VIDEO ON DEMAND

COEN 233-COMPUTER NETWORKS

AUTHOR

ANKIT ANAND

W1652456

AANAND4@SCU.EDU

UNDER THE GUIDANCE OF

DR. KEYVAN MOATAGHED

AUDIENCE

This study paper defines the notion of video on demand, which is one of the most recent concepts used globally to watch internet content for various purposes such as entertainment, educational, marketing or distribution, personal development, etc. This research project on Video on Demand (VoD) explores the significance, topology, architecture, protocols and workflow, content delivery, scalability, and future trends in VoD systems. It also covers topics such as content management system, monetization models, integration with live streaming and social media. Through case study the research provides insight of Netflix VoD technology. The project aims to benefit students, researchers, and instructors by offering a comprehensive resource in the field of VoD.

Depending on the focus and scope of the research, primary audience consists of students, fellow researchers, and instructors within the academic community. To a student seeking foundational knowledge, a fellow researcher looking for comparative analysis and further research opportunities, or an instructor in need of teaching material, this research project provides a comprehensive resource that can be adapted to suit different knowledge levels. It aims to foster learning, inspire innovation, and contribute to the academic discourse surrounding VoD technologies. Here is how each group can benefit from this research:

- **Students:** This research project serves as a valuable resource for students studying computer science, multimedia systems, network engineering, or related disciplines. It provides a comprehensive overview of VoD technologies, including architectures, protocols, content delivery strategies, scalability approaches, and future trends. Students can use this information for their coursework, projects, or to gain a deeper understanding of the concepts and principles underlying VoD systems.
- **Fellow Researchers:** Researchers in the fields of computer science, multimedia systems, and network engineering can find this research project as a reference for their own studies. The detailed exploration of VoD architectures, protocols, content delivery, scalability, and other topics offer insights that can inspire further research or validate existing findings. By building upon this research, fellow researchers can contribute to the knowledge base and drive advancements in VoD technologies.
- **Instructors:** Instructors teaching courses related to multimedia systems, networking, or digital media will find this research project beneficial as a teaching aid. The organized outline and in-depth analysis provide instructors with a structured framework to cover various aspects of VoD in their lectures or as supplemental material for student assignments. Additionally, instructors can use the comparative analysis and case studies as real-world examples to illustrate key concepts and engage students in discussions.

By targeting students, fellow researchers, and instructors, the research project aims to facilitate learning, knowledge dissemination, and academic discourse surrounding the field of Video on Demand. Through this research, the focus is to contribute to the academic community and inspire future generations of students and researchers to explore and innovate VoD technologies.

However, this research paper should not act as sole source of information on VoD for any target audience group. It should be complemented with the group's own research, knowledge, and references. Also, this paper might sound accurate and comprehensive to some, but it cannot be same for another group or person.

TABLE OF CONTENT

TABLE OF FIGURES

1. INTRODUCTION AND OUTLINE

Welcome to this research paper on Video on Demand (VoD).

The emergence of Video on Demand (VoD) has transformed the media consumption landscape, delivering convenience and flexibility in accessing a diverse array of video content. As streaming platforms proliferate and on-demand services gain popularity, it has become imperative to comprehend the underlying technologies and considerations in VoD systems. The primary aim of this research project is to comprehensively explore multiple facets of VoD, encompassing architecture, protocols, content delivery, scalability, challenges, and future trends.

To achieve a thorough understanding, the research scrutinizes various architectural approaches such as client-server, peer-to-peer, and content delivery networks, evaluating their scalability, performance, etc. In addition, a detailed examination of streaming protocols, video codecs, adaptive bitrate streaming techniques, and content delivery strategies, illuminating the technologies that enable seamless video playback and efficient bandwidth utilization is done.

Another crucial area of focus is content management, encompassing acquisition, licensing, metadata extraction, and recommendation algorithms. Monetization models, live streaming integration, social media integration, market analysis and QoS is also explored to provide a comprehensive panorama of the VoD ecosystem.

By means of a case study this research project aims to furnish invaluable insights and recommendations for enhancing VoD systems. By identifying challenges and surveying future trends, the research paper tries to contribute to the ongoing development of VoD technologies, thereby continuously enhancing user experiences.

Students, researchers, and instructors seeking a deeper comprehension of VoD systems will benefit from this research project. It serves as a valuable repository, elucidating the technologies, architectures, protocols, content management strategies, and user experience enhancements within the ever-growing Video on Demand ecosystem.

The rest of the research paper is organized as follows:

2. Introduction covers the history and overview of Video on Demand (VoD), its significance in the media industry and objectives and scope of the research project.

3. VoD Architecture and topology covers the VoD architectures such as Client-server architecture, Peer-to-peer (P2P) architecture, Content delivery networks (CDNs) and topologies employed such as centralized topology, distributed topology, and hybrid topology.

4. VoD protocols and workflow covers various streaming protocols along with their pros and cons, video codecs and compression techniques, adaptive Bitrate Streaming (ABR) technique and its advantages.

5. Content Delivery in VoD covers content ingestion and transcoding processes, storage and caching mechanisms, content distribution strategies (e.g., edge caching, multi-CDN).

6. Scalability in VoD Systems covers different scaling types and strategies, SVC and its role in scalability.

7. VoD CMS covers all the aspects of content management in terms of a VoD system and dives into recommendation algorithms.

8. VoD monetization models discusses the types of monetization strategies followed by industry giants and provides an insight of it.

9. VoD and other platforms discusses other platforms like OTT and live streaming in brief and covers integration of VoD with live streaming and social media.

10. VoD market analysis covers market trends, dynamics and challenges to the same. It also provides insight on key market players of VoD.

11. QoS discusses advantages of QoS and use cases in VoD industry.

12. Case study – Netflix gives a detail on system design of the industry giant.

13. Conclusion and future aspects covers the summary of research project and future trends of VoD system including probable integration of AI, ML, VR, AR.

2. Introduction to Video on Demand

The era of scheduling lives around television programs or repeatedly renting DVDs for movie marathons is long gone. The advent of Video on Demand (VoD) and the widespread accessibility of high-speed broadband internet have revolutionized the way people enjoy media and entertainment content.

According to **Conviva**'s report, Video on Demand (VoD) is experiencing significant growth in the broadcasting industry, with a year-by-year increase of 155% in viewership. On average, users are spending 17.1 minutes per session consuming VOD content.

A survey conducted by **Hubspot** indicates that people show a preference for video content over other forms, such as infographics, blog posts, and emails.

According to **Cisco Systems'** projections, by 2023, approximately 70% of the population is expected to possess mobile devices. As a result, the desire for 4K/UHD video content is anticipated to increase twofold.

What is video on demand? How can we define VoD? What is history of VoD?

These questions will be answered in this chapter.

2.1 What is VoD?

So, again arrives a question, *what is VoD?*

According to Wikipedia, Video on demand (VOD) is a media distribution system that allows users to access to videos, television shows and films without a traditional video playback device and a typical static broadcasting schedule.

Britannica defines Video on Demand as, "a technology for delivering video content, such as movies and television shows, directly to individual customers for immediate viewing, regardless of broadcast schedules."

Techopedia says, Video on demand (VoD) is a system that allows users to select and watch video content of their choice on their TVs or computers. Video on demand is one of the dynamic features offered by Internet Protocol TV. VoD provides users with a menu of available videos from which to choose. The video data is transmitted via Real-Time Streaming Protocol.

So, fundamental definition of VoD is a technology that fulfils user ability to watch video of his/her choice at user's time of convenience in any part of the world. Video on Demand refers to a technology and service that allows users to access and stream video content instantly and on-demand. It offers a convenient and flexible way of consuming movies, TV shows, documentaries, and other forms of video entertainment.

In simple terms, Video on Demand (VoD) streaming enables individuals to watch videos instantly from a collection of content. They can access these videos on their TV, computer, mobile phone, or any other device. VoD services are typically provided by cable, telephone, or broadband providers, giving viewers the ability to control the playback of videos, including options to pause, play, rewind, or fast forward, according to their preferences.

2.2 Reasons for shift towards VoD

The VoD has been on a growing spree as the favourite method for consuming entertainment and media. There are several advantages of VoD which are the major factors for the seen growth.

1. Convenient and Flexible Access: VoD offers users the convenience of accessing a vast library of content on their own terms, without being constrained by scheduled programming. It provides the flexibility to watch videos whenever desired, aligning with the busy lifestyles of modern audiences.

2. Personalized viewing experience: VoD platforms provide a tailored and individualized viewing experience by leveraging advanced algorithms and data analytics. Through personalized recommendations based on user preferences and viewing history, viewers are exposed to relevant and engaging content that enhances their overall satisfaction.

3. Extensive content selection: VoD services present a diverse range of content choices, encompassing movies, TV series, documentaries, and original programming. These platforms cater to a wide array of interests, genres, and niche preferences, ensuring there is an abundance of options to suit every viewer's specific tastes.

4. Global accessibility: VoD transcends geographical limitations, enabling viewers worldwide to access content from different regions. Leading platforms expand their services globally, providing a rich and diverse selection of international content. This fosters cultural exchange and promotes a global viewing community.

5. Technological advancements: Technological progress and improved internet connectivity have significantly contributed to the rise of VoD. High-speed broadband and advanced streaming capabilities allow users to seamlessly enjoy high-quality videos. Adaptive streaming techniques and video compression technologies optimize the viewing experience by reducing buffering and latency issues.

6. Mobility and multi-device compatibility: The proliferation of mobile devices empowers users to access VoD content on smartphones, tablets, laptops, or smart TVs. This mobility enables viewers to watch videos on the go or from various locations, providing the flexibility to consume content wherever and whenever they prefer.

7. Cost-effectiveness and value for money: VoD services often offer cost-effective subscription plans, allowing users to access a vast content library at a reasonable price. This eliminates the need for traditional cable or satellite subscriptions, which may include unwanted channels or content. Users can pay for the specific content they want, ensuring value for their investment.

8. Original and exclusive content: VoD platforms invest in the production of high-quality original programming, creating exclusive content that sets them apart. Viewers can access critically acclaimed TV series, movies, and documentaries that rival traditional studio productions, enhancing the overall value and appeal of VoD services.

2.3 History of VoD

Video on Demand (VoD) emerged as a possibility around 1995, challenging the notion that streaming video or television shows with uncompromised quality over carrier-provided bandwidth on copper cables was impossible. Broadcasting a television program requires a minimum bandwidth of 200 Mbps, significantly higher than the data transmission capacity of a speech signal over a copper phone wire. However, advancements in video compression techniques such as Discrete Cosine Transform (DCT) and Asymmetric Digital Subscriber Line (ADSL) made VoD feasible.

DCT, a lossy compression technique proposed in 1972, was utilized in video coding guidelines like motion-compensated DCT calculations for H.26x designs after 1988 and MPEG groups after 1991. ADSL increased the data transfer capacity of a phone line from around 100 kbps to 2 Mbps. The combination of DCT and ADSL technologies reduced the required transmission capacity for a TV signal from around 200 Mbps to 2 Mbps. This breakthrough in DCT and ADSL paved the way for the practical implementation of VoD services in the 1990s.

In Japan, a service program for an 'Integrated Network System' was introduced around 1986. However, it was not practically viable until the development of DCT and ADSL technologies in the mid-1990s. Initial VoD systems utilized tapes as the continuous source of video streams.

In November 1992, Bell Atlantic announced its plans to conduct trials for VoD. IBM developed a video server named Tiger Shark, which was chosen by Bell Atlantic. In April 1993, the system became the first VoD over ADSL to be deployed in the lab, serving 50 video streams. Simultaneously, Digital Equipment Corporation was developing a scalable video server capable of handling a variable range of video streams.

Starting in September 1994, a VoD service became a significant part of the Cambridge Digital Interactive Television Trial in England. MPEG-1 encoded video was streamed over an ATM network from an ICL media server to set-top boxes designed by Acorn Online Media, reaching 250 homes and several schools connected to the Cambridge Cable system. The trials began with a speed of 2 Mbit/s to the home and gradually increased to 25 Mbit/s. Despite being a notable achievement, sourcing content posed a significant challenge, leading to the closure of the trial in 1996.

Around 1998, Kingston emerged as the leading organization in the UK to develop and offer commercial Video on Demand (VoD) services. They pioneered the integration of TV and internet access through a set-top box utilizing IP delivery over ADSL. In 2005, TV providers like NTL and Telewest launched their own versions of VoD services in the UK, competing against the traditional TV giant, BSkyB. In response, BSkyB introduced their broadband service called skyBy broadband, later renamed as Sky Anytime on PC, which was deployed on January 2, 2006.

Sky Anytime on PC utilized a peer-to-peer approach, utilizing Kontiki technology, to facilitate high-capacity multi-point downloads of video content. Rather than relying solely on central Sky servers, the content was distributed from other users who had downloaded the same content onto their systems.

VoD services are now available in various parts of the United States, including small towns, with the US having the highest global adoption rate of VoD. In 2010, 80% of American internet users had watched videos online, with 42% of mobile phone users preferring video download apps over web browsers. Due to the extensive coverage of satellite TV broadcasters, it is impractical for most of them to offer streaming services. However, networks like EchoStar or Dish Network provided VoD setups for their satellite network subscribers who owned personal video recorders (PVRs). Once programs were downloaded onto a user's PVR, they could be watched, played, paused, and searched at the viewer's convenience.

VoD is also common in higher-end hotels, where systems are in place to store and deliver content downloaded directly from the internet. In Europe, there were 142 paying VoD services operational by the end of 2006, which increased to 650 by 2009, according to the European Audiovisual Observatory.

During the Consumer Electronics Show in Las Vegas, the CEO and president of Sezmi demonstrated a new set-top box with a one-terabyte hard drive, which could be used for video-on-demand services offered by satellite TV and other broadband providers.

The creation of VoD involves extensive negotiations to establish a financial model that serves the interests of both content creators and cable providers, while offering attractive content to viewers at an acceptable price point. The economic feasibility of a VoD model depends on factors such as video or movie purchase rates and the revenue split between Hollywood and cable operators. Sending a movie over a communication signal instead of a cable or fibre optic line eliminates the cost of laying extensive kilometers of physical infrastructure.

3. VoD Architecture and topology

Video-on-Demand (VoD) is an essential technology for various multimedia applications like movie-on-demand, distance learning, digital libraries, videoconferences, and interactive TV. Typically, a VoD system operates on a client/server model where multimedia content is stored on a central server and distributed to clients across a network. However, the success of on-demand services as a profitable business relies on the system's ability to handle real-time constraints and resource requirements as it grows. If the system's catalogue is too small, it will not be appealing, but if the customer base grows too large and the infrastructure cannot support it, the business plan fails. Thus, providing multimedia content from a central server to many distributed clients is technically challenging due to the significant bandwidth demands of high-quality media streams and is economically inefficient.

So, to tackle the above challenges, a robust architecture, and a reliable network topology is required. Though, in first glance, both architecture and topology might sound similar but there are differences in the two terminologies used here. VoD topology refers specifically to the network infrastructure and connectivity, while VoD architecture encompasses the broader system design and components involved in delivering VoD services. The topology defines the network structure, while the architecture encompasses the network topology and the various components that make up the VoD system.

3.1 VoD Topology

The choice of network topology in a Video on Demand (VoD) system is flexible and depends on the underlying network infrastructure and the preferred architecture. VoD systems can be implemented using different network topologies based on factors such as scalability, geographic distribution, and available resources. Here are three most common topologies used in VoD systems:

3.1.1 Centralized topology

In this type of topology, the video content is stored and managed in a centralized server or data centre. The server acts as the central point of control for delivering the content to the viewers. The centralized VoD topology provides a centralized framework for controlling and managing the VoD system. It streamlines content organization and metadata management, facilitating the smooth delivery of video content to viewers. However, as the number of viewers and content library grows, scalability can become a concern since the central server handles all requests, potentially leading to performance limitations. The below figure, fig 3.1, illustrates the discussed topology.

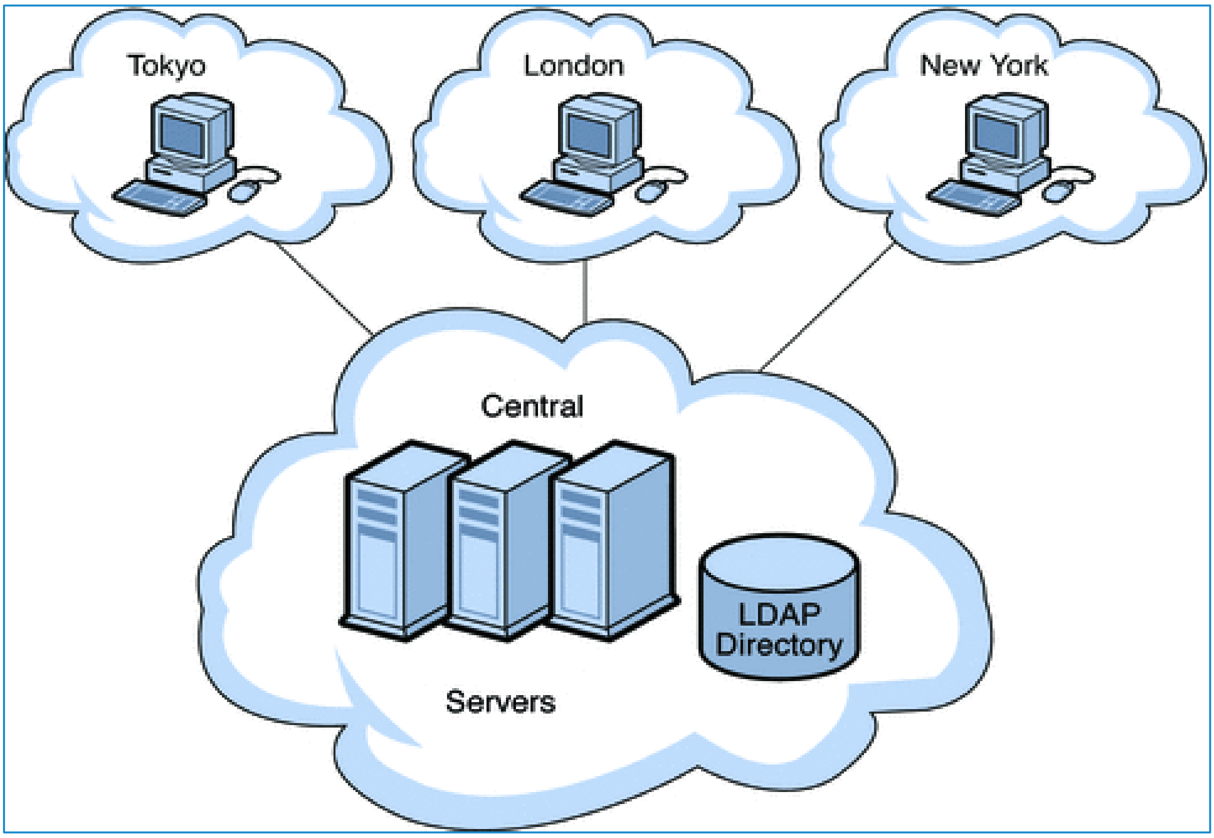


Fig 3.1: Centralized topology

3.1.2 Distributed topology

In a distributed VoD topology, the video content is stored and replicated across multiple servers or nodes that are strategically placed in various locations. When a viewer requests a specific video, the request is routed to the nearest or most appropriate server that houses the requested content. This distributed setup enables faster content delivery, reduces network congestion, and enhances the overall viewing experience. This topology also offers scalability as new servers can be added to the network to handle increased demand or accommodate expanding content libraries and handles fault tolerance as fault at one server location does not affect the other servers. Figure fig 3.2 shows a distributed topology.

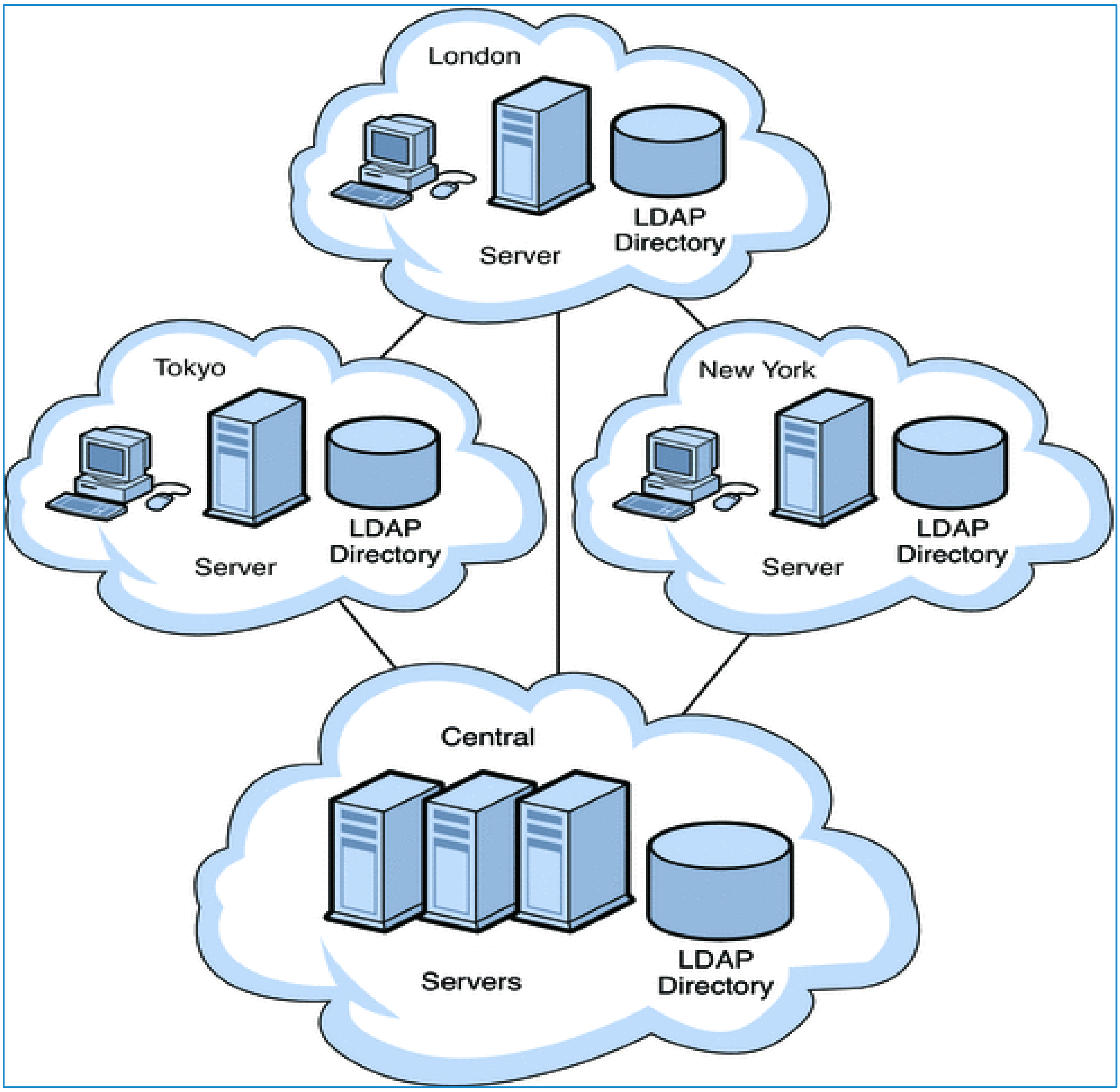


Fig 3.2: Distributed topology

3.1.3 Hybrid topology

In a hybrid VoD topology, a combination of network topologies is employed to optimize content delivery and system performance. It may combine elements of a centralized, distributed or any other topology depending on the specific needs of the VoD system. The primary objective of a hybrid topology is to capitalize on the advantages offered by different architectures, creating a VoD system that is adaptable and scalable. By integrating multiple topologies, the hybrid approach enhances the availability of content, accelerates delivery speeds, optimizes the utilization of resources, and guarantees a smooth viewing experience for users.

