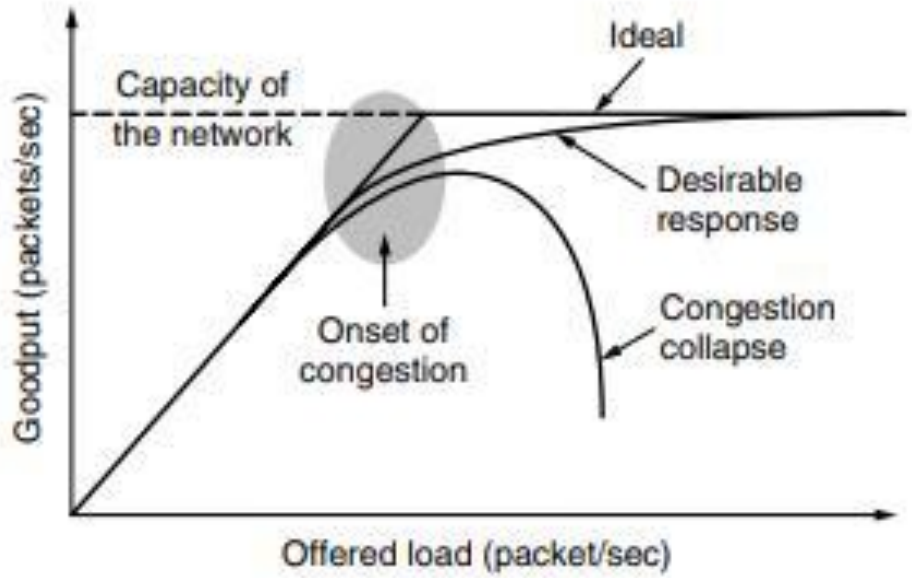
Quality of Service

What is congestion? Impact of congestion



- Packet queues at links start to grow
- Packets start dropping
- Sources start re-transmitting
- After a while only re-transmissions occupy the network
- Network resources start getting utilized in useless work (packets in queues that get timed out and re-transmitted)
- "Goodput" goes to nearly zero

Congestion control:

- What is congestion control?
  - ✓making sure the network is able to carry the offered traffic.
- Two solutions come to mind:
  - Increase the resources
  - Decrease the load
- Best way to handle congestion:

  The host can get a ‘‘slow down’’ message either because the receiver cannot handle the load or because the network cannot handle it.

Approaches to congestion control

- The most basic way to avoid congestion is to build a network that is well matched to the traffic that it carries.
- Other ways to control congestion:

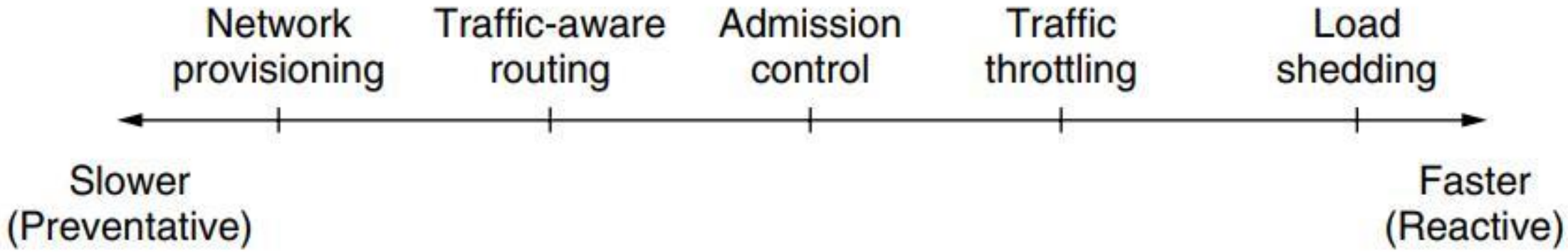


Fig.2 Timescales of approaches to congestion control

Network Provisioning:

- Dynamic allocation of resources, such as routers, links, and bandwidth, to adapt to congestion or upgrade heavily utilized components, driven by long-term traffic patterns.
- Ensures fault tolerance and optimal performance in a network infrastructure.

Traffic Aware Routing:

- Dynamic strategy that adapts network routes based on real-time traffic conditions
- Utilizes methods like adjusting path weights and incorporating live updates
- Enables efficient data transmission and avoids congestion hotspots.

