Journal of Communications and Information Networks
Vol.1, No.1, Jun. 2016
DOI: 10.11959/j.issn.2096-1081.2016.010

Cloud-based video communication and networking: an architectural overview

YAN Zhisheng, CHEN Changwen

State University of New York at Buffalo, Buffalo NY 14260, USA

Abstract: Cloud-based video communication and networking has emerged as a promising new research paradigm to significantly improve the quality of experience for video consumers. An architectural overview of this promising research area was presented. This overview with an end-to-end partition of the cloud-based video system into major blocks with respect to their locations from the center of the cloud to the edge of the cloud was started. Following this partition, existing research efforts on how the principles of cloud computing can provide unprecedented support to 1) video servers, 2) content delivery networks, and 3) edge networks within the global cloud video ecosystems were examined. Moreover, a case study was exemplfied on an edge cloud assisted HTTP adaptive video streaming to demonstrate the effectiveness of cloud computing support. Finally, by envisioning a list of future research topics in cloud-based video communication and networking a coclusion is made.

Key words: cloud, video communication, video networking, survey

Citation: YAN Z S, CHEN C W. Cloud-based video communication and networking: an architectural overview[J]. Journal of communications and information networks, 2016, 1(1): 27-43.

1 Introduction

Recent years have witnessed the significant booming of a wide variety of networked video services, such as video streaming, video conferencing and video gaming. Such booming is not only due to the advances in terminals, e.g., desktops, laptops, smartphones, etc., but also owning to the omnipresent video services offered by various commercial and non-commercial companies. For example, videos have even become the panacea that would help traditional news media to retain their regular customers. According to a recent survey[1], global consumer internet video traffic will account for an impressive 80% of all consumer Internet traffic in 2019.

However, the user experience of current video services is far from satisfactory. We identify three critical issues in today's networked video services. First, the limited computing and storage resources of thin-client devices hinder the broader applicability of many video services, especially considering the big data nature of videos. For example, many video games require sophisticated graphics and computing support, which prevents them from the participation of most game users. Second, dynamic and heterogeneous network environment is still a serious time bomb that may cause unstable experience and unpredictable events. Both backbone congestion and last-hop uncertainty could lead to various annoying glitches. Finally, there is an inherent mismatch between the traditional Qos

Manuscript received Feb. 05, 2016; accepted Apr. 29, 2016
This work is supported by National Science Foundation Grants (No.ECCS-1405594).

(Quality of Service) based Internet applications and the QoE (Quality of Experience) driven video services. Profound attentions need to be paid to addressing the users' QoE in the networked video systems designs. Therefore, a paradigm shifting new framework is needed to develop and deploy next-generation Internet video services.

Recently, cloud computing technologies have emerged as such a promising paradigm that can provide abundant storage and computing resources for Internet-based video services. Such resources can be dynamically allocated to individual applications in an on-demand and real-time fashion. This offers a viable solution to resolve the thin-client problem in traditional localized video processing paradigm. More importantly, the virtualized resource allocation mechanism in cloud computing provides a prime opportunity to remedy the heterogeneity in service types, network environments, device categories, and QoE/QoS demands.

This gives birth to a new research paradigm, i.e., cloud-based video communication and networking. A natural question one would ask is how should the principles of cloud computing be seamlessly applied to individual tasks in the ecosystem of video communication and networking in order to enhance the users' QoE. To answer this question, we first summarize the key challenges of cloud-based video communications and networking as follows.

- Service heterogeneity. With the support of cloud and participation of diverse commercial enterprizes, the nature of video services can be drastically different, demanding dedicated technical treatment.
- Terminal diversity. The fast-evolving advances of computing, storage, and display technologies have speeded up the adoption of broad types of consumer terminals, from most common desktops, laptops, smartphones, to relatively new gaming computers and phablets.
- Resource management. The cloud is considered to be able to manage huge amount of data and network loads from massive number of clients in a dynamic, on-demand and real-time manner as this is the single foundation for cloud computing.
- Network dynamics. The backbone networks face the unprecedented congestion in this big data era which has not been seen before. Besides, the wireless and mobile access networks are still the top gripe of many mobile video users.
- User experience. The emphasis on system-level evaluation will be shifted from the objective QoS metrics into the more subjective QoE assessment.
- Security and privacy. The abundance of user data stored in the cloud is threatened by all sources of attacks from both individuals and groups.

There have been an array of technical advances made in the past few years to tackle these challenges. In this article, we present an overview of the emerging cloud-based video communication and networking system. We first revisit the structured partition of cloud-based video communication and networking system from an end-to-end perspective in Section 2. We then review in Section 3 to Section 5 the state-of-art technologies that aim at addressing the aforementioned challenges using the end-to-end architecture. We focus on high-level design concepts and intend to provide practical insights for future designs and implementations. In Section 6, we examine into the details of a recent case study on edge-cloud assisted adaptive HTTP streaming. Through this practical example, we can demonstrate the effectiveness and efficiency of the proposed cloud-based video communication and networking framework. Finally, we discuss a number of potential new research topics and conclude this article in Section 7.

2 Cloud-based video communication and networking: architectural components

The emerging cloud-computing technologies[2-5] have been envisioned to be capable of significantly improving video experience. In this section, we introduce the end-to-end modular architecture of cloud-based video communication and network systems. This perspective will be adopted as the guidance to survey existing research efforts and suggest future cloud video research. We highlight the design objective and considerations with an emphasis on the improving users' QoE. The modular end-to-end view of the cloud video system is shown in Fig.1. The system is comprised of four essential components in the cloud-based video communication and networking ecosystems, i.e., backend video servers, content distribution networks, edge networks and video consumers. In this framework, the first three components could employ various cloud computing technologies in order to enhance the QoE of the video consumers.

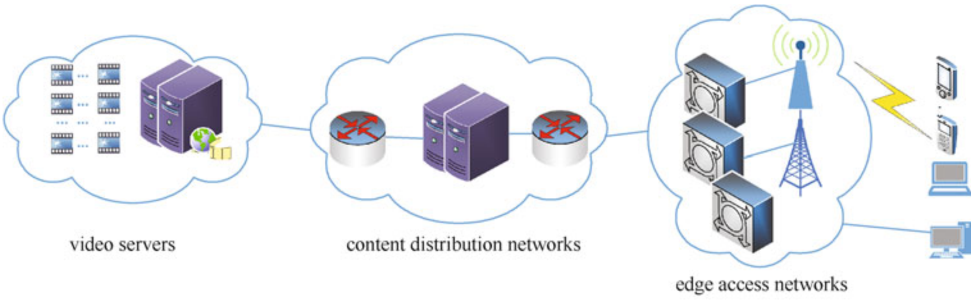


Figure1 An end-to-end architecture of cloud-based video communication and network

Backend video servers are in charge of primary provisioning of video services. These video servers are managed by content providers who are responsible for preparing video contents for distribution and consumption. Content providers host the video sources not only from professional producers acquired with advanced equipment, e.g., movies trailers, but also from everyday Internet users via their own camera or smart devices, e.g., user generated videos in YouTube. These extremely heterogeneous video sources in terms of genre, capturing device, and uploader can place demanding burdens for content providers to appropriately prepare the videos. Effective video encoding, rendering, and processing will definitely play a critical role on guaranteeing the quality of the to-be-sent videos, which will directly impact the QoE of users receiving the contents. Besides, it is also necessary to strike a dedicated balance between the video quality/users QoE and the cost of computing and storage in the backend video cloud.

