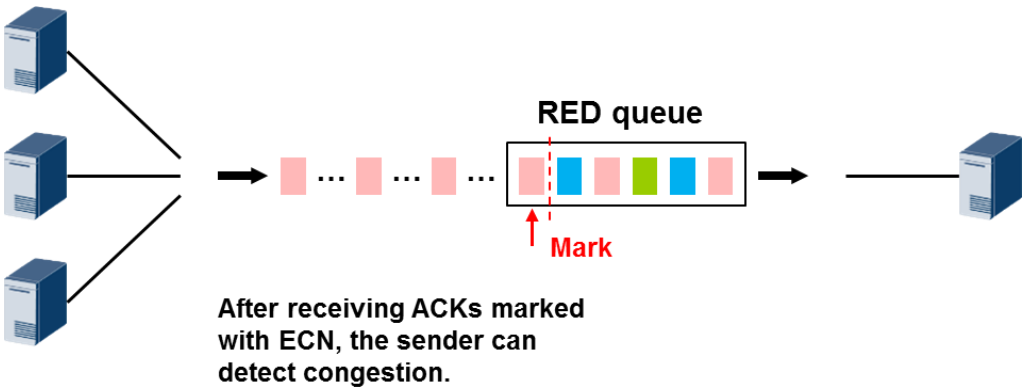
1. If packet loss occurs in mice flows, RTO is likely to occur. As a result, the bandwidth usage and FCT of mice flows severely deteriorate. Packet loss has little impact on throughput for elephant flows, so FCTs of elephant flows are insensitive to packet loss.
2. Large buffer is used to absorb burst traffic, reduce packet loss, and guarantee the throughput. In this way, the bandwidth is allocated evenly and the FCT is optimized.
3. The bandwidth bottleneck and large buffer required in the DC are clarified.
4. To determine whether the buffer is enough, we need to ensure that the link is underflow and there is no RTO for mice flows. In a scenario of 1000 concurrent flows, each 10GE port requires the buffer of about 20 MB.
5. The buffering requirement increases proportionally with services after bandwidth upgrade.

4 Buffering Requirement Based on the ECN

4.1 Functions of ECN

ECN is TCP/IP extension in RFC3168. Similar to packet loss, ECN is a way to feed back congestion information. The difference from congestion feedback for packet loss is that data packets that exceed the ECN threshold are tagged with the ECN mark, which is shown in Figure 4-1

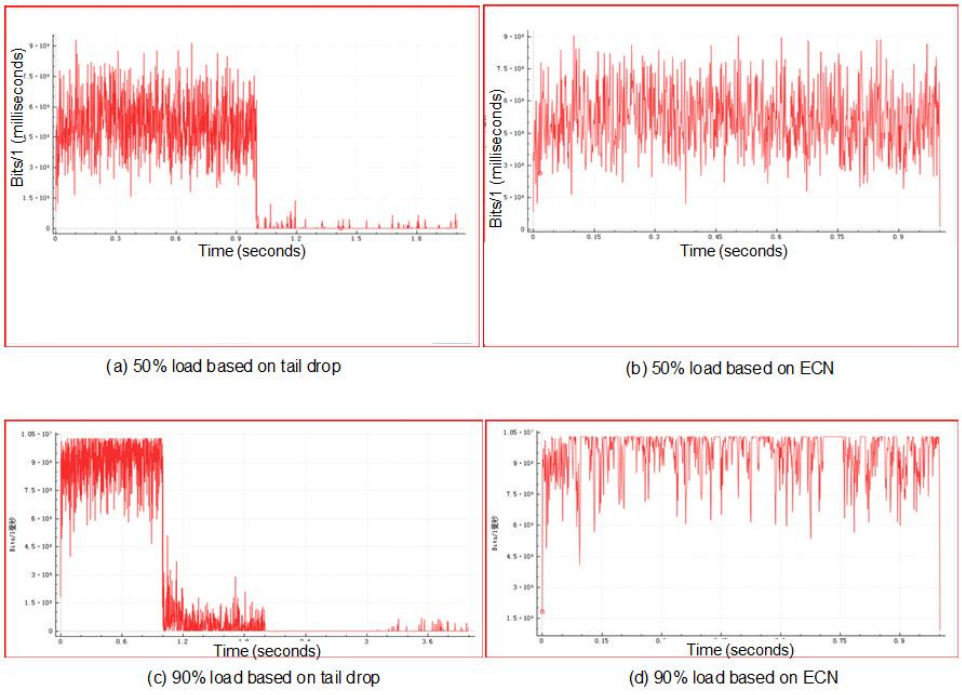
Figure 4-1 ECN mark



The following describes ECN functions:

