Wenwu Zhu, Chong Luo, Jianfeng Wang, and Shipeng Li

Multimedia Cloud Computing

An emerging technology for providing multimedia services and applications

This article introduces the principal concepts of multimedia cloud computing and presents a novel framework. We address multimedia cloud computing from multimedia-aware cloud (media cloud) and cloud-aware multimedia (cloud media) perspectives. First, we present a multimedia-aware cloud, which addresses how a cloud can perform distributed multimedia processing and storage and provide quality of service (QoS) provisioning for multimedia services. To achieve a high QoS for multimedia services, we propose a media-edge cloud (MEC) architecture, in which storage, central processing unit (CPU), and graphics processing unit (GPU) clusters are presented at the edge to provide distributed parallel processing and QoS adaptation for various types of devices. Then we present a cloud-aware multimedia, which addresses how multimedia services and applications, such as storage and sharing, authoring and mashup, adaptation and delivery, and rendering and retrieval, can optimally utilize cloud-computing resources to achieve better quality of experience (QoE). The research directions and problems encountered are presented accordingly.

Digital Object Identifier 10.1109/MSP.2011.940269
Date of publication: 19 April 2011

INTRODUCTION

Cloud computing is an emerging technology aimed at providing various computing and storage services over the Internet [1], [2]. It generally incorporates infrastructure, platform, and

software as services. Cloud service providers rent data-center hardware and software to deliver storage and computing services through the Internet. By using cloud computing, Internet users can receive services from a cloud as if they were employing a super computer. They can store their data in the cloud instead of on their own devices, making ubiquitous data access possible. They can run their applications on much more powerful cloud-computing platforms with software deployed in the cloud, mitigating the users' burden of full software installation and continual upgrade on their local devices.

MULTIMEDIA PROCESSING IN A CLOUD IMPOSES GREAT CHALLENGES.

With the development of Web 2.0, Internet multimedia is emerging as a service. To provide rich media services, multimedia computing has emerged as a noteworthy technology to generate, edit, process, and search media contents, such as images, video, audio, graphics, and so on. For multimedia applications and services over the Internet and mobile wireless networks, there are strong demands for cloud computing because of the significant amount of computation required for serving millions of Internet or mobile users at the same time. In this new cloud-based multimedia-computing paradigm, users store and process their multimedia application data in the cloud in a distributed manner, eliminating full installation of the media application software on the users' computer or device and thus alleviating the burden of multimedia software maintenance and upgrade as well as sparing the computation of user devices and saving the battery of mobile phones.

Multimedia processing in a cloud imposes great challenges. Several fundamental challenges for multimedia computing in the cloud are highlighted as follows.

1) *Multimedia and service heterogeneity*: As there exist different types of multimedia and services, such as voice over IP (VoIP), video conferencing, photo sharing and editing, multimedia streaming, image search, image-based rendering, video transcoding and adaptation, and multimedia content delivery, the cloud shall support different types of multimedia and multimedia services for millions of users simultaneously.

2) *QoS heterogeneity*: As different multimedia services have different QoS requirements, the cloud shall provide QoS provisioning and support for various types of multimedia services to meet different multimedia QoS requirements.

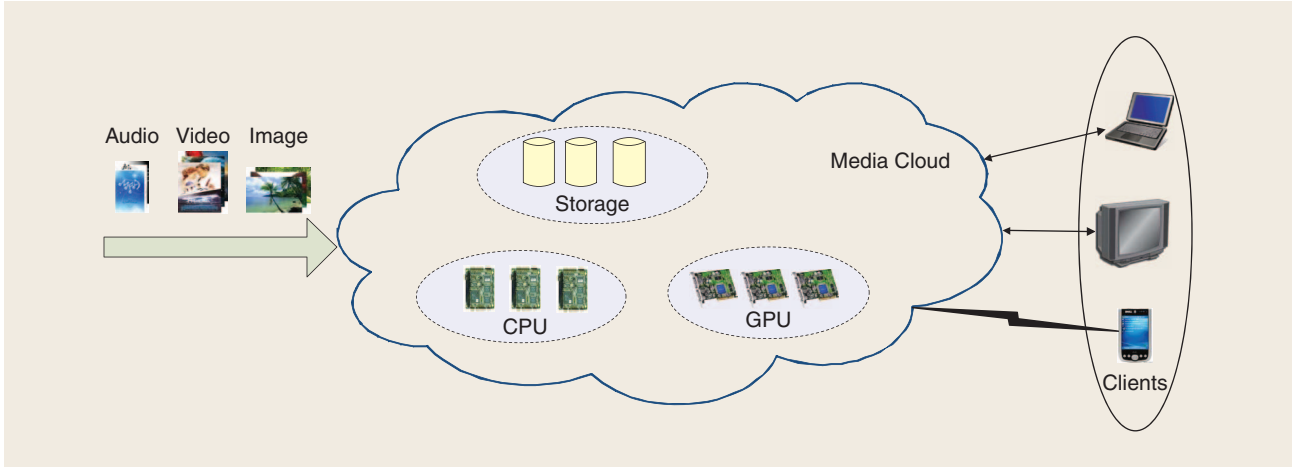
3) *Network heterogeneity*: As different networks, such as Internet, wireless local area network (LAN), and third generation wireless network, have different network characteristics, such as bandwidth, delay, and jitter, the cloud shall adapt multimedia contents for optimal delivery to various types of devices with different network bandwidths and latencies.

4) *Device heterogeneity*: As different types of devices, such as TVs, personal computers (PCs), and mobile phones, have different capabilities for multimedia processing, the cloud shall have multimedia adaptation capability to fit different types of devices, including CPU, GPU, display, memory, storage, and power.

For multimedia computing in a cloud, simultaneous bursts of multimedia data access, processing, and transmission in the cloud would create a bottleneck in a general-purpose cloud because of stringent multimedia QoS requirements and large amounts of users' simultaneous accesses at the Internet scale. In today's cloud computing, the cloud uses a utilitylike mechanism to allocate how much computing (e.g., CPU) and storage resources one needs, which is very effective for general data services. However, for multimedia applications, in addition to the CPU and storage requirements, another very important factor is the QoS requirement in terms of bandwidth, delay, and jitter. Therefore, using a general-purpose cloud in the Internet to deal with multimedia services may suffer from unacceptable media QoS or QoE [3]. Mobile devices have limitations in memory, computing power, and battery life; thus, they have even more prominent needs to use a cloud to address the tradeoff between computation and communication. It is foreseen that cloud computing could become a disruptive technology for mobile applications and services [4]. More specifically, in mobile media applications and services, because of the power requirement for multimedia [5] and the time-varying characteristics of the wireless channels, QoS requirements in cloud computing for mobile multimedia applications and services become more stringent than those for the Internet cases. In summary, for multimedia computing in a cloud, the key is how to provide QoS provisioning and support for multimedia applications and services over the (wired) Internet and mobile Internet.

