Multimedia Networking

School of Informatics
Institute for Computing Systems Architecture

(Networked) Multimedia Applications

- Definition: Networked applications that employ audio or video

- Commonly used nowadays (e.g., YouTube, BBC iPlayer, Netflix, Skype)

- It would take an individual more than 5 million years to watch the amount of video that will cross global IP networks each month in 2021*.
- IP video traffic will be 82 percent of all the consumer Internet traffic by 2021*.

*Source: Cisco Visual Networking Index: Forecast and Methodology, 2016-2021

(Networked) Multimedia Applications

- Traditional elastic application like email, file transfer and web-browsing
  - **Are delay tolerant but loss in-tolerant**

- While multimedia applications
  - Can tolerate loss
  - But **not delay (and jitter)**, especially for interactive audio/video communication and live streaming
    - For conversational voice, delays **≤ 150ms** for best user experience while **> 400ms** unacceptable

Outline

- Multimedia data (audio and video)
  - Features
  - Compression
  - Service requirements
  - Design issues and protocols

Multimedia Data

-Properties of Audio Data

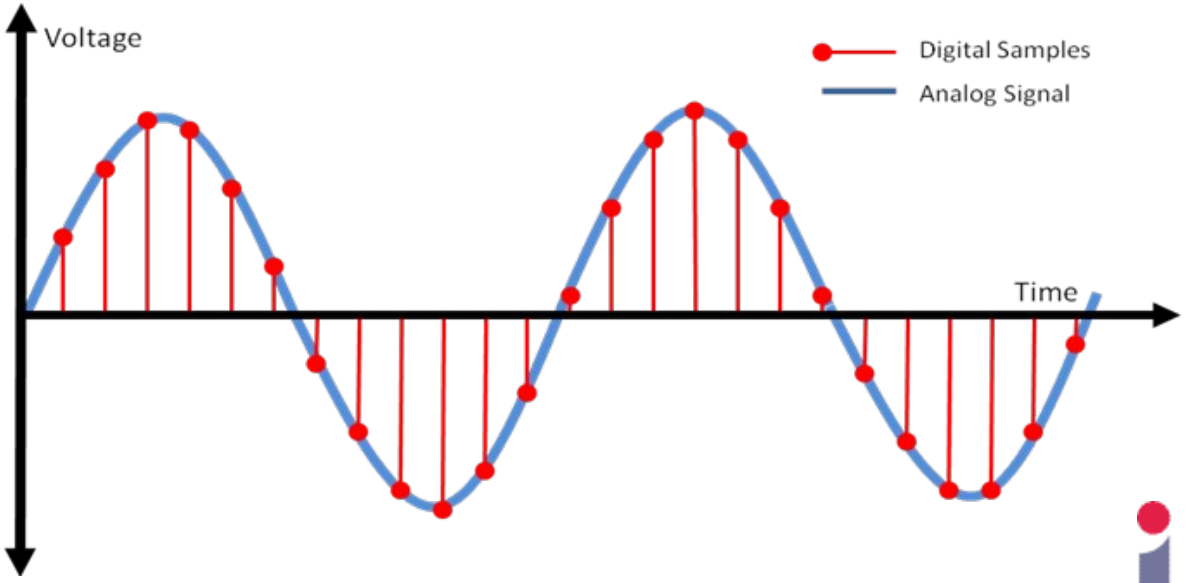
-Properties of Video Data

THE UNIVERSITY OF EDINBURGH

School of Informatics
Institute for Computing Systems Architecture

Properties of Audio Data
(Pulse Code Modulation)

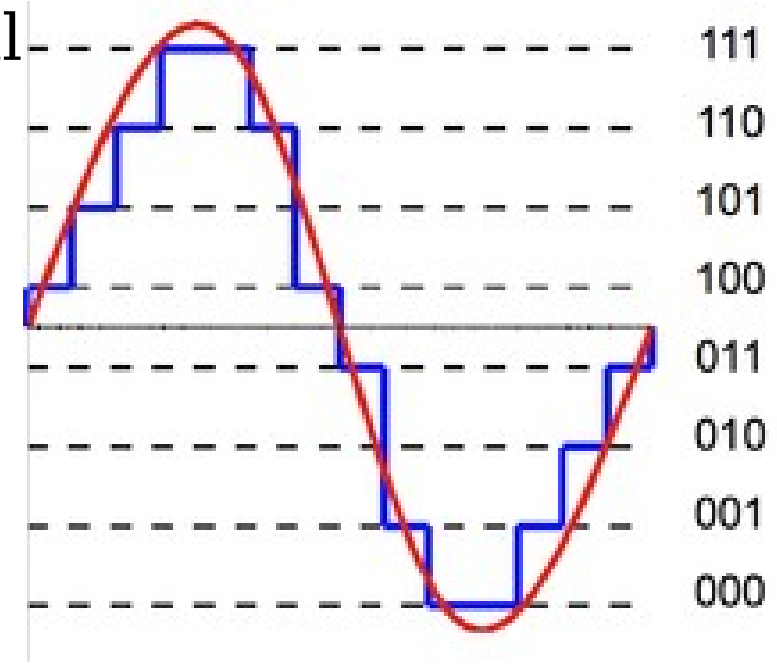
-Digital Audio data has much less bandwidth requirements than that of Video.
-Analog audio signal is **sampled** at a fixed rate (e.g. 8000 samples/sec, Human hearing and voice range is about 20 Hz to 20 kHz, most sensitive at 2 to 4 KHz.). The value of each sample is a real value.

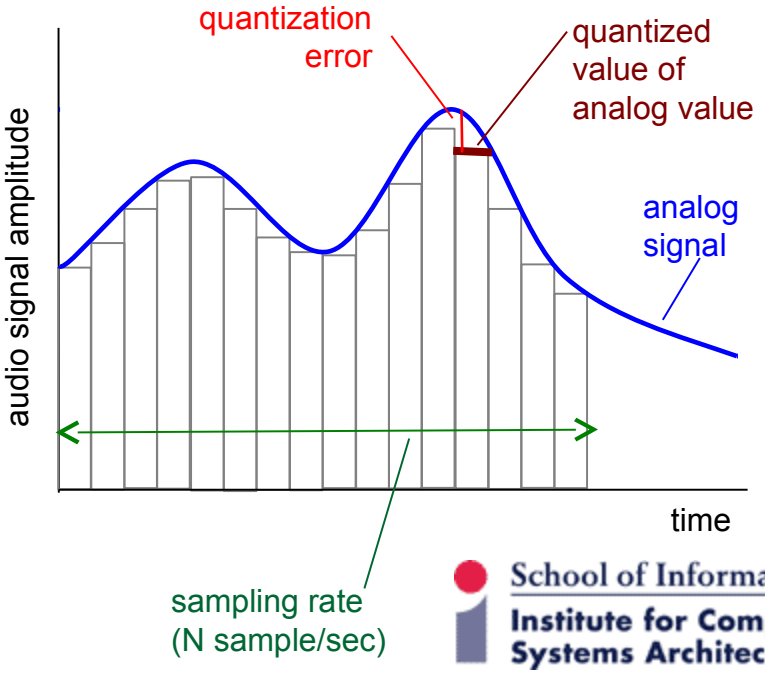


Properties of Audio Data
(Pulse Code Modulation)

-Each sample value is then rounded to one of the finite number of values. This operation is called **quantization**.

-The finite values are called quantization values, and are always in power of 2, e.g. 256 quantization val





THE UNIVERSITY OF EDINBURGH
School of Informatics
Institute for Computing Systems Architecture

Properties of Audio Data
(Pulse Code Modulation)

-Number of bits required to send a sample is thus a Byte in our example.
-Once each sample is assigned a quantization value, a digital signal is formed by concatenating the corresponding quantized samples. In our example **64000 bps** is the rate of digital signal.
-When playing back the digital signal via an audio speaker, it is decoded back into audio signal; but due to loss in sampling and quantization steps… **original signal cannot be recovered**.

Properties of Audio Data
(Pulse Code Modulation)

-**Audio CD** also uses PCM: sampling rate of 44100 samples per second and 16 bits per sample on a linear scale

-705.6 Kbps (mono) and 1.411 Mbps (stereo).

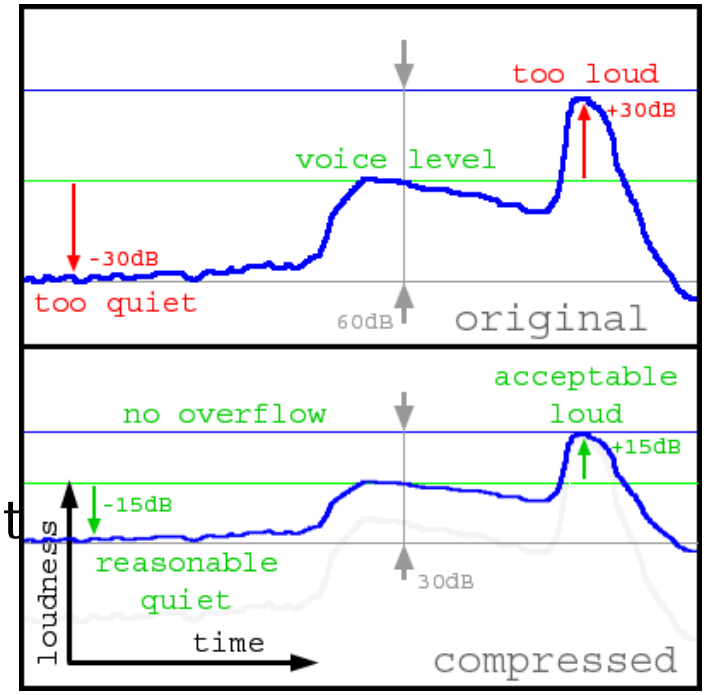
-**PCM**-encoded speech and music is **rarely used** in Internet.

-The reason is to have reduced bit rate.

Properties of Audio Data
(Audio Compression Overview)

- Audio transmitted on the Internet is typically compressed with a reduced bit rate.
- Human speech offers a great potential for compression to as low as **2.4 Kbps**
- MPEG 1 audio layer 3 (**MP3**): popular audio compression technique for music streaming encodes at different layers, the most common is 128 Kbps rate.
- **Ears more sensitive than eyes** Audio glitches need to be kept minimal to ensure good user experience

MP3 compression



THE UNIVERSITY OF EDINBURGH
School of Informatics
Institute for Computing Systems Architecture

Properties of Audio Data
(MP3 Compression)

Also Known as Perceptual Coding

-Remove frequencies that human ear cannot hear (i.e. very high and low frequencies are removed).

-Also if both a louder and softer sounds are played at same time, the filtering ignores the softer ones (i.e. simultaneous/temporal masking). Our ears and minds cannot separate events that are close in time.

