Adaptive Mobile Cloud Computing to Enable Rich Mobile Multimedia Applications

Shaoxuan Wang and Sujit Dey, *Senior Member, IEEE*

***Abstract*—With worldwide shipments of smartphones (487.7 million) exceeding PCs (414.6 million including tablets) in 2011 [1], and in the US alone, more users predicted to access the Internet from mobile devices than from PCs by 2015 [2], clearly there is a desire to be able to use mobile devices and networks like we use PCs and wireline networks today. However, in spite of advances in the capabilities of mobile devices, a gap will continue to exist, and may even widen, with the requirements of rich multimedia applications. Mobile cloud computing can help bridge this gap, providing mobile applications the capabilities of cloud servers and storage together with the benefits of mobile devices and mobile connectivity, possibly enabling a new generation of truly ubiquitous multimedia applications on mobile devices: Cloud Mobile Media (CMM) applications.**

In this paper, we look at early trends, and opportunities and benefits for new CMM applications and services. We analyze the challenges imposed by mobile cloud computing that need to be addressed to make CMM applications viable, including response time, user experience, cloud computing cost, mobile network bandwidth, and scalability to large number of CMM users, besides other important cloud computing issues like energy consumption, privacy, and security. We illustrate the challenges using Cloud Mobile Gaming (CMG), an approach that enables rich multiplayer Internet games on mobile devices, where compute intensive tasks like graphic rendering are executed on cloud servers in response to gaming commands on a mobile device, and the resulting video has to be streamed back to the mobile device in near real time, making it one of the most challenging CMM applications. Subsequently, we focus in this paper on developing adaptive mobile cloud computing techniques to address the CMG challenges. Specifically, we propose a rendering adaptation technique, which can dynamically vary the richness and complexity of graphic rendering depending on the network and cloud computing constraints, thereby impacting both the bit rate of the rendered video that needs to be streamed back from the cloud server to the mobile device, and the computation load on the CMG servers. Experiments conducted on a cellular network demonstrate that our proposed technique can significantly improve user experience, and ensure scalability of the CMG approach in terms of both network bandwidth and server computational need.

***Index Terms*—Mobile Computing, Cloud Computing, Cloud Gaming, Multimedia Applications.**

I. Introduction

OVER the last few years, there has been an increased number of applications that have "migrated to the cloud", and new cloud-based applications that have become popular. Most of the early adopters of cloud have been enterprise applications and IT departments; according to Juniper Research, revenue from mobile enterprise cloud-based applications and services is expected to rise from nearly $2.6 billion in 2011 to $39 billion in 2016 [3].

Similar motivations that have driven mobile enterprise cloud services are also driving adoption of mobile consumer cloud services: the ability to access media from anywhere: any device, platform, and network. According to Juniper Research, revenues from consumer cloud mobility services, initially driven by cloud based music and video storage and download services like the ones recently launched by Amazon's Cloud Drive and Apple's iCloud, are expected to reach $6.5 billion per year by 2016 [3].

Besides such storage and download services, a big boost to mobile consumer cloud services will come from a major shift in the mobile applications market, from primarily native applications to ones based on mobile cloud computing: utilizing the computing and storage resources available in the cloud, thereby enabling the use of cutting edge multimedia technologies that are much more computing and storage intensive than what mobile devices can offer, and thus enabling much richer media experiences than what current native applications can offer. While according to MarketsAndMarkets.com, the global mobile applications market is expected to be worth $25.0 billion by 2015 [4], use of mobile cloud computing will enable more powerful applications, and hence more significant growth. Also, initiatives such as GSMA's OneAPI [5], which will allow access to network information, regardless of operator, via Web applications rather than device clients, will further motivate and ease development of cloud-based mobile applications. And finally, mobile cloud computing based applications can simultaneously avail of not only cloud resources, but also the unique resources of mobile devices, like user location and device sensors, that will make such applications more powerful than either server or PC-based applications, or current native mobile applications.

In this paper, we focus on Cloud Mobile Media (CMM) applications and services, which will enable mobile users to not only access rich media from any mobile device and platform, but even more importantly, which will enable mobile users to engage in new, rich media experiences, through the use of mobile cloud computing, that are not possible otherwise from their mobile devices. CMM will also enable service providers and network operators to offer services much more efficiently, with lower cost and better user experience. As more consumers adopt smartphones and tablets as one of their primary media experience platforms, CMM has the potential of significantly boosting the revenue of cloud Software-as-a-Service (SaaS) providers. Some of the media rich CMM services will require new and

Manuscript received March 06, 2012; revised December 10, 2012; accepted December 19, 2012. Date of publication January 16, 2013; date of current version May 13, 2013. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Xiaoqing Zhu.

The authors are with the Mobile Systems Design Lab, Electrical and Computer Engineering Department, University of California, San Diego, La Jolla, CA 92093 USA (e-mail: shaoxuan@ece.ucsd.edu; dey@ece.ucsd.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.2013.2240674

richer platform and infrastructure capabilities as explained in the next sections, thereby providing a new set of revenue opportunities for cloud platform and infrastructure providers. And finally, CMM offers new opportunities for mobile network operators to close the growing gap between growth in data usage and data revenue by offering innovative CMM services and experiences, outside of conventional application stores where their participation has not been strong so far.

In the rest of the paper, we discuss in Section II different types of possible CMM applications including their advantages, and also challenges that will be faced in making them successful. In Section III, we elaborate on a few major challenges of CMM applications: user experience including response time, cloud computing cost, mobile network bandwidth, and ability to scale to a large number of CMM users. We illustrate the challenges using Cloud Mobile Gaming (CMG), one of the most compute and mobile bandwidth intensive CMM applications. Subsequently in Section IV, we propose an adaptive mobile cloud computing technique to address the challenges associated with CMG. We show that it is possible to dynamically vary the richness and complexity of graphic rendering in the cloud servers, depending on the mobile network and cloud computing constraints, thereby impacting both the bit rate of the rendered video that needs to be transmitted from the cloud server to the mobile device, and the computation load on the CMG servers [6]. We propose an adaptation process whose offline steps determine in advance complexity and rendering models, which are used by the online step to adapt rendering settings in real time to meet changing communication and computation constraints. We present experimental results demonstrating the ability of the proposed approach to dynamically address changing network conditions to ensure user experience, as well as ensure scalability by significantly reducing cloud computing costs and network bandwidth needed. We conclude in Section V with suggestions for future research for mobile cloud computing to efficiently enable future Cloud Mobile Media applications.

II. Cloud Mobile Media Architecture and Applications

Utilizing available cloud computing and storage resources, we expect a heterogeneous set of Cloud Mobile Media services and applications to emerge, with different types of consumer experiences and advantages enabled. In this section, we first describe the typical end-to-end control and data flow architecture of CMM applications. Next, we categorize the existing and expected CMM applications, and analyze for each category the cloud infrastructure and platform needs, advantages and user experiences enabled, and challenges to make the applications successful.

Fig. 1 shows the overall architecture, including end-to-end flow of control and data between the mobile devices and the Internet cloud servers, for a typical CMM application. Though a CMM application may utilize the native resources of the mobile device, like GPS and sensors, it primarily relies on cloud computing Infrastructure as a Service (IaaS) and Platform as a Service (PaaS) resources, like elastic computing resources and storage resources, located in Internet public, private, or federated (hybrid) clouds. A typical CMM application has a small footprint client on the mobile device, which provides the appropriate user interfaces (gesture, touchscreen, voice, text based)

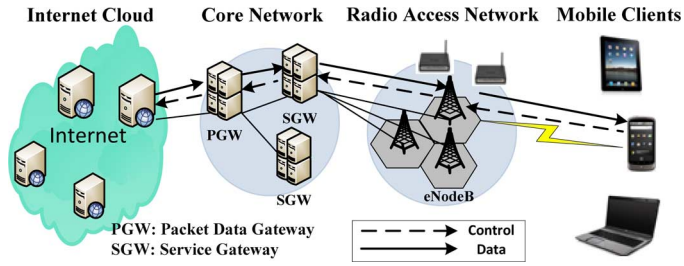


Fig. 1. Cloud Mobile Media architecture, showing control and data flows.

to enable the user to interact with the application. The resulting control commands are transmitted uplink through cellular Radio Access Networks (RAN) or WiFi Access Points to appropriate gateways located in operator Core Networks (CN), and finally to the Internet cloud. Subsequently, the multimedia data produced by the cloud, either as a result of processing using the cloud computing resources, and/or retrieval from cloud storage resources, is transmitted downlink through the CN and RAN back to the mobile device. The CMM client then decodes and displays the results on the mobile device display. From the above description, and as shown in Fig. 1, a typical CMM application will be highly interactive, with some types of applications needing near real-time response times. Note that for certain types of CMM applications, the control and data flow may deviate from that shown in Fig. 1. For example, for CMM applications like Cloud based Media Analytics described later, the application may not always be initiated by a mobile CMM client (like in Fig. 1), and may collect data from both the client and the cloud to provide analytics to other CMM applications.

Table I summarizes the different categories of mobile multimedia applications that already are, or can potentially be, driven by the use of the cloud, including storage, download and synchronization applications, audio and video streaming applications, interactive applications like multi-way video conferencing, interactive advertisements, and mobile remote desktop, rich rendering based applications like mobile multi-user gaming and augmented reality, and cloud based media analytics that will provide better understanding of user preferences and experiences, and drive personalized mobile services. For each category of CMM applications, we list the IaaS and PaaS features that will be needed, including some which are available today, and some that need to be developed. We also list the advantages of each category of CMM applications, including what multimedia experience can be enabled that cannot be supported currently, and the challenges that need to be addressed to make the application category successful.