To meet multimedia's QoS requirements in cloud computing for multimedia services over the Internet and mobile wireless networks, we introduce the principal concepts of multimedia cloud computing for multimedia computing and communications, shown in Figure 1. Specifically, we propose a multimedia-cloud-computing framework that leverages cloud computing to provide multimedia applications and services over the Internet and mobile Internet with QoS provisioning. We address multimedia cloud computing from multimedia-aware cloud (media cloud) and cloud-aware multimedia (cloud media) perspectives. A multimedia-aware cloud focuses on how the cloud can provide QoS provisioning for multimedia applications and services. Cloud-aware multimedia focuses on how multimedia can perform its content storage, processing, adaptation, rendering, and so on, in the cloud to best utilize cloud-computing resources, resulting in high QoE for multimedia services. Figure 2 depicts the relationship of the media cloud and cloud media services. More specifically, the media cloud provides raw resources, such as hard disk, CPU, and GPU, rented by the media service providers (MSPs) to serve users. MSPs use media cloud resources to develop their multimedia applications and services, e.g., storage, editing, streaming, and delivery.

On the media cloud, we propose an MEC architecture to reduce latency, in which media contents and processing are pushed to the edge of the cloud based on user's context or



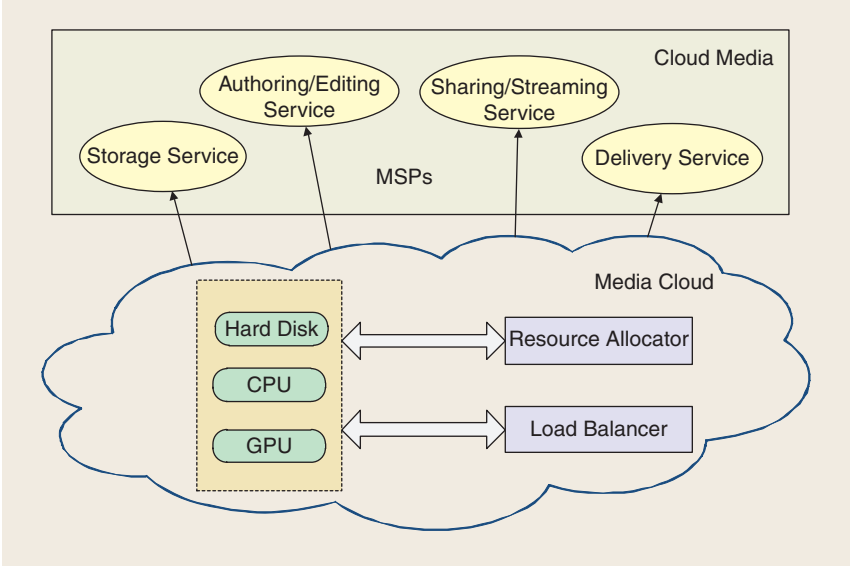
[FIG1] **Fundamental concept of multimedia cloud computing.**

profile. In this architecture, an MEC is a cloudlet with data centers physically placed at the edge. The MEC stores, processes, and transmits media data at the edge, thus achieving a shorter delay. The media cloud is composed of MECs, which can be managed in a centralized or peer-to-peer (P2P) manner. First, to better handle various types of media services in an MEC, we propose to place similar types of media services into a cluster of servers based on the properties of media services. Specifically, we propose to use the distributed hash table (DHT [6]) for data storage while using CPU or GPU clusters for multimedia computing. Second, for computing efficiency in the MEC, we propose a distributed parallel processing model for multimedia applications and services in GPU or CPU clusters. Third, at the proxy/edge server of the MEC, we propose media adaptation/transcoding for media services to heterogeneous devices to achieve high QoE.

On cloud media, media applications and services in the cloud can be conducted either completely or partially in the cloud. In the former case, the cloud will do all the multimedia computing, e.g., for the case of thin-client mobile phones. In the latter case, the key problem is how to allocate multimedia-computing (e.g., CPU and GPU) resources between the clients and cloud, which will involve client–cloud resource partition for multimedia computing. In this article, we present how multimedia applications, such as processing, adaptation, rendering, and so on, can optimally utilize cloud-computing resources to achieve high QoE.

RELATED WORKS

Multimedia cloud computing is generally related to multimedia computing over grids, content delivery network (CDN), server-based computing, and P2P multimedia computing. More specifically, multimedia computing over grids addresses infrastructure computing for multimedia from a high-performance computing (HPC) aspect [7]. The CDN addresses how to deliver multimedia at the edge so as to reduce the delivery latency or maximize the bandwidth for the clients to access the data. Examples include Akamai Technologies, Amazon CloudFront, and Limelight Networks. YouTube uses Akamai's CDN to deliver videos. Server-based multimedia computing addresses desktop computing, in which all multimedia computing is done in a set of servers, and the client interacts only with the servers [8]. Examples include Microsoft Remote Display Protocol and AT&T Virtual Network Computing. P2P multimedia computing refers to a distributed application architecture that partitions multimedia-computing tasks or workloads between peers. Examples include Skype, PPlive, and Coolstream. The media cloud presented in this article addresses how the cloud can provide QoS provisioning for multimedia computing in a cloud environment.



[FIG2] **The relationship of the media cloud and cloud media services.**

To our knowledge, there exist very few works on multimedia cloud computing in the literature. IBM had an initiative for cloud computing (http://www.ibm.com/ibm/cloud/resources.html#3). Trajkovska et al. proposed a joint P2P and cloud-computing architecture realization for multimedia streaming with QoS cost functions [9].

MULTIMEDIA CLOUD COMPUTING IS GENERALLY RELATED TO MULTIMEDIA COMPUTING OVER GRIDS, CONTENT DELIVERY NETWORK, SERVER-BASED COMPUTING, AND P2P MULTIMEDIA COMPUTING.

MULTIMEDIA-AWARE CLOUD

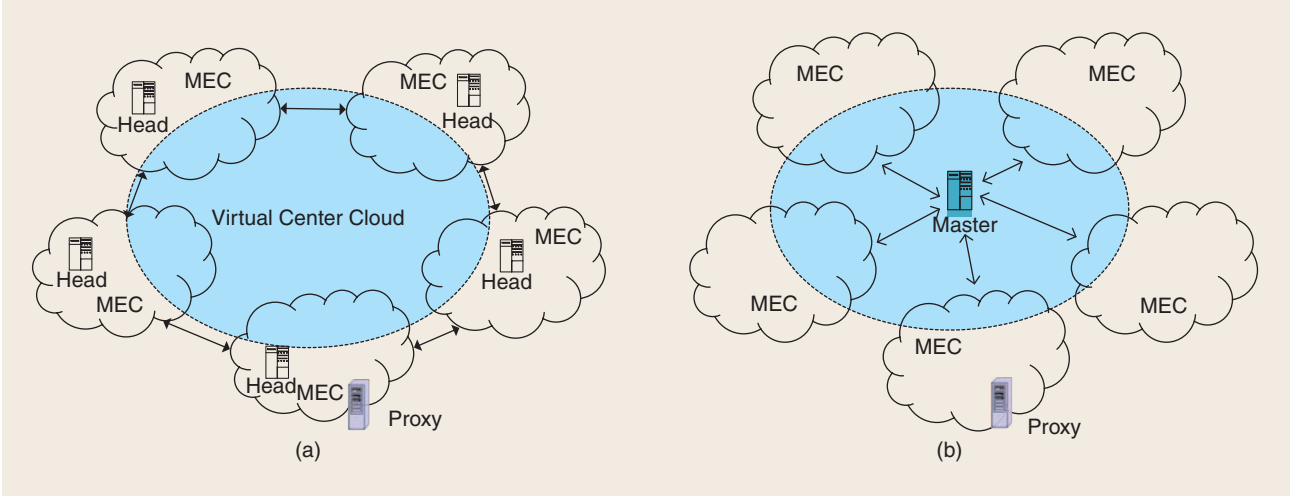
The media cloud needs to have the following functions: 1) QoS provisioning and support for various types of multimedia services with different QoS requirements, 2) distributed parallel multimedia processing, and 3) multimedia QoS adaptation to fit various types of devices and network bandwidth. In this section, we first present the architecture of the media cloud. Then we discuss the distributed parallel multimedia processing in the media cloud and how the cloud can provide QoS support for multimedia applications and services.

