Intel® Technology Journal

APRIL 2015

Video-aware Wireless Networks

Publisher
Cory Cox

Managing Editor
Stuart Douglas

Content Architects
David Ott and Jeff Foerster

Program Manager
Stuart Douglas

Technical Editor
David Clark

Technical Illustrators
MPS Limited

Technical and Strategic Reviewers
David Ott
Jeff Foerster
Xiaoqing Zhu
Douglas S. Chan

Intel Technology Journal

 ISBN 978-1-934053-66-9, ISSN 1535-864X

Intel Technology Journal
Volume 19, Issue 1

Publisher: Cory Cox

Managing Editor: Stuart Douglas

Library of Congress Cataloging in Publication Data:

Printed in China

10 9 8 7 6 5 4 3 2 1

First printing: April 2015

INTEL® TECHNOLOGY JOURNAL
VIDEO-AWARE WIRELESS NETWORKS

Articles

TOWARDS REALIZING VIDEO AWARE WIRELESS NETWORKS

Contributors

Jeffrey Foerster
Intel Corporation

David Ott
Intel Corporation

Ozgur Oyman
Intel Corporation

Yiting Liao
Intel Corporation

Srinivasa Somayazulu
Intel Corporation

Xiaoqing Zhu
Cisco Systems

Douglas S. Chan
Cisco Systems

Chris Neisinger
Verizon

This introductory article provides background on the formation of a unique industry-academic research initiative focused on the efficient delivery of video content across wireless networks. A multidisciplinary team of researchers from five leading US and international universities partnered with Intel, Cisco, and Verizon to explore novel approaches in this challenging area. Known as Video Aware Wireless Networks, or VAWN, the research program included experts in the areas of wireless communications, networking, video quality, and video processing. A key focus was optimizing the end user perceptual quality of experience while trying to minimize network resources. This article summarizes the market trends that motivated this research, provides industry perspectives from Verizon, Cisco, and Intel on the opportunities and challenges for realizing an end-to-end optimized video delivery system, and discusses standards and technical transfer considerations. The goal for this article and this special Intel Technical Journal issue is to summarize the key lessons and research ideas from this industry-university initiative and to motivate further work within the industry to realize many of these new ideas in practice.

Introduction

No matter which way you look at it, the trend in video content delivery over wireless networks would seem to be clear: up, up, and up. Whether it's streamed video, Internet video, video on demand, personal video streaming, video sharing applications, video conferencing, live video broadcasting, video Twitter*, or video blogging, wireless network users have never been more interested in consuming and sharing digital video content over wireless networks. And by all indications, this interest will continue to expand rapidly as more users are enabled with more capable wireless devices.

To ground these observations in numbers, consider the *Cisco Visual Networking Index* mobile forecast for 2014–2019.[1] According to measurement data, global mobile data traffic grew by 69 percent in 2014 to 2.5 exabytes per month (one exabyte equals one billion gigabytes). Projections for 2019 anticipate an order of magnitude growth to a staggering 24.3 exabytes. Most importantly for our purposes, mobile video data is expected to grow at a compound annual growth rate (CAGR) of 66 percent between 2014 and 2019 to 17.4 exabytes. As a percentage of global mobile traffic data, mobile video data is expected to grow from 55 percent in 2014 to 72 percent in 2019. Whether seen in absolute terms (total number of bytes) or relative terms (percentage of total traffic), it's clear that video traffic plays a major role in current wireless networks and that its role in future mobile networks can hardly be understated.

The implications of this growth are significant. Demand for digital video content will continue to stress the capacity of both downlinks and uplinks in future wireless communications networks as users find new ways to consume and share video content. For instance, future graphics applications might utilize a mix of video and graphics content to provide an increasingly immersive effect for virtual reality, gaming, navigation, and other contexts yet to be invented. User device manufactures will need solutions to the problem of managing a rich user experience despite finite network resources, complexities that arise from variations in device resolution and processing capabilities, and the ever-present pressure of limited battery life. In general, the trend is for device platforms to become more capable and for users to demand video of increasingly higher quality (such as HD and 4k content).

"Demand for digital video content will continue to stress the capacity of both downlinks and uplinks in future wireless communications networks as users find new ways to consume and share video content."

In 2010, Intel, Cisco, and Verizon jointly established a research program entitled *Video Aware Wireless Networks (VAWN)* to explore innovative solutions to these challenges through university research (see Figure 1). In particular, the program looked at three key vectors for enabling future video content delivery over wireless networks: *video transport optimizations, video processing optimizations*, and *novel network architectures for video delivery*. The first, video transport optimizations, looks at how wireless network elements, both within the network and at network endpoints, could be modified or redesigned to better serve the transport needs of digital video. The second, video processing optimizations, looks at how digital video itself and processing mechanisms at the source and destination could be redesigned to better comprehend the context of wireless transport. The last vector, novel network architectures, asks whether significant

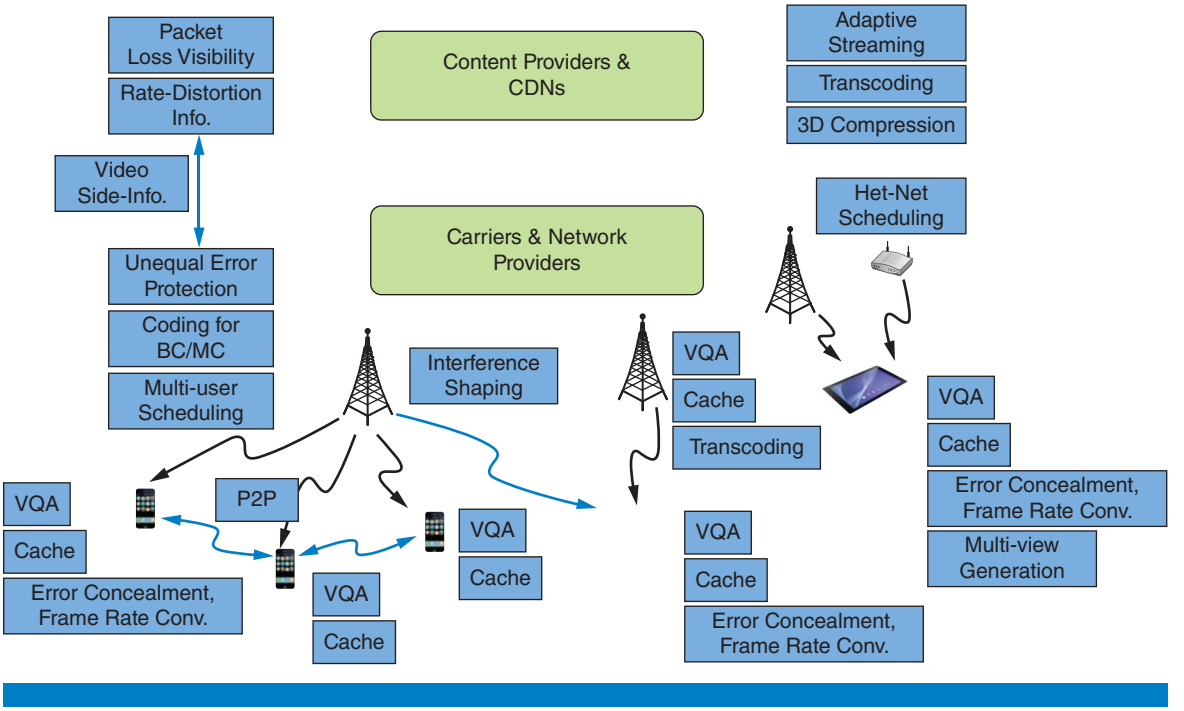


Figure 1: Video Aware Wireless Networks (VAWN)
(Source: Intel Corporation, 2014)

advantages could be achieved if wireless network architectures were redesigned and what that experimental redesign would look like. All three vectors address two key grand challenges within the program: *increasing the capacity of future wireless networks to deliver digital video content* and *improving the quality of experience for users accessing digital video over future wireless networks*.

Universities participating in this initiative included Cornell University, Moscow State University, the University of California at San Diego, the University of Southern California, and the University of Texas at Austin. Working with multiple industry sponsors (Intel, Cisco, and Verizon), research was understood to be precompetitive and approaches were freely discussed by all participants throughout the program duration of three years. During that time, lead principal investigators, along with their collaborators and graduate students, published over 100 papers in leading conferences and journals and generated a wide spectrum of results that speak to program challenges.

"…researchers and sponsors present an overview of key research results and recommendations for the industry on digital video transport over wireless networks."

In this edition of the ITJ, VAWN researchers and sponsors present an overview of key research results and recommendations for the industry on digital video transport over wireless networks.

Wireless Carrier Challenges and Opportunities

Wireless operators are experiencing tremendous network growth as a result of consumer demand for video services. This has been fueled by the advent of smartphones, tablets, high resolution screens, and the changing content on the Internet. As devices have improved, the demand for quality has also increased. For example, YouTube content has changed from home movie quality (cute cat videos) to professionally created mobile-specific content. Periodic blocking and stalling of this professional quality content is no longer tolerated. This leads to customer expectations of wire-line performance on a wireless network; customers want immediate delivery whenever they request it and wherever they happen to be. They also expect simple and affordable service offerings.

The operator is also under increasing pressure to improve efficiency. The unprecedented data growth over the past 10 years has come at a time when increased spectral efficiency technology improvements from 2G to 3G to 4G LTE have been sufficient to keep up with the demand and drive the telecom industry ecosystem. However, the spectral efficiency gains beyond 4G will be minimal. The wireless operator must find other methods that increase efficiency in order to keep up with consumer demand and maintaining profitability.

The far right side of the chart in Figure 2 illustrates a potentially unstable ecosystem. The cost of the network exceeds the customer's willingness to pay. Clearly the challenge is to find additional techniques to significantly improve efficiency and push the crossover as far to the right as possible. This improved efficiency will enable a stable ecosystem that will support the growth of wireless well into the future.

"The challenge for the operator is to find network management techniques that increase network efficiency while maintaining or improving customer experience."

The challenge for the operator is to find network management techniques that increase network efficiency while maintaining or improving customer experience.

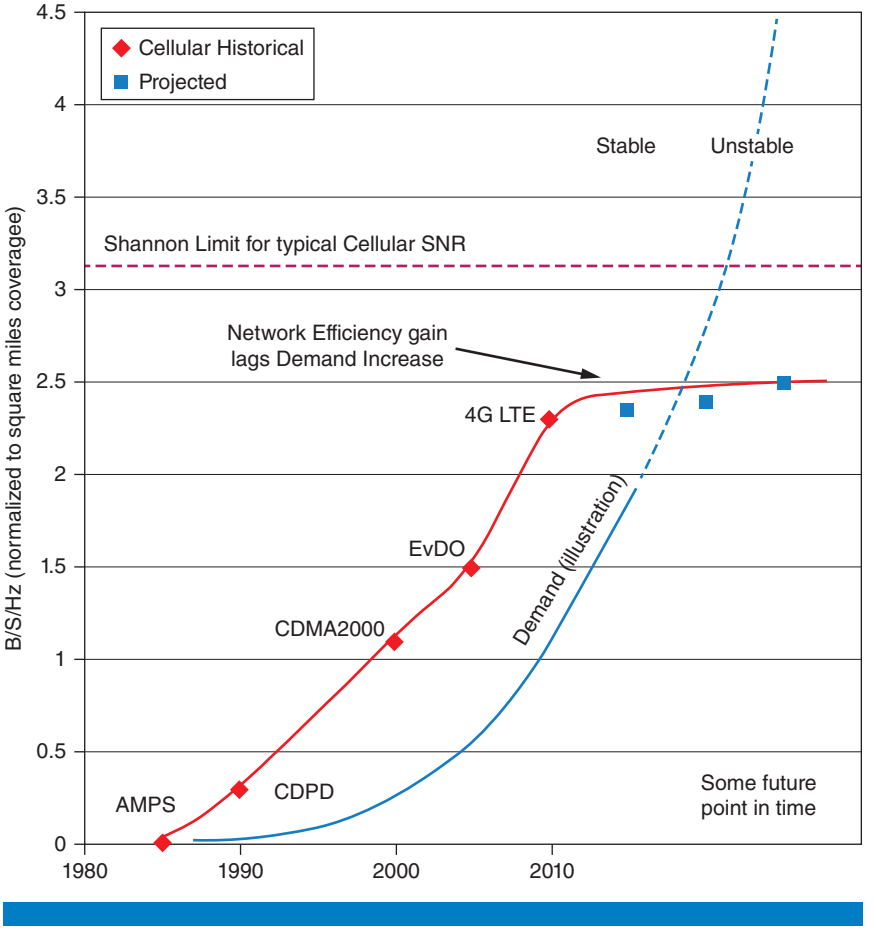


Figure 2: Network capacity demands and cellular technology trends (Source: Verizon, 2014)

Similar situations appeared when the fixed Internet transitioned toward video content. Prior to the introduction of content delivery networks (CDNs), the core was becoming uneconomical to augment. The answer was not to build a bigger core network to handle the traffic, but rather to change the architecture to be more efficient. The Internet edge had unused capacity but the core could not pass the traffic. CDNs rebalanced this to effectively increase the overall efficiency and capacity of the Internet. The purpose of the VAWN initiative is to find technology and architecture opportunities for unlocking capacity in the wireless Internet.

