Chapter 16
Video Transcoding Services in Cloud Computing Environment

Sampa Sahoo, Bibhudatta Sahoo and Ashok Kumar Turuk

Abstract Nowadays, online video consumption is an outstanding source of infotainment. Current social media era allows people to communicate with others around the world via Facebook, LinkedIn, YouTube and other platforms by sharing/sending photos, videos over the Internet. The proliferation of viewing platforms, file formats, and streaming technologies generate the need for video transcoding. The transcoding process ensures that video content can be consumed from any networks and devices, but it is a time-consuming, computation-intensive method and requires high storage capacity. The rise of video distribution and consumption makes the video service providers face unpredictable CAPEX and OPEX, for delivering more videos across multi-screens and networks. A cloud-based transcoding is used to overcome the limitations with on-premise video transcoding. The virtually unlimited resources of the cloud transcoding solution allow video service providers to pay as they use today, with the assurance of providing online support to handle unpredictable needs with lower cost. This chapter is designed to discuss various techniques related to cloud-based transcoding system. Various sections in this chapter also present the cloud-based video transcoding architecture, and performance metrics used to quantify cloud transcoding system.

16.1 Introduction

The Internet is now an important part of entertainment media, i.e., a user can watch a video of their choice or watch live events or matches through the Internet. The volume of media assets is increasing rapidly due to the growth of on-line viewing, social

S. Sahoo (✉) · B. Sahoo · A. K. Turuk
NIT Rourkela, Rourkela 769008, India
e-mail: sampaa2004@gmail.com

B. Sahoo
e-mail: bibhudatta.sahoo@gmail.com

A. K. Turuk
e-mail: akturuk@gmail.com

B. S. P. Mishra et al. (eds.), *Cloud Computing for Optimization: Foundations, Applications, and Challenges*, Studies in Big Data 39,
https://doi.org/10.1007/978-3-319-73676-1_16

media and mobile outlets. Current social media era allows people to communicate with others around the world via Facebook, LinkedIn, YouTube and other platforms by sharing/sending photos, videos over the internet. The variation in video quality, file sizes, and compression codecs makes the job of media professionals critical to maintaining it. Growth in other technologies like internet connectivity, increase in bandwidth put additional pressure. More recently, a significant new data modality has emerged due to unstructured data from video and images. A plethora of videos generated by digital devices demands attention in the current Big Data market. The video uploading rate of most popular online video archiving systems YouTube is around 100 video per minute. In recent time, one of the research challenges is analysis and processing of video data to support anytime anywhere viewing irrespective of devices, networks, etc. [1]. Video service providers use broadcast-quality video streaming services to reach local and worldwide audiences irrespective of networks and devices. Streaming is the process of fragmenting the video files into smaller pieces and delivered it to the destination. Streaming use buffering technology to collect several packets before the file being played. For example, Imagine a glass filled with water with a hole at the bottom, then there is a constant stream of water drainage as long as there is enough water in the glass. The streaming technology applies to both live streams and progressive downloads for audio and video on demand. Streaming can be done at the streaming server or by renting streaming service provided by streaming service providers who can host the video on the cloud. Streaming service can be rented on an hourly or monthly basis, i.e., the user needs to pay only for the resources consumed.

Cloud computing is used to provide ubiquitous, on-demand network access such that user can access computing resources anywhere and at any time through the Internet. It is built on the base of distributed computing, grid computing, and virtualization, which delivers the on-demand computing resources over the Internet on a pay-for-use basis. Users can access the resources without considering the installation details and use it as required with paying only for the used units. The cloud service model can be Software as a service (SaaS), Platform as a service (PaaS), Infrastructure as a service (IaaS). SaaS is a subscription model to access software running on Service Providers servers. Few examples of SaaS applications are Google Apps, Box, Dropbox. PaaS provides a computing platform for development and deployment of web applications without buying and maintaining software and infrastructure required for it. Examples of PaaS include Google App Engine, Microsoft Azure Services, Force.com, etc. Infrastructure resources like storage, network, and memory are delivered as on-demand service in the IaaS cloud service model. Examples of IaaS providers are Amazon Web Services, Rackspace, etc. [2–6]. A virtualization technique used by cloud computing allows splitting of coarse-grained physical resources into fine-grained virtual resources to be allocated on-demand [7]. The cloud-based video delivery architecture allows storage and distribution of single, high quality, high bitrate video files in multiple formats without the expense of purchasing and maintaining own infrastructure [8]. In Video-on-Demand (VoD) cloud, the content provider rent resources (e.g. storage and bandwidth) from the cloud provider. The resources can be rescaled based on fluctuating demands of the user to satisfy some

Quality of Service (QoS) (e.g. video start-up delay). There are several advantages of moving video streaming services to the cloud and are listed as follows:

(i) The company converts upfront capital expenditure (CAPEX) to operating expense (OPEX). Cloud eliminates the massive capital investment for on-premise hardware (e.g. servers, storage arrays, networks) and software. It also puts an end to investment for continuously expanding and upgrading on-premise infrastructures.
(ii) In the cloud, a user pays as they go for processing intensive services such as encoding (transcoding), asset management and storage of video streaming. The payment can be paid on a transaction basis, monthly subscription or as an annual fee. For example, a file-based video content can be transcoded on an hourly, pay-as-you-go basis in the Amazon Web Services (AWS) marketplace.
(iii) To build a massively scalable encoding on-premise platform with support for latest devices and players is costly and not trivial. But, the infrastructure needed for this can be upgraded by the cloud provider without the knowledge of user very easily and with less effort.
(iv) A user can work from anywhere in the world at any time.
(v) The time-consuming uploads, downloads or inefficient bandwidth use is eliminated and thus making cloud time-efficient.
(vi) The cloud offers flexibility and scalability. On-premise scaling-up/down of resources is not hassle free sometimes. As resources can't be added on the fly content providers, need to start the whole process of buying and maintaining the new resources. Whereas in the cloud, the addition of new resources is simple and quick as the user only need to change its requirement details. The cloud service provider will accordingly either reduce or add resources for the user and charge for the same.
(vii) Whether Video delivery is on-demand or live streaming cloud ensures high quality and stability.

