

Module 3

Quality of Services QoS/Part 1

Prepared by Dr. Manal Mizher

Lecture 5

جامعة الجميع الذكية

تسعى لتكون الكيان الأول الذي يتبنى مبدأ المجتمع التعليمي
التعاوني ويقدم العلم مقابل العمل لتنمية وبناء المجتمعات.

Network programming

Course ID: CYB615



جامعة الجميع الذكية
EVERYONE'S SMART UNIVERSITY

1. What is Quality of Service (QoS) in Networking?
(Congestion, Common Services for QoS)
2. How Does QoS Work?
3. Why is QoS Important?
4. Types of Network Traffic & Flow classes
5. QoS Tools
6. Techniques to Improve QoS
7. What Best Practices Are Involved in QoS?
8. Advantages of QoS

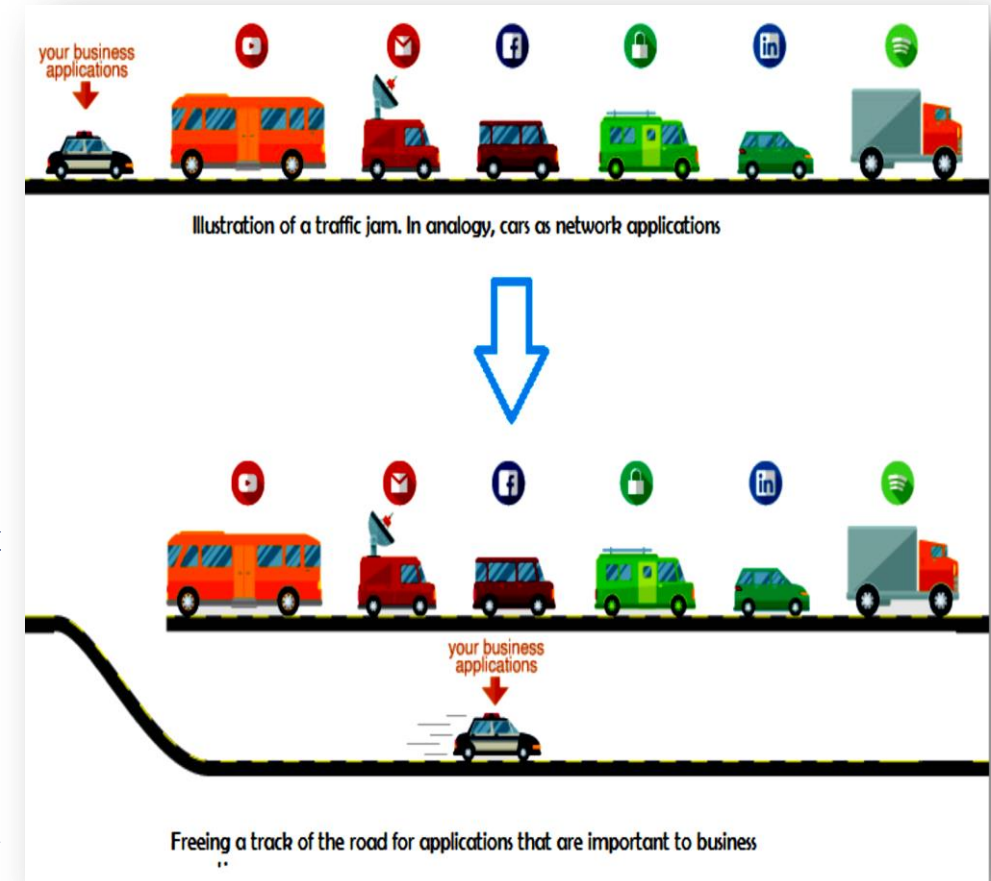
What is Quality of Service (QoS) in Networking?

- Quality of service (QoS) is the use of mechanisms or technologies that work on a network to control traffic and ensure the performance of critical applications with limited network capacity.
- It enables organizations to adjust their overall network traffic by *prioritizing specific high-performance applications*.
- QoS is typically applied to networks that carry traffic for resource-intensive systems (heavy used).

- What is meant by congestion? Why the congestion occur in network?

Network node or link is carrying more data than it can handle.

- 1) in cases of traffic overloading when a link or network node is handling data in excess of its capacity.
- 2) user sends data into the network at a rate greater than that allowed by network resources.
- 3) the switches in a network have a limited buffer size to store arrived packets.
- 4) As the data rate can be different among different networks, congestion may happen to both wide area network (WAN) and local area network (LAN).