1. ECN reduces or even avoids packet loss, thereby reducing the RTO of mice flows. This is the essence of ECN. Compared with congestion feedback due to buffer overflow and packet loss, ECN enables the sender to detect traffic congestion and decrease the sending rate as early as possible, in order to avoid the occurrence of RTO. For the buffer of the same device, we simulate that the network works at 50% and 90% loads, and compare the TCP based on tail drop and TCP based on ECN. Figure 4-2 shows the result.

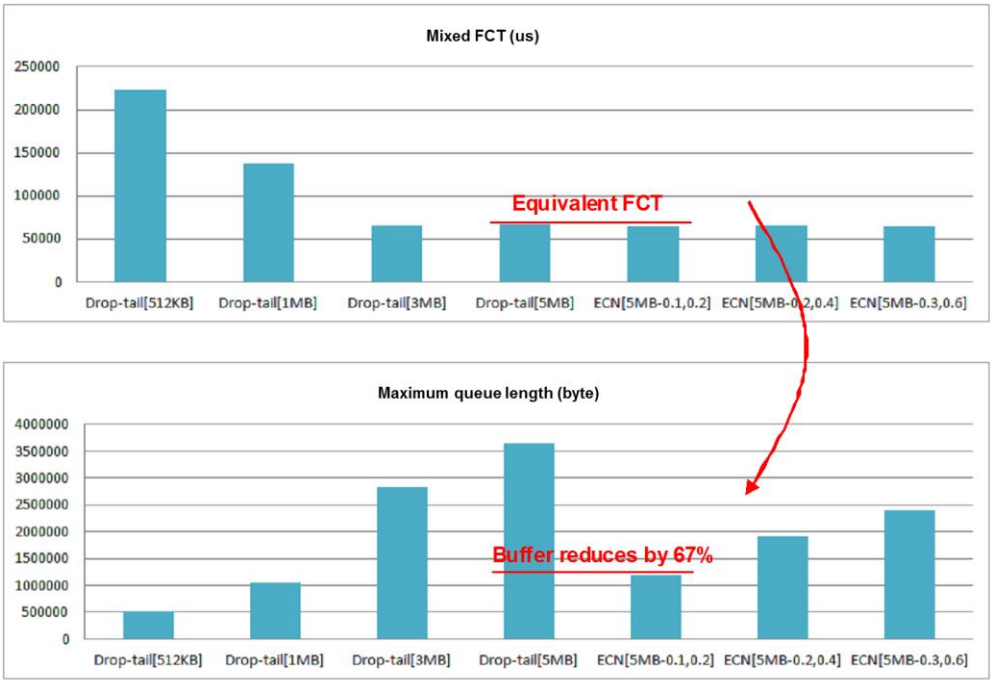
Figure 4-2 ECN reduces RTO



As shown in the preceding figure, under the same load, RTO is more likely to occur on the TCP based on tail drop, resulting that the link is idle. However, there is no RTO on the TCP based on ECN.

2. ECN reduces the buffering requirement. From the perspective of FCT optimization, ECN enables the sender to detect traffic congestion and decrease the sending rate as early as possible, in order to avoid the occurrence of RTO and guarantee the FCT of mice flows. It is different from the TCP based on tail drop. In this case, the large buffer is used to avoid the RTO of mice flows, reducing the buffering requirement. We compare the network performance in the scenario where the TCP network based on tail drop is implemented in different buffers and the scenario where the TCP network based on ECNs is implemented with different ECNs. Figure 4-3 shows the result.

