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

E-LETTER

Vol. 8, No. 5, September 2013

IEEE COMMUNICATIONS SOCIETY

CONTENTS

Message from MMTC Chair 3
EMERGING TOPICS: SPECIAL ISSUE ON MULTIMEDIA STREAMING OVER MOBILE NETWORKS 4
Guest Editors: Sanjeev Mehrotra, Microsoft Research, USA, sanjeevm@microsoft.com 4
Mohamed Hefeeda, Simon Fraser University, Canada, mhefeeda@cs.sfu.ca 4
Recent and Future Trends in Mobile Video Streaming 6
Nabil J. Sarhan 6
Wayne State University, Detroit, USA 6
nabil@ece.eng.wayne.edu 6
Dynamic Adaptive Streaming over HTTP (DASH) Standardization at MPEG and 3GPP 9
Ozgur Oyman, Intel Labs, Santa Clara, CA, USA, ozgur.oyman@intel.com 9
Energy-Efficient On-Demand Streaming in Mobile Cellular Networks 12
Yong Cui, Xiao Ma*, Jiangchuan Liu† and Yayun Bao** 12
** Tsinghua University, Beijing, P.R. China, cuiyong@tsinghua.edu.cn, max11@mails.tsinghua.edu.cn, jimmybao0730@gmail.com* 12
† Simon Fraser University, Burnaby, BC, Canada, jcliu@cs.sfu.ca 12
A Marketplace for Mobile Applications Supporting Rich Multimedia Feeds 16
Ngoc Do[1], Ye Zhao[1], Cheng-Hsin Hsu[2], and Nalini Venkatasubramanian[1] 16
[1] University of California, Irvine, USA, {nmdo, yez, nalini}@ics.uci.edu 16
[2] National Tsing Hua University, Hsin-Chu, Taiwan, chsu@cs.nthu.edu.tw 16
Adaptive Streaming in Mobile Cloud Gaming 20
Shervin Shirmohammadi[1,2] 20
[1]University of Ottawa, Canada, shervin@eecs.uottawa.ca 20
[2] Istanbul Sehir University, Turkey, shervinshirmohammadi@sehir.edu.tr 20
INDUSTRIAL COLUMN: SPECIAL ISSUE ON VIDEO AWARE WIRELESS NETWORKS 24
Guest Editor: Jeffrey R. Foerster, Intel, USA; Michelle Gong, Google, USA 24
jeffrey.r.foerster@intel.com; michellegong@google.com 24
Video QoE Models for the Compute Continuum 26
Lark Kwon Choi[1, 2], Yiting Liao[1], and Alan C. Bovik[2] 26
[1]Intel Labs, Hillsboro, OR, USA, {lark.kwon.choi, yiting.liao}@intel.com. 26
[2]The University of Texas at Austin, Austin, TX, USA, bovik@ece.utexas.edu 26
Perceptual Optimization of Large-Scale Wireless Video Networks 30

Robert W. Heath Jr., Alan C. Bovik, Gustavo de Veciana, Constantine Caramanis, and Jeffrey G. Andrews 30
Wireless Networking and Communications Group, The University of Texas at Austin 30
{rheath,bovik,gustavo,cmcaram,jandrews}@ece.utexas.edu 30
Dynamic Adaptive Streaming over HTTP: Standards and Technology **33**
Ozgur Oyman, Intel Labs, Santa Clara, CA, USA, ozgur.oyman@intel.com 33
Utsaw Kumar, Intel Architecture Group, Santa Clara, CA USA, utsaw.kumar@intel.com ... 33
Vish Ramamurthi, Intel Labs, Santa Clara, CA USA, vishwanath.ramamurthi@intel.com ... 33
Mohamed Rehan, Intel Labs, Cairo, Egypt, mohamed.m.rehan@intel.com 33
Rana Morsi, Intel Labs, Cairo, Egypt, ranax.a.morsi@intel.com 33
Cross-Layer Design for Mobile Video Transmission **37**
Laura Toni, Dawei Wang**, Pamela Cosman**, Laurence Milstein*** 37
**EPFL, Lausanne, Switzerland, laura.toni@epfl.ch* 37
***University of California San Diego, USA,{daw017, pcosman, lmilstein}@ucsd.edu* 37
Timely Throughput Optimization of Heterogeneous Wireless Networks **42**
Sina Lashgari and A. Salman Avestimehr 42
Cornell University, Ithaca, NY 42
sl2232@cornell.edu, avestimehr@ece.cornell.edu 42
Video Caching and D2D Networks **45**
Giuseppe Caire, Andreas F. Molisch, and Michael J. Neely 45
Dpt. of EE, University of Southern California 45
{caire, molisch, mjneely}@usc.edu 45
MMTC OFFICERS **42**

Message from MMTC Chair

Dear MMTC colleagues:

Wish you all a pleasant summer and new semester starts with a refreshed spirit. It has been a great honor for me to serve as the Asia vice-chair for this vital ComSoc Committee during the period 2012-2014. Supported by our community, our TC has made significant accomplishments in 2013

First, ICME 2013 has been a great success in San Jose, California. The conference received 622 valid submissions, among which 79 were accepted as oral presentations and 110 were accepted as posters. As one of the sponsoring TCs, we have contributed one TPC co-chair (Dr. Weisi Lin from Nanyang Technological University), TPC area chairs and TPC committee members. On behalf of Dr. Jianwei Huang, I have chaired our TC meeting at ICME with more than 20 community leaders. The discussions and resulted suggestions at the TC meeting would shape how MMTC will be further improved in our operations.

Second, supported by our IG leaders, Dr. Chonggang Wang and I have received more than 10 proposals on special issues. We are going through the list of proposals right now and hope to consolidate them into 4-6 high-quality proposals. Our next step will be to connect each team with a senior mentor to ensure a high rate of acceptance. In the meantime, we welcome new proposals from our community on emerging topics in multimedia communications.

Finally, I would like to offer an invitation to all the MMTC members to join the technical program committee of ICC 2014 Selected Area on Communication Symposium on Cloud Computing Track. The CFP can be found at http://www.ieee-icc.org/2014/Cloud_Computing.pdf. Representing MMTC, I will serve as the symposium chair for this SAC. If it aligns with your research interest, please kindly drop me a note to my email (ygwen@ntu.edu.sg).

I would like to thank all the IG chairs and co-chairs for the work they have already done and will be doing for the success of MMTC and hope that any of you will find the proper IG of interest to get involved in our community!

Yonggang Wen
Asia Vice-Chair of Multimedia Communications TC of IEEE ComSoc
Assistant Professor, School of Computer Engineering, Nanyang Technological University, Singapore

EMERGING TOPICS: SPECIAL ISSUE ON MULTIMEDIA STREAMING OVER MOBILE NETWORKS

Guest Editors

Sanjeev Mehrotra, Microsoft Research, USA, sanjeevm@microsoft.com

Mohamed Hefeeda, Simon Fraser University, Canada, mhefeeda@cs.sfu.ca

With recent advances in technology, mobile networks such as Wi-Fi, 3G networks (e.g., UMTS, EVDO), and 4G networks (LTE) are increasingly being used to access the Internet as they allow for ubiquitous connectivity. At the same time, mobile devices such as smartphones, tablets, and even small form factor full-featured laptops are making it practical and convenient for users to utilize a range of applications from anywhere at any time.

The rapid adoption of both mobile networks and mobile devices makes it practical for users to enjoy a range of multimedia applications such as multimedia streaming and allowing for new multimedia scenarios such as gaming and virtual reality. At the same time it also presents unique challenges since mobile networks are inherently different from classical networks in terms of available capacity, delay, loss, and variability. In addition, mobile devices often have stringent size, computational power, and energy constraints.

This special issue of E-Letter focuses on recent advances in the broad area of multimedia streaming over mobile networks. It is an honor to have five papers contributed by authors from leading academic and industrial labs describing their recent research work in the mobile multimedia area.

In the first article, titled "*Recent and Future Trends in Mobile Video Streaming",* Nabil J. Sarhan from Wayne State University presents an overview of video streaming over mobile networks, and analyzes recent trends in mobile video streaming, and highlights several challenges in this area. He also presents an overview of adaptive streaming technologies, such as DASH, to deal with some of these challenges, and discusses methods used for content distribution in mobile networks.

The second article, *"Dynamic Adaptive Streaming over HTTP (DASH) Standardization at MPEG and 3GPP"* by Ozgur Oyman from Intel Labs goes into further detail regarding adaptive streaming using DASH and its specific implementation in 3GPP. The author also considers further DASH enhancements specifically relevant to mobile networks.

In the third paper, *"Energy-Efficient On-Demand Streaming in Mobile Cellular Networks",* Yong Cui, Xiao Ma, Jiangchuan Liu and Yayun Bao, authors from Tsinghua University and Simon Fraser University present work on an energy efficient design for media streaming applications over mobile cellular networks. Lyapunov optimization is used to optimally schedule traffic to multiple clients to minimize total energy cost to users subject to users' QoE constraints. The optimization considers both the energy cost required when receivers are in high energy radio states as well as the additional cost of tail states for each transmission burst.

In the fourth paper, *"A Marketplace for Mobile Applications Supporting Rich Multimedia Feeds"* by Ngoc Do, Ye Zhao, Cheng-Hsin Hsu, and Nalini Venkatasubramanian, authors from University of California, Irvine and from National Tsing Hua University consider how to enhance the quality of service for multimedia applications by creating a marketplace where mobile users trade their residual data plan using short-range networking technology such as Bluetooth and Wi-Fi Direct. The proposed marketplace is realized using a Lyapunov optimization framework.

Shervin Shirmohammadi jointly affiliated with the University of Ottawa and Istanbul Sehir University presents the final paper, *"Adaptive Streaming in Mobile Cloud Gaming"* which considers a relatively new application for multimedia streaming which is online mobile gaming. In the mobile gaming scenario, challenging network conditions and potential data limits may be present on mobile networks. In addition, stringent CPU, memory, and battery life constraints may be present on the device. The author discusses adaptation techniques to overcome these challenges.

As can be seen from the five papers, mobile networks and devices provide both opportunities for new multimedia streaming applications as well unique networking and device challenges which have to be overcome in order for multimedia streaming to work well. The high demands of multimedia applications coupled with the stringent constraints of mobile networks and devices require novel adaptation and optimization techniques to allow for good performance. Thus it provides an active and fruitful area for research.

While this special issue is far from delivering complete coverage of the vast area of multimedia streaming over mobile networks, we hope the five papers give an overview of some of the exciting research going on in the field as well as provide an opportunity to explore and collaborate in related fields.

We would like to thank the authors for their contribution in making this special issue a success as well as thank the E-letter board for making this issue possible.

Sanjeev Mehrotra is a Principal Architect at Microsoft Research in Redmond, WA, USA. He received his Ph.D. in electrical engineering from Stanford University in 2000. He was previously the development manager for the audio codecs and DSP team in the core media processing technology team and led the development of WMA audio codec. He is the inventor of several audio and video codec technologies in use today and developed the initial prototype version of Microsoft's Smooth Streaming. His work has shipped in numerous Microsoft products and standard codecs such as H.264 AVC. His research interests include multimedia communications and networking, multimedia compression, and large scale distributed systems. Dr. Mehrotra is an author on over 30 refereed journal and conference publications and is an author on over 70 US patent applications, out of which 35 have been issued. He was general co-chair and TPC chair for the Hot Topics in Mobile Multimedia (HotMM) workshop at ICME 2012 and is currently an associate editor for the Journal of Visual Communication and Image Representation. He is a senior member of IEEE and is a recipient of the prestigious Microsoft Gold Star Award for his work on WMA codec.

Mohamed Hefeeda received his Ph.D. from Purdue University, USA in 2004, and M.Sc. and B.Sc. from Mansoura University, Egypt in 1997 and 1994, respectively. He is an associate professor in the School of Computing Science, Simon Fraser University, Canada, where he leads the Network Systems Lab. His research interests include multimedia networking over wired and wireless networks, peer-to-peer systems, mobile multimedia, and cloud computing. In 2011, he was awarded one of the prestigious NSERC Discovery Accelerator Supplements (DAS), which were granted to a selected group of researchers in all Science and Engineering disciplines in Canada. His research on efficient video streaming to mobile devices has been featured in multiple international news venues, including ACM Tech News, World Journal News, SFU NEWS, CTV British Columbia, and Omni-TV. He serves on the editorial boards of several premier journals such as the ACM Transactions on Multimedia Computing, Communications and Applications (TOMCCAP), and he has served on many technical program committees of major conferences in his research area, such as ACM Multimedia. Dr. Hefeeda has co-authored more than 70 refereed journal and conference papers and filed more than 14 patents from his research. He is a senior member of IEEE.

Recent and Future Trends in Mobile Video Streaming

Nabil J. Sarhan
Wayne State University, Detroit, USA
nabil@ece.eng.wayne.edu

1. Introduction

Video streaming services to mobile and wireless devices, including smart phone and tablets, have recently grown greatly in popularity. Many wireless carriers currently offer on-demand and live video streaming services. Unfortunately, the distribution of video streams faces a significant scalability challenge due to the high server and network requirements. Additionally, the mobile and wireless environment imposes other significant challenges.

This paper discusses the current and future status of mobile video and the main challenges in mobile video streaming. It also discusses recent and future trends in the distribution of video streams to mobile devices.

The rest of the paper is organized as follows. Section 2 summarizes the current and future status of mobile data traffic. Subsequently, Section 3 discusses the main challenges in mobile video streaming. Sections 4 and 5 discuss the recent and future trends in mobile video streaming, respectively.

2. Trends in Mobile Traffic

A 2013 report by Cisco on global mobile data traffic forecast demonstrates a huge increase in mobile traffic in general and mobile video traffic in particular [1]. The trends of mobile data traffic in 2012 can be summarized as follows:

- The global mobile data traffic grew by 70 percent and was approximately 12 times the size of the entire Internet in 2000.
- Nearly a third of that traffic was offloaded onto fixed networks through Wi-Fi or femtocell.
- Mobile video traffic was 51 percent of the mobile data traffic.
- The speeds of mobile network connections more than doubled.
- Although 4G connections represent less 1% percent of mobile connections, they account for nearly 15 percent of the mobile data traffic.

The report's forecast for 2017 is not less dramatic.

- Two-thirds of the global mobile data traffic will be video.
- The traffic generated by mobile-connected tablets will exceed that of the entire mobile network in 2012.
- 4G connections will represent 10 percent of all connections and will account for 45 percent of the traffic.

With the rising popularity of mobile video, the higher connection speeds, and the wider spectrum of devices offered and used, the development of high quality video streaming services will become increasingly more important in the future.

3. Main Challenges in Mobile Video Streaming

Mobile video streaming faces significant challenges. First, delivering video streaming to a huge number of users is very demanding of server and network resources. Video streams require high data transfer rates and thus high bandwidth capacities and must be received continuously in time with minimal delay. Second, supporting heterogonous receivers with varying capabilities is hard to achieve efficiently. Mobile devices vary greatly in their capabilities, including screen resolution, computational power, and download bandwidth. Third, the unique characteristics of the wireless and mobile environment should be considered. These characteristics include noise, multi-path interference, mobility, and subsequently dynamic network conditions, and great variations in the actual download bandwidth over time.

4. Recent Trends in Distribution of Mobile Video Streams

The most common approach currently used for addressing the scalability challenge in video streaming is *Content Delivery Networks* (CDNs). As illustrated in Figure 1, the content in the origin server(s) is automatically stored in surrogate servers, located in many cities around the world. Therefore, the user's request for streaming a video is transparently transferred to a surrogate server close to the user's geographical location. The delivery of the content by a server close to the user leads to fast and reliable video streaming and reduces the contention on the Internet [2]. An accounting mechanism is typically employed to relay access information and detailed logs to the origin server(s).

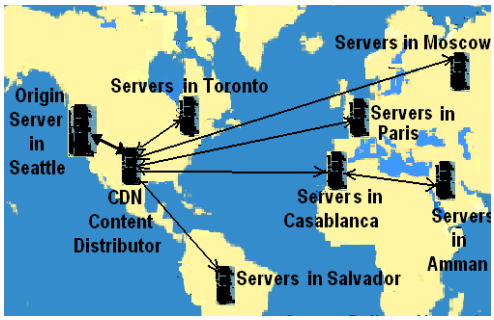


Figure 1: An Illustration of a Simple Content Delivery Network with Many Distributed Servers

To address the dynamic network and channel conditions and to support devices with varying resources (such as download bandwidth and screen resolution), many mobile video streaming services that are provided by wireless carriers started to adopt *Dynamic Adaptive Streaming over HTTP* (DASH), also known as MPEG-DASH. DASH uses HTTP because it is widely deployed in all Web servers and thus provides ease of development. In addition, HTTP allows the use of CDNs and offers firewall friendliness. The MPEG-DASH standard, published as ISO/IEC 23009-1:2012 in April 2012, allows interoperability between devices and servers of various vendors [3].

With DASH, videos are stored with different levels of quality, and users dynamically switch to the appropriate quality level based on the current device state and their preferences. More specifically, the video is encoded into video streams with different bitrates. Subsequently, the video streams are partitioned into segments. These segments are then hosted on origin server(s), along with metadata information describing the relationships among segments and how they collectively form a media presentation. This information is referred to as *Media Presentation Description* (MPD). The client fully controls the streaming session by making HTTP requests to selected segments at different times. Therefore, most of the intelligence is at the client, freeing the server from containing state information for the clients. For further details about DASH, please refer to [3, 4] and ISO/IEC 23009-1:2012.

5. Future Trends in Video Distribution

As discussed earlier, the main approach that has been used for addressing the scalability challenge is CDNs. Another approach is *Peer-to-Peer* (P2P). While the first approach requires maintaining a huge number of geographically distributed servers, the second still relies heavily on central servers [5, 6]. Both these approaches mitigate the scalability problem but do not eliminate it because the fundamental problem is due to unicast delivery [6]. Multicast is highly effective in delivering high-interest and high-usage content and in handling flash crowds. Recently, there has been a growing interest in enabling native multicast to support high-quality on-demand video distribution, IPTV, and other major applications [6, 7].

Multicast-based video delivery can be used to provide a true solution for the scalability problem. This delivery can be done on a *client-pull* or a *server-push* fashion, depending on whether the channels are allocated on demand or reserved in advance, respectively. The first category includes *stream merging* [8] (and references within), which reduces the delivery cost by aggregating clients into larger groups that share the same multicast streams. The achieved degrees of resource sharing depend greatly on how the waiting requests are scheduled for service. The second category consists of *Periodic Broadcasting* techniques [9] (and references within), which divide each video file into multiple segments and broadcast each segment periodically on dedicated server channels.

In contrast with stream merging, periodic broadcasting is cost-performance effective for highly popular content but leads to channel underutilization when the request arrival rate is not sufficiently high. Stream merging works well even for lower arrival rates. The most effective stream merging technique is *Earliest Reachable Merge Target* (ERMT) [10]. It is a near optimal hierarchical stream merging technique, which allows streams to merge unlimited number of times, leading to a dynamic merge tree. Specifically, a new user or a newly merged user group snoops on the closest stream that it can merge with if no later arrivals preemptively catch them [10]. To satisfy the needs of the new user, the target stream may be extended, and this extension may change that stream's merging target [8].

The practical use of stream merging has been hindered by few complications. First, user interactions (such as pause and jump) cause requests to leave ongoing streams, thereby negatively impacting the stream merging process and complicating the server design. Second, supporting heterogonous receivers becomes complicated. Third, supporting video advertisements along with premium video content is also complicated when stream merging is employed.

Fortunately, recent work has addressed most of these challenges [8, 9, 11, 12], and thus multicast-based video delivery is becoming more viable.

6. Conclusion

The interest in mobile video has increased dramatically, and mobile video is expected to account for two-thirds

of the global mobile data traffic in 2016. In this paper, we have discussed the main challenges in mobile video streaming. We have also argued that while DASH coupled with CDN will offer the best solution for mobile video streaming in the near future, multicast-based video delivery will likely become more viable in the long term.

References

[1] Cisco, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2012–2017," *White Paper*, February 6, 2013.

[2] G. Pallis and A. Vakali, "Insight and perspectives for content delivery networks," *Commun. ACM*, vol. 49, pp. 101-106, 2006.

[3] I. Sodagar, "The MPEG-DASH Standard for Multimedia Streaming Over the Internet," *IEEE MultiMedia*, vol. 18, pp. 62-67, 2011.

[4] T. Stockhammer, "Dynamic adaptive streaming over HTTP --: standards and design principles," in Proceedings of ACM Conference on Multimedia Systems, San Jose, CA, USA, 2011.

[5] W. Chuan, L. Baochun, and Z. Shuqiao, "Diagnosing Network-Wide P2P Live Streaming Inefficiencies," in *INFOCOM 2009, IEEE*, 2009, pp. 2731-2735.

[6] V. Aggarwal, R. Caldebank, V. Gopalakrishnan, R. Jana, K. K. Ramakrishnan, and F. Yu, "The effectiveness of intelligent scheduling for multicast video-on-demand," in Proceedings of ACM Conference on Multimedia, Beijing, China, 2009.

[7] S. Ratnasamy, A. Ermolinskiy, and S. Shenker, "Revisiting IP multicast," *SIGCOMM Comput. Commun. Rev.*, vol. 36, pp. 15-26, 2006.

[8] B. Qudah and N. J. Sarhan, "Efficient delivery of on-demand video streams to heterogeneous receivers," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 6, pp. 1-25, 2010.

[9] P. Gill, L. Shi, A. Mahanti, Z. Li, and D. L. Eager, "Scalable on-demand media streaming for heterogeneous clients," *ACM Transactions Multimedia Computing, Communication, and Applications*, vol. 5, pp. 1-24, 2008.

[10] D. Eager, M. Vernon, and J. Zahorjan, "Optimal and efficient merging schedules for video-on-demand servers," in Proceedings of ACM Conference on Multimedia (Part 1), Orlando, Florida, 1999.

[11] K. Nafeh and N. J. Sarhan, "Design and Analysis of Scalable and Interactive Near Video-on-Demand Systems," in Proceedings of IEEE International Conference on Multimedia and Expo, San Jose, California, 2013.

[12] N. J. Sarhan and M. S. Al-Hadrusi, "Waiting-Time Prediction and QoS-based Pricing for Video Streaming with Advertisements," in Proceedings of IEEE International Symposium on Multimedia, 2010.

Nabil J. Sarhan received the Ph.D. and M.S. degrees in Computer Science and Engineering at the 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. Dr. Sarhan is an internationally recognized expert in multimedia systems and multimedia computing and networking and has published extensively in top conferences and journals. His research projects have been primarily sponsored by the National Science Foundation. His main research areas include video streaming and communication, computer and sensor networks, automated video surveillance, multimedia systems design, energy-efficient systems, cross-layer optimization, and storage.

Dr. Sarhan is the Chair of the Interest Group on Media Streaming of the IEEE Multimedia Communication Technical Committee (MMTC). 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, invited panel and speaker co-chair, publicity chair, track chair, and program committee number. He served as the Co-Director of the IEEE MMTC Review Board. He has served as a consulting expert in cases involving patent infringement in mobile video streaming.