-As a sound becomes quiter and quiter,humans are able only to make out less and less detail. The encoder thus not saves every single detail of quite sounds (i.e. minimum audition threshold)

-Reduction by a factor of 10 (33MB songs on CD can be compressed to about 3MB)

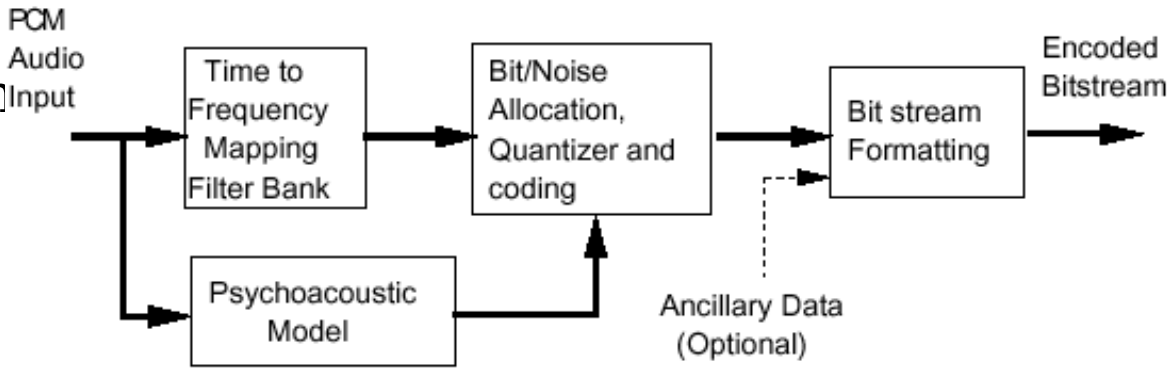
Properties of Audio Data
(MP3 Compression steps)

1. Filter Bank: It divides sounds into sub-bands of frequency

2. Psychoacoustic model: It utilizes the concept of auditory masking, that determines what can and cannot be heard in each sub-band.

3. Quantization and Coding: Bit code allocation to remaining samples.

4. Bit ıll the inform



MPEG/Audio Encoder

Properties of Video Data (High bit rate requirement)

- High bit rate compared to music streaming and image transfers.
  - From **100 Kbps** for low quality video conferencing to over **3 Mbps** for streaming HD movies
  - Internet video streaming and downloads will grow to more than **80%** of all consumer Internet traffic by 2019 (Data via Cisco)

	Bit rate	Bytes transferred in 67 min
Facebook Frank	160 kbps	80 Mbytes
Martha Music	128 kbps	64 Mbytes
Victor Video	2 Mbps	1 Gbyte

Comparison of bit rate requirements of three internet applications.

Properties of Video Data
(Video Compression Overview)

- Video is a sequence of images that are typically displayed at a constant rate, typically referred to as frames per second (fps), e.g., **24 or 30 fps**

Properties of Video Data
(Video Compression Overview)

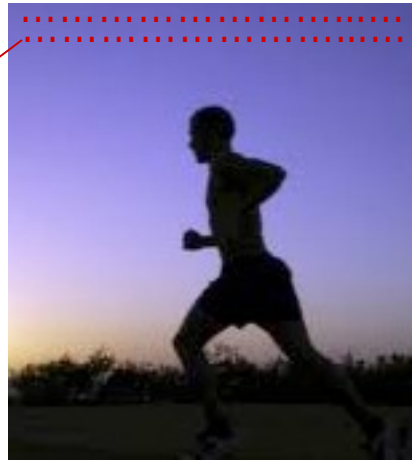
- Image is an array of pixels, each encoded into a number of bits to represent luminance and color.

- Two types of redundancies in Video
  - **Spatial redundancy:** An image having large portion with white spaces or similar color can be compressed much without quality degradation.
  - **Temporal redundancy:** When an image and a subsequent image are same, no need to encode the second image,

Properties of Video Data
(Video Compression Overview)

- Video can exploit both spatial and temporal redundancy
- Multiple versions are created of same video for adaptive delivery depending on network conditions (end-to-end available bandwidth) and host characteristics

spatial coding example:
instead of sending N values of same color (all purple), send only two values: color value (purple) and number of repeated values (N)

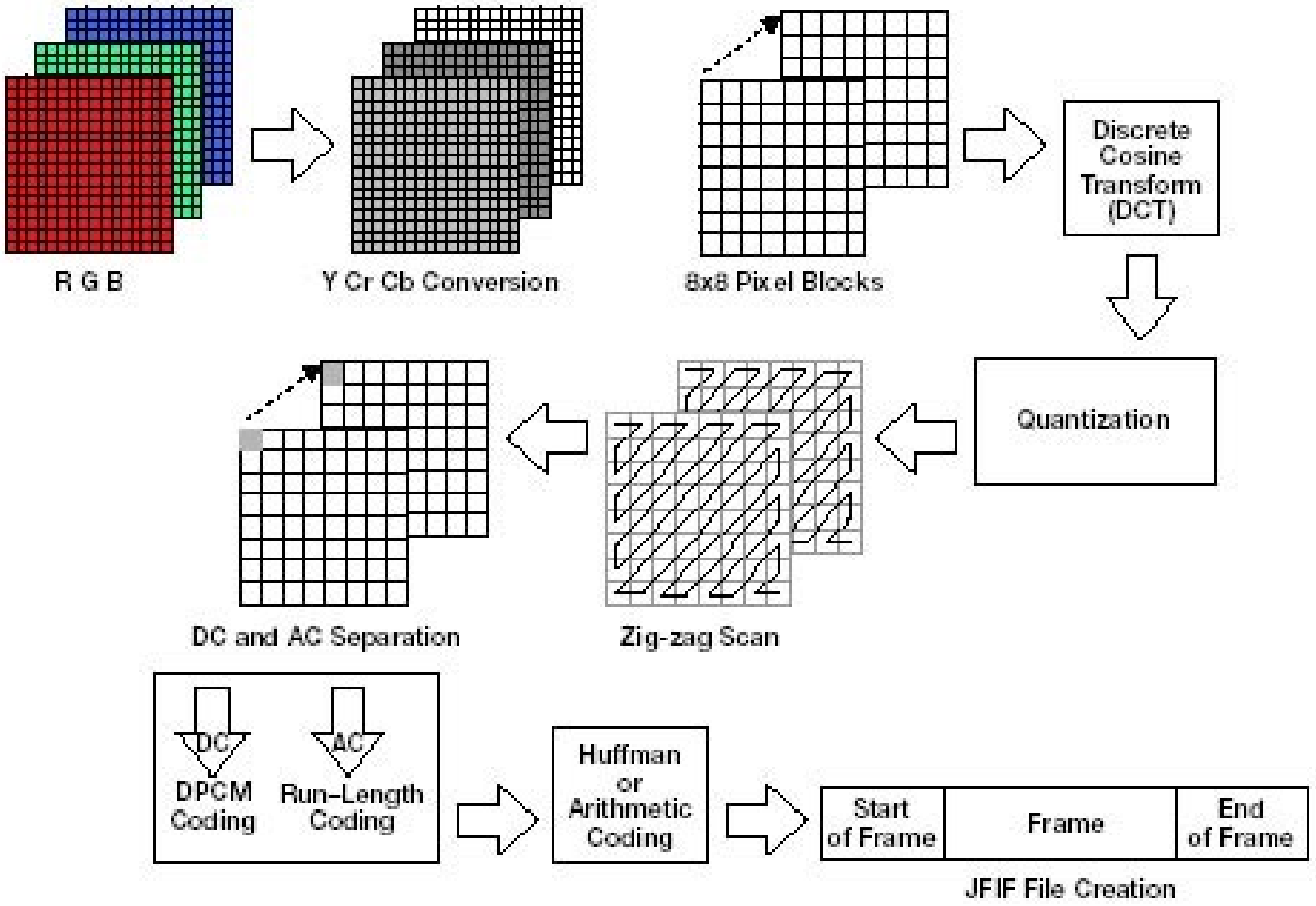


frame i

temporal coding example:
instead of sending complete frame at i+1, send only differences from frame i

frame i+1

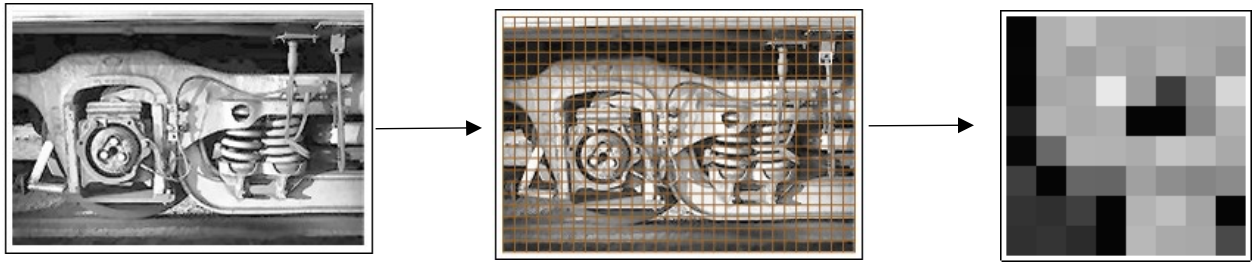
Properties of Video Data
(Spatial Compression: JPEG)



THE UNIVERSITY OF EDINBURGH
School of Informatics
Institute for Computing Systems Architecture

Properties of Video Data
(Spatial Compression: JPEG)

1. Preprocessing



A

RGB to YCbCr

Divide into 8X8 blocks

5	176	193	168	168	170	167	165
6	176	158	172	162	177	168	151
5	167	172	232	158	61	145	214
33	179	169	174	5	5	135	178
8	104	180	178	172	197	188	169
63	5	102	101	160	142	133	139
51	47	63	5	180	191	165	5
49	53	43	5	184	170	168	74

A

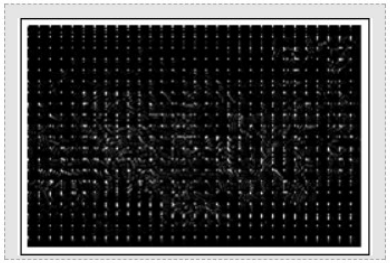
-122	49	66	41	41	43	40	38
-121	49	31	45	35	50	41	24
-122	40	45	105	31	-66	18	87
-94	52	42	47	-122	-122	8	51
-119	-23	53	51	45	70	61	42
-64	-122	-25	-26	33	15	6	12
-76	-80	-64	-122	53	64	38	-122
-78	-74	-84	-122	57	43	41	-53

B

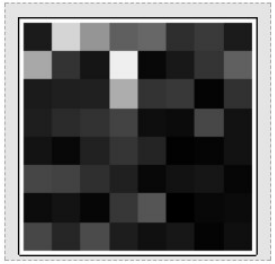
Subtract 127 from each pixel intensity

Properties of Video Data
(Spatial Compression: JPEG)

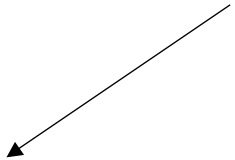
2. Transformation : C=UBU' (where U is an 8x8 special block and B is the last block in the Pre-processed step), DCT pushes high intensity values to left upper corner.



