TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 24/24 pp1615–1632
DOI: 10.26599/TST.2024.9010010
Volume 29, Number 5, October 2024

Resilient TCP Variant Enabling Smooth Network Updates for Software-Defined Data Center Networks

Abdul Basit Dogar, Sami Ullah, Yiran Zhang, Hisham Alasmary, Muhammad Waqas*, and Sheng Chen

Abstract: Network updates have become increasingly prevalent since the broad adoption of software-defined networks (SDNs) in data centers. Modern TCP designs, including cutting-edge TCP variants DCTCP, CUBIC, and BBR, however, are not resilient to network updates that provoke flow rerouting. In this paper, we first demonstrate that popular TCP implementations perform inadequately in the presence of frequent and inconsistent network updates, because inconsistent and frequent network updates result in out-of-order packets and packet drops induced via transitory congestion and lead to serious performance deterioration. We look into the causes and propose a network update-friendly TCP (NUFTCP), which is an extension of the DCTCP variant, as a solution. Simulations are used to assess the proposed NUFTCP. Our findings reveal that NUFTCP can more effectively manage the problems of out-of-order packets and packet drops triggered in network updates, and it outperforms DCTCP considerably.

Key words: software defined data center networks; network updates; DCTCP; out-of-order packets; packet drop; SDN

1 Introduction

Computer networks are dynamic, complex, and have diverse critical infrastructures. In order to maintain the availability, correctness, and performance of networks in an efficient manner, network operators frequently involve in various tasks of updating the routing policies, changing the security policies, recovering from link failures, migrating flows, etc. These tasks are termed as network updates[1]. In a modern data center network (DCN), network updates are becoming even more frequent as some new scenarios of network updates include: (1) virtual machines migration among physical servers, (2) reconfiguration between a load balancer and its backend servers, (3) examining the functionality and compatibility of a new switch on-boarding through moving traffic, (4) installation of a

• Abdul Basit Dogar is with Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China and also with Department of Informatics and Systems, University of Management and Technology, Lahore 54660, Pakistan. E-mail: abasitdogar@gmail.com.
• Sami Ullah is with Department of Computer Science, Shaheed Benazir Bhutto University, Sheringal, Upper Dir 18050, Pakistan. E-mail: sami@sbbu.edu.pk.
• Yiran Zhang is with School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China. E-mail: yiranzhang@bupt.edu.cn.
• Hisham Alasmary is with Department of Computer Science, College of Computer Science, King Khalid University, Abha 61421, Saudi Arabia. E-mail: alasmary@kku.edu.sa.
• Muhammad Waqas is with Department of Computer Engineering, Faculty of Information Technology, University of Bahrain, Sakheer 32038, Bahrain, and also with School of Engineering, Edith Cowan University, Perth WA 6027, Australia. E-mail: engr.waqas2079@gmail.com.
• Sheng Chen is with School of Electronics and Computer Science, University of Southampton, Southampton, SO17 1BJ, UK. E-mail: sqc@ecs.soton.ac.uk.
* To whom correspondence should be addressed.
Manuscript received: 2023-09-15; revised: 2023-12-08; accepted: 2024-01-09

new firmware version at the switch, and so on[2–4]. Network updates in software-defined DCN (SD-DCN) may have various additional reasons and circumstances, which inevitably lead to reroute the traffic[5, 6].

Rapidly expanding companies have employed the software-defined network (SDN), which provides considerable advantages over traditional DCN in managing data transfer. For example, Microsoft[7] and Google[8] interconnect their data centers with SDN to achieve high network performance. Google reports that the link utilization can attain near about 100%, whereas traditional networks can only achieve an average of 30% to 40%[8]. An SDN-enabled network often requires frequent and fast network updates to manage rapid flow rescheduling decisions and utilize a virtually centralized controller to fulfill the different requirements at the data plane. Specifically, the procedures for updating the rules are performed over non-synchronized machines, and implementing the properties of consistency with rules dependencies demands frequent and fast updates to avoid forwarding anomalies and exacerbated network performance[9–13].

During the execution of successive network updates, various processes are carried out in a non-blocking manner, thereby ensuring the consistency properties is crucial[14]. In SDN setups, frequent flow rescheduling, if not carried out carefully, may cause major network update issues referred to as network confusions[1], such as link congestion, network policy violation, forwarding blackhole, and forwarding looping. These issues lead to inconsistencies in network updates. As a result, problems such as out-of-order packets and packet drops occur. In order to preserve network consistency properties during updates, the majority of previous research has focused on link congestion[2, 7, 9, 11, 12, 15–18], forwarding blackhole and forwarding looping[2, 7, 9, 11, 19–22], and policy violations[10, 22].

In view of the aforementioned discussion on the network update scenarios and consistency properties maintenance, we notice the key observations as follows.

(1) Almost every scenario involving network updates requires rerouting of the flows or traffic, which may lead to the issues of out-of-order packets and packet drops.

(2) If the rescheduling of flows is not handled carefully, it may cause inconsistent network updates, resulting in transient congestion and rerouting. Therefore, out-of-order packets and packet drops can occur.

According to Ref. [23], observations show that transmission control protocol (TCP) traffic accounts for 99.91% of the traffic in data centers. However, the rerouting violates TCP's assumption and has a negative repercussion on TCP traffic, resulting in severe network performance degradation. When rerouting, out-of-order packets and packet drop are serious issues. The predicted difficulties of TCP can be classified as follows when the network is updated:

(1) Certain disorders may result in the suppression of window size and unusual retransmission. Many researchers attempt to address the difficulties using timer or DUPACK threshold estimate to solve the problems.

(2) TCP invariably begins with a "slow start". When rerouting a flow, a network gives the new route with a "quick start" that can result in out-of-order packets and severe congestion, particularly if the updates are frequent and inconsistent or if the flow scheduling is imperfect.

Given these issues related to TCP, the performance of TCP is inevitably poor whenever the network is being updated. As pointed out previously, network update occurs frequently. For example, Hedera[3] updates the network every 5 s, and the authors of Ref. [3] believe even sub-second and possibly sub- 100 ms networking updates are achievable, as evidenced by the work of Ref. [24], which updates the network using 1 ms, 5 ms, and 1000 ms. However, TCP performs significantly worse when updates occur frequently.

There exist many TCP variants. However, to the authors' best knowledge, so far no one has studied how TCP variants will react to the aforementioned problems during the network updates or reroutes. To further investigate the issues related to TCP variants, in this paper, we first perform extensive experiments using cutting-edge TCP variants including DCTCP[23], CUBIC[25, 26], and BBR[27] to observe their behaviors during network updates. Our results show that CUBIC, BBR and DCTCP face the serious problems during inconsistent and frequent network updates, thereby confirming that the existing TCP variants are incapable of dealing with frequent and inconsistent network updates effectively.

1.1 Motivation

Motivated by this experimental investigation of TCP variants, we propose a TCP modification based on

DCTCP, named network update friendly TCP (NUFTCP), to overcome the aforementioned issues when networks are updated in SD-DCN[26]. The key is to understand how existing TCP designs, particularly DCTCP, can be extended to handle network updates more gracefully. The fundamental challenges of smoothing network updates are handling packet reordering and avoiding packet drops, which may reduce the window size even though there is no actual congestion. By limiting window size and delaying duplicated acknowledgments (ACKs), NUFTCP can perform better than DCTCP, especially when the updates of networks are inconsistent and frequent. We investigate how the proposed NUFTCP can cope with frequent and inconsistent network updates in detail and demonstrate through simulations that our model accurately captures the essences of network updates. Our NUFTCP remains resilient at frequent and inconsistent network updates, and it outperforms DCTCP. The main contributions of this paper are summarized as follows.

(1) Extensive experiments are conducted and the results are analyzed to better understand and pinpoint "flaws" in the current TCP designs utilized in data centers when networks are updated frequently and inconsistently.

(2) We focus on the issues with out-of-order packets and dropped packets caused by network updates, which substantially impacts the real-time transmission of data.

(3) We propose a novel NUFTCP, a TCP extension of DCTCP that gracefully handles network updates.

(4) Simulations are used to evaluate NUFTCP, and the results obtained reveal that it achieves more satisfactory performance in SD-DCN during inconsistent and frequent network updates.

The remaining sections of the paper are organized as follows. In Section 2, we present extensive experiments and detailed analysis of existing TCP solutions with an example to demonstrate why gracefully handling network updates are crucial in SD-DCNs. Section 3 is devoted to our proposed NUFTCP design. We use simulation based evaluations to demonstrate that NUFTCP works well in different scenarios in Section 4. In Section 5, the relevant literatures are reviewed, and in Section 6, we conclude this paper.

2 Analysis of TCP Variants in Network Updates

The performance of TCP is investigated in the context of inconsistent and frequent network updates. Specifically, we perform a series of tests to see how different TCP variants perform. The experimental results are analyzed and explained.

2.1 Experimental setup

The topology we employed in the experiments is depicted in Fig. 1. The network comprises four Intel Xeon X5650 servers with six cores running at 2.67 GHz that function as both senders (Server 1 and Server 2) and receivers (Server 3 and Server 4). We utilize the Linux kernel version 4.9. The senders and receivers are connected to the two switches, H3C S6800 (Switch 1) and H3C S6300 (Switch 2). Both the switches implement OpenFlow 1.3 and have a 10 Gbps link speed. They are connected by two ports. As a result, there are two ways for a sender to reach a receiver. A third port connects the SDN H3C VCF Controller to the switch.

The actual path of each flow is controlled by the controller. In the experiments, two groups of flows are used, and they are $\{f_n\}$: Server 1 → Server 3 and $\{f_m\}$: Server 2 → Server 4. Switch 1 receives these flows and then forwards to Switch 2. The network is shown in its starting condition in Fig. 1. The two groups of flows are forwarded as follows in their normal state: $\{f_n\}$: Server 1 → Server 3 through link l_n (Switch 1 to Switch 2); $\{f_m\}$: Server 2 to Server 4 through link l_m (Switch 1 to Switch 2).