Dr. Sarhan is the recipient of the 2008 IEEE Southeastern Michigan Section Outstanding Professional of the Year Award and the Wayne State University 2009 President's Award for Excellence in Teaching.

Dynamic Adaptive Streaming over HTTP (DASH) Standardization at MPEG and 3GPP

Ozgur Oyman, Intel Labs, Santa Clara, CA, USA (ozgur.oyman@intel.com)

1. Introduction on DASH

HTTP adaptive streaming, which has recently been spreading as a form of internet video delivery with the recent deployments of proprietary solutions such as Apple HTTP Live Streaming, Microsoft Smooth Streaming and Adobe HTTP Dynamic Streaming, and is expected to be deployed more broadly over the next few years. Several important factors have influenced this transition to HTTP streaming, including: (i) broad market adoption of HTTP and TCP/IP protocols to support majority of the internet services offered today, (ii) HTTP-based delivery avoids NAT and firewall traversal issues, (iii) HTTP-based (non-adaptive) progressive download solutions are already broadly deployed today, and these can conveniently be upgraded to support adaptive streaming, and (iv) ability to use standards HTTP servers and caches instead of specialized streaming servers, allowing for the reuse of the existing infrastructure and thereby providing better scalability and cost effectiveness.

In the meantime, the standardization of HTTP Adaptive Streaming has also made great progress with the recent completion of technical specifications by various standards bodies. In particular, Dynamic Adaptive Streaming over HTTP (DASH) has recently been standardized by Moving Picture Experts Group (MPEG) and Third Generation Partnership Project (3GPP) as a converged format for video streaming [1,2], and the standard has been adopted by other organizations including Digital Living Network Alliance (DLNA), Open IPTV Forum (OIPF), Digital Entertainment Content Ecosystem (DECE), World-Wide Web Consortium (W3C) and Hybrid Broadcast Broadband TV (HbbTV). DASH today is endorsed by an ecosystem of over 50 member companies at the DASH Industry Forum.

2. Technical Overview of DASH

The scope of both MPEG and 3GPP DASH specifications [1,2] includes a normative definition of a media presentation or manifest format (for DASH access client), a normative definition of the segment formats (for media engine), a normative definition of the delivery protocol used for the delivery of segments, namely HTTP/1.1, and an informative description on how a DASH client may use the provided information to establish a streaming service. This section will provide a technical overview of the key parts of the DASH-based server-client interfaces which are part of MPEG and 3GPP DASH standards. More comprehensive tutorials on various MPEG and 3GPP DASH features can be found in [3]-[5].

The DASH framework between a client and web/media server is depicted in Figure 1. The media preparation process generates segments that contain different encoded versions of one or several of the media components of the media content. The segments are then hosted on one or several media origin servers along with the *media presentation description (MPD)*, that characterizes the structure and features of the media presentation, and provides sufficient information to a client for adaptive streaming of the content by downloading the media segments from the server over HTTP. The MPD describes the various representations of the media components (e.g., bitrates, resolutions, codecs, etc.) and HTTP URLs of the corresponding media segments, timing relationships across the segments and how they are mapped into media presentation.

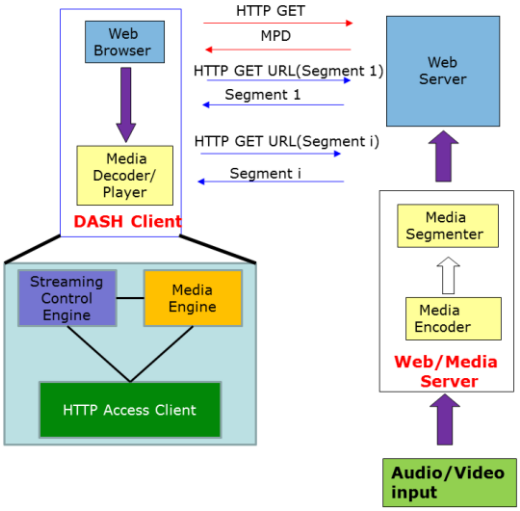


Figure 1 – Server-client interface for DASH-based streaming.

The MPD is an XML-based document containing information on the content based on a hierarchical data model as depicted in Figure 2: Each Period consists of one or more Adaptation Sets. An adaptation set contains interchangeable / alternate encodings of one or more media content components encapsulated in Representations, e.g., an adaptation set for video, one

for primary audio, one for secondary audio, one for captions, etc. In other words, Representations encapsulate media streams that are considered to be perceptually equivalent. Typically, dynamic switching happens across Representations within one Adaptation Set. Segment Alignment permits non-overlapping decoding and presentation of segments from different Representations. Stream Access Points (SAPs) indicate presentation times and positions in segments at which random access and switching can occur. DASH also uses a simplified version of XLink in order to allow loading parts of the MPD (e.g., periods) in real time from a remote location. The MPD can be static or dynamic: A dynamic MPD (e.g., for live presentations) also provides segment availability start time and end time, approximate media start time, and the fixed or variable duration of segments. It can change and will be periodically reloaded by the client, while a static MPD is valid for the whole presentation. Static MPD's are a good fit for video-on-demand applications, whereas dynamic MPD's are used for live and PVR applications.

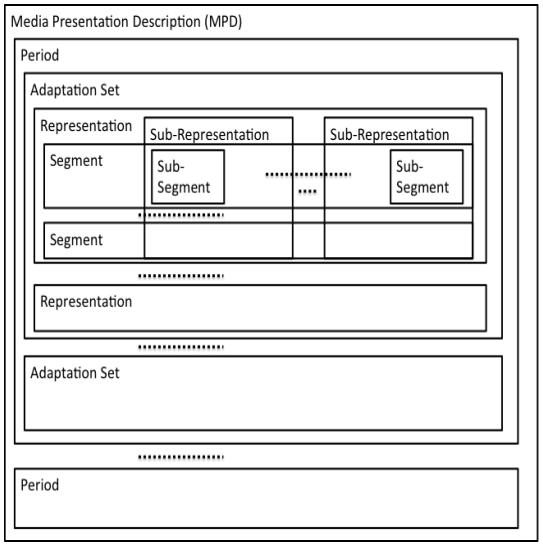


Figure 2 – DASH Media Presentation Description (MPD) hierarchical data model.

A DASH segment constitutes the entity body of the response when issuing a HTTP GET or a partial HTTP GET requests and is the minimal individually addressable unit of data. DASH segment formats are defined for the ISO Base Media File Format and MPEG2 Transport Stream formats. A Media Segment contains media components and is assigned an MPD URL Element and a start time in the media presentation. Segment URLs can be provided in the MPD in the form of exact URLs (segment list) or in the form of templates constructed via temporal or numerical indexing of segments. Dynamic construction of URLs is also possible, by combining parts of the URL (base URLs) that appear at different levels of the hierarchy. Each media segment also contains at least one SAP point, which is a random access or switch-to point in the media stream where decoding can start using only data from that point forward. An Initialization Segment contains initialization information for accessing media segments contained in a Representation and does not itself contain media data. Index segments, which may appear either as side files, or within the media segments, contain timing and random access information, including media time vs. byte range relationships of sub-segments.

DASH provides the ability to the client to fully control the streaming session, i.e., it can intelligently manage the on-time request and smooth playout of the sequence of segments, potentially adjusting bitrates or other attributes in a seamless manner. The client can automatically choose initial content rate to match initial available bandwidth and dynamically switch between different bitrate representations of the media content as the available bandwidth changes. Hence, DASH allows fast adaptation to changing network and link conditions, user preferences and device states (e.g., display resolution, CPU, memory resources, etc.). Such dynamic adaptation provides better user quality of experience (QoE), with higher video quality, shorter startup delays, fewer re-buffering events, etc.

3. MPEG and 3GPP DASH Standards

At MPEG, DASH was standardized by the Systems Sub-Group, with the activity beginning in 2010, becoming a Draft International Standard in January 2011, and an International Standard in November 2011. The MPEG-DASH standard [1] was published as ISO/IEC 23009-1:2012 in April, 2012. In addition to the definition of media presentation and segment formats standardized in [1], MPEG has also developed additional specifications [8]-[10] on aspects of implementation guidelines, conformance and reference software and segment encryption and authentication. Toward enabling interoperability and conformance, DASH also includes profiles as a set of restrictions on the offered media presentation description (MPD) and segments based on the ISO Base Media File Format (ISO-BMFF) [7] and MPEG-2 Transport Streams [6], as depicted in Figure 3. In the meantime, MPEG DASH is codec agnostic and supports both multiplexed and un-multiplexed encoded content. Currently, MPEG is also pursuing several core experiments toward identifying further DASH enhancements, such as signaling of quality information, DASH authentication, server and network assisted DASH operation,

controlling DASH client behavior and spatial relationship descriptions.

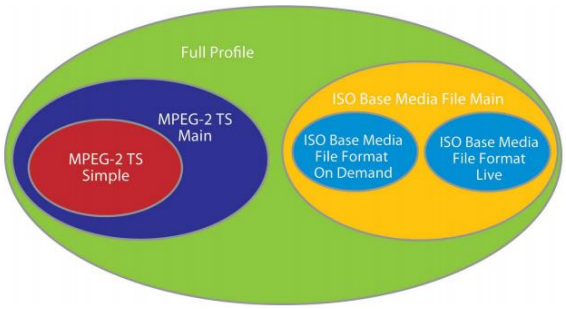


Figure 3 – MPEG DASH Profiles

At 3GPP, DASH was standardized by the 3GPP SA4 Working Group, with the activity beginning in April 2009 and Release 9 work with updates to Technical Specification (TS) 26.234 on the Packet Switched Streaming Service (PSS) [11] and TS 26.244 on the 3GPP File Format [12] completed in March 2010. During Release 10 development, a new specification TS 26.247 on 3GPP DASH [2] has been finalized in June 2011, which is a compatible profile of MPEG-DASH based on the ISOBMFF format. In conjunction with a core DASH specification, 3GPP DASH also includes additional system-level aspects, such as codec and Digital Rights Management (DRM) profiles, device capability exchange signaling and Quality of Experience (QoE) reporting. Since Release 11, 3GPP has been studying further enhancements to DASH and toward this purpose collecting new use cases and requirements, as well as operational and deployment guidelines. Some of the documented use cases in the related Technical Report (TR) 26.938 [14] include: Operator control for DASH, e.g., for QoE/QoS handling, advanced support for live services, DASH as a download format for push-based delivery services, enhanced ad insertion support, enhancements for fast startup and advanced trick play modes, improved operation with proxy caches, multimedia broadcast and multicast service (MBMS) [13] assisted DASH services with content caching at the UEs, handling special content over DASH and enforcing specific client behaviors, and use cases on DASH authentication.

REFERENCES

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

[2] 3GPP TS 26.247: "Transparent end-to-end packet switched streaming service (PSS); Progressive download and dynamic adaptive streaming over HTTP (3GP-DASH)"

[3] I. Sodagar, "The MPEG-DASH Standard for Multimedia Streaming Over the Internet", IEEE Multimedia, pp. 62-67, Oct.-Dec. 2011.

[4] T. Stockhammer, "Dynamic Adaptive Streaming over HTTP: Standards and Design Principles", Proc. ACM MMSys2011, San Jose, CA, Feb. 2011.

[5] O. Oyman and S. Singh, "Quality of experience for HTTP adaptive streaming services", IEEE Commun. Mag., vol. 50, no:4, pp. 20-27, Apr. 2012.

[6] ITU-T Rec. H.222.0 | ISO/IEC 13818-1:2013: "Information technology - Generic coding of moving pictures and associated audio information: Systems"

[7] ISO/IEC 14496-12: "Information technology - Coding of audio-visual objects - Part 12: ISO base media file"

[8] ISO/IEC 23009-2: "Information Technology – Dynamic adaptive streaming over HTTP (DASH) – Part 2: Conformance and Reference Software"

[9] ISO/IEC 23009-3: "Information Technology – Dynamic adaptive streaming over HTTP (DASH) – Part 3: Implementation Guidelines"

[10] ISO/IEC 23009-4: "Information Technology – Dynamic adaptive streaming over HTTP (DASH) – Part 4: Segment Encryption and Authentication"

[11] 3GPP TS 26.234: "Transparent end-to-end packet switched streaming service (PSS); Protocols and codecs"

[12] 3GPP TS 26.244: "Transparent end-to-end packet switched streaming service (PSS); 3GPP file format (3GP)"

[13] 3GPP TS 26.346: "Multimedia Broadcast Multicast Service (MBMS); Protocols and codecs"

[14] 3GPP TR 26.938: "Improved Support for Dynamic Adaptive Streaming over HTTP in 3GPP"

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.

Energy-Efficient On-Demand Streaming in Mobile Cellular Networks

Yong Cui, Xiao Ma*, Jiangchuan Liu† and Yayun Bao**
**Department of Computer Science and Technology, Tsinghua University, Beijing, P.R. China*
†School of Computer Science, Simon Fraser University, Burnaby, BC, Canada
**cuiyong@tsinghua.edu.cn, max11@mails.tsinghua.edu.cn, jimmybao0730@gmail.com*
†jcliu@cs.sfu.ca

I. Introduction

With the rapid development of wireless access technologies and mobile terminals (e.g. tablet, smartphone), end users are able to enjoy abundant applications with heavy workload on mobile platforms from anywhere at anytime. However, due to limited terminal size and portable requirement, mobile devices are difficult to provide sufficient computation and battery capacity, hence making the Quality of Experience (QoE) of mobile platform applications hard to guarantee [1].

Among the different sorts of heavy-workload applications, media streaming is a representative one [2]. Since media content needs to be transmitted, decoded and played on mobile devices, the energy of both wireless network interface, processor and screen is cost during the entire process. Recent literature has shown that transmission costs up to 50% of total energy on average [3, 4]. In the literature, media content encoding/decoding [5, 6] or traffic shaping [7, 8] are usually employed to reduce transmission energy cost. The former one needs to modify the media encoding/decoding pattern to reduce the amount of transmission or change the traffic pattern in a more energy-efficient way, while the latter one is transparent to both source and destination and is more flexible to multiple video/audio encoding/decoding formats.

In the last one decade, 4G LTE standardized by 3GPP [9] has witnessed a rapid development. With the advantages as high speed, wide coverage range, etc., it has attracted an increasing number of researchers to devote themselves to it. However, recent literature has shown that both 3G and 4G LTE are not adapt for media streaming transmission from energy consumption perspective [10, 11]. The power control mechanisms of 3G (RRC) and 4G (DRX/DTX) both have their inherent limitations that make the wastage of energy even more severe. For instance, the inappropriate setting of the inactivity timer could increase tail energy cost by nearly 60% percentage in 3G network [12]. Moreover, real-environment experiments have shown that the signal strength of cellular radio can also significantly influence the unit transmission energy cost per bit. According to Bartender [13], when the signal is weak, the energy consumed per bit is as much as 6x higher than when it is strong. Therefore, the traffic scheduling objective is to make transmission happens during superior network condition and to avoid too many tail states.

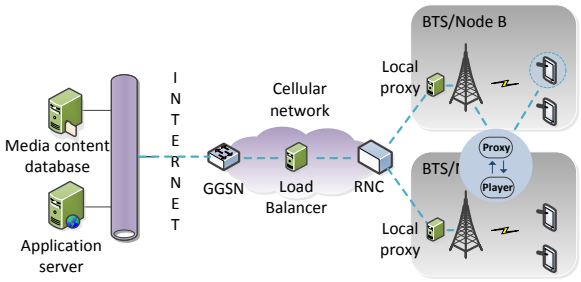


Figure 1. System Architecture

In this paper, we consider energy efficient design of media streaming applications in mobile cellular networks. Particularly, since the energy saving in our solution is achieved by traffic scheduling, to ensure sufficient schedulable time interval, on demand video streaming is considered as the objective application since the server possesses the entire media content at the beginning of the scheduling process. We employ proxy-based traffic shaping in our solution to avoid transmission during high energy-consuming radio state and to reduce unnecessary occurrence of tail energy state. To ensure high quality of experience (QoE) of streaming user, we propose a deadline-based strategy. However, since the performance is based on accurate signal strength prediction and it may cause too many tail states, we further employ Lyapunov Optimization as a tool to solve multi-client traffic scheduling problem to achieve the objective of minimizing total energy cost and guaranteeing users' QoE simultaneously. Simulation result has shown that our scheme can achieve a feasible tradeoff between the QoE of media streaming applications and the energy saving of mobile clients.

II. System Architecture

In this section, we intend to illustrate the basic idea of our scheduling mechanism: traffic shaping by proxy between media streaming server and client to make transmission happens during superior network condition; and control the number of bursts to avoid

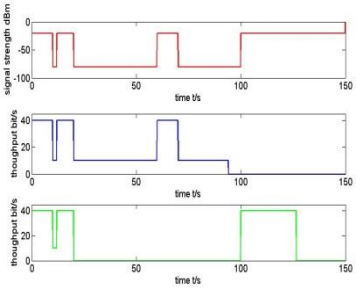


Figure 2. An example of traffic shaping

too much energy wasted by tail state. Figure 1 illustrates our basic system architecture. It mainly consists of a Load Balancer (LB), several Local Proxies (LPs) and client-side proxy (CP) locating on each mobile client.

Local proxy locates behind each base station (BTS)/Node B and is responsible for shaping the streaming traffic directed to each client within the coverage area of the BTS according to the signal strength of each client. Therefore, client-side proxy should transmit its signal strength variation to the local proxy hence the proxy can make scheduling decision in a more energy-efficient way. Since caching streaming data on local proxy may cause congestion from network operator perspective, a load balancer is involved in mobile backbone network to dynamically schedule traffic among different BTSs.

While deferring traffic to avoid bad-signal periods is a simple idea, it has many potential problems. First, media streaming is delay-sensitive, how to ensure user's quality of experience during the entire scheduling process is a problem. Second, after traffic shaping, the original continuous traffic may be broken into several traffic bursts, which may cause more time of cellular radio spent on tail state. Third, consider multiple clients locating within the coverage area of one base station and require media content simultaneously, the signal strength based scheduling may lead to traffic distribution imbalance on base station hence the channel resource is not sufficiently utilized. In this paper, we intend to design a longterm online optimization to achieve high QoE and low energy consumption and make the system stable.

Figure 2 provides a simple example of our traffic shaping mechanism. The sub-figure on the top is the signal strength variation from 0s to 180s. Without traffic shaping, the entire transmission may perform like the figure in the middle. Since signal strength also has an impact on throughput, the variation trend of throughput is consistent with that of signal strength. There is only one tail state since the transmission is not interrupted. If traffic shaping is enabled, the transmission from 20s to 100s is deferred to avoid bad-signal period, two tail states are incurred. If the energy cost by one additional tail state is lower than the energy wasted by transmitting during bad-signal period and the QoE can be guaranteed throughout the entire process, traffic shaping obtains positive profit. Furthermore, if the user is extremely power hungry, the energy saving can be achieved even more under sacrificing a portion of QoE.

III. Solution Analysis

During the entire transmission process on local proxy, the cached media content cannot be infinitely delayed since otherwise the user-side playback may be stuttered. If the video/audio playback rate is obtained by local proxy, the deadline of each data packet can be predicted by the proxy. We define the schedulable time duration of a data packet to be the time from entering local proxy to the deadline. Therefore, a simple strategy would be to choose the time slot within the schedulable time duration in which the signal strength is better. However, this strategy may need to predict the signal strength in the near future, the effect highly relies on the accuracy of the prediction. Moreover, it does not consider the overhead brought in by additional tail states (since one continuous flow is broken into multiple bursts). In the literature, there are several related work following this prediction-based strategy [13, 14]. Since the strategy is not practical and has many disadvantages, we try to propose an online traffic scheduling mechanism that only relies on current state information.

Assume there are N $(1, 2, ..., N)$ mobile clients located within the coverage area of one BTS. To model the transmission process of client i, we denote the amount of data on the local proxy to be transmitted to client i at the moment t to be $Q_i(t)$, the amount of data belonging to client i that arrives proxy at the moment t to be $A_i(t)$. Since the throughput and signal strength vary with time, we denote the throughput at the moment t to be $w_i(t)$, the transmission energy cost per bit to be $P_i(t)$. Therefore, the decision should be made to determine whether transmitting data to client i at time slot t or not. Define a decision parameter $\alpha_i(t)$ to be,

$$\alpha_i = \begin{cases} 1 & transmit\ Q_i(t) \\ 0 & idle \end{cases} \quad (1)$$

Therefore, we can define the amount of data that leaves local proxy and prepares to transmit to client i to be $b_i(t) = \alpha_i(t) w_i(t) \tau$, Where τ is the duration of one time slot. Hence, the energy cost of client i at time slot t can be denoted by $E_i(t) = \alpha_i(t)(P_i(t)\tau + P_{tail} t_{interval})$, where P_{tail} is the average tail power and $t_{interval} \in (0, T_{tail})$ is the

duration of the tail state. Note that we both consider the transmission energy and tail energy cost.

As a result, the average energy cost of the N clients over a long time duration can be depicted as follows,

$$\overline{E} = \limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{i=1}^{N} E_i(t) \quad (2)$$

To depict the requirement of user's QoE, the average queue length *Q(t)* of the *N* clients throughout a long time duration should be stable,

$$\overline{Q} = \limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{i=1}^{N} Q_i(t) < \infty \quad (3)$$

Having the equations above, our energy-QoE tradeoff problem can be formulated by:

$$\min \ E \quad s.t.\ Q < \infty \quad (4)$$

To achieve system stability in the long run, we employ Lyapunov optimization in control theory to address the problem formulated by Eq.4. We first define a Lyapunov function $L(Q(t)) = \frac{1}{2}\sum_{i=1}^{N} Q_i^2(t)$ to depict the congestion level of the queue. The larger the *L* function is, the more the clients will be suffered from congestion. To keep the system stable, we involve Lyapunov drift $\delta(Q(t))$ to depict the difference of the *L* functions between two adjacent time slots,

$$\delta(Q(t)) = L(Q(t+1)) - L(Q(t)) \quad (5)$$

Therefore, we can minimize the following equation to tradeoff between delay and energy,

$$\delta(Q(t)) + VE(t) \quad (6)$$

Where *V* is a non-negative weight that is chosen as desired to affect an energy-delay tradeoff. The more the V is, the more the energy saving effect will be and vice versa.

By bounding the upper-bound of Eq.6, it can be derived that we can minimize the following simplified equation to minimize the upper-bound of Eq.6,

$$VE(t) - \sum_{i=1}^{N} Q_i(t) b_i(t) \quad (7)$$

As a result, the traffic scheduling algorithm can be easily designed based on the analysis above. At every time slot *t*, choose an appropriate decision vector $\alpha(t) = \{\alpha_1(t), \alpha_2(t), ..., \alpha_N(t)\}$ to minimize Eq.7. At the next time slot *t+1*, update $Q_i(t+1) = Q_i(t) + A_i(t) - b_i(t)$ based on the decision made before. The QoE and energy saving can be simultaneously well-guaranteed since Lyapunov optimization has provided such good characteristics.

