AVB shaping and network reservations

LINKÖPING
UNIVERSITY

Motivation for traffic shaping

- A basic best-effort Ethernet network provides variable service (varying packet delivery latency and sometimes packet get dropped)
- Why does it happen?
  - Overbooking / Overloading. Sources are pushing data that is greater than the capacity of the network elements (links, buffers)
  - Traffic jams due to bursts of traffic from sources
  - Traffic piles up at network "intersections" (i.e., switches/bridges) to go out on the same port

Prioritization

- Prioritize Traffic: Handle packets carrying deterministic traffic in high priority output queues
  - Implement a range of output queues from low- to high-priority on each output port.
  - Prioritized packets are sent even when lower-priority packets are waiting.
  - A burst of high-priority traffic interrupts lower-priority traffic.
  - Good for high-priority traffic. ***Bad/unfair*** to lower-priority traffic.

Contracts and reservations

- Use the privilege of prioritization according to an agreed “contract”
  - Each prioritized source agrees to a contract (aka, a “reservation”) with the network that provides service acceptable to that source. This agreement could be made dynamically or be “engineered” into the network ahead of time.
  - Reservation ensures that prioritized traffic won’t overwhelm best-effort (low-priority) traffic.
  - Traffic shaping is at the heart of a reservation

Shaping approached (time slots, rates, credit)

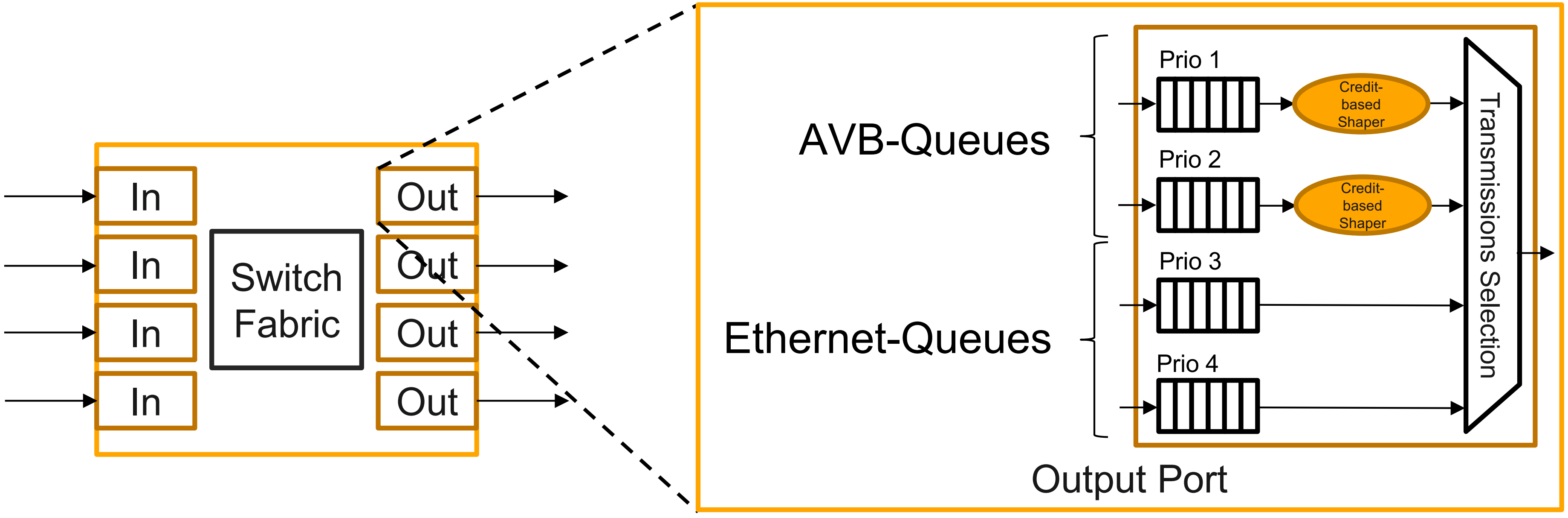
- Credit Based Shaper (CBS): Provides a smooth stream of packets within a maximum data rate
  - Also called FQTSS (Forwarding and Queuing for Time Sensitive Streams) in 802.1Qav, now part of 802.1Q-2014
- 802.1Qbv: Time Aware Shaper (TAS): Provides access to queue at specified times
- P802.1Qcr: Asynchronous Traffic Shaper (ATS): Provides immediately delivery of packets, up to a specified burst size, and within a maximum data rate

First approach: AVB

- 2 ms maximum delay
  - the maximum delay between a musician doing "something" and hearing that same "something" is 10 ms
  - the transit time of sound from monitor speakers to the musician, plus digital signal processing (DSP) delays, plus mixer delays, plus more DSP delays uses up 8 ms
  - network gets 2 ms
- maximum synchronization error less than 10 microseconds

Quality of Service with AVB

- Credit based shaper delays messages to avoid bursts
  - To avoid buffer overflow and message loss
  - To guarantee some bandwidth to lower priority traffic



AVB Credit based shaper

- Space out the high priority stream frames as far as possible
- The spaced out traffic prevents the formation of long bursts of high priority traffic, which typically arise in traffic environments with high bandwidth streams
- Bursts are responsible for significant QoS reductions of lower priority traffic classes
  - Can completely block the transmission of the lower priority traffic for the transmission time of the high priority burst
  - Increases maximum latency of this traffic and thereby also the memory demands in the bridges and end stations.

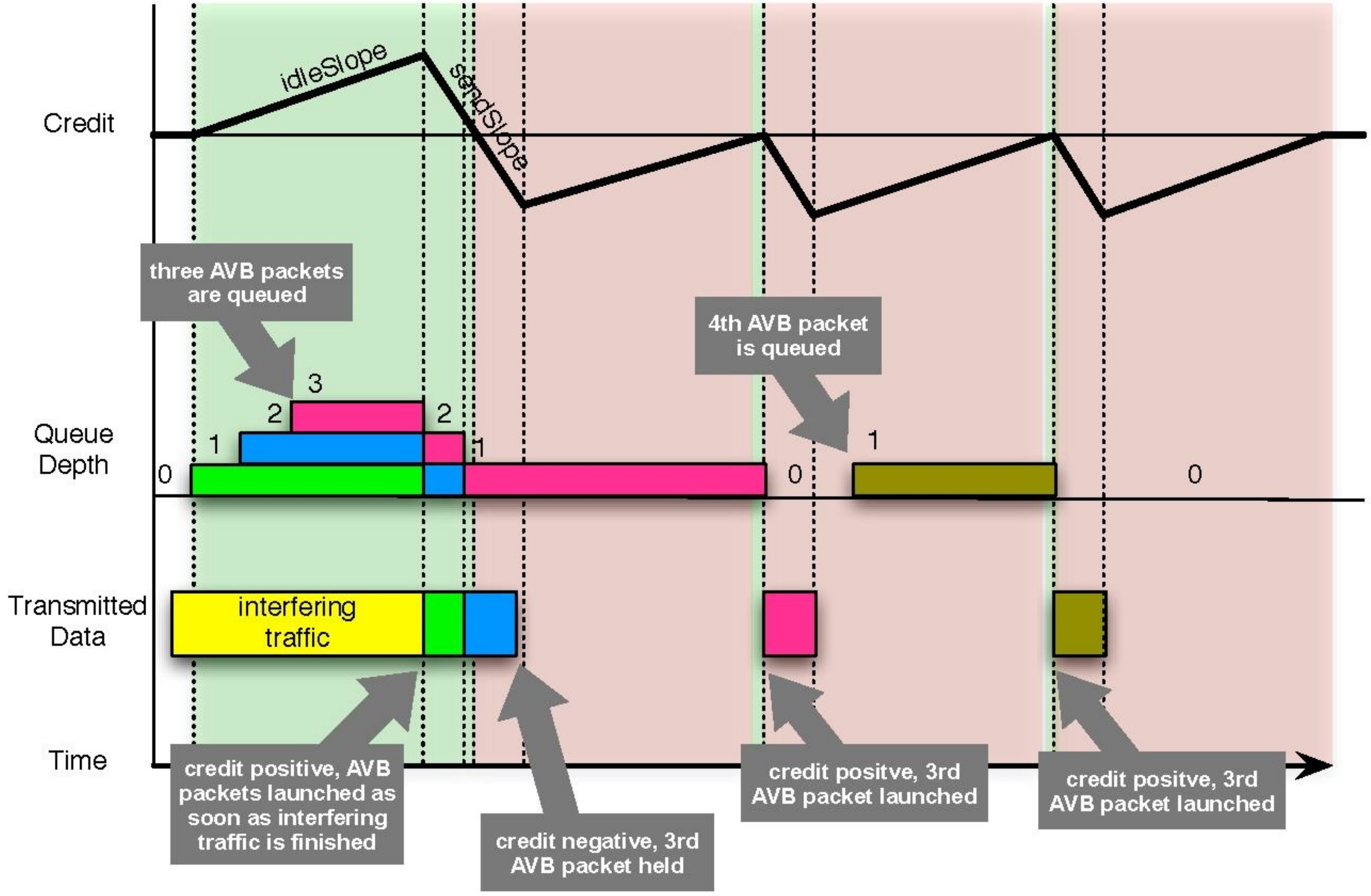
- Long bursts increase the interference time between high priority stream frames from different streams (which arrive from different ports) inside a bridge.
  - This increases the maximum latency of high priority stream frames and again the memory requirements in bridges
- Another task of the shaper is to enforce the bandwidth reservations. This enforces, on the one hand, that every AVB stream is limited to its reserved bandwidth in the talker, and, on the other hand, that the overall AVB stream bandwidth of each port (in talker and bridges) is limited to the reserved amount