CDN (Content Distribution Networks) are expected to efficiently distribute the video content in the backbone Internet, typically from backend video servers to edge access networks in different locations. CDN providers build a massive networks of distribution servers across different geographic regions. These servers normally exploit caching and/or prefetching techniques in order to distribute the videos over the Internet with minimal delay. The key challenge for achieving satisfactory QoE lies in minimizing the delay caused by the long and diverse distribution path and possible network congestion on the way. Furthermore, how to cost-efficiently and QoE-effectively deploy and regulate the CDN has become especially challenging due to the vast volume of videos with non-negligible redundancy and diverse types of user request patterns.

Edge networks are the last-hop access networks implemented by network operators. An edge network usually delivers the video content from wired gateways, wireless access points (in WiFi), base stations (in cellular networks) to a wired/wireless video terminal. The last-hop access is an essential part in the entire end-to-end video delivery pipeline. It often directly impacts the QoS and QoE of users. The wireless and mobile edge access always presents tricky issues considering the high dynamics in the wireless networks. Such dynamics could lead to annoying QoE degradation such as video re-buffering and video quality fluctuation. These QoE impairments will get even more complicated due to the diversity of terminals, network modes, and QoS/QoE requirements among different users.

In order to enhance the effectiveness of the these three components and eventually enrich users' video experience, cloud services have been introduced in the preparation, distribution, and access of video contents by reshaping a pool of computing, storageand networking resources. Cloud service providers build data centers that house a plethora of regular servers/workstations as well as video-dedicated CPU/GPU arrays to empower the efficient video processing. Besides, sizable storage capacity are also equipped for video storage. With these video-friendly resources, it is becoming possible to improve users QoE by making proper use of rich cloud resources. It is also important to note that the boundary between cloud services providers and the above three architectural components of video delivery system are becoming vague nowadays. For instance, video services and cloud services can be deployed jointly by one single provider, e.g., Amazon.

In the remaining of this article, we adopt this end-to-end architectural framework to examine emerging research on how cloud computing can support the three components in order to enhance users' QoE in various video services.

3 Cloud computing in video servers

In this section, we examine how the cloud resource can provide support in backend video servers. Due to the mismatch between the computation-intensive video services and the resource-limited thin clients, an increasing number of video services will need to shift the video preparation tasks from traditional single device to the cloud in order to achieve an effective video content provisioning.

We illustrate a general modular design of the cloud-assisted video servers in Fig.2. A service client first sends a service request to the cloud video servers. The network server in the cloud handles the request and delivers the request to the load balancer. The load balancer then schedules dedicated media GPU/CPU in the cloud, i.e., the video processor, based on the current resource allocation in the cloud and the characteristics of the request. Finally, each individual video processor takes over the processing tasks of content provision assigned to them and efficiently completes the tasks in a parallel fashion via virtualized instances of machines. The prepared video content will then be transmitted to the frontend users via the network server using standard network protocols.

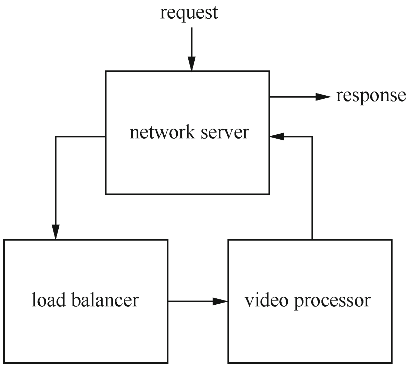


Figure 2 A general modular design of video servers in the cloud

3.1 Video encoding in the cloud

Video encoding has been recognized as a key task that consumes substantial computation resources. Its task is to compress the raw video data captured by the cameras or video recorders into encoded video bitstreams. To enable cloud-supported video encoding, the video processor in Fig.2 will be changed to a suite of standard video encoders that can usually be implemented in a parallel way.

CloudStream[6] is a typical cloud-based video processor that can deliver high-quality streaming videos through parallelization. It encodes the original video in real time to a scalable coded bitstream. The encoding framework is characterized by a multilevel encoding parallelization with Hallsh-based Mapping and Lateness-first Mapping. Essentially, video contents are split into slice or group of pictures and these sub-videos are encoded in different computing nodes in the cloud. The authors explored the complexity of video content and decided the minimum number of computing nodes to be used for each video. The parallel algorithms are able to minimize encoding speed without sacrificing the encoded video quality.

In addition to virtualize and parallel encoding, video encoding in the cloud can also utilize some specific characteristics and requirements of a video codec. For example, the authors in Ref.[7] proposed a cloud-based framework to augment the rate control design in the video codec. This work is able to improve the compression efficiency in the cloud environment. Experimental results demonstrate that such a joint rate control and cloud infrastructure achieve better visual quality.

Another interesting idea for cloud-based encoding is introduced in Ref.[8]. The authors completely bypassed the current practices of pixel based image encoding. Instead, they proposed a model based method to describe images by image descriptors and then exploit the large-scale image database in the cloud to reconstruct the new images by using their descriptors. Such a method can achieve an extremely high compression efficiency (1 000:1) compared with traditional intra-frame image coding. It is of strong interests to explore whether or not this novel idea can be used in video encoding where the description of video frames may be a more difficult task.

3.2 Video game rendering in the cloud

Another type of video services that needs advanced computation power for content provision is the emerging cloud gaming services. In fact, the video processing of cloud gaming is even more complicated than the video encoding/streaming services. Recall that video encoding services demands arrays of encoders in the video processor block in Fig.2. In contrast, the video processor block for cloud gaming services[9,10] will be replaced by multiple functionalities as shown in Fig.3. The user requests from the clients will be delivered to individual video game processor via the load balancer. First, these game actions are signaled to game logic for inferring the future moves. The instructions of new game actions are transmitted to the GPU to render future gaming images. Finally, these images are encoded by standard encoders and delivered to the clients.

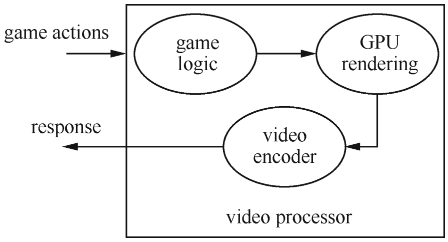


Figure 3 The video processor in cloud gaming services

One fundamental issue in cloud video gaming services is the delay between the moment a user performs an action and the moment the respective rendered gaming videos displayed on the terminals. Such an interaction delay consists of three main parts[11]: (1) network delay which is required to deliver a user command to the server and return a image to the client, (2) processing delay during which the video server renders the gaming image based on user command, (3) playout delay for the client to decode and render the video frame on the devices.

To address the latency issue, outatime[12] utilized the idea of speculative execution to predict the possible user action via a Markov model. Multiple predicted video frames will be rendered an entire RTT time ahead in order to reduce the processing delay in the cloud. The authors proposed a set of techniques to increase the prediction accuracy. For example, they approximates the correct behavior state of users by applying error compensation on the (mis) predicted frame. They also used rollback and checkpoint of users' behavior state to prevent the error propagation of states. Finally, the states of user behavior were compressed to save the network bandwidth.

