IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.20XX.DOI

A Survey of Traffic Shaping Technology in Internet of Things

HAO FU[1], MUYI SUN[1], BINGYU HE[1], JINHUA LI[1] AND XIAOMIN ZHU.[2]
[1]Qilu University of Technology (Shandong Academy of Sciences), Jinan 250306, China
[2]Shandong Institute of Bigdata, Jinan 250306, China

Corresponding author: Hao Fu (mail:fuhao15725150062@163.com).

This work was supported by the Shandong Provincial Natural Science Foundation of China under Grant No. ZR2020MF057, the National Natural Science Foundation of China under Grant No. 92067108, and the Beijing Nova Program of Science and Technology under grant Z191100001119113.

ABSTRACT In the era of the Internet of Everything, the burst of traffic will bring many problems. Traffic shaping, as a common means to limit burst traffic rate, achieves "peak and valley reduction" to smooth the output rate, avoid network congestion, guarantee the quality of service (QoS) and improve the overall network transmission efficiency. This paper first explains the basic concept of traffic shaping, the related algorithms in traffic shaping, and compares the differences between traffic shaping and traffic policing. It then describes the research on traffic shaping in software-defined networks and the results from the combination of the two. This is followed by an introduction to relatively new technology for industry, time-sensitive networks, and an exploration of the functions of various types of traffic shapers in time-sensitive networks. This is followed by a description of case studies of traffic shaping in IoT scenarios and an overview of these traffic shaping schemes, summarizing the usefulness of the shaping schemes. Finally, the future of traffic shaping is explored.

INDEX TERMS Traffic shaping, Internet of Things, Industrial Internet, Sensor network

I. INTRODUCTION

With the rapid development of the Internet and the proliferation of network users, user-oriented network applications occupy most of the network resources, which seriously consumes bandwidth and affects the utilization of the network. Today, a large number of P2P (peer-to-peer) downloads, BT downloads, and other applications consume a large number of network resources, which can easily cause network congestion and affect the overall network quality [1]. In the face of these preemptive download applications, the traditional Internet best-effort service cannot guarantee the transmission of critical services. For the industrial Internet, the data collected by sensors can be divided into real-time data and non-real-time data according to business requirements; the traditional network cannot guarantee the degree of data priority, which can easily cause data to compete for network resources, thus triggering network congestion and affecting the transmission of real-time data.

In the actual Internet environment, users' demand for network resources such as data transmission time, transmission rate, and bandwidth-delay varies greatly and fluctuates greatly in a particular case. At present, network bandwidth resources are very scarce and cannot be solved for a long time in the future; and the limited network resources cannot meet the above-mentioned user services with sudden changes in demand for bandwidth resources. Therefore, it is inevitable that the needs of users will not be met by network resources. Therefore, traffic shaping is needed to manage the network traffic at the output to avoid wider network fluctuations, thus alleviating network congestion and improving the quality of service and overall performance of the network. [2].

Traffic shaping emerged to solve the congestion problem caused by bursts of network traffic to the network and to smooth the peaks and valleys of data transmission [3]. Traffic shaping is an important technique used to provide QoS guarantees for computer networks. Traffic shaping shapes the traffic that does not conform to the settings or irregularities to match the bandwidth of the upstream devices when transmitting data to the downstream devices, avoiding interface bottlenecks that cause message loss and congestion. The idea of traffic shaping is to classify the incoming packets and send those that do not need to be shaped directly, while those that need to be shaped enter the queue and wait for some shaping algorithm to manage the output, thus achieving control over the traffic rate. The typical role of traffic shaping is used to limit the traffic rate at the output of the network so that such messages are sent outward at a more uniform rate [4].

The network traffic control through traffic shaping technology can limit some non-critical service traffic that maliciously seizes bandwidth resources in the actual network environment, guarantees the transmission of critical service traffic, ensure the fairness of bandwidth usage so the entire network does not experience congestion, and improves the QoS of the network. From the economic aspect, the use of network traffic shaping technology can reduce the pressure of network expansion, thus saving the cost of network deployment and operation cost [5] [6]. Therefore, network traffic shaping has important research values.

This paper provides a review of common traffic restriction techniques in networks, focusing on traffic shaping techniques and research descriptions related to industrial and vehicular IoT networks. There are several existing surveys (e.g., [7] [8]) on the traffic shaping in industrial IoT. The existing surveys are carried out mainly about time-sensitive networks. Specifically, they focus on the development of TSNs, the principles and principles of the four shapers in TSNs.In contrast, this paper first surveys the traditional traffic shaping techniques and describes the relevant traffic shaping algorithms. Then the study of two techniques, software-defined networks and time-sensitive networks, in traffic shaping is explained. The relevant research into these two techniques for IoT aspects is presented, as well as the ideas that can be provided for industrial IoT aspects. This is followed by a description of current traffic shaping applications in the IoT, showing the uses of traffic shaping in the IoT.

We organize the remainder of this paper as follows. In Section II, we review the knowledge related to traffic shaping, and in Section III, we compare traffic shaping and traffic policing, and then describe several common algorithms and variants of algorithms. In Section IV, the research on traffic shaping in SDN frameworks and TSN standards is presented. In Section V, we describe the application of traffic shaping in different IoT environments.

II. BACKGROUND KNOWLEDGE

A. BASIC OVERVIEW

Traffic is the flow of data that can be uniquely identified by the transmission protocol between the sender and receiver, and is divided into message flow and packet flow in computer networks. As the most basic traffic unit in computer networks, each message consists of a message header and a data field, respectively. For transport layer messages, the message header contains some basic information required by the network, including a list of information called the five elements: (source IP address, destination IP address, source port, destination port, and protocol) [9].

The Internet has a wide range of applications and the data generated is variable, and the service index specified by QoS cannot provide a guarantee for traffic that exceeds the specified rate. Through traffic shaping management input downstream equipment transmission rate, reasonable shaping and control group inflow, so that the rate irregular (such as burst data flow) data transmission according to the transmission set rate, change the burst traffic output characteristics, fill the valley rate and cut the peak rate, smooth the data output rate to reduce the rate of jitter, circumvent network congestion, so that bandwidth resources are fully used.

B. QOS AND TRAFFIC SHAPING

Currently, bandwidth resources are relatively scarce and increasing bandwidth can be a dramatic increase in deployment costs, and QoS guarantees can bring some optimization to the limited bandwidth resources. QoS hopes to provide better service assurance for network communications by using basic technical means and is a technique commonly used to avoid problems such as congestion and network delays. QoS is a series of service requests to be met by the network when transmitting data streams, which can be specifically quantified as performance indicators such as bandwidth, latency, jitter, loss rate, and throughput. As a common technique in QoS, traffic shaping smoothes outbursts of network traffic by limiting the rate of traffic at the output egress, keeping data transmission at a stable rate and avoiding network congestion. Compared with other rate limiting techniques in QoS, traffic shaping achieves fewer dropped messages and makes full use of bandwidth resources, so traffic shaping can be a good way to improve the QoS of the network. In practical applications, traffic shaping can be used in conjunction with different models in QoS [1]. It is illustrated in the literature [10] that networks with traffic shaping and QoS have a higher quality than before. Table 1 shows the service model for QoS.

TABLE 1. QoS three service models

	Advantages	Disadvantages
Best-Effort service	The model is simple and is implemented via a FIFO queue.	No distinction between message types and no service guarantees.
Integrated service	Provide end-to-end QoS services with guaranteed bandwidth and latency.	Poor scalability and low utilization of bandwidth resources.
Differentiated service	Good scalability, low resource consumption, different data differentiation to meet different QoS requirements.	The mechanism is complex and there is no absolute fairness.

C. IRREGULAR TRAFFIC AND TRAFFIC SHAPING

The main purpose of traffic shaping techniques is to solve The main purpose of traffic shaping techniques is to solve the network congestion problem caused by bandwidth bottlenecks generated by various irregular traffic in the network such as burst traffic, or low bandwidth utilization caused by rates below a specified value. Changes to the network are impossible to avoid completely. For the traditional Internet, the rapid development of network technology has led to the emergence of a variety of network applications, a variety of applications will generate a variety of network traffic with different characteristics, most of these flows are not regular, and the traffic transmitted in the network may offset each other to make the network smooth, or may be superimposed on each other to produce greater network fluctuations

Secondly, another factor that causes network traffic bursts is the special data flow-variable bit rate (VBR) type [11]. The output rate of VBR traffic shows fluctuating variation, and VBR traffic grabs a lot of resources at the time when bandwidth is free and has the characteristic of an intermittent burst. In general, the transmission rate of VBR traffic is much higher than the average rate. Network nodes often have to reserve bandwidth resources much higher than the actual transmission rate according to the peak rate of VBR traffic to meet the QoS requirements of VBR communication. This way of allocating bandwidth greatly wastes network bandwidth resources and increases the cost of network deployment, and there is still a large rate of jitter in VBR traffic, which is an important reason for network congestion [9].

D. QUEUE SCHEDULING

Queue scheduling is one of the core technologies for traffic control and is used to solve the problems that arise when multiple service flows are sharing network resources. The traditional scheduling algorithms are, 1) First-in, first-out (FIFO) algorithm: arriving packets are inserted into the queue according to time, and then packets are removed from the front of the queue. 2) Priority queuing algorithm (PQ): arriving packets are assigned to different queues according to priority rules, each queue has a different priority, and then packets are selected for transmission from the head of the non-empty queue with the highest priority. 3) Fair queueing (FQ) algorithm: It uses polling to access all queues in the buffer and sends the first packet of each queue. It ensures that each service flow shares the resources fairly. 4) Weighted fair queuing (WFQ) algorithm: It is a flow-based queuing algorithm that assigns a weight to each service flow according to its different priorities, and the weight determines the bandwidth of the output port occupied by the service flow. 5) Class-based queuing (CBQ) algorithm: The traffic shares the bandwidth in a balanced way after being grouped by class. The classification can be based on different parameters.