Different video streaming services used in practice are storage, transcoding, content delivery. This study mainly focuses on video transcoding service. The Cisco white paper [9] discusses why and how the cloud can support video delivery to multiscreen, i.e., more devices. London Olympic, 2012 is considered as a milestone that takes traditional viewing to a new level, i.e., shifting towards connected devices like tablets, smartphones, etc. for anytime anywhere viewing. The challenge lies in creating an effective multiscreen offer with the consideration of a different combination of devices, networks, service platforms, etc. The cloud architecture used for both homogeneous and heterogeneous environments could reduce the potential cost from 36 percent to 13 percent compared to traditional video architecture. One of the ways to manage the cost of multiscreen access is to allow a temporary bandwidth increase for premium consumers, who are ready to pay extra for the better experience. The paradigm shift will also help video service providers to reduce CAPEX/OPEX as well as coping with the growth of online video industry. According to statistics presented in [10], 462 million active Internet users are there in India, which is 34.8% of the whole population and 13.5% of the world Internet users. Despite a significant percentage of Internet users, the average connection speed in India is 3.6 Mbps, the lowest in the Asia-Pacific region. The average peak connection speed is also the lowest in India with 26.1 Mbps. In Asia-Pacific region, South Korea has the highest average connection speed, i.e., 27 Mbps, and average peak connection speed in Singapore is the highest with 157.3 Mbps [11]. From the report we can see that country-wise, there is variation in Internet speed. So, a single format of a video for all will not be sufficient. The advancement of mobile devices, tablets, PC support multiple formats, and it adds additional challenges to the video service providers. So, there is a need of video format conversion that will satisfy the demand for the various user devices and network. The format conversion process is known as transcoding.

An online video consumed by a user on multiple screens, such as digital TVs, smartphones, and tablets, need to be conveyed in a device suitable format. Video content providers require many formats of a single video file to provide service to users with varying need. It is practically impossible to prepare a video in all formats as it requires large content storages. There is also continuous development in the field of coding and encoding technology, codecs, etc. So, there is a need for a solution that will convert a video into the required format with less effort and cost [12, 13]. The conversion of video from one digital format to another format termed as transcoding. Video transcoding method helps video content providers to generate all the possible formats of a single video [14]. To provide such a transcoding capability, video content providers need enormous computing powers and storage space. Video service providers want their videos to look good and playable irrespective of devices or platforms. The proliferation of video distribution and consumption makes the video service providers to face unpredictable CAPEX and OPEX, to deliver more videos across multi-screens. Video transcoding solution requires enormous computing power and must deal with issues related to cost, scalability, video quality, delivery flexibility, and ubiquity. Cloud computing has emerged as a front-runner to give a solution to time-consuming and resource-aware transcoding process.

Video transcoding was initially employed to reduce the video file size, but now the priority has changed, i.e., transcoding is not only used to reduce the file size but also make the video viewable across platforms, enable HTTP streaming and adaptive bitrate delivery. Transcoding may results degradation in video quality. So, it is desirable to start with a high-quality mezzanine file and carefully do the transcoding based on specific target formats and content types. Transcoding a large, high-definition video to a diverse set of screen sizes, bit rates, and quality requires a lot of time, computation, storage capacity. To overcome the difficulty associated with the transcoding process content providers are using the cloud services. The cumbersome transcoding is simple, take less time and pocket-friendly in the cloud as compared to in-house process. A user only needs to specify its requirements and subscribe the services provided by the cloud which is only a single click away. The rest of the task, i.e., resource allocation and time-consuming transcoding process will be performed in the background. Finally, the user is charged only for the resources consumed without much overhead. Before discussing transcoding in the cloud further first, we present various terms used in video transcoding in next Sect. 16.1.1.

16.1.1 Terms in Video Transcoding

A video consists of video sequences where each video sequence consists of Group of Pictures (GOPs). Each GOP has several video frames. Usually, transcoding is performed at GOP level or frame level. Few more terms related to video transcoding are

(i) Codec: The method used by a system to determine the amount of change between frames is called a codec. Codec stands for Compressor-Decompressor, and it either encode or decodes the video.

(ii) Bitrate (Data rate): It is the amount of data that is used for each second of video (Kbps, Mbps, etc.). Bit rate can be constant (CBR), i.e., the same amount of data every second or variable (VBR), i.e., the amount of data is adjusted depending on changes between frames.

(iii) Resolution: Resolution is the actual size of the video (1 frame) measured in pixels. For example 1920×1080 resolution $= 2{,}073{,}600$ pixels. Let each pixel uses 24 RGB color bit, then the size of one frame is $2{,}073{,}600\times24 = 0.25$ MB (1 MB = 8388608bit).

(iv) Frame rate: Number of frames shown every second is known as frame rate. Popular frame rates are 24 fps, 30 fps, 50/60 fps, etc. If 1 frame size is 0.25 MB then bandwidth (data rate) requirement of a video @60 fps is 15 MB/s, whereas @24 fps is 6 MB/s.