MEDIA-CLOUD-COMPUTING ARCHITECTURE

In this section, we present an MEC-computing architecture that aims at handling multimedia computing from a QoS perspective. An MEC is a cloudlet with servers physically placed at the edge of a cloud to provide media services with high QoS (e.g., low latency) to users. Note that an MEC architecture is like the CDN edge servers architecture, with the difference being that CDN is for multimedia delivery, while MEC is for multimedia computing. Using CDN edge servers to deliver multimedia to the end users can result in less latency than direct delivery from the original servers at the center. Thus, it can be seen that multimedia computing in an MEC can produce less multimedia traffic and reduce latency when compared to all multimedia contents that are located at the central cloud. As depicted in Figure 3(a) and (b), respectively, an MEC has two types of architectures: one is where all users' media data are stored in MECs based on their user profile or context, while all the information of the associated users and content locations is communicated by its head through P2P; the other one is where the central administrator (master) maintains all the information of the associated users and content locations, while the MEC distributedly holds all the content data. Within an MEC, we adopt P2P technology for distributed media data storage and computing. With the P2P architecture, each node is equally important and, thus, the MEC is of high scalability, availability, and robustness for media data storage and media computing. To support mobile users, we propose a cloud proxy that resides at the edge of an MEC or in the gateway, as depicted in Figure 3(a) and (b), to perform multimedia processing (e.g., adaptation and transcoding) and caching to compensate for mobile devices' limitations on computational power and battery life.

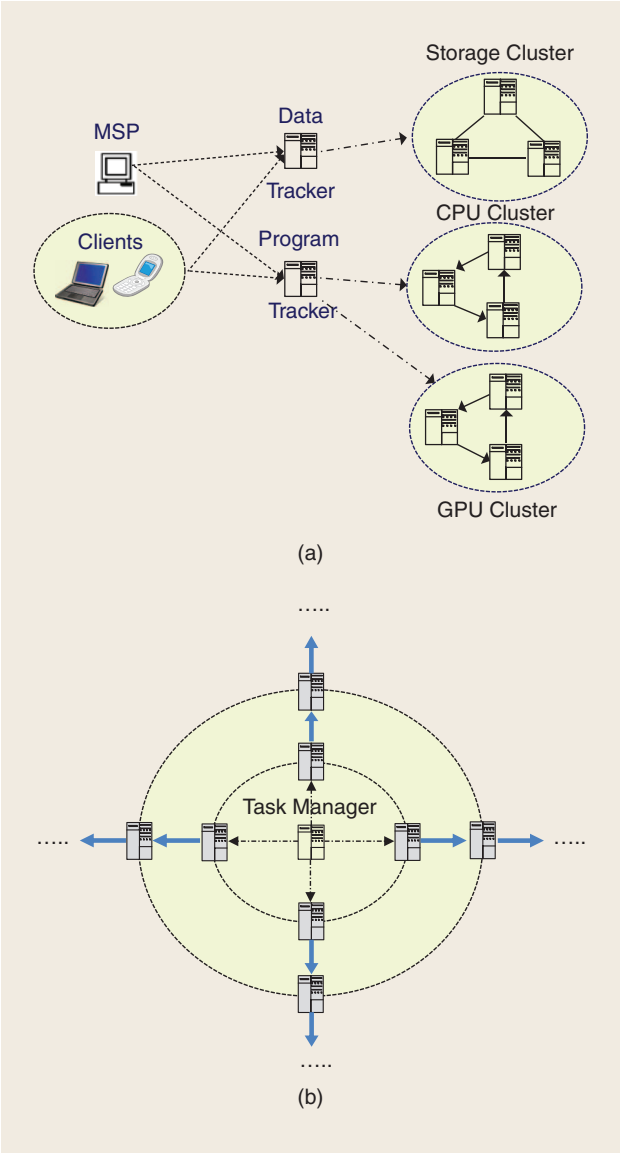
DISTRIBUTED PARALLEL MULTIMEDIA PROCESSING

Traditionally, multimedia processing is conducted on client or proprietary servers. With multimedia cloud computing, multimedia processing is usually on the third party's cloud data centers (unless one likes building a private cloud, which is costly). With multimedia processing moved to the cloud, one of the key challenges of multimedia cloud computing is how the cloud can provide distributed parallel processing of multimedia data for millions of users including mobile users. To address this problem, multimedia storage and computing need to be executed in a distributed and parallel manner. In the



[FIG3] Architecture of (a) P2P-based MEC computing and (b) central-controlled MEC computing.

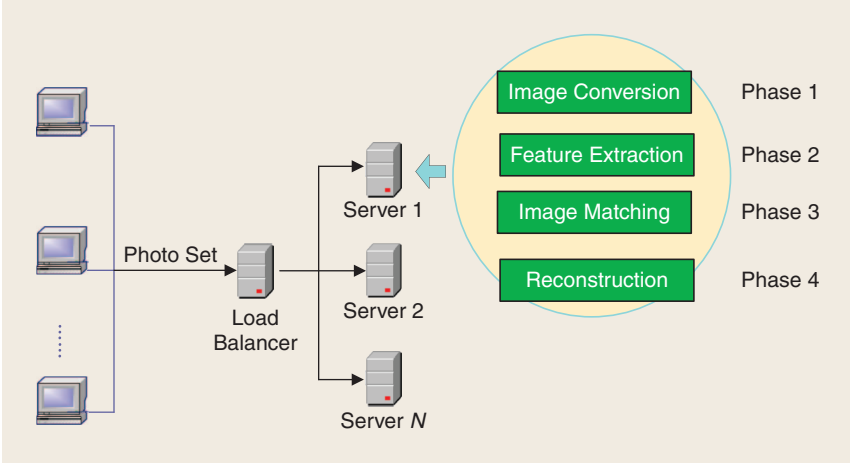
MEC, we propose to use DHT for media data storage in the storage cluster and employ multimedia processing program parallel execution model with a media load balancer for media computing in CPU or GPU clusters. As illustrated in Figure 4(a), for media storage, we assign unique keys associated with data to the data tracker, which manages how media data will be distributed in the storage cluster. For media computing, we use the program tracker or the so-called load balancer to schedule media tasks distributedly, and then the media tasks are executed in CPU or GPU clusters in the MEC. Moreover, we use parallel distributed multimedia processing for large-scale multimedia program execution in an MEC, as illustrated in Figure 4(b). Specifically, first, we can perform media load balancing at the users' level [10]. Second, we can do parallel media processing at the multimedia task level. In other words, our proposed approach not only brings distributed load balancing at users' levels but also performs multimedia task parallelization at the multimedia task levels. This is demonstrated in the following case study.