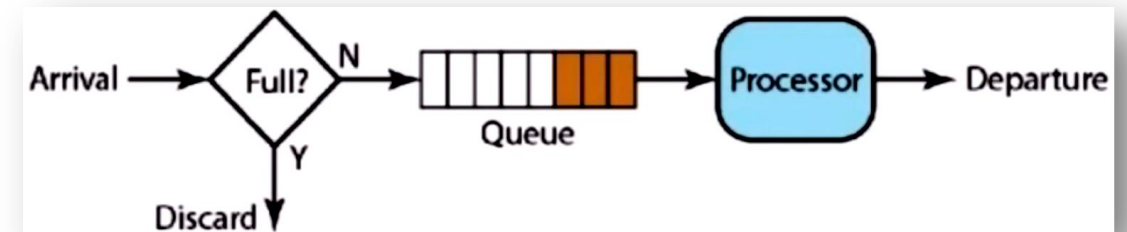
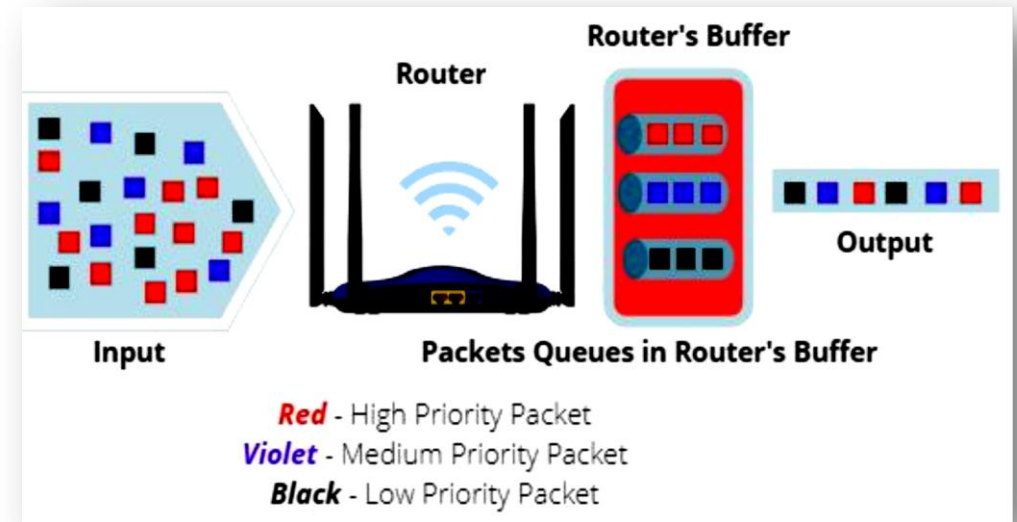


- *internet protocol television (IPTV), online gaming, streaming media, videoconferencing, video on demand (VOD), and Voice over IP (VoIP).*



What is QoS queue?

- QoS provides switches and routers with a mechanism to queue and service higher priority traffic before lower priority traffic.
- QoS also provides a mechanism to drop lower priority traffic before higher priority traffic, during periods of congestion.



- Some organizations may only need to:
 - configure bandwidth limits for specific services.
 - fully configure interface and security policy bandwidth limits for all their services.
- Traffic can be classified by :
 - 1) port
 - 2) Internet protocol (IP) addresses
 - 3) application
 - 4) user

Why is QoS Important?

- 1) When computer used a telephone line to access the internet (dial up).
- 2) Interactive applications carrying audio and video content need to be delivered at high speed, without packet loss or variations in delivery speed.
- 3) Real-time traffic, such as VoIP and videoconferencing.
- 4) QoS helps businesses prevent the delay of these sensitive applications, ensuring they perform to the level that users require.
- 5) Organizations need to deploy techniques and technologies that guarantee the best possible service for specified users.

Why is QoS Important?

- 6) QoS is also becoming increasingly important as the Internet of Things (IoT) continues to grow.

For example, in the manufacturing sector, machines now leverage networks to provide real-time status updates on any potential issues. Therefore, any delay in feedback could cause highly costly mistakes in IoT networking. QoS enables the data stream to take priority in the network and ensures that the information flows as quickly as possible.

<https://www.simplilearn.com/iot-devices-article>

Understanding how QoS network software works is depend on defining the various types of traffic that it measures. These are:



Bandwidth

Bandwidth: The speed of a link. QoS can tell a router how to use bandwidth. For example, assigning a certain amount of bandwidth to different queues for different traffic types.



Delay

Delay: The time it takes for a packet to go from its source to its end destination. This can often be affected by queuing delay, which occurs during times of congestion and a packet waits in a queue before being transmitted. *QoS enables organizations to avoid this by creating a priority queue for certain types of traffic.*



Loss

Loss: The amount of data lost as a result of packet loss, which typically occurs due to network congestion. *QoS enables organizations to decide which packets to drop in this event.*



Jitter

Jitter: The irregular speed of packets on a network as a result of congestion, which can result in packets arriving late and out of sequence. This can cause distortion or gaps in audio and video being delivered.

Reliability: lack of reliability means losing a packet or acknowledgement, which means entails retransmission. Different application programs need different levels of reliability.

ATM Forum Traffic Management 4.0 ATM Service Category	Typical Use
Constant Bit Rate (CBR)	Real-time, QoS guarantees , uncompressed video and audio
Real-Time Variable Bit Rate (rt-VBR)	Tight limit on delay i.e. used for time sensitive applications, real time, LIVE TV Transmission having video information
Non-Real-Time Variable Bit Rate (nrt-VBR)	remote terminal access applications, Used for text/image transfer, messaging, retrieval and distribution applications
Available Bit Rate (ABR)	Resource exploitations, feedback control, LAN to LAN connection for information transfer
Unspecified Bit Rate (UBR)	Best effort, no guarantees
(No equivalent)	Burst level feedback control

- **There are three categories of QoS tools:**
 - 1) Classification and marking tools
 - 2) Congestion avoidance tools
 - 3) Congestion management tools
- **QoS networking technology works by:**
 - 1) marking packets to identify service types.
 - 2) then configuring routers to create separate virtual queues for each application, based on their priority (specific flows in network traffic).
 - 3) as a result, bandwidth is reserved for critical (applications, websites, or traffic flow) that have been assigned priority access.