III. TRAFFIC SHAPING AND TRAFFIC POLICING

A. TRAFFIC POLICING

In the network, when data is transmitted from high-speed links to low-speed links, the bandwidth will bottleneck at the low-speed link interface because of the abrupt change in transmission rate, which leads to serious data shortage; especially for data with low latency requirements. Traffic Policing is typically applied to solve the emergence of the above problem. Traffic Policing works on the input side of the data, by monitoring the specifications of certain traffic entering the network and limiting it to a certain range, thus protecting the network resources and the operator's interests.

Traffic policing usually uses the committed access rate (CAR) to limit the traffic and bursts of data entering or exiting a particular connection to a network.CAR uses token buckets for traffic control. Figure 1 shows the basic process of traffic control using CAR: first, the incoming messages are classified according to the set rules. The messages that do not meet the rules are sent directly, and the messages that meet the rules are matched with the tokens in the token bucket, and the messages are sent when the number of tokens is sufficient for the messages, and the messages are discarded if the number is insufficient.

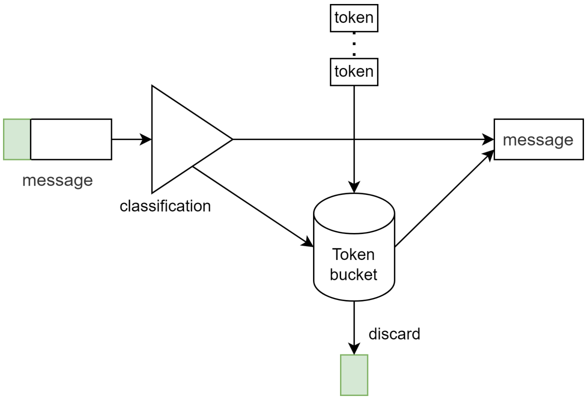

FIGURE 1. Traffic policing

B. TRAFFIC SHAPING

Traffic shaping enables such messages to be sent outward at a more even rate by limiting the traffic and bursts for a particular connection flowing out of a given network. Traffic shaping is typically accomplished using buffers and token buckets. When messages are sent too fast, they are first cached in the buffers and then sent evenly under the control of token buckets. As shown in Figure 2 generic traffic shaping (GTS) is shaping irregular traffic. It caches the portion of the stream that exceeds the traffic specification and then sends out the cached packets at the appropriate time, thus smoothing and moderating the output traffic so that the output rate matches that of the downstream network devices.

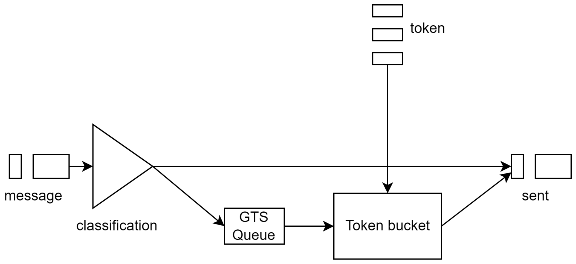

FIGURE 2. Traffic shaping

C. TRAFFIC POLICING AND TRAFFIC SHAPING COMPARISON

Traffic Policing Propagation Burst. When the traffic rate reaches the configured maximum rate, the excess traffic is discarded. The result is an output rate that appears as a sawtooth with crests and slots. Contrary to policing, Traffic Shaping holds excess packets in the queue and then schedules the excess packets for later data transmission. The result of traffic shaping is a smooth packet output rate [12]. A comparison of the flow output is shown in Figure 3.

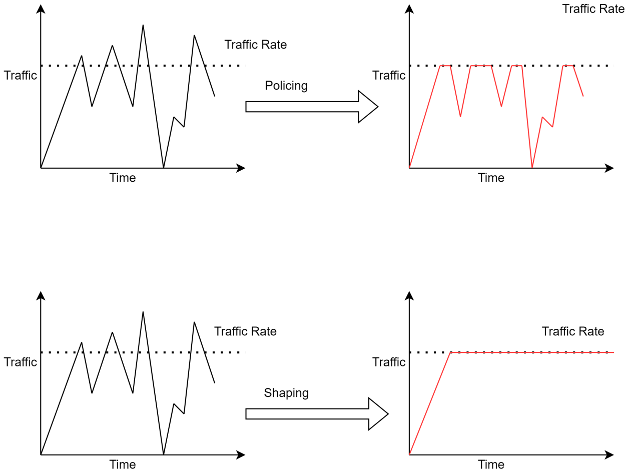

FIGURE 3. Comparison of traffic output rates(by cisco document 19645)

TABLE 2. COMPARISON OF TRAFFIC SHAPING AND TRAFFIC POLICING

	Traffic Shaping	Traffic Policing
Purpose	Buffer and queue excess packets at the committed rate.	Discard packets that exceed the committed rate without caching.
Applicable on Inbound	No	Yes
Applicable on Outbound	Yes	Yes
Burst	Control bursts, use leaky buckets to delay flow and smooth output rates.	Propagation burst. No smoothing.
Advantage	Cache excess packets with low probability of packet loss and avoid retransmission.	Control output rate through packet loss. Avoid delays due to queuing.
Disadvantage	May cause delays, especially for deep queues.	Discard redundant packets and limit TCP window size.

D. ALGORITHMS IN TRAFFIC SHAPING

1) LEAKY BUCKET ALGORITHM

The leaky bucket algorithm [9] [13] is a common algorithm used for traffic shaping and rate limiting in networks, the main purpose of which is to limit the transmission rate of data in the network and to achieve smooth burst traffic. The leaky bucket rate-limiting algorithm turns bursts of traffic with irregular input rates into a regular stream of equal-gap data. The rate of the output traffic is made fixed regardless of the size of the arriving traffic. The leaky bucket algorithm delivers the irregular data stream to a packet queue controlled by a leaky bucket queue controller. The queue receives the packets to be forwarded, waits for scheduling to obtain the forwarding timing, and the packets in the queue are fed into the leaky bucket. The packet input rate into the bucket is compared with the input rate at the bottom of the bucket, and the corresponding forwarding action is performed. 1) The packet arrival rate is lower than the output rate of the bucket, and the bucket is invalid. 2) The arrival rate is higher than the output rate of the bucket, and the capacity set for the bucket can meet the demand of burst traffic is further analyzed; if the capacity of the bucket is exceeded, this part of the data is discarded, otherwise, the data smaller than the output rate is not If the capacity of the bucket is exceeded, the data is discarded, otherwise, the data less than the output rate is sent without delay. Using the leaky bucket algorithm, the burst traffic can be sent to the network smoothly after the action of the leaky bucket buffer. The advantage of the leaky bucket algorithm is that it limits the output rate and ensures that the packet rate sent to the network is not higher than the rate required by the network.

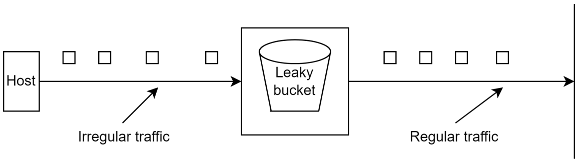

FIGURE 4. Leaky bucket algorithm

The purpose of traffic shaping is to smooth out the output traffic and thus avoid network congestion. However, when the incoming rate reaches a peak, network congestion is likely to occur due to limited bandwidth resources. This leads to many non-conforming frames. To address this problem, the literature [14] gives the scheme to set a receding time in the leakage bucket of the traffic shaping mechanism. The schemes given in the literature [14] for calculating the backoff time are pseudorandom backoff(PB) time, exponential backoff(EB) time, and random backoff(RB) time. Experimental results show that EB and RB perform better than PB.

2) TOKEN BUCKET ALGORITHM

The token bucket algorithm is one of the most frequently used algorithms in traffic shaping and rate limiting [15] [16] [17]. Unlike the leaky bucket algorithm, the token bucket algorithm can control the amount of data transmitted to the network and respond quickly to bursts of data, while avoiding data loss.The size of the token bucket is fixed to keep generating tokens at a specified constant rate. If a token is not used, or if the rate of token generation is greater than the rate of use, tokens will be added continuously until the bucket is filled. The tokens generated after that will overflow from the bucket. The number of tokens in the token bucket will never exceed the bucket's capacity. The transmission of grouped data requires the consumption of tokens, and the number of tokens consumed varies for different packet sizes. The average rate of data flowing out of the queue is controlled by managing the rate at which tokens are generated. The token bucket algorithm uses a token bucket to manage when the queue controller sends traffic, allowing bursts of traffic while keeping the transmission rate sent smoothly. If a token exists in the token bucket, the traffic is allowed to be sent; otherwise, the traffic is not allowed to be sent. Thus, if the token bucket has a sufficient number of tokens and the burst limit is properly configured, the traffic can be transmitted at a steady peak rate all the time.

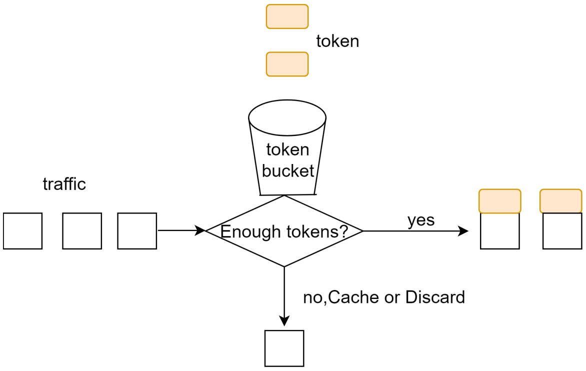

FIGURE 5. Token bucket algorithm

3) TWO TOKEN BUCKET ALGORITHM

To enable token buckets to work better in environments with large variations in burst size and burst rate, two token bucket algorithms are defined in the IETF's RFC document-the single-rate tri-color marking algorithm and the two-rate tri-color marking algorithm, both of which evaluate by marking messages with red, yellow, and green colors [15]. Based on the color of the marker on the message, QoS determines the discard priority of the message. Both the single-rate tri-color marking algorithm and the dual-rate tri-color marking algorithm can cope with bursty data, with the former focusing on bursts in message size and the latter on bursts in rate.