As discussed before, *Mobile Cloud Storage* is the most commonly used category of CMM application/service today, with offerings from Amazon, Apple, Dropbox, Funambol, and Google, among others. These services provide diverse capabilities, including storing documents, photos, music and video in the cloud, accessing media from any device anywhere irrespective of the source of the media and/or the device/platform used to generate the media, and synchronizing data/media across multiple devices a typical user owns. According to ARCchart [7], the demand for mobile cloud storage will grow exponentially as the volume of user generated content using

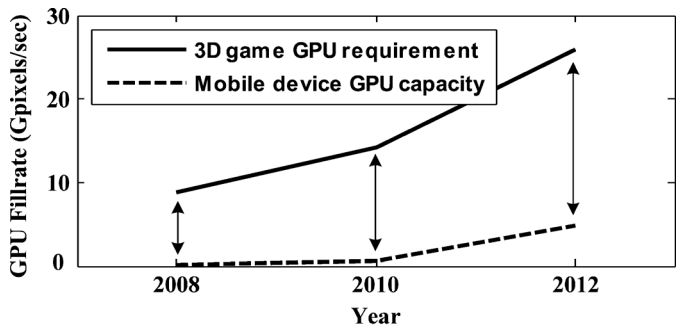
TABLE I
ANALYSIS OF DIFFERENT CATEGORIES OF POTENTIAL CLOUD MOBILE MEDIA APPLICATIONS, INCLUDING CLOUD CAPABILITIES NEEDED, ADVANTAGES DERIVED BY HAVING THE APPLICATIONS BASED ON CLOUD, AND CHALLENGES TO MAKE THE APPLICATIONS SUCCESSFUL

	Mobile Cloud Storage	**Audio/Video Streaming**	**Interactive Services**	**Cloud based Rendering**	**Media Analytics**
Sample applications /services enabled on mobile devices	Storing and accessing Photos, Music, Files	Streaming audio, video; Cloud DVR	Video Chat; Remote Desktop; Interactive advertisements	Mobile Gaming; Augmented Reality; Telemedicine	Differentiated Services/Billing; Personalized services
IaaS, PaaS features needed	Cloud Storage with high availability and integrity	Cloud Transcoding, Transrating, Caching	Cloud Transcoding, Transrating	Multi-core GPUs; Efficient cloud rendering;	Cloud media usage/QoE probes; Media classification engines
Advantages	Ubiquitous access from any device; Synchronizing between devices	Low capex; High scalability with demand	Easier support for multiple devices/platforms	Enables highest quality rendering; Multi-player, multi-platform	Unified analytics for media usage across devices and networks
Challenges	Ensuring content security, privacy; Additional wireless traffic	Cloud service cost; Cloud energy, cooling costs	Response time; Video quality; Cloud service cost; Cloud energy, cooling costs	Response time; User experience; Cloud service cost; Additional wireless traffic; Cloud energy, cooling costs	Data protection; privacy

mobile devices grow, and camera resolutions of mobile devices continue to increase. To enable mass adoption of such services, the PaaS providers will need to ensure high availability and integrity of data, and the SaaS provider will need to ensure content security and user privacy. However, a major impact of the mass adoption of this category of CMM service will be significant increase in mobile data traffic, and potentially larger data bills for mobile subscribers, issues that will need to be addressed for these services to be used ubiquitously, including using cellular networks.

Audio and video streaming based services can benefit by utilizing cloud computing resources to perform compute intensive tasks of encoding, and transcoding and transrating needed to adjust to different devices and networks. For on-demand video, computing costs can be reduced by caching popular videos at different resolutions and bit rates. Besides lower initial capital expenses, the advantage of cloud based audio/video services is the use of elasticity in cloud computing resources to more cost-effectively handle variable peak demands. However, to support the expected increase in demand for mobile video [8], the cloud will need to provide server architectures and tools that can enable massively concurrent transcoding/transrating implementations, at efficiencies and cost points that can be enabled today with custom hardware solutions. The concern is particularly critical for the use of public clouds to offer video services, as current public cloud capabilities and price structures, including network bandwidth costs, may make operating expenses of video services very high.

The *Interactive Services* category is expected to be a rapidly growing segment of mobile multimedia applications, including mobile video conferencing, mobile remote desktop, and interactive mobile advertisements. As shown in Table I, besides the typical consideration of lower capital expense, use of the cloud will lead to easier support for multiple devices and operating platforms. One of the biggest challenges of such applications will be the potentially high latency and packet loss of the wireless network that may be experienced by the video stream, both from and to the mobile device, thereby potentially affecting the very low response time requirements of such interactive applications, and also the video quality, depending on what transport protocol is used. As in audio/video streaming, the operating cost can also be a concern, till cloud capabilities and pricing structures are improved as suggested in the previous paragraph.



	2008	2010	2012
3D game GPU requirements	Call of Duty 4	Call of Duty 7	Battlefield 3
	8.8 GPixels/sec	14.16 GPixels/sec	25.9 GPixels/sec
Smartphone GPU capability	Iphone 3G	Iphone 4	Iphone 5
	0.135 pixels/sec	0.5 Gpixels/sec	4.875 Gpixels/sec

Fig. 2. Growing gap between (recommended) GPU requirement of rendering-based applications and GPU capability of mobile devices.

A promising category of CMM applications that has the potential of significantly enhancing the media experience of mobile users is *Cloud based Rendering*. Despite the progress in the capabilities of mobile devices, there is a widening gap with the growing requirements of the latest 3D and multi-view rendering techniques and that can be supported by today's and near future mobile devices, including tablets. Fig. 2 shows this widening gap from 2008 to 2012, in terms of the recommended GPU requirements [9] of the most demanding games in those years, Call of Duty 4 in 2008, Call of Duty 7 in 2010, and Battlefield 3 in 2012, and the GPU capabilities [10] of the popular smartphones in those years, iPhone 3G, iPhone 4, and iPhone 5 respectively. Cloud based Rendering can bridge this gap by allowing rendering to be executed in the cloud, instead of on the mobile device, thereby potentially enabling mobile users to play the same

rich Internet games available to high-end PC users [11], or participate in rich augmented reality and/or multi-view immersive experiences that are being developed primarily with PC users in mind. Moreover, analogous to transcoding, we expect a new capability to emerge that we term trans-rendering—the ability to automatically adjust rendering to different device and platform capabilities, thus relieving game and augmented reality application developers from the expensive cycle of developing device and platform specific mobile versions. The ability to enable such rich experiences on all mobile devices and platforms, coupled with the inherent advantages of ubiquity and location information associated with the use of mobile devices, will have the potential to drive a new generation of cloud based mobile media applications.

As summarized in Table I, enabling cloud mobile rendering applications will require additional capabilities from cloud infrastructure and platforms, as well as addressing some significant challenges. Clouds will need to include architectural provisions, like massively multi-core GPUs, and software support for developing highly concurrent cloud based rendering applications like cloud gaming engines, as each cloud rendering session will need to be supported by a separate cloud rendering instance. Moreover, in applications like Cloud Mobile Gaming (CMG), in response to gaming commands from the mobile device, not only does rendering need to be performed in the cloud, but the rendered video will need to be encoded and delivered over wireless networks to the mobile device, all in near real time, as user experience for such applications is highly dominated by fast response time [11]. Also, the need to transmit the rendered video for each rendering/gaming session can mean significant additional bandwidth cost leading to high operating expenses for the service provider, significant additional traffic on the wireless networks possibly leading to overloading of the networks, loss of video quality, and additional cost for mobile subscribers due to tiered data plans.

Because of the high levels of multimedia computation and transmission involved, the above three categories of CMM applications can lead to a significant increase in cloud energy and cooling costs, which will need to be addressed through new developments in green computing and sustainability research.

Cloud Media Analytics: While data analytics is playing an ever important role in the related Internet, Media, and Telecommunications market segments, CMM applications will offer significant opportunities for unified media analytics that can be utilized by cloud platform providers, mobile network operators, and CMM service providers to offer differentiated services with deterministic quality of experience (QoE) levels, and more personalized services depending on user interests. Cloud media analytics can expand on current data analytics and reporting capabilities to include the rich multimedia content that will be consumed by CMM users, and benefit from a more comprehensive view of media consumption across different cloud platforms and mobile networks, and from different types of devices. Besides, cloud media analytics services will be able to collect and utilize information from the devices, including location information. Example of cloud media analytics may include ability to understand user preferences and intent from the cloud media activities of users. Similarly, cloud media analytics may be able to provide new cloud access and performance metrics, including quality of experience, which can be used to improve the cloud designs and offerings. As noted in Table I, cloud media analytics will need new PaaS capabilities, like cloud media usage probes, cloud quality of experience measurement techniques, and media classification engines which can automatically classify the category of media activities, like videos watched by different users. However, cloud media analytics will need to address significant concerns that will naturally arise regarding the ability to maintain user privacy, and protection and sharing of media analytics data collected.

In summary, we envision media-rich cloud based mobile applications to emerge, besides many current mobile media services migrating to the cloud. These developments can lead to new and efficient mobile media experiences, and thereby revenue growth opportunities. However, as pointed in this section, several technical and eco-system challenges will need to be addressed, including ensuring high availability, data integrity and user privacy, lowering energy consumption and cooling costs, ensuring response time and user experience over wireless networks, and reducing cloud service cost associated with high computing and bandwidth needed by CMM applications, and thereby ensuring service scalability. In the next section, we take a deeper look at the last two challenges, as they are crucial to understand and address to make CMM applications attractive to end users, as well as economically viable for service providers, thus leading to mass adoption and monetization.

