5.1 Properties of Video and Audio

Properties of Video

1. Video is a sequence of images displayed at constant rate e.g. 24 images/sec.
2. Digital image is array of pixels. Each pixel represented by bits.
3. The most important characteristic of video is its **high bit rate**. Video on Internet usually ranges from 100 kbps for low-quality video conferencing to over 3 Mbps for streaming High-Definition (HD) movies.
4. Video can be compressed, therefore trading off video quality with bit rate.
5. A video is a sequence of images, typically being displayed at a constant rate, for example, at 24 or 30 images per second.
6. An uncompressed, digitally encoded image consists of an array of pixels, with each pixel encoded into a number of bits to represent luminance and colour.
7. Multiple versions of same video can be created using different compression techniques.

Properties of Audio

1. Digital audio has significantly lower bandwidth requirements than video.
2. The analog audio signal is sampled at some fixed rate. Example: Telephone: 8,000 samples/sec and CD music: 44,100 samples/sec.
3. Each of the samples is then rounded to one of a finite number of values. This operation is referred to as quantization. e.g., $2^8 = 256$ possible quantized values.
4. Each of the quantization values is represented by a fixed number of bits. Example: 8,000 samples/sec, 256 quantized values → 64,000 bps.
5. Receiver converts bits back to analog signal (some quality reduction).

5.2 Types of Multimedia Network Applications

- Classes of Multimedia applications :
  1) Streaming stored streaming

2) Conversational voice/video-over-IP

3) Streaming live audio/video

1) Streaming Stored Audio/Video :

- Streaming means a user can use the file after the downloading has started.
- The files are compressed and stored on a server. A client down loads the files through the Internet. This is also called as clients on-demand audio/video. The stored audio files can be songs, orchestra, historical recordings and famous lectures. The stored video files can be movies, television shows, documentaries, historical events music video clips.

a) Stored media

- The contents are pre-recorded and is stored at Internet server. The user has the facility to pause, rewind, fast forward over the multimedia contents. If the file is larger than the server's memory, it may fetch and send a block at a time. The media player starts fetching and playing the file, block by block.

b) Streaming

- In streaming a user requests a file containing (compressed) audio/video, the contents of the file must be played out as they received. The user playing out audio/video from one location in the file while it is receiving later portion of the file from the server.
- Streaming avoids down loading of entire file hence the delay is reduced considerably. In most entertainment applications, in order to meet high play out rates when large number of browsers are accessing the server simultaneously special servers, called streaming servers are to be used

c) Continuous playout

- The original timing of recording should be followed while playing out any multimedia content. Hence for continuous playout delay constraint on data delivery is to be considered.
- Delay in data delivery affect continuous playout of audio/video. This delay is more significant in live

and interactive sessions. Such as Internet telephony, Video conferencing.

2) Streaming Live Audio/Video

- Streaming live audio/video is similar to traditional broadcasting of radio and TV programs. But in this application the transmission takes place through the Internet.
- Because of streaming live audio/video it is now possible to broadcast a live audio or video transmission over the Internet. Some stations broadcast over the air and over the Internet simultaneously.
- The streaming live audio/video application is sensitive to delay also it can not accept retransmission. In this application the communication is multicast and live.

3) Real-time Interactive Audio/Video

- In real-time interactive audio/video application users can communicate with each other in real-time mode. Typical examples of real-time interactive audio/video are video conferencing and voice over IP(VOIP).
- It also supports services which traditional circuit-switched network can not provide. For example - web phone integration, real time communication in group, directory series etc. Real time data requires the preservation of the time relationship between packets of session. Jitter is introduced in real-time data by the delay between packets.
- Packet jitter is the variation in packet delays within same packet stream. This deteriorate the video quality. To prevent jitter, time stamping of packets is done. Also playback buffer is required for real time traffic.

5.2.1 UDP Streaming

- Server transmits at rate appropriate for client, often: send rate = encoding rate = constant rate.
- Transmission rate can be unaware of network congestion level.

- Short playout delay (2-5 seconds) is used to remove network jitter.
- Error recovery is implemented application-level, time permitting.
- Encapsulation of audio/video chunks in RTP (Real-Time Transport Protocol, RFC 3550) and then in UDP. It needs a control connection in parallel to pause, resume reposition, etc: Real-Time Streaming Protocol (RTSP, RFC 2326).
- There is an important issue that UDP may not go through firewalls.