Credit based shaper

$$\text{idleSlope} = \frac{\text{reservedBytes}}{\text{classMeasurementInterval}}$$

$$= \text{reservedBandwidth}.$$

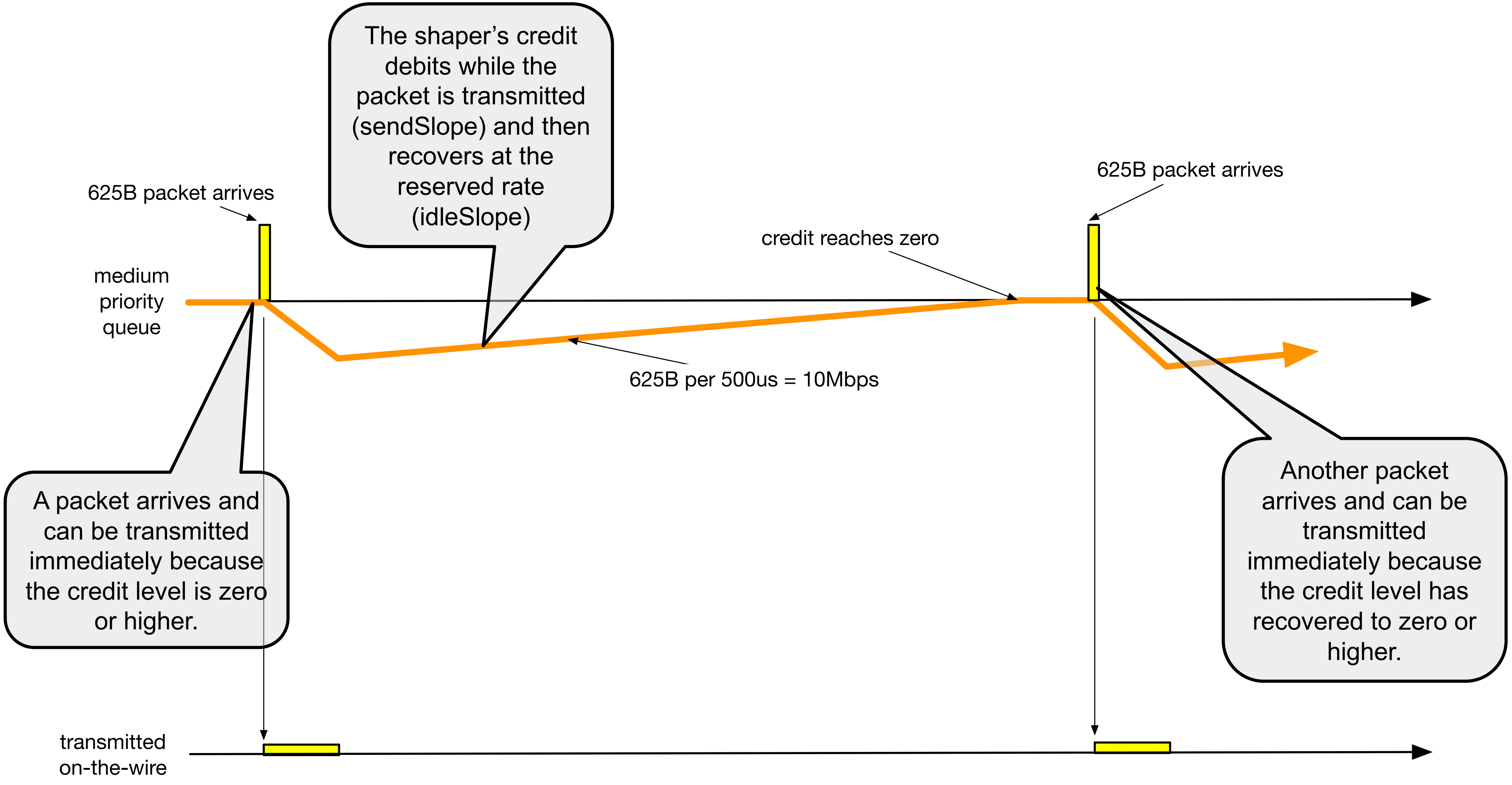
$$\text{sendSlope} = \text{idleSlope} - \text{portTransmitRate}.$$


Credit
idleSlope
sendSlope
three AVB packets are queued
4th AVB packet is queued
Queue Depth
0
1
2
3
2
1
0
1
0
Transmitted Data
interfering traffic
Time
credit positive, AVB packets launched as soon as interfering traffic is finished
credit negative, 3rd AVB packet held
credit positve, 3rd AVB packet launched
credit positve, 3rd AVB packet launched

Credit calculation rules

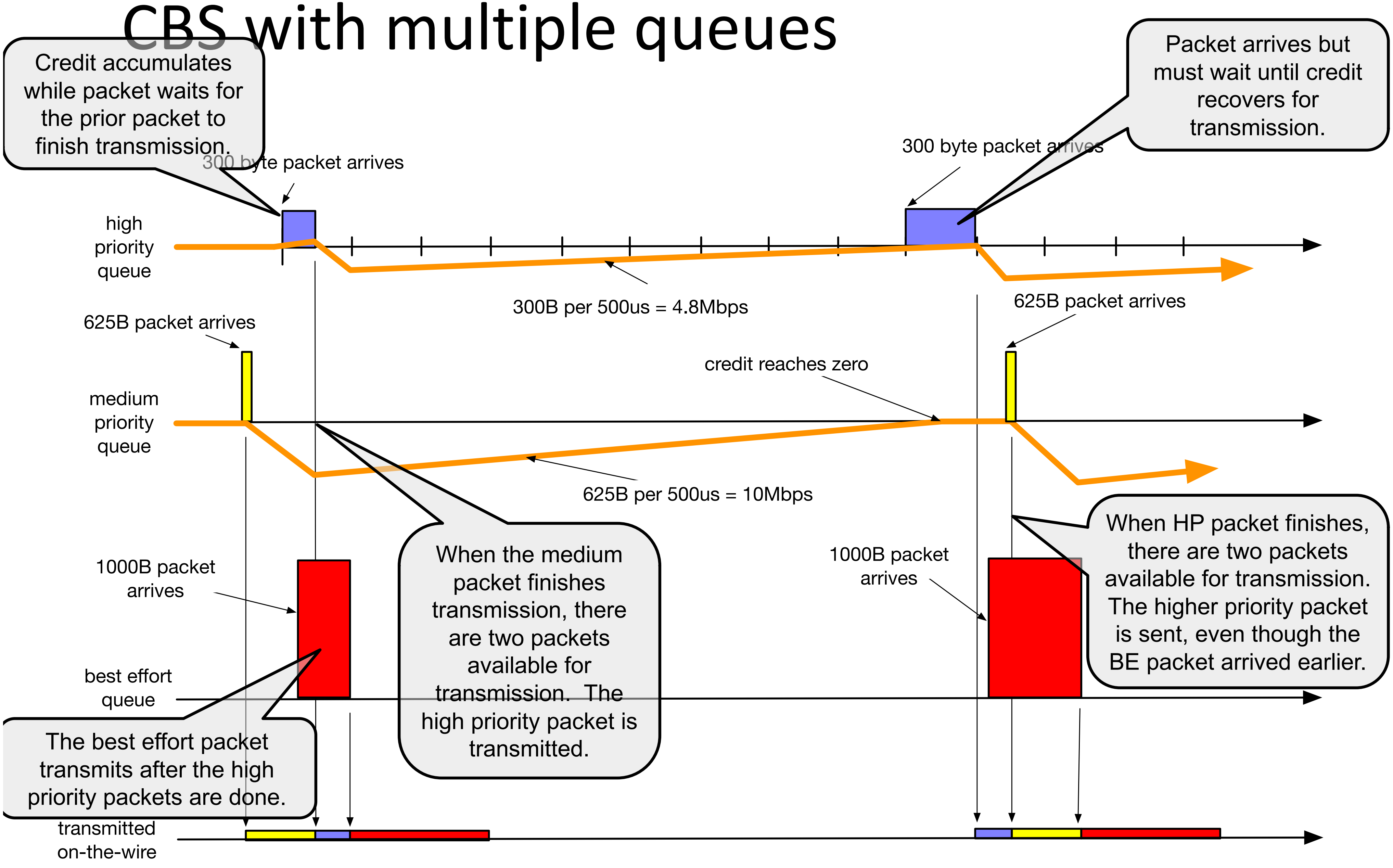
- If there is positive credit but no frame to transmit, the credit is set to zero
- During the transmission of a frame, the credit is reduced with the send slope.
- If the credit is negative and no frame is in transmission, the credit is accumulated with the idle slope until zero credit is reached.
- If there is a frame in the queue that cannot be transmitted because another frame is in transmission, the credit is accumulated with the idle slope. In this case, the credit is not limited to zero.