Admission Control:

- Sometimes it is not possible to increase capacity.
- Only way is to decrease the load.
- In a virtual-circuit network, new connections can be refused if they would cause the network to become congested.

Traffic Throttling

- When congestion is imminent the network can deliver feedback to the sources whose traffic flows are responsible for the problem.
- The network can request these sources to throttle their traffic, or it can slow down the traffic itself.

Load Shedding:

- When routers are being inundated by packets that they cannot handle, they just throw them away.

Congestion control and quality of service

- The techniques we looked at in the previous sections are designed to reduce congestion and improve network performance.
- However, there are applications (and customers) that demand stronger performance guarantees from the network than ‘‘the best that could be done under the circumstances.’’
- Multimedia applications in particular, often need a minimum throughput and maximum latency to work.
- An easy solution to provide good **quality of service** is to build a network with enough capacity for whatever traffic will be thrown at it.

Quality of Service

- **Quality of service** mechanisms let a network with less capacity meet application requirements just as well at a lower cost.

Four issues must be addressed to ensure quality of service:

1. What applications need from the network.
2. How to regulate the traffic that enters the network.
3. How to reserve resources at routers to guarantee performance.
4. Whether the network can safely accept more traffic.

Modules of Quality of Service

- Application Requirements
- Traffic Shaping
- Packet Scheduling
- Integrated Service
- Differentiated Service

Application Requirements:

Application	Bandwidth	Delay	Jitter	Loss
Email	Low	Low	Low	Medium
File sharing	High	Low	Low	Medium
Web access	Medium	Medium	Low	Medium
Remote login	Low	Medium	Medium	Medium
Audio on demand	Low	Low	High	Low
Video on demand	High	Low	High	Low
Telephony	Low	High	High	Low
Videoconferencing	High	High	High	Low

Fig. Stringency of applications' quality-of-service requirements

- To accommodate a variety of applications, networks may support different categories of QoS.

1. Constant bit rate (e.g., telephony).
2. Real-time variable bit rate (e.g., compressed videoconferencing).
3. Non-real-time variable bit rate (e.g., watching a movie on demand).
4. Available bit rate (e.g., file transfer).

Traffic Shaping

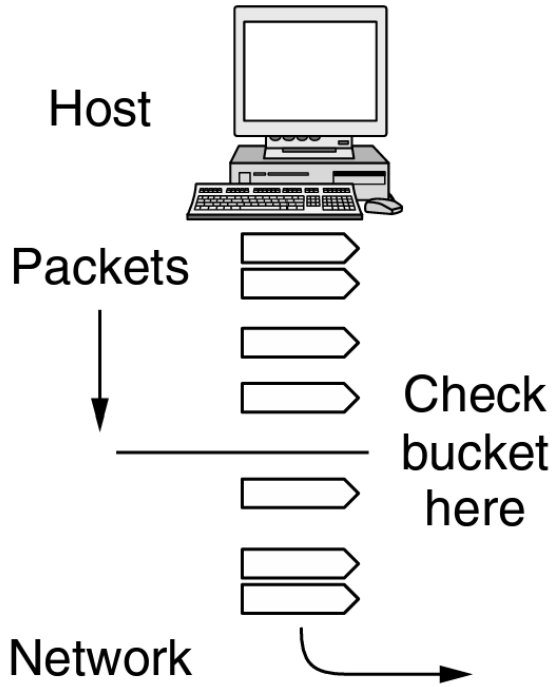
- Traffic shaping is a technique for regulating the average rate and burstiness of a flow of data that enters the network.
- The goal is to allow applications to transmit a wide variety of traffic that suits their needs, including some bursts, yet have a simple and useful way to describe the possible traffic patterns to the network.
- Traffic shaping reduces congestion and thus helps the network live up to its promise.