"The purpose of the VAWN initiative is to find technology and architecture opportunities for unlocking capacity in the wireless Internet."

Current measures of quality for wireless networks often revert to average speed tests. But speed is only a surrogate for quality. Speed tests are relatively simple to implement, report, and advertise. However, average speed is not always a good measure for true quality of experience (QoE). Average speed is a great way of indicating the quality of a network if the application is FTP. But the wireless Internet handles nearly zero FTP traffic. Since the majority of the traffic on the wireless network is video, it makes much better sense to measure the quality of a network by some methods that actually convey the user-perceived quality of video. Methods to improve network efficiency must be better balanced by placing more emphasis on QoE and reducing the

"…QoE metrics need to be technically accurate enough to drive real network improvements, yet easily communicated to the lay consumer."

"The rapidly growing demand for wireless video content, calls for a fundamental change in the design of networking technologies and solutions."

importance of average speed. Accompanying solutions will also be needed for the advertising positioning problem: using speed as a surrogate for quality makes it easier to craft a 6-second sound bite advertising network service. In general, consumers do not understand latency, jitter, frame error rate (FER), or macro block pixilation. As such, the industry needs to develop standard methods and procedures for conveying true quality of experience. These QoE metrics need to be technically accurate enough to drive real network improvements, yet easily communicated to the lay consumer.

Networking Challenges and Opportunities

The rapidly growing demand for wireless video content, along with elevated user expectations for video quality and viewing experience, calls for a fundamental change in the design of networking technologies and solutions. It will no longer suffice for the network device (such as the router, switch, or gateway) to simply act as a "dumb pipe" and diligently pump packets from point A to point B at the highest possible speed, regardless of what type of information those packets carry. Instead, it is our vision that intelligent participation from the network—when done right—can help to alleviate the video-induced bandwidth crunch over wireless networks while maximizing user QoE. Much of the research effort in the VAWN program has focused on figuring out how to do it right.

One way to realize such a vision is to open up a richer set of APIs (or information exchange channels) between network devices and video application endpoints. This will allow network devices at resource bottlenecks to make informed decisions to optimize for users' quality of experience in lieu of optimizing for conventional quality of service metrics (for instance, bandwidth, loss, delay, and delay jitter). For example, in the presence of unforeseen network congestion, a video-aware router can choose to only drop video packets marked by endpoints at relatively low priority, either within an individual video flow or across multiple video flows, or only drop non-video packets, so as to mitigate the impact of packet drops on perceived video quality at the receiver. Alternatively, the information can flow along the reverse direction: from the network to the video endpoint. A video sender can benefit from explicit congestion notifications from network nodes that proactively learn and predict impending congestion, thereby adapting more swiftly to avoid congestion-induced packet losses in the first place. As discussed further in the section "Existing and Emerging Standards," success in deploying such cross-layer approaches depends not only on the ingenuity of the specific technical solutions themselves, but also on successful adoptions of international standards that push for this direction in general.

Interestingly, while near-term transition to a QoE-optimized network needs to overcome foreseeable hurdles in standardization, the emerging software-defined networking (SDN) technologies may open up even greater opportunities for QoE-based network optimization in the long run. In SDN, the network controller is implemented as software. It is not only logically, but also physically separated from data-plane operations of packet forwarding that remains lightning fast and straightforward. In addition, one network controller

may have a centralized view of the global network topology within its own domain. Such a novel architecture may therefore afford greater flexibility, visibility, and simplicity in defining and instantiating packet treatment policies per application or per flow. The networking industry is currently busy defining a new standard interface between network controllers and data-forwarding devices; the timing cannot be better for transporting the thinking and learning gained from the VAWN program to the yet-to-be-defined arena of SDN.

"The networking industry is currently busy defining a new standard interface between network controllers and data-forwarding devices…"

Embedding intelligence within the network goes beyond improving existing networking functionalities. Technological advances have continued to improve the capability of computing and storage resources at ever-decreasing unit prices. It therefore makes perfect sense to push further the idea of CDNs and to augment today's network edge nodes (such as WiFi* access points and cellular base stations) with affordable computing and storage resources. Intelligent network edge nodes can be leveraged to improve the wireless video transport in many ways: distributed content caching and pre-fetching, on-the-fly video transcoding based on wireless link quality of each client, intelligent soft handoff following users' mobility patterns, and radio-aware resource management among competing video clients. The resulting networking infrastructure will be a diverse, distributed, and heterogeneous platform—fog computing—that complements today's centralized cloud computing paradigm.

Mobile Device and End-to-end Considerations

It's important to understand the end-to-end system in order to ensure a high quality of experience while at the same time minimizing the network resources needed to deliver that experience. Figure 3 shows three main elements

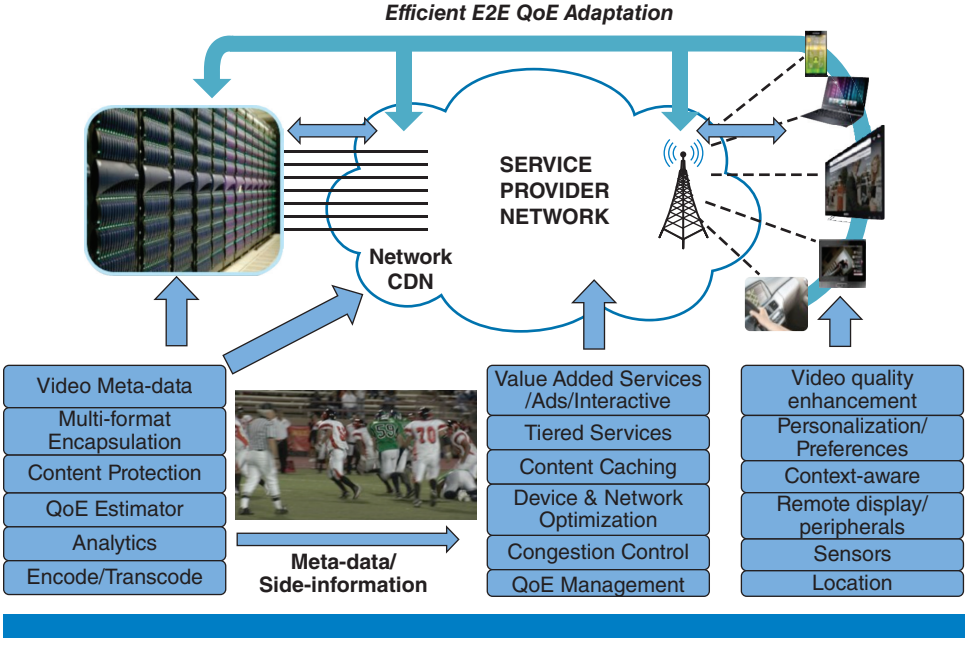


Figure 3: Opportunities in an end-to-end system (Source: Intel Corporation, 2014)

(simplified) to deliver a video streaming solution: (i) server, (ii) network, and (iii) client. Each element has a different responsibility in the end-to-end system, but can be improved through close cooperation between the elements which is one of the primary objectives of this university research and one of the significant challenges in the industry to realize.

"Adaptive streaming applications are primarily driven by the client that requests a particular bandwidth representation from the server…"

Adaptive streaming applications are primarily driven by the client that requests a particular bandwidth representation from the server, which is then streamed in an appropriate format over the network to the client. When selecting the bandwidth representation, the client needs to estimate the available network bandwidth and can also take into account the amount of content in the buffer, battery life, environmental conditions, and user/device context information. The primary role of the server on the left is to host the cloud-based application, store content coming from a content provider, manage content protection requirements, and encapsulate the content into the appropriate format that the client application can understand. Today, there are three main adaptive streaming formats, Apple HTTP Live Streaming (HLS), Microsoft Smooth Streaming (MSS), and Adobe HTTP Dynamic Streaming (HDS). In addition, there is a new standard emerging, MPEG DASH, which will hopefully reduce and simplify the number of formats needed in the future. However, in the intermediate term, there will be a need for the servers to support multiple formats, and several server implementations do this through dynamic switching between formats in order to limit how many versions of the content need to be stored. In addition to the multiple formats that need to be supported in the cloud server, adaptive streaming inherently means that multiple bit rates need to be supported at the server. This can be done in one of two ways: (i) multiple versions of the content are stored each with a different bit rate, or (ii) the content is transcoded in real time to the requested bit rate from the mobile device. Approach (i) requires more memory while approach (ii) is more compute intensive. The tradeoff between memory and compute resources in the cloud is still an ongoing debate and the future will likely include both, depending on where the content is stored (in a data center or closer to the edge of the network) and the popularity of the content (frequently requested content might be best stored in multiple bit rates while less frequently requested content might be better to transcode on demand in order to free up more memory). As the cost of both memory and compute resources continue to go down, this tradeoff space will likely change over time.

"…it's important that the network have information on the content being streamed and the status of the client."

In order to efficiently stream video content over the network, especially when multiple users are vying for the same congested resources, it's important that the network have information on the content being streamed and the status of the client. This allows it to make appropriate decisions when dividing up and prioritizing resources fairly among the competing clients and applications. This is where industry cooperation and standards need to come into play to create a framework for information sharing so that the network can optimize the user experience for all users while managing its limited resources. These standards, enabling information exchange between the video

server, network, and clients, has been a major focus within Intel and in the industry to help realize some of the opportunities highlighted by university research.

On the client, adaptive streaming requires the client to accurately estimate the available network bandwidth in order to request the appropriate segment bit rate to be sent by the server. This approximation becomes a challenge, since the available bandwidth is impacted by the other users in the network, the wireless radio channel conditions, and the impact of loss in the network—including the response of TCP back-off. The two main factors adversely impacting a user's quality of experience for streaming applications at the client are number of rebuffering events (frame freezing) and poor video quality rendering. Therefore, the client's objective is to request the highest quality representation from the video server while maintaining a very low probability of rebuffering. In order to do this, the client needs to make decisions based on several inputs: (i) the rate-distortion characteristics of the content, (ii) the status of the current playback buffer, and (iii) the current available network bandwidth. The rate-distortion characteristics of the content could either come from the server in the presentation description (to be discussed more in the section "Existing and Emerging Standards") or be predicted by the client. To predict this at the client, some of the new non-reference video quality metrics could be useful, which received considerable focus by university researchers within the program. Motivated by the university research at UT Austin in this area, and assisted by interns, Intel has developed its own non-reference quality estimator along with a bit-rate estimator which could be used at the client. Figures 4 through 6 explain these algorithms and performance measured against a video quality database. Somewhat surprising is the fact that the non-reference video quality metrics were able to achieve correlations above 0.90, performing almost as well as reference-based quality metrics.