5.2.2 HTTP Streaming

- In HTTP streaming, the video is simply stored in an HTTP server as an ordinary file with a specific URL. When a user wants to see the video, the client establishes a TCP connection with the server and issues an HTTP GET request for that URL.
- The server then sends the video file, within an HTTP response message, as quickly as TCP congestion control and flow control allow.
- On the client side, the bytes are collected in a client application buffer. Once the number of bytes in this buffer exceeds a predetermined threshold, the client application begins playback-specifically, it periodically grabs video frames from the client application buffer, decompresses the frames, and displays them on the user's screen.
- The use of HTTP over TCP also allows the video to traverse firewalls and NATs more easily.
- Streaming over HTTP also avoids the need for a media control server, such as an RTSP server, reducing the cost of a large-scale deployment over the Internet.
- Due to all of these advantages, most video streaming applications today-including YouTube and Netflix-use HTTP streaming (over TCP) as its underlying streaming protocol.

underlying streaming protocol.

5.2.3 Adaptive Streaming and DASH

- DASH stands for Dynamic, Adaptive Streaming over HTTP.

Server Side Operation

- Video is encoded into several different versions, with each version having a different bit rate and, correspondingly, a different quality level.
- The client dynamically requests chunks of video segments of a few seconds in length from the different versions. When the amount of available bandwidth is high, the client naturally selects chunks from a high-rate version; and when the available bandwidth is low, it naturally selects from a low-rate version.
- The client selects different chunks one at a time with HTTP GET request messages.
- The HTTP server also has a **manifest file**, which provides a URL for each version along with its bit rate

Client Side Operation

- Periodically measures server-to-client bandwidth
- Consulting manifest, requests one chunk at a time :
  - Chooses maximum coding rate sustainable given current bandwidth
  - Can choose different coding rates at different points in time (depending on available bandwidth at time)

Intelligence at Client

- Client determines :

1. When to request chunk (so that buffer starvation, or overflow does not occur)
2. What encoding rate to request (higher quality when more bandwidth available)
3. Where to request chunk (can request from URL server that is "close" to client or has high available bandwidth)

5.2.4 Content Distribution Networks

- **Challenge :** How to stream content (selected from millions of videos) to hundreds of thousands of simultaneous users ?

Option 1 : single, large "mega-server"