Traffic shaping algorithms:

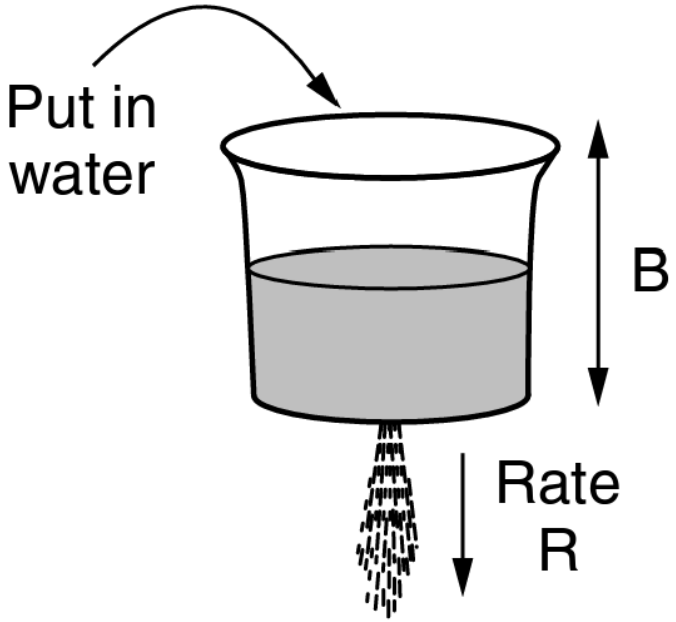
- Leaky Bucket
- Token Bucket

Leaky Bucket

- The leaky bucket algorithm is used to control rate in a network.
- It is implemented as a single server queue with content service time.
- If the bucket overflows, then packets are discarded.
- In the algorithm the input rate can vary but the output rate remains constant.
- The algorithm serves bursty traffic into fixed rate traffic by averaging the data rate.



a) Leaky bucket with Packet



b) Leaky bucket with water

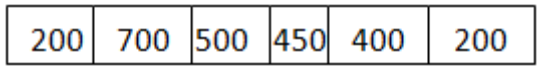
Algorithm

Step1: Initialize a counter to n at the tick of the clock.

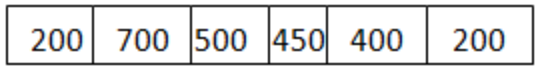
Step2: Repeat until n is smaller than the packet size of the packet at the head of the queue.

- Pop a packet out of the head of the queue, say P.
- Send the packet P, into the network
- Decrement the counter by the size of packet P.

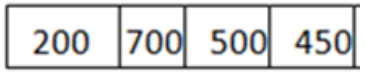
Step3: Reset the counter and go to step 1.

- **Example:** Let n=1000
- Packet= 

- Since n > size of the packet at the head of the Queue, i.e. n > 200
  Therefore, n = 1000-200 = 800
  Packet size of 200 is sent into the network.



- Now, again n > size of the packet at the head of the Queue, i.e. n > 400
  Therefore, n = 800-400 = 400
  Packet size of 400 is sent into the network.



- Since, n < size of the packet at the head of the Queue, i.e. n < 450
  Therefore, the procedure is stopped.
- Initialize n = 1000 on another tick of the clock.
  This procedure is repeated until all the packets are sent into the network.

Token Bucket

- The token bucket algorithm limits the number of tokens that can be in the bucket at any given time, representing the maximum capacity or permission available to the system.
- Tokens are added to the bucket at a fixed rate over time, starting with an empty bucket.
- When an event occurs, it requests a token from the bucket.
- If a token is available, it is removed from the bucket, allowing the event to occur.
- If no tokens are available, the event is blocked or delayed until a token becomes available.
- After each event, the algorithm checks whether the bucket has exceeded its capacity, and if so, additional tokens are discarded.
- This ensures that the bucket is not too full and the system remains controlled.

Algorithm:

- Step1: A token is added every Δt time.
- Step2: The bucket can hold at most b-token. If token arrive when bucket is full it is discarded.
- Step3: When the packet of m bytes arrived m tokens m tokens are removed from the bucket and the packet is sent to the network.
- Step4: If less, then n tokens are available no tokens are removed from the bucket and the packet is considered to be non conformant.
- Step5: The conformant packet may be enqueued for subsequent transmission when sufficient token have been accumulated in the buffer.
- Step6: If C is the maximum capacity of the bucket and q is the arrival rate and M is the maximum output rate then Brust length S can be calculated as

$$C + qS = MS$$

Example: Consider a frame relay network having a capacity of 1Mb of the data is arriving at the rate of 25mbps for 40msec. The token arrival rate is 2mbps and the capacity of the bucket is 500kb with maximum output rate 25mbps. Calculate

1 The total Burst length

2 The total output time

Solution: Here C is capacity of bucket = 500kb

M = 25mbps

q= 2mbps

1. S= 500/((25-2)*1000)=21.73msec~=22
2. For 22msec the output rate is 25msec after that the output rate becomes 2mbps i.e. token arrival rate.