Credit-Based Shaping (802.1Q-2014 §34)



625B packet arrives
medium priority queue
A packet arrives and can be transmitted immediately because the credit level is zero or higher.
The shaper's credit debits while the packet is transmitted (sendSlope) and then recovers at the reserved rate (idleSlope)
625B per 500us = 10Mbps
credit reaches zero
625B packet arrives
Another packet arrives and can be transmitted immediately because the credit level has recovered to zero or higher.
transmitted on-the-wire

CBS with multiple queues



Credit accumulates while packet waits for the prior packet to finish transmission.

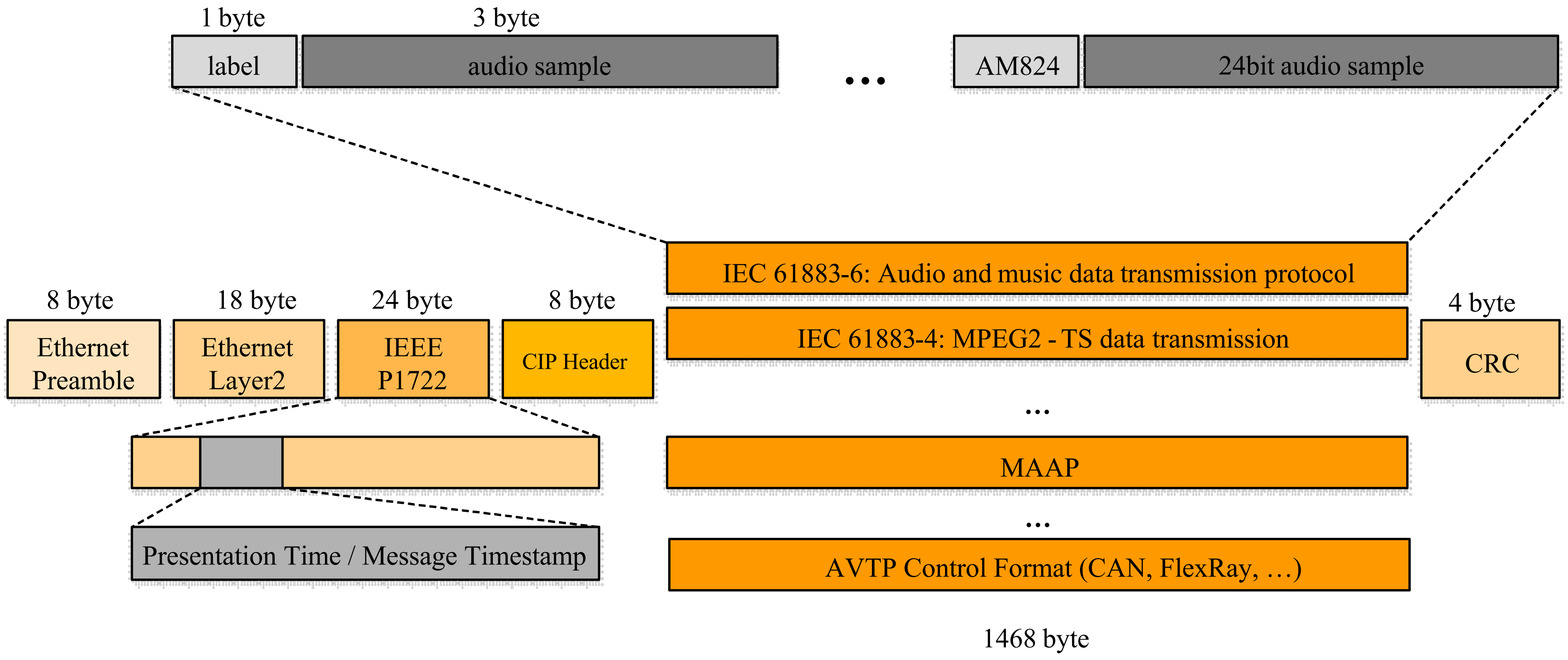
Packet arrives but must wait until credit recovers for transmission.

When the medium packet finishes transmission, there are two packets available for transmission. The high priority packet is transmitted.

When HP packet finishes, there are two packets available for transmission. The higher priority packet is sent, even though the BE packet arrived earlier.

The best effort packet transmits after the high priority packets are done.

AVB transport protocol (IEEE 1722)

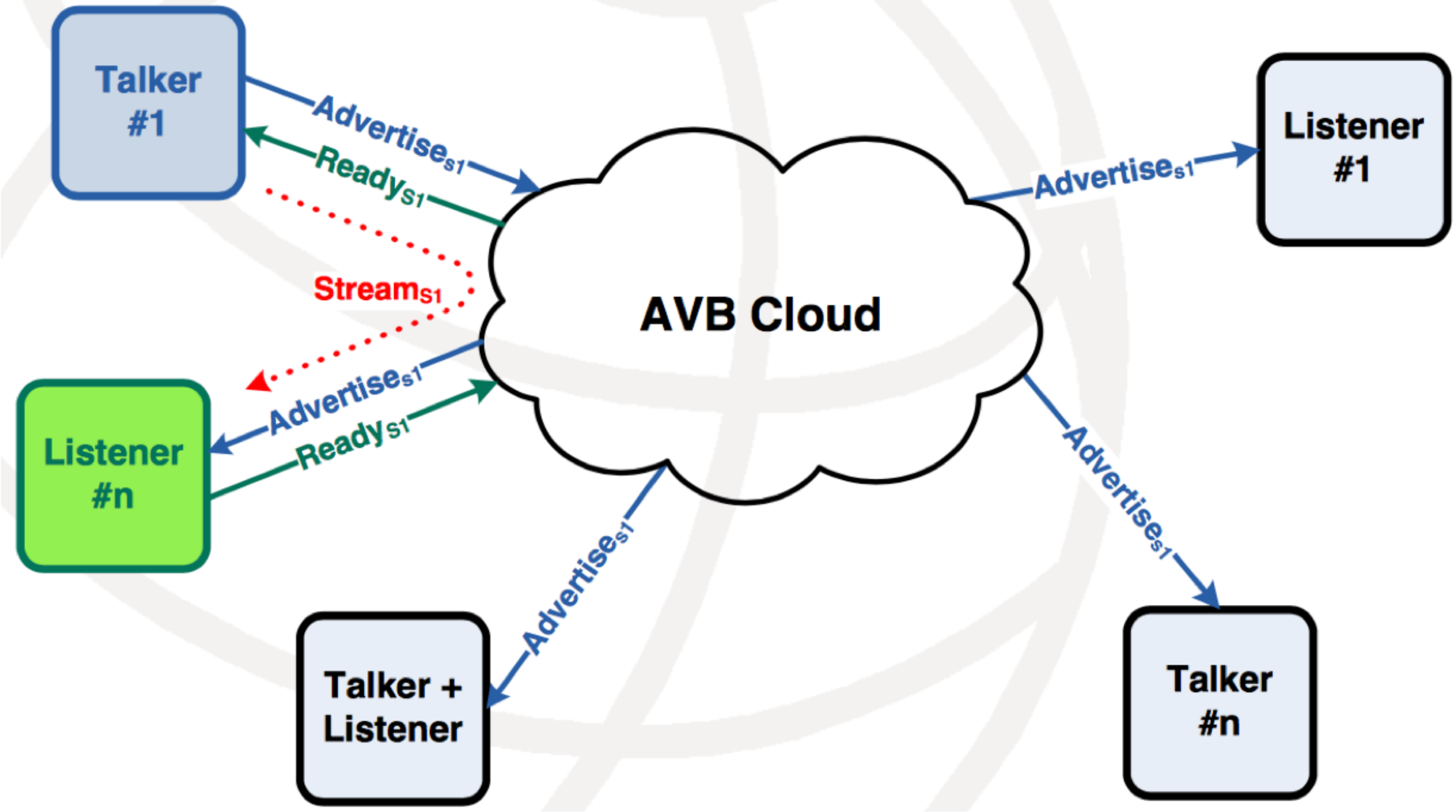

1 byte
3 byte
label
audio sample
…
AM824
24bit audio sample
8 byte
18 byte
24 byte
8 byte
Ethernet Preamble
Ethernet Layer2
IEEE P1722
CIP Header
IEC 61883-6: Audio and music data transmission protocol
IEC 61883-4: MPEG2 - TS data transmission
…
MAAP
…
AVTP Control Format (CAN, FlexRay, …)
4 byte
CRC
Presentation Time / Message Timestamp
1468 byte

