Workgroup: DetNet
Internet-Draft:
draft-hp-detnet-tsn-queuing-mechanisms-
evaluation-01
Published: 19 December 2023
Intended Status: Informational
Expires: 21 June 2024
Authors: Jinjie. Yan Yufang. Han Shaofu. Peng
ZTE Corporation ZTE Corporation ZTE Corporation
Yuehong. Gao
Beijing University of Posts and Telecommunications

Analysis and Evaluation for TSN Queuing Mechanisms

Abstract

TSN technology standards developed in the IEEE 802.1TSN Task Group define the time-sensitive mechanism to provide deterministic connectivity through IEEE 802 networks, i.e., guaranteed packet transport with bounded latency, low packet delay variation, and low packet loss.This document summarizes and evaluates various queuing technologies of TSN as reference information for Scaling Deterministic Networks Requirements[I-D.ietf-detnet-scaling-requirements] and Enhancing Deterministic Forwarding.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on 21 June 2024.

Table of Contents

1. Introduction

Time sensitive networking (TSN) makes it possible to carry data traffic of time-critical and/or mission-critical applications over a bridged Ethernet network shared by various kinds of applications with different Quality of Service(QoS) requirements, i.e., time and/or mission critical TSN traffic and non-TSN best effort traffic. TSN provides guaranteed data transport with bounded low latency, low delay variation, and extremely low data loss for time and/or mission critical traffic. By reserving resources for critical traffic, and applying various queuing and shaping techniques, TSN guarantees a worst-case end-to-end latency for critical data, and achieves zero congestion loss for critical data traffic. TSN also provides ultra-reliability for data traffic via a data packet level reliability mechanism as well as protection against bandwidth violation, malfunctioning, malicious attacks, etc.

At present, TSN series standards are basically mature and provide queuing or scheduling algorithms that support different delay accuracies, such as frame preemption ([IEEE802.3br] and [IEEE802.1Qbu]), CBS ([IEEE802.1Qav]), CQF ([IEEE802.1Qch]), ECQF ([IEEE802.1Qdv]), TAS ([IEEE802.1Qbv]), ATS ([IEEE802.1Qcr]), etc. These mechanisms provide QoS capabilities for different application scenarios, such as CBS guarantees the upper bound of latency while ensuring rate, ATS provides low latency services for emergency flows. These two can be classified as mechanisms with bounded latency. CQF can provide delay jitter independent of the number of hops, while TAS can provide extremely low jitter through precise calculations. These two can be classified as mechanisms for jitter control. The following sections will analyze these queueing technologies one by one.

2. TSN queuing and shaping technologies

2.1. Frame Preemption

2.1.1. Frame Preemption Overview

Frame preemption mechanism was introduced to mitigate negative effects of the guard band reserved by the TAS. As it requires modifications of both management (IEEE 802.1) and Ethernet MAC (IEEE 802.3) functions, two working groups jointly proposed required changes to both standards. Therefore, the frame preemption is described in two different standard documents: [IEEE802.1Qbu] and [IEEE802.3br].

[IEEE802.3br], also named Interspersing Express Traffic, differentiates two types of traffic: preemptable (also called mPacket) and express. The type of a frame is identified by examining

the VLAN tag defined by IEEE 802.1Q. Frames arriving from the MAC client are serviced either by preemptable MAC (pMAC) or express MAC (eMAC). If both frames arrive at the same time, express traffic is serviced first as it has higher priority. In the case when express frame arrives while preemptable frame is already being transmitted on egress port, if certain conditions are met, it will interrupt current transmission . After express traffic has been serviced, the transmission of interrupted frame is resumed and different parts of the interrupted frame are re-assembled by a MAC Merge Sublayer (MMS) that is a part of the modified Ethernet MAC which supports frame preemption.

It is important to note that frame fragmentation works on link-by-link basis, i.e., each switch forwards preemptable frame only after it is fully re-assembled. This is clearly different from end-to-end packet fragmentation that is commonly used in IP networks. This ensures compatibility with the devices that do not support frame preemption mechanism.