Single rate three color market (SrTCM) algorithm, evaluated by the following three parameters: committed access rate (CIR), the rate at which the tokens in the token bucket are filled; committed burst size (CBS), the capacity of the token bucket, the maximum size that can allow burst traffic; and excess burst size (EBS). The SrTCM algorithm consists of two token buckets. One of them is bucket C. Tc is the number of tokens in bucket C. CBS is the bucket depth, which represents the capacity of bucket C. The other bucket is bucket E. Te is the number of tokens in bucket E, and EBS is the bucket depth, which represents the capacity of bucket E. The other bucket is bucket E, and Te is the number of tokens in bucket E. And EBS capacity is larger than CBS. start with both buckets full of tokens. The tokens are filled at the set CIR rate, first adding tokens to bucket C and then adding tokens to bucket E when bucket C is full. When both buckets are full, the newly generated tokens are discarded. In this algorithm, the packets are color-marked through markers. When the message length B is less than Tc, the message is marked in green; when Te>B>Tc, the message is marked in yellow; when B>Te, the message is marked in red. The color-marked messages are allowed to flow into the network, and the packets are discarded according to the red, yellow and green levels when congestion occurs in the network.

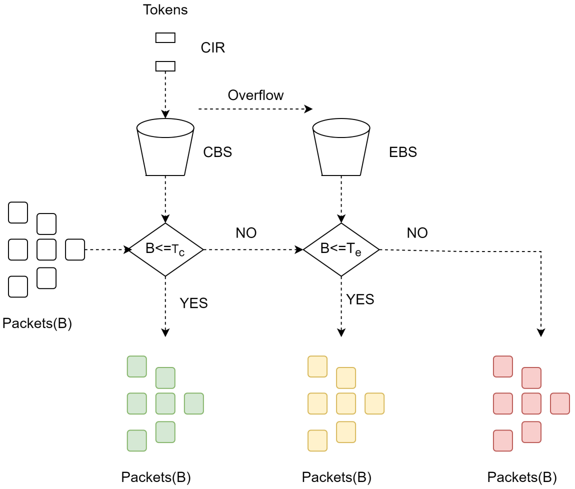

FIGURE 6. SrTCM

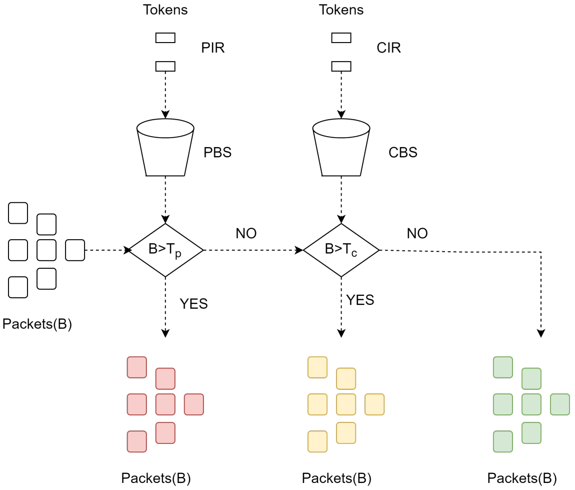

FIGURE 7. TrTCM

The two rate three color market (TrTCM) algorithm evaluates based on the following four parameters: CIR, CBS, peak information rate (PIR), and peak burst size (PBS). The first two parameters are the same as the SrTCM algorithm, while the PIR parameter is only present on the switch and its value is greater than or equal to the set CIR value. If the PIR is greater than the CIR, the rate is limited to something in between. The two buckets are the same as the single-rate tricolor token algorithm, except that the two token buckets of the dual-rate tricolor token algorithm have different token generation rates, bucket C and bucket P. Bucket C has a capacity of CBS, a token generation rate of CIR, and Tc represents the number of tokens in the bucket; bucket P has a capacity of PBS, a token generation rate of PIR, and Tp represents the number of tokens in the bucket. The initial states are both full. The rate of the message is marked as red when it is greater than the PIR, green when it is less than the CIR, and yellow before both. Marked packets are allowed to flow into the network, and packets are discarded according to the red, yellow, and green levels in case of network congestion.

TABLE 3. COMPARISON OF SrTCM AND TrTCM

	SrTCM	TrTCM
The structure of the barrel	Single barrel or double barrel.	Double bucket.
Token addition	Simple, constant rate (CIR) add token for both single and dual buckets.	complex, the two buckets add tokens at different rates (CIR and PIR, respectively).
Message processing process	Simple, the message length is compared with the number of tokens in the bucket, and if there are enough tokens, it is forwarded, otherwise it is discarded.	Complex, the message rate is compared with CIR and PIR, <CIR, the message is green; >PIR, the message is red; between the two messages are yellow.
Merit	Focus on bursts in message size and borrow tokens from the EBS bucket when the message length is too long.	Focus on rate bursts, and borrow tokens from the PBS bucket when the message rate is too high.

The comparison of single and dual rate tri-color tagging algorithms is shown in Table 4. compared to the dual-rate tri-color tagging algorithm, the single-rate algorithm tri-color tagging algorithm is easier to implement and has become a more common traffic tagging method in the industry today. There is also some variability in the performance results exhibited using different implementations. A reasonable debit method can be improved in the burst traffic processing performance, data forwarding smoothing degree, mixed packet forwarding performance of small and large packets, packet loss rate, and other performance indicators. However, there exists certain burst traffic with a very large rate, and its performance cannot be compared with the dual-rate tri-color marking algorithm. The two algorithms have mutual advantages and disadvantages and cannot be simply replaced; the appropriate marking algorithm should be selected according to the actual network environment and traffic characteristics.

4) RESEARCH ON TOKEN BUCKET ALGORITHM

The traffic shaping algorithm has been developed quite maturely and the token bucket algorithm is widely used. To provide a better quality of service and adapt to more complex network environments, many researchers have made a lot of efforts on shaping algorithms. Among them, a multi-level token redistribution traffic shaper is proposed in the literature [9]. The traditional single token bucket can only shape traffic for a single service level, and the scheme designs a tree-shaped hierarchical traffic shaper according to the three major service levels of Diff-Serv networks. The literature [18] gives an adaptive traffic shaper to solve the problem of network congestion caused by traffic bursts caused by rate mismatch between optical and wireless domains and different service classifications in the wireless-optical broadband access network (WOBAN), which sets a token bucket (TB) for each service according to different service classes, while dynamically adjusts the token generation rate r of the TB and the number of tokens b in the bucket; the results show that the scheme can reduce the pressure on network convergence devices while improving QoS. literature [19] gives a multi-token bucket algorithm supporting multi-queue shaping for the congestion problem that exists in packet switching systems of high-speed data networks. Dynamic support for multi-service token bucket traffic shaping algorithm is proposed in the literature [20]. The multi-token bucket algorithm proposed in the above scheme can reduce packet loss and congestion problems well in the face of complex network environments and different levels of services. To solve the bottleneck problem caused at the edge gateway due to the difference in bandwidth between enterprise intranet and WAN, a smart token bucket filter (STBF) mechanism is proposed in the literature [21] to effectively solve the problem of unfair traffic competition caused by a bottleneck in a limited bandwidth.

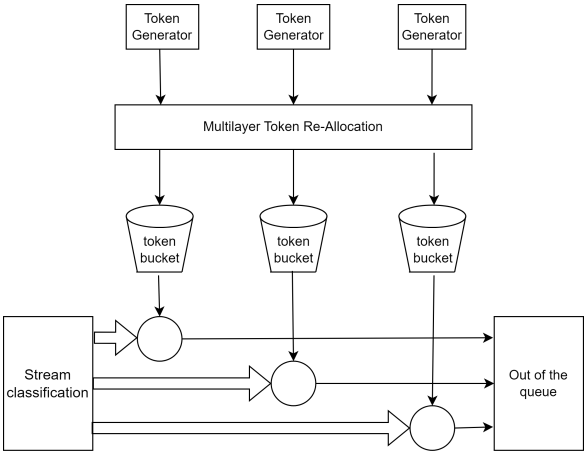

FIGURE 8. Multilayer Token Re-Allocation [9]

In this section, we first describe traffic policing and traffic shaping separately and compare these two rate limiting techniques to highlight the advantages of traffic shaping. This is followed by a description of two common algorithms-the leaky bucket algorithm and the token bucket algorithm; finally, several improved algorithms are reviewed for the token bucket algorithm, which are better able to adapt to complex network environments.

IV. RESEARCH ON TRAFFIC SHAPING IN SDN AND TSN

Industrial Internet of Things, industrial control networks, in-vehicle networks, and other IoT directions are developing rapidly, and traditional technologies can hardly meet the needs of data transmission in terms of determinism, real-time, and reliability; and have average performance in terms of scalability. Therefore, two different technologies, SDN and TSN have emerged to address the shortcomings of traditional technologies. SDN has the concept of network programmability, which performs well in terms of flexibility and scalability, and the programmability in SDN can provide development extensions for industrial needs with centralized control functions. TSN is designed for low latency and high availability to ensure deterministic data transmission [22].

A. TRAFFIC SHAPING COMBINED WITH SDN

The software-defined networking (SDN) concept was originally proposed by Nicira Networks based on its early developments at UCB, Stanford, CMU, and Princeton [23]. The goal of SDN is to provide open, user-controlled management of the forwarding hardware in the network, which brings applications and network services, and devices closer together, while it separates the network device control and data It separates the control and data planes of the network devices, breaking the tight coupling between the control and data planes in the traditional network architecture, and realizing the controllability, security, and economy of network resources. Currently, OpenFlow is the most popular SDN protocol/standard with a set of design specifications [24].