   Therefore

Leaky Bucket	Token Bucket
When the host has to send a packet , packet is thrown in bucket.	In this, the bucket holds tokens generated at regular intervals of time.
Bucket leaks at constant rate	Bucket leaks at constant rate
Bursty traffic is converted into uniform traffic by leaky bucket.	If there is a ready packet , a token is removed from Bucket and packet is send.
In practice bucket is a finite queue outputs at finite rate	If there is no token in the bucket, then the packet cannot be sent.

Packet Scheduling

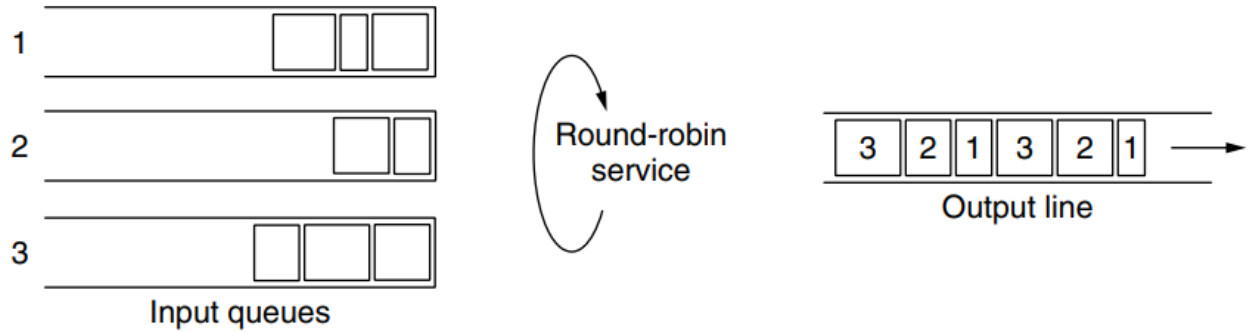
- Packet scheduling algorithms allocate bandwidth and other router resources by determining which of the buffered packets to send on the output line next.
- Three different kinds of resources can potentially be reserved for different flows:
- 1. Bandwidth. 2. Buffer space. 3. CPU cycles.