1. Single point of failure
2. Point of network congestion

3. Long path to distant clients
4. Multiple copies of video sent over outgoing link

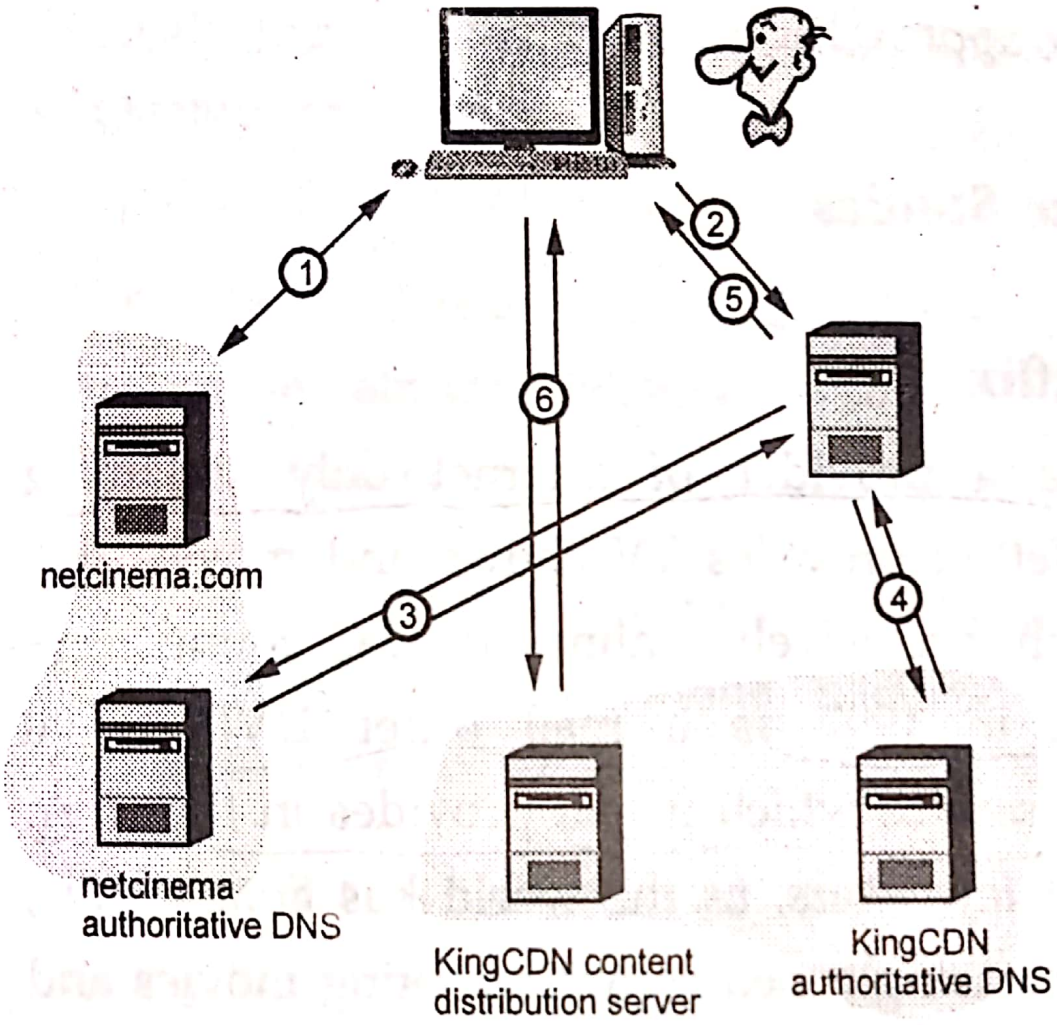
- Quite simply, this solution doesn't scale.
- Challenge : how to stream content (selected from millions of videos) to hundreds of thousands of simultaneous users ?

Option 2 : store/serve multiple copies of videos at multiple geographically distributed sites (CDN)

1. **Enter deep :** Push CDN servers deep into many access networks close to users. It is used by Akamai, 1700 locations
2. **Bring home :** Smaller number (10's) of larger clusters in POPs near (but not within) access networks. It is used by Limelight.
3. Google uses both, in addition to its mega data centers responsible for serving dynamic content

CDN Operation

- Bob (client) requests video http://video.netcinema.com/6Y7B23V actually stored in a KingCDN content distribution server.
- Fig. 5.2.1 shows CDN operation.



1. Bob gets URL for video http://video.netcinema.com/6Y7823V from netcinema.com web page

2. Resolve video.netcinema.com via Bob's local DNS that relays to netcinema's authoritative DNS server

3. netcinemas DNS returns a1105.kingcdn.com

4&5. Resolve a1105.kingcdn.com via KingCDN's authoritative DNS, which returns IP address of KingCDN distribution server with video

6. Request video from KingCDN server, streamed via HTTP

Fig. 5.2.1 CDN operation

CDN cluster selection strategy

Challenge : How does CDN DNS select "good" CDN node to stream to client

1. CDN learns the IP address of the client's local DNS via the client's DNS lookup
2. CDN can then implement a selection strategy to dynamically direct clients to a "suitable" server cluster or data center

Possible strategies :

1. Pick IP address of CDN node geographically closest to client
2. Pick IP address of CDN node with shortest delay (or min # hops) to client (CDN nodes periodically ping access ISPs, reporting results to CDN DNS)
3. Always pick the same IP address, but make sure this IP address is an IP anycast address associated with all CDN nodes

Alternative : Let client decide!

1. Give client a list of several CDN servers
2. Client pings servers, picks "best"
3. Netflix approach

5.3 Case Studies

5.3.1 Netflix

- Netflix is a provider of internet only streaming media. Netflix provides TV shows and movies that you watch exclusively online. It was founded in California in 1997 as a mail order DVD rental service, a service which it still provides in the USA. In the last few years, as the world has gone online, and as increasingly people are watching movies and TV shows online, Netflix has transformed into a powerhouse of online entertainment.
- Netflix owns very little infrastructure and uses 3rd party services and CDNs.
- Netflix uses video distribution using a CDNs and adaptive streaming over HTTP.