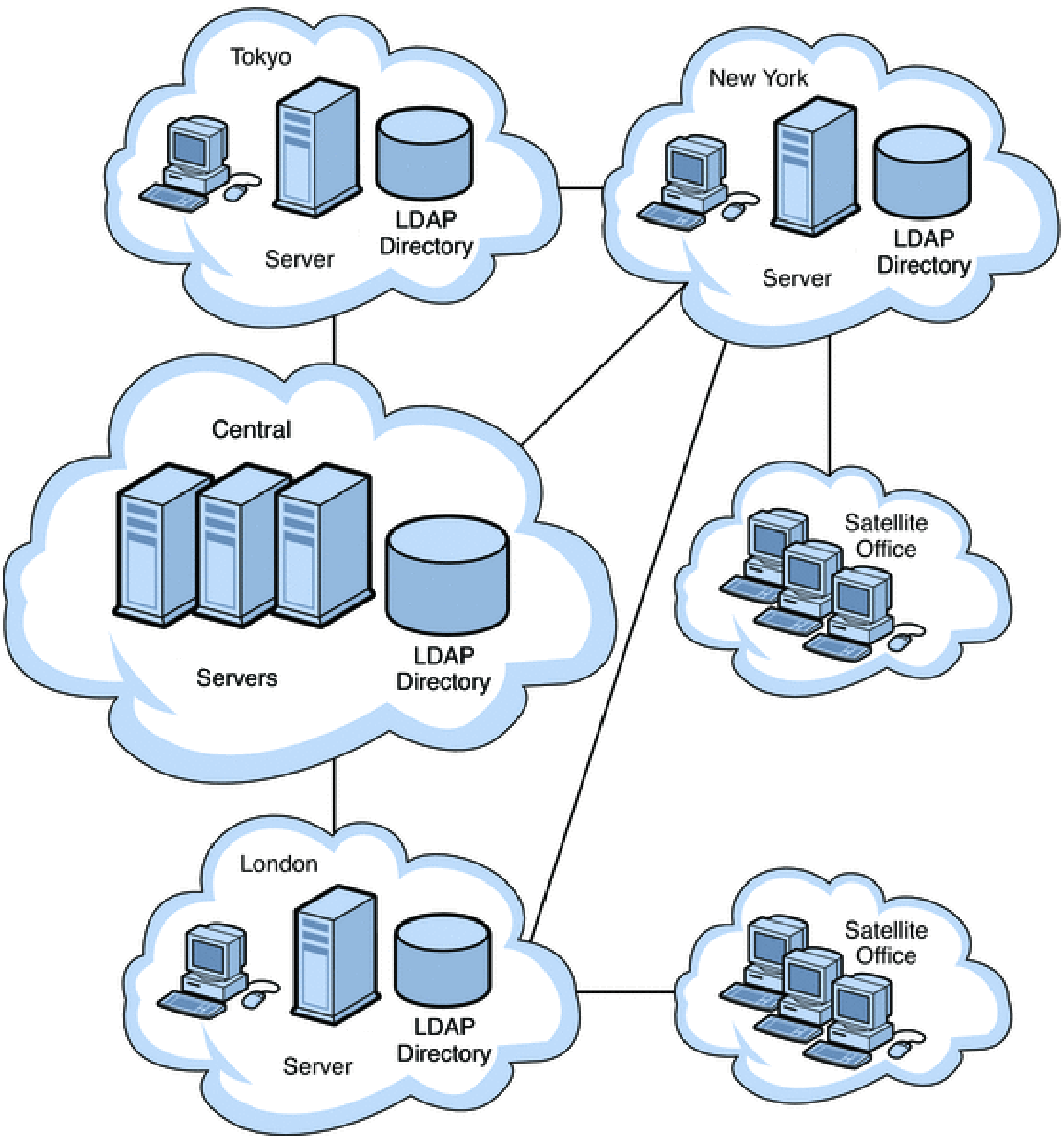


Fig 3.3: Hybrid topology

3.2 VoD architecture

3.2.1 Client-server architecture

The client-server architecture is a prevalent design choice for Video on Demand (VoD) systems. In this architecture, video content is stored and managed on a central server as shown in below picture. When a user requests a specific video, the client device sends a request to the server, which processes the request, retrieves the video, and streams it back to the client for viewing. The central server plays a vital role in managing the video library, user authentication, content access rights, and ensuring a smooth viewing experience. While the client-server architecture offers centralized control and streamlined content management, scalability challenges may arise as the number of clients and the video library size increase. Techniques such as load balancing and server clusters can be employed to address scalability concerns. The client-server VoD architecture provides a structured and manageable framework for delivering on-demand video content to viewers, making it a widely adopted solution in the industry.

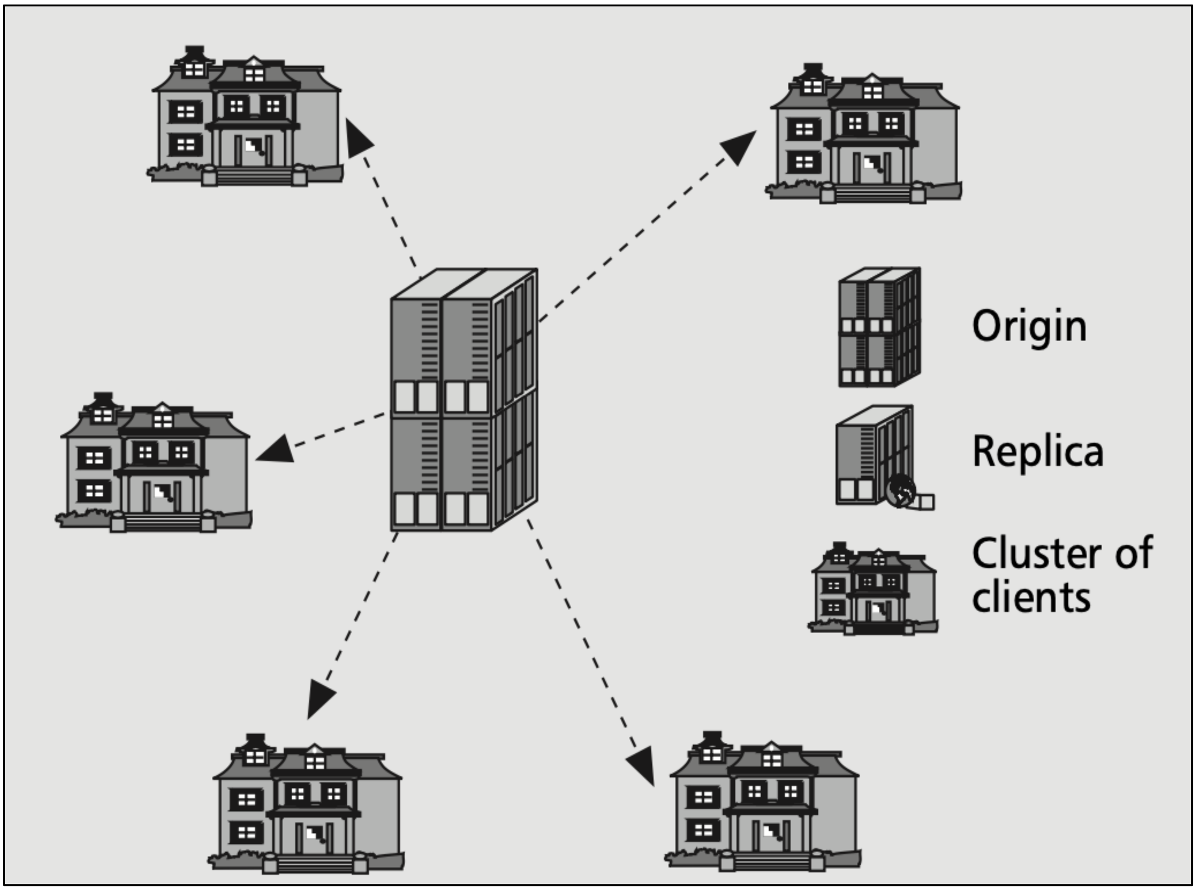


Fig 3.4: Client-server architecture

3.2.2 Peer-to-peer (P2P) architecture

The peer-to-peer (P2P) architecture is a notable approach for Video on Demand (VoD) systems. In this architecture, as shown below, video content is distributed across multiple user devices, forming a network of peers. When a user requests a video, it is obtained from other peers who have already downloaded or partially downloaded the same content. The P2P architecture offers advantages such as decentralized content distribution, load balancing, and improved scalability. However, challenges related to content availability, content quality, and network congestion must be addressed. P2P VoD architectures are an area of active research and continue to evolve to enhance content delivery efficiency and provide a seamless viewing experience for users.

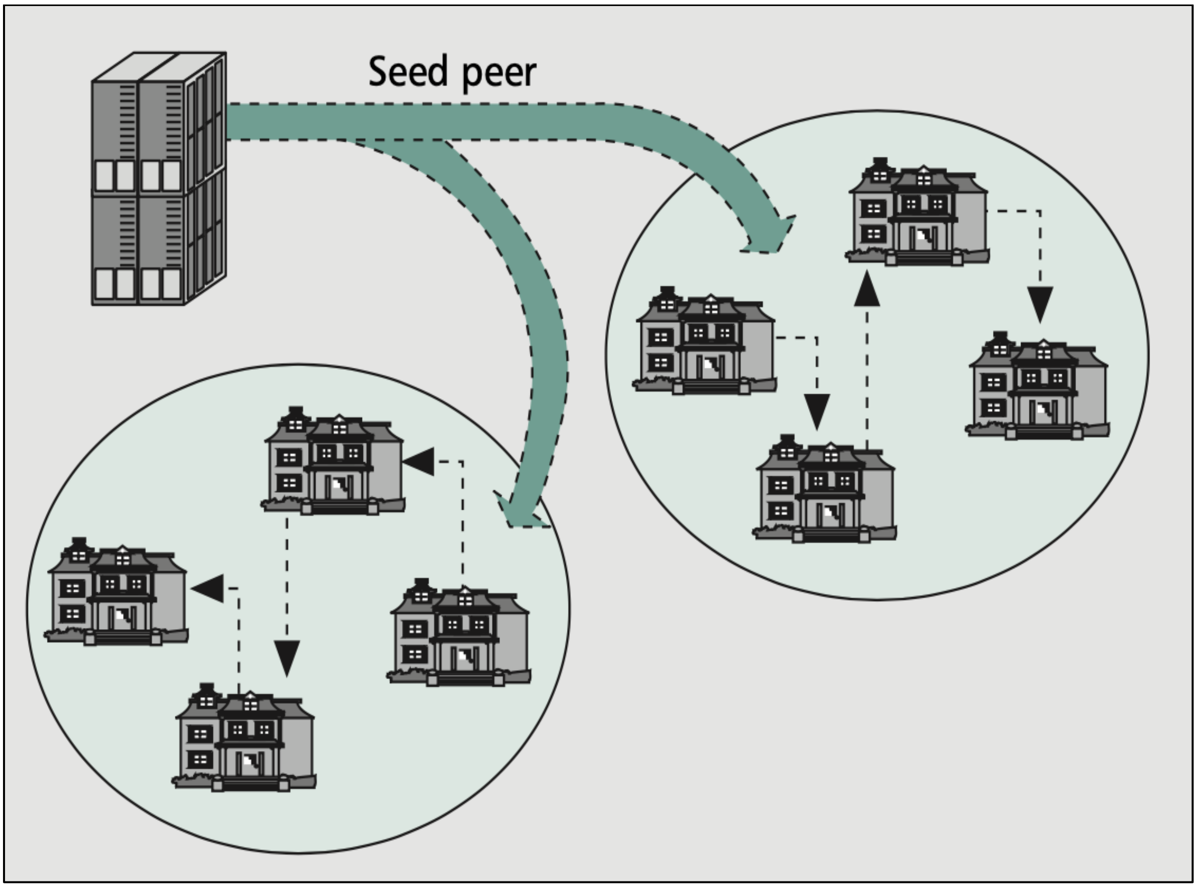


Fig 3.5: P2P architecture

3.2.3 Content delivery networks (CDNs)

CDN architecture in VoD systems leverages a distributed network of edge servers strategically placed in different geographical locations. Video content is uploaded to an origin server, which then replicates and distributes the content to multiple edge servers within the CDN network as illustrated in below figure. When a user requests a specific video, the request is directed to the nearest edge server, reducing latency, and optimizing content delivery. The edge server retrieves the requested video from its local cache and streams it to the user's device for seamless viewing. CDNs employ caching techniques to store popular or frequently accessed videos on edge servers. This optimizes content retrieval, reduces the load on origin servers, and enhances overall system efficiency. CDN architecture in VoD systems enhances user experience by reducing buffering time, improving streaming quality, and ensuring a smooth playback experience. The proximity of edge servers to users minimizes latency and delivers video content with minimal interruptions.

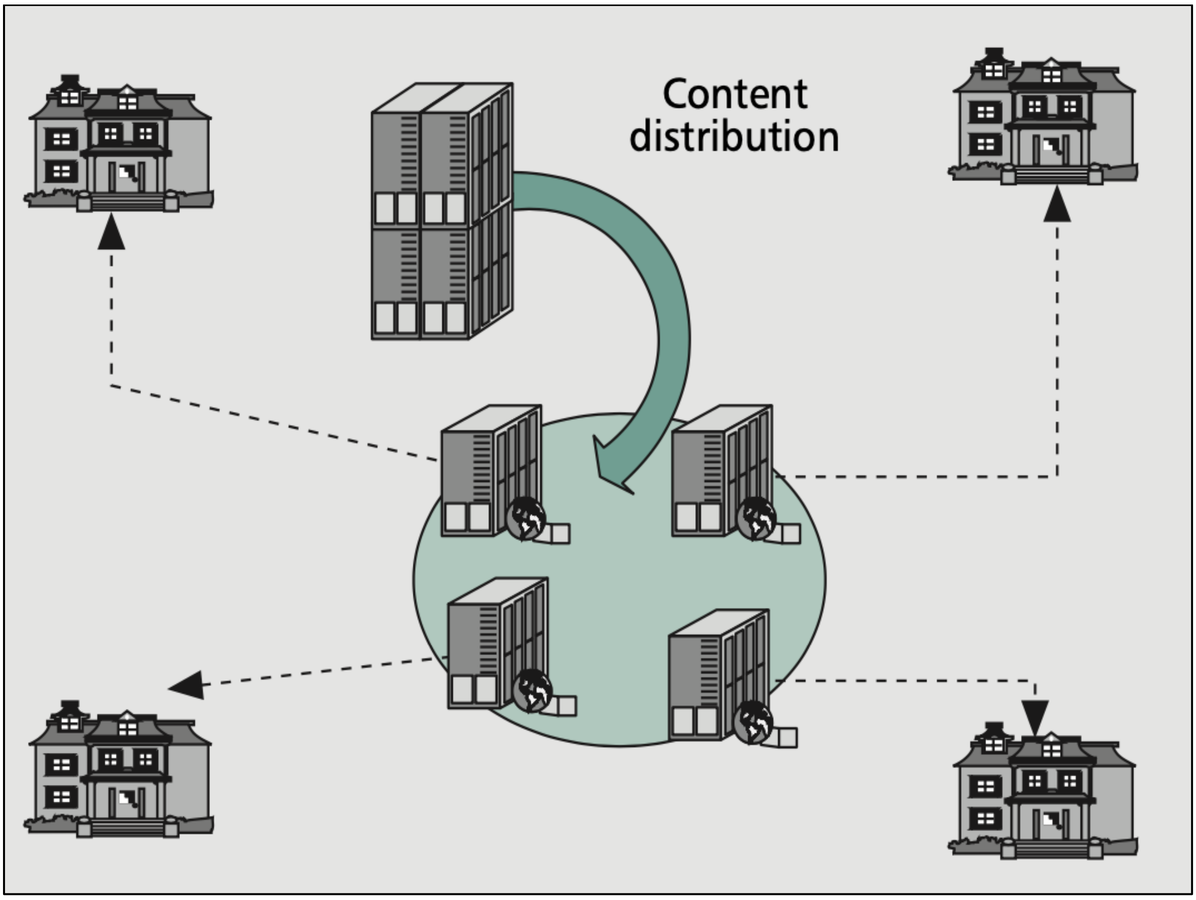


Fig 3.6: CDN architecture

4. VoD protocols and workflow

A protocol refers to a collection of guidelines that dictate the transmission of data between communicating systems. These protocols are structured in a layered manner to create a protocol stack, enabling each layer to perform specific tasks and collaborate effectively. The foundation is established by the lowest layer, with subsequent layers adding increasing levels of complexity as they build upon one another. "VoD protocol" typically refers to the specific communication protocols used to facilitate the streaming and delivery of video content to end-users. These protocols enable the transmission, control, and synchronization of video data over networks. Various protocols are employed in different stages of the VoD workflow, including content acquisition, encoding, storage, and playback.

4.1 Streaming protocols

A streaming protocol is a collection of guidelines that specify how data is transmitted between devices or systems over the Internet. These protocols establish the rules and procedures for effective communication, ensuring seamless data transfer from one device or system to another. There are many protocols available, and every streaming protocol has a distinct purpose and offers specific advantages, such as minimal latency or excellent adaptability. In this research paper 9 protocols are discussed which can be fall into one of the below three more generic categories:

Legacy protocols utilize Basic Authentication, which involves the use of usernames and passwords, to establish connections with email clients, calendars, and web services.

HTTP-Based protocols, functioning as request-response protocols, enable users to interact with web resources such as HTML files by transmitting hypertext messages between clients and servers.

Modern protocols, often open-source and not yet universally adopted, represent state-of-the-art technologies that address the limitations of previous video streaming protocols.

The most used streaming protocols are:

- **HTTP live streaming (HLS):** HLS has emerged as the prevailing streaming protocol in today's landscape. Initially introduced by Apple to replace Flash on iPhones, HLS enjoys widespread compatibility with a wide range of devices including desktop browsers, smart TVs, set-top boxes, Android, and iOS mobile devices, and HTML5 video players. This broad compatibility enables content creators to reach a larger audience with their streams. HLS leverages Adaptive Bitrate (ABR) technology, ensuring optimal video quality for viewers.
    - Type: HTTP based
    - Pros: High compatibility, secure streaming, and ultra-high quality video streams.
    - Cons: High latency and poor ingest.
- **Dynamic adaptive streaming over HTTP (MPEG-DASH):** MPEG-DASH, created by the Moving Pictures Expert Group (MPEG), emerged as an alternative to the HLS standard. Being an open-source standard, MPEG-DASH offers extensive customization options for various audio and video codecs. Like HLS, MPEG-DASH incorporates adaptive-bitrate streaming capabilities, enabling viewers to receive the optimal video quality based on their network's capabilities.
    - Type: HTTP based
    - Pros: High adaptability and allows user more customization option.
    - Cons: Limited compatibility and on verge of getting obsolete.
- **WebRTC:** WebRTC, an open-source initiative, enables the real-time delivery of video streams to viewers with minimal latency. Originally designed for text-based chat applications and Voice over IP (VoIP) services, its acquisition by Google has propelled its adoption by video chat and conference app developers. The WebRTC protocol offers a low-latency streaming solution that leverages peer-to-peer streaming (P2P) technology. This protocol is utilized by popular applications such as Google Meet, Discord, Houseparty, WhatsApp, and Facebook Messenger.
    - Type: Modern type.
    - Pros: Flexibility and real time latency.
    - Cons: Limited support.
- **Secure reliable transport (SRT):** The SRT protocol, developed by Haivision, is an open-source standard that has gained recognition for its security, reliability, compatibility, and low-latency streaming capabilities. It has become the preferred choice for organizations associated with the SRT Alliance. Unlike other protocols, SRT is codec-agnostic, enabling developers to use it in conjunction with any desired audio and video codecs. This flexibility empowers developers to customize their streaming solutions according to their specific requirements.
    - Type: Modern type.
    - Pros: High security, low latency, and high compatibility.

    - Cons: Limited support.
- **Real-time messaging protocol (RTMP):** RTMP, a protocol originally created by Adobe, was designed to facilitate the transfer of audio and video files between a streaming server and the Adobe Flash Player. However, as Flash has been gradually phased out, the primary use of RTMP has shifted away from delivering content directly to viewers. Instead, it is now commonly employed for ingesting live streams using RTMP-enabled encoders. In this scenario, the video feed captured by the encoder is transmitted to the streaming platform via the RTMP protocol. Subsequently, the content is delivered to end users using the widely adopted HLS protocol.
    - Type: Legacy protocol.
    - Pros: High flexibility low latency, and high compatibility.
    - Cons: Limited support and low bandwidth.
- **Real-time streaming protocol (RTSP):** RTSP, initially created for entertainment purposes, is a legacy protocol primarily utilized for establishing and controlling media sessions, such as those involving TV shows and movies, between endpoints. It shares similarities with HLS but cannot independently transmit live streaming data. Instead, it relies on RTSP servers working in conjunction with RTP and other protocols to accomplish streaming tasks. While RTSP supports low-latency streaming, it is not widely compatible with most devices and browsers. As a result, it is commonly employed for delivering low-latency streaming to specific small audiences through dedicated servers, making it a standard choice for video surveillance and CCTV systems.
    - Type: Legacy protocol.
    - Pros: Segmented streaming and highly customizable.
    - Cons: Low popularity and HTTP incompatible.
- **Transmission control protocol (TCP):** TCP, being one of the original and fundamental networking protocols, plays a vital role in numerous essential internet applications including the World Wide Web (HTTP), email, and File Transfer Protocol (FTP). Its primary focus is on ensuring precise delivery rather than prioritizing speed, offering a dependable and consistent end-to-end byte stream across potentially unreliable networks. This reliability is achieved through positive acknowledgment with retransmission (PAR). TCP is compatible with various streaming protocols such as RTMP, RTSP, HLS, and MPEG-DASH.
    - Type: Internet protocol.
    - Pros: Limited errors and high on reliability.
    - Cons: Slow speed and heavy protocol.
- **User datagram protocol (UDP):** UDP, a connectionless protocol with minimal overhead, operates without the need for recipients to acknowledge the arrival of every data packet, which can introduce some unreliability. Stateless in nature, UDP is particularly well-suited for transmitting data to many clients. It supports multicast for service discovery and broadcasting. With its low rate of retransmission delays, UDP proves to be an excellent choice for real-time applications like VoIP, online gaming, and live video streaming. UDP is compatible with protocols such as SRT, WebRTC, RTSP, and RTP.
    - Type: Internet protocol.
    - Pros: High speed, connectionless and lightweight protocol.
    - Cons: Only does basic error checking and inaccurate.

- **Session initiation protocol (SIP):** Session Initiation Protocol (SIP) serves as a signalling protocol utilized for initiating, managing, modifying, and concluding real-time communications among Internet Protocol (IP) devices. It facilitates the functionality of Voice over Internet Protocol (VoIP) by establishing the communication messages exchanged between endpoints and overseeing various aspects of a call. SIP extends its support to voice calls, video conferencing, instant messaging, and media distribution.
    - Type: Signalling protocol.
    - Pros: High flexibility, high customization offered and easy maintenance.
    - Cons: New and niche and features are still being developed.

	RTMP	WebRTC	FTL	SRT
Pros	• Stability • Low buffering • Wide platform support	• No plugins needed • Supports new codecs • Ultra-low latency	• Sub-second latency	• Supported codecs • High quality • Stream stability • Sub-second latency
Cons	• Relatively high latency • Possible security issues • Old codecs	• Still in development • Instability due to sub-second latency	• Lower quality • Weak platform support	• Weak platform support • No playback
Video Codec	H.264	VP8, VP9, H.264 (AV1 in progress)	H.264	Codec-agnostic
Audio Codec	AAC	Opus	Opus	Codec-agnostic
Latency	3 - 30 sec.	Less than 1 sec.	Less than 1 sec.	Less than 1 sec.

Graphic by Restream.io

Fig 4.1: A comparison of streaming protocols

Note: Mixer, a streaming platform owned by Microsoft, had developed a **Faster Than Light (FTL)** protocol. However, Mixer faced challenges in scaling its operations and decided to shut down the platform. This decision was influenced by the platform's inability to compete effectively with other streaming platforms in the market. So, for comparison purpose it has been mentioned in the table.

4.2 Video codecs and compression techniques

Codec is a technology, either in the form of hardware or software, that is responsible for compressing and decompressing substantial volumes of data. A codec performs the task of converting data from one format to another, with encoding and decoding functions being its key components. During a communication session, the encoder compresses the media file, while the decoder handles the decompression process at the receiving end. Numerous codecs exist, each specifically designed to encode several types of media such as video and audio. The figure below explains the basic working of a codec.

Below are some of popular video codecs:

- **H.264/AVC (advanced video coding):** H.264 is a widely adopted video codec known for its high compression efficiency. It delivers good video quality while minimizing file size, making it suitable for streaming, video conferencing, and broadcasting.
- **H.265/HEVC (high-efficiency video coding):** HEVC is a successor to H.264 and offers even higher compression efficiency. It provides better video quality at lower bitrates, reducing bandwidth requirements. HEVC is commonly used for 4K UHD video streaming and high-quality video delivery.
- **VP9:** VP9 is an open-source video codec developed by Google. It offers efficient compression and high video quality, particularly for web-based video content. VP9 is commonly used in browsers, online video platforms, and streaming services.
- **AV1:** AV1 is an advanced open-source video codec developed by the Alliance for Open Media (AOMedia). It provides efficient compression and high video quality, competing with established codecs like H.264 and HEVC. AV1 is gaining popularity in streaming platforms and video-on-demand services.
- **MPEG-2:** MPEG-2 is an older video codec widely used for DVD video compression and broadcast television. It offers good video quality but has lower compression efficiency compared to newer codecs.
- **MPEG-4:** MPEG-4 is a versatile video codec that supports various multimedia applications, including streaming, video conferencing, and mobile video. It offers a balance between file size and video quality.