"…adaptive streaming requires the client to accurately estimate the available network bandwidth in order to request the appropriate segment bit rate to be sent by the server."

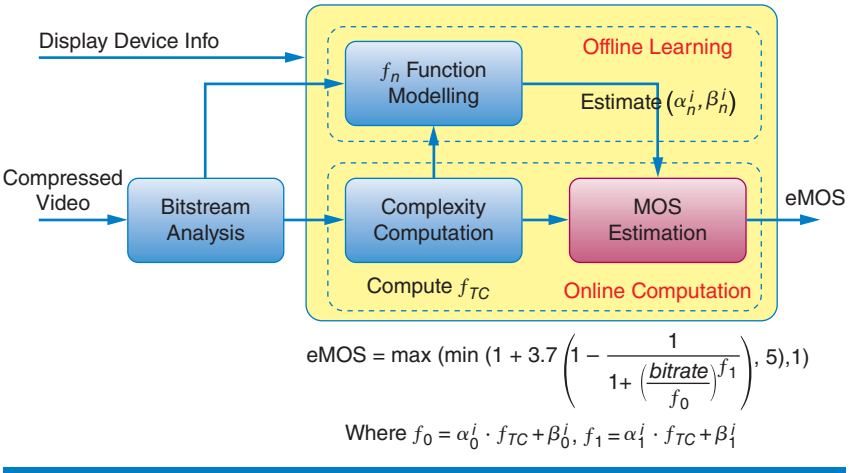


Figure 4: Non-Reference (NR) Mean Opinion Score (MOS) subjective quality estimator framework
(Source: Intel Corporation, 2014)

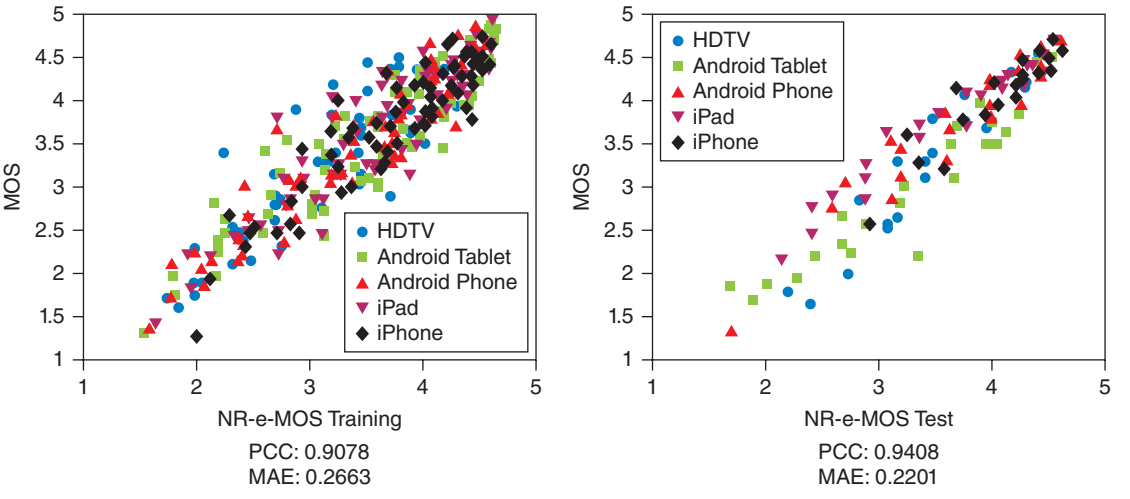


Figure 5: Performance evaluation
(Source: Intel Corporation, 2014)

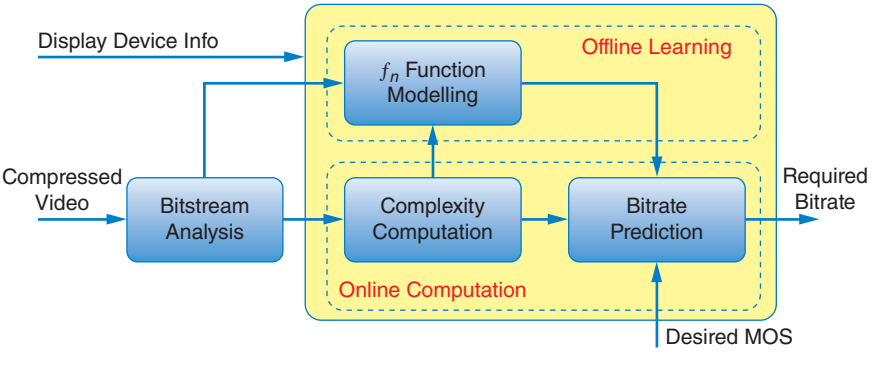


Figure 6: Bit-rate estimator framework based upon target desired quality (MOS) using the NR MOS estimator
(Source: Intel Corporation, 2014)

In order for the client to deliver the best experience possible for a video streaming application, it needs to be cross-layer aware. This means that it is not sufficient just to estimate the available bandwidth and then stream at the maximum bit rate possible (as many existing solutions approach the problem). Rather, the client needs to use information from the application (playout buffer status and video quality characteristics), from the user and device context (environment lighting conditions and device battery state), from the radio (physical layer throughput), and from the transport layer (congestion state and back-off state). This requires close cooperation among elements on the platform, including hardware, firmware, and software. Figure 7 shows an example of how a client platform could become cross-layer aware. University researchers in VAWN have helped to identify and quantify the benefits of the various cross-layer optimization opportunities. Ideally, these ideas need to

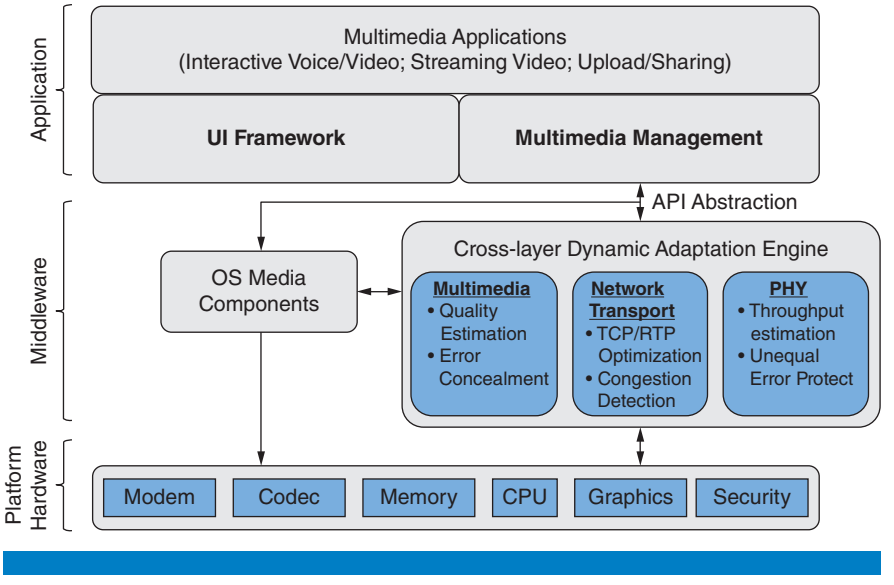


Figure 7: Example of a cross-layer aware platform implementation framework (Source: Intel Corporation, 2014)

be realized on a platform with minimal additional complexity and minimal information sharing between the blocks. This continues to be an active area of research.

Existing and Emerging Standards

The standardization of adaptive streaming over HTTP has made great progress recently, with the technical specifications being completed 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.[2][3] 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.

The scope of both MPEG and 3GPP DASH specifications[2][3] includes a normative definition of a media presentation or manifest format (for an access client), a normative definition of the segment formats (for a media engine), a normative definition of the delivery protocol used for the delivery of segments (namely, HTTP/1.1), and an informative description of how a DASH client may use the provided information to establish a streaming service. In addition to the definition of media presentation and segment formats standardized in the MPEG and 3GPP DASH specifications[2][3], MPEG has also developed additional specifications[4][5][6] addressing implementation guidelines, conformance and reference software, and segment encryption and

authentication. Further discussion of MPEG and 3GPP standardization can be found in articles by Sodagar[7], Stockhammer[8], and Oyman and Singh[9].

"DASH provides a client with the ability to fully control a streaming session."

DASH provides a client with the ability to fully control a streaming session. That is, it can intelligently manage an on-time request and the smooth playout of a sequence of segments, potentially adjusting bit rates 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 bit-rate representations of media content as available bandwidth changes. Hence, DASH allows fast adaptation to changing network and link conditions, as well as user preferences and device states (for example, display resolution, CPU, and memory resources). Such dynamic adaptation provides better user quality of experience (QoE) with higher video quality, shorter startup delays, fewer rebuffering events, and so on.

Intel has been involved in DASH standardization within both 3GPP SA4 and MPEG collaborations and has contributed to the development activity on numerous fronts. These standards enable the necessary exchange of information between a video server, mobile network, and the end client so that each element in the end-to-end delivery system can make the more informed decisions on how best to manage the information flow. This, in turn, enables delivery of the best possible user experience while balancing the needs of every user across the network. With the ultimate goal of delivering the best DASH-based multimedia streaming experience, Intel has also been developing, prototyping, and demonstrating a number of DASH differentiation features. These prototypes serve to demonstrate how intelligent video adaptation algorithms can be constructed using DASH. These solutions are applicable to multimedia frameworks for many OS platforms, including Android*, Windows*, Chrome*, and Tizen*.

Promising New Ideas

The VAWN program committee appreciates the many new ideas and approaches generated by academic researchers throughout the life of the program. In this section, we highlight several promising new ideas that emerged and captured our attention. This list is by no means comprehensive as subsequent articles in this issue of the *ITJ*, which describe program research more fully, will demonstrate.

"…there is considerable potential in the general idea of managing video applications over wireless networks using end user perceptual experience…"

First, we believe there is considerable potential in the general idea of managing video applications over wireless networks using end user perceptual experience as a key control input. This is true for both video streaming and interactive video and is well-aligned with current industry trends of focusing on the end user experience. Program research has shown that it is possible to estimate and predict the quality of a user's video experience through both reference-based and non-referenced-based algorithms. This result is noteworthy and remarkable since, at the beginning of the research initiative, it was unclear whether non-reference-based metrics could achieve even a correlation score of 0.7 when

compared to subjective scoring. In fact, later results far exceeded this target with generic non-reference scores going above 0.8 and codec-specific optimized solutions above 0.9. The effectiveness of these approaches was remarkably close to reference-based solutions. In addition, promising algorithms for wireless network resource allocations based on these metrics have been shown to yield significant gains compared to the current state of the art. Such results are highlighted in various articles to follow within this journal.

A second area of significant opportunity and far-reaching implications is the introduction of distributed, intelligent nodes within the network, especially at the network edge. Several of the universities demonstrated the potential value of caching content at the edge of the network, and even on mobile devices. While gains from the approach are dependent on the popularity of content that is cached and the ability to predict content popularity, the approach has been shown to be promising. Intelligent nodes could also process video content by transcoding or transrating it directly at the edge of the network. Such nodes could leverage platforms with significant compute capabilities and storage, a major focus within the industry ecosystem.

"A second opportunity is the introduction of distributed, intelligent nodes within the network, especially at the network edge."

Another key idea highlighted by the program is the need for devices to communicate directly with one another in order to share content and to improve overall network efficiency. In addition to program research on this topic, device-to-device (D2D) communications is an active area of development within the industry, including standards efforts in both Wi-Fi and cellular/LTE. Understanding the most efficient and robust method for devices to directly communicate to deliver a high quality of experience for multiple applications remains an open research question. This includes the need to avoid impacting overall network efficiency, a key part of research on future 5G networks.