As the number of tenants in the cloud grows and control channels intersect under the same physical network creating interference, virtualized SDNs have a crisis in control channel fairness. The computing resources in the cloud data center should be guaranteed to be relatively independent for each tenant. For network isolation, cloud data centers use network virtualization (NV). However, NV is not user-programmable; for this, SDN data plane and control plane separation is used in conjunction with NV, enabling each user to build and program their virtual network simply. Since many control channels are co-located in the same physical network, the throughput of each control channel has a significant impact on the performance of SDN-NV. As the number of tenants increases, the time to install the flow rule forwarding setup also increases correspondingly. This is because the control channels sharing the physical network interfere with each other and have serious fairness issues. Also, the data transmission in the data plane is delayed in sending because of the increase in forwarding setup time, which indicates that the throughput of the control channel can seriously affect the performance of SDN-NV.

To address the above mentioned fairness issues, an adaptive control channel shaping scheme Sincon (control channel adaptive traffic shaping scheme) for SDN-NV is proposed in the literature [25]. To achieve fairness in the control channel, the scheme uses traffic shaping to constrain the maximum throughput of the control channel, and to prevent data loss, a buffer queue is added to the control channel to store temporary data. On the other hand, Sincon implements an automatic adaptation function to adjust the throughput limit. When the usage is close to the throughput limit, the throughput limit is reasonably increased. The results show that Sincon implements traffic shaping for the control channel, and the fairness of the control channel throughput is improved.

The inequity problem also appears in video technology, the rapid increase in mobile video traffic, and most of the video providers are now using HTTP adaptive streaming technology; however, using this protocol currently suffers from link instability, bandwidth inefficiency, and unfair competition among multiple applications. A dynamic SDN-based traffic shaping technique (DASH-SDN) is given in the literature [26] to address these problems. The video stream has ON-OFF (data transmission has interval, ON refers to the interval of data transmission, and OFF refers to the interval in idle state and shows periodic characteristics.) Transmission mode characteristics, this solution uses the SDN controller to achieve network monitoring, traffic inspection, and bandwidth management to improve the quality of service of video in many HTTP video player long idle (OFF) moments, improve bandwidth resource utilization and reduce the power consumption of mobile video playback.

In terms of bandwidth allocation, the SDN architecture is used to converge shaping techniques to meet the QoS requirements in the network, in response to the excellent scalability and operation of SDN. In the literature [27], a new SDN model is proposed and two traffic shaping algorithms ("packet tagging, queuing and forwarding to queue" and "bandwidth allocation") are introduced to implement weighted fair queueing (WFQ) techniques to reduce congestion and smooth service flow. In this paper, WFQ is used to assign queues to data streams and assign weights to each stream through shaping techniques to solve the single queue limitation of the FIFO algorithm, while using buffer management to achieve whether packets enter the queue or not; the above methods can mitigate latency and congestion problems, meet the demand for real-time services, and achieve the purpose of improving QoS. Cloud games can solve the amusement demand of users with poor device performance; the computing operation of the game is completed in the data center, and the data center sends the processing results down to the users, so guaranteeing the user experience has high requirements for bandwidth. A traffic management and shaping approach using SDN architecture is proposed in the literature [28] to solve the bandwidth allocation problem in cloud gaming data center networks. The scheme proposed in the literature uses the SDN's controller to fairly optimize the flow distribution in the data center network path, which improves the utilization of the remaining bandwidth and improves the user's gaming experience while guaranteeing the lowest latency.

The above-mentioned solutions can bring some ideas to the design of the industrial field, which also faces the problem of bandwidth resource allocation and unfair resource allocation in the transmission of industrial data. Because of the advantages of flexibility and scalability, SDN has attracted much attention in the industrial field. Meanwhile, the data plane and control plane of SDN are separated, and the control plane in SDN can be used to achieve comprehensive and centralized control of industrial networks.

B. TRAFFIC SHAPING IN TSN

The continuous development of the times and the update of technology have produced many new applications such as 5G, autonomous driving, unmanned delivery vehicles, smart factories, etc. Traditional technologies have great differentiation and problems in compatibility. The aforementioned scenarios have requirements for low latency, low jitter, no congestion, no packet loss, and high robustness in data transmission, which are not well met by traditional technologies [7] [29]. At the same time, since traditional Ethernet is a shared transmission medium [30], data from different sources must be queued when passing through the same switch or routing node, and congestion occurs when traffic changes abruptly making the queuing time ambiguous and deterministic less than guaranteed. Therefore, the IEEE Time Sensitive Networking Group has been developing a set of standards to lay the foundation for deterministic Ethernet.TSN has received much attention in the industry as a new approach to address deterministic and real-time data transmission.

To provide good real-time and deterministic guarantees to the network, the IEEE 802.1TSN task group has developed a new substandard based on specific requirements. Its purpose is to establish a generic time-sensitive mechanism for Ethernet protocols to ensure the time determinism of network data transmission and to meet the vehicle and industrial requirements for real-time performance and high reliability [31]. Traffic shaping is a key technology for TSN and plays an important role in facing the real-time and deterministic problems of time-sensitive flow transmission [32]. The current mainstream shaping techniques for TSN are credit-based shaper (CBS), time awareness shaper (TAS), cyclic queuing forwarding (CQF), and frame preemption [8].

1) Credit-based Shaper:CBS was originally used for buffering information streams during real-time transmission of audio and video signals over Ethernet to address packet loss due to bursty media stream congestion, ensuring that high-priority data are prioritized and that "credit values" are used to schedule data transmission so that lower-priority service data has the opportunity to be output. However, CBS has uncertainty about the delay caused by buffer queue queuing.
2) Time Awareness Shaper:The purpose of TAS is to better ensure the real-time transmission of time-sensitive flows. By setting up different channels at the switching node according to different traffic categories, the corresponding channels are opened within a specified period of time and only the corresponding channels are allowed to transmit time-sensitive flows. This can prevent the mutual interference of traffic from the time level and ensure the transmission timeliness of traffic.
3) Cyclic Queuing Forwarding:CQF features frame synchronization inbound and outbound, CQF allows LAN bridges to synchronize with frame transmissions in a single cycle to obtain zero blocking packet loss as well as bounded latency and the ability to exist independently of the network topology. CQF is used to solve the problem of deterministic real-time signaling.
4) Frame Preemption:The frame preemption mechanism allows the user to interrupt the transmission of lower priority data one or more times with higher priority data, and then resume the transmission of lower priority data after the higher priority data has been transmitted. Before performing frame preemption, data frames are classified into high-speed frames and low-speed frames according to the degree of latency sensitivity. The delay-sensitive frames are called high-speed frames and the rest are called low-speed frames, where the low-speed frames are also called preemptable frames. Frame preemption can reduce the queuing blocking time of high-speed frames, thus effectively reducing the latency of high-speed frames. But at the cost of increasing the latency of low-speed frames. Uncertainty in the timing of low-speed frame slicing can lead to fluctuations in the time when high-speed frames are blocked, thus introducing some delay jitter.

TABLE 4. COMPARISON OF TSN TRAFFIC SHAPING ALGORITHMS

Traffic shaper	Time synchronization	Characteristic	Existing problem
CBS	Non-essential	The transmission of low priority data is also available.	Buffering queues cause uncertainty with large average delay.
TAS	Need	The time-sensitive traffic is not disturbed by dividing the time channel.	The utilization rate of the communication bandwidth has decreased.
CQF	Need	The caching and sending of the data traffic alternate in the two queues to ensure the certainty of the delay.	Data buffer occupies two queues.
Frame preemption	Non-essential	Allow high-priority frames to interrupt the transmission of medium-to-low-priority frames.	Section timing was uncertain, resulting in time-delay jitter.

In industrial networks, the number of sensors in a factory is large, and the data traffic transmitted is large, which has a certain demand for bandwidth. At the same time, industrial traffic is divided into real-time data flow and non-real-time data flow, and the two types of traffic are differentiated and have different requirements for transmission. Real-time traffic has a periodic pattern, and the superposition of different moments of traffic has an impact on the peak rate. Traffic shaping of TSN works at the data link layer and can be shaped and scheduled by identifying the priority and timestamp characteristics of the traffic to ensure the real-time and deterministic nature of real-time data.

In the context of IEEE TSN, three new traffic shaping mechanisms, called time-aware shaper (TAS), burst-limiting shaper (BLS), and peristaltic shaper (PS), are considered in the literature [33]. End-to-end delay experiments were conducted for these three shapers in the worst-case environment, and the results showed that TAS was the shaper that provided the best low delay and low jitter performance. In addition in real industrial environments, some urgent but non-periodic real-time service traffic may occur and TAS cannot guarantee the transmission of such traffic, and recalculating the schedule for traffic would be extremely costly. Therefore, a novel enhancement to TAS referred to as eTAS, is proposed in the literature [34]. This approach defines a new immediate forwarding scheduling rule for the above mentioned emergencies without affecting the non-emergency real-time services and guarantees real-time performance.

The TAS algorithm in TSN can provide a real-time transmission guarantee for traffic, which is well suited for industrial Internet with periodic traffic characteristics, but no guarantee can be made for transmission determinism. Therefore, TSN investigates an algorithm called CQF to achieve transmission determinism. The problem of configuring a CQF-based network is defined in the literature [35]. The authors of the literature developed a constraint programming (CP) formulation for multi-CQF, and a simulated annealing (SA)-based metaheuristic solution. The CQF variants were evaluated for their advantages and disadvantages and also compared with TAS. However, most of the algorithms are difficult to implement in large-scale deterministic networks (LDNs). A large-scale deterministic network flow shaping (LFS) mechanism is proposed in the literature [36] for guaranteeing deterministic worst-case delays and zero packet loss for time-sensitive flows in LDNs. The results show that the method effectively improves bandwidth resource utilization and also has good schedulability.

Currently, different traffic shaping algorithms in TSNs are usually implemented as separate parts. In the literature [37] a traffic shaping engine (TSE) intended for the next generation of time sensitive networking (TSN) compliant network/system on chip (NoC/SoC) devices, targeting especially the automotive industry, is proposed as a generic interface management shaper. The generic hardware architecture proposed in the literature allows the selection of arbitrary shapers and even the combination of shapers in some cases. Flexible processing is achieved by this generic structure, abstracting the management of the traffic shaping process of the TSN. The scheme achieves minimizing latency while maximizing throughput and QoS.

