Quality of Service

So what is it?

- Quality of service (QoS) refers to any technology that manages data traffic to reduce packet loss, latency and jitter on a network.
- QoS controls and manages network resources by setting priorities for specific types of data on the network.

QoS

- Enterprise networks need to provide predictable and measurable services as applications -- such as voice, video and delay-sensitive data -- to traverse a network.
- Organizations use QoS to meet the traffic requirements of sensitive applications, such as real-time voice and video, and to prevent the degradation of quality caused by packet loss, delay and jitter.

QoS parameters

- **Packet loss-**
  - —This happens when network links become congested, and routers and switches start dropping packets.
  - —When packets are dropped during real-time communication, such as in voice or video calls, these sessions can experience jitter and gaps in speech.
  - —Packets can be dropped when a queue, or line of packets waiting to be sent, overflows.

QoS parameters (2)

- **Jitter-**
  - —This is the result of network congestion, timing drift and route changes.
  - —Too much jitter can degrade the quality of voice and video communication.
- **Latency-**
  - —This the time it takes a packet to travel from its source to its destination.
  - —Latency should be as close to zero as possible. If a voice over IP call has a high amount of latency, users can experience echo and overlapping audio.

QoS parameters (3)

- **Bandwidth-**
  - —This is the capacity of a network communications link to transmit the maximum amount of data from one point to another in a given amount of time.
  - —QoS optimizes the network performance by managing bandwidth and giving high priority applications with stricter performance requirements more resources than others.

Why is QoS important?

- Without QoS, network data can become disorganized, clogging networks to the point where performance degrades or, in certain cases, the network shuts down completely.
- Quality of service is important because enterprises need to provide stable services for employees and customers to use.

What are the benefits of QoS?

- It ensures the availability of an organization's network and the applications that run on that network.
- It provides the safe, efficient transfer of data over that network.
- QoS also allows organizations to use their existing bandwidths more efficiently, instead of upgrading network infrastructure to expand bandwidth.

What are the benefits of QoS? (2)

- Mission-critical applications have access to the resources they require.
- Administrators can manage traffic better.
- Organizations can reduce costs by eliminating the need to purchase new network infrastructure.
- User experience is improved.

Models deploying QoS

- Three models exist to implement QoS: Best Effort, Integrated Services and Differentiated Services.
- **Best Effort**
  - A QoS model where all the packets receive the same priority, and there is no guaranteed delivery of packets.
  - Best Effort is applied when networks have not configured QoS policies or when the infrastructure does not support QoS.

Models deploying QoS (2)

- **Integrated Services(IntServ)**
  - A QoS model that reserves bandwidth along a specific path on the network.
  - Applications ask the network for resource reservation, and network devices monitor the flow of packets to make sure network resources can accept the packets.
  - Implementing IntServ requires IntServ-capable routers and uses the Resource Reservation Protocol (RSVP) for network resource reservation.
  - IntServ has limited scalability and high consumption of network resources.

Models deploying QoS (3)

- **Differentiated Services (DiffServ)**
  - A QoS model where network elements, such as routers and switches, are configured to service multiple classes of traffic with different priorities.
  - Network traffic must be divided into classes based on a company's configuration
  - For example, voice traffic can be assigned a higher priority than other types of traffic.
  - Packets are assigned priorities using Differentiated Services Code Point (DSCP) for classification.
  - DiffServ also uses per-hop behavior to apply QoS techniques, such as queuing and prioritization, to packets.

Techniques To Improve QoS

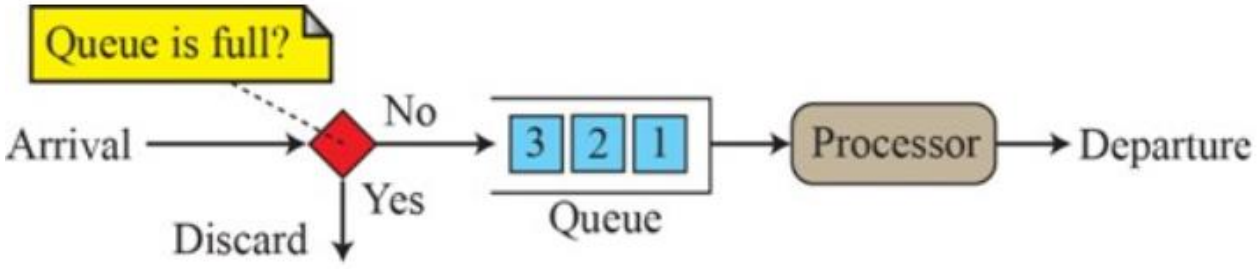
- Scheduling
  - —FIFO QUEUING
  - —PRIORITY QUEUING
  - —WEIGHTED FAIR QUEUING
- Traffic Shaping
  - —LEAKY BUCKET
  - —TOKEN BUCKET
- Resource Reservation
- Admission Control

Scheduling

- Packets from different flow arrive at switch or router for processing.