From the above calculation, we can see that even a video with few frames need a significant amount of bandwidth. For a movie or long duration video, it will be even more. A video is also demanding a substantial amount of storage. Different ways to deal with massive storage and bandwidth requirements are: buy and use infrastructure or convert video into a format (transcode) that will consume less storage and bandwidth. As transcoding is time-consuming and computationally intensive, cloud-based video conversion is preferable to reduce a provider's expense.

16.2 Video Transcoding Process

Transcoding (Encoding) is the process of changing the input video file from one format to another for video delivery to different programs and devices without losing originality [15]. Transcoding is commonly used as an umbrella term that covers some digital media tasks, such as transrating, trans-sizing [16]. Transcoding is typically the change of codec, bitrate or resolution. The change of bitrates, i.e., 5 Mbps to 3 Mbps or 1.8 Mbps, etc. are known as transrating, and change of resolution (video frame size), i.e., 1080p–720p is referred as trans-sizing. Transcoding involves following two steps: first decoding is done to convert input video file to uncompressed format and after that re-encoding is done to generate data format supported by the user's device [17]. Screen size 1080p racking up the pixel dimensions to 1920×1080 for

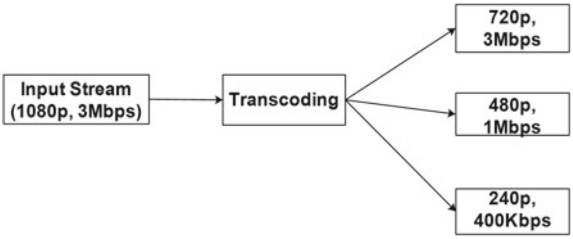


Fig. 16.1 Transcoding Process

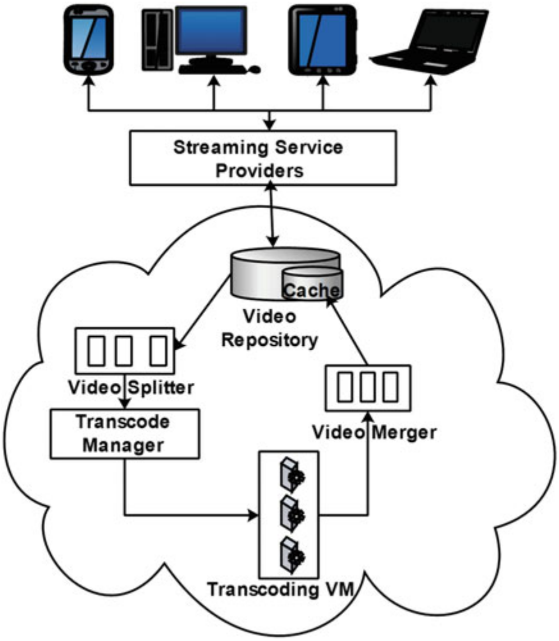
full high definition. Here p is for progressive scanning, meaning each line is drawn in sequence upon screen refresh. 3 Mbps represents bit rate. Figure 16.1 shows that how transcoding will generate different formats. The encoded video file can be delivered on-demand or live. The video file can be transferred entirely before playing it (downloading) or stream to the user device. Video content delivery depends on the distance between user location and media server containing the requested file. If the distance is less content delivery is fast, but if distance is more user experience choppiness, loading lag, poor quality. The advantage of transcoding in the cloud is lower cost, virtually unlimited scalability, and elasticity to counter peak demand in real-time. The cloud transcoding solution allows video service providers to pay-as-you-use, with the assurance of providing online support to handle unpredictable needs [18, 19]. Video transcoding service in cloud uses popular cloud service model that include Software as a Service (SaaS), Platform as a Service (PaaS), Infrastructure as a Service (IaaS), and hybrid model [20].

Video service providers want their videos to look good and playback irrespective of devices or platforms. The proliferation of video distribution and consumption makes the video service providers to face unpredictable CAPEX and OPEX, to deliver more videos across multi-screens (smartphones, PCs, TVs, tablets, etc.) and network (2/3/4G, mobile, broadcast, etc.). Encoding is not just compression; it means having to choose and accelerating, not declining as the number of renditions are needed to support the diversity of user devices and networks. The encoding solutions, whether in-house infrastructure, the third party need to deal with cost, scalability, video quality, delivery flexibility, and ubiquity issues. One of the solution to solve these problems is cloud-based transcoding. The advantage of transcoding in the cloud is lower cost, virtually unlimited scalability, and elasticity to counter peak demand in real-time. The cloud transcoding solution allows video service providers to pay as they use today, with the assurance of providing online support to handle unpredictable needs [21]. Operators, service providers, and content providers see the benefits of using standard servers in the cloud and want to move away from special appliances or dedicated hardware. Pushing video to the cloud, in real-time requires a high-speed, highly available network. The selection of cloud for any application depends on following primary requirements: bandwidth, storage, cost, security, and accessibility. The total bandwidth required for a video stream varies depending on the number of frames or images being captured per second as well as the quality of the images being captured. The availability and affordability of bandwidth are not consistent from city to city, country to country. Bandwidth and storage requirement calculation is an essential step in the planning or design of any cloud-based video delivery system.