A comparison of the performance of four traffic shapers in TSN is done in the literature [38], and the effects produced by different combinations of shapers are also explored. The literature [39] mentions that to help practitioners how to choose between stream-based TAS, class-based TAS and frame preemption, the above three mechanisms are compared. The results illustrate that stream-based scheduling is well-suited for isochronous traffic but cannot cope with the loosened talker model of cyclic traffic. Class-based scheduling is easy to plan but due to its large class-window, the bandwidth efficiency is low. From the above literature, it can be concluded that different shaping algorithms have different application scenarios. The TAS algorithm is used for scenarios with periodicity and strict requirements on transmission delay and jitter. For scenarios with periodic requirements but with maximum delay requirements, the CQF algorithm is suitable to be chosen. For bursty ones with high delay requirements the frame grabbing algorithm is selected. In practical applications, complex scenes using one algorithm is unable to cope with them, so a reasonable combination of multiple algorithms is chosen to not only better adapt to complex scenes but also to improve real-time transmission performance. In the literature [40] two shapers, credit-based shaper and time awareness shaper are explored and the question of how ready TSN and OPC UA are and what the requirements of the end devices and applications supporting these two new technologies for the implementation of TSN for industrial automation is explored.

TSN works at the data link layer and can be deployed on any Ethernet according to the requirements.TSN guarantees real-time and priority data by adding precise timestamps to the Ethernet frames. The TAS in TSN requires a high degree of clock accuracy, and any timing deviation can cause errors in the network. The emerging asynchronous traffic shaping (ATS) ensures determinism for real-time applications while eliminating the dependence on synchronous communication [41] [42]. The performance of TAS and ATS is compared in the literature [43], where the latency and packet loss of both shapers are evaluated in an industrial network environment. The results show that ATS has good performance in the face of sporadic traffic, while it does not require time synchronization of network nodes; while TAS can reach the bounded delay limit only with a suitable configuration. Bounds on the end-to-end worst-case delay and node backlog size are calculated for each class of deterministic networks for CBS and ATS in the literature [44]. Similarly in the literature [45] CBS and ATS are evaluated. The results in the paper show that ATS provides better real-time guarantees than CBS for non-periodic traffic in a highly loaded network scenario.

In this section, two relatively new technologies-SDN and TSN are described. Traffic shaping schemes using their frameworks are introduced in SDN networks to solve channel inequities, bandwidth allocation problems, etc.; at the same time, the SDN framework provides design ideas for industrial networks and in-vehicle networks, etc. due to its flexibility, programmability, scalability, and separation of control and data. In TSN standard, it can well meet the real-time and deterministic needs of the industry, which is a breakthrough of traditional network technology, while breaking the longstanding device compatibility problem. In short, two relatively novel technologies can assure the development of industrial IoT.

V. APPLICATION OF TRAFFIC SHAPING IN IOT

Traffic shaping is used in real-world network environments to limit burst traffic rates, relieve network congestion, conserve resources, improve bandwidth resource utilization and enhance overall network quality. In addition to these common functions, traffic shaping is also used in home IoT environments to protect privacy and prevent outside attacks.

A. APPLICATION OF TRAFFIC SHAPING IN SENSOR NETWORK

Sensor network (SN) can monitor, sense, and collect information on various environments or monitoring objects in real-time, and has a wide range of applications in defense and military, environmental monitoring, traffic management, factory automation, and many other areas. According to the way sensor nodes report information, the working models of sensor networks can be divided into four types: on-demand query model, peri- odic report model, event-driven model, and mixed model. The mixed model is a combination of the periodic reporting model and the event-driven model.

Wireless multimedia sensor networks (WMSN) consist of hybrid distributed sensor nodes and multimedia sensor nodes that are combined with miniature cameras to retrieve, store and process real-time multimedia data. In WMSN, bandwidth resources are scarce. Multimedia video transmission requires reliable bandwidth requirements used to meet low latency, while VBR real-time routing to meet diversity requirements suffers from communication problems such as network congestion, link failures, high real-time traffic, and limited bandwidth because of intermittent bursty characteristics. To address these two problems literature [46] gives an efficient dynamic traffic shaping mechanism for real-time video streaming. traffic integer dynamically changes the sending rate of subsequent network traffic according to the buffer size, rate at the input, and output rate. Also, an efficient real-time video streaming routing protocol (RTVP) over WMSN is proposed. The results show that the mechanism guarantees high QoS performance in terms of transmission rate, end-to-end delay, optimization complexity, smooth routing reliability, and fidelity criteria of multimedia streams.

Fixed-rate shaping methods are not suitable for application in sensor networks with uneven traffic. Therefore, using traffic prediction methods to preset the output rate to improve the traffic shaping performance becomes a new solution. In the literature [47], a new traffic shaping algorithm-traffic shaping algorithm with variable weight combination forecast (TSAV)-is presented to address the shortcomings of existing traffic shaping algorithms for sensor network applications, which can plan the shaping rate to smooth the output packet flow. Experiments show that the TSAV algorithm can accurately predict the traffic when applied to sensor networks, reduce the packet discard rate and increase the network throughput at the same time, and improve the QoS performance of sensor network information transmission.

Energy efficiency is a key issue in remote monitoring, in-vehicle networks and manufacturing industries that are increasingly using sensors for automation and intelligence. In manufacturing industries where data is collected by sensors to control machine work, multimedia data transmission requires the use of real-time Ethernet, which requires large bandwidth resources and is inefficient in terms of energy saving. In the literature [48] a traffic shaping algorithm called energy efficient ethernet with prediction (EEEP) strategy is proposed to save energy in the case of multimedia industrial traffic with similar characteristics. The results demonstrate that the algorithm can achieve further significant energy savings at the cost of limited delay in packet transmission.

Network coding is an excellent technique to increase the throughput of the network while saving resources. However, the existing network coding aware routing suffers from decoding failure problems and does not consider the energy and load of the actual sensor nodes. To address these problems, traffic-shaped network coding aware routing (TSCAR) is proposed in the literature [49]. In TSCAR, the decoding failure problem is solved by designing a generic network coding condition. The results show that TSCAR not only improves the throughput of the network but also makes the network lifetime of wireless sensors effectively extended.

1) IN-VEHICLE NETWORK

The in-vehicle network is a complex mesh of sensors, controls, and actuators in the early days of the automobile, which were connected by point-to-point wiring. To reduce the number of in-vehicle wires for data sharing and fast exchange, and to improve reliability and other aspects of the rapidly developing computer networks, the automotive electronic network system based on CAN, LAN, LIN, MOST, etc. were implemented, i.e. the in-vehicle network. The bandwidth requirements of modern and future automotive applications are constantly approaching the limits of existing in-vehicle network (IVN) technology. Sensors are connected to the IVN system from inside and outside of the vehicle, where they collect data from the vehicle's surroundings and transmit it to the cloud for analysis and processing and subsequent command to the IVN system. This data needs to be transmitted in real-time and with deterministic latency requirements.

For these requirements, many domestic and foreign research institutions have made great contributions. Among them, audio video bridging technology (AVB) is an audio video streaming service technology defined by the IEEE 802.1AVB working group (AVBTG), which is now widely used in other video transmission areas such as in-vehicle network systems. AVB is based on traditional Ethernet and provides the fundamental guarantee of deterministic transmission of audio-video streams by a CBS algorithm for traffic shaping. In [50], based on the delay analysis calculation of the CBS flow model, an improved CBS frame model is given to calculate the queuing transmission delay of audio and video traffic by considering the discrete characteristics of the actual queuing and scheduling of data frames in AVB networks, and the worst-case delay calculation formula is derived by comparing the frame model with the delay results of the flow model. The results show that the end-to-end delay obtained in the CBS frame model is smaller compared to the streaming model. In the literature [51], a simple traffic shaping strategy, non-blocking jitter reduction (NJR), is proposed to improve the real-time performance of the network and eliminate queue jitter. In particular, it is a greedy traffic shaping method that does not introduce any additional end-to-end delay. Scheduling and shaping in TSN as a component of future automotive Ethernet is explained in the literature [52] for both TAS and ATS shaping techniques. The results show that TAS can guarantee the worst delay for high priority flows, while ATS can guarantee the average delay for all flows.

In addition to the above needs, how to further reduce the cost and safety of the in-vehicle network is also a key issue. Safety considerations require many sensors to collect data from the surrounding environment to analyze and react to road conditions. Ethernet can integrate sensors to build active safety and autonomous driving systems which have an increasing demand for bandwidth. In the literature [53], an asynchronous traffic class called urgency based scheduler (UBS) is proposed to provide low latency guarantees in switched Ethernet while maintaining low implementation complexity. It is used to meet the stringent reliability and real-time requirements of such security systems. In terms of cost and resource consumption, the use of traffic shaping is desired to reduce the network resources required, thereby reducing costs. For the selected IP/Ethernet based network topology, a new traffic shaping algorithm, simple traffic smoother (STS), is proposed in the literature [54], where STS reduces the peak transmission rate by sending packets with a certain time distance between each frame. STS can provide the best QoS guarantees with low resources.

2) INDUSTRIAL INTERNET

For industrial Internet, the factory is full of sensors, which need to collect a huge amount of data on the network of transmitting data has high requirements, and the data of the factory can be divided into real-time data and non-real-time data according to the business requirements; real-time data often has the characteristics of periodicity in a long time the traffic generated is relatively smooth, but in a smaller time may make the transmission rate of data significantly lower than the peak rate. Not reaching the set rate, resulting in a waste of bandwidth resources. While for non-real-time data its transmission using bandwidth free moments has intermittent bursty characteristics, making sudden changes in traffic at a certain time; at the same time when periodic traffic from a large number of sensors with different periods and frame sizes is aggregated in the gateway, the aggregated traffic may burst and far exceed the average value of the aggregated traffic, thus generating network congestion [55].