During the network update, routes change frequently, and the traffic on link l_n shifts to link l_m and vice versa. Then, the network state is as follows: $\{f_n\}$: Server 1 to Server 3 via link l_m (Switch 1 to Switch 2); $\{f_m\}$: Server 2 to Server 4 via link l_n (Switch 1 to Switch 2). Another scenario is when link l_n failure happens, and the network state is as follows: $\{f_n + f_m\}$: Server 1 and Server 2 to Server 3 and Server 4 via link l_m (Switch 1

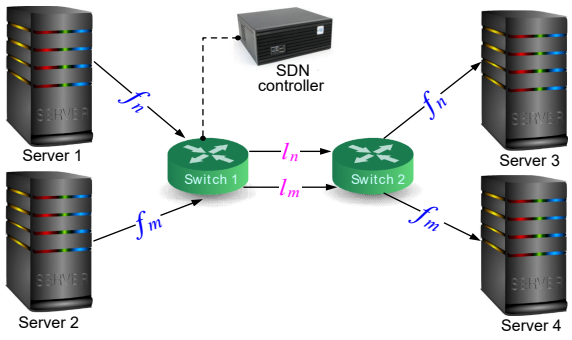


Fig. 1 Network topology for network updates with TCP variants.

to Switch 2). The scenario of link l_m failure is similar.

To investigate the behavior of TCP variants during frequent network updates, we reroute the flow traffic initiated by Server 1 towards link l_m and the flow traffic initiated by Server 2 towards link l_n. These rerouting occur each time the controller initiates a network update. During the five-minute experiment, the duration for each route change is at most one second. The other system parameters for the experiment are specified in Table 1. The experiment settings emulate unpredictable and frequent network update situations, which may result in link congestion, dropped packets, and out-of-order packets. Due to excessive retransmissions and reduced flow throughput, there will be a significant reduction in throughput.

2.2 Analysis for consistent network updates

We first investigate the performance of the TCP variants in consistent network updates. The requirement for consistency demands that the transmission of flows is seamlessly toward new paths of routing or rerouting, in order to enforce the implementation rules with their interdependence in the flow tables across multiple switches on a routing path[28]. Therefore, the most important aspect of maintaining consistency is the correct order of updates[29], and the updates follow the right sequencing are called consistent updates.

The impact of network update frequency on the achievable network performance of the TCP variants is investigated in Fig. 2. It can be seen that given the network update frequencies of ten seconds, five seconds, three seconds and one second, the corresponding network performance degradations are more than two percents, five percents, ten percents and thirty percents, respectively. In practice, frequent DCN updates often occur, initiated by operators, software or in unusual cases of failure[2]. The controller and data plane must quickly update in real-time because of the recurring flow dynamics[30]. Since tens of thousands of flows may occur in just a few milliseconds, high efficiency in network updates is crucial[7].

Table 1 Parameters of experimental system.

Parameter	Value
Packet MTU size	1500 B
Queue type	RED
Traffic flow pattern	Pre-defined
Small buffer size	200 kB
Large buffer size	1.0 MB
Link capacity	10 Gbps
Measurement time	120 s – 300 s
ACK delay threshold	2 packets
ACK delay timeout	200 ms
Reroute time	1 s

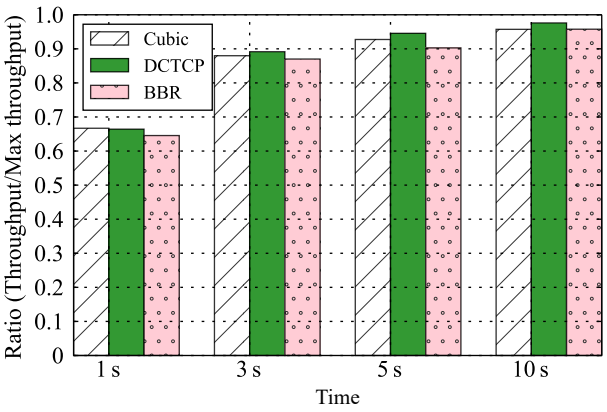


Fig. 2 Effect of network update frequency on performance of TCP variants. Network updates are consistent.

Below we examine how the TCP variants behave during consistent network updates, in terms of the total number of packet drops, average throughput, variation in congestion window size cwnd and average number of retransmissions. In the experiment, by defining a threshold, the SDN controller initiates a network update with reference to rerouting every second. Using the same topology with two scenarios of large buffer size and small buffer size, we run the experiment.

(1) Number of retransmissions

Figures 3 and 4 plot the cumulative distribution functions (CDFs) of the numbers of average retransmissions for the three TCP variants during consistent network updates with large and small buffer sizes, respectively. There are different reasons for retransmissions, e.g., timeouts, damaged packet data, out-of-order packets, etc. For the case of large buffer and at the CDFs of 20%, 80%, and 100%, BBR requires 100, 200, and 280 retransmissions, and DCTCP requires 180, 400, and 600 retransmissions, while CUBIC imposes 380, 580, and 800 retransmissions. For the case of small buffer and at the CDFs of 20%, 80%, and 100%, BBR requires 130, 300, and 500 retransmissions, and DCTCP requires 200, 350, and 550 retransmissions, while CUBIC imposes 180, 350, and 550 retransmissions. It is evident that BBR in the both cases outperforms the other two TCP variants. For BBR, larger buffer improves the performance and reduces the number of

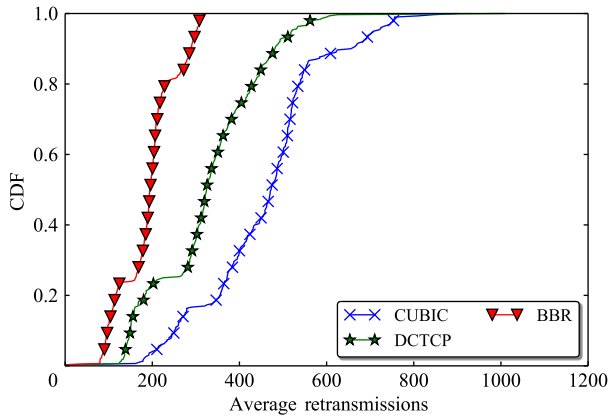


Fig. 3 Numbers of average retransmissions during consistent network updates for TCP variants with large buffer size.

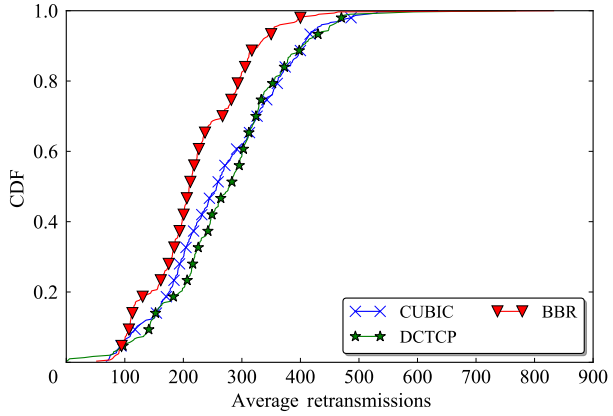


Fig. 4 Numbers of average retransmissions during consistent network updates for TCP variants with small buffer size.

retransmissions, also see Ref. [31]. BBR attempts to find the optimal operating point during network updates by estimating the bandwidth and round-trip propagation delay to take care of bandwidth and round-trip time. BBR also ignores packet loss as a congestion signal[27, 32]. It can also be seen that with the small buffer size, the retransmission performance of DCTCP and CUBIC improve.

(2) Throughput

Figures 5 and 6 depict the CDFs of the average throughput for the three TCP variants during consistent network updates with the large and small buffer sizes, respectively. For the large buffer and at the CDFs of 20%, 80%, and 100%, CUBIC can reach around 3.5 Gbps, 4.8 Gbps, and 6 Gbps, and DCTCP achieves around 2.8 Gbps, 4.7 Gbps, and 5 Gbps, while BBR reaches around 3.3 Gbps, 4.5 Gbps, and 5 Gbps. It appears that CUBIC has the edge in this case. For the

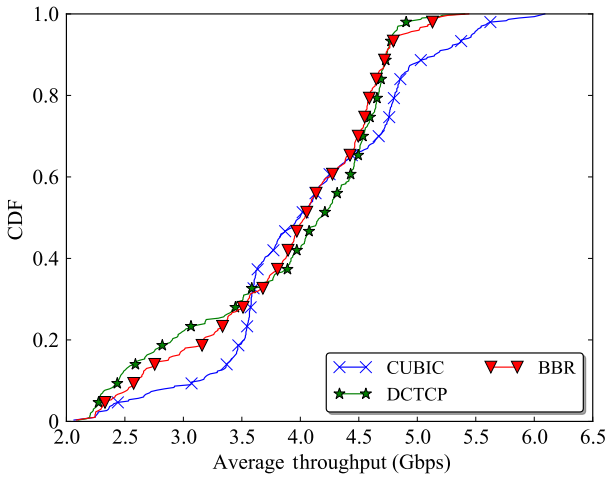


Fig. 5 Average throughput during consistent network updates for TCP variants with large buffer size.

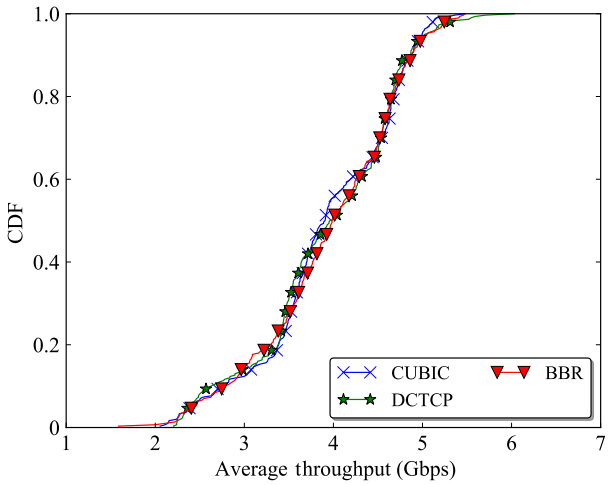


Fig. 6 Average throughput during consistent network updates for TCP variants with small buffer size.

small buffer scenario, the three TCP variants have similar performance. Specifically, at the CDFs of 20% and 80%, the three schemes achieve around 3.3 Gbps and 4.8 Gbps, while at the CDF of 100%, CUBIC and BBR achieve about 5.2 Gbps, but DCTCP has a slight edge reaching near 6 Gbps.