2.1.2. Frame Preemption Analysis

As explained earlier, the main motivation behind the frame preemption mechanism is to reduce the length of guard band enforced by the TAS. Without frame preemption, reserved guard band must match the transmission time of the largest low-priority frame. In the case of 100 Mbps Ethernet, the worst-case time would be around 125us (transmission time of the largest Ethernet frame), which represents a huge bandwidth penalty.Frame preemption allows reducing the guard band down to approximately 12us which is tenfold improvement. It can also be combined with other queue technologies to minimize the interference delay from low priority packets.

2.2. CBS(Credit-Based Shaper)

2.2.1. CBS(CBS Overview)

CBS proposed by [IEEE802.1Qav] divides time-sensitive services which need to be transmitted preferentially into two classes: class A and class B, and sets a certain bandwidth for them. Through priority mapping, TSN flows with different priorities enter different queues for scheduling respectively. As described in Section 8.6.8.2 of [IEEE802.1Qav], the credits of each class increase according to the idle slope (as the guaranteed rate), and decrease according to the send slope (usually equal to idle slop minus port transmit rate), both of which are parameters of the CBS.

TSN flows are gently sent to the network by credit evaluation to deal with data burst and aggregation, CBS can limit burst traffic and prevent audio and video streams arriving at the same time from

different terminals, which generates significant buffering congestion, resulting in packet loss.

2.2.2. CBS Analysis

CBS sets the pre-configuration of bandwidth limit for each traffic class. Typically set 75% of the maximum bandwidth for bandwidth intensive applications such as audio and video.

CBS does not rely on time synchronization, but still rely on frequency synchronization. So it can be applied in scenarios such as cross clock domains, non strict time synchronization, and asynchronous clocks.

The disadvantage of CBS is that the average latency will increase, although the combination of CBS and SRP (Stream Reservation Protocol) can limit the latency of each bridge to less than 250us. The paper[AVB-Latency] analyzes that in small-scale networks using FE (Fast Ethernet, 100Mbps) ports, CBS can guarantee a worst-case latency of less than 2 milliseconds for Class A and less than 50 milliseconds for Class B under a maximum of 7 hops. However, other papers[ClassA-Latency-Calc] shows the conclusion is not valid, it indicates that there is still a problem of delay degradation in CBS due to burstiness cascade. In general, the more hops, the worse the delay degradation . In large-scale networks, the number of network hops is usually large, such as 15 or more hops, which poses great challenges for the deployment of CBS independently. The upper bound of latency can not meet the requirements of many services which need low latency.

2.3. TAS(Time-Aware Shaping)

2.3.1. TAS Overview

In industrial IoT application scenarios, some time-sensitive streams will carry critical information. These streams require highly predictable delay and jitter in transmission. If the delay or jitter exceeds the threshold, it may cause serious consequences. At the same time, most of these streams are transmitted according to a certain time period, and streams with this characteristic are called Scheduled Traffic.

For the Scheduled Traffic, CBS transmission algorithm can not meet the requirements, because in CBS algorithm, if a low priority frame is already being transmitted, then that transmission will complete before a higher priority frame can access the transmission medium, so there could be a delay of up to a maximum-sized frame before a high priority transmission can start. If such delays occur at every hop, then the accumulated latency could be unacceptably large.

To address this issue, [IEEE802.1Qbv] proposes the TAS mechanism. As Scheduled Traffic is a periodic stream, it is possible to determine the time when each packet of streams arrives at each network device after Scheduled Traffic starts to transmit. As long as sufficient bandwidth is reserved for Scheduled Traffic on these devices in advance, it can ensure that other non-scheduled traffic will not interfere with the transmission of Scheduled Traffic.

TAS provides a scheduling mechanism of gate operations, which is based on high-precision clock synchronization. Each port of the TSN bridge has a gate control list (GCL) for opening or closing operations, and the 8 queues at these ports need to be associated with each of the 8 Transmission Gates respectively. Each entry in the GCL corresponds to a gate operation, and then packets are selected from the queue for transmission based on the gate control list. The gate control list contains two items: GateState and TimeInterval. GateState is used to set the state of Transmission Gate corresponding to queues, there are two states for each Transmission Gate: Open and Closed. "Open" means that packets in the associated queue can be transmitted according to the corresponding transmission algorithm, while "Closed" means that packets in the associated queue are not allowed to be transmitted. TimeInterval indicates the duration of the gate state. After TimeInterval ticks have elapsed since the completion of the previous gate operation in the GCL, control passes to the next gate operation.