III. Major Challenges: User Experience, Cost, and Scalability

CMM applications, unlike other cloud applications, will need to overcome the challenges of the wireless network, including limited bandwidth and impact on user experience. Moreover, many of the CMM applications will be very compute and network bandwidth intensive, and hence will have major implications on cloud and network costs incurred per user, and the ability to scale to millions of users as mobile cloud computing becomes popular. In this section, we discuss in more details the above two challenges, while providing some empirical data we collected.

A. Impact of Wireless Network Factors on User Experience

To study the impact of wireless networks on the quality of CMM applications, we conducted experiments with a Cloud Mobile Gaming (CMG) application we have developed, which as described in Section II is a highly interactive cloud based rendering application: gaming commands are transmitted uplink from the mobile device to the cloud servers, and the rendered video needs to be streamed downlink from the server to the mobile client in near real time.

Since this application is highly sensitive to response time, we measured uplink delay, downlink delay, and round-trip response time, as shown in Fig. 3. The experiments were conducted under different network conditions – Fig. 3 shows data samples collected under three different conditions: when the network was not loaded (data collected at mid night), when the network was loaded (data collected at 5 pm), and when the network was loaded and the signal conditions were not strong (data collected at 6 pm, and inside a building).

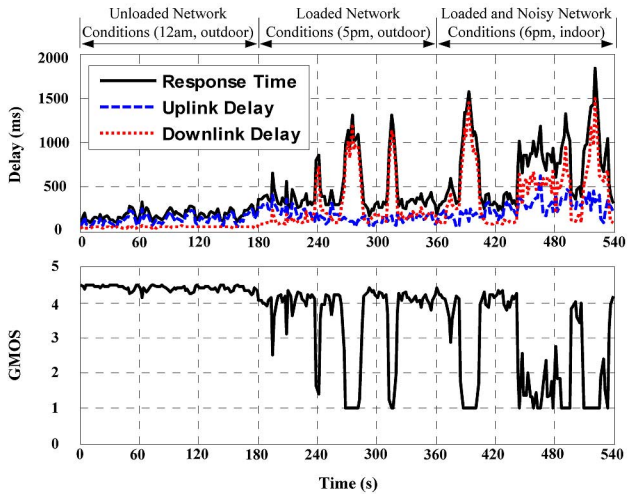


Fig. 3. Cloud Mobile Gaming: Delays, response time, and user experience.

In our previous work [11], [12], we had developed Game Mean Opinion Score (GMOS) as a new metric for Mobile Gaming User Experience (MGUE) in the Cloud Mobile Gaming approach, and developed techniques to quantitatively measure GMOS in a real-time gaming session. The value of GMOS ranges from 1 to 5, with the following opinion scales: 5 (excellent), 4 (good), 3 (fair), 2 (poor), and 1 (bad). MGUE is recognized as acceptable only if GMOS is above 3. In Fig. 3, we also report the user experience obtained as measured by GMOS.

From Fig. 3, we can observe that when the network is not loaded and the signal strength is strong, CMG application can achieve low response time. However when the 3G network is loaded, or when the user is in noisy network condition with poor signal strength, there are significant increases in uplink, downlink, and round-trip response time, leading to significant adverse impact on the quality of gaming experience, as measured by the GMOS scores reported.

The above experiments indicate that for CMM applications to be successful, serious attention has to be given to (a) address challenges imposed by mobile networks like latency and response time, and (b) ensure good user experience.

B. Cloud Service Cost and Scalability

One of the primary advantages of using cloud services is to eliminate capital expenses, and depend on the elasticity of cloud computing, and the cloud utility or on-demand pricing model, to scale to varying capacity needs. However, as we show in this section, there will be challenges faced by computing and bandwidth intensive CMM applications like cloud based mobile gaming, in terms of prohibitively high operating expenses when using on-demand cloud pricing models.

Table II shows the cloud pricing structures of three popular cloud service providers (whose names have been withheld to maintain anonymity), including computing price, storage price, and network price. It also shows the cost per hour of a VGA resolution cloud mobile gaming session of the popular Multiplayer Online Role-Playing Game (MMORPG), World of Warcraft, (WoW), assuming each session needs 1 GB cloud storage space, 700 kbps cloud network bandwidth, and up to 8000 MIPS

TABLE II
CLOUD SERVICE COST FOR CLOUD MOBILE GAMING USING DIFFERENT CLOUD PLATFORMS (NOV. 2012)

Cloud Provider	Computing Price ($/MIPS)	Storage Price ($/GB/sec)	Network Price ($/kb)	Cost for a WoW Session ($/hour)
Provider 1	2.58e-9	4.24e-8	1.50e-8	0.113
Provider 2	3.70e-9	3.45e-8	1.37e-8	0.141
Provider 3	3.70e-9	5.01e-8	1.50e-8	0.145

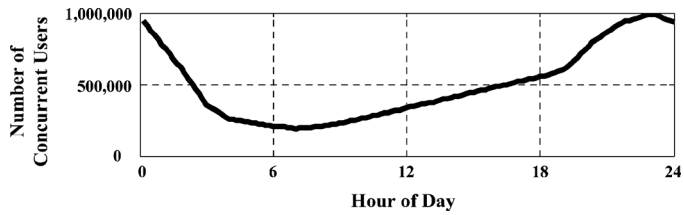


Fig. 4. Daily concurrent user pattern for game WoW.

TABLE III
CLOUD MOBILE GAMING OPERATING COST FOR GAME WOW

Video Resolution	480p	720p	1080p
Hourly cost per session ($)	0.113	0.190	0.272
Daily operating cost for Figure 4 usage (million $)	1.343	2.020	3.141

cloud computing capacity. Assuming average playing time of 23 hours/week [13], from Table II the monthly operating expense for a cloud mobile gaming provider using public cloud platforms will be at least $10/month per WoW player. Considering typical subscription prices (for example, current price of WoW prepaid card is $15/month), this level of operating expense will be too high, even to support VGA resolution.

Fig. 4 shows our estimate of concurrent WoW online gamers according to hours of day in China, one of the large online gaming markets. Our estimate is based on a study showing daily usage patterns for WoW gamers [14], and extrapolating with the number of WoW peak concurrent users in China, that has been steady for several years at 1 million [15]. Fig. 4 highlights the advantage of using the rental model offered by cloud computing, instead of owning servers: peak concurrent session demands can be very high, and provisioning with service provider's own servers can lead to significant capital costs. However, since each gaming session has to execute its own rendering and encoding session in the cloud, supporting the significant number of concurrent sessions with cloud servers can be very challenging. Table III shows the cost to support WoW cloud mobile gaming per session per hour for different resolutions, as well as the daily operating cost for the concurrent user profile shown in Fig. 4, assuming each similar resolution gaming session has similar computing, storage, and bandwidth needs. From Table III we see that the daily operating expense can be very high, up to $3.14 M for WoW, which puts a question mark on the scalability of cloud mobile gaming, as the level of concurrency that need to be supported may be even much higher to support all the other popular games. Similar analysis can be performed for other computation and bandwidth intensive CMM applications listed in Table I.

Clearly, techniques need to be developed to address the cloud cost and scalability challenges faced by CMM services in using public clouds.

C. Mobile Network Cost and Scalability

Besides the potentially high cloud operating expenses, and cloud scalability concern, CMM applications can have very high demand on wireless network bandwidth, having implications on the capacity of the mobile networks, in particular during peak demand periods, potentially negatively impacting network latency, packet loss, and response time, with the consequent negative impact on user experience. Moreover, the high wireless bandwidth requirement may prohibitively increase the wireless data bills of mobile users, making CMM applications impractical.

Consider a cloud mobile gaming application. For a game like WoW, the video bit rate needed (from cloud to device) for high quality experience on a smartphone (VGA resolution) is approximately 700 Kbps, and on a tablet like iPAD (1024×768 resolution) is 1.5-2 Mbps. Assuming average playing time of 23 hours per week [13], the monthly consumption will be 29 GB for smartphones and 62 GB for tablets. If a subscriber uses cellular data access (3 G/4 G) only 25% of time, the cellular data consumption will be 7.25 GB/month for smartphones and 15.5 GB/month for tablets. Even with the best current mobile data plans in the US, the data cost of a mobile user will be \$72/month for smartphone usage and \$155/month for tablet usage. Clearly, such monthly data usage costs will not be acceptable for most users of CMM applications. Hence, techniques will need to be developed to significantly reduce the wireless network bandwidth needed for CMM applications.

Among all CMM applications, cloud based rendering applications are the most communication bandwidth and computation intensive. The challenges of ensuring acceptable user experience, low cloud and mobile network costs, and scalability, will be more critical for cloud based rendering applications than most other CMM applications. Hence in the next section, we focus on addressing the above challenges for cloud based rendering, and specifically Cloud Mobile Gaming.

IV. Rendering Adaptation Approach to Address Cloud and Network Challenges

To address the challenges of ensuring high user experience, low cloud cost and network bandwidth, and high scalability for cloud based mobile rendering applications, in this section we propose an innovative rendering adaptation approach, which can dynamically vary the richness and complexity of graphic rendering depending on the network and server conditions, thereby impacting both the bit rate of the rendered video that needs to be transmitted from the cloud server to the mobile device, and the computation load on the cloud servers. While the proposed rendering adaptation approach can be widely used for any cloud based rendering applications, we use Cloud Mobile Gaming (CMG) as a running example to introduce and validate our approach in details.

