Visual computing and platform services in multimedia clouds: Challenges and solutions

D.A. Milovanovic, Z.S. Bojkovic

***Abstract*—**This paper focuses on recent advances in cloud computing for multimedia. Multimedia computing has emerged as technology to generate, edit, process, and search media contents (such as images, video, audio, graphics) in order to provide rich media services. Several fundamental challenges for multimedia computing in the cloud are highlighted as multimedia/service/network/QoS/device heterogeneity. Key research of cloud-assisted visual computing and streaming include encoding/decoding with cloud computing resources, and transcoding with balanced cloud and edge resources for mobile media. The optimization problem is identified as three-way trade-off between the caching, transcoding, and bandwidth costs. We outline a mobile cloud-based platform solution for live transcoding and streaming-as-a-service using the standard MPEG-DASH dynamic adaptive streaming over HTTP and adaptive client framework.

***Keywords*—**multimedia communications, media processing, mobile cloud computing

I. INTRODUCTION

Cloud computing is computation and resource provisioning technology exploited for a wide range of multimedia applications, including adaptive media processing, storage and transmission. The interest in cloud computing for multimedia has recently increased dramatically. As cloud computing offers low cost, high scalability, enhanced reliability, and device independence, it can be used to efficiently deploy multimedia services. Cloud system incorporates infrastructure, platform, and software as services. Performing processing and storage on the cloud reduces the demands on Internet user devices, especially mobile devices, which have limited energy, storage, and computational capability [1, 2].

Cloud computing has become a popular phrase since 2007. Framework of cloud computing is divided into three layers: ***IaaS*** infrastructure layer (infrastructure as a service includes resources of computing and storage), ***PaaS*** platform layer (platform as a service considered as a core layer in the cloud computing system, which includes the environment of parallel programming design, distributed storage and management system), and ***SaaS*** application layer (software as a service provides some simple software and applications, as well as costumer interfaces to end users). Users use client software or a browser to call services from providers through the Internet, and pay costs according to the utility business model. The core features of cloud computing are *virtualization* (provides resource pool where all bottom layer hardware devices is virtualized in order to improve the efficiency to use resources), *reliability, usability and extensibility; large-scale* (cloud computing system normally consists of thousands of servers and PCs) and *autonomy* (system automatically configures and allocates the resources of hardware, software and storage to clients on-demand, and the management is transparent to end users).

As a development and extension of cloud computing and mobile computing, **mobile cloud computing**, as a new phrase, has been devised since 2009. The main objective is to provide a convenient and rapid method for users to access and receive data from the cloud, accessing cloud computing resources effectively by using mobile devices. The major comes from the characters of mobile devices and wireless networks, as well as their own restriction and limitation. In mobile cloud computing environment, the limitations of mobile devices, quality of wireless communication, types of application, and support from cloud computing to mobile are all important factors that affect assessing from cloud computing [3, 4, 5].

The paper is organized as follows. Challenges of media processing and multimedia computing in multimedia-aware cloud and cloud-aware multimedia systems are highlighted in the first part. Next, a standard solution of video transcoding and adaptive streaming based on MPEG DASH is presented as cloud-based media platform services.

II. MULTIMEDIA CLOUD COMPUTING

Multimedia cloud computing is an emerging area that involves services on the cloud, multimedia communications and applications. Cloud is referred to a shared pool of configurable computing resources that can allow ubiquitous on-demand access (Fig. 1). For example, a straightforward use is to rely the powerful, scalable computing power of cloud to facilitate multimedia content encoding and processing. Next, the cloud can be used to speed up and/or improve the quality of multimedia communications. The provisioning of cloud infrastructure, resource allocation, network routing, and QoE management are important challenges. Media applications in the cloud can be conducted either completely or partially in the cloud. In the former case, the cloud will do all the multimedia computing. In the latter case, the key problem is client–cloud resource partition for multimedia computing [1].

D.A.Milovanovic is with the University of Belgrade, Studentski Trg 1, 111000 Belgrade, Republic of Serbia (e-mail: dragoam@gmail.com).

Z.S.Bojkovic is with the University of Belgrade, Studentski Trg 1, 111000 Belgrade, Republic of Serbia (e-mail: z.bojkovic@yahoo.com).