(3) Variation in congestion window

Figures 7 and 8 characterize the variations in the average congestion window size cwnd by the three TCP variants with large and small buffer sizes, respectively. For the large buffer and at the CDFs of 20%, 80%, and 100%, the cwnd sizes of BBR are 150 kB, 280 kB, and 350 kB, and DCTCP has the cwnd sizes of 220 kB, 450 kB, and 650 kB, while CUBIC has the cwnd sizes of 400 kB, 750 kB, and 820 kB. For the small buffer and at the CDFs of 20%, 80%, and

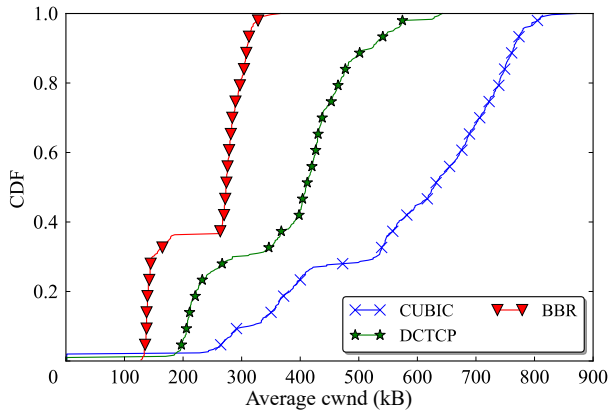


Fig. 7 Variations in average congestion window size during consistent network updates for TCP variants with large buffer size.

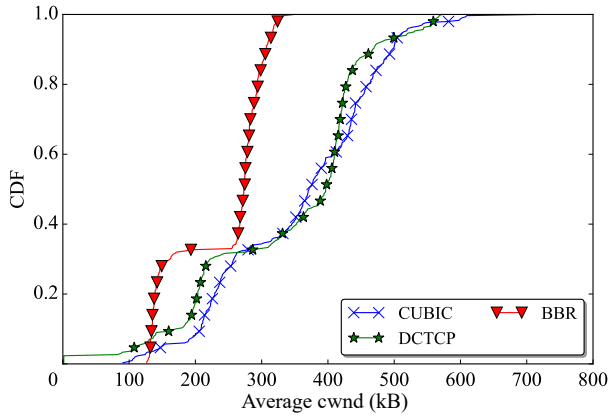


Fig. 8 Variations in average congestion window size during consistent network updates for TCP variants with small buffer size.

100%, BBR has the cwnd sizes of 140 kB, 280 kB, and 330 kB, and DCTCP has the cwnd sizes of 200 kB, 430 kB, and 550 kB, while CUBIC has the cwnd sizes of 230 kB, 450 kB, and 700 kB. It is seen that BBR has the smallest cwnd and the buffer size has little impact on its congestion window size. This is because BBR controls the congestion window by setting cwnd to two times the estimated bandwidth-delay product (BDP)[31]. DCTCP uses an effective multiplicative reduction technique to adjust cwnd based on the level of network congestion. Specifically, DCTCP uses the explicit congestion notification (ECN) to estimate the predicted proportion of the marked packets and appropriately modifies its cwnd size[23]. CUBIC has the largest cwnd because it has a tendency to fully fill a buffer.

(4) Number of packet drops

DCTCP and CUBIC have no dropped packets in consistent network updates with a large buffer. Therefore, we perform an experiment with the small buffer size, and the numbers of packet drops experienced by the three TCP variants are shown in Fig. 9. The results indicate that BBR has the worst performance and DCTCP has the best performance. More specifically, BBR drops 18 336 packets, CUBIC drops 7865 packets, and DCTCP drops merely 3693 packets, respectively.

When packet loss is observed, BBR ignores it as a congestion indication and does not perform back off. Hence, it has no mechanism for congestion detection and reaction to congestion[31]. Due to the fact that CUBIC has an ACK-based congestion control mechanism, it cannot deal with packet drop. DCTCP is also unable to control packet loss if there are severe and short-lived traffic bursts[23]. Therefore, traditional TCP variants are unable to deal efficiently with the packet drop problem in frequent consistent network updates.

2.3 Analysis for inconsistent network updates

Inconsistencies arise when data plane state changes violate policies or update rules due to varying delays in coordination or no coordination across multiple switches or between controller and switch[28]. Delayed updates in the network may cause inconsistent network updates. To evaluate TCP variants, we use switches with an SDN controller to produce an update delay inconsistency of 100 ms, with a large buffer size.

(1) Number of retransmissions

As can be seen from Fig. 10, BBR exhibits the best retransmission performance, while CUBIC has the worst retransmission performance, which is similar to Fig. 3. Specifically, at the CDFs of 20%, 80%, and

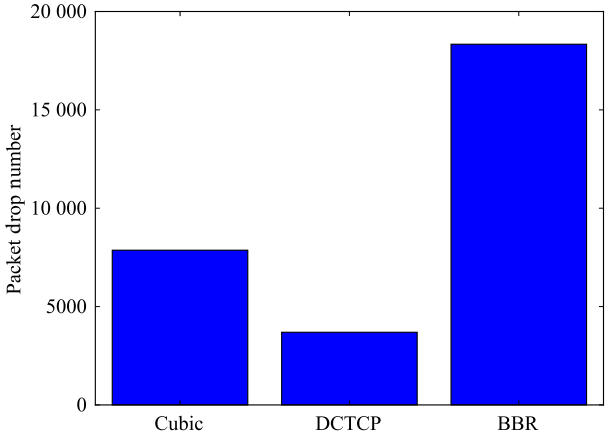


Fig. 9 Average numbers of packet drops during consistent network updates for TCP variants with small buffer size.

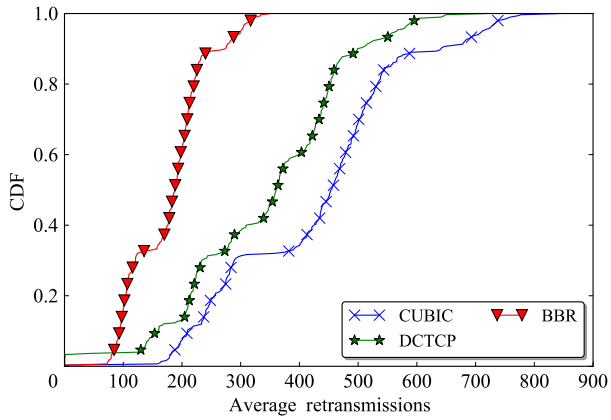


Fig. 10 Numbers of average retransmissions during inconsistent network updates for TCP variants with large buffer size.

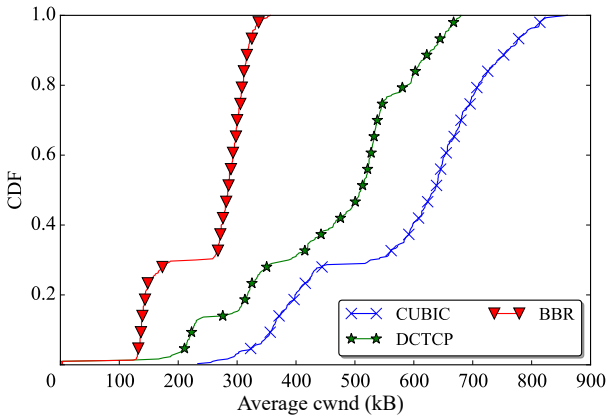


Fig. 12 Variations in average congestion window size during inconsistent network updates for TCP variants with large buffer size.

100%, BBR requires 100, 200, and 320 retransmissions, and DCTCP requires 200, 450, and 650 retransmissions, while CUBIC imposes 250, 520, and 800 retransmissions. The retransmission performance of BBR and DCTCP are slightly worst than the corresponding performance under consistent network updates given in Fig. 3, but CUBIC has slightly better performance.

(2) Throughput

Figure 11 shows the throughput of the three TCP variants during inconsistent network updates. All the three TCP variants exhibit similar performance, which is different from Fig. 5. In particular, for the CDF above 40%, the throughput of all the three TCP variants are very close.

(3) Variation in congestion window

Figure 12 represents the average congestion window size variations in inconsistent network updates with a large buffer size, which are similar to Fig. 7. As mentioned in Subsection 2.2, BBR controls the congestion window by setting cwnd to two times the BDP and it shows the least window size growth, while CUBIC uses an ACK-based congestion control mechanism to maximally grow the window size. DCTCP is ECN-based and it squeezes the window size based on ECN-marked packets.

(4) Number of packet drops

According to Fig. 13, CUBIC shows the worst performance with 623 packets dropped, and the BBR shows the best performance with 11 packets dropped, while DCTCP drops 187 packets. BBR with a large buffer size has less packet loss and retransmissions because the inflight cap limits the usage of the buffer at around one BDP, which prevents packet loss most of

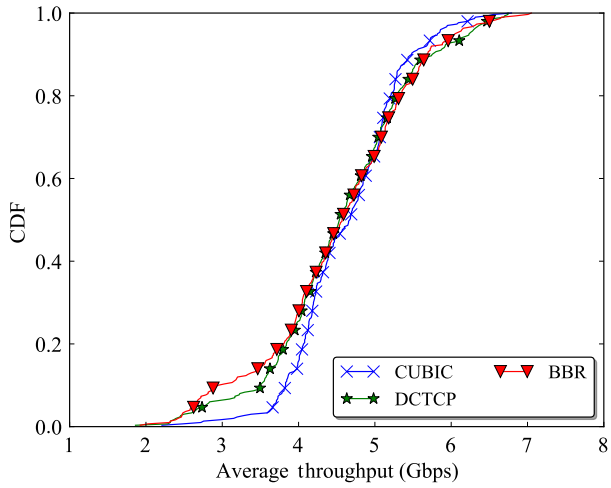


Fig. 11 Average throughput during inconsistent network updates for TCP variants with large buffer size.