Since transmission operation is an "atomic operation", in order to avoid the situation that the packet in corresponding queue can not be completely transmitted before the gate is closed, TAS defines an advanced check mechanism. If a packet cannot be fully transmitted within the remaining time of the corresponding gate operation state is open, this packet will not be transmitted until the next time when the gate is opened.

In order to ensure that the remaining non-scheduled traffic cannot affect the transmission of scheduled traffic, TAS uses a guard band (Guard Band) mechanism long enough to stop the transmission of non-scheduled traffic in advance of the protected time slot to be certain that the last non-scheduled transmission has completed before scheduled traffic transmission starts. In the worst case, the last non-scheduled transmission would start a maximum-sized frame transmission before the start of the scheduled traffic "window". In effect, a guard band is created before the time that the scheduled traffic transmission is due to start; transmission of non-scheduled traffic is not permitted between the start of the guard band and the start of the scheduled traffic window. The simplest approach for the guard band is to be as long as a maximum-sized frame transmission time.

2.3.2. TAS Analysis

The premise of TAS is that all terminals and network devices need to achieve nanosecond clock synchronization across the network (such as [IEEE1588], [IEEE802.1AS])to ensure that the GCL time of all outgoing ports is synchronized. Appropriate transmission "windows" can be arranged for the scheduled traffic at each outgoing port to achieve that the traffic can obtain extremely low transmission delay by accurate calculation. But when the network topology scale is large, that is, there may be a large number of nodes and links, it is usually difficult to achieve real-time synchronization, that limits the deployment of TAS; At the same time, large-scale networks carry a massive number of application flows, which will be a great challenge for TAS that relies on precise calculations and complex configurations.

On the other hand, the transmission window reserved for deterministic flows through GCL is usually exclusive. During the time period when the gate state of the queue associated with scheduled flows is open, even if the scheduled traffic does not arrive as expected, the transmission opportunities during this period will not be shared with other non-scheduled flows. Therefore, the bandwidth utilization in this scenario is insufficient.

2.4. CQF(Cyclic Queuing and Forwarding)

2.4.1. CQF Overview

CQF follows the gate operations of the TAS mechanism: when the gate is open, the packets in the queue are allowed to be forwarded to the next node; when the gate is closed, incoming packets are buffered in the queue before they are allowed to transmit. CQF simplifies the design of TAS by installing fixed configurations on the GCL. Time in CQF networks is divided into cycles with equal value T, and there are two queues performing enqueue and dequeue operations in a cyclic manner under the control of RX GCL and TX GCL. When the packets enter the queue Q1 in cycle duration T(cycle c), the receiving gate of Q1 opens. Meanwhile, the output sending gate of Q2 opens, packets are transmitted to the next hop. When the next cycle(c+ 1) starts, the output sending gate of Q1 opens and sends the packets received in the previous cycle, the receiving gate of Q2 is open and starts to receive new packets. This cyclic queuing and forwarding mode can achieve transmission in a fixed duration that does not exceed 2T on per hop. CQF could provide the deterministic latency relies on two principles. First, the upstream and downstream nodes are perfectly synchronized, and the rotation of the upstream sending cycle and the downstream receiving cycle must be consistent. Second, a packet received at the cycle must be sent at the next cycle in a node. Thus, the predictable end-to-end latency only depends on the cycle size and

path length, and regardless of topology. CQF is useful for applications that do not require very small latency and jitter, but which are still real-time and require bounded worst-case latency.

2.4.2. CQF Analysis

CQF can provide deterministic services with a maximum jitter of no more than 2T. The key issue is how to select the size of the cycle T and calculate the start time of the flow. The length of the queue is directly related to the size of cycle. If the cycle is too small, the queue is short too. Although the single hop queuing delay for traffic transmission is very small, there is not enough space to buffer more incoming flows, which can lead to a large number of flows that can not be scheduled; If the cycle is too large, it also means that queuing delay will become too large on per hop, which will result in a large end-to-end worst-case delay. Some traffic with requirements of low latency can not be scheduled, and larger queue lengths will also require more buffer resources. Due to limited underlying hardware resources, the difficulty and cost of hardware implementation are directly proportional to the buffer size.

