Packet Loss

- We have assumed that the queue is capable of holding an infinite number of packets.
- In reality a queue preceding a link has finite capacity, although the queuing capacity greatly depends on the router design and cost.
- Because the queue capacity is finite, packet delays do not really approach infinity as the traffic intensity approaches 1.
- Instead, a packet can arrive to find a full queue.
- With no place to store such a packet, a router will drop that packet; that is, the packet will be lost.

Packet Loss

- From an end-system viewpoint, a packet loss will look like a packet having been transmitted into the network core but never emerging from the network at the destination.
- The fraction of lost packets increases as the traffic intensity increases.
- Therefore, performance at a node is often measured not only in terms of delay, but also in terms of the probability of packet loss.

End-to-End Delay

- Consider the total delay from source to destination.
- Suppose there are N-1 routers between the source host and the destination host.
- Let's also suppose that the network is uncongested (so that queuing delays are negligible), the processing delay at each router and at the source host is d_{proc}, the transmission rate out of each router and out of the source host is R bits/sec, and the propagation on each link is d_{prop}.
- The nodal delays accumulate and give an end-to-end delay, $d_{end\text{-}end} = N (d_{proc} + d_{trans} + d_{prop})$

  where, once again, dtrans = L/R, where L is the packet size.

Throughput in Computer Networks

- Another critical performance measure in computer networks is end-to-end throughput.
- To define throughput, consider transferring a large file from Host A to Host B across a computer network. (This transfer might be, a large video clip from one peer to another in a P2P file sharing system.)
- The instantaneous throughput at any instant of time is the rate (in bits/sec) at which Host B is receiving the file.
- If the file consists of F bits and the transfer takes T seconds for Host B to receive all F bits, then the average throughput of the file transfer is F/T bits/sec.

QUALITY OF SERVICE

- The Internet was originally designed for best-effort service with of guarantee of predictable performance.
- Best-effort service is often sufficient for a traffic that is not sensitive to delay, such as file transfers and e-mail.
- Such a traffic is called elastic because it can stretch to work under delay conditions; it is also called available bit rate because applications can speed up or slow down according to the available bit rate.
- The real-time traffic is generated by some multimedia applications.
- The real-time traffic is delay sensitive and therefore requires guaranteed and predictive performance.
- Quality of service (QoS) is an internetworking issue that refers to a set of techniques and mechanisms that guarantees the performance of the network to deliver predictable service to an application program.

QUALITY OF SERVICE

Data-Flow Characteristics

- If we want to provide quality of service for an Internet application, we first need to define what we need for each application.
- Traditionally, four types of characteristics are attributed to a flow: reliability, delay, jitter, and bandwidth.
- Reliability: Reliability is a characteristic that a flow needs in order to deliver the packets safe and sound to the destination.
- Delay: Source-to-destination delay is another flow characteristic.
- Jitter: Jitter is the variation in delay for packets belonging to the same flow.
- Bandwidth: Different applications need different bandwidths. In video conferencing we need to send millions of bits per second to refresh a color screen while the total number of bits in an e-mail may not reach even a million.

QUALITY OF SERVICE

Sensitivity of Applications

Table 1.1 *Sensitivity of applications to flow characteristics*

Application	*Reliability*	*Delay*	*Jitter*	*Bandwidth*
FTP	High	Low	Low	Medium
HTTP	High	Medium	Low	Medium
Audio-on-demand	Low	Low	High	Medium
Video-on-demand	Low	Low	High	High
Voice over IP	Low	High	High	Low
Video over IP	Low	High	High	High

[Behrouz A Forouzan, Firouz Mosharraf, "Computer Networks: A top down Approach", McGraw Hill Education]

QUALITY OF SERVICE

- For those applications with a high level of sensitivity to reliability, we need to do error checking and discard the packet if corrupted.
- For those applications with a high level of sensitivity to delay, we need to be sure that they are given priority in transmission.
- For those applications with a high level of sensitivity to jitter, we need to be sure that the packets belonging to the same application pass the network with the same delay.
- For those applications that require high bandwidth, we need to allocate enough bandwidth to be sure that the packets are not lost.
- Several scheduling techniques are designed to improve the quality of service.
- Various service models in Quality of Service (QOS) are also designed. [Integrated Services (IntServ), Differentiated Services (DiffServ)]

QUALITY OF SERVICE

Flow Control to Improve QoS

- IP datagram has a ToS field that can informally define the type of service required for a set of datagrams sent by an application.