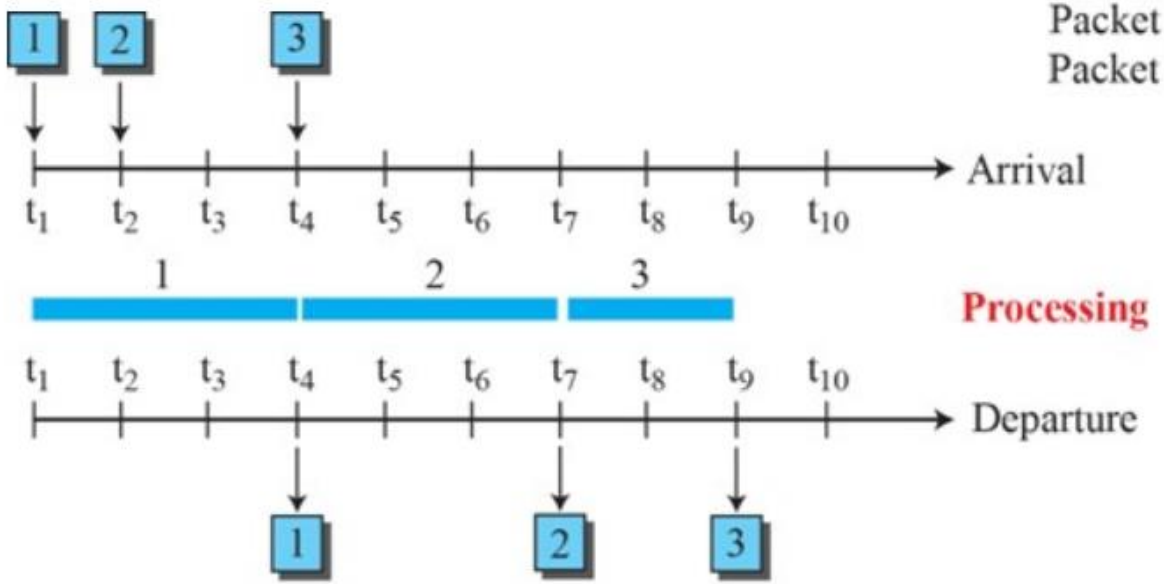
- A good scheduling technique treats the different flow in a fair and appropriate manner.

FIFO Queuing



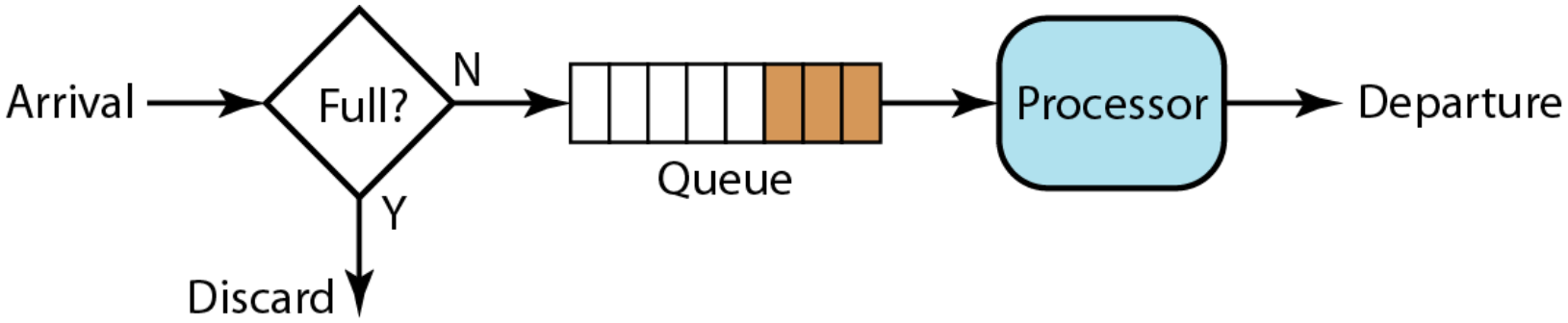
a. Processing in the router

Required processing time
Packet 1: three time units
Packet 2: three time units
Packet 3: two time units