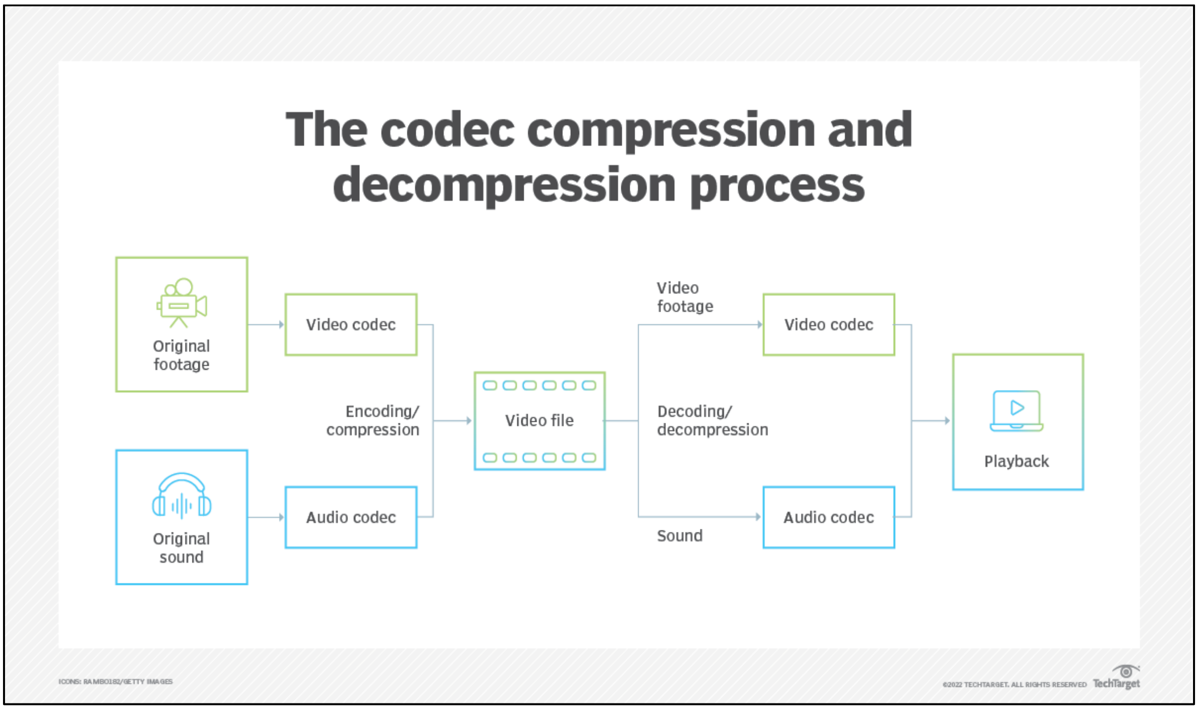


Fig 4.2: A codec

4.3 Adaptive Bitrate Streaming (ABR)

Adaptive bitrate streaming (ABR) is a dynamic process that adjusts the compression level and video quality of a stream in response to the available bandwidth. When streaming video or consuming media content online, the data is transmitted in smaller segments, allowing viewers to watch in real time without the need to wait for the entire content to load.

The video bitrate of a stream refers to the speed at which video data is transmitted to a device and is measured in megabits per second (Mbps). It is essential to consider two key factors when implementing ABR:

1. **Video resolution:** The resolution determines the number of pixels within a frame, impacting the clarity of the images, especially on larger screens.
2. **Frame rate:** The frame rate, typically 24 to 60 frames per second (fps), influences the smoothness of playback, particularly in live sports or other fast-motion streams.

Higher bitrates contribute to better video streaming quality. However, if the bitrate exceeds the user's internet connection bandwidth, buffering issues may arise. Buffering occurs when the video player cannot download the video file quickly enough to maintain normal playback speed, resulting in freezing or the appearance of a loading icon.

The main advantages of using employing ABR in streaming strategy are:

- **Enhanced viewer experience:** ABR enables broadcasters to provide viewers with the highest possible video quality while minimizing interruptions such as buffering, ensuring an optimal viewing experience.
- **Quicker start-up:** ABR offers faster streaming start-up times. By initially delivering a low bitrate stream and dynamically adjusting based on bandwidth estimation and buffering, video playback can commence promptly. This eliminates concerns of losing viewers due to slow loading times.
- **Improved mobile streaming:** ABR significantly enhances the mobile viewing experience on smartphones and tablets. Previously, streaming video on mobile devices posed challenges due to data plan limitations and device processing power. ABR overcomes these limitations, enabling smooth video playback without buffering.
- **Cost-effective solution:** ABR utilizes HTTP-based streaming technologies, which are compatible with standard web servers and Content Delivery Networks (CDNs). This eliminates the need for specialized servers or maintaining persistent connections, resulting in a cost-effective scaling solution without compromising video quality.

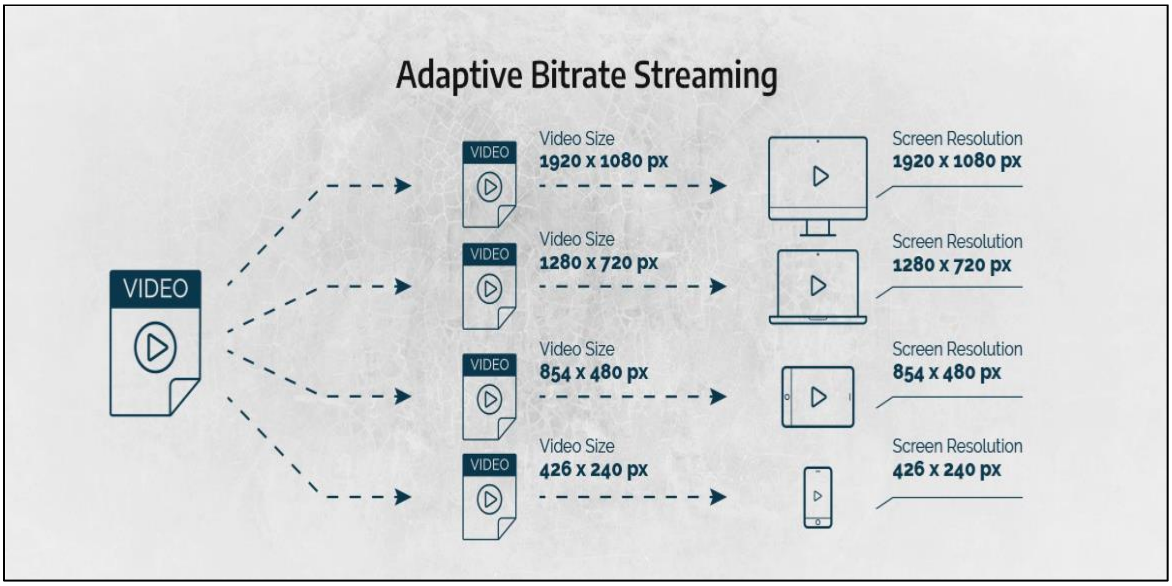


Fig 4.3: ABR Illustration

5. Content Delivery in VoD

Content delivery in VoD refers to the process of distributing media content to viewers over the internet. It involves the efficient and reliable delivery of video files from the content provider's servers to the end-user devices, ensuring a seamless streaming experience. Content delivery in VoD involves various processes and mechanisms to ensure efficient distribution of media content to viewers. Some of the key aspects of content delivery will be discussed in following subsections.

5.1 Content ingestion and transcoding processes

For any content to be delivered, they need to be ingested and transcoded. Content ingestion involves the acquisition and preparation of media files for distribution. It includes tasks such as receiving, encoding, and transcoding the content into various formats suitable for streaming. Transcoding ensures compatibility with different devices and network conditions.

5.1.1 Content ingestion

Ingestion is a fundamental process in media management, often overlooked compared to encoding or streaming. It plays a vital role in organizing and preparing media data for efficient delivery within a digital library. Ingestion ensures that each piece of media content is properly catalogued and stored, ready to be seamlessly accessed when requested by users. It involves capturing, organizing, and assigning metadata to media assets, enabling effective content search, retrieval, and delivery. Ingestion ensures that the media library is well-organized, enabling smooth and efficient content delivery whenever needed. Ingestion is the crucial initial stage in the media processing pipeline. It involves transferring media files from a source to the processing system. This includes steps such as file transfer, transcoding to a suitable format, and extraction of associated metadata. Ingestion is like gathering ingredients, preparing them for processing, and documenting essential details for subsequent stages.

As shown in below diagram, the system ingests new multimedia content by uploading video files. These files are then encoded into multiple video formats with different resolutions and bitrates. The uncompressed audio from the multimedia files is analysed by an audio features engine, generating predefined audio features through frame-based analysis. The resulting feature vectors are fed into a pipeline of classifiers, consisting of individually trained models that capture specific content characteristics. The system also captures user-provided metadata, such as video title, description, and tags. All the captured metadata, encoded video file locations, feature vectors, and classification results are stored in a database object.

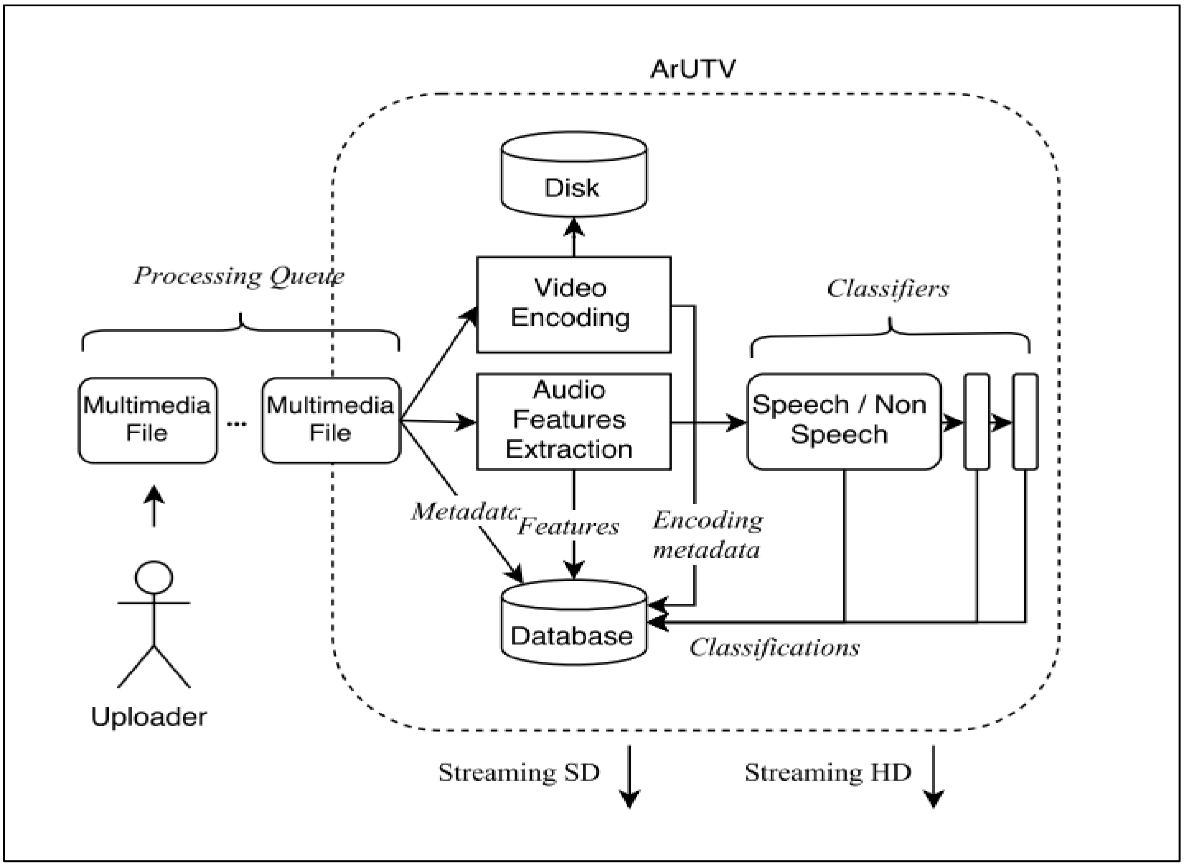


Fig 5.1: Multimedia content ingestion

5.1.2 Content transcoding

Transcoding is the process of converting one type of digital media, either audio or visual, to another format. It involves unencoding an already encoded media file to modify its size or bitrate. The primary purpose of transcoding is to increase compatibility with different playback devices, making the media accessible to a broader audience. It allows users to watch content on various devices, such as televisions and smartphones. Transcoding also involves changing the audio and/or video codec used for compression, ensuring compatibility with different streaming protocols. Transcoding is a broad term that encompasses various tasks related to digital media. However, it is commonly utilized for the following purposes:

Transrating: It involves adjusting the bitrate of a file, which determines the data transmission speed across a network. For instance, a 4K video stream at 13Mbps can be converted into multiple renditions with lower bitrates, such as HD at 6Mbps, 3Mbps, 1Mbps, or 600kbps.

Transizing: It focuses on resizing the video frame, such as reducing a resolution from 3840x2160 (4K UHD) to 1920x1080 (1080p) or 1280x720 (720p).

A simple illustration for transcoding is shown in below figure.

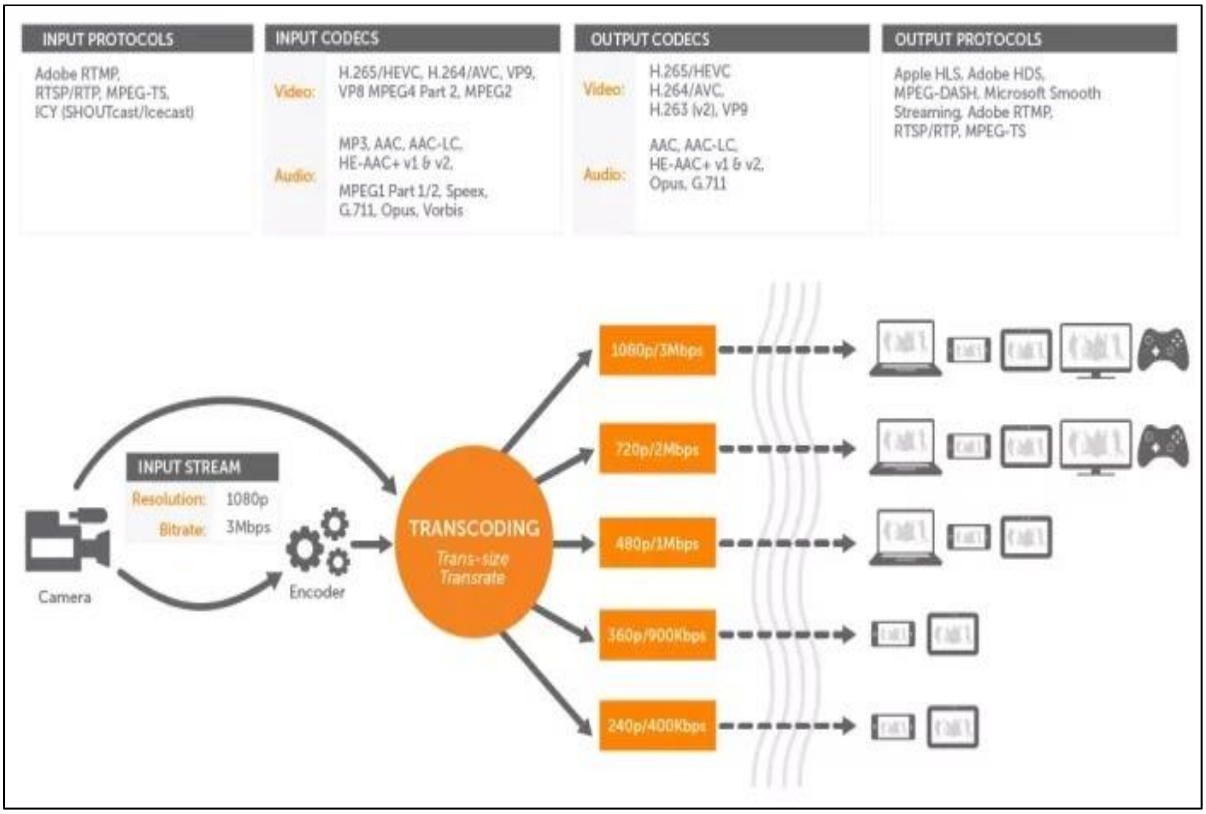


Fig 5.2: Transcoding process

It is worth mentioning that transcoding process should not be confused with similar terms like **encoding** and **transmuxing.**

5.2 Storage and caching mechanisms

Storage and caching mechanisms play a crucial role in the efficient delivery of VoD content. These mechanisms involve storing and managing media files, metadata, and other related data to ensure quick and reliable access for viewers.

Storage mechanisms typically involve using dedicated servers, cloud-based storage solutions, or content delivery networks (CDNs) to store the media files securely. The choice of storage solution depends on factors such as scalability, redundancy, and cost-effectiveness.

Caching is a technique that involves storing a portion of data in a high-speed storage layer to facilitate faster retrieval of that data in future requests. It acts as a temporary storage for frequently accessed or transient data, allowing subsequent requests to be served more quickly compared to accessing the data from its primary storage location. Caching enables efficient reuse of previously retrieved or computed data, optimizing performance and reducing the need for repeated computations or accessing slower storage systems. By keeping frequently accessed data closer to the users or applications, caching improves response times, reduces latency, and enhances overall system performance.

A cache stores data in fast-access hardware like RAM and is often combined with software components. Its main objective is to enhance data retrieval performance by minimizing the need to access the slower underlying storage layer.

Caches prioritize speed over capacity and typically store a portion of data temporarily, unlike databases that store complete and durable data.

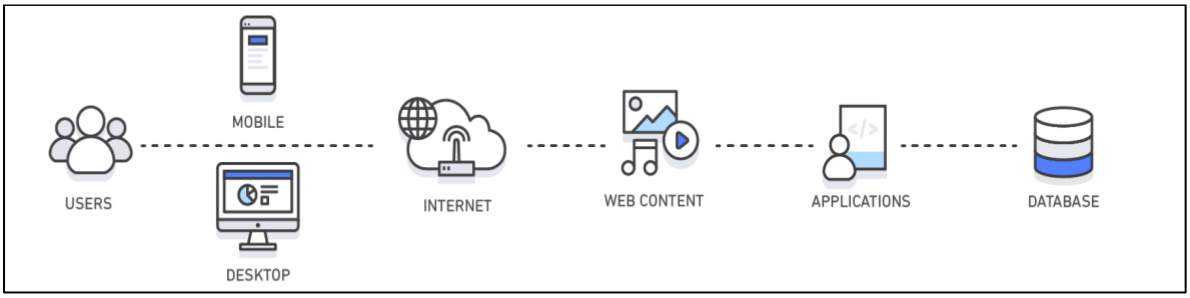


Fig 5.3: Caching

There are several benefits of caching in VoD systems. It improves performance of application, reduces load on backend, limits database cost to an extent, removes database hotspots, keeps application data predictable even in case of high load and increases Read Throughput (IOPS)

5.3 Content distribution strategies

Content distribution strategies in a VoD system serve the purpose of ensuring the effective and dependable delivery of media content to end-users. These strategies are developed to streamline the distribution process and enhance the overall satisfaction of users. Content distribution strategies involve the implementation of diverse techniques and technologies that aim to enhance the availability, scalability, bandwidth usage, reliability, and performance of content delivery.

Some of the most used strategies for content distribution in VoD are listed below,

5.3.1 Edge caching

The websites or applications might be slow at times and the media might keep buffering which can be frustrating for user. To overcome this issue there are several ways, but one prominent way is to cache at the **edge**. The internet is an extensive network of interconnected computers and servers, facilitating data sharing, processing, and communication. The **"edge"** denotes the outermost part of this network, where devices like routers, laptops, and mobile phones connect to access and interact with the internet. The emergence of robust edge devices has spurred the advancement of **edge computing**, a technology that offers solutions to address performance constraints found in traditional architectural patterns. **Edge computing** represents a shift from traditional cloud-computing patterns. Unlike the centralized approach, where computations are conducted in an individual location, **edge computing** enables processing at the network's edge, closer to end-users. This proximity reduces latency and eliminates the need for data to traverse across long distances, enhancing overall performance and user experience.

Caching has long been utilized to enhance performance even before the introduction of edge computing. Previously, caches were stored on servers or devices within the network, like web servers or content delivery networks (CDNs).

However, with the emergence of edge computing, the concept of the edge has evolved from a mere point of network access to a place for data processing and storage. In the context of **edge caching**, it has become a valuable location to store frequently accessed data or content. **Edge caching** works by storing frequently accessed data or content at locations closer to end-users, such as edge servers or devices located at the network edge. When a user requests a piece of content, the edge cache checks if the requested data is already available in its local storage. If the data is present, it can be delivered directly to the user without needing to fetch it from a centralized server. Below is a simple diagrammatical perspective of edge caching.

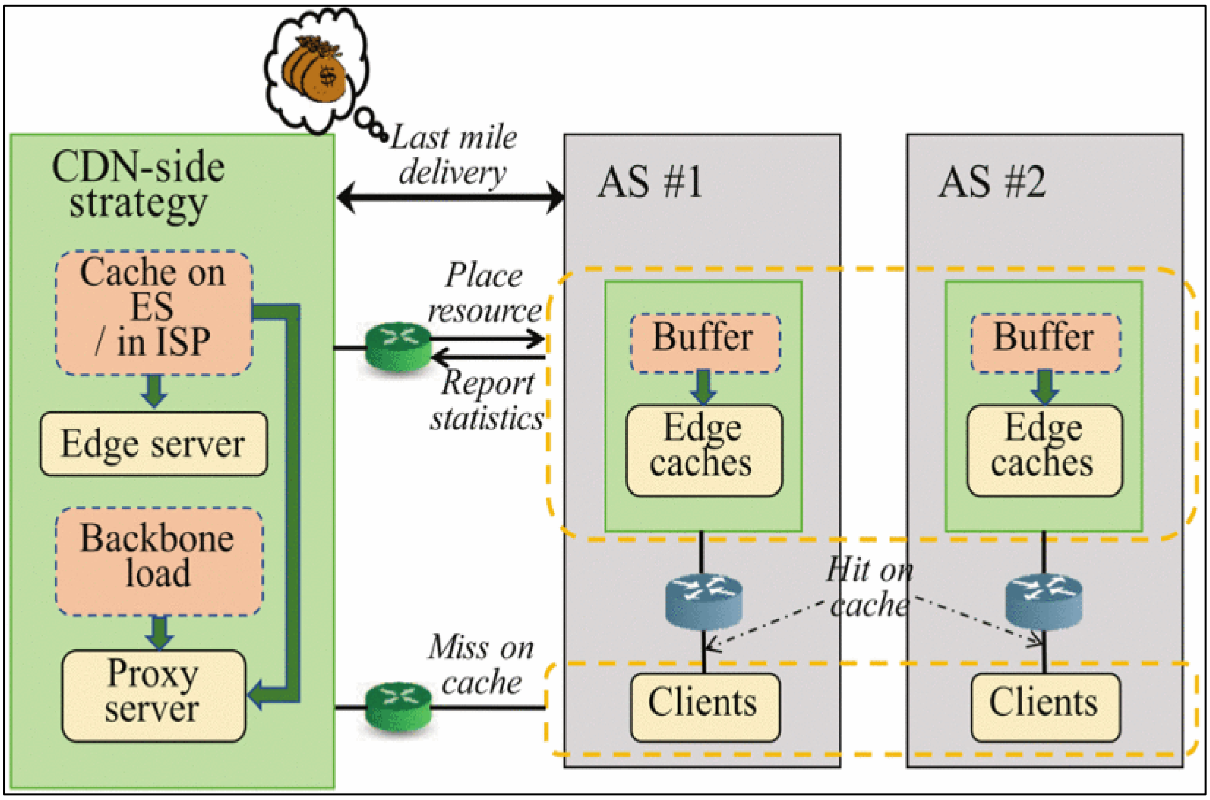


Fig 5.4: Edge caching

Edge caching can reduce latency and increase speed of system by several folds, it makes the system more secure, makes VoD system more reliable and can reduce overall cost. However, the capacity of edge servers may be limited, and the infrastructure is more complex and is still in its infancy The data and content to be cached must be selected wisely as it may lead to inconsistency.