A. Overview of Proposed Rendering Adaptation Approach

Graphic rendering is the process of generating an image from a graphic scene file, which usually contains geometry, viewpoint, texture, lighting, and shading information as a description of the virtual scene. It is configurable by a set of rendering parameters. The term "rendering setting" usually denotes a setting which consists of different values of these rendering parameters. We next introduce the Communication Complexity (CommC) and Computation Complexity (CompC) associated with each rendering setting in the CMG approach. The proposed rendering adaptation approach will dynamically vary the CommC and CompC of graphic rendering by changing the rendering settings, such that the resulting video bit rate and computation need can meet the available network bandwidth and cloud server computing resource respectively, thereby achieving network and server scalability for the CMG approach while ensuring a good MGUE for each user.

The Communication Complexity (CommC) of a rendering setting denotes the level of how much bit rate is needed to deliver CMG video with this rendering setting. While the video bit rate is determined by the video compression ratio used, it is largely affected by the video content complexity. To quantitatively measure CommC, we define the value of CommC of a rendering setting as the ratio of the bit rate need of this rendering setting to the minimum bit rate need among all the possible rendering settings, for the same video compression ratio.

The Computation Complexity (CompC) of a rendering setting indicates the level of GPU computation resource needed to render the game with this rendering setting. Similar to quantifying CommC, we define the value of CompC of a rendering setting as the ratio of the GPU utilization using this rendering setting to the minimum GPU utilization of all the possible rendering settings.

We propose two key principles how rendering adaptation can be used to affect CommC and CompC. The first principle is to reduce the number of objects in the graphic scene file, as not all of these objects are necessary for playing the game. For example, in a Massively Multiplayer Online Role-Playing Game (MMORPG), a player mainly manipulates one object, his avatar, in the gaming virtual world. Many other unimportant objects (e.g., flowers, small animals, and rocks) or far way avatars will not affect the user playing the game. Removing some of these unimportant objects in graphic scene file will not only reduce the load of graphic computation but also reduce the video content complexity, and thereby CommC and CompC. The second key principle for rendering adaptation is related to the complexity of rendering operations. In the rendering pipeline, many operations are applied to improve the graphic reality. The complexities of these rendering operations directly affect CompC. More importantly, some of the operations also have significant impact on content complexity, and thereby CommC, such as adjusting texture detail. If we can scale these operations, we will be able to adapt CommC and CompC as needed.

The above principles of rendering adaptation look promising to let the CMG application scale the video bit rate need and server computation need by dynamically adapting the rendering setting with proper CommC and CompC. However, since the number of different rendering settings possible may be very large, finding the optimal rendering setting for given available cloud server computing resource and network bandwidth may be time consuming. On the other hand, to be effective, rendering adaptation should be performed in real time in response to rapid changes in network and server conditions. To resolve the above conflict, we

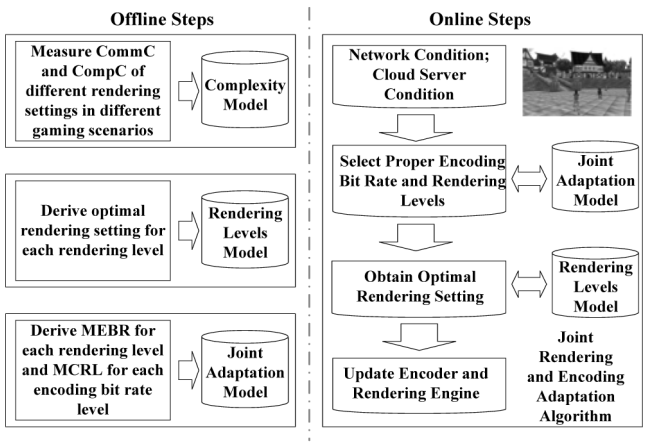


Fig. 5. Proposed rendering adaptation methodology.

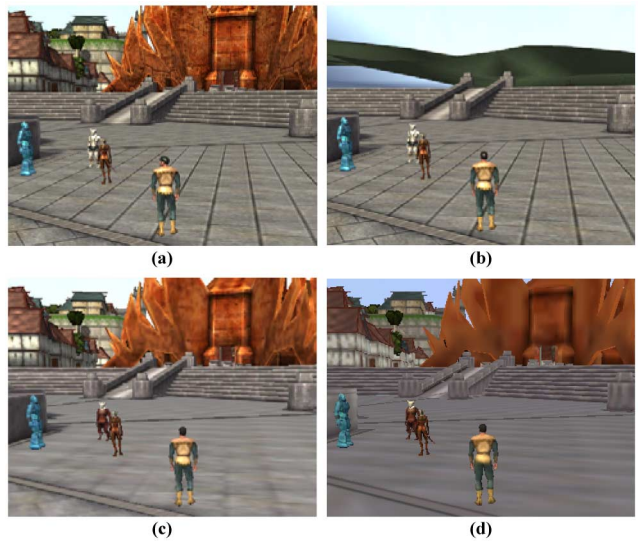


Fig. 6. Screenshots of game "PlaneShift" in different settings of view distance and texture detail (LOD). (a) 300 meters view and high LOD. (b) 60 meters view and high LOD. (c) 300 meters view and Medium LOD. (d) 300 meters view and low LOD.

propose to partition the rendering adaptation approach into two parts: offline and online steps. The offline steps will characterize and pre-determine the optimal rendering settings for different levels of CommC and CompC, thereby allowing the online steps to select and vary the rendering settings in real time in response to the fluctuations of network and server resources.

Fig. 5 gives an overview of the proposed rendering adaptation approach which involves the above mentioned offline and online steps. In the first offline step, rendering parameters are identified which can affect the communication and computation complexities of the game. Subsequently, for each possible rendering setting involving the selected parameters, CommC and CompC values are derived. This will result in a *complexity model*, which is a mapping of rendering settings to CommC and CompC values. Next, several rendering levels are selected, each of which reflects a certain CommC and a certain CompC. Then using the complexity model, optimal rendering settings are derived that meet the CommC and CompC targets of each rendering level, leading to a *rendering levels model*.

During an online gaming session, our adaptation technique can select in real time a proper rendering level and the corresponding optimal rendering setting, using the rendering levels model. However, since the mobile network bandwidth can vary very frequently, use of rendering adaptation alone may lead to frequent varying of rendering levels, which is not desirable from a user experience perspective. Therefore, we develop an online Joint Rendering and Encoding bit rate Adaptation (JREA) algorithm, which addresses the challenges of fluctuating and bandwidth constrained wireless network by judiciously utilizing the power of changing the video source through rendering adaptation, with large impact on network bandwidth needed, together with adapting the video encoding bit rate to address relatively small but frequent network bandwidth fluctuations. Adapting both rendering and video encoding jointly will necessitate understanding the optimal values (leading to a *joint adaptation model*) of encoding bit rate or rendering level that can be used when encoding or rendering is adapted respectively. Also shown in Fig. 5 are the online steps. Depending on the network and server conditions, JREA decides if the rendering level and encoding bit rate level needs to be adapted. If either of them is changed, it will check the joint adaptation model to decide if the other one needs to be changed correspondingly. If rendering level is to be changed, it will select the optimal rendering setting based on rendering levels model and update the game engine consequently to effect the rendering level change.

We have given a brief overview of our proposed rendering adaptation approach. Next we will present a detailed methodology of how we design and enable rendering adaptation for Cloud Mobile Gaming, including the three offline modeling steps and the online adaptation algorithm.

B. Adaptive Rendering Parameters and Settings

The first step in enabling dynamic game rendering adaptation in the CMG approach is to identify the adaptive rendering parameters and adaptive rendering settings. A game may have many different rendering parameters, but only a few of them have obvious impacts on CommC or CompC. An "adaptive rendering parameter" must be able to adapt at least one of CommC or CompC. An "adaptive rendering setting" is a set of values for the adaptive rendering parameters which affect CommC, CompC or both.

As discussed in Section IV-A, reducing the number of objects in the graphic scene file or reducing the complexity of rendering operations could lead to the decreases in CommC and CompC. Based on the above principles, we identify four common parameters which we believe are suitable for rendering adaptation in most 3D games: 1) *Realistic effect*: Realistic effect basically includes four parameters: color depth, anti-aliasing, texture filtering, and lighting mode. Each of the four parameters only affects part of graphic rendering. Varying any one of them may not reduce the graphic rendering load. Thus when we reduce/increase the realistic effect, we vary all four parameters. 2) *Texture detail*: This is also known as Level of Detail (LOD). It refers to how large and how many textures are used to present objects. The lower texture detail level, the lower resolution the textures have. As shown in Figs. 6(a), 6(c), 6(d), the surfaces of objects get blurred as we decrease the texture detail. 3) *View distance*: This parameter determines which objects in the camera view will be included in the resulting frame, and

TABLE IV
ADAPTIVE RENDERING PARAMETERS AND EXPERIMENT SETTINGS

Parameters	Experiment Values		
Realistic Effect	H(High)	M(Medium)	L(Low)
color depth	32	32	16
multi-sample factor	8	2	0
texture-filter factor	16	4	0
lighting mode	Vertex light	Lightmap	Disable
Texture Down Sample Rate (Texture Detail)	0, 2, 4		
View Distance (meter)	300, 100, 60, 40, 20		
Enabling Grass (Environment Detail)	Y(Yes), N(No)		

thereby should be sent to the display list for graphic rendering. For example, Figs. 6(a) and 6(b) compare the visual effects in two different view distance settings (300 m and 60 m) in the game PlaneShift (PS) [16]. 4) *Environment detail*: Many objects and effects (grass, flowers, and weather) are applied in modern games, to make the virtual world look more realistic. However they are not really necessary for users playing the game. Therefore, we could eliminate some of these objects or effects to reduce CommC and CompC if needed.