- Scheduling
  - FIFO Queuing
  - Priority Queuing
  - Weighted Fair Queuing

- Traffic Shaping or Policing

QUALITY OF SERVICE

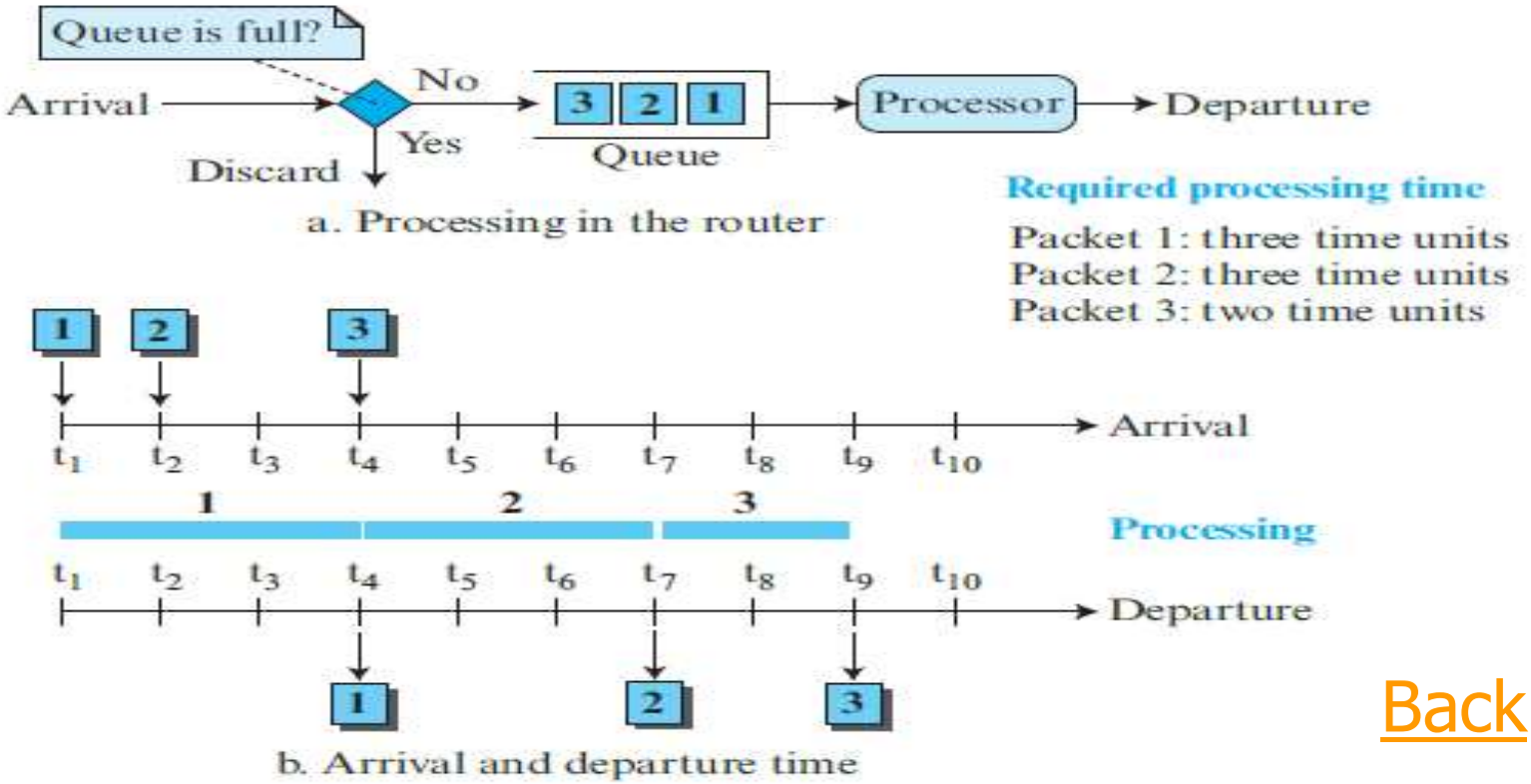
- Scheduling
  - FIFO Queuing
    - FIFO queuing is the default scheduling in the Internet.
    - In first-in, first-out (FIFO) queuing, packets wait in a buffer (queue) until the node (router) is ready to process them.
    - If the average arrival rate is higher than the average processing rate, the queue will fill up and new packets will be discarded. Fig
    - The only thing that is guaranteed in this type of queuing is that the packets depart in the order they arrive.
    - With FIFO queuing, all packets are treated the same in a packet-switched network. No matter if a packet belongs to FTP, or Voice over IP, or an e-mail message, they will be equally subject to loss, delay, and jitter.