DASH

- The four major components of Netflix are :
  1. Registration and payment servers,
  2. Amazon cloud,
  3. Multiple CDN providers, and
  4. Clients.
- The basic architecture of the Netflix video-streaming platform is shown in Fig. 5.3.1.

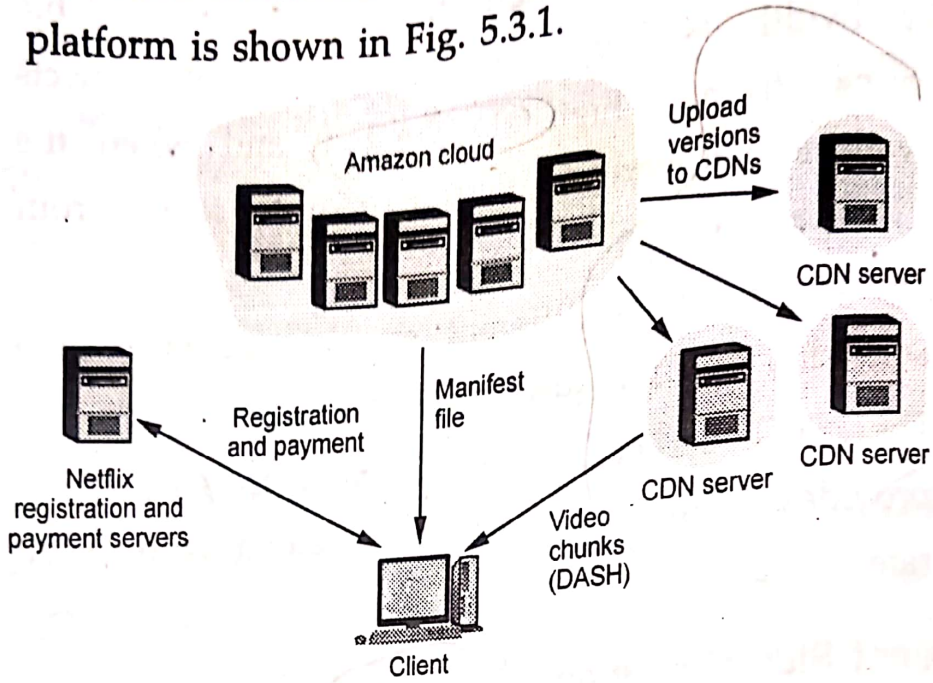


Fig. 5.3.1 Netflix operation

- Registration and payment servers handle registration of new accounts and capture credit-card payment information.
- Amazon (3rd party) cloud services includes :
  1. Netflix uploads studio master to Amazon cloud
  2. Create multiple version of movie (different encodings) in cloud
  3. Upload versions from cloud to CDNs
  4. Cloud hosts Netflix web pages for user browsing
- To deliver the movies to its customers on demand, Netflix makes extensive use of CDN technology.
- Netflix uses three 3rd party CDNs host/stream content simultaneously: Akamai, Limelight, Level 3

5.3.2 YouTube

- YouTube is an American video-sharing website headquartered in San Bruno, California. YouTube is the world's largest video-sharing site. YouTube now operates as one of Google's subsidiaries.
- YouTube makes extensive use of its own private CDN technology to distribute its videos.
- Google has installed server clusters in many hundreds of different locations.

- Google uses DNS to redirect a customer request to a specific cluster.
- YouTube employs HTTP streaming. The HTTP byte range request to limit the flow of transmitted data after a target amount of video is prefetched.
- YouTube does not employ adaptive streaming (such as DASH), but instead requires the user to manually select a version.
- YouTube often makes available different versions for a video, each with a different bit rate and corresponding quality level.
- YouTube processes each video uploaded, converting it to a YouTube video format and creating multiple versions at different bit rates. This processing takes place entirely within Google data centers.

5.3.3 Kankan

- Kankan is a video-on-demand service with a web interface as well as a client application.
- Kankan, the online video streaming service of download service provider Xunlei, offers paid monthly subscriptions that include online video-related premium offerings and other services.
- Kankan is an entirely different approach for providing video on demand over the Internet at a large scale
- Kankan uses P2P delivery instead of client-server (via CDNs) delivery. P2P video streaming is very similar to BitTorrent file downloading.
- When a peer wants to see a video, it contacts a tracker (which may be centralized or peer-based using a DHT) to discover other peers in the system that have a copy of that video. This peer then requests chunks of the video file in parallel from these other peers that have the file.
- The Kankan protocols for communication between peer and tracker, between peer and DHT, and among peers are all proprietary.
- Kankan uses UDP for distributing video chunks among peers.
- Kankan is also available on mobile platforms including Android and iOS (including iPhone and Kankan HD for iPad).