16.3 Cloud-Based Video Transcoding Architecture

The system architecture of the cloud-based video transcoding service adopted from [13, 19, 22, 23] is shown in Fig. 16.2. The architecture consists of many components like a streaming service provider, a video repository, splitter, transcode manager, transcoding VMs (Servers), video merger and caching storage. The *streaming service providers* like YouTube, Netflix accepts user's request and checks if required video is present in *video repository* or not. If the video is present in its desired format, then starts streaming the video. However, if the coveted video is in another format than the one requested by the user, online transcoding is done using cloud resources. The service provider charged according to the amount of resource reserved in clouds. For online transcoding first, the video is split into several segments, or chunks by *video splitter*. The video segments are mapped to transcoding VMs by the transcode manager to be transcoded independently. The video segments can be of equal size, an equal number of frames, equal number of GOPs, equal-size with an odd number of intraframes or different size with an equal number of intraframes [24, 25].

Fig. 16.2 Cloud-based Transcoding Architecture



The *transcode manager* dispatches the transcoding jobs (video segments) to appropriate VMs. The goal of the manager's mapping of jobs to VMs is to satisfy some user related QoS like minimum startup delay, cost, etc. It is the responsibility of the manager to scale the capacity of the transcoding service up and down by adding and removing VM instances based on user demands.

The *transcoding VM* is used to transcode the source videos into targeted videos with desired video specification concerning format, resolution, quality, etc. with certain QoS constraints. Each transcoding VM is capable of processing one or more simultaneous transcoding task. There are two possibilities to do transcoding inside a VM. First, all segments of a video are transcoded in a single VM and second different parts of a video in different VMs simultaneously. Since the first approach is centralized user may have to wait for a video segment if it is not transcoded at the time of the request arrival. But it eliminates the overhead of maintaining different segments from the different VMs. The advantage of the second strategy is that always there is something to serve the user, as several segments are transcoded at a time. This method suffers if VM does not supply the asked segment at the time of the request.

Video Merger is used to place all the video streams in the right order to create the resulting transcoded stream. After the transcoding operation, a copy of the transcoded video is stored in the video repository to avoid repetition of the video transcoding process for further request of a video.

From the literature study, we found that video access rate follows a long tail distribution, i.e., only a few videos (i.e., Popular) are frequently accessed, whereas the user rarely streams, many other videos. All the possible forms of popular and frequently accessed videos are stored in cache storage. The unpopular video requested by the user is transcoded online and served to the user. Along with this essential component, some researchers used prediction module for transcoding on the fly to reduce waste of bandwidth and terminal power [23, 26, 27]. Xiaofei et al. use prefetching and social network aware transcoding [28]. Adnan et al. and Zixia et al. proposed stream-based and video proxy based transcoding architecture respectively [22, 29]. The overall process must be completed before the deadline, i.e., the delivery time of the video frame.

Figure 16.3 shows the overall working of a cloud-transcoding process. When a user request arrives at the streaming service provider the video repository invoked to check whether the video is present in required form or not. If the video is in inquired form, then it is directly served to the user. If required form is missing in the video repository, then online transcoding is implemented to produce demanded form. The first step of the online transcoding process is to break the video into several segments for easy processing. Video splitter is used to divide original video into several video chunks. These video chunks are forwarded to one of the transcoding VMs. The transcoding VMs convert this chunks into requested format. In the end, video merger is used to arrange the transcoded video chunks of a video and store them in the repository so that new request of the same video format can be directly served. As soon as all the operations are over, the streaming service provider start streaming the requested video.

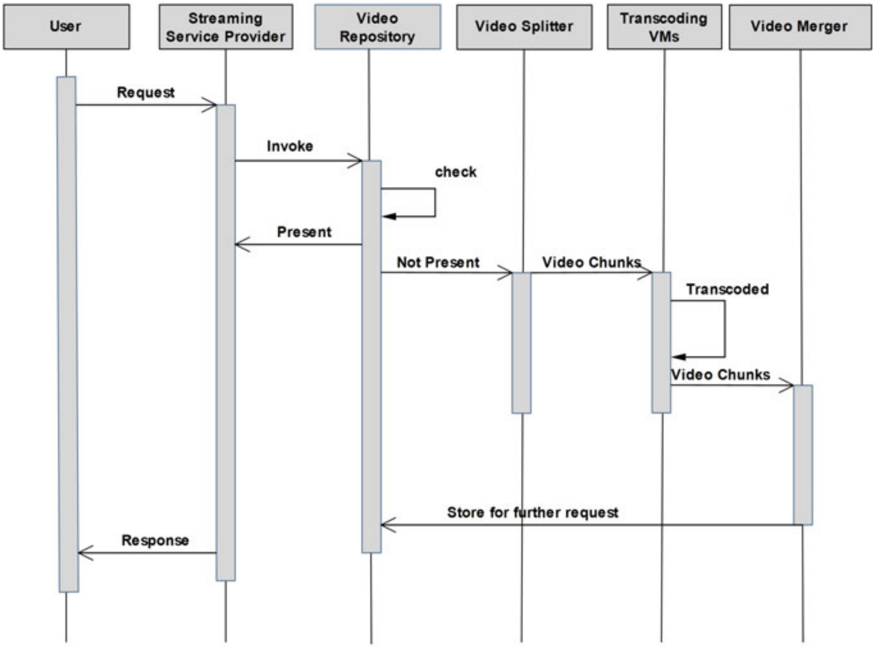


Fig. 16.3 Working of Cloud-based Transcoding Process

Table 16.1 Time in different phases of a transcoding process

Time	Definition
Reach Time (RT)	Time for a request to reach from user to the streaming server
Check Time (CT)	Time spent by the streaming server to check video repository for required video format
Split Time (ST)	Time to split the original video into several video chunks
Transcode Time (TT)	Time to transcode video segments into requested format
Merging Time (MT)	Time for arranging and storing transcoded segments of a video
Response Time (RT)	Time to give response to user by the streaming service provider