It is necessary to carefully select the cycle size which needs to be large enough to accommodate all deterministic traffic, and in addition, the cycle includes a time duration called dead time (DT), which is the sum of delays 1, 2, 3 and 4 defined in Figure 1 of [RFC9320]. The value of DT ensures that the last packet of a cycle on the upstream node can be fully transmitted to the buffer of the same cycle on the downstream node. In the case of LAN, DT is relatively small compared to cycle T and is considered negligible, so only two buffer queues can run well. But in some deterministic networks, a single hop over a long distance can produce a large delay. Considering that the optical transmission speed in fiber is 200000km/s, the propagation delay of some long-distance links can be in the order of a few milliseconds, which is much larger than in LAN, and cannot be ignored. In order to cover the DT, more buffer queues need to be introduced.

On the other hand,the dead time (DT) reduces the available time for deterministic stream transmission within the cycle time, that make it impossible to deliver high-bandwidth services with extremely low jitter. Meanwhile, like TAS, classic CQF also rely on nanosecond clock synchronization across the entire network, where all network nodes align their cycle boundaries, and they cooperate with each other. This pattern limits the application of CQF in networks where precise time synchronization cannot be deployed.

In addition, a large amount of deterministic traffic demands will produce more fluctuations when dynamic services join or leave, which requires corresponding resource scheduling algorithms to allocate

resources appropriately among multiple flows to avoid transmission conflicts. For example, in some complex aggregation situations, a large number of traffic with periodic characteristics may be gathered at a certain intermediate node. If the scheduling result is not appropriate, it will result in traffic congestion in one queue of the intermediate aggregation node, while the other part of the queue is idle, the traffic distribution is not balanced enough, which also exacerbates the probability of traffic conflicts. Currently CQF rely on overprovision to solve this problem, but this will result in a small scale of supported flows. Therefore, it is necessary to introduce optimized traffic planning design with path calculation and resource reservation, such as planning through a centralized controller, but it also puts higher requirements on the algorithm.

2.5. ECQF(Enhancements to Cyclic Queuing and Forwarding)

2.5.1. ECQF Overview

ECQF specifies procedures, protocols and managed objects for enhanced CQF, to avoid the requirement for system clock synchronization. ECQF specifies a transmission selection procedure that organizes frames in a traffic class output queue into logical bins that are output in strict rotation at a fixed frequency. It ensures that each bin will be emptied before the next bin is due for transmission. There are two ways of filling the bins: Time-based CQF stores received frames into bins based on the time of reception of the frame. Count-based CQF stores received frames into bins based on per-stream-per-output-queue byte counter state machines, and is recommended to use only on the boundary nodes of the frequency locking domain. Bin selection method can be configured based on an input-output port pair, or be configured for specific streams. It also provided multiple cycle model for different services, and the processing of flow aggregation/ disaggregation based on count-based CQF.

ECQF is based on the following principles:

1. A Bridge output queue using ECQF is notionally divided into bins. The bins are enabled for output serially, at a fixed interval T_C, which same (or nearly the same) value is used for some number of Bridges along the path of a stream, said path constituting a ECQF segment of a network. At any given instant in time, a particular output bin can be available for accepting frames for later transmission, or enabled for transmitting frames to the associated medium, or neither, but never both.

2. Each stream utilizing a ECQF segment is allocated a certain number of bit times per transmission interval T_C. Steps are taken to ensure that no bin contains frames for any stream that will take, in total, longer than that stream’s allocated bit

times to transmit. Resource reservation ensures that the total bit times allocated over all streams passing through a ECQF queue do not exceed T_C, even including possible interference from other queues on the port.

3. Frames assigned to the same bin at ingress to a ECQF Segment remain together in the same bin at each hop along the ECQF segment. Two methods are provided to accomplish this, time-based bin assignment and count-based bin assignment.

2.5.2. ECQF Analysis

ECQF does not rely on time synchronization. Time-based CQF need frequency lock and frequency synchronization. The number of CQF cycles in two Bridges that are frequency locked must be the same, over an arbitrarily long interval of time. Count-based CQF can be more relaxed. The cycle phase difference between two nodes is allowed. However, when the sum of clock jitter and phase difference exceeds N cycles (N is the number of selected bins), time-based CQF will not work.