Shi, et al. attempted to reduce the rendering delay in the cloud from a different perspective[13]. They proposed to explore the graphics rendering states, such as view point, pixel depth, and camera motion, to release the burden of video encoding and thereby improve the system-level performance. The basic idea is to pick the key frames in the game and apply a 3D image warping algorithm to interpolate other intermediate frames. Thus less encoding tasks are needed and the encoding quality may be enhanced since more bits can be allocated to the key frames.

In addition to these application-layer solutions that inspects into the nature of video gaming, VGRIS[14] represents another direction of cloud-gaming approach that aims at efficiently managing the GPU resources in the cloud. The authors implement a parallel virtualization framework with an agent for each VM and a centralized scheduling controller. They also simplify the usage of the framework by building a suite of APIs for easy and efficient access of resource scheduling in the cloud.

3.3 Video processing in the cloud

Apart from video encoding and rendering in the video servers, there are other types of video processing that can be implemented in the cloud to enhance end users' QoE.

One typical example of such video processing is video transcoding, which converts the encoded videos into another version, e.g., a new resolution and/or bitrate. Most existing efforts exploited a Map/Reduce based method[15,16] for video transcoding, where input video sequences are divided into segments, and mapped to multiple computing servers. Each sub-task are performed in parallel with processing results concatenated to the final output sequences. This way, the efficient cloud-based encoding can be accomplished. Then, one important research issue is how to design the parallel algorithm to optimally balancing computing resources and encoding quality.

Similarly, Ref.[17] also segments the video sources into group of pictures and then assign each group to individual cloud server independently. The key idea of this work is to predict the load of transcoding services based on the load history such that the optimal number of encoding nodes can be derived. The allocation and deallocation of virtual machines is done in a horizontal fashion. The simulation results show that such a proactive resource allocation can achieve satisfactory performance if the prediction accuracy is sufficiently high. One potential issue is how to guarantee an accurate load prediction in the practical environment with heterogeneous services models.

More recently, a novel cloud-based dynamic scheduling methodology for video transcoding is studied under the emerging MPEG DASH environment[18]. MPEG DASH is the dominant video streaming technique in the contemporary Internet, where videos are split into chunks and encoded into different bitrate for dynamic and adaptive delivery. This work monitors the workload on each encoder in the cloud and selects the fastest video processors to execute those high-priority jobs, thereby resulting in improved performance. The system-level real-world results are able to demonstrate that this strategy can improve

QoE by increasing video encoding time and ensuring playback smoothness.

Another example of video processing in the cloud is to support the emerging free viewpoint video rendering for mobile devices. In Ref.[19], a cloud-based free viewpoint rendering framework for mobile devices over cellular networks is presented. In this framework, the cloud in video servers assists the mobile devices for computing-intensive free viewpoint rendering. A resource allocation scheme that balances the rendering allocation between cloud and client based on users QoE is also proposed. Experimental results show that the improvement of such cloud-assisted framework against traditional strategies can be achieved.

4 Cloud computing in CDN

In this section, we review emerging cloud computing technologies in CDN. Nowadays, video content are mostly distributed via CDN to the end users. The CDN are normally created on top of high-performance cloud that are also responsible for the distribution of videos. There is a fundamental tradeoff between the QoE/QoS of video services and the cloud resources or cloud services costs consumed in the CDN. This is further complicated by the geographically distributed nature of CDN servers. As such, various types of network-layer strategies, such as routing, bandwidth allocation, load balancing, content caching, and content replication, have been developed recently.

4.1 Video routing and bandwidth allocation in the cloud

In order to efficiently and reliably deliver the video packets over the CDN with acceptable QoS requirement, such as latency and loss rate, it is essential to optimize the data flow in the CDN cloud. One traditional topic is packets routing. The nature of video services that handles this particular type of data introduce new opportunity to improve the routing performance. Compared with source-routing based approaches, the cloud-assisted routing can offer more cost-effective provisioning by utilizing better knowledge of the network as well as the powerful computation capability.

vSkyConf[20] is a cloud-based multi-party video conferencing platform. It was proposed to address the inefficiency of traditional non-cloud video conferencing, where the video chat images are encoded in the local devices. By monitoring the latency and bandwidth on the connections from/to neighboring clients, a vSkyConf client would be able to construct a connectivity topology from the point of view of this client. The client can then compute multiple routing paths for outgoing streams to other clients. The Amazon EC2 based experiments have verified the superior performance over unicast solutions. Similarly, several multi-path provisioning algorithms for cloud-assisted SVC (Scalable Video Coding) based streaming have been proposed in Ref.[21]. The unique consideration was to utilize the separability of SVC videos and to optimize the routing of different layers over multiple paths.

One different yet very interesting work tried to jointly optimize the request mapping and response routing which were usually considered independently[22]. The authors formulate a general workload management optimization by considering delay, location diversity of electricity and bandwidth costs. A distributed near-optimal algorithm is designed and evaluated to prove the effectiveness of the solution. One concern of such a complex algorithm is the lack of practical value considering that only trace-driven rather than real-world evaluations are carried out.

Apart from optimally routing the packets through CDN, another dimension that researchers have investigated in CDN cloud is to reserve/allocate bandwidth resources in the CDN cloud in order to minimize the cloud resource cost while still achieving desired QoE for video users. For example, the authors in Ref.[23] predicted the short-term bandwidth demand of video streaming users according to the load history in order to accomplish statistical multiplexing in video bandwidth. Another example[24] is the study of bandwidth allocation across geo-distributed cloud data centers. By borrowing the Nash Bargaining Theory, the authors derived an optimal tradeoff between cloud operational cost and the QoE in video streaming services.

4.2 Video caching and replication in the cloud

With the popularity of video services and massive video requests all over the world, it is evident that a significant portion of these requests will be overlapped. In fact, the video access over the Internet manifests a long-tail effect, i.e., a small number of videos account for most of the video access. By removing duplicated storage of popular videos that may be accessed by vast number of users, the efficiency of cloud storage in the distribution networks can be significantly enhanced. Furthermore, the band width and latency performance of video services can also be improved if one can appropriately place the popular content across proper geographic regions. Therefore, two emerging research topics arise from meeting these requirements: (1) video caching that studies which videos should be prefetched in the cloud storage, (2) video replication that investigates how to store each video in different cloud.

In Ref.[25], the authors demonstrated a simple yet effective scheme of cloud caching for RSS (Really Simple Syndication) based video delivery. In traditional RSS delivery system, all the users who retrieve the same RSS feed and the associated content will need to download the same files. In order to remove such a repeated download pattern and avoid unnecessary data delivery, the authors proposed to cache all the requested files in the cloud. Once a duplicated request occur, the cached content can be delivered directly from the cloud nearer to the user without a new fetching from the RSS server. However, this may introduce numerous duplicated contents. Hence, a scheme to remove the content duplication may be needed to further improve this scheme.