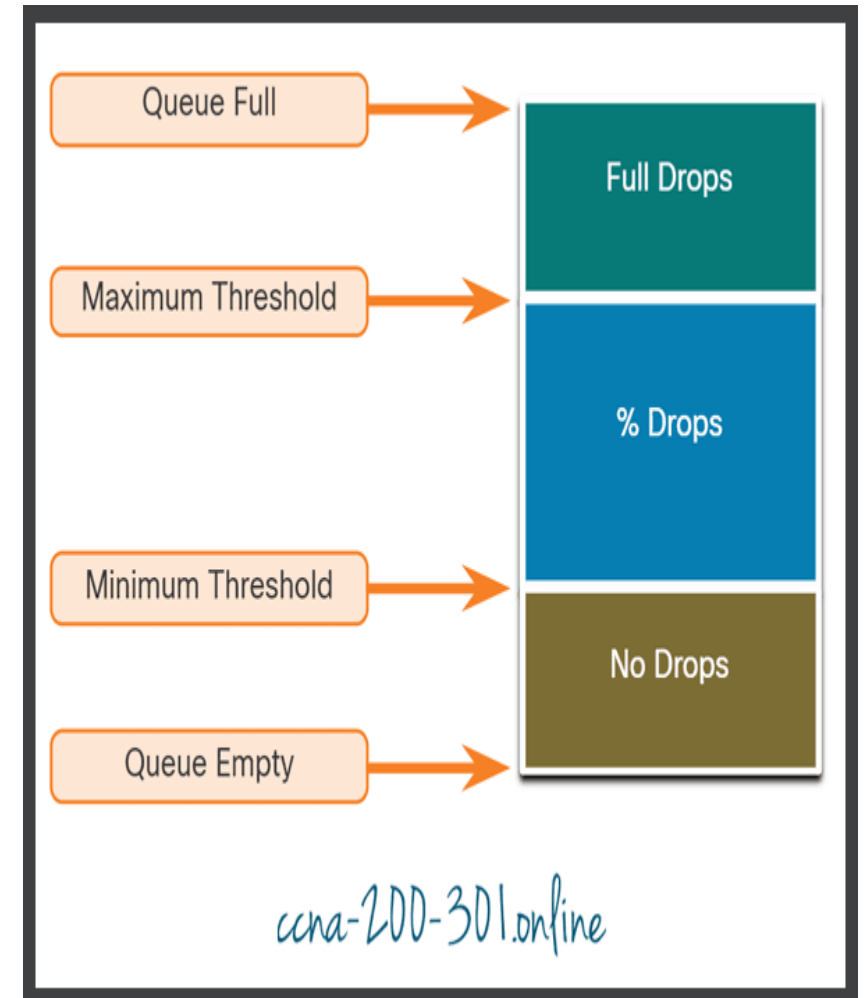


- QoS classification tools categorize packets by examining the contents of the frame, and packet headers.
- Marking tools allow the QoS tool to change bits in the packet header for easier classification.
- Most QoS tools classify traffic, which allows for each class of traffic to receive a different level of treatment from other traffic classes. These different types or classes of traffic are typically called ***service classes*** in QoS terminology. For instance:
 - Marking tools change the marking in voice packets to ensure that a classification tool can differentiate a voice packet from a data packet.
 - Classification tools place voice and data traffic in separate queues.

- **Classification with NBAR** – NBAR classifies packets that are normally difficult to classify.

For instance, some applications use dynamic port numbers, so a statically configured match command, looking for a particular UDP or TCP port number, simply could not classify the traffic. NBAR can look past the UDP and TCP header, and refer to the host name, URL, or MIME type in HTTP requests. This deeper examination of the packet contents is sometimes called deep packet inspection.

- Congestion avoidance tools monitor network traffic loads in an effort to anticipate and avoid congestion at common network and internetwork bottlenecks before congestion becomes a problem.
- 1) These tools can monitor the average depth of the queue.
 - 2) When the queue is below the minimum threshold, there are no drops.
 - 3) As the queue fills up to the maximum threshold, a small percentage of packets are dropped.
 - 4) When the maximum threshold is passed, all packets are dropped.



- For example, Cisco IOS QoS includes weighted random early detection (WRED) as a possible congestion avoidance solution:
 - 1) The WRED algorithm allows for congestion avoidance on network interfaces by providing buffer management and allowing TCP traffic to decrease, or throttle back, before buffers are exhausted (**connection-oriented/heavy/Stream**).
 - 2) Using WRED helps avoid tail drops and maximizes network use and TCP-based application performance.
 - 3) There is no congestion avoidance for User Datagram Protocol (UDP)-based traffic, such as voice traffic. In case of UDP-based traffic, methods such as queuing and compression techniques help to reduce and even prevent UDP packet loss (**connectionless/lightly/Datagram <= MTU**).