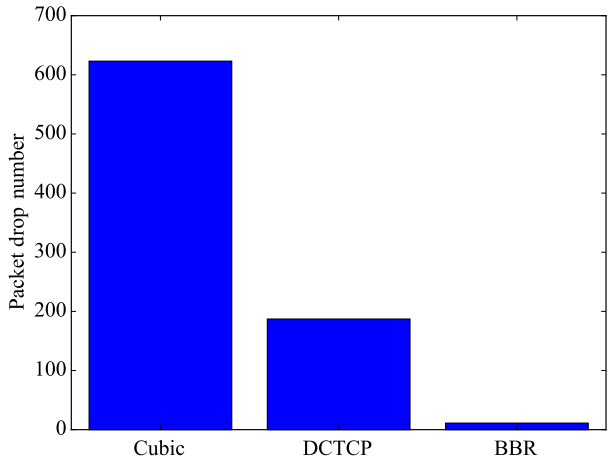


Fig. 13 Average numbers of packet drops during inconsistent network updates for TCP variants with large buffer size.

the time[31]. It can easily be inferred that the situation would worsen with a small buffer size during inconsistent network updates. Recalling that DCTCP and CUBIC have no dropped packets in consistent network updates with a large buffer, it can be seen that the problem of packet drops becomes much serious in inconsistent network updates.

2.4 Conclusions of TCP variants analysis

In the scenario of consistent and frequent network updates with large buffer size, DCTCP and CUBIC have no dropped packets but they impose higher average number of retransmissions and larger growth in cwnd size than BBR. In the same scenario with small buffer size, BBR has the largest number of dropped packets and DCTCP has the smallest number of dropped packets. This is because the sender in DCTCP estimates the buffer size by maintaining an estimate of fraction of packets that are experiencing congestion. The estimated fraction of packets α, can be updated after one RTT (round-trip time) for every window of data as follows.

$$\alpha = (1-g)\times\alpha + g\times F \tag{1}$$

where the term $(1-g)\times\alpha$ represents the decay of previous value of α, while the term $g\times F$ represents the new information from the feedback or increase in α due to the current packet being marked as congested. The gain factor g controls the relative weight of these two terms. Consequently, Eq. (1) is used to update the value of α based on the current value of α and the feedback received from ACKs and negative acknowledgments (NAKs). Thus, upon reception of every packet, the sender receives ECN-marks when the queue length is higher than threshold value K. Moreover, the value of estimation close to 1 indicates high congestion and close to 0 indicates low congestion. While the performance of the three TCP variants are closer in terms of retransmission. It can be observed from the experimental results that BBR has the smallest cwnd size. The reason is that BBR controls the cwnd based on the bandwidth delay product, BDP, as follows:

$$\text{cwnd} = 2\times\text{BDP} \tag{2}$$

where BDP can be calculated as

$$\text{BDP} = B_r\times\text{RTT}_{\min} \tag{3}$$

Here, B_r is the smallest data rate (bottleneck data rate) and $\text{RTT}_{\min}$ is the minimal RTT. Conversely, the DCTCP calculates the cwnd as

$$\text{cwnd} = \text{cwnd}\times\left(1-\frac{\alpha}{2}\right) \tag{4}$$

Hence, when α is close to 0, indicating low congestion, reducing the window slightly. As the DCTCP senders reducing cwnd gently, therefore, its average cwnd is larger than BBR.

The situation is similar in terms of retransmission and cwnd size, when updating the network inconsistently with large buffer size, but all the three TCP variants suffer from the problem of dropped packets. This indicates that the problem of out-of-order packets is more serious when the network is updated inconsistently and frequently. It may also be reasonably inferred from the results that the performance of TCP variants will deteriorate in the case of small buffer size.

3 NUFTCP design

The previous simulation experiments have revealed that the state-of-the-art TCP variants are ineffective, particularly when the network is updated inconsistently and frequently. This motivates us to design a better solution that is capable of coping with network updates smoothly. Specifically, we develop the NUFTCP design, which aims to mitigate the problem of dropped packets during network updation by restricting the window size, managing the queue and handling out-of-order packets in SD-DCN. More specifically, two modifications are introduced. The first one restricts the size of the transmit side window to mitigate dropped packets in inconsistent network environments. This is because the transmission window on the sending side decides the amount of data that can be sent before the receiver sends an acknowledgment, ensuring that it does not exceed the receiver's buffer. The second one diminishes out-of-order packets by delaying duplicated ACKs on the recipient side. The main symbols utilized in the design are listed in Table 2.

3.1 Limit the window size and queue management

(1) Why NUFTCP cares about packet drop not congestion avoidance

NUFTCP does not focus on ordinary congestion avoidance; instead, it avoids severe congestion, which leads to packet drop. A TCP protocol keeps adding packets to the queue until a packet is dropped. Then TCP extends queues to accommodate transient traffic bursts, and hence the average queue length is quite

Table 2 Main notations.

Symbol	Meaning
$\{f_i\}$	Set of flows
$\{s_i\}$	Set of data size
$\{t_i\}$	Set of round-trip times (RTTs)
$\{r_i\}$	Set of flows' throughput
$\hat{t}$	RTT without queuing latency
c	Link capacity
l	Queue length
$l_{\max}$	Maximum queue length
τ	Queue latency
w	Window size
$\overline{s}$	Data size acknowledged
α	Estimated marked packets
F	Fraction of marked packets
BDP	Bandwidth delay product
B_r	Smallest data rate
$\text{RTT}_{\min}$	Minimum RTT
cwnd	Congestion window

long. Therefore, the rate of dropped packets is incredibly low until a queue is full, thereafter it is possible that all the incoming packets from all flows are lost. Consequently, packet drop slows down all the TCP flows. In the worst-case scenario, a packet drop can result in significant data corruption or possibly the loss of a link completely. For this reason, we are more apprehensive about packet drops. NUFTCP utilizes the inherent negative feedback to prevent the problem of packet drops triggered by severe congestion.

(2) Relationship among data size, flow throughput, and RTT

Since TCP is not aware of the congestion of the underlying network, it relies on a time limit, which shows the duration a sender waits before retransmitting a lost segment to estimate whether the forwarded segment is dropped or not. Specifically, if an acknowledgment is not received by the sender side within the defined time limit, it is considered a segment drop. In the context of standard TCP, the time limit is referred to as the retransmission timeout (RTO). The RTO is not a fixed value, but rather it is dynamically adjusted based on the estimated RTT. The RTO is maximum amount of time that TCP will wait for an ACK for a segment before retransmitting it. The RTO is typically set to a few times the estimated RTT, which is the time it takes for a segment to travel from the sender to the receiver and back. It is helpful to account for factors such as network delays and processing times. Further, if TCP retransmits a segment too many times without receiving an ACK, it assumes that the connection is congested and will slow down its sending rate. This is known as congestion control. The reason TCP relies on time limits or timeouts to estimate congestion is because it does not have direct access to network information such as buffer occupancy and queue lengths. Therefore, the RTO is one of the common ways to determine the packet drop. Accordingly, TCP keeps track of RTT for each segment. A TCP connection receives an estimated maximum data size that the receiver side can accommodate. Let RTT be t and flow throughput be r. Then data size s transmitted in t is given by $s = r \times t$. In other words, the flow throughput is proportional to the data size and inversely proportional to the RTT. When a link begins to exhibit congestion, the queue length for the link is increased, which induces the rise in the queuing latency of the packets going through the link. This implies that the RTT of the corresponding flow will be longer and the throughput will be lower in order to alleviate the congestion. Therefore, when we transmit the data size in one RTT, the throughput is controlled by negative feedback. There will be no packet drop as long as the buffer size of the queue is sufficiently large.

(3) Queue management

We will use the example of Fig. 14 to illustrate queue management. Prior to the update, assume that we have $s_1 = r_1 \times t_1$ and $s_2 = r_2 \times t_2$, where s_1, r_1, and t_1 are related to flow f_n, while s_2, r_2, and t_2 are related to flow f_m. In this case, $r_1 = r_2 = c$ and $t_1 = t_2 = \hat{t}$.

Assume that when the network first commences updating, it offers an inconsistent network state. Subsequently, the network state of f_n is updated and its route is altered, while the network state of f_m remains unchanged. Specifically, flow f_n is routed to the intermediate node that f_m is currently using. Both the flows now utilize the same link, and the queue is overburdened, resulting in packet loss. For the sake of simplicity, suppose that the queue length (in bytes) in the stable state is l. Dividing the queue length by the link capacity yields the queuing latency $\tau = l/c$. After the inconsistent network update, $s_1 = r_1' \times t_1'$ and $s_2 = r_2' \times t_2'$ with $t_1' = t_2' = \hat{t} + \tau$ and $r_1' + r_2' = c$. Hence, we have $l = \sum_i s_i - c \times \hat{t}$. Because $\sum_i s_i \leqslant 2 \times c$, we have $l \leqslant c \times \hat{t}$. Thus, the maximal queue length should be $l_{\max} < l + \sum_i s_i$ (if the queue length in the previous RTT is more than l, the queue length will not be increased).