Aggregated switch networks provide cost-effective regulation of IoT traffic by aggregating traffic. Aggregated network traffic is prone to bursts and partial loss of traffic when transmitted at the same time. The problem caused by bursty traffic can be solved by traffic shaping. The current traditional shaping techniques are queuing shaping in a single switch and the efficiency of the service is limited. In the literature [56] it is suggested to use multiple switches for cooperative traffic shaping to accommodate more traffic. The results show that the above scheme can effectively reduce the queue length. IoT gateways generate a form of congestion known as the massive access problem (MAP) while satisfying the data demand, and in the literature [57] it is demonstrated that the quasi-deterministic-transmission-policy (QDTP) traffic shaping approach can greatly alleviate the above problem.

In the industrial Internet, multimedia traffic from sensors is divided into real-time traffic and non-real-time traffic according to business requirements, and real-time traffic has a certain demand for bandwidth. The priority of real-time packets is higher than that of non-real-time packets. To reduce conflicts with real-time packets transmitted by other nodes, each local node needs to smooth its non-real-time packet flows. In the literature [58] a traffic shaping framework is given to improve the connectivity carrying capacity of nodes. The developed framework aims to exploit the delay tolerance difference between real-time and non-real-time traffic to smooth less delay-sensitive traffic and thus increase the capacity of the network nodes. The literature [59] provides an in-depth study of input rate regulation schemes from the perspective of smoothing and regulation effects of input traffic. The smoothing effect is characterized by the difference between the interval of the packet departure process and the input rate regulation mechanism. Hard real-time traffic (HRT) is periodic and must have guaranteed deadlines, while soft real-time traffic (SRT) is aperiodic, time-limited and has a lower priority than HRT. There are scheduling problems for the two types of traffic with different performance goals, for which an easy-to-implement and low-complexity traffic shaping scheme is proposed in the literature [60], which not only satisfies the timing requirements of HRT traffic but also preserves the feasibility of SRT traffic and improves the response time [61] [62].

B. TRAFFIC SHAPING IN VIDEO STREAM

Video surveillance systems have evolved to the stage of network video surveillance. In network video surveillance systems, the transmission of real-time video requires low packet loss, low time delay, and low bit rate jitter. These several features combine to determine the quality of the real-time video. The quality of real-time video in network surveillance systems is easily affected by network bandwidth. To address this problem, a closed-loop code rate regulation strategy based on RCTP control messages is proposed in the literature [63], which adjusts the encoder code rate by analyzing the feedback information of RTCP receiver packets to self-adapt to the network carrying capacity; meanwhile, a stream traffic shaping strategy is proposed to solve the problem of high burstiness of H.264 streams in the transmission process, avoiding network congestion and jitter, thus reduce the packet loss rate. The results show that the proposed scheme can improve the video transmission quality and reduce the packet loss problem in the network.

HTTP adaptive streaming (HAS) is a streaming video technology that is widely used on the Internet. However, it has many drawbacks; when multiple user applications in the same network compete for the bandwidth it can significantly degrade the user quality of experience (QoE). To address this problem literature [64] gives a strategy called RWTM (receive window tuning method), which is based on TCP flow control, the scheme uses flow control in the gateway to limit the maximum rate at which the traffic can be sent so that the client can receive it in time, thus improving the client's QoE.

Adaptive video streaming in the wireless network in the face of multiple applications sharing the same link underperformed. Mobile video streaming has an ON-OFF (data transmission is spaced, with ON referring to intervals of data transmission and OFF referring to intervals that are in an idle state and exhibit periodic characteristics.) The characteristics of the transmission pattern can hinder the accuracy of the available bandwidth estimation in the network and thus affect the network performance. For this reason, a new video rate selection scheme XMAS (efficient mobile adaptive streaming) is given in the literature [65] for efficient video streaming in wireless networks. A client-based traffic shaping scheme is also described, which acts as a restriction on packet transmission from the server and the transmission of data is maintained in a constant state. The results show that XMAS achieves an average video rate increase of up to 20% while significantly reducing the re-buffering rate.

Multiple applications in the same network competing for bandwidth may make the application unstable in addition to reducing QoE. The instability is due to the ON-OFF characteristics of video streams that lead to uncertainty in the estimation of available bandwidth, and applications from different users can overlap between the two states, leading to misjudgment of available bandwidth and causing the application to oscillate between different video files. A server-based traffic shaping approach is proposed in the literature [66] that can significantly reduce this oscillation and does not degrade bandwidth utilization. A mechanism based on traffic decomposition is proposed in the literature [67], which divides the bandwidth in the gateway, formulates the desired range of bandwidth required by each user, and constrains the client's bit rate to the target range. This mechanism achieves the best quality of experience for a large number of users.

In addition to the above issues, TCP congestion control algorithms have an impact on QoE. The literature [68] compares the results produced by combining each of the two traffic shaping methods (hierarchical token bucket shaping method (HTBM) and RWTM) in the gateway with each of the four TCP congestion algorithms (NewReno, Vegas, Illinois, and Cubic) in the server. The results show that Illinois with RWTM has the best QoE without causing congestion.

C. TRAFFIC SHAPING FOR PRIVACY DEFENSE IN IOT

Privacy protection is a key issue in the IoT environment; in the current era of intelligence, many smart devices are connected to the network every day, and smart devices collect personal data for analysis and processing; although the data are protected by encryption, there is still a leakage crisis of private information [69]. While smart furniture brings convenience to people, it also increases the problem of personal privacy leakage. In the literature [70], a stochastic traffic padding (STP) traffic shaping algorithm is introduced to equate the upload and download of traffic during user activity, and inject upload and download traffic during other periods to interfere with the eavesdropper's judgment and prevent the leakage of user activity. A scheme that can automatically integrate traffic shaping behavior into the program code of different traffic shaping profiles while minimizing intrusions is proposed in the literature [71].

IoT devices are commonly connected using wireless, which leads to a concise deployment of the devices and the risk of exposure of private information despite the encryption of the data [72]. In literature [73] a new MAC-layer traffic shaping defense against device fingerprinting attacks is proposed, moreover, this method does not increase the latency of the network. In literature [74], a traffic shaper satisfying a first-come-first-served queuing discipline that outputs traffic dependent on the input using a DP (differential privacy) mechanism is proposed by building a rigorous event-level DP model. In literature [75], a lightweight programmable privacy framework called privacy guard is designed and developed to address the problem of sensitive information leakage and protect the privacy and security of users.

This section gives an introduction to the scenario application of traffic shaping in an IoT environment. The approach of traffic shaping in sensor networks, in-vehicle networks, industrial Internet, and video transmission is explored. From these cases, it can be seen that traffic shaping techniques can well control the rate of traffic, circumvent the congestion problem caused by burst traffic, and improve the overall performance of the network; at the same time, it can save cost, reduce the pressure of network equipment, and provide quality QoS for various applications. Finally, in the home IoT scenario, shaping techniques are used to regulate the moment of traffic appearance to interfere with the identification of intruders and achieve privacy protection.

VI. TRAFFIC SHAPING IN THE FUTURE INTERNET OF THINGS

The development of traffic shaping technology has been quite mature, and the shaping algorithm has been continuously applied and developed. However, due to the lack of network resources and the difficulty of bandwidth technology breakthrough, making traffic shaping for the rate needs to make a fixed setting. For the industrial Internet, the data generated by its sensors generally has a periodic pattern, which can adapt to the setting of shaping technology. But for the traditional Internet, its data traffic is mostly irregular, and it is not easy to set the rate of traffic at a reasonable value. Whether it is the industrial Internet or the traditional Internet, the shaping rate is set too high to easily cause network congestion, and too low to cause a waste of bandwidth resources and not utilize the data transmission, the fixed rate approach does not adapt well to the current environment [76].

Nowadays, the booming field of artificial intelligence, such as machine learning and deep learning, has brought new ideas to the development of other fields. By using algorithms of machine learning and deep learning combined with shaping techniques, a new scheme is designed. Firstly, explore the bandwidth idleness in the network link through a neural network and so on, and find out the maximum transmission rate of the current link according to the bandwidth utilization; then adjust the rate of shaping technique adaptively according to its rate, so that the utilization of bandwidth resources can be maximized while keeping the output rate smooth and avoiding network bandwidth jitter and congestion.

The multi-token bucket algorithm is generated by different service hierarchies in the network. It is used to address bandwidth fairness, reduce packet loss and improve network performance. In an industrial network environment, most of the sensors in the plant transmit audio and video data and control commands with a relatively simple data structure, and the data in the plant is divided into real-time data and non-real-time data with certain priority requirements. Therefore, the hierarchical multi-token bucket algorithm scheme can be applied to the industrial Internet, where separate shapers are equipped with different priority services for scheduling, and the traffic with the same traffic characteristics after shaping is aggregated to the parent shaper for further scheduling. The token generation rate and a bucket capacity of the traditional token bucket algorithm are fixed at the time, which cannot adapt to the changeable network environment. By combining classification and class clustering algorithms in machine learning to dynamically adjust the token generation rate and the number of tokens in the bucket. The new scheme can dynamically adapt to the network environment and the class of service traffic. The hierarchical shapers are tree-like, and the shaped traffic is aggregated at the parent node. With the same layer of shapers, tokens can be borrowed from each other to deal with unexpected situations to improve network performance and avoid congestion. In the face of complex network environments, a single shaping algorithm cannot well solve all the needs of the network. The collection of multiple shaping algorithms becomes an intelligent shaper. Different network environments are analyzed and combined shapers are selected to solve the problem.