- Only when an interface is congested will the queuing theory begin to work. For instance, system supports class-based weighted fair queuing (CBWFQ) and low latency queuing (LLQ):
 - 1) **CBWFQ:** is a scheduling mechanism where each of the CBWFQ queues is assigned a weight, and the packets are served from the queues based upon the weight of the queue.
 - 2) **LLQ:** usually used in voice and interactive video. During configuration, all the applications of LLQ type can occupy no more than 33% of the total bandwidth.

Congestion control in TCP (An example)

- Packet from a sender may pass through several routers before reaching its final destination.
- Router has a buffer that stores the incoming packets, processes them, and forwards them.
- If a router receives packets faster than it can process, congestion might occur and some packets could be dropped.
- When a packet does not reach the destination, no acknowledgement is sent for it.
- The sender has no choice but to retransmit the lost packet. This may create more congestion and more dropping of packets, which means more retransmission and more congestion.
- A point may then be reached in which the whole system collapses and no more data can be sent. TCP therefore needs to find some way to avoid this situation.
- If the network cannot deliver the data as fast as they are created by the sender, it needs to tell the sender to slow down. In other words, in addition to the receiver, the network is a second entity that determines the size of the sender's window in TCP.
- If the cause of the lost segment is congestion, retransmission of the segment does not remove the cause – it aggravates it.

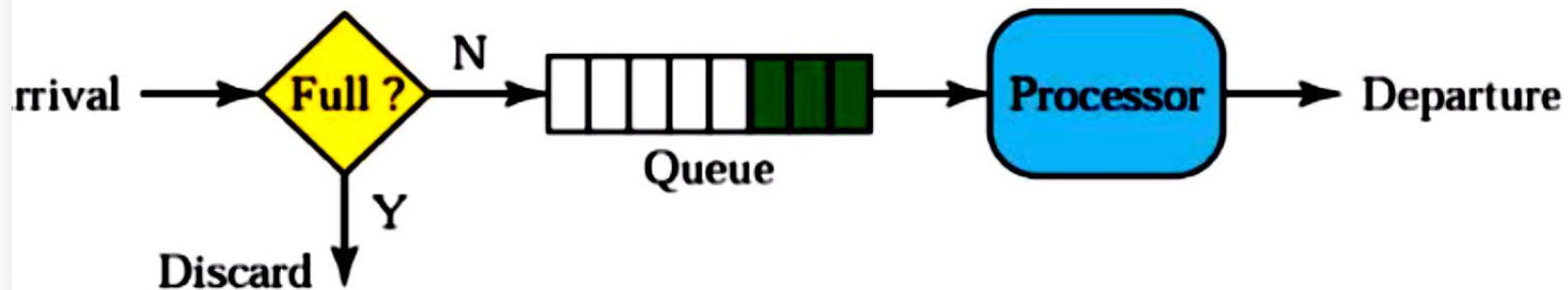
QoS Tools	Description
Classification and marking tools	•Sessions, or flows, are analyzed to determine <u>what traffic class they belong to</u>. •When the traffic class is determined, the packets are marked. •<u>Marking involves setting some bits inside a data link or network layer header</u>, with the goal of letting other devices' QoS tools classify traffic based on the marked values.
Congestion avoidance tools	•Traffic classes are allotted portions of network resources, as defined by the QoS policy. •The QoS policy also identifies how some traffic may be selectively dropped, delayed, or re-marked to avoid congestion. •The primary congestion avoidance tool is WRED and is used to regulate TCP data traffic in a bandwidth-efficient manner before tail drops caused by queue overflows occur.
Congestion management tools	•When traffic exceeds available network resources, traffic is queued to await availability of resources. •Common Cisco IOS-based congestion management tools include CBWFQ and LLQ algorithms.

- Some techniques that can be used to improve the quality of service. The four common methods: scheduling, traffic shaping, resource reservation, and admission control.

- **Scheduling**
 - FIFO Queuing
 - Priority Queuing
 - Weighted Fair Queuing
- **Traffic Shaping**
 - Leaky Bucket
 - Token Bucket
 - Combination of Leaky Bucket and Token Bucket.
- **Resource Reservations**
- **Admission Control**