5.4 Dimensioning Best Effort Network

- An approach to improving the quality of multimedia application is to deploy enough link capacity so that congestion doesn't occur, multimedia traffic flows without delay or loss.
  - Low complexity of network mechanisms (use current "best effort" network)
  - High bandwidth costs
- The main challenges are :
  1. Network dimensioning : how much bandwidth is "enough?"
  2. Estimating network traffic demand : needed to determine how much bandwidth is "enough" (for that much traffic)

5.5 Providing Multiple Classes of Service

- In any multimedia application audio/video packets are delay sensitive but by internet all packets are treated equally i.e. QoS offered is same for all applications. This causes congestion in traffic followed by delay and loss of packets.
- Analyzing varying network scenarios principles of Quality of Services (QoS) needed for multimedia applications are derived.

Principle 1 : Packet marking allows a router to distinguish among packets belonging to different classes of traffic.

Modified principle 1 : Packet classification allows a router to distinguish among packets belonging to different classes of traffic.

Principle 2 : A degree of isolation is desirable among traffic flows, so that one flow is not adversely affected by another misbehaving flow.

Principle 3 : For isolating flows, it is desired to use resources like BW and buffers as efficiently as possible.

Principle 4 : A call admission process is needed where flows declare their QoS requirement.

5.6 Scheduling and Policing

- Scheduling and policing studies various mechanisms used to provide guaranted QoS.

5.6.1 Scheduling Mechanisms

- Packets from different networks are to be multiplexed and queued in buffer for transmission on a link. The methods of selecting queued packets for transmission are called as **link scheduling discipline**. It plays an important role in providing better QoS.

Different queuing methods are -

1. First-In-First Out (FIFO)
2. Priority queuing
3. Round robin queuing
4. Weighted Fair Queuing (WFQ)

1) First-In-First Out (FIFO)

- In FIFO link scheduling mechanism, packets arriving are queued for transmission on link on first come first serve basis. Fig. 5.6.1 shows this mechanism.
- The packets arriving are stored in buffer as they arrive, when the buffer is full the packets are discarded.
- The operation of FIFO queuing mechanism is illustrated in Fig. 5.6.2.

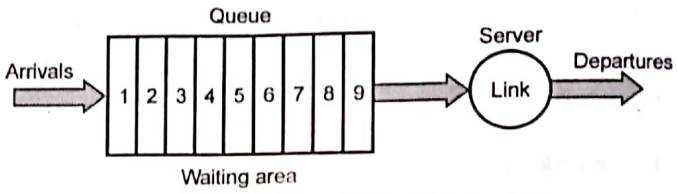


Fig. 5.6.1 FIFO queuing mechanism

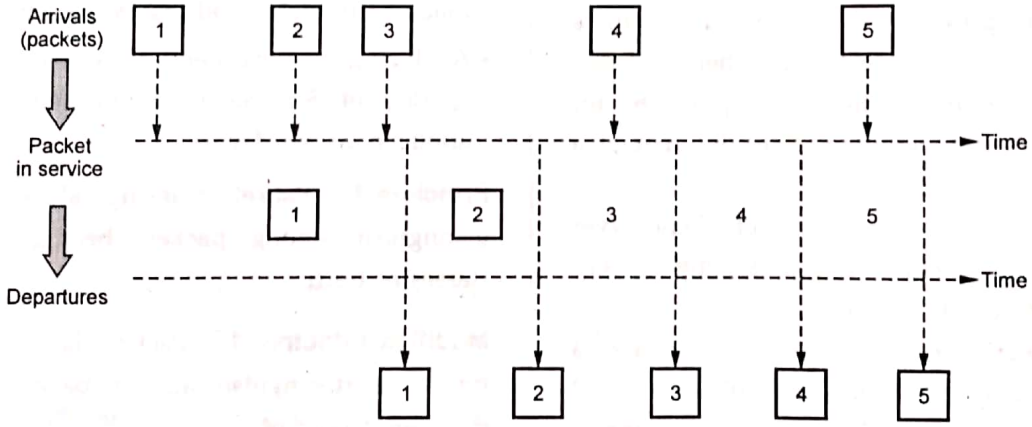


Fig. 5.6.2 FIFO queuing operation

- Arriving packets are shown in numbers on time axis. The packets are then processed by servers, definite time is required to service a packet. Then the packets are sent for departure on link. The packets leane in sequence as they arrived.