A fourth idea is that of *cross-layer optimization.* University researchers demonstrated and quantified the value of cross-layer optimizations by showing how information within each layer of the OSI stack could be shared with different elements of the end-to-end network. Although cross-layer optimization research has existed in the academic world for decades, it hasn't been fully embraced in the context of video and wireless networks. This is due to the complexity of creating standards to share information across the network and the need for efficient algorithms capable of using this information. University researchers in the VAWN program helped to motivate the value and need for standards changes. In fact, several changes have been realized in both 3GPP-based cellular standards and MPEG DASH standards that address this vision, as discussed in the previous section.

"…researchers demonstrated and quantified the value of cross-layer optimizations…"

Finally, researchers in VAWN have significantly increased our knowledge and understanding of 3D video challenges and opportunities. Although a comprehensive understanding of how to objectively quantify 3D video quality remains an open problem, several good initial steps have been taken and achievements have been made. Eventually, it should be possible to apply the

same optimization techniques that have been developed for 2D video to 3D video. Since compression is critical for all video applications, it's important to understand different compression options in terms of bandwidth and quality tradeoffs. University researchers in VAWN have identified promising ideas using a combination of depth map information and current multi-view video encoding optimizations. Although 3D video is still early in market adoption, our improved understanding of compression and 3D video quality impairments will be invaluable as research in this area continues.

Industry Technology Transfer Considerations

Although university researchers were highly successful in generating new ideas and approaches, and in demonstrating potential gains with respect to wireless network capacity and improved quality of user experience, there exist many challenges to realizing these promising techniques in real wireless networks. The VAWN program committee referred to this as the challenge of *technical transfer* of research results. Here, we highlight several areas that the committee identified as key obstacles to transferring approaches into industry practice.

As mentioned in the previous section, considerable advancements were made by VAWN researchers in video quality assessment and the development of objective metrics that effectively predict user experience of a given segment of natural scene video. This includes both reference and non-reference approaches to quality metrics. To facilitate industry technical transfer, however, there is a pressing need for a large-scale, open access database of video segments that can be used to further refine and optimize metrics for real devices and real network contexts. Segments in the database must span a variety of content types, include both pristine video and distortions, and have a complete set of subjective scores for use as a basis of comparison when evaluating an objective metric. Another challenge in this arena is the need to reduce the complexity of some quality assessment metrics for various applied contexts. Academic research within VAWN was correctly focused on unconstrained estimates of video quality. For example, estimates are independent of any particular codec. But the application of quality metrics is often highly applied and focused on a specific codec, application, and device. We believe that the particulars of a given context may create opportunities for simplifying algorithms, and thus facilitating deployment. For instance, most content today is encoded using the H.264 video coding standard and soon will be moving to H.265. Focusing metrics for this particular codec and then making additional simplifications for specific hardware platforms (such as handheld, tablet, or laptop) should be possible. The VAWN program committee also believes that video quality metrics will eventually need to be standardized in order to give both content vendors and service providers confidence in their accuracy for predicting user perceived quality. In general, further work will be needed before these new perceptual based video quality metrics can become pervasive in the industry.

"…there is a pressing need for a large-scale, open access database of video segments…"

"Another challenge is the need to reduce the complexity of some quality assessment metrics for various applied contexts."

The need for industry standards extends beyond the sphere of quality metrics to other areas of program research as well. It was mentioned earlier that standards are critical to communications systems since single vendors rarely make every element of an end-to-end system, and standards are needed to enable interoperability. A key approach explored in the VAWN program that relies on such interoperability is cross-layer optimization. The communications industry is still largely segmented according to the standard OSI stack; a set of vendors focus on hardware (for example, a radio chip), a different set of vendors focus on the operating system software stack, and yet a different set of vendors focus on applications. This segmentation creates a significant challenge for realizing cross-layer optimizations, which require standard APIs to enable information exchange between the different layers of the OSI stack. Without such information exchange, it is difficult for contributing layers within the end-to-end system to make more informed decisions. While some of these APIs are being discussed in standards, clearly more will be needed. In addition, more work will be needed to develop algorithms and implementations that exploit new information at various layers and to manage the interaction of layers. To illustrate the latter, rate adaptation decisions are currently made at the radio layer in response to wireless channel conditions, at the transport layer in response to congestion and errors, and at the application layer in response to end-to-end throughput estimates. A cross-layer optimization approach could be used to better coordinate the whole stack.

"This segmentation creates a significant challenge for realizing cross-layer optimizations…"

A third challenge to technical transfer of research results is how to modify the equipment in today's wireless networks, which has already been widely deployed (and at great expense), yet often does not support new uses of information, cross-layer information, or information sharing. Perhaps incremental advancements within this problem sphere could be made by software-defined networking (SDN) approaches that promise increased programmability and more opportunity to orchestrate network changes across clusters of nodes. Broadly, this could enable intelligent algorithms within the network that act on new information being shared between elements of the end-to-end system (for example, video server and client devices) in new ways. As video traffic over wireless networks continues to grow, the efficient management of traffic will become increasingly important and provide incentive for better optimizing video applications in an end-to-end manner. As equipment deployments support new programmability, opportunities for application-specific and cross-layer optimizations will likewise increase.

"A third challenges is how to modify the equipment in today's wireless networks…"

Conclusions

The Video Aware Wireless Networks university research initiative was created to investigate fundamental improvements in the delivery of video data over tomorrow's wireless networks. Given the complexity of this end-to-end challenge, program researchers found it essential to have industry perspectives and feedback throughout the course of the research.

Furthermore, it was helpful that Intel, Cisco, and Verizon play different roles within the ecosystem, and thus could contribute a range of complementary viewpoints. Another observation is that the research greatly benefited from the multidisciplinary approach; it was decided early on that this kind of system level research would greatly benefit from bringing together a multidisciplinary team, including experts in wireless communications, networking, video quality, and video processing. This unique combination of industry-led, multidisciplinary university research not only led to valuable lessons for this particular problem space, but also created new connections among the universities as well as new connections across different disciplines within the universities. As the complexity of communication systems and next generation networks grows and focus moves towards optimizing end user quality of experience, we believe that industry collaboration and bringing together experts across disciplines will become increasingly important.

References

[1] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2014–2019, http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.html

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

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

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

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

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

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

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

[9] Oyman, O. and S. Singh, "Quality of experience for HTTP adaptive streaming services," *IEEE Communications Magazine.*, vol. 50, no. 4, pp. 20–27, Apr. 2012.

[10] Lark Kwon Choi, Yiting Liao, Barry O'Mahony, Jeffrey R Foerster, and Alan C. Bovik, "Extending The Validity Scope of ITU-T P.1202.2," *Workshop on Video Processing and Quality Metrics for Consumer Electronics* (VPQM 2014).

[11] Ramamurthi, V. and O. Oyman "Link Aware HTTP Adaptive Streaming for Enhanced Quality of Experience", *IEEE Globecom 2013, Atlanta, GA, Dec. 2013*.

[12] Foerster, Jeff, "Transforming Mobile Multimedia Delivery," *ACM MobiCom 2013*, Panel: The Evolution of Wireless Video - Technology and Applications, invited presenter and panelist.

[13] Ramamurthi, V. and O. Oyman, "Video-QoE Aware Radio Resource Allocation for HTTP Adaptive Streaming," *IEEE ICC 2014 - Communication QoS, Reliability and Modeling Symposium*.

[14] Zheng Lu, V. S. Somayazulu, H. Moustafa, "Context-Adaptive Cross-Layer TCP Optimization for Internet Video Streaming", *ICC 2014, Communication Software, Services, and Multimedia Applications Symposium*.

[15] Foerster, J. and M. Gong, "Industrial Column: Special Issue on Video Aware Wireless Networks," Guest editor and organizer for invited section for *IEEE Multimedia Communications Technical Committee (MMTC) E-Letters*, published in Sep. 2013.

[16] Oyman, O., U. Kumar, V. Ramamurthi, M. Rehan, and M. Morsi, "Dynamic Adaptive Streaming over HTTP: Standards and Technology," invited paper for *IEEE Multimedia Communications Technical Committee (MMTC) E-Letters*, published in Sep. 2013.

[17] Oyman, O., "Dynamic Adaptive Streaming over HTTP (DASH) Standardization at MPEG and 3GPP," invited paper for *IEEE Multimedia Communications Technical Committee (MMTC) E-Letters*, published in Sep. 2013.

[18] Lark Kwon Choi, Yiting Liao, and Alan C Bovik, "Video QoE Models for the Compute Continuum," invited paper for *IEEE Multimedia Communications Technical Committee (MMTC) E-Letters*, published in Sep. 2013.

[19] O'Mahony, Barry A, Hassnaa Moustafa, Henry Bruce, Suman Sharma, and Shantidev Mohanty, "Enriching the Content Access Experience with Context-Awareness," *SWPC 2013*.

[20] Oyman, Ozgur, "Optimizing DASH Delivery over Wireless Networks," *IEEE COMSOC MMTC E-Letter*, March 2013.

[21] Kumar, U. and O. Oyman, "QoE Evaluation for Video Streaming over eMBMS," *Journal of Communications*, 2013.

[22] Marek, D., J. Gromada, H. Moustafa, J. Forestier, "A context-aware architecture for IPTV services personalization," in *Journal of Internet Services and Information Security (JISIS)*, Vol 3. No 1. Feb 2013.

[23] Diallo, M., H. Moustafa, H. Afifi, and N. Marechal, "Context-aware QoE for audio-visual service groups," *IEEE Comsoc MMTC E-Letter*, Vol.8 No. 2, March 2013.

[24] Oyman, Ozgur, "Video Streaming Enhancements for LTE Advanced," invited talk at UCSD *ITA Workshop*, March 2013

[25] Yiting Liao, "Achieving QoE across the compute continuum: how compression, content, and devices interact," presented at *VPQM 2013*.

[26] S. Song, H. Moustafa, H. Afifi, "Modeling an NGN authentication solution and improving its performance through clustering," *PROCEEDINGS of IEEE Globecom 2012*, Dec 2012.

[27] Gupta, V. ; Somayazulu, S. ; Himayat, N. ; Verma, H. ; Bisht, M. ; Nandwani, V., "Design Challenges in Transmitting Scalable Video over Multi-Radio Networks," 8th Broadband Wireless Access Workshop, *Globecom 2012*.

[28] Foerster, Jeff, Ozgur Oyman, Yiting Liao, Mohamed Rehan, and Wafaa Taie, "Enhanced Adaptive Streaming over LTE-Advanced Wireless Networks," *IEEE Asilomar Conference on Signals, Systems, and Computers*, 2012.

[29] Foerster, Jeff, "Video over Wireless: Disrupting Media Panel," Texas Wireless Summit, presentation and panelist 2012.

[30] Yiting Liao and Audrey Younkin, "Understanding and Adapting Video Streaming Quality of Experience for the Compute Continuum," presentation at IDF 2012, San Francisco.

[31] Wang, D., V. S. Somayazulu, J. R. Foerster, "Efficient Cross-layer Resource Allocation for H.264/SVC Video Transmission over Downlink of an LTE System," *IEEE WoWMoM Workshop on Video Everywhere*, June 2012.

[32] Oyman, Ozgur, "Optimizing HTTP Adaptive Streaming for Enhanced Service Capacity and QoE," invited talk at Cisco Adaptive Media Transport Workshop.

[33] Foerster, Jeff, "Optimizing Media Delivery in the Future Mobile Cloud," *ICC 2012* workshop on Realizing Advanced Video Optimized Wireless Networks.

[34] C-C Chiu, S.-Y. Chien, C.-H. Lee, V. S. Somayazulu, and Y.-K. Chen, "Hybrid Distributed Video Coding With Frame Level Coding Mode Selection," *IEEE International Conference on Image Processing*, Oct. 2012.

[35] Oyman, Ozgur, "Video Optimization R&D for LTE Advanced," presentation at USU-MIT-Verizon-Intel Wireless Technology Workshop.

