Cloud Mobile Media: Reflections and Outlook

Yonggang Wen, *Senior Member, IEEE*, Xiaoqing Zhu, *Member, IEEE*, Joel J. P. C. Rodrigues, *Senior Member, IEEE*, and Chang Wen Chen, *Fellow, IEEE*

***Abstract*—This paper surveys the emerging paradigm of *cloud mobile media*. We start with two alternative perspectives for cloud mobile media networks: an end-to-end view and a layered view. Summaries of existing research in this area are organized according to the layered service framework: i) cloud resource management and control in infrastructure-as-a-service (IaaS), ii) cloud-based media services in platform-as-a-service (PaaS), and iii) novel cloud-based systems and applications in software-as-a-service (SaaS). We further substantiate our proposed design principles for cloud-based mobile media using a concrete case study: a cloud-centric media platform (CCMP) developed at Nanyang Technological University. Finally, this paper concludes with an outlook of open research problems for realizing the vision of cloud-based mobile media.**

***Index Terms*— Cloud Computing, Cloud Media, Mobile Media, Content Distribution Network, Quality of Experience, Cloud-Centric Media Network and Media Analytics.**

I. Introduction

GROWING popularity of mobile devices (e.g., smartphones and tablets), together with ubiquitous wireless Internet, has fueled an increasing user demand on rich media experience on the go. This trend, in turn, is triggering an exponential growth of mobile traffic, dominated by video contents. According to [1], mobile video will increase 25-fold between 2011 and 2016, accounting for over 70% of total mobile data traffic by 2016. However, user experience with mobile video is severely constrained by three fundamental challenges. First, the limited on-board resources of mobile devices are ill-fitted for intense media coding and processing tasks [2]. Second, the inherently time-varying and unreliable wireless channel limits the communication bandwidth between mobile devices and back-end content delivery systems. Finally, the relatively static mechanism of system resource provision in existing back-end systems cannot react fast enough to flash-crowd demands for popular content [3]. This calls for the design of a new paradigm for resolving the tussle between the growing demand for mobile media applications and the aforementioned limitations of existing media delivery networks.

The guiding principle of mobile media network design, following a general rule for network architecture suggested by Clark [4], is driven by a fundamental *trade-off* between network cost and quality of service (QoS). On one hand, the cost—including an initial capital expenditure (CAPEX) and a re-occurring operating expenditure (OPEX)—to build and operate a mobile media network should be kept low. Such cost borne by service providers will ultimately translate into the service price, which would seriously affect the penetration of media services among mobile Internet users. On the other hand, end users demand an enhanced quality of service (or, quality of experience) for what they pay for. Balancing this design tussle requires new ideas and/or emerging technologies for the next-generation mobile media network.

Recently, the emerging cloud-computing technology [5] offers a natural solution to reduce the cost of deploying and operating mobile media networks. Under the cloud-computing paradigm, system resources can be allocated dynamically to meet the elastic application demand in a real-time manner. For example, computing resources (CPUs or GPUs) in data centers can be instantiated into virtual machines (VMs), whose capacity can be dynamically configured for specific media applications (e.g., transcoding, rendering, etc.). As such, the emerging cloud-computing paradigm has started to transform mobile media experience, resulting in a new area of research, *cloud mobile media* [6], [7], [8], [9]. Under this new paradigm, the cost reduction comes inherently from cloud computing.

However, the emerging paradigm of cloud mobile media posits new technical challenges. For example, cloud computing platforms are often built upon off-the-shelf equipments whose performance and reliability are lagging behind specially designed carrier-grade media systems [10]. Moreover, security and privacy concerns [11], [12], [13] are pervasive in cloud computing. Finally, these challenges originated from cloud computing are further complicated by the mobility management for mobile devices and users. In this paper, we suggest a list of technical challenges for cloud mobile media network:

- *Scalability*: the system should be able to handle a large number of contents, users and devices.

Manuscript received September 21, 2013; revised January 23, 2014; accepted March 20, 2014. Date of publication April 03, 2014; date of current version May 13, 2014. The work of Y. G. Wen was supported by NTU Start-up Grant, MOE Tier-1 Grant (RG 31/11), EIRP02 Grant from Singapore EMA, and the Singapore National Research Foundation under is IDM Futures Funding Initiative and administered by the Interactive & Digital Media Programme Office, Media Development Authority. The work of J. Rodrigues was supported in part by *InstitutodeTelecomunicações*, Next Generation Networks and Applications Group (NetGNA), Covilhã Delegation, by National Funding from the FCT - *FundaçãoparaaCiênciaeTecnologia* through the Pest-OE/EEI/LA0008/2013 Project, and by the AAL4ALL (Ambient Assisted Living for All), project co-funded by COMPETE under FEDER via QREN Programme. The work of C. W. Chen was supported in part by US NSF Grants 0915842 and 0964797, the Gift Grants from Microsoft Research and Huawei Technologies, Inc. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Charles D. (Chuck) Creusere.

Y. G. Wen is with the School of Computer Engineering, Nanyang Technological University, Singapore (e-mail: ygwen@ntu.edu.sg).

X. Q. Zhu is with Advanced Architecture & Research, Cisco Systems Inc., San Jose, CA 95134 USA (e-mail: zhuxq@alumni.stanford.edu).

J. J. P. C. Rodrigues is with Instituto de Telecomunicações, University of Beira Interior, Covilhã, Portugal (e-mail: joeljr@ieee.org).

C. W. Chen is with the State University of New York, Buffalo, NY 14260 USA (e-mail: chencw@buffalo.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2014.2315596

- *Heterogeneity*: the system should be able to accommodate contents in diverse formats, users with diverse preferences, and devices with diverse forms. Networks are also heterogenous.
- *Reliability*: the system should be designed with calculated redundancy to offer almost non-interruptive services in presence of system failures, as well as issues related to unreliable wireless channels.
- *Usability*: the system should be designed to make it convenient for all possible users with a wide range of technology capabilities. The user interface should be intuitive, easy to learn, and tailored to mobile devices with limited interactive options.
- *Security*: Digital rights management (DRM) and privacy are serious concerns in any cloud mobile media solutions.

In this article, we report a literature survey (date to late 2012) of the cloud mobile media networks, targeted for well-informed researchers in cloud computing, media systems, and mobile computing. We first propose a structured decomposition of a cloud mobile media system, including two alternative views (i.e, an end-to-end view and a layered view). This architectural principle is then used as a framework to survey existing efforts in this research area, ranging from resource management and control, media platform services, cloud systems and applications, to system verticals (privacy, security, and economics). We further substantiate our framework with a proof-of-concept (POC) cloud-centric media platform, to demonstrate its feasibility and effectiveness. The research on cloud-based mobile media is still at its infancy and the experience of mobile media can be substantially enhanced with sophisticated technologies and solutions. Finally, we suggest a list of potential research problems and issues to be investigated.

The rest of this paper is organized as follows. In Section II, we present two alternative views of cloud mobile media network, including a layered view and an end-to-end view. The layered approach will serve as the framework to present existing efforts on cloud mobile media research. In Section III, we survey the existing research in the area of resource management and control for cloud mobile media network. In Section IV, we present a structured view of current research thrusts in cloud-based media services. In Section V, we sample a list of cloud-based systems and applications in the area of cloud mobile media, and briefly discuss the verticals in designing cloud mobile media networks, including privacy, security, and economics. Following this survey, we introduce in Section VI a proof-of-concept prototype of a cloud-centric media platform (CCMP), developed at Nanyang Technological University. In Section VII, we present a systematic outlook for future research topics in cloud mobile media networks.

II. Cloud Mobile Media Network: A System Perspective

Cloud mobile media network was envisioned to leverage the emerging cloud-computing technologies [14], [15], [16], [17] to enhance mobile media experience. Previous research [8] defined media cloud from either a cloud-centric view or a media-centric view (e.g., media-aware cloud and cloud-aware multimedia). In this paper, from a system perspective, we aim to decompose the cloud mobile media system into a set of participatory modules. Specifically, we present two alternative viewpoints for the cloud mobile media architecture, including an end-to-end perspective and a layered perspective. These two architectural viewpoints will serve as blueprints to survey existing research efforts and guide future system research.

A. An End-to-End Workflow Model

In Fig. 1, we illustrate a schematic end-to-end view of the cloud mobile media system. The system consists of three participatory stakeholders in the digital media value chain, including content providers, media cloud service providers, and content consumers. Moreover, unique to the mobile cloud media paradigm, a mobile cloud edge is included in the workflow to emphasize the daunting challenge of radio resource management in this end-to-end architecture.

Content providers are responsible for creating media contents for distribution and consumption. Media contents could be generated by professional producers with sophisticated digital cameras, or regular Internet users who capture videos and/or images with their own (mobile) devices. In this paper, we are particularly interested in media contents generated by mobile devices (e.g., smartphones or tablets with camera). Media contents captured by these mobile devices present overwhelming technical challenges in processing, transmitting and analyzing them, for traditional media systems, demanding new solutions that would embrace latest advances in information and communication technology (ICT) domain, in particular, cloud-computing technologies.

Media cloud service providers pull together a pool of shared ICT resources, including computing, storage, and networking, and allocate them elastically for various media-related tasks in response to their real-time application demands. Computing capacity could come from a diverse set of resources, for instance, data centers that houses a fleet of general-purpose rack/blade servers of commercial grade and CPU/GPU arrays that are dedicated for image or video processing. These computation facilities are often sided with super-size storage capacity that are distributed across different locations and can be request on demand. The storage space could come from dense provisioning (e.g., storage area network) or sparse provisioning (built-in disks with servers). These computing and storage resources are interconnected by a network fabric to formulate a pool of system resources, as shown in Fig. 1.

This pool of ICT resources can be dynamically reconfigured to complete tasks in media networks, for example, media processing (encoding/decoding/trancoding), media distribution, media rendering and media analytics, to name a few. Compared to the static resource allocation in traditional media systems, the cloud-based media network can scale up and down to meet dynamic demand, with a reduced cost and a better QoS for media experience. For example, the cloud-based media network can better deal with the notorious flash-crowd phenomenon [3] in media systems, when a lot of users are interested in one unpredictably particular piece of content within a very short time. In such a case, any amount of statically allocated resources would be oversubscribed, resulting in a deteriorating user

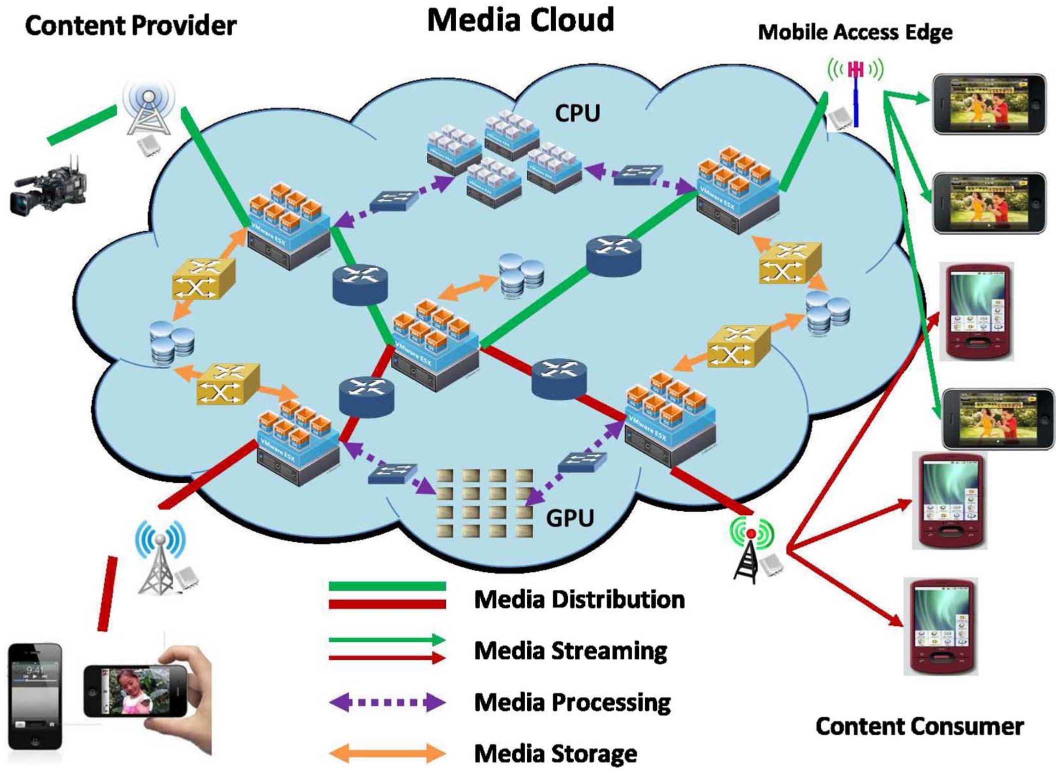


Fig. 1. An end-to-end view of a cloud mobile media network: content providers, media cloud service providers, mobile access cloud edge (possibly integrated with service providers) and content viewers are interconnected via an underlying network infrastructure supported by ICT resources.

experience. Moreover, it is also technically feasible to dedicate smaller clouds of ICT resources for specific media applications, offering niche services (e.g., content delivery [18]).

