MULTIMEDIA COMMUNICATIONS TECHNICAL COMMITTEE
IEEE COMMUNICATIONS SOCIETY
http://www.comsoc.org/~mmc

E-LETTER

Vol. 8, No. 6, November 2013

IEEE COMMUNICATIONS SOCIETY

CONTENTS

Message from MMTC Chair

Dear MMTC colleagues:

In this brief editorial, I would like to summarize the status of our activities related to the sponsorship of workshops and conferences.

As you probably know, we are technically sponsoring conferences and workshops whose scope is in the area of multimedia communications and that see the involvement of the MMTC community in the organizing or technical program committees. Indeed, as far as there is interest in the event from our members and there are at least three colleagues that are willing to act as liaison between the conference and MMTC by serving in the TCP or in the organizing committees, then we give our formal support. This support is also enough to get the ComSoc Technical sponsorship. In the last year we have already endorsed 5 conferences and we have received a couple of new requests which are in the process of being evaluated. Additionally, this year we have organized the workshop on "Hot Topics in 3D" in co-junction with ICME 013 and we are co-organizing the workshops on "Cloud Computing Systems, Networks, and Applications" (CCSNA) and "Quality of Experience for Multimedia Communications" (QoEMC) to be held next December jointly with Globecom 2013, in Atlanta. With reference to the workshops held jointly with our major reference conferences (i.e., ICC, Globecom and ICME), **I would like to invite you to take an active role by suggesting new proposals**. If you are interested in send an email to me and I will be happy to provide all the necessary information and support.

Additionally, I would like to take this occasion to make you aware of a standardization activity sponsored by IEEE SA that you may be interested in. It is the IEEE P1907.1 Standard for Network-Adaptive Quality of Experience (QoE) Management Scheme for Real-Time Mobile Video Communications, which defines a mechanism for managing the end-to-end quality of real-time video user experience. It clearly within the MMTC scope and then you may be interested in. Also in this case, to get involved contact me or send an email to the other officers you may find in the relevant webpage http://grouper.ieee.org/groups/1907/1/

I would like to thank all of you that worked towards the organization of the successful MMTC workshops and conferences and invite to continue supporting the activities of our prosperous Committee!

Luigi Atzori
Europe Vice-Chair of Multimedia Communications TC of IEEE ComSoc
Researcher, University of Cagliari, Italy (l.atzori@diee.unica.it)

EMERGING TOPICS: SPECIAL ISSUE ON CLOUD COMPUTING FOR MULTIMEDIA

Cloud Computing for Multimedia

Guest Editor: Nabil J. Sarhan, Wayne State University, USA

nabil@ece.eng.wayne.edu

The interest in cloud computing for multimedia has recently increased dramatically. As cloud computing offers low cost, high scalability, enhanced reliability, and device independence, it can be used to efficiently deploy multimedia services. Performing processing and storage on the cloud reduces the demands on user devices, especially mobile devices, which have limited energy, storage, and computational capability. Therefore, new powerful multimedia applications have become possible.

This special issue of MMTC E-Letter focuses on recent advances in cloud computing for multimedia. It includes five high-quality articles, spanning a variety of topics, namely cloud mobile multimedia, cloud content-based image retrieval, cloud object classification, cloud gaming, and competitive bandwidth reservation of cloud resources.

In the first article, titled "Addressing User Experience, Cost and Scalability Challenges of Cloud Mobile Multimedia Applications*",* Dey, Wang, and Liu from the University of California, San Diego discuss the main challenges in enabling Cloud Mobile Multimedia (CMM) applications. These applications employ cloud computing to provide rich multimedia experiences that are not possible otherwise from mobile devices. They also discuss solutions for addressing these challenges.

The second article, titled "Content Based Image Retrieval on Cloud" and authored by Lu, Cai and Wen from Nanyang Technological University, Singapore, provides a survey of recent work on applying the MapReduce framework for content-based image retrieval on the cloud. As MapReduce is typically used to process large datasets by a distributed system, it can be employed for searching for similar images from the collection of all Internet images.

The third article, titled "Utilizing the Cloud for Image-Based Food Recognition", is a collaborative study between the University of Ottawa, Canada and Istanbul Şehir University, Turkey. The authors Pouladzadeh, Bakirov, Shirmohammadi, and Bulut consider the use of a cloud-based system for the automatic recognition of food in images captured by user smartphones. This approach can help users, including those who are overweight or suffering from obesity, to track their calorie intakes. The authors develop a new classification scheme for food recognition based on Support Vector Machine (SVM).

In the fourth article, titled "Cloud Gaming: From Concept to Reality", Wu and Xue from Sun Yat-sen University, China provide insights on important performance issues and design alternatives of cloud gaming systems. In particular, they provide and analyze in-depth experimental results of CloudUnion, a leading cloud gaming system in China.

The last article, titled "Competitive Bandwidth Reservation via Cloud Brokerage for Video Streaming Applications" and authored by Jin and Kwok from the University of Hong Kong, explores the problem of competitive resource procurements in a cloud broker market. In particular, it models the pricing scheme of the cloud broker and tenant surplus. It also presents a non-cooperative game to model such competitive resource procurements.

We hope that this issue will be both informative and a pleasure to read.

Finally, we would like to thank all the authors for their great contributions and the MMTC E-Letter Board for all their support and for making this special issue possible.

Nabil J. Sarhan received the Ph.D. and M.S. degrees in Computer Science and Engineering at Pennsylvania State University and the B.S. degree in Electrical Engineering at Jordan University of Science and Technology. Dr. Sarhan joined Wayne State University in 2003, where he is currently an Associate Professor of Electrical and

Computer Engineering and the Director of Wayne State Multimedia Computing and Networking Research Laboratory. His main research areas are video streaming and communication, computer and sensor networks, automated video surveillance, multimedia systems design, energy-efficient systems, and cross-layer optimization. Dr. Sarhan is the Chair of the Interest Group on Media Streaming of the IEEE Multimedia Communication Technical Committee. He is an Associate Editor of the IEEE Transactions on Circuits and Systems for Video Technology. Dr. Sarhan has been involved in the organization of numerous international conferences in various capacities, including chair, technical program committee co-chair, publicity chair, track chair, and technical program committee number. He served as the Co-Director of the IEEE Multimedia Communication Technical Committee Review Board. Dr. Sarhan is the recipient of the 2008 Outstanding Professional of the Year Award from the IEEE Southeastern Michigan Section and the 2009 President's Award for Excellence in Teaching from Wayne State University.

Addressing User Experience, Cost and Scalability Challenges of Cloud Mobile Multimedia Applications

Sujit Dey, Shaoxuan Wang, Yao Liu
Mobile Systems Design Lab, ECE Department, University of California, San Diego
{dey, shaoxuan, yal019}@ece.ucsd.edu

1. Introduction

Clearly, more and more mobile applications, both enterprise and consumer, are migrating to the cloud. The benefits are numerous – from supporting Bring Your Own Device needs in the enterprises, to scaling cost efficiently to user demands using the elasticity of the cloud resources. In this paper, we discuss a new trend emerging – enabling a new class of Cloud Mobile Multimedia (CMM) applications [1][2] which will enable new, rich media experiences through the use of cloud computing, that are not possible otherwise from their mobile devices. However, CMM applications will also bring new sets of challenges for cloud adoption and migration not seen before, most importantly with regards to the quality of user experience, and the cost and scalability of the services. We discuss the CMM specific challenges, and argue the need to address them effectively for mass adoption of CMM applications. We briefly describe some of the solutions we are developing to address these challenges which are showing promising results.

2. Challenges to Enable CMM Applications

We first describe the architecture and data/control flow of typical CMM applications, followed by resulting challenges. Though a CMM application may utilize the native resources of the mobile device, like GPS and sensors, it primarily relies on cloud computing and storage resources. A typical CMM application has a small footprint client on the mobile device, which provides the appropriate user interfaces (touchscreen, voice, gesture, text based) 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 the mobile Core Network, 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, it is clear that a typical CMM application will be highly interactive, with some needing near real-time response times.

In the context of the above CMM architecture, we next describe the primary challenges CMM applications face. Foremost, unlike other cloud applications, CMM applications 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.

User Experience: Response Time and Media Quality. We have implemented several CMM applications, including cloud based mobile video streaming [1], cloud rendering based applications like Cloud Mobile Gaming (CMG) [2][3], and Cloud Mobile Desktop (CMD) [4], and evaluated the impact of wireless networks, both cellular and WiFi, on the user experience of such applications.

We illustrate the challenges using the CMG application. In CMG, 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. Figure 1 shows the uplink delay, downlink delay, and round-trip response time under different network conditions (180 seconds data samples for each). It also shows the overall user experience, including video quality besides response time, as measured by a metric GMOS that was developed and validated in [5]. (GMOS score above 4.0 indicates very good experience, 3.0 - 4.0 indicates acceptable experience, and below 3.0 indicates unacceptable experience). Figure 1 shows significant increase in uplink and downlink delays, and round-trip response time, when the network is congested and/or the user is in poor signal conditions, leading to significant

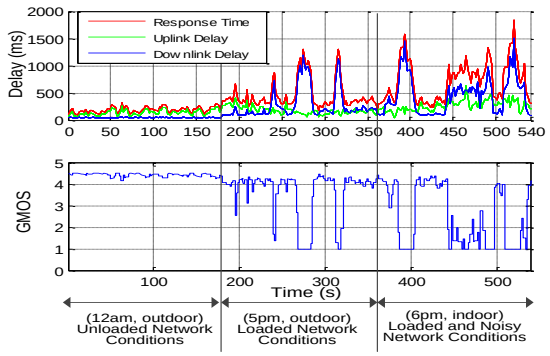


Figure 1: Delays, response time, and user experience of CMG.

adverse impact on user experience. We have reported similar observations for the other CMM applications we have evaluated [1][2][3][4]. We conclude 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.

Cost and Scalability.

Cloud Provider	Computing Price ($/MIPS)	Storage Price ($/GB/sec)	Network Price ($/kb)	Cost for WoW Session ($/hour)
Cloud Provider 1	3.75e-9	3.85e-8	1.50e-8	0.100
Cloud Provider 2	4.41e-9	3.66e-8	1.50e-8	0.112
Cloud Provider 3	4.95e-9	3.85e-8	1.38e-8	0.120

Table 1: Operating cost for CMG using different public clouds.

A top motivation for developing, or migrating to, cloud based applications is to eliminate capital expenses of servers and/or provisioning for peak demands, and instead use public clouds with on-demand pricing models that allow cost efficient scaling to varying user demands. However, our analysis has shown that use of public clouds for computing and bandwidth intensive CMM applications, like cloud based mobile gaming and rendering, can lead to prohibitively high operating cost.

Table 1 shows the cloud pricing structures of three popular cloud service providers (whose names have been withheld to maintain anonymity), including CPU price, storage price, and network bandwidth price. It also shows the cost per hour of a VGA resolution CMG session of the popular game World of Warcraft, (WoW), assuming each session needs 1GB cloud storage space, 600kbps cloud network bandwidth, and up to 5000 MIPS cloud computing capacity. Assuming average playing time of 23 hours/week [10], from Table 1 the monthly operating expense for a CMG provider using public clouds will be at least $10/month per WoW player, which according to typical game subscription prices is too high, and will be more prohibitive to support higher resolution mobile devices. Our analysis using daily usage patterns of gamers show that CMG based on public clouds is not scalable [1][2]. Clearly, there is a need to develop new cloud architectures and techniques to address the cost and scalability challenges faced by CMM applications in using public clouds.

Besides, CMM applications can have high demand on mobile network bandwidth, adversely affecting the capacity of mobile networks and carrier economics, as well as the data bills of mobile users. Hence, techniques will need to be developed to significantly reduce the wireless network bandwidth needed for CMM applications.

	User Experience	Cloud Cost and Scalability	Mobile Bandwidth and Capacity	Device Scaling
a. User Experience Modeling and Monitoring	✓		✓	
b. Network and Device Aware Application Adaptation	✓	✓	✓	✓
c. Mobile Network Cloud	✓		✓	
d. Mobile Cloud Scheduling	✓	✓	✓	

Table 2: Approaches to address CMM challenges.

3. Proposed Approaches to Address User Experience, Cost, and Scalability

In this section, we discuss a set of techniques that we believe can address the CMM challenges of ensuring high user experience, low cloud cost and high scalability, low mobile network bandwidth and high network capacity, and ability to scale to heterogeneous devices and platforms. Table 2 summarizes the approaches we will discuss, and which of the four challenges a specific approach can address. Next, we will briefly discuss each of the above approaches.

User Experience Modeling, Measurement, and Monitoring.

As discussed in Section 2, the user experience associated with CMM applications can be severely affected by wireless network factors. Moreover, if media adaptation techniques are deployed to scale the content generated and streamed from the cloud servers to address network and device constraint, like adapting video bit rate or adapting the richness of graphic rendering, they may improve response time, but adversely affect video/graphics quality. Moreover, any CMM server over-utilization encountered, and characteristics of the mobile device, may add to delay, and hence affect user experience. In [3][5], we developed user experience models for CMM applications, that takes into account the effects of different network, video and rendering parameters, as well as cloud server and mobile device factors, on the response time and visual quality (video and graphics) experienced by the end user. Through subjective testing, the models were validated with high accuracy. Subsequently, a server-client prototype of the CMM user experience model allows real-time quantitative measurement of changing network conditions and consequent user experience score when a CMM

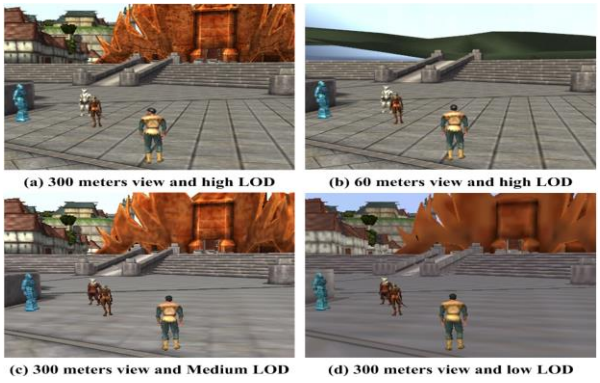


Figure 2: Screenshots of game PlaneShift in different settings of view distance and texture detail.

application runs over any wireless network (like shown by the GMOS score in Section 2), and when any of the proposed approaches to address CMM challenges are developed and applied.

Network and Device Aware Application Adaptation. CMM applications will need to seamlessly support heterogeneous wireless networks and devices to enable an important benefit: ubiquitous and consistently high user experience. To enable this, CMM applications need to be scalable and adaptive to different wireless networks and conditions, and device capabilities like screen resolution. Moreover, CMM applications should be able to scale to different demand levels, reducing bandwidth and cloud costs.

Content scaling techniques, like video transrating and transcoding, can be used to address network bandwidth constraints and device capabilities. Video encoding in the cloud can benefit from knowledge of the CMM content being rendered in the cloud server to perform content aware encoding, leading to significant improvement in overall user experience [6]. However, video rate adaptation techniques may not be always feasible for cloud based applications like CMG. Unlike other delay sensitive applications like real-time video streaming and video conferencing, CMG is also much less tolerant to loss in video quality and frame rate. Hence, we need to develop other techniques that can reduce the network bandwidth needed for CMM applications.

For cloud based mobile rendering applications, we have developed novel adaptive rendering techniques, which can adjust the content complexity and hence the video encoding bit rate needed to varying network conditions, as well as scale significantly bandwidth and computing costs [2]. One way content complexity is adapted is by reducing the number of objects rendered, for example, by changing the rendering view distance. Figure 2(a)(b) shows video frames rendered with two different view distance settings (300m and 60m) in the game PlaneShift (PS) [7]. Note that the resulting video frame of Figure 2(b) has significantly less complexity than the one of Figure 2(a), and hence needs much less encoding bits for the same compression level (video quality) as the frame of Figure 2(a). Also, computing resources needed to render the frame in Figure 2(b) is significantly less, with the above two reductions leading to much less cloud cost. The second rendering adaptation technique is related to the complexity of rendering operations. For example, Figure 2(a)(c)(d) show the results of rendering with progressively reduced texture detail, with the resulting video frames needing progressively less encoding bits for the same video quality level, and needing progressively less computation and hence cloud cost.

We have developed a prototype system for cloud mobile rendering applications like CMG, using live monitoring of user experience, adaptive rendering used to address large fluctuations in network bandwidth and/or need to scale computing needs, and adaptive video encoding to address smaller but more frequent bandwidth fluctuations without noticeable loss in video quality. Experiments conducted with cellular networks

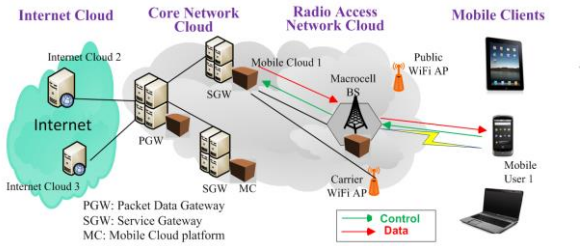


Figure 3: Mobile Network Cloud architecture.

show the ability of the system to (a) produce acceptable user experience (GMOS > 3.0) even for rapidly fluctuating mobile network conditions, (b) reduce bandwidth need by up to 3X for acceptable user experience, (c) reduce computation need by 4X without affecting user experience [2].

Mobile Network Cloud.
As discussed in Section 2, a critical challenge for CMM applications is the network latency and response time 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 mobile networks. To address the above concerns, we propose the development of Mobile Network Clouds, bringing the benefits of cloud computing and storage to the edge of the mobile networks. A Mobile Network Cloud will consist of computing and storage resources supplementing the gateways in the Core Network (CN) and base stations

(BS) in the RAN, and possibly carrier WiFi access points, so that content processing and retrieval can be performed at the edge of the mobile network, as opposed to in Internet Clouds, thereby reducing round trip network latency, as well as reducing congestion in the mobile CN and RAN. Figure 3 shows a LTE network based Mobile Network Cloud, with gateway nodes and BSs supplemented by small-scale Mobile Cloud platforms including computing and storage.

Since there are thousands of base stations and access points, the proposed Mobile Network Cloud is a massively distributed network of much 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. In [8], we investigated the use of Mobile Network Clouds, consisting of small caches in the RAN BSs to improve the latency of video delivery to mobile devices, and the capacity of networks to support concurrent video requests. We concluded that conventional caching techniques are not as effective due to the relatively small RAN cache sizes. We developed RAN aware caching techniques, which make use of video preferences of active users in a cell. Our simulation results show that RAN caching, together with a backhaul scheduling approach, can improve the probability of video requests that can meet initial delay requirements by almost 60%, and the number of concurrent video requests that can be served by up to 100%, as compared to fetching every mobile video requested from Internet Clouds [9].

In [10], we consider a hierarchical Mobile Network Cloud like shown in Figure 3, where the SGW and PGW nodes are also supplemented by caches, which can better support user mobility across cells. Our investigations show that the hierarchical cache architecture can enhance cache hit ratio of video requests by almost 25% compared to caching only in the RAN, without increasing the total size of caching used [10]. When considering mobility of users across cells, up to 50% gain in capacity can be obtained [10].

To support the increasingly popular Adaptive Bit Rate (ABR) streaming, we supplemented RAN caches with limited processing capabilities that can be used for transrating to obtain a requested video with the right bit rate if a higher bit rate version is available in the RAN cache, instead of having to fetch it from the Internet Cloud [11]. Our experimental results show that the use of our proposed ABR aware RAN caching and processing architecture and algorithms can increase the capacity of mobile networks by almost 2X with almost similar user experience obtained by ABR streaming alone [11]. While the above research demonstrates the effectiveness of Mobile Network Cloud for efficient delivery of mobile video, we believe it can be highly effective to address the response time and capacity challenges of other CMM applications, like CMG.

Mobile Cloud Scheduling.

One of the biggest challenges for computing and bandwidth hungry CMM applications is ensuring scalability for large number of simultaneous users. The scalability challenge comes from the prohibitive cloud costs that may be incurred to handle the desired number of simultaneous CMM sessions, as well as the limited capacity of the mobile networks. Hence, a new problem of *mobile cloud scheduling* [12] needs to be addressed, which can simultaneously consider the mobile network resources as well as the cloud computing and storage resources when making resource allocation decisions, such that the number of simultaneous CMM users is maximized, while minimizing the cloud cost. In [12], we have proposed an approach, which also utilizes alternative access networks, like WiFi, when available to a CMM user. Preliminary results show that mobile cloud scheduling can significantly increase the number of simultaneous users, while maximizing aggregate user experience and minimizing cloud cost [12]. In the future, mobile cloud scheduling will need to leverage evolving heterogeneous access networks (HetNet), and Mobile Network Clouds besides Internet Clouds, to maximize the number of concurrent CMM sessions that meet desired user experience levels, while minimizing cloud costs.

4. Conclutions

In this article, we have discussed the challenges that need to be addressed to make CMM applications successful, and suggested new technology directions to address them. The results of our initial research in the proposed directions have been promising. We believe there is significant research that still needs to be performed to make rich CMM applications viable, along the lines we have discussed in this article.

References

[1] S. Dey, “Cloud Mobile Media: Opportunities, Challenges, and Directions”, in Proceedings of IEEE ICNC, Jan. 2012.

[2] S. Wang, S. Dey, “Adaptive Mobile Cloud Computing to Enable Rich Mobile Multimedia Applications”, *IEEE Transactions on Multimedia*, vol. 15, no. 4, pp. 870- 883, Jun. 2013.

[3] Y. Liu, S. Wang, S. Dey, “Modeling, Characterizing, and Enhancing User Experience in Cloud Mobile Rendering”, in Proceedings of IEEE ICNC, Maui, Jan. 2012.

[4] S. Dey, Y. Liu, S. Wang, Y. Lu, “Addressing Response Time of Cloud-based Mobile Applications”, In Proceedings ACM MobileCloud '13, July 2013.

[5] S. Wang, S. Dey, "Cloud Mobile Gaming: Modeling and Measuring User Experience in Mobile Wireless Networks," ACM SIGMOBILE MC2R, vol. 16, issue 1, Jan. 2012.

[6] S. Wang, S. Dey, "Addressing Response Time and Video Quality in Remote Server Based Internet Mobile Gaming," in Proceedings of IEEE WCNC, March 2010.

[7] Planeshift, http://www.planeshift.it/

[8] H.Ahlehagh and S.Dey "Video Caching in Radio Access Network", in *Proceedings of IEEE WCNC*, April 2012.