To further improve caching efficiency, it is essential to understand the user behavior in terms of video access. Hence, many efforts have been made to study the video caching via OSN (Online Social Networks), where people exhibit a tractable pattern of interaction. For example, the authors in Ref.[26] collect the video viewing statistics in OSN and learn the user behavior including user interest and hot viewing period of a day in order to facilitate a more effective video caching. Similarly, another scheme[27] has also been developed to utilize the sharing levels in OSN to predict the possibility that a shared video may be accessed by the potential recipients. If a video is shared via direct recommendation rather than public sharing, it will be most likely to be viewed by the recipient of this sharing activity. Such video should be cached with higher preference.

Regarding video replication, the key objective is to satisfy the caching requirement with minimal cloud storage and bandwidth cost, i.e., how to cost-efficiently place cached content in particular servers of particular regions. AREN[28] is an adaptive replication scheme that minimizes the number of SLA (Service Level Agreement) violations. It also tracks the state of bandwidth reservation and cloud caching to minimize the cost. AREN was shown to successfully prevent the vast majority of SLA violation under heavily load situations while reducing nearly seven-fold of cloud storage and also increasing 20% aggregate bandwidth. Another example described in Ref.[29] aimed at jointly achieving energy-efficient and bandwidth-efficient replication in the data centers across the CDN.

Similar to video caching, video replication in OSN inherits certain characteristics. In Ref.[30], the authors carried out a large-scale measurement on a popular microblog platform and revealed three important factors impacting the propagation of videos: (1) social connection, (2) geographical location, (3) temporal period. The replication strategy employs a hybrid cloud and peer method, where a video recommended by a friend will be cached by the peer while a video is popular in a particular location during a time period will be cached in the local CDN cloud. An improvement of 30%~40% cache hit ratio is observed against conventional replication schemes.

5 Cloud computing in edge networks

Edge access networks are the last-hop communication premises in the end-to-end architecture for video delivery to the end users. Wireless/mobile edge networks have attracted special research attentions due to the inherent mismatch between the prevalence of mobile video services and the notorious unreliability of lossy wireless channels. In order to fill in this gap, video adaptation has been widely proposed in the edge cloud, which acts as an intermediate agent between users and video servers. The edge cloud can adapt the video content in real-time to balance the competing obligations in QoS/QoE requirement, cloud services cost, and the last-hop wireless/mobile link quality.

A general schematic flow of the edge cloud for video adaptation is illustrated in Fig.4. Based on the services requirements and service cost that can be obtained from the video servers, together with the last-hop channel condition that can be fed back from the video client, the principle adaptation logic in the edge cloud would be able to perform appropriate QoE/QoS-driven and channel-aware video adaptation. This shall systematically improve the video services.

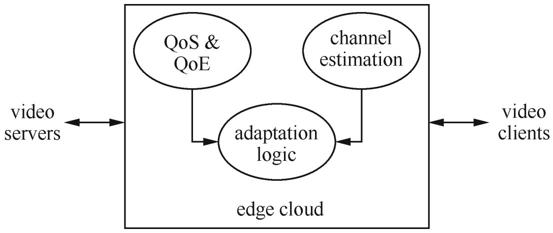


Figure 4 A general structure for video adaptation in edge cloud

5.1 Video adaptation in the edge cloud

A wide range of video adaptation tasks can be executed in the edge cloud. One crucial application is to adapt the video rendering in online gaming such that the interactive latency can be guaranteed under the last-hop wireless channel conditions.

In Ref.[31], the authors developed a rendering adaptation technique that can adapt the game rendering parameters to meet the requirement of last-hop communication and computation. The rendering options for adaptation included texture detail, environment detail, frame rate, realistic effect and view distance. These variables were manipulated and searched within the categories to seek an optimal combination while considering the constraints of total cost of communication and computation. Similarly, the authors in Ref.[32] explored the necessity of specialized GPU-intensive cloud in the edge networks to speed up interactive gaming services. They observed that such a configuration can complement the relative performance gap of general-purpose cloud in the backend, but would yield large distribution and deployment cost of edge cloud. Hence, they reached the conclusion that a hybrid architecture with distributed edge cloud and a centralized backend cloud is needed for today's video gaming services which were computationally highly extensive.

The other common type of video adaptation in edge cloud is to transcode the streamed video on the fly in order to match the device and channel status of the mobile terminals. Such a idea has been validated by some genuine experimental studies in various network environments in Ref.[33]. Cloud transcoder[34] is a complete transcoding solution in the edge cloud to bridge the gap between source videos in the Internet and hardware support in the mobile devices. These gaps usually include different resolution support, codec hardware, and file format, between the two ends. Since such a real-time transcoding may introduce high pressure on the edge cloud, Cloud Transcoder uses data reuse among users to cache those duplicated videos for future access. Besides, Cloud Transcoder will also keep the transcoded version of multiple videos for a larger coverage of video caching.

5.2 HTTP adaptive streaming in the edge cloud

Video streaming over cellular networks has become one of the most prevalent mobile services. Due to its inherent scalability and versatility, HAS (HTTP Adaptive Streaming) has been widely recognized as the dominant technology for mobile video delivery. In HAS, the video source is pre-encoded in several bitrate versions and each version is split into small segments. The client adaptively requests the video segment at each switching point based on per-segment throughput measurement and estimation. In this way, the user is expected to receive the most proper video version and achieve satisfactory QoE under current channel conditions. One particular issue in this type of client-driven HAS is that it could cause serious performance degradation when multiple clients are competing within the same bottleneck[35,36]. Since a client is unaware of other clients in the shared bottleneck, it may overestimate or underestimate its own bandwidth and thereby cause the decision dilemma.

Little work has been done to design HAS over multi-client cellular networks. Some schemes[37,38] combined the designs of rate adaptation and resource allocation in order to allow channel-aware playback. These schemes depend on the customized low-layer cellular scheduler, which needs to modify the standard cellular infrastructure with proportional fair scheduler. Others[39,40] aimed at optimizing the utility function of cellular users. Although the system performance is improved, the client-side decisions are completely overwritten, which could cause issues when client device has certain limitations. Furthermore, the utility function lacks desired connection with user experience. Therefore, there is a significant potential to move up the design space to a hybrid edge cloud and client adaptation to enhance the QoE of multiple mobile users. To illustrate such potential design, we will present a recently developed scheme on how to address this tricky issue as a successful example for edge cloud assisting video services.

6 Case study: edge cloud assisted HTTP adaptive streaming

This is one of the first schemes to study edge cloud assisted HAS in order to improve the QoE over multi-client HAS in mobile cellular networks[41-45]. In this section, we briefly introduce a complete solution based on edge cloud, called Prius[45], which is the first successful hybrid edge cloud and client adaptation framework for HAS. Prius has been designed to take full advantage of the new capabilities empowered by recent advances in edge cloud computing and has shown promising performance under various mobile environments.

6.1 Motivations

Although broad consensus has been reached for client-driven HAS in wired or single-client wireless networks, the understanding of rate adaptation strategies for mobile cellular videos is still quite limited. The reason behind this observation is that edge cloud introduces new effects on HAS over cellular networks while little is known regarding the optimal adaptation under this new context.