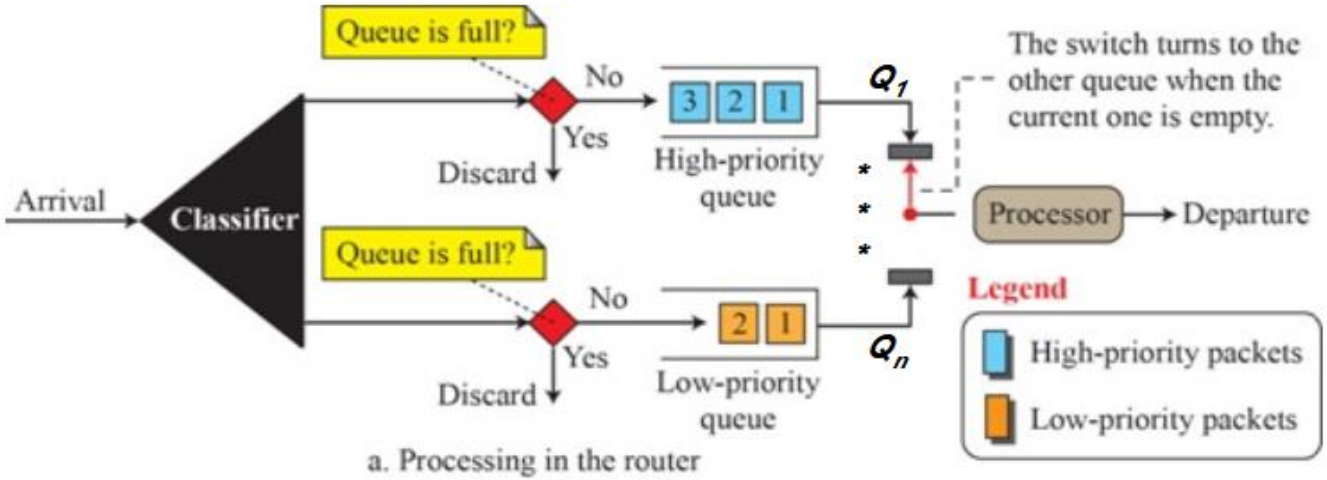
b. Arrival and departure time

FIFO Queuing (2)

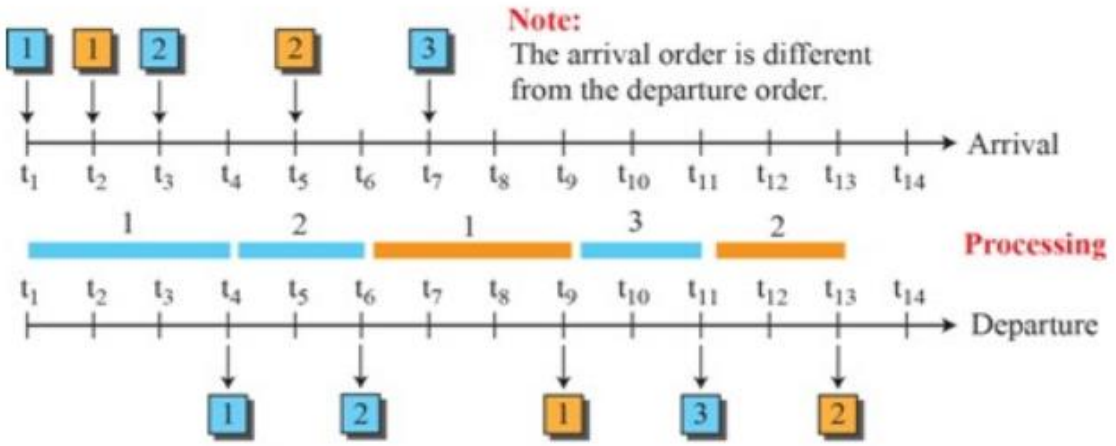


- Packets wait in a buffer (queue) until the node is ready to process them.
- If avg. arrival rate>avg. processing rate=>new packets will be discarded.