U: Fixed block



C: DCT block



-27.500	-213.468	-149.608	-95.281	-103.750	-46.946	-58.717	27.226
168.229	51.611	-21.544	-239.520	-8.238	-24.495	-52.657	-96.621
-27.198	-31.236	-32.278	173.389	-51.141	-56.942	4.002	49.143
30.184	-43.070	-50.473	67.134	-14.115	11.139	71.010	18.039
19.500	8.460	33.589	-53.113	-36.750	2.918	-5.795	-18.387
-70.593	66.878	47.441	-32.614	-8.195	18.132	-22.994	6.631
12.078	-19.127	6.252	-55.157	85.586	-0.603	8.028	11.212
71.152	-38.373	-75.924	29.294	-16.451	-23.436	-4.213	15.624

Properties of Video Data
(Spatial Compression: JPEG)

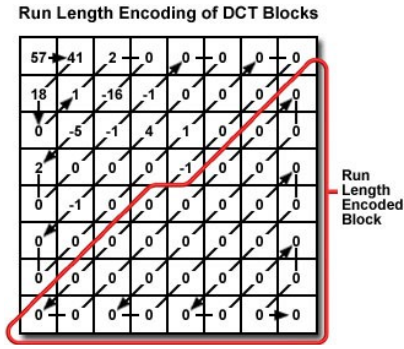
3. Quantization : Elements near zero will converted to zero and other elements will be shrunk so that their values are closer to zero. Divide each element in set C by corresponding element in set Z (a predefined set) and round off the resultant value.

$$C = \begin{bmatrix} -27.50 & -213.47 & -149.61 & -95.28 & -103.75 & -46.95 & -58.72 & 27.23 \\ 168.23 & 51.61 & -21.54 & -239.52 & -8.24 & -24.50 & -52.66 & -96.62 \\ -27.20 & -31.24 & -32.28 & 173.39 & -51.14 & -56.94 & 4.00 & 49.14 \\ 30.18 & -43.07 & -50.47 & 67.13 & -14.12 & 11.14 & 71.01 & 18.04 \\ 19.50 & 8.46 & 33.59 & -53.11 & -36.75 & 2.92 & -5.80 & -18.39 \\ -70.59 & 66.88 & 47.44 & -32.61 & -8.20 & 18.13 & -22.99 & 6.63 \\ 12.08 & -19.13 & 6.25 & -55.16 & 85.59 & -0.60 & 8.03 & 11.21 \\ 71.15 & -38.37 & -75.92 & 29.29 & -16.45 & -23.44 & -4.21 & 15.62 \end{bmatrix}$$

$$Z = \begin{bmatrix} 16 & 11 & 10 & 16 & 24 & 40 & 51 & 61 \\ 12 & 12 & 14 & 19 & 26 & 58 & 60 & 55 \\ 14 & 13 & 16 & 24 & 40 & 57 & 69 & 56 \\ 14 & 17 & 22 & 29 & 51 & 87 & 80 & 62 \\ 18 & 22 & 37 & 56 & 68 & 109 & 103 & 77 \\ 24 & 35 & 55 & 64 & 81 & 104 & 113 & 92 \\ 49 & 64 & 78 & 87 & 103 & 121 & 120 & 101 \\ 72 & 92 & 95 & 98 & 112 & 100 & 103 & 99 \end{bmatrix}$$

$$Q = \begin{bmatrix} -2 & -19 & -15 & -6 & -4 & -1 & -1 & 0 \\ 14 & 4 & -2 & -13 & 0 & 0 & -1 & -2 \\ -2 & -2 & -2 & 7 & -1 & -1 & 0 & 1 \\ 2 & -3 & -2 & 2 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & -1 & -1 & 0 & 0 & 0 \\ -3 & 2 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

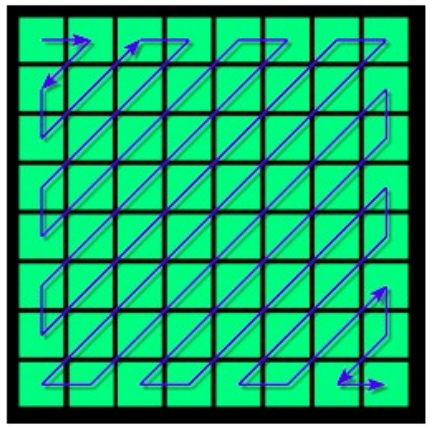
Properties of Video Data
(Spatial Compression: JPEG)



3. Encoding : Original Image: 160*240*8 = 307,200 bits

Transformed Image: 85,143 bits, (saving of over 70%)

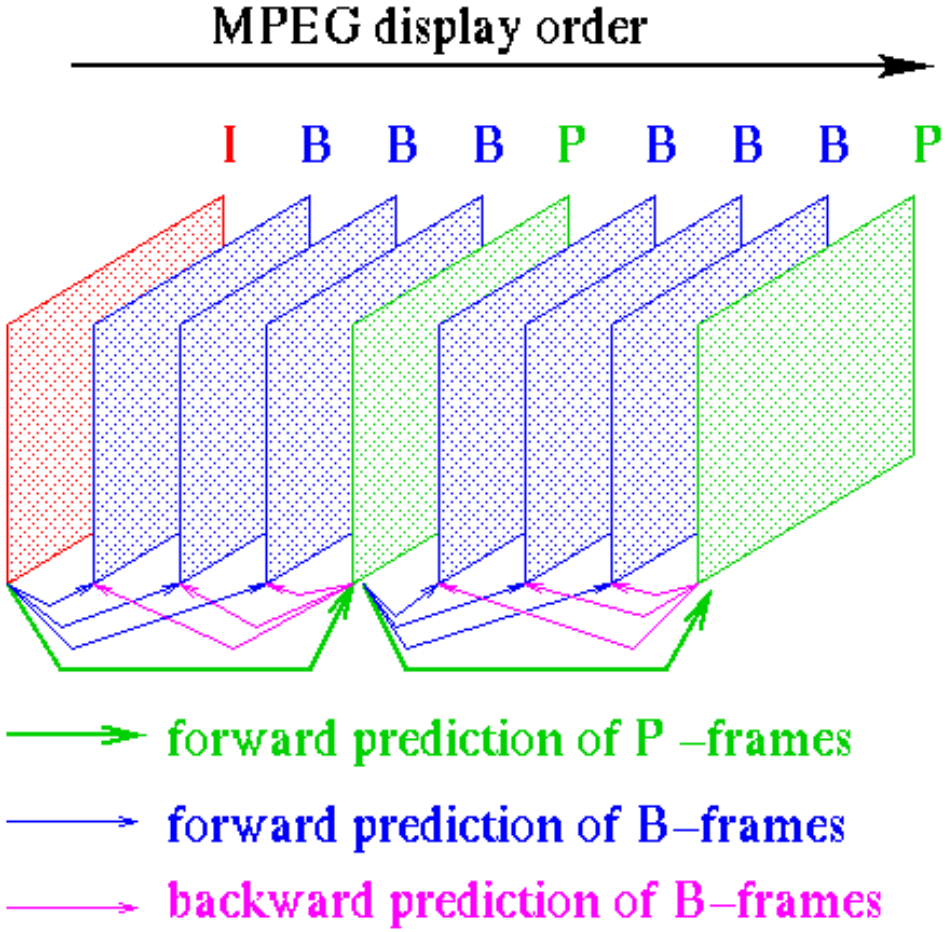
$$\begin{bmatrix} -2 & -19 & -15 & -6 & -4 & -1 & -1 & 0 \\ 14 & 4 & -2 & -13 & 0 & 0 & -1 & -2 \\ -2 & -2 & -2 & 7 & -1 & -1 & 0 & 1 \\ 2 & -3 & -2 & 2 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & -1 & -1 & 0 & 0 & 0 \\ -3 & 2 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$



(Most a few non-zero values and then string of zeros represented as EoF)

Properties of Video Data
(Temporal Compression: MPEG)

1. **Intra-frame (I)**
2. **Predicted frame (P)**
3. **Bi-directional frame (B)**



Properties of Video Data
(Temporal Compression: MPEG)

- MPEG (Moving Picture Expert Group) is broken up in **Group of Pictures (GoP)**
- GoP consists of I,P and B frames.
- Transmit order IPBBBPBBB
- The first frame must be **I frame**: It consists of the the entire picture.

P frame

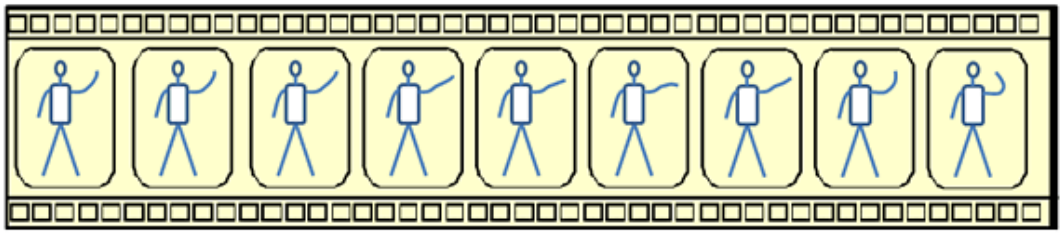
- P frame contains the difference in information from the I frame.
- So it may consist some 40 to 50 % of the information that I frame originally contains

- **B frame** shows how the things move in between (across a macro block 16x16)
  - It relates both to the I and P frame
  - It contains information, how to render the objects in macroblocks (like their directions or vectors)

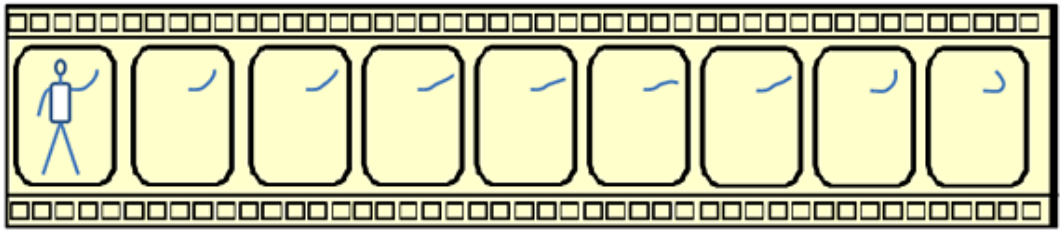
Properties of Video Data
(Temporal Compression: MPEG)

- I and P frames in a GoP are generated and transmitted before B frames
- But each frame has given a sequence number
- The decoder at receiving end then positions them correctly before de-compression.

GoP



IBBBPBBBP



Only the differences between frames are encoded for each GoP

Reference: http://joshweeklyatsit.blogspot.co.uk/2016/03/interframe-and-intra-frame.html

Networked Multimedia Applications

School of Informatics
Institute for Computing Systems Architecture

3 Types

1.Streaming stored audio/video (e.g., Netflix, Hulu, Kankan, BBC iPlayer, YouTube, YouKu, Dailymotion)

2.Streaming live audio/video (e.g., Internet radio, IPTV)

3.Conversational voice and video over Internet (e.g., Skype, Google Talk)

1. Streaming Stored Audio and Video (Features)

- Media is **pre-recorded**.
  - Online movie watching sites (Netflix, Hulu, Lovefilm, Kankan, etc.)
  - Catch up TV (BBC iPlayer, etc.)
  - Online video sharing sites (YouTube, Dailymotion, YouKu, etc.)

- Streaming stored video contains **both video and audio** components
  - –Streaming stored audio alone much lower bit rates, hence less challenging
  - –Streaming stored video sometimes also referred to as video-on-demand **(VoD)**

1. Streaming Stored Audio and Video (Storage/ sharing)

- Multimedia content stored on servers, possibly in **different encodings.**

- Content is often placed on a content distribution network **(CDN)** , rather than a single data centre, for faster access.

- With **P2P** streaming applications, peers hold different chunks of content and they collectively form the “server”. The chunks arrive from different peers speeding up sharing.