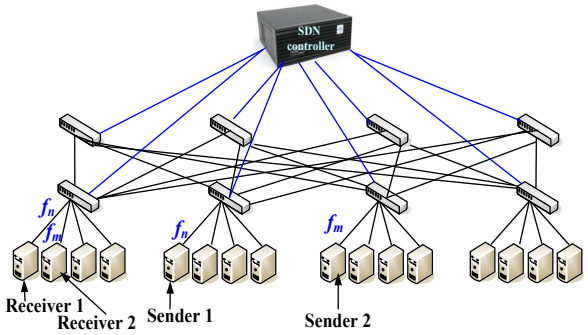


Fig. 14 Topology of a software defined data center network.

Therefore, we can draw the conclusion:

$$l_{\max} < 3 \times c \times \hat{t} \tag{5}$$

For example, if $c = 10$ Gbps and $\hat{t} = 250$ μs, $l < 937.5$ kB. It implies that if the buffer size is 1 MB, no packets should drop. Equation (5) ensures that packets are not dropped and network stability is maintained throughout network updates. The conclusion can be extended to more general cases. Algorithm 1 describes the queue management process of NUFTCP.

(4) How NUFTCP limits window size

Benefiting from negative feedback, NUFTCP aims to limit the window size w when a congestion occurs. As aforementioned, congestion avoidance protocol grows the window size constantly and linearly, which may cause large number of packet drops in severe congestion situation. NUFTCP limits the maximum window size to avoid this issue. When there is a congestion, the window size w is nearly equal to one TCP flow, i.e., $w \approx s$. If the flow does not face congestion, w may be significantly greater than s. TCP uses an exponentially weighted moving average (EWMA) to keep track of the fluctuations of RTT over time, and it places approximately 20 percent of the weight on the most recent RTT measurement. Therefore, NUFTCP keeps track of the average actual data size acknowledged $\bar{s}$ in the past $\hat{t}$ using EWMA. The upper window size limit is then adjusted to $1.5 \times \bar{s}$ in the following $\hat{t}$, where the factor 1.5 is employed because the window size must be able to grow gently.

Since NUFTCP is based on DCTCP, packets will not be discarded due to random early detection (RED), which is founded on statistical probabilities and is more balanced than the tail drop. However, DCTCP will decrease the slow start threshold when the transmit side receives packets with an explicit congestion expected (ECE) flag, which is part of an explicit congestion notification (ECN) protocol. If NUFTCP does not receive a packet with the ECE flag in 10 consecutive RTTs, it will reset the slow start threshold. This is because to prevent the excessive window size limit, avoiding congestion, and proactively addressing potential packet loss in the network. Algorithm 2 presents the congestion control and window size adjustment of NUFTCP.

Algorithm 1 NUFTCP queue management algorithm

1: **Input:** l (queue length in bytes), c (link capacity), $\hat{t}$ (time ween RTTs), S_i (Data size for flow f_i), t_i (Transmission time for flow f_i), τ (Queuing latency)
2: **Output:** $l_{\max}$ (maximum queue length)
3: $l_{\max} \leftarrow 0$ ▷ Initializing variable
4: $l \leftarrow \sum_i s_i - c \times \hat{t}$ ▷ Update queue length
5: $l_{\max} \leftarrow 3 \times c \times \tau$ ▷ Calculating maximum queue length
6: **if** $l_{\max} > l$ **then**
7: Set $l \leftarrow l_{\max}$ ▷Updating queue length
8: Drop packets until queue length is reduced to $l_{\max}$
9: **end if**
10: $s_i \leftarrow s_i - \left(\frac{\tau \times s_i}{l_{\max}}\right)$ ▷ Calculating data size for each flow (s_i):
11: Transmit data for each flow (s_i)
12: Repeat steps 5–11 until queue is empty

3.2 Delay duplicated ACKs

NUFTCP can rectify the out-of-order issue by delaying the delivery of duplicate ACKs during the network update. The duplicated ACK will only be returned if it is postponed for period T and the accompanying data packet has not yet been delivered. A duplicated ACK is rejected in all the other cases. Here, T is a preconfigured time threshold that equals to the maximal RTT difference on different paths, typically smaller than 1 ms or 2 ms. Since the window size limit avoids the packet drop, the impact of delaying the fast recovery is tolerable.

To be able to correctly delay duplicate ACKs, NUFTCP needs to first identify network update. The cooperation of switches, like ECN, is necessary for NUFTCP. A version number is assigned to each flow table entry in the switches. Each time the controller recalculates a flow table entry, the version number will be raised by one. Each packet that matches the entry will be marked with the current version number of the entry (if the version number is greater than the one carried in the packet). The first 4 bits in the TTL field are used for version number tagging because it is

Algorithm 2 Congestion control and window size adjustment in NUFTCP

1: **Variables:** s (Average actual data size acknowledged in the past using EWMA), ECE_flag (Explicit congestion expected flag)
2: **Output:** w (Window size)
3: $w \leftarrow s$ ▷ Initialize congestion window w
4: SST $\leftarrow s$ ▷ Initialize slow start threshold
5: ECE_flag_counter = 0 ▷ Initialize ECE flag counter
6: **while** true **do**
7: **if** congestion detected **then**
8: $w \leftarrow s$ ▷ Update window size
9: **end if**
10: **if** congestion is not detected **then**
11: **if** $w <$ SST **then**
12: $w \leftarrow \min(w+1, \text{SST})$ ▷ Gradually increase window size
13: **else**
14: $w \leftarrow \min(w+1.5s, w_{\max})$ ▷ Update w using weight = 1.5
15: **end if**
16: **end if**
17: **if** an ECE_flag is received **then**
18: SST $\leftarrow s$ ▷ Update slow start threshold
19: ECE_flag_counter $\leftarrow 0$ ▷ Reset ECE_flag_counter
20: **else**
21: ECE_flag_counter++ ▷ Increment ECE_flag_conter
22: **end if**
23: **if** ECE_flag_counter == 10 **then**
24: SST $\leftarrow s$ ▷ Reset slow start threshold
25: ECE_flag_counter $\leftarrow 0$ ▷ Reset ECE_flag_counter
26: **end if**
27: $w \leftarrow \min(w, \text{SST})$ ▷ Update window size
28: Transmit data using window size w
29: **end while**
30: Repeat steps 6–29 until congestion is resolved or slow start threshold is reached

generally futile for intra-data center traffic whose maximal hops are smaller than 15. By tracking TTL modifications, NUFTCP has the capacity to explicitly identify network updates and defer repeated ACKs. Table 3 presents the performance comparison of BBR[27], CUBIC[25], DCTCP[23], and the proposed NUFTCP. Algorithm 3 describes the duplicate ACK delay mechanism of NUFTCP.

4 Evaluation

The currently available network updating methods[2, 3, 7, 9–11, 20, 29] take into account the violations of consistency properties and network update frequency. Hence the comparison between NUFTCP and these methods is unsuitable, as NUFTCP is concerned with how dropped packets and out-of-order packets issues of network updates influence TCP performance. For this reason, we compare NUFTCP with the state-of-the-art TCP variant DCTCP, which has been shown to outperform other TCP variants as demonstrated in the experiments of Section 2.

4.1 Simulation network topology and parameters

We assess NUFTCP using NS3 based simulations under two different conditions of large and small buffer sizes. We use SD-DCN topology to compare NUFTCP and DCTCP. As illustrated in Fig. 14, we utilize an advanced conventional 2-tier Clos network topology in DCN attached through an SDN controller. The network makes use of equal-cost multipath routing (ECMP). Each link has a 10 Gbps capacity with $\hat{t} \approx 500\ \mu s$ RTT without queuing latency. The large and small buffers have 1 MB and 0.5 MB in size, respectively. The controller transmits the updated information towards the edge switches every 25 ms, assuming one of the aggregation layer switches is down (or upgrading). Flows passing precisely via the down (or upgrading) switches must modify their transmitting pathways backward and forward in this format. NUFTCP strategy utilizes four bits of the packet header's TTL field to indicate the version number, and hence in our simulations, the utmost multitude of distinct versions is sixteen.

Our network architecture states that f_n flows are transmitted from Sender 1 to Receiver 1 and f_m flows are transmitted from Sender 2 to Receiver 2 simultaneously. Because there is inconsistency across switches during the update, we simulate update delay of $0 - 1$ ms in the switches. The variation in update time is produced at random, similar to Ref. [24].

4.2 Results and discussion

(1) Packet drops and out-of-order packets

A randomized delay is used to preserve inconsistency throughout the network updates. In the scenario of large buffer size, we execute several simulations and find that the numbers of packet drops caused by DCTCP are 1188, 1294, 1159, and so on. The average number of dropped packets is 1213, which is displayed in Fig. 15. By contrast, NUFTCP never drops a packet as can be seen from Fig. 15. As shown in Fig. 16, the

Table 3 Performance comparison of different TCP variants.