5.3.2 Content delivery network (CDN)

A Content Delivery Network (CDN) is a network of distributed servers that store cached copies of website content based on users' geographic locations. In a Content Delivery Network (CDN), the original content is stored on the origin server, while multiple distributed edge servers are in various locations worldwide.

CDNs use edge locations to store cached copies of content, ensuring that visitors can access the website's static assets from the closest edge server, minimizing the distance between the user and the server. This caching process reduces latency and improves scalability as subsequent requests for static assets are served from the edge servers rather than the origin server. For instance, in the below figure, if a user from the UK requests a website hosted in the USA, they will be served from the closest edge location, like the London edge location. This significantly speeds up content delivery compared to making a complete request to the origin server, which would result in higher latency.

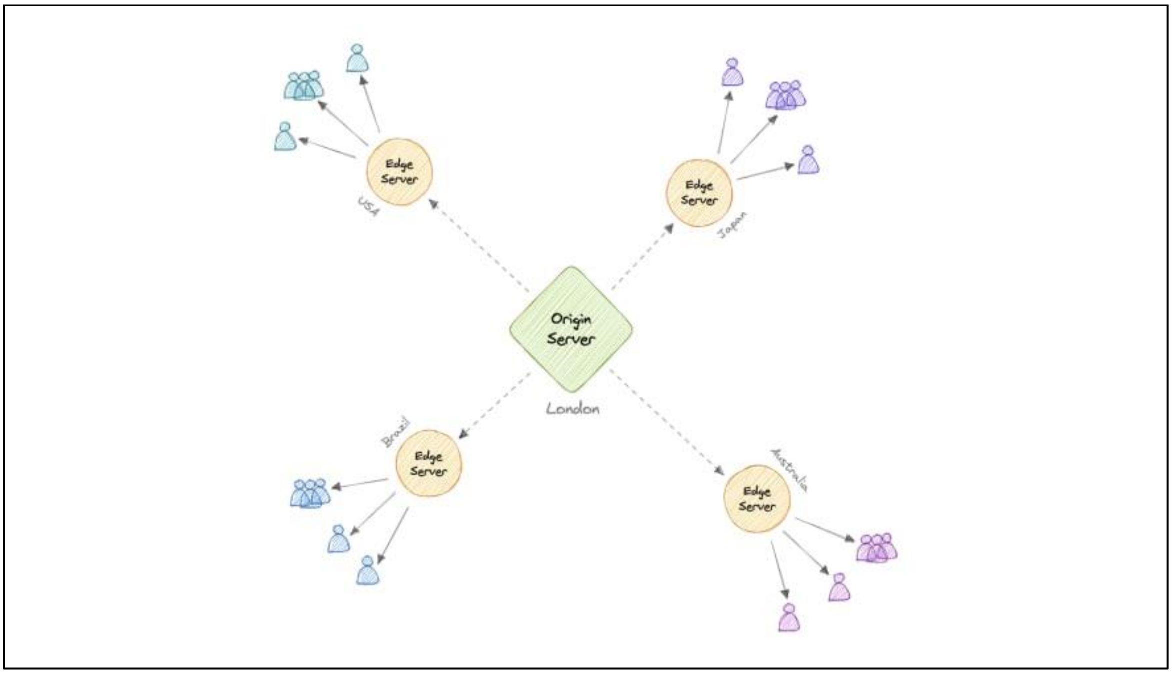


Fig 5.5: CDN

CDN provides better connectivity, better scalability, user satisfaction, better response to network spikes, low latency, secure network, and improved content delivery. However, in this strategy, the cost might be on higher side, and reachability might be restricted where the servers are not installed.

5.3.3 Multi-CDN

A Multi CDN is a solution that integrates multiple Content Delivery Networks (CDNs) from various providers into a unified network, offering increased coverage and improved website performance compared to using a single CDN. Since different CDN vendors operate diverse networks of edge servers or Points of Presence (PoPs), along with unique features and functions, their performance, security, and cost benefits may vary based on factors such as traffic, workload, and geographic location. Multi CDN enables organizations to optimize their content delivery by directing specific traffic and workloads to the most suitable CDNs, allowing them to use "the best CDN for the job."

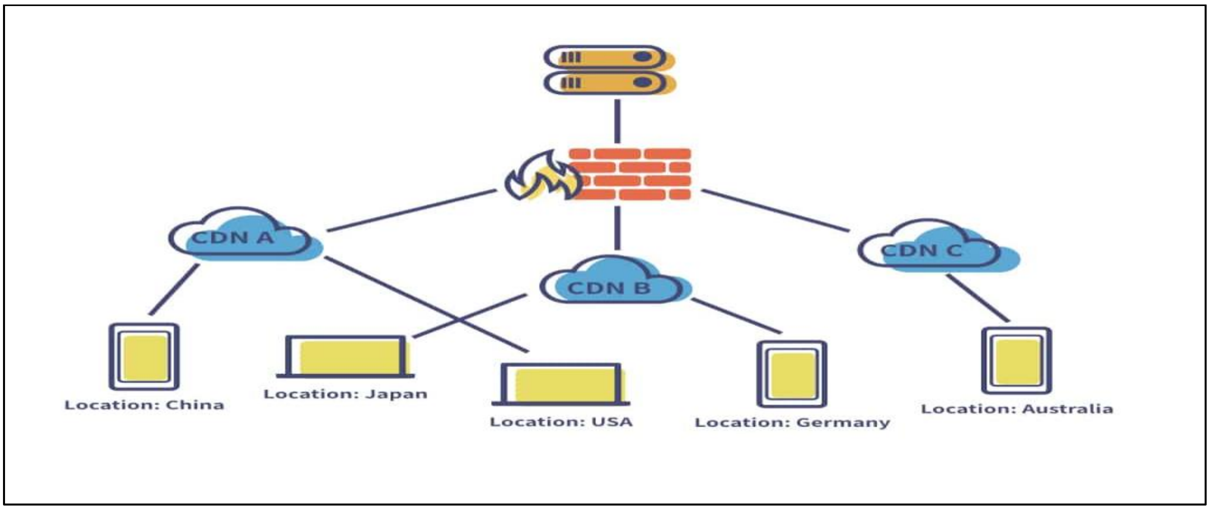


Fig 5.6: Multi-CDN

5.3.4 Peer-to-peer distribution

P2P distribution leverages the resources of individual users to distribute content. It allows users to download and stream content from other users' devices, reducing the burden on central servers and improving scalability. Each peer acts as both a consumer and distributor of content, making it possible to create a decentralized and self-organizing network.

When a user requests a specific piece of content, the P2P system locates available copies of that content from other peers within the network. The content is then downloaded simultaneously from multiple sources, enhancing the download speed and reducing the load on individual servers. As more peers join the network, the content's availability and distribution efficiency improve, creating a scalable and robust system for content delivery.

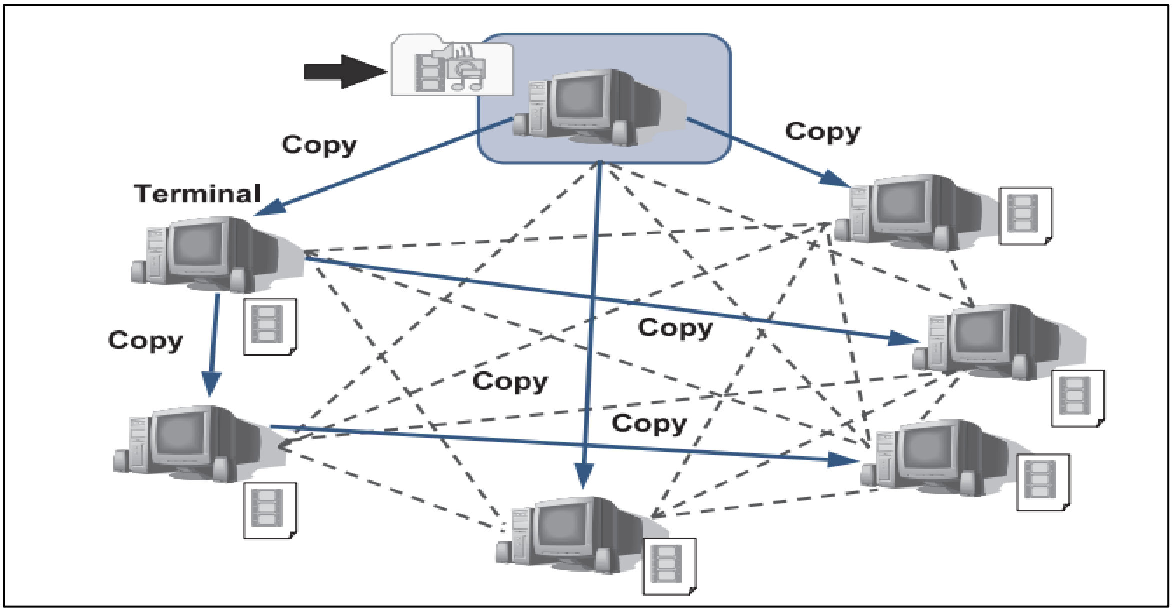


Fig 5.7: P2P distribution

Note: The content distribution strategies are related to VoD architecture but are not similar. The strategies might have similarity with architecture, but both the aspects are of different purpose in VoD system. VoD architecture refers to the overall design and structure of the Video on Demand system, including various components like content storage, encoding, transcoding, metadata management, and user interfaces. It encompasses the entire system's organization and infrastructure to efficiently manage and deliver video content to viewers. On the other hand, content delivery strategies are specific techniques and methods used to ensure efficient and reliable distribution of media content to end-users. These strategies include edge caching, multi-CDN, adaptive bitrate streaming, and other techniques to optimize content delivery, reduce latency, and enhance the user experience.

6. Scalability in VoD Systems

Scalability in Video on Demand (VoD) systems refers to the system's ability to handle increasing amounts of traffic, users, and content without compromising performance. A scalable VoD system can effectively expand its capacity and resources to accommodate higher demand, ensuring a seamless user experience even during peak times. In this chapter different scaling techniques and related topics will be covered.

6.1 Horizontal and vertical scaling

In a Video on Demand (VoD) system, horizontal and vertical scaling are employed to manage increased demand and enhance system performance. Horizontal scaling, or scaling out, entails adding more servers to distribute the load and increase capacity. This approach handles higher traffic by spreading the workload across multiple servers, making it useful for accommodating sudden spikes in demand. On the other hand, vertical scaling, or scaling up, involves upgrading the resources of a single server, such as CPU and memory, to handle a larger load. The combination of these scaling techniques creates an efficient and responsive VoD system that can handle varying traffic loads and ensure a smooth user experience while accessing video content.

However, in VoD industry horizontal scaling is more prevalent. To accommodate increasing number of users, VoD giants like Netflix, Amazon Prime, etc. Use horizontal scaling for their networking infrastructure.

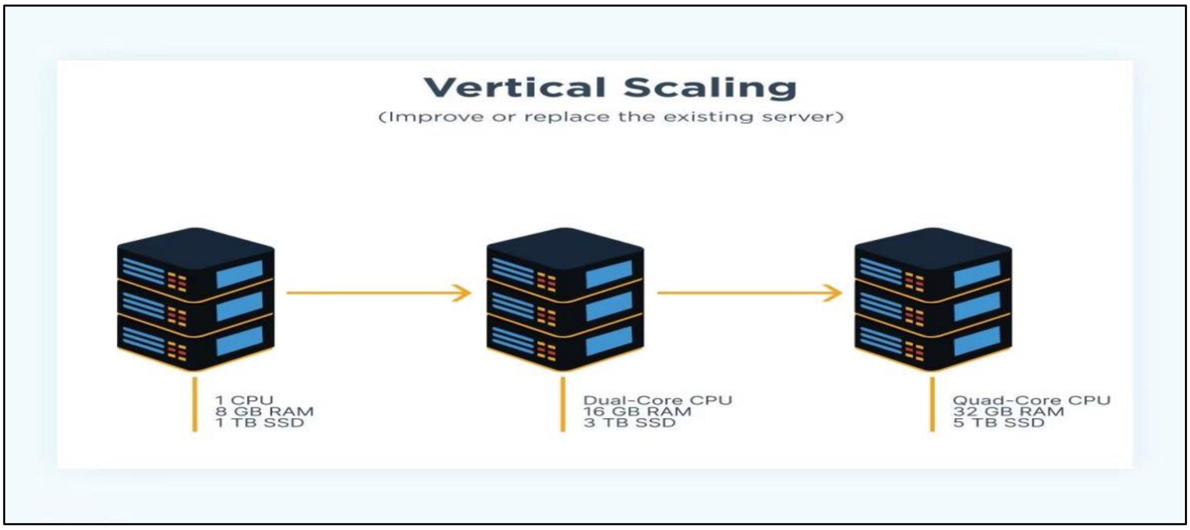


Fig 6.1: Vertical scaling

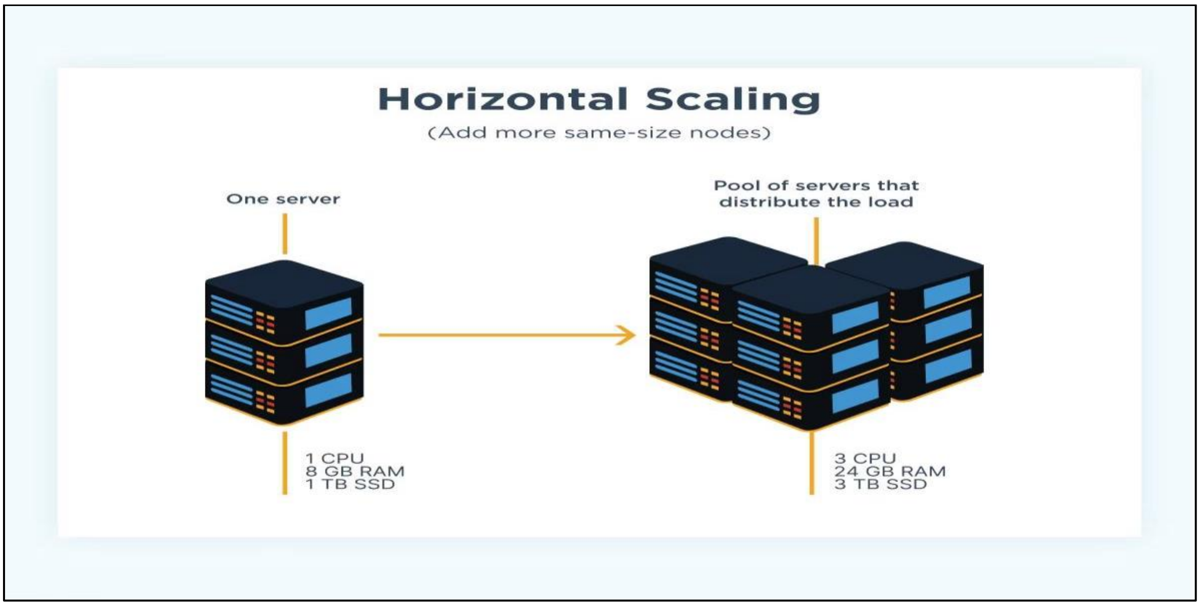


Fig 6.2: Horizontal scaling

6.2 Load balancing techniques

Load balancing is a method used to evenly distribute user requests across multiple available servers. Its purpose is to ensure that requests are directed to servers that are in working condition. In other words, load balancing helps distribute the workload across different servers. For instance, if there are 10 servers and 1000 users, the ideal scenario would be for each server to handle the requests of 100 users. Load balancing is achieved using load balancers, which facilitate the process of evenly routing user requests to the available servers. The below figure captures load balancing concisely.

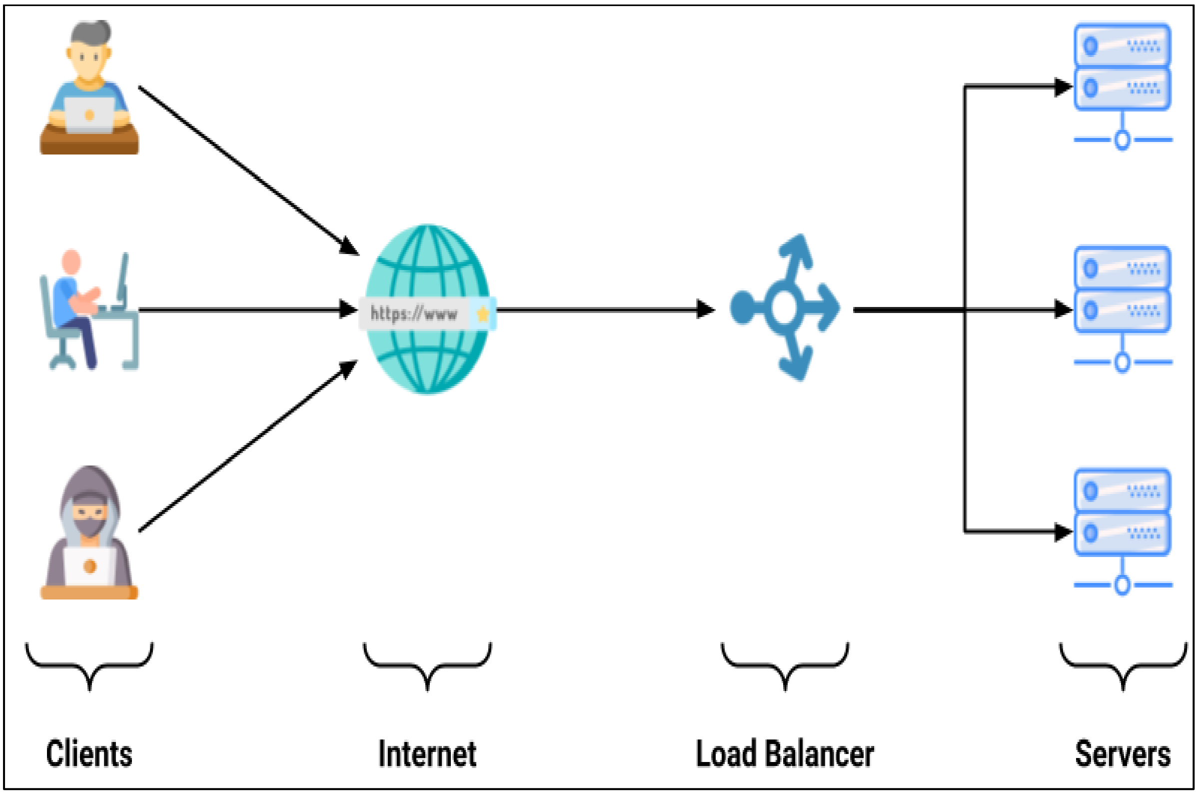


Fig 6.3: Load balancing

Load balancers are responsible for directing user requests only to healthy servers, meaning those that are responding correctly. To achieve this, load balancers use various algorithms for routing requests among the available servers:

- **Round Robin:** This algorithm treats all servers as if they were placed in a circle. Each request is routed to the next server in the cycle, ensuring an even distribution of requests among servers.
- **Least Connection:** Requests are directed to the server with the fewest active connections at the given moment. This helps distribute the load based on the current server load.
- **Least Time:** The request is sent to the server that has the fastest response time and the least number of active connections, optimizing performance.
- **Hashing:** Requests are routed based on a specific key, such as the user's IP address, request URL, or requested. The key is hashed, and the request is directed to the server corresponding to the hashed value.
- **IP Hashing:** In this algorithm, the client's IP address is hashed using a hash function, and the request is directed to the server associated with the hashed value.

Load balancing plays a vital role in VoD systems by effectively distributing user requests among the available servers in a VoD system and optimizing resource usage. It ensures that the system can handle higher demand, maintain performance, and deliver a smooth user experience. As user traffic increases, load balancing enables the system to scale smoothly without compromising responsiveness or reliability.

6.3 Content replication

Media files and data are duplicated and stored in multiple locations, often referred to as replicas. These replicas are placed strategically in different geographic locations or data centres to reduce the distance between the content and end-users. Content replication is a fundamental strategy that enhances scalability in VoD platforms by improving performance, load distribution, availability, and flexibility. It enables platforms to handle increased demand, ensuring a seamless user experience as the user base grows and the system expands.

6.4 Scalable video coding (SVC) and its role in scalability

Scalable Video Coding (SVC) is a video coding standard that plays a crucial role in achieving scalability in video streaming and transmission. SVC allows video content to be encoded and delivered in multiple layers, providing flexibility in adapting to different network conditions and end-user devices. Here's how SVC contributes to scalability:

6.4.1 Spatial scalability

Spatial scalability in video coding is achieved using a pyramid approach, where pictures of different spatial layers are independently encoded with specific motion parameters. This helps improve coding efficiency compared to simulcast, which transmits separate video streams for each layer. To further enhance efficiency, inter-layer prediction mechanisms are introduced to eliminate redundancies between layers. These mechanisms are switchable, allowing encoders to freely choose a reference layer for an enhancement layer, optimizing the use of inter-layer prediction. For effective inter-layer prediction, the temporal prediction structures of spatial layers should be temporally aligned. These inter-layer prediction concepts include techniques for motion parameter and residual prediction.

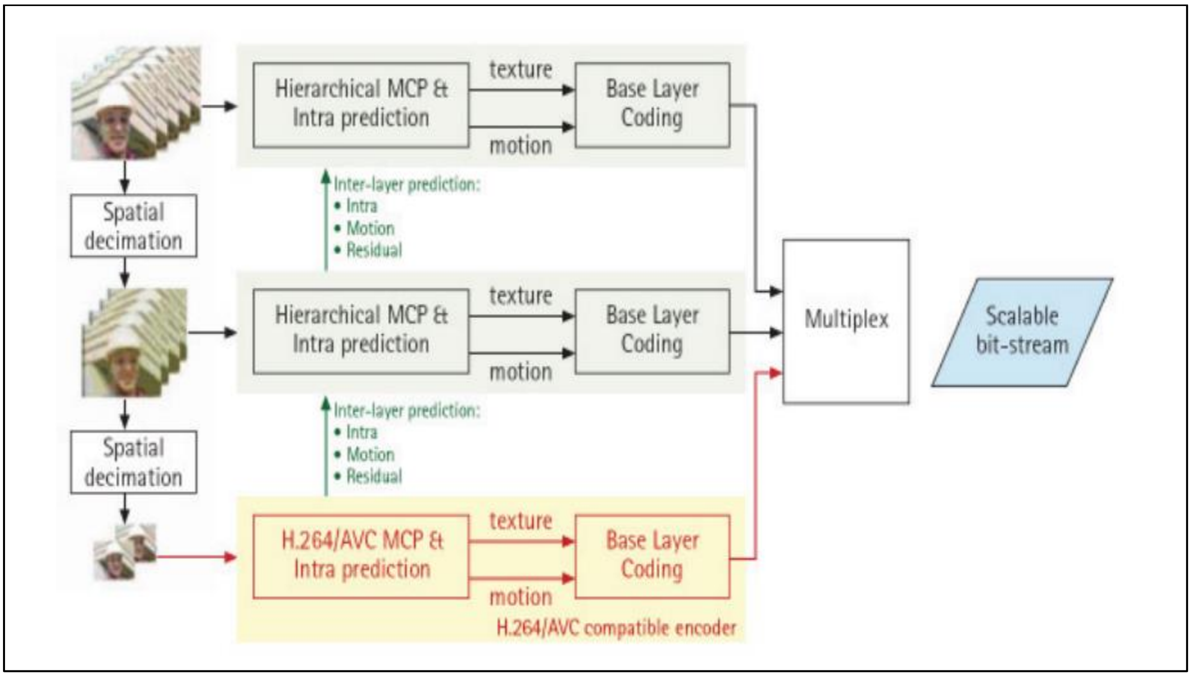


Fig 6.4: Spatial scalable architecture of scalable extension of H.264/AVC

6.4.2 Temporal scalability

Temporal scalable bitstreams can be generated using a hierarchical prediction structure without modifying H.264/AVC. The structure consists of four dyadic hierarchy stages, providing four temporal scalability levels. The first picture of a video sequence is an intra-coded IDR picture, followed by key pictures coded at regular intervals. A key picture and the pictures between it and the previous key picture form a Group of Pictures (GOP). Key pictures can be intra-coded or inter-coded using previous key pictures as reference for motion compensated prediction, while the rest of the GOP is hierarchically predicted. For instance, in below figure, layers 0, 1, 2, and 3 have 3, 5, 9, and 18 frames, respectively.