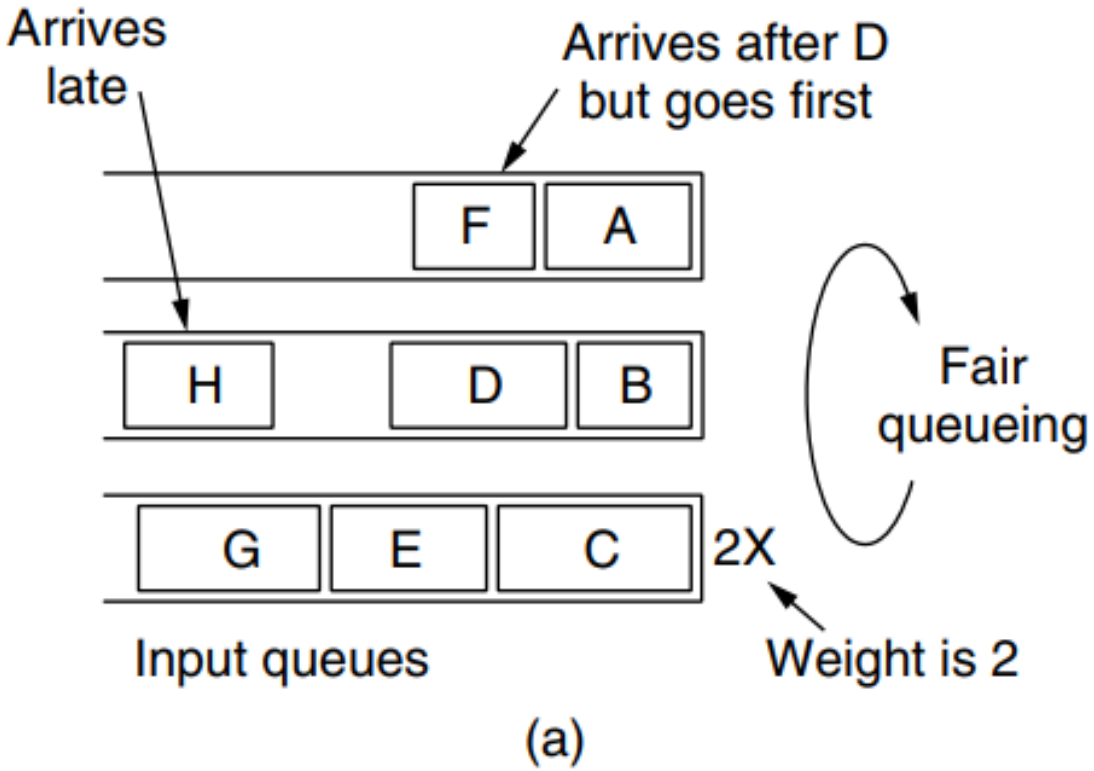
Types of Packet Scheduling Algorithm

1) **FIFO:**
    - Router buffers packets in a queue for each output line until they can be sent, and they are sent in the same order that they arrived.
    - FIFO routers usually drop newly arriving packets when the queue is full. Since the newly arrived packet would have been placed at the end of the queue, this behavior is called tail drop.

2) Round-robin Fair Queueing

- The essence of this algorithm is that routers have separate queues, one for each flow for a given output line.
- When the line becomes idle, the router scans the queues round-robin. It then takes the first packet on the next queue. In this way, with n hosts competing for the output line, each host gets to send one out of every n packets.
- It is fair in the sense that all flows get to send packets at the same rate.





Packet	Arrival time	Length	Finish time	Output order
A	0	8	8	1
B	5	6	11	3
C	5	10	10	2
D	8	9	20	7
E	8	8	14	4
F	10	6	16	5
G	11	10	19	6
H	20	8	28	8

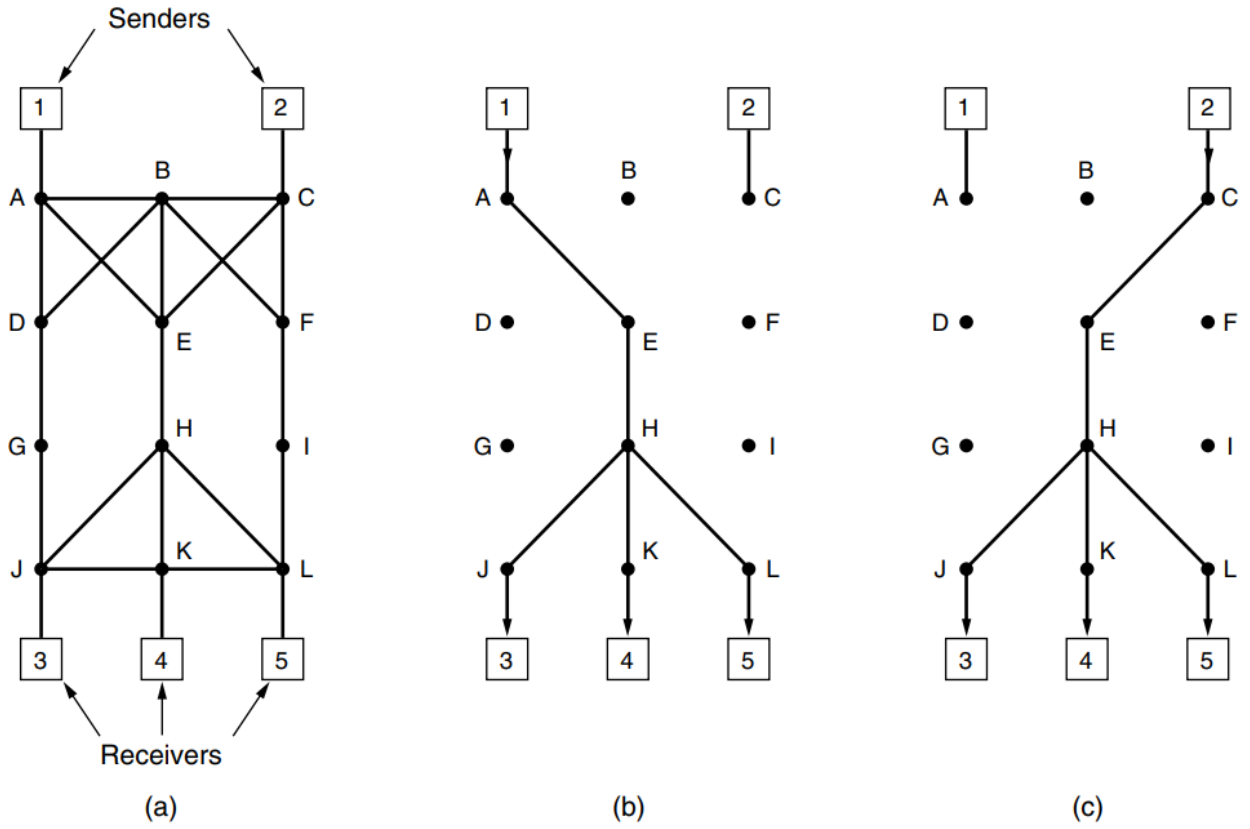
(b)