If we need to provide different services to different classes of packets, we need to have other scheduling mechanisms.

QUALITY OF SERVICE

[Behrouz A Forouzan, Firouz Mosharraf, "Computer Networks: A top down Approach", McGraw Hill Education]

Figure *FIFO queue*



Back

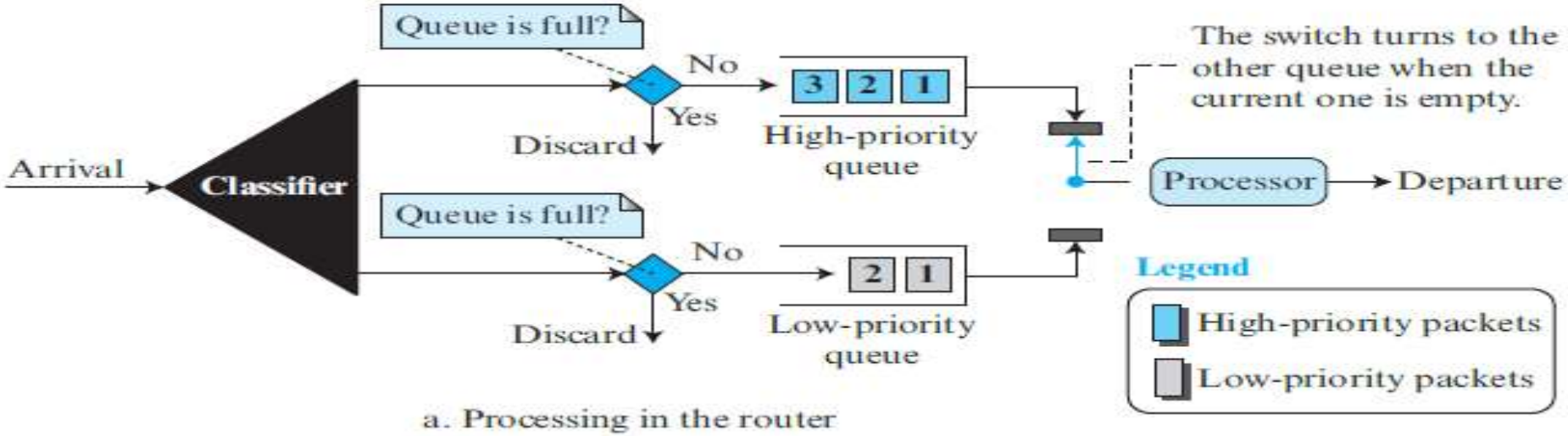
- Packets 1 and 2 need three time units of processing, but packet 3, which is smaller, needs two time units.
- This means that packets may arrive with some delays but depart with different delays.
- If the packets belong to the same application, this produces jitters. If the packets belong to different applications, this also produces jitters for each application.