[36] Oyman, O. and S. Singh, "On Capacity-Quality Tradeoffs in HTTP Adaptive Streaming over LTE Networks," UCSD *ITA Workshop 2012*.

[37] Somayazulu, S., Shao-Yi Chien, P-K Tsung, Y-K Chen et al., "Power Optimization of Wireless Video Sensor Nodes in M2M Networks," *17th Asia-South Pacific Design Automation Conference*, Jan. 2012.

[38] Singh, S., O. Oyman, A. Papathanassiou, D. Chatterjee and J. G. Andrews, "Video Capacity and QoE enhancements over LTE," *ICC 2012*.

[39] Foerster, Jeff, "Optimizing Media Delivery in the Mobile Cloud," PCCA workshop on The Mobile Cloud, 2011.

[40] Oyman, Ozgur, "Adaptive Video Streaming over Wireless Networks," lecture session at IDF 2011 in San Francisco, CA.

Author Biographies

Dr. Jeff Foerster (jeffrey.r.foerster@intel.com) is currently a principal engineer in the Wireless Communications Research Lab in Intel Labs. Jeff was one of the principal organizers of the Video Aware Wireless Networks (VAWN) university research program and currently manages a team focused on future emerging wireless technologies. Prior to joining Intel, he worked on Broadband Wireless Access (BWA) systems and standards (IEEE 802.16). He received his BS, MS, and PhD from the University of California, San Diego, where his thesis focused on adaptive interference suppression and coding techniques for CDMA systems. Jeff is an IEEE Fellow.

David Ott (david.e.ott@intel.com) is a research director for Intel's University Research Office. His role is to identify opportunities for innovative technology development in the areas of computer security and communications. These form the basis for collaborative university research programs that bring together top academic researchers worldwide to explore new approaches. David Ott joined Intel in 2005 and has worked in a variety of technical roles over the

years focusing on enterprise computing, software aspects of Intel platforms, performance analysis, and computer security. David holds an MS and a PhD in Computer Science from the University of North Carolina at Chapel Hill.

Ozgur Oyman (ozgur.oyman@intel.com) is currently a senior systems architect at Intel's Mobile and Communications Group, and is in charge of video over LTE research and standardization. He has been with Intel since 2005. He serves as the prime delegate for Intel in the 3GPP SA4 working group that specializes in mobile multimedia services, related codecs, protocol stacks, and file formats. He's held several editorship and rapporteurship roles for 3GPP SA4 and MPEG DASH standardization, and has also been the principal Intel representative at the DASH Industry Forum. He holds a PhD and MS from Stanford University and a BS from Cornell University (all in electrical engineering).

Yiting Liao (yiting.liao@intel.com) is a research scientist in Wireless Communications Lab at Intel Labs in Hillsboro, Oregon, focusing on multimedia communication technologies and wireless body area networks. She received her PhD in Electrical and Computer Engineering from the University of California, Santa Barbara and MS and BS in Electrical Engineering from Tsinghua University, China. She has published 20 journal/conference papers and book chapters, and holds over 10 pending patents. Her research interests include image and video quality assessment, video optimization techniques and QoE enhancement over wireless networks, machine learning and protocol design for wearables and body area networks.

V. Srinivasa Somayazulu (Zulu) (v.srinivasa.somayazulu@intel.com) is a senior research scientist with the Wireless Communication Research Lab within Intel Labs. His research interests are in the areas of multimedia communications over wireless networks, low-power wireless body area and personal area networks (WBAN/WPAN) and sensor communications, as well as platform architecture optimizations for energy efficiency and QoE improvements for multimedia communications. He has previously contributed to the development of cross-layer optimized solutions for real-time video conferencing, wireless display solutions over Wi-Fi and WiGiG interfaces, Wimedia (UWB) PHY/MAC specifications and UWB spectrum regulations.

Xiaoqing Zhu (xiaoqzhu@cisco.com) is a technical leader in the Chief Technology and Architecture Office (CTAO) at Cisco Systems Inc. She received the B.Eng. degree in Electronics Engineering from Tsinghua University, Beijing, China. She earned both an MS and a PhD in Electrical Engineering from Stanford University. Prior to joining Cisco, Dr. Zhu interned at IBM Almaden Research Center in 2003, and at Sharp Labs of America in 2006. Her research interests span multimedia applications, networking, and wireless communications. She received the best student paper award at ACM Multimedia 2007. She also won the best presentation award at the IEEE Packet Video Workshop in 2013.

Douglas S. Chan (douglas.chan@ieee.org) is a Visiting Research Scholar in Electrical Engineering and Computer Science at the University of California, Berkeley. His work is affiliated with the Swarm Lab and Berkeley Wireless Research Center. Prior to that (2006–2014), Doug was a Technical Leader at Cisco Systems where he worked in R&D on wireless LAN products and their standardizations. Later at Cisco, Doug was a member of Cisco Advanced Architecture and Research (Enterprise Networking Labs) where he investigated emerging paradigms like fog computing, data analytics, and the Internet of Things. Doug received his M.Eng. and PhD in Electrical Engineering from Cornell University. Doug is a Senior Member of IEEE and has received recognitions from the IEEE Standards Association for his contributions to the 802.11n and 802.11y wireless LAN standards.

Chris Neisinger (chris.neisinger@verizonwireless.com) is the executive director in the CTO Organization at Verizon. In this role, Mr Neisinger and his organization are responsible for the technical evolution of the wireline and wireless networks. This includes evaluation and development of new technologies and architectures. He has led Verizon's efforts in 3G and 4G wireless architecture development, IMS and VoLTE, cloud, analytics, and M2M. His current focus on emerging technologies including SDN and NFV and how to use these new architectures to create new services and business opportunities. Chris has held various engineering and operational roles during his 20 years in the wireless industry. Chris' previous industry experience includes leadership positions in joint ventures in Hungary and Malaysia. Chris holds a Bachelor of Science degree in Electrical Engineering from Washington State University.

PERCEPTUAL OPTIMIZATION OF LARGE SCALE WIRELESS VIDEO NETWORKS

Contributors

Robert W. Heath Jr.
The University of Texas at Austin

Alan C. Bovik
The University of Texas at Austin

Gustavo de Veciana
The University of Texas at Austin

Constantine Caramanis
The University of Texas at Austin

Jeffrey Andrews
The University of Texas at Austin

Chao Chen
The University of Texas at Austin

Michele Saad
Intel Corporation

Zheng Lu
The University of Texas at Austin

Amin Abdel Khalek
Freescale Semiconductor Inc.

Sarabjot Singh
The University of Texas at Austin

One of the greatest challenges facing wireless networks is the support of ubiquitous high quality video streaming. The problems begin with even defining what high quality means, since popular quantitative metrics are often misleading. Transmission of video differs from other types of data in that video data is delay-sensitive: every video packet must be delivered by a certain deadline to support continuous playback. Although real-time video is most sensitive to rate fluctuations and delays, stored video streaming is also challenging since it requires a long period of relatively consistent network performance, despite many diverse factors that make wireless network performance highly variable. The goal of this article is to provide an overview of the key challenges associated with wireless video transmission and to survey recent progress made by the authors in their collaborative research with Intel and Cisco in the Video Aware Wireless Networks program.

Introduction

The next generation of video networks will deliver unicast and multicast video content to mobile users, leveraging rapidly expanding wireless networks. To achieve the highest performance, wireless video transmission systems should be designed to optimize the video quality perceived by the end user. This task is challenging in several ways, which we classify in four overarching categories:

1. *Perceptual video quality is subjective and hard to quantify.* It is well known that the peak signal-to-noise ratio (PSNR) does not correlate well with perceptual video quality as measured by large, human subjective studies.[1] On the LIVE video database[2], the linear correlation between PSNR and the perceptual quality is only 0.40.[3] In recent years, several high-accuracy *image* quality assessment models such as SSIM[4] and VIF[5] have been proposed. These operate by extracting perceptually relevant features from images and then using them to form predictions of perceptual quality. The experience of watching a video, however, is very different from that of viewing still images (pictures). Image quality predictors are not entirely adequate for assessing the perceptual quality of videos. For example, the high performance image quality predictor MS-SSIM only achieves a 0.72 correlation against human scores on the LIVE database.[2][3][4] Moreover, nearly all existing video quality predictors are full reference. In a practical video transmission system, reference videos are generally not available.

2. *Wireless networks have highly variable throughput.* Most video services require playback concurrent with transmission. If the end-to-end throughput falls below the video source rate in a specified time interval, then the amount of video that is buffered at the receiver is less than the amount of video that is played out. These throughput variations can be caused by a myriad of factors including mobility (leading to different path loss, shadowing, or fading), bursty interference from other nearby users, or congestion at the serving access point or base station. Once all the video data buffered at the receiver has been exhausted, the playback process stalls, which has a negative effect on the viewer's quality of experience (QoE).[6] Similarly, in real-time streaming, the video reception simply experiences an outage, which is equally unacceptable to QoE. Thus, the video source data rate needs to be somehow adapted according to the time-varying throughput. Although there exist adaptive video streaming protocols such as MPEG-DASH[7], which allow a video user to switch the video data rate every several seconds, finding an adaptation policy that optimizes QoE is complicated, since lowering the source rate also adversely affects QoE, and the current buffer state also should be factored in. All in all, adaptation of video streaming to a highly stochastic wireless network is a major challenge.

"Wireless networks have highly variable throughput."

3. *There are delay constraints, particularly for real-time video.* Supplying real-time video service introduces additional difficulties in algorithmic design. As compared with stored video, real-time video applications, such as video telephony, necessitate tighter delay constraints. Channel variations over fine time scales may also cause playback interruptions. Hence, a video transmission scheme operating in this environment must exploit adaptation mechanisms at the lower layers. For example, at the media access control (MAC) layer, transmission resource allocation amongst users can be adapted over intervals of tens of milliseconds. At the physical (PHY) layer, modulation and coding schemes can be adapted to minimize the impact of video packet errors on perceptual quality. Since the video packets in a video stream are predictively encoded, the impact of fine time-scale adaptations on perceptual video quality is difficult to characterize, which makes the design of fine time-scale adaptation schemes difficult. Moreover, unlike stored video streaming scenarios where video content is stored at the server before transmission, real-time video is generated at the time of transmission. Thus the video content to be transmitted in the future is not known. Such uncertainty further increases the difficulty of optimizing video transmission algorithms.
4. *Networks are becoming increasingly more dense and complex.* The application-agnostic paradigm of current data networks is not able to leverage the spatio-temporal bursty nature of video. Since video streams occupy much larger bandwidths than "ordinary" data applications, even moderate variations in the video user population being served may potentially cause huge fluctuations in video traffic. This may cause the traffic loads in

"Networks are becoming increasingly more dense and complex."

different wireless networks to become highly unbalanced, thereby reducing the utilization efficiency of the existing capacity. To address this problem, a coordination algorithm can be introduced that seeks to dynamically balance the load of video traffic across heterogeneous wireless networks. Devising a good load-balancing strategy, however, is challenging due to the following: 1) the complex geometric distribution of wireless nodes makes the load distribution difficult to model and the design of optimal load balancing policies intractable, and 2) the broadcast nature of the wireless medium causes interference between wireless networks. This interference depends on the load distribution and affects the data throughput to video users. The impact of traffic load on interference needs to be accurately modeled and properly considered in the design of effective load-balancing algorithms.

"This article 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."