1. Streaming Stored Audio and Video (3 Requirements)

–Streaming

- Video plays out after a few seconds of the beginning of the download
- Simultaneous playout of earlier part and download of later part
- This is called streaming (as opposed to download-and-play)

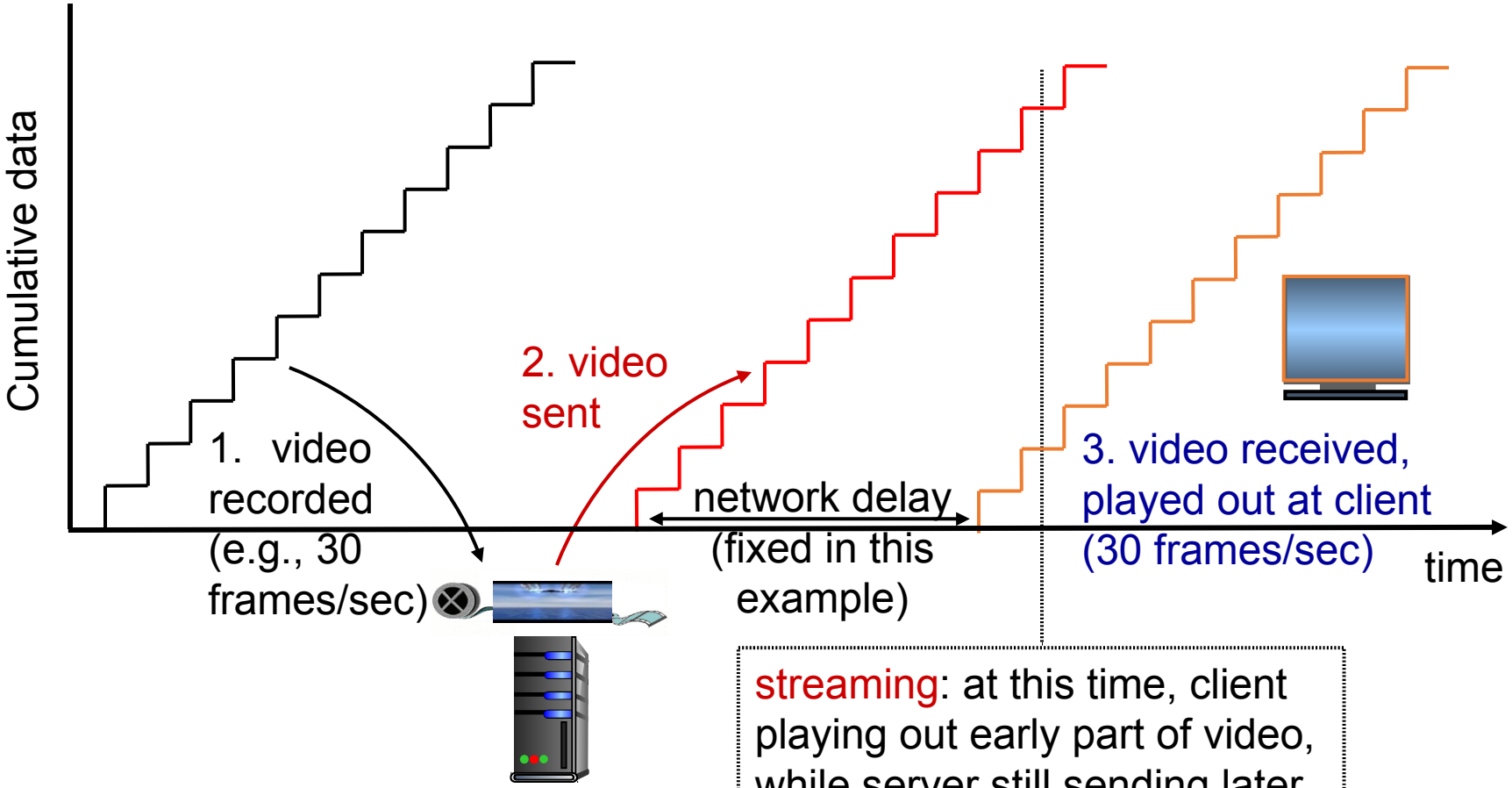
–Interactivity

- Pause, reposition forward, reposition backward, fast-forward
- Response within few seconds

–Continuous

- *When playout* begins, it proceeds according to the original time of recording.
- When not receiving frames in time causes stalls on the client side and degrades user experience

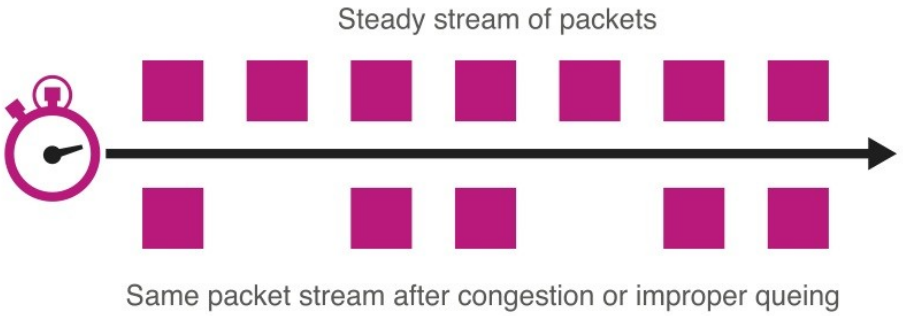
1. Streaming Stored Video
(Streaming: no jitters)



streaming: at this time, client playing out early part of video, while server still sending later part of video

1. Streaming Stored Audio and Video (Continuos)

- Continuous playout and interactivity require low and **stable delays**
- However it is common for Internet paths to exhibit variable end-to-end delays and available bandwidth
  - –This is especially the case for long paths with many links; long paths also increase the initial playout delay
  - –Retransmissions of lost or dropped packets with TCP



Packet Jitter

1. Streaming Stored Audio and Video (Continuos)

- Most important feature is **average throughput**

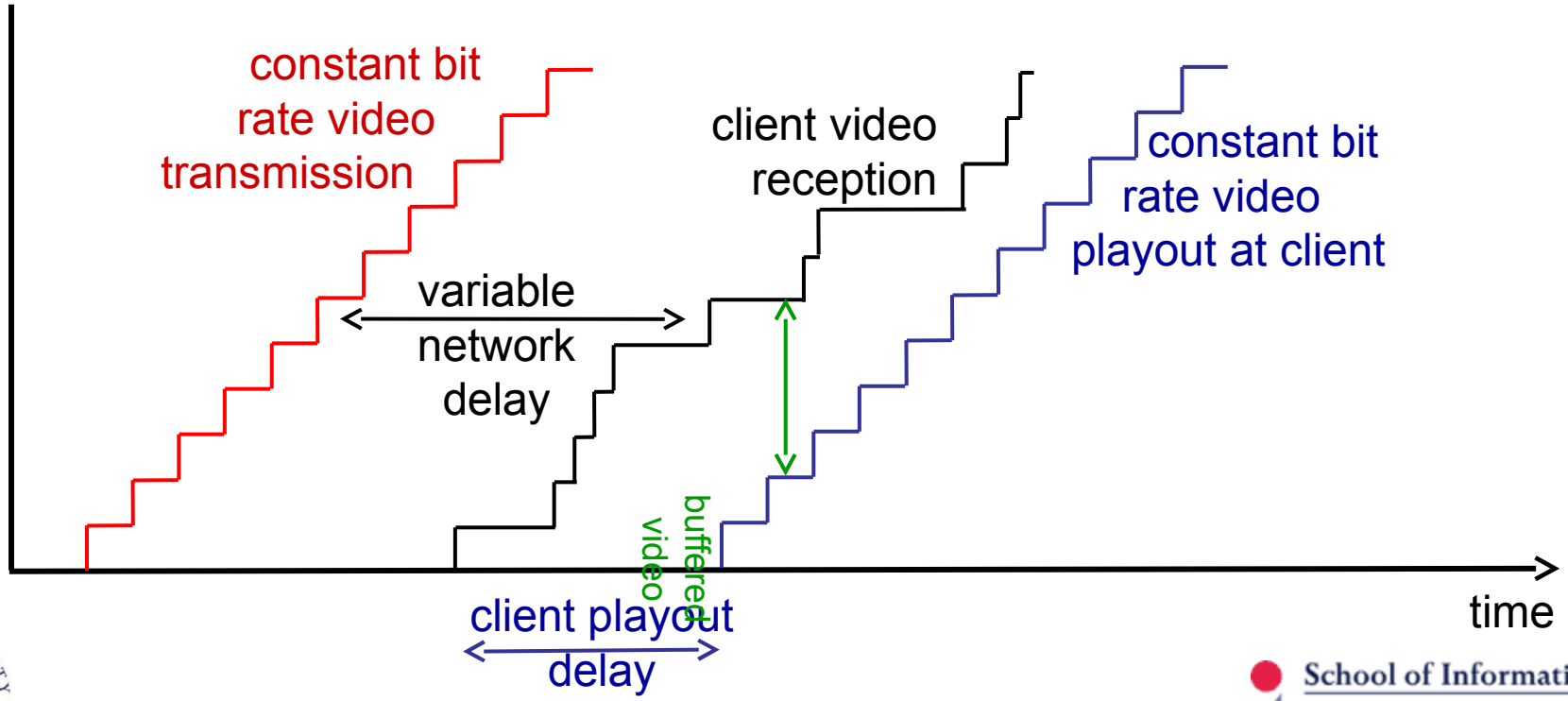
–With buffering and prefetching continuous playout can be maintained even if throughput fluctuates.

–Avg. throughput should be at least equal to the bit rate of video itself.

1. Streaming Stored Video

(Client-Side Buffering)

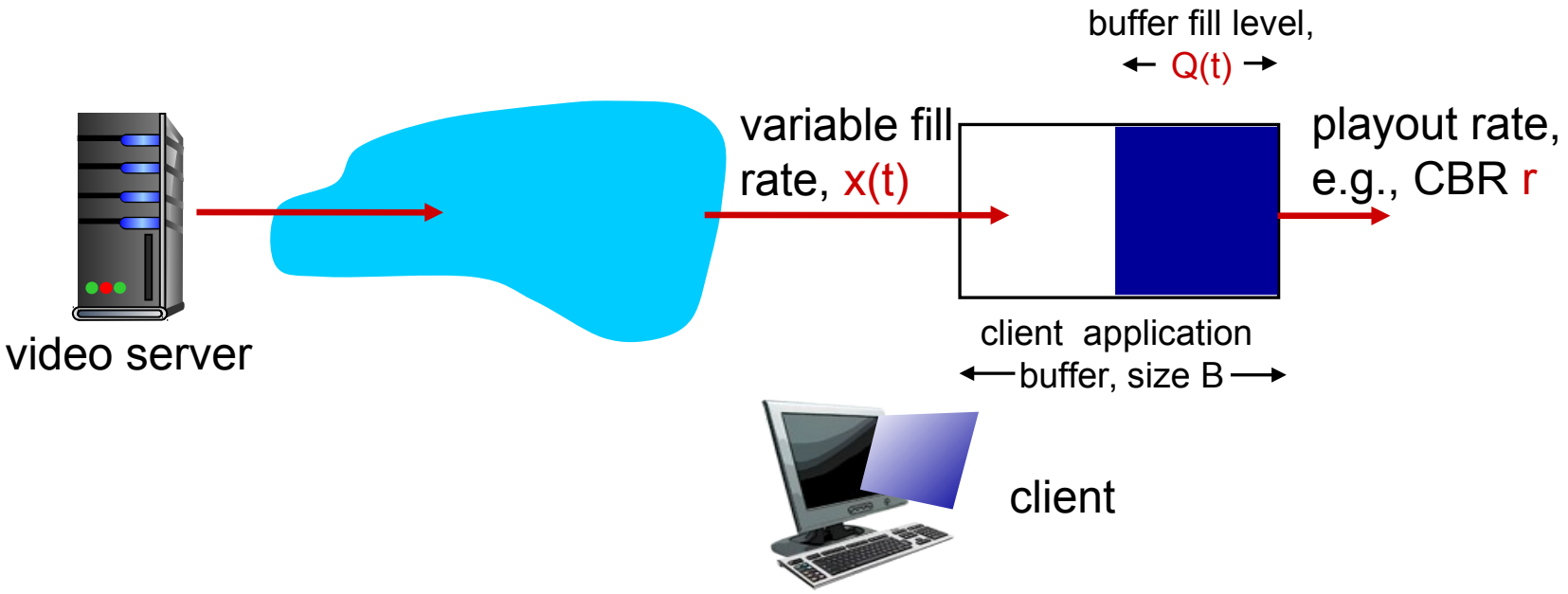
- Common technique employed in all forms of streaming video systems to absorb network delay variability and enable continuous playout
  - Higher the delay variability, longer the playout delay needed