- AVTP control format added in 2016
- EtherType: 0x22F0

Stream Reservation Protocol

- 802.1Qat (now rolled into 802.1Q)
- One of the core protocols of AVB
- Allows sources (talkers) to advertise streams to sinks/users (listeners) through the network
- Also allows to withdraw
- Gives end stations the tool to automatically configure the network to deliver content to the right users
- Multiple Stream Registration Protocol (MSRP)
- Multiple VLAN Registration Protocol (MVRP)
- MSRP and MVRP are in turn based on the Multiple Registration Protocol (MRP)

SRP advertise and ready frames

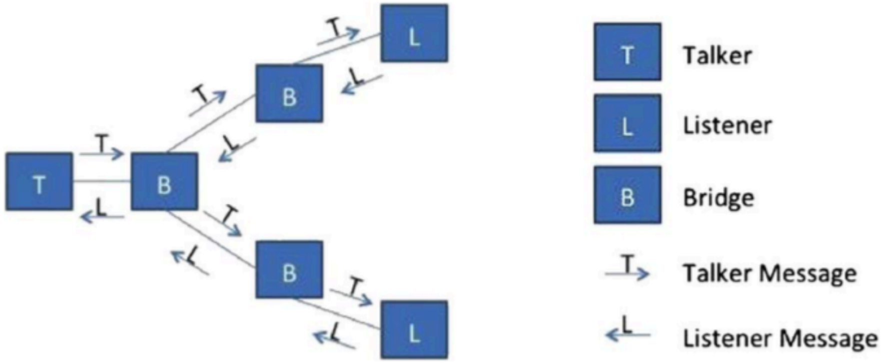


Talker advertise message format

- stream ID (MAC address associated with the talker plus a 16 bit ID)
- stream DA
- VLAN ID
- priority (determines traffic class)
- rank (emergency or nonemergency)
- traffic specification (TSpec): max frame size; maximum number of frames per class interval
- accumulated latency

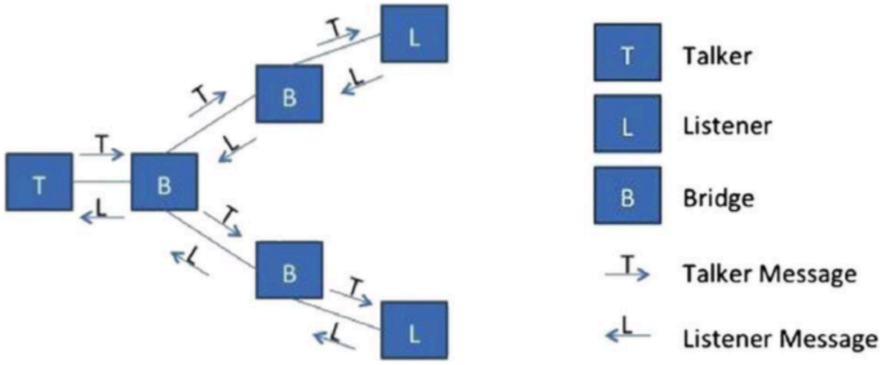
Forwarding of stream announce

- Talker send advertise message
- Each switch/bridge evaluates whether reservation can be made
  - whether sufficient bandwidth is available on each port
  - whether sufficient memory is available to guarantee no packet loss
  - reservation is not made; only when receiving listener message
  - forwards the talker message, after updating the accumulated the hop count



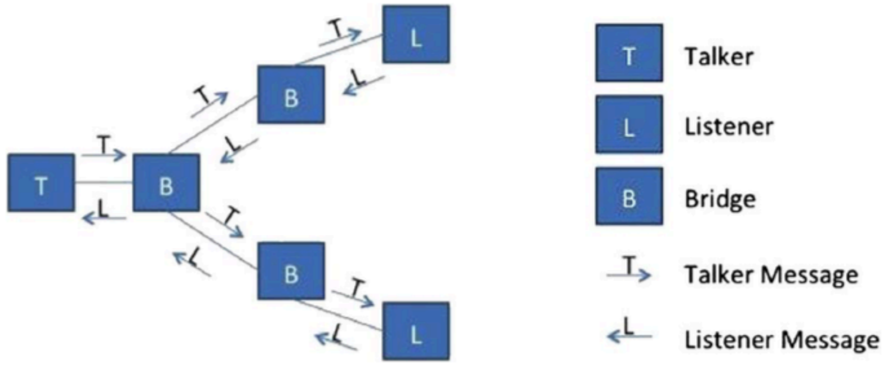
Reservation failures

- If any device on the path from talker to listener determines that the stream cannot be supported, it changes the type of the message from <u>talker advertise</u> to <u>talker failed</u>
- Then adds additional information to the message
  - bridge ID where the failure occurred;
  - reservation failure code to identify the reason for the failure.
  - Allows network engineer to pinpoint the location of the issue



Listeners

- Listeners send listener message if they want the stream
- Listener communicates the status of the stream by sending either a listener ready if it received a talker advertise or a listener asking failed if it received a talker failed



Reservations made

- When bridges receive a listener ready (or ready failed) message for a valid stream on a given port, they make a reservation on that port
  - update the bandwidth on the traffic shaper for the queue
  - update available bandwidth for the given port
  - adding the port to the forwarding entry for the stream DA

$$\begin{aligned}\text{idleSlope} &= \frac{\text{reservedBytes}}{\text{classMeasurementInterval}}\\ &= \text{reservedBandwidth}.\end{aligned}$$

Listener messages propagated back

- Listener message propagated back toward the talker
- Talker receives a listener ready message, it may begin transmitting
- If talker receives ready failed, it knows that at least one listener has requested the stream but the corresponding reservation could not be created

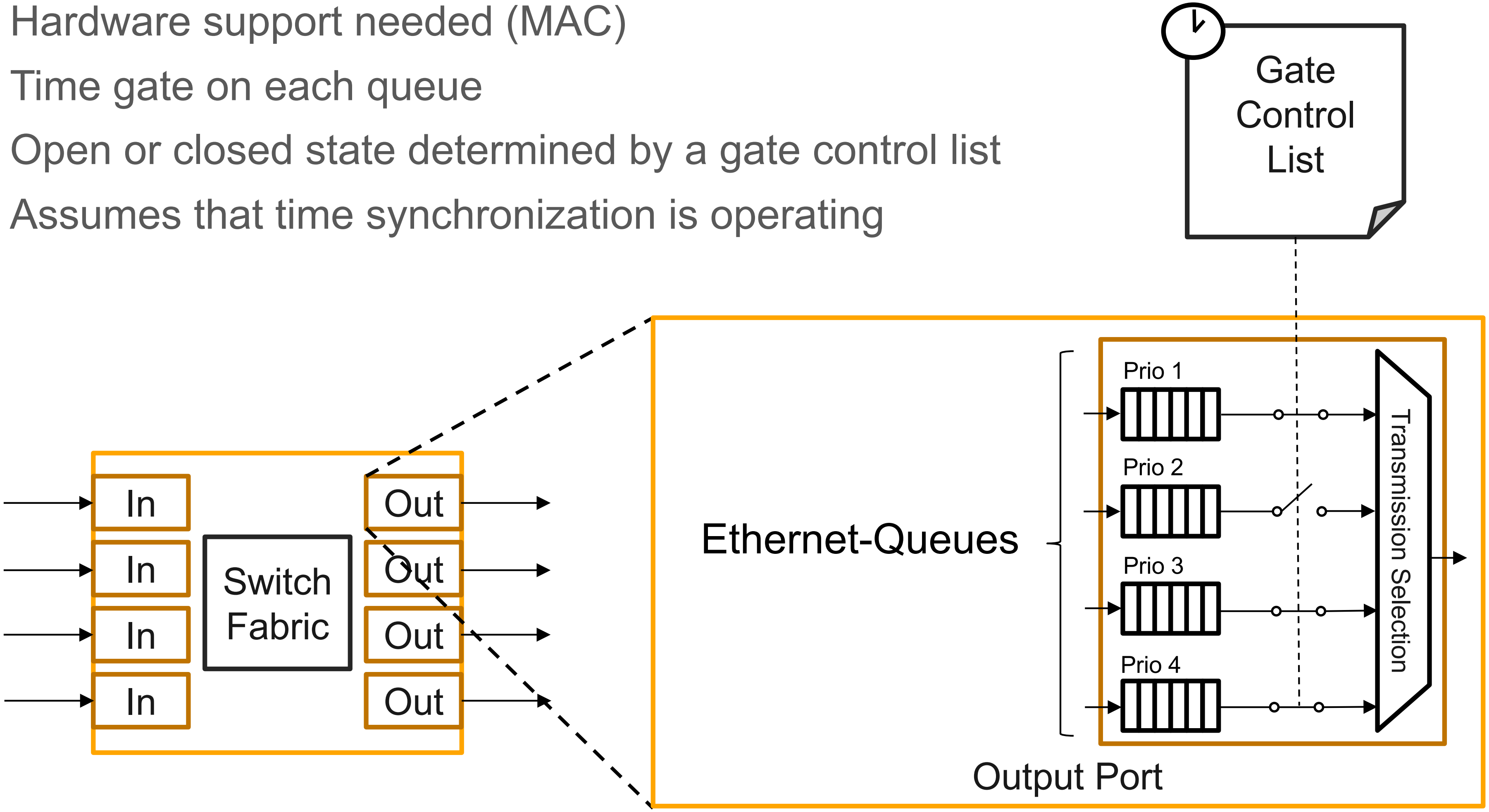
For engineered networks

- Use SRP to establish data paths and bandwidth reservations once
  - Then program components with the resulting configuration
- “Manual” static configuration or network design tool