Content consumers watch videos on different media outlets (e.g., TV, laptop, smartphone, and tablet), via wireless Internet. The design for this use case faces a list of technical challenges, including:

- Mobile devices are inherently resource constrained.
- The connectivity exposed to mobile devices are usually inferior to their desktop counterparts.
- The expectations of mobile users are increasingly higher, because features like mobility support, interactive support, come by naturally in non-media related applications.

This tussle between limited resources and high demand renders mobile media experience with a decent QoS an unparalleled challenge. We are seeing more and more solutions that leverage the emerging cloud-computing technologies to provide additional system sources to enhance viewing experience over wireless Internet (cf. Section III).

Mobile cloud edge serves an important role in connecting resource-constrained mobile devices with the resource-rich cloud infrastructure. Examples of the mobile cloud edge include base stations, WiFi access points, and other wireless edge devices. A key component to enable seamless interaction between the cloud and the mobile devices is the access scheme via various wireless gateways. It is through such wireless gateways that the mobile devices are able to offload the limitation in computing and storage to the cloud. The current cloud mobile media systems adopt a variety of connection protocols as their wireless gateways, including WiFi, Bluetooth, WiMAX, and 3G/4G LTE. These wireless gateways are commonly used in mobile cloud computing to support the contemporary convergent computing that bridge between mobile computing and cloud computing [19]. For cloud mobile media applications, the demand for broadband access and continuous connectivity to ensure adequate quality-of-experiences (QoE) for mobile media consumers shall impose significant challenges. During certain extended media applications, unlike wired networks, the mobile users may be moving across several local wireless access cells and demanding seamless switching of media gateway (access point) from one cell to another.

Moreover, wireless gateways to cloud often consist of heterogeneous radio access networks. The heterogeneity of today's wireless access networks imposes additional challenges for efficient access and resource management across multiple radio access technologies [20]. An intelligent approach needs to be designed, so as to maintain an always-on high-quality broadband mobile connectivity by exploiting available mobile specific information in users' location, context, and request services [21]. An alternative approach is well illustrated by a new network concept–cognitive wireless cloud—that handles the spectrum sharing amongst these heterogeneous radio access networks [22]. In this conceptual system, both radio access networks and the user terminal are assumed to use common frequency bands and therefore need to have a cognitive spectrum sensing function to find the vacant frequency bands to operate. In addition to these innovative solutions proposed for implementation at the wireless gateways, an unconventional ap-

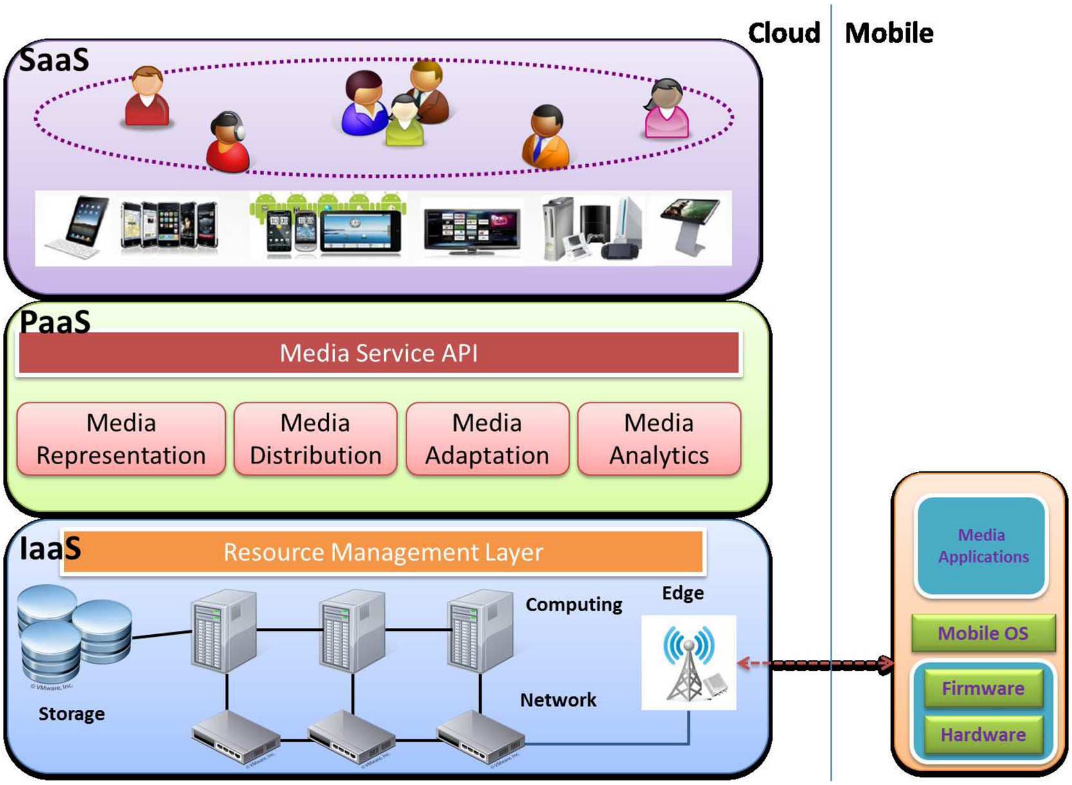


Fig. 2. A layered view of cloud mobile media system, consisting of three service models (IaaS, PaaS and SaaS). Note that the three layers have no binding as in the traditional network layering architecture. In this view, the three layers merely follow from the three service models in cloud computing.

proach to facilitate better gateway for mobile devices is to bring the cloud closer to mobile users. Instead of moving the entire collection of cloud facilities to mobile users, the approach of cloudlet [23], [24] advocates the use of a trusted, resource-rich computer or a cluster of computers that's well-connected to the Internet and available for use by nearby mobile devices. This cloudlet resembles a "data center in a box" capable of self-managing, Internet connectivity without WAN delays, and access control for device/client set up [23].

In addition to the networking function needed for the mobile devices, more and more mobile cloud edge devices are providing computing and storage resources that can be dynamically allocated for specific multimedia centric tasks. In [25], a cross-layer architecture consisting of a pair of proxies has been developed to offer seamless mobility support to wireless devices executing multimedia applications. These proxies can be considered as special mobile cloud edge servers that enable the adaptive and concurrent use of different network interfaces during the communications of multimedia applications. A cloud computing environment is used as the infrastructure to dynamically set up (and release) the proxies on the server-side, in accordance with the pay-as-you-go principle of cloud based services. In [26], a hierarchical video caching edge at the radio access network has been developed to achieve dual goals: (1) reducing the need to bring requested videos from cloud internal content delivery networks (CDNs), thus reducing overall backhaul traffic, and (2) improving video quality of experience and increasing network capacity to support more simultaneous video requests. With this hierarchical caching at cloud edge, the overall network capacity can be improved by enabling mobile media users from multiple cell sites to share caches at higher levels of the hierarchy, thereby improving overall cache hit ratio, without increasing the total cache size used. More recently, an integrated approach to media streaming via HTTP has been developed which is capable of significantly enhancing video streaming performance from cloud servers to mobile users via an innovative proxy design [27], [28]. The details of this approach will be reviewed in Section IV.

B. A Layered Service Model

The cloud mobile media system, inherited from the definition of Cloud Computing, can be also understood in a layered service model, as illustrated in Fig. 2. Note that, in this layered model, there is no binding between two interfacing layers; while in the Internet layered model, service binding is enforced between interfacing layers. Specifically, media services in the PaaS layer can run over either on cloud infrastructure, or on raw ICT infrastructure, or on a hybrid of both resources. Similarly, media applications in the SaaS layers can leverage either media service provided by the PaaS layer, or other media services not exposed via cloud API, or a hybrid of both categories, or even run over traditional server architecture. The layered view only provides a conceptual hierarchy to under the complexity of the cloud mobile media architecture.

1) Infrastructure-as-a-Service: In the *IaaS* layer, ICT resources are pooled together from a hybrid cloud infrastructure (e.g., public cloud, private cloud, and community cloud). Enabled by virtualization technology [29], cloud service provider can allocate these resources in a fine-granular manner, to meet specific user demand. These resources can be exposed to media applications and/or media services in its raw format, with specific service level agreement (SLA). Within this model, a key component is a distributed resource management protocol that oversees all the available resources in the cloud infrastructure. Intelligent algorithms should be developed in research to address the list of technical challenges in cloud resource management, in connection with an enhanced mobile media experience. This will be the main focus of Section III.

2) Platform-as-a-Service: In the *PaaS* layer, various media services are encapsulated into a layer of middleware, running over raw ICT resources or cloud ICT resources. In the context of cloud mobile media, we classify potential media-specific services into the following four categories:

- *Media Representation*: this service refers to encoding, decoding, and transcoding media contents in various container formats.
- *Media Distribution*: this service deals with acquiring media contents into the cloud, moving them across different processing units and streaming them to mobile devices. Examples of media distribution services include content caching, content pre-fetching, content storage, content routing, and etc.
- *Media Adaption*: this service refers to algorithms and mechanisms that modify the original media contents in different domains, for example, contents (tagging, annotation), formats, rates, etc. Other typical services in this category include media mashup, media rendering, etc.
- *Media Analytics*: this service refers to algorithms and mechanism that are derived from media contents and in data analytics nature. The most popular service is the content search service and its associated data-mining algorithms. Other data-mining algorithms, for example, content recommendation, user-behavior analysis, location-based service, are also included in this category.

Other non-media specific services can be encapsulated in this layer. Essential services for any cloud mobile media application include digital right management (DRM), Authentication/Authorization/Accounting (AAA) service, social group management, etc. An in-depth survey of existing cloud-based media services will be covered in Section IV.

These media services are encapsulated into a set of media service APIs, which are in turn exposed to media applications. Application developers can leverage this programming environment to develop and to orchestrate their innovative applications. As such, it serves the purpose of platform-as-a-service. Moreover, we suggest that a media service orchestration module be built in the PaaS layer, providing mechanisms for integrating different media services, including services provided by both an internal media system and external systems, into an integrated media application for both content providers and content consumers. For example, a social TV application [30] would draw a lot of individual media services together, including a buddy service from a social networking provider and a media streaming service.

3) Software-as-a-Service: In the *SaaS* layer, mobile media contents and applications are consumed by viewers in their mobile devices. Typically, these applications consist of a light-weight client running on mobile devices, supported by media services running in the cloud. The design principle of cloud-based mobile media experience is to optimally leverage the strengths of both the mobile client and the cloud infrastructure/service, with an objective to provide the best possible user experience at the lowest possible cost. Cloud media applications are abundant on the market, enabled by the two leading cloud platforms (i.e., Google's Application Engine and Microsoft's Azure). We have built a multi-screen mobile social TV application (cf. Section VI), on top of a private cloud platform at NTU. More cloud media systems and applications will be analyzed in Section V, together with system challenges (e.g., privacy, security, and economics).

In this rest of this paper, we adopt this layered framework to survey existing research efforts in the aforementioned three layers. We characterize the technical challenges encountered by each layer and suggest potential solutions.

III. Resource Management and Control

As mobile devices get lighter and thinner, their computational and storage capabilities can hardly keep up with users' growing demand for a media-rich experience. Increasingly, mobile media services such as online gaming and video conferencing are hosted inside cloud computing centers due to the availability of abundant resources at relatively low cost. Such a paradigm shift imposes the intriguing problem of resource control and management. To meet the dynamic demands from media flows, novel solutions are needed to shift computational and storage loads from mobile devices to the cloud, to perform load balancing within a cloud, and to allocate resources across multiple clouds. This section reviews some of the ongoing work and existing solutions in this area.

A. Joint Mobile-Cloud Resource Management

One way to subsidize the limited computational resources of mobile devices is to offload computationally expensive modules to be executed inside the cloud. The offloading can take on various forms.