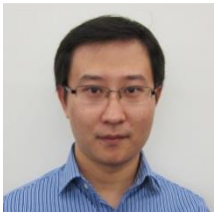
[9] H.Ahlehagh and S.Dey, "Video Aware Scheduling and Caching in the Radio Access Network", To appear in the IEEE/ACM Transactions on Networking.

[10] H. Ahlehagh, S. Dey, "Hierarchical Video Caching in Wireless Cloud: Approaches and Algorithms", in Proceedings of IEEE ICC, June 2012.

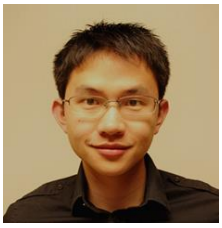
[11] H. Ahlehagh and S. Dey, "Adaptive Bit Rate Capable Video Caching and Scheduling", in Proceedings of IEEE WCNC, April 2013.

[12] S. Wang, Y. Liu, S. Dey, "Wireless Network Aware Cloud Scheduler for Scalable Cloud Mobile Gaming", in Proc. of IEEE ICC, Jun. 2012.

Sujit Dey is a Professor in the Department of Electrical and Computer Engineering, University of California, San Diego, 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 also serves as the Faculty Director of the von Liebig Entrepreneurism Center. He is affiliated with the Qualcomm Institute, and the UCSD Center for Wireless Communications. He served as the Chief Scientist, Mobile Networks, at Allot Communications from 2012-2013. He founded Ortiva Wireless in 2004, where he served as its founding CEO and later as CTO till its acquisition by Allot Communications in 2012. Prior to Ortiva, he served as the Chair of the Advisory Board of Zyray Wireless till its acquisition by Broadcom in 2004. Prior to joining UCSD in 1997, he was a Senior Research Staff Member at the NEC Research Laboratories in Princeton, NJ. He received his PhD. Degree in Computer Science from Duke University in 1991. Dr. Dey 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. He has been the recipient of six IEEE/ACM Best Paper awards, and has chaired multiple IEEE conferences and workshops.



Shaoxuan Wang received his Ph.D degree in Computer Engineering from University of California, San Diego. He is currently a scientist, senior staff engineer in Broadcom Corp. Dr. Wang is the co-inventor of 1 US and 1 international patents, with several others pending.



Yao Liu is currently a PhD student at University of California San Diego. His research interests include mobile multimedia, wireless communication, and mobile cloud computing. His industry experiences include interning at Qualcomm R&D in 2010 and Yahoo Inc. in 2013.

Content Based Image Retrieval on Cloud

Haifeng Lu, Jianfei Cai and Yonggang Wen
School of Computer Engineering, Nanyang Technological University, Singapore
{*hflu, asjfcai, ygwen*}*@ntu.edu.sg*

1. Introduction

Ever since the invention of digital camera, taking photos is no longer the privilege to professionals. One can generate volumes of photos with content of persons, buildings, nature scenes, etc. From this personal image library, how do we find out the photos, which are visually similar? We human can try to scan through the whole library and figure out the similar ones. But when the number of photos is large, such manual way is inefficient. Content based image retrieval (CBIR) is the exact application of computer vision techniques for computers to solve this problem. Interested readers can refer to the latest survey on CBIR [1] to get more information.

As the technology advances, the capability of CBIR systems grows rapidly. In [2], Sivic and Zisserman borrowed the ideas from text retrieval and achieved to search a database with 5k photos. From that, many works in improving the size of manageable database emerged. In [3], Nister and Stewenius introduced tree structure for organizing image features and successfully managed a database with 50k images. In [4], Jégou *et al*. used descriptor compression technique to find the similar images from a 10m-image database within merely 50ms.

Nowadays, almost every smart phone is equipped with a digital camera. People can take photos at any time anywhere. With the services provided by Flickr [5], Facebook [6] and Instgram [7], these photos can be uploaded to internet for archiving and sharing. It is reported that there are 350m photos uploaded to Facebook each day [8]. The above mentioned solutions are based on the assumptions that all the images are stored on one machine and the descriptors of the images can be fully loaded into memory on the machine. However, due to the consideration of data availability, these uploaded images are typically stored across different machines. And taking account of the amount of images, hundreds of millions, even if using compression, fully loading the descriptors into memory on one machine is impossible. Thus, searching from such a large quantity of images imposes new challenges upon existing CBIR systems. Fortunately, we now have cloud computing [9] and MapReduce [10] framework for such big-data applications.

After first introduced by Google [10], the MapReduce quickly becomes the de facto standard framework for processing extremely large datasets. Derived from [10] and [11], Apache Hadoop [12] is an open source implementation of Google's MapReduce and Google File System (GFS). Ever since its first release, Hadoop has attracted many industry partners (including Yahoo! and Facebook) together with academic interests[1]. There are already some initiative works on applying Hadoop on CBIR systems. This survey is dedicated to capture this new trend in CBIR.

The rest of the survey is organized as follows. Section 2 briefly introduced the idea of MapReduce. The recently works on applying Hadoop on CBIR systems are reviewed in Section 3. Finally, Section 4 concludes this survey.

2. MapReduce

Inspired by functional programming, MapReduce is a programming model for processing extremely large datasets by exploiting data independence to do automatic distributed parallelism. In Hadoop, a basic MapReduce job usually consists of a Map function and a Reduce function. The data is distributed in blocks to all the participants of a Hadoop cluster using Hadoop Distributed File System (HDFS).

After a job is launched, Hadoop system automatically generates as many mappers as there are data blocks to process. A mapper can be treated as one processing unit. It reads the data assigned to it iteratively as a (key, value) pair. The process logic is defined in Map function. If necessary, the mapper generates output (key, value) pairs for reducers. These intermediate (key, value) pairs are collected by the system according to the key. Values with the same key are grouped together. These groups are then sent to reducers to generate final results.

Hadoop system transparently handles the partitioning of the input data, scheduling the tasks execution across different machines, managing the communication between these machines and gathering/disseminating (key, value) pairs from/to these machines. The user only needs to define the

[1] Please refer to [13]-[15] for surveys of applying Hadoop on different research problems.

processing logic in Map and Reducer functions. This mechanism greatly simplifies the procedure of writing a distributed program. As long as the problem can be modelled by this MapReduce framework, one can harness the power of cloud computing easily to process the big data.

3. Content based Image Retrieval on Cloud

There are different ways to build a content based image retrieval system [1]. The differences mainly lie in the ways of similarity comparison. In this survey, we focus on applying MapReduce framework on two leading ways of similarity comparison: Full Representation (FR) and Bag-of-Words representation (BoW).

Full representation

In full representation, the approximate nearest neighbors for the features in the query image are obtained by searching all the features extracted from the dataset using an efficient search method (e.g. Kd-Tree (Kdt) [16]). Obviously, when the number of images in the dataset is huge, the storage needed for storing all the features is large. Also, due to RAM limits on one machine, the size of manageable dataset typically cannot go beyond a few million images.

To break through the constraints of one machine, the authors in [17] proposed distributed Kd-Tree (DKdt) for parallelization using MapReduce paradigm. The basic idea is described as follows.

Since a single Kdt for the entire dataset does not fit on one machine, in DKdt, the Kdt is divided into root and leaves, where the root stores the root tree and the leaf machines store the leaf trees (refer to Fig. 1). At query time, the root machine dispatches the search to a subset of the leaf machines. The leaf machines find the nearest neighbors within their subtree and send them back to the root machine. The root machine then sorts and outputs the final images.

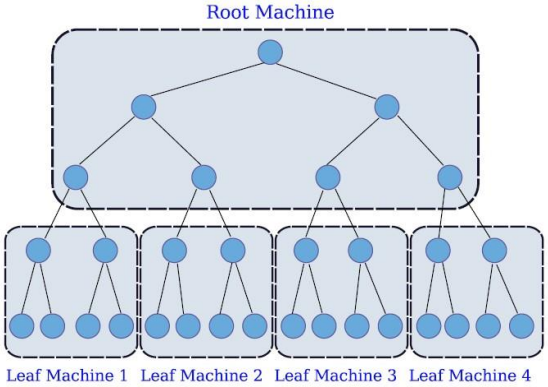


Fig. 1 The architecture of distributed Kd-Trees [17]

In the phase of building DKdt, there are two tasks: 1) build the root tree; 2) build the leaf trees. Note that in Kd-Trees, the root tree performs as a feature "distributor". It directs a feature to certain leaf trees. While in building the leaf trees, each feature in dataset goes through the root tree and finds which leaf machine it belongs to. After that, each leaf machine builds its own Kd-Tree. Clearly, features are independent and leaf machines are also independent. By exploring the independence, MapReduce framework can be applied to build the root tree and the leaf trees, as shown in Algorithms 1 and 2.

```
function MAP (key, val)
// key is the id of a feature, val is the feature
    if key mod skip == 0 then
        // For every skip features, only one
        // feature will be passed to
        // REDUCE to generate root tree.
        output (0, val)
    end if
end function

function REDUCE (key, vals)
// Gather all the features passed from MAP
// Since MAP output key is 0 for all, only
// one REDUCE will be started.
    root = buildTree(vals)
    // generate root tree from the features
    store(root)
    // store the tree information on disk
end function
```

Alg. 1 MapReduce flow for building root tree

```
function MAP (key, val)
// key is the id of a feature, val is the feature
    root = loadRootTree();
    indexID = searchRoot (root, val)
    // a feature traverse the root tree and find
    // its corresponding leaf tree id.
    output(indexID, val)
    // a feature is passed REDUCE based on
    // its leaf tree id
end function

function REDUCE (key, vals)
// key is the leaf tree id, vals is all the
// features belonging to this leaf tree.
// Different leaf trees will be generated by
// by different REDUCE
    leaf = buildTree(vals)
    store(key, leaf);
end function
```

Alg. 2 MapReduce flow for building leaf trees

In the phase of searching similar images, each feature of the query image goes through the root tree first. Instead of finding one leaf tree, as long as the distance is below a certain threshold *s*, the corresponding leaf machines are included in the searching process. After that, these leaf machines find the similar images from their own Kdts. Similar to the phase of building Dkdt, here the features and leaf machines are also independent. The MapReduce flow for searching is described in Alg. 3.

```
function Map (key, val)
// key is the id of a feature, val is the feature
    root = loadRoot()
    indexIDs = searchRoot(root, val, s)
    // search the root tree, find leaf tree ids
    // which have split values smaller than s
    for id in indexIDs do
    // dispatch the feature to the leaf trees
        output(id, val)
    end for
end function

function REDUCE (key, val)
// key is the id of leaf tree, val is the query
// feature
     leaf = loadLeaf(key)
     nn = searchKNN (val)
     matchedImages = match (nn)
     // find the corresponding images from
     // these nearest features
     output (key, matchImages)
end function
```

Alg. 3 MapReduce flow for searching

It has been reported to use such DKdt based CBIR system to manage a dataset with 100m images using 2048 machines. The search time for each query image is just a fraction of a second.

Bag of words

Unlike full representation, in bag of words, each image is represented by a histogram of occurrences of quantized features, and search is done efficiently using certain data structures (e.g. Inverted File [18])

A typical CBIR system using bag of words consists of five components: feature extraction, vocabulary construction, feature quantization, index building and searching. Next, we will introduce how to apply MapReduce framework on each of these components.

Feature extraction

In feature extraction, the local features are extracted from each image in the dataset. As shown in Alg. 4, it is straightforward to apply MapReduce here since the extractions on different images are independent.

We did a feature-extraction test on 5,000 images and each image contains 2,000 – 3,000 SIFT features. By using ten 6-core nodes, it took 6 minutes 18 seconds to finish the extraction. Compared with using 1 node, which took 1 hour 24 minutes, 10 times speedup can be achieved.

Vocabulary construction

There are many ways to construct feature vocabulary. In this survey, we only introduce two algorithms: 1) k-means [19]; 2) extended cluster pruning [20].

```
function MAP (key, val)
// key is image id, val is image
     features = extract(val)
     // extract local features from image
     Store features
end function
```

Alg. 4 MapReduce flow for feature extraction

In k-means, each feature is compared with the centroids to determine which centroid it belongs to. After all the features are assigned to a certain centroid, each centroid is updated by averaging features in that cluster. This process is repeated until the difference between new centroid and old centroid is below a threshold. Considering that the features and centroids are independent, the MapReduce flow for k-means can be implemented as Alg. 5 [19].

In [21], the authors modified this MapReduce flow for k-means by adding a combiner between MAP and REDUCE. This combiner can significantly reduce the amount of data passing from MAP to REDUCE.

Following [21], we did a k-means test on 400,000 features extracted from 100 images. We set *k* to 10,000 and the maximal iteration to 10. By using ten 6-core nodes, it took 20 minutes to finish. We repeated the test on different numbers of nodes used in MapReduce. The results are summarized in Table 1. From this table, we can see that the obtained speedup is almost linear with the number of nodes used in MapReduce.

The drawback for the above implementation is that k-means algorithm is iterative while MapReduce does not naturally support iterative algorithm. The only way to mimic the procedure is to create MapReduce

job repeatedly, which introduces much job-launching overhead. The authors in [20] also noticed this drawback and they applied the MapReduce framework on extended clustering pruning (eCP) algorithm which is iterative.

```
function MAP (key, val)
// key is id of a feature, val is the feature
    find the centroid (id) which is closest to val
    output (id, val)
end function

function REDUCE (key, vals)
// key is the centroid id, vals are the features in
// this center
     centroid = mean(vals)
     // update the centroid by averaging the
     // features
     store(key, centroid)
end function
```

Alg. 5 MapReduce flow for k-means

Table 1. Test results of k-means using different number of nodes.

No. of nodes	1	5	20
Time	3hrs 7mins	38mins	12mins

In eCP, the centroids are picked randomly from all the features in dataset and organized into a tree structure. All the features in dataset then traverse the tree and are assigned to the closest centroid at the bottom of the tree. When all the features are assigned, the tree and the assignment are recorded. Porting eCP algorithm to MapReduce is similar to k-means. The only difference is that in k-means, the closest centroid is found by comparing the distance to all the centroids while in eCP, the comparison is logarithmic complexity on the centroid tree.

Feature quantization, index building and searching
After all the features are assigned to certain centroid, a feature can be approximated by the assigned centroid. By counting the occurrence of quantized features in an image, we can obtain the feature histogram of the image. When all the histograms in the dataset are ready, we start to build the inverted file. For feature quantization, we can use Map function to distribute different features to different machines. For index building, we can use Map function to distribute different items in the histograms to different machines and use Reduce function to gather the output from Map function to integrate into a complete inverted file. The algorithm is similar to Alg. 5. Readers are encouraged to find out more details in [19].

Since the indexed file is rather compact that can be fit into one machine, the searching phase remains unchanged. The MapReduce framework can be used to dispatch different queries to different machine and gather query results. By applying MapReduce framework, [19] is able to manage 1m images on 3 machines and [20] is able to manage 100m images on 108 machines for content based image retrieval.

4. Conclusion
In this survey, we show how MapReduce helps large-scale content based image retrieval systems. From the introduced works, we can see that as long as we can figure out the data independency, the process is able to port to MapReduce framework and harness the power of cloud computing. This idea is also applicable on other multimedia applications such as video transcoding [22] and video streaming [23]

Although MapReduce framework makes it easier to design applications running on the cloud, the overhead incurred in job launching stage cannot be ignored especially when an application requires real-time processing. Such applications may require other frameworks such as Storm [24], which is similar to MapReduce framework but tailored for real-time computing.

References

[1] R. Datta, D. Joshi, J. Li and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys*, vol. 40, no. 2, pp. 1-60, 2008

[2] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proc. of ICCV*, pp. 1470-1477, 2003

[3] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Proc. of CVPR*, pp. 2161-2168, 2006

[4] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in Proc of CVPR, pp. 3304-3311, 2010

[5] Flickr, www.flickr.com

[6] Facebook, http://www.facebook.com

[7] Instgram, http://www.instgram.com

[8] Quora, "How many photos are uploaded to facebook

each day?" http://goo.gl/mNG14i

[9] P. Mell and T. Grance, "The NIST definition of cloud computing (draft)," *NIST special publication*, vol. 800, no. 145, pp. 7, 2011

[10] J. Dean and S. Ghemawat, "MapReduce: simplified data processing on large clusters," *Communcionts of the ACM*, vol. 51, no. 1, pp. 107-113, 2008

[11] S. Ghemawat, H. Gobioff. And S. T. Leung, "The google file system," *ACM SIGOPS Operating System Review*, vol 37, no. 5 pp. 29-43, 2003

[12] Apache Hadoop, http://hadoop.apache.org

[13] J. Dittrich and J. A. Quian é-Ruiz, "Efficient big data process in hadoop MapReduce," *in Proc. of VLDB*, pp. 2014-2015, 2012

[14] X. Kong, "Survey on scientific data processing using hadoop MapReduce in cloud environments," in *Proc. of ICEBEG*, pp. 917-920, 2012

[15] Q. Zou, X. B. Li, W. R. Jiang, Z. Y. Lin, G. L. Li, and K. Chen, "Survey of MapReduce frame operation in bioinformatics," *Briefings in bioinformatics*, 2013

[16] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp91-110, 2004

[17] M. Aly, M. Munich, and P. Perona, "Distributed kd-trees for retrieval from very large image collections, " in *Proc. of BMVC*, 2011

[18] J. Zobel and A. Moffat, "Inverted files for text search engines, " *ACM Computing Surveys*, vol. 38, no. 2, pp. 6, 2006

[19] J. S. Hare, S. Samangooei, and P. H. Lewis, "Practical scalable image analysis and indexing using hadoop," *Multimedia Tools and Applications*, pp. 1-34, 2012

[20] D. Moise, D. Shestakov, G. Gudmundsson, and L. Amsaleg, "Indexing and searching 100 iamges with MapReduce," in *Proc. of ICMR*, pp. 17-24, 2013

[21] W. Z. Zhao, H. F. Ma, and Q. He. "Parallel k-means clustering based on MapReduce." *Cloud Computing*, pp. 674-679. Springer Berlin Heidelberg, 2009

[22] A. Garica, H. Kalva, and B. Furht, "A study of transcoding on cloud environments for video content delivery," in *Proc. of MM workshop on Mobile cloud media computing*, pp.13-18, 2010

[23] M. Kim, S. H. Han, J. J, Jung, H. Lee, and O. Choi, "A robust cloud-based service architecture for multimedia streaming using hadoop," in *Proc. of MUIC*, pp. 365-370, 2014

[24] Q. Anderson, *Storm Real-Time Processing Cookbook*, Packt Publishing Ltd.

Haifeng Lu received his B.Eng degree in computer science and engineering from the University of Science and Technology of China in 2008. Since 2009, he has been a PhD student at School of Computer Engineering, Nanyang Technological University. His research interests include network coding, rateless coding, cloud computering. Currently, he works as project officer at Rapid-Rich Object Search (ROSE) Lab, NTU.

Jianfei Cai (S'98-M'02-SM'07) received his PhD degree from the University of Missouri-Columbia. Currently, he is the Head of Visual & Interactive Computing Division at the School of Computer Engineering, Nanyang Technological University, Singapore. His major research interests include visual information processing and multimedia networking. He has published over 100 technical papers in international conferences and journals. He has been actively participating in program committees of various conferences. He had served as the leading Technical Program Chair for IEEE International Conference on Multimedia & Expo (ICME) 2012 and he currently sits in the steering committee of ICME. He was an invited speaker for the first IEEE Signal Processing Society Summer School on 3D and high definition high contrast video process systems in 2011. He is also an Associate Editor for IEEE T-IP and T-CSVT, and a senior member of IEEE.

Yonggang Wen received the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 2008. He is currently an Assistant Professor with the School of Computer Engineering, Nanyang Technological University, Singapore. Previously, he was with Cisco, San Jose, CA, USA, as a Senior Software Engineer and a System Architect for content networking products. He has also been a Research Intern with Bell Laboratories, Murray Hill, NJ, USA, Sycamore Networks, Chelmsford, MA, USA, and a Technical Advisor to the Chairman at Linear A Networks, Inc., Milpitas, CA, USA. His current research interests include cloud computing, mobile computing, multimedia networks, cyber security, and green ICT.

Utilizing the Cloud for Image-Based Food Recognition

Parisa Pouladzadeh[1], Aslan Bakirov[2], Shervin Shirmohammadi[1,2], and Ahmet Bulut[2]
[1] Distributed and Collaborative Virtual Environments Research (DISCOVER) Lab
University of Ottawa, Canada
{ppouladzadeh | shervin}@discover.uottawa.ca
[2] Data Science Lab, Istanbul Şehir University, Turkey
aslanbakirov@std.sehir.edu.tr {shervinshirmohammadi | ahmetbulut}@sehir.edu.tr

1. Introduction

Obesity and overweightness is now a serious health problem in the world. According to the World Health Organization, the number of obese persons in the world has surpassed one billion, and will increase to 1.5 billion by 2015 [2]. Also, in 2013 the American Medical Association officially classified obesity as a disease that requires medical treatments and has dangerous health consequences [2]. Due to the severity of the situation, much effort is being spent on building tools that can mitigate this problem. One such tool is to provide the ability to automatically or at least semi-automatically measure daily calorie in-take. This is very important in order for a dietician or a doctor to be able to manage and treat a patient's obesity. Among the automated approaches for capturing calorie in-take, the most accurate approach is one that identifies, for each meal or food item, the exact types of food eaten, and their portion sizes [3]. One way to do so in a practical manner and without the need for specialized equipment is to use the person's smartphone to capture a picture of the food, and then use image processing and machine learning to extract the required information (types of food, and their portion sizes) from the food image. The amount of calories can then be calculated from the food type and portion size using readily available nutritional fact tables such as [4].

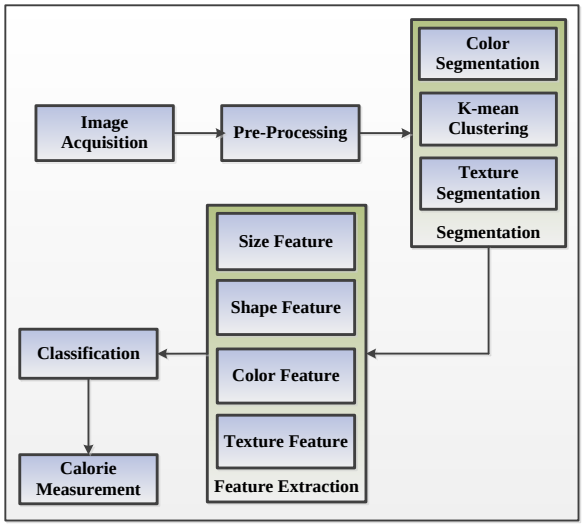


Figure 1. Calorie Measurement System. [5]

This can be done using image segmentation, to identify individual food items and their size, and machine learning, to identify the specific food type for a given portion. We have previously presented a system based on this approach and we have shown the specific steps needed to measure the amount of calories in the food from its image taken with a smartphone [5][6][7], as shown in Figure 1. For the classification part, we use a Support Vector Machine (SVM). The SVM's accuracy depends on how well and with how many images it is trained. As such, the SVM needs to be constantly updated with incoming food images captured by the patients. As the number of food images increases, it becomes computationally longer to train the SVM from scratch on a smartphone or a personal computing device, such that beyond a certain number of images, the smartphone or computing device will no longer have enough resources to re-train the SVM from scratch. In such situations, the SVM can only be incrementally updated. These incremental updates, as opposed to re-training the SVM from scratch, will lead to reduced accuracy in the SVM.