IV. Simulation

In this section, we evaluate our scheme on C++ platform. The throughput at the beginning of each time slot follows a uniform distribution on [0, 200*bps*]. The number of users is *N = 20*. The average tail power is set to 0.7*W*. The duration of each time slot τ is 100*ms* and data packet size is 100*bits*.

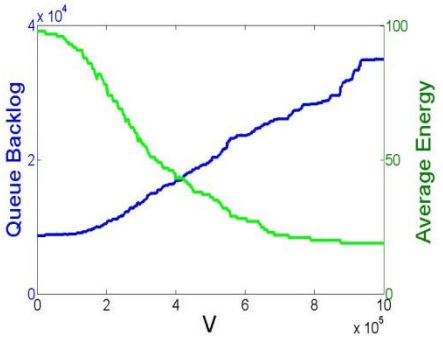


Figure 3. The influence of tradeoff parameter V

Figure 3 illustrates the variation of the time-averaged queue backlog and the average energy consumption with the change of tradeoff parameter *V*. We can see from the figure that the average energy consumption falls quickly at the beginning and then tends to decrease slowly while the time-averaged queue backlog grows linearly with *V*. It verifies our theoretical analysis that the parameter *V* can significantly influence the tradeoff between energy saving and QoE of media streaming application.

V. Conclusion

Mobile streaming service has become increasingly popular on mobile handset platforms in the last one decade. However, a large amount of energy is wasted during media content transmission due to the inherent limitations of power control mechanisms of cellular radio. This paper introduces an energy-aware traffic shaping mechanism that saves media streaming energy consumption by re-scheduling traffic to avoid transmission during bad network condition and to decrease tail energy wastage. To achieve load-balancing from network operator perspective, we develop a network-wide long-term optimization mechanism in an online fashion to avoid rush hour on BTS/node B in cellular networks. Simulation results have approved that our approach is able to achieve a feasible tradeoff between the QoE of media streaming applications and the energy saving of mobile clients.

References

[1] N. Thiagarajan, G. Aggarwal, A. Nicoara, D. Boneh, and J. P. Singh, "Who killed my battery?: analyzing mobile browser energy consumption," in Proceedings of the 21st international conference on World Wide Web. ACM, 2012, pp. 41–50.

[2] R. Trestian, A.-N. Moldovan, O. Ormond, and G.-M. Muntean, "Energy consumption analysis of video

streaming to android mobile devices," in Network Operations and Management Symposium (NOMS), 2012 IEEE. IEEE, 2012, pp. 444–452.

[3] M. Hoque, M. Siekkinen, and J. Nurminen, "Energy efficient multimedia streaming to mobile devices: A survey," 2012.

[4] H. Zhu and G. Cao, "On supporting power-efficient streaming applications in wireless environments," Mobile Computing, IEEE Transactions on, vol. 4, no. 4, pp. 391–403, 2005.

[5] Y. Eisenberg, C. E. Luna, T. N. Pappas, R. Berry, and A. K. Katsaggelos, "Joint source coding and transmission power management for energy efficient wireless video communications," Circuits and Systems for Video Technology, IEEE Transactions on, vol. 12, no. 6, pp. 411–424, 2002.

[6] S. Mehrotra, W.-g. Chen, and K. Kotteri, "Low bitrate audio coding using generalized adaptive gain shape vector quantization across channels," in Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on. IEEE, 2009, pp. 9–12.

[7] S. Sharangi, R. Krishnamurti, and M. Hefeeda, "Energy-efficient multicasting of scalable video streams over wimax networks," Multimedia, IEEE Transactions on, vol. 13, no. 1, pp. 102–115, 2011.

[8] C.-H. Hsu and M. M. Hefeeda, "Broadcasting video streams encoded with arbitrary bit rates in energy-constrained mobile tv networks," Networking, IEEE/ACM Transactions on, vol. 18, no. 3, pp. 681–694, 2010.

[9] "3gpp lte," http://www.3gpp.org/LTE.

[10] F. Qian, Z. Wang, A. Gerber, Z. M. Mao, S. Sen, and O. Spatscheck, "Top: Tail optimization protocol for cellular radio resource allocation," in Network Protocols (ICNP), 2010 18th IEEE International Conference on. IEEE, 2010, pp. 285–294.

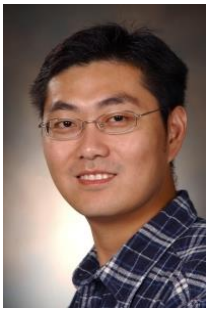
[11] J. Huang, F. Qian, A. Gerber, Z. M. Mao, S. Sen, and O. Spatscheck, "A close examination of performance and power characteristics of 4g lte networks," in Proceedings of the 10th international conference on Mobile systems, applications, and services. ACM, 2012, pp. 225–238.

[12] N. Balasubramanian, A. Balasubramanian, and A. Venkataramani, "Energy consumption in mobile phones: a measurement study and implications for network applications," in Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference. ACM, 2009, pp. 280–293.

[13] Schulman, V. Navda, R. Ramjee, N. Spring, P. Deshpande, C. Grunewald, K. Jain, and V. N. Padmanabhan, "Bartendr: a practical approach to energy-aware cellular data scheduling," in Proceedings of the sixteenth annual international conference on Mobile computing and networking. ACM, 2010, pp. 85–96.

[14] S. Chandra and A. Vahdat, "Application-specific network management for energy-aware streaming of popular multimedia formats." in USENIX Annual Technical Conference, General Track, 2002, pp. 329–342.

Yong Cui, Ph.D., Professor in Tsinghua University, Council Member in China Communication Standards Association, Co-Chair of IETF IPv6 Transition WG Softwire. Having published more than 100 papers in refereed journals and conferences, he is also the winner of Best Paper Award of ACM ICUIMC 2011 and WASA 2010. Holding more than 40 patents, he is one of the authors in RFC 5747 and RFC 5565 for his proposal on IPv6 transition technologies. His major research interests include mobile wireless Internet and computer network architecture.



Jiangchuan Liu is an Associate Professor in the School of Computing Science at Simon Fraser University, British Columbia, Canada. His research interests are in networking, in particular, multimedia communications, peer-to-peer networking and cloud computing. He is a co-recipient of IEEE IWQoS'08 and IEEE/ACM IWQoS'2012 Best Student Paper Awards, IEEE Communications Society Best Paper Award on Multimedia Communications 2009, IEEE Globecom'2011 Best Paper Award, and ACM Multimedia'2012 Best Paper Award. He is a recipient of the NSERC 2009 Discovery Accelerator Supplements (DAS) Award, and a recipient of 2012 Research Excellence Award from SFU Faculty of Applied Science.

A Marketplace for Mobile Applications Supporting Rich Multimedia Feeds

Ngoc Do[1], Ye Zhao[1], Cheng-Hsin Hsu[2], and Nalini Venkatasubramanian[1]
[1] Department of Information and Computer Science, University of California, Irvine, USA
[2] Department of Computer Science, National Tsing Hua University, Hsin-Chu, Taiwan
{nmdo, yez, nalini}@ics.uci.edu, chsu@cs.nthu.edu.tw

1. Introduction

Mobile devices are pervasive today; multimedia applications executing on smartphones and tablets are also commonplace. Rich content involving images, voice, audio, video, graphics, and animations is a part and parcel of the mobile experience for a wide range of applications ranging from entertainment to crisis response. The large volumes of information being captured, exchanged, disseminated through wired and wireless networks result in network congestion, packet drops and consequently low Quality of Service/Experience for end-users. Often a single network alone is incapable of supporting a large number of rich feeds.

For example, current cellular providers are not able to support massive live video broadcast of popular sporting events (such as World Cup Soccer games) to a large number of diverse devices. Recent efforts have indicated that combining cellular infrastructures with ad hoc network capabilities offer additional scalability [1]. Similarly, in a disaster situation, surge loads and damages to infrastructure often cause a loss in network capacity when it is critically needed. Multimodal citizen reports through participatory sensing on mobile phones, social media and the Internet can aid situational awareness. The use of multiple networks concurrently has also been shown to help fast dissemination of rich alerts in that situation [2]. The ability to share mobile Internet access, that may be spotty, unavailable or expensive, is critical in each of these cases.

Mobile Internet usage is also influenced by the fact that a large fraction of mobile operators, today, only offer tiered data plans. It is therefore tricky for mobile users to "select" contracts, e.g., (1) light mobile users may want to avoid data plans all together, (2) heavy mobile users may accidently exceed the monthly quotas and be charged at higher rates, and (3) most mobile users may waste their residue quotas every month. Volume-based access plans are generally unsuitable for rich multimedia feeds; the ability to share network access across devices offers additional flexibility to users. However, mobile devices are limited in resources; one such key consumable resource that impacts the desire and ability to share access is the available battery capacity on the mobile host offering the shared access.

While users may be motivated to share mobile Internet access and utilize their local resources in dire situations (e.g., emergencies), users need to be incentivized to share access in more general scenarios. We envision a *marketplace* where mobile users trade their residual data plan quotas over short-range networks, such as Bluetooth and WiFi Direct to enable a more flexible data plan quota usage [3]. Such a marketplace also allows cellular operators to: (1) extend the cellular network coverage and (2) offload some of the traffic load from the crowded cellular networks – the latter is possible because the short-range networks run on different frequency bands causing virtually no interference to the cellular networks and providing additional access networks (which are not managed by cellular operators).

There are multiple challenges in creating the basic functionality to enable such a marketplace that we will discuss in this short article. Firstly, we will highlight a generalized system architecture that spans multiple providers, network types and entities. The entities of this ecosystem include mobile devices, mobile hotspots (those mobile devices providing connectivity to a backbone network for Internet access), brokers, service and content providers. We argue that a control framework that controls low level information flow reliably is required to enable shared access – we believe that Lyapunov based control theoretic framework can provide a good basis for this. In this short article, we also discuss non-functional challenges that dictate the viability of the proposed scheme – security, pricing and payment are some key issues.

2. An Architecture and Control Framework for Enabling Shared Mobile Internet Access

Figure 1 illustrates the high-level architecture of the considered marketplace, in which mobile users who need Internet access, called *mobile clients*, hire nearby mobile users, called *mobile hotspots*, to transport mobile data for a small fee. As an illustrative example, mobile clients C1 and C2 hire mobile hotspots H1 and H2 for Internet access. To join the system, mobile users

register at a proxy and billing server, which is managed by cellular operators or third party companies. The mobile clients make a monetary deposit to the proxy and billing server before they can gain Internet access from mobile hotspots. They are charged for their data usage transferred through the mobile hotspot's data connection. For each request from a mobile client, the mobile hotspot may charge the client three fees: (1) data plan fee: for the used cellular quota, (2) resource fee: for the local resources, such as energy and storage, and (3) SLA fee: for setting up a Service-Level Agreement (SLA) with the cellular operators for transferring data plan quotas. The considered marketplace works for various mobile applications, e.g., video upload/download, video streaming, Web browsing, and Online Social Network (OSN) updates.

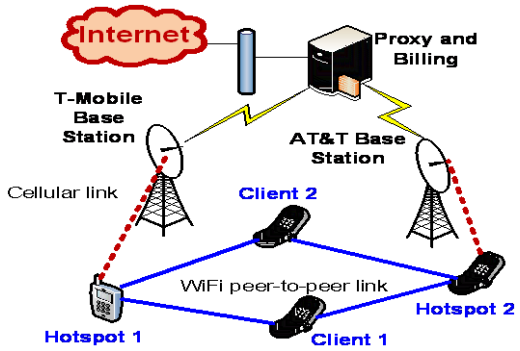


Figure 1 The proposed marketplace.

A Potential Solution using the Lyapunov Framework. To realize the proposed marketplace, a control mechanism that allows for reliable exchange of content between devices is essential. We present high-level software architecture in Figure 2 to enable the content exchange. To illustrate the flow of information in a concrete scenario, we consider video upload applications, in which each video is divided into multiple segments to better adapt to the network dynamics (the intuition here is similar to that of Dynamic Adaptive Streaming over HTTP (DASH) [4]). When a mobile client wants to upload a video, it first sends a request for each video segment and hires a mobile hotspot to transfer the segment to the Internet. A mobile hotspot invokes the *Client Request Admission* module to decide if it would admit the request based on the mobile hotspot's current workload and optimization objectives (revenue maximization for example). Then the mobile hotspot sends a reply to the client with a segment transfer delay and a cost/price to serve the request. The client may receive multiple replies from surrounding mobile hotspots. The client uses the *Hotspot Selection* module to choose the hotspot with the most preferred trade-off between segment transfer delay and cost, and transmits the video segment to the mobile hotspot. The incoming video segments transmitted via the *Data Transfer* module at both mobile hotspot and client. The mobile hotspot and client also employ the AAA (Authentication, Authorization and Accounting) module for secured connection establishment and payment.

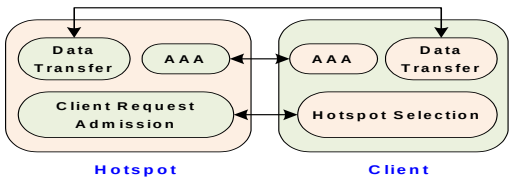


Figure 2 Software components.

One of the key functionalities in the system is provided by the Client Request Admission module running on each mobile hotspot, which admits or rejects the incoming requests from multiple mobile clients in order to: (a) maximizing the long term revenue (measured as average revenue over time), (b) ensuring overall stability of the system (implying no buffer overflow instances), and (c) providing a distributed and practical implementation. We develop the admission control algorithm using a Lyapunov optimization framework. It makes admission decisions based on the characteristics of the incoming requests, their potential to generate increased revenue, and the current set of ongoing commitments made by the mobile hotspot. The Lyapunov approach provides a meaningful theoretical underpinning for stability analysis of the dynamic execution environment [5].

3. Potential Research Challenges

There are plenty of other challenges to make the data plan marketplaces into reality. We briefly discuss some of them in the following.

Multihoming support for multimedia applications. Multimedia applications require low delay and high bandwidth; a difficult challenge for cellular networks. One promising approach is allowing mobile devices to hire multiple nearby mobile APs and WiFi APs for higher aggregate bandwidth, more stable connectivity, and lower latency. Concurrently leveraging multiple access networks is known as multihoming in the literature, e.g., for high-quality video streaming [6]. However, further study is required to efficiently apply the multihoming techniques in data plan marketplaces. Moreover, for real-time multimedia applications, it is desired to have a comprehensive control framework for timely exchanges of delay-sensitive multimedia feeds.

Dynamic pricing. Instead of assuming that each mobile hotspot owner will manually set a price, a

possible approach is to have a dynamic pricing mechanism based on residual traffic quotas, battery levels, network congestion levels, and degree of competition. For example, when a mobile AP's residual traffic quota is high, the owner may be willing to sell the service at a lower rate, compared to another mobile AP that has almost used up its dataplan quota. A dynamic pricing mechanism, perhaps based on game theory, will allow mobile hotspots to adapt prices based on their conditions. Note that embedding the game theoretic solution within a real system is not necessarily straightforward. Additionally, the lack of popular micro-payment mechanisms may slow down the deployment of data plan marketplaces. We believe that a credit-based solution may be employed initially, and virtual currency mechanisms such as BitCoin [7] and Square [8] should be explored in the longer run.

Mobility support. Mobile clients and hotspots are often moving, the ability to support continued service in spite of this movement is essential in a mobile service marketplace. One possibility is to leverage mobile host trajectories in order to: (i) improve the reliability of mobile Internet access by reducing the number of likely disconnections during data transfers and (ii) increase the performance of mobile Internet access by performing proactive handoff operations. We envision a distributed technique for achieving these two goals: (i) a lightweight client that runs on individual mobile devices to collect local device conditions and the neighboring network environment and (ii) optimization logics that run on a broker for optimal decisions to adapt to device mobility.

Security support. Several practical security mechanisms, such as encryption and digital signatures [9], can be applied in the data plan marketplaces to avoid data manipulation by malicious mobile APs. Integrating these security mechanisms is no easy task as mobile devices are resource-constrained, and the overhead of adding potentially complex security mechanisms must be taken into consideration. A scalable mechanism that allows users to choose the most appropriate security level depending on their residual resources and the nature of the data transfers is desirable. Another key open issue is that concerning user privacy. For example, a mobile device may not want to reveal its geographical location, but selecting a mobile AP inherently indicates that this mobile device is very close to that mobile AP. Mechanisms to keep mobile devices (and mobile APs) anonymous for better privacy is an interesting direction of research.

In this short article, we present our vision of building up a marketplace where mobile devices trade their resources and residual data plan quotas. Other types of resources may also be traded among resource-constrained mobile devices, and more complex ecosystems can be gradually built up. For example, a mobile device with abundant battery level may sell computational power to near-by mobile devices, or even provide them a wireless charging service for a small fee. Similarly, public spaces (e.g., malls, airports) today deploy expensive WiFi network infrastructures to provide the temporary occupants with Internet access; one can envision offering incentives (e.g., coupons, discounts) to those mobile devices that volunteer to serve as mobile hotspots in this case. In general, we see a great potential in creating mobile marketplaces–however, there are many challenges that need to be addressed before such ecosystems can be widely deployed and accepted.

References

[1] N. Do, C. Hsu, J. Singh, and N. Venkatasubramanian, "Massive Live Video Distribution Using Hybrid Cellular and Ad Hoc Networks," in Proc. of IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM'11), Lucca, Italy, June 2011, pp. 1–9.

[2] N. Do, C. Hsu, and N. Venkatasubramanian, "HybCAST: Rich Content Dissemination in Hybrid Cellular and 802.11 Ad Hoc Networks," in Proc. of IEEE International Symposium on Reliable Distributed Systems (SRDS'12), Irvine, CA, October, 2012, pp. 352–361.

[3] N. Do, C. Hsu, and N. Venkatasubramanian, "CrowdMAC: A Crowdsourcing System for Mobile Access," in Proc. of ACM/IFIP/USENIX International Conference on Middleware (Middleware'12), Montreal, Canada, December 2012, pp. 1–20.

[4] T. Stockhammer, "Dynamic Adaptive Streaming over HTTP - Standards and Design Principles," in *Proc. of MMSys*, pp. 133-144, 2011.

[5] L. Georgiadis, M. Neely, and L. Tassiulas. Resource Allocation and Cross-Layer Control in Wireless Networks. Foundations and Trends in Networking, 1(1-144): 752–764, April 2006.

[6] N. Freris, C. Hsu, J. Singh, and X. Zhu, "Distortion-aware scalable video streaming to multi-network clients," IEEE/ACM Transactions on Networking, vol. 21, no. 2, pp. 469-481, April 2013.

[7] "Bitcoin web page," http://bitcoin.org, 2013.

[8] "Starbucks and Square: Creating a virtual currency," http://money.msn.com/technology-investment/post.aspx?post=28c7d2d6-8d8e-4cf5-aace-9d613b0629d9", 2013.

[9] Stallings, Cryptography and Network Security: Principles and Practices, 3rd ed. Prentice Hall, 2003

Ngoc Do is currently a PhD candidate at University of California Irvine. He worked as a research intern at Deutsche Telekom R&D Labs in 2010, and Alcatel-Lucent Bell Labs in 2011. His research interests include multimedia, wireless communication, mobile computing, crowdsourcing, social networks and algorithms.

Ye Zhao received B.S. from the School of Telecommunication Engineering at Beijing University of Posts and Telecommunications, China, in 2004. He received M.S. from the department of Electrical and Electronics Engineering at Imperial College London, United Kingdom, in 2005. He is currently a PhD candidate in the Department of Information and Computer Science, University of California Irvine. His research interests include online social networks, content dissemination with distributed and P2P systems, wireless and mobile systems.

Cheng-Hsin Hsu received the Ph.D. degree from Simon Fraser University, Canada in 2009, the M.Eng. degree from University of Maryland, College Park in 2003, and the M.Sc. and B.Sc. degrees from National Chung Cheng University, Taiwan in 2000 and 1996. He is an Assistant Professor at National Tsing Hua University, Taiwan. He was with Deutsche Telekom Lab, Lucent, and Motorola. His research interests are in the area of multimedia networking and distributed systems. He and his colleagues won the Best Technical Demo Award in ACM MM'08, Best Paper Award in IEEE RTAS'12, and TAOS Best Paper Award in IEEE GLOBECOM'12. He served as the TPC Co-chair of the MoVid'12 and MoVid'13, the Proc. and Web Chair of NOSSDAV'10, and on the TPCs of ICME, ICDCS, ICC, GLOBECOM, MM, MMSys, and NOSSDAV.

Nalini Venkatasubramanian is currently a Professor in the School of Information and Computer Science at the University of California Irvine. She has had significant research and industry experience in the areas of distributed systems, adaptive middleware, pervasive and mobile computing, distributed multimedia and formal methods and has published extensively in these areas. As a key member of the Center for Emergency Response Technologies at UC Irvine, She is the recipient of the prestigious NSF Career Award, an Undergraduate Teaching Excellence Award from the University of California, Irvine in 2002 and multiple best paper awards. Prof. Venkatasubramanian has served in numerous program and organizing committees of conferences on middleware, distributed systems and multimedia and on the editorial boards of journals. She received and M.S and Ph.D in Computer Science from the University of Illinois in Urbana-Champaign. Her research is supported both by government and industrial sources such as NSF, DHS, ONR, DARPA, Novell, Hewlett-Packard and Nokia. Prior to arriving at UC Irvine, Nalini was a Research Staff Member at the Hewlett-Packard Laboratories in Palo Alto, California.

Adaptive Streaming in Mobile Cloud Gaming

Shervin Shirmohammadi[1,2]
[1] *Distributed and Collaborative Virtual Environments Research (DISCOVER) Lab*
University of Ottawa, Canada, shervin@eecs.uottawa.ca
[2] *Machine Learning Lab (MLL)*
Istanbul Sehir University, Turkey, shervinshirmohammadi@sehir.edu.tr

1. Introduction

Cloud gaming leverages the well-known concept of cloud computing to provide online gaming services to players, including mobile players who have more computational or download restrictions than players with dedicated game consoles or desktop computers. The idea in cloud gaming is to process the game events in the cloud and to stream the game to the players. Cloud gaming can be single player, where a user plays the game on his/her own, or multiplayer, where multiple geographically distributed users play with or against each other. Since it uses the cloud, scalability, server bottlenecks, and server failures are alleviated in cloud gaming, helping it become more popular in both research and industry, with companies such as OnLive [1], StreamMyGame [2], Gaikai [3], G-Cluster [4], OTOY [5], Spoon [6], CiiNOW [7], and others providing commercial cloud gaming services.

One of the challenges in cloud gaming is adaptive streaming of the game to the players. By adaptive, we mean that the server has to adapt the game content to the characteristics and limitations of the underlying network or client's end device. These include limitations in the available network bandwidth, or limitations in the client device's processing power, memory, display size, battery life, or the user's download limits as per his/her mobile subscription plan. While some of these restrictions are becoming less problematic due to rapid progress in mobile hardware technologies, battery life in particular and download limit to some extent are still problems that must be seriously considered. Also, consuming more bandwidth or computational power means consuming more battery. In this article, we present some approaches for adaptive streaming in cloud gaming to reduce battery, processing, and bandwidth consumption. We start this by looking at the types of streaming that can be done in cloud gaming.