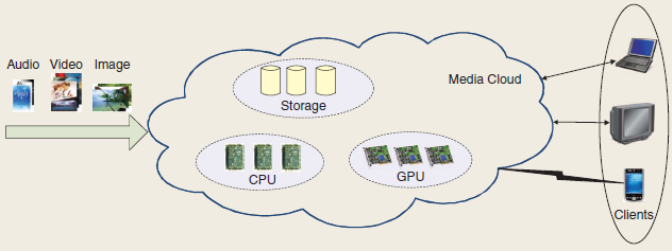

Fig. 1 Fundamental concept of multimedia cloud computing [1].

Multimedia processing in a cloud imposes great challenges [6, 7, 8]:

Multimedia and service heterogeneity. The cloud will support different types of multimedia and multimedia services for millions of users simultaneously (video conferencing, photo sharing and editing, multimedia streaming, image search, image-based rendering, video transcoding and adaptation, and multimedia content delivery).

QoS heterogeneity. The cloud will provide QoS provisioning and support for various types of multimedia services to meet different multimedia QoS requirements (different multimedia services have different QoS requirements).

Network heterogeneity. The cloud will adapt multimedia contents for optimal delivery to various types of devices with different network bandwidths and latencies (Internet, wireless local area network, and third generation wireless network, have different network characteristics, such as bandwidth, delay, and jitter).

Device heterogeneity. The cloud will have multimedia adaptation capability to fit different types of devices, including CPU, GPU, display, memory, storage, and power (TVs, personal computers, and mobile phones, have different capabilities for multimedia processing,).

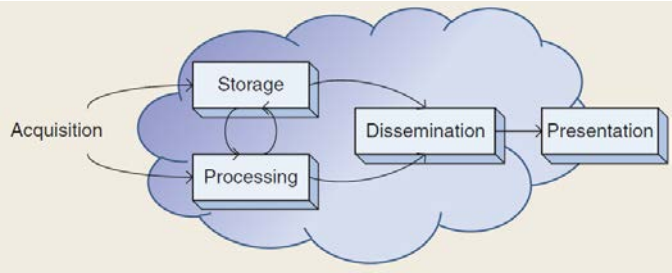

Fig. 2 A typical media life cycle [1].

In a *media-centric view*, **cloud-aware multimedia** addresses how multimedia services and applications, such as storage and sharing, authoring and mashup, adaptation and delivery, and rendering and retrieval, can optimally utilize cloud-computing resources to achieve better quality of experience (QoE). The emergence of cloud computing will have a profound impact on the entire life cycle of multimedia contents. As shown in Figure 2, a typical media life cycle is composed of acquisition, storage, processing, dissemination, and presentation. Before the cloud-computing era, media storage, processing, and dissemination services were provided by different service providers with their proprietary server farms. Now, various service providers have a choice to be users of public clouds. The *pay-as-you-go* model of a public cloud would greatly facilitate small businesses and multimedia fanciers. For small businesses, they pay just for the computing and storage they have used, rather than maintaining a large set of servers only for peak loads. For individuals, cloud utility can provide a potentially unlimited storage space and is more convenient to use than buying hard disks.

In a *cloud-centric view*, **multimedia-aware cloud** addresses how a cloud can perform distributed multimedia processing and storage and provide quality of service (QoS) provisioning for multimedia services. The media cloud needs to have the following functions: QoS provisioning and support for various types of multimedia services with different QoS requirements, distributed parallel multimedia processing, and multimedia QoS adaptation to fit various types of devices and network bandwidth. Figure 3 depicts the relationship of the media cloud and cloud media services. More specifically, the media cloud provides raw resources, such as hard disk, CPU, and GPU, rented by the media service providers (MSPs) to serve users. MSPs use media cloud resources to develop their multimedia applications and services, e.g., storage, editing, streaming, and delivery.

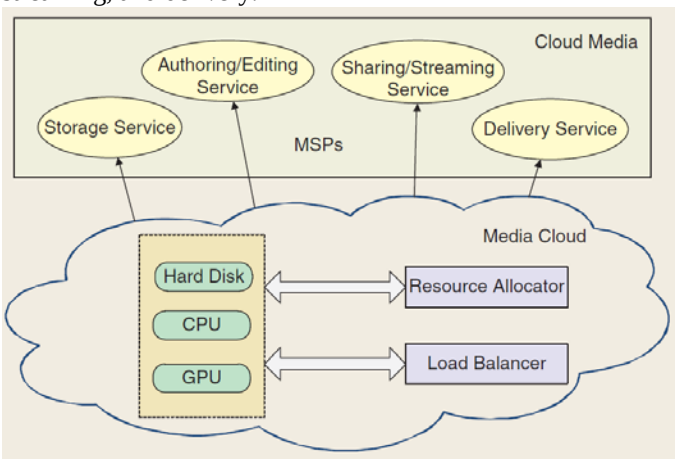

Fig. 3 The relationship of the media cloud and cloud media services [1].