This article 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 four interconnected research vectors, corresponding to the above four challenges. They are briefly summarized as:

1. *Video quality modeling and prediction.* This included the development of new models and algorithms for full-reference[8], reduced-reference[9], and no-reference prediction[10][11][12] that achieve high correlation against human judgments of quality recorded in large-scale subjective experiments. A dynamic system model was also created that captures the QoE on long videos.[13][14] These models have been used to drive adaptation algorithms developed as part of the other research vectors on spatio-temporal network adaptation.
2. *Rate-adaptive transmission of stored video.* This focused on the development of stored video streaming techniques that not only maximize the delivered QoE, but also reduce risk of receiver buffer starvation.[15][16][17][18][19]
3. *Cross-layer design for real-time video transmission.* Work here led to the realization of a series of cross-layer designs that optimize video perceptual quality for both single-user and multiuser real-time video transmission.[20][21][22][23][24][25][26][27][28][29][30]
4. *Load management in heterogeneous wireless networks.* This research leveraged stochastic geometric models to develop tractable load-balancing strategies capable of greatly enhancing the video capacity of heterogeneous networks.[31][32][33]

Progress made on each of these vectors is described in subsequent sections of this article.

Video Quality Assessment

The following section describes three types of video quality assessment. The techniques operate on the observed possibly distorted video sequence and vary based on what is known (if anything) about the original video source.

Full Reference Video Quality Assessment

All the aspects of the work developed herein benefit by the development of objective perceptual-theory–based algorithms that can evaluate the integrity of generic video signals delivered to human end users. The MOVIE index[8] is a full-reference video quality assessment algorithm that is able to evaluate the quality of a video sequence, relative to an undistorted (presumably pristine) reference version of the same video. While prior full-reference video quality models have made little direct use of computed motion information, thus limiting their effectiveness, the MOVIE framework captures separate spatial and temporal perceptual primitives that are used to conduct video quality evaluation.

"The MOVIE index is a full-reference video quality assessment algorithm…"

The first stage of the framework is a Gabor filter decomposition of the test and reference videos by a bank of spatio-temporal bandpass channels. This linear decomposition mimics the response behavior of simple cells in primary visual cortex area V1, which are highly selective to spatial and temporal frequency and orientation, and which are well-modeled as space-time localized linear filters. The behavior of neurons in extracortical visual area MT, which is fed by space-time responses from area V1 and which is believed to handle a large portion of perceptual motion processing, is also characterized by directional sensitivity but in a different way.

Following the initial stage of spatio-temporal linear decomposition of a video, a spatial quality measurement (similar in nature to the SSIM index) is obtained from the Gabor filter responses of the test and reference videos. This defines the Spatial MOVIE index. Temporal quality evaluation begins by computing motion information in the form of optical flow vectors from the reference video and used for temporal evaluation. The Gabor responses along with optical flow are then used to define a Temporal MOVIE index, which is highly sensitive to the distortions that are temporal in nature. This stage of processing deploys a model of processing in visual brain area MT to penalize errors in locally computed motion vectors. The overall MOVIE index is then defined as the product of the spatial and temporal MOVIE indices reflecting the approximately separable space-time processing of visual data in the brain. A diagram that summarizes the MOVIE framework is shown in Figure 1.

Reduced Reference Video Quality Assessment

Video RRED[9] is a reduced-reference approach to video quality assessment, where one does not require an entire reference video to be available to evaluate the quality of a test video. Instead, only partial information from a reference video is assumed. Video RRED is a family of algorithms that scales video quality prediction accuracy with the amount of information that is received from the reference video. The amount of information taken from the reference can range from a single number per frame to the entire reference video.

"Video RRED is a family of algorithms that scales video quality prediction accuracy with the amount of information that is received from the reference video."

Video RRED relies on a statistical model of wavelet coefficients, obtained via a steerable pyramid decomposition of frames and frame differences to perform quality evaluation. A Gaussian scale mixture model is used to describe the wavelet coefficients of frames and frame differences, and this model is used to

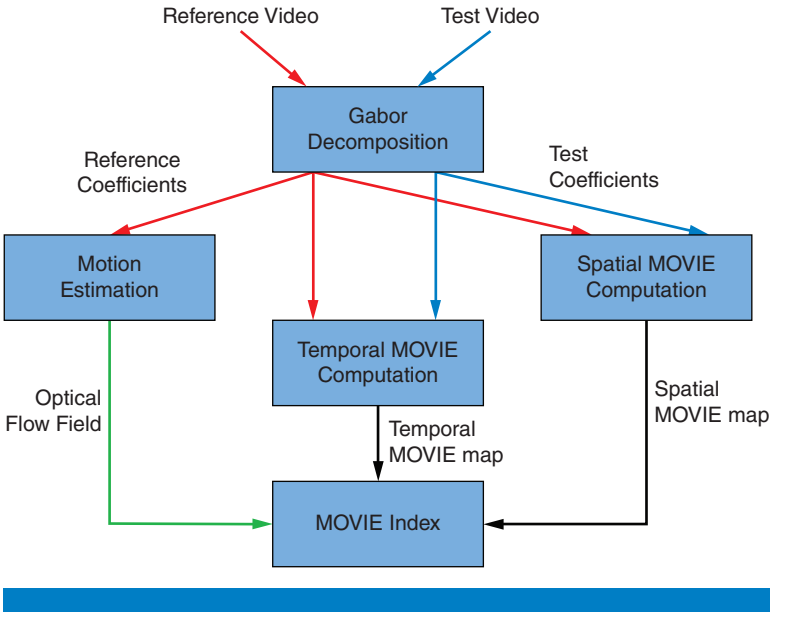


Figure 1: Diagram from Seshadrinathan and Bovik[8] summarizing the MOVIE index framework. The first stage is a linear Gabor decomposition of the reference and test videos. Temporal and spatial quality evaluation are then performed using the Gabor response coefficients. The temporal evaluation also utilizes optical flow measurements from the reference video and a neural model of motion error sensitivity. The spatial and temporal MOVIE indices are then combined into a final quality score (Source: K. Seshadrinathan and A. C. Bovik, 2010)

compute conditional entropic differences between reference and test videos. Again, two separable subindices are defined: a spatial video RRED index obtained by computing entropic differences between reference and test videos on a frame-by-frame basis, and a temporal video RRED index obtained by calculating entropic differences from consecutive frame differences. A final video RRED score is obtained by multiplying the spatial and temporal indices.

"Video BLIINDS is a blind/no-reference video quality evaluation algorithm…"

No Reference Video Quality Assessment

1) Video BLIINDS: Video BLIINDS[12] is a blind/no-reference video quality evaluation algorithm that requires no reference video nor any information from a reference to predict the quality of a test video. While most no-reference image and video quality assessment algorithms require some knowledge about the type of distortion afflicting an image or video, Video BLIINDS is a generic, non-distortion-specific algorithm that does not require any previous distortion knowledge. The approach relies on a spatio-temporal model of video scenes in the discrete cosine transform (DCT) domain, and on a model that characterizes the type of motion occurring in the scenes, to predict video quality. The Video BLIINDS approach relies on a natural video statistics (NVS) model of local DCT coefficients obtained from consecutive frame differences. It was shown that the statistics of spatially local DCT coefficients change systematically with the level of distortion perceived in a video. A generalized Gaussian family of

distributions was found to be an appropriate model fit to the DCT coefficients of differenced-frames. The parameters of the model are then used as indicators of video quality. Further, a model of video motion content, based on a two dimensional structure tensor of a video's motion vectors, is defined and used in conjunction with the NVS features to train a learning machine that predicts human quality scores. The model of motion content is motivated by the experiments of Suchow and Alvarez[34], which suggest that large coherent motion silences/masks transient temporal change, such as flicker, which is a common description of many temporal video distortions. Consequently, a 2D structure tensor of the motion vectors computed from the test video are used to determine the degree of local coherency of motion present in the video. Figure 2 summarizes the flow of the Video BLIINDS algorithm.

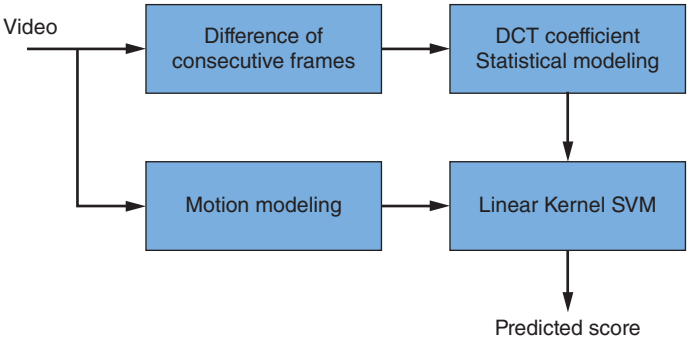


Figure 2: Diagram summarizing the Video BLIINDS framework in Saad and Bovik[12]
(Source: The University of Texas at Austin, 2014)

2) NIQE: While no-reference image and video quality assessment algorithms require that a model be first trained on a data corpus from which a mapping to quality scores is learned, we envision that future developments in quality assessment will do away with the need for training completely. Training implies important shortcomings in a video quality model, including dependency on the spatial sizes of the videos in the training data corpus and the number of features extracted to represent an image and video. Ideally, the data corpus would need to be comprehensive in regards to both video content and in possible distortions that might occur. Towards the goal of eliminating training, a new "completely blind" image quality model called NIQE[11] was designed, which simply computes a distance between a feature vector extracted from a test image relative to a set of feature vectors extracted from a predefined set of pristine images. The larger this distance, the higher the perceived level of distortion present in the image. While no training is required for NIQE to predict a quality score, a set of predefined pristine images (along with their extracted features) is used as a global pristine reference relative to which a distance is computed. The features extracted by NIQE are based upon the classical natural scene statistics (NSS) model provided by Ruderman[35], which is a model of spatial domain pixel intensities transformed by a processes of local mean removal and divisive normalization.

"Towards the goal of eliminating training, a new "completely blind" image quality model called NIQE was designed…"

PSNR	SSIM	VQM	MOVIE	Video RRED	V-BLIINDS
0.671	0.650	0.745	0.807	0.826	0.750

Table 1: Full-reference and reduced-reference median SROCC correlations on every possible combination of train/test set splits (subjective DMOS vs. predicted DMOS) on the LIVE VQA database. Eighty percent of the video content used for training. The training and test content were always kept separate
(Source: The University of Texas at Austin, 2014)

Table 1 summarizes the results of MOVIE, Video RRED, and Video BLIINDS along with widely used PSNR measure, the SSIM index (a popular full-reference IQA index)[36], and VQM (a standardized reduced-reference VQA index).[37] The results are reported in terms of the Spearman rank order correlation coefficient between the predicted scores and the subjective ratings on the LIVE VQA database.[38] Since Video BLIINDS is trained on a portion of the LIVE VQA database and tested on the remaining videos, we report results for all the algorithms on the same portions of the database used for testing.

The LIVE VQA database contains 10 uncompressed reference videos of natural scenes and 15 distorted versions of each of the 10 reference videos. The distortions present in the database are simulated MPEG-2 and H.264 compression distortions, and simulated error-prone IP and wireless network distortions. Each video was viewed and assessed by 38 subjects in a single stimulus study with hidden reference removal and scored on a continuous quality scale.

Modeling Time-Varying Subjective Quality

Newly developed HTTP-based video streaming technologies enable flexible rate adaptation under varying channel conditions. Accurately predicting the users' quality of experience (QoE) on rate-adaptive HTTP video streams is thus critical for rate adaptation. An important aspect of understanding and modeling QoE is predicting the time-varying subjective quality (TVSQ), that is, the up-to-the-moment subjective quality of a video as it is played.

"An important aspect of understanding and modeling QoE is predicting the time-varying subjective quality (TVSQ), that is, the up-to-the-moment subjective quality of a video as it is played."

We propose to predict TVSQ in two steps.[13][14] The two steps capture the spatio-temporal characteristics of the video and the hysteresis effects in human behavioral responses, respectively. In the first step, quality-varying videos are partitioned into one-second-long video chunks and the perceptual quality of each chunk is predicted using the method described by Soundararajan and Bovik.[9] In the second step, a Hammerstein-Wiener (HW) model is used to predict the TVSQ . The model is illustrated in Figure 3. The core of the HW model is a linear filter, which is intended to capture the memory of the perceptual quality curve. At the input and output of the HW system, two nonlinear static functions are employed to model potential nonlinearities in the human responses.