To maintain high accuracy for the SVM, we propose to periodically re-train the SVM from scratch in the cloud. Due to the cloud's resources and scalability, it will be practical to do this re-training even with a large number of food images in the millions. The rest of this paper describes how we propose to do this and what performance improvements to expect. We start this discussion by first explaining how our classification component works.

2. Food Type Classifier

The SVM is trained as shown in Figure 2. We use a large number of existing food portion images to train the SVM. A set of four features (colour, size, shape, texture) for each food portion as well as the name of the food portion will be fed as input to the SVM. The output is the SVM model, which can then be used to recognize food portions from other food images. Since the SVM algorithm operates on numeric attributes, we first need to convert the data into numerical format. To do so, each attribute is scaled linearly to the range of [-1; +1]. After scaling the dataset, we have to choose a

kernel function for creating the model. For the RBF kernel model, the C and γ parameters have to be set, which are adjusted based on the feature values. We have shown that using all four features at the same time will lead to a higher accuracy as opposed to using the features individually, and we have been able to get an average accuracy of 92% for a dataset of 200 images, 100 used for training and 100 used for testing [5].

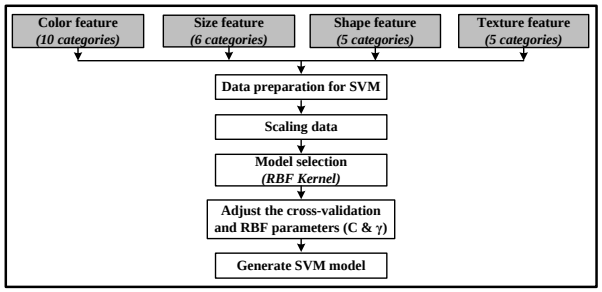


Figure 2. SVM training phase. [5]

We can see that, as the number of training images increase, the computation time needed to train the SVM also increases. To make our system scalable in terms of number of images, which can easily be in the millions, we propose to use cloud computing to train the classifier, as discussed next.

3. Training the Classifier in the Cloud

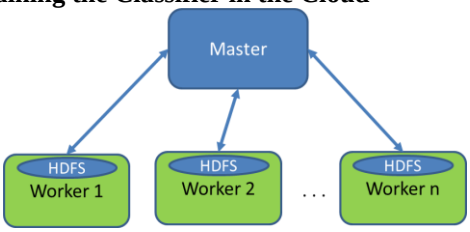


Figure 3. Cloud configuration with each worker running the Hadoop Distributed File System (HDFS).

To train the classifier in the cloud, we can utilize a cluster of computing nodes, as shown in Figure 3, and use a MapReduce [8] execution framework to distribute and send the SVM training job to the workers in parallel. A MapReduce job consists of a map task that does the operation on the data in each worker, and a summary reduce task that does the final computation based on the results gathered from all workers. In our case, the data are the four features (colour, size, shape, texture) of each food image used for training the SVM. In the cloud, assuming each worker has equal computing capacity, the above data is equally distributed between the workers. Each worker then runs its given map task in serial. Once all map tasks are finished, the interim results produced by the map tasks from each worker are co-located around a key identifier at a destination node, where they are to be reduced to a final result by the reduce task.

4. Implementation and Performance Results

For our implementation of the cloud configuration shown in Figure 3, we used four workers each running Apache Hadoop, as an implementation of MapReduce, and each using the Hadoop Distributed File System (HDFS) to store data. HDFS is a distributed, fault-tolerant, and scalable file system that requires a namenode to accept and serve requests. In our implementation, we designated the master node to be the namenode and the workers acting as the datanodes. In order to build the SVM on Hadoop, we used a cascade-SVM implementation [9] and the image features were stored on our HDFS cluster.

We then trained the SVM with 1000, 2000, and 3000 images in the cloud, and used the resulting SVM model from each to test the accuracy of the model using 1000 test images. The results are shown in Table 1, where we can see a very small increase in training time each time we add a large number of images to the training set. This shows that the system will be scalable for a large number of food images. We can also see from Table 1 that the accuracy of the system increases as we increase the number of training images, as expected.

Table 1. SVM training time and model accuracy

Number of images	Training time (sec)	Accuracy of the model
1000	7.1	79.7%
2000	7.9	82.2%
3000	8.6	86.0%

Another interesting question is by how much does the accuracy improve if we train the SVM from scratch compared to if we update the SVM as new images come in? To answer this question, we must first design a method to update the SVM online, which is subject to future work.

5. Conclusions

In this article, we presented a design to utilize the cloud in order to increase the accuracy and training speed of an image-based food classifier system. We showed how the classifier can be applied to food images, and how it can be implemented in the cloud using a MapReduce method and using features from the images. Preliminary results confirm the scalability of the system in the cloud: an important contribution towards the feasibility of food recognition and calorie in-take systems which have to deal with millions of food images. For future work, it will be interesting to see by how much the proposed cloud system increases the accuracy of the classification.

References

[1] World Health Organization, "The World Health Organization warns of the rising threat of heart disease and stroke as overweight and obesity rapidly increase", September 2005, [Online] http://www.who.int/mediacentre/news/releases/2005/pr44/en/

[2] The American Medical Association, "AMA Adopts New Policies on Second Day of Voting at Annual Meeting", June 18 2013, [Online] http://www.ama-assn.org/ama/pub/news/news/2013/2013-06-18-new-ama-policies-annual-meeting.page

[3] R. Steele, "An Overview of the State of the Art of Automated Capture of Dietary Intake Information", Critical Reviews in Food Science and Nutrition, 2013.

[4] Health Canada, "Nutrient Value of Some Common Foods", 2008, [Online] http://www.hc-sc.gc.ca/fn-an/alt_formats/pdf/nutrition/fiche-nutri-data/nvscf-vnqau-eng.pdf

[5] P. Pouladzadeh, S. Shirmohammadi, and T. Arici, "Intelligent SVM Based Food Intake Measurement System", IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications, Milan, Italy, July 15-17 2013.

[6] P. Pouladzadeh, G. Villalobos, R. Almaghrabi, and S. Shirmohammadi, "A Novel SVM Based Food Recognition Method for Calorie Measurement Applications", Proc. International Workshop on Interactive Ambient Intelligence Multimedia Environments, in Proc. IEEE International Conference on Multimedia and Expo, Melbourne, Australia, July 9-13 2012, pp. 495-498.

[7] G. Villalobos, R. Almaghrabi, B. Hariri, and S. Shirmohammadi "A Personal Assistive System for Nutrient Intake Monitoring", Proc. ACM Workshop On Ubiquitous Meta User Interfaces, in Proc. ACM Multimedia, Scottsdale, Arizona, USA, November 28-December 1 2011, pp. 17-22.

[8] M. Bhandarkar, "MapReduce programming with apache Hadoop", IEEE International Symposium on Parallel and Distributed Processing, Atlanta, Georgia, USA, April 19-23 2010.

[9] T. Kraska, A. Talwalkar, J.Duchi, R. Griffith, M. Franklin, and M.I. Jordan, "MLbase: A Distributed Machine Learning System", Proc. Biennial Conference on Innovative Data Systems Research, Asilomar, California, USA, January 6-9 2013.

Parisa Pouladzadeh received her MSc from University of Ottawa in 2012, where her thesis was nominated for a best thesis award. Currently she is a PhD student in the School of Electrical Engineering and Computer Science at the University of Ottawa, working on food recognition systems. Her other research interests include image processing, artificial intelligence and classification.

Aslan Bakirov is a Research Assistant in the Data Science Lab, Istanbul Şehir University, Turkey. His main research interest is in large scale data intensive distributed systems. Aslan holds a BSc from Bogazici University, Turkey, and an MSc from Fatih University, Turkey.

Shervin Shirmohammadi received his Ph.D. degree in Electrical Engineering from the University of Ottawa, Canada, where he is currently a Full Professor at the School of Electrical Engineering and Computer Science. He is Co-Director of both the Distributed and Collaborative Virtual Environment Research Laboratory (DISCOVER Lab), and Multimedia Communications Research Laboratory (MCRLab), conducting research in multimedia systems and networking, specifically in gaming systems and virtual environments, video systems, and multimedia-assisted biomedical engineering. The results of his research have led to more than 200 publications, over a dozen patents and technology transfers to the private sector, and a number of awards and prizes. He is Associate Editor-in-Chief of IEEE Transactions on Instrumentation and Measurement, Associate Editor of ACM Transactions on Multimedia Computing, Communications, and Applications, and was Associate Editor of Springer's Journal of Multimedia Tools and Applications from 2004 to 2012. Dr. Shirmohammadi is a University of Ottawa Gold Medalist, a licensed Professional Engineer in Ontario, a Senior Member of the IEEE, and a Professional Member of the ACM.

Ahmet Bulut received his PhD degree in Computer Science from University of California, Santa Barbara. Between 2005-2007, he worked at Citrix Online, and between 2007-2009 he was a senior researcher at Like.com. Since 2010, he has been Assistant Professor at Istanbul Şehir University, Turkey, and conducts research in service platforms for cloud computing, information and communication technologies for smart cities, intelligent transportation systems, and sensor network applications.

Cloud Gaming: From Concept to Reality

Di Wu and Zheng Xue
Department of Computer Science, Sun Yat-sen University, China
wudi27@mail.sysu.edu.cn, xuezh@mail2.sysu.edu.cn

1. Introduction

With recent advances in cloud computing, the idea of cloud gaming, which enables users to play games in the cloud, is not a concept any more, but becomes a reality. *Cloud gaming* [1] is a new type of online gaming, in which games are stored, synchronized, and rendered in the remote servers and delivered to players using streaming technology.

Unlike traditional PC games, cloud gaming offers many novel features: Firstly, with cloud gaming, players are relieved from expensive hardware investment and constant upgrades. A thin client (e.g., set-top box, laptop, mobile device) with a broadband Internet connection is enough to play any video games; Secondly, cloud gaming allows games to be platform independent and players don't need to worry about the compatibility issues when playing games. It is possible to play games on any operating system (e.g., Mac, Linux, Android) or device (e.g., PC, mobile phone, smart TV); Thirdly, cloud gaming allows users to start playing games instantly, without the need to download and install the game images; Finally, cloud gaming makes copyright protection much easier, as games can only run on remote servers. For game publishers, cloud gaming is an attractive form for digital rights management (DRM).

There have been a number of large-scale cloud gaming systems being developed and deployed in the past few years, such as OnLive [2], Gaikai [3], CloudUnion [4], etc. Among them, GamingAnywhere [5] was the first open-source cloud gaming system. However, there still lacks a comprehensive understanding of the operational cloud gaming systems and it is unclear how to make better design choices to provide good user QoE.

In this paper, we conduct an in-depth measurement study on a leading cloud gaming system in China, namely, CloudUnion [4]. We develop a customized measurement platform to measure the CloudUnion platform from both global and local perspectives. The quantitative results obtained from our measurements shed lights on important performance issues and design alternatives of cloud gaming systems.

2. Measurement Platform

Our measurements of CloudUnion can be divided into two categories: *Active Measurement* and *Passive Measurement*. The active measurement is used to gain a global view of the entire CloudUnion infrastructure and its internal mechanism. The passive measurement is used to obtain a deeper understanding of traffic pattern and gaming experiences from the perspective of game players.

The measurement of CloudUnion is challenging because the CloudUnion's protocol is proprietary. In order to understand the underlying protocol of CloudUnion, we had to collect a large amount of Wireshark traces from multiple gaming sessions and analyze the communications between the client and servers in the cloud platform. Based on our understanding of the CloudUnion's signaling protocols, we developed a customized crawler to automatically query the critical components of the CloudUnion's infrastructure (e.g., the portal server, gateway servers) and retrieve important information about the infrastructure. In our passive measurements, we captured all the traffic exchanged between the gaming client and remote servers in the cloud (e.g., portal server, gateway server, gaming server). To ease packet analysis, we developed our own customized packet analyzer to analyze the various fields and contents in the CloudUnion packets.

3. Measurement Results

As a leading cloud gaming service provider, CloudUnion [4] was the first to launch cloud gaming services in China and its subscribers have exceeded 300,000 as of July 2012.

Architecture of CloudUnion Platform.

Our measurements show that the CloudUnion's infrastructure can be briefly illustrated in Fig. 1.

A portal server is responsible for user registration, authentication and bandwidth test. After a user logs into the system, the portal server will return a list of gateway servers in different data centers, from which the user should manually choose one to start gaming. Normally, a user will choose a data center in a nearby region. After selecting a preferred game, the request will be routed to the gateway server of the selected data center. Upon receiving a user request, the cloud gaming platform will launch a dedicated gaming server to run the game specified in the request and stream the gaming video to the user client. All the user inputs

(keyboard and mouse events) are sent to the gaming server directly, and game frames are encoded with x264 codecs.

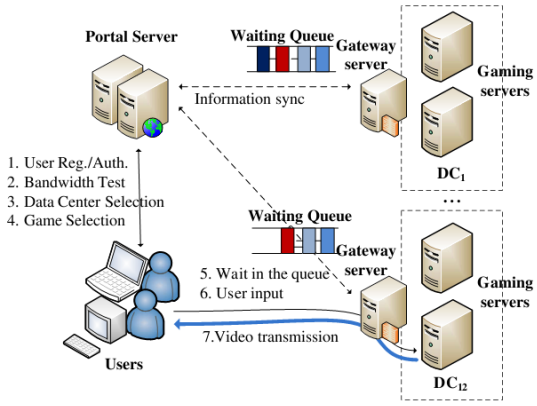


Fig. 1. The architecture of CloudUnion's platform

Queueing Phenomenon in CloudUnion.
When the capacity of a data center cannot satisfy the demand timely, user requests routed to that data center will be held in a waiting queue.

Our trace analysis shows the portal server keeps track of the status of the waiting queue of each data center. By querying the portal server with the CloudUnion's protocol, we can obtain queueing information of each data center, including the number of user requests in the waiting queue, the position of a user request in the queue, the estimated waiting time of a user request, etc. To automate the querying job, we developed a customized crawler to query the portal server every 30 seconds. The crawler was continuously running for 40 days (from Mar 24, 2013 to May 2, 2013) and logged all the queueing information of data centers.

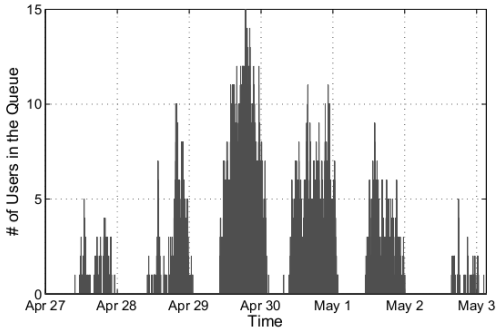


Fig. 2. The number of users in the waiting queue of a data center during the period from Apr 27, 2013 to May 2, 2013

Fig. 2 plots the number of user requests in the waiting queue of a data center over one week (from Apr 27, 2013 to May 2, 2013). We can observe that the queueing phenomenon occurred every day and became more serious during the period of Apr 29 - May 1, 2013, which are the Labor Days in China. The above results show that the current server provisioning strategy used by CloudUnion is not elastic as expected, and cannot provision enough number of gaming servers in a timely manner.

Inter-chunk Arrival Time.
Fig. 3 illustrates the distribution of inter-chunk arrival time for downlink flows. In our paper, the *inter-chunk arrival time* is defined as the interval between two consecutive chunk arrivals. For real-time gaming, the inter-chunk arrival time has significant impacts on the user experience (e.g., response delay).

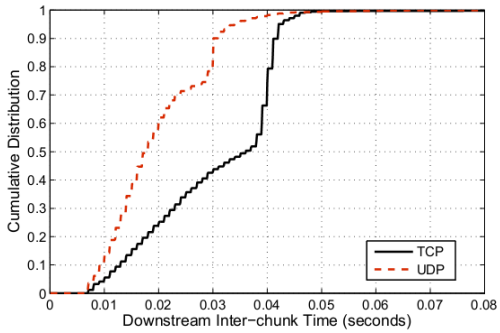


Fig. 3. Downstream Inter-chunk Arrival Time

Fig. 3 shows that UDP can achieve a much lower inter-chunk arrival time compared with TCP. By using UDP, the inter-chunk arrival time is no more than 0.03 second with a probability of 90%.

Resource Utilization.
CloudUnion adopts a kind of thin-client design, with computation-intensive jobs being executed on the remote server. To evaluate how such a cloud-assisted design relieves the load on the local computer, we compare the CPU and RAM usage of two gaming modes: *cloud gaming* and *local gaming*. For the cloud gaming mode, only the CloudUnion client is run on the local computer, while the original game software is executed on the remote cloud server; for the local gaming mode, we directly run the original game software on the local computer. We choose a popular online role-playing game called World of Warcraft (WoW) [6] and run the game under two gaming modes separately. The local computer has the same configuration as that of remote gaming server. By logging the information about CPU and RAM usage of the game software, we are able to monitor the load status on the local computer under two gaming modes.

Under the cloud gaming mode, the CPU usage on the local computer can decrease from around 30% to less than 10% (as shown in Fig. 4). For the memory usage, the cloud gaming mode can reduce the RAM usage on the local computer from 1.3GB to around 100MB. It is because that the local computer only needs to handle video/audio decoding and user input/output, while all

the computation-intensive tasks are offloaded to the remote gaming server. Our results provide a quantitative comparison between two gaming modes and confirm that a thin-client design is not only feasible but also cost-effective for cloud gaming systems.

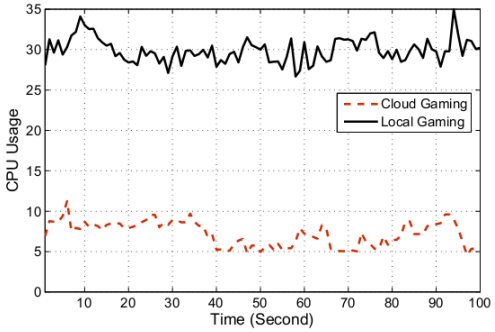


Fig. 4. CPU Usage under cloud gaming and local gaming modes

Video Latency.

Video Latency is defined as the difference between the time the client sends a player's command to the server and the time the generated video frame is decoded and presented on the screen. Video latency has significant impacts on the interactivity of cloud gaming. The direct measurement of video latency is very difficult due to the proprietary nature of the CloudUnion system. Instead, we adopted a method in [7] to measure the video latency.

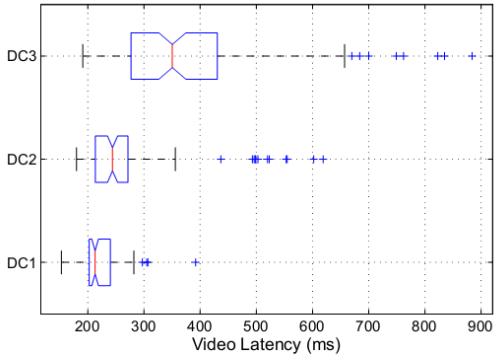


Fig. 5. Video latency when selecting different data centers

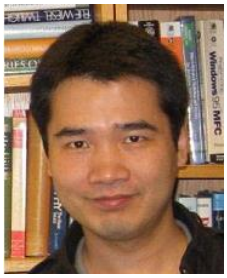
Fig. 5 shows the difference of video latency when selecting different data centers. The video latency exhibits significant spatial diversity, with DC1 being the lowest and DC3 being the highest. The average video latency ranges from 210 ms to 350 ms.

4. Conclusion

In this paper, we conducted a detailed measurement study of a popular cloud gaming system in China, namely, CloudUnion, whose subscribers have surpassed 300,000. Through passive and active measurements, we are able to characterize the CloudUnion system from different angles, including platform architecture, traffic pattern, user behavior, gaming latency, etc. Our observations will be valuable for the design and deployment of future cloud gaming systems.

References

[1] R. Shea, J. Liu, E. C. Ngai, and Y. Cui, "Cloud gaming: Architecture and performance," IEEE Network, vol. 27, no. 4, 2003.
[2] Onlive. Homepage. http://www.onlive.com/.
[3] Gaikai. Homepage. http://www.gaikai.com/.
[4] CloudUnion. Homepage. http://www.yxyun.com/.
[5] C.-Y. Huang, C.-H. Hsu, Y.-C. Chang, and K.-T. Chen, "GamingAnywhere: An open cloud gaming system," in Proceedings of ACM Multimedia Systems 2013, Feb 2013.
[6] WoW. Homepage. http://us.battle.net/wow/en/.
[7] K. Chen, Y. Chang, P. Tseng, C. Huang, and C. Lei, "Measuring the latency of cloud gaming systems," in Proceedings of the 19th ACM international conference on Multimedia. ACM, 2011, pp. 1269–1272.



Di Wu received the B.S. degree from the University of Science and Technology of China in 2000, the M.S. degree from the Institute of Computing Technology, Chinese Academy of Sciences, in 2003, and the Ph.D. degree in Computer Science and Engineering from the Chinese University of Hong Kong in 2007. From 2007 to 2009, he was a postdoctoral researcher in the Department of Computer Science and Engineering, Polytechnic Institute of NYU, advised by Prof. Keith W. Ross. He has been an Associate Professor in the Department of Computer Science, Sun Yat-Sen University, China, since July 2009. He was the winner of IEEE INFOCOM 2009 Best Paper Award, and is a member of the IEEE, the IEEE Computer Society, the ACM, and the Sigma Xi. His research interests include multimedia communication, cloud computing, peer-to-peer networking, Internet measurement, and network security.

Xue Zheng is a graduate student in the Department of Computer Science, Sun Yat-sen University, Guangzhou, China. He received his B.S. degree from Sun Yat-sen University in 2012. His research interests include cloud computing, data center network, content distribution, network measurement, cloud-assisted mobile computing. His advisor is Prof. Di Wu.