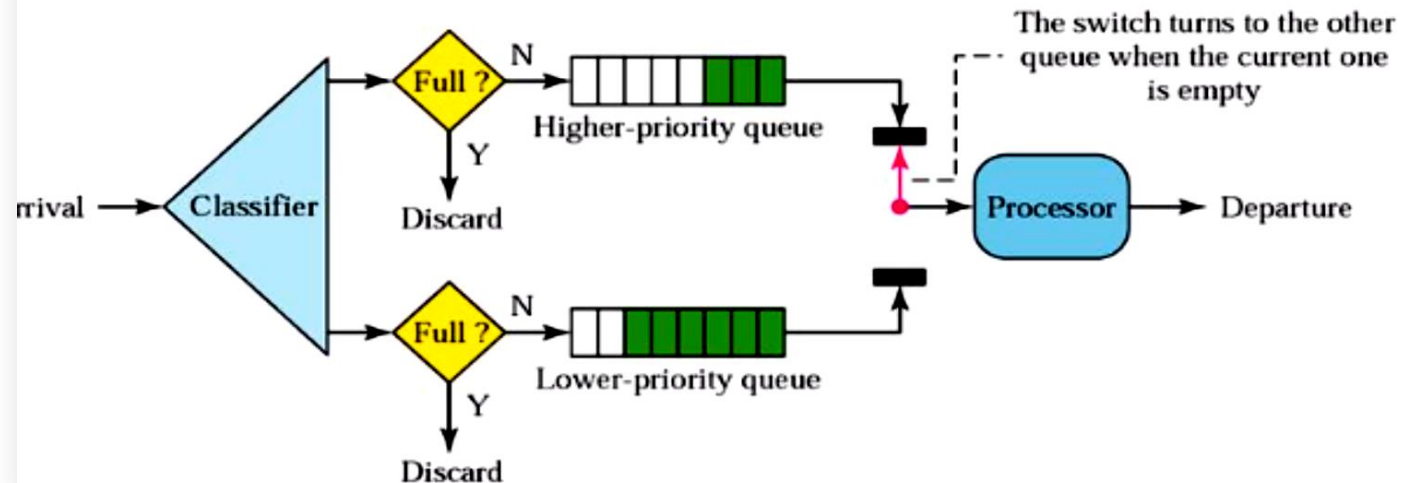
- Packets from different flows arrive at a switch or router for processing. A good scheduling technique treats the different flows in a fair and appropriate manner. Several scheduling techniques are designed to improve the quality of service:
 - 1) FIFO queuing
 - 2) priority queuing
 - 3) weighted fair queuing.

- **FIFO Queuing:**
 - First-in first-out queuing
 - Packets wait in a buffer (queue) until the node (router or switch) is ready to process them.
 - If average arrival rate is higher than average processing rate, the queue will fill up and new packets will be discarded.



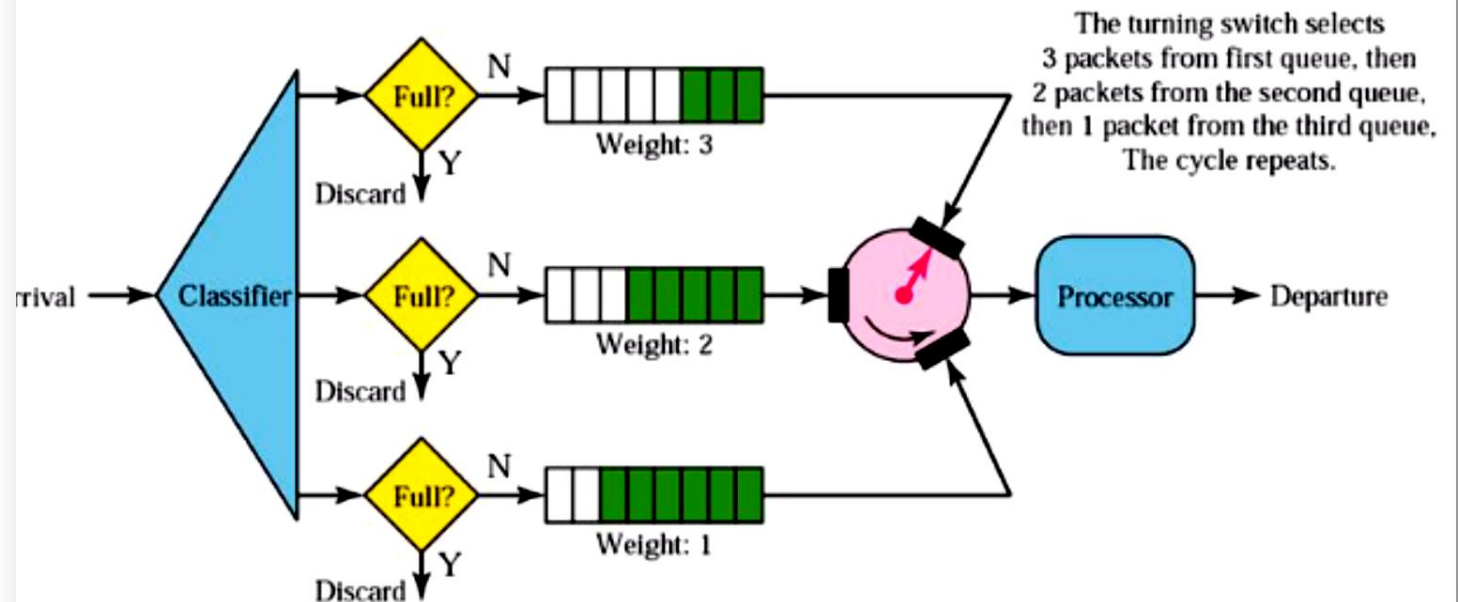
Scheduling (Priority Queuing)

- Packets are first assigned to a priority class.
- Each priority class has its own queue.
- Packets in highest-priority queue are processed first. Packets in lowest-priority queue are processed last.
- System does not stop serving a queue until it is empty.
- Good for multimedia traffic.
- Starvation is possible: If there is a continuous flow in a high-priority queue, the packets in lower-priority queues will never have a chance to be processed.