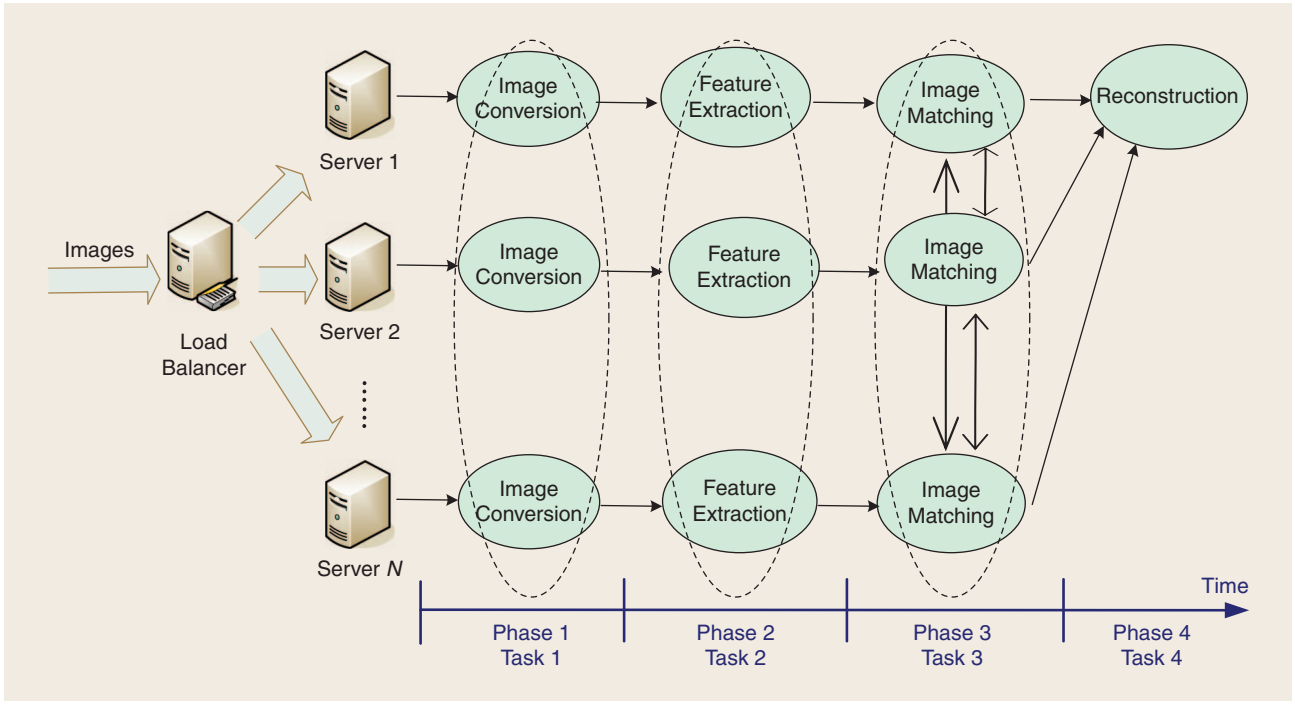
[FIG4] (a) Data and program trackers for multimedia service. (b) Distributed parallel multimedia processing.

CASE STUDY

To demonstrate what we discussed earlier, we will use Photosynth as an example to illustrate how multimedia parallel processing works and how it outperforms the traditional approach. Photosynth (http://www.photosynth.net/) is a software application that analyzes digital photographs and generates a three-dimensional (3-D) model of the photos [11], [12]. Users are able to generate their own models using the software on the client side and then upload them to the PhotoSynth Web site. Currently, computing Photosynth images is done in a local PC. The major computation tasks of Photosynth are image conversion, feature extraction, image matching, and reconstruction. In the traditional approach, the aforementioned four tasks are performed sequentially in a local client. In the cloud-computing environment, we propose that the four computation tasks of Photosynth be conducted in an MEC. This is particularly important for mobile devices because of their low computation capability and battery constraints. To reduce the computation time when dealing with a large number of users, we propose cloud-based parallel synthing with a load balancer in which the PhotoSynth algorithms are run in a parallel pipeline in an MEC. Our proposed parallel synthing consists of user- and task-level parallelization. Figure 5 shows the user-level parallel synthing in which all tasks of synthing from one user are allocated to one server to compute, but the tasks from all users can be done simultaneously in parallel in the MEC. Figure 6 shows the task-level parallel synthing in which all tasks of synthing from one user are allocated to N servers to compute



[FIG5] User-level parallel synthing in an MEC.



[FIG6] **Task-level parallel synthing in the MEC.**

in parallel. More specifically, the tasks of image conversion, feature extraction, and images matching are computed in N servers. Figure 6 shows that, theoretically, when the communication overhead is omitted, image conversion, feature extraction, and image matching can save N times less computation time using task-level parallelization with N servers than that with the one-server case.

We performed a preliminary simulation in which we ran the parallel photosynthing code in an HPC cluster node with nine servers. Tables 1 and 2 list the simulation results for 200 and 400 images, respectively. From Tables 1 and 2, we can see that for the two-server case we can achieve a 1.65 times computational gain over the traditional sequential approach, and for the nine-server case, we can achieve a 4.25 times computational gain over the traditional approach.

[TABLE 1] SIMULATION RESULTS OF 200 IMAGES PARALLEL SYNTHING IN AN HPC CLUSTER.

	ONE SERVER	TWO SERVERS		NINE SERVERS	
TASKS	**TIME (MS)**	**TIME (MS)**	**GAIN**	**TIME (MS)**	**GAIN**
IMAGE CONVERSION	65,347.25	41,089.77	1.59	15,676.95	4.17
FEATURE EXTRACTION	34,864.50	18,140.31	1.92	7,414.13	4.70
IMAGE MATCHING	47,812.50	30,103.78	1.59	11,510.03	4.15
TOTAL TIME/GAIN	148,024.25	89,333.86	1.66	34,601.11	4.28

[TABLE 2] SIMULATION RESULTS OF 400 IMAGES PARALLEL SYNTHING IN AN HPC CLUSTER.

	ONE SERVER	TWO SERVERS		NINE SERVERS	
TASKS	**TIME (MS)**	**TIME (MS)**	**GAIN**	**TIME (MS)**	**GAIN**
IMAGE CONVERSION	150,509.30	91,233.67	1.65	32,702.28	4.60
FEATURE EXTRACTION	101,437.00	51,576.00	1.97	20,391.11	4.97
IMAGE MATCHING	207,696.33	135,965.17	1.53	55,194.38	3.76
TOTAL TIME/GAIN	459,642.63	278,774.84	1.65	108,287.77	4.24

MEDIA CLOUD QoS

Another key challenge in the media cloud/MEC is QoS. There are two ways of providing QoS provisioning for multimedia: one is to add QoS to the current cloud-computing infrastructure within the cloud and the other is to add QoS middleware between the cloud infrastructure and multimedia applications. In the former case, it focuses on the cloud infrastructure QoS, providing QoS provisioning in the cloud infrastructure to support multimedia applications and services with different media QoS requirements. In the latter case, it focuses on improving cloud QoS in the middle layers, such as QoS in the transport layer and QoS mapping between the cloud infrastructure and media applications (e.g., overlay).