Competitive Bandwidth Reservation via Cloud Brokerage for Video Streaming Applications

Xin Jin and Yu-Kwong Kwok
The University of Hong Kong, Hong Kong SAR
{*tojinxin, ykwok*}*@eee.hku.hk*

1. Introduction

The Infrastructure-as-a-Service (IaaS) view of cloud computing is widely adopted by several large cloud providers, which has fundamentally changed the operation of many industries [1-3]. Indeed, large cloud providers such as Amazon Web Services [4], Windows Azure [5**Error! Reference source not found.**], and Google App Engine [6] offer Internet-scale distributed computing facilities, where tenant users can dynamically reserve cloud resources including CPU, memory, and bandwidth so as to satisfy their own service requirements [7].

In such a multi-tenant cloud computing environment, cloud brokers exploit demand correlation among tenants and obtain volume discounts from cloud providers via tenant demand aggregation. Therefore, tenants dynamically procure resources via cloud brokerage services due to lower offered price rates. Therefore, we consider resource procurements from cloud brokers, and tackle the problem of tenant demand competition with a realistic broker pricing policy. In a practical cloud market, resource demands and prices will be cleared at an equilibrium level, where tenant consumers maximize their surplus and cloud brokers optimize the collected revenue given optimal demand responses of tenant consumers.

In this paper, our specific contributions are three-fold. Firstly, we build a general game model to realistically capture broker pricing scheme design. Tenant surplus (i.e., tenant utility minus dollar cost) is realistically formulated to model tenant rationality. Secondly, to relax the impractical assumption of complete information, we propose a dynamic game based bounded rationality to attain Nash equilibrium in a distributed manner by using only local information. Thirdly, we present evaluation results to validate our analytical model and obtain insightful observations.

2. Game Model for Tenant Competition

We consider a cloud system with multiple cloud brokers and a large number of tenant users. Denote by N the number of tenant users in the cloud system. The number of cloud brokers is M. The broker i sells the cloud resources at price rate p_i.

Pricing Model.

The commodity sold in the cloud market is in the units of bandwidth. To model prices offered by cloud broker i, we consider a realistic pricing function where demands affect prices:

$$p_i(\mathbf{d}_i)=\alpha+\beta\cdot\left(\sum_{j=1}^{N} d_{ij}\right)^{\tau},\ \forall i\in\{1,\cdots,M\}, \qquad (1)$$

where d_{ij} is the amount of resources reserved by tenant j from cloud broker i, and $\mathbf{d}_i=[d_{i1},\cdots,d_{ij},\cdots,d_{iN}]^T$ is the vector of all resource demands at broker i. This practically reflects the situation that the price increases with the growth of aggregate demand at one cloud broker due to the limited amount of cloud resources reserved from cloud providers.

Tenant Surplus.

Denote by l_{ij} the network delay due to tenant j's resource procurements from cloud broker i. L represents the maximum experienced network delay in the entire cloud system. Then, the utility of unit bandwidth resource can be modeled as

$$b_{ij}=\ln\left(1+\left(L-l_{ij}\right)\right), \qquad (2)$$

where $L\geq l_{ij}$ and L represents the maximum tolerated delay by tenant consumers. Then, the total utility obtained by tenant user j is $\sum_{i=1}^{M} b_{ij}\cdot d_{ij}$, with the financial cost of $\sum_{i=1}^{M} b_{ij}\cdot p_i(\mathbf{d}_i)$. Therefore the surplus of tenant j can be formulated as follows:

$$\begin{aligned}\pi_j(\mathbf{s}_j) &= \sum_{i=1}^{M} b_{ij}\cdot d_{ij}-\sum_{i=1}^{M} d_{ij}\cdot p_i(\mathbf{d}_i)\\ &=\sum_{i=1}^{M} b_{ij}\cdot d_{ij}-\sum_{i=1}^{M} d_{ij}\cdot\left(\alpha+\beta\cdot\left(\sum_{j=1}^{N} d_{ij}\right)^{\tau}\right),\end{aligned}$$

where $\mathbf{s}_j=[d_{1j},\cdots,d_{ij},\cdots,d_{Mj}]^T$ is a vector of tenant user j's demands from all the cloud brokers.

Static Game and Nash Equilibrium.

Based on the tenant surplus formulation in the above, we can formulate a non-cooperative game among competing tenant users. The players in this game are all the tenant users. The strategy of each player (e.g., tenant user j) is the demand vector of resources reserved from different cloud brokers (i.e., $\mathbf{s}_j$ for tenant j). The payoff of each tenant user j is the surplus earned from the usage of cloud resources (i.e., $\pi_j(\mathbf{s}_j)$). We use Nash equilibrium to solve the game. The Nash equilibrium of a game is a solution concept in which no player can increase his own payoff by unilaterally changing its own strategy. The Nash equilibrium can be obtained by solving the best response function, which is the optimal strategy of one player given the others' strategy choices. That is, the best response function of tenant j can be formulated as:

$$BR_j(\mathbf{S}_{-j})=\text{argmax}_{\mathbf{s}_j}\pi_j(\mathbf{S}), \quad (4)$$

where $\mathbf{S}=[d_{ij}]$, $\forall 1\le i\le M$, *and* $1\le j\le N$ denotes the strategy matrix of all tenant users and $\mathbf{S}_{-j}=[d_{ik}]$ with $i\neq j$ represents the strategy matrix of all tenants except tenant j. To this end, we can obtain the Nash equilibrium by solving the following equation array:

$$\frac{\partial\pi_j(\mathbf{s}_j)}{\partial d_{ij}} = -\beta\cdot\tau\cdot\sum_{i=1}^{M} d_{ij}\cdot\left(\sum_{j=1}^{N} d_{ij}\right)^{\tau-1} = 0. \quad (5)$$

In the following theorem, we investigate the analytical solution of Nash equilibrium for the special case of M=1. That is, $b_{ij}=b_j$ and $d_{ij}=d_j$, $\forall i$.

THEOREM 1 For the special case of M=1, there exists a unique Nash equilibrium given by

$$d_j^*=\left(\frac{b_j-\alpha}{\beta\cdot\tau\cdot Q^{\tau-1}}-\frac{Q}{\tau}\right)^+, \forall 1\le j\le M, \quad (6)$$

where $Q=\dfrac{\sum_{j=1}^{N} b_j-\alpha\cdot N}{\beta\cdot(N+\tau)}$ and $(x)^+=\max(x,0)$.

Proof. From Equation array 5, we get

$$\frac{\partial\pi_j(\mathbf{s}_j)}{\partial d_j} = b_j-\alpha-\beta\cdot\left(\sum_{j=1}^{N} d_j\right)^{\tau} -\beta\cdot\tau\cdot d_j\cdot\left(\sum_{j=1}^{N} d_j\right)^{\tau-1} = 0. \quad (7)$$

Summing up the left side and the right side of the above equations, we have

$$\sum_{j=1}^{N} b_j-\alpha\cdot N-\beta\cdot N\cdot\left(\sum_{j=1}^{N} d_j\right)^{\tau}-\beta\cdot\tau\cdot\left(\sum_{j=1}^{N} d_j\right)^{\tau}=0.$$

Suppose that $Q=\sum_{j=1}^{N} d_j$. We can readily get

$$Q=\left(\frac{\sum_{j=1}^{N} b_j-\alpha\cdot N}{\beta\cdot(N+\tau)}\right)^{1/\tau}. \quad (9)$$

Substitute Q into Equation 7, we obtain the unique Nash equilibrium:

$$d_j=\frac{b_j-\alpha}{\beta\cdot\tau\cdot Q^{\tau-1}}-\frac{Q}{\tau}. \quad (10)$$

However, this is on that condition that

$$d_j=\frac{b_j-\alpha}{\beta\cdot\tau\cdot Q^{\tau-1}}-\frac{Q}{\tau}\ge 0; \quad (11)$$

otherwise, the best response of tenant j is d_j=0. To sum it up, we obtain the unique Nash equilibrium:

$$d_j^*=\max\left(\frac{b_j-\alpha}{\beta\cdot\tau\cdot Q^{\tau-1}}-\frac{Q}{\tau},0\right). \quad (12)$$

Dynamic Game and Stability Analysis.

In a practical cloud system, one tenant user may not be aware of the strategies and surplus of the other tenant users. Therefore, each tenant user has to learn others' strategies and pricing behaviors based on the interaction history. To this end, we propose distributed learning algorithms for dynamic demand adjustments so as to gradually achieve Nash equilibrium for competitive resource procurements. In tenant demand competition, tenant users can adjust the resource demands from different cloud brokers towards the most promising direction (i.e., the direction of marginal profit function). Therefore, the adjustment of the optimal demand level is calculated in a dynamic game for tenant j:

$$d_{ij}(t+1)=d_{ij}(t)+\delta_j\cdot d_{ij}(t)\cdot\frac{\partial\pi_j(\mathbf{s}_j)}{\partial d_{ij}}=\Gamma\left(d_{ij}(t)\right), \ \forall i, \forall j, \quad (13)$$

where $d_{ij}(t)$ is the demand of tenant j from cloud broker i at time slot t and δ_j is the strategy updating step size (i.e., the learning rate) of tenant j. $\Gamma\left(d_{ij}(t)\right)$ is the self-mapping function of the dynamic game. The dynamic

game defined by Equation 13 is proposed under the notion of bounded rationality where the tenant users cannot adapt their strategies to the optimal demand levels immediately.

3. Performance Evaluation

In this section, we present our evaluation results. We consider a cloud system with one cloud broker and two tenant users procuring bandwidth from the broker (i.e., M=1 and N=2). In the pricing model, we use α=0 and β=1. The impact of τ is explored by varying its values. By default, we have τ=1. For tenant surplus, we have the maximum incurred delay L=30000. The impact of network delay is examined by varying l_{ij}.

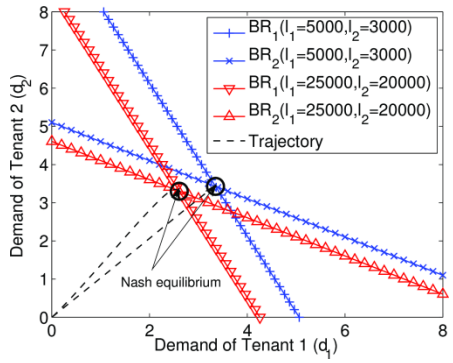


Figure 1: Illustration of Nash equilibrium with two tenant users: best response functions.

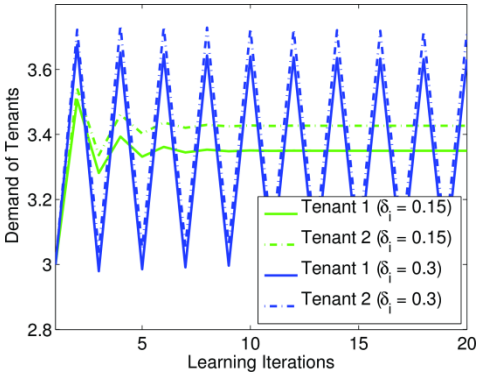


Figure 2: The impact of learning rate on the convergence of the dynamic game.

We first examine Nash equilibrium and the impact of network delay in Fig. 1 for the special case of two tenant users. Here, we investigate the impact of network delay on the equilibrium demand levels. With the decrease of network delay (i.e., better service quality), the corresponding tenant user would like to procure more resources from the cloud broker. On the other hand, the network delay of one tenant user affects the other's procurement of cloud resources. This clearly explains the impact of network delay and the interactions among tenants for resource procurements, when a large number of tenants coexist in the cloud system. We also show the trajectories of the competitive strategies learning of the tenant users in Fig. 1 for the special case of δ_i=0.05 . It shows the convergence of the dynamic game in distributed learning. Fig. 2 shows that, when learning rate is large (e.g., 0.3), the dynamic game may never converge.

4. Related Work

Pricing has been discussed for more than a decade by computer scientists for network resource allocation [8]. Recently, cloud resource pricing is widely adopted as the dominant resource allocation scheme in a cloud computing environment with multi-tenancy. Therefore, there already exist some studies on pricing scheme design and tenant resource procurements. Wang *et al.* [9] examine the importance of cloud resource pricing from the perspective of economics. Due to the coexistence of spot pricing and usage based pricing, Wang *et al.* [10] investigate optimal data center capacity segmentation between both pricing schemes with the objective of total cloud revenue maximization. Niu *et al.* [11, 12] propose a pricing scheme to better leverage the demand correlation among tenant consumers with VoD traffic and argue the necessity of brokers in a free cloud market. Most recently, Xu *et al.* [13, 14] propose centralized schemes so as to maximize the revenue of the cloud provider. Wang *et al.* further discuss optimal resource reservation with multiple purchasing options in IaaS clouds in [15]. While the above studies acknowledge the dominant role of the cloud provider and brokers in pricing, they ignore the competitive cloud resource procurements and its impact on broker revenue and pricing, which is the key problem we aim to solve in this paper.

5. Conclusion

In this paper, we explore the problem of competitive cloud resource procurements in a cloud broker market. We realistically model the pricing scheme of the cloud broker and tenant surplus. We propose a non-cooperative game to model such competitive resource procurements. We then conduct equilibrium analysis under the assumption of perfect information. To relax the assumption of perfect information, we propose the adoption of dynamic game to reach Nash equilibrium in a distributed manner by using local information only. The results revealed insightful observations for practical pricing scheme design. In the future, we would like to extend our model to the more general case of an interrelated market formulated by the cloud provider, brokers, and tenant consumers with strategic interactions.

References

[1] C. Joe-Wong, S. Sen, T. Lan, and M. Chiang,

"Multi-Resource Allocation: Fairness-Efficiency Tradeoffs in a Unifying Framework," in Proc. of INFOCOM, March 2012.

[2] H. Ballani, P. Costa, T. Karagiannis, and A. Rowstron, "Towards Predictable Datacenter Networks," in Proc. of SIGCOMM, August 2011.

[3] Z. Liu, M. Lin, A. Wierman, S. H. Low, and L. L. H. Andrew, "Greening Geographical Load Balancing," in Proc. of SIGMETRICS, June 2011.

[4] Amazon EC2, 2013, http://aws.amazon.com/ec2/.

[5] Windows Azure Pricing Calculator, 2013, http://www.windowsazure.com/en-us/pricing/calculator/.

[6] Google App Engine, 2013, https://appengine.google.com/.

[7] A. Ghodsi, M. Zaharia, B. Hindman, A. Konwinski, S. Shenker, and I. Stoica, "Dominant Resource Fairness: Fair Allocation of Multiple Resource Types," in Proc. of USENIX NSDI, March 2011.

[8] S. Shenker, D. Clark, D. Estrin, and S. Herzog, "Pricing in Computer Networks: Reshaping the Research Agenda," SIGCOMM Computer Communication Review, vol. 26, no. 2, pp. 19–43, April 1996.

[9] H. Wang, Q. Jing, R. Chen, B. He, Z. Qian, and L. Zhou, "Distributed Systems Meet Economics: Pricing in the Cloud," in Proc. of USENIX HotCloud, June 2010.

[10] W. Wang, B. Li, and B. Liang, "Towards Optimal Capacity Segmentation with Hybrid Cloud Pricing," in Proc. of ICDCS, June 2012.

[11] D. Niu, C. Feng, and B. Li, "A Theory of Cloud Bandwidth Pricing for Video-on-Demand Providers," in Proc. of INFOCOM, March 2012.

[12] D. Niu, C. Feng, and B. Li, "Pricing Cloud Bandwidth Reservations under Demand Uncertainty," in Proc. of SIGMETRICS, June 2012.

[13] H. Xu and B. Li, "Maximizing Revenue with Dynamic Cloud Pricing: The Infinite Horizon Case," in Proc. of IEEE ICC, June 2012.

[14] H. Xu and B. Li, "A Study of Pricing for Cloud Resources," ACM SIGMETRICS Performance Evaluation Review, Special Issue on Cloud Computing, March 2013.

[15] W. Wang, B. Li, and B. Liang, "To Reserve or Not to Reserve: Optimal Online Multi-Instance Acquisition in IaaS Clouds," in Proc. of ICAC, June 2013.

Xin Jin received his BEng degree in communication engineering from University of Electronic Science and Technology of China, Chengdu, China, in 2008. He received his Ph.D. degree in Electrical and Electronic Engineering from the University of Hong Kong in 2013. His main research interests are incentive provision, and performance modeling of distributed systems including P2P networks, and cloud computing.

Yu-Kwong Kwok is a Professor in the Electrical and Electronic Engineering Department at the University of Hong Kong (HKU). He received his B.Sc. degree in computer engineering from HKU in 1991, the M.Phil. and Ph.D. degrees in computer science from the Hong Kong University of Science and Technology (HKUST) in 1994 and 1997, respectively. Before joining HKU in August 1998, he was a Visiting Scholar at Purdue University from August 1997 to July 1998. During his sabbatical leave year (from August 2004 to July 2005), he served as a Visiting Associate Professor at the University of Southern California (USC). From 2007 to 2009, he worked as an Associate Professor at the Colorado State University (CSU). He is an Associate Editor for the IEEE Transactions on Parallel and Distributed Systems. He also serves as a member of the Editorial Board for the International Journal of Sensor Networks. From March 2006 to December 2011, he served on the Editorial Board of the Journal of Parallel and Distributed Computing as the Subject Area Editor in Peer-to-Peer (P2P) Computing. He is a Senior Member of the ACM and the IEEE. He is also a member of the IEEE Computer Society and the IEEE Communications Society. He received the Outstanding Young Researcher Award from HKU in November 2004. In January 2010, one of his journal papers was ranked #4 among top ten All-Time Most Cited Papers published in the IEEE Transactions on Parallel and Distributed Systems, based on Scopus and Google Scholar citation counts as of October 2009. In April 2013, he got the Outstanding Reviewer Service Award from the IEEE Computer Society because as of 2013 he was the All-Time Most Prolific Reviewer for the IEEE Transactions on Parallel and Distributed Systems. His recent research endeavors are mainly related to incentive, dependability, and security issues in wireless systems, P2P applications, and clouds.

INDUSTRIAL COLUMN: SPECIAL ISSUE ON MULTIMEDIA COMMUNICATIONS IN FUTURE WIRELESS NETWORKS

Multimedia Communications in Future Wireless Networks

Guest Editor: Farah Kandah, University of Tennessee at Chattanooga, USA

farah-kandah@utc.edu

The emerge of the wireless network technologies and its ability to provide seamless connectivity anywhere any time show a significantly increase in the demand on this promising wireless network. Users start tending to use their mobile devices on daily basis due to their availability, low-cost, and flexibility. Among different types of Internet traffic, multimedia streaming is growing at an exponential rate due to the huge increase of multimedia content on the Internet and the widespread availability of mobile devices. However, providing excellent quality of service and experience are still challenging due to different limitations that are facing future wireless networks, such as the insufficiency of the wireless medium resources, the heterogeneity of the access technology, and the energy constraints.

This special issue of E-Letter focuses on multimedia communications in future wireless networks. It is the great honor of the editorial team to have six leading research groups, from both academia and industry laboratories, to report their solutions in addressing these challenges and share their latest results.

In the first article titled, "*Optimizing HTTP Adaptive Streaming over Mobile Cellular Networks",* A. Beck, S. Benno, and I. Rimac from Bell Labs / Alcatel-Lucent investigated the HTTP Adaptive Streaming (HAS) web based video delivery method and presented two techniques to tackle the HAS streaming challenges occurred due to wireless networks conditions such as interference, fading, handoffs, *etc*. The first technique presented is WiLo, a robust HAS rate determination algorithm (RDA) which suited for wireless networks and low delay applications in providing stable video output for individual HAS clients. The second proposed technique is the Adaptive Guaranteed Bit Rate (AGBR), which aims to maximize the aggregate quality of all HAS streaming and data flows served from the same base station. Their findings were supported through experimental results showing the ability of WiLo rate determination algorithm in providing stable video output, as well as the ability of AGBR scheduler to optimize aggregate utility over all HAS video.

T. Dagiuklas from Hellenic Open University and I. Politis from University of Patras authored the second article, "*Multimedia Optimization over Mobile Clouds*". The authors investigated the multimedia in mobile cloud computing and discussed different challenges that need to be addressed to fully exploit the potential of mobile cloud computing for multimedia services. The first challenge discussed was the computational offloading due to the heterogeneity of access networks and the distance between the mobile device and the cloud. Another challenge discussed that facing the code delivery networks which is used by most existing Multimedia-streaming applications is the server-based load balancing which can be enhanced by the use of software-defined networking. And finally the authors discussed the intelligent mobile management and QoE management that need to be defined and addressed to ensure seamless uninterrupted services due to network heterogeneity.

The third article is contributed by C. Greco *et al.* from Télécom ParisTech, and the title is *"Network Coding for Advanced Video Streaming over Wireless Networks"*. In this work, the authors presented two network coding technique to enhance the video streaming services in wireless networks. The authors presented their Multiple Description Coding (MDC) technique, which provides a graceful degradation in the presence of losses in the stream and the Multi-View Coding (MVC) technique that aims to provide a new and interactive 3D video service to the users. Through experimentations the authors show potential enhancement in video streaming over wireless networks with the use of network coding.

S. Lederer *et al.* from Alpha Adria-Universität Klagenfurt presented the fourth article, *"Adaptive Multimedia Streaming over Information-Centric Networks in Mobile Networks using Multiple Mobile Links"*. In this work the authors presented the usage of Content-Centric Networking (CCN) instead of HTTP in MPEG-DASH and its impact in mobile environments. The authors evaluated the performance of DASH over CCN using multiple mobile links based on real-world bandwidth traces from mobile networks. Their results showed an improvement in the performance with higher average media bitrate compared to experiments using single available link.

The fifth article is "*Sender-Side Adaptation for Video Telephony over Wireless Communication Systems*", from Ma *et al.* at InterDigital Communications, Inc. With the increase in the popularity of video telephony applications developed for smart phones by different vendors the orientation of the video captured and sent by the sending device may not align correctly with the orientation of the receiving device's display which leads to the video orientation misalignment problem. To overcome this challenge the authors presented a sender-side video orientation adaptation method to improve the user experience and the performance in the network-wide system.

The last article of this special issue is from V. Ramamurthi and O. Oyman at Intel Labs, with the title *"HTTP Adaptive Streaming (HAS): QoS-Aware Resource Allocation over LTE"*. The authors discussed the problem of providing good video Quality of Experience (QoE) to large number of users with limited resource in modern wireless networks. To address this issue, the authors presented the Proportional Fair with Barrier Frames (PFBF) algorithm to improve QoE outage based video capacity of the system and the Re-Buffering Aware Gradient (RAGA) algorithm with consideration of the re-buffering constraints. Through simulations, the authors showed significant improvements in QoE through Video-QoE aware radio resource allocation based on simple cross layer feedback.