School of Informatics
Institute for Computing
Systems Architecture

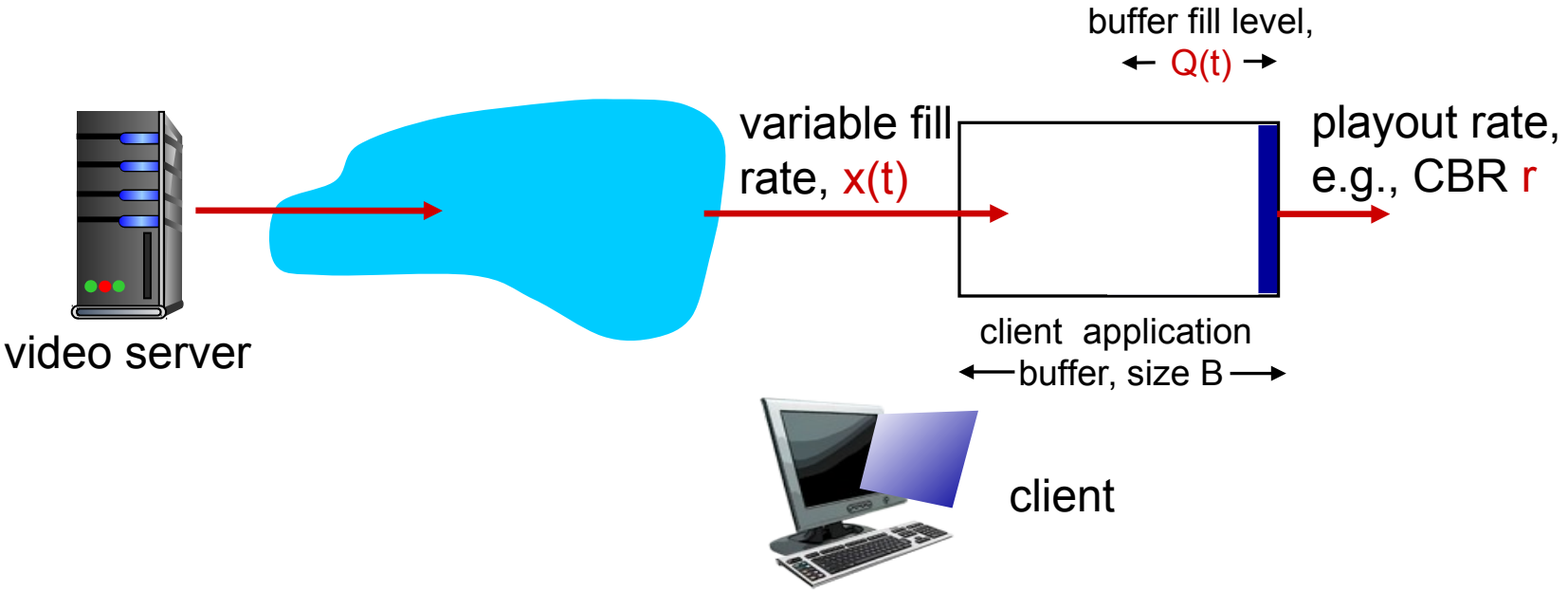
1. Streaming Stored Video

(A different view of Client-Side Buffering)



1. Streaming Stored Video

(A different view of Client-Side Buffering)

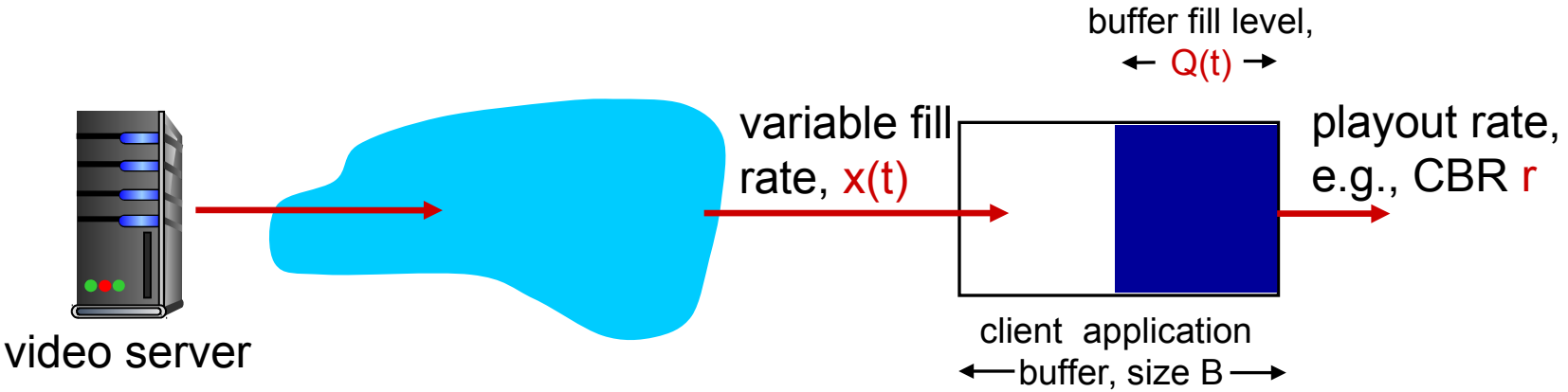


1. Initial fill of buffer until playout begins at t_p
2. Playout begins at $t_{p,}$
3. Buffer fill level varies over time as fill rate x(t) varies and playout rate r is constant

THE UNIVERSITY OF EDINBURGH
School of Informatics
Institute for Computing Systems Architecture

1. Streaming Stored Video

(A different view of Client-Side Buffering)



Key parameters: average fill rate (x_{avg}), playout rate (r)

- **x_{avg} < r**: buffer eventually empties (causing freezing of video playout until buffer again fills)
- **x_{avg} > r**: buffer will not empty, provided initial playout delay is large enough to absorb variability in x(t)
  - **initial playout delay tradeoff**: buffer starvation less likely with larger delay, but larger delay until user begins watching

THE UNIVERSITY OF EDINBURGH
School of Informatics
Institute for Computing Systems Architecture

1. Streaming Stored Video

(3 Categories)

- Three types, depending on mode of transport:

1. UDP streaming
2. HTTP streaming
3. Adaptive HTTP streaming

- Majority of current systems use (adaptive) HTTP streaming
- Optimisation strategies
  - Client buffering and prefetching
    - A common characteristic of the three categories to mitigate the effect of varying end-to-end delay and bandwidth.
  - Adapting video quality to available bandwidth
  - CDN based distribution

1. Streaming Stored Video

(Category1: UDP Streaming)

- Audio/video chunks are placed inside Real-Time Transport Protocol (RTP) or other similar protocol packets, which become the payload in UDP packets
- Server transmit rate matches with client video consumption rate.
  - Example: If client video consumption rate is 2Mbps, and each packet contains 8000 bits, then server should transmit each UDP packet at 8000b/2Mbps=4msec.
  - transmission rate can be oblivious to congestion levels
  - A small buffer at client side is used to hold (2-5sec) video to remove network jitter

1. Streaming Stored Video

(Category1: UDP Streaming)

- For interactivity (i.e. pause, reposition, resume etc.), a separate parallel "control" connection to the server via Real-Time Streaming Protocol (RTSP) is used.
- Limitations:
  - Due to unpredictable and varying amount of available bandwidth, constant-rate streaming may fail to provide continuous playout.
  - Freezing and skipped frames
  - Requirement for media control server (e.g., RTSP server) to support interactivity, increases cost and complexity with scalability
  - UDP traffic maybe blocked by many firewalls

1. Streaming Stored Video

(Category2: HTTP Streaming)

- Multimedia file is stored at HTTP server as an ordinary file.
- Retrieved via URL
  - TCP connection with server
  - HTTP GET request for the URL
  - HTTP response message (the video file)
  - Client buffers incoming video, when it reaches a certain level the playout begins.
- YouTube and Netflix used HTTP streaming over TCP

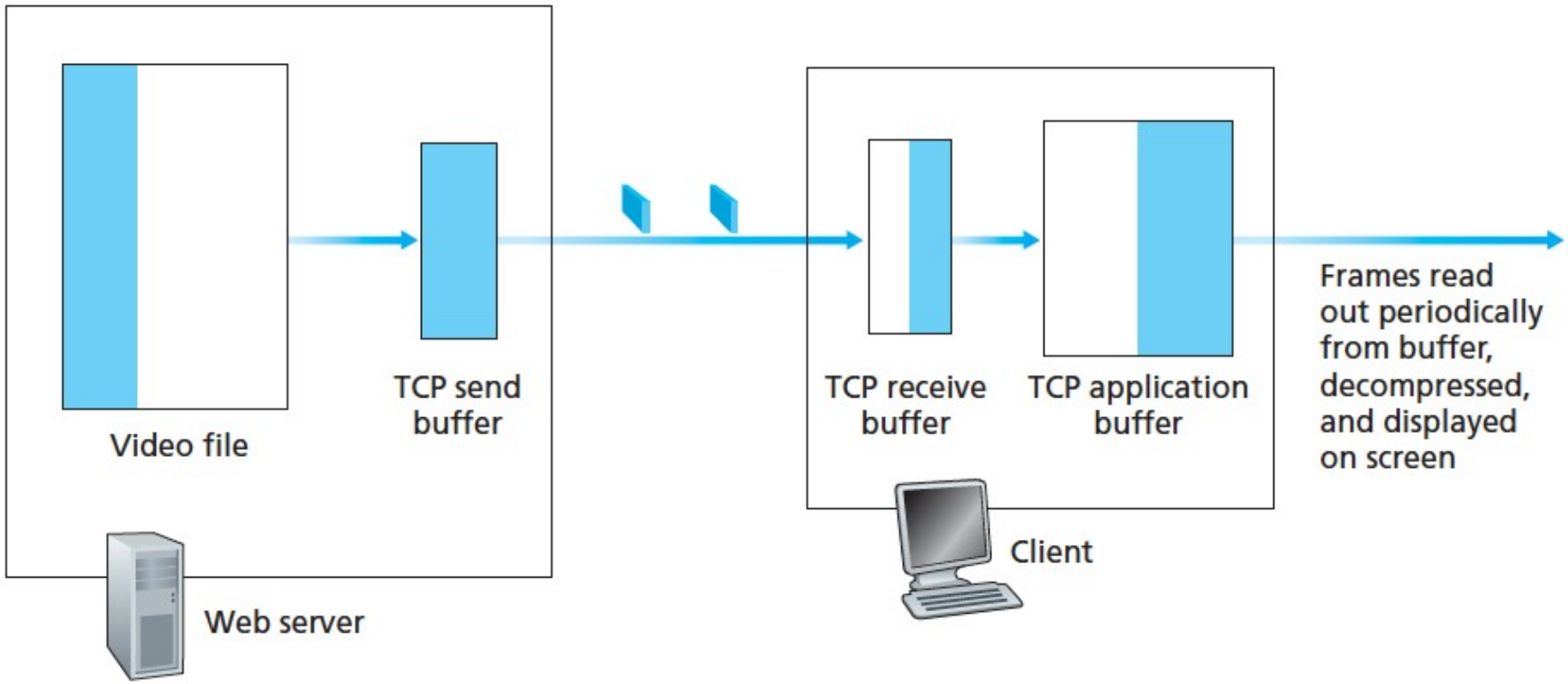
1. Streaming Stored Video

(Category2: HTTP Streaming)

- Features
  - HTTP/TCP passes more easily through firewalls and NATs
  - Can manage without a media control server (like RSTP), thus scalable
  - Fill rate fluctuates due to TCP congestion control, retransmissions (in-order delivery)
  - However keep sending bits at maximum possible rate that TCP allows
  - A form of prefetching from client perspective to handle jitter.
  - Additionally using a larger playout delay smooth TCP delivery rate
  - Early Termination and Repositioning of the video (interactivity)
  - HTTP byte range header (HTTP get message)
    - Server forgets about earlier request and start sending bytes from the point specified in the header.