Figure 4-3 ECN reduces the buffering requirement



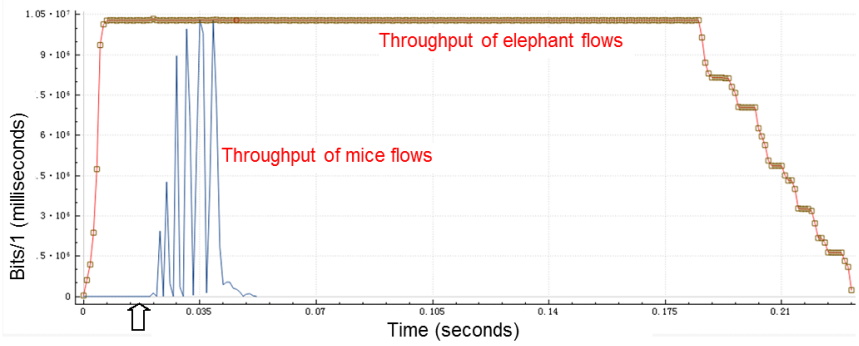
As shown in the preceding figure, to achieve the same FCT of large buffer based on tail drop, the maximum queue length is reduced by 67%. This indicates that the buffer required by the switch can be reduced by 67% after ECN is enabled. In addition, the ECN threshold continues to increase after the link is underflow, which will deteriorate the FCT. This is because the increasing threshold increases the latency of mice flows.

4.2 ECN Threshold Setting

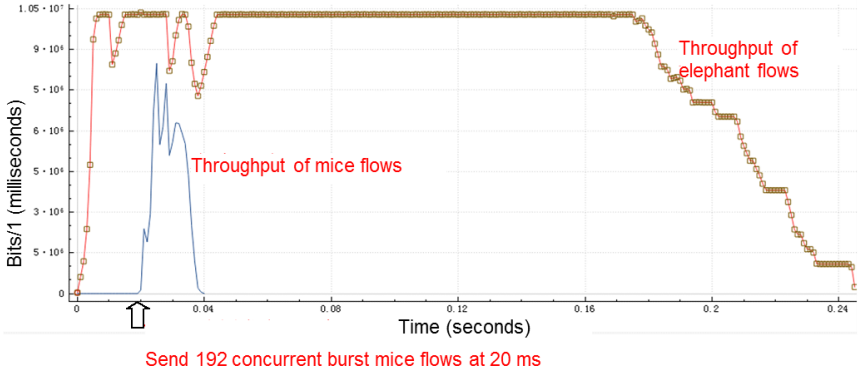
Before setting the ECN threshold, we need to clarify the criteria for setting the appropriate ECN threshold. From the perspective of network performance, the setting of the ECN threshold must ensure the DCN performance. (1) The link is underflow. (2) Low latency is ensured. (3) Priority-based Flow Control (PFC) is triggered as less as possible or not triggered. From the perspective of applications, ECN threshold is set to optimize the FCT as much as possible.

On a device without differentiation scheduling of elephant and mice flows, there is a conflict between targets. The settings of the ECN threshold need to be different according to different service scenarios. The following uses the scenario where burst traffic is transmitted in a TOR switch as an example to simulate and test the criteria for ECN threshold setting. We use the topology in Figure 3-1 and perform 48 elephant flows under 100% load. At 20 ms, we send 192 concurrent mice flows simultaneously and test network performance when the ECN threshold is 1BDP and 15%BDP (about BDP/$\sqrt{48}$). Figure 4-4 shows the throughput of these two scenarios, and Figure 4-5 and Figure 4-6 show the performance data.