While this special issue is far from delivering a complete coverage of this exciting research area, we hope that the six invited letters give the audiences an overview of interesting research and current activities in this field, and provide them an opportunity to explore and collaborate in the related fields. Finally, we would like to thank all the authors for their contributions in succeeding this special issue and the E-Letter Board for making this special issue possible.

Farah Kandah received the B.S. and M.S. degrees in Computer Science from the Hashemite University – 2002 and the University of Jordan - 2005, Jordan, respectively, and the Ph.D. degree in Computer Science from North Dakota State University, Fargo, ND, in 2012. He is an assistant professor of the Computer Science and Engineering department of the University of Tennessee at Chattanooga, Chattanooga, TN. His research interest include wireless Ad-hoc and Mesh networks, sensor networks, resource allocation, QoS provisioning, cloud computing and security and privacy in wireless networks. He has multiple publications in highly reputable international conferences and journals. His service includes Co-Chairing the Network and Information Security Symposium in ChinaCom'2012, serving in Technical Program Committee at multiple IEEE international conferences, and serving as a reviewer for many IEEE/ACM prestigious international journals.

Optimizing HTTP Adaptive Streaming over Mobile Cellular Networks

Andre Beck, Steve Benno, Ivica Rimac
Bell Labs / Alcatel-Lucent, USA/Germany
{andre.beck, steven.benno, ivica.rimac}@alcatel-lucent.com

1. Introduction

HTTP Adaptive Streaming (HAS) is a popular web-based video delivery method that relies on client-side algorithms to adapt the video bit rate dynamically to match network conditions. We found, however, that existing HAS implementations are unstable in challenging network conditions, such as those typically found in wireless networks. Fading, interference, roaming, handoffs, and large roundtrip times cause sudden changes in throughput, which cause existing HAS algorithms a tendency to overshoot, undershoot, oscillate, or worst of all, allow the buffer to starve, which severely impacts perceived user experience. We develop two techniques to tackle the above HAS streaming challenge in wireless networks, which can be deployed independently or in combination.

Our first technique is WiLo [1], a robust HAS rate determination algorithm (RDA) that is well suited for wireless networks and low-delay applications. The design of WiLo follows the principle of selecting a quality level that averages out the bandwidth peaks and valleys rather than trying to follow them. The result is an RDA that is robust and stable in dynamic network conditions, and works well even with small buffers for low-latency applications, such as live streaming. WiLo is a client-side only algorithm replacement and does not require any infrastructure or network support; hence, it can be easily adopted in popular HAS clients.

The second technique we propose, which we term AGBR (adaptive guaranteed bit rate), is a novel scheduler in wireless base stations. The goal of AGBR is to maximize the aggregate quality of all HAS and data flows served from the same base station. We formulate this objective as a utility maximization problem that separately takes into account different utility functions for video and data flows, which we use to derive our AGBR algorithm to control the over-the-air throughout. We show that the proposed algorithm can achieve required fairness among video flows as well as automatically adapt video quality with increasing congestion thereby preventing data flow throughput starvation.

2. WiLo Rate Determination Algorithm

There exists a basic tradeoff between average video bit rate and stable video quality. Motivated by the instability of existing clients and the fact that instability reduces quality of experience (QoE) [2], WiLo is designed with the philosophy of providing a stable output and avoiding buffer starvation while providing a high QoE.

To accomplish this, our WiLo RDA measures the bandwidth used for downloading each chunk, $bw_C[n]$, and averages the instantaneous measurements using a 60-second rectangular window to compute the sliding average $\mu_{bw}[n]$ and the standard deviation of the samples in the window, $\sigma_{bw}[n]$. The bandwidth estimate used in the decision logic is $\widehat{bw}[n] = \mu_{bw}[n] - 0.5\sigma_{bw}[n]$. The large sliding window is used to track long-term trends in available bandwidth. The standard deviation is used to make $\widehat{bw}[n]$ more conservative when there are large fluctuations in bandwidth and less conservative when the network is stable.

Buffer Fullness is the amount of video stored in the client's buffer, measured in seconds, waiting to be played. Larger buffers can absorb more network jitter but add latency, which could be a problem for live streams. WiLo increases its rate only if the buffer fullness is above the upper threshold and $\widehat{bw}[n]$ is greater than the next bit rate, and only if both of these conditions are met for a sustained period of time, called the hangover period.

Large sliding windows, however, are slow to respond to sudden changes. To compensate, heuristics and dynamic thresholds are used to make timely quality level changes. To prevent the long term average from getting out-of-sync with instantaneous network conditions, the RDA will not increase its rate if the instantaneous bandwidth is less than the smoothed bandwidth. This prevents the client from increasing its bit rate when the available bandwidth is decreasing, which we demonstrate is a problem with [3].

For a direct comparison between the WiLo RDA and the latest (v5.1) Microsoft Silverlight Smooth Streaming client [3] we ran several experiments in a wireless lab consisting of a commercial end-to-end LTE network (including radio access network and evolved packet core) and laptops with LTE USB modems running both the Microsoft HAS client and our WiLo client. We ran both clients on different laptops side-by-side in our wireless lab so that both clients experience the same wireless conditions. Both clients were competing with 8 FTP flows for a maximum best effort bandwidth of approximately 28 Mbps.

In the experiment in Figure 1, the network starts with no added noise (30 dB SINR) for the first minute, then

noise is added to create 6 dB SINR. The noise is added for 10 seconds and then removed for 10 seconds. The intervals are then increased to 15, 20, and 30 seconds during the course of the session. This scenario tests the client's ability to absorb mid-term fluctuations that are within its buffer size, a common occurrence in wireless networks.

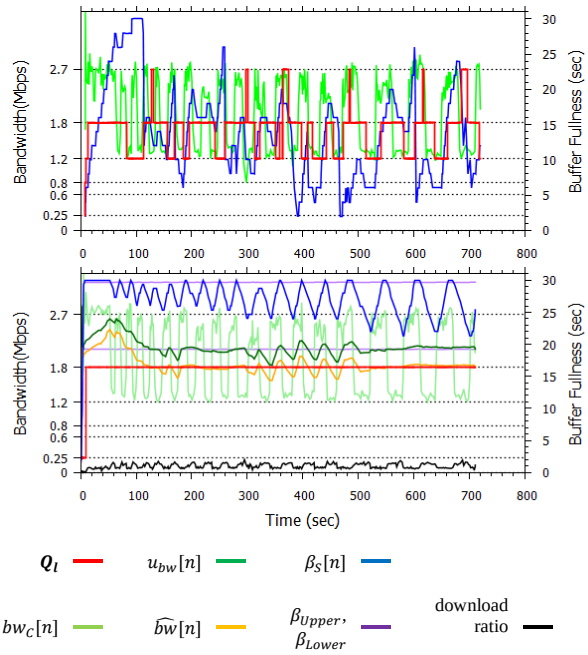


Figure 1. Changes in HAS bandwidth in LTE networks; Silverlight (top), WiLo (bottom).

Figure 1 shows that Silverlight reacts too quickly to increases in bandwidth, which causes severe oscillations in the quality level Q_l. It is out of phase with the bandwidth changes, which contributes to its unstable behavior. Silverlight has an average bit rate of 1.81 Mbps and 52 transitions. In contrast, WiLo is able to absorb the fluctuations in bandwidth and, after the initial transition, provides a constant quality level for the rest of the session. WiLo with a 30 second buffer has an average bit rate of 1.71 Mbps, and only 1 transition, an order of magnitude less than Silverlight.

3. Optimized Base Station Scheduler

The best-effort (BE) scheduler in an eNode-B provides equal radio resources to all active flows, which yields low throughput for clients with poor radio conditions. In contrast, the guaranteed bit rate (GBR) scheduler (as described e.g., in [4]), tries to ensure that a GBR bearer at least achieves the set target throughput as the user experiences varying radio conditions. This is achieved by allocating more radio resources when conditions deteriorate.

When applying GBR to the HAS flows, resources are prioritized for these flows, resulting in much less video quality variation (and higher QoE). GBR allows a video "floor" to be applied in order to ensure that HAS video rates do not drop below a specified threshold. In addition, a maximum bit rate (MBR) can also be set to prevent video throughputs from exceeding the desired rate when resources are available. While in principle GBR and MBR offer a means to control video quality, there are number of caveats that make it impractical. At the time the flow is admitted, a user may have reasonable radio conditions allowing for a good GBR target. However, if the radio conditions worsen, maintaining the GBR can consume an excessive amount of resources leaving the BE data flows starved. Conversely, setting the GBR target too low can make it irrelevant because even the data flows will achieve that throughput. Thus, it is unclear how to pick an appropriate value for GBR that will result in a reasonable amount of resources being devoted to the GBR flow. Network operators shy away from using GBR for this reason.

To mitigate the problem, we propose an adaptive guaranteed bit rate (AGBR) scheduler. AGBR treats the initial GBR setting received at the base station from the core network as a nominal value. It dynamically adapts the actual target between a minimum and a maximum value over time taking into account the radio conditions of the different video flows and the congestion conditions of the air-interface. To adapt the GBR targets, we developed an optimization algorithm at the base station that distributes the resources available for GBR flows with the objective of maximizing a metric that is representative of the sum of the qualities of the different HAS flows. Specifically, we use the alpha-proportional fairness metric [5] with the quality treated as being equal to the throughput, and minimum and maximum values for the throughput. The solution to this optimization problem then drives the GBR setting to the eNode-B scheduler.

In general, the optimization approach guarantees that:

1) When a large amount of resources are available for video flows then all flows' targets are set at the maximum (A_{max}) so that all users enjoy good quality.

2) When the amount of resources starts to diminish, the GBR targets are set differently for different users depending on the radio conditions; the target values will be chosen so as to maximize an aggregate throughput measure with a fairness criterion suited for video and different from the underlying scheduler.

3) When resources are limited so that some users' targets are down to the minimum (A_{min}), then their targets are kept at that level by providing additional resources borrowed from better flows so that as many flows as possible stay above the minimum.

4) If not enough resources are available, each client is scaled down proportionally to its A_{min} value, while the total amount of resources is still chosen such that some resources remain for the data clients to share.

Mathematical details of the above optimization algorithm can be found in [6]. Figure 2 shows the total utility associated with each of the algorithms as a function of the number of video clients. The utility function reflects the value an average user would attach to a video service. The GBR scheduler is better than the BE scheduler for a small number of video clients, but the AGBR scheduler yields the largest total utility for a higher number of clients.

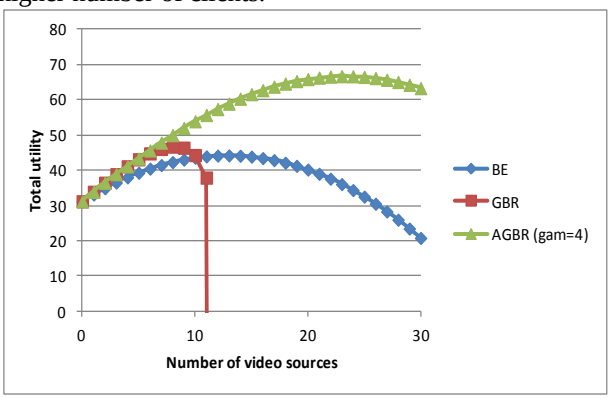


Figure 2: Total utility as a function of the number of clients, in case all SINR values are equal.

4. Conclusion

We have presented two techniques, WiLo and AGBR, which can significantly improve HTTP adaptive streaming over wireless networks. Our experimental results show that under typical conditions found in wireless networks the WiLo rate determination algorithm provides stable video output for individual HAS clients, and the AGBR scheduler enables the base station to optimize aggregate utility over all HAS video and best effort flows.

References

[1] S. Benno, A. Beck, J. Esteban, L. Wu, R. Miller, "WiLo: A Rate Determination Algorithm for HAS Video in Wireless Networks and Low-Delay Applications," to appear in *IEEE Globecom 2013 Workshop - Control Techniques for Efficient Multimedia Delivery*, Dec. 2013.

[2] B. Krogfoss, A. Agrawal, and L. Sofman, "Analytical method for objective scoring of HTTP Adaptive Streaming," in *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, June 2012, pp. 1-6.

[3] (2013, June) Microsoft © SilverlightTM Release History. [Online]. http://www.microsoft.com/getsilverlight/locale/en-us/html/Microsoft%20Silverlight%20Release%20History.htm

[4] M. Andrews, L. Qian, A. Stlyar, "Optimal Utility Based Multi-user Throughput Allocation subject to Throughput Constraints," In Proceedings of IEEE INFOCOM'05, (Vol. 4, pp. 2415-2424).

[5] M. Uchida, J. Kurose, "An Information-Theoretic Characterization of Weighted -Proportional Fairness," in Proceedings of IEEE INFOCOM'09, (pp. 1053-1061).

[6] D. De Vleeschauwer, H. Viswanathan, A. Beck, S. Benno, G. Li, R. Miller, "Optimization of HTTP Adaptive Streaming Over Mobile Cellular Networks", in *Proc. of IEEE INFOCOM*, 2013.

Andre Beck is a member of technical staff at Alcatel-Lucent Bell Labs in Naperville, Illinois. His current research interests include next-generation content delivery networks and distributed carrier clouds. He received B.S. and M.S. degrees in computer science and business administration from the University of Mannheim, Germany.

Steve Benno is a member of technical staff at Alcatel-Lucent Bell Labs in Murray Hill, New Jersey. He obtained his BSEE from Rutgers University College of Engineering in Piscataway, New Jersey, his MSEE from Columbia University School of Engineering and Applied Science in New York, New York, and his Ph.D. from Carnegie Mellon University in Pittsburgh, Pennsylvania. Dr. Benno has worked on a variety of projects including speech processing for wireless networks, mobile phone accessibility to TTY/TDD users, Voice Over IP softphone, and content delivery networks, which is his current area of research.

Ivica Rimac is a senior researcher at Bell Labs, the research organization of Alcatel-Lucent. Ivica joined Bell Labs in 2005 after receiving his Ph.D. in electrical engineering and information technology from Darmstadt University of Technology, Germany. His field of research is computer networking and distributed systems where he has co-authored numerous papers and patents in the areas of content distribution and delivery.

Multimedia optimization over mobile clouds

Tasos Dagiuklas[1] and Ilias Politis[2]
1. *Hellenic Open University, Patras 26335, Greece*
2. *Dept. of Electrical & Computer Engineering, University of Patras, 26500, Greece*

1. Introduction

Recently there is an abundance of applications for mobile devices, expanding from the entertainment and games to news and social networks. This blooming of mobile applications is fueled by the pervasiveness of mobile networking according to which, the user is accessing applications and services seamlessly, regardless of the location.

Nevertheless mobile networking is inherently suffering from limitations such as the scarceness of the wireless medium resources, the heterogeneity of the access technologies and the limited energy supply [1].

To provide rich media services, multimedia computing has emerged as a technology to generate, edit, process, and search media contents, such as images, video, audio, graphics, and so on. Typical types of cloud-based services are the following: Infrastructure as a Service (IaaS), Network as a Service (NaaS), Platform as a Service (PaaS), Identity and Policy Management as a Service (IPMaaS), Data as a Service (DaaS), Software as a Service (SaaS) [2].

Existing cloud-computing technologies are not particularly media/video capable. Any modern server, whether in the cloud or not, can run a video application and even deliver a few video streams. Handling of multiple media flows in terms of encoding, processing and streaming is a much larger problem that stresses computing infrastructure due to large data and bandwidth requirements.

Moreover, the mobile cloud computing poses several challenges, where the cloud serves as a complement to the mobile device, allowing the offloading of data and computation load to mobile clouds that provide resources for storage and processing, away from the mobile device itself [3].

2. Technical challenges of multimedia mobile cloud

There are several key challenges that need to be addressed in order to fully exploit the potential of mobile cloud computing for multimedia services, as shown in Figure 1. Although the integration of the mobile and cloud computing designate the advantages of both technologies (i.e., storage and processing outsourcing, dynamic QoS provisioning, support of seamless mobility), the mobile device limitations and wireless networks unreliability comprise a variety of technical challenges that are still under investigation from both the industry and academic worlds.

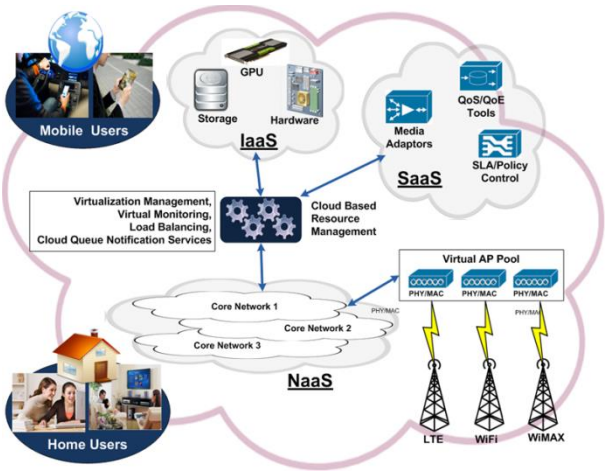


Figure 1. The multimedia mobile cloud paradigm

Computational offloading

A key operation in a mobile cloud would be the offloading and execution of computing-intensive application to the cloud with least energy and time cost. Due to the heterogeneity of the access networks and the physical distance between the mobile device and the cloud, potential problems of increase latency and bandwidth fluctuation may exist. Currently, three offloading methods are under consideration:

- Client –Server communication: the mobile device offloads tasks and applications to a surrogate device using protocols such as the remote procedure calls (RPC), remote method invocation (RMI) and sockets. Although, RPC and RMI are supported by well-defined and stable application programming interfaces (APIs), they require that the services are already installed in the involved devices, which poses restrictions for the mobile cloud when the user is in the vicinity of mobile devices that do not support the particular service.
- Virtualization: in this method the memory image of a virtual machine is transferred without interruption to the destination server. Since the transaction does not require the interruption of the operating system or its applications, it provides a near seamless migration. The advantages of this method include the fact that no code changes are required during the offloading and the isolation of the server due to the virtualization provides a level of security. However, the synthesis of virtual machines is time demanding and power

consuming.

- Mobile agents: this method partitions and distributes the work load to one or more surrogate servers, increasing the performance. On the other hand, there are still grey areas with this methodology in terms of the management of the mobile agents of the surrogates and the lack of fault tolerant mechanisms.

Use of SDN to optimize media delivery

Future networks meet cloud via network function virtualization and software defined networking (SDN). Network virtualization brings many benefits to network operators in terms of reduced equipment cost, network configuration optimization, multi-tenancy support allowing services to be tailored to multiple users [4], [5]. The availability of open APIs for management and data plane control, (e.g. OpenFlow, OpenStack, OpenNaaS etc), provide an additional degree of integration of network functions virtualization. Using SDN, network routing can be configured dynamically on per-flow basis in order to provide QoS differentiation [6]. In this respect, SDN can be used to provide media optimization:

- Content Delivery Networks (CDN): Most of the existing multimedia streaming applications (e.g. live and on-demand streaming) over the Internet rely on CDNs, and load balancing mechanisms. However, such approach on Internet poses several limitations since, only server-based load balancing is possible. This disadvantage can be alleviated by using SDN where load balancing can be considered as a network primitive along with virtualization associated with video processing and streaming functionalities. This is a multi-objective optimization problem requiring the design of cloud-friendly middleware [7], [8]
- Exploit Path Diversity for Video Streaming: Multiple descriptions coding (MDC) is a technique [9] that encodes a source into multiple descriptions supporting multiple quality levels. In the Internet, each description should be sent over different routes. That is feasible through P2P. In SDN, each MDC description can be considered as a different flow and therefore, descriptions can be placed on disjoint paths between the server and the client

Mobility management

Another important issue that the multimedia mobile cloud needs to address is the intelligent mobile management for ensuring seamless uninterrupted services to the mobile user across heterogeneous networks. Currently the mobile devices (smartphones tablets, etc.) are equipped with multiple wireless interfaces (i.e., WiFi, 3G, LTE, etc.). It is important to determine which interface is more suitable for offloading an application or transmit data.

Within the context of mobile clouds, mobility management is executed in the cloud rather than the mobile device. Mobile cloud enables the offering of network as a service, hence allowing functions such as carrier bandwidth aggregation, network sharing infrastructure and optimization of network resources in terms of baseband processing and joint radio resource management across heterogeneous networks. Moreover multi-homing can be realized through the use of a pool of access points (physical and data link layer functionalities) of all the network technologies available in the location of the mobile user. Hence, allowing the concurrent video transfer across heterogeneous links.

QoE management

Another technical challenge that needs to be addressed as media mobile clouds are becoming essential part of the multimedia networking architecture is the optimal QoE managements. Although QoE management for multimedia cloud applications is currently focusing on the provisioning of available resources, or selecting the best candidate network for media delivery, the research for ensuring the user experience and supporting an optimal cloud wide resource allocation scheme is not yet mature [10]. In particular, the cloud is expected to perform resource allocation by considering also the allocation of the rendering process that is performed by the cloud instead of the mobile device. The challenges for QoE management over multimedia clouds are [11]:

- QoE modeling – yet there are no satisfying QoE models for cloud applications.
- QoE monitoring and control – it necessitates the use of deep packet inspection techniques that tries to identify packets associated with premium content. Towards this end data mining techniques are required in order to handle multiple users and services in large scale.
- QoE adaptation – real time adaptation of the user's experience requires the knowledge of the users' physical location, the status of the ongoing session (i.e., traffic, available resources, etc.)
- QoE management signaling – new interfaces need to be defined between several network entities and the mobile node.
- Service level agreements – policy management allows operators to granularly control the availability and QoE of different services.

3. Conclusion

Mobile multimedia cloud computing emerges as the new trend of the future as it comprises the advantages of the mobile multimedia computing and the cloud computing in an effort to provide optimal QoE for real time and on demand video services, or video conferences, etc., to the mobile user. This article, illustrates the key technological challenges such as the computational offloading, the optimized media delivery over SDN and mobility management and the QoE managements that need to be addressed in order for mobile multimedia clouds to provide seamless multimedia services to users anywhere, anytime and using any mobile device.

References

[1] Zhu, Wenwu, Chong Luo, Jianfeng Wang, and Shipeng Li. "Multimedia cloud computing." Signal Processing Magazine, IEEE 28, no. 3 (2011): 59-69.

[2] Wang, Shaoxuan, and Sujit Dey. "Adaptive Mobile Cloud Computing to Enable Rich Mobile Multimedia Applications." (2013): 1-1.