The most straightforward way is to execute mobile applications inside the cloud in a cloned virtual machine environment. CloneCloud [31], for instance, combines static analysis and dynamic profiling to partition applications automatically at a fine granularity. The system aims to optimize execution time and energy use for a target computation and communication environment. Similarly, ThinkAir [32] provides a framework for migrating smart phone applications to the cloud by means of virtualization and method-level computation offloading. In [33], a single application is partitioned into multiple components called "weblets". "Weblets" are either executed on the mobile device, or migrated to the cloud. The intelligent decision is based on the status of the device, including CPU load, memory, battery level, network connection quality, and user preferences. The work in [34] further studies the best strategy to choose between mobile execution and cloud execution within an energy-minimized framework.

In addition to offloading computational burdens, smart phone applications can resort to the cloud as a back-end for data storage. The data access patterns often depend on user locations. For example, a restaurant recommendation application is usually used to get information about nearby restaurants. This motivates the design of "WhereStore", a location-based data storage system [35]. It uses filtered replication along with the location history of each device to distribute items between smart phones and the cloud. Alternatively, the work in [36] creates a storage cloud using edge devices, based on Peer-to-Peer resource provisioning. This approach combines all end-user edge devices—mobile phones, cable modems, and set-top-boxes (STBs)—into one flexible and highly scalable storage system. By keeping the data close to the users, it improves data availability while reducing data retrieval latency.

Finally, contemporary applications such as free viewpoint video may utilize the cloud as a back-end for both computation and data storage. The application needs to access large amounts of multi-view video data in order to render a novel views on the mobile device. When the cloud is used for storage only and the rendering of a free viewpoint video is executed on a mobile device, the user experience may become unacceptable due to high latency over the wide area network. In [37], Miao *et al.* have presented a resource allocation scheme that jointly considers rendering in the cloud and rendering by mobile clients. The scheme strikes a balance between quality-optimal cloud rendering and delay-optimal client rendering.

B. Intra- and Inter-Cloud Resource Scheduling

When a single cloud is responsible for multimedia mobile applications, the resource scheduling problem often revolves around cost minimization while guaranteeing some level of quality of service. In [38], the authors formulate intra-cloud resource scheduling problem based on a queuing model of the system. For both single-class and multi-class service scenarios, the proposed scheme aims at minimizing response time and resource cost. Alternatively, in [39], the optimization framework jointly considers multiple distribution paths and placement of web server replicas in a cloud-based CDN.

For resource sharing across multiple clouds, research efforts often focus on dynamically balancing between loads and the demands of geographically diverse users. Analysis of a large collection of YouTube video request traces reveals that partitioning the network solely based on social relationships leads to unbalanced access [40]. Instead, a novel social graph partitioning algorithm is proposed to preserve the social relationships with more balanced network partitions. An optimization framework is presented in [41] for live streaming applications. The proposed scheme adaptively leases and adjusts cloud server resources with fine granularity—on an hourly basis—so to accommodate temporal and spacial dynamics of global demands. Geographically diverse cloud-computing platforms can also benefit video-on-demand (VoD) applications. The scheme in [42] optimally decides video replication and user request dispatching in a hybrid cloud of on-premise servers and geo-distributed cloud data centers. In [43], the resources of multiple clouds are pooled together in an intelligent manner to ensure the bandwidth guarantee for individual subscribers to Netflix-like services.

C. A Unified Optimization Framework

On-going works in resource control and management for mobile cloud computing often share a similar mathematical formulation, as a constrained optimization problem. In this subsection we propose to abstract various formulations into a unified optimization framework for resource management and control in mobile cloud media. Specifically, its objective is to minimize the total cost of ownership for a cloud media network including both upfront cost (i.e., CAPEX) and re-occurring cost (i.e., OPEX). In addition, the optimization is subject to two categories of constraints including:

- *Capacity constraints*: the resource usage (e.g., computing, bandwidth, and storage) should be within the maximum capacity that the cloud infrastructure can provide. Mathematically, two alternative formulations can be used: (i) bounded capacity constraint, in which the resource usage is strictly less than a predetermined threshold, and (ii) penalized capacity constraint, in which the resource usage can be larger than a predetermined threshold, with a penalty function associated with the over-provisioned capacity.
- *QoS/QoE constraints*: the system performance, in the format of QoS or QoE, has to be guaranteed to provide decent user experience. A chosen QoS/QoE metric (e.g., delay, response time, mean opinion score, etc.) is normally a function of system status, resource allocation, and etc.

Finally, the formulation should take randomness from different sources, for example, user mobility, wireless channel variations, etc., into consideration.

Mathematically, we can express the optimization framework as

$$\begin{aligned} \min_{\vec{s}\in\mathbf{S}} \quad & C_{tot} = \mathbb{E}_{\vec{\mathrm{R}}}\{C_{CAPEX} + C_{OPEX}\} \\ \text{s.t.} \quad & f_i(\vec{s}) \leq F_i, i = 1, 2, \cdots, m, \\ & g_j(\vec{s}, \vec{\lambda}, \vec{\mathrm{R}}) \geq G_j, j = 1, 2, \cdots, n, \end{aligned}$$

where $\vec{s}$ is the resource-allocation vector, $\mathbf{S}$ is the space of resource-allocation vector, $f_i(\cdot)$ refers to counting function for resource of type i and F_i denotes the maximum capacity for resource of type i, $g_j(\cdot)$ refers to the measurement of QoS/QoE of type j and G_j denotes the required QoS/QoE metrics, $\vec{\lambda}$ indicates the user request parameters, $\vec{\mathrm{R}}$ refers to the set of random parameters in the system.

Standard approaches (e.g., linear programming, non-linear programming, dynamic programming, and geometric programming) or emerging techniques (e.g., stochastic optimization, and robust optimization) can be sought to solve this optimization problem. Interested readers might refer to classical textbooks [44], [45], [46] for more in-depth discussion.

IV. Cloud-based Media Platform Services

In this section, we investigate the list of cloud-based media platform services, which are encapsulated in the service layer and offered via application programming interfaces (APIs) for application development. The set of media platform services is diverse in nature, embracing growing needs from applications. For example, in [8], the authors introduced the cloud-

TABLE I
CLOUD MEDIA SERVICES

Service		Dimension	Approach and Reference
Representation	Transcoding	Format	Hadoop-based approach [50], [51], [52] SHARC for 3D content [53]
	Encoding/Decoding	Format	MapReduced-based approach [54], [55] Rate control with SVC [56] Cloud-codebook encoding [57]
Distribution	Content Acquisition	Existence	acquisition from mobile devices [58], [59], [60]
	Content Delivery	Existence	Cloud CDN solutions [61], [62], [18], [63], [64] social media delivery [65], [66] edge content delivery [67], [68], [69]
	Media Streaming	Existence	proxy-based streaming [70] DASH-based streaming [27], [28]
Adaptation	Media Metadata	Knowledge	semantic adaption [71]
	Media Mashup	Knowledge	Popcorn Maker [72]
	Media Rendering	Knowledge	proxy-based 3D rendering [73]
Analytics	Content Analysis	Knowledge	MapReduce-based solution [74] RanKloud [75]
	Content Recommendation	Knowledge	context-information survey [76] context-aware algorithm [77]
	Media Retrieval	Knowledge	Content-based image retrieval as a service [78]

aware multimedia, which provides multimedia services, such as, storage and sharing, authoring and mash-up, adaptation and delivery, rendering and retrieval. Moreover, they include not only media-specific services but also non-media-related services that are crucial to support salient media applications. The diverse nature of media platform services demands a systematic framework, currently absent in the research literature to categorize them and provide guidelines for further research and development. In this paper, we first introduce a systematic framework to categorize media manipulation in three axis. Following that, we propose a system framework, as illustrated in Fig. 2, to categorize media services in the context of (mobile) cloud media network. The framework classifies media-related platform services into four categories, each of which will be explained in Section IV-B–V-E.

A. Systematic Framework for Media Manipulation

This subsection introduces a systematic framework to understand different dimensions in which digital media can be manipulated. Specifically, as illustrated in Fig. 3, media content can be modified in three orthogonal dimensions, namely:

- *Existence Dimension*: in the existence dimension, media is labelled with its temporal and spatial properties. Specifically, it specifies where and when a piece of media exists.
- *Format Dimension*: in the format dimension, media is represented by a sequence of bits that are encoded with a chosen codec (e.g., MPEG-4, scalable video coding (SVC), H.264, etc.) at a specific bit rate. Other possible format metrics (e.g., frame rate, etc.) can also be introduced to characterize a media content.
- *Knowledge Dimension*: in the knowledge dimension, media is associated with its semantic meaning and possible contextual metadata (e.g., tagging, user record, etc.).

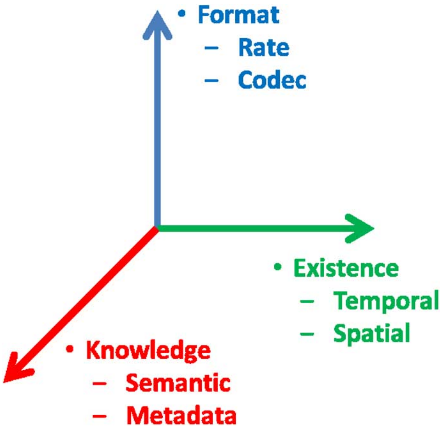


Fig. 3. A framework for manipulating mobile media in three dimensions: Existence, Format, and Knowledge.

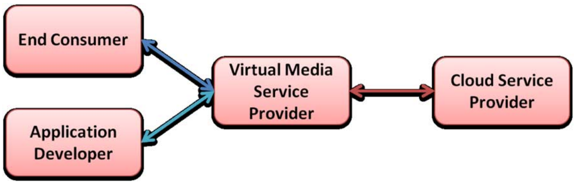


Fig. 4. A new business model enabled by integration between cloud computing and mobile media.

Media platform services, as defined in this paper, refer to mechanisms and algorithms to modify media in its three orthogonal dimensions. One media service could modify the media in one or more dimensions. Examples of media platform services include, but not limited to,

- Transcoding service [47] that converts the media form one codec to another one.
- Automatic media tagging service [48] that adds new metadata to the media.
- Adaptive media streaming service [49] that changes the playback rate and also transfers the content from the streaming engine to the mobile devices.

In the following subsections we elaborate on the four categories of media platform services as summarized in Table I.

B. Media Representation

The introduction of cloud computing into mobile media offers new opportunities to transform the existing media representation research. For example, in the context of mobile cloud media, computation-intensive media processing tasks (e.g., encoding, decoding, transcoding) can be offloaded from the resource-limited mobile devices to virtual machines in the cloud [34]. As a result, algorithms previously considered infeasible to mobile devices can now be made practical.

Cloud-assisted media representation has been an active area of research. Key research thrusts in this category include, but are not limited to,

- 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 [79]) 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). We believe that this area of research will accelerate over the next few years and these fundamental challenges will be the main focus for further research.

In this subsection, we briefly survey cloud-based platform services for media representation, specifically, transcoding, and encoding/decoding algorithms for media contents.

1) Transcoding: 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. For example, in [50], a Hadoop-based cloud is used in the transcoding of media contents, to supply the variety of existing media formats requested by end-users. Another example is CloudStream [51], a cloud-based video proxy that employs a multi-level transcoding parallelization framework with two mapping options (Hallsh-based mapping and Lateness-first mapping). The system employs a multifold objective to optimize transcoding speed and reduce transcoding jitters while preserving quality of the encoded video. This approach often focuses on the parallelization framework, while leaving the actual transcoding tasks to off-the-shelf transcoders (e.g., the *ffmpeg*[1] tool). For instance, in [52], transcoding was achieved through the use of Hadoop streaming jobs that utilize the ffmpeg tool. 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.

In addition to the aforementioned versatile approach, researchers have also developed cloud-based solutions to transcode specific media contents (e.g., 3D contents). In this case researchers often adopt a holistic (or horizontal) view of the problem, and provide an end-to-end solution. For example, in [53] SHARC was presented as a solution to enable scalable support of real-time 3D applications in a cloud computing environment. The solution uses a scalable pipelined processing infrastructure. It consists of three processing units including a virtualization server network for running 3D virtual applications, a graphics rendering network for processing graphics rendering workload with load balancing, and a media streaming network for transcoding rendered frames into H.264/MPEG-4 media streams and streaming the media streams to a cloud user. Moreover, a novel approach is to leverage a GPU array for media processing, compared to a CPU array as in the standard cloud architecture. Given the computation-intensive nature of video and image, GPU-based cloud solution would be a more cost-efficient solution. As a result, an interesting research subject is to investigate the effectiveness of a hybrid architecture including both CPU and GPU.