Other 802.1 Real-Time Packet Scheduling

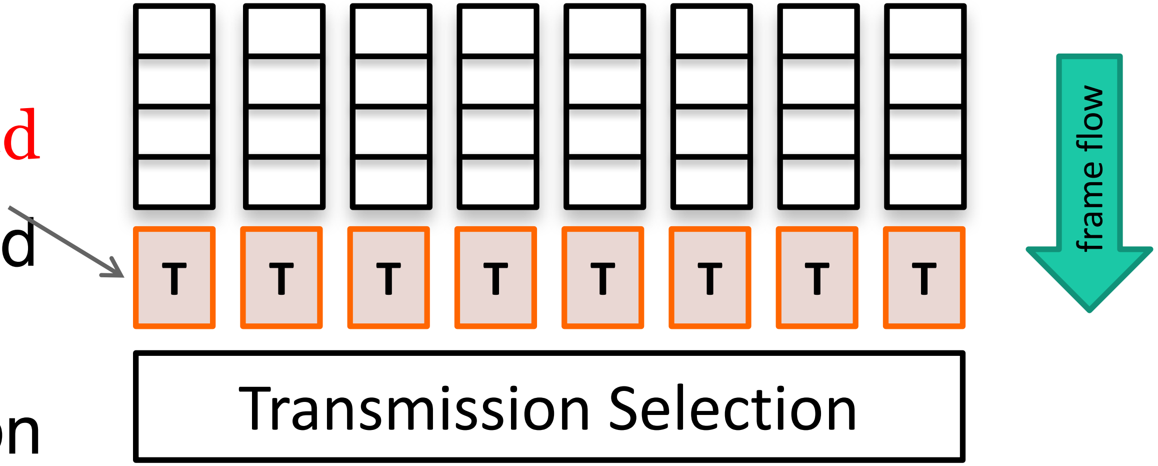
IEEE 802.1Qbv (TAS: Time Aware Shaper)

- Hardware support needed (MAC)
- Time gate on each queue
- Open or closed state determined by a gate control list
- Assumes that time synchronization is operating

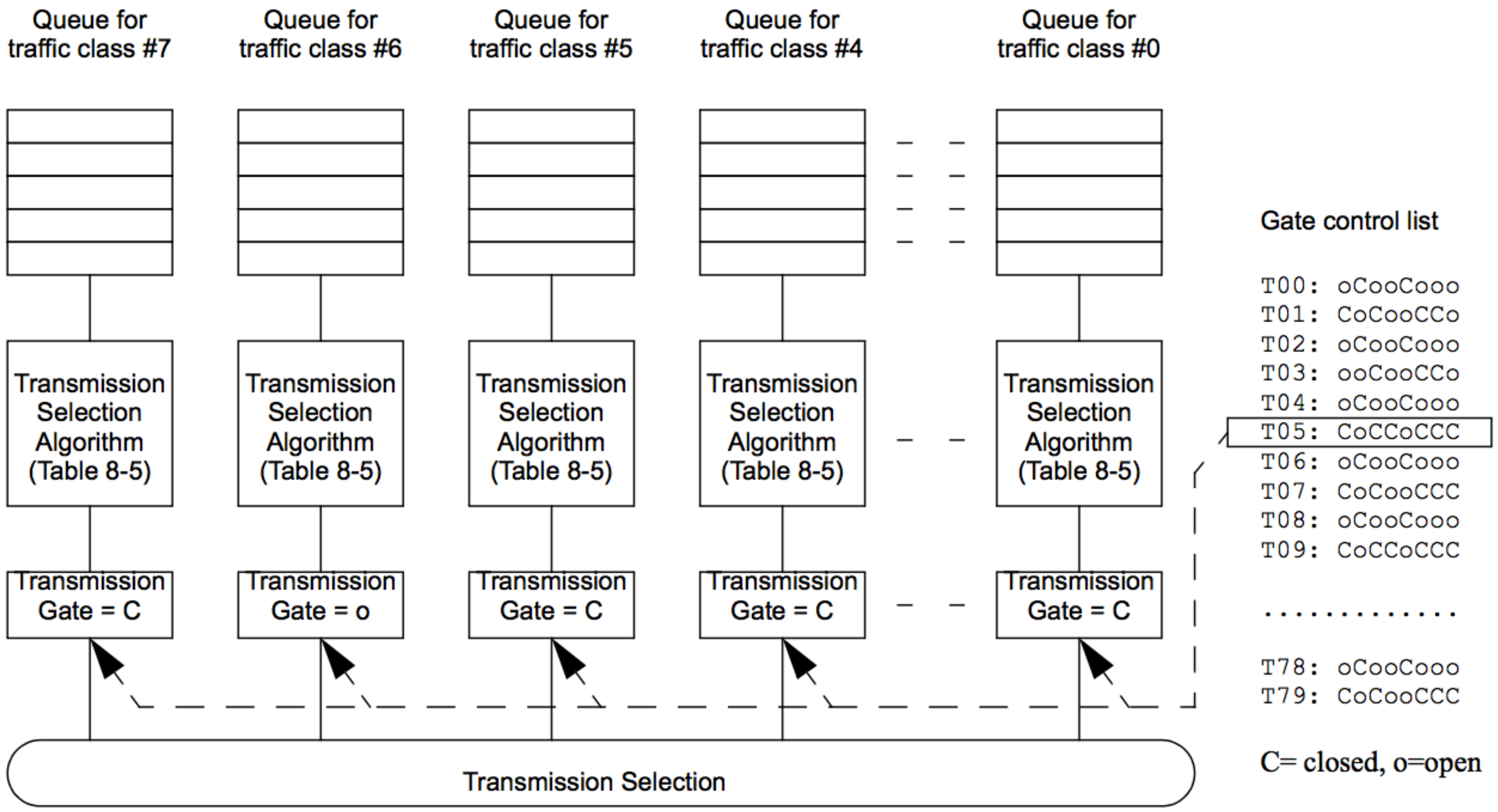


Scheduled Traffic

- Reduces latency variation for Constant Bit Rate (CBR) streams, which are periodic with known timing
- Time-based control/programming of the 8 bridge queues (802.1Qbv)
- Time-gated queues
- Gate: Open or Closed
- Periodically repeated time-schedule
- Time synchronization is needed



Queuing and Packet Prioritization


Queue for traffic class #7
Queue for traffic class #6
Queue for traffic class #5
Queue for traffic class #4
Queue for traffic class #0
Transmission Selection Algorithm (Table 8-5)
Transmission Selection Algorithm (Table 8-5)
Transmission Selection Algorithm (Table 8-5)
Transmission Selection Algorithm (Table 8-5)
Transmission Selection Algorithm (Table 8-5)
Transmission Gate = C
Transmission Gate = o
Transmission Gate = C
Transmission Gate = C
Transmission Gate = C
Transmission Selection