C. Derivation of the Complexity Model

Having defined adaptive rendering parameters and settings, we next use the game Planeshift (PS) as an example to explain how to derive the complexity model, where we have also elaborately studied how different adaptive rendering settings affect the CommC and CompC. Subsequently, we also have studied the impacts on CommC and CompC when video encoding setting, or video resolution, or server GPU is changed. This will help to demonstrate that the key concept that communication complexity and computation complexity can be affected by different rendering settings is broadly applicable, no matter what kind of video resolution or video encoding setting, and no matter what kind of graphic GPU is used.

1) Characterizing CommC and CompC: Four adaptive rendering parameters are selected for game PS, with their possible values shown in Table IV. We conduct experiments to characterize CommC and CompC for every possible rendering setting obtained using the values of parameters in Table IV. The experiments are conducted on a desktop server which integrates a NVIDIA Geforce 8300 graphic card. Video resolution used is VGA. The video codec used is X264, and its encoding method is set to Variable Bit Rate (VBR). The Quantization Parameter (QP) is 25, while the encoding frame rate is 15 fps and the size of Group of Pictures (GOP) is 30. We have randomly selected several different gaming scenes. In each test scene, for each rendering setting, we let the game avatar roam in the gaming world along the same route. We measure the average compressed video bit rate and GPU utilization in each experiment test to calculate the CommC and CompC.

Fig. 7 shows some representative data points from the experiments. For each adaptive rendering parameter, we first present the sample results of CommC and CompC in two figures respectively. In each of these figures, each plot represents a rendering setting where only one of the rendering parameters is varied

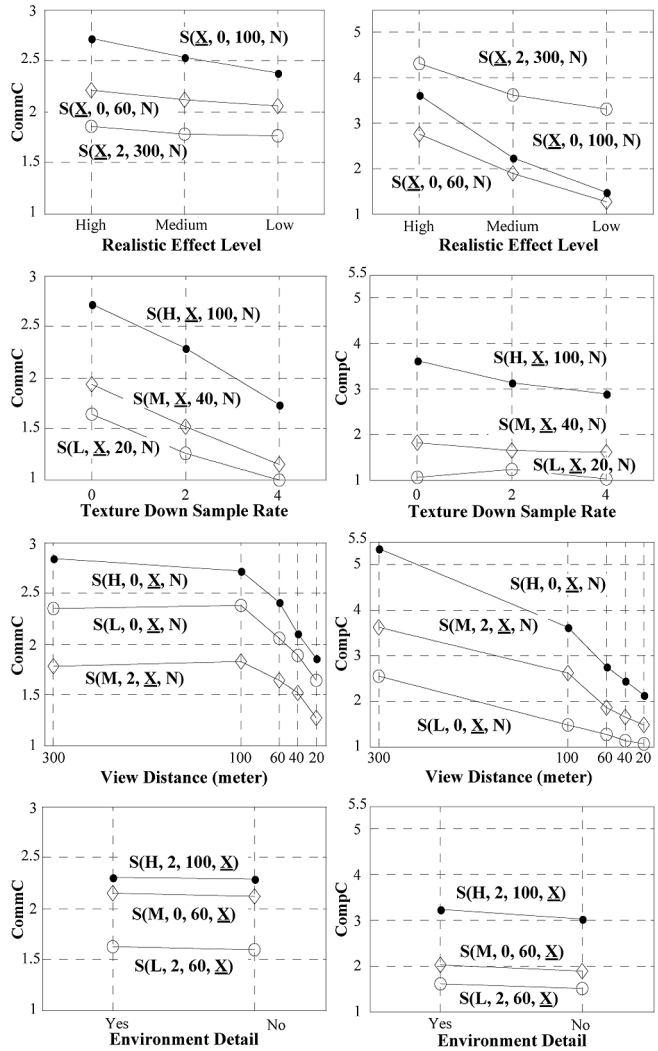


Fig. 7. Sample data to show how CommC and CompC vary for different rendering parameters, for game PS in VGA resolution.

(marked by "X" in the associated setting), while keeping the other parameters to fixed values shown in the rendering setting tuple. From Fig. 7, we have the following observations: 1) Realistic effect has high impact on CompC. But it has low impact on CommC, because realistic effect has little impact on content complexity of game video. 2) Texture down sample rate significantly affects CommC, as the highest CommC is almost 3. However, texture detail almost does not affect CompC. This is because the reduced textures in different levels for an object are pre-calculated and loaded in the memory, so that the graphic pipeline can load the textures quickly without any additional computing cost. 3) View distance will significantly affect both CommC and CompC. While its impact on CompC is almost linear, impact on CommC becomes clear only below a certain point (100 meters). 4) The impact of enabling environment details on CommC and CompC is limited (up to 9%), mainly because the effect of environment details in game PS is very simple such that varying environment details has low impact on frame content complexity and computation complexity.

In Fig. 8, we present a visual representation of the complexity model. Each point in Fig. 8 denotes an adaptive rendering setting, with the corresponding values on the x- and y-axes de-

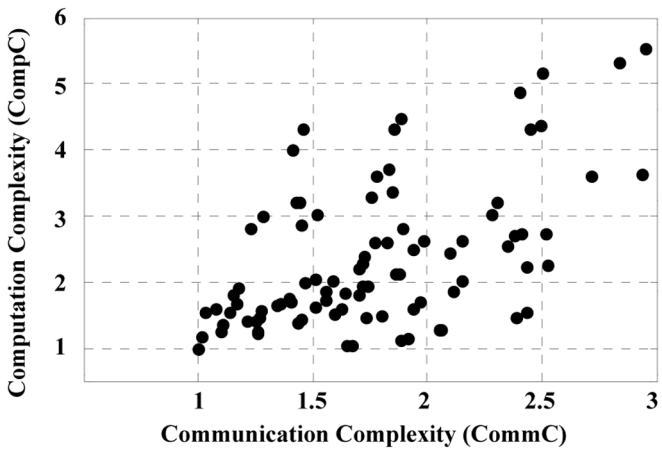


Fig. 8. Complexity model of game PS.

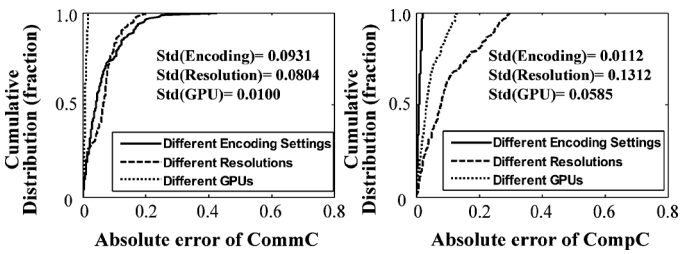


Fig. 9. Absolute error distribution and standard deviation of measured CommC and CompC in the test with different video encoding settings, different video resolutions, and different CPUs, for game PS.

noting the CommC and CompC values for that rendering setting. From Fig. 8, we observe that for game PS, the maximum CompC and CommC of the adaptive rendering settings can be as high as about 5 and 3.

2) Characterizing CommC and CompC With Different Encoding Settings, Video Resolutions, and GPUs: The complexity model we presented above was derived using a certain video encoding setting, video resolution, and GPU. We next investigate the impact of using different video encoding and resolution settings, and different GPUs, on the complexity model. We have conducted experiments and measured CommC and CompC of each rendering setting in three test cases: a) using various encoding settings (different QP and GOP settings), b) using three different resolutions (QVGA, CIF, and VGA), and c) using three different GPUs (Intel GMA4500, NVIDIA 8300, and NVIDIA GTX580). Fig. 9 shows absolute error distribution and standard deviations of measured CommC and CompC in these test cases. From Fig. 9, we can observe that the overall variations of CommC and CompC in these different test cases are not significant. Hence, we can conclude that the offline modeling step does not need to characterize the CommC and CompC and create different complexity models for different video resolutions, or video encoding settings, or different platforms.

D. Derivation of the Rendering Levels Model

In this section, we introduce how we leverage the complexity model to derive the rendering levels model, which provides an optimal rendering setting for each rendering level. Each rendering level has two dimensions: 1) CommC rendering level,

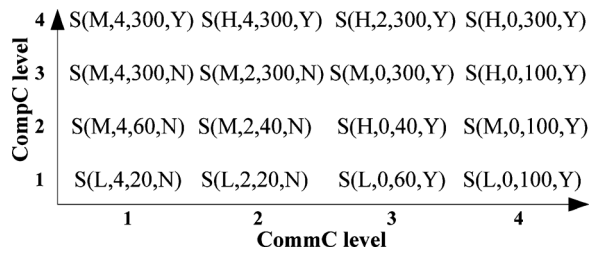


Fig. 10. Rendering levels model of game PS.

reflecting the level of network bandwidth need of that rendering level; 2) CompC rendering level, reflecting the level of computation need of that rendering level. We use L_{ij} to denote a rendering level, whose CommC and CompC rendering levels are i and j, and target complexities are $CommCL_i$ and $CompCL_j$ respectively. For each game, depending on the range of CommC and CompC values derived during the complexity modeling step, the CommC and CompC levels can be set in many ways. In our work, we evenly divide the possible ranges (from the maximum value to minimum value) of CommC and CompC into m and n levels, $CommCL_i$ (give range of i) and $CompCL_j$ (give range of j). Once having target $CommCL_i$ and $CompCL_j$ rendering levels, we can find an optimal rendering setting s_{ij} for each rendering level L_{ij} from all adaptive rendering settings set, such that the root mean squared error between the target complexities ($CommCL_i$ and $CompCL_j$) of L_{ij} and measured complexities of setting s_{ij} obtained from complexity model is minimized. If we apply the above approach with the complexity model (Fig. 8) of game PS, we can obtain the rendering levels model of game PS, as shown in Fig. 10. For instance, if CommC rendering level and CompC rendering level are both 2, the optimum rendering setting to be used by the graphic engine is (M, 2, 40, N): medium realistic effect, 2 for texture down sample rate, 40 meters for gaming view distance, and not enabling the environment details.