Priority Queuing



a. Processing in the router

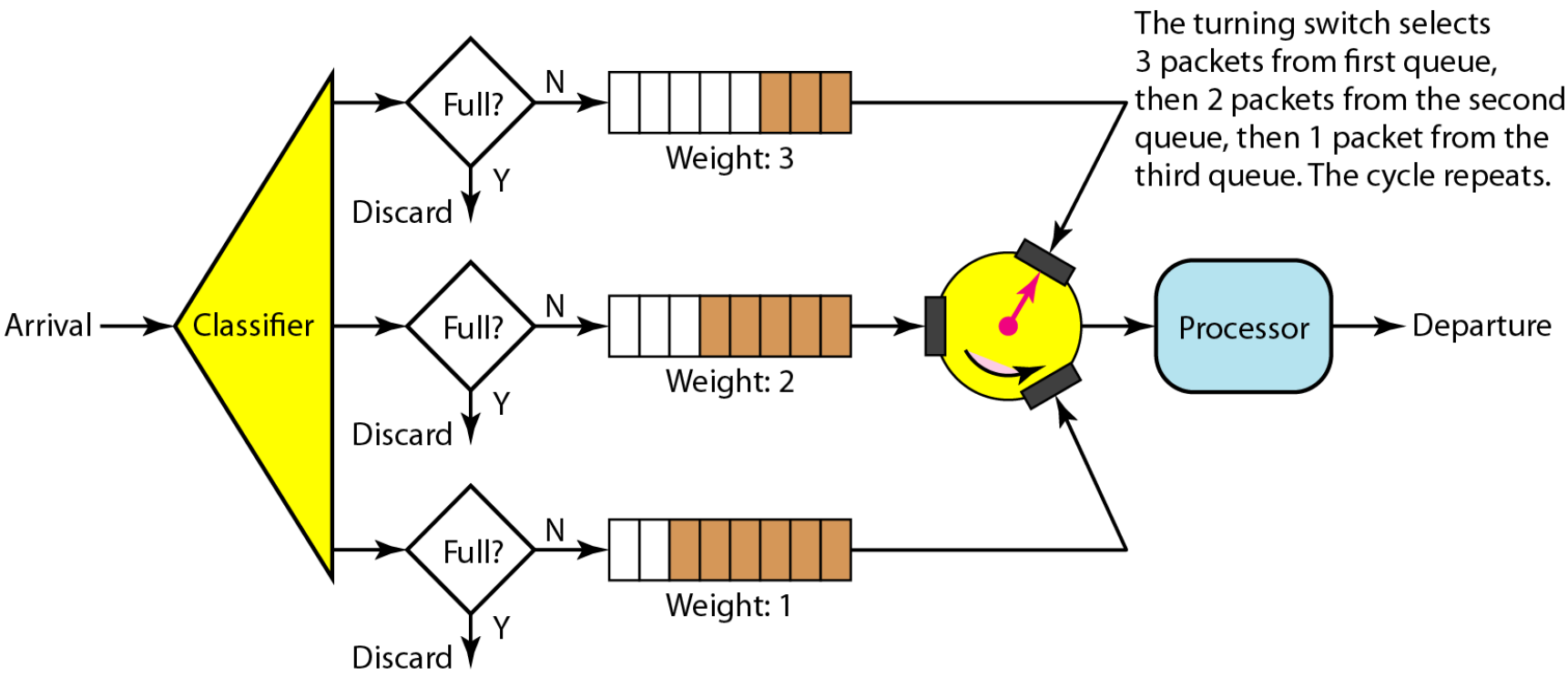


b. Arrival and departure time

Priority Queuing (2)

- Packets are assigned to a priority class
- Each class has its own queue.
- Higher class packets are processed first.
- Drawbacks of PQ
  - If the amount of highest priority traffic is excessive then the lower priority queue may not get any service until the highest priority traffic is served completely.
  - During this while, the queues allocated to lower priority traffic may overflow.

Weighted fair queuing



Weighted fair queuing (2)

- Packets are still assigned to different classes and admitted to different queues.
- The queues are weighted based on the priority.
- Higher priority means higher weight.
- If weight are 3,2,1 then 3 packets are processed from q1,2 from q2,and 1 from q1 in a round robin fashion.

Traffic Shaping or Policing

- To control the amount and the rate of traffic.

- **Traffic Shaping** is to control the traffic when it leaves the network.
- **Traffic Policing** is to control the traffic when it enters the network.
- Two techniques can shape or police the traffic: **leaky bucket** and **token bucket**.