Table 16.1 shows time for different phases of online video transcoding process, starting from user request reach to a streaming service provider to response sent by the service provider. Let D be the deadline of a frame to be delivered then total time taken to transcode a frame must be less than or equal to D as shown in Eq. 16.1.

$$RT + CT + ST + TT + MT + RT \leq D \tag{16.1}$$

16.4 Video Transcoding Techniques in Cloud

The cumbersome transcoding is simple, take less time and pocket-friendly in the cloud as compared to in-house process. A user only needs to specify its requirements and subscribe the services provided by the cloud which is only a single click away. The rest of the task, i.e., resource allocation and time-consuming transcoding process will be performed in the background. Finally, the user is charged only for the resources consumed without much overhead. Following are different transcoding techniques in the cloud.

A. **Pre-transcoding**: One of the ways to deal with the video transcoding problem is pre-stored multi-version videos. This process stores various transcoded versions of each video to serve all user's requests and different types of client devices. The main drawback of this approach is the usage of large storage and computation resources. Even though pre-transcoding is a widely used method in VoD industry, but it induces a high cost. The increase in cost is due to provisioning and upgradations of fast-growing transcoding infrastructure, storage overhead to maintain all versions of all videos. It becomes cost prohibitive for small and medium size streaming service providers. The explosive growth of video streaming demands on a broad diversity of the client devices makes it ineffective [13, 30].

B. **On-demand Transcoding**: On-demand/real-time/online video transcoding performs conversion only when a particular format is not present in the storage. This method reduces the expenses of storage as it eliminates the need to store all versions. The on-demand resource reservation in the cloud makes the video transcoding process simpler and less costly. A leading content delivery network provider Akamai uses on-demand transcoding. Along with all the advantages, some challenges associated with this method are over/under-provisioning of cloud resources. The reserved resources must be future ready. Transcoder performs on-demand transcoding only for the segment that is not present in storage but requested by the user [30]. A QoS and cost aware on-demand transcoding of video streams using cloud resources presented in [13] with the aim of minimizing incurred cost of stream providers. Both on-demand and pre-transcoding are done to reduce the transcoding overhead. Pre-transcoding of video based on popularity based prediction is done into specific formats and resolution when the transcoding request falls (e.g. at night). A user is given several choices if its required format is not in local storage. If a user disagrees with the options, on-demand transcoding is done to generate user-specified form [26]. In [19] QoS-aware online video transcoding in the cloud is presented. Akamai advocates transcoding of multiple files in parallel for a fast turnaround time. Akamai the leading content delivery network (CDN) service provider of media and software distribution, uses stream packaging feature to detect the set of stream formats required by the user device and do the formatting and packaging of video streams at the network edge servers on-demand. The computation at edge server eliminates the shortcomings of pre-transcoding and centralized server transcoding process as it reduces additional storage cost and content management overhead [31]. Transcoding improves throughput (bandwidth) while preserving quality.

C. **Partial Transcoding**: In partial transcoding, some segments of a video are converted to other formats and stored in cloud storage. Based on the user viewing pattern rest segments of a video is transcoded on-demand. The purpose of doing so is to reduce operational cost, i.e., transcoding cost plus storage cost. Another approach is to store multiple versions of a popular segment and only one highest quality version for unpopular segments. If a user request for a segment that is not in the cloud storage, do the transcoding in real-time. Authors in [32] proposed partial transcoding scheme for content management in media cloud. Initially, a partial subset of video contents in different bitrates is stored in local storage that can be directly consumed. If the user specified format is not present in local storage, then online transcoding is done. The purpose of this approach is to reduce long-run overall cost. In [30] partial transcoding proposed based on user viewing pattern.

D. **In-network Transcoding**: Here, transcoding service can be placed on nodes inside the network, i.e., the introduction of a transcoding service into a network infrastructure. But, it requires routers that support this service processing. Further service placement of routers requires redesigning of the underlying network architecture. Thus, the practical implementation of this is not applicable to the existing network architecture [30].

E. **Bit-rate Reduction Transcoding**: A bitrate is the number of bits processed per unit time i.e. usually bits per second. The data rate for a video file is the bitrate. Example: bit rate of a standard definition DVD is 6 megabits per second (mbps) whereas video for phones is in kilobits per second (kbps). As the high rate video demands high network bandwidth, the video stream bit rate is reduced to ensure smooth streaming. This process is also known as transrating. Video segmentation is used to perform bit rate reduction transcoding in a distributed computing environment. The distributed transcoder do transcoding of different segments parallelly on several machines to speed-up transcoding process [25]. Cloud computing is an extension of distributed computing so the parallel bit-rate reduction transcoder can be implemented with or without any modification. This paper [13] introduces bit-rate, spatial and temporal resolution reduction transcoding techniques.

   Video data at 1920×1080 pixels, 60 frames per second means original frame size is 1920×1080 pixels and each pixel is sampled 60 times per second. Spatial resolution ascertains information related to each frame and temporal resolution defines the change between frames represented by frames per second. Example 1080 HD is a case of spatial resolution or containing more pixels, but 720 HD is a case of temporal resolution or containing more frames per second. Based on the reduction in the temporal or spatial domain following type of transcoding are possible.