2. Game Streaming in the Cloud

There are currently three types of game streaming in the cloud: graphics streaming, video streaming, and their combination.

In graphics steaming, the game objects are represented by 3D models and textures, and these are streamed to players as needed. Rendering of the game is done at the client, but the game logic runs in the cloud and the state of the game (position and orientation of objects, as well as actions and events) is streamed to clients as an update message every 10 to 280 msec, depending on game genre and activity levels [8]. The advantage of graphics streaming is that except for the textures, the 3D object models and the update messages are small and do not require much bandwidth. So once textures are received, which happens only one time, game streaming uses very little bandwidth. Depending on the client's storage capacity and the nature of the game, textures can be even preloaded at the client when the game is installed, reducing bandwidth consumption even more during game play.

In video streaming, the cloud not only executes the game logic, but also the game rendering. The resulting game scene is then streamed to clients as video, typically in the 20 frames per second range, and in some cases up to 50 frames per second [9]. The advantage here is that as long as the client can display video, which pretty much all smartphones and tablets and most other mobile devices today do, the user can play the game without needing 3D graphics rendering hardware or software. The disadvantage is that video is much bulkier than 3D graphics or updates, and requires substantial bandwidth [10], and although bandwidth is becoming more affordable, the battery life problem due to bandwidth consumption, mentioned in section 1, has to be dealt with.

It is also possible to use a hybrid approach and to simultaneously mix graphics streaming with video streaming, as is done in CiiNO, for example.

With the above explanations in mind, let us now present some approaches for doing adaptive streaming in cloud gaming for graphics steaming and video streaming, respectively.

3. Adaptations in Graphics Streaming

The network and client limitations mentioned in section 1 essentially imply that we can stream only a limited amount of information at a time. Streaming more information than that will not be possible since either the network cannot transport it or the client does not have enough capacity for it. As such, one approach to

adaptive streaming is to prioritize game objects and to first stream the most important objects in the context of gameplay. Traditionally, this has been accomplished by using Area of Interest management or similar region based techniques, whereby objects that are closer to the player or within the player's viewing scope, strictly in terms of distance, are streamed first and with higher quality [11][12]. But such approaches do not consider the context under which the game is played. For example, when fighting an enemy in a jungle, the enemy has a higher priority than the trees and surrounding bushes, and should be updated with higher quality in terms of both resolution and update frequency, even if the enemy is further away from the bushes and trees. Distance-based approaches on the other hand would simply render with higher quality whatever is closer to the player, which is not necessarily the most important object for the player. Therefore, it is more logical to adopt a context aware approach whereby we classify actions in a game as activities like walking, running, aiming, shooting, etc., and we determine how important a specific game object is in the context of that activity, building a one-time apriori importance matrix [13]. Using this matrix, an object selection threshold can then be set, according to the bandwidth or client limitations (for example download limit), so that less important objects in the scene compared to the threshold will not be streamed, freeing up resources for more important objects. It has been shown that, compared to traditional distance based approaches, such an approach leads to a higher quality of gameplay in terms of game score [13] and that the object selection threshold can be set automatically using optimization [14].

Of course not streaming less important objects at all could have a negative effect in the visual quality of the game, as can be seen in Figure 1(I). It is therefore preferable to still stream the less important objects, but with a lower graphics quality, as shown in Figure 1(II).

To stream an object, we must consider that an object consists of two elements: 3D mesh model, and image textures. 3D mesh models are relatively small and in most cases can be downloaded quickly. In cases where the mesh model itself is large, progressive mesh streaming approaches can be used such that with each update message, the visual quality of the object increases [15].

Textures are the most bandwidth and rendering consuming elements of objects. To facilitate texture streaming, we can design adaptive approaches that would stream textures efficiently according to the mobile device's battery life and/or bandwidth restrictions and the importance of textures, with textures having different importance levels. For instance, in the previous enemy in the jungle example, the textures of the enemy can have importance level 1 (most important), bushes and trees can have importance level 2, and the sky with clouds or the water in the lake can have importance level 3 (least important). We can then design an optimization algorithm that would stream textures according to their importance and bandwidth/client restrictions, in a progressive manner so that with each update message the resolution and quality of the texture improves up to a certain threshold determined by the said restrictions [16].

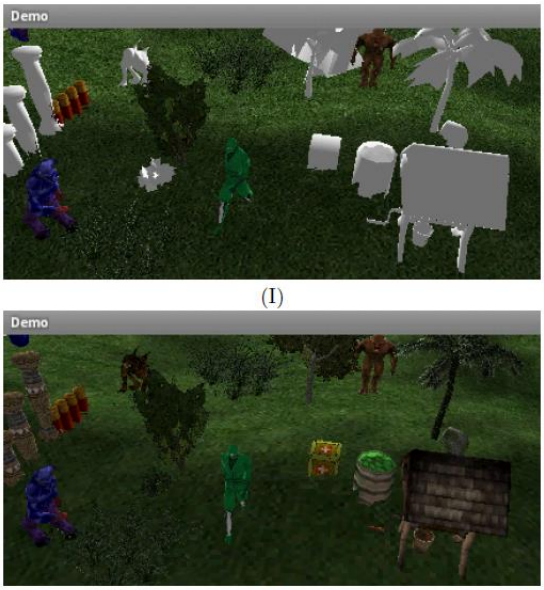


Figure 1. Game scene with the less important objects (I) not screamed at all (indicated in white), and (II) streamed with lower quality. [16]

We can further save battery life by limiting lighting effects and doing smarter control of game object brightness [17]. Specifically, we can choose the lighting effects that are the most computationally expensive effects, such as specular highlights, reflection, transparency, and shadows, and not apply them to the less important objects. In addition, we can make darker the less important objects and save display energy, hence increasing battery life without significantly affecting the gaming experience. It has been shown that such approaches can lead to 20% to 33% extension of battery life [17].

The above approaches save battery life by smarter streaming of the game graphics. To further save battery life, we can enhance the above approaches with streaming techniques whereby the wireless interface of the mobile device is turned off for a time during which the status of an object can be more or less estimated and

so there is no need to transmit update messages for that object. Turning off the interface saves more battery life. Dead-reckoning has been used for decades in games and simulations to determine when an update message needs to be received for an object. This can be used to design dead-reckoning based game streaming protocols that allow the wireless interface to be turned off when no updates are expected, and turned on only when updates are expected. Such approach can save as much as 36.5% of battery life [19].

4. Adaptations in Video Streaming

We can apply some of the above approaches for graphics streaming to video encoding at the server side, and stream a more efficient video to the client. For example, we can use the concept of object importance and not encode the less important objects in the video, leading to as much as 8% bandwidth saving for the client and 7% reduction of encoding time in the cloud [20]. We can also apply smarter brightness control to the video, such that the less important objects appear darker, saving display energy and increasing battery life with little negative effect on the player's experience.

For a more efficient streaming technique, we can combine the idea of object importance with visual attention models, such as Judd's SVM-based model that has been trained using a large database of eye tracking data. Visual attention models are used when it is essential to understand where humans look at in a scene. In gaming, the player most of the time looks at the most important part of the scene, so such saliency models can be used, in combination with object importance models, to label scene parts with various priority levels. We can then encode each scene part with a different quality level, according to the priority of that part, and stream the resulting video to the client. Specifically, in each frame of the gameplay, we can consider the importance of each object and visual saliency features to decide which regions of the frame are more important for the accomplishment of the player's current activity [21]. Then, we encode each region of the game frame with a different QP value proportional to the importance of that region. To make the technique less time consuming, in practice only the first frames of each GOP can be analyzed to find the appropriate QP values.

This is shown if Figure 2, where the image on the top has been encoded normally with a QP value of 30, but the image at the bottom has been encoded with Saliency + Importance approach using three QP values of 30, 35 and 40 for the high, medium and low importance macro-blocks, respectively. In the figure, S1, P1, and P2 represent the high quality regions of the saliency map, and the high and medium quality regions of the importance map, respectively. It has been shown that such an approach can reduce the streaming bandwidth by as much as 50% [21].

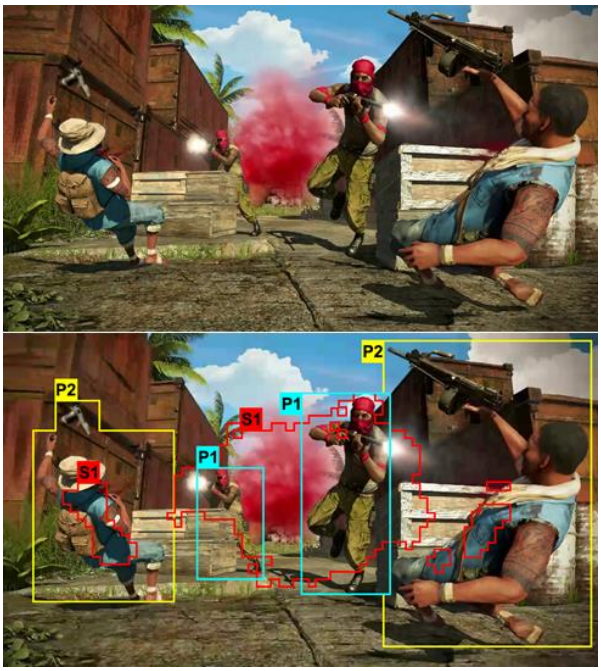


Figure 2. Game frame encoded by a single QP value (top) and with three QP values (bottom). [21]

5. Conclusion

In this article, we gave an overview of different game streaming techniques in the context of mobile cloud gaming. We presented application-level adaptation approaches, specifically for game streaming, which reduce battery or bandwidth consumption and therefore cope with the limited battery life and download limit of mobile players. As cloud gaming become more popular, supporting mobile players becomes more important and approaches such as those presented in this article need to be implemented to improve the quality of mobile gaming.

References

[1] OnLive, [Online] http://www.onlive.com/

[2] Stream My Game, [Online] http://www.streammygame.com/

[3] GaiKai, [Online] http://www.gaikai.com/

[4] G-Cluster, [Online] http://www.gcluster.com/

[5] OTOY, [Online] http://www.otoy.com/

[6] Spoon, [Online] http://www.spoon.net/

[7] CiiNOW, [Online] http://www.ciinow.com/

[8] S. Ratti, B. Hariri, and S. Shirmohammadi, "A Survey of First-Person Shooter Gaming Traffic on the Internet",

IEEE Internet Computing, Vol. 14, No. 5, September/October 2010, pp. 60-69.

[9] C.Y. Huang, C.H. Hsu, Y.C. Chang, and K.T. Chen, "GamingAnywhere: an open cloud gaming system", *Proc. ACM Multimedia Systems*, February 27 to March 1 2013, Oslo, Norway, pp. 36-47.

[10] M. Claypool, D. Finkel, A. Grant, and M. Solano, "Thin to win? Network performance analysis of the OnLive thin client game system," *Proc. ACM/IEEE Workshop on Network and Systems Support for Games (NetGames)*, Venice, Italy, Nov. 22-23 2012, pp. 1-6.

[11] F. Li, R. Lau, D. Kilis, L. Li, "Game-on-demand: An Online Game Engine Based on Geometry Streaming," *ACM Transactions on Multimedia Computing, Communications and Applications*, Vol. 7, Issue 3, August 2011, Article No. 19.

[12] R. Waters, D. Anderson, J. Barrus, D. Brogan, M. Casey, S. McKeown, T. Nitta, I. Sterns, W. Yerazunis, "Diamond Park and Spline: A Social Virtual Reality System with 3D Animation, Spoken Interaction, and Runtime Modifiability", *Presence*, 6(4):461-480. 1997.

[13] H. Rahimi, A. Nazari, and S. Shirmohammadi, "Activity-Centric Streaming of Virtual Environments and Games to Mobile Devices", *Proc. IEEE Symposium on Haptic Audio Visual Environments and Games*, Qinhuangdao, Hebei, China, October 15-16 2011, pp. 45-50.

[14] M. Hemmati, S. Shirmohammadi, H. Rahimi, and A. Nazari, "Optimized Game Object Selection and Streaming for Mobile Devices", *International Conference of Information Science and Computer Applications*, Bali, Indonesia, November 19-20 2012, in *Advances in Information Technology and Applied Computing*, Volume 1, pp. 144-149.

[15] W. Cheng, W.T. Ooi, S. Mondet, R. Grigoras, and G. Morin, "Modeling progressive mesh streaming: Does data dependency matter?"*ACM Transactions on Multimedia Computing, Communications and Applications*, 7(2), Article No. 10 , February 2011.

[16] M. Hosseini, J. Peters, and S. Shirmohammadi, "Energy-Budget-Compliant Adaptive 3D Texture Streaming in Mobile Games", *Proc. ACM Multimedia Systems*, Feb 27 to Mar 1 2013, Oslo, Norway, pp. 1-11.

[17] M. Hosseini, A. Fedorova, S. Shirmohammadi, and J. Peters, "Energy-Aware Adaptations in Mobile 3D Graphics", *Proc. ACM Multimedia*, October 29-November 2 2012, Nara, Japan, pp. 1017-1020.

[18] M. Hosseini, D.T. Ahmed, and S. Shirmohammadi, "Adaptive 3D Texture Streaming in M3G-based Mobile Games", Proc. ACM Multimedia Systems, Chapel Hill, North Carolina, USA, February 22-24 2012, pp. 143-148.

[19] C. Harvey, A. Hamza, C. Ly, and M. Hefeeda, "Energy-Efficient Gaming on Mobile Devices using Dead Reckoning-based Power Management", *Proc. ACM/IEEE Workshop on Network and Systems Support for Games (NetGames)*, Taipei, Taiwan, November 2010, 6 pages.

[20] M. Hemmati, A Javadtalab, A. Nazari, S. Shirmohammadi, and T. Arici, "Game as Video: Bit Rate Reduction through Adaptive Object Encoding", *Proc. ACM Workshop on Network and Operating Systems Support for Digital Audio and Video*, February 27 2013, Oslo, Norway, pp. 7-12.

[21] H. Ahmadi, S. Khoshnood, M.R. Hashemi, and S. Shirmohammadi, "Efficient Bitrate Reduction Using A Game Attention Model in Cloud Gaming", *Proc. IEEE Symposium on Haptic Audio Visual Environments and Games*, Istanbul, Turkey, October 26-27 2013, 6 pages.

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.

INDUSTRIAL COLUMN: SPECIAL ISSUE ON VIDEO AWARE WIRELESS NETWORKS

Guest Editors: Jeffrey R. Foerster, Intel, USA; Michelle Gong, Google, USA
jeffrey.r.foerster@intel.com; michellegong@google.com

Video content is quickly dominating mobile network traffic, and, in fact, 2012 was the first year that mobile video traffic consumed more than 50% of the total traffic over mobile networks, according to Cisco VNI [1]. By 2017, Cisco predicts that 2/3 of all the total mobile traffic will be video. Although it will be important to continue advancing wireless networks, from LTE to LTE-Advanced, as well as network infrastructure from macro-cells to small-cells and increasing use of WiFi off-loading, this will not be sufficient to guarantee a good Quality of Experience (QoE) and cost-effective transmission of this video traffic. For the past three years, Intel, Cisco, and Verizon have been sponsoring university research to find new ways to better manage this video traffic by applying greater intelligence throughout the end-to-end delivery chain and working to identify other industry partners to realize some of these ideas.

This special issue of E-Letter presents some of the latest research results on how wireless networks could be optimized for more efficient delivery of video traffic and how the end QoE can be measured and enhanced. We included a diverse set of papers covering a number of different opportunities, from perceptual-based QoE measurement and monitoring tools to cross-layer optimization to content caching to managing throughput in a heterogeneous network, as well as industry and standards progress to realize a better end-to-end video delivery system for mobile networks.

One of the most challenging problems in video aware networking is to measure end user QoE across multiple devices. In the first paper, titled "Video QoE Models for the Compute Continuum", the authors introduce existing video quality assessment (VQA) models and describe their efforts on estimating QoE via a content and device-based mapping algorithm. The paper also describes a new dynamic system model of time varying subjective quality that captures temporal aspects of QoE.

The second contribution, titled "Perceptual Optimization of Large-Scale Wireless Video Networks", summarizes the authors' recent work on developing new models for perceptual video quality assessment, and using these models to adapt video transmission based on perceptual distortion.

Great progress has been made by various standard organizations. In particular, the third paper, titled "Dynamic Adaptive Streaming over HTTP: Standards and Technology", summarizes the standardization efforts on Dynamic Adaptive Streaming over HTTP (DASH), by Moving Picture Experts Group (MPEG) and Third Generation Partnership Project (3GPP). The authors also describe research challenges and recent findings in optimizing DASH delivery over wireless networks.

In the next paper, titled "Cross-Layer Design for Mobile Video Transmission", the authors consider the robust design of a cross-layer optimized system for arbitrary mobility. This paper shows the trade-off in performance when considering only application level information, only physical layer information, and a combination of both (e.g., a cross-layer approach) for a range of mobile speeds / Doppler spreads. Clearly, for delivering high-quality video to mobile devices, the system must work in a variety of channel conditions.

It's clear that future wireless networks will consist of multiple heterogeneous connections to mobile devices including cellular and WiFi. As a result, it will be important to stream video over the best network connection while also balancing the overall network capacity as well as the delivered QoE. As such, the next paper, titled "Timely Throughput Optimization of Heterogeneous Wireless Networks", addresses this important problem of how to optimally allocate resources to maximize capacity for delay-sensitive video applications over heterogeneous networks.

The significant increase in wireless traffic projected over the next several years will require rethinking the network infrastructure and, in particular, how the network edge as well as mobile devices themselves can help reduce the traffic and/or improve the efficient delivery of the traffic. Simply increasing the number of base stations comes at a significant cost, and, therefore, motivates the need for finding ways to better leverage some of the existing devices and infrastructure. Thus, the final paper in this special issue, titled "Video

Caching and D2D Networks", is focused on caching video content at the edge of the network and even directly on mobile devices which has the potential of significantly increasing the overall capacity of the network for delivering video-on-demand content.

While this special issue is far from a complete coverage on this exciting research area, we hope that the six invited papers provide a sketch of the recent research activities in the area. We would like to thank all the authors for their contribution and hope these articles will stimulate further research on video aware wireless networks.

Jeffrey R. Foerster joined Intel in August 2000 and is currently a Principal Engineer in the Wireless Communications Lab and leads a team focused on Wireless Multimedia Solutions, which includes topics on joint source-channel coding, adaptive streaming, and end-to-end video network optimizations. His past research has included Ultra-wideband (UWB) technology and related regulations, 60 GHz system design, and wireless displays. Jeff has published over 30 papers including journals, magazine, and conferences, and has been an invited panelist and presenter at several conferences. Prior to joining Intel, he worked on Broadband Wireless Access (BWA) systems and standards. He received his B.S., M.S., and Ph.D. degrees from the University of California, San Diego, where his thesis focused on adaptive interference suppression and coding techniques for CDMA systems. Jeff is a Fellow of the IEEE.

Michelle X. Gong is currently working on machine learning and data mining at Google. Prior to joining Google, she worked as a Sr. Research Scientist at Intel Labs and as a System Architect at Cisco Systems. Michelle has explored a wide range of research areas including wireless networking, video over wireless, high-throughput Wi-Fi, 60GHz systems, mobile systems, machine learning, and data mining. She has authored 2 book chapters, 25 conference and journal papers, and about 60 pending or issued patents. She is a TPC member of many IEEE and ACM conferences. She is an editor of ACM Mobile Computing and Communications Review and an editor of IEEE MMTC E-letter. Michelle received her PhD in Electrical Engineering from Virginia Tech in 2005.

Video QoE Models for the Compute Continuum

Lark Kwon Choi[1, 2], Yiting Liao[1], and Alan C. Bovik[2]
[1]Intel Labs, Intel Corporation, Hillsboro, OR, USA
[2]The University of Texas at Austin, Austin, TX, USA
{*lark.kwon.choi, yiting.liao*}*@intel.com, bovik@ece.utexas.edu*

1. Introduction

Video traffic is exponentially increasing over wireless networks due to proliferating video technology and the growing desire for anytime, anywhere access to video content. Cisco predicts that two-thirds of the world's mobile data traffic will be video by 2017 [1]. This imposes significant challenges for managing video traffic efficiently to ensure an acceptable quality of experience (QoE) for the end user. Since network throughput based video adaptation without considering user's QoE could lead to either a bad video service or unnecessary bandwidth waste, QoE management under cost constraints is the key to satisfying consumers and monetizing services [2].

One of the most challenging problems that needs to be addressed to enable video QoE management is the lack of automatic video quality assessment (VQA) tools that estimate perceptual video quality across multiple devices [2]. Researchers have performed various subjective studies to understand essential factors that impact video quality by analyzing compression or transmission artifacts [3], and by exploring dynamic time varying distortions [4]. Furthermore, some VQA models have been developed based on content complexity [5] [6]. In spite of these contributions, user QoE estimation across multiple devices and content characteristics, however, remains poorly understood.

Towards achieving high QoE across the compute continuum, we present recent efforts on automatically estimating QoE via a content and device-based mapping algorithm. In addition, we investigate temporal masking effects and describe a new dynamic system model of time varying subjective quality that captures temporal aspects of QoE. Finally, we introduce potential applications of video QoE metrics, such as quality driven dynamic adaptive streaming over HTTP (DASH) and quality-optimized transcoding services.

2. Improving VQA model for better QoE

VQA models can be generally divided into three broad categories: full-reference (FR), reduced-reference (RR), and no-reference (NR). Some representative high performing algorithms include: MultiScale-Structural SIMilarity index (MS-SSIM) [7] which quantizes "perceptual fidelity" of image structure; Video Quality Metric (VQM) [5] which uses easily computed visual features; Motion-based Video Integrity Evaluation (MOVIE) [8] which uses a model of extra-cortical motion processing; Video Reduced Reference spatio-temporal Entropic Differencing (V-RRED) [9] which exploits a temporal natural video statistics model; and Video BLIINDS [10] which uses a spatio-temporal model of DCT coefficient statistics and a motion coherence model.

The success of the above VQA metrics suggests that disruptions of natural scene statistics (NSS) can be used to detect irregularities in distorted videos. Likewise, modeling perceptual process at the retina, primary visual cortex, and extra-striate cortical areas are crucial to understanding and predicting perceptual video quality [11].

In addition, the quality of a given video may be perceived differently according to viewing distance or display size. Similarly, the visibility of local distortions can be masked by spatial textures or large coherent temporal motions of a video content. In this regard, modern VQA models might be improved by taking into account content and device characteristics. This raises the need to understand QoE for video streaming services across multiple devices, thereby to improve VQA models of QoE across the compute continuum.

3. Achieving high QoE for the compute continuum

How compression, content, and devices interact

To investigate perceived video quality as a function of compression (bitrate and resolution), video characteristics (spatial detail and motion), and display device (display resolution and size), we executed an extensive subjective study and designed an automatic QoE estimator to predict subjective quality under these different impact factors [2].