Fig. (a) Weighted Fair Queueing. (b) Finishing times for the packets.

Integrated Services

- Integrated service provides the capability for IP applications to request and reserve bandwidth using the ReSerVation Protocol (RSVP) and quality of service (QoS) APIs.
- Integrated service policies use the RSVP and the Resource Reservation Setup Protocol API (RAPI) (or qtoq socket API) to guarantee an end-to-end connection. This is the highest level of service that you can designate; however, it is also the most complex service.
- RSVP—The Resource reSerVation Protocol:This protocol is used for making the reservations; other protocols are used for sending the data.

Example:
Hosts 1 and 2 are multicast senders, and hosts 3, 4, and 5 are multicast receivers. In this example, the senders and receivers are disjoint, but in general, the two sets may overlap.

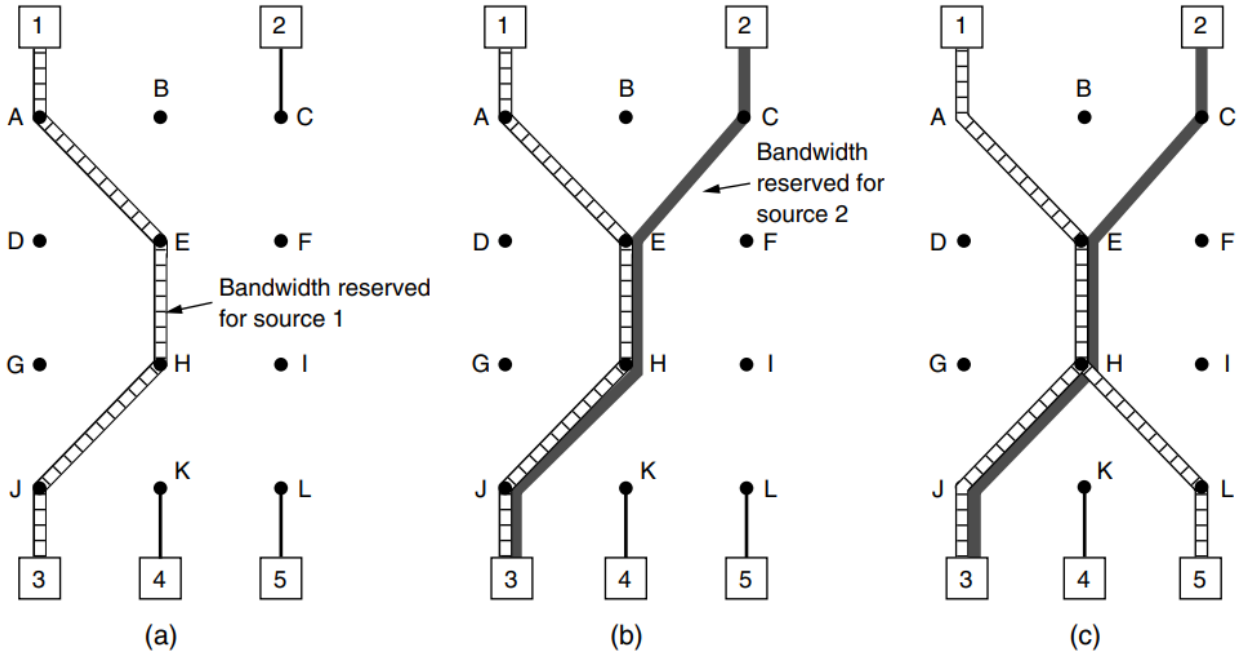


(a) A network. (b) The multicast spanning tree for host 1. (c) The multicast spanning tree for host 2.