1. Streaming Stored Video

(Category2: HTTP Streaming)



Streaming stored video over HTTP/TCP

1. Streaming Stored Video

(Category 3: Adaptive HTTP Streaming via DASH)

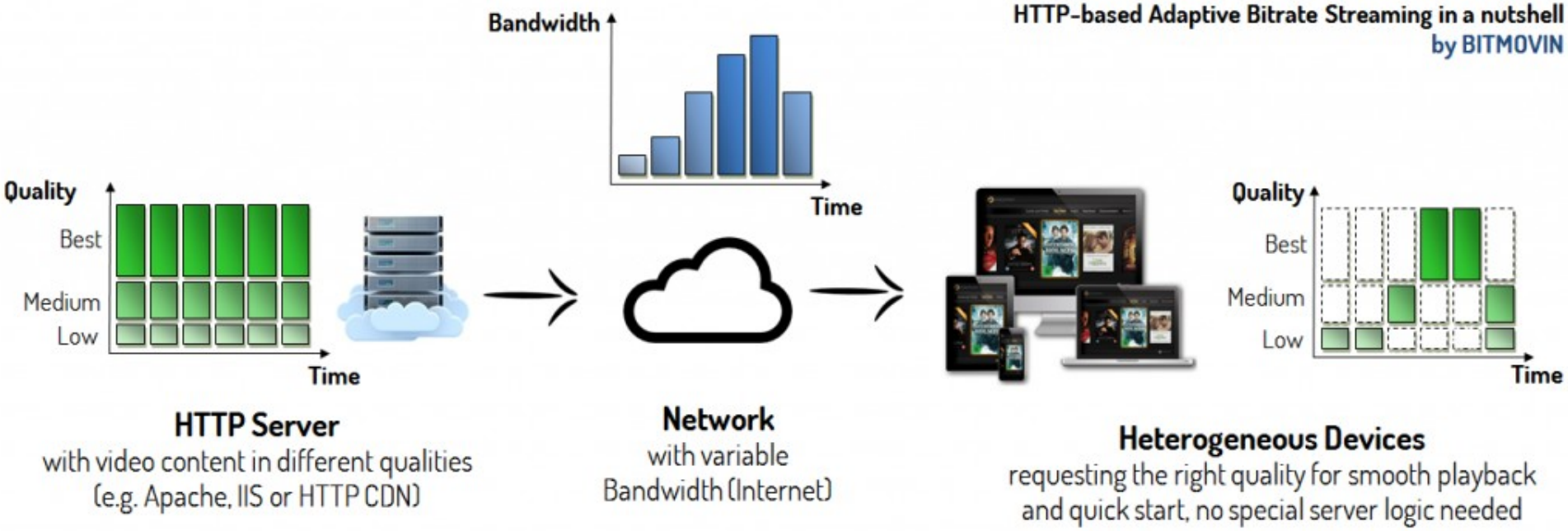
- In HTTP streaming (YouTube at its inception)
  - Every user receives same encoding of video despite differences in bandwidth and bandwidth variation in time.
- *DASH: Dynamic, Adaptive Streaming over HTTP*
- *Server:*
  - divides video file into multiple chunk
  - each chunk stored, encoded at different rates (e.g. 3G, Fiber)

–*manifest file:* provides URLs for different chunks

- *Client:*
  - periodically measures server-to-client bandwidth
  - consulting manifest, requests one chunk at a time
  - chooses maximum coding rate sustainable given current bandwidth via a *rate determination algorithm*
  - can choose different coding rates at different points in time (depending on available bandwidth at any given point in time)

1. Streaming Stored Video

(Category 3: Adaptive HTTP Streaming via DASH)



Reference: https://bitmovin.com/dynamic-adaptive-streaming-http-mpeg-dash/

1. Streaming Stored Video

(Category 3: Adaptive HTTP Streaming via DASH)

- *Features:*
  - Reduces startup delays and buffering stalling
  - By-passes NATs and Firewalls by using HTTP
- *"Intelligence"* at client: client determines

–*when* to request chunk (so that buffer starvation, or overflow does not occur)

–*what encoding rate* to request (higher quality when more bandwidth available)

–*where* to request chunk (can request from URL server that is "close" to client or has high available bandwidth)

- A byproduct: improved server-side scalability

Content Distribution Network

School of Informatics
Institute for Computing Systems Architecture

CDN

(Content Distribution Challenge)

- Many Internet video comapnies are distributing on-demand multi-Mbps streams to millions of users

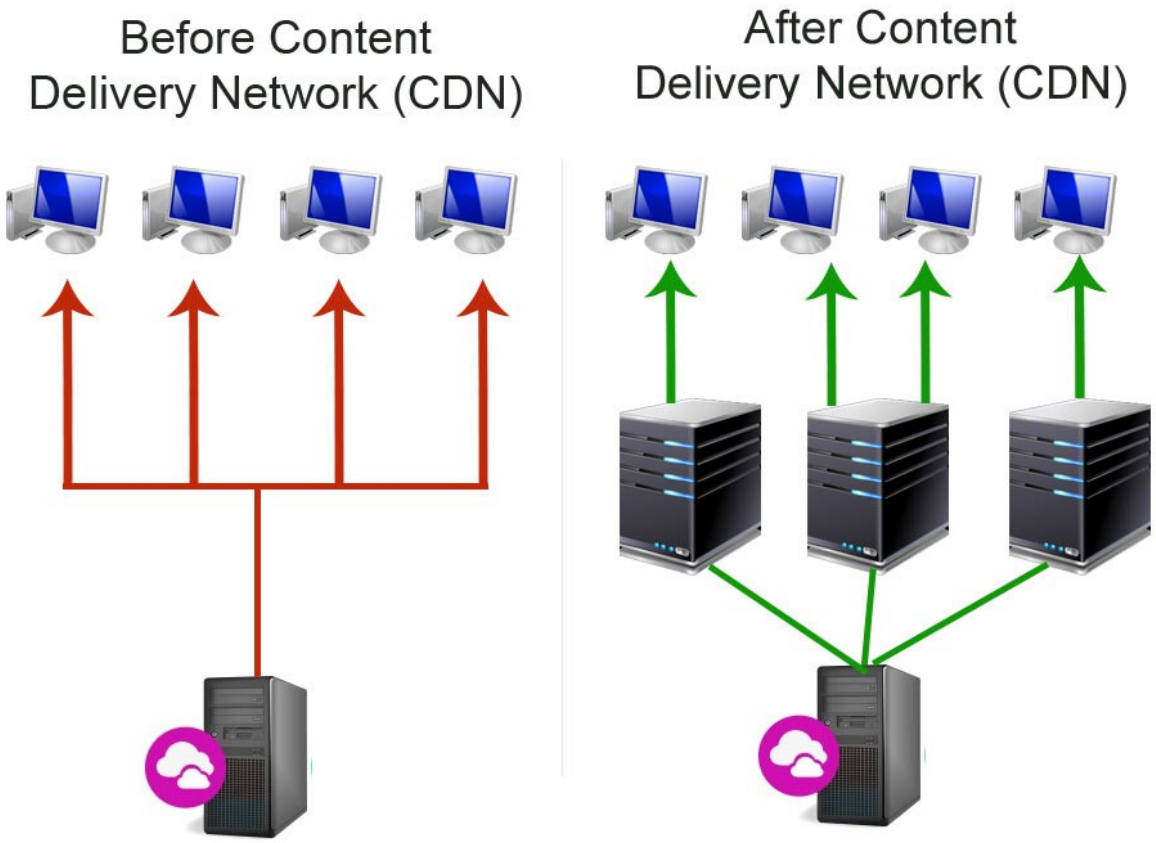
- Challenge: how to stream content (selected from millions of videos) to hundreds of thousands of simultaneous users while providing continous playout and high interactivity?

- Option 1: single, large “mega-server/ massive data center”
  - –single point of failure
  - –point of network congestion
  - –long path to distant clients : if client is far from data center

School of Informatics
Institute for Computing Systems Architecture

CDN

(Without and With CDN)



CDN

(A better approach)

- A CDN consists of multiple server clusters distributed geographically around the world
  - Stores copies of videos
  - Web contents (including documents, images and audios).
  - Serves a user request for content from the “closest” CDN location with the requested content
- CDN can be
  - **Private CDN:** owned by the content provider itself (e.g. Google's CDN distributed YouTube videos and other types of content)
  - **Third-party CDN:** that distributed content on behalf of multiple content providers (Akamai's CDN distributes Netflix and Hulu content, among others)

CDN

(A better approach)

Two approaches for placement of CDN server clusters

1. Enter deep: Push CDN servers deep into many access networks of ISP, by deploying server clusters in access ISPs all over the world

- Close to users, improves user-perceived delay and throughput by decreasing number of links and routers (used by Akamai, 1700 locations)

- Task of maintaining and managing the clusters becomes challenging

2. Bring home: Builds large clusters at a smaller number (10's) of key locations and connecting these clusters using a private high speed network (used by Limelight).

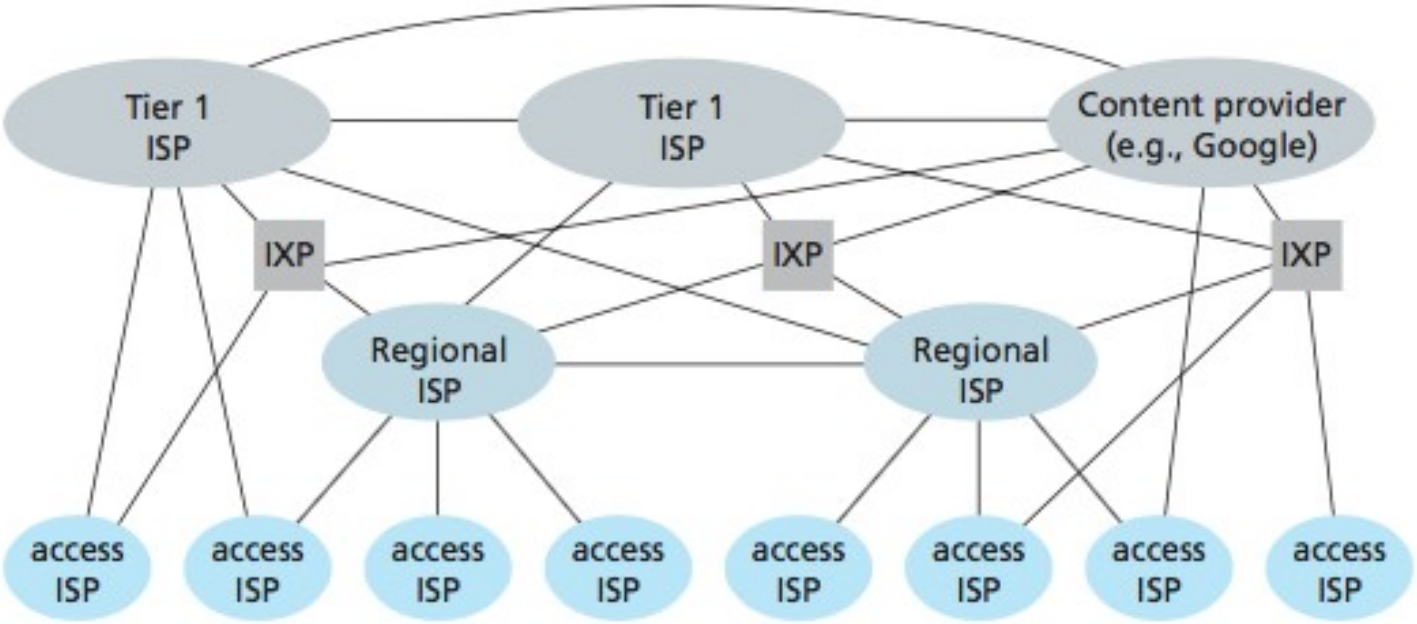
- Near Tier 1 ISPs.