Gate control list

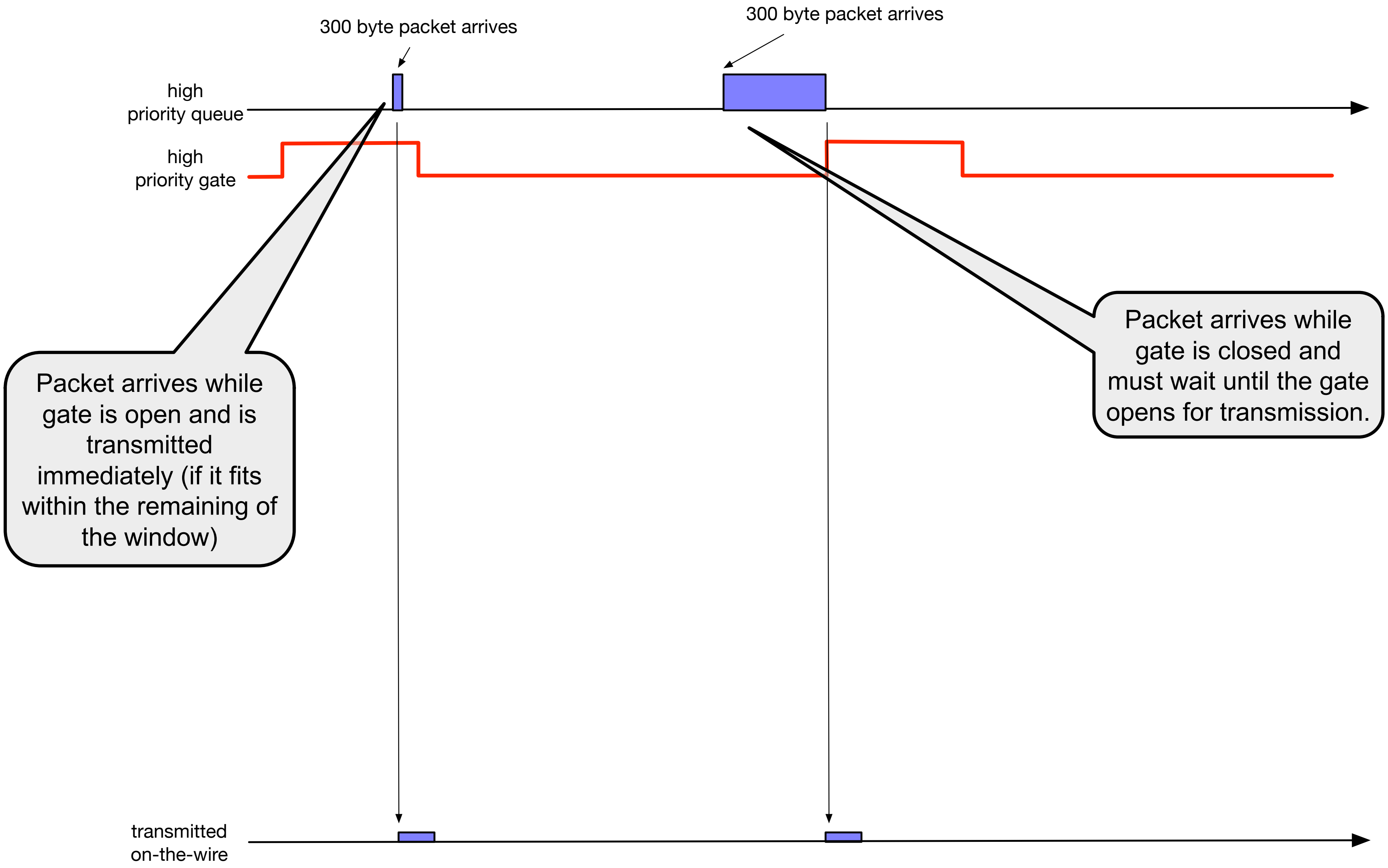
```
T00: oCooCooo
T01: CoCooCCo
T02: oCooCooo
T03: ooCooCCo
T04: oCooCooo
T05: CoCCoCCC
T06: oCooCooo
T07: CoCooCCC
T08: oCooCooo
T09: CoCCoCCC

.............

T78: oCooCooo
T79: CoCooCCC
```

C= closed, o=open

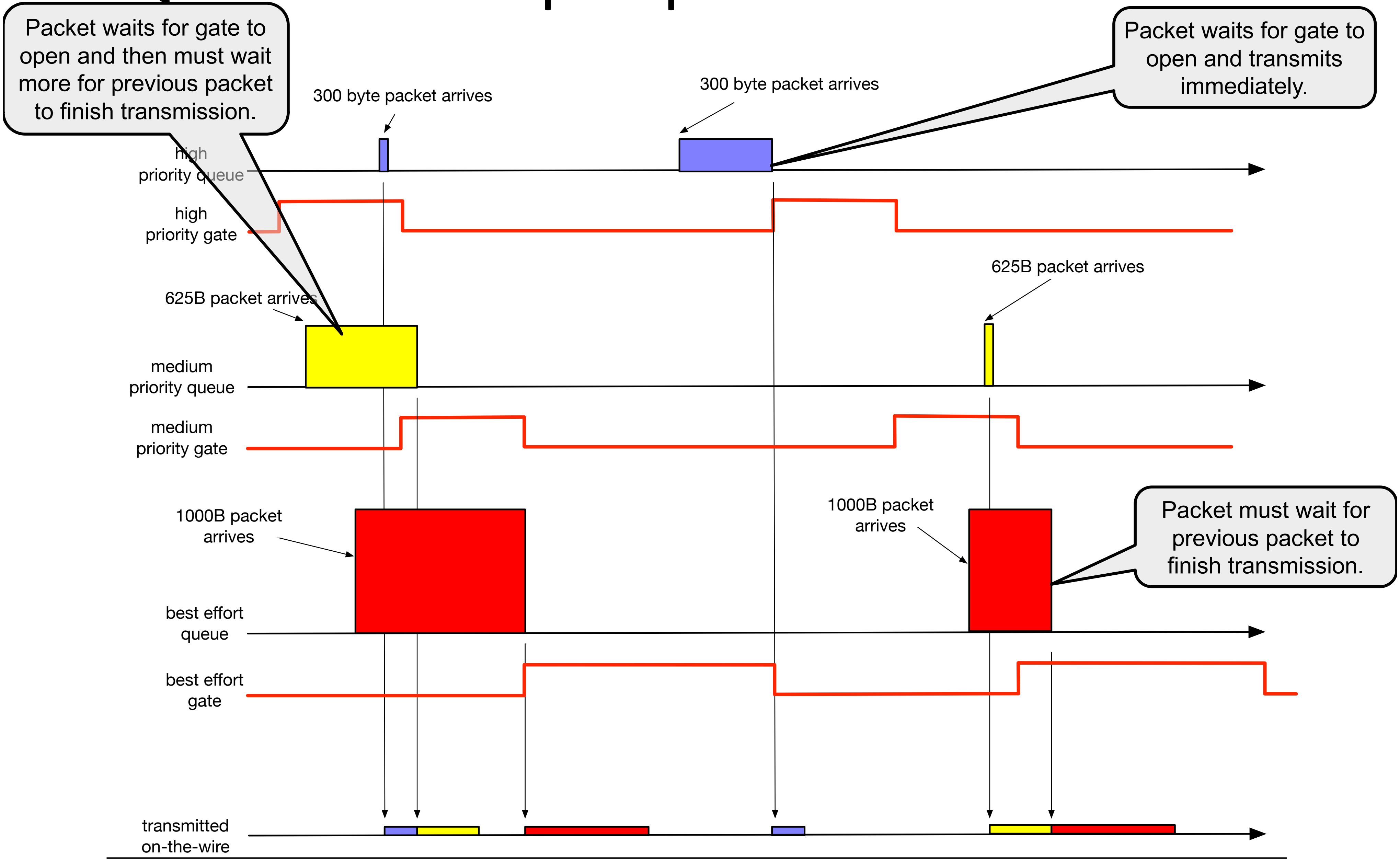
Time-Aware Shaper (802.1Qbv)



Packet arrives while gate is open and is transmitted immediately (if it fits within the remaining of the window)

Packet arrives while gate is closed and must wait until the gate opens for transmission.

Qbv with multiple queues



Packet waits for gate to open and then must wait more for previous packet to finish transmission.

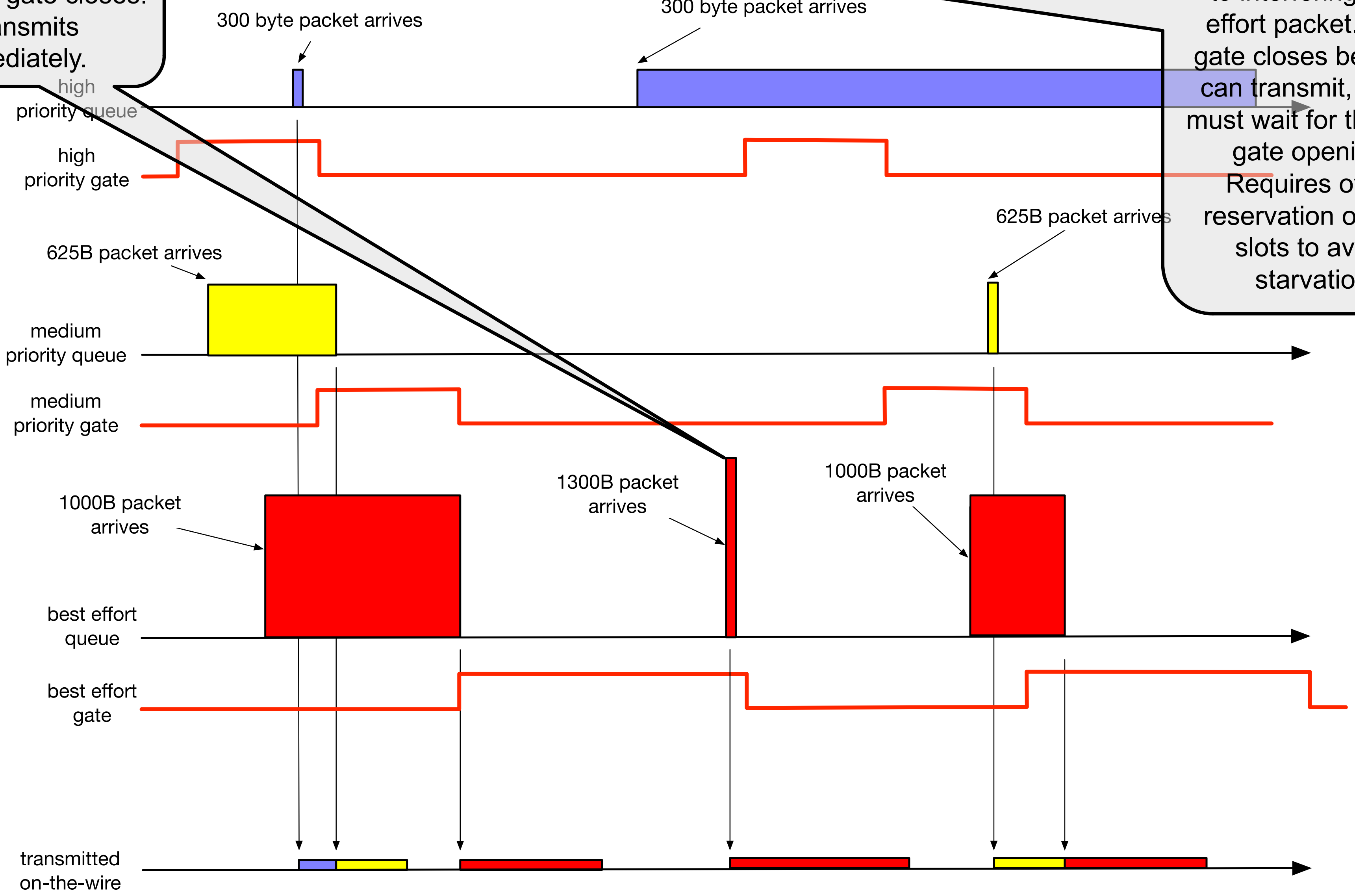
Packet waits for gate to open and transmits immediately.