- A better scheduling method is weighted fair queuing.
- For example, if the weights are 3, 2, and 1, three packets are processed from the first queue, two from the second queue, and one from the third queue.

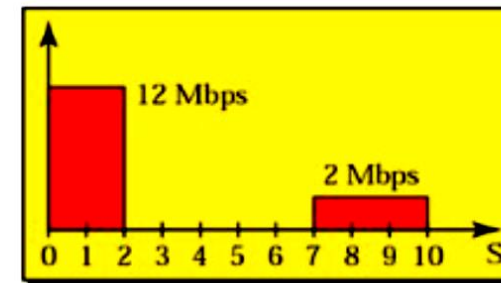
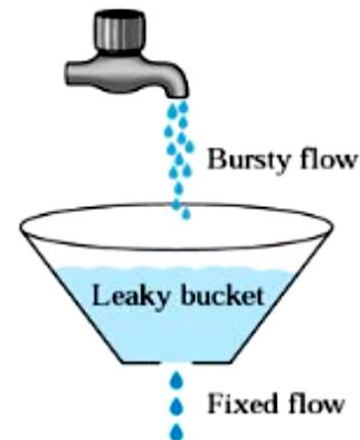
- Packets are assigned to different classes and admitted to different queues.
- System processes packets in each queue in round-robin fashion with the number of packets selected from each queue based on the corresponding weight.



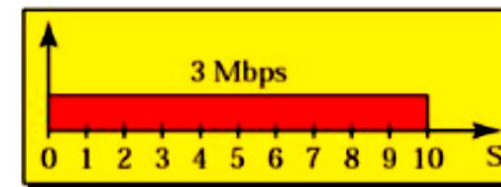
- Traffic shaping: Mechanism to control the amount and rate of the traffic sent to the network.
- Two Techniques for Traffic Shaping
 1. Leaky Bucket
 2. Token Bucket

Traffic Shaping (Leaky Bucket)

- Input rate varies; Output rate is fixed.
- Using FIFO queue, if traffic consists of fixed-size packets, the process removes a fixed number of packets from the queue at each tick of the clock. If traffic consists of variable-length packets, the fixed output rate must be based on the number of bytes or bits.



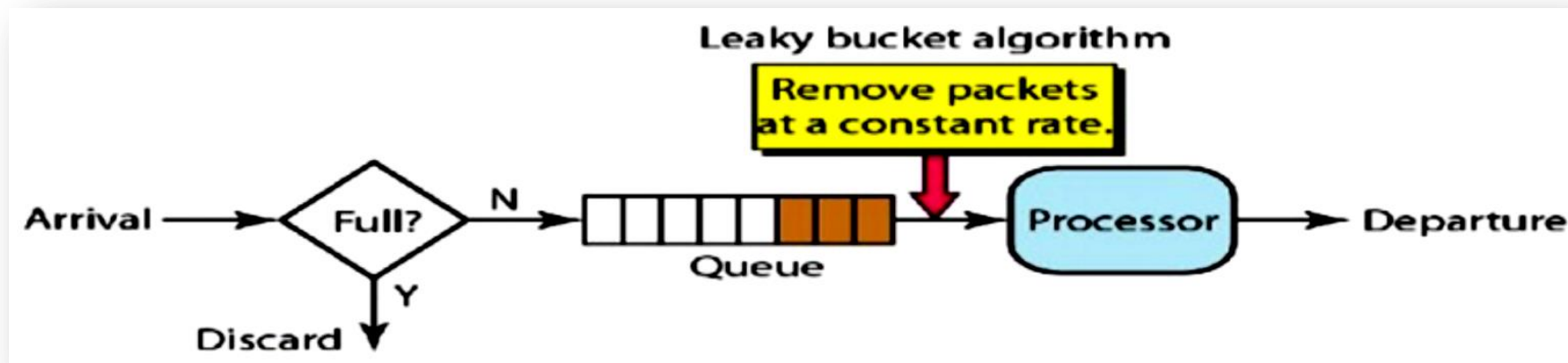
Bursty data



Fixed-rate data

Leaky Bucket implementation(example)