In the ideal case, every stream would have a T_C value chosen so that exactly one frame of a stream is transmitted on each cycle T_C. Multiple values of T_C can be applied to a single output port. Streams are allocated to, and thus use up the bandwidth available to, each cycle separately. There are many ways to allocate buffer space to individual frames. Allocating bandwidth to a slower cycle times uses more buffer space, because frames dwell for a longer time. On the contrary, allocating bandwidth to a faster cycle time may get the optimal bounded latency, which may be somewhat oversubscribed. If the end-to-end latency requirements of the streams permit (but the case is not always like this), a stream can be assigned to a slower cycle. This will reduce the overprovision factor. Overprovision reduces the utilization of network resources.

Different CQF priority levels may operate simultaneously on one output port and have different maximum frame sizes, and some may enable preemption, different priority levels may have different amounts of time during one cycle that cannot be allocated to stream transmission. The burst may be limited at edge. For a new stream to be admitted, it must be true that the available transmission times over all of the CQF levels on all of the output ports through which the stream travels have not been exhausted.

2.6. ATS(Asynchronous Traffic Shaping)

2.6.1. ATS Overview

In order to solve the problem of zero congestion and packet loss in the transmission of aperiodic data, and to further optimize the

bandwidth utilization of services without strict requirements for time synchronization, [IEEE802.1Qcr] defines an asynchronous traffic shaping device ATS.

ATS is designed based on Urgency-Based Scheduler ([UBS]). First,it identify the packets through the stream_handle (a sub-parameter of the stream identification function in [IEEE802.1CB]) and priority (the priority field in the VLAN tag) and match it into the corresponding stream filter, which specifies the stream gate and scheduler for the packets. The specified stream gate assigns internal priority, in this way, different degrees of delay guarantee can be provided in different nodes on the transmission path, it make the allocation of latency more flexible. The packets that has been assigned an internal priority enter the specified scheduler for shaping, which uses the interleaved algorithm based on the token bucket, and then assigned a eligible time, which is the expected transmission time of the packets. After the shaper, packets enter the corresponding shared queue (per incoming port plus traffic class) according to the internal priority and wait to be sent. The transmission selection algorithm is based on strict priority that transmits packets from the queues in the order from higher priority to lower priority sequentially. If the eligible time of the first packet in the shared queue is less than the current time, then the first packet can be sent directly and executes the transmission selection algorithm from the higher priority. Otherwise, turn to the next higher priority.

2.6.2. ATS Analysis

ATS adopts a principle called Rate-Controlled Service Disciplines (RCSD), which is a non work-conserving packet service discipline. It consists of two parts: the rate controller implements the rate control policy, and the scheduler implements packet scheduling based on some scheduling policy, such as static priority, first come first served, or earliest deadline. By separating the rate controller and the scheduler, RCSD effectively decouples the bandwidth of each stream from its delay bound, therefore, RCSD can support low latency and low bandwidth service.

The advantage of ATS is that when packets enter the queue, packets are assigned an eligible time, it allows urgent flows can be transmitted preferentially. ATS also has the concept of a scheduler group, where multiple ATS schedulers can belong to a single ATS scheduler group. The ATS scheduler does not rely on binding to hardware queues. From the delay analysis formula of ATS, it can be concluded that the allocation of internal priority and the assignation of scheduler directly determine the delay boundary of flows. In the stream gate component of each hop, different internal priority can be assigned to packets instead of external priority,

that can more flexibly allocate the service level. Therefore, the ATS scheduler can perform flexible shaping for per flow or aggregated flows. ATS can be placed on each hop, then the network will not generate large burst accumulation, and the performance will be improved.

The ATS scheduler state machine operation is based on the ATS scheduler clocks, which is an implementation specific local system clock function. There is no need to require nodes in the network to achieve time synchronization.

A large number of flow aggregations will occur in a complex network topology, and it is necessary to consider flow aggregation strategies at intermediate nodes in the network. The end-to-end delay upper bound provided by ATS is generally inversely proportional to the service rate and may be larger.

Scaling deterministic networks require a large number of services to be carried, and the cost of interleaved regulators (IR) maintained in per hop is high. Meanwhile, it is necessary to pay attention to the problem of IR head-of-line blocking(HOL) in large-scale networks.