To collect data for model parameterization and validation, a database of rate-varying video sequences was built to simulate quality fluctuations commonly encountered in video streaming applications. Then, a subjective study was

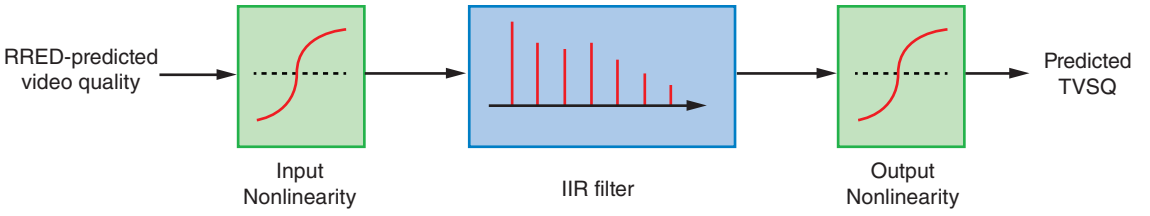


Figure 3: Proposed Hammerstein-Wiener model for TVSQ prediction[13]
(Source: Chen et al., 2013. Used with permission of the IEEE)

conducted to measure the TVSQs of these video sequences. Specifically, the videos in the database were each viewed and scored by 25 subjects. Video sequences were displayed to the viewers on an HDTV monitor. During the play of each video, a continuous scale sliding bar was displayed near the bottom of the screen. The sliding bar was marked with five labels: *Bad*, *Poor*, *Fair*, *Good*, and *Excellent*. The subject could continuously move the bar via a mouse to express his/her judgment of the video quality as each video was played. The position of the bar was sampled and recorded automatically in real time as each frame was displayed (30 fps). Difference mean opinion scores of the TVSQ were then derived from the collected samples from viewers. This database is useful for developing and validating TVSQ models and thus is important in its own right, since it may contribute to future research efforts.

Using the new database, a dynamic system model was proposed to predict the average TVSQ per second of a given video. Experimental results showed that the proposed model reliably tracks the TVSQ of video sequences that exhibit time-varying qualities (see Figure 4). The dynamic system model has a simple,

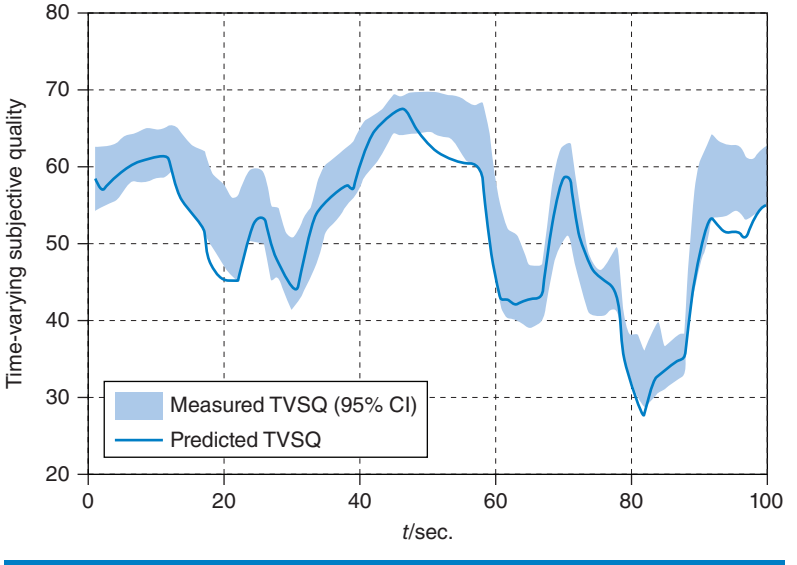


Figure 4: The performance of the dynamic system model for TVSQ prediction[14]
(Source: Chen et al. 2014. Used with permission from the IEEE)

easily realized structure and is computationally efficient. As such, it is quite suitable for online TVSQ-optimized rate adaptation.

In our subjective study, we also asked the subjects to feed back their subjective judgment of the overall quality of the entire video. We found that the average qualities of these videos only have limited correlation with their subjective qualities. Actually, the worst parts of a video tend to dominate the overall quality of an entire video. Indeed, as shown by Chen et al.[18], the marginal distribution and the variation of the varying video qualities can be used to achieve a good accuracy in overall quality prediction. A summary of the lessons learned is shown in Table 2.

	Lesson Learned
Video quality assessment	Natural scene and video statistics-based models enable accurate quality prediction in the absence of explicit reference information.
Time-varying subjective quality modeling	Time-varying subjective quality can be predicted using simple dynamic system
Subjective study on rate-adaptive videos	The worst parts of a video tend to dominate the overall quality of an entire video

Table 2: Summary of lessons learned in video quality assessment (Source: The University of Texas at Austin, 2014)

Stored Video Transmission

"Stored video streaming is the workhorse underlying much of the video data being delivered over today's networks."

Stored video streaming is the workhorse underlying much of the video data being delivered over today's networks. In this section we discuss four key themes. We first consider QoE-optimized rate adaptation. When videos are stored and available in several representations corresponding to different perceptual quality levels at the expense of different amounts of data (rate), a client may adapt its requested representation to match the available resources. We discuss a holistic approach to rate adaptation that accounts for multiple facets of user's QoE and that fits well into state-of-the-art protocol frameworks such as DASH. Maintaining a consistent QoE for video streaming in a network with variability in resources as well as in load requires some form of admission control. In our second theme we introduce a measurement-based approach to admission control aimed at maintaining a consistent user experience. While rate adaptation combined with measurement-based admission control may allow one to mitigate the effects of uncertainty and variability in capacity wireless users will see, this remains a major challenge. An alternative approach is to attempt to get knowledge of the future variations they will see and design a network around such prediction. Since users' mobility is fairly predictable, there is an increasing availability of coverage and throughput maps capturing the performance users are likely to see. The third theme we discuss presents some preliminary ideas in this direction. In our fourth theme, we present an overall system-level performance evaluation of QoE enhancements that improved streaming and resource allocation

algorithms would yield for an LTE-based wireless system. This provides insights into potential achievements with current wireless cellular technology.

A cross-layer overview of our stored video transmission approaches discussed in the sequel is shown in Figure 5. We use text boxes to denote key functional components of each approach organized by network layers and components (that is, server, Base Station (BS), and client). We also use arrows to denote information exchanges within a single layer and across layers. Note that our approaches mainly involve three layers: APP, MAC, and PHY. The APP layer provides video-specific context such as quality-rate tradeoff and playback buffer status, and performs video rate/quality adaptation based on cross-layer information. The PHY layer provides information on available wireless capacity. The MAC layer performs resource allocation based on cross-layer information. More details of the approaches can be found in each subsection below.

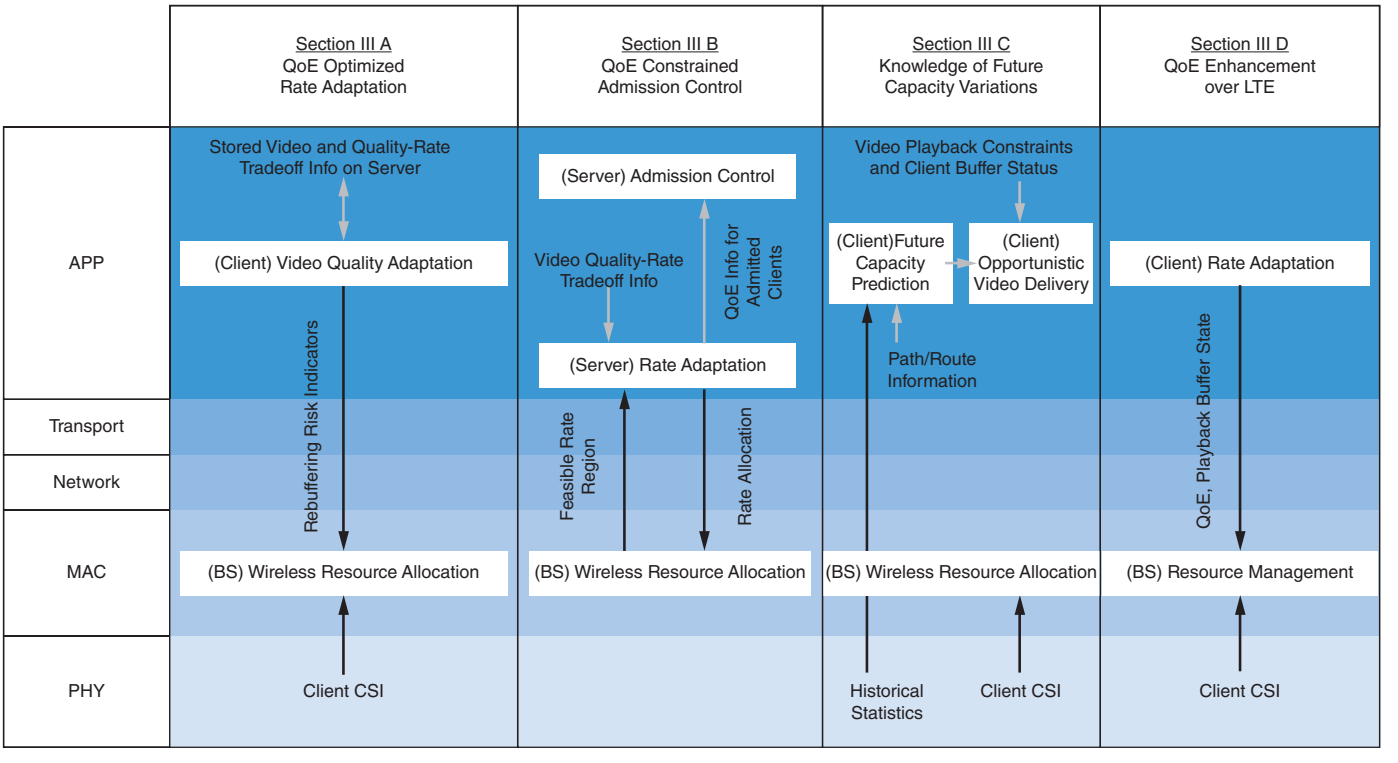


Figure 5: Cross-layer overview for our stored video transmission approaches. Text with a box represents a functional module associated with a video transmission approach. Text with no box represents information being exchanged. A gray arrow indicates information exchanges within a single layer. A black arrow indicates information exchanges across layers
(Source: The University of Texas at Austin, 2014)

QoE-Optimized Rate Adaptation: Mean-Variance-Rebuffering Tradeoff

We view the stored video transmission optimization problem for a network as that of fairly maximizing the video clients' quality of experience (QoE) subject to network constraints. There are four key factors determining the QoE of a

video client: (1) average video quality, (2) temporal variability in video quality, (3) time spent rebuffering (including startup delay), and (4) cost to the video client and video content provider. So technically, we focus on solving an optimization problem sketched below:

$$max \sum_{i \in N} U_i^E(Mean\ Quality_i - Quality\ Variability_i)$$

subject to $Rebuffering_i$, $Cost_i$, and Network constraints, where N is the set of video clients supported by the network and U_i^E is a "nice" concave function chosen in accordance with the fairness desired in the network.

"A comprehensive solution to this problem requires two components: (1) network resource allocation, and (2) rate and thus quality adaptation."

A comprehensive solution to this problem requires two components: (1) network resource allocation, and (2) rate and thus quality adaptation. The allocation component decides how network resources are allocated across video clients. The adaptation component decides how the video clients adapt their video quality in response to the allocated resources, the characteristics of the video being streamed, possibly users' QoE preferences, and so forth.

While there certainly is a substantial amount of work in this area, we proposed[15] a preliminary new approach to this problem, where the QoE model only addressed average video quality and quality variation, and the model involved a strong assumption of synchrony, that is, the downloads of segment for each video client start at the beginning of a (network) slot and finish by the end of the slot. We first derived the optimal solution of an offline convex optimization, in which we assumed all time-varying quantities were known. Then, we proposed an asymptotically optimal online algorithm, AVQ, which required almost no statistical information about the system. Finally, we modified AVQ to obtain a practical low complexity algorithm PAVQ.