QUALITY OF SERVICE

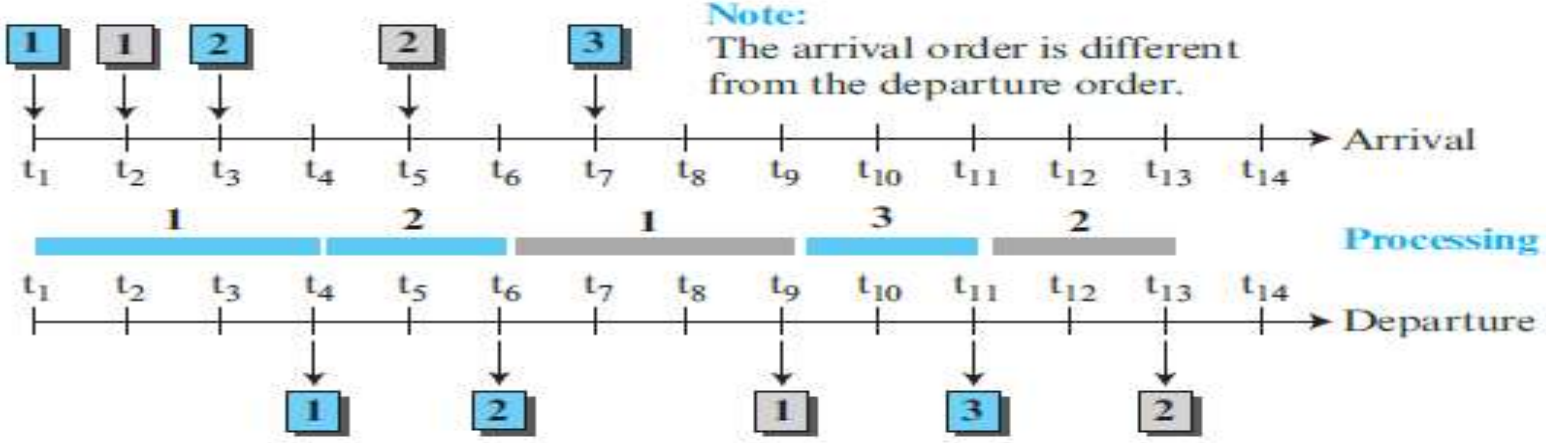
- Scheduling
  - Priority Queuing
    - Queuing delay in FIFO queuing often degrades quality of service in the network. A frame carrying real-time packets may have to wait a long time behind a frame carrying a small file.
    - Can solve this problem by using multiple queues and priority queuing.
    - In priority queuing, packets are first assigned to a priority class. Each priority class has its own queue. The packets in the highest-priority queue are processed first. Packets in the lowest-priority queue are processed last.
    - A packet priority is determined from a specific field in the packet header:
      - the ToS field of an IPv4 header, the priority field of IPv6, a priority number assigned to a destination address, or a priority number assigned to an application (destination port number), and so on.

QUALITY OF SERVICE

Figure *Priority queuing*



a. Processing in the router



b. Arrival and departure time

[Behrouz A Forouzan, Firouz Mosharraf, "Computer Networks: A top down Approach", McGraw Hill Education]

QUALITY OF SERVICE

- Scheduling
  - Priority Queuing
  - A priority queue can provide better QoS than the FIFO queue because higher-priority traffic,such as multimedia, can reach the destination with less delay.
  - Drawback : Starvation
    - If there is a continuous flow in a high-priority queue, the packets in the lower-priority queues will never have a chance to be processed. This is a condition called starvation. Severe starvation may result in dropping of some packets of lower priority.