Feature	BBR[27]	CUBIC[25]	DCTCP[23]	Proposed NUFTCP
Objective	High bandwidth utilization	Fairness and high throughput	Data center congestion control	Addresses packet drop problem during frequent consistent network updates in SD-DCN
Congestion control type	Delay-based	Window-based	ECN-based	ECN-based on inherent negative feedback
Congestion signal	RTT	Packet loss	ECN markings	Inherent negative feedback
Response to congestion	Proactive, aims to prevent congestion	Reactive, reduces window size upon congestion	Proactive, utilizes ECN	Proactive, utilizes inherent negative feedback
Awareness of frequent network updates	Low	Low	Moderate	High
Throughput behavior	Aims for high throughput	Tends to oscillate	Balances throughput and low latency	High throughput
Throughput optimization	Yes	Yes	Yes	Yes
Buffer utilization	Efficient use of buffers	May lead to buffer bloat	Designed to avoid buffer bloat	Designed to avoid buffer bloat
Buffer bloat mitigation	Yes	No	Yes	Yes
Scalability	Yes	Yes	Yes	Yes
Use case	Broadband connections	General-purpose	Data center environments	SD-DCN environments

Algorithm 3 NUFTCP duplicate ACK delay mechanism

Input: Ack (Acknowledgment packet), T (Preconfigured time threshold for delaying duplicate ACKs), packet_received (Flag indicating whether the corresponding data packet has been received), current_version (Current version number of the flow table entry), packet_version (Version number carried in the ACK packet), t (time since ACK reception)
Output: processed_ACK (Boolean flag indicating whether the ACK was processed or not)
while processed_ACK == False **do**
  if current_version > packet_version **then**
    ACK indicates a network update
  end if
  if If an ACK is duplicate **then**
    if If $t > T$ and packet_received == False **then**
      processed_ACK ← True
      packet_received ← True
    else
      processed_ACK ← False ▹ Discard the ACK
    end if
  end if
  if ACK is not a duplicate **then**
    processed_ACK ← True
  end if
end while
Return processed_ACK flag

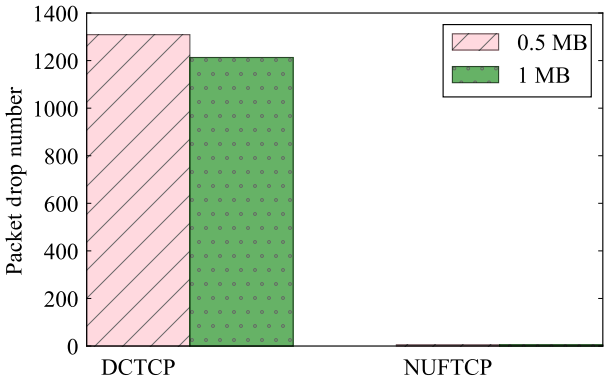


Fig. 15 Comparison of packet drops for DCTCP and NUFTCP during network updates with large and small buffer sizes.

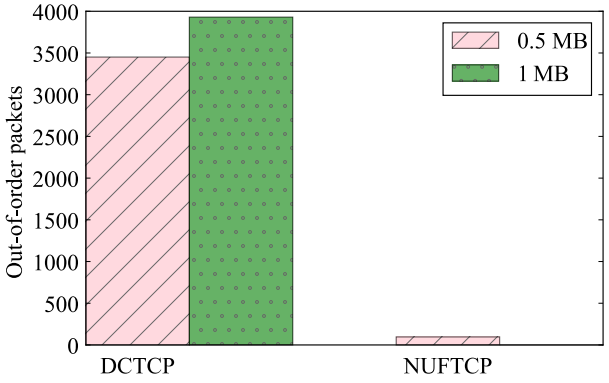


Fig. 16 Comparison of out-of-order packets for DCTCP and NUFTCP during network updates with large and small buffer sizes.

typical number of out-of-order packets for DCTCP is 3930, which is very high. NUFTCP on the other hand performs much better than DCTCP with no out-of-order packets.

In the scenario of small buffer size, after several simulations, the numbers of packet drops caused by DCTCP are found to be 1381, 1366, 1181, and etc. The average number of packet drops by DCTCP is 1309 as depicted in Fig. 15. Because a small buffer may lead to early packet drops, DCTCP exhibits a higher number of packet drops than in the large buffer scenario. Again NUFTCP does not suffer from packet loss. In Fig. 16, DCTCP exhibits a high number of out-of-order packets (3451) but is less than in the large buffer case. It is rare for NUFTCP to suffer from the problem of out-of-order packets and in this case, it only has 96 out-of-order packets.

(2) Throughput

For the case of large buffer size, Figs. 17 and 18 depict the throughput achieved by DCTCP and NUFTCP, respectively, during the network update. Observe from Fig. 17 that for DCTCP, because of packet drops in the beginning, both its flows' throughput are reduced to almost zero. Then after some time, one of its flow throughput is reduced to almost zero owning to packet drops occurring again. Finally, there is a prolonged converging time of its two flows' throughput. As can be seen from Fig. 18, NUFTCP by contrast offers more consistent throughput since packets are not dropped and out-of-order problem is dealt with effectively.

In the scenario of small buffer size, the throughput achieved by DCTCP and NUFTCP during the network update are depicted in Figs. 19 and 20, respectively. It can be seen from Fig. 19 that DCTCP performs even worst than in the large buffer case, and there exists a long period of almost zero throughput for its two flows. NUFTCP on the other hand exhibits more consistent and stable throughput throughout the whole update

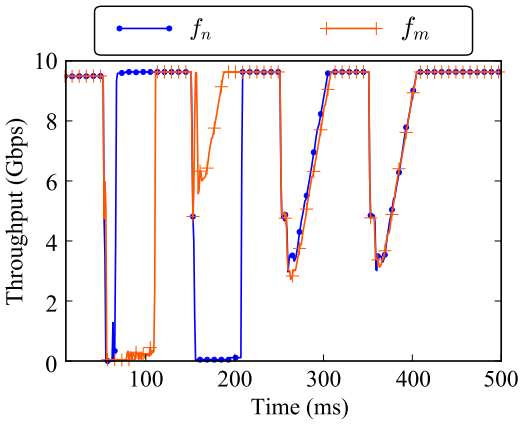


Fig. 17 DCTCP throughput during network updates with large buffer size.

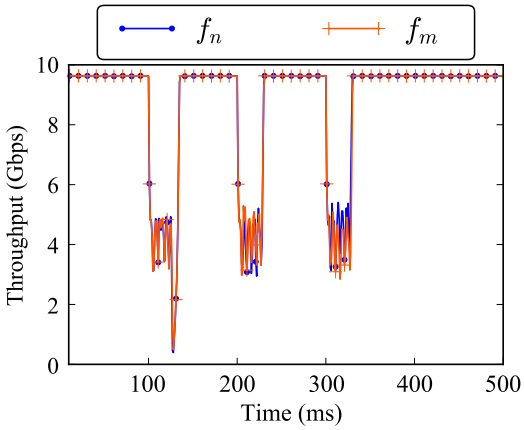


Fig. 18 NUFTCP throughput during network updates with large buffer size.

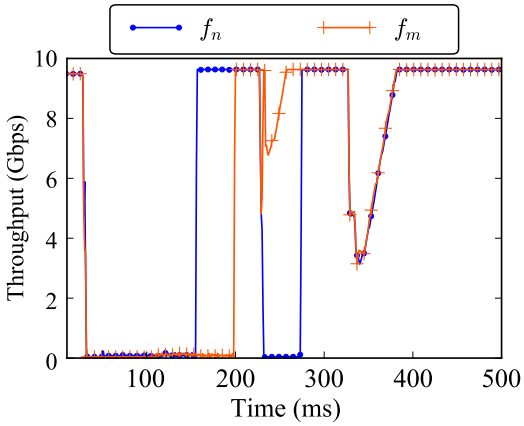


Fig. 19 DCTCP throughput during network updates with small buffer size.

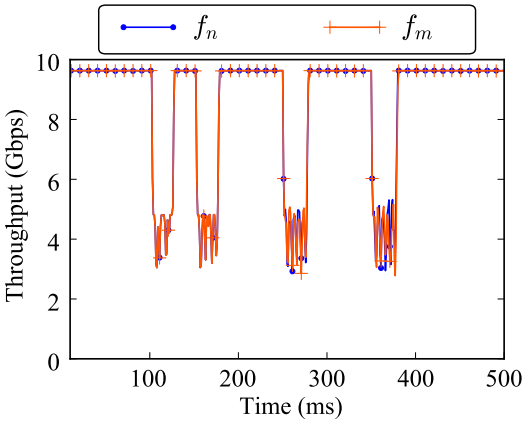


Fig. 20 NUFTCP throughput during network updates with small buffer size.

period, just as in the case of large buffer. This clearly demonstrates that NUFTCP can effectively maintain a stable network throughput during network update.

(3) Discussion

The aforementioned experimental results again confirm that the state-of-the-art TCP variants, such as DCTCP (Data Center TCP), suffer from the serious problem of packet drops and out-of-order packets during frequent and inconsistent network updates, which significant degrades their achievable throughput, particularly during the early period of update. The simulation results also validate that our proposed NUFTCP achieves its design goals. Specifically, NUFTCP can effectively deal with the issues of packet drops and out-of-order packets, and consequently it achieves better and more consistent throughput during inconsistent and frequent network updates. Based on this evaluation, we may draw the conclusion that our NUFTCP design offers an effective means for handling inconsistent and frequent network updates in SD-DCN. However, NUFTCP is unable to completely mitigate the issue of TCP Incast congestion. Incast congestion refers to a decrease in throughput when multiple senders simultaneously communicate with a single receiver, thereby exceeding the receiver's buffer capacity. Hence, our future work will seek to address this issue.

5 Related Work

5.1 TCP designs

All TCP designs offer some kinds of congestion algorithms to avoid and resolve congestion problems and to attempt to mitigating the issues of out-of-order packets and packet drops. The relevant contemporary TCP implementations are reviewed in this subsection.

Several delay based-TCP variants[33–35], have been proposed, which rely on packet delay measurements as a signal of congestion[36]. These schemes aim to reduce queue lengths and congestion at routers. However, queuing delays in data centers, are comparable to sources of noise in the system, thus, unable to provide a reliable congestion signal. Moreover, delay signals are not accurate enough to compute appropriate congestion window to reduce congestion at routers[37]. TCP Reno[38, 39] presents a fast recovery mechanism using available bandwidth designated by the arrival of DUPACK but it is intolerant to connections with long delays. TCP NewReno[40, 41] is a loss-based congestion algorithm that is an extension of TCP Reno with a modified fast recovery algorithm. CUBIC[25] introduces a cubic function of elapsed time for window growth when the loss occurs, and hence it improves the friendliness of binary increase congestion control (BIC). DCTCP[23] alters ECN, so that switches mark packets in accordance with the current queue length and senders modify the size of their send window according to the estimated fraction of marked packets. Linux TCP now has a novel congestion control technique called bottleneck bandwidth and RTT (BBR)[27]. BBR finds a better operating point that takes care of bandwidth and RTT by estimating the round-trip propagation delay and bandwidth, and it sets cwnd to a small multiple of the estimated BDP. TCP-PR[42] and TCP-RR[43] were developed for persistent out-of-order packets but they are not suitable for the DCN update scenario. TCP-RR relies on DSACK, which is not supported by all servers, and TCP-PR needs to maintain tables in memory which imposes high computation costs.