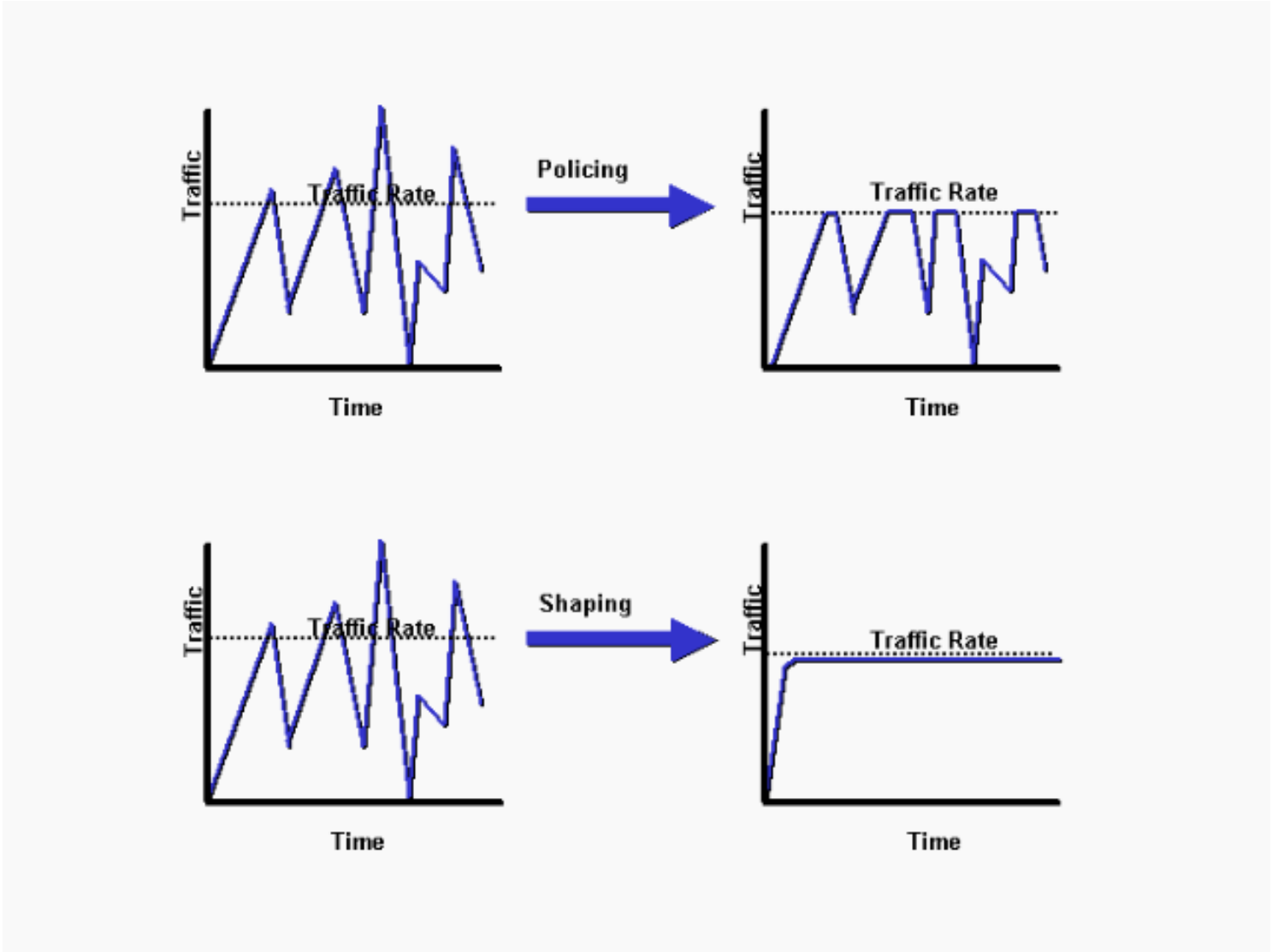
Bursty Traffic Policing VS. Shaping

- **Traffic Policing** propagates **bursts**.
  - —When the traffic rate reaches the configured maximum rate, excess traffic is dropped (or remarked).
  - —The result is an output rate that appears as a saw-tooth with crests and troughs.
- **Traffic Shaping** In contrast to policing, **retains excess packets** in a **QUEUE** and then schedules the excess for later transmission over increments of time.
  - —The result of traffic shaping is a smoothed packet output rate.

Bursty Traffic Policing VS. Shaping (2)

- Shaping implies the existence of a queue and of sufficient memory to buffer delayed packets, while policing does not buffer excess packets.

Bursty Traffic Policing VS. Shaping (3)



Leaky Bucket

- A water bucket leaks (outputs) in a **constant** rate of water regardless of the input flow of water.

- Hence, a network can **regulate** the output data rate of its **bursty** input traffic rate.

- The input rate varies but the o/p remains constant ,similarly ,network can smooth out bursty traffic

Leaky bucket (2)