The cloud infrastructure QoS is a new area where more research is needed to provide stringent QoS provisioning for multimedia applications and services in the cloud. In

this article, we focus on how a cloud can provide QoS support for multimedia applications and services. Specifically, in an MEC, according to the properties of media services, we organize similar types of media services into a cluster of servers that has the best capability to process them. An MEC consists of three clusters: storage, CPU, and GPU clusters. For example, media applications related to graphics can go to the GPU cluster, normal media processing can go to the CPU cluster, and storage types of media applications can go to the storage cluster. As a result, an MEC can provide QoS support for different types of media with different QoS requirements.

THE MEDIA CLOUD PROXY IS DESIGNED TO DEAL WITH MOBILE MULTIMEDIA COMPUTING AND CACHING FOR MOBILE PHONES.

To improve multimedia QoS performance in a media cloud, in addition to moving media content and computation to the MEC to reduce latency and to perform content adaptation to heterogeneous devices, a media cloud proxy is proposed in our architecture to further reduce latency and best serve different types of devices with adaptation especially for mobile devices. The media cloud proxy is designed to deal with mobile multimedia computing and caching for mobile phones. As a mobile phone has a limited battery life and computation power, the media cloud proxy is used to perform mobile multimedia computing in full or part to compensate for the mobile phone's limitations mentioned above, including QoS adaptation to different types of terminals. In addition, the media cloud proxy can also provide a media cache for the mobile phone. Future research directions include cloud infrastructure QoS and cloud QoS overlay for multimedia services, dynamic multimedia load balancing, and so on.

CLOUD-AWARE MULTIMEDIA APPLICATIONS

The emergence of cloud computing will have a profound impact on the entire life cycle of multimedia contents. As shown in Figure 7, a typical media life cycle is composed of acquisition, storage, processing, dissemination, and presentation.

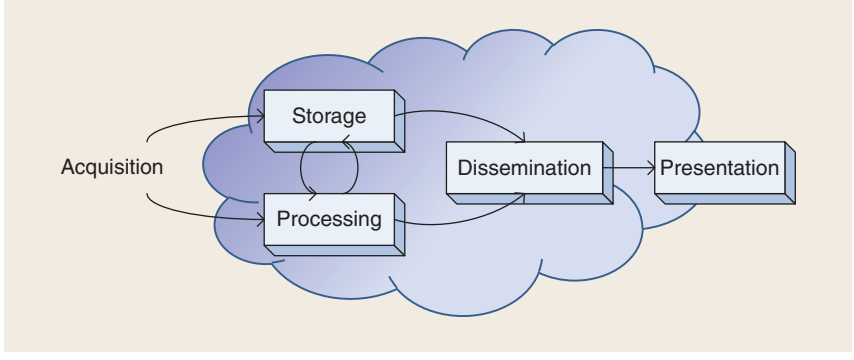
For a long time, high-quality media contents could only be acquired by professional organizations with efficient devices, and the distribution of media contents relied on hard copies, such as film, video compact disc (VCD), and DVD. During the recent decade, the availability of low-cost commodity digital cameras and camcorders has sparked an explosion of user-generated media contents. Most recently, cyber-physical systems [13] offer a new way of data acquisition through sensor networks, which significantly increases the volume and diversity of media data. Riding the Web 2.0 wave, digital media contents can now be easily distributed or shared through the Internet. The huge success of YouTube demonstrates the popularity of Internet media.

Before the cloud-computing era, media storage, processing, and dissemination services were provided by different service providers with their proprietary server farms. Now, various service providers have a choice to be users of public clouds. The "pay-as-you-go" model of a public cloud would greatly facilitate small businesses and multimedia fanciers. For small businesses, they pay just for the computing and storage they have used, rather than maintaining a large set of servers only for peak loads. For individuals, cloud utility can provide a potentially unlimited storage space and is more convenient to use than buying hard disks.

In the following, we will present storage and sharing (storage and dissemination), authoring and mashup (storage and processing), adaptation and delivery (processing and dissemination), and media rendering, respectively.

STORAGE AND SHARING

Cloud storage has the advantage of being "always-on" so that users can access their files from any device and can share their files with friends who may access the content at an arbitrary time. It is also an important feature that cloud storage provides a much higher level of reliability than local storage. Cloud storage service can be categorized into consumer- and developer-oriented services. Within the category of consumer-oriented cloud storage services, some cloud providers deploy their own server farm, while some others operate based on user-contributed physical storage. IOmega (www.iomega.com) is a consumer-oriented cloud storage service, which holds the storage service on its own servers and, thus, offers a for-pay only service. AllMyData (www.amdwebhost.com) trades 1 GB of hosted space for 10 GB of donated space from users. The major investment is software or services, such as data encrypting, fragmentation and distribution, and backup. Amazon S3 (https://s3.amazonaws.com/) and Openomy (http://www.openomy.com/) are developer-oriented cloud storage services. Amazon S3 goes with the typical cloud provisioning "pay only



[FIG7] A typical media life cycle.

for what you use." There is no minimum fee and startup cost. They charge for both storage and bandwidth per gigabyte. In Openomy, the files are organized purely by tags. A publicly callable application programming interface (API) and strong focus on tags rather than the classic file system hierarchy is the differentiating feature of Openomy. The general-purpose cloud providers, such as Microsoft Azure (http://www.microsoft.com/windowsazure/), also allow developers to build storage services on top of them. For example, NeoGeo Company builds its digital asset management software neoMediaCenter.NET based on Microsoft Azure.

Sharing is an integral part of cloud service. The request of easy sharing is the main reason the multimedia contents occupy a large portion of cloud storage space. Conventionally, multimedia sharing happens only when the person who shares the contents and the person who is shared with are both online and have a high-data-rate connection. Cloud computing is now turning this synchronous process into an asynchronous one and is making one-to-many sharing more efficient. The person who shares simply uploads the contents to the cloud storage at his or her convenience and then sends a hyperlink to the persons being shared with. The latter can then access the contents whenever they like, since the cloud is always on. Sharing through a cloud also increases media QoS because cloud–client connections almost always provide a higher bandwidth and shorter delay than client–client connections, not to mention the firewall and network address translation traversal problems commonly encountered in client–client communications. The complexities of cloud-based sharing mainly reside in naming, addressing, and access control.