Fourteen source videos with a wide range of spatial complexity and motion levels were used for the study. They are in a 4:2:0 format with a 1920 × 1080 resolution. Most videos are 10~15 second long, except Aspen Leaves (4s). To obtain a desired range of video quality, the encoding bitrate and resolution sets for each video were chosen to widely range from 110kbps at 448 × 252 to 6Mbps at 1920 × 1080 based on assumed realistic video content and display devices. 80 and 96 compressed videos were displayed on a 42 inch HDTV and four mobile devices (TFT tablet, Amoled phone,

Retina tablet, and Retina phone), respectively, and about 30 participants were recruited for each device to rate the videos by recording opinion score using the single-stimulus continuous quality evaluation (SSCQE) [12] method.

MS-SSIM was used since it delivers excellent quality predictions and is faster than MOVIE or VQM. Figure 1 shows the plots of MS-SSIM against MOS for each device along with the best least-squares linear fit. The Pearson linear correlation coefficient (LCC) between MS-SSIM and MOS is 0.7234 for all data points, while LCC using device-based mapping is 0.8539, 0.8740, 0.7989, 0.8329, and 0.8169 for HDTV, TFT tablet, Amoled phone, Retina tablet, and Retina phone, respectively. Furthermore, device and content-specific mapping between MS-SSIM and MOS shows very high LCC (mean: ~ 0.98) as illustrated in Figure 2. To validate the proposed methods on a different VQA database (DB), we also analyzed the models using the TUM VQA DB [13]. LCC between MS-SSIM and MOS using the device and content-specific mapping for TUM VQA DB shows similar results (mean: ~0.98, standard deviation: 0.016). Results indicate that human perception of video quality is strongly impacted by device and content characteristics, suggesting that device and content-based mapping could greatly improve the prediction accuracy of video quality prediction models.

We then designed a MOS estimator to predict perceptual quality based on MS-SSIM, a content analyzer (spatial detail (S), motion level (M)), a device detector (display type (D), and resolution (R)) [3]. The predicted MOS is calculated as,

$$e\_MOS = \alpha \times \text{MS-SSIM} + \beta \qquad (1)$$

where α and β are functions of the four impact factors S, M, D, and R above. Using the proposed predictor and estimated values of α and β, LCC between the estimated MOS and actual MOS is 0.9861. Future work includes building a regression model to calculate α and β based on the impact factors and extending the video data set to better validate the designed predictor.

Temporal masking and time varying quality

The visibility of temporal distortions influences video QoE. Salient local changes in luminance, hue, shape or size become undetectable in the presence of large coherent object motions [14]. This "motion silencing" implies that large coherent motion can dramatically alter the visibility of visual changes/distortions in video. To understand why it happens and how it affects QoE, we have developed a spatio-temporal flicker detector model based on a model of cortical simple cell responses [15]. It accurately captures the observers'

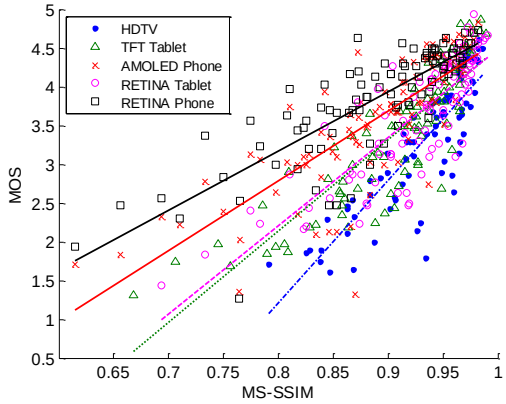


Figure 1 Device-based MS-SSIM and MOS

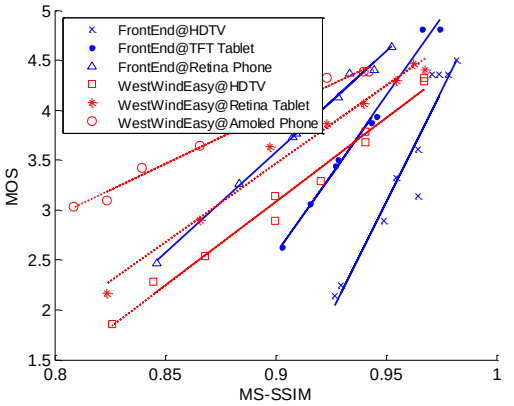


Figure 2 Device and content-based MS-SSIM and MOS mapping

perception of motion silencing as a function of object motion and local changes. In addition, we have investigated the impact of coherent object motion on the visibility of flicker distortions in naturalistic videos. The result of a human experiment involving 43 subjects revealed that the visibility of flicker distortions strongly depends on the speed of coherent motion. We found that less flicker was seen on fast-moving objects even if observers held their gaze on the moving objects [16]. Results indicate that large coherent motion near gaze points masks or 'silences' the perception of temporal flicker distortions in naturalistic videos, in agreement with a recently observed motion silencing effect [14].

Time varying video quality has a definite impact on human judgment of QoE. Although recently developed HTTP-based video streaming technology enables flexible rate adaptation in varying channel conditions, the prediction of a user's QoE when viewing a rate adaptive HTTP video stream is not well understood. To solve this problem, Chao *et al.* have proposed a dynamic system model for predicting the time varying subjective quality (TVSQ) of rate adaptive videos [17]. The model first captures perceptual relevant spatio-temporal features of the video by measuring short time subjective quality using a high-performance RR VQA

model called V-RRED [9], and then employs a Hammerstein-Wiener model to estimate the hysteresis effects in human behavioral responses. To validate the model, a video database including 250 second long time varying distortions was constructed and TVSQ was measured via a subjective study. Experimental results show that the proposed model reliably tracks the TVSQ of video sequences exhibiting time-varying level of video quality. The predicted TVSQ could be used to guide online rate-adaptation strategies towards maximizing the QoE of video streaming services.

Application of video QoE models

Recently developed QoE models open up opportunities to improve cooperation between different ecosystem players in end-to-end video delivery systems, and to deliver high QoE using the least amount of network resources. We have shown that in an adaptive streaming system, DASH clients can utilize quality information to improve streaming efficiency [18]. The quality-driven rate adaptation algorithm jointly optimizes video quality, bitrate consumption, and buffer level to minimize quality fluctuations and inefficient usage of bandwidth, thus achieving better QoE than bitrate-based approaches [18]. Another usage of the QoE model is to allow transcoding services to determine the proper transcoding quality on a content-aware and device-aware fashion. The QoE metric helps the transcoder to achieve the desired QoE without over consuming bandwidth. Furthermore, content-specific and device-specific video quality information may facilitate service providers to design more advanced multi-user resource allocation strategies to optimize overall network utilization and ensure a good QoE for each end user.

Acknowledgment

The authors thank Philip Corriveau and Audrey Younkin for collaboration on the subjective video quality testing.

References

[1] Cisco Systems, Inc., "Cisco visual networking index: Global mobile data traffic forecast update, 2012-2017," Feb. 2013.

[2] Y. Liao, A. Younkin, J. Foerster, and P. Corriveau, "Achieving high QoE across the compute continuum: How compression, content, and devices interact," in *Proc. of VPQM*, 2013.

[3] K. Seshadrinathan, R. Soundararajan, A.C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process.*, vol.19, no.6, pp.1427-1441, June 2010.

[4] K. Moorthy, L. K. Choi, A. C. Bovik, and G. de Veciana, "Video quality assessment on mobile devices: Subjective, behavioral and objective studies," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 652-671, Oct. 2012.

[5] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcast.*, vol. 50, no. 3, pp. 312-322, Sep. 2004.

[6] J. Korhonen and J. You, "Improving objective video quality assessment with content analysis," in *Proc. of VPQM*, 2010.

[7] Z. Wang, E. Simoncelli, and A.C. Bovik, "Multi-scale structural similarity for image quality assessment," *Ann. Asilomar Conf. Signals, Syst., Comput.*, 2003.

[8] K. Seshadrinathan and A.C. Bovik, "Motion-tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. Image Process*, vol. 19, no. 2, pp. 335-350, Feb. 2010.

[9] Soundararajan, and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 4, pp. 684-694, Apr. 2013.

[10] M. Saad and A.C. Bovik, "Blind quality assessment of videos using a model of natural scene statistics and motion coherency," Invited Paper, *Ann Asilomar Conf Signals, Syst, Comput.*, 2012.

[11] A.C. Bovik, "Automatic prediction of perceptual image and video quality," *Proceedings of the IEEE*, to appear, 2013,

[12] ITU-R Rec. BT. 500-11, "Methodology for the subjective assessment of the quality of television pictures," 2002.

[13] C. Keimel, A.Redl and K. Diepold, "The TUM high definition video data sets," in *Proc. of QoMEX*, 2012.

[14] J. W. Suchow and G. A. Alvarez, "Motion silences awareness of visual change," *Curr. Biol*., vol. 21, no. 2, pp.140-143, Jan. 2011.

[15] L. K. Choi, A. C. Bovik and L. K. Cormack, "A flicker detector model of the motion silencing illusion," *Ann. Mtng. Vision Sci. Soc.*, 2012.

[16] L. K. Choi, L. K. Cormack, and A. C. Bovik, "On the visibility of flicker distortions in naturalistic videos," in *Proc. of QoMEX*, 2013.

[17] C. Chen, L. K. Choi, G. de Veciana, C. Caramanis, R. W. Heath Jr, A. C. Bovik, "A dynamic system model of time-varying subjective quality of video streams over HTTP," in *Proc. of ICASSP*, 2013.

[18] Y. Liao, J. Foerster, O. Oyman, M. Rehan, Y. Hassan, "Experiment results of quality driven DASH," *ISO/IEC JTC1/SC29/WG11*, M29247, 2013.

Lark Kwon Choi received his B.S. degree in Electrical Engineering from Korea University, Seoul, Korea, in 2002, and the M.S. degree in Electrical Engineering and Computer Science from Seoul National University, Seoul, Korea, in 2004, respectively. He has worked for KT (formerly Korea Telecom) as a senior engineer from 2004 to 2009 on IPTV platform research and development. He has contributed on IPTV standardization in International Telecommunication Union (ITU-T) and Telecommunications Technology Association (TTA). He is currently pursuing his Ph.D. degree as a member of the Laboratory for Image and Video Engineering (LIVE) at The University of Texas at Austin under Dr. Alan. C. Bovik's supervision. His research interests include image and video quality assessment, spatial and temporal visual masking, video QoE, and motion perception.

Yiting Liao received the B.E. and M.S. degrees in Electronic Engineering from Tsinghua University, China in 2004 and 2006, respectively, and the Ph.D. degree in Electrical and Computer Engineering from the University of California, Santa Barbara, in 2011.

She is currently a Research Scientist with Intel Corporation, Hillsboro, OR. Her current research interests include video quality assessment and user experience evaluation, and video optimization techniques over wireless networks.

Alan C. Bovik (S'80–M'81–SM'89–F'96) holds the Keys and Joan Curry/Cullen Trust Endowed Chair at The University of Texas at Austin, where he is a Professor in the Department of Electrical and Computer Engineering and in the Institute for Neuroscience, and Director of the Laboratory for Image and Video Engineering (LIVE).

He has received a number of major awards from the IEEE Signal Processing Society, including: the Best Paper Award (2009); the Education Award (2008); the Technical Achievement Award (2005), the Distinguished Lecturer Award (2000); and the Meritorious Service Award (1998). Recently he was named Honorary Member of IS&T (2012) and received the SPIE Technology Achievement Award (2012). He was also the IS&T/SPIE Imaging Scientist of the Year for 2011.

Professor Bovik has served in many and various capacities, including Board of Governors, IEEE Signal Processing Society, 1996-1998; Editor-in-Chief, *IEEE Transactions on Image Processing*, 1996-2002; Overview Editor, *IEEE Transactions on Image Processing*, 2009-present; Editorial Board, The Proceedings of the IEEE, 1998-2004; Senior Editorial Board, *IEEE Journal on Special Topics in Signal Processing*, 2005-2009; Associate Editor, IEEE Signal Processing Letters, 1993-1995; Associate Editor, *IEEE Transactions on Signal Processing*, 1989-1993; He founded and served as the first General Chairman of the *IEEE International Conference on Image Processing*, held in Austin, Texas, in November, 1994.

Dr. Bovik is a registered Professional Engineer in the State of Texas and is a frequent consultant to legal, industrial and academic institutions.

Perceptual Optimization of Large-Scale Wireless Video Networks

Robert W. Heath Jr., Alan C. Bovik, Gustavo de Veciana,
Constantine Caramanis, and Jeffrey G. Andrews
Wireless Networking and Communications Group, The University of Texas at Austin
{*rheath,bovik,gustavo,cmcaram,jandrews*}*@ece.utexas.edu*

1. Introduction

The next generation of video networks will deliver unicast and multicast of video content to mobile users, leveraging rapidly expanding wireless networks. Video networks must operate with high video network capacity, essentially maximizing the number of video flows that can be supported. Unfortunately, the application-agnostic paradigm of current data networks is not suited to meet rising video demands. Nor is the uniform coverage and capacity goal of cellular planning well suited for leveraging the spatio-temporal, bursty nature of video. We believe that video networks at every time-scale and layer should operate under the premise that distortion in the observed video stream, and in particular, perceptual distortion as would be perceived by a human consumer, should be the ultimate measure of error at the destination.

This paper summarizes key findings from a three-year project on video aware wireless networks with the objective of increasing (and defining a baseline) video capacity by at least 66x to meet projected capacity demands. Our research falls into two interconnected research vectors, summarized in Fig. 1. The work on video quality defined full-reference, reduced-reference, and no-reference models that achieve good correlation with subjective experiments. The models have been used to drive adaptation algorithms in the second research vector on spatio-temporal network adaptation. The work on network adaptation leverages aggressive deployment of small-cell infrastructure and exploits properties of stored-video streaming and real-time video to enable video-aware scheduling. The remainder of this letter summarizes select results in each research thrust.

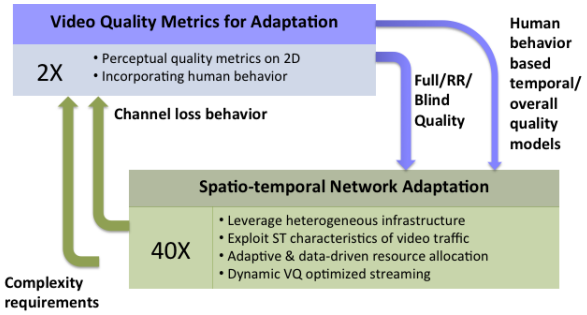


Fig. 3. Research directions and capacity gains.

2. Perceptual Video Quality Assessment

As discussed in a companion paper in this issue, a number of powerful new video quality assessment (VQA) models have been developed that deliver quality predictions that correlate closely with human quality judgments as measured on the Video Quality Expert Group (VQEG) FRTV Phase 1 database and on the LIVE VQA database [1]. The performance of these algorithms is boosted by the use of motion measurements [2] and/or natural video statistics, and depends on the amount of information available (if any) regarding the reference video(s) being tested. Efficacy is still high when little or no reference information is available; in particular "no reference" (NR) or blind models have great potential for assessing video traffic in wireless video networks.

Quality of Experience

Newly developed HTTP-based video streaming technology enables flexible rate-adaptation in varying channel conditions. The users' Quality of Experience (QoE) of rate-adaptive HTTP video streams, however, is not well understood. Therefore, designing QoE-optimized rate-adaptive video streaming algorithms remains a challenging task. An important aspect of understanding and modeling QoE is to be able to predict the up-to-the-moment subjective quality of video as it is played. In [3], we proposed a dynamic system model to predict the time-varying subjective quality (TVSQ) of rate-adaptive videos transported over HTTP. The new model effectively predicts the time-varying subjective quality of rate-adaptive videos in an online manner, making it possible to conduct QoE-optimized online rate-adaptation for HTTP-based video streaming. Fig. 2 shows that our dynamic system model can accurately predict the TVSQ.

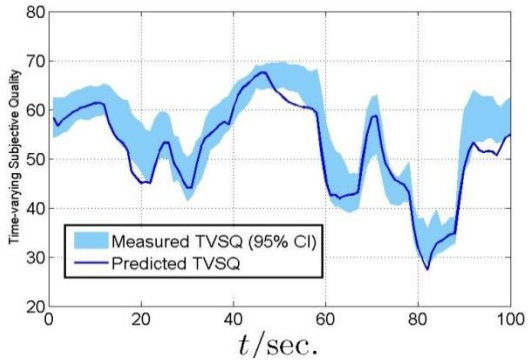


Fig. 2. The performance of the dynamic system model for TVSQ prediction.

A New Mobile Video Quality Database

Reference databases with mean opinion scores are important to allow researchers to compare competing VQA algorithms. We built a database of rate-varying video sequences called the LIVE Mobile Video Quality Database that simulate quality fluctuations commonly encountered in video streaming applications [4], [5]. We conducted a large scale subjective study on which time-varying subjective judgments of video quality were collected using two types/sizes of wireless display devices (smartphone and tablet). We envision that this free, publicly available database will prove useful for developing and validating visual quality models for quality-varying long videos.

3. Spatio Temporal Interference Management

Interference Shaping for Improved Quality of Experience for Real-Time Video Streaming

Bursty co-channel interference is a prominent cause of wireless throughput variability, which leads to annoying video quality variations. In [6], we propose and analyze a network-level resource management algorithm termed interference shaping to smooth video quality variations, by decreasing the peak rate of co-channel best effort users. The proposed algorithm is designed to maximize the H-MS-SSIM index [7], which incorporates a hysteresis (or 'recency') effect in predicting the perceived video quality. In Table I, we compare the performance of our IS method with a transmission scheme that does not incorporated IS. We utilized the coefficient of variation of Q (CoQV) defined by $\sqrt{Var[Q]}/E[Q]$ as a normalized measure of the fluctuation of the video quality. Table I reveals that our algorithm improves the average predicted quality with reduced predicted quality fluctuations.

Table I COMPARISON OF COQV AND AVERAGE MS-SSIM.

	Single Interference		Multiple Interference	
	CoQV	MS-SSIM	CoQV	MS-SSIM
Without IS	0.0694	0.9099	0.0119	0.9851
With IS	0.0097	0.9859	0.0076	0.9965

Multi-User Rate Adaptation for Stored Video Transport Over Wireless Systems

It has long been recognized that frequent video quality fluctuations could significantly degrade the QoE, even if the average video quality is high. In [8], we develop an online multi-user rate-adaptation algorithm (NOVA) to maximize the weighted sum of average quality and quality variations. The algorithm only requires minimal statistical information about the wireless channel dynamics and the rate-quality characteristics. For the wireless cellular downlink with fixed number of users, the algorithm is asymptotically optimal. Capacity gains with the proposed algorithm are in the range of 2x. In Fig. 3, we compare the performance of NOVA against that of PF-RM (which uses proportional fair resource allocation and buffer aware rate matching for quality adaptation) and PF-QNOVA (which uses proportional fair resource allocation and NOVA's quality adaptation) in a wireless network supporting N video clients. NOVA provides significant network capacity gains.

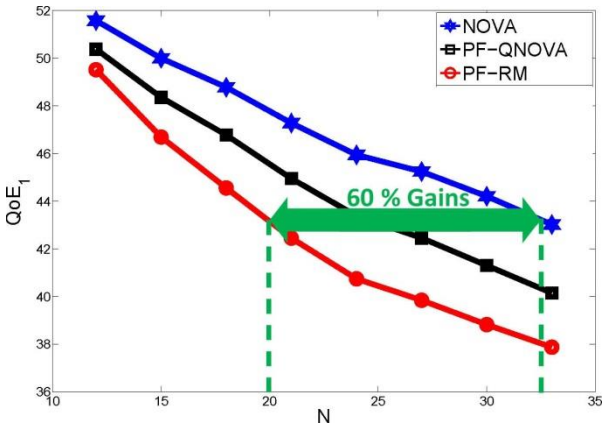


Fig. 3. The QoE (mean quality+variation) of NOVA.

MIMO Video Adaptation

In [11], we introduce an architecture for real-time video transmission over multiple-input multiple-output (MIMO) wireless communication systems using loss visibility side information of video packets. To jointly capture video quality and network throughput, we define the optimization objective as the throughput weighted by the loss visibility of each packet, a metric coined *perceived throughput*. We use the loss visibility side information to classify video packets and transmit them through different subchannels of the MIMO channel. When tested on H.264-encoded video sequences, the proposed architecture achieves the same video quality (SSIM [12]) at a 17 dB reduction in transmit power for a 2x2 MIMO system, giving a 2-4x capacity gain over a baseline MIMO system. Fig. 4 demonstrates the video quality gains achieved over a range of antennae. Our prioritized transmission methodology only requires an SNR of 3 dB to achieve a video quality of 0.9 on 2x2 MIMO systems. By comparison, the non-prioritized method requires 20 dB. Furthermore, gains in excess of 10 dB are achieved over a wide range of antenna configurations.

4. Conclusions

In this paper we summarized some of our recent work on developing new models for perceptual video quality assessment, and using these models to adapt video transmission based on perceptual distortion. Our

adaptive algorithms give capacity gains on the order of 2-4x depending on the definition of capacity and the baseline. A major finding not discussed here is that capacity gains of 40x or more could be achieved through aggressive deployment of small-cell infrastructure [13]. These capacity gains come on top of the other gains from adaptive algorithms. In further work, we are developing models that better describe the quality of experience and using these models to develop more advanced algorithms.

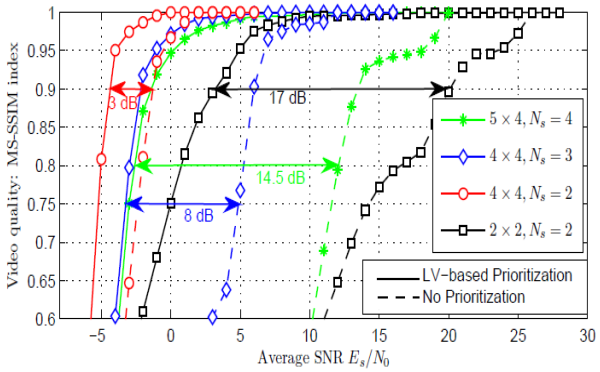


Fig. 4. Comparison of the loss visibility-based prioritization vs. non-prioritized MIMO precoding.

Acknowledgements

This work is supported by the Intel-Cisco Video Aware Wireless Networks (VAWN) Program.

References

[1] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1427–1441, Jun. 2010.

[2] K. Seshadrinathan and A.C. Bovik, "A structural similarity metric for video based on motion models," *IEEE Int'l Conf Acoust., Speech, Signal Process.*, Honolulu, HI, April 2007.

[3] C. Chen, L.K. Choi, G. de Veciana, C. Caramanis, R.W. Heath, Jr., and A. C. Bovik, "A dynamic system model of time-varying subjective quality of video streams over HTTP ," *IEEE Int'l Conf Acoust, Speech Signal Process*, Vancouver, British Columbia, May 2013.

[4] A.K. Moorthy, L.K. Choi, A.C. Bovik and G. de Veciana, "Video quality assessment on mobile devices: Subjective, behavioral, and objective studies," *IEEE J. Sel Topics Signal Process.*, Special Issue on New Subjective and Objective Methodologies for Audio and Visual Signal Processing, vol. 6, no. 6, pp. 652-671, Oct. 2012.