[3] Luo, Hongli, and Mei-Ling Shyu. "Quality of service provision in mobile multimedia- a survey." Human-centric computing and information sciences 1.1 (2011): 1-15.

[4] Open Networking Foundation, Available: http://opennetworking.org

[5] J. Gabrielsson, O. Hubertsson, I. Mas and Robert Skog, "Cloud Computing in Telecommunications", Ericsson Review, 2010

[6] H. E. Egilmez, S. Civanlar, and A. Murat Tekalp, "An Optimization Framework for QoS-Enabled Adaptive Video Streaming Over OpenFlow Networks", *IEEE Trans. On Multimedia,* Vol. 15, No. 3, April 2013

[7] D. Mendyk, "Cloud Computing & Telco Data Centers: Coping with XaaS", Light Reading Insider, Vol. 9, No. 11, November 2009

[8] J. Gabrielsson, O. Hubertsson, I. Mas and Robert Skog, "Cloud Computing in Telecommunications", Ericsson Review, 2010

[9] W. Yao, A. R. Reibman, and S. Lin. "Multiple description coding for video delivery", *Proc. of the IEEE*, Vol. 93, no. 1, 2005

[10] X. Nan, Y. He, and L. Guan, "Optimal resource allocation for multimedia cloud based on queuing model," Multimedia and Signal Processing, 2011.

[11] Hobfeld, Tobias, Raimund Schatz, Martin Varela, and Christian Timmerer. "Challenges of QoE management for cloud applications." Communications Magazine, IEEE 50, no. 4 (2012): 28-36.

Tasos Dagiuklas received the Engineering Degree from the University of Patras-Greece in 1989, the M.Sc. from the University of Manchester-UK in 1991 and the Ph.D. from the University of Essex-UK in 1995, all in Electrical Engineering.

Currently, he is employed as Assistant Professor at the Department of Computer Science, Hellenic Open University, Greece. He is the Leader of the Converged Networks and Services Research Group (http://cones.eap.gr). Past positions include Assistant Professor at TEI of Mesolonghi, Department of Telecommunication Systems and Networks, Greece, Teaching Staff at the University of Aegean, Department of Information and Communications Systems Engineering, Greece and senior posts at INTRACOM and OTE, Greece. Dr Dagiuklas is a Vice-Chair for IEEE MMTC QoE WG and Key Member of IEEE MMTC MSIG and 3DRPC WGs. He has been involved in several EC R&D Research Projects under FP5, FP6 and FP7 research frameworks, in the fields of All-IP network and next generation services. He has served as TPC member to more than 30 international conferences. His research interests include Future Internet architectures and media optimization across heterogeneous networks. He has published more than 120 papers at international journals, conferences and standardization in the above fields.

Dr. Dagiuklas is a Senior Member of IEEE and Technical Chamber of Greece.

Ilias Politis (M'05) received his BSc in Electronic Engineering from the Queen Marry College London in 2000, his MSc in Mobile and Personal Communications from King's College London in 2001 and his PhD in Multimedia Networking from University of Patras Greece in 2009.

Currently he is a Postdoctoral Researcher at the Wireless Telecommunications Lab of the Electrical and Computer Engineering Department at the University of Patras, Greece. He is also an Adjunct Lecturer at the Dept. of Telecommunication Systems and Networks, Technical Educational Institute of Mesolonghi, Greece. He has been involved in FP7-ICT-ROMEO and FP7-ICT-FUTON projects and several national funded research projects. His research interests include immersive multimedia, quality of experience modeling, 3D video and multimedia networking.

Dr. Politis is a member of FITCE and of the Technical Chamber of Greece.

Network Coding for Advanced Video Streaming over Wireless Networks

Claudio Greco, Irina D. Nemoianu, Marco Cagnazzo, Béatrice Pesquet-Popescu*
Institut Mines-Télécom, Télécom ParisTech, CNRS LTCI
{greco,nemoianu,cagnazzo,pesquet}@telecom-paristech.fr

1. Introduction

During the last few years, thanks to the availability of low-cost high-capacity wireless connections, and with the increased computational power of mobile devices, the majority of services provided to mobile users has shifted from text-based to multimedia.

These days, mobile video services are proliferating at an astonishing pace: everything –from movies and TV shows to clips from ordinary users– is available to whoever is connected to the Internet, whether with a laptop, a tablet, or a smartphone. The new frontier of networking lies in this new paradigm: "Video Anywhere at Anytime".

Even though wireless technology has greatly advanced during the past years, a great deal of improvement is still needed in the domain of mobile video networking. Wireless networks still have a significantly lower capacity and a higher expected packet loss rate than wired networks, resulting in generally unreliable time- and location-varying channel conditions. Also, mobile terminals often rely on battery power, which is a scarce resource, and are far less dependable than Internet servers, routers, and clients.

This calls for video streaming techniques that on one hand reduce the bit-rate needed for transmission at a given video quality, and on the other hand are capable to provide a graceful degradation in presence of losses.

2. Network Coding

One of the fundamental assumptions of classical networking is that multi-hop data transfers are handled at intermediate nodes by forwarding the received messages without modifying their content. If more data flows share an intermediate node in their path, this will simply assign each of them a priority (scheduling) and an output link through which to be sent (routing). This view has been challenged with the introduction of the Network Coding (NC) paradigm [1,2], in which each message sent on a node's output link is a linear mixture, in a finite field, of the messages received on the node's input links. Such a strategy of packet mixing (or "coding"), together with means of decoding at the receiver, has been shown to outperform traditional routing by improving the throughput, minimizing the delivery delay, and reducing the impact of losses.

In this letter, we summarize some of our main contributions in the context of NC for high-quality video distribution services over wireless networks. In particular, we present our efforts of integrating NC with advanced video coding techniques such as Multiple Description Coding (MDC), which is used to provide a graceful degradation in the presence of losses in the stream, and Multi-View Coding (MVC), which is used to provide new and interactive 3D video services to the users. We also discuss how the overhead due to the use of NC can be reduced, thus better accommodating the relatively small MTU used in wireless networks.

3. Joint MDC/NC Streaming over Wireless Overlay

Multiple description coding is based on splitting a media stream into a certain number of sub-streams, known as descriptions. Any description can be independently decoded, but the quality increases with the number of descriptions and can be expected to be roughly proportional to the bit-rate sustained by the receiver. MDC is considered a valuable tool to cope with packet losses in wireless networks [3].

In our work [4], we proposed to use MDC jointly with NC to allow instant decoding of received video packets. We first formulated the problem of broadcasting a video stream encoded in multiple descriptions over a wireless network in terms of finding an optimal set of coding coefficients; then, we introduced an objective function that takes into account the effects on the total distortion of decoding a given number of descriptions.

The optimal encoding coefficients are selected via a distributed maximization of the objective function, which the nodes in the network operate based on up-to-date information about the topology. This information is gathered through a wireless overlay construction and maintenance cross-layer protocol we had previously proposed for real-time streaming of MDC video [5,6].

Our experimental results (Fig. 1) show that this approach consistently outperforms the well-known random linear network coding technique [7].

Arguably, this is due to the limits on the generation size imposed by the delay constraints that severely affect the performance of the reference technique.

4. Scheduling for Streaming MDC/MVC Content over Wireless Networks

While the method presented in the previous section benefits from a transmission overlay that supports the exchange of information among nodes, we present here another contribution wherein the optimization is performed without any feedback from the receivers. Namely, we propose a framework for video delivery in wireless networks that combines Expanding Window Network Coding [8], and a novel Rate-Distortion Optimized scheduling algorithm that optimizes the order in which the video packets are included in the coding window. We successfully applied this framework to both MDC [9] and Multi-View streams [10].

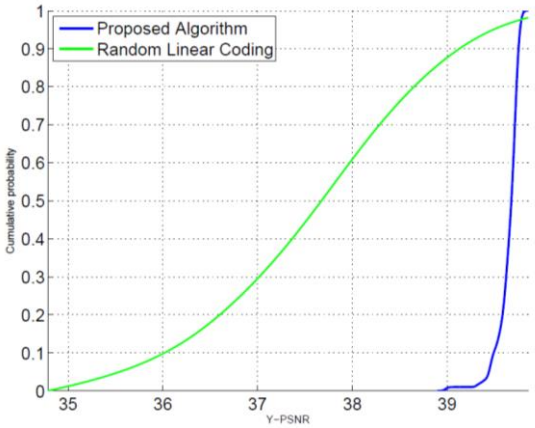


Figure 2 Comparison of PSNR cumulative distribution functions for video sequence "foreman" CIF, 30 fps, 1.8 Mbps

Expanding Window Network Coding (EWNC) is a NC strategy that progressively increases the size of the coding window by using a lower-triangular mixing matrix. The order of inclusion in the coding window is crucial as, by using Gaussian elimination at the receiver side, this method provides instant decodability of data packets.

Since the communication could be abruptly interrupted, due to the neighbors' mobility or disconnection, the scheduling has to be such that the expected video quality is maximized at each sending opportunity. However, imposing the optimal scheduling on all nodes would completely eliminate diversity, thus defeating the purpose of using NC. To address this challenge, we proposed to provide the nodes with a simplified RD model of the stream, so that parts of the video with similar RD properties are considered equivalent for the scheduling purpose (*clustering*). This provides them with a degree of freedom in the choice of the schedule, yet results on each node in a scheduling just slightly less performing than the optimal one.

Applied to both MDC and MVC streams, this strategy has shown to achieve a much higher video quality than both non-NC approaches, and NC approaches based on exact RD optimization or random scheduling (Fig.2).

5. Low-Overhead Network Coding

One of the commonly mentioned drawbacks of network coding is its high overhead. Since the decoder needs to invert the exact mixing matrix in order to be able to reconstruct the original packets, the senders have to include, in each mixed packet, the coefficient used in the combination.

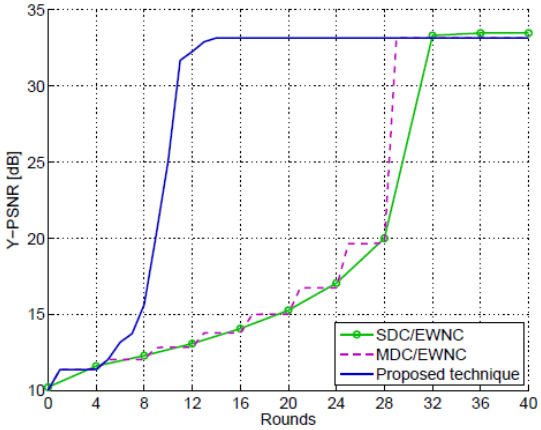


Figure 3 Comparison of average PSNR of the decoded sequences for two sources and 10% packet loss rate

In our recent work [11] we have argued that, using a combination of channel coding and a limited *a priori* knowledge of the sources, it is possible to reconstruct the original messages with high probability even if the combination coefficients are not sent.

This work is placed in the context of Blind Source Separation [12] –a well-established domain of research– but has to deal with the additional constraint that the sources are defined in a finite field, a very challenging addition that so far only few works have addressed [13].

Most BSS techniques rely on entropy minimization as a tool to distinguish between original sources (typically structured, thus carrying low entropy) and linear mixtures (less structured, and therefore with

higher entropy). Unfortunately, in the case of video content, the encoded bit-stream has typically a distribution very close to uniform, *i.e.*, a very high entropy.

Our main idea is to increase the discriminating power of the algorithm by preprocessing the sources with an error-detecting code. The entropy minimization is then performed at the receiver, restraining the estimation of the entropy to the solutions that are admissible in the sense that the reconstructed source is a codeword. This eliminates several solutions that, even if they present low entropy and could be mistakenly identified as sources by merely entropy-based techniques, cannot be admitted as they are not part of the code. Ideally, the code should be such that only the original sources belong to the code, whereas any other possible mixtures do not. This is in practice unfeasible, but we design a code such that the probability of a mixture accidentally being a codeword is very low.

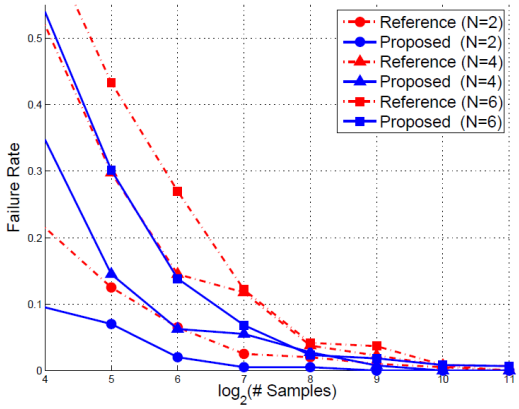


Figure 4 Comparison between the entropy-only BSS method and the proposed technique for finite field GF(4) as a function of the number of samples in the mixture

Our experimental results show that the proposed technique consistently outperforms the entropy-based method, especially in the case of sources with a small number of available samples, which is more critical for the entropy-based methods, making our BSS method more suitable for practical wireless applications, where the number of samples is typically limited by the size of a packet.

6. Conclusion

In this letter we have shown some applications of network coding to multimedia streaming on wireless networks. Our results confirm that NC has the potential for improving the video streaming services on wireless networks, by increasing the throughput and reducing the delay with a slight packet overhead.

References

[1] R. Ahlswede, N. Cai, S.-Y. Li, and R. Yeung, "Network information flow," IEEE Transactions on Information Theory, vol. 46, no. 4, pp. 1204–1216, Jul. 2000.

[2] S.-Y. R. Li, R. W. Yeung, and N. Cai, "Linear network coding," IEEE Transactions on Information Theory, vol. 49, no. 2, pp. 371–381, Feb. 2003.

[3] V. K. Goyal, "Multiple description coding: Compression meets the network," IEEE Signal Processing Magazine, vol. 18, no. 5, pp. 74–93, Sept. 2001.

[4] I.-D. Nemoianu, C. Greco, M. Cagnazzo, and B. Pesquet-Popescu, "A framework for joint multiple description coding and network coding over wireless ad-hoc networks", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Kyoto, Japan, March 2012.

[5] C. Greco and M. Cagnazzo, "A cross-layer protocol for cooperative content delivery over mobile ad-hoc networks", Inderscience International Journal of Communication Networks and Distributed Systems, vol. 7, no. 1–2, pp. 49–63, June 2011.

[6] C. Greco, M. Cagnazzo, and B. Pesquet-Popescu, "Low-latency video streaming with congestion control in mobile ad-hoc networks", IEEE Transactions on Multimedia, vol. 14, no. 4, pp. 1337–1350, Aug. 2012.

[7] P. Chou, Y. Wu, and K. Jain, "Practical network coding," in Allerton Conference on Communication Control and Computing, 2003.

[8] D. Vukobratovic and V. Stankovic, "Unequal error protection random linear coding for multimedia communications," in Proc. of IEEE Workshop on Multimedia Signal Processing, Saint-Malo, France, Oct. 2010.

[9] C. Greco, I.-D. Nemoianu, M. Cagnazzo, and B. Pesquet-Popescu, "A network coding scheduling for multiple description video streaming over wireless networks", Proceedings of the European Signal Proc. Conference, Bucharest, Romania, Aug. 2012.

[10] I.-D. Nemoianu, C. Greco, M. Cagnazzo, and B. Pesquet-Popescu, "Multi-View Video Streaming over Wireless Networks with RD-Optimized Scheduling of Network Coded Packets", Proc. of Visual Communications and Image Processing, San Diego,

CA, USA, Nov. 2012.

[11] I.-D. Nemoianu, C. Greco, M. Castella, B. Pesquet-Popescu, and M. Cagnazzo, "On a practical approach to source separation over finite fields for network coding applications", Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, May 2013.

[12] P. Comon, "Independent component analysis, a new concept?" Signal Processing (Elsevier Science), vol. 36, no. 3, pp. 287–314, Apr. 1994.

[13] A. Yeredor, "Independent component analysis over Galois Fields of prime order," IEEE Transactions on Information Theory, vol. 57, no. 8, pp. 5342–5359, Aug. 2011

Irina Delia Nemoianu (S'11) received her engineering degree in Electronics Telecommunications and Information Technology in 2009, from the "Politehnica" Institute, Bucharest, Romania and her PhD degree in Signal and Image Processing in 2013, from Télécom ParisTech, France. Her research interests include advanced video service, wireless networking, network coding, and source separation in finite fields.

Claudio Greco (M'13) received his his laurea magistrale in Computing Engineering (with honors), equivalent to an M.Sc., from the Federico II University of Naples, Italy in 2007, and his Ph.D. in Signal and Image Processing in 2012, from Télécom ParisTech, France, defending a doctoral thesis on robust broadcast of real-time video over wireless network. He is currently post-doctoral fellow at INRIA Rocquencourt on a shared project with Telecom-ParisTech and the L2S research unit.
His research interests include multiple description video coding, multi-view video coding, mobile ad-hoc networking, cooperative multimedia streaming, cross-layer optimization for multimedia communications, and network coding.

Marco Cagnazzo (SM'11) obtained the Laurea (equivalent to the M.S.) degree in Telecommunications Engineering from Federico II University, Napoli, Italy, in 2002, and the Ph.D. degree in Information and Communication Technology from Federico II University and the University of Nice-Sophia Antipolis, Nice, France in 2005.
Since February 2008 he has been Associate Professor at Télécom ParisTech (Paris), within the Multimedia team. His current research interests are scalable, robust, and distributed video coding, 3D and multi-view video coding, multiple description coding, network coding and video delivery over MANETs. He is the author of more than 80 scientific contributions (peer-reviewed journal articles, conference papers, book chapters).

Beatrice Pesquet-Popescu (F'13) received the engineering degree in Telecommunications from the "Politehnica" Institute in Bucharest in 1995 (highest honours) and the Ph.D. thesis from the Ecole Normale Supérieure de Cachan in 1998. Since Oct. 2000 she is with Télécom ParisTech first as an Associate Professor, and since 2007 as a Full Professor, Head of the Multimedia Group.
In 2013-2014 she serves as a Chair for the Industrial DSC Standing Committee.) and is or was a member of the IVMSP TC, MMSP TC, and IEEE ComSoc TC on Multimedia Communications. She is currently (2012-2013) a member of the IEEE SPS Awards Board. Beatrice Pesquet-Popescu serves as an Editorial Team member for IEEE Signal Processing Magazine, and as an Associate Editor for several other IEEE Transactions. She holds 23 patents and has authored more than 260 book chapters, journal and conference papers in the field. She is a co-editor of the book to appear "Emerging Technologies for 3D Video: Creation, Coding, Transmission, and Rendering", Wiley Eds., 2013. Her current research interests are in source coding, scalable, robust and distributed video compression, multi-view video, network coding, 3DTV and sparse representations.

Adaptive Multimedia Streaming over Information-Centric Networks in Mobile Networks using Multiple Mobile Links

Stefan Lederer, Christopher Mueller, Reinhard Grandl and Christian Timmerer
Multimedia Communication (MMC) Research Group, Institute of Information Technology (ITEC),
Alpen-Adria-Universität Klagenfurt, Klagenfurt, Austria
bitmovin GmbH, Klagenfurt, Austria
{firstname.lastname}@itec.aau.at, {firstname.lastname}@bitmovin.net

1. Introduction

From a content perspective, multimedia is omnipresent in the Internet, e.g., producing 62% of the total Internet traffic in North America's fixed access networks[1]. Today's dominant streaming systems are based on the common approach of leveraging existing, cost-efficient and scalable HTTP-based Internet infrastructures, which are consequently based on the Transmission Control Protocol (TCP) and the Internet Protocol (IP). Especially the adaptive multimedia streaming (AMS) via HTTP is gaining more and more momentum and resulted in the standardization of MPEG-DASH [1], which stands for Dynamic Adaptive Streaming over HTTP. The basic idea of AHS is to split up the media file into segments of equal length, which can be encoded at different resolutions, bitrates, etc. The segments are stored on conventional HTTP Web server and can be accessed through HTTP GET requests from the client. Due to this, the streaming system is pull based and the entire streaming logic is on the client side. This means that the client fully controls the bitrate of the streaming media on a per-segment basis, which has several advantages, e.g., the client knows its bandwidth requirements and capabilities best.

A variety of revolutionary Internet architectures have been proposed in the last decade [1] and some of them seem to overcome current limitations of today's Internet. One of these new Internet architectures is the Information-Centric Network (ICN) approach which moves the focus of traditional end-to-end connections to the content, rather than on addressing its location. One of the most promising representatives of ICN is Content-Centric Networking (CCN)[4], which is also the basis for our work. CCN could eventually replace IP in the future, but it is also possible to deploy it on top of IP. In comparison to IP, where clients set up connections between each other to exchange content, CCN is directly requesting content pieces without any connection setup. This means that a client which wants to consume content simply sends an interest for it to the network, which takes care of routing it to the actual origin as well as responding with the actual data, wherever the content may be located.

ICN and adaptive multimedia streaming have several elements in common, such as the client-initiated pull approach, the content being dealt with in pieces as well as the support of efficient replication and distribution of content pieces within the network. As ICN is a promising candidate for the Future Internet (FI) architecture, it is useful to investigate its suitability in combination with AMS systems and standards like MPEG-DASH as shown in [5][6].

As the mobile multimedia traffic currently growths by more than 100 % per year, reaching a total share of 66% of the total mobile traffic in 2016[1], mobile video streaming and AMS is becoming more and more important, which will be also be the case in next generation networks. In this context, the purpose of this paper is to present the usage of CCN instead of HTTP in MPEG-DASH, and the performance thereof in mobile environments. As CCN is not based on classical host-to-host connections, it is also possible to consume content from different origin nodes as well as over different network links in parallel, which can be seen as an intrinsic error resilience feature w.r.t. the network. This is a useful feature of CCN for adaptive multimedia streaming within mobile environments since most mobile devices are equipped with multiple network links. Thus, we evaluate the performance of DASH over CCN using multiple mobile links based on real-world bandwidth traces from mobile networks.

Figure 5. Bandwidth Traces [7].

2. Real-World Mobile Streaming Evaluation

We evaluated the streaming performance of DASH over CCN in mobile networks, using three different real-world bandwidth traces, which have been recorded during several freeway car drives as depicted in Figure 5. These traces have been the basis for previous evaluations of AMS systems [7][7]. As overall performance indicator for the comparison of the different systems we used the ***average bitrate*** of the transferred media stream. The ***number of quality switches*** has been used to measure the variances during the streaming sessions, where large values potentially decrease the Quality of Experience (QoE)[9]. The

Table 1. Comparison Mobile Bandwidth Traces Evaluations.