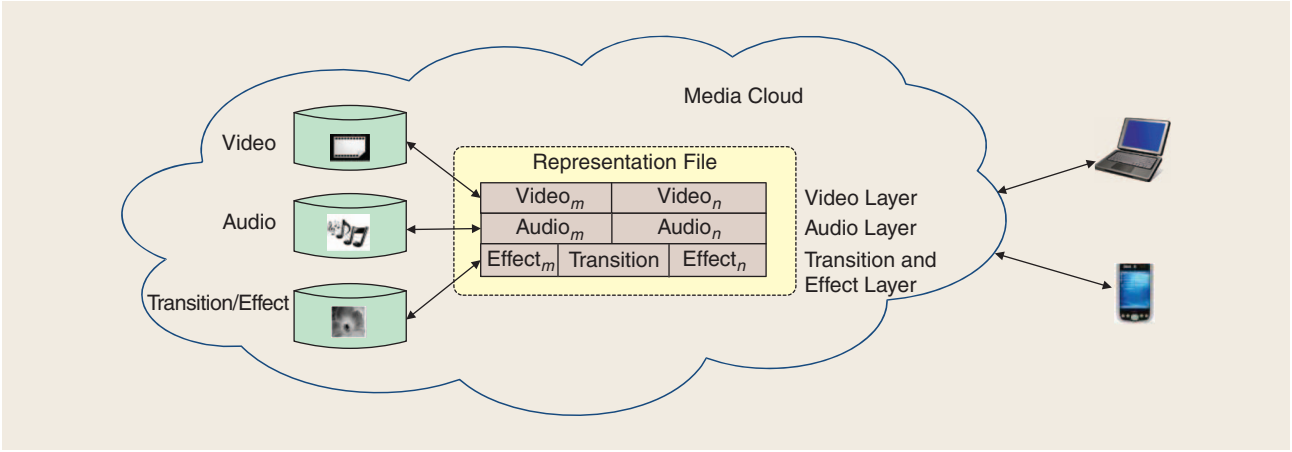
THE ABILITY TO PERFORM ONLINE MEDIA COMPUTATION IS A MAJOR DIFFERENTIATING CHARACTERISTIC OF MEDIA CLOUD FROM TRADITIONAL CDNS.

Instantaneous music and video sharing can be achieved via streaming. Compared to the conventional streaming services operating through proprietary server farms of streaming service providers, cloud-based streaming can potentially achieve much a lower latency and provide much a higher bandwidth. This is because cloud providers own a large number of servers deployed in wide geographical areas. For example, streamload targets to those who are interested in rich media streaming and hosting. They offer unlimited storage for paying users, and the charging plan is based on the downloading bandwidth. In this article, we propose a cloud-based streaming approach. In the architecture of cloud-based streaming, compared with the architecture of video streaming in [14], the media cloud/MEC has a distributed media storage module for storing the compressed video and audio streams for millions of users and a QoS adaptation module for different types of devices, such as mobile phones, PCs, and TVs. While our media cloud/MEC-based streaming architecture presents a possible solution to cloud streaming QoS provisioning, there are still many research issues to be tackled. Considering the cloud network properties, a new cloud transport protocol may need to be developed. In addition, the high bandwidth demand of multimedia contents calls for dynamic load balancing algorithms for cloud-based streaming.

AUTHORING AND MASHUP

Multimedia authoring is the process of editing segments of multimedia contents, while mashup deals with combining multiple segments from different multimedia sources. To date, authoring and mashup tools are roughly classified into two categories: one is offline tools, such as Adobe Premiere and Windows Movie Maker, and the other is online services, such as Jaycut (http://www.jaycut.com). The former provides more editing functions, but the client usually needs editing software maintenance. The latter provides fewer functions, but the client need not bother about its software maintenance.

Authoring and mashup are generally time consuming and multimedia contents occupy large amount of storages. A cloud can make online authoring and mashup very effective, providing more functions to clients, since it has powerful computation and storage resources that are widely distributed geographically. Moreover, cloud-based multimedia authoring and mashup can avoid preinstallation of editing software in clients. In this article, we present a cloud-based online multimedia authoring and mashup framework. In this framework, users will conduct editing and mashup in the media cloud. One of the key challenges in cloud-based authoring and mashup is the computing and communication costs in processing multiple segments from single source or multiple sources. To address this challenge, inspired by the idea of [15], we present an extensible markup language (XML)-based representation file format for cloud-based media authoring and mashup. As illustrated in Figure 8, this is not a multimedia data stream but a description file, indicating the organization of different multimedia contents. The file can be logically regarded as a multilayer container. The layers can be entity layers, such as video, audio, graphic, and transition and effect layers. Each segment of a layer is represented as a link to the original one, which maintains associated data in the case of being deleted or moved, as well as some more descriptions. The transmission and effects are either a link with parameters or a description considering personalized demands. Since a media cloud holds large original multimedia contents and frequently used transmission and effect templates [16], it will be beneficial to use the link-based presentation file. Thus, the process of authoring or mashup is to edit the presentation file, by which the computing load on the cloud side will be significantly reduced. In our approach, we will select an MEC to serve authoring or mashup service to all heterogeneous clients including mobile phone users. By leveraging edge servers' assistance in the MEC with proxy to mobile phone, it allows mobile editing to achieve good QoE. Future research on cloud-based multimedia authoring and mashup needs to



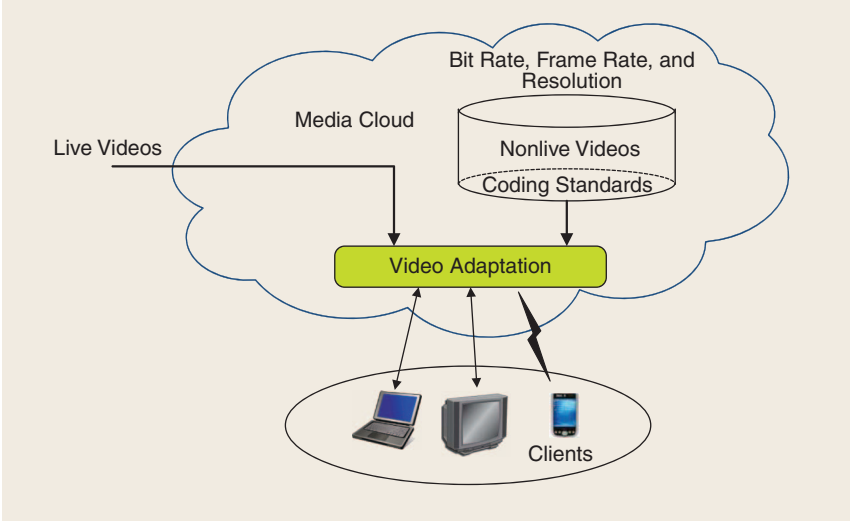
[FIG8] **Cloud-based multimedia authoring and mashup.**

tackle distributed storage and processing in the cloud, online previewing on the client, especially for mobile phones and slate devices.