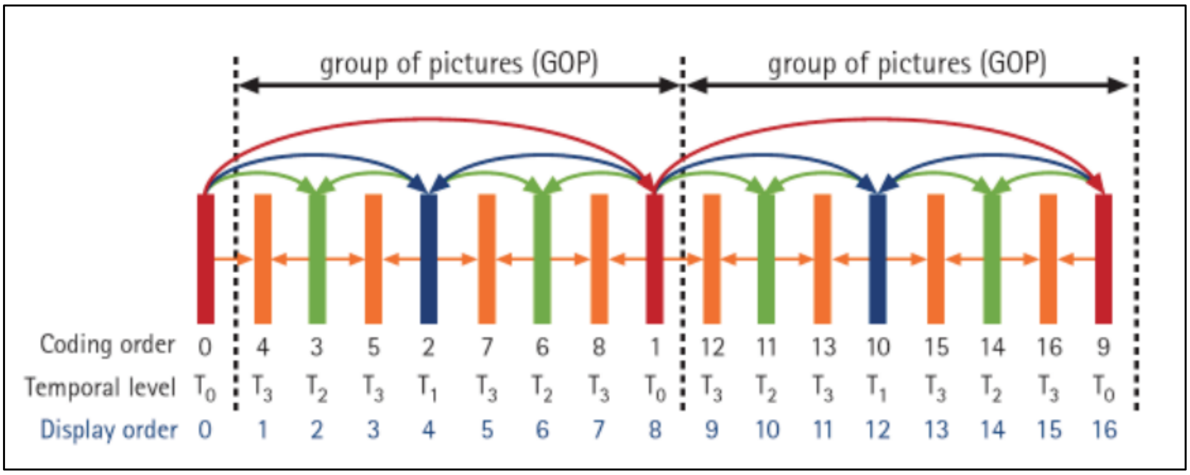


Fig 6.5: Temporal scalable architecture of Scalable extension of H.264/AVC

6.4.3 SNR scalability

In the scalable extension of H.264/AVC, two types of SNR scalability are provided: coarse-grain SNR scalability (CGS) and medium-grain SNR scalability (MGS). CGS employs the same inter-layer prediction mechanisms used for spatial scalability. On the other hand, MGS enhances the granularity for SNR scalability, enabling bitstream adaptation at the network adaptation layer (NAL) unit level. This allows more precise adjustments in SNR scalability, providing greater flexibility in managing video quality based on network conditions.

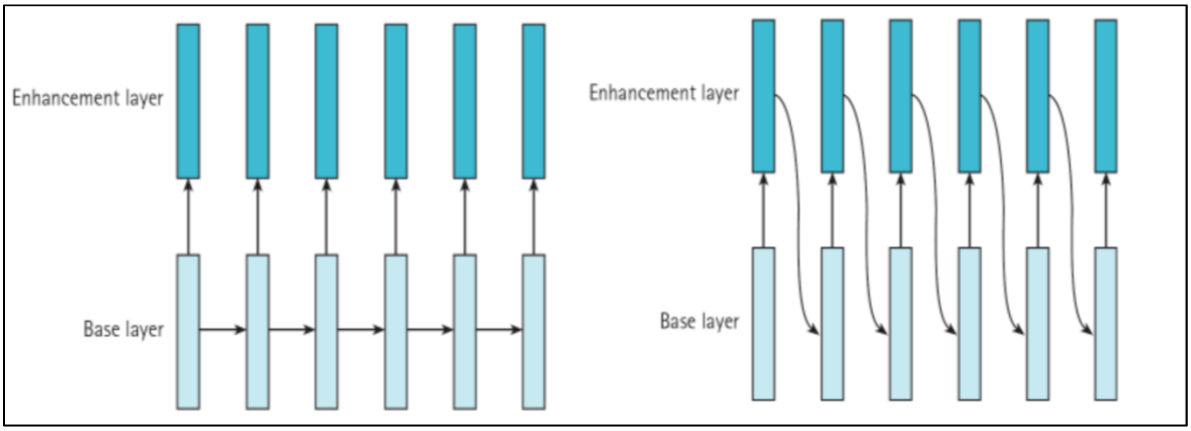


Fig 6.6: SNR scalable architecture for MPEG4, MPEG2.

7. VoD Content Management System

A video content management system (also referred to as a video CMS) is a software platform that facilitates the organization, storage, management, and distribution of video content. This platform is specifically designed to address the unique needs associated with managing video files in a variety of formats, resolutions, and bitrates. Video content management systems (CMSs) facilitate the upload, editing, categorization, and publication of video content for various purposes. These purposes can range from public-facing videos such as promotional videos, tutorials to private videos such as training materials or internal communications or entertainment videos. Some of the topics which come under video CMS are discussed below.

7.1 Content acquisition and licensing processes

To ensure legal compliance and ensure the content is made available to users for streaming, content acquisition and licensing procedures are essential. These processes involve the identification of content, the sourcing of content from various sources, and the negotiation of licenses. Licensing terms may vary depending on the type of content, the length of availability, the geographic area, and the distribution rights. To avoid any copyright infringement, content licensing agreements should focus on the ownership and usage of copyright. Content owners are typically remunerated through royalty payments, which are based on the use and consumption of their material on the platform. Licensing agreements should provide information on payment terms and methods, as well as the implementation of digital rights management (DRM) technologies to ensure that only authorized users can access and deliver copyrighted content. Clearances and releases for third-party materials are essential to prevent copyright violations. Legal experts are also involved in the review of licensing agreements to ensure that terms and conditions are compliant with legal requirements and protect the rights of both content owners and VoD platforms. Below is a diagram showing how the flow of content acquisition and licensing process by Adobe.

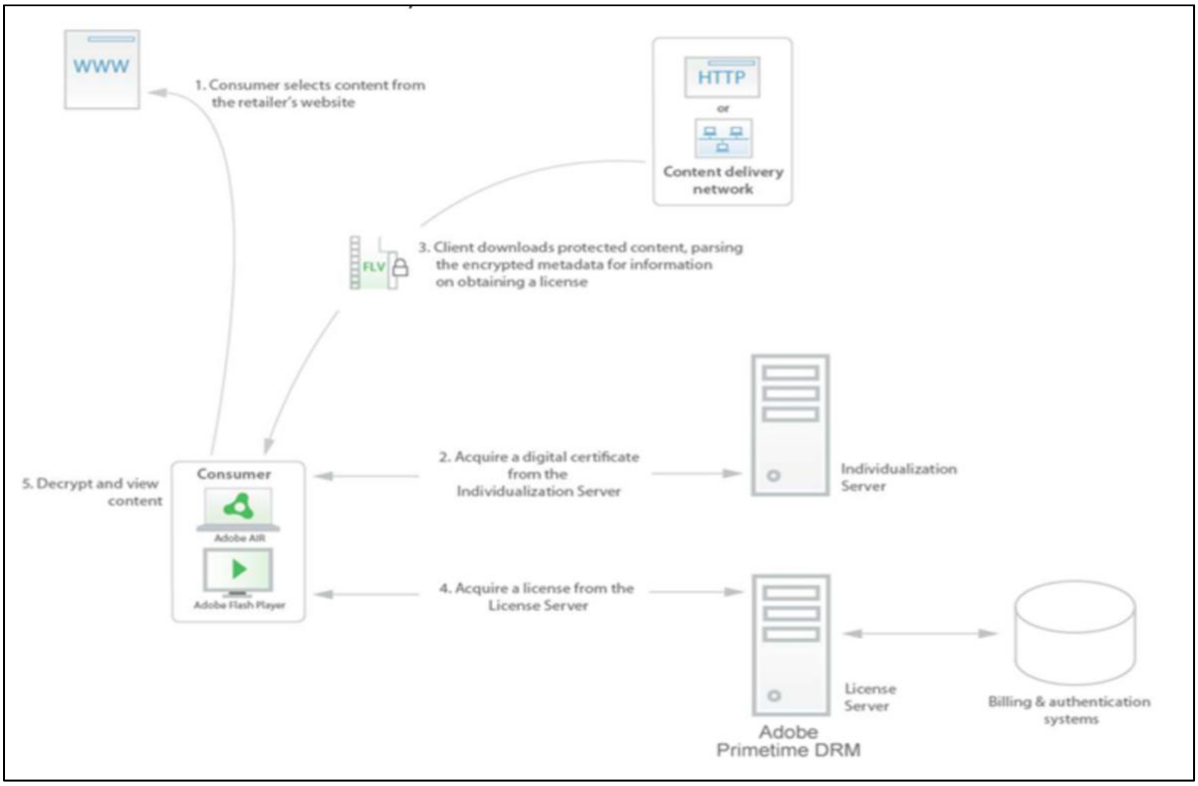


Fig 7.1: Content acquisition and licensing in Adobe

7.2 Metadata extraction and enrichment techniques

What is video metadata?

Video metadata is the data or information associated with a particular video. All digital content has its own metadata for categorization and searchability. For instance, photographs taken with modern cameras contain predefined metadata, such as the date and time, the filename, and the geolocation. For videos, descriptive metadata are employed, which enables user to determine the primary attributes of the file. Inserting video metadata is primarily for the purpose of providing search engines with textual descriptions of the video content, for improved accessibility and legibility. Video metadata enables search engines to identify which videos are most pertinent to a user's search queries, thus providing accurate results. Additionally, platforms such as Netflix utilize machine learning algorithms to read video metadata and make personalized video recommendations.

Metadata can be divided into six or more distinct types, depending on the type of content and industry. However, only three types of metadata are applicable to videos:

1. **Administrative metadata:** Administrative metadata contains all the necessary information for video administrators to manage a video, including hardware and software, the file creator's name, and the creation date.
2. **Structural metadata:** Structural metadata is used to organize and categorize video assets.
3. **Descriptive metadata:** Descriptive metadata is used to optimize metadata for search engines.

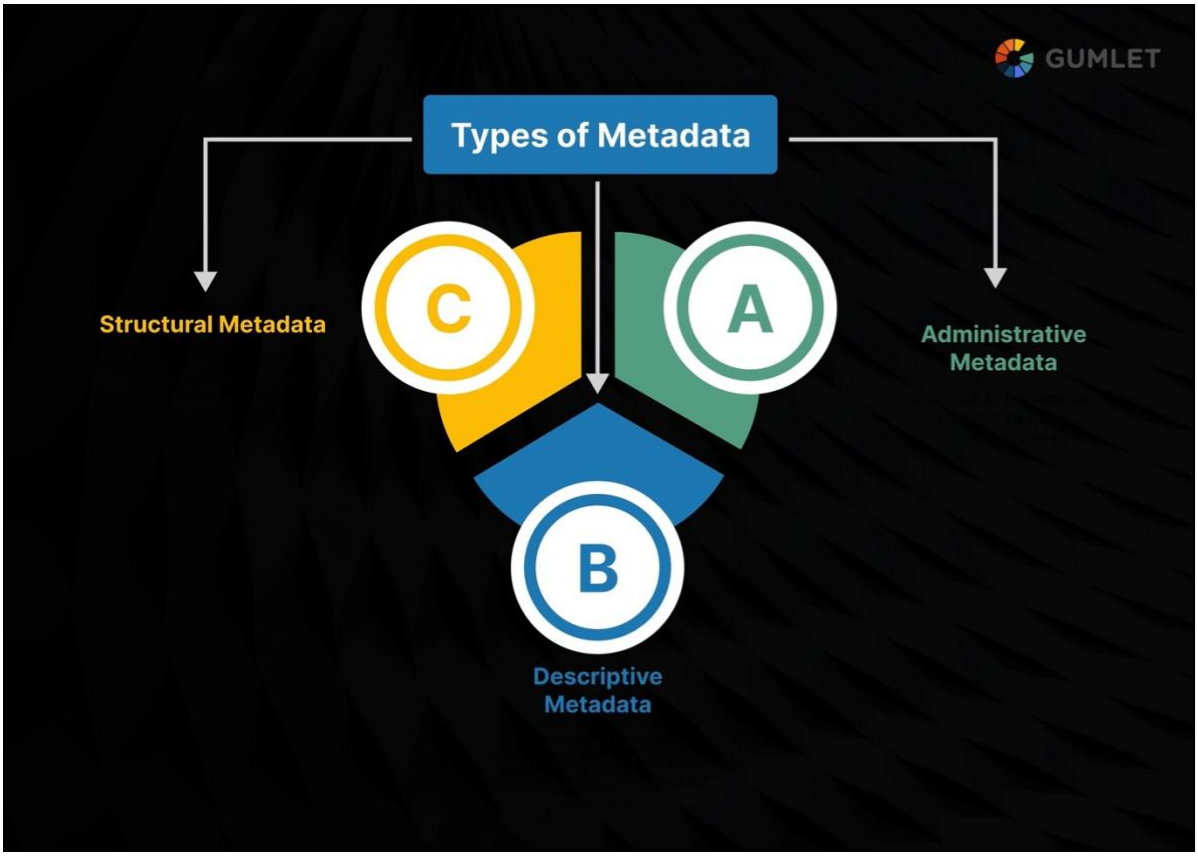


Fig 7.2: Type of video metadata

Below figure encapsulates importance of video metadata in brief.

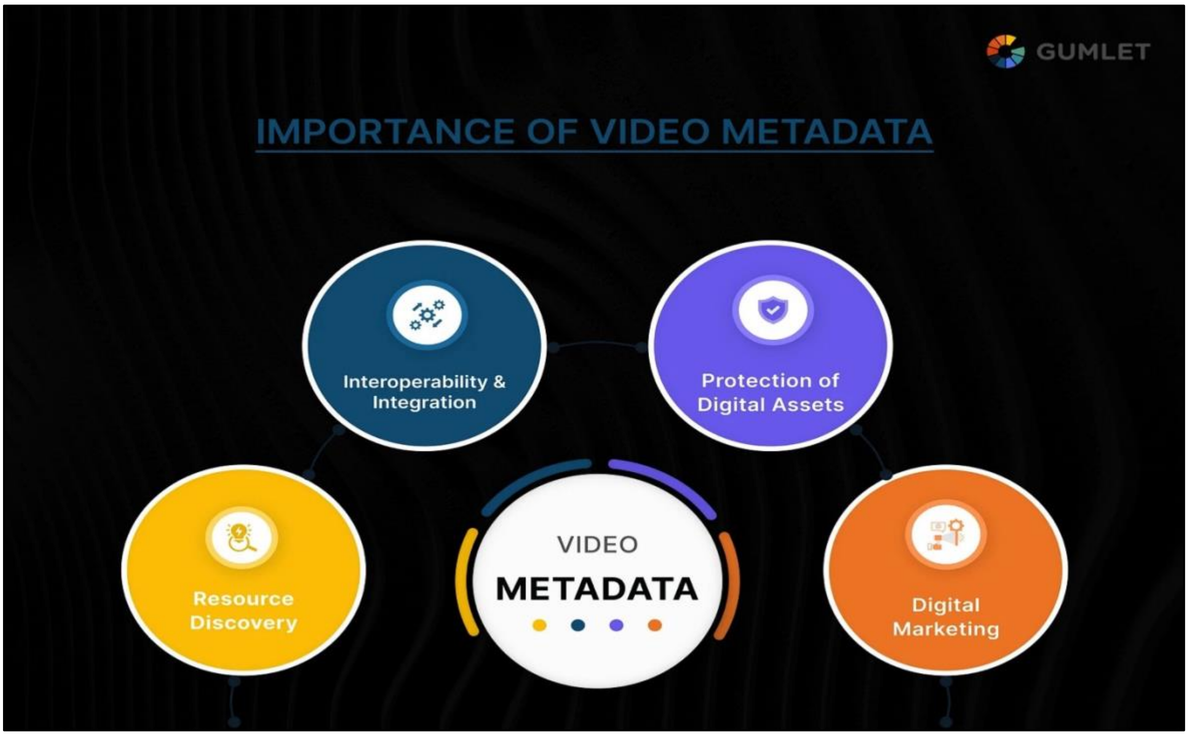


Fig 7.3: Importance of video metadata

In a Video CMS, metadata extraction and enrichment techniques serve as crucial elements to elevate the organization, discoverability, and user experience of video content. These methods involve capturing relevant details from videos and augmenting them with additional data to provide a deeper context and structure to the content. Here are some key aspects of these techniques:

- **Automated metadata extraction:** Video CMS platforms utilize automated algorithms and machine learning methods to retrieve metadata from video files. This includes technical information like video duration, resolution, frame rate, video codec, audio codec, and more.
- **Speech-to-text transcription:** Leveraging speech recognition technology, video CMS converts spoken words in videos into text, creating searchable transcripts that enhance content discoverability and accessibility.
- **Facial recognition:** Sophisticated facial recognition algorithms identify individuals in videos and associate them with metadata, making it effortless to search and categorize content based on featured people.
- **Object recognition:** Employing object recognition technology, video CMS detects and identifies objects or scenes within videos, resulting in descriptive metadata that improves content understanding and ease of searching.
- **Scene detection:** Video CMS can analyse videos to detect scene changes, enabling the division of content into chapters or segments for smoother navigation.
- **Sentiment analysis:** By employing sentiment analysis techniques, video CMS assesses the emotions and tone expressed in videos, offering valuable insights into the content's mood and sentiment.
- **Geolocation tagging:** Geolocation data extracted from videos can be used to add location-based metadata, empowering users to discover videos related to specific geographic areas.
- **Content categorization:** Videos can be categorized based on genres, topics, or themes, simplifying the process for users to find content that aligns with their interests.
- **User-generated tags:** Users can contribute tags or keywords to videos, allowing for community-driven metadata enrichment and enhanced content discovery.
- **Integration with external databases:** Video CMS platforms can seamlessly integrate with external databases or APIs to access additional metadata, such as movie or TV show details from IMDb or other sources.

By employing metadata extraction and enrichment techniques, Video CMS platforms optimize content management, searchability, and personalization, enriching the user experience. Leveraging these methods, video platforms can foster higher user engagement, provide more tailored content recommendations, and elevate overall user satisfaction.

7.3 Content categorization and tagging

Tags and categories are two essential elements that can be utilized to organize information on a VoD portal and facilitate user discovery through search and navigation. **Tags** are a useful feature that can be used to organize data on a portal and facilitate user discovery. They are words or sentences that can be affixed to any piece of content on a website. **Categories** are analogous to tags in that they are intended to be used by administrators, rather than by users. Categories can be hierarchically organized and grouped together in dictionaries. Whereas tags are an informal way for users to organize content, categories are designed to enable administrators to structure content in a formal, hierarchical manner. Content can be tagged and categorized in a variety of ways. Some examples are manual tagging, automated tagging, user generated tags, taxonomy-based tagging, trending tags, etc.

Content categorization and tagging enables user to find relevant content easily and quickly. It also helps in content personalization, distribution, and monetization.

7.4 Content recommendation systems and algorithms

A recommendation system, also known as a recommender system, is a type of machine learning which uses data to anticipate, filter, and discover what people are searching for among an ever-increasing number of choices. A recommendation system is an AI algorithm, often powered by machine learning and Big Data, that uses various data points, such as past viewed contents, search history, and demographics, to suggest additional content to users. These systems are valuable in helping users discover content they might not have found otherwise. By analysing user interactions like clicks, likes, and purchases, recommender systems can predict personalized consumer interests and provide relevant recommendations. This capability makes them popular among content and product providers, as they can drive users to a wide range of offerings, from books and videos to health classes and clothing.

There are number of recommender algorithms and techniques available, but they generally fall into three main categories:

7.4.1 Collaborative filtering

Collaborative filtering algorithms in recommender systems suggest items based on user preferences gathered from multiple users. By analysing past interactions between users and items, these algorithms can predict future interactions. The approach involves identifying similarities in user preference behaviour, such as items purchased or rated previously, and using this information to build a predictive model. If multiple users have made similar decisions in the past, the recommender system assumes they will likely agree on additional future selections. For example, if the system finds that two users share similar tastes in movies, it might recommend a movie to one user that it knows the other user already enjoys.

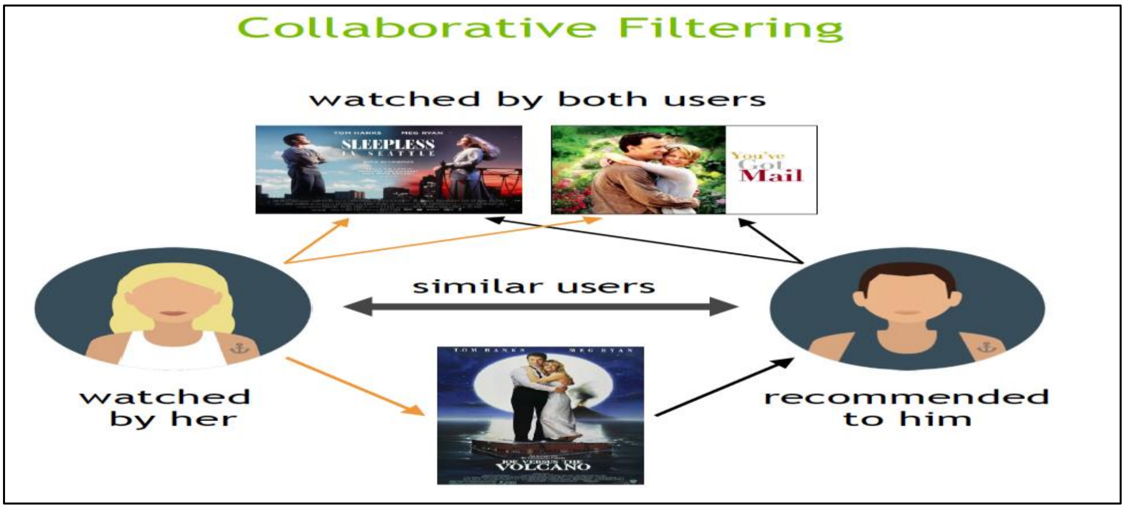


Fig 7.4: Collaborative filtering

7.4.2 Content filtering

Content filtering in recommender systems relies on the attributes or characteristics of an item to recommend other items that align with the user's preferences. This approach is based on analysing similarities between user and item features. By considering information about the user, such as age or preferences, and information about items they have interacted with, the system models the likelihood of a new interaction. For instance, if a content filtering recommender system observes that a user enjoyed movies like "You've Got Mail" and "Sleepless in Seattle," it might suggest another movie with similar genres or cast, such as "Joe Versus the Volcano." The recommendation is based on matching specific attributes of the items to user’s past preferences.

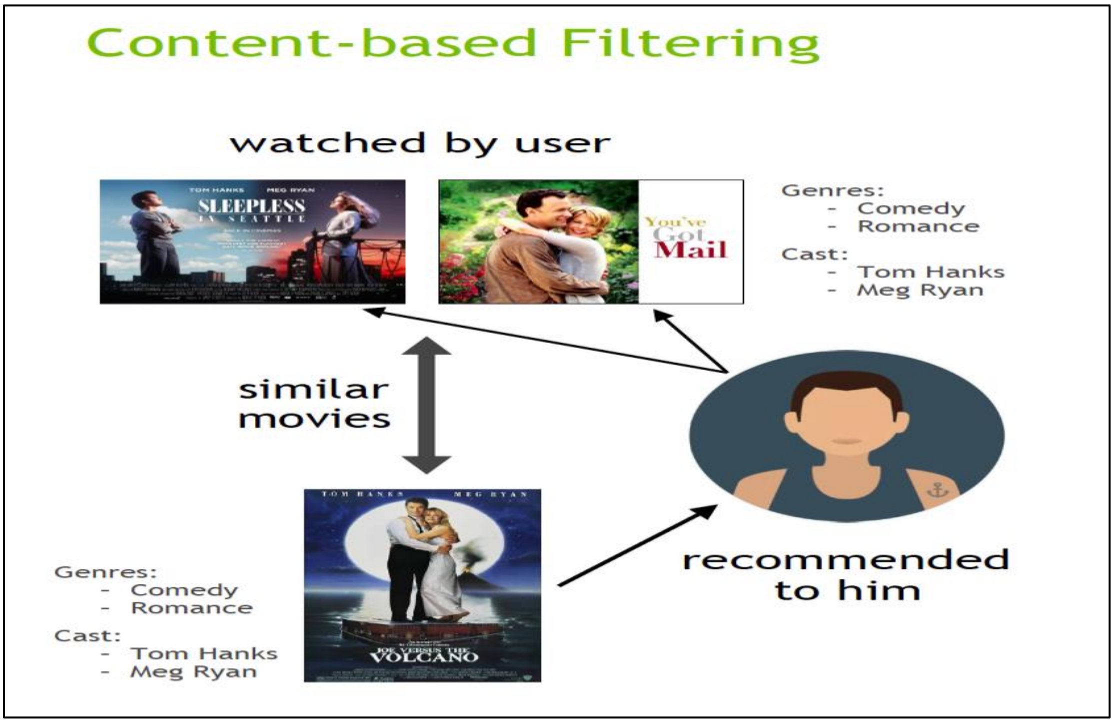


Fig 7.5: Content-based filtering

7.4.3 Context filtering (hybrid system)

Context filtering involves incorporating users' contextual information into the recommendation process. For instance, Netflix presented an approach during NVIDIA GTC that frames recommendations as contextual sequence predictions. This method utilizes a sequence of contextual user actions, along with the current context, to forecast the likelihood of the next action. In the case of Netflix, they used sequences containing details like the user's country, device, date, and time when they watched a movie to train a model for predicting their next movie choice. This context-based approach aims to offer more accurate and relevant recommendations to users based on their unique preferences and behaviour in specific situations.