[5] A.K. Moorthy, L.K. Choi, A.C. Bovik and G. DeVeciana, LIVE Mobile Video Quality Database. Online: http://live.ece.utexas.edu/research/quality/live_mobile_video.html.

[6] S. Singh, J.G. Andrews, G. de Veciana, "Interference shaping for improved quality of experience for real-time video streaming," *IEEE J. Sel. Areas in Commun.*, vol.30, no.7, pp. 1259-1269, Aug. 2012.

[7] *K. Seshadrinathan and A.C. Bovik*, **"Temporal hysteresis model of time-varying subjective video quality,"** *IEEE Int'l Conf. Acoust., Speech Signal Process*, May 22-27, Prague, Czech Republic, 2011.

[8] V. Joseph and G. de Veciana, "Jointly optimizing multi-user rate adaptation for video transport over wireless systems: Mean-fairness-variability tradeoffs," *IEEE INFOCOM*, March 2012.

[9] Z. Wang, E. Simoncelli and A.C. Bovik, "Multi-scale structural similarity for image quality assessment," *Asilomar Conf. Signals, Syst, Comput.*, Pacific Grove, CA, Nov. 2003.

[10] A.A. Khalek, C. Caramanis, and R.W. Heath, Jr., "Loss visibility optimized real-time video transmission over MIMO systems," *IEEE Trans. Circ. Syst., Video Technol.*, submitted.

[11] Z Wang, A.C. Bovik, H.R. Sheikh and E.P. Simoncelli, "The SSIM index for image quality assessment," Online: http://www.cns.nyu.edu/~ lcv/ssim.

[12] J. G. Andrews, "Seven Ways that HetNets are a Cellular Paradigm Shift", *IEEE Communications Magazine*, Vol. 51, No. 3, pp. 136-44, Mar. 2013.

Robert W. Heath Jr. is a Professor in the ECE Department at UT Austin. He received the Ph.D. in EE from Stanford University. He holds the David and Doris Lybarger Endowed Faculty Fellowship in Engineering, is a registered Professional Engineer in Texas, and is an IEEE Fellow.

Al Bovik. See photo and biosketch in another letter in this issue.

Gustavo de Veciana is a Professor in the ECE Department at UT Austin. He received the Ph.D. in EECS from U.C. Berkeley. He is the Joe J. King Endowed Professor in Engineering, and is an IEEE Fellow.

Constantine Caramanis is an Associate Professor in the ECE Department at UT Austin. He received his Ph.D. in EECS from MIT. He is a member of the IEEE.

Jeffrey G. Andrews is a Professor in the ECE Department at UT Austin. He received the Ph.D. in EE from Stanford University. He holds the Brasfield Endowed Faculty Fellowship in Engineering, and is an IEEE Fellow.

Dynamic Adaptive Streaming over HTTP: Standards and Technology

Ozgur Oyman, Intel Labs, Santa Clara, CA, USA (ozgur.oyman@intel.com)
Utsaw Kumar, Intel Architecture Group, Santa Clara, CA USA (utsaw.kumar@intel.com)
Vish Ramamurthi, Intel Labs, Santa Clara, CA USA (vishwanath.ramamurthi@intel.com)
Mohamed Rehan, Intel Labs, Cairo, Egypt (mohamed.m.rehan@intel.com)
Rana Morsi, Intel Labs, Cairo, Egypt (ranax.a.morsi@intel.com)

1. Introduction on DASH

HTTP adaptive streaming (HAS), which has recently been spreading as a form of internet video delivery with the recent deployments of proprietary solutions such as Apple HTTP Live Streaming, Microsoft Smooth Streaming and Adobe HTTP Dynamic Streaming, is expected to be deployed more broadly over the next few years. In the meantime, the standardization of HTTP Adaptive Streaming has also made great progress with the recent completion of technical specifications by various standards bodies. More specifically, the Dynamic Adaptive Streaming over HTTP (DASH) has recently been standardized by Moving Picture Experts Group (MPEG) and Third Generation Partnership Project (3GPP) as a converged format for video streaming [1,2], and the standard has been adopted by other organizations including Digital Living Network Alliance (DLNA), Open IPTV Forum (OIPF), Digital Entertainment Content Ecosystem (DECE), World-Wide Web Consortium (W3C) and Hybrid Broadcast Broadband TV (HbbTV). DASH today is endorsed by an ecosystem of over 50 member companies at the DASH Industry Forum.

2. DASH Standards in MPEG and 3GPP

The scope of both MPEG and 3GPP DASH specifications [1,2] includes a normative definition of a media presentation or manifest format (for DASH access client), a normative definition of the segment formats (for media engine), a normative definition of the delivery protocol used for the delivery of segments, namely HTTP/1.1, and an informative description on how a DASH client may use the provided information to establish a streaming service.

At MPEG, DASH was standardized by the Systems Sub-Group, with the activity beginning in 2010, becoming a Draft International Standard in January 2011, and an International Standard in November 2011. The MPEG-DASH standard [1] was published as ISO/IEC 23009-1:2012 in April, 2012. In addition to the definition of media presentation and segment formats standardized in [1], MPEG has also developed additional specifications [5]-[7] on aspects of implementation guidelines, conformance and reference software and segment encryption and authentication. Toward enabling interoperability and conformance, DASH also includes profiles as a set of restrictions on the offered media presentation description (MPD) and segments based on the ISO Base Media File Format (ISO-BMFF) [4] and MPEG-2 Transport Streams [3]. Currently, MPEG is also pursuing several core experiments toward identifying further DASH enhancements.

At 3GPP, DASH was standardized by the 3GPP SA4 Working Group, with the activity beginning in April 2009 and Release 9 work with updates to Technical Specification (TS) 26.234 on the Packet Switched Streaming Service (PSS) [8] and TS 26.244 on the 3GPP File Format [9] completed in March 2010. During Release 10 development, a new specification TS 26.247 on 3GPP DASH [2] has been finalized in June 2011, in which ISO-BMFF based DASH profiles were adopted. In conjunction with a core DASH specification, 3GPP DASH also includes additional system-level aspects, such as codec and Digital Rights Management (DRM) profiles, device capability exchange signaling and Quality of Experience (QoE) reporting. Since Release 11, 3GPP has been studying further enhancements to DASH and toward this purpose collecting new use cases and requirements, as well as operational and deployment guidelines. Some of the documented use cases in the related Technical Report (TR) 26.938 [11] include: Operator control for DASH, e.g., for QoE/QoS handling, advanced support for live services, DASH as a download format for push-based delivery services, enhanced ad insertion support, enhancements for fast startup and advanced trick play modes, improved operation with proxy caches, multimedia broadcast and multicast service (MBMS) [10] assisted DASH services with content caching at the UEs, handling special content over DASH and enforcing specific client behaviors, and use cases on DASH authentication.

3. Research Challenges for Optimizing DASH Delivery in Wireless Networks

As a relatively new technology in comparison with traditional streaming techniques such as Real-Time Streaming Protocol (RTSP) and HTTP progressive

download, deployment of DASH services presents new technical challenges. In particular, enabling optimized end-to-end delivery of DASH services over wireless networks requires developing new algorithms, architectures, and signaling protocols for efficiently managing the limited network resources and enhancing service capacity and user QoE. Such development must also ensure access-specific (e.g., 3GPP-specific) optimizations for DASH services, which clearly require different approaches and methods compared to those for traditional streaming techniques. In the remaining of this paper, we highlight some of the specific research problems we are tackling in this space and summarize our related findings.

a) Link-Aware DASH Adaptation

One of the key elements in rate adaptation in DASH is the estimation of available network bandwidth by the DASH client. Traditional approaches use either application or transport layer throughput for this purpose. However using higher layer throughputs alone can potentially have adverse effects on user QoE when the estimated value is different from what is available at lower layers. This is typically the case when operating on wireless links whose characteristics fluctuate with time depending on the physical environment as well as load on the system. A lower estimate of available network bandwidth results in lower video quality and a higher estimate can result in re-buffering To alleviate this problem, we proposed in [13], a Physical Link Aware (PLA) HAS framework for enhanced QoE. This is a cross-layer approach in which physical layer goodput information is used as a complement to higher layer estimates , such as those based on video segment fetch and download times. A link-aware approach allows us to track wireless link variations over time at a finer scale and thus provide more opportunistic video rate adaptation to improve the QoE of the user. Simple enhancements were proposed in [13] that use the link level goodput to improve user QoE in terms of enhanced startup video quality and reduced re-buffering percentage. Our simulation results in Fig. 1 show that 2-3 dB improvement in startup quality (in terms of PSNR) can be obtained for 90% of users using PLA approach as opposed to physical layer unaware (PLU) approach. Fig. 2 shows the reduction in re-buffering that can be obtained using PLA.

b) Server-Assisted DASH

One of the most common problems associated with video streaming is the clients' lack of knowledge of server and network conditions. For instance, clients usually make requests based on their bandwidth unaware of the servers' capacity or the number of clients streaming from the same server at the same time. Thus, clients tend to request the highest possible bitrates based on their perception regardless of the servers' condition. This often causes clients to encounter playback stalls and pauses. Consequently, the clients' QoE is dramatically affected. The same issue applies in cases where some greedy clients tend to eat up network bandwidth and stream at higher quality, leaving the rest of the clients to suffer much poorer QoE. A possible quick fix would be to allocate bandwidth equally among streaming clients. However, another potential problem might arise in the case where clients are streaming different types of content, e.g., fast vs. slow motion content, where simply sharing bandwidth equally does not necessarily imply same QoE for both clients. This is due to the fact content with high speed motion need to be streamed at higher bitrates to achieve the same QoE achieved by clients streaming content with slower motion.

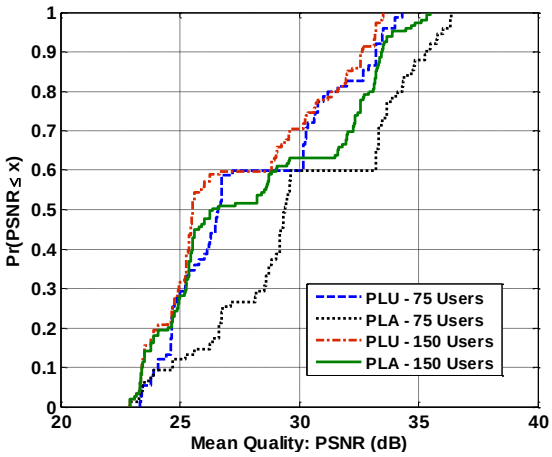


Fig. 1: Startup quality improvement using physical link awareness.

We attempt to solve these problems by introducing a feedback mechanism between the server and clients. The clients notify the server of their perceived QoE so far. This is in the form of statistics sent by the client regarding the client's average requested bitrate, average quality and number of stalling events. The server in return advises each client about the bandwidth limit it can request. In other words, the server can notify each client which representation can be requested at any given time. In our implementation, the server provides the clients with a special binary code, Available Representation

Code (ARC). This ARC assigns a bit (representation access bit), which can be either '0' or '1', for each representation. The server uses the information sent by the clients in addition to its knowledge regarding the content streamed by each client to dynamically update the ARC and notify the clients accordingly as shown in Figure 3. The selection of which representation to be enabled or disabled is subject to server-based algorithms. We observed that the use of server-assisted feedback has resulted in a significant reduction in re-buffering at the client side at the expense of little or no perceivable quality loss.

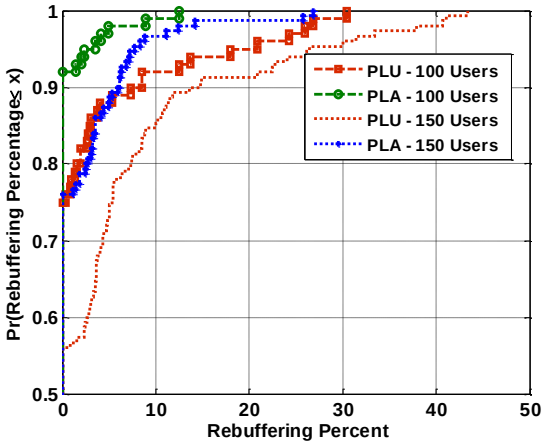


Fig. 2: Reduced Re-buffering percentage using physical link awareness.

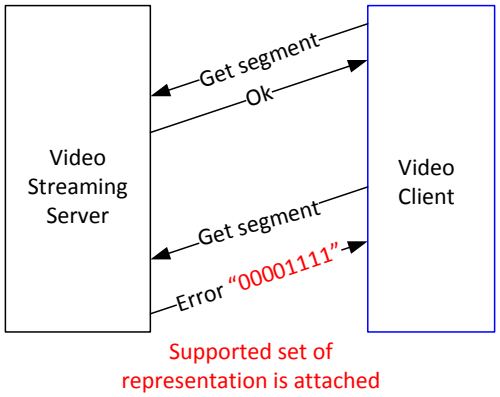


Fig. 3: DASH Server-assisted feedback

c) DASH Transport over eMBMS

As the multicast standard for LTE, e-MBMS was introduced by 3GPP to facilitate delivery of popular content to multiple users over a cellular network. If a large number of users are interested in the same content, for e.g., live sports, news clips, etc., multicast transmission can significantly lower the cost and make better use of the spectrum as compared to unicast transmission. Our paper in [14] studies and analyzes the quality of experience (QoE) at the end user during live video streaming over e-MBMS using a comprehensive end-to-end e-MBMS streaming framework that is based on H.264/AVC encoded video content delivered using the File Delivery over Unidirectional Transport protocol, combined with application layer (AL) forward error correction (FEC). The study involves QoE evaluation in terms of startup delay, re-buffering percentage and PSNR metrics and provides performance evaluations to characterize the impact of various MBMS streaming, transport and AL-FEC configurations on the end user QoE. In addition, for simulating Media Access Control-Protocol Data unit losses, we proposed a new Markov model and showed that the new model captures the coverage aspects of e-MBMS in contrast to the RAN endorsed model [15], which assumes the same two state Markov model for all the users. A new decoding strategy is also proposed that takes into consideration the systematic structure of AL-FEC codes. We show that this strategy outperforms the decoding strategy when a decoding failure leads to the loss of the whole source block.

4- 4. Summary and Future Research

This paper gave an overview of the latest DASH standardization activities at MPEG and 3GPP and reviewed a number of research vectors we are pursuing with regards to optimizing DASH delivery over wireless networks. We believe that this is an area with a rich set of research opportunities and that further work could be conducted in the following domains: (i) Development of evaluation methodologies and performance metrics to accurately assess user QoE for DASH services. (ii) DASH-specific QoS delivery, that involves developing new policy and charging control (PCC) guidelines and QoS mapping rules over radio access network and core IP network architectures. (iii) QoE/QoS-based adaptation schemes for DASH at the client, network and server (potentially assisted by QoE feedback reporting from clients), to jointly determine the best video, transport, network and radio configurations toward realizing the highest possible service capacity and end user QoE. (iv) Transport optimizations over heterogeneous network environments, where DASH content is delivered over multiple access networks such as WWAN unicast (e.g., 3GPP PSS [8]), WWAN broadcast (e.g., 3GPP MBMS [10]) and WLAN (e.g., WiFi) technologies.

REFERENCES

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

[2] 3GPP TS 26.247: "Transparent end-to-end packet switched streaming service (PSS); Progressive download and dynamic adaptive streaming over HTTP (3GP-DASH)"
[3] ITU-T Rec. H.222.0 | ISO/IEC 13818-1:2013: "Information technology - Generic coding of moving pictures and associated audio information: Systems"
[4] ISO/IEC 14496-12: "Information technology - Coding of audio-visual objects - Part 12: ISO base media file"
[5] ISO/IEC 23009-2: "Information Technology – Dynamic adaptive streaming over HTTP (DASH) – Part 2: Conformance and Reference Software"
[6] ISO/IEC 23009-3: "Information Technology – Dynamic adaptive streaming over HTTP (DASH) – Part 3: Implementation Guidelines"
[7] ISO/IEC 23009-4: "Information Technology – Dynamic adaptive streaming over HTTP (DASH) – Part 4: Segment Encryption and Authentication"
[8] 3GPP TS 26.234: "Transparent end-to-end packet switched streaming service (PSS); Protocols and codecs"
[9] 3GPP TS 26.244: "Transparent end-to-end packet switched streaming service (PSS); 3GPP file format (3GP)"
[10] 3GPP TS 26.346: "Multimedia Broadcast Multicast Service (MBMS); Protocols and codecs"
[11] 3GPP TR 26.938: "Improved Support for Dynamic Adaptive Streaming over HTTP in 3GPP"
[12] O. Oyman and S. Singh, "Quality of experience for HTTP adaptive streaming services," IEEE Commun. Mag., vol. 50, no:4, pp. 20-27, Apr. 2012.
[13] V. Ramamurthi and O. Oyman, "Link Aware HTTP Adaptive Streaming for Enhanced Quality of Experience," 2013 IEEE GLOBECOM Symp. on Comms. Software, Services and Multimedia, accepted for publication.
[14] U. Kumar, O. Oyman and A. Papathanassiou, "QoE Evaluation for Video Streaming over eMBMS," Journal of Communications, vol:8, no:6, June 2013.
[15] 3GPP Tdoc S4-111021, "Channel modeling for MBMS," 3rd Generation Partnership Project (3GPP), 2011.

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.

UTSAW KUMAR is a wireless systems engineer with the Standards and Advanced Technology division of the Mobile and Communications Group at Intel. He is author or co-author of over 10 technical publications. He holds Ph.D. and M.S. degrees from University of Notre Dame, and a B.Tech. degree from Indian Institute of Technology in Kanpur, India.

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.

MOHAMED REHAN obtained his Ph.D. from the University of Victoria, Victoria, British Columbia, Canada in the area of video coding in 2006. He received his B.S. in communications and his M.S. in computer graphics in 1991 and 1994 respectively from the department of electronics and communications, Cairo University, Cairo, Egypt. Dr. Rehan is currently working as a senior research scientist at Intel-labs Egypt. His research focus includes video streaming and broadcasting over wireless networks.
Dr. Rehan has been involved in multimedia research and development for more than 20 years. He has an extensive set of publications in video coding and multimedia. He has also been involved in the research and development of several products related to video coding and multimedia applications including video standards.

RANA MORSI received her BSc. Degree in Computer Science and Engineering in 2012 at the German University in Cairo with a grade of Excellent with highest honors. Rana currently works as a software and wireless applications consultant at Intel labs-Egypt as part of the mobile video streaming team. Rana has several contributions to video streaming using DASH protocol. Other research interests include Mobile Augmented Reality and Natural Langauage Processing.

Cross-Layer Design for Mobile Video Transmission

Laura Toni, Dawei Wang**, Pamela Cosman** and Laurence Milstein***
**EPFL, Lausanne, Switzerland, **University of California San Diego, USA*
laura.toni@epfl.ch, daw017@ucsd.edu, pcosman@ucsd.edu, lmilstein@ucsd.edu

1. Introduction

The research results summarized in this paper are typical of a broader set of results that a aim to optimize physical/application cross-layer designs for either real-time or archival video, where the channel is doubly selective (i.e., exhibits both time and frequency selectivity). The scenario of most interest to us is when there is arbitrary mobility between the transmitter and the receiver; the relative velocity between the two is, in general, changing with time over the duration of the video. This model results in a channel that is nonstationary, since the coherence time of the channel varies with time, and so it does not lend itself to various results in the literature. For example, a standard proof of the ergodic capacity for a fast fading channel that requires the channel to be both stationary and ergodic does not apply.

To approach this type of problem, our philosophy is to design the systems to yield performance that is robust over a wide range of Doppler spreads. That is, rather than designing a system to yield optimal performance for a specific set of operating conditions, we design systems that perform satisfactorily over a wide range of operating conditions, but which might not be optimal for any set of such conditions.

The information used in these designs are channel state information (CSI) at the physical layer, and distortion-rate (DR) information at the application layer. The combination of this physical and application layer information enables us to segregate bits into different importance levels, and then protect the bits in each level more or less according to its importance class. The basic physical-layer waveform that we use is a multicarrier waveform, and among the techniques that we use to achieve this unequal error protection (UEP) are forward-error correction (FEC) and mapping more important bits to subcarriers that are experiencing larger channel gains.

In what follows, we summarize our key results in two areas, video resource allocation for systems operating at arbitrary mobility, and slice mapping for non-scalable video for systems operating at low mobility.

2. Resource allocation

We study a multiuser uplink video communication system where a group of K users is transmitting scalably-encoded video with different video content to a base station. The frame rate for the videos is the same, and the video frames of each user are compressed for each group of pictures (GOP). Also, the number of frames in a GOP is the same for all users. The system operates in a slotted manner, with the slot starting and ending epochs aligned for all users, and where the slot duration is the same as the display time of one GOP.

On the physical-layer side, we consider a time-varying orthogonal frequency division multiplexed (OFDM) system with equally spaced subcarriers spanning the total system bandwidth. We assume a block-fading model in the frequency domain, and a contiguous group of D_f subcarriers, defined as a subband, experiences the same fading realization, whereas different subbands fade independently.

On the application-layer side, to minimize the sum of the mean-square errors (MSE) across all users, the base station collects the distortion-rate (DR) information for every user. For each bitstream, the most important video information (e.g., motion vector, macroblock ID's) is contained in a substream called the base layer. One or more enhancement layers are added such that the MSE distortion decreases as additional enhancement bits are received by the decoder.

The source encoder ranks the packets based on their importance in the GOP. If an error occurs in the transmission, the entire packet and all the other packets with lower priority are dropped, but all the previous packets, which have higher priority and which have already been successfully received by the decoder, are used for decoding the video.

In Figures 1 and 2, the performance of the cross-layer algorithm, described in detail in [1], is presented. This algorithm is optimized by jointly using the CSI of each subcarrier, and the DR curve of each video. The intent is to strike a balance between giving users a number of subcarriers that is proportional to their individual needs (as determined by each user's own DR curve) and assigning each subcarrier to that user whose channel gain is largest for that particular subcarrier. Further, two comparison curves are shown in Figures 1 and 2, one of which uses just the CSI for the resource

allocation (i.e., it makes no use of the application-layer information), and the other one of which uses only the DR curves for the resource allocation (i.e., it makes no use of the physical-layer information).

Consider Fig. 1, which corresponds to three video users competing for bandwidth in increments of OFDM subcarriers, of which there are 16. The ordinate of Fig. 1 is the average peak-signal-to-noise ratio (PSNR) and the abscissa is the normalized Doppler spread (which is the inverse of the number of consecutive channel symbols that experience a highly correlated fade). The three solid curves show the error-free performance of the system (and so represent an upper bound to the actual system performance), whereas the three dashed curves incorporate the effects of channel noise and fading. Within each set, the three curves correspond to the cross-layer algorithm, the application-layer algorithm, and the physical-layer algorithm.