Another 3D video transcoding scheme has also been developed for heterogeneous mobile users with limited 2D display capabilities [80]. In this case the transcoding and adaptive 3D video transmission are necessary to extract/generate the required data content and represent it with appropriate formats and bit rates for the heterogeneous terminal devices. This scheme is able to obtain any desired view, either an encoded one or a virtual one, and compress it with more universal H.264/AVC for mobile terminals with 2D display only. The key idea is to appropriately utilize motion information contained in the bit stream to generate candidate motion information for efficient transcoding. Finally, this scheme was not only developed for mobile cloud media, but also can be easily adopted between cloud and mobile gateways.

Further research in both vertical and horizontal integration is highly demanded, with a dual objective to provide a real-time media transcoding service at low cost, while providing a decent QoS/QoE to the end users.

2) Encoding/Decoding: 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. For example, in [54] and [55], the authors proposed a Split&Merge architecture, generalizing the MapReduce paradigm that rationalizes the use of resources by exploring on-demand computing. In addition, the performance of cloud-assisted media encoding/decoding algorithms can be improved by incorporating media-specific considerations (e.g., rate control and SVC). An example of this approach is in [56], which proposed a video encoding infrastructure, coupled with a rate control scheme, to improve the coding efficiency on cloud environments. Experimental results demonstrated that the proposed video encoding infrastructure and rate control scheme gained better visual quality.

However, in cloud mobile media, encoding/decoding task offloading imposes additional challenges to the network connectivity between the mobile device and the cloud infra-

[1] FFmpeg is a complete, cross-platform solution to record, convert and stream audio and video. Online URL: http://www.ffmpeg.org/

structure. 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, as shown in [34].

Another potential idea in cloud-assisted media encoding is to leverage the huge number of videos and images as a random code book to encode the media contents on the mobile device. Authors from [57] have suggested such a solution. However, this solution is still in its early stage, due to the obvious limitation that the code book is not necessarily ubiquitously accessible and the decoding process would add extra traffic in retrieving codewords (i.e., reference images) from the cloud. Efficient algorithms for content lookup, routing, and retrieval are required to make this scheme practical.

C. Media Distribution

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. In this subsection, we first survey existing research in cloud-assisted media distribution and then suggest potential future research thrusts in this area.

1) Content Acquisition: In this step, media contents are acquired into the media cloud from their generating devices. Traditionally videos are often captured with cameras and then uploaded into servers from where they are acquired into the media cloud. Lately, with the technological advances in smartphones and wireless networks, user-generated videos can be captured by smartphones and then uploaded into the media cloud, preferably via wireless connectivity. However, direct acquisition from mobile devices imposes extra challenges, due to varying wireless channels and mobility. In [58], [59], [60], the authors proposed to upload media files from mobile devices via a collaborative wireless network (i.e., an ad-hoc mobile cloud). The multi-path opportunity in such a network offers tremendous potential to optimize the content uploading process. Various design choices can be adopted (e.g., inter-path packet coding, and packet allocation) to reduce the file uploading delay and the energy consumed by the mobile devices.

We notice that existing research in this category often focuses on the networking part of the problem. Little attention has been paid to jointly optimize media encoding and content acquisition in the cloud mobile media. We believe that the interplay between media encoding and content acquisition could result in better QoS and higher resource efficiency. For example, it is beneficial to distribute the media encoding task across both mobile device and the cloud in response to the wireless channel condition.

2) Media Delivery: Media delivery is predominately accomplished via a content distribution network (CDN), which is a large distributed system of servers deployed in multiple data centers, serving contents to end users with high availability and high performance (e.g., fast response, high throughput). Pioneered by Akamai, CDN lends itself an inherent demand for cloud-based solutions, overcoming the possible shortcomings of static resource provisioning. Existing research efforts in cloud-based CDN can be classified into two categories: (i) commercial solutions and (ii) academic research. Each is explained as follows.

In commercial solutions, CDN services [81] are being upgraded with cloud-based solutions. Specifically, commercial solutions can be classified into three categories [82], as follows:

- CDN Reseller: In this solution, 3rd party service providers first acquire a bulky CDN service from an original CDN provider (e.g., Akamai, Level-3 Communications), and resell the service in small pieces to clients with low demand or short duration. Examples in this categories include Distribution Cloud, VPS.net, Brightbox CDN, Rackspace, etc.
- Content Broker: In this solution, content providers upload their contents to some 3rd party websites (e.g., Youtube, Facebook), which serve as central exchange places for end users to consume these contents. These websites either have their own in-house CDN or outsource content delivery tasks through a long-term contract with CDN providers.
- Cloud CDN: In this solution, an emerging trend is for cloud service providers to sell CDN services. Examples in this category include Amazon CloudFront and CloudFlare.

The basic principle of these commercial services is to provide content distribution service in a "pay-per-use" model, offering the flexibility for the users to benefit from a fine-grained resource provisioning in a cloud environment. Due to their proprietary nature, we have not been able to obtain technical details of their solutions in public.

Compared to the business model in commercial deployment, academic research in cloud-based CDN focuses on improving performance, aiming to provide better QoE or QoS, while conserving the cost - including both CAPEX and OPEX. For example, in [61], the authors proposed and implemented the cloud downloading scheme, which caches the unpopular contents in the cloud, via the intra-cloud data transfer acceleration. The commercial system (named "VideoCloud") confirmed that this system achieves high-quality video content distribution by using cloud utilities to guarantee the data health and enhance the data transfer rate. In another case, authors in [62], [18] proposed and implemented a content-delivery-as-a-service (CoDaaS) scheme, which is built on a hybrid media cloud, and offers an elastic private virtual content delivery service with an agreed quality of service (QoS) to UGC providers. The scheme aims to distribute the emerging user-generated contents (UGCs), which are long-tail in nature. The preliminary results validate all the required features for UGC delivery and verify its comparative performance advantages. A similar example is the content distribution network cloud architecture (CDNCA) in [63]. The proposed solution is based not only on QoS criteria (e.g., round trip time, network hops, loss rate), but also on the quality of experience (QoE). Simulation results, based on OPNET, show that CDNCA yields significant improvement over traditional approaches. In another work [64] the authors presented experimental results of streaming distribution in a hybrid architecture consisting of mixed connections among P2P and cloud nodes that can inter-operate together. The QoS of a streaming service can be efficiently improved, by strategically placing certain distribution network nodes into the cloud provider's infrastructure, taking advantage of the reduced packet loss and low latency among its data centers.

In addition, the CDN service can also be tailored for specific media applications, for example, the emerging social media applications. In [65], the authors proposed efficient proactive algorithms for dynamic, optimal scaling of a social media application in a geo-distributed cloud. The key contribution is an online content migration and request distribution algorithm with a suit of salient features, including: i) future demand prediction, ii) one-shot optimal content migration and request distribution, and iii) look-ahead mechanism for optimization adjustment. Theoretical analysis verifies the effectiveness of the proposed algorithm. Similarly, in [66], the authors proposed a cloud-based social media service model that can store and manage massive social media contents. Using the universal plug and play (UPnP) technology, the approach arranges media services fit for users' tendency in the order of priority and provides the results to the users, with the objective to provide an efficient management for complex social media contents and media services appropriated for users in real time.

Finally, media cloud can be built over equipment located in network edges. For example, in [67], [68], [69], the authors proposed Media Cloud, a home gateway service for classifying, searching, and delivering media across the cloud that interpolates with UPnP. In this architecture media cloud services are located in a home gateway, which has access to the home network and the Internet. The home gateway can communicate with devices located in the home environment and also provide search services, content delivery and filtering to friends and family outside home domain.

3) 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. For example, in [70], the authors proposed a novel cloud-assisted architecture for supporting low-latency mobile media streaming applications such as online gaming and video conferences. A media proxy at the cloud is used to calculate the optimal media adaption decisions on behalf of the mobile sender, based on past observations of packet delivery delays of each stream. The proxy-based intelligent frame skipping problem is formulated within the Markov decision process (MDP) framework and is solved using the stochastic dynamic programming (SDP) approach. Simulation results indicate that the optimal policy consistently outperforms greedy heuristic schemes.

Another recently developed approach is to consider the special case of streaming, namely 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. Recently, novel research efforts that anticipate and investigate the key challenges to support HAS users requesting videos simultaneously from multiple video servers have been carried out [27]. One of the key issues is the design of data scheduling as video data from multiple servers may not arrive in proper sequential order at the receiver. A chunk of video from one server will become useless if its dependent previous chunk from another server cannot arrive in time. Another key issue is related to client rate adaptation because the optimal streaming rate depends on multiple cloud servers' bandwidth as well as mutual dependencies of the video chunks requested from them. The authors innovatively address these challenges with their development of "cloudDASH". This cloudDASH 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. Their implementation with PlanetLab networking research platform shows the effectiveness of the cloudDASH scheme.

Extension of DASH to the mobile terminal receivers faces a different set of challenges. For example, another recent effort to enable the DASH application for mobile users was also published in [28]. The solution to these challenges lies in the design of a novel proxy design for DASH service over cellular wireless networks aiming at significantly enhancing QoE in video streaming. This wireless proxy for DASH, named "WiDASH", can be located at the edge between cloud edge and wireless access networks. By performing a rate adaption algorithm at this novel proxy, a joint optimization of multiple concurrent DASH flows going through the base station can be achieved without sacrificing the scalability of the video streaming server. The WiDASH scheme exploits simultaneously the potential of split-TCP and parallel-TCP in which the original TCP connection from DASH server to a wireless user is split into one wired TCP connection and multiple wireless TCP connections. An adaptive control theory is then applied to the multiple-input multiple-output optimal rate controller resulting from the parallel architecture. The core innovation of the WiDASH is this rate controller based on the multidimensional adaptive control theory. The WiDASH scheme is able to minimize a convex cost function of QoE defined by weighted sum of several relevant video quality metrics in distortion, quality variation, and playback jitter. Unlike existing schemes of adopting DASH for wireless users that experience great fluctuation in video quality, this WiDASH scheme achieves much enhanced performance with smoother video flows and noticeably improved average visual quality.

In observance of these existing research efforts, we believe that future research in CDN service for mobile media should focus on cost-optimized mobile media distribution via cloud. This thrust could embrace various tasks, including but not limited to, distribution tree design, distributed storage and content caching, distributed content routing, etc. Sample research topics range from distributed tree algorithm, with respect to various cloud pricing models, erasure-based cloud storage mechanisms for media, distributed content routing and discovery algorithm, and core-to-edge distribution.

D. Media Adaptation

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: in this service, media contents are supplemented with tagging, annotation, and other metadata (e.g., content overlay, etc.).
- Media Mashup: it combines multiple media contents into a new content, usually serving a specific purpose. For example, the open-source Popcorn Maker [72] allows the user to integrate contents from multiple sources into streaming videos.
- Media Rendering: rendering is the process of generating an image from a model (or models in what collectively could be called a scene file), by means of computer programs. In mobile cloud media, rendering is normally conducted at the client sides (e.g., geometry modeling, texture mapping, and so on). However, as the media industry evolves, the clients do not have sufficient resources for complicated rendering tasks. For example, 3D rendering is resource-hungry, exceeding the capability of the latest mobile phones, especially in real-time applications. In this case, one could leverage the cloud resource for video rendering in mobile experience.

However, we have noticed that research on cloud-assisted media adaption is rare with few examples. One remotely related example is from [73], wherein the authors proposed a rendering proxy to perform 3D video rendering for the mobile phone. The rendering proxy could be substantiated with a virtual machine in the cloud.

In this area, one key research topic is context-aware media rendering of 2D/3D contents, graphical contents and immersive contents. The context could be diverse, such as networking conditions, outlet capability, user preference, device capability, and environment context. Moreover, the adaption is often constrained by on-board resources availability in mobile devices. Therefore, previous research has often taken a constrained optimization approach. Various themes have been pervasive in this area of research, for example, distributed rendering design [8], trade-off between energy and experience in mobile device, and energy-efficient graphic rendering on mobile devices, etc. With introduction of cloud computing, this approach can be extended with an elastic computing model in which computing resources can be dynamically reconfigured at a cost. New technical solutions and business models would result from this elastic resource provisioning mechanism.