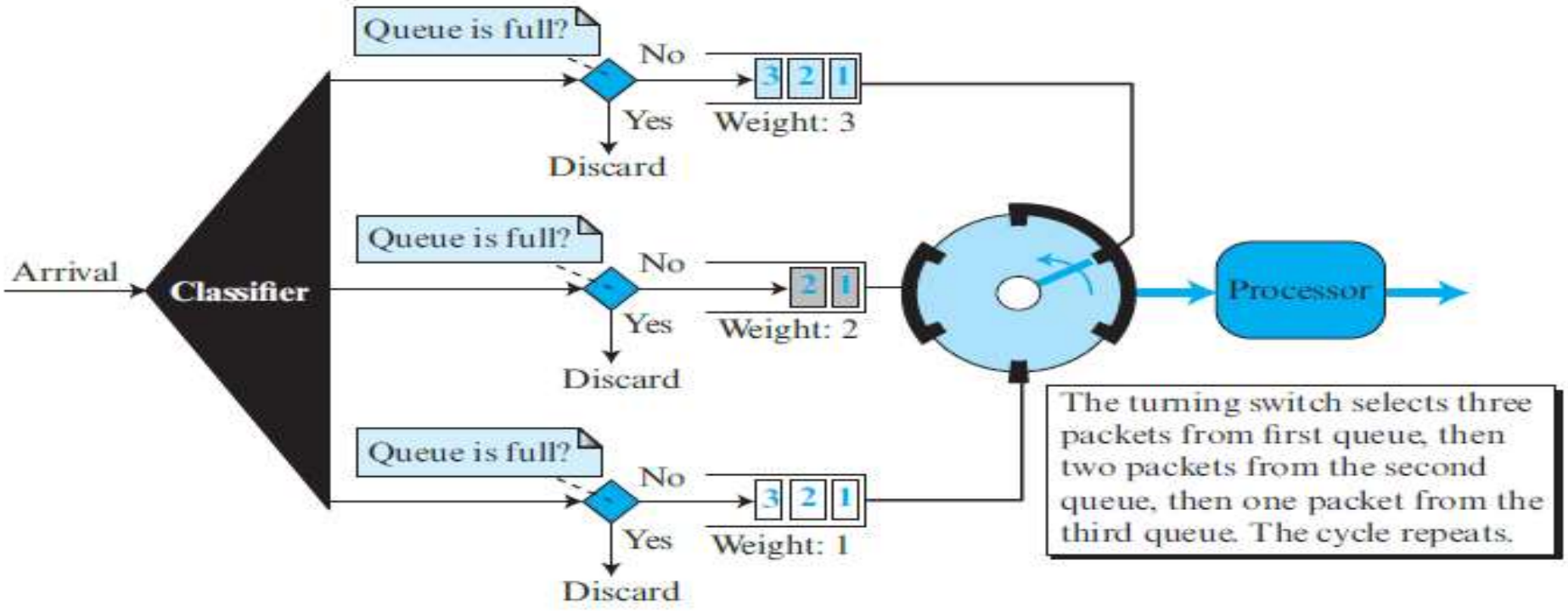
QUALITY OF SERVICE

- Scheduling
  - Weighted Fair Queuing
  - Packets are still assigned to different classes and admitted to different queues.
  - Queues are weighted based on the priority of the queues; higher priority means a higher weight.
  - The system processes packets in each queue in a round-robin fashion with the number of packets selected from each queue based on the corresponding weight.
  - 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 : fair queuing with priority.
  - A fraction of time is devoted to serve each class of packets, but the fraction depends on the priority of the class.

QUALITY OF SERVICE

[Behrouz A Forouzan, Firouz Mosharraf, "Computer Networks: A top down Approach", McGraw Hill Education]

Figure *Weighted fair queuing*



- If the throughput for the router is R, the class with the highest priority may have the throughput of R/2, the middle class may have the throughput of R/3, and the class with the lowest priority may have the throughput of R/6. [3/6=1/2, 2/6=1/3, 1/6]
- This situation is true if all three classes have the same packet size, which may not occur. Packets of different sizes may create many imbalances in dividing a decent share of time between different classes.

QUALITY OF SERVICE

Traffic Shaping or Policing

- To control the amount and the rate of traffic is called traffic shaping or traffic policing.
- The first term is used when the traffic leaves a network; the second term is used when the data enters the network.
- Two techniques can shape or police the traffic: leaky bucket and token bucket.