2. Priority queuing

- In priority queuing the packets arriving are classified into two or more priority classes. Each priority class has a separate queue. The packet from highest priority queue is selected for transmission. Among same priority class the packet is selected on FIFO manner. Fig. 5.6.3 shows priority queuing model.

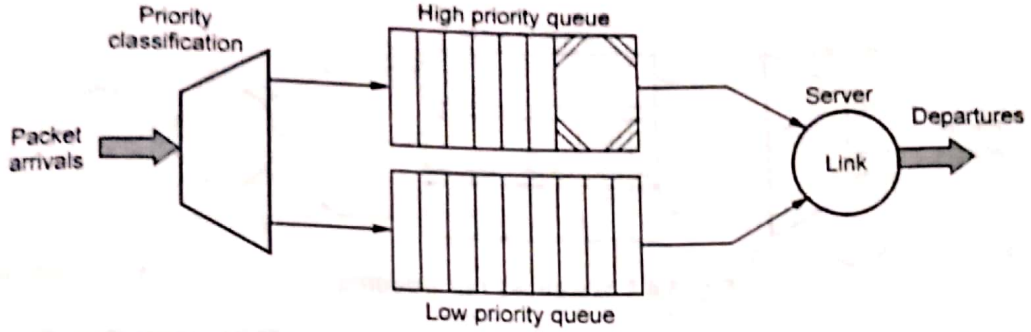


Fig. 5.6.3 Priority queuing model

The operation of priority queuing model is shown in Fig. 5.6.4.

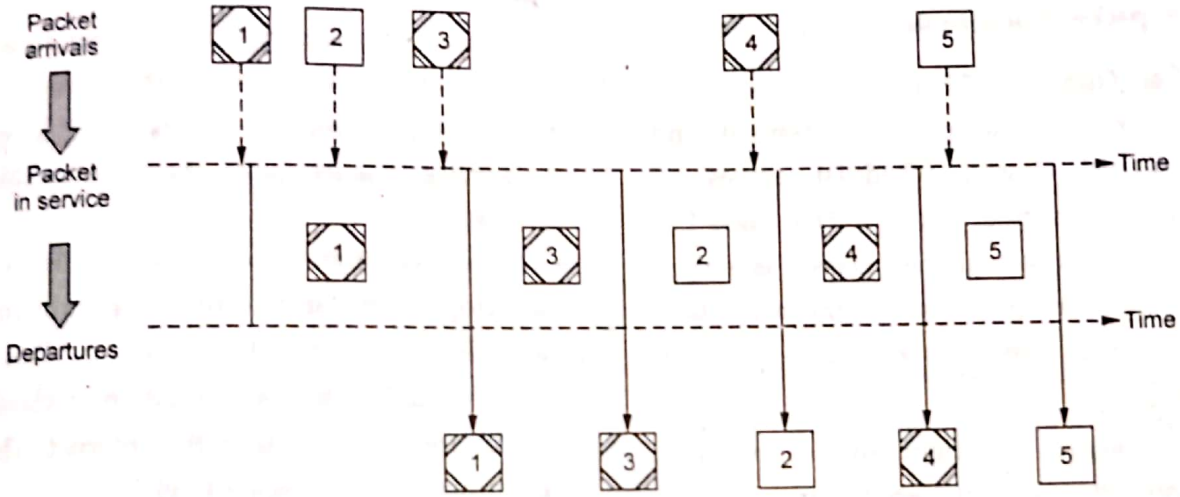


Fig. 5.6.4 Priority queuing model operation

Two priority classes are shown in Fig. 5.6.4 packets 1, 3, 4 are from high priority class and packets 2, 5 are from low priority class. High priority class packets are transmitted first than the low priority class packets.

Round robin queuing

In round robin queuing packets are again sorted into classes i.e. round robbin scheduler alternately switches transmitting link among the classes.