It should be noted that the modeling results presented in Figs. 7 and 8 are the average costs measured from the off-line analysis of a large number of CMG runs in the ourdoor gaming scenarios. Though the CommC and CompC of each level may vary at run-time depending on the gaming scenarios, the optimal rendering settings in Fig. 10 is derived using the average costs (Figs. 7 and 8). It should also be noted that the ability to reduce video content complexity by adapting view distance may be limited in some gaming scenarios. For instance, in the indoor scenario where the objects rendered are all in a relatively short distance, we may not want to further reduce view distance to reduce the video content complexity. Although it will become difficult to affect video content by adapting view distance in the indoor scenario, the other rendering adaptive parameters are still capable to vary the content complexity of gaming video. And more importantly, a higher rendering level in Fig. 10 still has higher costs than a lower rendering level. This property makes the rendering settings shown in Fig. 10 useful for the rendering adaptation algorithm described later.

E. Derivation of the Joint Adaptation Model

Having derived rendering levels model, we can come up with an online rendering adaptation technique. However, as we

TABLE V
MINIMUM ENCODING BIT RATE FOR EACH COMMC RENDERING LEVEL FOR GAME PS IN VGA RESOLUTION

CommC Rendering Level	4	3	2	1
MEBR	300	250	200	150

TABLE VI
MAXIMUM COMMC RENDERING LEVEL FOR EACH ENCODING BIT RATE FOR GAME PS IN VGA RESOLUTION

Encoding Bit Rate (kbps)	700	600	500	400	300	250	200	150
MCRL	4	3	3	2	2	2	1	1

mentioned in Section IV-A, frequently varying the rendering settings to address fluctuating mobile network bandwidth may not be acceptable from a user experience perspective. Therefore, we develop an online Joint Rendering and Encoding bit rate Adaptation (JREA) algorithm. In JREA algorithm, rendering adaptation technique is used to address cloud server computing resource constraint by varying its CompC rendering level, while rendering and encoding bit rate adaptation technique will be jointly utilized to address network bandwidth constraint by varying CommC rendering level and video bit rate. However, it is imperative to know how to optimally select video encoding bit rates and CommC rendering levels such that MGUE is maximized. In fact, for each CommC rendering level, there is a Minimum Encoding Bit Rate (MEBR) that is acceptable for the resulting video quality. And similarly, for each bit rate we use for gaming video, there is a Maximum CommC Rendering Level (MCRL) that provides the video quality which has minimum impacts on user gaming experience. We next explain how we obtain MEBR and MCRL to derive the joint adaptation model.

During the offline experiment, for each bit rate level, the average video PSNR of compressed game video is measured. The Minimum Encoding Bit Rate (MEBR) is the minimum bit rate which can at least provide the user minimum acceptable PSNR [11]. Table V shows the MEBR for each CommC rendering level for game PS in VGA resolution. Thus in the online joint adaptation algorithm, when encoding bit rate used is lower than the MEBR associated with the current CommC rendering level being used, the rendering level will be adapted to a lower level to get a lower required MEBR, such that the user perceived video quality is acceptable.

Similar to the method used to derive the MEBR values for each rendering level, the following method is used to derive the MCRL for each video encoding bit rate used. For each encoding bit rate, offline experiment selects different CommC rendering levels. Then for each CommC rendering level, it measures the average PSNR of compressed video among different test scenes. For each encoding bit rate, the MCRL is the maximum CommC rendering level in which the resulting video PSNR is at least higher than the excellent video quality threshold [11] below which user will feel the impacts due to the video quality. Table VI shows the MCRL for each encoding bit rate for game PS in VGA resolution. During the gaming session, depending on the encoding bit rate used, the CommC rendering level (MCRL) is periodically updated from Table VI, such that rendering setting used is maximized while user perceived video quality remains unimpaired.

F. Online JREA Algorithm

The motivation for developing an online Joint Rendering and Encoding bit rate Adaptation (JREA) algorithm is presented in Section IV-A. We next describe the steps of the JREA algorithm, which decides when and how to switch the CommC rendering level, CompC rendering level, and the encoding bit rate during a gaming session, in response to the current network conditions and server utilization:

1) The first step is to decide the encoding bit rate K used to encode the rendered video. During a short time interval λ, if the network Round trip Delay *(RDelay)* keeps increasing and its average value is greater than minimum Acceptable RDelay (RD_A)[11], JREA algorithm will reduce encoding bit rate K. On the other hand, if for a significant time T_1, *RDelay* remains below RD_A and there is no packet loss, it will increase the encoding bit rate.
2) The second step is to check and update CommC rendering level I. After the first step, the new encoding bit rate may be below the Minimum Encoding Bit Rate (MEBR) for the current CommC rendering level I, which will lead to an unexpected user experience as we discussed before. If this happens, JREA algorithm has to reduce CommC rendering level to reduce the Minimum Encoding Bit Rate. On the other hand, if the CommC rendering level has not been changed for over a certain significant period T_2, it will be updated and changed to the Maximum CommC Rendering Level (MCRL) depending on the current encoding bit rate.
3) The third and last step is to decide on CompC rendering level J, depending on server utilization ServUtil. If ServUtil is over U_1, the lower CompC rendering level is selected. Otherwise, if CompC rendering level has not been changed for more than time T_3, and ServUtil is below U_2, it increases CompC rendering level J by 1.

Next, based on the new selected CommC and CompC rendering levels JREA algorithm will use the optimal rendering settings from Fig. 10 to update the game graphic engine, while it uses the new selected video bit rate to update the video encoder.

G. Experimental Validation

We next report on experiments conducted to verify the effectiveness of the proposed rendering adaptation technique. For the experiments reported in this paper, we use the values 3 seconds, 20 seconds, 60 seconds, 60 seconds, 90%, and 40% for the parameters $\lambda, T_1, T_2, T_3, U_1$, and U_2 respectively. The game used is PlaneShift, its RD_A is 440 ms [11]. The rendering levels model and joint adaptation model for game PS are pre-calculated off-line using the methods described in Sections IV-D and IV-E. It should be noted that our proposed JREA algorithm can leverage any encoding bit rate adaptation technique with our proposed rendering adaptation technique. In this paper, for the purpose of experimental results, we use an encoding bit rate adaptation technique [17] which has been shown to produce better video quality by being aware of the gaming content.

1) Addressing Network Effects on User Experience: To evaluate the effectiveness of our proposed technique in addressing

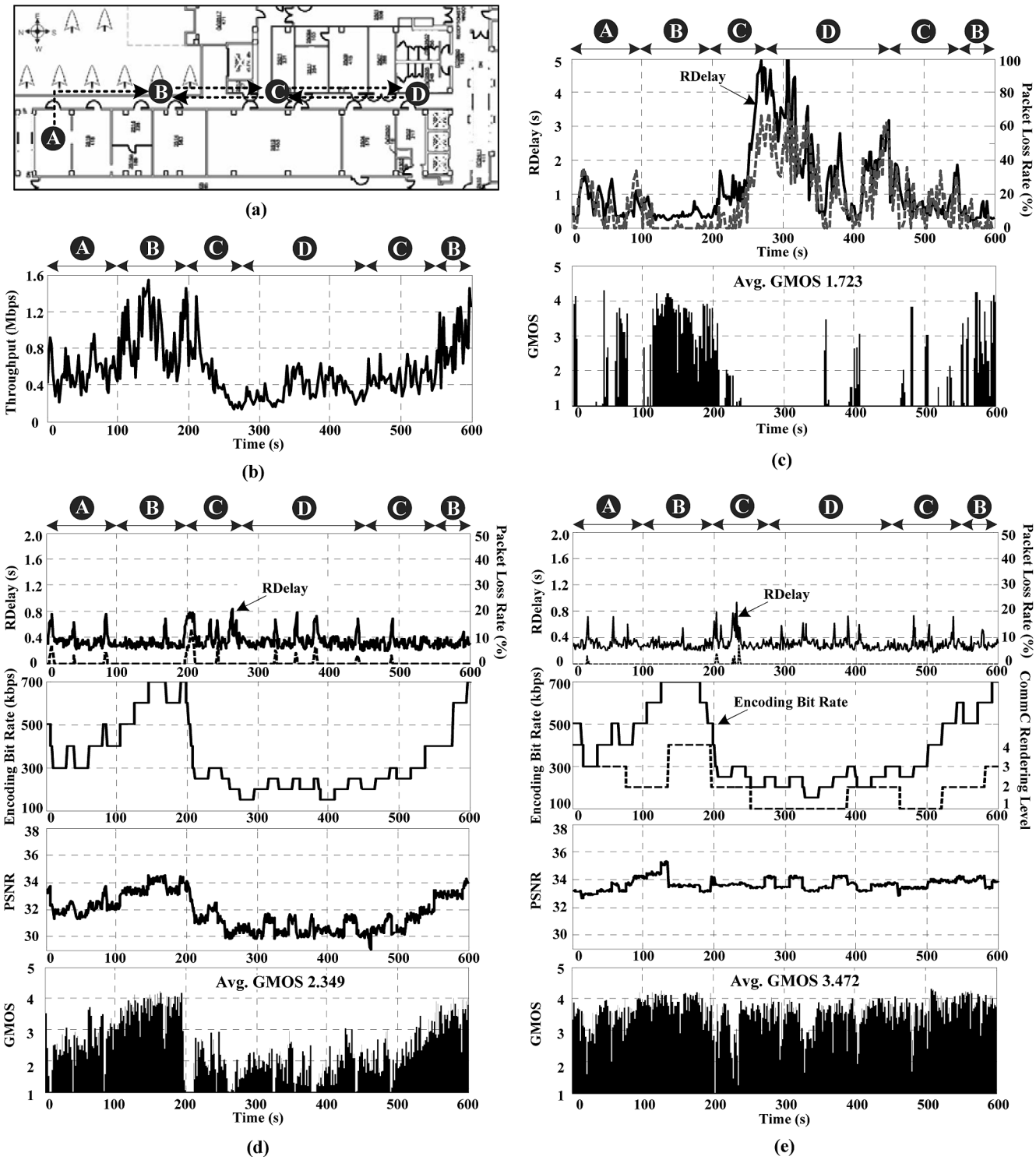


Fig. 11. Experiment results to demonstrate the effectiveness of proposed rendering adaptation technique to address the challenge of ensuring MGUE. (a) Test environments. (b) Network bandwidth measurement results. (c) CMG without any adaptation technique. (d) CMG with encoding bit rate adaptation technique only. (e) CMG with our proposed Joint Rendering and Encoding bit rate Adaptation (JREA) technique.