Example of reservation request along the spanning tree:

Here host 3 has requested a channel to host 1. Once it has been established, packets can flow from 1 to 3 without congestion. Now consider what happens if host 3 next reserves a channel to the other sender, host 2, so the user can watch two television programs at once.

A second path is reserved, as illustrated in Fig. (b). Note that two separate channels are needed from host 3 to router E because two independent streams are being transmitted. Finally, in Fig.(c), host 5 decides to watch the program being transmitted by host 1 and also makes a reservation.



(a) Host 3 requests a channel to host 1. (b) Host 3 then requests a second channel, to host 2. (c) Host 5 requests a channel to host 1.

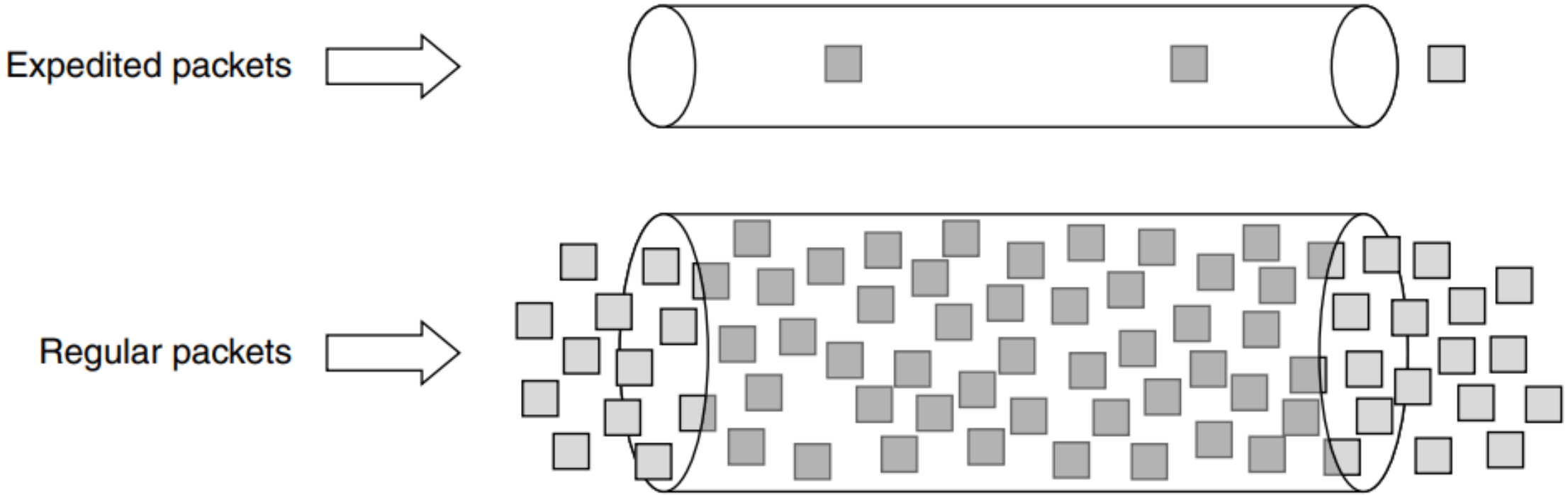
Differentiated Services

- A simpler approach to quality of service, one that can be largely implemented locally in each router without advance setup and without having the whole path involved. This approach is known as class-based (as opposed to flow-based) quality of service.
- Differentiated services can be offered by a set of routers forming an administrative domain. The administration defines a set of service classes with corresponding forwarding rules. If a customer subscribes to differentiated services, customer packets entering the domain are marked with the class to which they belong.

Expedited Forwarding

- Two classes of service are available: regular and expedited. The vast majority of the traffic is expected to be regular, but a limited fraction of the packets are expedited. The expedited packets should be able to transit the network as though no other packets were present. In this way they will get low loss, low delay and low jitter service—just what is needed for VoIP.

Expedited packets experience a traffic-free network



Assured Forwarding

- Assured forwarding specifies that there shall be four priority classes, each class having its own resources. The top three classes might be called gold, silver, and bronze. In addition, it defines three discard classes for packets that are experiencing congestion: low, medium, and high. Taken together, these two factors define 12 service classes.

A possible implementation of assured forwarding