Fig. 5.6.5 shows round robin queuing of two classes.

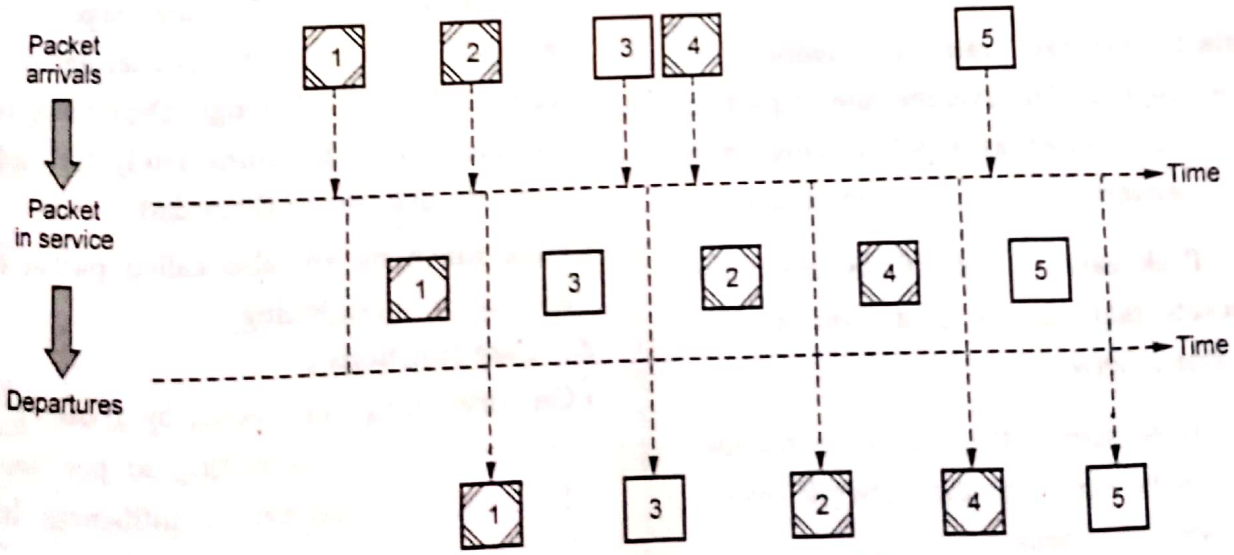


Fig. 5.6.5 Round robin queuing model

- Packets 1, 2, 4 are from class-1 and packets 3, 5 are from class-2 on transmission of packet 1, the link scheduler looks for class-2 packets i.e. packet 3 then for class-1 i.e. packet 3 and so on.

4. Weighted Fair Queuing (WFQ)

- In WFQ the arriving packets are classified and queued in several classes, schedular serves all classes of queues in circular manner. First class-1 is served then class-2 is served then class-3 is served and the service pattern moves on to the next class. And thus the service pattern is repeated. Fig. 5.6.6 shows the WFQ.
- Each class is assigned a weight and as per this weight the service time of that class is varying.

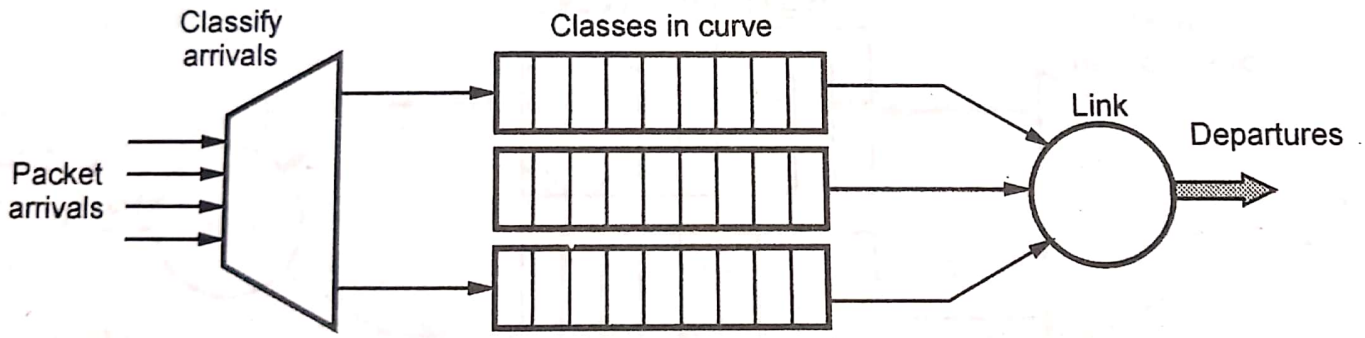


Fig. 5.6.6 Weighted fair queuing

5.6.1.1 Policing

- Policing is the regulation of the rate at which packet flow is injected into the network.