- Content owners such as media companies and e-commerce vendors pay CDN operators to deliver their content to their end users. In

School of Informatics
Institute for Computing Systems Architecture

CDN

(CDNs deployed near Point of Presence)



Google has Bring Home CDNs deployed near Tier 1 ISPs

THE UNIVERSITY OF EDINBURGH

School of Informatics
Institute for Computing Systems Architecture

CDN

(CDN operation)

Suppose a content provider NetCinema, employs the 3rd-party CDN KingCDN to distribute videos to its customers

1. User visits the web-page at NetCinema and clicks the video link http://video.netcinema.com/6Y7B23V

2. User's host sends DNS query for video.netcinema.com

3a. The users LDNS relays DNS query to authoritative DNS for NetCinema.

3b. By seeing the string 'video', instead of returning an IP address, the DNS server returns KingCDN's domain e.g. a1105.kingcdn.com to LDNS

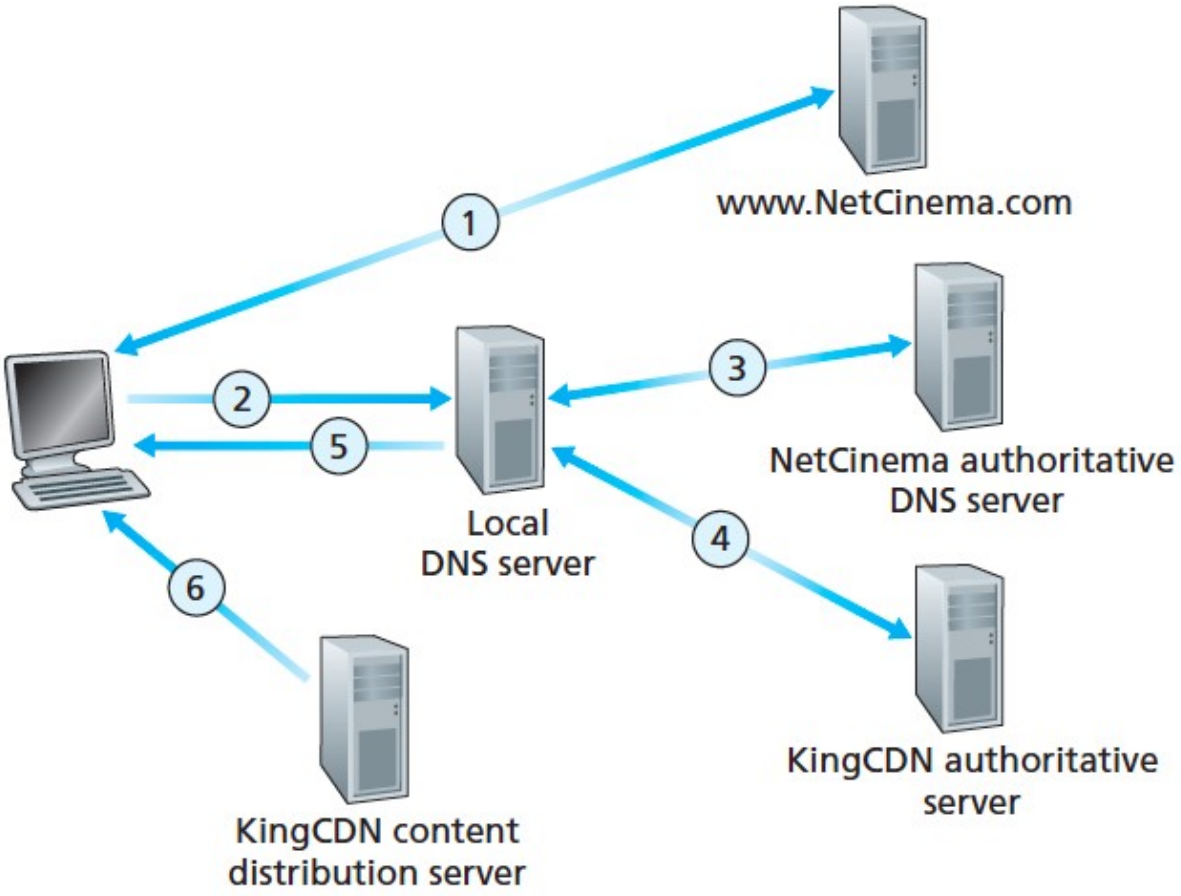
4a. The LDNS query enters into KingCDN's private DNS structure.

4b. KingCDN's DNS system eventually returns the IP addresses of the content server to LDNS.

5. The LDNS forwards IP address of the content server to client.

6. Which then established TCP connection with the server and issues HTTP GET request for the video (manifest file helps client choose appropriate version of the video).

CDN
(CDN operation)



DNS directs user request to CDN

THE UNIVERSITY OF EDINBURGH
School of Informatics
Institute for Computing Systems Architecture

CDN

(Selection Strategies)

Problem: how does CDN DNS select "good" CDN cluster to stream requested content to client?

Some possible **solution** strategies:

1). Pick the CDN cluster **geographically** closest to client's LDNS server, several drawbacks:

- Closest geographically does not mean closest in network terms (i.e path)
- May not be a good choice with remote local DNS servers
- Network dynamics not taken into account (ignores the variation in delay and available bandwidth over time)

2) Measurement based selection of best CDN cluster.

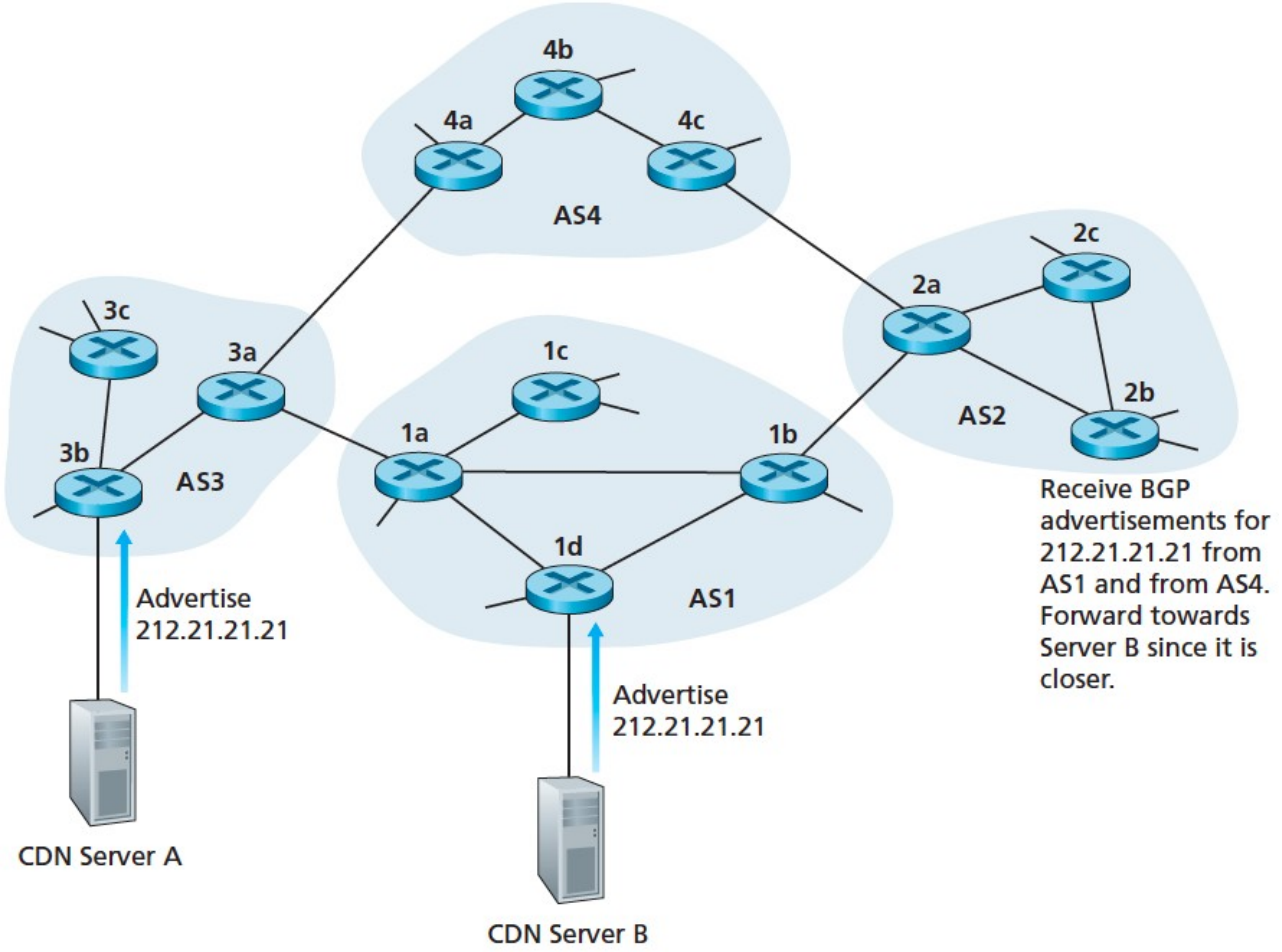
- Active measurement of delay and loss through dedicated probes.
- Passive monitoring of recent/ongoing traffic between clients and CDN servers (observing delay suffered by SYNACK and ACK messages during 3-way handshake)

School of Informatics
Institute for Computing Systems Architecture

CDN
(Selection Strategies)

- **3. IP anycast:** The routers in the Internet route the client's packet to the “closest” cluster, as determined by BGP.
  - In IP anycast, the CDN company assigns the same IP address to each of its cluster locations
  - When a BGP router receives multiple route advertisements for this same IP address, it treats these advertisements as providing different paths to the same physical location.
  - Following standard operating procedures, the BGP router will then pick the “best” (e.g. closes as determined by AS-hop counts) route to the IP address according to its route selection mechanism.
  - This approach has the advantage of finding the cluster that is closest to the client rather than the cluster that is closest to the client's LDNS.
  - It does not take into account the dynamic nature of the Internet over short time scales.