First, standard throughput based adaptations[46,47] cannot accurately capture the bandwidth variations in cellular networks. It has been shown that these schemes largely over/underestimate the bandwidth share unless the downloading of a segment would saturate the end-to-end bandwidth[35,36]. This may be attributed to the well-known mismatch between per-segment throughput and real bandwidth share. The strong dynamics of cellular bottleneck makes this issue even trickier because channel condition can degrade suddenly just when the high-layer estimation/smoothing/probing approximates the bandwidth share. Thus the instable and unfair playback will be inevitably introduced. Moreover, measurement studies[48] have shown that the prevalent assumption of TCP fairness may not be true in cellular networks. This indicates that multiple clients would not achieve fair performance even though they somehow learn the channel information and request the true bandwidth share. Finally, since the multi-client playback on top of TCP bandwidth share may not be fair over cellular networks, it is desired to proactively adjust bitrate of each client. For example, requesting a bitrate lower than the extremely high bandwidth of a client may be fairer than pushing its bitrate to the unfair bandwidth share. This is because it saves some channel resources for other clients with lower bandwidth share, which will contribute to an overall QoE fairness.

We observe that the fundamental reason behind the aforementioned issues in cellular networks is that the clients are oblivious to the bottleneck radio channel and cannot coordinate with each other to determine a fair bitrate under the unfair bandwidth share. The emerging edge cloud presents a new opportunity to remedy these issues. As the low-layer RAN (Radio Access Network) information (e.g., instant channel state) is available in edge cloud, the variation of link capacity can be predicted and utilized towards rate adaptation. Besides, edge cloud is a centralized entity that sits within the RAN, e.g., at the base station, and thereby can perform cell-wide joint adaptation. Moreover, thanks to the computational support of edge cloud, sophisticated QoE models and optimal rate adaptation for multiple clients can be developed to maintain fair perception of multiple users. In this research, we develop Prius, a HAS system with edge cloud support in order to maximally enhance all users' QoE and QoE fairness fora multi-client mobile cell.

6.2 Architecture

Prius adopts a similar system architecture as shown in Fig.1, where multiple clients are receiving the streams the segmented videos at a set of potentially different bitrates. Prius consists of a Prius client and a Prius adaptation module for the hybrid adaptation. Since edge cloud is assumed to be computationally powerful and can access RAN information available in the base station, e.g., CQI, Prius overlays a layer of adaptation intelligence at the edge cloud on top of the individual clients' bandwidth-irrelevant bitrate requests. In particular, as it is unlikely to accurately capture the bandwidth share in cellular links, Prius clients request a bitrate without concerning the bandwidth. However, other device related constraints, such as display, computation ability, and battery power, are all considered. This client request is in fact a upper bound of the bitrate request, which reflects the maximum bitrate that can be supported by the Prius client. At the edge cloud, the Prius adaptation module can explore the channel knowledge of multiple streams for joint adaptation while meeting the local client-side requests.

The operational process of Prius is summarized as follows. Initially, the HAS server sends out the MPD (Media Presentation Description) so that Prius adaptation module and clients will have the knowledge of available video representations. At each adaptation period that equals to the segment length, Prius clients request a video segment at a certain bitrate version based on its local factors unrelated to bandwidth. Unlike conventional client-side adaptation where the cellular networks simply forward the client requests to the video server, the edge cloud will intercept the requests. Prius adaptation module will then overwrite the adaptation decisions based on client-side request and execute cell-wide optimization of multiple clients, where both low-layer CQI and high-layer playback information are properly utilized. Client playback information for adaptation, such as buffer and QoE status, is embedded in the periodic feedback from clients.

The bitrate adaptation results are then delivered to the video server for streaming the next segment. In this way, the users are able to enjoy the video with optimized and fair QoE while no modification is needed in the video servers. Through this design,the proposed architecture can also be implemented friendly without modifying current infrastructure of cellular schedulers and the standard request-response mechanism of HAS.

6.3 Designs

Prius adaptation module in the edge cloud consists of three components: QoE evaluator, channel estimator, and bitrate adapter. We now briefly describe the functionality and key operations of these components.

QoE evaluator assesses the expected QoE of a user given a potential bitrate selection. We propose to use a sophisticated QoE metric, QoE continuum[49], by leveraging the rich computing resource in the edge cloud.

Psychological studies have discovered that the strength of human memory decays exponentially with respect to time[50]. For example, the visual perception of most recent video frames plays a more important role toward forming one's subjective QoE thanthat of less recent frames. Hence, we exploit this effect to model the QoE continuum and measure QoE in a temporally continuous fashion. We first assume only one frame can be played at one moment and derived Q_k, the QoE continuum at moment k, as the weighted summation of instantaneous user experience over all previous moments until the measuring moment, i.e.,

$$Q_k = \gamma Q_{k-1} + (1-\gamma) q_k \ , \tag{1}$$

where $k(k>0)$ is an index indicating a certain moment, q_k is the instantaneous user experience at moment k, Q_{k-1} is the QoE continuum at the previous moment $k-1$, and γ is the characterization constant of the memory strength. Thus, with a given initial QoE Q_0, we could iteratively fit in each instantaneous experience $(q_1, q_2, \cdots, q_k)$ and finally output the QoE continuum at moment k. Note that Q_k, $|q_k|$ and γ all belong to (0,1]. With this formulation, we can capture the QoE from all previous moments until the current measuring moment. q_k is decided by the player status and bitrate level. Detailed modeling can be found in Ref.[49].

Channel Estimator predicts CQI (Channel Quality Index) of last-hop link by using available CQI history in edge cloud. We propose a single exponential smoothing based scheme to estimate the CQI of the upcoming adaptation period. Single exponential smoothing has been commonly used in the forecasting of time series data without systematic trend[51], which is especially true for the highly dynamic radio channel dictated by user movement, signal fading, shadowing, etc. The estimated CQI of the upcoming period t', $CQI_{\text{est},t'}$, can be expressed as

$$CQI_{\text{est},t'} = \alpha CQI_{\text{avg},t} + (1-\alpha) CQI_{\text{est},t} \ , \tag{2}$$

where $CQI_{\text{avg},t}$ is the average CQI of the period t, $\alpha(0<\alpha<1)$ is the smoothing factor. Since the optimal α shall depend on channel variation pattern and such patterns could be largely distinct under different moving status,it is critical to experimentally study the impacts of α. Details of the choice for α will be discussed in the evaluation results.

Based on the estimated CQI, the edge cloud can finally obtain the estimated maximum transmission rate of the upcoming period via the 3GPP mapping table $f(\cdot)$, i.e.,

$$R_{t'} = \frac{f(CQI_{\text{est},t'})}{\tau} \ . \tag{3}$$

Bitrate Adapter takes the QoE Evaluator and Channel Estimator as the internal functions and employs both functions to optimize the bitrate selection for the upcoming period.

Given a video, which is segmented into T-second chunks, with a set of different bitrate versions $\mathcal{V}=\{br_1, br_2,\cdots, br_M\}$ and N clients with client-bounded bitrate $br_{i,\text{bound}}$, each with current QoE continuum Q_i, current buffer status B_i, and estimated maximum transmission rate R_i, the problem of bitrate adaptation is to determine the bitrate version $r_i(r_i\in\mathcal{V})$ for all clients $i\in\mathcal{N}$ such that the average QoE continuum for all users at the end of next adaptation period is maximized without exceeding shared resource constraint and client-side bound. Mathematically, bitrate adapter solves the following problem,

$$\begin{aligned} \max_{(\boldsymbol{r})} \quad & \frac{1}{N}\sum_{i=1}^{N} Q_{i,t+T} \ , \\ \text{s.t.} \quad & \sum_{i=1}^{N} \varphi_i \leqslant 1 \ , \\ & r_i \leqslant br_{i,\text{bound}} \ , \end{aligned} \tag{4}$$

where $\varphi_i = \frac{r_i}{R_i}$ is the radio resource share of client i.