3. Evaluation of TSN queuing mechanism with the requirements of scaling Deterministic networks

The following requirements are described in [I-D.ietf-detnet-scaling-requirements].

3.1. Tolerate Time Asynchrony

- CBS: Does not rely on time synchronization, but still rely on frequency synchronization.

- TAS: Packets must be sent in a specific fixed timeslot. Non-synchronized network nodes can cause packets to not be sent completely in the expected transmission gate window and will have to wait for a dedicated window in the next period, resulting in a delay. This delay can affect end-to-end latency if it accumulates on every node in the path. Therefore, in the calculation of GCL, an accurate arrival time of a flow at each node needs to be learned. The premise is that all terminals and network devices need to achieve nanosecond clock synchronization across the network (such as [IEEE1588], [IEEE802.1AS])to ensure that the GCL time of all outgoing ports is synchronized.

- CQF: Rely on nanosecond clock synchronization across the entire network, where all network nodes share the same hardware scheduling timeslots as cycles and align their cycle boundaries, cooperating with each other.

- ECQF: Does not rely on time synchronization. Time-based CQF need frequency lock (frequency synchronization too). Count-based CQF can be more relaxed.

- ATS: Based on the ATS scheduler clocks, which is an implementation specific local system clock function. No need to require nodes in the network to achieve time synchronization, but still need frequency synchronization or careful bandwidth management constraints.

3.2. Support Large Single-hop Propagation Latency

- CBS: Link delay does not affect the rate based shaping logic of CBS. Traffic is assumed to arrive asynchronously.

- TAS: In the calculation of GCL, both the propagation delay and processing delay in the path has been taken into account. Even if a large single-hop propagation delay exists, a feasible solution can still be obtained through proper scheduling. All nodes in the path are independently configured with their own GCL. Therefore, the propagation delay of a single-hop link only impacts on the determination of TAS transmission gate window position by precise calculation on the outgoing port of the nodes, it is independent of the value of the link delay.

- CQF: The propagation delay must be much smaller than cycle time or even considered negligible, so 2-buffer mode can work. The longer the propagation or processing delay results in the larger DT that would reduce available time in a cycle.

- ECQF: The cycle phase difference between two nodes is allowed. CPAP detect message covers link propagation delay.

- ATS: Link delay does not affect asynchronous traffic shaping on per hop.

3.3. Accommodate the Higher Link Speed

- CBS: More buffer space is required to server more service bursts accordingly.

- TAS: More precise time control (smaller TimeInterval of GCL) is required.

- CQF: More buffer space is required for a specific length of cycle duration. Smaller cycle size may be choosed, but with a much smaller available zone due to the impact of DT.

- ECQF: More buffer space is required for a specific length of cycle duration. Smaller cycle size may be choosed, but may need a more

accuracy time based determination of receiving cycle in the case of clock jitter.

- ATS: More buffer space is required to server more service bursts accordingly.

3.4. Be Scalable to The Large Number of Flows and Tolerate High Utilization

- CBS: Shaping of CBS is based on serveral traffic class for aggregated flows. May need re-shaping (ATS) to avoid burstiness cascade for each class. Set the pre-configuration of bandwidth limit for each traffic class. Best-effort flows can use the unused portion of the reserved bandwidth of TSN flows.

- TAS: GCL calculation for all flows in the control plane is NP-hard problem. On the outgoing port, TAS maintain queues per traffic class. The TimeInterval of GCL, with dedicated bandwidth, reserved for a TSN flow is exclusive. During the specific TimeInterval, if that TSN flow does not send packets, the bandwidth is waste, can not used by other flows.The guard band may cause no packets on the outgoing ports to be sent within the time interval of the guard bandwidth even if the packets are ready to be sent in their queues, the available bandwidth within these time intervals is wasted.

- CQF: Transmission gates are associated with each aggregated queue. Stream filtering and policing actions per stream should be placed on each node. There is also overprovision issues. The cycle duration includes a time zone called dead time (DT) contributed by Output delay, Link delay, Frame preemption delay, Processing delay, which can not be used to send packets. So, CQF can only support fewer flows.

- ECQF: Transmission gates are associated with each aggregated queue. Stream filtering and policing actions per stream should be placed on each node. Count-based CQF needs to maintain states per flow. Cycle size is always far less than burst interval, so overprovision (caused by burst/cycle) may cause low utilization.