Packet must wait for previous packet to finish transmission.

Qbv – "Slot Slop"

Best effort packet arrives just before the best effort gate closes. It transmits immediately.

The high priority packet waits for its gate to open. But it still can't transmit due to interfering best effort packet. The gate closes before it can transmit, and it must wait for the next gate opening. Requires over-reservation of time slots to avoid starvation



802.1Qbv observations

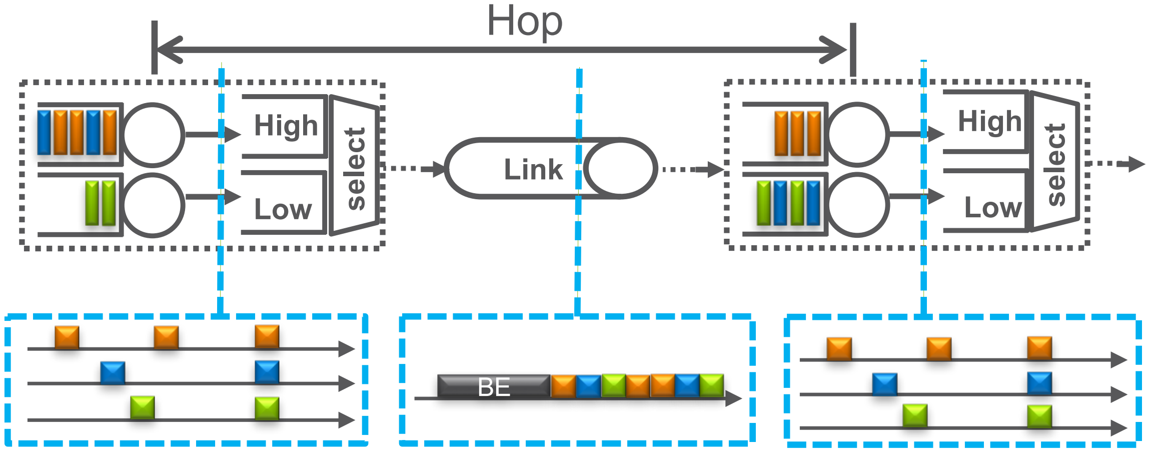
- Pre-defined time access to queues
- Suitable for highly engineered networks
- Suitable for carrying streams with common and regular structure (e.g., sensors that send small packets at very regular periodicity)

802.1Qbv observations

- Engineering the network can be difficult: depending on stream makeup, queue scheduling can be difficult to optimize or create
- Slot slop has to be avoided through careful engineering
  - "Guard" time slots (reduces network efficiency)
  - Eliminate non-engineered traffic
  - Use MACs with frame preemption capability (802.1Qbu-2016 / 802.3br)
- Likely need to synchronize software execution on nodes to avoid missing open Qbv windows

Asynchronous Traffic Shaping

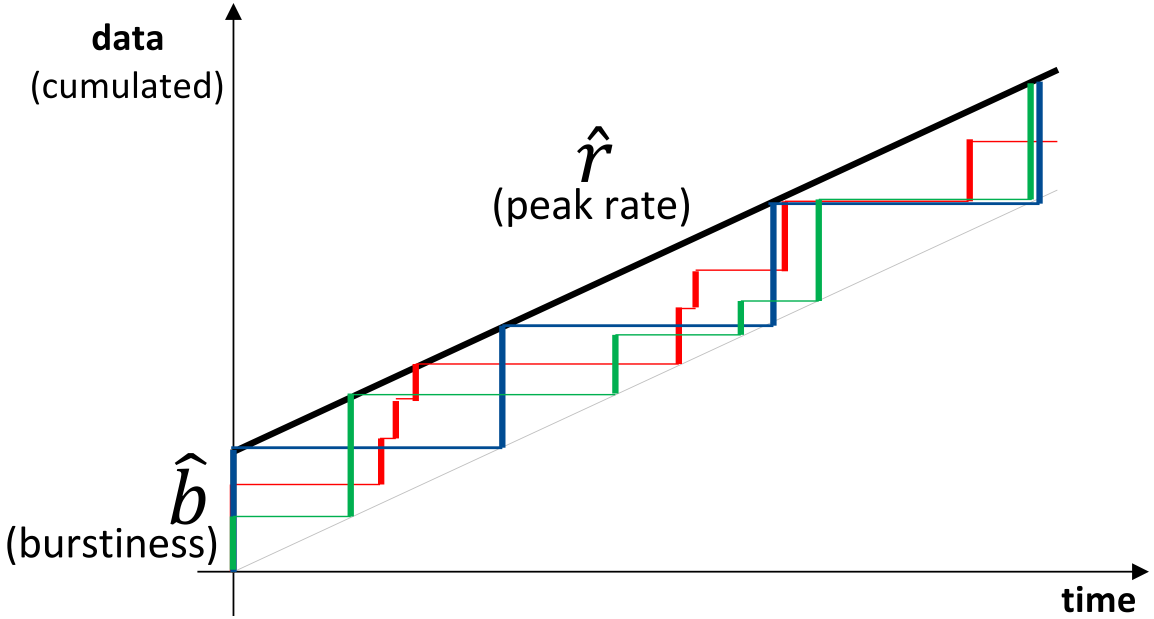
- Zero congestion loss without time synchronization
- Asynchronous Traffic Shaping (P802.1Qcr ATS)
  - Smoothen traffic patterns by re-shaping per hop
  - Prioritize urgent traffic over relaxed traffic