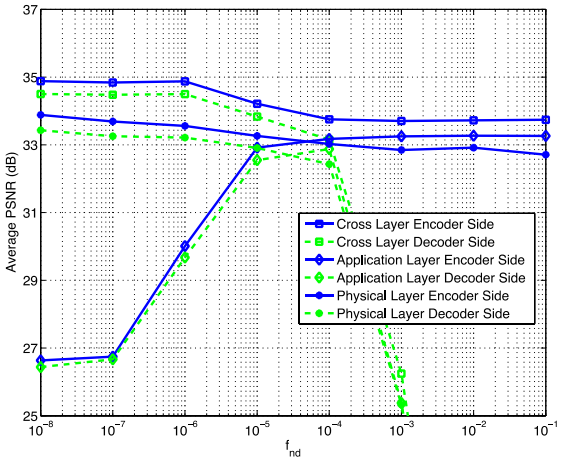


Figure 1. Ls = 100, 16 Subcarriers, PSNR versus Normalized Doppler Spread.

Note that for low Doppler spreads, the cross-layer and the physical-layer algorithms start out with relatively high PSNRs, and those PSNRs remain high as the Doppler spread increases until a point is reached where they abruptly degrade for any additional increase in the Doppler spread. This is because both algorithms make initial subcarrier assignments based upon which user has the strongest channel at each subcarrier location. As the Doppler spread increases, the benefit of time diversity helps system performance, but beyond a certain point, the Doppler spread becomes too large and the performance of both systems degrades very rapidly. On the other hand, at low Doppler spreads, the application-layer algorithm performs poorly, because it does not make any use of the CSI when subcarriers are allocated, but rather assigns subcarriers to users in a random manner. As the Doppler spread increases, the application-layer algorithm's performance increases because of the effect of the time diversity, and then it, like the other two algorithms, degrades very rapidly as the Doppler spread gets even larger.

The key limitation to good performance at high Doppler spreads is the need to track the variations of the channel sufficiently fast. Otherwise, the CSI will be outdated when it is used by the receiver. Since the estimates are typically obtained by the use of unmodulated pilot symbols, one can compensate for rapid fading by decreasing the spacing between the pilot symbols. However, this will result in a loss of throughput, since the pilot symbols contain no information. To see the effect of this tradeoff between channel outdating and loss of throughput, consider Fig 2, which is the same as Fig. 1 except that in Fig. 1, the pilot spacing is 100 symbols, and in Fig. 2, the pilot spacing is 25 symbols. It is seen that all three systems can now function properly at Doppler spreads that are about an order of magnitude larger than in the system of Fig. 1.

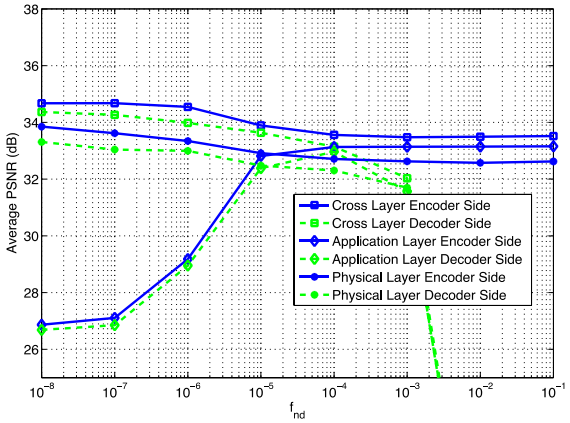


Figure 2. Ls = 25, 16 Subcarriers, PSNR versus Normalized Doppler Spread.

3. Mapping of video slices to OFDM subcarriers

We considered transmission of non-scalably encoded video sequences over an OFDM system in a slowly varying Rayleigh faded environment. The OFDM waveform consists of Nt subcarriers, which experience block fading in groups of M consecutive subcarriers. That is, we have N groups of M subcarriers each, where the fading in a given group is perfectly correlated, and the fading experienced by different groups is independent from group to group. Note that N is a measure of the potential diversity order of the system, while M is a measure of the coherence bandwidth of the system.

We make use of a slice loss visibility (SLV) model that can evaluate the visual importance of each slice. The

cross-layer approach, taking into account both the visibility scores available from the bitstream and the CSI available from the channel, makes use of the difference in importance levels of the bits that comprise the video.

In our model, the ith slice of frame k is encoded into Lk(i) bits and has a priority level Vk(i). The Vk(i) values range from 0 to 1, where Vk(i) = 0 means that the slice, if lost, would produce a glitch that would likely not be noticed by any observer, and Vk(i) = 1 means that the loss artifact would likely be seen by all users. So, each encoded slice is characterized by the pair (Vk (i), Lk (i)), where k = 1, . . . , J , and J is the number of frames per GOP.

We consider various scenarios, focusing on the availability of instantaneous CSI and SLV parameters. We consider all possible combinations of knowing the instantaneous CSI and not the SLV, knowing the SLV and not the instantaneous CSI, knowing both pieces of information, or knowing neither. The main point of the approach is that if the sender has at least one of the two types of information, then the algorithm can exploit that information. In particular, we consider two types of exploitation: the first is forward error correction using different channel code rates for different slices or different subcarriers, and the second is slice-to-subcarrier mapping, in which the algorithm maps the visually more important slices to the better subcarriers. Note that the UEP FEC could, in principle, make use of the information of either the SLV or the instantaneous CSI, or both. That is, heavier error protection could be provided to specific slices (because they are more important) or to specific subcarriers (because they are not reliable). In contrast, the slice-to- subcarrier mapping operation requires both the SLV and instantaneous CSI information. If the instantaneous CSI is available from a feedback channel, the subcarriers of the resource block can be ordered from the most reliable to the least reliable, and if, in addition, the SLV information is available, then the most important slices can be allocated (mapped) to the most reliable subcarriers.

To illustrate typical results, we consider two baseline algorithms as a means of comparison: sequential and random. In both of these, we assume that slice importance is not known, and so no packet is more important than any other. The sequential algorithm sequentially allocates the slices of each frame to the resource block (RB). This means that the first slice of the first frame of the considered GOP is allocated to the first subcarrier. When no more information bits are available in the first subcarrier, the algorithm starts allocating the current frame to the next subcarrier. Once the slices of the first frame of the GOP are allocated, the second frame is considered. The random algorithm allocates each slice of the GOP to a random position of the RB.

The results show that the UEP approach gives a respectable gain over the baselines, and the slice mapping-to-subcarriers approach gives an even larger gain over the baselines. Applying both approaches at the same time produces a negligible gain over just doing the subcarrier mapping. As a consequence, the cross-layer algorithm that we discuss below corresponds to slice mapping as described above, with equal error protection. We refer to this design as "Scenario B" to be consistent with the terminology in [2]. In Fig. 3, the best VQM score, which is a common perceptually-based metric for quantifying video quality, whereby a score of zero is the best and a score of unity is the worst (see [3]), is plotted as a function of signal-to-noise ratio (SNR), denoted by γ, for systems with (N,M) = (32,4). That is, we have 128 subcarriers, and they are divided into 32 groups with 4 subcarriers in each group. Note that for each γ value, we provided the best VQM optimized over the whole video sequence. As expected, the general behavior is that the VQM decreases with increasing mean SNR (i.e., with increasing channel reliability). More important, for all the considered mean SNRs, Scenario B outperforms the baseline algorithms, and the gain is as much as to 0.28 in VQM score (for γ = 13 dB).

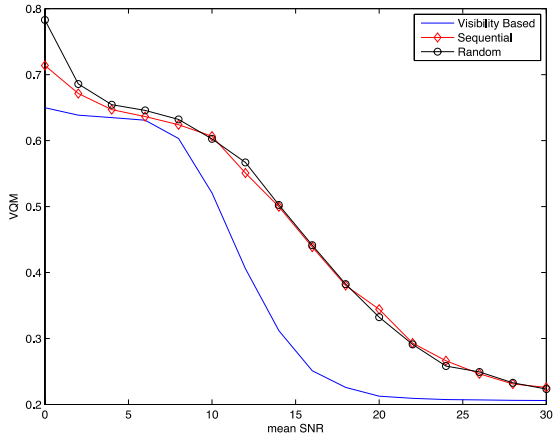


Figure 3. VQM vs. mean SNR for both visibility-based and baseline algorithms for systems with (N, M) = (32, 4).

We next consider the case of a variable number of independent subbands, and we again compare the visibility-based algorithm for Scenario B with the baselines. Fig. 4 depicts the system performance when (N,M) = (8,16) for the same video. From the figure, it can be observed that, even reducing the number of

independent subbands, the visibility-based optimization in Scenario B, when compared to the baseline algorithms, still achieves a large gain in terms of VQM. When only 2 independent channels are considered, as shown in Fig. 5, as expected, due to the limited opportunity for time diversity offered by the channel, all the algorithms lead to almost the same performance.

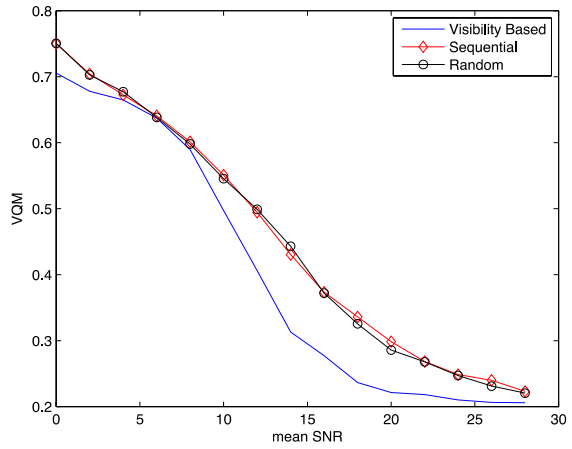


Figure 4. VQM vs. mean SNR for both visibility-based and baseline algorithms for systems with (N, M) = (8, 16).

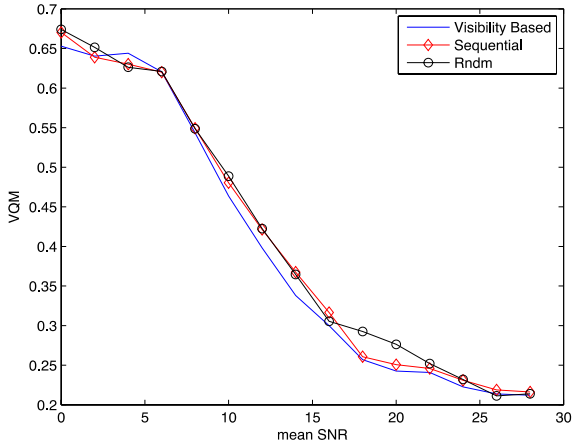


Figure 5. VQM vs. mean SNR for both visibility-based and baseline algorithms for systems with (N, M) = (2, 64).

4. Conclusions

The first example that we presented was chosen to demonstrate the design philosophy that was described in the Introduction, namely to design for robustness rather than for localized optimality. From either Fig. 1 or Fig. 2, it can be seen that performance curves of the physical-layer algorithm and the application-layer algorithm cross one another, with the former yielding better performance at low Doppler, and the latter yielding better performance at higher Doppler. However, the cross-layer algorithm yields the best performance at all Doppler spreads.

The purpose of presenting the second example was to illustrate robustness in a different context. The goal here was to have a system design that would yield satisfactory performance over a wide range of channel gains. From Figs. 3-5, it is seen that at virtually all average channel gains, the cross-layer design yields better performance than do either of the two baseline approaches, in some cases by very significant amounts.

REFERENCES

[1] D. Wang, L. Toni, P. C. Cosman, and L. B. Milstein, "Uplink Resource Management for Multiuser OFDM Transmission Systems: Analysis and Algorithm Design.," IEEE Transactions on Communications, pp. 2060-2073, May 2013.

[2] L. Toni, P. C. Cosman, and L. B. Milstein, "Channel Coding Optimization Based on Slice Visibility for Transmission of Compressed Video over OFDM Channels," IEEE Journal on Selected Areas in Communications, pp. 1172-1183, August 2012.

[3] M. H. Pinson and S. Wolf, "A New Standardized Method for Objectively Measuring Video Quality," IEEE Transactions on Broadcasting, pp. 312-322, September 2004.

Acknowledgment: This research was supported by the Intel-Cisco Video Aware Wireless Networks program, and the National Science Foundation under Grant No. CCF-0915727.

Dawei Wang (S'11) received the B.Eng. degree in electronic engineering (First Class Honors) from the Hong Kong University of Science and Technology (HKUST), Kowloon, Hong Kong SAR, China, in 2008, and the M.S. and Ph.D. degrees, in 2011 and 2013, respectively, from the University of California, San Diego, La Jolla, CA. He was also a visiting student at the University of Minnesota, Minneapolis. His industry experience includes an internship at Intel Corporation, Hillsboro, OR, in 2011. His research interests are in the areas of communication theory and video processing.

Laura Toni (S'06-M'09) received the M.S. degree (with honors) in electrical engineering and the Ph.D. degree in electronics, computer science and telecommunications from the University of Bologna, Italy, in 2005 and 2009, respectively. In 2005, she joined the Department of Electronics, Informatics and Systems at the University of Bologna, to develop her research activity in the area of wireless

communications. During 2007, she was a visiting scholar at the University of California at San Diego, CA, working on video processing over wireless systems. Since December 2012, she has been a Post-doctoral fellow in the Signal Processing Laboratory (LTS4) at the Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland. Her research interests are in the areas of image and video processing, wireless communications, and underwater communications.

Pamela Cosman (S'88-M'93-SM'00-F'08) obtained her B.S. with Honor in Electrical Engineering from the California Institute of Technology in 1987, and her M.S. and Ph.D. in Electrical Engineering from Stanford University in 1989 and 1993, respectively. She was an NSF postdoctoral fellow at Stanford University and a Visiting Professor at the University of Minnesota during 1993-1995. In 1995, she joined the faculty of the department of Electrical and Computer Engineering at the University of California, San Diego, where she is currently a Professor and Vice Chair. She was the Director of the Center for Wireless Communications from 2006 to 2008. Her research interests are in the areas of image and video compression and processing, and wireless communications. She was an associate editor of the IEEE Communications Letters and of the IEEE Signal Processing Letters, as well as the Editor-in-Chief (2006-2009) and a Senior Editor of the IEEE Journal on Selected Areas in Communications. She is a member of Tau Beta Pi and Sigma Xi.

Laurence B. Milstein (S'66-M'68-SM'77-F'85) received the B.E.E degree from the City College of New York, New York, in 1964, and the M.S. and Ph.D. degrees in electrical engineering from the Polytechnic Institute of Brooklyn, Brooklyn, NY, in 1966 and 1968, respectively. From 1968 to 1974, he was with the Space and Communications Group, Hughes Aircraft Company, and from 1974 to 1976, he was a member of the Department of Electrical and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY. Since 1976, he has been with the department of Electrical and Computer Engineering, University of California, San Diego, where he is currently the Ericsson Professor of Wireless Communications Access Techniques and former Department Chair, working in the area of digital communication theory with special emphasis on video transmission over mobile channels.

Timely Throughput Optimization of Heterogeneous Wireless Networks

Sina Lashgari and A. Salman Avestimehr
Cornell University, Ithaca, NY
sl2232@cornell.edu, avestimehr@ece.cornell.edu

1. Introduction

With the evolution of wireless networks towards heterogeneous architectures, including pico, femto, and relay base stations, and the growing number of smart devices that can connect to several wireless technologies (e.g. cellular and WiFi), it is promising that the opportunistic utilization of heterogeneous networks can be one of the key solutions to help cope with the phenomenal growth of video demand over wireless networks. This motivates an important problem: How to optimally utilize network heterogeneity for the delivery of video traffic?

In this paper, we focus on this problem in the context of real-time video streaming applications, such as live broadcasting, video conferencing and IPTV, that require tight guarantees on timely delivery of the packets. In particular, the packets for such applications have strict per-packet deadline; and if a packet is not delivered successfully by its deadline, it will not be useful anymore. As a result, we focus on the notion of timely throughput, proposed in [1,2], which measures the long-term average number of "successful deliveries" (i.e., delivery before the deadline) for each client as an analytical metric for evaluating both throughput and quality-of-service (QoS).

We consider the downlink of a wireless network with N Access Points (AP's) and M clients, where each client is connected to several out-of-band AP's, and requests timely traffic. We study the maximum total timely throughput of the network, which is the maximum average number of packets delivered successfully before their deadline, and the corresponding scheduling policies for the delivery of the packets via all available AP's in the network.

This is a very challenging scheduling problem since, at each interval, the number of possible assignments of clients to AP's, which should be considered in order to find the optimal solution, grows exponentially in M (in fact, it grows as N^M). To overcome this challenge, we propose a *deterministic relaxation* of the problem, which converts it to a network with deterministic delays in each link. We show that the solution to the deterministic problem tracks the solution to the Main Problem very well, and establish a bound on the worst-case gap between the two. Furthermore, using a linear-programming (LP) relaxation, we show that the deterministic problem can itself be approximated efficiently (efficient in both approximation gap as well as computational complexity). Hence, via the proposed deterministic approach, one can find an efficient approximation of the original problem.

2. Problem Formulation

We consider the downlink of a network with M wireless clients, denoted by $Rx_1, \dots, Rx_M$, and N Access Points, denoted by $AP_1, \dots, AP_N$. All AP's are connected via reliable links to the Backhaul Network (see Fig.1).

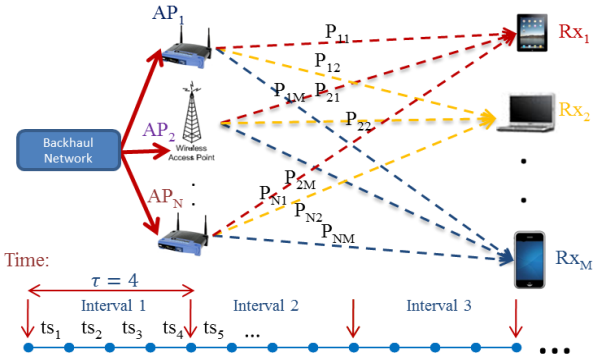


Figure 1. Network Model.

As shown in Fig. 1, time is slotted and transmissions occur during time-slots. Furthermore, time-slots are grouped into intervals of length τ, corresponding to the inter-arrival time of the packets (e.g., 1/30 seconds for video frames). Each AP is connected via wireless channels to a subset (possibly all) of the clients. These wireless links are modeled as packet erasure channels that, for simplicity, are assumed to be i.i.d over time, and have fixed, but possibly different success probabilities[1]. The success probability of the channel between AP_i and Rx_j is denoted by p_{ij}, which is the probability of successful delivery of the packet of Rx_j when transmitted by AP_i during a time-slot. If there is no link between AP_i and Rx_j, $p_{ij} = 0$.

At the beginning of each interval, a new packet for each client arrives. Each packet will then be assigned to one of the AP's for delivery during that interval. At each time-slot (of that interval), each AP can choose any of the packets that are assigned to it for transmission. At the end of that time-slot the AP will know if the packet has been successfully delivered or

[1] The i.i.d. assumption can be relaxed by considering a Markov model for channel erasures. Also, a more realistic fading model can replace the packet erasure model. These extensions are discussed in detail in [3].

not. If the packet is successfully delivered, the AP removes it from its buffer; otherwise it remains for possible future transmissions. At the end of the interval, all packets that are not delivered will be discarded (since they pass their deadlines). The *scheduling policy* determines how the packets are assigned to AP's at the beginning of each interval, and which packet each AP transmits at each time-slot. A scheduling policy η decides causally based on the entire past history of events up to the point of decision-making. We denote the set of all possible scheduling policies by $\mathcal{S}$.

When using a particular scheduling policy η, the total timely throughput, denoted by $T^3(\eta)$, is defined as the long-term average number of packets delivered successfully before their deadlines. In other words, if $N_j(r,\eta)$ denotes a binary RV, which is 1 if and only if Rx_j successfully receives its packet at interval r, then

$$T^3(\eta) \triangleq \lim_{r\to\infty} sup\ \frac{\sum_{k=1}^{r}\sum_{j=1}^{M} N_j(k,\eta)}{r}.$$

Main Problem.

Our objective is to find the maximum achievable total timely throughput, denoted by C_{T^3},

$$C_{T^3} \triangleq \sup_{\eta\in\mathcal{S}} T^3(\eta),$$

and the corresponding optimal policy. Characterizing C_{T^3} is challenging, since the dimension of the corresponding optimization problem at each interval (i.e., the number of all possible assignments) grows exponentially as N^M. As we discuss next, we propose a deterministic relaxation of the problem to overcome this challenge. The main idea is to first reduce the problem to a closely connected integer program, which is a special case of the generalized assignment problem (see e.g., [5]). Then, we show that the corresponding integer program can be approximated using a linear-programming (LP) relaxation. We prove tight guarantees on the worst-case loss from the above two-step relaxation, hence providing an efficient approach to approximate C_{T^3}, and find a near-optimal schedule.

3. Deterministic Relaxation

Consider the case that AP_i has only one packet to send, and wants to transmit that packet to Rx_j. Thus, AP_i persistently sends that packet to Rx_j until the packet is delivered. The number of time-slots expended for this packet to be delivered is a Geometric random variable with parameter p_{ij}, and average $\frac{1}{p_{ij}}$. In other words, a memory-less erasure channel with success probability p_{ij} can be viewed as a pipe with *random* delay (distributed as a geometric random variable) to deliver a packet. To simplify the Main Problem, we relax each channel into a pipe with *deterministic* delay equal to the inverse of its success probability. Therefore, for any packet of Rx_j, when assigned to AP_i for transmission, we associate a deterministic delay of $1/p_{ij}$ for its delivery. This means that each packet assigned to an AP can be viewed as an object with a weight (representing the delay for its delivery), where the weight varies from one AP to another (since p_{ij}'s for different i's are not necessarily the same). On the other hand, each AP has τ time-slots during each interval to send the packets assigned to it. Therefore, we can view each AP during an interval as a bin of capacity τ. Hence, our new problem can be viewed as a packing problem: we want to maximize the number of objects that we can fit in those N bins of capacity τ. We call this problem the *Relaxed Problem (RP)*, and denote its solution, i.e. the maximum possible number of packed objects, by C_{det}. It is easy to see that RP is an integer program, which is a special case of the generalized assignment problem (see e.g., [5]).

4. Main Results

Our main contribution in this paper is two-fold. First we show that the Main Problem can be approximated via its deterministic relaxation, discussed above. This is stated in the following theorem (the reader is referred to [3,4] for the proof).

Theorem 1. $-2\sqrt{N\left(C_{det}+\frac{N}{4}\right)} < C_{T^3} - C_{det} < N.$

Note that the number of AP's (i.e., N), is typically small (compared with M), hence in the large throughput regime, Theorem 1 implies a good approximation of C_{T^3} via solving its deterministic relaxation (i.e., C_{det}). Moreover, the approximation gap in Theorem 1 is a worst-case bound, and via numerical analysis (see [3] for more details) one can observe that the gap between the Main Problem and RP is in most cases much smaller. Hence, the solution to RP tracks the solution to the Main Problem very well, even for a limited number of clients.

Our second contribution is to show that the deterministic relaxation of the main problem can be solved efficiently (efficient in both approximation gap as well as computational complexity) via a linear-programming (LP) relaxation, as stated below.