    - Leaky Bucket
    - Token Bucket

QUALITY OF SERVICE

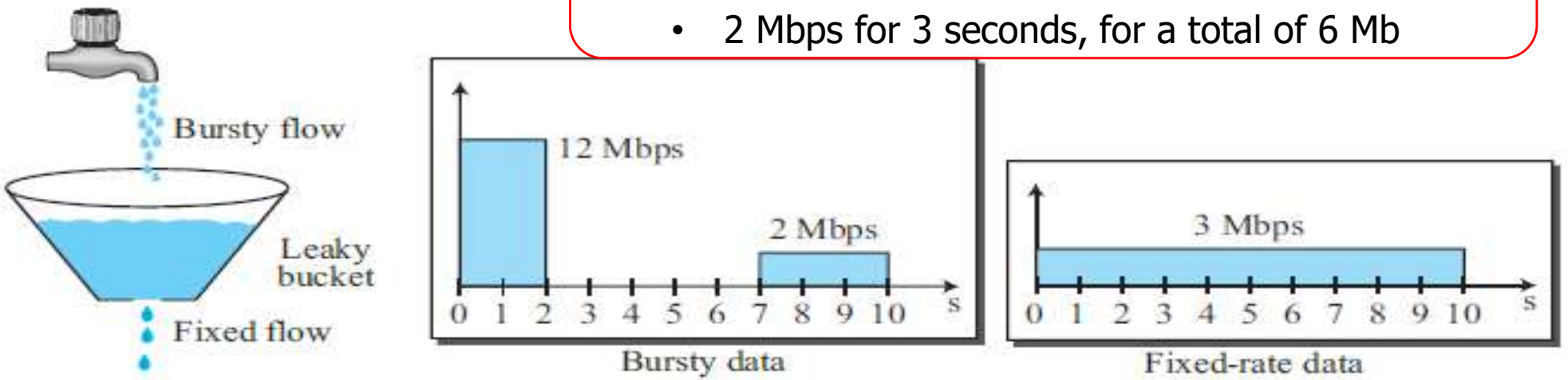
Traffic Shaping or Policing

- Leaky Bucket
- The input rate can vary, but the output rate remains constant.
- The rate at which the water leaks does not depend on the rate at which the water is input unless the bucket is empty.
- If the bucket is full, the water overflows.

Network has committed a bandwidth of 3 Mbps for a host

- 12 Mbps for 2 seconds, for a total of 24 Mb
- Host is silent for 5 seconds
- 2 Mbps for 3 seconds, for a total of 6 Mb

Figure *Leaky bucket*



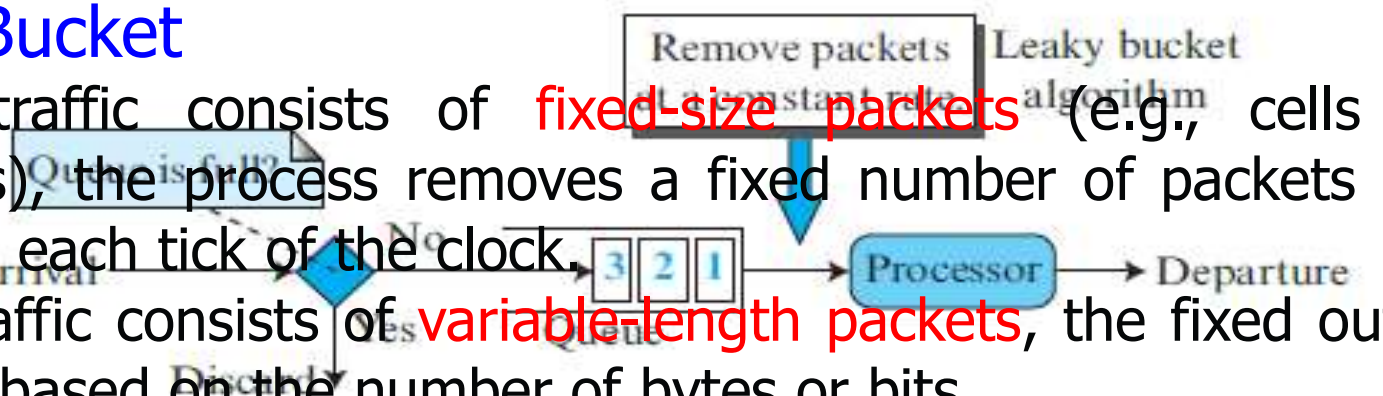
[Behrouz A Forouzan, Firouz Mosharraf, "Computer Networks: A top down Approach", McGraw Hill Education]

QUALITY OF SERVICE

Traffic Shaping or Policing

- Leaky Bucket
- If the traffic consists of fixed-size packets (e.g., cells in ATM networks), the process removes a fixed number of packets from the queue at each tick of the clock.
- If the traffic consists of variable-length packets, the fixed output rate must be based on the number of bytes or bits.