A different avenue of research in media adaptation has focused on maintaining semantics of the media while converting the original media content for display or rendering at the mobile devices. This is particularly suitable for mobile devices to access small scale media datasets such as private cloud designed for personal services. In general, the total number of semantic concepts will be numerous in the real world and in the general media database. When the application is limited to consumer photos served in a personal cloud, the concepts appearing in consumer domain encompass only a small fraction of the general concepts. It has been shown that most consumer photos are relevant to one of the 12 events as defined in [83]. One successful semantics-based media adaptation scheme that targets on consumer photos can be easily extended to personal cloud-based media adaptation [71]. With limited number of event definitions in the proposed media adaptation system, the semantic analysis can be designed to effectively utilize the user provided semantic keywords to extract the semantically important objects from a given photo. This scheme provides the mobile users with the most desired image content, which integrates the content semantic importance with user preferences and provides perceptually optimized display on mobile devices under limited mobile display constraints. One particular feature of this scheme that is shown to bridge the semantic gap for adaptation is the Bayesian fusion approach to properly integrate low level features with high level semantics. It is straightforward to extent this adaptation scheme developed for consumer photo database to the broader application of cloud media-based adaptation for mobile terminals. In this extension, one needs to design cloud-edge server interface that replaces the personal consumer photo database in the original scheme. One possible solution is to design such server as cloudlet that is placed near the consumers when such request is made via mobile devices [24].

E. Media Analytics

Recent years have witnessed an explosive growth of multimedia data and metadata, due to higher processor speeds, faster networks, wider availability of mass-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.

1) Content Analysis and Metadata Mining: Cloud computing, owing to its elastic resource provisioning and distributed computing paradigm, renders itself a natural solution to large-scale content analysis and metadata mining applications. For example, the MapReduce framework can be readily applied for the multi-dimensional data analysis problem, ubiquitous in multimedia applications. Existing works have been observed in this area to provide cloud-based content analysis and mining services in a Platform-as-a-Service (PaaS) environment. For example, in [74] researchers described media cloud services including automatic tagging, in which feature information is extracted by image or video analysis technologies (such as color recognition, shape recognition, scene analysis, character recognition, and speech recognition). In [75] the authors presented an overview of recent developments in the area of scalable multimedia and social media retrieval and analysis in the cloud to build a scalable data processing middleware, called "RanKloud", specifically sensitive to the needs and requirements of multimedia and social media analysis applications. RanKloud includes a tensor-based relational data model to support the complete lifecycle (from collection to analysis) of the data, involving various integration and other manipulation steps. With the growing adoption of cloud computing in mainstream media analytics research, we believe that more and more work will emerge in this area. Interested readers might refer to [84] for a list of latest works in this area.

2) Content Recommendation: The vast volume of media data and metadata provides a golden opportunity to develop more accurate content recommendation in mobile media experience. The opportunity is further made practical with the advanced computing capability in the cloud environment. As a result of these two drivers, existing work in cloud-based content recommendation is prolific. For example, in [76], the authors investigated several relevant recent developments in ubiquitous media services, especially in the area of content recommendation. It is desired for the service providers to offer highly relevant recommendations of media content to the users by exploring user-user, user-media, and user-context relationships. Another example is a context-aware recommendation service based on the cloud model [77], in which a cloud-computing paradigm was proposed as the next generation infrastructure to support a highly scalable service oriented architecture, with an objective to provide personalized content recommendation services that recommend content relevant to the user irrespective of his location or access devices.

3) Media Retrieval: Multimedia retrieval—for example, content-based image retrieval (CBIR)—can leverage the capability of cloud computing to enhance the performance of various building blocks, such as, feature extraction, similarity measurement, and relevance feedback. In [78], the authors presented the system architecture of a Content-Based Image Retrieval system as a web service. The proposed solution is composed of two parts, a client running a graphical user interface for query formulation and a server where the search engine explores an image repository. The separation of the user interface and the search engine follows a Software-as-a-Service (SaaS) model, a type of cloud computing design where a single core system is online and available to authorized clients.

However, we noticed that existing research in media analytics mostly focus on leveraging cloud infrastructure to conduct traditional media analysis tasks. In fact, the interplay between the cloud infrastructure and media analytic algorithms could be exploited further, potentially resulting in new research topics. For example, new algorithms could be developed to adapt the media analytics in the media platform service layer, by taking considerations of different power consumption dynamics in mobile devices and cloud services, and various resource pricing models offered by the cloud service provider. We expect that such a new paradigm could reveal new insights in system design and service deployment.

V. Novel Cloud Systems and Applications

The advent of cloud-based mobile media not only brings about individual technical challenges discussed in the previous few sections, but also gives rise to the need of novel system architectures and opens up opportunities for novel applications. This section will sample a few such instances, followed by discussions on some overarching issues in enabling mobile applications on the cloud.

A. Platforms, Frameworks, and Architecture

There already exist several popular cloud-based parallel programming frameworks, such as Google's MapReduce [79] and Microsoft's Dryad [85]. However, they typically lack the support for arbitrarily complex workloads and iterative operations, hence can be problematic for implementing media encoding and processing functions. A new framework is presented in [86], called P2G. It is designed specifically for developing and processing distributed real-time multimedia data. P2G supports arbitrarily complex dependency graphs with cycles, branches, and deadlines. It provides both data- and task-parallelism. The framework is implemented to scale transparently with available (heterogeneous) resources, therefore naturally fits into the cloud computing paradigm.

A cross-layer architecture is described in [25]. It allows wireless devices to execute multimedia applications with seamless mobility. A cloud computing environment is used as the infrastructure to dynamically set up (and release) the proxies on the server side, in accordance with the pay-as-you-go principle of cloud based services. For high-quality multimedia applications, the authors in [87] proposed a novel IP multimedia subsystem (IMS) framework. The proposed scheme supports heterogeneous networking with quality-of-service policies. It also builds upon MapReduce to enhance computing capability of the system.

B. Novel Applications

The convergence of technologies in cloud computing, media processing, and mobile communications gives rise to many novel applications. In [88], for instance, the authors describe a cloud-based mobile location search service. Users can find out where they are in a visually intricate environment (e.g., on the busy streets of Hong Kong) by capturing a short video clip using their mobile devices. The system then matches the scale-invariant feature transform (SIFT) points extracted from the clip to those from a repository already tagged and stored in the cloud. The proposed scheme applies spatio-temporal pruning of SIFT points in the video repository, as well as principal component analysis (PCA) projecting and indexing of SIFT points in the cloud. Consequently, it can achieve robust SIFT feature matching in video sequences; its computational complexity scales with a large video repository.

The abundant resources of the cloud can also be leveraged to improve the efficacy of conventionally labor-intensive tasks. A framework is presented in [89] for performing video quality evaluations via crowd-sourcing. By hosting video servers inside the cloud and re-directing subjective viewing tasks to mechanical turks in an automated fashion, such an approach can significantly expand the pool of subjective viewing test participants in a cost-effective manner. At the same time, it achieves comparable inter-lab correlations with regard to conventional tests in a standardized environment.

Other typical examples of cloud mobile media applications range from cloud-based multimedia conferencing [90] to social media sharing [91], [92] and cloud interactive learning [93].

C. PST: Privacy, Security and Trust

Mobile cloud media, inherited from the wide perception about cloud computing, is also suffering from concerns on privacy, security and trust management. Fortunately, these concerns are not unique to the mobile cloud media system. As such, a lot of existing solutions can be readily adopted into the mobile cloud system (e.g., fine-grained access control [94], secure service admission [95]). At the same time, combining multimedia with cloud renders some new perspectives in security. In [96], the authors conduct an in-depth survey on recent multimedia storage security research activities in the new cloud computing paradigm. Four topics, namely data integrity, data confidentiality, access control, and data manipulation in the encrypted domain, were investigated in this survey.

Another topic of relevance is the emerging named data network (NDN) [97] for content-centric network, in which content security is a built-in nature of any data in the system. For example, in [98], the authors proposed a name-based trust and security protection mechanism for content-centric network. The scheme is built with identity-based cryptography (IBC), where the identity of a user or a device can act as a public key string. The trust of a content is seamlessly integrated with the verification of the content's integrity and authenticity with its name or prefix. For scalable deployment, the authors further propose to use a hybrid scheme that combines conventional public-key infrastructure (PKI) with IBC.

D. Economics: Cost Control and Revenue Maximization

The integration of cloud computing and media service enables new business models by transforming the mobile media value chain. Specifically, as illustrated in Fig. 4, media service providers can rent infrastructure resources from cloud service providers (e.g., Amazon, Google, and Microsoft) and, at the same time, provide metered media services to end users and/or application developers. With both east and west interfaces, the virtual media service provider normally operates with a dual objective: i) to control the cost and ii) to maximize the revenue.

The process of controlling the cost can be divided into two fundamental steps, including increasing resource utilization and reducing the price paid to the cloud service providers. These two steps can be individually optimized or jointly optimized, to reduce the operating cost for the virtual media service providers. To increase the resource utilization, one can design smart resource allocation algorithms to avoid potential resource wastage by renting cloud resources in response to applications' demands. At the same time, the virtual service provider can leverage the alternative price models offered by the cloud service providers, to reduce its price paid. For example, AWS offers three alternative price models including on-demand, reserved, and spot instances, each of which exhibits different price elasticity. Therefore, it is feasible to optimize the resource acquisition process to reduce the total cost.

The process of maximizing the revenue can be accomplished by intelligently pricing the media service provided to the end users and the application developers. The details of cloud service pricing strategies are beyond the scope of this paper.

Finally, it is appealing to jointly optimize cost reduction and revenue maximization. One example in [99] considers a new type of service where VoD providers make reservations for bandwidth guarantee from the cloud at negotiable prices to support continuous media streaming. It has been proved that the market has a unique Nash equilibrium where the bandwidth reservation price for a VoD provider critically depends on its demand correlation to the market.

VI. Cloud-Centric Media Platform: From Theory to Practice

At NTU, we have pioneered in putting the aforementioned system architecture into practice and completed a prototype system of a cloud-centric media platform that supports a multi-screen mobile social TV application. In this section, we briefly describe our system architecture and functional design, with verified performance in response time and bandwidth saving. The platform is tailored toward interactive TV experience.

A. System Architecture

The design of the cloud-centric media platform for a multi-screen mobile social TV application is guided by one fundamental principle, i.e., reducing the capability requirement for content sources and media outlets, by leveraging the emerging cloud computing paradigm to offload computing-intensive tasks into the cloud infrastructure. The specific system architecture, as illustrated in Fig. 5, is adapted from the generic cloud media service platform in Section II, with a holistic extension. The whole system consists of three participatory stakeholders, including:

- *Content Sources*: it integrates all possible content sources in a typical TV experience, including: (i) Live TV Video Streams: it contains not only the most traditional live video streams from TV channels, but also the streams from any third-party live content brokers such as Hulu, (ii) On-Demand TV Video Streams: it could be either over-the-top (OTT) contents from VoD distributors such as Netflix, managed contents from service providers, or the user generated contents (UGCs) from digital video recorder (DVR) and personal video recorder (PVR), and (iii) any other 3rd party contents (e.g., contents stored at personal cloud).
- *Media Cloud*: a layered service model is introduced in the media cloud, including: (i) Infrastructure-as-a-Service: In this service model, there is a resource pool powered by virtualization technology in a set of data centers. Those resources can be utilized on demand to provide elastic computation and storage capability to the upper layers, (ii) Platform-as-a-Service: The media service platform offers a solution stack as a service, through a proprietary inter-cloud messaging bus protocol. Specifically, a wide variety of tools, libraries, and interfaces are provided in this model. The offerings facilitate the deployment of the media

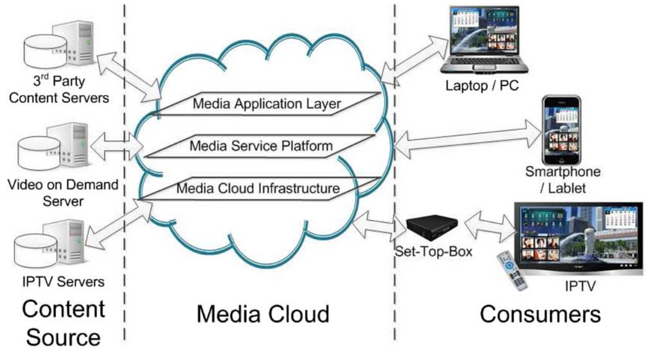


Fig. 5. A system architecture for multi-screen mobile social TV over the cloud-centric media platform.

cloud on resource allocation, cloud clone management, media service discovery, etc., and (iii) Software-as-a-Service: The media application layer follows a software delivery model. It instantiates the multi-screen cloud social TV system as a service, where the software and the major associated data are hosted in the cloud, and only a thin client is required on the client side.