Hybrid recommender systems leverage the strengths of both collaborative filtering and content filtering approaches to build a more inclusive and effective recommendation system. By combining these two types of algorithms, hybrid systems can offer a more comprehensive and accurate set of recommendations to users based on their preferences, behaviour, and item attributes. The hybrid approach aims to overcome the limitations of individual methods and provide users with more diverse and personalized recommendations, enhancing their overall user experience.

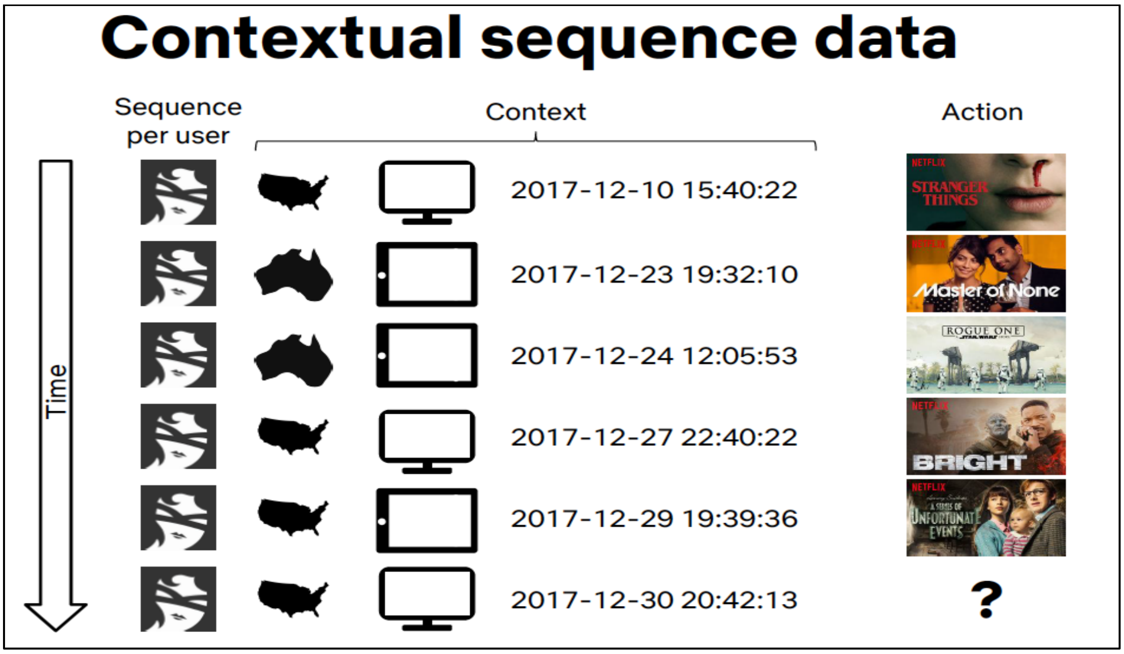


Fig 7.6: Context filtering

These algorithms are extremely useful for VoD service providers. Implementing effective recommendation systems can lead to several benefits for them:

- **Enhanced customer retention:** By continuously catering to customer preferences and providing personalized recommendations, VoD companies can foster loyalty and retain customers who feel understood and valued.
- **Increased sales:** Research indicates that accurate product recommendations can result in significant upselling revenue, with potential sales boosts ranging from 10% to 50%. Recommendation strategies like 'you might also like' suggestions, abandoned cart reminders, and highlighting popular purchases can drive sales growth. Same works for video and media industry too.
- **Influencing customer habits and trends:** Consistent delivery of relevant content through recommendations can trigger cues that build strong customer habits and influence usage patterns, encouraging repeat interactions and purchases.
- **Improved productivity:** Tailored suggestions for resources and materials can significantly speed up the work of analysts and researchers, saving up to 80% of their time, enabling them to focus on valuable insights.
- **Increased content:** Large VoD platforms with extensive inventories can utilize recommendation systems to suggest relevant products to customers, leading to increased content and improved user experience.

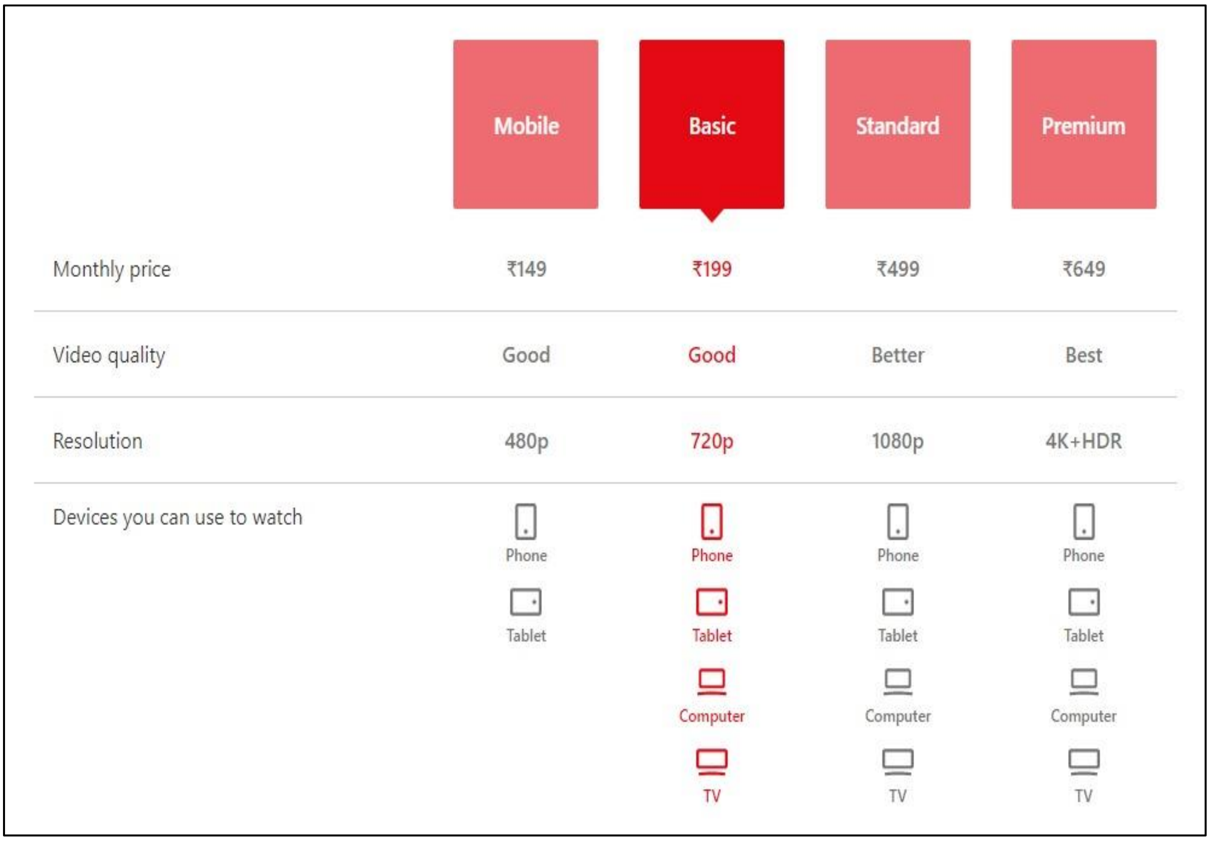
By using recommendation systems, VoD service providers can optimize their customer interactions, drive subscriptions, and improve overall efficiency, resulting in a more satisfied and engaged customer base.

8. VoD Monetization Models

Monetization models in VoD platforms are crucial for generating revenue and shaping user experiences. They include subscription-based, transactional (pay-per-view, rentals), advertising-supported, and hybrid/freemium models. This study compares these strategies to understand their implications for content providers and viewers, helping VoD platforms create sustainable and engaging ecosystems. Selecting the right model is vital for navigating the evolving digital content landscape and fostering business success. The popular monetization models are as discussed below:

8.1 Subscription-based models

Subscription Video on Demand (SVOD): It is a VOD monetization model where users pay a recurring subscription fee to access a content provider's library. Examples include Netflix, Hotstar, Hulu, HBO, and Amazon Prime Video. With SVOD, users are not shown ads as they have already paid the subscription fee upfront, ensuring an ad-free viewing experience. The subscription can be billed monthly or yearly, granting users unlimited access to the platform's content throughout the subscription period.



	Mobile	Basic	Standard	Premium
Monthly price	₹149	₹199	₹499	₹649
Video quality	Good	Good	Better	Best
Resolution	480p	720p	1080p	4K+HDR
Devices you can use to watch	Phone Tablet	Phone Tablet Computer TV	Phone Tablet Computer TV	Phone Tablet Computer TV

Fig 8.1: Example of a SVOD pricing model - Netflix India

8.2 Transactional models

Transactional Video on Demand (TVOD): It is a distribution model where customers pay for individual VoD content. It involves two sub-categories: **electronic sell-through (EST)**, allowing permanent access to purchased content, and **download to rent (DTR)**, enabling limited-time access upon renting. TVOD services include platforms like Apple iTunes Store, Google Play Store, and VOD rental services provided by multichannel television providers (cable or satellite). Customers pay for each movie or TV show they watch, offering a flexible and on-demand viewing experience.

Fig 8.2: An example of TVOD – YouTube giving option of EST & DTR

8.3 Advertising-supported models

Advertising Video On Demand (AVOD): It is a monetization model that offers free access to video content to viewers, supported by displaying advertisements during the viewing experience. This model allows content providers to generate revenue through advertising instead of charging a subscription fee. AVOD is commonly employed by streaming platforms that offer free content in exchange for showing ads to viewers. YouTube, Hulu, Dailymotion, etc. Are example of platforms using AVOD. YouTube deploys several types of ads on its videos.

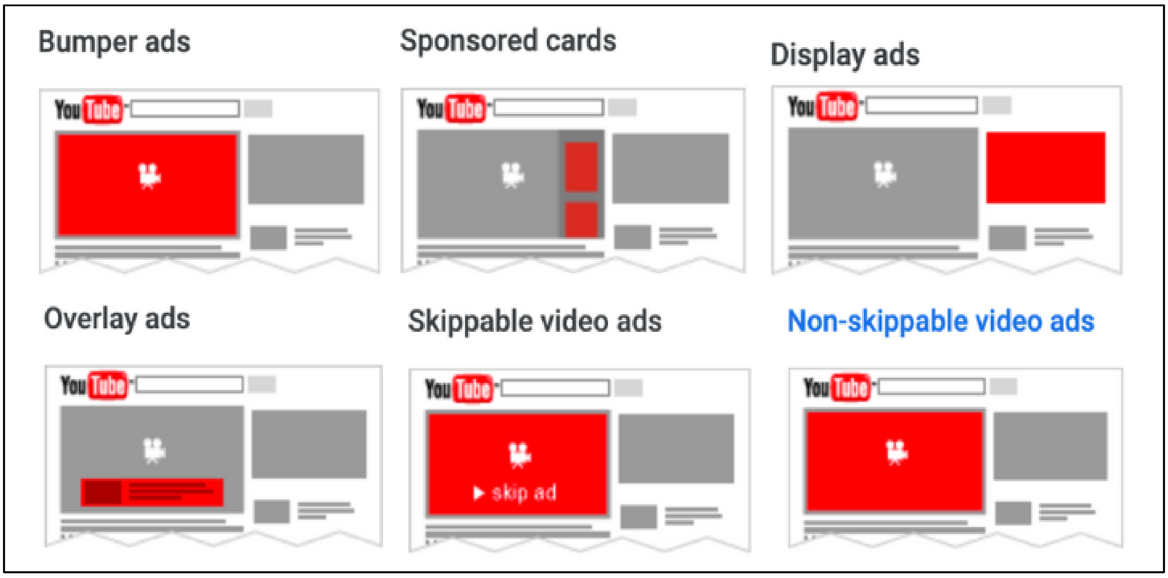


Fig 8.3: Types of ad YouTube displays on its videos

8.4 Premium models

Premium Video on Demand (PVOD): It is a hybrid model that offers users the option to pay a premium price for early access to specific content, either ahead of other SVOD or TVOD customers. PVOD can be compared to an online movie theatre, where viewers can pay an additional fee to watch a movie before it becomes available to the general SVOD subscriber pool. Disney's Mulan serves as an example of PVOD, where subscribers to Disney+ had the opportunity to watch the movie on its first-day release by paying an extra fee. However, they retained access to the movie until it was made available to all Disney+ subscribers.

Fig 8.4: Depiction of PVOD concept

8.5 Near video on demand model

Near Video on Demand (NVOD) is a broadcasting approach that enables television viewers to access a wide range of on-demand videos with a high degree of interactivity between the viewer and the content. Unlike traditional TV viewing, NVOD allows for instant access to recorded movies, video programs, games, and other services. The system offers convenient response times, ranging from seconds to hours, for program selection and content delivery. NVOD incorporates interactive advertising content during video interludes, allowing viewers to watch new movies at lower prices by accepting ad inserts in their subscription packages. Service providers can offer various price levels based on the number of commercials viewers are willing to tolerate, offsetting reduced billing with ad revenues earned. NVOD is commonly used by cable or satellite providers, allowing audiences to watch their preferred shows and films on different channels at various timings, resembling the pre-scheduling of movies in theatres. Viewers can choose their preferred start time, making NVOD comparable to the way movies are presented in theatres with multiple screening slots. NVOD is popular in those countries who lag in terms of good internet connectivity. DirecTV and Dish Network are examples of NVOD.

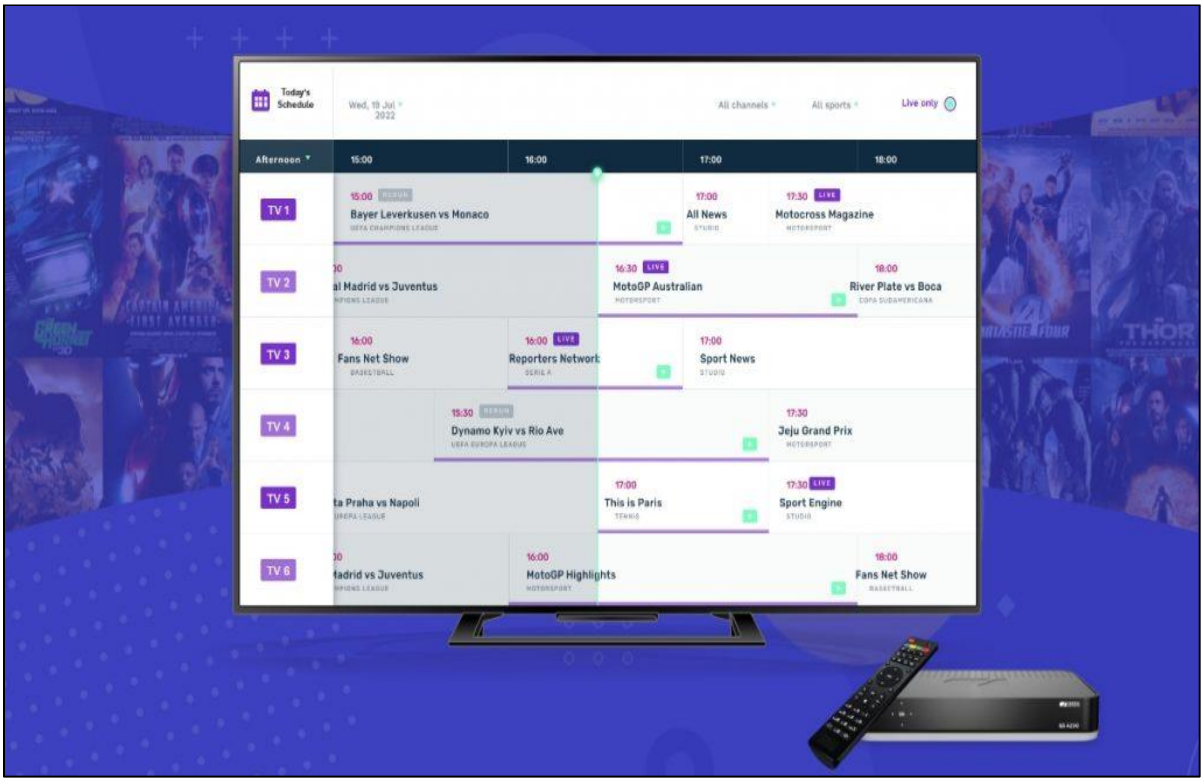


Fig 8.5: An example of NVOD

8.6 Hybrid and freemium models

Many popular streaming services employ hybrid models like the "Freemium Model," which combines AVOD and SVOD elements. The Freemium approach attracts users with free content, enticing them to become paying subscribers for an ad-free experience or access to premium content. An example from Disney+ Hotstar, shown in figure below, in India illustrates this hybrid model with two plans - Super and Premium. The Super plan includes advertisements, while the Premium plan offers an ad-free viewing experience. Another version of the freemium model allows users to access a portion of the content library for free, with the option to subscribe to unlock premium content. Some sports streaming services even provide delayed free streaming of games, encouraging users to pay for live access. The freemium VOD model effectively leverages both advertising and subscription revenues to cater to different user preferences and maximize monetization opportunities.

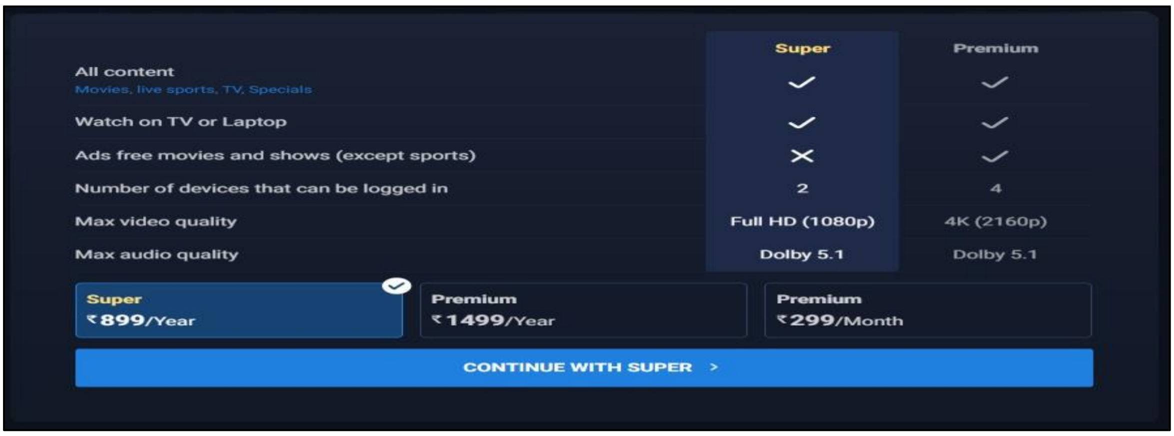


Fig 8.6: Disney+Hotstar freemium model

9. VoD and other video content platforms

In today's digital age, video consumption has undergone a significant transformation with the emergence of three popular content delivery methods: Over-the-Top (OTT), Video on Demand (VoD), and Live Streaming. These approaches have revolutionized video access, offering convenience and a diverse range of entertainment options.

9.1 OTT and Live streaming

Over-The-Top (OTT): OTT refers to the distribution of video content over the internet directly to viewers, bypassing traditional cable or satellite TV providers. OTT platforms deliver content through the internet to various devices, such as smartphones, tablets, smart TVs, and computers. These platforms offer both live streaming and Video on Demand services. OTT services allow users to access content on-demand and often provide personalized recommendations based on user preferences. Examples of OTT platforms include Hulu, YouTube, and HBO Max.

Live Streaming: Live streaming involves the real-time broadcasting of video content over the internet as it happens. It enables viewers to watch events, shows, or presentations as they occur, providing a sense of immediacy and interactivity. Live streaming is commonly used for events like sports matches, concerts, conferences, and live news coverage. Unlike VoD and OTT, where content is pre-recorded, live streaming requires a continuous feed of data in real-time. Platforms like Twitch, YouTube Live, and Facebook Live are popular for live streaming.

9.2 VoD Integration

VoD services can be integrated with platforms like live streaming and social media to enrich user experience and expand VoD library. The integration of live streaming can give user flexibility to access live events and highlights and the integration with social media can open new methods of content searching and discovery. It empowers user with social sharing features, user-generated content, and crowd-sourced recommendations.

9.2.1 VoD and Live streaming integration

Incorporating live events into VoD platforms offers numerous opportunities for content enhancement and organization. Live events can be recorded in their entirety and made available as VoD content, empowering viewers to access the event at their convenience. Platforms also have the flexibility to trim the content, removing any unnecessary segments from the beginning or end of the transmission. Moreover, during the live event, highlights can be generated to share clips or important moments with the audience. These highlights are treated as separate VoD content, allowing platforms to create as many clips as needed from the live stream.

To further enhance content organization, live events can be divided into smaller pieces known as chapters. Each chapter serves as a distinct VoD content unit, enabling better content storage and consumption organization once the event concludes. This approach ensures that viewers can easily access and enjoy specific sections of the live event independently, enhancing their overall viewing experience.

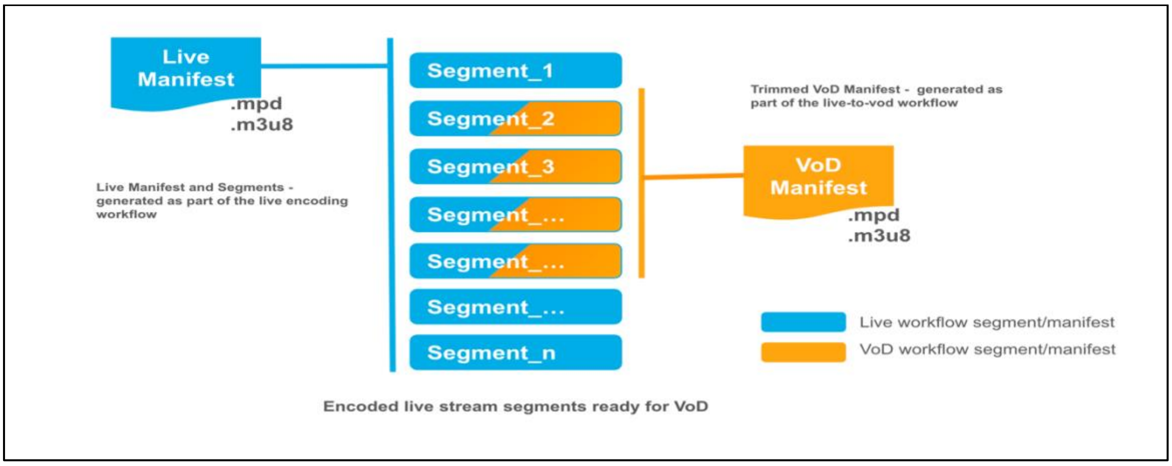


Fig 9.1: Bitmovin Live-to-VoD workflow

9.2.2 VoD and Social Media Integration

VoD platforms offer a range of features to facilitate social sharing and engagement, such as buttons and features to easily share videos with friends and followers. User-generated content is also encouraged through social media integration, which allows users to review, comment, and rate content to help other users discover latest content. Additionally, social interactions can provide insights into trending and popular videos, which can influence other users' viewing choices. Content providers can use social media to promote their VoD content with sponsored posts and targeted advertising, as well as influencer marketing, to drive traffic to their platform. Incorporating social media into a VoD platform provides users with the ability to interact in real-time with content creators and influencers, as well as viewers. This can be achieved through the organization of live Q&A sessions and watch parties, as well as interactive polls, which can be used to increase user engagement during live stream events or the launch of latest content. VoD platforms can provide social login options, enabling users to sign in with their social media credentials, thus simplifying the registration process and providing a seamless user experience. Data insights and analytics can be gained from social media integration, allowing for the optimization of content strategies, optimization of marketing campaigns, and the delivery of personalized recommendations to users.

Fig 9.2: Social media integration

10. VoD Market Analysis

The Video-on-Demand (VoD) market is projected to experience substantial growth, increasing from USD 104.42 billion in 2023 to USD 173.27 billion by 2028, with a CAGR of 10.66% during the forecast period (2023-2028) as per Mordor Intelligence's report.