F. **Spatial-Resolution Reduction Transcoding**: The spatial resolution indicates the encoded dimension size of a video. In spatial resolution reduction transcoding macro-blocks of an original video are removed or combined without sacrificing the video quality. Spatial resolution reduction transcoding produces an output video with a smaller resolution, but with same frame rate than the original video. In [24] spatial resolution reduction transcoding implemented in multi-core distributed system. Since the spatial resolution reduction method reduces the resolution, the transcoder has to deal with less number of bits. Since a virtual machine (VM) can have multiple cores, this process can be extended to the cloud environment.

G. **Temporal-Resolution Reduction Transcoding**: In a temporal resolution, transcoding frames are dropped to support frame rate of the client device and reduce the required transcoding time. To reduce the transcoding time a temporal resolution reduction transcoding is used for the cloud environment in [22]. Here the transcoder drops a small proportion of video frames from a video segment of continuous video content delivery to the user.
H. **Scalable Transcoding**: Scalable coding consists of a base layer (minimum bit-rate) and several enhanced layers (gradually increase the bit rate). Depending on the link capability of the user device one or more enhanced layer delivered along with a base layer. There is always something to play (i.e., base layer). In [28] an adaptive video streaming is proposed that can adjust the streaming flow with scalable video coding technique based on the feedback of link quality. This paper also discusses the advantage of this method like an efficient use of bandwidth. The video encoding strategy presented in Zencoder a cloud-based video encoding service, white paper [33] maintain a base format (e.g. MP4) that is playable on a broad range of devices. Then decide the number and type of renditions. For example, short duration (e.g. 1 min) videos can be easily downloadable, so only a few versions are sufficient. But long duration videos like movies cant be downloadable in a single go so need extra attention. As the user always expect a high-quality video, many renditions are required that can be used according to available network bandwidth and the user device.

Video transcoding can be performed in the client device or on-premise architecture or in any third party architecture like a cloud. The compute intensive and time-consuming transcoding job suffers from low processing power and energy sources of the client device (e.g. smart phone). So, it is not feasible to perform transcoding on client devices [13]. The in-house architecture suffers from scalability issues. Let there is infrastructure for a particular rendition sets. After sometimes a new set of renditions is required. To satisfy the new demand, the process of buying hardware and installing software will be time-consuming, cumbersome and costly. This process may also suffer from over/under-provisioning of resources which will have an adverse impact on the companys investment. Apart from economic inefficiency, there is also wastage of resources as most of the time servers are in an idle state [13, 19, 27]. Forecasting transcoding capacity is difficult due to the proliferation of delivery options, the unpredictability of the volume of video assets, a significant lead time required to acquire and prepare the necessary infrastructure [18]. To overcome the difficulties of in-house infrastructure to deal with the flash crowd, over/under-provisioning of resources, video service providers can look into cloud technology. The scalable cloud architecture allows on-demand resource reservation with less maintenance and cost. The compute intensive and time-consuming video transcoding can use on-demand cloud resources to reduce the expenses of video service providers [13, 19].

16.5 Performance Metrics

Performance metrics determined the benefit/lack of system designs. These are the metrics used to evaluate the working of a system in all conditions (favourable or unfavourable). The following section discusses various performance metrics used by researchers to assess their cloud transcoding system and listed in Table 16.2.

(i) Encoding Delay (Transcoding Time): It is the total time used to convert a video from one format to another.

(ii) Prefetching Delay: This is the time used to pre-fetch video segments for a user based on his/her social network service activity.

(iii) Watching (Click-to-play Delay): This is a time to wait for a user from clicking a video link to the first streaming arrival.

(iv) Transcoding Rate/Speed: It is the number of transcoding operation completed per unit time.

(v) Number of streams with Jitter (Transcoding Jitter): It is the measure of streams with jitter where jitter is the time difference in stream inter-arrival time.

(vi) Number of dropped frames: It is the count of frames being dropped to avoid deadline violation.

(vii) Queue waiting time: Waiting time of a stream in the queue to be encoded is known as queue waiting time.

(viii) Launch Latency: It is the time between the opening of a VM and its become ready to provide service.

(ix) End-to-End Delay: The end-to-end delay is the time between streaming provider starts to deliver streaming media to the time a user device starts playing it.

(x) Utilization Rate: It is defined as the ratio between used transcoding capacities to the total transcoding capacity of the system. Lower utilization rate indicates the larger idle time of the transcoding system.

(xi) Average Response Time: It is the difference between the time of transcoding request to the time of first transcoding response of the system.

(xii) Operational Cost: The overall operational cost is the combination of storage, computing and transcoding cost. The storage and computing cost is adopted from Amazon S3 and Amazon EC2 On-Demand instances respectively [32].

(xiii) VM Provisioning Cost (Cost): It is the cost to acquire resources in the cloud for transcoding process. To reduce the VM provisioning cost, the number of servers required to transcoding process should also be minimized [23].