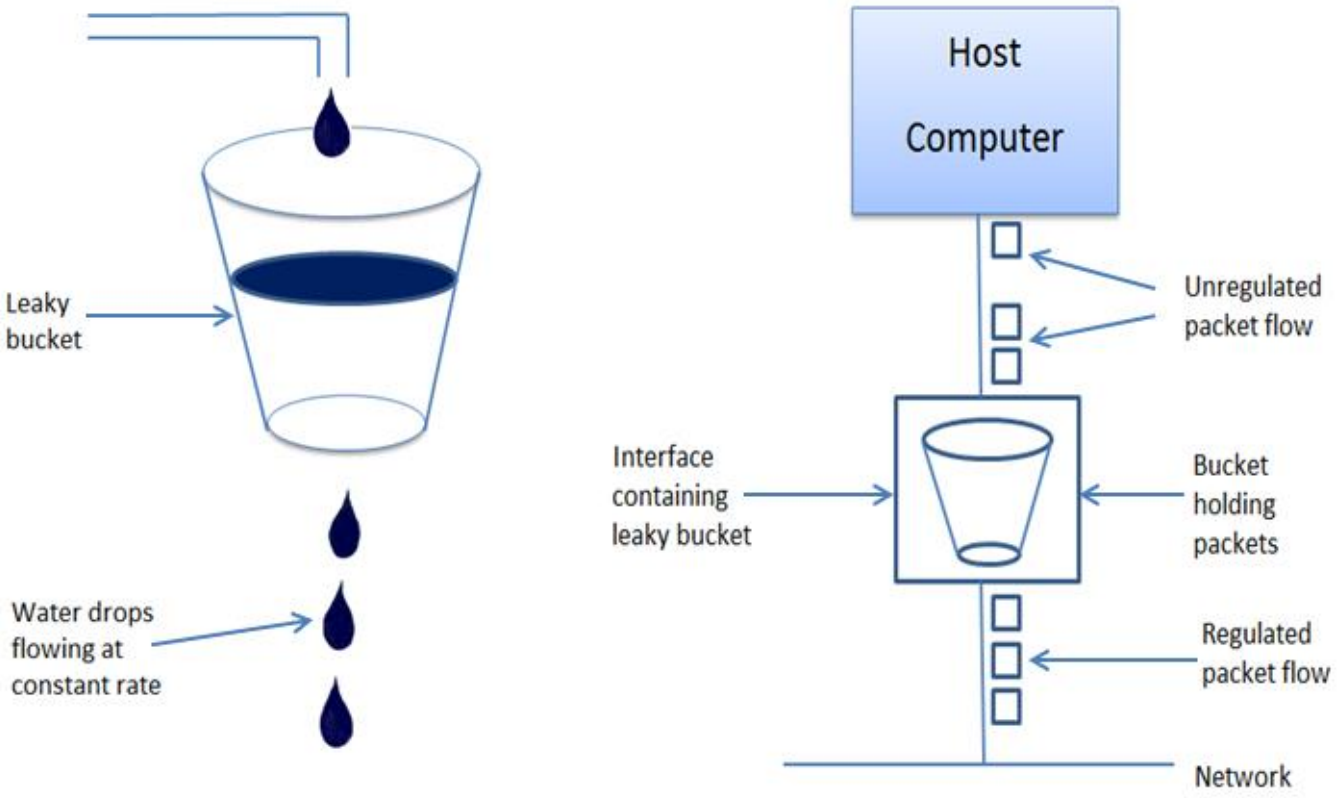
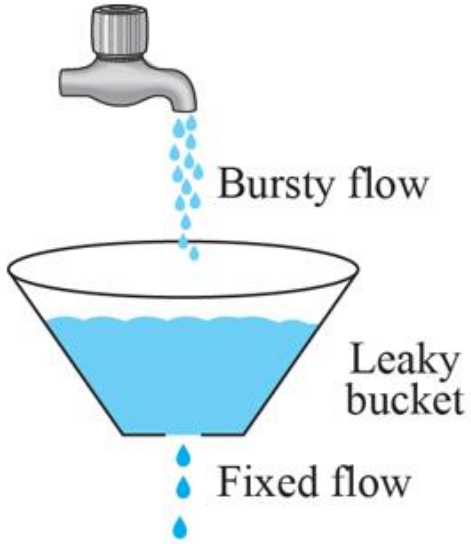
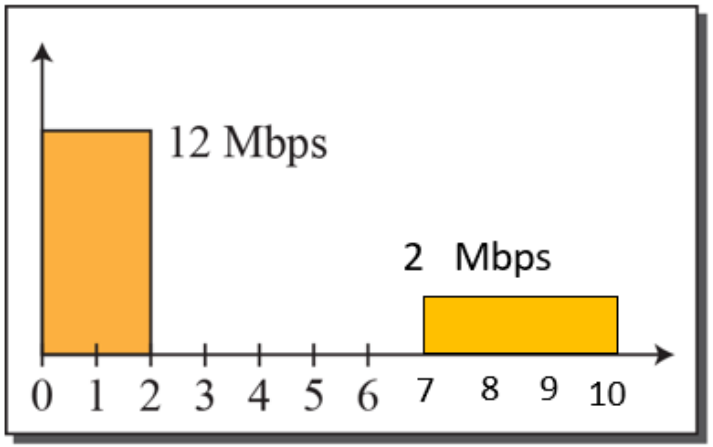