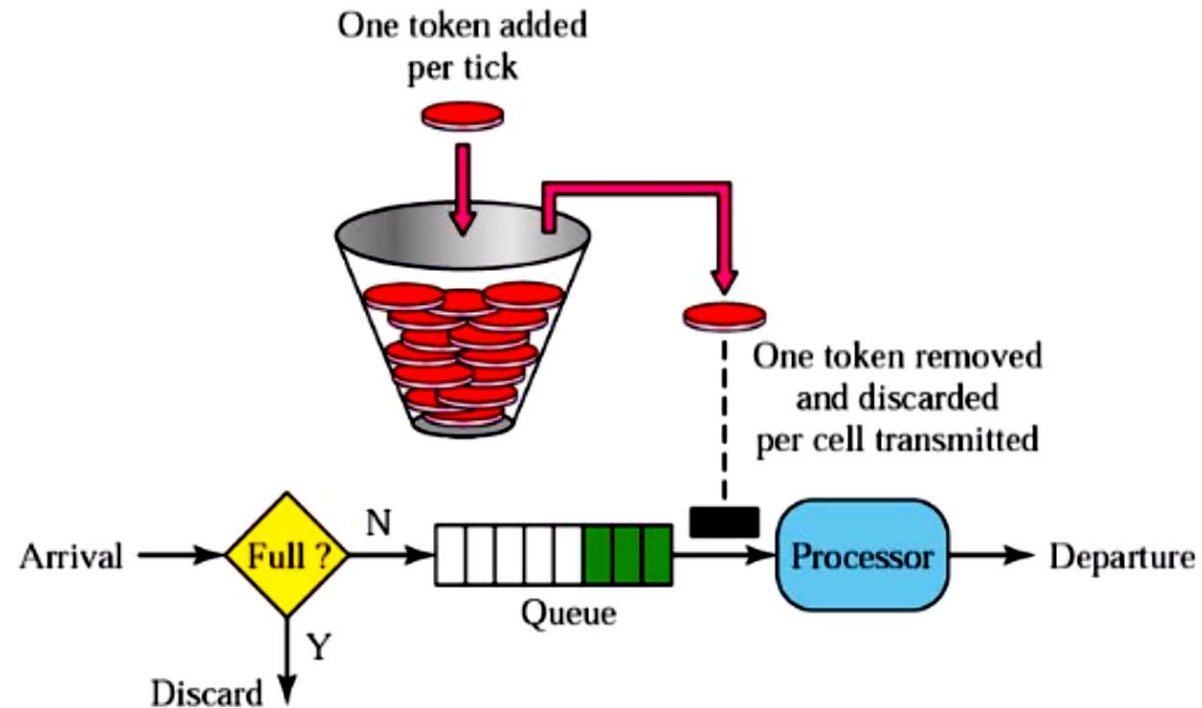
- Assume that the network has committed a bandwidth of 3 Mbps for a host.
- the host sends a burst of data at a rate of 12 Mbps for 2 s, for a total of 24 Mbits of data.
- The host is silent for 5 s and then sends data at a rate of 2 Mbps for 3 s, for a total of 6 Mbits of data.
- In all, the host has sent 30 Mbits of data in 10s.
- The leaky bucket smooth's the traffic by sending out data at a rate of 3 Mbps during the same 10 s.



very restrictive

- Leaky bucket does not credit an idle host.
- If a host is not sending for a while, bucket becomes empty.
- The idle time of a host is not considered in leaky bucket.
- In token bucket, idle hosts accumulate credit for the future in the form of tokens.
- The token bucket can easily be implemented with a counter. The token is initialized to zero. Each time a token is added, the counter is incremented by 1.
- Each time a unit of data is sent, the counter is decremented by 1. When the counter is zero, the host cannot send data.

Traffic Shaping (Token Bucket)