Figure: *Leaky bucket implementation*



The following is an algorithm for variable-length packets:

1. Initialize a counter to n at the tick of the clock.
2. If n is greater than the size of the packet, send the packet and decrement the counter by the packet size. Repeat this step until the counter value is smaller than the packet size.
3. Reset the counter to n and go to step 1.

A leaky bucket algorithm shapes bursty traffic into fixed-rate traffic by averaging the data rate. It may drop the packets if the bucket is full.

[Behrouz A Forouzan, Firouz Mosharraf, "Computer Networks: A top down Approach", McGraw Hill Education]

QUALITY OF SERVICE

Traffic Shaping or Policing

- Token Bucket
- The leaky bucket is very restrictive. It does not credit an idle host.
- For example, if a host is not sending for a while, its bucket becomes empty. Now if the host has bursty data, the leaky bucket allows only an average rate.
- The time when the host was idle is not taken into account.
- The token bucket algorithm allows idle hosts to accumulate credit for the future in the form of tokens.

QUALITY OF SERVICE

Traffic Shaping or Policing

- Token Bucket
- Assume the capacity of the bucket is c tokens and tokens enter the bucket at the rate of r token per second.
- The system removes one token for every cell of data sent.

- The maximum number of cells that can enter the network during any time interval of length t is shown below.

  Maximum number of packets = rt + c

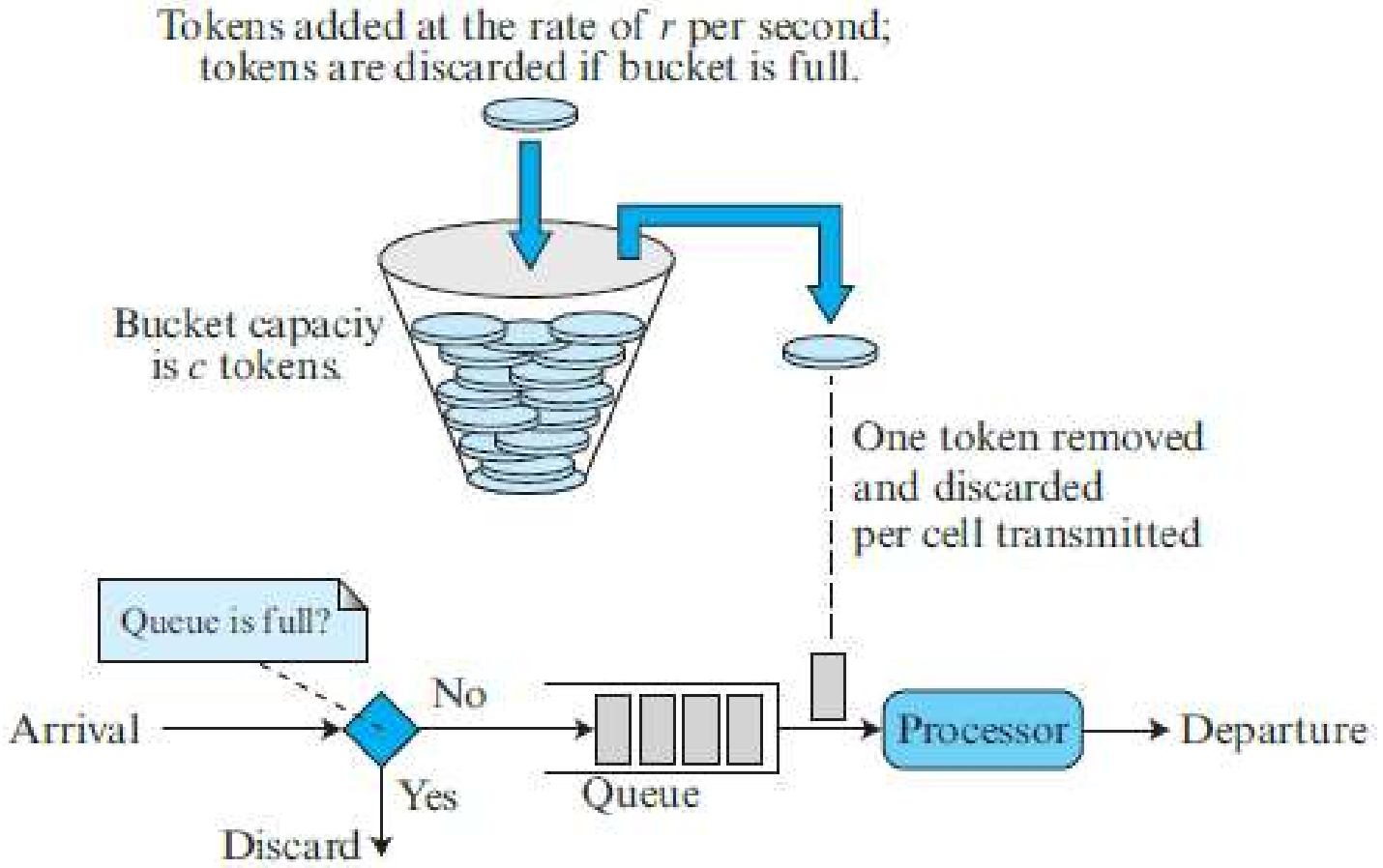
- The maximum average rate for the token bucket is shown below.