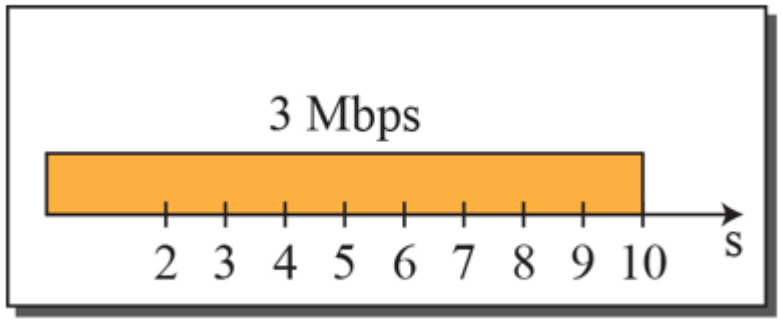
Fig: Leaky Bucket Algorithm

Leaky bucket (3)



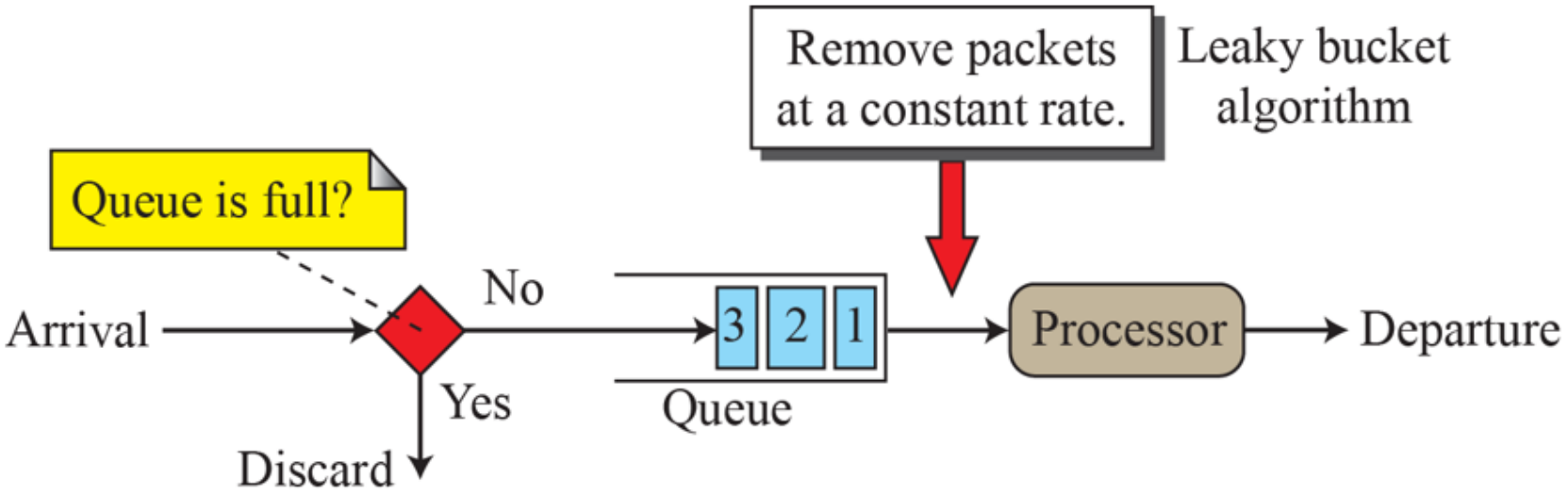


Bursty data



Fixed-rate data

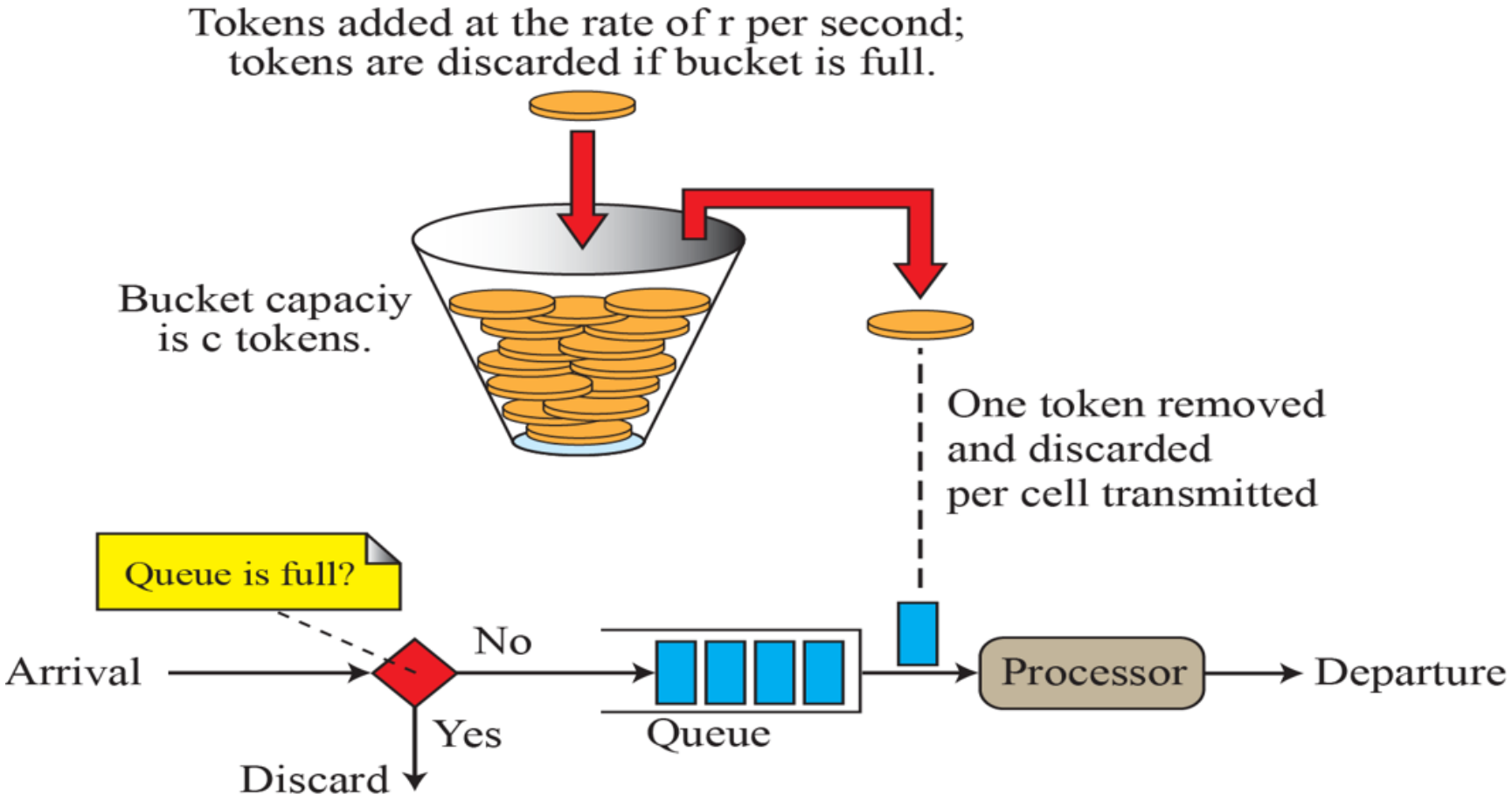
Leaky bucket implementation



Token Bucket

- Leaky bucket does not taken into a/c idle host , if a host is not sending for a while , its bucket becomes empty.
- If the host has bursty data ,leaky bucket allows only avg. rate.
- Token bucket takes into a/c the idle time , with each clock tick the tokens are added to bucket ,when the data needs to be send ,it collects token from bucket and then send the data packet consisting of data=no. of tokens

Token Bucket (2)



Token Bucket (3)

- The bucket gets tokens at a certain rate (data unit per sec, du/s).
- A token is permission for the source to send a certain number of du's into the network.
- To send a packet, remove from the bucket a number of tokens equal in representation to the packet size in du's.

Token Bucket (4)

- If not enough tokens are in the bucket to send a packet, the packet either:
  - —queued waiting until the bucket has enough tokens (in the case of a shaper), OR
  - — discard/marked-down (in the case of a policer).
- If the bucket fills to its specified capacity (max burst size) , newly arriving tokens are discarded.
- A token bucket permits burstiness, but bounds (shape/police) it.

Token Vs Leaky

TOKEN BUCKET	LEAKY BUCKET
Token dependent.	Token independent.
If bucket is full token are discarded, but not the packet.	If bucket is full packet or data is discarded.
Packets can only transmitted when there are enough token	Packets are transmitted continuously.
It allows large bursts to be sent faster rate after that constant rate	It sends the packet at constant rate
It saves token to send large bursts.	It does not save token.

Resource Reservation

- 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.
- A QoS model called Integrated Services, which depends heavily on resource reservation to improve the quality of service.

Admission Control

- Admission control refers to the mechanism used by a router or a switch to accept or reject a flow based on predefined parameters.
- Before a router accepts a flow for processing, it checks the flow specifications to see if its capacity can handle the new flow.
- It takes into account bandwidth, buffer size, CPU speed, etc., as well as its previous commitments to other flows.

Admission Control

- Admission control in ATM networks is known as Connection Admission Control (CAC), which is a major part of the strategy for controlling congestion..