Name	Average Bitrate	Average Switches	Average Unsmoothness
Unit	[kpbs]	[# of Switches]	[Seconds]
Microsoft [7]	1522	51	0
Adobe [7]	1239	97	64
Apple [7]	1162	7	0
DASH VLC [7]	1464	166	0
Improved DASH VLC [8]	2341	81	0
DASH SVC [8]	2738	101	0
DASHoverCCN[5]	1326	160	0

smoothness of the streaming session is represented by the ***number of unsmooth seconds***, which should be zero since it is the goal of an AMS to prevent stalls, as they cause lower QoE[10]. For the evaluation, we integrated the CCNx (www.ccnx.org) implementation in the DASH VLC Plugin and used the same DASH content as well as experimental setup as for previous evaluations [8][7] to provide an objective comparison.

Table 1 shows the results of the evaluation of DASH over CCN compared to previous evaluations of proprietary systems, i.e., Microsoft Smooth Streaming (MSS), Apple HTTP Live Streaming (HLS), and Adobe Dynamic HTTP Streaming (ADS). In terms of ***average bitrate*** of the transferred media stream, DASH over CCN can definitely compete with the systems of Apple as well as Adobe, and it is close to MSS and an early version of our DASH VLC Plugin [7]. However, it cannot compete with improved DASH clients presented in [8], which leverages advanced HTTP/1.1 (cf. "Improved DASH VLC") or adopts a scalable video codec (cf. "DASH SVC"). DASH over CCN got a relatively high ***number of average switches,*** which indicates that the used adaptation logic of [8] needs more adjustments to the characteristics of CCN. However, the main goal of AMS was reached as the ***number of unsmooth seconds*** is zero and thus there was no stall in any of the three streaming sessions.

Figure 6 gives a detailed view of the results for one of the bandwidth traces (track three). Since these settings have been also used in previous evaluations of other systems [8][7] (c.f. Table 1), one can compare the results and figures of those papers with this work. DASH over CCN starts at the lowest representation to minimize the startup delay and quickly selects higher bitrate representations as soon as a minimum buffer fill level. As one can see, the adaptation follows the available bandwidth very well, maintaining a high buffer fill level over major parts of the streaming session. However, it is not able to choose higher representations than 2800 kbps (except the wrong adaptation in sec. 135) which is a result of limitations of the CCNx implementation. These include high overhead for packet headers and poor support for pipelining of CCN interests as well as DASH segments, which causes

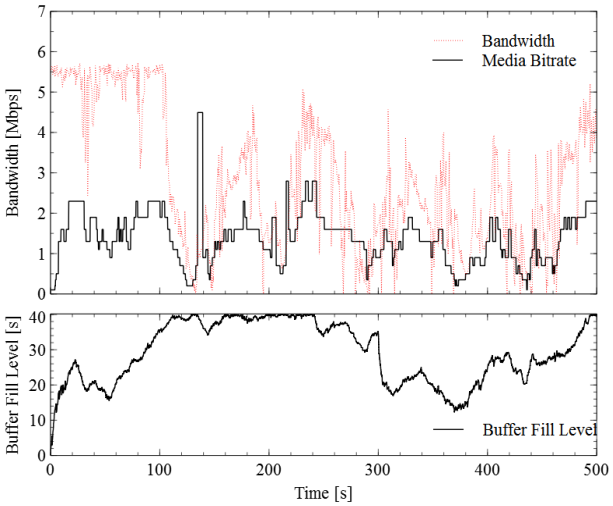


Figure 6: Evaluation Result Track Three [4].

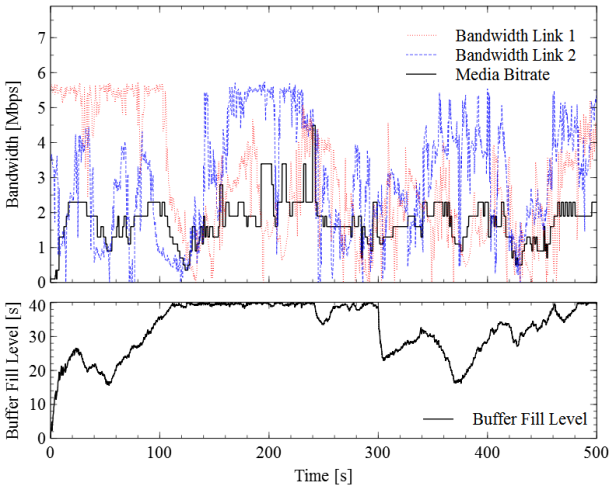


Figure 7: DASH over CCN over multiple mobile links [4].

reduced media throughput in case of higher network delays, as in the given mobile scenario (RTT = 150 ms).

3. Multilink Transmission

CCN is agnostic to the used network link and can switch between multiple links. The decision which link should be used is done by the CCN strategy layer based on the routing information and the performance of each link in the past. As soon as the performance in terms of throughput on link is lower than on an alternative link, the subsequent interests will be sent out on that alternative link. This behavior offers also the possibility to react very fast to link failures, and is useful for client devices offering multiple links like today's mobile phones. The bandwidth of such wireless links heavily depends on the location and the signal strength, which may even lead to outages when out of range of, e.g., a WiFi or 3G base station. This is a major problem for IP-

based traffic, which is bound to the underlying network link. Although there are solutions for this problem, e.g., 802.21, they are not widely deployed. When combining DASH and CCN, it is possible to enable adaptive video streaming handling both: bandwidth and network link changes. That is, CCN handles the network link decision and DASH adapts the video stream to the current bandwidth situation.

We evaluated the performance of DASH over CCN in the presence of multiple network links, which was done by using two of our real-world mobile bandwidth traces (traces two and three) and the evaluation setting described in [4]. Figure 7 shows the actual evaluation results in terms of media bitrate and buffer fill level during the streaming session. The DASH over CCN client constantly chooses the link with the higher bandwidth, which can be seen, e.g., in seconds 110, 200 or 270. Using both links together, the average bitrate was 1710 kbps. Comparing this result with experiments using only one available link, the average bitrate was 1324 kbps for link 1 and 1490 kbps for link 2 respectively. Thus, DASH over CCN using both links achieves ~29 % and ~15 % higher media bitrate than using link 1 and 2 separately. Additionally, the buffer fill level is higher over the whole streaming session than in case of only one network link.

4. Conclusion

This paper proposed and evaluated the combination of CCN with DASH for its usage in mobile environments, especially for devices equipped with multiple network links. Different evaluations based on real-world mobile bandwidth traces showed that DASH over CCN definitely can compete with the systems of major industry players like Adobe and Apple, but cannot compete with optimized DASH clients. Furthermore, the DASH over CCN streaming using multiple links as well as its benefits have been evaluated, showing the seamless switching between the links and resulting in an higher average media bitrate, compared to experiments using only one of the available links. Future work may concentrate on optimizing the underlying CCNx implementation as well as a more efficient utilization of all available links to combine the available bandwidths.

5. Acknowledgements

This work was supported by the EC in the ALICANTE(FP7-ICT-248652) and SocialSensor(FP7-ICT- 287975) projects and performed in the Lakeside Labs research cluster at AAU.

References

[1] Sandvine, "Global Internet Phenomena Report 1H 2013," *Sandvine Intelligent Broadband Networks*, 2013.

[2] Sodagar, "The MPEG-DASH Standard for Multimedia Streaming Over the Internet," in *IEEE MultiMedia*, vol. 18, no. 4, pp. 62–67, 2011.

[3] J. Pan, S. Paul, and R. Jain, "A Survey of the Research on Future Internet Architectures," *IEEE Communications Magazine*, Vol. 49, Issue 7, pp. 26 – 36, 2011.

[4] V. Jacobson, D. Smetters, J. Thornton, M. Plass, N. Briggs and R. Braynard, "Networking named content", *in Proc. of the 5th int. Conf. on Emerging Networking Experiments and Technologies* (CoNEXT '09). ACM, New York, NY, USA, 2009, pp. 1-12.

[5] S. Lederer, C. Mueler, B. Rainer, C. Timmerer, and H. Hellwagner, "Adaptive Streaming over Content Centric Networks in Mobile Networks using Multiple Links", in *Proceedings of the IEEE International Conference on Communication (ICC)*, Budapest, Hungary, June, 2013.

[6] R. Grandl, K. Su and C. Westphal, "On the Interaction of Adaptive Video Streaming with Content-Centric Networking", in *Proceedings of the 20th Packet Video Workshop 2013*, San Jose, USA, December, 2013.

[7] C. Mueller, S. Lederer and C. Timmerer, "An Evaluation of Dynamic Adaptive Streaming over HTTP in Vehicular Environments", *in Proc. of the 4th Workshop on Mobile Video (MoVid12)*, Feb. 2012.

[8] C. Mueller, D. Renzi, S. Lederer, S. Battista and C. Timmerer, "Using Scalable Video Coding for Dynamic Adaptive Streaming over HTTP in Mobile Environments", *in Proc. of the 20th European Signal Processing Conf. 2012*, Bucharest, Romania, August 27-31, 2012.

[9] P. Ni, R. Eg, A. Eichhorn, C. Griwodz and P. Halvorsen, "Spatial Flicker Effect in Video Scaling", *in Proc. of the Third Int. Workshop on Quality of Multimedia Experience (QOMEX'11)*, Mechelen, Belgium, Sept. 2011, pp. 55-60.

[10] T. Hossfeld, M. Seufert, M. Hirth, T. Zinner, T. Phuoc and R. Schatz, "Quantification of YouTube QoE via Crowdsourcing", *in Proc. of IEEE Int. Symp. on Multimedia (ISM) 2011*, 2011, pp.494-499.

Stefan Lederer is assistant professor at the Institute of Information Technology (ITEC), Multimedia Communication and head of business at bitmovin GmbH. He received his M.Sc. (Dipl.-Ing.) in Computer Science in Mar'12 and his M.Sc. (Mag.) in Business Administration in Jul'13, both from the Alpen-Adria-Universität (AAU) Klagenfurt. His research topics include transport of modern/rich media, multimedia adaptation, QoS/QoE as well as future internet architectures, where he published more than 15 papers. He participated in several EC-funded Projects (ALICANTE, SocialSensor) and in the MPEG-DASH standardization, where he contributed several open source tools (DASHEncoder, DASHoverCCN VLC plugin) and datasets.

Christopher Müller is assistant professor at the Institute of Information Technology (ITEC), Multimedia Communication Group and head of research at bitmovin GmbH. He received his M.Sc. (Dipl.-Ing.) from the AAU Klagenfurt. His research interests are multimedia streaming, networking, and multimedia adaptation; where he published more than 20 papers. He gained practical expertise in various companies (Infineon, Dolby Laboratories Inc. LA, etc.) and participated in the MPEGDASH standardization, contributed several open source tools (VLC plugin, libdash) and participated in several EC-funded projects (ALICANTE, SocialSensor).

Reinhard Grandl is a researcher scientist at bitmovin GmbH, focusing on multimedia adaptation, future internet architectures and streaming. He gained his knowledge from his university background at the AAU Klagenfurt, in the Institute of Networked and Embedded Systems, as well as research positions in Europe and the USA. He is currently working towards his Ph.D. in computer science.

Christian Timmerer is an assistant professor in the Institute of Information Technology (ITEC) of the AAU Klagenfurt and head of research at bitmovin GmbH. His research interests include immersive multimedia communication, streaming, adaptation, and Quality of Experience. He was the general chair of WIAMIS'08, ISWM'09, EUMOB'09, AVSTP2P'10, WoMAN'11 and has participated in several ECfunded projects, notably DANAE, ENTHRONE, P2P-Next, ALICANTE, and SocialSensor. He also participated in ISO/MPEG work for several years, notably in the area of MPEG-21, MPEG-M, MPEG-V, and DASH/MMT. He received his PhD in 2006 from the Alpen-Adria-Universität Klagenfurt. Publications and MPEG contributions can be found under research.timmerer.com, follow him on twitter.com/timse7, and subscribe to his blog blog.timmerer.com

Sender-Side Adaptation for Video Telephony over Wireless Communication Systems

Liangping Ma, Yong He, Gregory Sternberg, Yan Ye, and Yuriy Reznik
InterDigital Communications, Inc. USA
{*liangping.ma, yong.he, gregory.sternberg, yan.ye, yuriy.reznik*}*@interdigital.com*

1. Introduction

Mobile video telephony is gaining significant traction due to the availability of highly efficient video compression technologies such as H.264/AVC [1] and HEVC [2] and the availability of high-capacity wireless access networks such as LTE/LTE-Advanced [3]. This is evidenced by the increasing popularity of video telephony applications developed for smart phones such as iPhones and Android phones. Compared to the traditional audio only communication, video telephony provides much richer content and better user experience.

However, if the video sender and the video receiver do not coordinate well, mismatches may occur, resulting in poor user experience and/or inefficient use of the network resource. The mismatches may be in video orientation, video aspect ratio, or video resolution. The video orientation mismatch occurs when the orientation of the transmitted video does not align correctly with the orientation of the display at the receiver. For example, the transmitted video is vertical, whereas the video display at the receiver is horizontal. This mismatch could be resolved by manually rotating the receiver until it aligns with the sent video at the cost of degraded user experience. The other two mismatches cannot be resolved by rotating the receiver. The video aspect ratio occurs if the aspect ratio of the transmitted video is different from that of the display at the receiver, even if the video orientations match. For example, the transmitted video is generated from a smart phone iPhone 4S (960×640) with an aspect ratio 960:640=3:2, whereas the aspect ratio of the display at the receiver (Samsung Galaxy S III) is 1280:720=16:9. The video resolution mismatch occurs when the resolution of the transmitted video is different from that of the display at the receiver. For example, the transmitted video has a resolution of 1080P (1920×1080), whereas the display at the receiver has a resolution of 720P (1280×720).

Desired solutions to these mismatch problems should be standard based, considering the heterogeneity of the mobile devices. The 3GPP multimedia telephony services for IMS (MTSI) is such an effort intended to resolve the video orientation mismatch without user intervention (i.e., manually rotating the receiver device). MTSI mandates that the sender signals the orientation of the image captured on the sender side to the receiver for appropriate rendering and projection on the screen [4]. The rendering and displaying could include cropping or rotating the video. However, the MTSI method, where the receiver adapts to the orientation of the sender, may not fully resolve the video orientation mismatch problem, as illustrated in Figure 1.With the knowledge of the video orientation of the transmitted video, the receiver can adapt to the captured video orientation by either (a) cropping and scaling up or (b) scaling down the received video to fit its own display. In Figure 1(a), portions of the image are lost, and in Figure 1(b), the video is down sampled and there are black bars around the displayed video. Both of them lead to sub-optimal user experience. Additionally, we note that in the examples shown in Figure 1, the MTSI approach is inefficient, because not all the video data delivered across the communication network are fully utilized by the receiver: either part of the video is thrown away or the whole video is down sampled. Figure 1 (c) and Figure 1 (d) show the inefficiency of the MTSI method for the case of aspect ratio mismatch and the case of resolution mismatch.

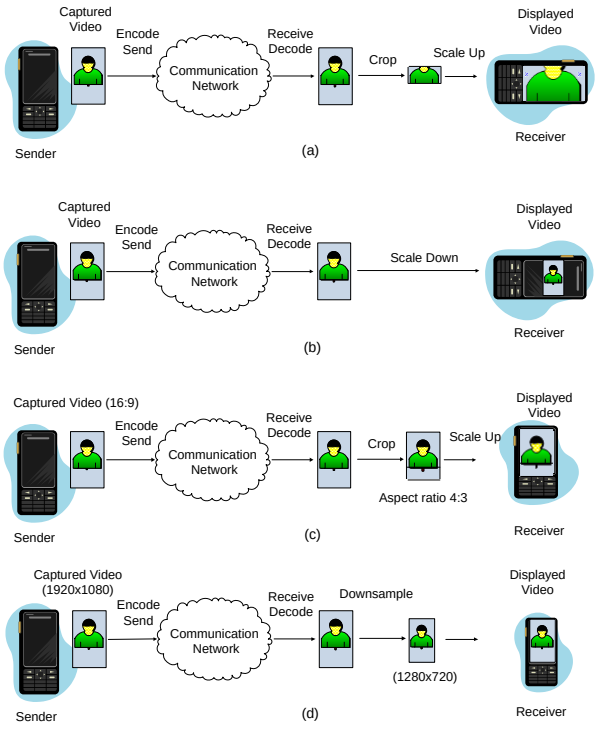


Figure 1. The MTSI method (receiver-side adaptation) leads to undesired user experience and/or inefficient use of the network resource.

In this paper, we propose a sender-side adaptation method which can solve all of the three aforementioned mismatch problems, resulting in better user experience, more efficient use of the network resources, and improved network-wide system performance.

2. Sender-Side Adaptation

The basic idea of our proposed method is to adapt the video processing and/or video capturing on the sender side to the display of the receiver. With the proposed method, every bit of video data delivered across the wireless communication system is fully utilized by the receiver.

In our proposed method, the receiver informs the sender of its desired video display orientation, the aspect ratio, and/or the width and height of the video to be displayed. Note that, by providing the desired width and height, the receiver also provides the desired aspect ratio. After obtaining such information, the video sender can use various adaptation techniques and we consider two of them here. In the first technique, the video sender crops the captured video according to the display orientation, the aspect ratio, and/or the resolution of the receiver, and encodes and transmits the cropped video as illustrated in Figure 2. Such cropping has the benefit of potentially saving a significant amount of network resource, which usually is precious in a wireless communication system. As an example, consider the scenario in Figure 3 (a). Let the image length be L pixels, and the width be W. Then, instead of sending encoded bits corresponding to LW raw pixels per image, we only need to send encoded bits corresponding to $W\big((W/L)W\big) = W^3/L$ raw pixels per image. Assuming the same video encoding efficiency, this represents a reduction of $\alpha = 1 - (W/L)^2$ in the encoded bit rate. Take the 1080P (1920×1080) resolution as an example, the reduction α is 68.36%. Alternatively, we can maintain the same encoded bit rate (thereby keeping the same traffic load to the communication system) during the video encoding process of the cropped images, which can significantly improve the objective video quality, generally resulting in better user experience.

In the second technique, the video sender adapts its video capturing to the display orientation, aspect ratio, or the resolution of the receiver. During video capturing, a subset of the image sensors is selected according to the orientation, aspect ratio, and/or resolution of the display of the receiver. This is illustrated in Figure 4. It is possible that video adaptively captured as such has the same resolution as the display at the receiver, since in practice the resolution of the image sensor array may be much higher than that of the video to be captured. For example, the Nokia Lumia 1020 smart phone features a sensor array of 41 Megapixels, much higher than the 1080P (1920×1080 = 2.07 Megapixels) resolution.

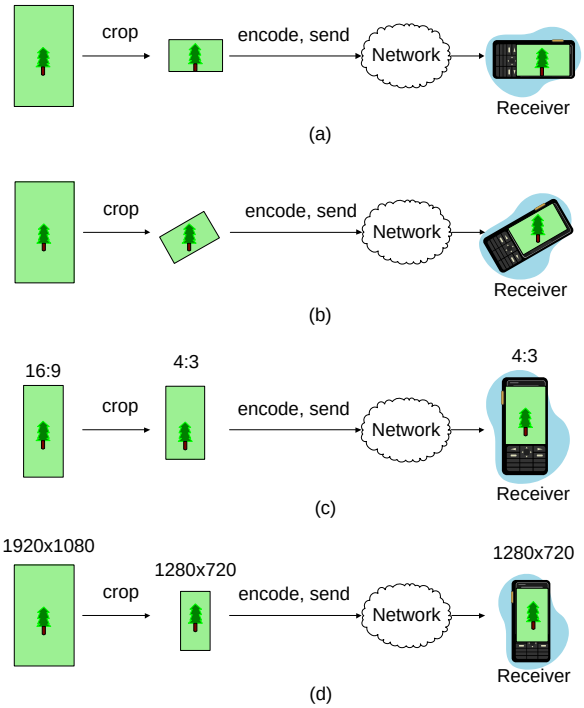


Figure 2 Cropping on the sender side.

To enable the aforementioned sender-side video adaptation techniques, the receiver can provide the sender with the following information: the height and width of the desired video pictures, and the *up direction* (as preferred by a user) in the video. The up direction is not necessarily opposite to the direction of gravity, e.g., when a phone is placed flat on a horizontal table. The up direction can be represented by an angle relative to the width of the display (denoted by A) as shown in Figure 5. After receiving the information, the video sender can find its own up direction, and then determine the picture that it needs to crop or capture. For example, the width is in the direction $-A$, and the height is in the direction (90 degrees $-A$). It can also decide how many pixels in the width direction and the height direction according to the width and height specified by the receiver. The angle is generally quantized at an appropriate granularity and signaled to the sender. The signaling occurs only if the angle has changed significantly.

Another benefit of the proposed method is that it can improve the network-wide system performance. For example, in the cropping technique, when a user

reduces its encoded bit rate, the network can release network resources from this user and assign it to other users that experience poor channel conditions. In doing so, the video quality of the other users is improved while the video quality of the first user remains the same, regardless of the antenna configuration.

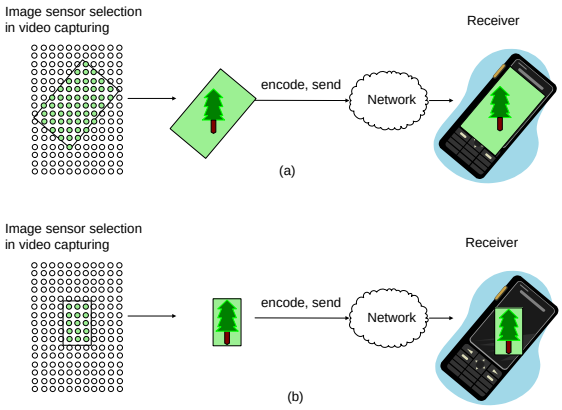


Figure 3 Adaptation in video capturing orientation

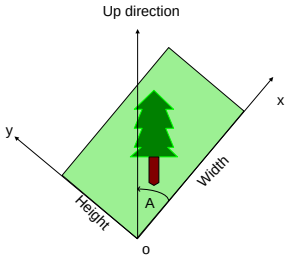


Figure 4 The desired video orientation, aspect ratio, and resolution for the receiver

In addition, if the network cannot provide enough resource for delivering the video at the desired resolution, the sender can generate a video of the same aspect ratio but at a lower resolution, and transmits the lower bit rate video. The receiver can then up sample the decoded video. This reduces packet losses in the network which result in error propagation – a cause of undesired user experience.

3. Conclusion

In this short paper we have proposed a sender-side video adaptation method that can significantly improve the user experience and the efficiency in using network resource for video telephony over wireless communication systems. The proposed method is attractive due to its effectiveness and simplicity.