Table 16.2 Performance metrics

Sl.No.	Reference no.	Performance metrics	Environment	Transcoding tool
1.	[27] [2015]	Cost	Cloud	Own Set-up
2.	[22] [2013]	Transcoding rate, Number of streams with Jitter, Queue waiting time	Cloud	Python with simpy framework
3.	[34] [2016]	Launch latency, End-to-End Delay, Utilization Rate, Average Response Time	Cloud	ffmpeg and Hi-cloud with Windows server 2008 operating system and Amazon EC2
4.	[32] [2015]	Operational Cost (storage cost + computing cost + transcoding cost)	Cloud	Own setup
5.	[29] [2011]	Transcoding Speed, Transcoding Jitters	Cloud	Own set-up
6.	[23] [2013]	VM Provisioning Cost	Cloud	Python with Simpy framework
7.	[26] [2012]	Average CPU Utilization, Cache Hit Rate, Data Transfer Rate	Cloud	ffmpeg
8.	[13] [2016]	Average startup delay, Average deadline miss-rate, Cost	Cloud	Cloudsim
9.	[28] [2013]	Pre-fetching delay, Watching delay, Encoding delay	Cloud	U-cloud server offered by Korean Telecom and JAVA based server application
10.	[19] [2017]	Transcoding time, Targeted Video Chunk Size	Cloud	Python, ffmpeg
11.	[35] [2014]	Energy	Cloud	Own set-up
12.	[30] [2017]	Storage and Transcoding Cost, Average Delay (startup)	Cloud	Own set-up

(xiv) Average CPU Utilization: It measures how efficiently CPU is utilized for transcoding process.

(xv) Cache Hit Rate: If the requested transcoded video is present in cache storage it is a hit. The cache hit rate indicates the ratio between a hit and total transcoded video request.

(xvi) Data Transfer Rate: Data Transfer Rate represents the transfer speed of the transcoded videos from the cloud to the user with the aim of saving user's energy consumption in retrieving the requested videos [26].

(xvii) Average Startup Delay: It is the delay occurred at the beginning of a video stream and represents the delay in receiving the first transcoded video stream [13].

(xviii) Average Deadline Miss-Rate: This delay occurs when a transcoding task misses its presentation deadline during streaming of a video [13].

(xix) Targeted Video Chunk Size: The average video chunk size influences the cloud resources, i.e., the number of CPU cores, so it should be carefully selected [19].

The study shows that transcoding time, i.e., the time taken by the system to transcode a video from one format to another format is an important metric. Other parameters used by researchers are transcoding rate/speed, i.e., the number of video transcoding per time unit. Some researchers used end-to-end delay (time difference between service requests in response), cost, etc. Operational cost, i.e., the cost used to acquire cloud resources such as storage, transcoding is also used by several researchers.

16.6 Conclusion

Initially, video transcoding performed to reduce file size, but now the perception has changed. Now transcoding is considered as a facility that provides options to choose. The options are picked to make the video viewable across platforms, devices, and networks. The cloud transcoding solution allows video service providers to pay as they use, with the assurance of providing online support to handle unpredictable needs, i.e., flash crowd. This paper discusses various aspects of performing a transcoding operation in cloud starting from why to shift the transcoding process to the cloud, to general architecture and metrics used to evaluate a cloud transcoding system. The system model designed for cloud-based video transcoding must consider the constraint, i.e., the online transcoding must be completed before the deadline of the frame. Future studies can be carried out to design and simulate a cloud-based transcoding model such that it will meet the constraint with less overhead. The overhead reduction may be concerning cost, scheduling, and provisioning of cloud resources, energy, response time, transcoding time, etc.

References

1. https://content.pivotal.io/blog/large-scale-video-analytics-on-hadoop
2. S. Sahoo, S. Nawaz, S.K. Mishra, B. Sahoo, Execution of real time task on cloud environment, in *2015 Annual IEEE India Conference* (INDICON) (New Delhi, 2015), pp. 1–5, https://doi.org/10.1109/INDICON.2015.7443778
3. S.K. Mishra, R. Deswal, S. Sahoo, B. Sahoo, Improving energy consumption in cloud, in *2015 Annual IEEE India Conference* (INDICON, New Delhi, 2015), pp. 1–6, https://doi.org/10.1109/INDICON.2015.7443710
4. S.K. Mishra, B.Sahoo, K.S. Sahoo, S.K. Jena, Metaheuristic approaches to task consolidation problem in the cloud, in *Resource Management and Efficiency in Cloud Computing Environments* (IGI Global, 2017), pp. 168–189, https://doi.org/10.4018/978-1-5225-1721-4.ch007
5. S. Sahoo, B. Sahoo, A.K. Turuk, S.K. Mishra, Real time task execution in cloud using mapreduce framework, in *Resource Management and Efficiency in Cloud Computing Environments* (IGI Global, 2017), pp. 190–209, https://doi.org/10.4018/978-1-5225-1721-4.ch008
6. S.K. Mishra, P.P. Parida, S. Sahoo, B. Sahoo, S.K. Jena, Improving energy usage in cloud computing using DVFS, in *International Conference on Advanced Computing and Intelligent Engineering (ICACIE)* (2016), http://hdl.handle.net/2080/2598
7. https://www.ericsson.com/res/thecompany/docs/publications/ericssonreview/2010/cloudcomputing.pdf
8. Kontron Whitepaper, Video optimization in the cloud (2014)
9. http://www.cisco.com/c/dam/en_us/about/ac79/docs/sp/Streaming_Under_the_Clouds.pdf
10. http://www.internetlivestats.com/internet-users/india/
11. https://www.akamai.com/uk/en/our-thinking/state-of-the-internet-report/state-of-the-internet-connectivity-visualization.jsp
12. S. Ko, S. Park, H. Han, Design analysis for real-time video transcoding on cloud systems, in *Proceedings of the 28th Annual ACM Symposium on Applied Computing* (SAC '13) (ACM, New York, 2013), pp. 1610–1615, https://doi.org/10.1145/2480362.2480663
13. X. Li, M.A. Salehi, M. Bayoumi, R. Buyya, CVSS: A cost-efficient and QoS-aware video streaming using cloud services, in *2016 16th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid)* (Cartagena, 2016), pp. 106–115, https://doi.org/10.1109/CCGrid.2016.49
14. S. Sahoo, I. Parida, S.K. Mishra, B. Sahoo, A.K. Turuk, Resource allocation for video transcoding in the multimedia cloud, in *International Conference on Advanced Computing, Networking, and Informatics (ICACNI)* (2017), http://hdl.handle.net/2080/2722
15. S. Sahoo, B. Sahoo, A.K. Turuk, An analysis of video transcoding in multi-core cloud environment. in *International Conference on Distributed Computing and Networking (ICDCN)* (2017), http://hdl.handle.net/2080/2643
16. https://www.wowza.com/blog/what-is-transcoding-and-why-its-critical-for-streaming (2015)
17. http://coconut.co/video-transcoding
18. http://download.sorensonmedia.com/PdfDownloads/LowRes/whitepaper.pdf (2011)
19. L. Wei, J. Cai, C.H. Foh, B. He, QoS-aware resource allocation for video transcoding in clouds. IEEE Trans. Circuits Syst. Video Technol. **27**(1), 49–61 (2017), https://doi.org/10.1109/TCSVT.2016.2589621
20. http://www.streamingmedia.com/Articles/Editorial/Featured-Articles/Buyers-Guide-to-Cloud-Based-Video-Encoding-and-Transcoding-2015-102483.aspx
21. http://download.sorensonmedia.com/PdfDownloads/LowRes/whitepaper.pdf (2011)
22. A. Ashraf, F. Jokhio, T. Deneke, S. Lafond, I. Porres, J. Lilius, Stream-based admission control and scheduling for video transcoding in cloud computing, in *2013 13th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing, Delft*, (2013), pp. 482–489, https://doi.org/10.1109/CCGrid.2013.21