Video adaption in a media cloud will take charge of collecting customized parameters, such as screen size, bandwidth, and generating various versions according to their parameters either offline or on the fly (Fig. 4).

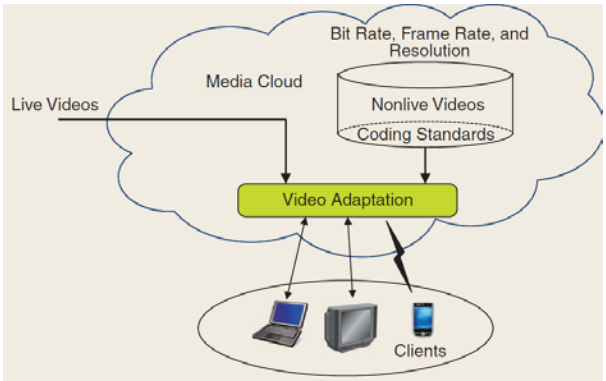

Fig. 4 Cloud-based video adaptation and transcoding [9].

A. *Transcoding in the cloud*

A simplified view of the main components of a typical cloud streaming chain is shown in Figure 5. The implementation of *rate-adaptive streaming* first requires a *transcoding* operation. For each video, the video provider would generate different representations, each of them characterized by a different bit-rate, resolution, sometimes frame rate and key frame period. Then, the video provider should package the video by aggregating the set of representations, by segmenting the representations and by creating the manifest file, which is the file that describes the playlist [10].

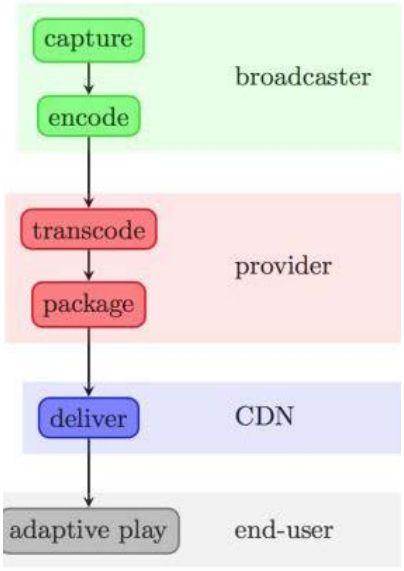

Fig. 5 The main elements of cloud streaming system.

One of the most prominent trends in cloud-assisted media **transcoding** research is to adopt the *MapReduce* framework for a parallel processing infrastructure. Specifically, the whole transcoding task, for example, in the unit of group of pictures (GOP), is decoupled into a set of parallel tasks, which are in turn mapped into a set of virtual machines for processing. It is expected that the interplay between the *Hadoop* framework and the transcoding algorithm would play an important role in further optimizing the cloud-assisted transcoder design. This vertical integration between computing and transcoding is an area of importance [2].

Media **encoding/decoding**, owing to its high computational complexity, stands out as a nature candidate for task offloading from mobile devices to the cloud infrastructure. The most common approach is to leverage the MapReduce paradigm for parallel processing, coupled with dynamic resource provisioning in the cloud for cost effectiveness. However, in cloud mobile media, encoding/decoding task offloading imposes additional challenges to the network connectivity between the mobile device and the cloud infrastructure. The gain in computing in the cloud could be offset by the additional bandwidth requirement in the transmission. As such, cloud-assisted media encoding/decoding design should be jointly optimized with network dynamics [2].

B. *Adaptive media streaming*

Media streaming refers to the process of transferring contents from the media cloud to the media outlets (e.g., smartphones and tablets). This process is often adaptive, in response to varying conditions in wireless channel, screen size, user preference, and resource availability, with an dual objective to provide better QoS/QoE and improve resource utilization.

The end users consume videos in different versions for different devices. Those video segments are originally published by the origin server owned by the content providers. They are transmitted to the users upon request via edge servers, which can strategically cache or transcode video segments to reduce the total cost of operating the adaptive streaming services within media cloud (Fig. 6).