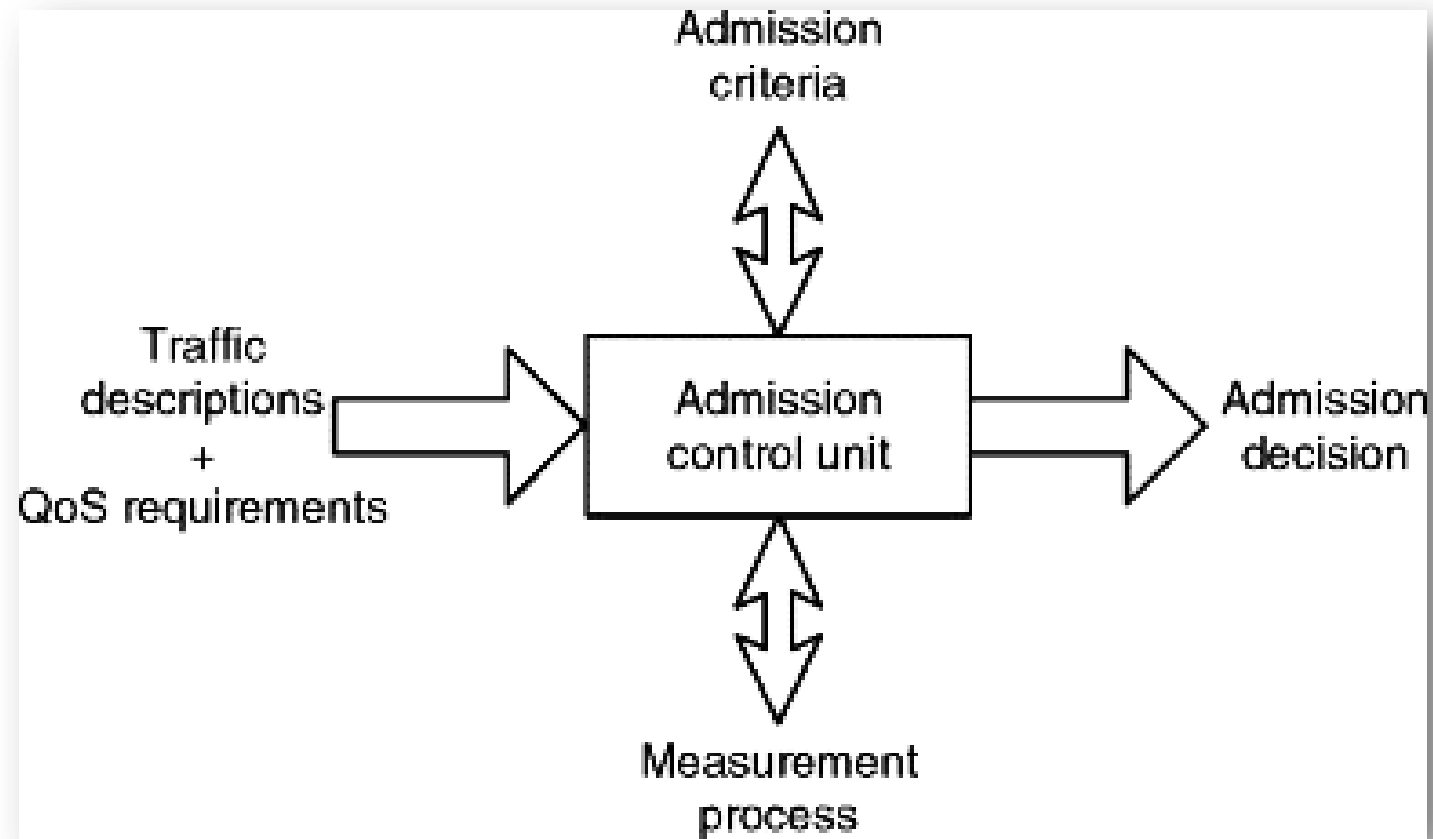


Combining Token Bucket and Leaky Bucket:

Leaky bucket is applied after the token bucket; the rate of the leaky bucket to be higher than the rate of tokens dropped in the bucket

- A flow of data needs resources such as a buffer, bandwidth, CPU time, and so on. The quality of service is improved if these resources are reserved beforehand. One QoS model called **Integrated Services**.
- **Integrated services or IntServ** is an architecture that specifies the elements to guarantee quality of service (QoS) on networks. IntServ can for example be used to allow video and sound to reach the receiver without interruption.
 - The Resource Reservation Protocol (RSVP)

- Admission control refers to the mechanism used by a router, or a switch, to accept or reject a flow based on predefined parameters called flow specifications.
 - Before a router accepts a flow for processing, it checks the flow specifications to see if its capacity (in terms of bandwidth, buffer size, CPU speed, etc.) and its previous commitments to other flows can handle the new flow.
- Admission control is a key component for QoS delivery in IP networks because:
 - 1) it determines the extent to which network resources are utilized.
 - 2) it determines whether the contracted QoS characteristics are actually delivered.



- 1) Ensure that maximum bandwidth limits at the source interface and security policy are not set too low to prevent excessive packet discard.
- 2) Consider the ratio at which packets are distributed between available queues and which queues are used by which services. This can affect latency levels, queue distribution, and packet assignment.
- 3) Only place bandwidth guarantees on specific services. This will avoid the possibility of all traffic using the same queue in high-volume situations.
- 4) Configure prioritization for all traffic through either type of service-based priority or security policy priority, not both. This will simplify analysis and troubleshooting.
- 5) Try to minimize the complexity of QoS configuration to ensure high performance.
- 6) To get accurate testing results, use the User Datagram Protocol (UDP), and do not oversubscribe bandwidth throughput.

- 1) **Unlimited application prioritization:** QoS guarantees that businesses' most mission-critical applications will always have priority and the necessary resources to achieve high performance.
- 2) **Better resource management:** QoS enables administrators to better manage the organization's internet resources. This also reduces costs and the need for investments in link expansions.
- 3) **Enhanced user experience:** The end goal of QoS is to guarantee the high performance of critical applications, which is summed up to delivering optimal-user-experience. Employees enjoy high performance on their high-bandwidth applications, which enables them to be more effective and get their job done more quickly.
- 4) **Point-to-point traffic management:** Managing a network is essential however traffic is delivered, be it end to end, node to node, or point to point. The latter enables organizations to deliver customer packets in order from one point to the next over the internet without suffering any packet loss.



- **Point to point communication :**

- 1) the channel is shared between two devices (TCP unicast).
- 2) There is dedicated link between two nodes.
- 3) the entire capacity is reserved between these connected two devices with the possibility of waste of network bandwidth/ resources.
- 4) there is one transmitter and one receiver.
- 5) the smallest distance is most important to reach the receiver.
- 6) provides security and privacy because communication channel is not shared.
- 7) Opposite, Multipoint (multicast, broadcast).

■ End to end communication :

- 1) indicates a communication happening between two applications (maybe you and your friend using Skype).
- 2) It doesn't care what's in the middle, it just consider that the two ends are talking with one another.
- 3) It generally is a Layer 4 (or higher) communication.

■ Hop by hop communication :

- 1) indicates the flow the communication follows.
- 2) Data pass through multiple devices with an IP address, and each is genetically named "hop".
- 3) Hop by Hop indicates analyzing the data flow at layer 3, checking all devices in the path.



- **Node to node communication :**

- 1) In telecommunications, is the movement of data from one node of a network to the next.
- 2) In the OSI model it is handled by the lowest two layers, the data link layer and the physical layer.

- **Port to port communication :**

AS sockets communications.