- *Video Outlets*: video outlet refers to the end device that users are using for video consumption.

B. Cloud Media Platform Services

The multi-screen cloud social TV application is built on the cloud-centric media technologies. The key idea of the cloud-centric media platform is to leverage the emerging cloud-computing technologies to transform the media value chain, including content acquisition, content distribution, media adaptation, and media analytics. All these media services are encapsulated in a layer of middleware and exposed through a set of well-designed application programming interfaces (APIs) for application development. The cloud social TV platform is an innovative application targeted to integrate the traditional “laid-back” video watching behavior with the emerging “lean-forward” social network experience, thus enabling more value-added content services through an immersive TV experience.

Specifically, the multi-screen mobile social TV application, developed in our lab at NTU, pioneers in providing shared video viewing experience, with asynchronous and synchronous interactive features, at the same time rendering an immersive video viewing experience over multiple media outlets (e.g., TV, PC, smartphone, and tablet). The application leverages seven classes of media platform services, including:

- *Content Management*: this service allows the users to acquire and manage video contents from multiple sources, including local DVR, OTT, live streams, etc. It is also possible for a viewer to customize his/her own channels by scheduling videos from different channels.
- *Video Streaming*: this service allows the viewer to watch the same video in different outlets, by adopting a cloud-based video transcoding service to conduct real-time format translation. Moreover, the playback rate can be adapted dynamically in response to the network condition, user preference, etc.
- *Social Networking Exchange*: this service leverages an XMPP messaging bus for the user to integrate major social networking services (e.g., Facebook, Google+) and create his/her private social networking.
- *Multi-Screen Orchestration*: the social TV platform embraces all possible media outlets (e.g., TV, PC, smartphone and tablet) into a seamless media experience, by using our proprietary session orchestration technology.
- *Metadata Publish and Video Overlay*: the asynchronous communication feature empowers an interactive video annotation application for the viewer to tag the content and share these illustrative metadata with other viewers.
- *Communication Modalities*: the synchronous communication feature provides the viewer with a capability to interact with other geo-remote viewers on live, via four communication modalities (i.e., text, voice, video, and avatar). One distinguished feature of our social TV is the avatar with emoticons, which provides an emulated social interaction in a 3D virtual environment.
- *Data Analytics*: in this subsystem, all the metadata and transaction logs are stored in a distributed database and various data-mining algorithms can be developed to extract insights about user behavior pattern and other possible value-added data services.

These features are built upon different media platform services in the cloud-centric media platform. As mentioned previously, these services are designed to be scalable and user-friendly. In the next subsection, we explain how the multi-screen session orchestration service is designed, followed by its scalable performance compared to other alternatives.

C. Interactive Media Applications

We demonstrate two interactive media applications of our multi-screen social TV system, including a virtual living-room TV application and a video teleportation application as an enhanced multi-screen experience, based on real scenarios.

Virtual Living-Room TV: on a weekend, Peter is enjoying his TV time via multi-screen cloud social TV. Among various content resources (e.g., video on local disk, video in the cloud, OTT videos, and live streaming), Peter chooses a video to watch. In the meantime, he finds his best friend Cathy is online from his friend list in Facebook and Google+. Peter says “hi” to Cathy via text chat box, and invites her to view the same video from the same point with the same playback rate. Via this social TV application, they can discuss the content they are both watching via video chat from different locations. They can even collaboratively edit the original content by inserting text, pictures, and audio, and generate new contents.

Video Teleportation: Peter suddenly realizes that he needs to attend a family reunion tonight, and he has to go to the supermarket at once. However, he does not want to give up the chance to enjoy a video and have a conversation with his best friend. The cloud social TV can help him to teleport this experience from TV screen to his mobile device, as follows. Peter first uses his smartphone to scan the TV screen, to obtain authentication for his smartphone (cf. Fig. 6(a)). Once his phone is authenticated, a synchronized image will be displayed

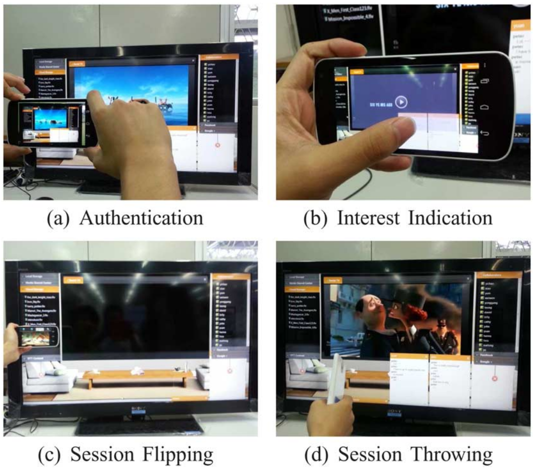


Fig. 6. Workflow for "video teleportation", an intuitive multi-screen orchestration protocol.

on his phone screen. Peter then indicates his interest in transferring the sessions on video chat and content viewing with him (cf. Fig. 6(b)). Finally, he simply flips his phone away from the TV to migrate the chosen sessions onto his smartphone without any interruption (cf. Fig. 6(c)). Later, when Peter returns home, he simply "throws" his smartphone towards the TV, and the sessions would immediately resume on the TV screen (cf. Fig. 6(d)).

VII. Outlook: Mobile Media Meets Cloud Computing

Trends toward mobile media are demanding new paradigms to transform viewer's experience. Looking into the future, videos are being consumed across all the devices (TV, laptop, and smartphone) at a record pace [100]. In addition, the second-screen consumption is rising in which smartphones and tablets are being used to complement the video consumption on the main screen. This paper can be used to support both arguments here. In both cases, smartphones have emerged as the *de facto* choice for videos and their related services (e.g., instant messages, contextual information, etc.). As a result, the tussle between the growing demand for mobile media and the inherent limitations in existing media network will become more prominent, dictating new approaches to resolve it.

We believe that cloud computing has emerged as a natural solution to transform mobile media network into a new paradigm of cloud mobile media network. Our survey on existing research efforts on this emerging area of research has suggested that it is still in its infancy stage. As such, we believe that more research efforts from our community should be dedicated to this important research subject. We believe that the following venues of research should be pursued.

First, the paradigm of *cloud mobile media* would enable service providers and network operators to offer media services to ever increasing mobile users with much improved efficiency. This objective can be manifested from different benefits, some of which are explained as follows,

- *Improved Performance.* Leveraging omnipresent clouds could improve the performance for mobile media network significantly. The MapReduce framework makes it possible to conduct real-time transcoding operations to adapt to the channel quality of wireless Internet and the different screen sizes of smartphones. As a result, it is high time to revisit all the media services (or algorithms) in this new context to optimize their performance.
- *Lower Cost.* Virtualized computing in cloud paradigm makes it possible to dynamically acquire resources to meet application demand. It also decouples the resource management from the resource acquisition, adding one more degree of freedom to optimize. One example is to leverage the alternative resource price models to reduce the operating cost of cloud media services.
- *Better QoS/QoE.* The mobile device can leverage the seemingly infinite cloud resources to extend its limited onboard resource. It is possible to judiciously allocate tasks of different natures into alternative resources, providing a better QoS/QoE. For example, rendering can be split between mobile devices and cloud proxies to balance the visual effect and the delay.
- *Human-Centric Social Media.* In the mobile domain, human-centric social media technologies are crucial, in which human factors should be taken into architectural and operational decisions. For example, with the growing number of Tweets about video, it would be extremely beneficial to mine these social media data to develop better cloud media services. To follow this trend, we expect research community to look into the combination of big data and multimedia networks.

These benefits of cloud mobile media can be further cast into research problems, similar to the existing research efforts surveyed in this paper.

Second, new mobile media applications and cloud computing platforms are driving each other for further innovations from both domains. More and more consumers are adopting mobile devices as one of their primary media experience platforms, expecting new classes of cloud-enabled mobile media applications. These growing demands in turn require new and more powerful cloud computing platform and infrastructure capabilities to support. The advancements in cloud computing will revert to trigger new media applications adopted into mobile context. We foresee that the interplay between these two forces will be the driving force of innovation in this emerging research area.

Finally, the cloud mobile media paradigm will also benefit from other emerging trends in ICT. Examples of related emerging technologies in ICT include, but are not limited to, software-defined networking (SDN) [101], named-data networking (NDN) [102], big data analytics [103]. For example, with the adoption of SDN, configuration and management of CDN services will become extremely convenient, further improving QoS and reducing cost; and cloud mobile media can benefit from processing the ever-growing metadata and transaction logs to provide operational guidelines.

Acknowledgment

The authors would like to thank the anonymous reviewers for their insightful comments to improve the quality of this paper.

REFERENCES

[1] "Global mobile data traffic forecast update, 2012-2017," Cisco Visual Networking Index, White Paper, Feb. 2013 [Online]. Available: http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.pdf

[2] M. Satyanarayanan, P. Bahl, R. Caceres, and N. Davies, "The case for VM-based cloudlets in mobile computing," *IEEE Pervasive Comput.*, vol. 8, no. 4, pp. 14–23, Oct. 2009.

[3] P. Wendell and M. J. Freedman, "Going viral: Flash crowds in an open CDN," in *Proc. 2011 ACM SIGCOMM Conf. Internet Measurement Conf. (IMC '11)*, New York, NY, USA, 2011, pp. 549–558.

[4] D. Clark, B. Lehr, S. Bauer, P. Faratin, R. Sami, and J. Wroclawski, "The growth of internet overlay networks: Implications for architecture, industry structure and policy," in *Proc. Telecommunications Policy Research Conf.*, 2005.

[5] P. Mell and T. Grance, in *The NIST Definition of Cloud Computing*, Sep. 2011.

[6] X.-S. Hua, G. Hua, and C. W. Chen, "ACM workshop on mobile cloud media computing," in *Proc. ACM MCMC'10*, Firenze, Italy, Oct. 2010.

[7] M. Tan and X. Su, "Media cloud: When media revolution meets rise of cloud computing," in *Proc. 6th IEEE Int. Symp. Service Oriented System Engineering (SOSE 2011)*, 2011.

[8] W. Zhu, C. Luo, J. Wang, and S. Li, "Multimedia cloud computing," *IEEE Signal Process. Mag.*, vol. 28, no. 3, pp. 59–69, May 2011.

[9] S. Dey, "Cloud mobile media: Opportunities, challenges, and directions," in *Proc. Int. Conf. Computing, Networking and Communications (ICNC), 2012*, Feb. 30-2, 2012, 2005, pp. 929–933.

[10] U. Hoelzle and L. A. Barroso, *The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines*, 1st ed. San Rafael, CA, USA: Morgan and Claypool, 2009.

[11] M. Zhou, R. Zhang, W. Xie, W. Qian, and A. Zhou, "Security and privacy in cloud computing: A survey," in *Proc. 6th Int. Conf. Semantics Knowledge and Grid (SKG), 2010*, 2010, pp. 105–112.

[12] Z. Wang, "Security and privacy issues within the cloud computing," in *Proc. Int. Conf. Computational and Information Sciences (ICCIS), 2011*, 2011, pp. 175–178.

[13] Y. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Cross-VM side channels and their use to extract private keys," in *Proc. 2012 ACM Conf. Computer and Communications Security (CCS '12)*, New York, NY, USA, 2012, pp. 305–316.

[14] B. Rimal, E. Choi, and I. Lumb, "A taxonomy and survey of cloud computing systems," in *Proc. 5th IEEE Int. Joint Conf. INC, IMS and IDC, 2009 (NCM'09)*, 2009, pp. 44–51.

[15] A. Beloglazov, R. Buyya, Y. Lee, and A. Zomaya *et al.*, "A taxonomy and survey of energy-efficient data centers and cloud computing systems," *Advances Comput.*, vol. 82, no. 2, pp. 47–111, 2011.

[16] P. Endo, G. Gonçalves, J. Kelner, and D. Sadok, "A survey on open-source cloud computing solutions," in *Proc. VIII Workshop em Clouds, Grids e Aplicações*, 2010, pp. 3–16.

[17] M. Zhou, R. Zhang, D. Zeng, and W. Qian, "Services in the cloud computing era: A survey," in *Proc. 4th IEEE Int. Universal Communication Symp. (IUCS), 2010*, 2010, pp. 40–46.

[18] Y. Jin, Y. Wen, G. Shi, G. Wang, and A. Vasilakos, "CoDaaS: An Experimental cloud-centric content delivery platform for user-generated contents," in *Proc. 2012 IEEE Int. Conf. Computing, Networking and Communications*. Maui, HI, USA: , 2012.