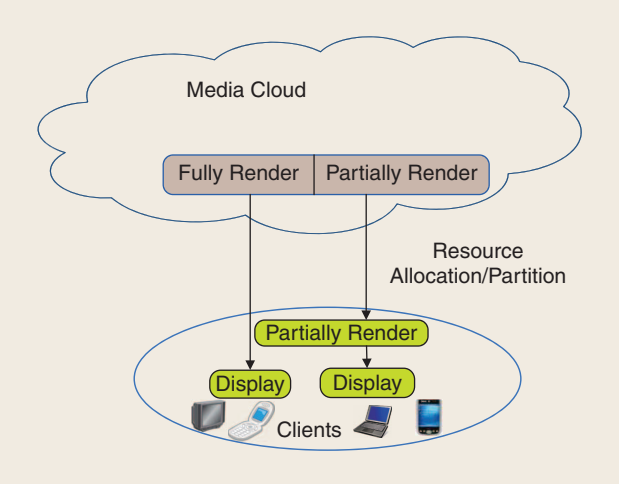
ADAPTATION AND DELIVERY

As there exist various types of terminals, such as PCs, mobile phones, and TVs, and heterogeneous networks, such as Ethernet, WLAN, and cellular networks, how to effectively deliver multimedia contents to heterogeneous devices via one cloud is becoming very important and challenging. Video adaptation [17], [18] plays an important role in multimedia delivery. It transforms input video(s) into an output video in a form that meets the user's needs. In general, video adaptation needs a large amount of computing and is difficult to perform especially when there are a vast number of consumers requesting service simultaneously. Although offline transcoding from one video into multiple versions for different conditions works well sometimes, it needs larger storage. Besides, offline transcoding is unable to serve live video such as Internet protocol television (IPTV) that needs to run in real time.

Because of the strong computing and storage power of the cloud, both offline and online media adaptation to different types of terminals can be conducted in a cloud. Cloudcoder is a good example of a cloud-based video adaptation service that was built on the Microsoft Azure platform [19]. The cloudcoder is integrated into the origin digital central management platform while offloading much of the processing to the cloud. The number of transcoder instances automatically scales to handle the increased or decreased volume. In this article, we present a framework of cloud-based video adaptation for delivery, as illustrated in Figure 9. Video adaption in a media cloud shall take charge of collecting customized parameters, such as screen size, bandwidth, and generating various versions according to their parameters either offline or on the fly. Note that the former needs more storage, while the latter needs dynamic video adaptation upon delivery. The ability to perform online media computation is a major differentiating characteristic of media cloud from traditional CDNs. We would like to point out that the processing capability at media-edge servers makes it possible for service providers to pay more attention to QoE under dynamic network conditions than just to some predefined QoS metrics. In the presented framework, adaptation for single-layer and multilayer video will be performed differently. If the video is of a single layer, video adaptation needs to adjust bit rate, frame rate, and resolution to meet different types of terminals. For scalable video coding, a cloud can generate various forms of videos by truncating its scalable layers according to the clients' network bandwidth. How to perform video adaptation on the fly can be one of the future research topics.



[FIG9] **Cloud-based video adaptation and transcoding.**



[FIG10] **Cloud-based multimedia rendering.**

MEDIA RENDERING

Traditionally, multimedia rendering is conducted at the client side, and the client is usually capable of doing rendering tasks, such as geometry modeling, texture mapping, and so on. In some cases, however, the clients lack the ability required to render the multimedia. For instance, a free viewpoint video [20], which allows users to interactively change the viewpoint in any 3-D position or within a certain range, is difficult to be rendered on a mobile phone. This is because the wireless bandwidth is quite limited, and the mobile phone has limited computing capability, memory size, and battery life. Shu et al. [21] proposed a rendering proxy to perform rendering for the mobile phone. For example, Web browsing is an essential mechanism to access the information on the World Wide Web. However, it is relatively difficult for mobile phones to render Web pages, especially the asynchronous JavaScript and XML (AJAX) Web page. Lehtonen et al. [22] proposed a proxy-based architecture for mobile Web browsing. When the client sends a Web request, the proxy fetches the remote Web page, renders it, and responds with a package containing a page description, which includes a page miniature image, the content, and coordinates of the most important elements on the page. In essence, in both examples, a resource-allocation strategy is needed such that a portion of rendering task can be shifted from the client to server, thus eliminating computing burden of a thin client.

Rendering on mobile phones or computationally constrained devices imposes great challenges owing to a limited battery life and computing power as well as narrow wireless bandwidth. The cloud equipped with GPU can perform rendering due to its strong computing capability. Considering the tradeoff between computing and communication, there are two types of cloud-based rendering. One is to conduct all the rendering in the cloud, and the other is to conduct only computational intensive part of the rendering in the cloud, while the rest will be performed on the client. In this article, we present cloud-based media rendering. As illustrated in Figure 10, the media cloud can do full or partial rendering, generating an intermediate stream for further client rendering, according to the client's rendering capability. More specifically, an MEC with a proxy can serve mobile clients with high QoE since rending (e.g., view interpolation) can be done in an MEC proxy. Research challenges and opportunities include how to efficiently and dynamically allocate the rendering resources between the client and cloud. One of the future research directions is to study how an MEC proxy can assist mobile phones on rendering computation, since the mobile phones have limitations in battery life, memory size, and computation capability.

Multimedia retrieval, such as content-based image retrieval (CBIR), is a good application example of cloud computing as well. In the following, we will present how CBIR can leverage cloud computing. CBIR [23] is used to search digital images in a large database based on image content other than metadata and text annotation. The research topics in CBIR include feature extraction, similarity measurement, and relevance feedback. There are two key challenges in CBIR: how to improve the search quality and how to reduce computation complexity (or computation time). As there exists a semantic gap between the low-level underlying visual features and semantics, it is difficult to get high search quality. Because the Internet image database is becoming larger, searching in such a database is becoming computationally intensive. However, in general, there is a tradeoff between quality and complexity. Usually, a higher quality is achieved at the price of a higher complexity and vice versa. Leveraging the strong computing capability of a media cloud, one can achieve a higher quality with acceptable computation time, resulting in better performance from the client's perspective.

SUMMARY

This article presented the fundamental concept and a framework of multimedia cloud computing. We addressed multimedia cloud computing from multimedia-aware cloud and cloud-aware multimedia perspectives. On the multimedia-aware cloud, we presented how a cloud can provide QoS support, distributed parallel processing, storage, and load balancing for various multimedia applications and services. Specifically, we proposed an MEC-computing architecture that can achieve high cloud QoS support for various multimedia services. On cloud-aware multimedia, we addressed how multimedia services and applications, such as storage and sharing, authoring and mash-up, adaptation and delivery, and rendering and retrieval, can optimally utilize cloud-computing resources. The research directions and problems of multimedia cloud computing were discussed accordingly.

In this article, we presented some thoughts on multimedia cloud computing and our preliminary research in this area. The research in multimedia cloud computing is still in its infancy, and many problems in this area remain open. For example, media cloud QoS addressed in this article still needs more investigations. Some other open research topics on multimedia cloud computing include media cloud transport protocol, media

cloud overlay network, media cloud security, P2P cloud for multimedia services, and so on.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their constructive comments that greatly improved the article. We acknowledge Lijing Qin and Lie Liu from Tsinghua University and Zheng Li from the University of Science and Technology of China (USTC) for their contributions to the section "Multimedia-Aware Cloud" when they were interns at Microsoft Research Asia. We also thank Jun Liao from Tsinghua University and Siyuan Tang from the USTC for their contributions to cloud-based parallel photosynthing case study when they were interns at Microsoft Research Asia. We also acknowledge Prof. Yifeng He from Ryerson University for proofreading the article and Dr. Jingdong Wang from Microsoft Research Asia for proofreading of CBIR.