CDN Cluster Selection Strategies

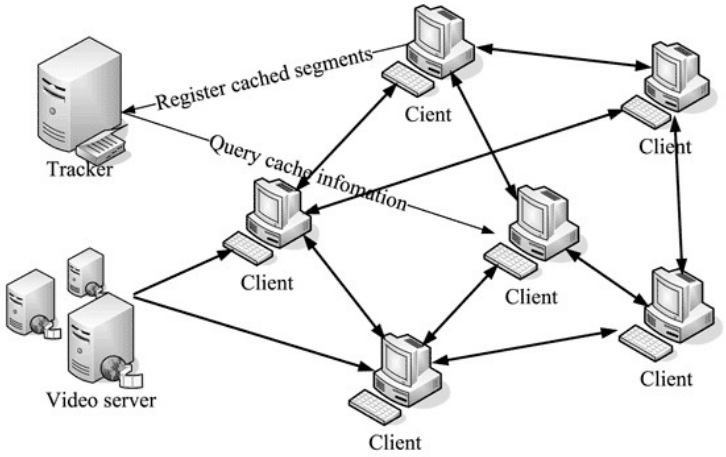


Using IP anycast to route clients to closest CDN cluster

CDN Cluster Selection Strategies

Other factors impacting cluster selection strategy

- Delay
- Loss
- Bandwidth performance
- Load on the cluster
- ISP delivery cost
- (i.e. contractual relationships between ISPs and the cluster operators



Peer-to-Peer (P2P) Assisted Streaming

P2P Video Streaming

Mostly used in Live Video Streaming: (e.g. P2PTV)

- Each user, while downloading a video stream, is simultaneously also uploading that stream to other users, thus contributing to the overall available bandwidth.

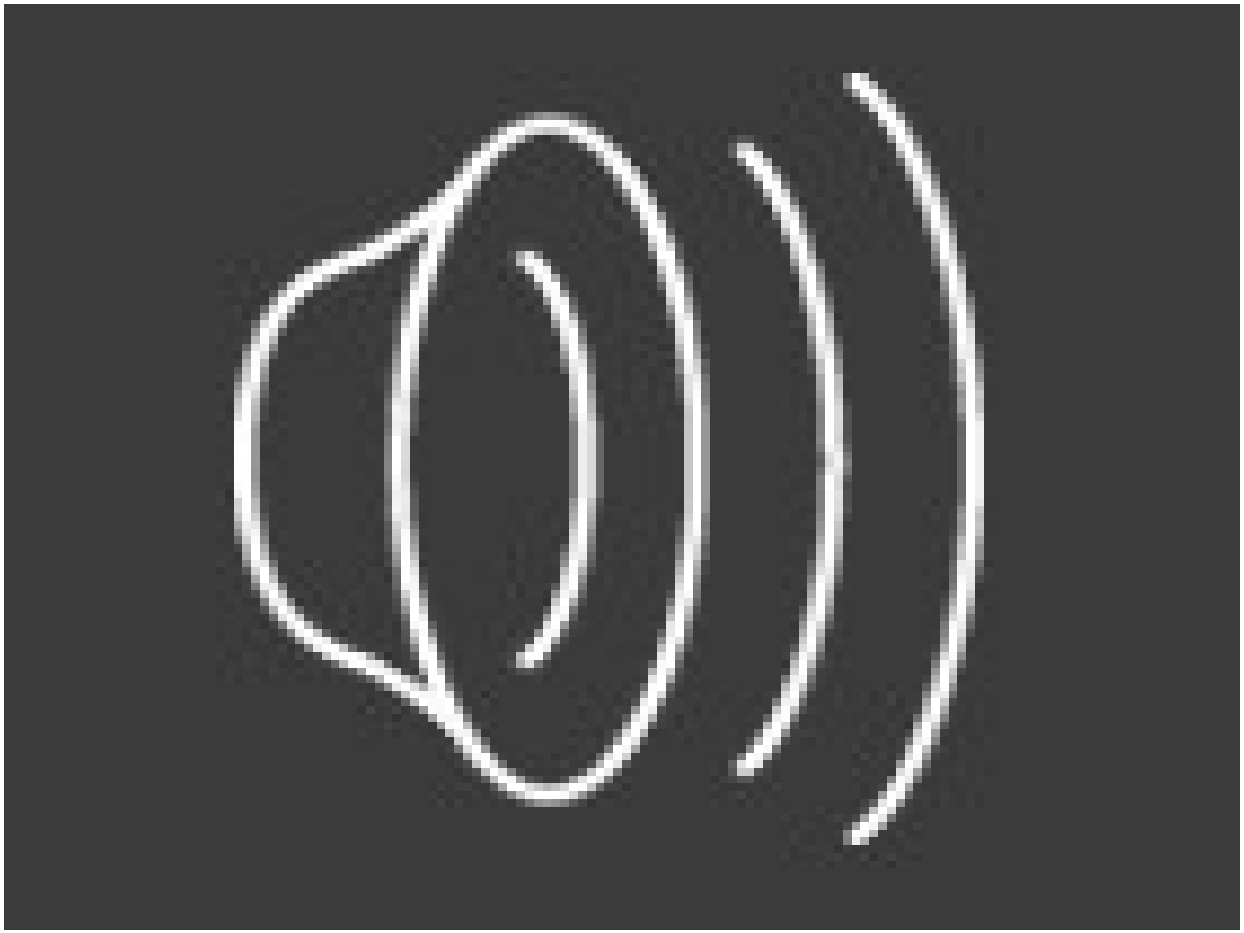
- The arriving streams are typically a few minutes time-delayed compared to the original sources.

- If a user wishes to view a certain channel, he is directed to the "tracker server" for that channel in order to obtain addresses of peers who distribute that channel; it then contacts these peers to receive the feed.

- The tracker records the user's address, so that it can be given to other users who wish to view the same channel.

- The need for a tracker can also be eliminated by the use of distributed hash table (DHT) technology, where the tracker is deployed at each peer node (out of scope of this study)

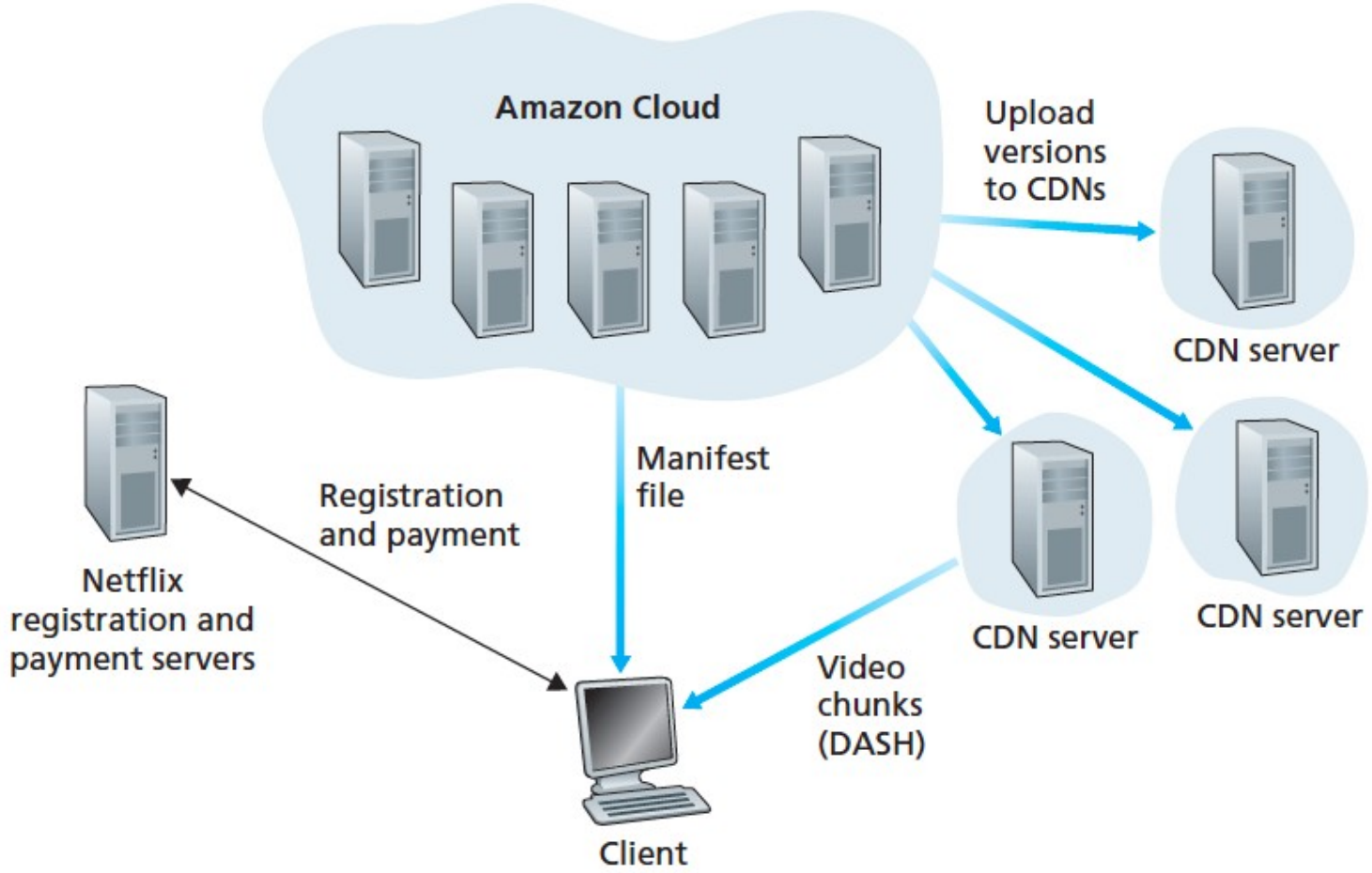
P2P Video Streaming



Case Study 1: Netflix

- Netflix
  - 30% downstream US traffic in 2011
  - Leading service provider for online movies and TV shows
  - Owns very little infrastructure, uses 3rd party services.
  - Rents servers, bandwidth, storage and database services from 3rd parties.
  - Employs both CDN and adaptive streaming over HTTP
  - Amazon cloud
- Four major components
  - **Registration and payment servers** (Own). Handles registration of new account and credit card payments.
  - **Content ingestion:** Before Netflix distributes a movie to its customers, it must first ingest and process the movie. Netflix uploads studio master to Amazon cloud
  - **Content processing:** Creates multiple versions of movie (different encodings), suitable for diverse client video players and data rates.
  - **Uploading versions to the CDNs:** Upload versions from cloud to CDNs
  - Multiple CDN providers and clients
  - **Three** 3rd party CDNs host/stream Netflix content: Akamai, Limelight, Level-3

Case Study 1: Netflix



Netflix video streaming platform

Case Study 2: YouTube

- Half a billion videos in its library and half a billion video views per day (A study conducted in 2011)
- World's largest video-sharing site
- Extensive use of **CDN** technology
- **Google uses private CDN** to distribute YouTube videos.
- Google has installed server clusters in different locations.
- Google uses DNS to redirect a customer request to a specific cluster.

Case Study 2: YouTube

- **Selection strategy:**
  - Mostly the cluster that results in lowest RTT to client.
  - Sometimes to balance load, a client is directed to more distant cluster
    - If a cluster does not have requested video instead of fetching it from somewhere else, **client is redirected** to another cluster.
- Previously YouTube used HTTP streaming and a small number of different versions for video, each with different bit-rate and quality.
- Recently YouTube has adopted adaptive HTTP streaming
- Once a video is received by a YouTube, Google creates its multiple versions at its data centers.

THE UNIVERSITY OF EDINBURGH
School of Informatics
Institute for Computing Systems Architecture

Case Study3: KanKan

- Netflix and YouTube both uses CDN and therefore have to bear the cost
- Kankan, **avoids CDN** by reducing its infrastructure and bandwidth costs
- Kankan, **is leading P2P-based video-on-demand provider** in China (has over 30 million unique users per month)
- When a peer wants to see a video it,
  - Contacts a tracker (centralized or peer-based DHT), to discover other peers in the system that have a copy of that video
  - The peer then requests chunks of the video file in parallel from the other peers that have the file.
  - Requests are preferentially made for the chunks to be viewed in near future (i.e. to have continuous playout).