The aforementioned TCP variants are ineffective to deal with the issues of out-of-order packets and packet drops occurred in inconsistent and frequent updates of SD-DCNs. By contrast, our proposed NUFTCP design is capable of dealing with the problems of out-of-order packets and packet drops effectively and, consequently, ameliorates TCP performance when networks are updated inconsistently and frequently.

5.2 Network updates

The literature of network updates provides state-of-the-art solutions for mitigating the problems of forwarding loops, forwarding blackhole, link congestion, and policy violation, which cause inconsistent network updates. Since inconsistent network updates may lead to the issues of out-of-order packets and packet drops, these solutions aim to maintain consistency.

A single switch can handle the difficulties of forwarding loop and forwarding blackhole, as mentioned in a method by Reitblatt et al.[20] To distinguish the old and new packets, they are stamped with version numbers to implement old and new rules. zUpdate[2] provides a solution to the problems of congestion, forwarding loop, and forwarding blackhole in the data center, and it uses ECMP to split the traffic equally using multiple redundant paths. Hedera[3] handles frequent network updates in data centers by allocating the paths for large flows based on the estimated demand using an annealing based algorithm.

SWAN[7] achieves high network capacity utilization of inter-DC links in SDN in the presence of traffic volume variations, and it leaves a small amount of scratch capacity on the link that can be used for updating. TRUS[12] provided a timely route updating technique that reduces network congestion while meeting the bandwidth needs of delay-sensitive traffic. Dionysus[11] supports fast and consistent network updates in SDN using dynamic scheduling during updates at switches individually. It can be applied to both SDN WAN and DCN environments. FLIP[10] proposes an algorithm that ensures forwarding correctness and forwarding policies using a fast and lightweight algorithm. Cupid[29] emphasizes on consistent flow tables and data plane updating to maintain the throughput of flows, and it outperforms Dionysus. ez-Segway[9] presents a decentralized consistent update mechanism, which completes network updates quickly by utilizing sophisticated coordinating actions in the switches.

The authors of Ref. [21] proposed suffix causal consistency (SCC) motivated by a consistency model for shared-memory systems for rule updates. The method ensures consistency properties to avoid blackhole loops, bounded loops etc. The approach of Ref. [22] is based on on temporal logic and model checking for data flow correctness verification and concurrent updates using Petri nets to make sure the absence of loops. The authors in Ref. [15] devised algorithms to mitigate transient congestion, reduce update time, and minimize control overhead. Their algorithms optimize the intermediate stages after finding the optimal route at each middle stage to minimize the temporary congestion efficiently. The authors of Ref. [16] proposed customizable update planner (CUP) which adopts the existing designs to achieve the congestion avoidance and optimize the update speed. CUP introduces generic linear programming models to schedule network updates to user-specified needs, and it offers a solution to the transient congestion problem. Hermes[17] provided a utility-aware network update system that maximizes the total utility by a rate-limiting scheme before the update. It ensures congestion-free property during network updates. The authors in Refs. [18, 19] used resource dependency graph to formulate network update problems, approximation algorithms to utilize bandwidth resources, spare-path-assisted algorithms for consistent flow migration, and rate-limiting-flow to resolve deadlocks. Their method ensures fast network updates with consistency properties. In Ref. [44], the authors emphasized that the real-time communication should remain invariant by diverting the traffic to uninvolved devices during network updates. The authors of Ref. [14] introduced a framework based on abstract algebra that enables controllers to combine the fast composition of numerous network updates with persistent and non-blocking modifications in the network by efficiently modeling the data plane operations.

Most of the aforementioned network updating methods focus on avoiding the violations of consistency properties in network updates. By contrast, our NUFTCP design is developed to alleviate the impact of inconsistent and frequent network updates on TCP performance so that network updates can happen smoothly.

6 Conclusion

The contribution of this paper has been twofold. Firstly, we have conducted comprehensive experiments to evaluate the performance of the state-of-the-art TCP variants in the presence of frequent and inconsistent network updates in SD-DCNs. Our findings have confirmed that current TCP variants are incapable of handling frequent and inconsistent network updates, and they suffer from the problems of out-of-order packets and packet drops, which leads to significant performance degradation in terms of network throughput. Secondly, we have proposed a network update friendly TCP modification, called NUFTCP, which is an extension to DCTCP. Our NUFTCP design can tackle the issues of packet drops and out-of-order packets throughout frequent and inconsistent network updates in SD-DCNs, which have not been resolved by the previous works. Our evaluation results have validated that NUFTCP performs substantially better than the state-of-the-art DCTCP, when the network is updated frequently and inconsistently. Our NUFTCP therefore offers a useful design to smoothly handle network updates in SD-DCNs.

Acknowledgment

This work was supportted by the King Khalid University through the Large Group Project (No. RGP.2/312/44).

References

[1] D. Li, S. Wang, K. Zhu, and S. Xia, A survey of network update in SDN, *Front. Comput. Sci. Sel. Publ. Chin. Univ.*,

vol. 11, no. 1, pp. 4–12, 2017.
[2] H. H. Liu, X. Wu, M. Zhang, L. Yuan, R. Wattenhofer, and D. Maltz, Zupdate: Updating data center networks with zero loss, in *Proc. SIGCOMM 2013*, Hong Kong, China, 2013, pp. 411–422.
[3] M. Al-Fares, S. Radhakrishnan, B. Raghavan, N. Huang, and A. Vahdat, Hedera: Dynamic flow scheduling for data center networks, in *Proc. NSDI 2010*, San Jose, CA, USA, 2010, pp. 1–15.
[4] T. Benson, A. Anand, A. Akella, and M. Zhang, MicroTE: Fine grained traffic engineering for data centers, in *Proc. CoNEXT 2011*, Tokyo, Japan, 2011, pp. 1–12.
[5] K. T. Foerster, S. Schmid, and S. Vissicchio, Survey of consistent software-defined network updates, *IEEE Commun. Surv. Tutorials*, vol. 21, no. 2, pp. 1435–1461, 2019.
[6] U. Haider, M. Waqas, M. Hanif, H. Alasmary, and S. M. Qaisar, Network load prediction and anomaly detection using ensemble learning in 5G cellular networks, *Comput. Commun.*, vol. 197, pp. 141–150, 2023.
[7] C. Y. Hong, S. Kandula, R. Mahajan, M. Zhang, V. Gill, M. Nanduri, and R. Wattenhofer, Achieving high utilization with software-driven WAN, *SIGCOMM Comput. Commun. Rev.*, vol. 43, no. 4, pp. 15–26, 2013.
[8] S. Jain, A. Kumar, S. Mandal, J. Ong, L. Poutievski, A. Singh, S. Venkata, J. Wanderer, J. Zhou, M. Zhu, et al., *SIGCOMM Comput. Commun. Rev*., vol. 43, no. 4, pp. 3–14, 2013.
[9] T. D. Nguyen, M. Chiesa, and M. Canini, Decentralized consistent updates in SDN, in *Proc. SOSR 2017*, Santa Clara, CA, USA, 2017, pp. 21–33.
[10] S. Vissicchio and L. Cittadini, FLIP the (flow) table: Fast lightweight policy-preserving SDN updates, in *Proc. INFOCOM 2016*, San Francisco, CA, USA, 2016, pp. 1–9.
[11] X. Jin, H. H. Liu, R. Gandhi, S. Kandula, R. Mahajan, M. Zhang, J. Rexford, and R. Wattenhofer, Dynamic scheduling of network updates, *SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 4, pp. 539–550, 2015.
[12] J. Zhu, J. Hua, M. Liu, Y. Li, and K. Cao, TRUS: Towards the real-time route update scheduling in SDN for data centers, *IEEE Access*, vol. 8, pp. 68682–68694, 2020.
[13] J. Zhang, B. Gong, M. Waqas, S. Tu, and Z. Han, A hybrid many-objective optimization algorithm for task offloading and resource allocation in multi-server mobile edge computing networks, *IEEE Transactions on Services Computing*, vol. 16, no. 5, pp. 3101–3114, 2023.
[14] G. Li, Y. R. Yang, F. Le, Y. S. Lim, and J. Wang, Update algebra: Toward continuous, non-blocking composition of network updates in SDN, in *Proc. INFOCOM 2019*, Paris, France, 2019, pp. 1081–1089.
[15] J. Zheng, H. Xu, G. Chen, H. Dai, and J. Wu, Congestion-minimizing network update in data centers, *IEEE Trans. Serv. Comput.*, vol. 12, no. 5, pp. 800–812, 2019.
[16] S. Luo, H. Yu, L. Luo, and L. Li, Customizable network update planning in SDN, *J. Netw. Comput. Appl.*, vol. 141, pp. 104–115, 2019.
[17] J. Q. Zheng, Q. F. Ma, C. Tian, B. Li, H. P. Dai, H. Xu, G. H. Chen, and Q. Ni, Hermes: Utility-aware network update in software-defined WAN, in *Proc. ICNP 2018*, Cambridge, UK, 2018, pp. 231–240.
[18] Y. Chen, H. Zheng, and J. Wu, Consistency, feasibility, and optimality of network update in SDNs, *IEEE Trans. Netw. Sci. Eng.*, vol. 6, no. 4, pp. 824–835, 2019.
[19] M. Waqas, M. Zeng, Y. Li, D. Jin, and Z. Han, Mobility assisted content transmission for device-to-device communication underlaying cellular networks, *IEEE Trans. Veh. Technol.*, vol. 67, no. 7, pp. 6410–6423, 2018.
[20] M. Reitblatt, N. Foster, J. Rexford, C. Schlesinger, and D. Walker, Abstractions for network update, *SIGCOMM Comput. Commun. Rev.*, vol. 42, no. 4, pp. 323–334, 2012.
[21] S. Liu, T. A. Benson, and M. K. Reiter, Efficient and safe network updates with suffix causal consistency, in *Proc. Fourteenth EuroSys Conf. 2019*, Dresden, Germany, 2019, p. 1–15.
[22] B. Finkbeiner, M. Gieseking, J. Hecking-Harbusch, and E. R. Olderog, Model checking data flows in concurrent network updates. International Symposium on Automated Technology for Verification and Analysis, Cham, Switzerland: Springer, 2019, pp. 515–533.
[23] M. Alizadeh, A. Greenberg, D. A. Maltz, J. Padhye, P. Patel, B. Prabhakar, S. Sengupta, and M. Sridharan, Data center TCP (DCTCP), *SIGCOMM Comput. Commun. Rev.*, vol. 40, no. 4, pp. 63–74, 2010.
[24] A. Basta, A. Blenk, S. Dudycz, A. Ludwig, and S. Schmid, Efficient loop-free rerouting of multiple SDN flows, *IEEE/ACM Trans. Netw.*, vol. 26, no. 2, pp. 948–961, 2018.
[25] S. Ha, I. Rhee, and L. Xu, CUBIC, *SIGOPS Oper. Syst. Rev*., vol. 42, no. 5, pp. 64–74, 2008.
[26] A. B. Dogar and Y. Zhang, NUFTCP: Towards smooth network updates in software-defined datacenter networks, in *Proc. 2021 17th Int. Conf. Network and Service Management* (*CNSM*), Izmir, Turkey, 2021, pp. 365–369.
[27] N. Cardwell, Y. Cheng, C. S. Gunn, S. H. Yeganeh, and V. Jacobson, BBR, *Commun. ACM*, vol. 60, no. 2, pp. 58–66, 2017.
[28] R. Mahajan and R. Wattenhofer, On consistent updates in software defined networks, in *Proc. Twelfth ACM Workshop on Hot Topics in Networks*, College Park, MD, USA, 2013, pp. 1–7.
[29] W. Wang, W. He, J. Su, and Y. Chen, Cupid: Congestion-free consistent data plane update in software defined networks, in *Proc. IEEE INFOCOM 2016 - The 35th Annual IEEE Int. Conf. Computer Communications*, San Francisco, CA, USA. IEEE, 2016, pp. 1–9.
[30] H. Xu, Z. Yu, X. Y. Li, C. Qian, L. Huang, and T. Jung, Real-time update with joint optimization of route selection and update scheduling for SDNs, in *Proc. IEEE 24th Int. Conf. Network Protocols* (*ICNP*), Singapore, 2016, pp. 1–10.
[31] M. Hock, R. Bless, and M. Zitterbart, Experimental evaluation of BBR congestion control, in *Proc. IEEE 25th Int. Conf. Network Protocols* (*ICNP*), Toronto, Canada, 2017, pp. 1–10.
[32] N. Cardwell, Y. Cheng, C. S. Gunn, S. H. Yeganeh, and V. Jacobson, BBR: congestion-based congestion control, *Queue*, vol. 14, no. 5, pp. 20–53, 2016.
[33] Verma, L. P. Sharma, V. K. Kumar, M. Kanellopoulos,