Theorem 2. *Denote any feasible packing for RP by a N-by-M binary matrix where element* (i,j) *is 1 if and only if packet requested by* Rx_j *is packed in bin i. Suppose that* $x^* = [{x_{ij}}^*]$ *is a basic optimal solution to the LP relaxation of RP. Then,* $C_{det} - \sum_{i=1}^{N}\sum_{j=1}^{M}\lfloor {x_{ij}}^* \rfloor \le N$.

Note that finding a basic optimal solution to an LP efficiently is straightforward (see e.g., [6]). Hence, Theorem 2 provides an efficient approximation algorithm for RP with additive gap of at most N.

Overall, Theorems 1 and 2 show that we can efficiently approximate the Main Problem, by solving an LP relaxation of its deterministic correspondent.

5. Numerical Results

In this section, we numerically demonstrate the impact of the proposed timely-throughput optimization approach, compared with a greedy scheduling approach that ignores the deadline requirements of the packets. We consider the network in Figure 2, where there is 1 Base Station (BS) in the middle, and 4 femto base stations around it, and a total of 100 receivers randomly located in the coverage area of BS and/or femtos. The erasure probabilities of the channels are assigned based on the AP-client distances.

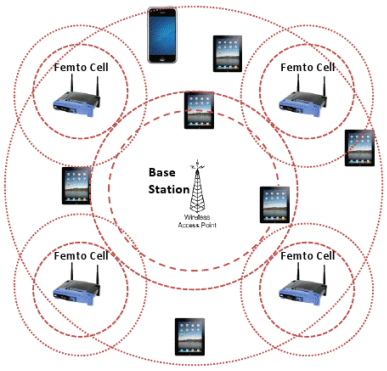


Figure 2. A network with 5 access points and 100 clients.

We consider 3 different scheduling algorithms. The first algorithm ignores the femtos, and only the BS is used for delivering the packets. In the second algorithm (Greedy Algorithm) each receiver connects to the AP that has the best channel to it. Finally, the third algorithm (DS Algorithm) considers the deterministic relaxation of the problem, solves its LP relaxation, and rounds the solution to obtain a scheduling policy that assigns clients to AP's. For 30 different realizations of this setup we run the network for 1000 intervals. If a receiver does not receive at least 50% of its packets before the deadline (during each period of 50 intervals), it will be considered to be in outage for that period. Table 1 demonstrates the average percentage of intervals that clients are in outage (averaged over all clients) for the aforementioned three algorithms. As we note, our proposed timely-throughput maximization approach that efficiently utilizes AP's for traffic delivery, results in large gains, as opposed to a simple greedy algorithm that ignores the deadline requirements of the clients.

6. Conclusion

We considered the timely-throughput maximization of heterogeneous wireless networks, and proposed a deterministic relaxation to efficiently approximate the problem. Our results can be extended in several directions, such as time-varying channels and traffic demands, imposing priority constraints on the clients, and considering fading channels (for which a timely-reward model can be used to formulate the problem). The reader is referred to [3] for these extensions.

Table 1

	Base-Station Only	Greedy Algorithm	DS Algorithm
Average Outage	0.6541	0.1095	0.042

References

[1] I-H. Hou, V. Borkar, and P.R. Kumar, "A theory of QoS for wireless," In Proc. of IEEE INFOCOM, 2009.
[2] I-H. Hou, A. Truong, S. Chakraborty, and P.R. Kumar, "Optimality of periodwise static priority policies in real-time communications," In Proc. of CDC, 2011.
[3] S. Lashgari and A.S. Avestimehr, "Timely Throughput of Heterogeneous Wireless Networks: Fundamental Limits and Algorithms," submitted to IEEE Transactions on Information Theory, available at http://arxiv.org/abs/1201.5173.
[4] S. Lashgari and A.S. Avestimehr, "Approximating the Timely Throughput of Heterogeneous Wireless Networks," In Proc. of IEEE ISIT, 2012.
[5] D. B. Shmoys and E. Tardos, "An approximation algorithm for the generalized assignment problem," *Mathematical Programming*, 62:461474, 1993.
[6] K Jain, "A factor 2 approximation algorithm for the generalized Steiner network problem," Combinatorica, Springer, 2001.

Sina Lashgari received his B.Sc. in EE from Sharif University of Technology in 2010. He joined Cornell University in 2010, where he is pursuing his Ph.D. in the School of Electrical and Computer Engineering. His research is focused on delay-constrained communication in heterogeneous networks, as well as interference management in wireless networks using delayed network-state information. In 2010 Sina received a Irwin D. Jacobs fellowship.

A. Salman Avestimehr is an assistant professor in the School of ECE at Cornell University. He received his Ph.D. in 2008 and M.S. degree in 2005 in EECS, both from the University of California, Berkeley. Prior to that, he obtained his B.S. in EE from Sharif University of Technology in 2003. He has received several awards including the ComSoc and IT Society Joint Paper Award in 2013, the PECASE award in 2011, the NSF CAREER award in 2010, and the AFOSR YIP award in 2010. Dr. Avestimehr has served as a Guest Associate Editor for the IEEE Transactions on Information, and has co-organized the 2012 North America Information Theory Summer School.

Video Caching and D2D Networks

Giuseppe Caire, Andreas F. Molisch, and Michael J. Neely
Dpt. of EE, University of Southern California
{*caire, molisch, mjneely*}*@usc.edu*

1. Introduction

Video transmission is the driving force for the growth in wireless data traffic; it is anticipated to grow by almost two orders of magnitude over the next five years. The use of new spectrum and better spectral efficiency of modulation schemes are not able to accommodate such a significant growth. Increased spatial reuse through deployment of new base stations, in particular femto-stations, may be a viable option; however, the necessity of *backhaul* from all of those new stations can be prohibitive and/or too costly. In our work we have thus suggested a new network structure that exploits a unique feature of wireless video, namely the high degree of (asynchronous) content reuse. Based on the fact that storage is cheap and ubiquitous in today's wireless devices, we suggest to replace *backhaul* by *caching*. Notice that caching is a well-known solution in current Content Distribution networks (CDNs). Nevertheless, CDNs are implemented in the Internet cloud, and this approach has not solved either the deficit in the wireless capacity and the bottleneck problem in the last mile backhaul. In contrast, our proposed approach consists of caching directly at the wireless edge (through dedicated helper nodes) and/or in the user devices. Therefore, our vision is radically new and represents a major step forward with respect to conventional CDNs. From the caching locations video is then transmitted through highly efficient short-distance wireless links to the requesting users. Thus, caching yields higher spectrum spatial reuse at much lower deployment (CapEX) and operating (OpEX) cost. Caches can be refreshed at a much lower rate through the standard cellular infrastructure (e.g., the local LTE base station) at off-peak times. Hence, our solution also offers a new usage opportunity of the existing cellular infrastructure. We envision a progression that starts with the widespread deployment of helper stations and culminates with caching on the user devices and (Base-station controlled) device-to-device (D2D) communications for efficient exchange of video files between users.

2. Content reuse

While traditional *broadcasting* is not accepted by customers, who prefer "on demand" viewing, time-accumulated viewing statistics still show that a few popular videos (YouTube clips, sports highlights, and movies) account for a considerable percentage of video traffic on the internet. The popularity distribution, which often follows a Zipf distribution, changes only on a fairly slow timescale (several days or weeks); it can furthermore be shaped by content providers, e.g., through pricing policies, or through offering of a large but limited library of popular movies and TV shows (as currently done by Netflix, Amazon Prime and iTune). It is thus possible for helper stations and devices to obtain popular content, e.g., through wireless transmission during nighttime, so that they are available when mobile devices demand the content. Due to the steep price drop in storage space, 2 TByte of data storage capacity, enough to store 1000 movies, cost only about 100$ and could thus be easily added to a helper station. Even on mobile devices, 64 GByte of storage could be easily dedicated to caching. When considering n users in a system, the size of the file library of interest (i.e., all the files that these users are *interested in*) grows – usually sublinearly – with n.

3. Helper Systems

We proposed a system nicknamed ``femto-caching'', where a number of dedicated helper nodes (e.g., Wifi or LTE femtocells) are deployed at fixed locations and are connected to power sources but not necessarily to a wired backhaul. Helpers have on-board caches, refreshed periodically by the system provider, through a variety of existing networks (cellular, wired, or a mix thereof). Users place their requests by connecting to the system video server, which re-directs the request to the subset of local helpers that contain the requested file. Then, the video streaming process takes place on the local high-rate links, allowing a large spatial spectrum reuse and avoiding the cluttering of the conventional cellular infrastructure. In this scenario, we have addressed two fundamental problems: 1) optimal file placement; 2) dynamic scheduling for video streaming from multiple helpers to multiple users.

Optimal File Placement:

The network formed by H helpers and N users can be represented by a bipartite graph $G(H,U,E)$ where an edge $(h,u)\ E$ indicates that helper h can communicate to user u. Each helper can contain M files (assumed of constant size for simplicity of exposition). Each link (h,u) is characterized by an average downloading time per bit . The goal is to place a library of m files into the helper caches such that the total average downloading time, for a given file request distribution, is minimized. In [1] we showed that this problem is NP-hard, but can be cast as a maximization of a sub-modular function subject to a matroid constraint, for which a simple greedy approach (place files one by one by maximizing

the objective function increase at each step) yields a solution within a factor ½ from the optimal. Furthermore, using intra-session coding (e.g., random linear network coding or quasi-MDS Raptor coding) yields a convex relaxation of the original combinatorial problem that can be solved via a linear program.

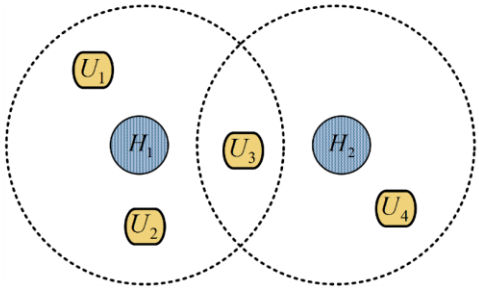


FIGURE 1: Non-trivial example of the file placement problem: users 1, 2 and 4 would like helpers 1 and 2 to both contain the same *M* most popular files, but user 3 would like helpers 1 and 2 to contain the *2M* most popular files.

Dynamic Scheduling for Video Streaming

Assume a certain network configuration with helpers, users, and set of requests. Here the problem is to devise a dynamic scheduling scheme such that helpers feed the video files sequentially (chunk by chunk) to the requesting users. In [2] we have solved this problem in the framework of Lyapunov Drift Plus Penalty (DPP) scheduling. We represent a video file as a sequence of chunks of equal duration. Each chunk may contain a different number of source-encoded bits, due to variable bit-rate (VBR) coding, and the same video file is encoded at different quality levels, such that lower quality levels correspond to less encoded bits. We pose the problem of maximizing a concave non-decreasing function of the users' long-term average quality indices, where the concavity imposes some desired notion of fairness between the users. The policy decomposes naturally into two distinct operations that can be implemented in a decentralized fashion: 1) Admission control; 2) Transmission scheduling. Admission control decisions are made at each streaming user, which decides from which helper to request the next chunk and at which quality index this shall be downloaded. This scheme is compatible with DASH (Dynamic Adaptive Streaming over HTTP), where the clients progressively download a video file chunk by chunk. In our system, the requested chunks are queued at the helpers, and each helper maintains a queue pointing at its served users. Users place their chunk requests from the helpers having the shortest queue pointing at them. Then, transmission scheduling decisions are made by each helper, which maximizes at each scheduling decision time its downlink weighted sum rate where the weights are provided by the queue lengths. The scheme provably achieves optimality of the concave objective function (network utility function) and can be implemented in a decentralized manner, as long as each user knows the lengths of the queues of its serving helpers, and each helper knows the individual downlink rate supported to each served user. Queue lengths and link rates represent rather standard protocol overhead information in any suitable wireless routing scheme. We have also implemented a version of such scheme on a testbed formed by Android smartphones and tablets, using standard WiFi MAC/PHY [8]. In [9] we have developed a related "video-pull" scheme for cache-aware scheduling (without the adaptive video component) in networks with general interference models and with tit-for-tat incentives on user participation.

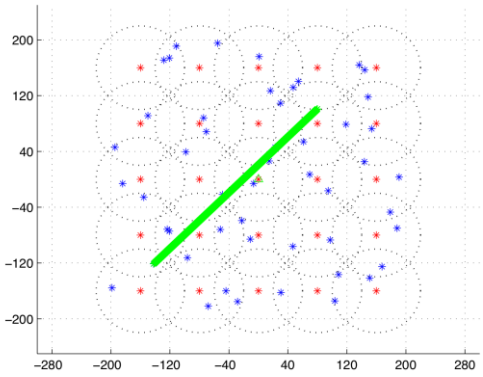


FIGURE 2: A simulation of our scheduling algorithm where a user moves across the network of helpers (green straight path) and the scheduler seamlessly ``discovers'' new serving helpers as the user moves in their proximity.

4. Device-to-device communications

We now consider a system where users cache video files on their own devices and exchange them through D2D communications; files that cannot be transmitted through D2D are sent through the traditional cellular downlink. Our overall goal is to minimize the download time of videos, or maximize the traffic that is handled by D2D links.

Consider a system with *n* users in a cell, and assume that all users employ the same transmit power, and thus have the same communication radius *r(n)*. We can divide our cell into "clusters" of size *r*, and assume that within each cluster, only a single link can be active at a time, but links on different clusters can operate on the same time/frequency resources. We then see that finding a good communication radius requires a tradeoff: as we decrease *r*, the number of clusters and thus the spatial reuse of the time/frequency resources) increases. On the other hand, decreasing *r* also decreases the probability that a user will find the file

among its neighbors (nodes within the communication distance) [3].

Now what files should be stored by the devices? If D2D capabilities would not exist, then clearly each user should store the most popular files, as this would increase the chance of a "self request", i.e., the probability that the user wants what is already in its cache. However, in the presence of D2D capabilities, different devices that can talk to each other should store different files; we can imagine the devices within one cluster to form a "central virtual cache" CVC, such that all files within the CVC can be efficiently distributed to requesting users. If the BS can centrally control what the devices cache, and the devices do not move, then each file should be stored only once in the CVC, i.e., the content of the caches of devices within one cluster should not overlap. If, however, it cannot be anticipated which device will be located in which cluster at the time of request, then a random caching strategy based on a particular probability density function is the best approach; in [4] we determine that the optimum such distribution has a waterfilling-like structure.

4.1 Optimal Scaling Laws

In [4] we defined the throughput-outage tradeoff of D2D one-hop caching wireless networks, where the throughput is defined as the minimum (over the users) of the long-term average throughput, and outage probability is the fraction of users that cannot be served by the D2D network. Our achievability strategy consists of random caching, according to the optimal caching distribution that maximizes the probability that a random request made according to the popularity probability , is satisfied within a cluster of given size *g*. Then, the network is partitioned into such clusters, and one link per cluster is activated in each time-frequency transmission resource according to an interference avoidance scheme (independent set scheduling). We can prove that in the regime where the library size *m* scales not faster than linearly with the number of users *n*, for any desired outage probability , the throughput scales as . Notice that when $M/m >> 1/n$, i.e., when the number of users is much larger than the ratio between the library size *m* and the cache size *M*, the caching network yields a very large gain with respect to a conventional system that broadcasts from a single base station. Remarkably, the throughput increases linearly with the device cache size *M*, i.e., we can directly tradeoff local memory (cheap and largely available resource) for bandwidth (expensive and scarcely available resource).

4.2 Comparisons

In [5], a completely different caching scheme based on coded multicast is proposed. In the coded multicast scheme users cache segments (packets) of the video files from the library, such that the total cache size corresponds to *M* files, but no file is entirely cached locally. In the presence of a certain set of demands, the base station sends a common coded message (coded multicast) such that all requesting users can reconstruct their demanded file. The scheme is best explained through a simple example: consider the case of $n = 3$ users requesting files from a library of $m = 3$ files, A, B and C. Suppose that the cache size is $M = 1$ file. Each file is divided into three packets, A1, A2, A3, B1, B2, B3 and C1, C2, C3, each of size $1/3$ of a file. Each user *u* caches the packets with index containing *u*. For example, user 1 caches A1,B1,C1. Suppose, without loss of generality, that user 1 wants A, user 2 wants B and user 3 wants C. Then, the base station will send the message [A2+B1, A3+C1, B3+C2] (sums are modulo 2 over the binary field), of size 1 file, such that all requests are satisfied. Interestingly, it can be shown that this scheme yields the same throughput scaling law of our D2D scheme, namely, . It follows that the two schemes must be compared on the basis of the actual throughput in realistic propagation conditions, rather than simply in terms of scaling laws. We made such comparison with the help of Monte Carlo simulation for a randomly placed network of caching user nodes deployed in a 1 sq.km area, with realistic Winner II channel models emulating a mix of indoor and outdoor links, with buildings and distance dependent blocking probability, line of sight propagation and shadowing. We also compared with today's state of the art schemes, i.e., broadcasting independently to the users from an LTE base station (no content reuse), and harmonic broadcasting the most popular files in the library [6]. Our simulations (see details in [7]) show that the D2D caching network yields significantly better performance in terms of throughput vs. outage tradeoff with respect to all other schemes (see Fig. 3).

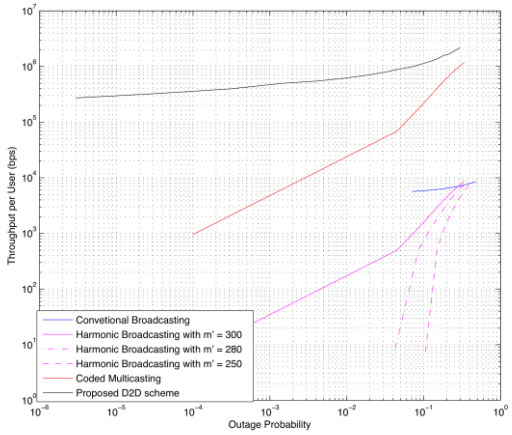


FIGURE 3: Throughput-outage performance of the D2D caching network and other competing schemes in a realistic propagation environment.

5. Summary and future work

In the framework of the VAWN project, we have developed a comprehensive framework for caching in wireless networks targeted to on-demand video streaming, which is a major killer application at the basis of the predicted x100 increase in wireless traffic demand in the next 5 years. We considered two related network architectures: caching in wireless helper nodes (femtocaching), such that users stream videos from helpers in their neighborhood, and caching directly in the user devices, such that users stream videos from other users via D2D connections. In both cases we have shown large potential gains and solved key problems in the design and analysis of such networks. Current and future work includes: 1) considering more advanced PHY schemes in femtocaching networks (e.g., helper nodes may have multiple antennas and use multiuser MIMO and advanced interference management schemes, beyond the simple WiFi-inspired schemes considered so far); 2) considering more advanced PHY schemes for D2D networks, beyond the simple spectrum reuse and interference avoidance clustering scheme used so far; 3) considering the possibility of limited multi-hop in D2D caching networks; Considering a combination of coded multicasting and D2D local communication, with the goal of combining the advantages of coded multicasting with these of D2D dense spectrum reuse.

In parallel, we are also implementing compelling testbed demonstrations of femtocaching networks and D2D networks for video streaming (a first example of which will be presented in Mobicom 2013 [8]).

References

[1] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless video content delivery through distributed caching helpers," Proc. of IEEE INFOCOM, pp. 1107-1115, 2012.

[2] D. Bethanabhotla, G. Caire and M. J. Neely, "Utility Optimal Scheduling and Admission Control for Adaptive Video Streaming in Small Cell Networks,'' Proc. of IEEE Int. Symp. On Inform. Theory, Istanbul Turkey, July 6-11, 2013.

[3] N. Golrezaei, P. Mansourifard, A. F. Molisch, A. G. Dimakis, "Base-Station Assisted Device-to-Device Communications for High-Throughput Wireless Video Networks", ArXiv Preprint 1304.7429.

[4] Mingyue Ji, G. Caire and A. F. Molisch, "Optimal Throughput-Outage Trade-off in Wireless One-Hop Caching Networks,'' Proc. IEEE Int. Symp. Inform. Theory, Istanbul Turkey, July 6-11, 2013.

[5] M. Maddah-Ali and U. Niesen, "Fundamental Limits of Caching, '' ArXiv Preprint 1209.5807.

[6] Li-Shen Juhn and Li-Ming Tseng, "Harmonic broadcasting for video-on-demand service," IEEE Trans. On Broadcasting, vol. 43, no. 3, pp. 268–271, 1997.

[7] Mingyue Ji, G. Caire and A. F. Molisch, "Wireless Device-to-Device Caching Networks: Basic Principles and System Performance," ArXiv Preprint 1305.5216.

[8] J. Kim, A. F. Molisch, G. Caire, and M. J> Neely "Adaptive Video Streaming for Device-to-Device Mobile Platforms," to appear in MobiCom 2013.

[9]M. J. Neely, "Optimal Peer-to-Peer Scheduling for Mobile Wireless Networks with Redundantly Distributed Data," IEEE Transactions on Mobile Computing, to appear (http://doi:ieeecomputersociety.org/10.1109/TMC2013.21).

GIUSEPPE CAIRE [S'92, M'94, SM'03, F'05] was born in Torino, Italy, in 1965. He received the B.Sc. in Electrical Engineering from Politecnico di Torino (Italy), in 1990, the M.Sc. in Electrical Engineering from Princeton University in 1992 and the Ph.D. from Politecnico di Torino in 1994. He is currently a professor of Electrical Engineering with the Viterbi School of Engineering, University of Southern California, Los Angeles, CA.

ANDREAS F. MOLISCH [F'05] (molisch@usc.edu) is Professor of Electrical Engineering at the University of Southern California. His research interests include cooperative communications, wireless propagation channels, MIMO, and ultrawideband (UWB). He is a Fellow of AAAS, Fellow of IET, and Member of the Austrian Academy of Sciences, as well as recipient of numerous awards.

MICHAEL J. NEELY received his PhD in Electrical Engineering from MIT in 2003. He joined the faculty of Electrical Engineering at the University of Southern California in 2004, where he is currently an Associate Professor. His research interests are in the areas of stochastic network optimization and queueing theory, with applications to wireless networks, mobile ad-hoc networks, and switching systems. He is the recipient of several awards.

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

Shiwen Mao	Director	Aburn University	USA
Guosen Yue	Co-Director	NEC labs	USA
Periklis Chatzimisios	Co-Director	Alexander Technological Educational Institute of Thessaloniki	Greece
Florin Ciucu	Editor	TU Berlin	Germany
Markus Fiedler	Editor	Blekinge Institute of Technology	Sweden
Michelle X. Gong	Editor	Intel Labs	USA
Cheng-Hsin Hsu	Editor	National Tsing Hua University	Taiwan
Zhu Liu	Editor	AT&T	USA
Konstantinos Samdanis	Editor	NEC Labs	Germany
Joerg Widmer	Editor	Institute IMDEA Networks	Spain
Yik Chung Wu	Editor	The University of Hong Kong	Hong Kong
Weiyi Zhang	Editor	AT&T Labs Research	USA
Yan Zhang	Editor	Simula Research Laboratory	Norway