The adaptation wisdom of bitrate adapter is that higher bitrate version is generally assigned to those users who currently possess a lower QoE continuum value and a better channel condition, while also avoiding significant bitrate variation. In other words, when a user enjoys good experience for a long time, his/her satisfaction will rise less than the one with bad previous experience if the video quality can be raised. Consequently, we can enhance not only the QoE continuum but also the fairness of users. We have proposed a heuristic algorithm to approximate the solution to the optimization problem by using a high-complexity dynamic programming algorithm. We will show the results of this efficient algorithm in the next section.

6.4 Evaluations

To verify the performance of Prius, we have built a ns-2 based environment based on the architecture in Fig.1 that includes video servers, core networks, cellular networks, the edge cloud with Prius adaptation module and Prius clients. We implement a 3.5G HSPA networks as the underlying cellular networks. The moving speed of users is 3 km/h under slow moving status (ITU pedestrian model) and 120 km/h under fast moving status (ITU vehicular model). Detailed configuration of the evaluations can be found in Ref.[45].

We compare the performance of Prius with several representative reference systems. To highlight the inherent advantages of edge cloud assisted hybrid adaptation, we first implement a conventional client-side rate adaptation algorithm, which requests the maximum bitrate that can be supported by current per-segment throughput. We also implement a typical centralized adaptation algorithm (referred as Instant) that captures the logic behind many existing works, where the joint adaptation maximizes the utility function dictated by selected instant bitrate, subject to the channel constraint, e.g., Ref.[52]. Finally, we implement the optimal dynamic programming algorithm of the adaptation problem to study the performance gap between the proposed heuristic algorithm and the theoretical upper bound.

We now describe here the procedures to evaluate the QoE continuum of the different HAS systems. We collect the actual QoE continuum (not the estimated $Q_{i,\,t+T}$) at all displaying frames from the video players. We consider the QoE continuum larger than 0.8 (corresponding to MOS 4.0) as "good" experience and show the probability of good QoE in Fig.5 and Fig.6. Under both slow and fast moving cases, Prius outperforms the reference systems. This is because Prius exploits the edge cloud to address the issue of shared bandwidth, which may cause playback instability for the case of the client-side adaptation that only uses end-to-end throughput based criterion. Furthermore, the bitrate adapter jointly considers the QoE continuum and channel resources whereas Instant algorithm only aggressively maximizes the instant bitrate, which is not necessarily able to ensure QoE continuum. From these evaluations, we also observe that the proposed QoE continuum based approach achieves a near-optimal performance, which validates the feasibility of the polynomial-time algorithm. Finally, the trend of QoE continuum versus α shows weighing more towards last CQI sample brings more accurate channel estimation in slow-moving channel with less channel variation and thus better QoE continuum. On the other hand, smoothing out the CQI samples with a smaller α will compensate for the frequent channel variation in fast-moving channel and will benefit the rate adaptation.

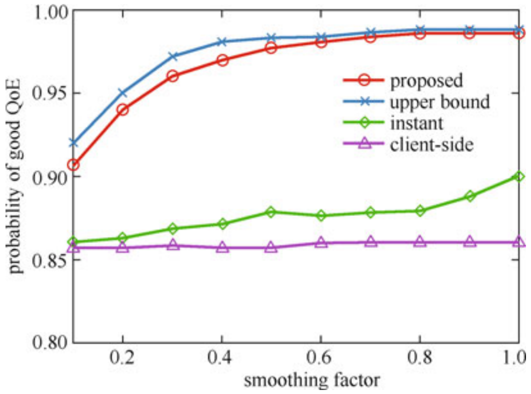


Figure 5 Probability of good QoE versus α under slow moving status

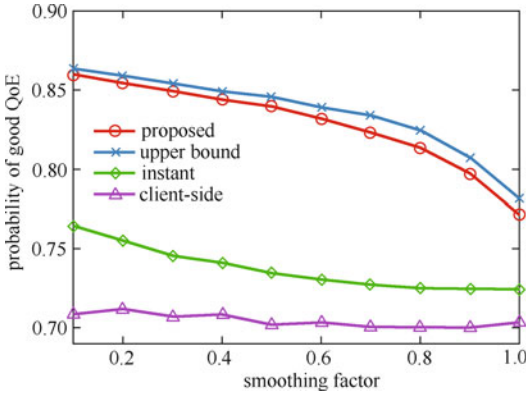


Figure 6 Probability of good QoE versus α under fast moving status

7 Concluding Remarks

Future video services are expected to offer the mobile users the possibility to view any desired video anywhere and anytime. At the same time, cloud computing will provide new design landscape to achieve this ultimate objective by providing substantial computing, storage and networking resources. We have witnessed such a emerging trend through tremendous ongoing research efforts as surveyed in this article. In order to realize the full potential of cloud-based video communication and networking systems with key requirements summarized in Section 1, we envision of a list of potential directions towards the ultimate objective.

- First, particular nature of video-based applications will need to be fully considered in the designs of cloud video systems. Heterogeneous video services usually exhibit rather diverse characteristics and services requirement. For example, an interactive service emphasizes latency more than the loss rate. Besides, the environmental context have substantial influence on a given service. For instance, a user moving around her smartphone may not experience a high QoE during the video conferencing whereas such a movement may be necessary in the video gaming case. It is absolutely necessary to highlight these application-layer features in the design of the cloud resource allocation and video adaptation.
- Second, the popularity of online social networks and the abundant information embedded in OSNs should be intelligently explored to understand the user behavior and thereby improving the QoE-driven cloud video systems. For example, by mining the video popularity history of streaming video sites, a cloud video system shall be able to predict the expected hits of a newly uploaded video and thereby build a more effective caching or replication scheme on top of the popularity of this

video. This should naturally save the storage cost and possibly reduce the delivery delay in CDN.

- Third, the newly promoted networking architecture may prompt us to fundamentally rethink the designs of cloud video systems. Emerging networking new paradigms, such as SDN (Software Defined Networking) and CCN/ICN (Content/Information Centric Networks) may reversal some commonly adopted design principles in the regular Internet based video system. By combining the new networking architecture and cloud-based video services, one may push the system-level performance bar to a surprisingly high level. For example, SDN adopts a highly convenient way for re-configuration and management of CDN services, which can definitely reduce the operation cost and thereby improving the video services.
- Forth, joint design of cloud resource allocation and video adaptation/preparation/distribution has not yet attracted research attention it deserves. Most existing works either focus on improving video services and QoE by assuming unlimited cloud resources or aim to maximally enhance cloud computing efficiency without considering particular use cases. We argue that by jointly investigating both topic, a more realistic cloud video system can be designed. Such an experimental research shall foresee higher practical impacts and will more likely to be adopted by industry for real-world deployment.
- Finally, although significant attentions have been attracted to the cloud video system designs for maximizing delivery performance, the security and privacy of video data and user information have not yet been addressed adequately. The huge amount of the data stored in the cloud further complicated this challenge. However, the gigantic data volume again presents the significant challenges in video data privacy. Different from textual data, raw video data is an intuitive format that anyone can understand and subject to all type of attacking if no proper protection is overlaid on the data. Therefore, it is imperative to resolve the issues related to the video data security and privacy in this cloud video era.