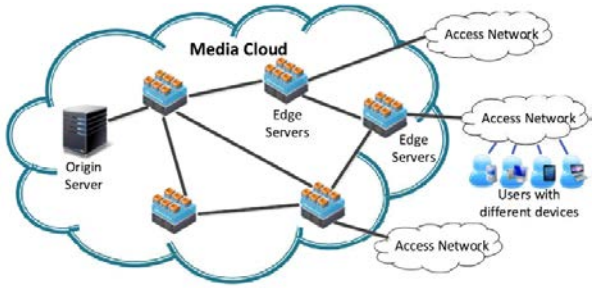

Fig. 6 End-to-end view of on-demand adaptive video streaming via media cloud [2].

Recently developed solution is to consider HTTP-based adaptive streaming (HAS), with extension beyond traditional Internet-based streaming in both server and client ends. It is assumed that under the emerging cloud mobile media paradigm, media servers can be placed inside the cloud while the consuming client uses mobile devices for media playout. However, extension to both cloud server and mobile terminals faces significant challenges. HAS essentially operates under the client-server architecture, which prevents it from taking full advantage of the abundant replicated video resources in the cloud. The MPEG DASH scheme can effectively solve the data scheduling problem by using multi-scale allocation of scalable coded video and network coding while imposing very light load onto HAS servers. It can also resolve the rate adaptation problem by introducing a multi-scale rate prediction to adapt the video bit rate to the inherent bandwidth dynamics of each server. Extension of MPEG DASH to the mobile terminal receivers faces a different set of challenges.

III. MOBILE CLOUD-BASED MEDIA PLATFORM SERVICES

In this section, we investigate the list of cloud-based media platform services [11, 12, 13, 14].

Media representation. Cloud-assisted media representation has been an active area of research. Key research thrusts in this category include *encoding/decoding* with cloud computing resources for mobile media, and *transcoding* with balanced cloud and edge resources for mobile media.

Media distribution refers to the process of moving media

contents from their sources, via a distribution network, to their consumers. This process can be logically decoupled into three sequential steps, including *content acquisition* from generation devices (e.g., smartphones or cameras), *content distribution* across content delivery networks, and *media streaming* to mobile devices or other media outlets.

Media adaptation service refers to algorithms and mechanisms that modify the semantic meaning of media contents. Typical adaption domains include, but are not limited to, *media metadata*, *media mashup*, and *media rendering*.

Media analytics. Recent years have witnessed an explosive growth of multimedia data and metadata, due to higher processor speeds, faster networks,wider availability ofmass-storage devices, and pervasive penetration of mobile devices. The enormous scale of multimedia data imposes great challenges in multimedia retrieval and mining. At the same time the explosion of the amount of data, number of mobile users, and availability of new resources (e.g., cloud computing) would lead to greater expectations for multimedia analytics, in terms of effectiveness and efficiency, for which existing analytics approaches and systems typically do not suffice. In this subsection, we present a brief survey of media analytic services with cloud support, categorized into the following three domains: *content analysis* and *metadata mining*, *content recommendation*, and *media retrieval.*

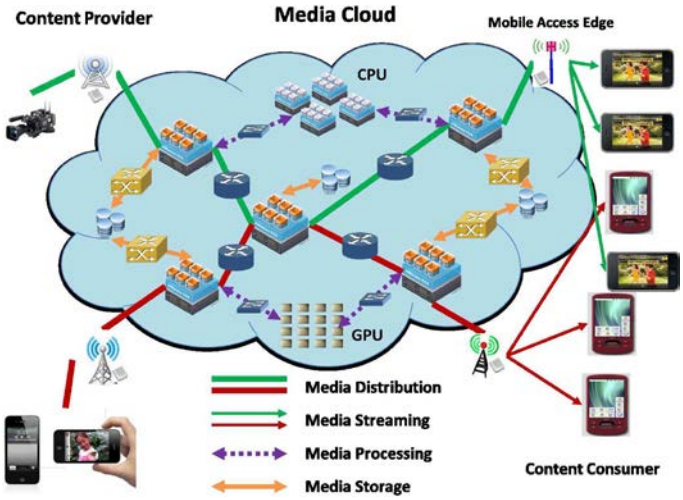

Fig. 7 An end-to-end view of a cloud mobile media network [2].

In a solution for cloud mobile media network, content providers, media cloud service providers, mobile access cloud edge (possibly integrated with service providers) and content viewers are interconnected via an underlying network infrastructure supported by ICT resources (Fig. 7). The rapid increase in the use of multimedia applications on mobile devices has led IT companies to evolve their technologies to cope with the multimedia requirements. Since mobile devices are inherently resource-limited, mobile cloud computing is emerging as a promising technology to enhance the capability of mobile devices to enable rich multimedia applications. The fundamental tension between resource-hungry multimedia streams and power-limited mobile devices has to be resolved, and is complicated by novel ways of operating mobile devices as both media clients and content providers. The objective is to minimize the total operational cost by optimally orchestrating multiple resources. The optimization problem is formulated by examining a three-way tradeoff between the caching, transcoding, and bandwidth costs, at each edge server [15, 16, 17].