network bandwidth constraints thereby ensuring user experience, we carried out multiple experiments using a 3G network and multiple test environments (locations, times) having different network conditions. In this section, we select one of the test environments to compare and discuss in details the effectiveness of our proposed rendering adaptation technique. Fig. 11(a) shows a test environment used in the UCSD campus, where the tests are conducted starting at an indoor location A (our lab), then at location B (outdoor), C (indoor), D (indoor), then from D back to C, B. The locations are selected as they display different network conditions: network bandwidth, and Signal to Interference plus Noise Ratio (SINR), received by the

mobile device at each location. In this case, the SINR of these four places from highest to lowest are B, A, C, D.

Fig. 11(b) shows the maximum mobile network throughput measured at the locations. Figs. 11(c)–11(e) present data collected from experiments when the game PS is played in VGA resolution in three scenarios: CMG without using any adaptation technique, CMG only using game aware video encoding bit rate adaptation technique [17], and CMG using our proposed JREA technique, respectively. We provide below a summary of the key observations from our experiments:

1) Fig. 11(c) shows the resulting round-trip delay, Packet Loss, and GMOS when CMG is used without any adaptation technique, with gaming video encoded and streamed at 700 kbps, adequate for good video quality at VGA resolution. At outdoor location B, as the wireless channel rate is high due to good SINR, the resulting network delay and packet loss are relatively low, and hence the GMOS score measuring the mobile gaming user experience is mostly above 3.0 (acceptable user experience threshold). However, at indoor locations, A, C, and D, due to the bad SINR, the wireless channel rate dips to a lower level. This causes network congestion, reflected by the high network delay and packet loss rate, leading to unacceptably poor gaming quality (GMOS scores below 3.0).
2) In the experiment only using encoding bit rate adaptation, the encoding bit rates used are adapted to the fluctuating network conditions. As shown in Fig. 11(d), the network congestion is almost eliminated, and as a result, network delay and packet loss are greatly improved at all the locations. However, the video quality (PSNR) deteriorates when video bit rate is lowered by encoding bit rate adaptation. When wireless network bandwidth is extremely low in bad SINR locations, like D, the user has very poor experience, reflected by poor GMOS, primarily due to the poor video quality.
3) Fig. 11(e) shows the results of applying the proposed rendering adaptation technique, including the resulting adaptive bit rates and CommC rendering levels used. In contrast to the results shown in Figs. 11(c), 11(d), 11(e) shows that application of our rendering adaptation technique can greatly improve the network delay and packet loss rate, while maintaining a good video quality (PSNR). Consequently, the user gaming experience, which includes response time, is significantly enhanced at all the locations, reflected by the relatively high and stable GMOS, dipping below 3.0 only very occasionally when the adaptation algorithm is responding to the channel rate variations.

2) Addressing Mobile Network Cost and Scalability: Our proposed CMG approach with rendering adaptation also addresses the challenges of reducing cost of using mobile network, and making the cloud mobile gaming approach scalable in terms of number of concurrent users that can use the network at congested times. For example, the average bit rate consumed by the CMG session with rendering adaptation is about 384 kbps for the results presented in Fig. 11(e), reduced from the original bit rate of 700 kbps without adaption (Fig. 11(c)), while also significantly enhancing user experience. Consequently, if we use the same daily usage assumption stated in Section III-C, the monthly mobile network data cost for a CMG smartphone user will be $39, almost half the cost when using CMG without our proposed technique. Secondly, the same mobile network can accommodate 2X more CMG users with application of our proposed technique.

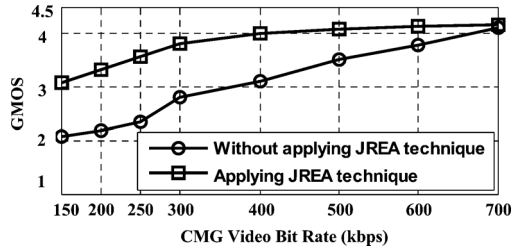


Fig. 12. Experiment results showing the effect of bit rate throttling on CMG user experience (GMOS) with and without proposed JREA technique.

The above analysis is based on an application based model where the mobile network cost is affected by the available network bandwidth. Alternatively, the CMG provider or mobile network operator can develop and apply policies limiting/throttling the CMG video bit rate allowed, depending on either user subscription levels or network congestion levels, to control costs or increase network capacity during congested times, respectively. We have conducted experiments using the same game PS, VGA resolution, to evaluate the impact of bit rate throttling policies on CMG user experience, with and without using our proposed technique. Fig. 12 shows that the average GMOS score of a CMG gaming session when limiting the CMG video bit rate to the rates specified. We observe that the adverse impact of bit rate throttling on CMG user experience can be significantly reduced when using our proposed adaptive CMG approach. Moreover, Fig. 12 shows the feasibility of enforcing significant bit rate throttling, while keeping the user experience to acceptable level $(\mathrm{GMOS} >= 3.0)$, thereby allowing the CMG providers and network operators to offer significant data cost savings to users, while also effectively managing network congestion. For example, by using adaptive CMG with rendering adaptation, the monthly data cost can be reduced to $15 by applying a 150 kbps data rate limit (assuming same monthly usage statistics and data costs as in Section III-B), while ensuring acceptable user experience (GMOS 3.0), not possible with original CMG. Similarly, instead of streaming CMG video at 700 kbps, almost 5X more users can be accommodated in a congested cell by enforcing a bit rate limiting policy of 150 kbps for each CMG user, while delivering acceptable user experience.

3) Addressing Cloud Server Cost and Scalability: To demonstrate the effectiveness of the proposed rendering adaptation technique to reduce cloud cost per CMG user, and ensure scalability in terms of number of concurrent users that can be served without additional cloud resources, we conduct experiments where we increase server GPU load by increasing the number of concurrent game engine tasks executed. The experimental server integrates a NVIDIA GTX580 graphic card. We initialize the CMG server with one game engine for a cloud mobile gaming session. After every 60 seconds, we start a new PS game engine for each new concurrent gaming session. Each game engine is configured to render 15 frames per second, but the actual rendered frame rate produced may drop below 15 if the GPU resource is over utilized. In the

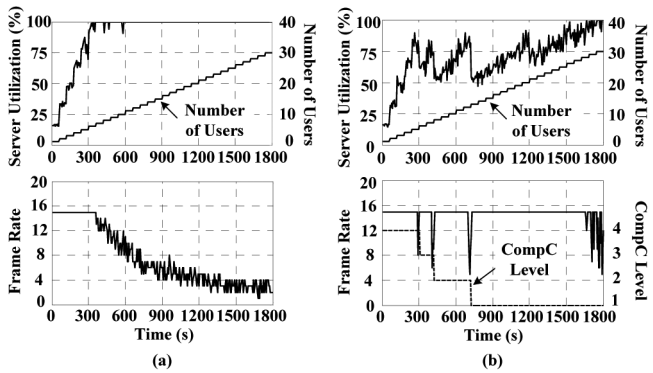


Fig. 13. Experiment results to demonstrate the effectiveness of proposed JREA technique in addressing cloud server cost and scalability. (a) Without applying JREA technique. (b) Applying JREA technique.

latter cases, the gaming user experience will suffer, with either gaming video appearing jerky, or the response time appearing slower than expected.

Fig. 13 presents the effects of increasing the number of concurrent gaming sessions on a CMG server, and the resulting server GPU utilization and the rendering quality achieved (rendering frames per second) for one sample gaming session, without and with applying rendering adaptation technique (Figs. 13(a) and 13(b) respectively). It should be noted that when the server utilization goes above the upper utilization threshold U_1, the adaptation technique will adapt the CompC rendering levels on all the gaming engines executed. We summarize below the following observations:

1) As shown in Fig. 13(a), without the use of our rendering adaptation technique, the CMG server can support only 6 CMG sessions with good quality (expected rendering rate of 15 fps), as the GPU utilization reaches 100% when we add the 7th game session and the rendered frame rate drops to as low as 12. With the addition of each new game engine task, the rendering frame rate keeps going down, to as low as 1 frame per second when the CMG server has to execute 30 concurrent game engines.
2) In contrast, as shown in Fig. 13(b), the CMG server can support up to 27 clients when using our proposed rendering adaptation technique, without deterioration in rendered quality (rendered frame rate). With appropriate adaptation of the CompC rendering levels used, rendering adaptation technique is able to ensure that the GPU is able to deliver the expected 15 fps for each gaming session, dipping a little below 15 only very occasionally when the adaptation algorithm is responding to the changes in server GPU loading. The above experimental results demonstrate that our proposed rendering adaptation technique is able to reduce the CMG computation need by about 5X without deteriorating the CMG user experience. This ability will be very helpful for CMG provider to reduce cloud service cost per user and ensure cloud service scalability.