We believe that major progress towards the ultimate objective can be made through the researches along these emerging directions as outlined above. It is time to seize these golden opportunities to enrich cloud-based video services and to enhance the end users' video QoE in such an omnipresent and immersive video communication and networking era.

References

[1] CISCO SYSTEMS. Cisco visual networking index: global mobile data traffic forecast update, 2014-2019[EB/OL]. http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.pdf.2015

[2] RIMAL BP, CHOI E, LUMB I. A taxonomy and survey of cloud computing systems[C]//Proc of IEEE International Joint Conference on INC, IMS and IDC, Seoul, Korea, c2009: 44-51.

[3] BELOGLAZOV A, BUYYA R, LEE YC, et al. A taxonomy and survey of energy-efficient data centers and cloud computing systems[J]. Advances in computers, 2011, 82(2): 47-111.

[4] ENDO PT, GONCALVES GE, KELNER J, et al. A survey on open-source cloud computing solutions[C]//Proc of Brazilian Symposium on Computer Networks and Distributed Systems, Gramado, Brazil, c2010: 3-16.

[5] ZHOU M, ZHANG R, ZENG D, et al. Services in the cloud computing era: A survey[C]//Proc of IEEE International Universal Communication Symposium, Beijing, China, c2010: 40-46.

[6] HUANG Z, MEI C, LI L E, et al. CloudStream: delivering high-quality streaming videos through a cloud-based SVC proxy[C]//Proc of IEEE INFOCOM Mini Conference, Shanghai, China, c2011: 201-205.

[7] ZHENG L, TIAN L, WU Y. A rate control scheme for distributed high performance video encoding in cloud[C]//Proc of IEEE International Conference on Cloud and Service Computing, Hong Kong, China, c2011: 131-133.

[8] YUE H, SUN X, YANG J, et al. Cloud-based image coding for mobile devices-toward thousands to one compression[J]. IEEE transactions on multimedia, 2013, 15(4): 845-857.

[9] SHEA R, LIU J, NGAI E, et al. Cloud gaming: architecture and performance[J]. IEEE network, 2013, 27(4): 16-21.

[10] HUANG C Y, HSU C H, CHANG Y C, et al. GamingAnywhere: an open cloud gaming system[C]//Proc of ACM Multimedia Systems Conference, Oslo, Norway, c2013: 36-47.

[11] CHEN KT, CHANG YC, TSENG PH, et al. Measuring the latency of cloud gaming systems[C]//Proc of ACM Internal Conference on Multimedia, Scottsdale, Arizona, USA, c2011: 1269-1272.

[12] LEE K, CHU D, CUERVO E, et al. Outatime: using speculation to enable low-latency continuous interaction for mobile cloud gaming[C]//Proc of ACM International Conference on Mobile Systems, Applications, and Services, Florence, Italy, c2015: 151-165.

[13] SHI S, HSU C H, NAHRSTEDT K, et al. Using graphics rendering contexts to enhance the real-time video coding for mobile cloud gaming[C]//Proc of ACM Internal Conference on Multimedi, Scottsdale, Arizona, USA, c2011: 103-112.

[14] QI Z, YAO J, ZHANG C, et al. VGRIS: virtualized GPU resource isolation and scheduling in cloud gaming[J]. ACM transactions on architecture and code optimization, 2014, 11(2): 17.

[15] LAO F, ZHANG X, GUO Z. Parallelizing video transcoding using map-reduce-based cloud computing[C]//Proc of IEEE International Symposium on Circuits and Systems, Seoul, Korea, c2012: 2905-2908.

[16] KO S, PARK S, HAN H. Design analysis for real-time video transcoding on cloud systems[C]//Proc of ACM Symposium on Applied Computing, Coimbra, Portugal, c2013: 1610-1615.

[17] JOKHIO F, ASHRAF A, LAFOND S, et al. Prediction-based dynamic resource allocation for video transcoding in cloud computing[C]//Proc of Euromicro International Conference on Parallel, Distributed and Network-Based Processing, Belfast, Northern Ireland, c2013: 254-261.

[18] MA H, SEO B, ZIMMERMANN R. Dynamic scheduling on video transcoding for MPEG DASH in the cloud environment[C]//Proc of ACM Multimedia Systems Conference, Singapore, c2014: 283-294.

[19] MIAO D, ZHU W, LUO C, et al. Resource allocation for cloud-based free viewpoint video rendering for mobile phones[C]//Proc of ACM International Conference on Multimedia, Scottsdale, USA, c2011: 1237-1240.

[20] WU Y, WU C, LI B, et al. vSkyConf: Cloud-assisted multi-party mobile video conferencing[C]//Proc of ACM SIGCOMM Workshop on Mobile Cloud Computing, Hong Kong, China, c2013: 33-38.

[21] ZHU Z, LI S, CHEN X. Design QoS-aware multi-path provisioning strategies for efficient cloud-assisted SVC video streaming to heterogeneous clients[J]. IEEE transactions on multimedia, 2013, 15(4): 758-768.

[22] XU H, LI B. Joint request mapping and response routing for geo-distributed cloud services[C]//Proc of IEEE INFOCOM, Turin, Italy, c2013: 854-862.

[23] NIU D, XU H, LI B, et al. Quality-assured cloud bandwidth auto-scaling for video-on-demand applications[C]//Proc of IEEE INFOCOM, Orlando, USA, c2012: 460-468.

[24] HE J, WU D, ZENG Y, et al. Toward optimal deployment of cloud-assisted video distribution services[J]. IEEE transactions on circuits and systems for video technology, 2013, 23(10): 1717-1728.

[25] WANG X, CHEN M. PreFeed: cloud-based content prefetching of feed subscriptions for mobile users[J]. IEEE systems journal, 2014, 8(1): 202-207.

[26] WANG X, KWON T T, CHOI Y, et al. Cloud-assisted adaptive video streaming and social-aware video prefetching for mobile users[J]. IEEE wireless communications, 2013, 20(3): 72-79.

[27] PAN B, WANG X, HONG C P, et al. Amvp-cloud: A framework of adaptive mobile video streaming and user behavior oriented video pre-fetching in the clouds[C]//Proc of IEEE International Conference on Computer and Information Technology, Chengdu, China, c2012: 398-405.

[28] SILVESTRE G, MONNET S, KRISHNASWAMY R, et al. Aren: a popularity aware replication scheme for cloud storage[C]//Proc of IEEE International Conference on Parallel and Distributed Systems, Singapore, c2012: 189-196.

[29] BORU D, KLIAZOVICH D, GRANELLI F, et al. Energy-efficient data replication in cloud computing datacenters[J].Cluster computing, 2015, 18(1): 385-402.