Figure 4-4 Throughput under different ECN thresholds



(a) ECN threshold = 1BDP



(b) ECN threshold = 15%BDP

Figure 4-5 FCTs under different ECN thresholds

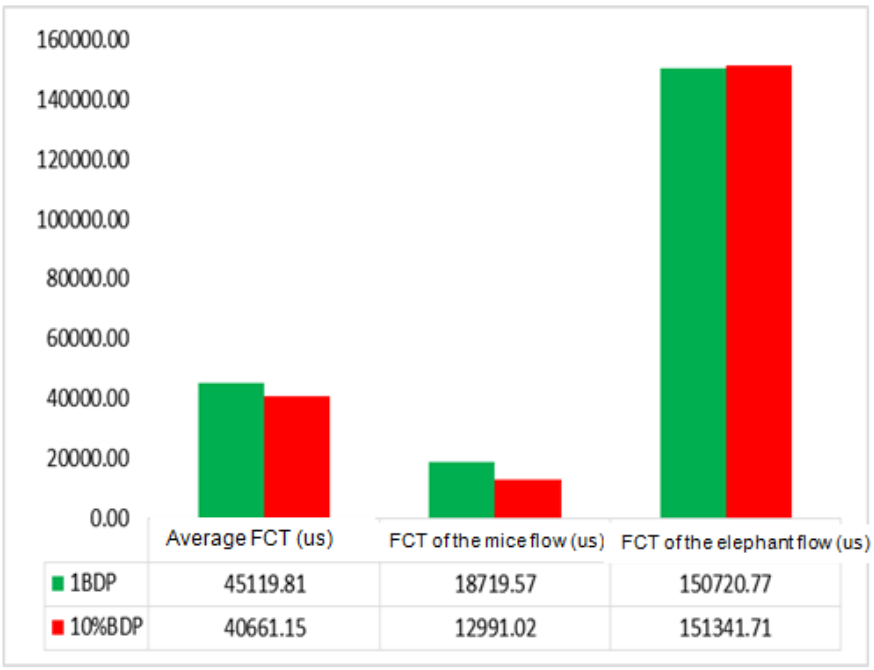
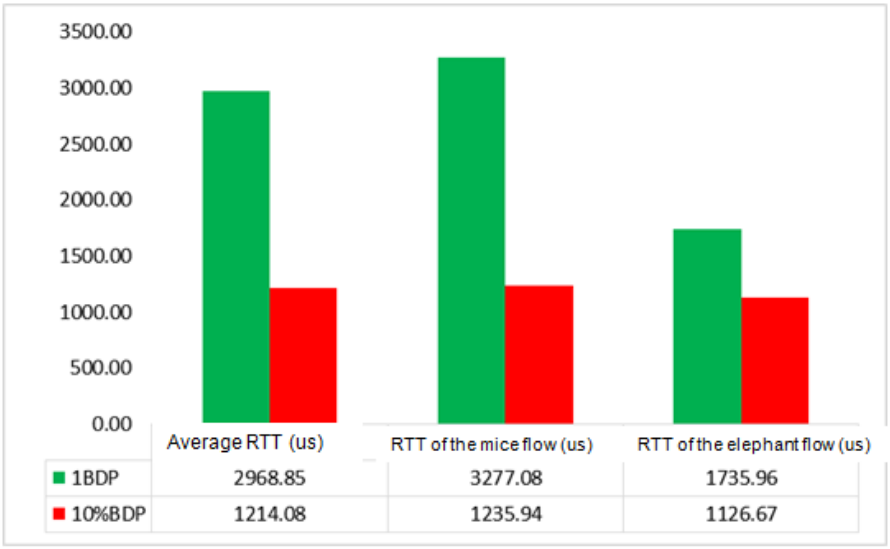


Figure 4-6 RTTs under different ECN thresholds



As shown in Figure 4-4, in the incast burst traffic scenario, when the ECN threshold is set to 1BDP, the link bandwidth is fully occupied. When the ECN threshold is set to 1BDP/ $\sqrt{n}$, the link is overflow. Therefore, **to ensure that the link is underflow, the ECN threshold should be set to 1BDP**. As shown in Figure 4-5 and Figure 4-6, after comparison between the lower threshold of 15%BDP and upper threshold of 1BDP, RTTs and average FCT of multiple flows are higher, and FCTs of mice flows deteriorate. When no packet loss occurs, although the link bandwidth is underflow, the queuing latency is long when the threshold is high, and the FCT

for latency-sensitive mice flows deteriorates severely. On the other hand, the link is overflow due to low threshold. However, micro-burst for mice flows does not last for a long time, and the FCT of elephant flows when the link is underflow deteriorates slightly. Therefore, From the perspective of FCT optimization, the threshold should be set to 15%BDP.