This growth is driven by factors such as seamless connectivity, widespread use of mobile devices for content consumption, and advanced smartphone capabilities. The rise of Over-the-Top (OTT) platforms like Netflix and Amazon Prime has contributed to the popularity of VoD services, supported by the accessibility of cloud platforms. The demand for digital media devices, faster internet access, and the growing adoption of mobile phones, fuelled by the popularity of social media platforms, are significant drivers of VoD market expansion. The pandemic further boosted the market, as viewers turned to streaming services for entertainment. Premium video advertising has also experienced growth, with viewers embracing ad-supported digital video content. However, concerns about content piracy and protection pose challenges to market growth and may lead to revenue loss.

The COVID-19 pandemic and subsequent lockdowns had a positive impact on the VoD industry, with an increase in streaming among consumers worldwide. Advanced networking technologies are expected to support the industry's expansion post-pandemic.

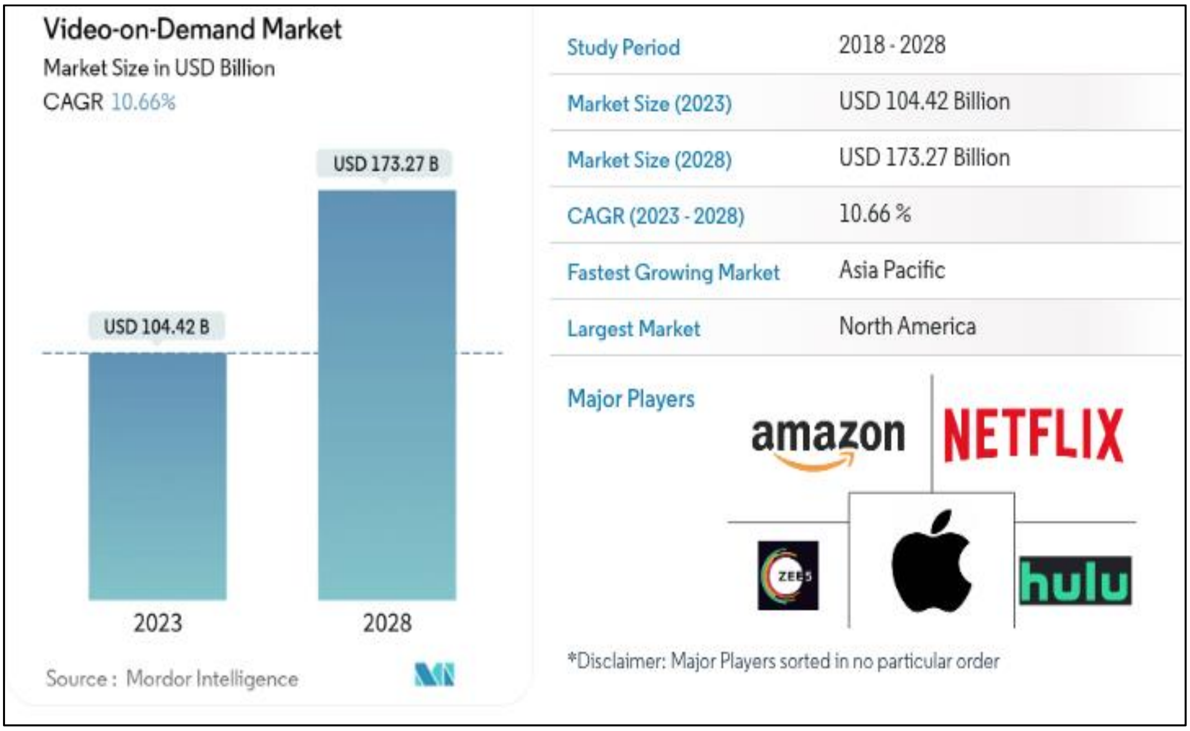


Fig 10.1: VoD market

10.1 Market trends

10.1.1 Market expansion because of increase in number of mobile internet users

The rise of binge-watching on streaming platforms has transformed the viewing experience for customers, with mobile devices playing a pivotal role in providing convenient access to content anywhere and anytime. The increasing availability of high-speed internet and widespread usage of smartphones, tablets, and PCs have fuelled instant satisfaction among users, allowing them to stay updated with the digital world without waiting.

Mobile data traffic has seen exponential growth, driven by data-capable devices and high-bandwidth applications like gaming and video streaming. More than half of the global population used mobile internet by the end of 2021, with a significant rise in low- and middle-income nations (LMICs). The adoption of 5G networks by operators is further boosting mobile-based internet usage. Erricson predicts a surge in 5G subscriptions by 2025, covering a substantial portion of the world's population and generating a significant share of mobile data traffic. The World Advertising Research Centre (WARC) projects that over 1.3 billion people will access the internet through smartphones and PCs by 2025.India and China lead in mobile internet monthly usage, driven by attractive data plans from service providers and changing video viewing habits among millennials. The growing number of smartphone subscriptions and increased data consumption per subscription contribute to the rapid growth in internet traffic.

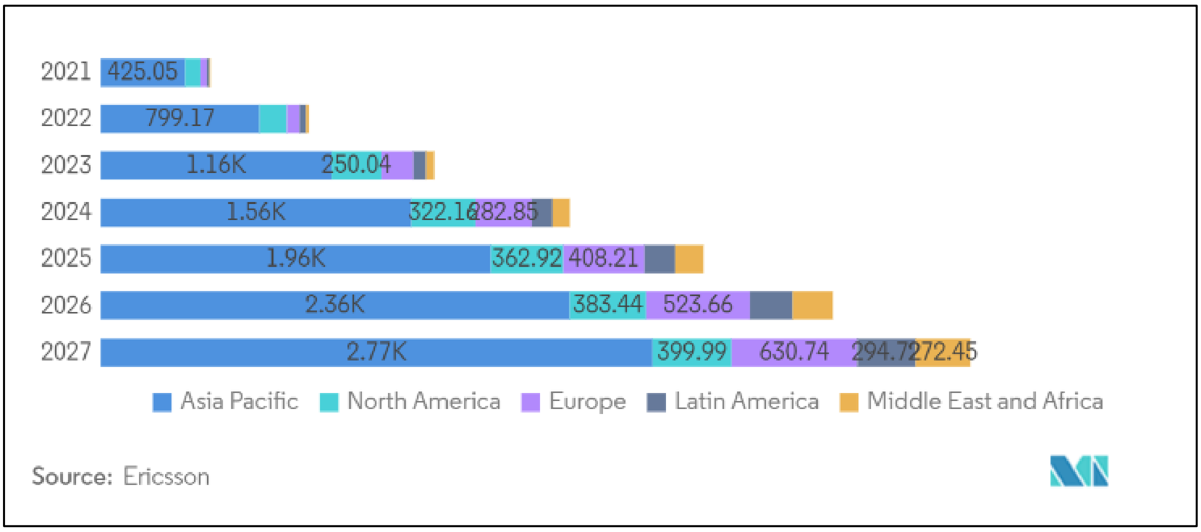


Fig 10.2: Forecast number of 5G users worldwide in millions

10.1.2 North America – the biggest market

The North American region, comprising the United States and Canada, has emerged as a significant market for Video on Demand (VoD) services. The United States, being the largest economy globally, has been a key driver of VoD adoption in the region. Factors such as flexibility, comfort, content personalization, diverse content availability, and the sheer volume of content have contributed to the popularity of VoD services in the region. According to nScreenMedia, the video streaming market in the United States has grown into a USD 16 billion industry over the past nine years, with significant impact from transactional and rental video markets. The country is witnessed transition from 4G to 5G, which led to increased smartphone transmission speeds, enabling higher-resolution videos and enhancing consumer experiences.

The COVID-19 pandemic further accelerated the demand for VoD services, as a substantial portion of the population stayed at home to curb the spread of the virus. Content producers have turned to video streaming platforms as an alternative to cope with theatre closures, leading to the launch of new services like AFN Now in November 2022.

In Canada, exciting growth opportunities for VoD services are fuelled by increased mobility trends, widespread internet proliferation, the rise of online streaming services, and collaborative efforts between content producers and streaming partners. These factors contribute to the rising demand for VoD services in the region.

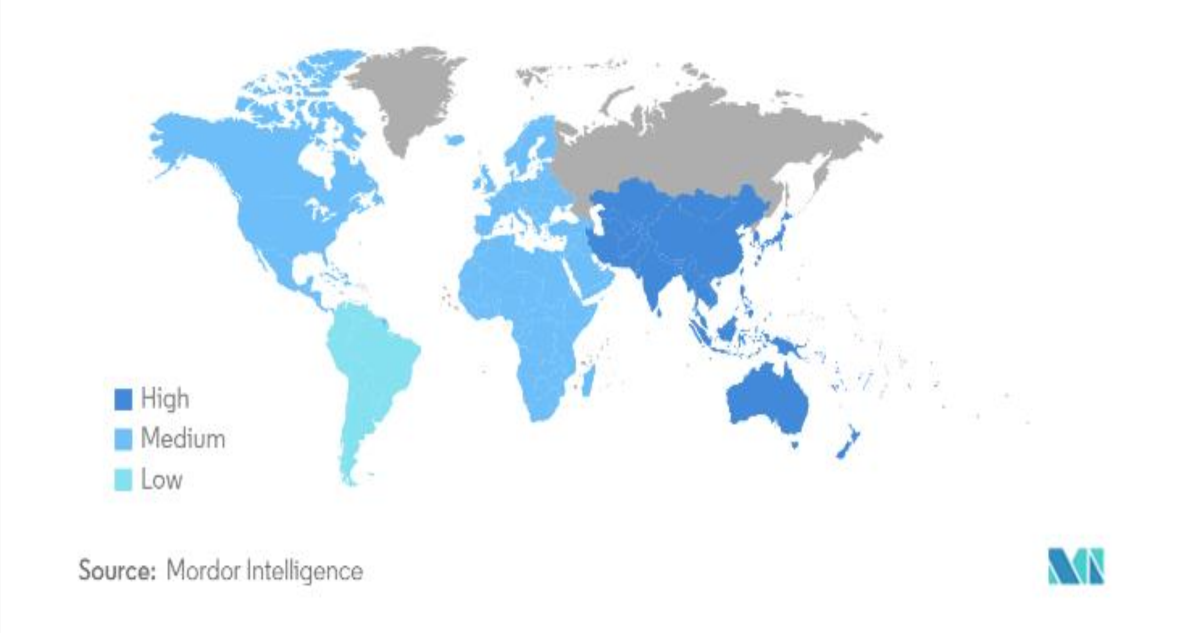


Fig 10.3: VoD market growth rate

10.2 Market dynamics

The VoD market has undergone remarkable growth and changes in recent times, influenced by various crucial factors. These include advancements in technology, shifting consumer preferences, the abundance of content, and the evolution of business models. Following are the detailed discussion on market dynamics:

Advancements in technology: The expansion of digital infrastructure, fast internet connectivity, and advancements in streaming technologies have played a vital role in fostering the growth of the VoD market. Enhanced video compression methods, CDNs, and adaptive streaming technologies have facilitated seamless access and streaming of high-quality video content across multiple devices.

Change in consumer choice: The demand for on-demand content has surged as consumer behaviour and preferences have changed. Users now prioritize convenience, flexibility, and personalized experiences, which VoD platforms provide through a wide selection of content accessible on various devices. The proliferation of mobile devices and smart TVs has contributed to the rapid adoption of VoD services.

Expansion in content library: The diverse selection of content, ranging from movies and TV shows to original series and user-generated content, has played a crucial role in fuelling the expansion of the VoD market. Leading platforms like Netflix, Amazon Prime Video, Disney+, and Hulu invest significantly in creating and obtaining content to appeal to and retain their audience. Moreover, the rise of specialized streaming services catering to specific genres or target groups has further broadened the array of content choices for viewers.

Intelligent business models: The VoD market has experienced a transition in its business models, moving away from pay-per-view and rental models towards subscription-based services. Platforms like Netflix and Amazon Prime Video, operating under SVOD, offer unlimited access to content for a fixed monthly or annual fee. This model ensures a steady revenue stream for providers and enables viewers to enjoy content without individual transaction costs. Additionally, AVOD and TVOD platforms coexist, offering diverse monetization approaches.

Increased competition among service providers: The VoD market is experiencing intense competition, with both established and new players from diverse industries vying for market share. This has resulted in content bidding wars, higher investments in original productions, and a focus on exclusive content offerings to differentiate their services. Market consolidation is also evident, with mergers and acquisitions among VoD providers and traditional media companies aiming to strengthen their digital presence. In India, Disney+ has merged with Indian giant Hotstar, Jio Cinema has announced many free contents, which used to be premium earlier, in India to kill market.

Global content: VoD services are now reaching beyond their domestic markets and are actively pursuing international audiences. To cater to diverse markets, they are employing localization strategies such as multilingual content, subtitles, and dubbing. Additionally, collaborations with local content creators and production studios are being fostered to enrich their regional content offerings.

Changes in regulations: The expansion of the VoD market has brought about regulatory complexities in different regions. Governments and regulatory bodies are dealing with issues like content censorship, taxation, licensing, and copyright enforcement. Diverse regulations across countries pose challenges for global VoD providers operating in multiple jurisdictions.

10.3 Key market players

- Netflix (U.S.)
- Amazon Inc., (U.S.)
- Google Inc., (U.S.)
- Akamai Technologies (U.S.)
- YouTube (U.S.)
- Apple Inc., (U.S.)
- Home Box Office, Inc. (U.S.)
- Cisco Systems, Inc. (U.S.)
- Roku Inc., (U.S.)
- iNDIEFLIX Group Inc (U.S.)
- Fandango (U.S.)
- Hulu, LLC (U.S.)
- Comcast (U.S.)
- Huawei Technologies Co., Ltd. (China)

- Fujitsu (Japan)
- CenturyLink (U.S.)
- Muvi LLC (U.S.)
- Vubiquity, inc. (U.S.)

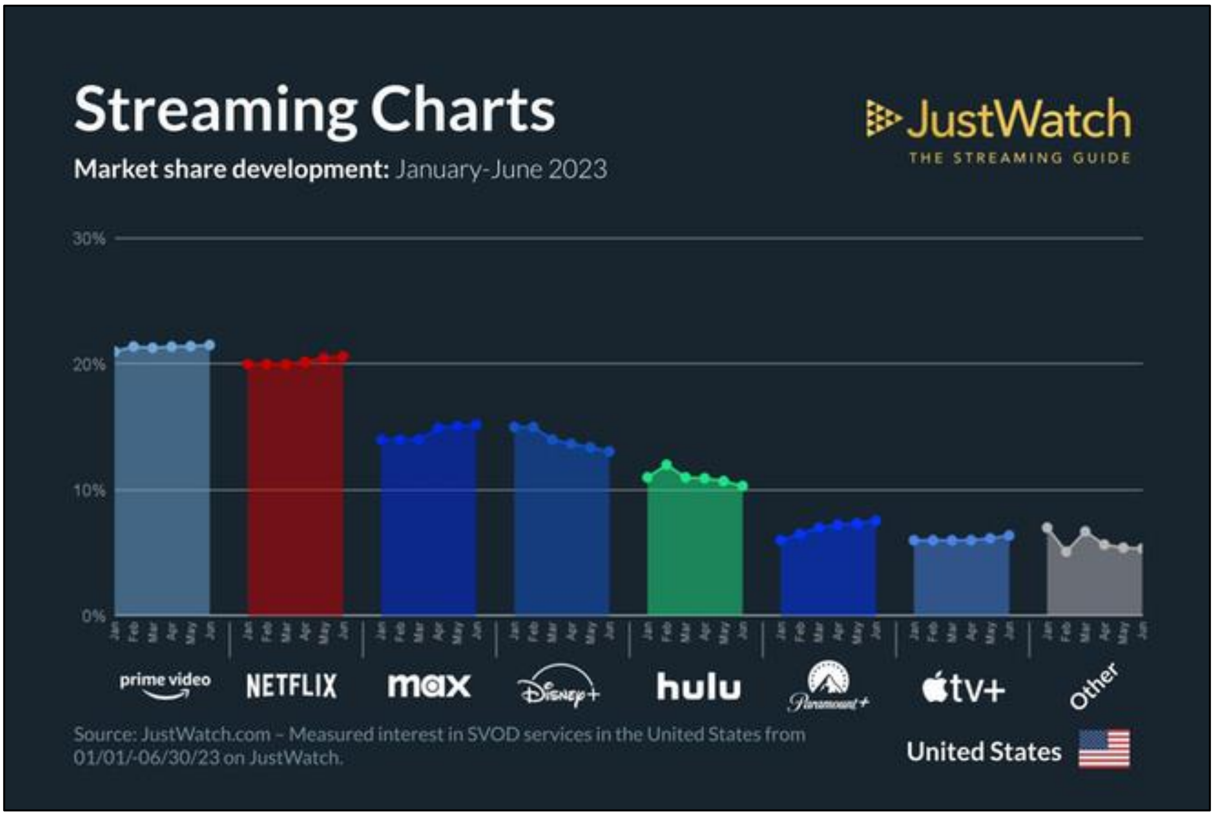


Fig 10.4: Market leaders

10.4 Challenges and limitations which can impact market negatively

- **Challenges related to licensing and distribution:** Obtaining content licenses and distribution rights can be intricate and costly for VOD providers. The competition for sought-after content often results in high licensing expenses, limiting its availability on certain platforms. Managing licensing agreements across regions and adhering to copyright regulations also poses further challenges for these operators.
- **Copyright infringement and piracy issues:** VOD providers continuously encounter challenges related to content piracy and unauthorized distribution. Illegal streaming platforms and file-sharing networks pose a threat to the revenue streams of legitimate VOD services, leading to copyright infringement issues and piracy concerns that require ongoing management and mitigation efforts.
- **Infrastructure issues:** Providing high-quality video streaming in regions with limited internet infrastructure or slower connection speeds can be a challenge. Inadequate broadband coverage, network congestion, and limitations in rural areas may hinder the expansion of the VOD market. Addressing these issues is essential to ensure an optimal streaming experience for users.

- **Cost management:** The proliferation of VOD platforms can lead to subscription fatigue for users as managing multiple subscriptions becomes burdensome and costly. Balancing affordability and content availability becomes crucial for VOD providers to ensure a seamless and immersive experience for users.
- **Upholding net neutrality policy:** The absence of net neutrality could lead to ISPs creating fast lanes for certain VoD services, giving them a competitive advantage over others. This could result in a skewed market, limited consumer options, and potential harm to smaller or emerging VoD platforms that may struggle to afford preferential treatment from ISPs. By upholding net neutrality principles, the VoD industry can maintain an open and accessible environment for content distribution, ensuring that users have the freedom to access a diverse range of VoD services without interference or bias from ISPs.

11 Quality of service (QoS)

QoS in reference to VoD means measurement and management of various parameters to ensure the delivery of high-quality video content to users. Measurements such as latency, bandwidth, latency variation and error rate fall under QoS. VoD is a real-time traffic which needs high bandwidth and therefore, QoS is necessary.

11.1 Advantages of QoS

Few of the important advantages of QoS in any VoD system is described below,

- **Minimizing latency:** QoS plays a crucial role in minimizing latency, which refers to the delay in data transmission across network devices. It achieves this by prioritizing essential traffic and allocating enough bandwidth. By doing so, QoS ensures that time-sensitive applications such as real-time video streaming and VoIP calls encounter minimal delays, resulting in enhanced responsiveness and an improved overall user experience.
- **Reduction of jitter:** For real-time applications like VoD applications, to maintain consistent audio and video quality, minimizing jitter is crucial. Jitter refers to the variation in packet delay, and it can negatively impact the user experience by causing flickering monitors, audio and video gaps, and other disruptions. QoS can effectively manage and reduce jitter by ensuring consistent and predictable transmission rates for time-sensitive data, thus minimizing the occurrence of these issues for end users.
- **Prevents loss of packets:** In situations where network traffic is high or resources are limited, packet loss can occur, leading to compromised communication quality and data transmission disruptions. QoS mechanisms prioritize essential packets and allocate sufficient resources to mitigate packet loss. By effectively managing network congestion and ensuring reliable delivery, QoS plays a crucial role in maintaining the integrity of data transmission.
- **Provides security:** By implementing QoS measures such as traffic shaping and bandwidth allocation, network security can be enhanced. QoS prioritizes critical traffic, ensuring important applications receive sufficient resources and protection from potential network attacks. Moreover, QoS can effectively mitigate Distributed Denial of Service (DDoS) attacks by blocking or throttling malicious traffic. This proactive approach to network management enhances overall security and ensures a smoother user experience.
- **Improves bandwidth management:** Bandwidth is the capacity of a network communication system to transmit data within a specific timeframe, usually measured in bits per second (bps). It is essential to understand that bandwidth is different from speed. QoS is specifically

designed to enhance bandwidth by prioritizing and allocating resources to data packets that require faster delivery times. This optimization ensures efficient data transmission and improved network performance, catering to the varying needs of different applications and users.

- **Manages traffic congestion:** In situations where data traffic becomes overwhelming and the connection speed fluctuates, routers utilize QoS to address this challenge. QoS helps manage data packets by categorizing them into different queues. These service-specific queues prioritize certain packets over others instead of handling all packets in a single queue and sending them out sequentially. By implementing QoS and organizing the packets into separate queues based on their priorities, routers can optimize data transmission, minimize delays, and ensure smoother network performance, even during periods of heavy traffic.
- **Traffic shaping:** Traffic shaping is a technique employed to regulate the flow of network traffic and avoid congestion. It involves managing the transmission rate of packets, which allows certain types of traffic to take precedence over others, ensuring a consistent quality of service. By controlling the flow of data in this way, traffic shaping helps maintain stable and efficient network performance, ensuring that critical applications receive the necessary resources and preventing bottlenecks that could degrade the overall user experience.

11.2 Use cases of QoS in VoD

QoS can help achieve below things in any VoD system:

- By closely monitoring QoS metrics, VoD service providers can effectively maintain and assess the established quality standards for video streaming performance. This monitoring enables them to adhere to Service Level Agreements (SLAs) and ensure that the agreed-upon levels of service are consistently met. Through continuous evaluation of QoS metrics, providers can identify potential issues, address performance gaps, and deliver a seamless and satisfactory streaming experience to their users.
- Examining QoS metrics is valuable in detecting potential network problems that may impact video streaming quality, like network bottlenecks, latency challenges, or limited bandwidth. This data proves beneficial for enhancing network infrastructure and overall streaming performance, ultimately leading to an improved user experience.
- VoD service providers can use QoS metrics to improve the delivery quality of video content and gain valuable performance insights across diverse regions and platforms.
- QoS aids in optimizing content delivery strategies, including choosing the most effective CDN and optimizing caching and edge server placement.
- Applying QoS on both upload and download traffic can help VoD service providers to ensure low latency traffic and optimized network performance, which in turn, will generate more revenue.

12. Case study – Netflix

Netflix, founded in 1997, has emerged as one of the most influential players in the global entertainment industry. From its humble beginnings as a DVD rental service to becoming a leading subscription-based video streaming platform, Netflix has transformed the way people consume content. Netflix has 180M+ subscribers in 200+ countries. In this section, key system designs of Netflix will be covered.

12.1 HLS design

Netflix works on two clouds namely: AWS and Open Connect. Open Connect is Netflix's own CDN.

Client: User devices like smart tv, smartphones, laptops

Open client: Netflix's own CDN is used to distribute its content to nearest server to user's location. When any user wants to stream any video, it is served to user from nearest location rather than original server.

Backend: AWS takes care of distribution, processing, and network traffic management.

Frontend: Netflix's frontend is written in ReactJS.

12.2 Architecture and working

Netflix receives top-tier videos and content directly from production houses. However, before delivering these videos to users, Netflix performs preprocessing to ensure optimal compatibility. With support for over 2200 devices, each requiring distinct resolutions and formats, Netflix adapts the content to meet the specific requirements of each device.