[30] WANG Z, SUN L, CHEN X, et al. Propagation-based social-aware replication for social video contents[C]//Proc of ACM Interna tional Conference on Multimedia, Nara, Japan, c2012: 29-38.

[31] WANG S, DEY S. Rendering adaptation to address communication and computation constraints in cloud mobile gaming[C]//Proc of IEEE Global Telecommunications Conference, Miami, USA, c2010: 1-6.

[32] CHOY S, WONG B, SIMON G, et al. A hybrid edge-cloud architecture for reducing on-demand gaming latency[J]. Multimedia systems, 2014, 20(5): 503-519.

[33] ISLAM S, GREGOIRE J C. Giving users an edge: a flexible cloud model and its application for multimedia[J]. Future generation computer systems, 2012, 28(6): 823-832.

[34] LI Z, HUANG Y, LIU G, et al. Cloud transcoder: bridging the format and resolution gap between internet videos and mobile devices[C] //Proc of ACM International Workshop on Network and Operating System Support for Digital Audio and Video, Toronto, Canada, c2012: 33-38.

[35] LI Z, ZHU X, GAHM J, et al. Probe and adapt: Rate adaptation for http video streaming at scale[J]. IEEE journal on selected areas in communications, 2014, 32(4): 719-733.

[36] JIANG J, SEKAR V, ZHANG H. Improving fairness, efficiency, and stability in http-based adaptive video streaming with festive[C]//Proc of ACM International Conference on Emerging Networking Experiments and Technologies, Nice, France, c2012: 97-108.

[37] CHEN J, MAHINDRA R, KHOJASTEPOUR M A, et al. A scheduling framework for adaptive video delivery over cellular networks[C]//Proc of ACM International Conference on Mobile Computing and Networking, Miami, USA, c2013: 389-400.

[38] JOSEPH V, DE VECIANA G. NOVA: QoE-driven optimization of DASH-based video delivery in networks[C]//Proc of IEEE INFOCOM, Toronto, Canada, c2014: 82-90.

[39] DE VLEESCHAUWER D, VISWANATHAN H, BECK A, et al. Optimization of HTTP adaptive streaming over mobile cellular networks[C]//Proc of IEEE INFOCOM, Turin, Italy, c2013: 898-997.

[40] EL ESSAILI A, SCHROEDER D, STEINBACH E, et al. QoE-Based traffic and resource management for adaptive http video delivery in lte[J]. IEEE transactions on circuits and systems for video technology, 2015, 25(6): 988-1001.

[41] PU W, ZOU Z, CHEN C W. Video adaptation proxy for wireless dynamic adaptive streaming over HTTP[C]//Proc of IEEE Packet Video Workshop, Munich, Germany, c2012: 65-70.

[42] YAN Z, XUE J, CHEN C W. QoE continuum driven HTTP adaptive streaming over multi-client wireless networks[C]//Proc of IEEE International Conference on Multimedia and Expo, Chengdu, China, c2014: 1-6.

[43] YAN Z, CHEN C W, LIU B. Admission control for wireless adaptive HTTP streaming: An evidence theory based approach[C]// Proc of ACM International Conference on Multimedia, Orlando, USA, c2014: 893-896.

[44] YAN Z, WESTPHAL C, WANG X, et al. Service provisioning and profit maximization in network-assisted adaptive HTTP streaming[C]//Proc of IEEE International Conference on Image Processing, Quebec, Canada, c2015: 2786-2790.

[45] YAN Z, XUE J, CHEN CW. Prius: hybrid edge cloud and client adaptation for HTTP adaptive streaming in cellular networks[J]. IEEE transactions on circuits and systems for video technology, 2016.

[46] MULLER C, LEDERER S, TIMMERER C. An evaluation of dynamic adaptive streaming over HTTP in vehicular environments[C]//Proc of ACM Workshop on Mobile Video, Chapel Hill, USA, c2012: 37-42.

[47] LIU C, BOUAZIZI I, GABBOUJ M. Rate adaptation for adaptive HTTP streaming[C]//Proc of ACM Conference on Multimedia Systems, San Jose, USA, c2011: 169-174.

[48] AGGARWAL V, JANA R, PANG J, et al. Characterizing fairness for 3G wireless networks[C]//Proc of IEEE Workshop on Local and Metropolitan Area Networks, Chapel Hill, USA, c2011: 1-6.

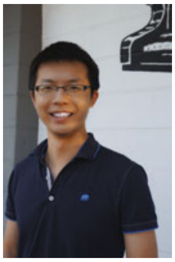
[49] XUE J, ZHANG DQ, YU H, et al. Assessing quality of experience for adaptive http video streaming[C]//Proc of IEEE International Conference on Multimedia and Expo Workshops, Chengdu, China, c2014: 1-6.

[50] BADDELEY A D. Essentials of human memory[M]. Lodon: Psychology Press, 1999.

[51] BROWN R G. Forecasting and prediction of discrete time series[M]. Courier Corporation, 2004.

[52] EL ESSAILI A, SCHROEDER D, STAEHLE D, et al. Quality-of-experience driven adaptive http media delivery[C]//Proc of IEEE International Conference on Communications, Budapest, Hungary, c2013: 2480-2485.

About the authors



YAN Zhisheng [corresponding author] is a Ph.D. student at Computer Science and Engineering Department, State University of New York at Buffalo. He received his B.S. and M.S. degrees from Shandong University and University of Science and Technology of China in 2010 and 2013, respectively. His research interests lie in the perception, processing and networking of multimedia content. Currently, his research is focused on mobile HTTP adaptive streaming and energy-saving mobile display. (Email: zyan3@buffalo.edu)

CHEN Changwen received his B.S. from University of Science and Technology of China in 1983, MSEE from University of Southern California in 1986, and Ph.D. from University of Illinois at Urbana-Champaign in 1992. He is currently an Empire Innovation Professor of Computer Science and Engineering at the University at Buffalo, State University of New York. He was Allen Henry Endow Chair Professor at the Florida Institute of Technology from July 2003 to December 2007. He was on the faculty of Electrical and Computer Engineering at the University of Rochester from 1992 to 1996 and on the faculty of Electrical and Computer Engineering at the University of Missouri-Columbia from 1996 to 2003.

He has been the Editor-in-Chief for IEEE Trans. Multimedia since January 2014. He has also served as the Editor-in-Chief for IEEE Trans. Circuits and Systems for Video Technology from 2006 to 2009. He has been an Editor for several other major IEEE Transactions and Journals, including the Proceedings of IEEE, IEEE Journal of Selected Areas in Communications, and IEEE Journal on Emerging and Selected Topics in Circuits and Systems. He has served as Conference Chair for several major IEEE, ACM and SPIE conferences related to multimedia, video communications and signal processing. His research is supported by NSF, DARPA, Air Force, NASA, Whitaker Foundation, Microsoft, Intel, Kodak, Huawei, and Technicolor.

He and his students have received 8 Best Paper Awards or Best Student Paper Awards over the past two decades. He has also received several research and professional achievement awards, including the Sigma Xi Excellence in Graduate Research Mentoring Award in 2003, Alexander von Humboldt Research Award in 2010, and the State University of New York at Buffalo Exceptional Scholar-Sustained Achievement Award in 2012. He is an IEEE Fellow and an SPIE Fellow. (Email: chencw@buffalo.edu)