VII. CONCLUSION

This paper begins with a general description of traffic shaping and provides an understanding of the principles and effects of traffic shaping by describing traffic shaping techniques and algorithms. Traffic shaping is a common means used to reduce peak traffic on the consumer Internet. This paper, however, focuses on traffic shaping techniques in the Internet of Things. With the rapid development of industrial digitalization and Internet of Everything technology, the transmission of IoT data requires a large number of bandwidth resources, which are relatively scarce, so using traffic shaping to limit the traffic becomes an economical way. At the same time, new demands emerge for scalability, determinism and real-time performance in industrial and in-vehicle IoT. And software-defined networks and time-sensitive networks have a bright future as emerging technologies that can meet these needs. Traffic shaping combined with these two technologies becomes the focus of our attention. In this paper, we explain the existing shaping schemes using the SDN framework and introduce their solutions for bandwidth resource allocation and inequity. This paper then illustrates the TSN standard, describes the four shapers in TSN, and explores the characteristics of different shapers. According to the investigation, it is found that a traffic shaping scheme using SDN architecture can achieve the allocation of bandwidth resources and break the inequity between competing applications. The shaper in TSN, on the other hand, can be well suited for industrial and vehicular IoT to meet the demand for real-time and determinism. Then, this paper presents the applications of traffic shaping in sensor networks, in-vehicle networks, industrial Internet, streaming video, and privacy aspects of IoT. Based on these applications it is found that these traffic shaping schemes can solve traffic burstiness, network congestion, bandwidth utilization, prioritization of service traffic, energy saving, and cost issues. Finally, this paper presents some views on the future of traffic shaping in IoT. We hope that the presentation and exploration of the research and application of traffic shaping techniques in IoT in this paper can provide some help to researchers in related fields.

REFERENCES

[1] S. K. Nair and D. C. Novak, "A traffic shaping model for optimizing network operations," *European Journal of Operational Research*, vol. 180, pp. 1358–1380, 2007.

[2] X. Zhang and T. Wang, "Elastic and reliable bandwidth reservation based on distributed traffic monitoring and control," *IEEE Transactions on Parallel and Distributed Systems*, vol. 33, pp. 4563–4580, 2022.

[3] Y. Wang, "CARMA+QOS realizing the adaptation of network's fluence," *Computer and Digital Engineering*, vol. 34, pp. 69–71, 2006.

[4] Y. Liu, "Research and implement of network traffic shaping based on leak bucket method and token bucket algorithm," Master's thesis, Northeast Normal University, Changchun,China, 2008.

[5] X. Zhang, Y. Wang, J. Zhang, L. Wang, and Y. Zhao, "RINGLM: A link-level packet loss monitoring solution for software-defined networks," *IEEE Journal on Selected Areas in Communications*, vol. 37, pp. 1703–1720, 2019.

[6] X. Zhang, Y. Wang, M. Yang, and G. Geng, "Toward concurrent video multicast orchestration for caching-assisted mobile networks," *IEEE Transactions on Vehicular Technology*, vol. 70, pp. 13 205–13 220, 2021.

[7] Z. Cao, Q. Liu, D. Liu, and B. Yan, "Survey of time-sensitive networking," *Application Research of Computers*, vol. 38, pp. 647–655, 2021.

[8] L. Zhang and P. Wang, "Survey of traffic shaping and scheduling in time-snesitive network," *Microelectronics & Computer*, vol. 39, pp. 46–53, 2022.

[9] X. Li, "Research based on multilayer token re-allocation traffic shaping," Master's thesis, National University of Defense Technology, Changsha,China, 2010.

[10] S. Budin, I. Riadi *et al.*, "Traffic shaping menggunakan metode HTB (hierarchical token bucket) pada jaringan nirkabel," *Buletin Ilmiah Sarjana Teknik Elektro*, vol. 1, pp. 144–152, 2019.

[11] B. Chen and Q. Yao, "Research on VBR traffic shaping with token bucket filter," *Journal of Zhejiang University(Engineering Science)*, vol. 38, pp. 56–59, 2004.

[12] W. M. Zuberek and D. Strzeciwilk, "Modeling traffic shaping and traffic policing in packet-switched networks," *Journal of Computer Sciences and Applications*, vol. 6, pp. 75–81, 2018.

[13] J. Turner, "New directions in communications (or which way to the information age?)," *IEEE Communications Magazine*, vol. 24, pp. 8–15, 1986.

[14] S. Lekcharoen, "Performance and evaluation of adaptive backoff schemes in traffic shaping over high speed network," in *Proc. IEEE APCC*, pp. 241–245, 2007.

[15] X. Li and Y. Guo, "Comparison between token bucket algorithms in QoS technology," *ZTE Communications*, vol. 13, pp. 56–60, 2007.

[16] J. Loeser and H. Haertig, "Low-latency hard real-time communication over switched ethernet," in *Proc. IEEE ECRTS*, pp. 13–22, 2004.

[17] S.-K. Kweon and K. Shin, "Achieving real-time communication over ethernet with adaptive traffic smoothing," in *Proc. IEEE RTAS*, pp. 90–100, 2000.

[18] R. Wang, W. Chi, and H. Zhang, "Adaptive traffic shaping policy based on token bucket algorithm of wireless-optical broadband access network," *Journal of Electronics & Information Technology*, vol. 39, pp. 1401–1408, 2017.

[19] M. Niu and L. Jiang, "Research and design based on multi-token bucket traffic shaping algorithm," *Microelectronics & Computer*, vol. 28, pp. 110–113, 2011.

[20] X. Zhang, P. Li, Q. Wang, and L. Du, "Multi-service token bucket algorithm of MAC QoS system in access network," *Computer Science*, vol. 36, pp. 68–70, 2009.

[21] Q. Li, "A mechanism of traffic shaping and bandwidth guarantee at edge gateways," *Microprocessors*, vol. 33, pp. 23–29, 2013.

[22] L. Silva, P. Pedreiras, P. Fonseca, and L. Almeida, "On the adequacy of SDN and TSN for industry 4.0," in *Proc. IEEE ISORC*, pp. 43–51, 2019.

[23] S. Ortiz, "Software-defined networking: On the verge of a breakthrough?" *IEEE Computer*, vol. 46, pp. 10–12, 2013.

[24] F. Hu, Q. Hao, and K. Bao, "A survey on software-defined network and OpenFlow: From concept to implementation," *IEEE Communications Surveys & Tutorials*, vol. 16, pp. 2181–2206, 2014.

[25] Y. Yoo, G. Yang, M. Kang, and C. Yoo, "Adaptive control channel traffic shaping for virtualized SDN in clouds," in *Proc. IEEE CLOUD*, pp. 22–24, 2020.

[26] I. B. Mustafa and T. Nadeem, "Dynamic traffic shaping technique for http adaptive video streaming using software defined networks," in *Proc. IEEE SECON*, pp. 178–180, 2015.

[27] R. Al-Haddad, E. S. Velazquez, A. Fatima, and A. Winckles, "A novel traffic shaping algorithm for SDN-sliced networks using a new WFQ technique," *International Journal of Advanced Computer Science and Applications*, vol. 12, 2021.

[28] M. Amiri, H. Al Osman, and S. Shirmohammadi, "Datacenter traffic shaping for delay reduction in cloud gaming," in *Proc. IEEE ISM*, pp. 569–574, 2016.

[29] A. Nasrallah, A. S. Thyagaturu, Z. Alharbi, C. Wang, X. Shao, M. Reisslein, and H. ElBakoury, "Ultra-low latency (ULL) networks: The IEEE TSN and IETF DetNet standards and related 5G ULL research," *IEEE Communications Surveys & Tutorials*, vol. 21, pp. 88–145, 2018.

[30] X. XU, "Time-sensitive network technology and its application in industrial network," *Telecommunications Network Technology*, pp. 1–5, 2018.

[31] H. Hu, Q. Li, H. Xiong, and B. Fang, "The delay bound analysis based on network calculus for asynchronous traffic shaping under parameter inconsistency," in *Proc. IEEE ICCT*, pp. 908–915, 2020.

[32] Y. Liu, "Research and implementation of time sensitive routing based on traffic shaping," Master's thesis, Beijing Jiaotong University, Beijing,China, 2021.

[33] S. Thangamuthu, N. Concer, P. J. L. Cuijpers, and J. J. Lukkien, "Analysis of ethernet-switch traffic shapers for in-vehicle networking applications," in *Proc. IEEE DATE*, pp. 55–60, 2015.

[34] M. Kim, D. Hyeon, and J. Paek, "eTAS: Enhanced time-aware shaper for supporting nonisochronous emergency traffic in time-sensitive networks," *IEEE Internet of Things Journal*, vol. 9, pp. 10 480–10 491, 2022.

[35] K. Alexandris, P. Pop, and T. Wang, "Configuration and evaluation of Multi-CQF shapers in IEEE 802.1 time-sensitive networking (TSN)," *IEEE Access*, vol. 10, pp. 109 068–109 081, 2022.

[36] G. Peng, S. Wang, Y. Huang, R. Huo, T. Huang, and Y. Liu, "Traffic shaping at the edge: Enabling bounded latency for large-scale deterministic networks," in *Proc. IEEE ICCW*, pp. 1–6, 2021.

[37] A. G. Marino, F. Fons, Z. Haigang, and J. M. M. Arostegui, "Traffic shaping engine for time sensitive networking integration within in-vehicle networks," in *Proc. IEEE VNC*, pp. 182–189, 2021.

[38] L. Zhao, P. Pop, and S. Steinhorst, "Quantitative performance comparison of various traffic shapers in time-sensitive networking," *IEEE Transactions on Network and Service Management*, vol. 19, pp. 2899–2928, 2022.

[39] D. Hellmanns, J. Falk, A. Glavackij, R. Hummen, S. Kehrer, and F. Dürr, "On the performance of stream-based, class-based time-aware shaping and frame preemption in tsn," in *Proc. IEEE ICIT)*, pp. 298–303. IEEE, 2020.

[40] A. Gogolev, R. Braun, and P. Bauer, "TSN traffic shaping for OPC UA field devices," in *Proc. IEEE INDIN*, vol. 1, pp. 951–956, 2019.