- ATS: ATS can perform flexible shaping for per flow or aggregated flows by maintaining interleaved regulators (IR) per “inport + traffic class”. When there are many ports, the cost is still high because it needs to maintain per flow states. ATS can achieve high bandwidth utilization.

3.5. Prevent Flow Fluctuation from Disrupting Service

- CBS: Each service flow of class A/B is permitted based on bandwidth reservation. The total amout of bandwidth reservation does not exceed the pre-configuration limit. However, flow fluctuation is more likely

to cause burstiness cascade for pure CBS, which makes the delay performance deteriorate seriously.

- TAS: The re-calculation of the GCLs are more complicated. The configuration of each outgoing port along the path is updated according to the new GCLs frequently.

- CQF: Requires corresponding flows setup algorithms to allocate resources appropriately among multiple flows to avoid transmission conflicts.

- ECQF: Time-based CQF: may ensure CQF flows to be protected.

Count-based CQF: may discard excess data above the contracted amount.

- ATS: The cost of interleaved regulators (IR) maintained per hop is high. The problem of IR head-of-line blocking should be considered.

3.6. Be Scalabcle to a Large Number of Hops with Complex Topology

- CBS: On each node the queueing delay is over-estiamted, basically inversely proportional to the idle slope. Thus the E2E delay is large. CBS does not limit the best latency, resulting in large jitter. More hops will make burst cascading more severe.

- TAS: Due to NP-hard problem, the GCL calculations and configurations are more complex, and may not meet the needs of large scale network. E2E queueing delay is negligible. E2E delay jitter is ultra-low.

- CQF: It is more difficult to select the cycle time. The end-to-end latency is proportion to cycle duration and hop count. Need making trade-offs between end-to-end delay and cycle duration

- ECQF: Need to select the cycle time from multi-CQF instances baesd on the trade-offs between end-to-end delay and cycle duration.

- ATS: Need to consider flow aggregation strategies at intermediate nodes. End-to-end delay upper bound provided by ATS is larger, basically inversely proportional to the reserved bandwidth.

3.7. Tolerate Failures of Links or Nodes and Topology changes

Not related to queuing mechanisms directly.

3.8. Support Multi-Mechanisms in Single Domain and Multi-Domains

Not related to a single queuing mechanism directly.

4. Evaluation results

According to the evaluation in section 3, the evaluation results of queuing mechanisms proposed in TSN are shown in the table below:

requiremens of scaling Deterministic Networks	evaluation results of TSN queuing mechanisms				
	CBS	TAS	CQF	ECQF	ATS
Tolerate Time Asynchrony	Yes	No	No	No	Yes
Support Large Single-hop Propagation Latency	Yes	Yes	No	Yes	Yes
Accommodate the Higher Link Speed	Yes	Partial	Partial	Partial	Yes
Be Scalable to The Large Number of Flows and Tolerate High Utilization	Partial	Partial	No	No	Partial
Prevent Flow Fluctuation from Disrupting Service	Partial	No	Partial	Partial	Partial
Be Scalabcle to a Large Number of Hops with Complex Topology	Partial	Partial	No	Partial	Partial
Tolerate Failures of Links or Nodes and Topology changes	Not directly related to queuing mechanisms				
Support Multi-Mechanisms in Single Domain and Multi-Domains	Not directly related to a single queuing mechanism				

Figure 1: Evaluation Results of TSN Queuing Mechanisms

5. Conclusion

Various applications in deterministic networks have different requirements for deterministic service indicator, and different queuing mechanisms can provide different levels of delay, jitter, and other guarantees. There may also be situations where network devices

provide multiple queuing mechanisms simultaneously. For example, network aggregation devices can use the mechanisms specified in [IEEE802.1Qbv] and [IEEE802.1Qcr] to forward traffic to different paths with different SLA at the same time. By providing multiple queuing mechanisms to meet diversified deterministic service requirements, this demand is particularly prominent in large-scale networks compared to small-scale environments.

This document uses the requirements of scaling deterministic networks to evaluate several existing queue mechanisms in TSN, analyze their characteristics, and provide a basis for selecting suitable queue mechanisms for services with different deterministic requirements. At the same time, the challenges faced by their deployment in scaling networks were also analyzed, and brings some thoughts to the design of several new queuing mechanisms proposed for enhanced deterministic forwarding.