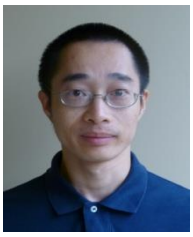
References

[1] ITU-T Recommendation H.264: Advanced video coding for generic audiovisual services, Nov. 2007.

[2] ITU-T H.265, "High Efficiency Video Coding," June, 2013.

[3] 3GPP TS 36.300, V11.6.0, "Evolved Universal Terrestrial Radio Access Network; Overall Description," Release 11, 2013.

[4] 3GPP TS 26.114 V12.1.0, "IP Multimedia Subsystem (IMS); Multimedia Telephony; Media handling and interaction (Release 12)," 2013.



Liangping Ma (M'05-SM'13) currently is working on network resource allocation for video QoE optimization and on cognitive radios at InterDigital. He was the principal investigator of two US government funded research projects. He was with San Diego Research Center Inc. (2005-2007) and Argon ST Inc. (2007-2009). He received his B.S. degree in physics from Wuhan University, China, in 1998, and his Ph.D. in electrical engineering from University of Delaware, US, in 2004. He has authored/co-authored more than 30 journal and conference papers.

Yong He is a member of technical staff in InterDigital Communications, Inc, San Diego, CA, USA. His early working experiences include various positions, including Principal Staff Engineer, at Motorola, San Diego, CA, USA, from 2001 to 2011, and Motorola Australia Research Center, from 1999 to 2001. He is currently active in video coding related standardization at MPEG, JCT-VC and 3GPP SA4 Working group. He received Ph.D. degree from Hong Kong University of Science and Technology, M.S. and B.S degrees from Southeast University, China.

Gregory Sternberg received his MSEE degree from the University of Pennsylvania (1996) and BSEE from the Pennsylvania State University (1994). He joined InterDigital in 2000 where he has developed algorithms for various 3GPP cellular systems for both technology and product development projects. Currently he is a Principal Engineer at InterDigital where he is leading a project related to Video Optimization over Wireless Networks. He holds more than 20 issued patents with many other patents pending and has co-authored several conference papers.

Yan Ye (M'08-SM'13) received her Ph.D. from the Electrical and Computer Engineering Department at University of California, San Diego in 2002. She received her M.S. and B.S. degrees, both in Electrical Engineering, from the University of Science and Technology of China, in 1997 and 1994, respectively. She currently works at Innovation Labs at InterDigital Communications. Previously she has worked at Image Technology Research at Dolby Laboratories Inc and Multimedia R&D and Standards at Qualcomm Inc. She has been involved in the development of various video coding standards, including the HEVC standard and its scalable extensions, the Key Technology Area of ITU-T/VCEG, and the scalable extensions of H.264/AVC. Her research interests include video coding, processing and streaming.

Yuriy A. Reznik (M'97-SM'07) is a Director of Engineering at InterDigital Communications, Inc., (San Diego, CA), where he leads R&D in multimedia coding and delivery over wireless networks. Previously, he worked at Qualcomm (2005-2011), RealNetworks (1998-2005), and also stayed as Visiting Scholar at Stanford University (2008). He holds a Ph.D. degree in Computer Science from Kiev University. He has authored/co-authored over 90 conference and journal papers, and co-invented over 20 issued US patents.

HTTP Adaptive Streaming (HAS): QoE-Aware Resource Allocation over LTE

Vishwanath Ramamurthi, and Ozgur Oyman
Intel Labs, Santa Clara, USA.
vishwanath.ramamurthi@intel.com, ozgur.oyman@intel.com

1. Introduction to HAS

Among various categories of Internet traffic, streaming video traffic is growing at an exponential rate due to an enormity of video content on the Internet and the widespread availability of mobile devices with high display abilities in the consumer market. According to the traffic forecast by Cisco [1], by 2017, two-thirds of the total mobile traffic will constitute video streaming. Such an enormous growth in streaming traffic over wireless brings tremendous challenges to service providers at various levels due to limited availability of wireless resources in terms of frequency, time, and space dimensions in addition to the volatile nature of wireless environment. The problem of providing good video Quality of Experience (QoE) to a large number of users with limited resources needs to be tackled from various directions including efficient video compression, improved wireless technology, adaptive streaming, etc. This paper highlights the use of intelligent radio resource allocation in conjunction with HTTP Adaptive Streaming (HAS) to improve user QoE over modern wireless networks such as 3GPP LTE (Long Term Evolution).

HAS has become a popular delivery platform for streaming video with proprietary deployments by prominent players like Apple (HTTP Live Streaming [2]), Microsoft (Smooth Streaming [3]) and Adobe (HTTP Dynamic Streaming [4]). Being a client-driven pull-based video adaptation approach, it has the ability to adapt to varying network conditions and deliver video efficiently using the available network resources. With several inherent advantages over traditional server-controlled solutions, HAS is expected to be broadly deployed over coming few years [5]. Meanwhile, it has also been standardized as Dynamic Adaptive Streaming over HTTP (DASH) by MPEG and 3GPP as a converged format for video streaming [6, 7] and is endorsed by an ecosystem of over 50 member companies at the DASH Industry Forum. It has also been adopted by various important organizations such as Digital Living Network Alliance (DLNA), Open IPTV Forum (OIPF), Digital Entertainment Content Ecosystem (DECE), World-Wide Web Consortium (W3C), Hybrid Broadcast Broadband TV (HbbTV) etc.

In HAS framework, the video content is divided into smaller segments which are pre-encoded at different adaptation levels and available at the content servers. The client plays the primary role in rate adaptation by choosing and requesting the appropriate video representation for each video segment. The client choses the adaptation level of the next video segment based on its state and also its estimate of the available link bandwidth. See [8, 9] for details on HAS state modeling and rate adaptation.

2. Video-Aware Wireless Resource Allocation

Wireless technology has been advancing rapidly to provide increasing bandwidths with Long Term Evolution (LTE) providing peak bandwidth of the order of several hundred Mbps and LTE-Advanced promising up to 1 Gbps peak bandwidth using several enhancements like carrier aggregation [10]. However the growth-rate in wireless bandwidth using technological advancements is still behind the growth-rate of video traffic. Also the average bitrates experienced by users is often much less than peak bitrates promised due to non-ideal channel conditions. Varying channel conditions nullify the benefits obtained using HAS-based rate adaptation because of mismatch between estimated link bandwidth and actual link bandwidth. This in turn results in poor video quality experience to users. Therefore wireless resources should be intelligently used to maximize user QoE [11].

Wireless resource allocation algorithms for cellular networks traditionally focus on opportunistic scheduling with some sort fairness guarantees [12, 13]. The Proportional Fair (PF) algorithm optimizes a metric which considers both the instantaneous channel quality information (CQI) and the average user-throughput [12]. In every scheduling time slot t (or more generally resource slot) the user which maximizes the following metric is chosen:

$$\text{PF}: \ j = \arg\max_{j \in N} \left[\mu_j(t) / R_j(t) \right] \qquad (1)$$

where $\mu_j(t)$ represents the peak throughput that could be obtained by user j in time slot t and $R_j(t)$ is the moving average service rate for user j. $R_j(t)$ is updated as follows based on scheduling decisions in each time slot:

$$R_j(t+1) = (1-\beta) R_j(t) + \beta \phi_j(t) \qquad (2)$$

Where $\phi_j(t) = \mu_j(t)$ if user j is scheduled in time slot t and $\phi_j(t) = 0$ otherwise. PF scheduling maximizes the sum of logarithms of average user service rates in the long run i.e.,

$$\text{PF}: \max \text{H}(\mathbf{R}) = \sum_{j} \log(R_j) \quad \text{s.t. } \mathbf{R} \in \mathbf{V} \tag{3}$$

where $\mathbf{R}$ = (R_1, R_2, R_3, …R_J) and $\mathbf{V}$ represents the capacity region of the system. Existing video-aware wireless resource allocation methods modify the optimization objective to video quality metrics like SSIM (Structural Similarity Index), PVQ (Perceived Video Quality) and MOS (Mean Opinion Score) [14-17] using mathematical modeling of these metrics in terms of average service rates. However, according to recent surveys [18], re-buffering has the highest impact on viewer experience. Considering this, an algorithm that modifies the PF objective to give high priority to users with low buffer levels was proposed in [9]. This algorithm known as Proportional Fair with Barrier Frames (PFBF) was shown to improve QoE-outage based video capacity of the system. However, modifying the objective to provide an emergency type response penalizes other users, thus decreasing the effectiveness of the algorithm. Our recent research has focused on periodic client media-buffer feedback based dynamic resource allocation. Such feedback has been incorporated into the DASH standard [19]. Unlike previous approaches, we add re-buffering constraints to the resource allocation problem in Eqn. (3) instead of modifying the optimization objective. To avoid re-buffering, we require that the rate of change of media buffer level has to be greater than a certain threshold for each client during each feedback period i.e.,

$$B_j^{i,\text{diff}} / \tau \geq \delta \quad \forall\, i, j \tag{4}$$

Where τ is the feedback reporting period, $B_j^{i,\text{diff}}$ is the difference in media buffer levels of user j during the i^{th} feedback reporting period and $\delta > 0$ is a design parameter to account for varying network conditions. We solve this constrained optimization problem using a token-based gradient algorithm called Re-buffering Aware Gradient Algorithm (RAGA) (see [20] for more technical details) , which maximizes a modified PF metric in every time slot:

$$\text{RAGA}: j = \arg\max_{j \in N} \left[e^{a_j(t) W_j(t)} \mu_j(t) / R_j(t) \right] \tag{5}$$

$W_j(t) > 0$ is a video-aware user token parameter that is updated based on rate of change of user-buffer level during each feedback reporting period τ and $a_j(t)$ is a parameter that is set based on the absolute buffer levels to improve convergence of the algorithm. When rate of media buffer change for a certain user is below threshold during a feedback period, the token parameter $W_j(t)$ for the user is incremented to increase its relative priority compared to other users. On the other hand when the rate of media buffer change is above the threshold, the user-token parameter $W_j(t)$ is decreased towards 0, thus moving towards the standard PF metric. $a_j(t)$ is set to unity for users with high buffer levels but for users with buffer levels below a threshold it is set to proportionately higher values for enforcing re-buffering constraints in a shorter time scale. Unlike [9], the scheduling priorities to users are continuously adjusted based not only on the absolute values of client media buffer levels but also on the rate of change of these buffer levels. Also it is friendly to non-video users.

3. Performance Evaluation

We evaluated the performance of RAGA on an LTE system level simulator developed in MATLAB. Our system consists of a 19-cell layout in which the center cell is surrounded by two layers of interfering cells generating full buffer traffic. 75 adaptive streaming video users are randomly distributed in the center cell. Each HAS client fetches video content from a server attached to the core network that is connected to the center cell through a backhaul network of very high (1 Gbps) bandwidth. The OFDMA-based LTE downlink air interface used has a bandwidth of 10 MHz and is the main bottleneck in the whole system. Half of this bandwidth is assumed to be reserved for the DASH-based video streaming service while the remaining half is assumed to be dedicated for other services. The parameter settings and assumptions on the LTE air interface are the same as in [8, 9] except for the scheduler modifications. Channel Quality Indicator (CQI) are delayed by 5ms, and HARQ retransmissions are delayed by 8 ms with a maximum of 4 retransmissions allowed. Each client randomly picks one of the five sample video clips whose rate-PSNR characteristics are shown in Fig. 1. Video traffic is simulated using the trace-based approach proposed in [21]. Video frame rate is set to 30 fps, GoP size to 16 frames, and segment size to 1 GoP. TCP Reno flavor is used as the transport protocol and no losses are assumed for TCP ACKs. An MTU of 1500 bytes was used for TCP packets and 40 bytes of header was also included in each TCP segment to account for NALU prefix and HTTP/TCP protocol headers. 100,000 LTE sub-frames were simulated to obtain performance statistics.

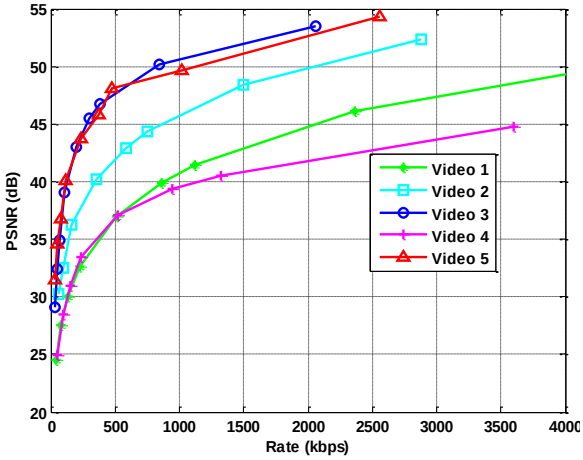


Fig. 1. Rate-PSNR curves of sample videos.

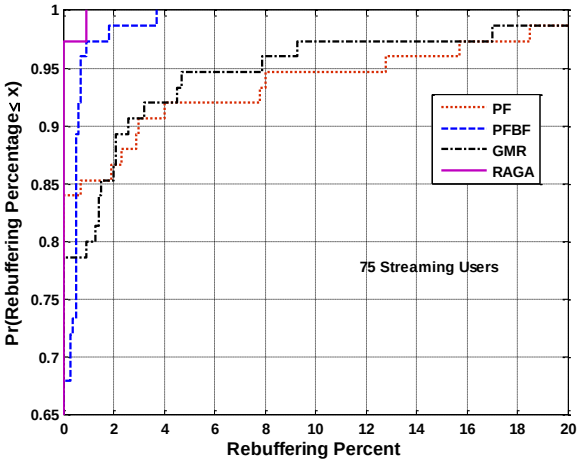


Fig. 2: CDF of Re-buffering Percentage.

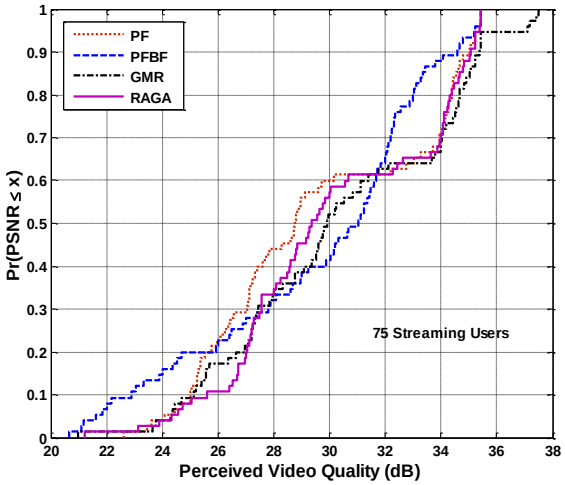


Fig. 3: CDF of Perceived Video Quality (PVQ).

We compare RAGA with standard PF, Proportional fair With Barrier for Frames (PFBF) [9], and GMR (Gradient with Min rate) [22] algorithms (with minimum rate set to the rate of the lowest representation level of the user's video).

Fig. 2 compares the CDFs of re-buffering percentage using PF, PFBF, GMR, and RAGA algorithms. PF, being QoE unaware, has the worst re-buffering performance with large number of users experiencing high re-buffering percentages. GMR is better than PF, but lags behind due to lack of dynamic cooperation between resource allocation and media buffer evolution. PFBF performs better than GMR by drastically reducing peak re-buffering percentages. But it has high number of users experiencing small re-buffering percentages due to inadvertently penalization of good users in emergency situations. RAGA has the lowest re-buffering percentage among all schemes with smallest number of users experiencing lowest peak re-buffering percentages.

Fig. 3 compares CDFs of Perceived Video Quality (PVQ) for the various schemes. PVQ is computed as the difference between the mean and standard deviation of PSNR. The PVQ using RAGA is better than PF scheduling for all users. GMR and PFBF appears to have marginally better PVQ than RAGA for some users but this is at a huge cost in terms re-buffering percentages. RAGA has the most balanced PVQ among all the schemes and also the lowest re-buffering percentages. RAGA shows significant reduction in re-buffering percentage for adaptive streaming users and better perceived video quality than other schemes.

Our results indicate that significant improvements in Video QoE could be obtained by Video-QoE aware radio resource allocation based on simple cross-layer feedback such as periodic media buffer feedback from adaptive streaming clients.

REFERENCES

[1] Cisco, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2012-2017."

[2] R. Pantos, "HTTP Live Streaming," IETF Draft, Oct 2012.

[3] Microsoft, "Smooth Streaming protocol.," http://msdn.microsoft.com/en-us/library/ff469518.aspx.

[4] Adobe, "Adobe HTTP Dynamic Streaming," http://www.adobe.com/products/httpdynamicstreaming/pdfs/httpdynamicstreaming_wp_ue.pdf.

[5] TDG-Research, http://www.tdgresearch.com.

[6] 3GPP TS 26.247, "Transparent end-to-end packet switched streaming service (PSS); Progressive download and dynamic adaptive streaming over HTTP (3GP-DASH)."

[7] ISO/IEC 23009-1, "Information technology — Dynamic adaptive streaming over HTTP (DASH) – Part 1: Media presentation description and segment formats."

[8] V. Ramamurthi and O. Oyman, "Link Aware HTTP Adaptive Streaming for Enhanced Quality of Experience," in Proc. 2013 IEEE GLOBECOM Symp. on Comms. Software, Services and Multimedia, accepted for publication.

[9] S. Singh, O. Oyman, A. Papathanassiou, D. Chatterjee, and J. G. Andrews, "Video capacity and QoE enhancements over LTE," in Proc. IEEE International Conference on Communications (ICC), 2012 pp. 7071-7076, 2012.

[10] 4GAmericas, "The Path to 4G: LTE and LTE-Advanced " Alcatel-Lucent, Oct. 2010.

[11] O. Oyman, J. Foerster, Y.-J. Tcha, and S.-C. Lee, "Toward enhanced mobile video services over WiMAX and LTE [WiMAX/LTE Update],"

Communications Magazine, IEEE, vol. 48, pp. 68-76, 2010.

A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system," in Proc. IEEE 51st Vehicular Technology Conference, Tokyo, vol. 3, pp. 1854-1858, 2000.

[12] M. Andrews, "A Survey of Scheduling Theory in Wireless Data Networks," in Wireless Communications. vol. 143, P. Agrawal, P. Fleming, L. Zhang, D. Andrews, and G. Yin, Eds., pp. 1-17, Springer New York, 2007.

[13] S. Thakolsri, S. Khan, E. Steinbach, and W. Kellerer, QoE-Driven Cross-Layer Optimization for High Speed Downlink Packet Access vol. 4, 2009.

[14] K. Piamrat, K. D. Singh, A. Ksentini, C. Viho, and J. Bonnin, "QoE-Aware Scheduling for Video-Streaming in High Speed Downlink Packet Access," in Proc. IEEE Wireless Communications and Networking Conference (WCNC), pp. 1-6, 2010.

[15] D. Bethanabhotla, G. Caire, and M. J. Neely, "Joint transmission scheduling and congestion control for adaptive streaming in wireless device-to-device networks," in Proc. Forty Sixth Asilomar Conference on Signals, Systems and Computers (ASILOMAR), pp. 1179-1183, 2012.

[16] V. Joseph and G. de Veciana, "Jointly optimizing multi-user rate adaptation for video transport over wireless systems: Mean-fairness-variability tradeoffs," in Proc. IEEE INFOCOM, pp. 567-575, 2012.

[17] Conviva, "Viewer Experience Report," 2012. (available online at http://www.conviva.com/vxr/)

[18] 3GPP, "TS 26.247: Transparent End-to-End Packet-Switched Streaming Service (PSS)—Progressive Download and Dynamic Adaptive Streaming over HTTP (3GP-DASH)," Release 11, 2013.

[19] V. Ramamurthi and O. Oyman, "Video-QoE Aware Radio Resource Allocation for HTTP Adaptive Streaming," Submitted to IEEE ICC 2014.

[20] P. Seeling and M. Reisslein, "Video Transport Evaluation With H.264 Video Traces," Communications Surveys & Tutorials, IEEE, vol. 14, pp. 1142-1165, 2012.

[21] M. Andrews, Q. Lijun, and A. Stolyar, "Optimal utility based multi-user throughput allocation subject to throughput constraints," in Proc. IEEE INFOCOM 2005, vol. 4, pp. 2415-2424 vol. 4, 2005.

VISHWANATH RAMAMURTHI received his B.S. degree in Electronics and Communication Engineering from Birla Institute of Technology, India, in 2003, his M.S. degree in communication engineering from the Indian Institute of Technology, Delhi, India, in 2005, and PhD in Electrical and Computer Engineering from the University of California Davis, CA, USA in 2009. He worked as a research scientist at the General Motors India Science Lab in 2006, as a research intern at Fujitsu Labs of America in 2009, and as a senior member of technical staff at the AT&T Labs from 2009 to 2012. Currently he is working as a Research Scientist with the Mobile Multimedia Solutions group at the Intel Labs in Santa Clara, CA, USA. His current research interests include video optimization over wireless networks, cross-layer design 4G/5G cellular networks, and cellular network modeling and optimization.

OZGUR OYMAN is a senior research scientist and project leader in the Wireless Communications Lab of Intel Labs. He joined Intel in 2005. He is currently in charge of video over 3GPP Long Term Evolution (LTE) research and standardization, with the aim of developing end-to-end video delivery solutions enhancing network capacity and user quality of experience (QoE). He also serves as the principal member of the Intel delegation responsible for standardization at 3GPP SA4 Working Group (codecs). Prior to his current roles, he was principal investigator for exploratory research projects on wireless communications addressing topics such as client cooperation, relaying, heterogeneous networking, cognitive radios and polar codes. He is author or co-author of over 70 technical publications, and has won Best Paper Awards at IEEE GLOBECOM'07, ISSSTA'08 and CROWNCOM'08. His service includes Technical Program Committee Chair roles for technical symposia at IEEE WCNC'09, ICC'11, WCNC'12, ICCC'12 and WCNC'14. He also serves an editor for the IEEE TRANSACTIONS ON COMMUNICATIONS. He holds Ph.D. and M.S. degrees from Stanford University, and a B.S. degree from Cornell University.

MMTC OFFICERS

CHAIR

Jianwei Huang
The Chinese University of Hong Kong
China

STEERING COMMITTEE CHAIR

Pascal Frossard
EPFL, Switzerland

VICE CHAIRS

Kai Yang
Bell Labs, Alcatel-Lucent
USA

Chonggang Wang
InterDigital Communications
USA

Yonggang Wen
Nanyang Technological University
Singapore

Luigi Atzori
University of Cagliari
Italy

SECRETARY

Liang Zhou
Nanjing University of Posts and Telecommunications
China

E-LETTER BOARD MEMBERS