[41] Z. Zhou, Y. Yan, M. Berger, and S. Ruepp, "Analysis and modeling of asynchronous traffic shaping in time sensitive networks," in *Proc. IEEE WFCS*, pp. 1–4, 2018.

[42] Z. Zhou, M. S. Berger, S. R. Ruepp, and Y. Yan, "Insight into the IEEE 802.1 Qcr asynchronous traffic shaping in time sensitive network," *Adv. sci. technol. eng. syst. j*, vol. 4, pp. 292–301, 2019.

[43] A. Nasrallah, A. S. Thyagaturu, Z. Alharbi, C. Wang, X. Shao, M. Reisslein, and H. Elbakoury, "Performance comparison of IEEE 802.1 TSN time aware shaper (TAS) and asynchronous traffic shaper (ATS)," *IEEE Access*, vol. 7, pp. 44 165–44 181, 2019.

[44] E. Mohammadpour, E. Stai, M. Mohiuddin, and J.-Y. Le Boudec, "Latency and backlog bounds in time-sensitive networking with credit based shapers and asynchronous traffic shaping," in *Proc. IEEE ITC 30*, vol. 2, pp. 1–6, 2018.

[45] B. Fang, Q. Li, Z. Gong, and H. Xiong, "Simulative assessments of credit-based shaping and asynchronous traffic shaping in time-sensitive networking," in *Proc. IEEE ICAIT*, pp. 111–118, 2020.

[46] A. A. Ahmed, "A real-time routing protocol with adaptive traffic shaping for multimedia streaming over next-generation of wireless multimedia sensor networks," *Pervasive and Mobile Computing*, vol. 40, pp. 495–511, 2017.

[47] C. Luo and W. Xie, "Sensor network traffic shaping algorithm by variable weight combination forecast," *Journal of Signal Processing*, vol. 29, pp. 1597–1603, 2013.

[48] A. Cenedese, M. Michielan, F. Tramarin, and S. Vitturi, "An energy efficient traffic shaping algorithm for ethernet-based multimedia industrial traffic," in *Proc. IEEE ETFA*, pp. 1–4, 2014.

[49] X. Shao, C. Wang, C. Zhao, and J. Gao, "Traffic shaped network coding aware routing for wireless sensor networks," *IEEE Access*, vol. 6, pp. 71 767–71 782, 2018.

[50] E. LI, F. HE, and H. XIONG, "End-to-end traffic latency computation using frame shaping model in AVB network," *Journal of Beijing University of Aeronautics and Astronautics*, vol. 43, pp. 1442–1449, 2017.

[51] R. I. Davis and N. Navet, "Traffic shaping to reduce jitter in controller area network (CAN)," *ACM SIGBED Review*, vol. 9, pp. 37–40, 2012.

[52] Z. Zhou, J. Lee, M. S. Berger, S. Park, and Y. Yan, "Simulating TSN traffic scheduling and shaping for future automotive ethernet," *Journal of Communications and Networks*, vol. 23, pp. 53–62, 2021.

[53] J. Specht and S. Samii, "Urgency-based scheduler for time-sensitive switched ethernet networks," in *Proc. IEEE ECRTS*, pp. 75–85, 2016.

[54] M. Rahmani, K. Tappayuthpijarn, B. Krebs, E. Steinbach, and R. Bogenberger, "Traffic shaping for resource-efficient in-vehicle communication," *IEEE Transactions on Industrial Informatics*, vol. 5, pp. 414–428, 2009.

[55] D. Daniel-Simion and G. Dan-Horia, "Traffic shaping and traffic policing impacts on aggregate traffic behaviour in high speed networks," in *Proc. IEEE SACI*, pp. 465–467, 2011.

[56] K. Honda, N. Shibata, R. Harada, Y. Ishida, K. Akashi, S. Kaneko, T. Miyachi, and J. Terada, "Cooperated traffic shaping with traffic estimation and path reallocation to mitigate microbursts in IoT backhaul network," *IEEE Access*, vol. 9, pp. 162 190–162 196, 2021.

[57] E. Gelenbe and K. Sigman, "IoT traffic shaping and the massive access problem," in *Proc. IEEE ICC*, pp. 1–6, 2022.

[58] A. Elwalid and D. Mitra, "Traffic shaping at a network node: Theory, optimum design, admission control," in *Proc. IEEE INFOCOM '97*, vol. 2, pp. 444–454, 1997.

[59] M. Sidi, W. Liu, I. Cidon, and I. Gopal, "Congestion control through input rate regulation," in *Proc. IEEE GLOCOM*, vol. 3, pp. 1764–1768, 1989.

[60] B. Gaujal and N. Navet, “Traffic shaping in real-time distributed systems: A low-complexity approach,” *Computer Communications*, vol. 22, pp. 1562–1573, 1999.

[61] X. Zhang, Y. Wang, G. Geng, and J. Yu, “Delay-optimized multicast tree packing in software-defined networks,” *IEEE Transactions on Services Computing*, pp. 1–14, 2021.

[62] S. Yao, M. Wang, Q. Qu, Z. Zhang, Y.-F. Zhang, K. Xu, and M. Xu, “Blockchain-empowered collaborative task offloading for cloud-edge-device computing,” *IEEE Journal on Selected Areas in Communications*, pp. 1–1, 2022.

[63] Z. Wang and Y. Wang, “Strategy of closed-loop rate control and traffic shaping based on RTCP,” *Computer Science*, vol. 38, pp. 100–102, 2011.

[64] C. B. Ameur, E. Mory, and B. Cousin, “Shaping HTTP adaptive streams using receive window tuning method in home gateway,” in *Proc. IEEE IPCCC*, pp. 1–2, 2014.

[65] S. Kim and C. Kim, “XMAS: An efficient mobile adaptive streaming scheme based on traffic shaping,” *IEEE Transactions on Multimedia*, vol. 21, pp. 442–456, 2019.

[66] S. Akhshabi, L. Anantakrishnan, C. Dovrolis, and A. C. Begen, “Server-based traffic shaping for stabilizing oscillating adaptive streaming players,” in *Proc. ACM NOSSDAV’13*, pp. 19–24, 2013.

[67] R. Houdaille and S. Gouache, “Shaping HTTP adaptive streams for a better user experience,” in *Proc. ACM MMSys’12*, pp. 1–9, 2012.

[68] C. Ben Ameur, E. Mory, and B. Cousin, “Combining traffic-shaping methods with congestion control variants for HTTP adaptive streaming,” *Multimedia Systems*, vol. 24, pp. 1–18, 2016.

[69] A. Engelberg and A. Wool, “Classification of encrypted IoT traffic despite padding and shaping,” in *Proc. the 21st Workshop on Privacy in the Electronic Society*, pp. 1–13, 2022.

[70] N. Apthorpe, D. Y. Huang, D. Reisman, A. Narayanan, and N. Feamster, “Keeping the smart home private with smart (er) iot traffic shaping,” *arXiv preprint arXiv:1812.00955*, 2018.

[71] D. Oehlert, S. Saidi, and H. Falk, “Code-inherent traffic shaping for hard real-time systems,” *ACM Transactions on Embedded Computing Systems*, vol. 18, pp. 1–21, 2019.

[72] K. Dziubinski and M. Bandai, “Bandwidth efficient iot traffic shaping technique for protecting smart home privacy from data breaches in wireless lan,” *IEICE Transactions on Communications*, pp. 2020–3182, 2021.

[73] M. Alyami, M. Alkhowaiter, M. Al Ghanim, C. Zou, and Y. Solihin, “MAC-Layer traffic shaping defense against WiFi device fingerprinting attacks,” in *Proc. IEEE ISCC*, pp. 1–7, 2022.

[74] S. Xiong, A. D. Sarwate, and N. B. Mandayam, “Network traffic shaping for enhancing privacy in IoT systems,” *IEEE/ACM Transactions on Networking*, vol. 30, pp. 1162–1177, 2022.

[75] M. Uddin, T. Nadeem, and S. Nukavarapu, “Extreme SDN framework for IoT and mobile applications flexible privacy at the edge,” in *Proc. IEEE PerCom*, pp. 1–11, 2019.

[76] A. V. Omelchenko, E. A. Rozdymakha, and O. V. Fedorovz, “Network traffic shaping based on prediction of polynomial trend self-similar time series,” in *Proc. IEEE 25th International Conference Radioelektronika (RADIOELEKTRONIKA)*, pp. 450–452. IEEE, 2015.

PLACE
PHOTO
HERE

HAO FU was born in Jining, Shandong, China, in 1998. He received the B.S. degree from the Qilu University of Technology, where he is currently pursuing the M.S. degree. His main research interests include computer networks, industrial internet and machine learning.

PLACE
PHOTO
HERE

MUYI SUN studied in the School of Information Science and Engineering of Shandong Agricultural University from 2015 to 2019, majoring in spatial information and digital technology, and obtained a bachelor’s degree in engineering. From 2021 to now, study computer technology in the Computing Department of Qilu University of Technology, studied for a master’s degree, studied under Professor Xinchang Zhang.Research direction: intelligent network.

PLACE
PHOTO
HERE

BINGYU HE was born in Anqing, Anhui, China,in 1999. He received the B.S. degree from Shandong Jiaotong University. He is currently pursuing the M.S. degree with the Qilu University of Technology. His main research interests include computer networks, networked control systems, and machine learning.

PLACE
PHOTO
HERE

JINHUA LI received the B.S. degree from the Shandong University of Science and Technology, China, in 2005. Presently, she is a senior engineer at the Qilu University of Technology (Shandong Academy of Sciences). She have developed several system such as video conference and edge data collection system. Her research interests include computer network and cloud computing.

PLACE
PHOTO
HERE

XIAOMIN ZHU received the Ph.D. degree from Institute of Computing Technology, Chinese Academy of Sciences, China, in 2010. He is Master Instructor of Qufu Normal University and is the director of the Shandong Institute of Bigdata. He has over 20 papers in research journals and conferences, and developed several systems including a big data analysis system and an edge computing platform. His research interests include networking, edge computing and big data.

• • •