and Dimitris, A novel delay-based adaptive congestion control TCP variant, *Computers and Electrical Engineering*, vol. 101, p. 108076, 2022.
[34] C. H. Chiang, Y. C. Chan, and P. L. Chen, Delay-based TCP with pacing and ECN for solving incast problem in data center networks, in *Proc. IET Int. Conf. Engineering Technologies and Applications* (*IET-ICETA*), Changhua, China, 2022.
[35] G. H. Kim and Y. Z. Cho, Delay-aware BBR congestion control algorithm for RTT fairness improvement, *IEEE Access*, vol. 8, pp. 4099–4109, 2020.
[36] N. Agarwal, M. Varvello, A. Aucinas, F. Bustamante, and R. Netravali, Mind the delay: The adverse effects of delay-based TCP on HTTP, in *Proc. 16th Int. Conf. Emerging Networking Experiments and Technologies*, New York, NY, USA, pp. 364–370, 2020.
[37] H. Ma and D. Xu, An INT-based TCP window modulator for congestion control in data center networks, *J. Netw. Comput. Appl.*, vol. 217, p. 103688, 2023.
[38] M. Allman, V. Paxson, and E. Blanton, TCP congestion control, https://doi.org/10.17487/ rfc5681, 2009.
[39] J. Mo, R. J. La, V. Anantharam, and J. Walrand, Analysis and comparison of TCP Reno and Vegas, *Proc. IEEE INFOCOM*, vol. 3, pp. 1556–1563, 1999.
[40] S. Floyd, A. Gurtov, and T. Henderson, The NewReno modification to TCP's fast recovery algorithm, https://datatracker.ietf.org/doc/rfc3782/, 2004.
[41] T. Henderson, S. Floyd, A. Gurtov, and Y. Nishida, The NewReno modification to TCP's fast recovery algorithm, https://dl.acm.org/doi/10.17487/RFC6582, 2012.
[42] S. Bohacek, J. P. Hespanha, J. Lee, C. Lim, and K. Obraczka, TCP-PR: TCP for persistent packet reordering, in *Proc. 3rd Int. Conf. Distributed Computing Systems*, Providence, RI, USA, 2003, pp. 222–231.
[43] M. Zhang, B. Karp, S. Floyd, and L. Peterson, RR-TCP: A reordering-robust TCP with DSACK, in *Proc. 11th IEEE Int. Conf. Network Protocols*, Atlanta, GA, USA, 2003, pp. 95–106.
[44] S. U. N. Prottoy, D. Saucez, and W. Dabbous, NUTS: Network updates in real time systems, in *Proc. SOSR 2019*, San Jose, CA, USA, 2019, pp. 160–161.

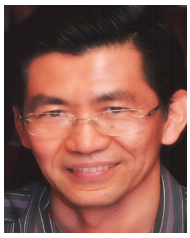
Abdul Basit Dogar is a PhD candidate at Department of Computer Science and Technology, Tsinghua University, Beijing, China. Prior to this, he worked at the Network Architecture Research Division Lab, Tsinghua University. Currently, he serves as a lecturer at Department of Informatics and Systems, University of Management and Technology (UMT), Lahore, Pakistan. From 2007 to 2015, he also worked as a lecturer and program coordinator at Department of Computer Science, Virtual University of Pakistan, where he mentored graduate research students. He later served as a researcher at IRIL, Al-Khawarizmi Institute of Computer Science (KICS), University of Engineering and Technology (UET), Lahore. He is a recipient of the gold medal and the Roll of Honor when persuing the BS degree in computer science, 2005. He received the MS degree in computer science in 2007 from COMSATS University, Pakistan, where he also worked as a research assistant at the Communication Networks Research Centre from 2005 to 2007. His current research interests include SDN, DCN, IoT, TCP variants, IPv4, and IPv6. He has published research papers in journals and conferences.

Yiran Zhang received the PhD degree from Tsinghua University in 2022. She is currently an assistant professor at Computer Science Department, Beijing University of Posts and Telecommunications. Her research interests include traffic management and control, and datacenter networks.

Sami Ullah received the BS degree in computer science from University of Malakand in 2007 and the MS degree in computer science from University of Agriculture, Peshawar, Pakistan, in 2012. He received the PhD degree in computer Science from Ghulam Ishaq Khan (GIK) Institute of Engineering Sciences and Technology, Pakistan in 2022. From 2007 to 2009, he served as a lecturer with University of Malakand, Pakistan. Since 2009, he has been with Shaheed Benazir Bhutto University, Pakistan, as a lecturer. His research interest includes vehicular ad hoc networks, network security, cognitive radio networks, Internet of Things, and Internet of Bodies.

Hisham Alasmary is an assistant professor at King Khalid University. He obtained the PhD degree from Department of Computer Science at University of Central Florida in 2020, and the MSc degree in computer science from The George Washington University, USA, in 2016. His research interests include software security, IoT security and privacy, ML/DL applications in information security, and adversarial machine learning.

Muhammad Waqas received the BSc and MSc degrees in electrical engineering (major in wireless communications) with Department of Electrical Engineering, University of Engineering and Technology, Peshawar, Pakistan in 2005 and 2009, respectively. He pursued the PhD degree with Department of Electronic Engineering at Tsinghua University, Beijing, China since 2019. From 2019 to 2022, he was a research associate at Faculty of Information Technology, Beijing University of Technology, Beijing, China and also affiliated with GIK Institute of Engineering Sciences and Technology, Pakistan. From 2022, he has been an assistant professor at Computer Engineering Department, College of Information Technology, University of Bahrain, Bahrain. In UK, he is currently a senior lecturer at School of Computing and Mathematical Sciences,University of Greenwich, London, UK. He has also been an adjunct senior lecturer at School of Engineering, Edith Cowan University, Australia, since 2021. He has more than 100 research publications in reputed journals and conferences with more than 2400 citations, an h-index of 25 and an i10 index of 63. He is an associate editor of *International Journal of Computing and Digital Systems*. He is also a guest editor of *Applied Sciences*. His current research interests are in the areas of wireless communication, vehicular networks, cybersecurity, and machine learning.



Sheng Chen received the BEng degree from East China Petroleum Institute, Dongying, China, in 1982, and the PhD degree from City University of London, in 1986, both in control engineering. In 2005, he was awarded the higher doctoral degree, Doctor of Sciences (DSc), from University of Southampton, Southampton, UK. From 1986 to 1999, he held research and academic appointments at Universities of Sheffield, Edinburgh and Portsmouth, all in UK. Since 1999, he has been with School of Electronics and Computer Science, University of Southampton, UK, where he holds the post of professor in intelligent systems and signal processing. His research interests include adaptive signal processing, wireless communications, modeling and identification of nonlinear systems, neural network and machine learning, evolutionary computation methods, and optimization.