In conclusion, **the ECN threshold needs to be adjusted dynamically according to services**. **When mice flows account for a large proportion of services, we need to focus on maintaining the low network latency and set a lower threshold. When elephant flows account for a large proportion of services, we need to focus on maintaining the throughput of the link and set a higher threshold (1BDP)**.

4.3 Buffer Size Based on ECN

The following uses a single TCP flow as an example to describe the maximum queue length of switches enabled with ECN. In the TCP slow-start phase, TCP overshoot causes the longest queue length. The following analysis is based on TCP slow start.

Assuming that the TCP initially starts with the window as 2 packets and the RTT is **T**. The latency for the switch to forward a data packet is δ. At 0, data packets 1 and 2 are sent, and the maximum queue length is 2. After 1RTT, the sender receives the ACK of packet 1 and the window size increases by 1. When the window size is 3, the sender sends data packets 3 and 4. After δ, the sender receives the ACK of packet 2 and the window size increases by 2. When the window size is 4, the sender sends data packets 5 and 6, and the maximum queue length increases by 1. As shown in Figure 4-7 and Figure 4-8, when the window size increases by 1, the maximum queue length increases by 1.

At some time, when the window size is $\mathbf{W_e}$ and the queue length is equal to the ECN threshold **h**, the packet is marked with ECN. After 1RTT, the sender receives the ACK marked with ECN. During the RTT, the sender continues to receive ACKs without ECN marked and the window size increases to $2w_e$. According to the rule that the queue length increases by 1 when the window size increases by 1, when the sender receives the ACK with ECN marked, the queue length increases to $W_e + h$. Then the sender responds to congestion, the window size decreases by half and the queue length is shorter

Figure 4-7 Single-flow window size and queue length

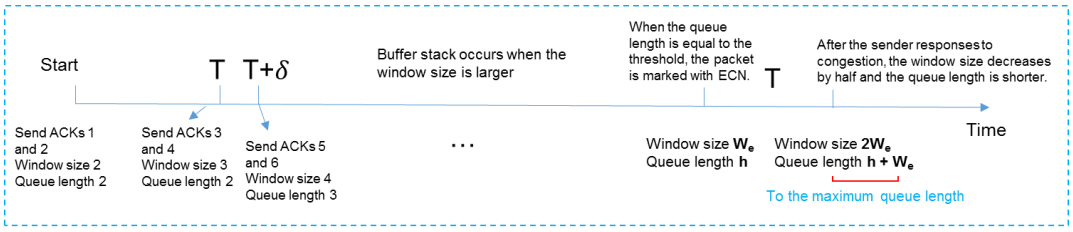


Figure 4-8 Single-flow window size and queue length

Time	Packets Acked	Window size	Packets sent	Queue length
0		2	1,2	2
T	1	3	3,4	2
$T+\delta$	2	4	5,6	3=2-1+2
$2T$	3	5	7,8	2
$2T+\delta$	4	6	9,10	3=2-1+2
$2T+2\delta$	5	7	11,12	4
$2T+3\delta$	6	8	13,14	5
$3T$	7	9	15,16	2
$3T+\delta$	8	10	17,18	3=2-1+2
$3T+2\delta$	9	11	19,20	4=3-1+2
$3T+3\delta$	10	12	21,22	5=4-1+2
$3T+4\delta$	11	13	23,24	6
$3T+5\delta$	12	14	25,26	7
$3T+6\delta$	13	15	27,28	8
$3T+7\delta$	14	16	29,30	9
		...		
		w_e		h
		w_e+1		$h+1$
		...		
		$2w_e = w_e + w_e$		$h+w_e$

Data source: Wu H, Ju J, Lu G, et al. Tuning ECN for data center networks[C]//Proceedings of the 8th international conference on Emerging networking experiments and technologies. ACM, 2012: 25-36.