AUTHORS

Wenwu Zhu (wenwuzhu@microsoft.com) received his Ph.D. degree from Polytechnic Institute of New York University in 1996. He worked at Bell Labs during 1996–1999. He was with Microsoft Research Asia's Internet Media Group and Wireless and Networking Group as research manager from 1999 to 2004. He was the director and chief scientist at Intel Communication Technology Lab, China. He is a senior researcher at the Internet Media Group at Microsoft Research Asia. He has published more than 200 refereed papers and filed 40 patents. He is a Fellow of the IEEE. His research interests include Internet/wireless multimedia and multimedia communications and networking.

Chong Luo (cluo@microsoft.com) received her B.S. degree from Fudan University in Shanghai, China, in 2000 and her M.S. degree from the National University of Singapore in 2002. She is currently pursuing her Ph.D. degree in the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China. She has been with Microsoft Research Asia since 2003, where she is a researcher in the Internet Media group. She is a Member of the IEEE. Her research interests include multimedia communications, wireless sensor networks, and media cloud computing.

Jianfeng Wang (wjf2006@mail.ustc.edu.cn) received his B.Eng. degree from the Department of Electronic Engineering and Information Science in the University of Science and Technology of China in 2010. He was an intern at Microsoft Research Asia from February to August in 2010. Currently, he is a master's student in MOE-Microsoft Key Laboratory of Multimedia Computing and Communication in USTC. His research interests include signal processing, media cloud, and multimedia information retrieval.

Shipeng Li (spli@microsoft.com) received his Ph.D. degree from Lehigh University in 1996 and his M.S. and B.S. degrees from the University of Science and Technology of China in 1991 and 1988, respectively. He has been with Microsoft Research Asia since May 1999, where he was a principal researcher and research area manager. From October 1996 to May 1999, he was with Sarnoff Corporation as a member of technical staff. He has authored and coauthored five books or book chapters and more than 200 journal and conference papers as well as holds more than 90 patents. His research interests include multimedia processing, retrieval, coding, streaming, and mobility.

REFERENCES

[1] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia. (2009, Feb. 10). Above the clouds: A Berkeley view of cloud computing. EECS Dept., Univ. California, Berkeley, No. UCB/EECS-2009-28 [Online]. Available: http://radlab.cs.berkeley.edu/

[2] R. Buyya, C. S. Yeo, and S. Venugopal, "Market-oriented cloud computing: Vision, hype, and reality for delivering it services as computing utilities," in *Proc. 10th IEEE Int. Conf. High Performance Computing and Communications,* 2008, pp. 5–13.

[3] K. Kilkki, "Quality of experience in communications ecosystem," *J. Universal Comput. Sci.*, vol. 14, no. 5, pp. 615–624, 2008.

[4] ABI Research. (2009, July). Mobile cloud computing [Online]. Available: http://www.abiresearch.com/research/1003385-Mobile+Cloud+Computing

[5] Q. Zhang, Z. Ji, W. Zhu, and Y.-Q. Zhang, "Power-minimized bit allocation for video communication over wireless channels," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 6, pp. 398–410, June 2002.

[6] P. Maymounkov and D. Mazieres, "Kademlia: A peer-to-peer information system based on the XOR metric," in *Proc. IPTPS02*, Cambridge, Mar. 2002, pp. 53–65.

[7] B. Aljaber, T. Jacobs, K. Nadiminti, and R. Buyya, "Multimedia on global grids: A case study in distributed ray tracing," *Malays. J. Comput. Sci.*, vol. 20, no. 1, pp. 1–11, June 2007.

[8] J. Nieh and S. J. Yang, "Measuring the multimedia performance of server-based computing," in *Proc. 10th Int. Workshop on Network and Operating System Support for Digital Audio and Video*, 2000, pp. 55–64.

[9] I. Trajkovska, J. Salvachúa, and A. M. Velasco, "A novel P2P and cloud computing hybrid architecture for multimedia streaming with QoS cost functions," in *Proc. ACM Multimedia*, 2010, pp. 1227–1230.

[10] G. Cybenko, "Dynamic load balancing for distributed memory multiprocessors," *J. Parallel Distrib. Comput.*, vol. 7, no. 2, pp. 279–301, 1989.

[11] N. Snavely, S. M. Seitz, and R. Szeliski, "Photo tourism: Exploring photo collections in 3D," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 835–846, 2006.

[12] N. Snavely, S. M. Seitz, and R. Szeliski, "Modeling the world from Internet photo collections," *Int. J. Comput. Vision*, vol. 80, no. 2, pp. 189–210, Nov. 2008.

[13] The National Science Foundation. (2008, Sept. 30). Cyber-physical systems. Program Announcements and Information. NSF, Arlington, Virginia [Online]. Available: http://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf08611

[14] D. Wu, Y. T. Hou, W. Zhu, Y.-Q. Zhang, and J. M. Peha, "Streaming video over the Internet: Approaches and directions," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 3, pp. 282–300, Mar. 2001.

[15] X.-S. Hua and S. Li., "Personal media sharing and authoring on the web," in *Proc. ACM Int. Conf. Multimedia*, Nov. 2005, pp. 375–378.

[16] X.-S. Hua and S. Li, "Interactive video authoring and sharing based on two-layer templates," in *Proc. 1st ACM Int. Workshop on Human-Centered MM 2006 (HCM'06)*, pp. 65–74.

[17] S. F. Chang and A. Vetro, "Video adaptation: Concepts, technologies, and open issues," *Proc. IEEE*, vol. 93, no. 1, pp. 148–158, 2005.

[18] J. Xin, C-W. Lin, and M-T. Sun, "Digital video transcoding," *Proc. IEEE*, vol. 93, no. 1, pp. 84–94, Jan. 2005.

[19] Origin Digital. (2009, Nov. 17). Video services provider to reduce transcoding costs up to half [Online]. Available: http://www.microsoft.com/casestudies/Case_Study_Detail.aspx?CaseStudyID=4000005952

[20] A. Kubota, A. Smolic, M. Magnor, M. Tanimoto, T. Chen, and C. Zhang, "Multiview imaging and 3DTV," *IEEE Signal Processing Mag.,* vol. 24, no. 6, pp. 10–21, Nov. 2007.

[21] S. Shi, W. J. Jeon, K. Nahrstedt, and R. H. Campbel, "Real-time remote rendering of 3D video for mobile devices," in *Proc. ACM Multimedia*, 2009, pp. 391–400.

[22] T. Lehtonen, S. Benamar, V. Laamanen, I. Luoma, O. Ruotsalainen, J. Salonen, and T. Mikkonen, "Towards user-friendly mobile browsing," in *Proc. 2nd Int. Workshop on Advanced Architectures and Algorithms for Internet Delivery and Applications* (ACM Int. Conf. Proc. Series, vol. 198), 2006, Article 6.

[23] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Machine Intell.,* vol. 22, no. 12, pp. 1349–1380, Dec. 2000.

[SP]