V. Future Directions to Address CMM Challenges

We have presented an adaptive mobile cloud computing approach to address the challenges associated with Cloud Mobile Gaming, one of the most computing and communication intensive Cloud Mobile Media applications. We will conclude this paper by discussing two additional new approaches which we believe can significantly help address the user experience, cost and scalability challenges associated with CMM applications.

As discussed in Section III, a critical challenge for CMM applications is the latency and jitter associated with the uplink and downlink transmissions between the mobile device and the Internet cloud servers. Moreover, the transmission of large amount of content between cloud servers and mobile devices, inherent in CMM applications, poses a major concern for the capacity of the networks to enable CMM applications. A promising direction will be to bring cloud computing to the edge of the mobile network, supplementing gateway nodes in the mobile Core Network (CN), and edge nodes like base stations in Radio Access Networks (RAN), and Femto and WiFi access points, with computing and storage resources, to form a true *Mobile Cloud*. With a Mobile Cloud architecture, content processing (like graphic rendering or video encoding) and retrieval can be performed at the edge of the mobile networks, as opposed to in Internet clouds, thereby reducing round trip network latency, as well as reducing congestion in the mobile CN and RAN.

While the direction of Mobile Cloud looks appealing, there are multiple challenges that need to be addressed. Since there are thousands of base stations and access points, the proposed Mobile Cloud will be a massively distributed network of smaller computing and storage resources, as opposed to the more centralized architecture of Internet clouds consisting of a few data centers with much larger computing and storage footprints. The above difference has interesting implications and challenges. For example, we investigated and demonstrated the use of Mobile Clouds, consisting of smaller caches in the RAN eNodeBs [18], and larger caches in the CN gateways [19], to improve the latency of video delivery to mobile devices, and the capacity of networks to support concurrent video requests. We observed that conventional Internet caching policies, like caching Most Popular Videos (MPV), or discarding the Least Recently Used (LRU) videos, are not as effective with the smaller sized RAN caches. Hence, we had to develop RAN-specific caching policies, which cache only videos most relevant to the mobile users present in the associated RAN cell, and can significantly improve the cache hit ratio. Similarly, though the presence of gateway caches help in addressing mobility of users from one cell to another, we need to address the challenge of determining how to partition caching of videos between the base station and gateway caches. In the future, it is promising to investigate the design and use of efficient Mobile Cloud architectures and algorithms for other CMM applications like Cloud Mobile Gaming, to improve response time and hence user experience.

One of the biggest challenges for mobile cloud computing, in particular when it comes to computing and bandwidth hungry CMM applications, is ensuring scalability for large number of simultaneous users, both from the high cloud costs that may be incurred, and the limited capacity of mobile networks. A promising direction is to develop *Mobile Cloud Scheduling* techniques, which can simultaneously consider the cloud computing and storage resources, together with the network availabilities for each CMM client, including the availability of alternative network accesses like Femto and WiFi to offload

CMM traffic, such that the number of simultaneous CMM users is maximized, while minimizing cloud cost. Note that current cloud scheduling techniques address the problem of efficiently assigning diverse cloud resources to heterogeneous requirements of application tasks, with the objective of ensuring fairness among requesting tasks [20] or reduce cloud cost [21]; however, these techniques do not consider the mobile network constraints, which is an important scalability challenge for CMM applications. Our initial work developing mobile cloud scheduling techniques for Cloud Mobile Gaming has shown promising results: the ability to significantly increase the number of simultaneous CMG users using available network resources, while reducing cloud cost [22]. In the future, mobile cloud scheduling techniques will need to be developed for other CMM applications, as well as consider capacity limited computing and storage resources in the Mobile Cloud.

REFERENCES

[1] Canalys, Smart Phones Overtake Client PCs, Feb. 2011. [Online]. Available: http://www.canalys.com/newsroom/smart-phones-overtake-client-pcs-2011.

[2] IDC, More Mobile Internet Users Than Wireline Users in the U.S. by 2015, Sep. 2011. [Online]. Available: http://www.idc.com/getdoc.jsp?containerId=prUS23028711.

[3] Juniper Research, Mobile Cloud: Smart Device Strategies for Enterprise & Consumer Markets 2011-2016, Jul. 2011. [Online]. Available: http://juniperresearch.com/.

[4] MarketsAndMarkets, World Mobile Applications Market—Advanced Technologies, Global Forecast (2010-2015), Aug. 2010. [Online]. Available: http://www.marketsandmarkets.com/.

[5] GSMA OneAPI. [Online]. Available: https://gsma.securespsite.com/access/Access%20API%20Wiki/Home.aspx.

[6] S. Wang and S. Dey, "Rendering adaptation to address communication and computation constraints in cloud mobile gaming," in *Proc. IEEE GLOBECOM*, Miami, FL, USA, Dec. 2010.

[7] ARCchart, The Mobile Cloud: Market Analysis and Forecasts, 2011.

[8] Cisco, Cisco Visual Networking Index: Forecast and Methodology, 2011-2016, 2012.

[9] 3D Game System Requirements. [Online]. Available: http://www.game-debate.com.

[10] Anandtechiphone Performance Reviews. [Online]. Available: http://www.anandtech.com.

[11] S. Wang and S. Dey, "Cloud mobile gaming: Modeling and measuring user experience in mobile wireless networks," in *Proc. ACM SIGMOBILE MC2R*, 2012, vol. 16, no. 1, pp. 10–21.

[12] S. Wang and S. Dey, "Modeling and characterizing user experience in a cloud server based mobile gaming approach," in *Proc. IEEE GLOBECOM*, Honolulu, HI, USA, Dec. 2009.

[13] WoW Basic Demographics. [Online]. Available: http://www.nickyee.com/daedalus/archives-/001365.php.

[14] Y.-T. Lee *et al.*, "World of warcraft avatar history dataset," in *Proc. ACM Multimedia Syst.*, 2011.

[15] Asia Times, Game still on at Tencent, Mar. 2010. [Online]. Available: http://www.atimes.com/atimes/China_Business/LC24Cb01.html.

[16] Planeshift. [Online]. Available: http://www.planeshift.it/.

[17] S. Wang and S. Dey, "Addressing response time and video quality in remote server based internet mobile gaming," in *Proc. IEEE WCNC*, Sydney, Australia, Mar. 2010.

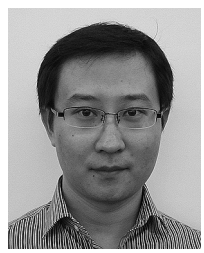
[18] H. Ahlehagh and S. Dey, "Video caching in radio access network," in *Proc. IEEE WCNC*, Paris, France, Apr. 2012.

[19] H. Ahlehagh and S. Dey, "Hierarchical video caching in wireless cloud: Approaches and algorithms," in *Proc. IEEE ICC Workshop on Realizing Advanced Video Optimized Wireless Networks*, Ottawa, ON, Canada, Jun. 2012.

[20] G. Lee, B. Chun, and R. Katz, "Heterogeneity-aware resource allocation and scheduling in the cloud," in *Proc. 3rd USENIX Workshop Hot Topics in Cloud Comput.*, Jun. 2011.

[21] R. Bossche, K. Vanmechelen, and J. Broeckhove, "Cost-optimal scheduling in hybrid IaaS clouds for deadline constrained workload," in *Proc. IEEE Int. Conf. Cloud Comput.*, Miami, FL, USA, Jul. 2010.

[22] S. Wang, Y. Liu, and S. Dey, "Wireless network aware cloud scheduler for scalable cloud mobile gaming," in *Proc. IEEE ICC*, Ottawa, ON, Canada, Jun. 2012.



Shaoxuan Wang received the B.S. degree in electrical engineering and the M.S. degree in wireless communication, both from Peking University, Beijing, China. He is currently working toward the Ph.D. degree in computer engineering in the Department of Electrical and Computer Engineering, University of California, San Diego, CA, USA.

He is the co-inventor of 1 US and 1 international patents, with several others pending.

Sujit Dey (SM'03) received the Ph.D. degree in computer science from Duke University, Durham, NC, USA, in 1991.

He is a Professor with the Department of Electrical and Computer Engineering, University of California, San Diego, CA, USA, where he heads the Mobile Systems Design Laboratory, which is engaged in developing innovative mobile cloud computing architectures and algorithms, adaptive multimedia and networking techniques, low-energy computing and communication, and reliable system-on-chips, to enable the next-generation of mobile multimedia applications. He serves as the Faculty Director of the von Liebig Entrepreneurism Center at UCSD, and as the Chief Scientist, Mobile Networks, at Allot Communications. He is affiliated with the California Institute of Telecommunications and Information Technology (Calit2), and the UCSD Center for Wireless Communications. He founded Ortiva Wireless in 2004, where he served as its founding CEO and later as CTO and Chief Technologist until its acquisition by Allot Communications in 2012. Prior to Ortiva, he served as the Chair of the Advisory Board of Zyray Wireless until its acquisition by Broadcom in 2004, and as an advisor to multiple companies including ST Microelectronics and NEC. Prior to joining UCSD in 1997, he was a Senior Research Staff Member at the NEC C&C Research Laboratories in Princeton, NJ, USA. He has co-authored close to 200 publications, including journal and conference papers, and a book on low-power design. He is the co-inventor of 17 US and 2 international patents, resulting in multiple technology licensing and commercialization.

Dr. Dey has been the recipient of several Best Paper awards, and has chaired multiple IEEE conferences and workshops.