It is noted that when the queue length reaches the threshold **h**, the value of window size $\mathbf{w_e}$ refers to inflight. Since the inflight data is transmitted in the link channel or the buffer queue, the channel capacity is BDP, we have $w_e \leq h + BDP$

To prevent packet loss, the maximum queue length $\mathbf{w_e + h}$ should not exceed the buffer size. Therefore, we have

$w_e + h \leq$ Buffer size

To ensure that the link is underflow and low queuing latency, the ECN threshold **h** is set to 1BDP. So the buffer size can be calculated as

Buffer size $\geq$ 3BDP

Therefore, **3BDP is required for the buffer based on ECN and the threshold is set to 1BDP**.

5 Buffering Requirement of Differentiated Scheduling of Elephant and Mice Flows

The intelligent buffer management based on differentiated scheduling of elephant and mice flows enables small buffer to achieve performance of large buffer or even better performance than large buffer.

5.1 Differentiated Scheduling of Elephant and Mice Flows

Differentiated scheduling of elephant and mice flows is a multi-queue QoS mechanism with the core of configuring two queues for the same type of flows. A high-priority queue is configured for the mice flow to enable preferential forwarding, and a normal queue is configured for the elephant flow. Based on the statistics about elephant and mice flows, the first N packets of each flow enter the high-priority queue, and the N+1 packets enter the normal queue. In this way, low latency is guaranteed for mice flows.

5.2 Achieving Performance of Large Buffer or Even Better Performance Based on Differentiated Scheduling of Elephant and Mice Flows

The performance of small buffer switches based on differentiated scheduling of elephant and mice flows is similar to or even better than that of large buffer switches.

- For elephant flows, the performance of large buffer switches is similar to that of small buffer switches based on differentiated scheduling of elephant and mice flows.
- The performance of mice flows is better. The result shows that the FCT of mice flows deteriorates with the increasing load in large buffer switches, while the FCT of mice flows remains unchanged in small buffer switches based on differentiated scheduling of elephant and mice flows.
- The FCT is optimized. The large buffer ensures that there is no RTO for mice flows. The FCT is optimized because elephant flows are overflow. However, large buffer causes long buffer length and high latency. High-priority queue is configured for mice flows and forwarding is performed to ensure that data packets are not lost and the latency is low, in order to optimize FCTs of mice flows.

In conclusion, small buffer switches based on the differentiated scheduling of elephant and mice flows have the same functions as large buffer, in order to achieve the performance similar to large buffer or even better performance than that of the large buffer.

5.3 Buffer Size Required Based on Differentiated Scheduling of Elephant and Mice Flows

The core advantage of small buffer switches based on differentiated scheduling of elephant and mice flows is to distinguish between elephant and mice flows, configure high-priority queues for mice flows, and forward packets preferentially to ensure no packet loss and low latency. To ensure the network performance, the buffer required needs to ensure that the outbound interface is underflow. Therefore, according to the conclusion in chapter 3, **at least 1BDP is required for the small buffer based on the differentiated scheduling of elephant and mice flows to ensure that the link is underflow**.

6 Conclusion

This white paper mainly analyzes buffering requirements in three mainstream DCN scenarios, including the TCP network based on tail drop, TCP network based on ECN, and TCP network enabled with differentiated scheduling based on elephant and mice flows.

On the TCP network based on tail drop, spine or leaf switches in the DC require large buffer of 20 MB for each 10 GE port. Large buffer required by spine and leaf switches is determined by the network topology and traffic model. The large buffer is used to absorb burst traffic, reduce packet loss, allocate bandwidth evenly, and optimize FCT. If the network bandwidth is upgraded, the buffering requirement increases proportionally.

On the TCP network based on ECN, 3BDP is enough for DC switches, and the ECN threshold is set to 1BDP.

On the TCP network enabled with differentiated scheduling based on elephant and mice flows, at least 1BDP is configured for DC switches to ensure that the link is underflow and the FCT is optimized.

7 Acronyms and Abbreviations

Acronym and Abbreviation	Full Name
DC	Data Center
ECN	Explicit Congestion Notification
FCT	Flow Completion Time
BDP	Bandwidth Delay Production
RTT	Round Trip Time
RTO	Retransmission Time Out
PFC	Priority-based Flow Control