[19] Fernando, Niroshinie, Loke, W. Seng, Rahayu, and Wenny, "Mobile cloud computing: A survey," *Future Gener. Comput. Syst.*, vol. 29, no. 1, pp. 84–106, Jan. 2013.

[20] Z. Sanaei, S. Abolfazli, A. Gani, and R. Buyya, "Heterogeneity in mobile cloud computing: Taxonomy and open challenges," *IEEE Commun. Surveys & Tutorials,* vol. 16, no. 1, pp. 1–24, 2013.

[21] A. Klein, C. Mannweiler, J. Schneider, and H. Schotten, "Access schemes for mobile cloud computing," in *Proc. 11th Int. Conf. Mobile Data Management (MDM), 2010*, 2010, pp. 387–392.

[22] H. Harada, "Cognitive wireless cloud: A network concept to handle heterogeneous and spectrum sharing type radio access networks," in *Proc. IEEE 20th Int. Symp. Personal, Indoor and Mobile Radio Communications, 2009*, 2009, pp. 1–5.

[23] M. Satyanarayanan, P. Bahl, R. Caceres, and N. Davies, "The case for VM-based cloudlets in mobile computing," *IEEE Pervasive Comput.* p. 1, 2011.

[24] T. Verbelen, P. Simoens, F. De Turck, and B. Dhoedt, "Cloudlets: Bringing the cloud to the mobile user," in *Proc. 3rd ACM Workshop Mobile Cloud Computing and Services (MCS '12)*, 2012, pp. 29–36.

[25] S. Ferretti, V. Ghini, F. Panzieri, and E. Turrini, "Seamless support of multimedia distributed applications through a cloud," in *Proc. IEEE 3rd Int. Conf. Cloud Computing (CLOUD), 2010*, Jul. 2010, pp. 548–549.

[26] H. Ahlehagh and S. Dey, "Hierarchical video caching in wireless cloud: Approaches and algorithms," in *Proc. IEEE Int. Conf. Communications (ICC), 2012*, 2012, pp. 7082–7087.

[27] W. Pu, Z. Zou, and C. W. Chen, "Dynamic adaptive streaming over http from multiple content distribution servers," in *Proc. 2011 IEEE Global Telecommunications Conf. (GLOBECOM 2011)*, 2011, pp. 1–5.

[28] W. Pu, Z. Zou, and C. W. Chen, "Video adaptation proxy for wireless dynamic adaptive streaming over http," in *Proc. 2012 19th Int. Packet Video Workshop (PV)*, 2012, pp. 65–70.

[29] J. Daniels, "Server virtualization architecture and implementation," *Crossroads*, vol. 16, no. 1, pp. 8–12, Sep. 2009.

[30] M.-J. Montpetit, "Your content, your networks, your devices: Social networks meet your TV experience," *Comput. Entertain.*, vol. 7, no. 3, pp. 34:1–34:3, Sep. 2009.

[31] B.-G. Chun, S. Ihm, P. Maniatis, M. Naik, and A. Patti, "CloneCloud: Elastic execution between mobile device and cloud," in *Proc. EuroSys'11*, Salzburg, Austria, Apr. 2011.

[32] S. Kosta, A. Aucinas, P. Hui, R. Mortier, and X. Zhang, "ThinkAir: Dynamic resource allocation and parallel execution in the cloud for mobile code offloading," in *Proc. IEEE Int. Conf. Computer Communications Mini-Conf. (INFOCOM'12)*, Houston, TX, USA, Mar. 2012.

[33] X. Zhang, A. Kunjithapatham, S. Jeong, and S. Gibbs, "Towards an elastic application model for augmenting the computing capabilities of mobile devices with cloud computing," *Mobile Netw. Applicat.*, 2011.

[34] Y. Wen, W. Zhang, and H. Luo, "Energy-optimal mobile application execution: Taming resource-poor mobile devices with cloud clones," in *Proc. IEEE Int. Conf. Computer Communications Mini-Conf. (INFOCOM'12)*, Houston, TX, USA, Mar. 2012.

[35] P. Stuedi, I. Mohomed, and D. Terry, "WhereStore: Location-based data storage for mobile devices interacting with the cloud," in *Proc. ACM MCS'10*, San Francisco, CA, USA, Jun. 2010.

[36] D. Neumann, C. Bodenstein, O. F. Rana, and R. Krishnaswamy, "STACEE: Enhancing storage clouds using edge devices," in *Proc. ACM ACE'11*, Karlsruhe, Germany, Jun. 2011.

[37] D. Miao, W. Zhu, C. Luo, and C. W. Chen, "Resource allocation for cloud-based free viewpoint video rendering for mobile phones," in *Proc. 19th ACM Int. Conf. Multimedia (MM '11)*, New York, NY, USA, 2011, pp. 1237–1240.

[38] X. Nan, Y. He, and L. Guan, "Optimal resource allocation for multimedia cloud based on queuing model," in *Proc. IEEE Int. Workshop Multimedia Signal Processing (MMSP'11)*, Hangzhou, China, Oct. 2011.

[39] F. Chen, K. Guoy, J. Liny, and T. L. Porta, "Intra-cloud lightning: Building CDNs in the cloud," in *Proc. IEEE Int. Conf. Computer Communications Mini-Conf. (INFOCOM'12)*, Houston, TX, USA, Mar. 2012.

[40] X. Cheng and J. Liu, "Load-balanced migration of social media to content clouds," in *Proc. NOSSDAV'11*, Vancouver, BC, Canada, Jun. 2011.

[41] F. Wang, J. Liu, and M. Chen, "CALMS: Cloud-assisted live media streaming for globalized demands with time/region diversities," in *Proc. IEEE Int. Conf. Computer Communications Mini-Conf. (INFOCOM'12)*, Houston, TX, USA, Mar. 2012.

[42] X. Qiu, H. Li, C. Wu, Z. Liy, and F. C. Lau, "Dynamic scaling of VoD services into hybrid clouds with cost minimization and QoS guarantee," in *Proc. IEEE Int. Packet Video Workshop (PV'12)*, Munich, Germany, May 2012.

[43] D. Niu, H. Xu, B. Li, and S. Zhao, "Quality-assured cloud bandwidth auto-scaling for video-on-demand applications," in *Proc. IEEE Int. Conf. Computer Communications Mini-Conf. (INFOCOM'12)*, Houston, TX, USA, Mar. 2012.

[44] D. P. Bertsekas, *Dynamic Programming and Optimal Control, Two Volume Set*, 2nd ed. Nashua, NH, USA: Athena Scientific, 2001.

[45] D. P. Bertsekas and S. E. Shreve, *Stochastic Optimal Control: The Discrete-Time Case*. Orlando, FL, USA: Academic, 1978.

[46] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge Univ. Press, 2004.

[47] A. Vetro, C. Christopoulos, and H. Sun, "Video transcoding architectures and techniques: An overview," *IEEE Signal Process. Mag.*, vol. 20, no. 2, pp. 18–29, Mar. 2003.

[48] T. Bailloeul, C. Zhu, and Y. Xu, "Automatic image tagging as a random walk with priors on the canonical correlation subspace," in *Proc. 1st ACM Int. Conf. Multimedia Information Retrieval*, 2008, pp. 75–82.

[49] B. Girod, M. Kalman, Y. Liang, and R. Zhang, "Advances in channel-adaptive video streaming," *Wirel. Commun. Mobile Comput.*, vol. 2, no. 6, pp. 573–584, 2002.

[50] A. G. Kunzel, H. Kalva, and B. Furht, "A study of transcoding on cloud environments for video content delivery," in *Proc. ACM MCMC'10*, Firenze, Italy, Oct. 2010.

[51] Z. Huang, C. Mei, L. E. Li, and T. Woo, "CloudStream: Delivering high-quality streaming video through a cloud-based H.264/SVC proxy," in *Proc. IEEE Int. Conf. Computer Communications Mini-Conf. (INFOCOM'11)*, Orlando, FL, USA, Mar. 2001.

[52] A. Garcia and H. Kalva, "Cloud transcoding for mobile video content delivery," in *Proc. IEEE Int. Conf. Consumer Electronics (ICCE), 2011*, Jan. 2011, pp. 379–380.

[53] W. Shi, Y. Lu, Z. Li, and J. Engelsma, "Scalable support for 3D graphics applications in cloud," in *Proc. IEEE 3rd Int. Conf. Cloud Computing*, 2010.

[54] R. Pereira, M. Azambuja, K. Breitman, and M. Endler, "An architecture for distributed high performance video processing in the cloud," in *Proc. IEEE 3rd Int. Conf. Cloud Computing (CLOUD), 2010*, Jul. 2010, pp. 482–489.

[55] R. Pereira and K. Breitman, "A cloud based architecture for improving video compression time efficiency: The split amp; merge approach," in *Proc. Data Compression Conf. (DCC), 2011*, Mar. 2011, p. 471.

[56] L. Zheng, L. Tian, and Y. Wu, "A rate control scheme for distributed high performance video encoding in cloud," in *Proc. Int. Conf. Cloud and Service Computing (CSC), 2011*, Dec. 2011, pp. 131–133.

[57] H. Yue, X. Sun, J. Yang, and F. Wu, "Cloud-based image coding for mobile devices - toward thousands to one compression," *IEEE Trans. Multimedia*, to be published.

[58] G. Zhang, Y. Wen, J. Zhu, and Q. Chen, "On file delay minimization for content uploading to media cloud via collaborative wireless network," in *Proc. IEEE WCSP'11*, 2011.

[59] J. Sun, Y. Wen, and L. Zheng, "On file-based content distribution over wireless networks via multiple paths: Coding and delay trade-off," in *Proc. IEEE INFOCOM, 2011* , 2011, pp. 381–385.

[60] Y. Wen, G. Zhang, and X. Zhu, "Lightweight packet scheduling algorithms for content uploading from mobile devices to media cloud," in *Proc. 2nd IEEE Workshop Multimedia Communications & Services—IEEE GLOBECOM 2011*, 2011.

[61] Y. Huang, Z. Li, G. Liu, and Y. Dai, "Cloud download: Using cloud utilities to achieve high-quality content distribution for unpopular videos," in *Proc. ACM Multimedia (MM'11)*, Scottsdale, AZ, USA, Nov. 2011.

[62] Y. Wen, G. Shi, and G. Wang, "Designing an inter-cloud messaging protocol for content distribution as a service (CoDaaS) over future internet," in *Proc. ACM CFI'11*, Seoul, Korea, Jun. 2011.

[63] H. A. Tran, A. Mellouk, and S. Hoceini, "QoE content distribution network for cloud architecture," in *Proc. 1st Int. Symp. Network Cloud Computing and Applications (NCCA), 2011*, 2011, pp. 14–19.

[64] J. Cervino, P. Rodriguez, I. Trajkovska, A. Mozo, and J. Salvachua, "Testing a cloud provider network for hybrid P2P and cloud streaming architectures," in *Proc. IEEE 4th Int. Conf. Cloud Computing*, 2011.

[65] Y. Wu, C. Wu, B. Li, L. Zhang, Z. Li, and F. C. Lau, "Scaling social media applications into geo-distributed clouds," in *Proc. IEEE Infocom 2012*, Mar. 2012.

[66] Y. Cui, M. Kim, H. gun Yoon, and H. Lee, "SMSS: Social media sharing system using UPnP in cloud computing environment," in *Proc. 3rd Int. Conf. Internet (ICONI)*, 2011.

[67] D. Daz-Sánchez, F. Almenares, A. Marin, and D. Proserpio, "Media cloud: Sharing contents in the large," in *Proc. IEEE Int. Conf. Consumer Electronics (ICCE'11)*, 2011.

[68] D. Daz-Sánchez, F. Almenarez, A. Marn, D. Proserpio, and P. A. Cabarcos, "Media cloud: An open cloud computing middleware for content management," *IEEE Trans. Consum. Electron.*, vol. 57, no. 2, May 2011.

[69] D. Daz-Sánchez, F. Almenarez, A. Marn, P. Arias, R. Sánchez-Guerrero, and F. Sanvido, "A privacy aware media gateway for connecting private multimedia clouds to limited devices," in *Proc. IFIP WMNC'11*, 2011.

[70] X. Zhu, J. Zhu, R. Pan, M. Prabhu, and F. Bonomi, "Cloud-assisted streaming for low-latency applications," in *Proc. Int. Conf. Computing, Networking and Communications (ICNC)*, Feb. 30–2, 2012, pp. 949–953.