6. IANA Considerations

This document has no IANA actions

7. Security Considerations

TBD.

8. Acknowledgements

TBD.

9. References

9.1. Normative References

[IEEE1588] "IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems", 2008, <https://ieeexplore.ieee.org/document/4579760>.

[IEEE802.1AS] "IEEE Standard for Local and Metropolitan Area Networks--Timing and Synchronization for Time-Sensitive Applications", 2020, <https://ieeexplore.ieee.org/document/9121845>.

[IEEE802.1Qav] "IEEE Standard for Local and metropolitan area networks -- Virtual Bridged Local Area Networks - Amendment 12: Forwarding and Queuing Enhancements for Time-Sensitive Streams", 2010, <https://ieeexplore.ieee.org/document/8684664>.

[IEEE802.1Qbu] "IEEE Standard for Local and metropolitan area networks -- Bridges and Bridged Networks -- Amendment

26:Frame Preemption", 2016, <https://ieeexplore.ieee.org/document/7553415>.

[IEEE802.1Qbv] "IEEE Standard for Local and metropolitan area networks -- Bridges and Bridged Networks - Amendment 25:Enhancements for Scheduled Traffic", 2016, <https://ieeexplore.ieee.org/document/8613095>.

[IEEE802.1Qch] "IEEE Standard for Local and metropolitan area networks -- Bridges and Bridged Networks - Amendment 29: Cyclic Queuing and Forwarding", 2017, <https://ieeexplore.ieee.org/document/7961303>.

[IEEE802.1Qcr] "IEEE Standard for Local and Metropolitan Area Networks--Bridges and Bridged Networks Amendment 34:Asynchronous Traffic Shaping", 2020, <https://ieeexplore.ieee.org/document/9253013>.

[IEEE802.1Qdv] "Draft Standard for Local and metropolitan area networks--Enhancements to Cyclic Queuing and Forwarding", 2023, <https://1.ieee802.org/tsn/802-1qdv/>.

[IEEE802.3br] "IEEE Standard for Ethernet-Amendment 5:Specification and Management Parameters for Interspersing Express Traffic.", 2016, <https://ieeexplore.ieee.org/document/7592835>.

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

9.2. Informative References

[AVB-Latency] "AVB Latency Math", 2010, <https://www.ieee802.org/1/files/public/docs2010/ba-pannell-latency-math-0910-v4.pdf>.

[ClassA-Latency-Calc] "Class A Bridge Latency Calculations", 2010, <https://www.ieee802.org/1/files/public/docs2010/ba-boiger-bridge-latency-calculations.pdf>.

[I-D.ietf-detnet-scaling-requirements] Liu, P., Li, Y., Eckert, T. T., Xiong, Q., Ryoo, J., zhushiyin, and X. Geng, "Requirements for Scaling Deterministic Networks", Work in Progress, Internet-Draft, draft-ietf-detnet-scaling-requirements-05, 20 November 2023, <https://datatracker.ietf.org/doc/html/draft-ietf-detnet-scaling-requirements-05>.

[IEEE802.1CB] "IEEE Standard for Local and metropolitan area networks--Frame Replication and Elimination for Reliability", 2017, <https://ieeexplore.ieee.org/document/8091139>.

[RFC9320] Finn, N., Le Boudec, J.-Y., Mohammadpour, E., Zhang, J., and B. Varga, "Deterministic Networking (DetNet) Bounded Latency", RFC 9320, DOI 10.17487/RFC9320, November 2022, <https://www.rfc-editor.org/info/rfc9320>.

[UBS] "Urgency-Based Scheduler for Time-Sensitive Switched Ethernet Networks", 2016, <https://ieeexplore.ieee.org/document/7557870>.

Authors' Addresses

Jinjie Yan
ZTE Corporation
China

Email: yan.jinjie@zte.com.cn

Yufang Han
ZTE Corporation
China

Email: han.yufang1@zte.com.cn

Shaofu Peng
ZTE Corporation
China

Email: peng.shaofu@zte.com.cn

Yuehong Gao
Beijing University of Posts and Telecommunications
China

Email: yhgao@bupt.edu.cn