  Maximum average rate = (rt + c)/t packets per second

QUALITY OF SERVICE

Token bucket limits the average packet rate to the network.

Figure *Token bucket*



[Behrouz A Forouzan, Firouz Mosharraf, "Computer Networks: A top down Approach", McGraw Hill Education]

QUALITY OF SERVICE

Traffic Shaping or Policing

- Token Bucket
- 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.

The token bucket allows bursty traffic at a regulated maximum rate.

QUALITY OF SERVICE

Traffic Shaping or Policing

- Combining Token Bucket and Leaky Bucket
- The two techniques can be combined to credit an idle host and at the same time regulate the traffic.

- The leaky bucket is applied after the token bucket; the rate of the leaky bucket needs to be higher than the rate of tokens dropped in the bucket.

QUALITY OF SERVICE

Service Models

Integrated Services (IntServ)

- To provide different QoS for different applications, IETF developed the integrated services (IntServ) model.
- In this model, which is a flow-based architecture, resources such as bandwidth are explicitly reserved for a given data flow.
- In other words, the model is considered a specific requirement of an application in one particular case regardless of the application type (data transfer, or voice over IP, or video-on-demand).
- What is important are the resources the application needs, not what the application is doing.

QUALITY OF SERVICE

Integrated Services (IntServ)

- The model is based on three schemes:
  1. The packets are first classified according to the service they require.
  2. The model uses scheduling to forward the packets according to their flow characteristics.
  3. Devices like routers use admission control to determine if the device has the capability (available resources to handle the flow) before making a commitment. For example, if an application requires a very high data rate, but a router in the path cannot provide such a data rate, it denies the admission.
- Integrated Services is a flow-based QoS model designed for IP. In this model packets are marked by routers according to flow characteristics.

QUALITY OF SERVICE

Differentiated Services (DiffServ)

- Packets are marked by applications into classes according to their priorities.
- Routers and switches, using various queuing strategies, route the packets.
- This model was introduced by the IETF (Internet Engineering Task Force) to handle the shortcomings of Integrated Services.
- Two fundamental changes were made:

  1. The main processing was moved from the core of the network to the edge of the network. This solves the scalability problem. The routers do not have to store information about flows. The applications, or hosts, define the type of service they need each time they send a packet.

QUALITY OF SERVICE

Differentiated Services (DiffServ)

2. The per-flow service is changed to per-class service. The router routes the packet based on the class of service defined in the packet, not the flow. This solves the service-type limitation problem.

- Differentiated Services is a class-based QoS model designed for IP. In this Model packets are marked by applications according to their priority.