[71] W. Yin, J. Luo, and C. W. Chen, "Event-based semantic image adaptation for user-centric mobile display devices," *IEEE Trans. Multimedia*, vol. 13, no. 3, pp. 432–442, 2011.

[72] Popcorn Maker [Online]. Available: https://popcorn.webmaker.org/

[73] S. Shi, W. J. Jeon, K. Nahrstedt, and R. H. Campbell, "Real-time remote rendering of 3D video for mobile devices," in *Proc. 17th ACM Int. Conf. Multimedia (MM '09)*, New York, NY, USA, 2009, pp. 391–400.

[74] K. Ota, H. Kubota, and T. Gotoh, "Media cloud service with optimized video processing and platform," *FUJITSU Sci. Tech. J.*, 2011.

[75] K. S. Candan, "RanKloud: Scalable multimedia and social media retrieval and analysis in the cloud," in *Proc. 9th Workshop Large-Scale and Distributed Informational Retrieval (LSDS-IR '11)*, New York, NY, USA, 2011, pp. 1–2.

[76] W. Yin, X. Zhu, and C. W. Chen, "Contemporary ubiquitous media services: Content recommendation and adaptation," in *Proc. IEEE First PerCom Workshop Pervasive Communications and Service Clouds*, 2011.

[77] K. S. Gopalan, S. Nathan, B. T. C. H, A. B. Channa, P. Saraf, and G. Shanker, "A cloud based service architecture for personalized media recommendations," in *Proc. 5th Int. Conf. Next Generation Mobile Applications and Services*, 2011.

[78] X. Giro-i Nieto, C. Ventura, J. Pont-Tuset, S. Cortes, and F. Marques, "System architecture of a web service for Content-Based Image Retrieval," in *Proc. ACM Int. Conf. Image and Video Retrieval (CIVR '10)*, New York, NY, USA, 2010, pp. 358–365.

[79] R. Lämmel, "Google's mapreduce programming model–revisited," *Sci. Comput. Program.*, vol. 70, no. 1, pp. 1–30, Jan. 2008.

[80] S. Liu and C. W. Chen, "A novel 3D video transcoding scheme for adaptive 3D video transmission to heterogeneous terminals," *ACM Trans. Multimedia Comput. Commun. Applicat.*, vol. 8, no. 3s, pp. 43:1–43:21, Oct. 2012.

[81] G. Pallis and A. Vakali, "Insight and perspectives for content delivery networks," *Commun. ACM*, vol. 49, no. 1, pp. 101–106, Jan. 2006.

[82] Y. Wen, P. Sun, and J. Cai, "Codaas: Content delivery as a service for user generated contents (invited paper)," *J. Internet Technol.*, vol. 14, no. 3, pp. 353–364, May 2013.

[83] A. Loui, J. Luo, S.-F. Chang, D. Ellis, W. Jiang, L. Kennedy, K. Lee, and A. Yanagawa, "Kodak's consumer video benchmark data set: Concept definition and annotation," in *Proc. Int. Workshop Multimedia Information Retrieval (MIR '07)*, New York, NY, USA, 2007, pp. 245–254.

[84] , D. Agrawal, K. S. Candan, and W. S. L. , Eds., *New Frontiers in Information and Software as Services: Service and Application Design Challenges in the Cloud*. Berlin, Germany: Springer, 2011.

[85] M. Isard, M. Budiu, Y. Yu, A. Birrell, and D. Fetterly, "Dryad: Distributed data-parallel programs from sequential building blocks," in *Proc. 2nd ACM SIGOPS/EuroSys European Conf. Computer Systems 2007 (EuroSys '07)*, New York, NY, USA, 2007, pp. 59–72.

[86] H. Espeland, P. Beskow, H. Stensland, P. Olsen, S. Kristoffersen, C. Griwodz, and P. Halvorsen, "P2G: A framework for distributed real-time processing of multimedia data," in *Proc. 40th Int. Conf. Parallel Processing Workshops (ICPPW), 2011*, Sep. 2011, pp. 416–426.

[87] J.-L. Chen, S.-L. Wu, Y. Larosa, P.-J. Yang, and Y.-F. Li, "IMS cloud computing architecture for high-quality multimedia applications," in *Proc. 2011 7th Int. Wireless Communications and Mobile Computing Conf. (IWCMC)*, Jul. 2011, pp. 1463–1468.

[88] Z. Ye, X. Chen, and Z. Li, "Video based mobile location search with large set of SIFT points in cloud," in *Proc. 2010 ACM Multimedia Workshop Mobile Cloud Media Computing (MCMC '10)*, New York, NY, USA, 2010, pp. 25–30.

[89] C. Keimel, J. Habigt, C. Horch, and K. Diepold, "Video quality evaluation in the cloud," in *Proc. IEEE Int. Packet Video Workshop (PV'12)*, Munich, Germany, May 2012.

[90] R. H. Glitho, "Cloud-based multimedia conferencing: Business model, research agenda, state-of-the-art," in *Proc. IEEE Conf. Commerce and Enterprise Computing*, 2011.

[91] C. Gadea, B. Solomon, B. Ionescu, and D. Ionescu, "A collaborative cloud-based multimedia sharing platform for social networking environments," in *Proc. 20th Int. Conf. Computer Communications and Networks (ICCCN'11)*, 2011.

[92] E. Vartiainen and K. Väänänen-Vainio-Mattila, "User experience of mobile photo sharing in the cloud," in *Proc. ACM MUM '10*, Limassol, Cyprus, Dec. 2010.

[93] S. Saranya and M. Vijayalakshmi, "Interactive mobile live video learning system in cloud environment," in *Proc. Int. Conf. Recent Trends in Information Technology (ICRTIT), 2011*, Jun. 2011, pp. 673–677.

[94] T. Ali, M. Nauman, F.-e. Hadi, and F. bin Muhaya, "On usage control of multimedia content in and through cloud computing paradigm," in *Proc. 5th Int. Conf. Future Information Technology (FutureTech), 2010*, May 2010, pp. 1–5.

[95] H. Liang, D. Huang, L. Cai, X. Shen, and D. Peng, "Resource allocation for security services in mobile cloud computing," in *Proc. IEEE Conf. Computer Communications Workshops (INFOCOM WKSHPS), 2011*, Apr. 2011, pp. 191–195.

[96] C.-T. Huang, Z. Qin, and C.-C. Kuo, "Multimedia storage security in cloud computing: An overview," in *Proc. IEEE 13th Int. Workshop Multimedia Signal Processing (MMSP), 2011*, Oct. 2011, pp. 1–6.

[97] V. Jacobson, D. K. Smetters, J. D. Thornton, M. F. Plass, N. H. Briggs, and R. L. Braynard, "Networking named content," in *Proc. 5th Int. Conf. Emerging Networking Experiments and Technologies (CoNEXT '09)*, New York, NY, USA, 2009, pp. 1–12.

[98] X. Zhang, K. Chang, H. Xiong, Y. Wen, G. Shi, and G. Wang, "Towards name-based trust and security for content-centric network," in *Proc. 2011 19th IEEE Int. Conf. Network Protocols (ICNP '11)*, Washington, DC, USA, 2011, pp. 1–6, IEEE Computer Society.

[99] D. Niu, C. Feng, and B. Li, "A theory of cloud bandwidth pricing for video-on-demand providers," in *Proc. IEEE Int. Conf. Computer Communications Mini-Conf. (INFOCOM'12)*, Houston, TX, USA, Mar. 2012.

[100] "The new multi-screen world: Understanding cross-platform consumer behavior, white paper, Google," Aug. 2012 [Online]. Available: http://services.google.com/fh/files/misc/multiscreenworld_final.pdf

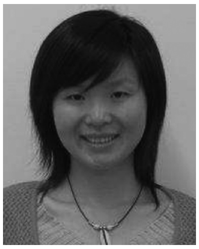
[101] "Software-defined networking: The new norm for networks, white paper, open networking foundation (ONF)," Apr. 2012 [Online]. Available: https://www.opennetworking.org/images/stories/downloads/white-papers/wp-sdn-newnorm.pdf

[102] L. Zhang, D. Estrin, J. Burke, V. Jacobson, J. D. Thornton, D. K. Smetters, B. Zhang, G. Tsudik, D. Massey, and C. Papadopoulos *et al.*, "Named data networking (ndn) project," Relatório Técnico NDN-0001, Xerox Palo Alto Research Center-PARC, 2010.

[103] J. Gantz and D. Reinsel, "Extracting value from chaos, White Paper, IDC," Jun. 2011 [Online]. Available: http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf

Yonggang Wen (S'00–M'08–SM'14) has been an assistant professor with school of computer engineering at Nanyang Technological University, Singapore, since 2011. He received his PhD degree in Electrical Engineering and Computer Science (minor in Western Literature) from Massachusetts Institute of Technology (MIT), Cambridge, USA. Previously he has worked in Cisco to lead product development in content delivery network, which had a revenue impact of 3 Billion US dollars globally. Dr. Wen has published over 90 papers in top journals and prestigious conferences. His latest work in multi-screen cloud social TV has been featured by global media (more than 1600 news articles from over 29 countries) and recognized with ASEAN ICT Award 2013 (Gold Medal) and IEEE Globecom 2013 Best Paper Award. He serves on editorial boards for IEEE Transactions on Multimedia, IEEE Access Journal and Elsevier Ad Hoc Networks.

Dr. Wen's research interests include cloud computing, green data center, big data analytics, multimedia network and mobile computing.



Xiaoqing Zhu is a Technical Leader at the Enterprise Networking Lab at Cisco Systems Inc. She received the B.Eng. degree in Electronics Engineering from Tsinghua University, Beijing, China. She earned both the M.S. and Ph.D. degrees in Electrical Engineering from Stanford University, California, USA. Prior to joining Cisco, Dr. Zhu interned at IBM Almaden Research Center in 2003, and at Sharp Labs of America in 2006. She received the best student paper award in ACM Multimedia 2007.

Dr. Zhu's research interests span across multimedia applications, networking, and wireless communications. She has served as reviewer, TPC member, and special session organizer for various journals, magazines, conferences and workshops. Previously, she contributed as guest editor to several special issues in IEEE Technical Committee on Multimedia Communications (MMTC) E-Letter, IEEE Journal on Selected Areas in Communications, and IEEE Transactions on Multimedia.

Joel J. P. C. Rodrigues (S'01–M'06–SM'06) received a PhD degree in informatics engineering, an MSc degree from the University of Beira Interior, and a five-year BSc degree (licentiate) in informatics engineering from the University of Coimbra, Portugal. He is currently a Professor in the Department of Informatics of the University of Beira Interior, Covilhã, Portugal, and a Researcher at the Instituto de Telecomunicações, Portugal. His main research interests include sensor networks, e-health, e-learning, vehicular communications, and mobile and ubiquitous computing.

He is the leader of NetGNA Research Group (http://netgna.it.ubi.pt), the Chair of the IEEE ComSoc Technical Committee on eHealth, the Past-chair of the IEEE ComSoc Technical Committee on Communications Software, Member Representative of the IEEE Communications Society on the IEEE Biometrics Council, Steering Committee member of the IEEE Life Sciences Technical Community, and officer of the IEEE 1907.1 standard. He is the editor-in-chief of the International Journal on E-Health and Medical Communications, the editor-in-chief of the Recent Advances on Communications and Networking Technology, and editorial board member of several journals, including IEEE Communications Magazine, IEEE Communications Surveys and Tutorials, Elsevier Journal of Computer Networks and Applications, Elsevier Computer Networks, Elsevier Journal of Vehicular Communications, Wiley Transactions on Emerging Telecommunications Technologies, and Wiley International Journal of Communications Systems. He has served as a guest editor for a number of journals and has been General Chair and TPC Chair of many international conferences. He is a member of many international TPCs and participated in several international conferences organization. He has authored or coauthored over 350 papers in refereed international journals and conferences, a book, and 2 patents. He had been awarded the Outstanding Leadership Award of IEEE GLOBECOM 2010 as CSSMA Symposium Co-Chair and several best papers awards.

Prof. Rodrigues is a licensed professional engineer (as senior member), member of the Internet Society, an IARIA fellow, and a senior member of ACM and IEEE.

Chang Wen Chen (F'04) is a Professor of Computer Science and Engineering at the University at Buffalo, State University of New York. He has been 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 Transactions on Multimedia since January 2014. He has also served as the Editor-in-Chief for IEEE Transactions on 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 the 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 received his BS 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 and his students have received seven 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 2009, 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.