23. F. Jokhio, A. Ashraf, S. Lafond, I. Porres, J. Lilius, Prediction-based dynamic resource allocation for video transcoding in cloud computing, in *2013 21st Euromicro International Conference on Parallel, Distributed, and Network-Based Processing*, (Belfast, 2013), pp. 254–261, https://doi.org/10.1109/PDP.2013.44
24. F. Jokhio, T. Deneke, S. Lafond, J. Lilius, Analysis of video segmentation for spatial resolution reduction video transcoding, in *2011 International Symposium on Intelligent Signal Processing and Communications Systems (ISPACS)* (Chiang Mai, 2011), pp. 1–6, https://doi.org/10.1109/ISPACS.2011.6146194
25. F. Jokhio, T. Deneke, S. Lafond, J. Lilius, Bit rate reduction video transcoding with distributed computing, in *2012 20th Euromicro International Conference on Parallel, Distributed and Network-based Processing* (Garching, 2012), pp. 206–212, https://doi.org/10.1109/PDP.2012.59
26. Z. Li, Y. Huang, G. Liu, F. Wang, Z. L. Zhang, Y. Dai. Cloud transcoder: bridging the format and resolution gap between internet videos and mobile devices, in *Proceedings of the 22nd international workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV '12)*, (ACM, New York), pp. 33–38, https://doi.org/10.1145/2229087.2229097
27. A. Alasaad, K. Shafiee, H.M. Behairy, V.C.M. Leung, Innovative schemes for resource allocation in the cloud for media streaming applications. IEEE Trans. Parallel Distrib. Syst. **26**(4), 1021–1033 (2015), https://doi.org/10.1109/TPDS.2014.2316827
28. X. Wang, M. Chen, T.T. Kwon, L. Yang, V.C.M. Leung, AMES-cloud: a framework of adaptive mobile video streaming and efficient social video sharing in the clouds. IEEE Trans. Multimed. **15**(4), 811–820 (2013), https://doi.org/10.1109/TMM.2013.2239630
29. Z. Huang, C. Mei, L.E. Li, T. Woo, CloudStream: delivering high-quality streaming videos through a cloud-based SVC proxy, in *2011 Proceedings IEEE INFOCOM* (Shanghai, 2011), pp. 201–205, https://doi.org/10.1109/INFCOM.2011.5935009
30. H. Zhao, Q. Zheng, W. Zhang, B. Du, H. Li, A segment-based storage and transcoding trade-off strategy for multi-version VoD systems in the cloud. IEEE Trans. Multimed. **19**(1), 149–159 (2017), https://doi.org/10.1109/TMM.2016.2612123
31. https://www.akamai.com/us/en/multimedia/documents/content/streaming-toward-televisions-future-4k-video-white-paper.pdf
32. G. Gao, W. Zhang, Y. Wen, Z. Wang, W. Zhu, Towards cost-efficient video transcoding in media cloud: insights learned from user viewing patterns. IEEE Trans. Multimed. **17**(8), 1286–1296 (2015), https://doi.org/10.1109/TMM.2015.2438713
33. https://www.brightcove.com/en/blog/2013/08/new-whitepaper-architecting-a-video-encoding-strategy-designed-for-growth
34. K.B. Chen, H.Y. Chang, Complexity of cloud-based transcoding platform for scalable and effective video streaming services, in *Multimedia Tools and Applications* (2016), pp. 1–18, https://doi.org/10.1007/s11042-016-3247-z
35. W. Zhang, Y. Wen, H.H. Chen, Toward transcoding as a service: energy-efficient offloading policy for green mobile cloud. IEEE Netw. **28**(6), 67–73 (2014), https://doi.org/10.1109/MNET.2014.6963807