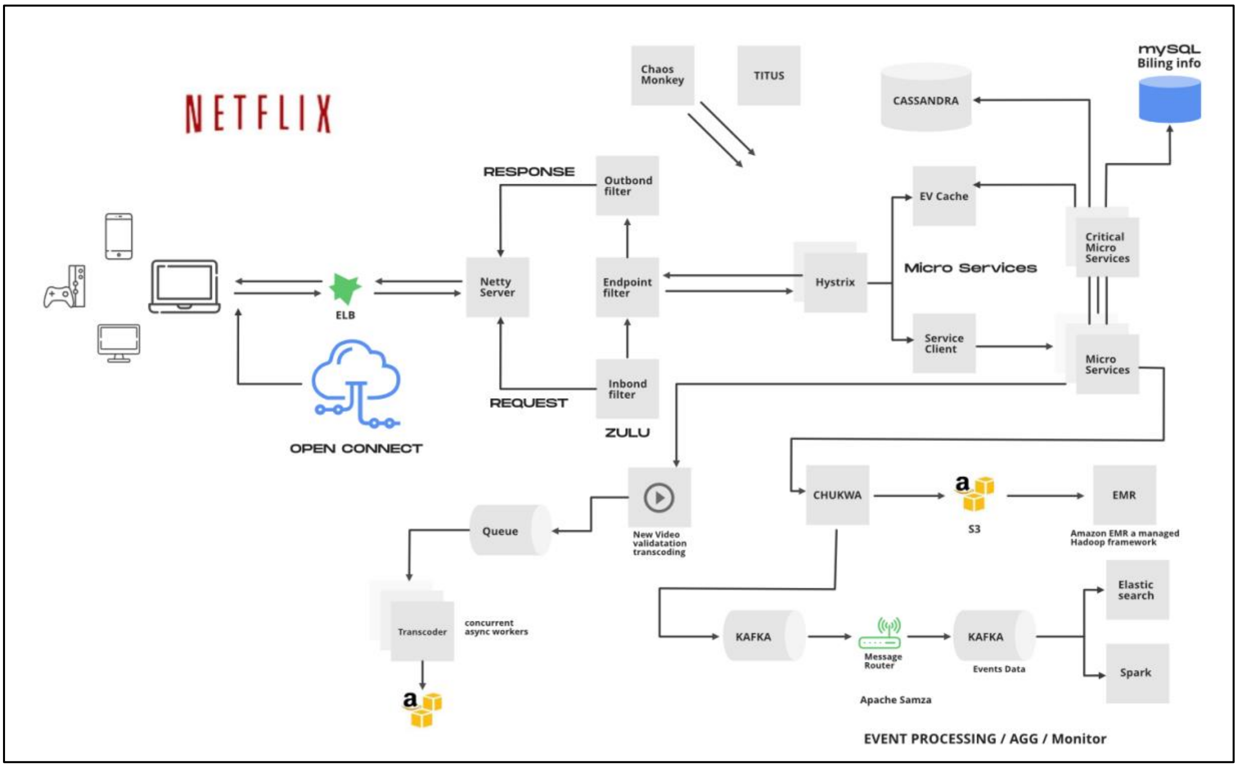


Fig 12.1: Netflix architecture

- **Elastic load balancer (ELB):** In Netflix, the Elastic Load Balancer (ELB) plays a crucial role in directing traffic to front-end services. It employs a two-tier load-balancing approach, starting with basic DNS-based Round Robin Balancing. When a request reaches the first load balancer, it evenly distributes the traffic across the designated zones using round-robin. The second tier consists of an array of load balancer instances, which further balances the requests among the instances within the same zone using the Round Robin Balancing technique. This ensures efficient distribution of workload across the servers, optimizing the overall performance.

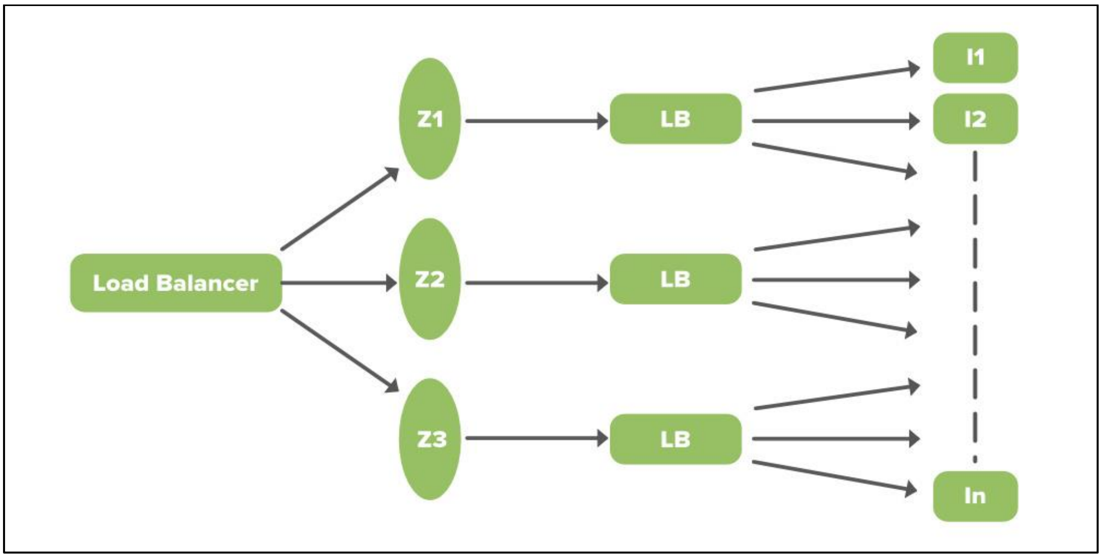


Fig 12.2: Netflix ELB

- **ZUUL:** ZUUL serves as a gateway service in the system, offering dynamic routing, monitoring, resiliency, and security capabilities. It efficiently handles various aspects of network protocols, web servers, connection management, and proxying tasks through the **Netty server**. When a request reaches the Netty server, it gets proxied to the **inbound filter**, which is responsible for tasks like authentication, routing, and request decoration. The request is then forwarded to the **endpoint filter**, which can either generate a static response or direct the request to the backend service (also known as the origin). Once the backend service processes the request and sends back the response, it goes through the **outbound filter**. This filter is responsible for tasks like content zipping, metric calculation, or manipulation of headers. Finally, the response is sent back to the Netty server and delivered to the client, completing the process.
- **Hystrix:** In a complex distributed system, the interdependencies between servers can lead to latency and system failures if one of the servers experiences issues. To address this challenge, the Hystrix library comes into play. Hystrix helps control interactions among distributed services by implementing latency and fault tolerance logic. By isolating access points between services, remote systems, and third-party libraries, Hystrix aims to prevent cascading failures in the system. It enables efficient handling of latency and failures from external dependencies, allowing for fast recovery and graceful degradation when possible. The library also supports near real-time monitoring, alerting, and operational control, making it a valuable tool in managing distributed systems. Additionally, Hystrix provides concurrency-aware request caching and automated batching through request collapsing.
- **Microservices:** Netflix's architectural style is based on the concept of microservices, where the system is composed of a collection of services. These microservices serve as the foundation for all the required APIs in applications and web apps. When a request is made to an endpoint, it calls upon other microservices to gather the necessary data. These microservices can also communicate with each other to retrieve data if needed. Once all the required data is gathered, a complete response is sent back to the endpoint. In a microservice architecture, the services are designed to be independent of each other. For instance, the video storage service is decoupled from the service responsible for transcoding videos. This decoupling allows for flexibility and scalability, making it easier to manage and develop individual services independently. Microservice architecture’s reliability is increased by using Hystrix and separating critical servers.

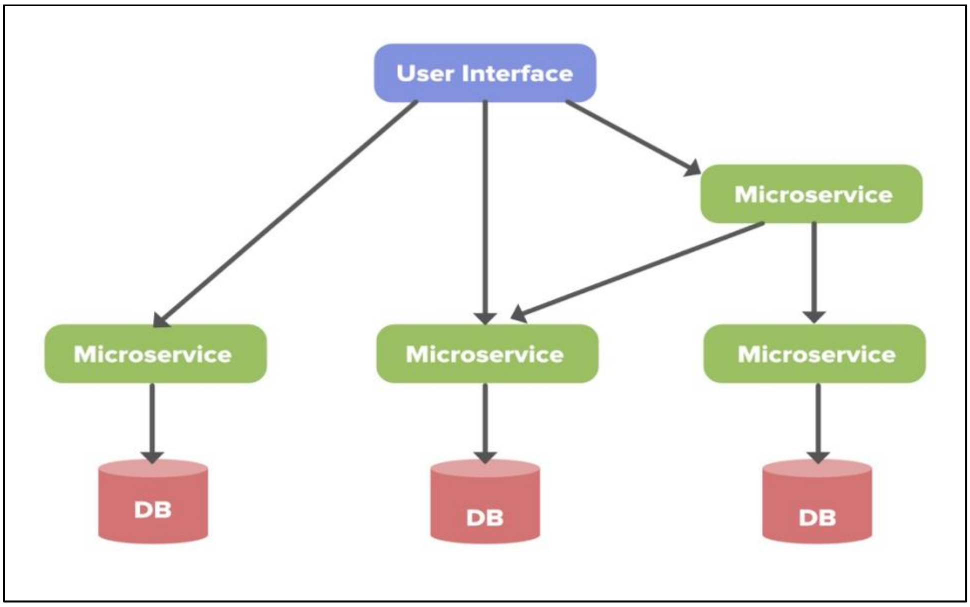


Fig 12.3: Microservice architecture

- **EV cache:** In many applications, frequently accessed data can be cached in various endpoints to improve response times. Netflix addresses this by utilizing its custom caching layer known as EV cache, which is built on top of Memcached. Multiple clusters of Memcached nodes and cache clients are deployed across numerous AWS EC2 instances. The data is shared within the same zone, and sharded nodes store multiple copies of the cache. When a write occurs, all nodes in all clusters are updated. However, for cache reads, the request is only sent to the nearest cluster and its nodes, not all clusters. If a node is unavailable, the read is directed to a different available node. This caching approach enhances application performance, availability, and reliability while reducing the load on the original server.

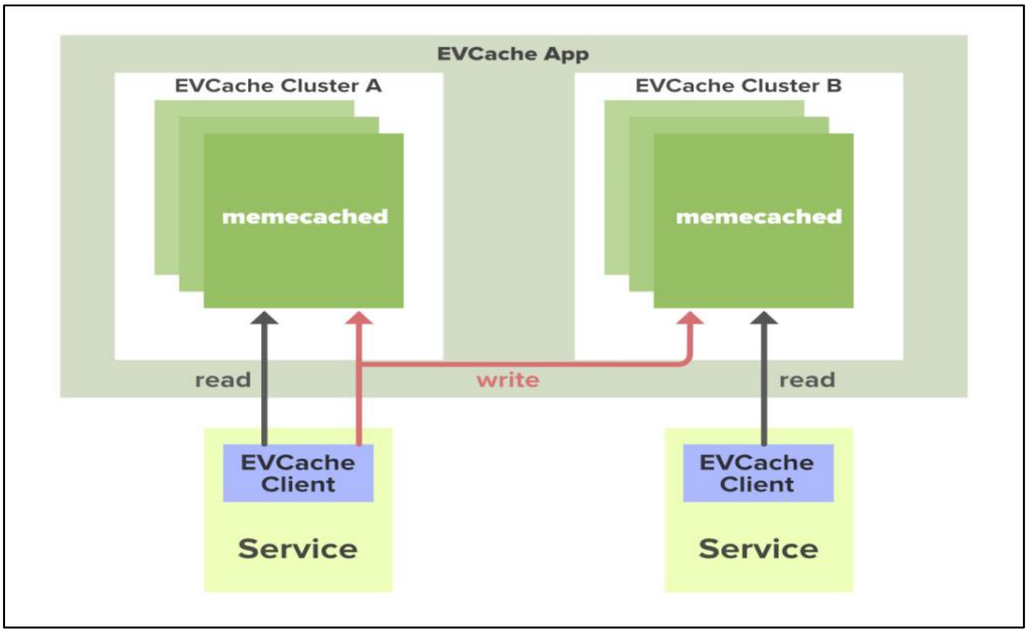


Fig 12.4: EV caching

- **Database:** Netflix employs MySQL (RDBMS) and Cassandra (NoSQL). MySQL is used to store information regarding user, billing, and transaction. Cassandra is used to store large viewing history.
- **Data processing:** Netflix utilizes Kafka and Apache Chukwa for data ingestion from different parts of the system, which includes nearly 500 billion data events consuming 1.3 PB/day and 8 million events consuming 24 GB/second during peak times. These events encompass error logs, UI activities, performance events, video viewing activities, and troubleshooting and diagnostic events. Apache Chukwa is an open-source data collection system built on HDFS and MapReduce framework. It offers Hadoop's scalability and robustness features and provides flexible toolkits for monitoring, displaying, and analysing results. Chukwa collects events from various system components, allowing monitoring, analysis, and event viewing through its dashboard. Events are written in the Hadoop file sequence format (S3), and the Big Data team processes these files into Hive in Parquet data format through batch processing at hourly or daily intervals. Chukwa also facilitates the upload of online events to EMR/S3 and provides traffic to Kafka for real-time data processing. Kafka is responsible for moving data from fronting Kafka to various sinks, including S3, Elasticsearch, and secondary Kafka. The routing of messages is achieved using the Apache Samza framework. Chukwa can send full or filtered streams, which may require further filtering on Kafka streams. To handle this, a router is employed to transfer data from one Kafka topic to another.
- **Elastic search:** In recent years, Netflix has experienced significant growth in its use of Elasticsearch, with approximately 150 clusters and 3,500 hosts running instances. Elasticsearch is employed for various purposes within Netflix, including data visualization, customer support, and error detection in the system. For instance, when customers encounter issues while playing a video, customer care executives utilize Elasticsearch to resolve the problem. The playback team accesses Elasticsearch to search for specific users and investigate why the video is not playing on their devices. This provides valuable insights into the events and information related to that particular user, helping to identify the cause of the video streaming error. Moreover, Elasticsearch is utilized by administrators to monitor resource usage and detect signup or login problems. It serves as a valuable tool for tracking important information and contributes to enhancing the overall performance and user experience on the platform.
- **Movie recommendation:** Netflix uses machine learning and Apache Spark for recommending movies to user. Apache spark also helps user to personalize content. Content filtering and collaborative filtering algorithms are used to build a system for recommendation. These algorithms are discussed in detail in CMS chapter.

13. Conclusion and future aspects

13.1 Conclusion

In conclusion, the research on VoD has highlighted the remarkable growth and transformation of this dynamic industry. VoD has become a dominant force in the digital entertainment landscape, offering viewers unprecedented flexibility, convenience, and personalized content experiences. The proliferation of streaming platforms, the availability of diverse content, and the rise of mobile devices have been instrumental in driving the adoption of VoD services worldwide. As consumer behaviour shifts towards on-demand content consumption, VoD providers continuously innovate to meet evolving preferences. The adoption of microservices architecture, caching mechanisms, and sophisticated QoS measures ensures smoother content delivery and improved user experiences. However, challenges remain, such as licensing complexities, content piracy, and network infrastructure limitations.

The research has shed light on architecture and topology used, different protocols employed, codecs and compression techniques, ABR. Content delivery is also discussed including ingestion, transcoding, storing, and caching mechanisms. After that content distribution strategies like edge caching, CDN, multi CDN, P2P are discussed. Scalability and its importance in VoD industry is covered including different techniques of scalability. VoD CMS is explained in deep including the topics which were within the scope of this research. VoD monetization models and their market examples are discussed. Other video content platforms are also discussed in brief and integration with VoD and its scope is laid out too. A detailed market analysis report is covered including market trends, dynamics and challenges. The vital role of QoS in maintaining consistent video streaming quality and optimizing network performance is covered. Technologies like EV cache, Hystrix, and Elasticsearch have proven instrumental in addressing these challenges and enhancing service reliability as in case of Netflix. The integration of VoD with social media, live streaming, and other monetization models further expands the scope of the industry and opens new avenues for engagement.

As the VoD market continues to evolve, it is crucial for providers to strike a balance between affordability and content availability, considering the rising concern of subscription fatigue. Localization efforts and international expansion strategies are essential for tapping into diverse markets and catering to global audiences effectively.

In summary, the VoD industry remains a captivating world of technological advancements, business innovation, and ever-changing consumer preferences. As streaming providers continue to navigate these dynamic waters, their commitment to delivering compelling content and enhancing user experiences will undoubtedly shape the future of digital entertainment for years to come.

13.2 Future aspects

The landscape of content consumption has been revolutionized by the emergence of VoD services. Platforms like Netflix, Hulu, and Amazon Prime have granted us unparalleled access to an extensive content library that can be enjoyed anytime, anywhere, and on various devices. However, the question remains: What lies ahead for the future of VoD services?

13.2.1 Integration of AI and ML

The Integration of artificial intelligence (AI) and machine learning is expected to revolutionize the landscape of streaming services. These advanced technologies enable personalized content recommendations based on our viewing habits, ensuring that we discover and enjoy content that aligns more closely with our interests and preferences. Furthermore, AI and machine learning can also be utilized to optimize video streaming quality. By analysing network conditions and user preferences in real-time, VoD services can dynamically adjust video resolution and bitrates to ensure a seamless and buffer-free viewing experience, even in fluctuating internet connectivity. Moreover, the future of VoD services might see increased collaboration between content creators and platforms.

13.2.2 Integration of VR/AR

The integration of augmented reality (AR) and virtual reality (VR) with VoD services creates lot of possibilities for transforming the way we consume content. With AR, users can experience an enhanced viewing experience by overlaying digital elements onto the real world. For instance, while watching historical documentaries, AR can superimpose relevant information, images, or animations to provide a more immersive understanding of the subject matter.

On the other hand, VR takes users to completely new virtual environments, making them feel as if they are part of the story. Imagine being able to explore fantastical worlds or attend live concerts from the comfort of your home through VR-enabled VoD platforms. This level of immersion can revolutionize entertainment, offering a deeper connection to the content and a more memorable viewing experience.

Furthermore, AR and VR integration can extend beyond content consumption to user engagement and interactivity. VoD platforms can leverage these technologies to enable virtual meet-ups, watch parties, or interactive games where users can socialize and engage with each other in shared virtual spaces. As AR and VR technologies continue to evolve and become more accessible, their integration with VoD services has the potential to revolutionize the entertainment industry. It will not only redefine the way we watch content but also enhance the overall user experience, making VoD platforms even more appealing and engaging to audiences worldwide.

13.2.3 Super aggregators

The rise of super aggregators in the VoD industry is set to reshape the landscape of content distribution and user experience. Super aggregators are platforms that bring together multiple streaming services and content offerings under a single umbrella, offering consumers a one-stop destination for all their entertainment needs. These super aggregators aim to simplify the content discovery process for users by providing a unified interface where they can browse and access a vast library of content from various streaming platforms. Instead of having to switch between multiple apps or subscriptions, users can conveniently find and watch their favourite movies, TV shows, documentaries, and original series from different providers on one platform.

For VoD providers, partnering with super aggregators can be beneficial as it creates new ways for reaching a wider audience and gaining exposure. By joining forces with these aggregators, content creators and streaming services can tap into a larger user base and gain access to potential subscribers who might not have discovered their content otherwise.

14. Acronyms

5G	5th Generation
AAC	Advanced audio coding
ABR	Adaptive bitrate
ADSL	Asymmetric digital subscriber line
AI	Artificial intelligence
AR	Augmented reality
ATM	Automated teller machine
AVC	Advanced video coding
AVOD	Advertising based video on demand
AWS	Amazon web services
bps	bits per second
CAGR	Compound annual growth rate
CCTV	Closed-circuit television
CDN	Content delivery network
CEO	Chief executive officer
CGS	Coarse grain SNR scalability
CMS	Content management system
Codec	coding decoding
CPU	Central processing unit
DASH	Dynamic adaptive streaming over HTTP
DB	Database
DCT	Discrete cosine transform
DRM	Digital rights management
DTR	Download to rent
DVD	Digital versatile disk
ELB	Elastic load balancer
EMR	Elastic map reduce
ES	Edge server
EST	Electronic sell through
EV	Eviction
FTL	Faster than light
FTP	File transfer protocol
GB	Giga Bytes
GOP	Group of pictures
HD	High definition
HDFS	Hadoop distributed file system
HLS	HTTP live streaming
HTML	Hypertext markup language
HTTP	Hypertext transfer protocol
ICL	International computers limited
IOPS	Input/Output operations per second
IP	Internet protocol
Kbps	Kilobits per second
LDAP	Lightweight directory access protocol
LMICs	Low and middle income countries
Mbps	Megabits per second

MCP	Motion compensated prediction
MGS	Medium grain SNR scalability
ML	Machine learning
MP3	MPEG audio layer 3
MPEG	Moving picture experts group
NAL	Network adaptation layer
NVOD	Near video on demand
OTT	Over the top
P2P	Peer to peer
PAR	Positive acknowledgment with retransmission
PB	Petabyte
PC	Personal computer
PoP	Point of presence
PVOD	Premium video on demand
PVR	Personal video recorder
Q&A	Question and answer
QoS	Quality of service
RAM	Random access memory
RDBMS	Relational database management system
RTMP	Real-time messaging protocol
RTSP	Real-time streaming protocol
SD	Standard definition
SIP	Session initiation protocol
SLA	Service level agreement
SNR	Sound to noise
SRT	Secure reliable transport
SSD	Solid state drive
SVC	Scalable video coding
SVOD	Subscription based video on demand
TB	Terabyte
TCP	Transmission control protocol
TV	Television
TVOD	Transaction based video on demand
UDP	User datagram protocol
UHD	Ultra-high definition
UI	User interface
UK	United Kingdom
USA	United States of America
VoD	Video on demand
VoIP	Voice over internet protocol
VR	Virtual reality
WARC	World advertising research center
WebRTC	Web real time communication

15. References

https://www.vplayed.com/blog/what-is-vod-how-video-on-demand-works/

https://www.dreamcast.in/blog/benefits-of-video-on-demand-streaming/

https://castr.com/blog/benefits-of-video-on-demand-streaming-services/

https://www.statista.com/outlook/dmo/digital-media/video-on-demand/video-streaming-svod/worldwide

https://hygraph.com/blog/video-streaming-architecture

https://ieeexplore.ieee.org/document/1613301

https://docs.oracle.com/cd/E19284-01/819-4439/acrjl/index.html#:~:text=In%20a%20central%20topology%2C%20most,1%20shows%20a%20central%20topology.

https://docs.oracle.com/cd/E19284-01/819-4439/acrjm/index.html

https://docs.oracle.com/cd/E19284-01/819-4439/6n6jehs99/index.html#fvybk

http://networks.ece.mcgill.ca/sites/default/files/thouin_IEEENetwork07.pdf

https://getstream.io/blog/streaming-protocols/

https://www.techtarget.com/searchunifiedcommunications/definition/codec

https://getstream.io/glossary/adaptive-bitrate-streaming/

https://restream.io/blog/streaming-protocols/

https://www.linkedin.com/pulse/content-ingestion-video-streaming-navigating-anton-gavrailov

https://www.researchgate.net/publication/283348434_Augmenting_Social_Multimedia_Semantic_Interaction_Through_Audio-Enhanced_Web-TV_Services

https://www.wowza.com/blog/what-is-transcoding-and-why-its-critical-for-streaming

https://aws.amazon.com/caching/

https://www.harperdb.io/post/edge-caching-explained-why-you-should-be-using-it

https://ieeexplore.ieee.org/document/8488647

What is a Content Delivery Network (CDN)? A youtube video by IBM technology. https://www.youtube.com/watch?v=Bsq5cKkS33I

https://www.ibm.com/topics/content-delivery-networks

https://dev.to/karanpratapsingh/system-design-content-delivery-network-cdn-bof

https://www.mlytics.com/blog/what-is-multi-cdn-and-do-i-really-need-it/

https://www.nec.com/en/global/techrep/journal/g06/n01/pdf/t060104.pdf

https://cloudcheckr.com/cloud-automation/horizontal-vertical-cloud-scaling/

https://afteracademy.com/blog/what-is-load-balancing-how-does-it-work/

https://www.researchgate.net/publication/221912506_Scalable_Video_Coding

https://blog.hubspot.com/website/video-content-management

https://experienceleague.adobe.com/docs/primetime/drm/overview/content-distribution-workflow/content-acquisition.html?lang=en

https://www.gumlet.com/learn/video-metadata/

analyticsinsight.net/everything-you-need-to-know-about-metadata-enrichment/

https://docs.oracle.com/en-us/iaas/data-catalog/using/enrich-metadata.htm

https://www.ibm.com/blog/metadata-enrichment-highly-scalable-data-classification-and-data-discovery/

https://help.liferay.com/hc/en-us/articles/360017873632-Tagging-and-Categorizing-Content

https://www.nvidia.com/en-us/glossary/data-science/recommendation-system/

https://ottverse.com/svod-avod-tvod-pvod-monetization-models/

https://www.vplayed.com/blog/near-video-on-demand/

https://cloudinary.com/guides/live-streaming-video/vod-streaming-versus-live-streaming-versus-ott-the-modern-video-economy

https://developer.bitmovin.com/encoding/docs/how-to-implement-a-live-to-vod-workflow-with-the-bitmovin-api

https://www.mordorintelligence.com/industry-reports/video-on-demand-market

https://www.enterpriseappstoday.com/news/video-on-demand-market-is-projected-to-grow-at-a-17-1-rate-through-the-forecast-period.html

https://www.linkedin.com/pulse/video-demand-vod-market-global-scenario-leading-players-ahirwar

https://www.mlytics.com/blog/qos-network-beginner-guide/

https://www.geeksforgeeks.org/system-design-netflix-a-complete-architecture/