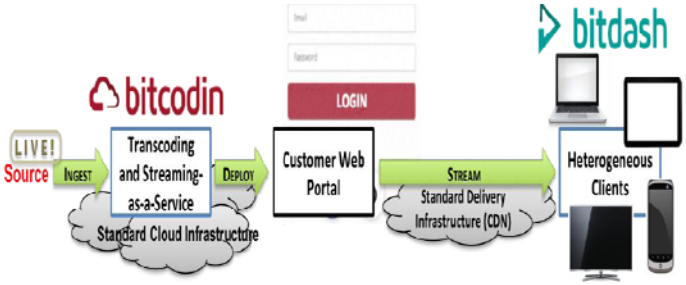

Fig. 8 High-level system architecture for live transcoding and streaming-as-a-Service [18].

The transcoding and streaming-as-a-service takes the live multimedia content as an input and transcodes it to multiple content representations in real-time on standard infrastructure-as-a-service (IaaS) cloud environments according to the requirements of the customer in terms of resolutions, bitrates, etc (Fig. 8). These requirements are expressed through an application programming interface (API) exposed to the customer. The resulting manifest describing the individual content representations and primary input for the streaming client is incorporated within the customer's Web portal offering the service to the actual clients (end users). The streaming is conducted utilizing standard CDN infrastructure. The heterogeneous clients request the multimedia segments based on the manifest received prior to the streaming and adapt themselves to the context conditions such as fluctuating network bandwidth.

A solution *bitcodin* is live transcoding and streaming-as-a-service platform using the MPEG-DASH standard which is used for both live 24/7 and event-based temporary services and *bitdash*, adaptive client framework. It comprises the following:

- the actual transcoding and streaming-as-a-service deployed on standard cloud infrastructure (e.g., Google, Amazon, Microsoft, etc.) taking the live source as input and providing multiple representations (e.g., resolution, bitrate, etc.) according to the MPEG-DASH standard as output (MPEG-DASH standard defines the media presentation description (MPD) as well as segment formats and deliberately excludes the specification of the client behavior, i.e., the implementation of the adaptation logic, which determines the scheduling of the segment requests, is left open for competition);
- the integration within the customer Web portal for the actual deployment;
- the streaming utilizing standard delivery infrastructure over a content distribution network (CDN); and
- the DASH client implementation integrated within heterogeneous devices [18].

IV. CONCLUDING REMARKS

In summary, for multimedia computing in a cloud, the key is how to provide QoS provisioning and support for multimedia applications and services over the (wired) Internet and mobile Internet. More specifically, in mobile media applications and services, because of the power requirement for multimedia and the time-varying characteristics of the wireless channels, QoS requirements in cloud computing for mobile multimedia applications and services become more stringent than those for the Internet cases.

Simultaneous bursts of multimedia data access, processing, and transmission in the cloud would create a bottleneck in a general-purpose cloud because of stringent multimedia QoS requirements and large amounts of users' simultaneous accesses at the Internet scale. Therefore, using a general-purpose cloud in the Internet to deal with multimedia services may suffer from unacceptable media QoS/QoE. Mobile devices have limitations in memory, computing power, and battery life; thus, they have even more prominent needs to use a cloud to address the tradeoff between computation and communication.

Cloud-assisted media representation has been an active area of research. Key research thrusts in this category include *encoding and decoding* with cloud computing resources for mobile media, and *transcoding* with balanced cloud and edge resources for mobile media. A common approach in this line of research is to leverage the emerging parallel programming paradigm (i.e., MapReduce/Hadoop) to provide an improved media processing capability. Under this framework, research challenges range from parallel algorithm design for cloud computing, to fundamental design trade-offs (e.g., computation complexity vs. media distortion, encoding performance vs. energy efficiency, distortion vs. delay). The objective is to minimize the total operational cost by optimally orchestrating multiple resources. The optimization problem is formulated by examining a three-way tradeoffs. Finaly, we outline a mobile cloud-based platform solution for live transcoding and streaming-as-a-service using the standard MPEG-DASH dynamic adaptive streaming over HTTP and adaptive client framework.