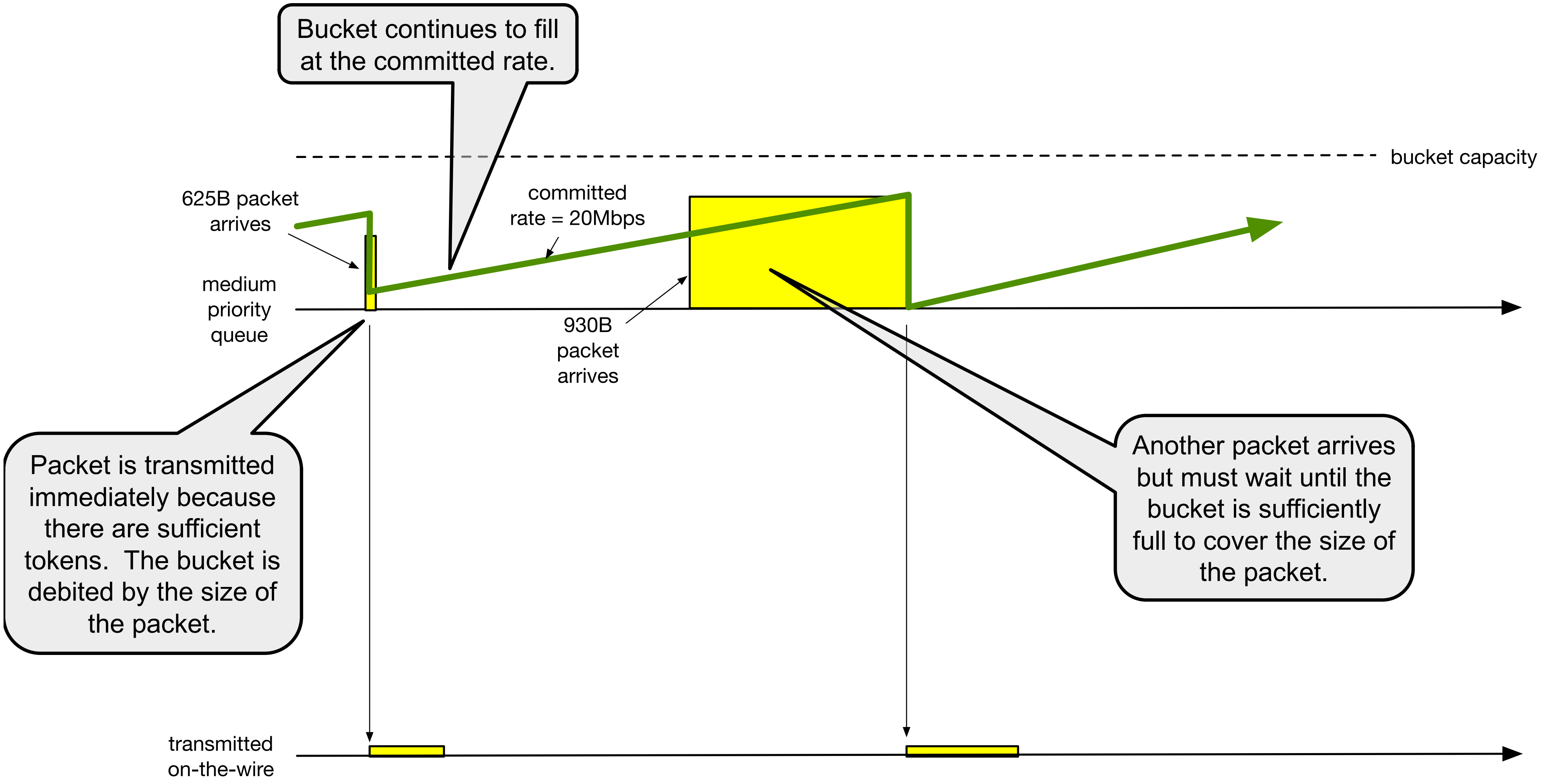
P802.1Qbr Asynchronous Traffic Shaping

- ATS is based on the token bucket algorithm
- Shaper has a token bucket that fills with tokens at a committed bit rate, until it reaches the maximum capacity
- Packets arrive at random times and with random size
- Shaper releases packets for transmission scheduling when the bucket holds tokens greater than or equal to the size of the packet, followed by incrementing the bucket size
- If there is an insufficient number of tokens to release a packet for transmission scheduling after a maximum residence time has elapsed, the shaper discards the packet

Traffic model



ATS operation

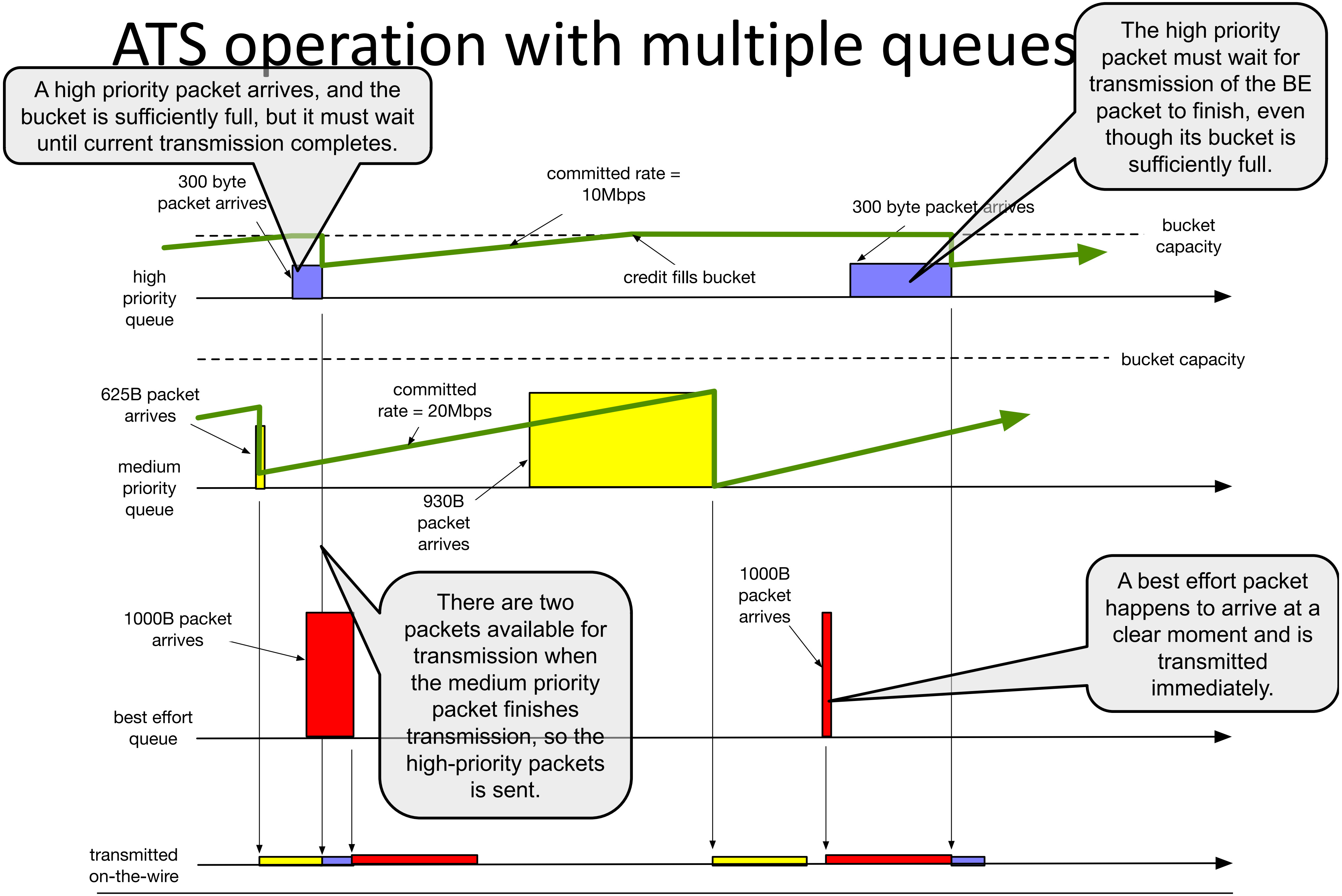


Bucket continues to fill at the committed rate.

Packet is transmitted immediately because there are sufficient tokens. The bucket is debited by the size of the packet.

Another packet arrives but must wait until the bucket is sufficiently full to cover the size of the packet.

ATS operation with multiple queues



A high priority packet arrives, and the bucket is sufficiently full, but it must wait until current transmission completes.

The high priority packet must wait for transmission of the BE packet to finish, even though its bucket is sufficiently full.

There are two packets available for transmission when the medium priority packet finishes transmission, so the high-priority packets is sent.

A best effort packet happens to arrive at a clear moment and is transmitted immediately.

Forwarding Processing Delay

- … or ”Store & Forward” delay inside a switch
- Delay between when the last CRC byte was received until the frame is put in an egress queue available to be selected for transmission (e.g., by CBS or TAS)
- 802.1 does not say anything about this delay (no definition and no maximum value for compliance)

Low port count switches

- While MAC receives the Ethernet frame, it transmits a byte stream to the switch core
- Switch core buffers the bytes in global memory (not yet known whether it's a valid frame)
- CRC is calculated over the stream on the fly for future comparison with the last 4 bytes (the FCS)
- Once MAC has signaled IPG to the switch core, the comparison is performed to determine whether or not the frame is valid
- Pointer to the buffer space (plus metadata extracted from header) is passed to the egress ports)
- Low tens of clock cycles for 1 Gbps switches (125 Mhz)

It is not as “simple” for all switches

- Higher speeds (e.g., 10 Gbps)
  - Cannot bump up the clock frequency to 1.25 GHz due to power limitations
- High port-count switches
  - Pack two 5-port switches to create an 8 port switch (but it creates a huge bottleneck for the cascaded connection)
  - Connect multiple 3 or 4 port switches in an internal ring topology (complicated analysis of the arbitration scheme that has to be created)
  - Use output port buffers in addition to global memory

Summary: Forwarding Processing Delay

- It is complicated to calculate a worst-case S&F delay, especially in case of many ports and in case of multi-Gpbs networks
- It is highly switch and vendor specific
- 802.1 does not specify internal switch implementations (after all, this is where the Intellectual Property is created)
- Solutions?
  - Make pessimistic assumptions (anyway, this delay is typically orders of magnitude lower than the dynamics of the application)
  - Ask vendors for performance numbers
  - Use measurement equipment

Summary

- AVB
  - Credit based shaping
  - Stream reservation protocol
  - or static configuration
- Other shapers
  - Time aware shaper (TAS) / Scheduled Traffic
  - Asynchronous Traffic Shaper
- Forwarding delay inside a switch is difficult to bound
  - Vendor specific; needs to be included in datasheet
  - Not standardized in IEEE 802.1