Criteria for policing

- Three important policing criterias are identified, these are :

1. Average rate 2. Peak rate 3. Burst size

1. Average rate : Average rate is defined as packets per time interval. The average rate of packets in a network can be limited as a policy. This limits the traffic in the network for a long period of time.

2. Peak rate : Peak rate is defined as maximum number of packets that can be sent over a short period of time over a network.

3. Burst size : Burst size is the maximum number of packets that can be sent into the network over a extremely short interval of time.

5.7 Diffserv (Differentiated Services)

- Some difficulties associated with RSVP and Intserv model are :

1. **Scalability :** Reservation requests cause significant overhead in large networks.
2. **Flexible service models :** The pre-specified service classes make them vulnerable to router crashes.

- The Differentiated Services (DiffServ) group has developed an architecture for providing scalable and flexible service differentiation. This architecture has the ability to handle different classes of traffic in different way within the internet. This approach is known as class-based QoS.

5.7.1 Functional Elements of Differentiated Service

- The Differentiated Services (DiffServes) architecture consists of two sets of functional elements :
  1. Edge functions
  2. Core functions

1. Edge functions :

- The packets arriving at the edge of network are marked. The mark of the packet defines the class of traffic to which it belongs. Depending on the mark, the packet may be immediately forwarded into the network, delayed or discarded.
- Edge functions are also called packet classification and traffic conditioning.

2. Core functions :

- On forwarding the packet by router it is then put on for next hop according to per hop behaviour. The per hop behaviour influences how router's buffer and BW are shared. It is a forwarding function of Diffserv.

- Fig. 5.7.1 shows a logical view of classification and marking function within the edge7 router.

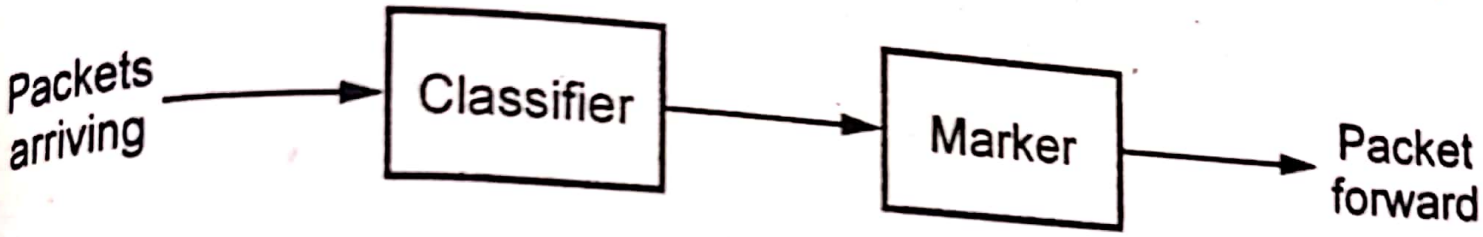


Fig. 5.7.1 Packet classification and marking

5.8 Per-Connection Quality-of-Service (QoS) Guarantees :

Resource Reservation and Call Admission

- Integrated service is a framework to provide guaranteed Quality of Service (QoS) to individual application sessions.
- A call step process involves following steps :
  1. Traffic characterization and specification of desired QoS.
  2. Signalling for call setup.
  3. Pre element call admission.

- The intserv architecture defines two major classes of service.
  a) Guaranteed service.
  b) Controlled load service.

5.8.1 Guaranteed Quality of Service

- The guaranteed QoS specification specifies queuing delays that a packet experience in a router.
- The actual delay is subject to the peak rate limitation of the input link and variations in the packet is transmission time.

5.8.2 Controlled-Load Network Service

- A session with controlled-load service will receive same QoS as from unloaded network, i.e. all the packets will successfully pass through router without loss and will experience a zero queuing delay.
- The controlled-load network service is developed mainly for real-time multimedia application over Internet.