REFERENCES

[1] W.Zhu, C.Luo,J.Wang, S.Li, "Multimedia cloud computing", *IEEE Signal Processing Magazine*, vol.28, no.3, pp.59-69, May 2011.

[2] Y.Wen, X.Zhu, J.J.P.C.Rodrigues, C.W.Chen, "Cloud mobile media: Reflection and outlook", *IEEE Trans. Multimedia*, vol.16, no.4, pp.885-902, June 2014.

[3] Z.Bojković, D.Milovanovic, "Mobile cloud analytics in Big data era", in Proc. 19th *International Conference on Computer Engineering and Applications* (CEA'16), Feb. 2016, Barcelona (invited paper), pp.91-94.

[4] Z.Bojkovic, D.A.Milovanovic, "Big data networking and platforms: Design and applications", presented at the Seminar at College of engineering, Karunagappally (IHRD), Kollam, Kerala, India, July 2016.

[5] D.Milovanovic, D.Kukolj, "Recent advances in UHD video coding technology", *IEEE Communications society, Multimedia communications Technical committee*, MMTC Communications - Frontiers, Vol.11, No.1, pp.50-55, Jan. 2016.

[6] K-T.Chen, A.C.Begen, C-F.Lai, (Eds), Emerging topics: *Special issue on cloud-aware multimedia systems, IEEE Communications society Multimedia communications Technical committee* MMTC E-LETTER vol.10, no.6, Nov.2015. Avalable online: *http://www.comsoc.org/~mmc*

[7] N.J.Sarhan, (Edt), Emerging topics: *Special issue on cloud computing for multimedia, IEEE Communications society Multimedia communications Technical committee* MMTC E-LETTER vol.8, no.6, Nov.2013. Avalable online: *http://www.comsoc.org/~mmc*

[8] Z.Li, C.Wu (Eds), Emerging topics: *Special issue on multimedia and cloud computing, IEEE Communications society Multimedia communications Technical committee* MMTC E-LETTER vol.8, no.1, Jan.2013. Avalable online: *http://www.comsoc.org/~mmc*

[9] G.Chan, P.Frossard, A.Vetro (Eds), *Distributed image processing, IEEE Signal Processing Magazine*, vol.28, no.3, May 2011.

[10] R.A-Pardo, G.Simon, A.Blanc, "Transcoding in the cloud: optimization and perspectives", *IEEE Communications society* MMTC E-LETTER vol.10, no.6, pp.12-15, Nov.2015.

[11] G.Bai, H.Shen, Y.Jin, D.D'Agostino, N.Achir, F.Grimaccia (Eds), "Mobile multimedia cloud computing", EURASIP *Journal on Wireless Communications and Networking*, 2016. Avalable online: *https://www.springeropen.com/collections/mmcc*

[12] H.T.Dinh, C.Lee, D.Niyato, P.Wang, "A survey of mobile cloud computing: architecture, applications, and approaches", *Wireless communications and mobile computing*, Wiley, 2011.

[13] N.Fernando, S.W.Loke, W.Rahayu, "Mobile cloud computing: A survey", *Future generation computer systems*, Elsevier, vol.29, no.1, pp.84-106, Jan. 2013.

[14] H.Qi, A.Gani, "Research on mobile cloud computing: Review, trend and perspectives", in Proc. Int. Conf. on *Digital Information and Communication Technology and it's Applications* (DICTAP), May 2012.

[15] J.Liu, W.Zhu, T.Ebrahimi, J. Apostolopoulos, X-S.Hua, C.Wu (Eds), *Introduction to the Special section on visual computing in the cloud: Fundamentals and applications, IEEE Trans. Circuits and systems for video technology*, vol.25, no.12, Dec.2015.

[16] X.Song, X.Peng, J.Xu, G.Shi, F.Wu, "Cloud-based distributed image coding", *IEEE Trans. Circuits and systems for video technology*, vol.25, no.12, pp.1926-1940, Dec.2015.

[17] Y.Jin, Y.Wen, C.Westphal, "Optimal transcoding and caching for adaptive streaming in media cloud: An analytical approach", *IEEE Trans. Circuits and systems for video technology*, vol.25, no.12, pp.1914-1925, Dec.2015.

[18] C.Timmerer, D.Weinberger, M.Smole, R.Grandl, C.Muler, S.Lederer, "Cloud-based transcoding and adaptive video Streaming-as-a-Service", *IEEE Communications society* MMTC E-LETTER vol.10, no.6, pp.7-11, Nov.2015.