A more comprehensive solution was proposed by Joseph and de Veciana[16], where we addressed all four QoE factors discussed earlier in our QoE model and presented a general optimization framework for stored video delivery optimization that factored heterogeneity in client preferences and QoE models, as well as capacity and video content variability. We developed a simple online algorithm NOVA (Network Optimization for Video Adaptation) to solve the multiuser joint resource allocation and quality adaptation problem. NOVA is an asynchronous algorithm in the sense that all video clients operate "at their own pace," which is better suited for DASH-like protocols than the previous synchronous solution. Moreover, NOVA carries out "cross-layer" joint optimization of resource allocation and quality adaptation in a distributed manner, where the network controller carries out resource allocation and video clients carry out their own quality adaptation. The cross-layer nature of NOVA is exhibited in Figure 5. We analyzed NOVA rigorously and validated its performance through extensive simulations.

"…NOVA carries out "cross-layer" joint optimization of resource allocation and quality adaptation in a distributed manner…"

Figure 6 exhibits the achieved QoE versus the number of users sharing a wireless base station. Our NOVA scheme can be compared to several

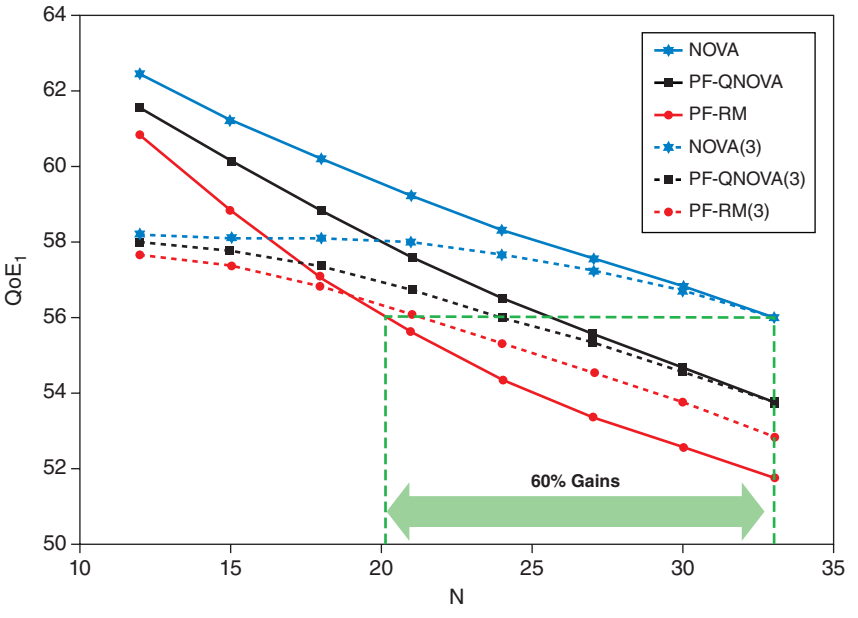


Figure 6: QoE gains from NOVA
(Source: The University of Texas at Austin, 2013)

variations and alternative algorithms. In particular PF-RM denotes a scheme that allocates wireless network resources using a proportionally fair (PF) scheduler and adapts the video rate to achieve rate matching (RM), that is, match the rate of the segment to the estimated available capacity. These preliminary results show that NOVA can support up to 60 percent more video clients given a requirement on average QoE. Figure 6 also exhibits the results achieved when only the quality adaptation mechanism of NOVA (QNOVA) is combined with PF resource allocation—see Joseph and de Veciana[16] for more results.

QoE-Constrained Admission Control and Rate Adaptation

In this section, we study the problem of maximizing the number of users in a wireless network that satisfy given constraints on QoE. We propose to characterize and predict the users' QoE using the second-order empirical cumulative distribution function (2nd-order eCDF) of the delivered video quality defined as

$$F^{(2)}(x; q) = \frac{1}{T} \sum_{t=1}^{T} \max \{x - q(t), 0\}, \quad (1)$$

where $q(t)$ represents the predicted quality[39] of the tth second of the video and T is the video length. Note that $\max \{x - q(t), 0\}$ is greater than zero if and only if $q(t)$ is less than x, so the 2nd-order eCDF captures for how long and by how much the predicted video quality falls below x. If we interpret x as the threshold below which users judge the video quality to be unacceptable, then the 2nd-order eCDF reflects the impact

of the unacceptable periods on the QoE. Since it has been recognized that the worst parts of a video tend to dominate the overall quality of an entire video[40][41][42][43][44], the 2nd-order eCDF can be used to predict the QoE. We reported[13][14] a large-scale subjective study where we found that the 2nd-order eCDF of video quality over time achieves the strong linear correlation (0.84) with the measured subjective quality of long video sequences. In comparison, the average video quality only achieves a correlation of 0.57. This lends strong support for eCDF as a good proxy for video QoE.

"We design adaptive video streaming algorithms that incorporate QoE constraints on the 2nd-order eCDFs of the video qualities seen by users."

We design adaptive video streaming algorithms that incorporate QoE constraints on the 2nd-order eCDFs of the video qualities seen by users.[18] In particular, we consider a wireless network in which a base station transmits videos to multiple users. The user population is dynamic: users arrive and depart from the network at random times. When a new user joins the network, the base station starts streaming a video to the user. A rate adaptation algorithm is employed to control the video data rate of all active video streams according to varying wireless channel conditions.

When the base station is shared by too many users simultaneously, the QoE served to each user can be poor. Instead of serving every user with poor QoE, it is preferable to satisfy the QoE constraints of existing users by selectively blocking newcomers. Therefore, in addition to rate adaptation algorithms, we introduce a new admission control strategy that is designed to maximize the number of video users satisfying the QoE constraints on their 2nd-order eCDFs (see Figure 7).[18] Although the admission control strategy damages the QoE of the blocked users, the overall percentage of users satisfying the QoE constraints among both admitted and blocked users can be significantly improved (see Figure 8 for example). Note that this approach involves cross-layer information sharing (see Figure 5).

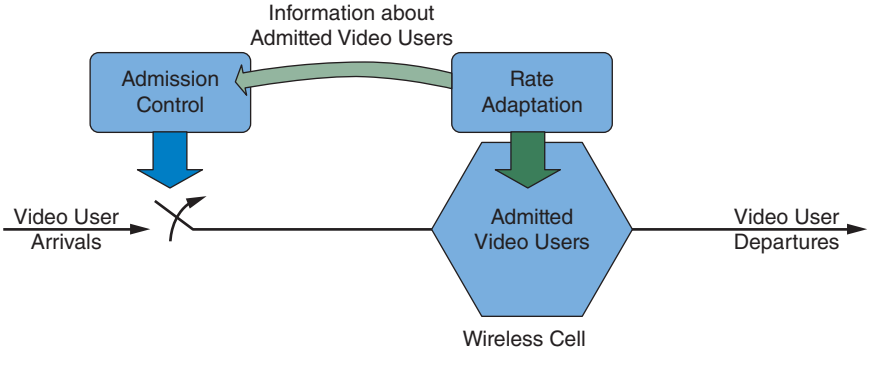


Figure 7: The proposed QoE-constrained video streaming system: admission control only affects newly arrived video users and the rate adaptation algorithm does its best to satisfy the QoE constraints of all admitted video users[18] (Source: Chen et al., 2013.[18] Used with permission from the IEEE)

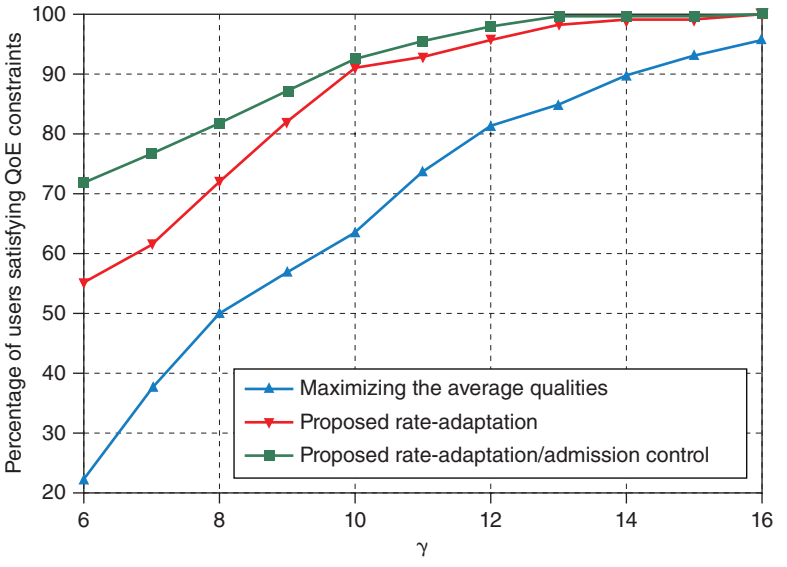


Figure 8: Simulation results of the proposed algorithms under different channel scaling parameter γ. Each data point is obtained by simulating 2000 video user arrivals. The y-axis shows the percentage of users satisfying the QoE constraints on 2nd-order eCDFs amongst both admitted and blocked users. The proposed rate adaptation algorithm and admission control algorithm is compared with the rate-adaptation algorithm, which maximizes the average quality of all arrived users (Source: The University of Texas at Austin, 2014)

Adaptive Video Transmission with Future Channel Knowledge

The increase in the wireless networks' overall capacity comes with higher degrees of capacity variability. Capacity variations are usually considered harmful to video transport (for example, they result in more rebuffering and quality jitters). We propose an approach[19] to cope with capacity variations that can even make them beneficial by exploiting the storage/buffer on mobile devices and knowledge of future capacity variations.

"We propose an approach to cope with capacity variations that can even make them beneficial…"

The key requirement is that the mobile user can predict the future capacity variations it will see. We posit that this is possible when users' future locations are known, and such knowledge can in turn be used to infer their future wireless coverage/capacity. For example, this is the case for users on public transportation buses/trains, or others using navigation systems in their cars.

Given the streaming video delivery requirements and knowledge of the future capacity variations, we propose a new class of cross-layer transmission policies that minimize system utilization while avoiding, if possible, rebuffering delays. In Lu and de Veciana[19] we study three cases. For the single-user anticipative case where all future capacity variations are known beforehand, we establish that the optimal transmission schedule is a Generalized Piecewise Constant Thresholding (GPCT) scheme. For the single-user case that is only partially anticipative, that is, where only a finite window of future capacity variations

is known, we propose an online Greedy Fixed Horizon Control (GFHC). Finally, we considered the general multiuser setting where we can exploit both future temporal and multiuser diversity. This approach involves exchanges of fine and coarse cross-layer information (such as predicted capacity); see Figure 5 for a summary.

Figure 9 shows some representative experimental results, comparing the system utilization and percent rebuffering time achieved by multiuser schemes exploiting knowledge of future capacity (MTP and MTO) to a baseline scheme with proportional rate allocation. MTP is an allocation based on multiuser thresholding under proportionally fair capacity, which allocates future capacity in a proportionally fair manner and has users perform the optimal GPCT scheme. MTO is an allocation based on multiuser thresholding under opportunistic capacity, which allocates future capacity in an opportunistic manner that ensures fairness via a token-based scheme. The higher the number of tokens, the more opportunistic the allocations are; but this is possibly the more unfair and thus results in an increased variability in the capacity allocated to users. The results show our schemes achieve up to 70-percent reduction of system utilization, or alternatively, twice or three times the number of users that can be supported under the same system utilization requirement. In terms of average percentage of rebuffering time, MTP always has the same rebuffering time as the baseline scheme. MTO with a higher token limit (see Lu and de Veciana[19] for details) achieves a lower system utilization but at the cost of a longer rebuffering time.

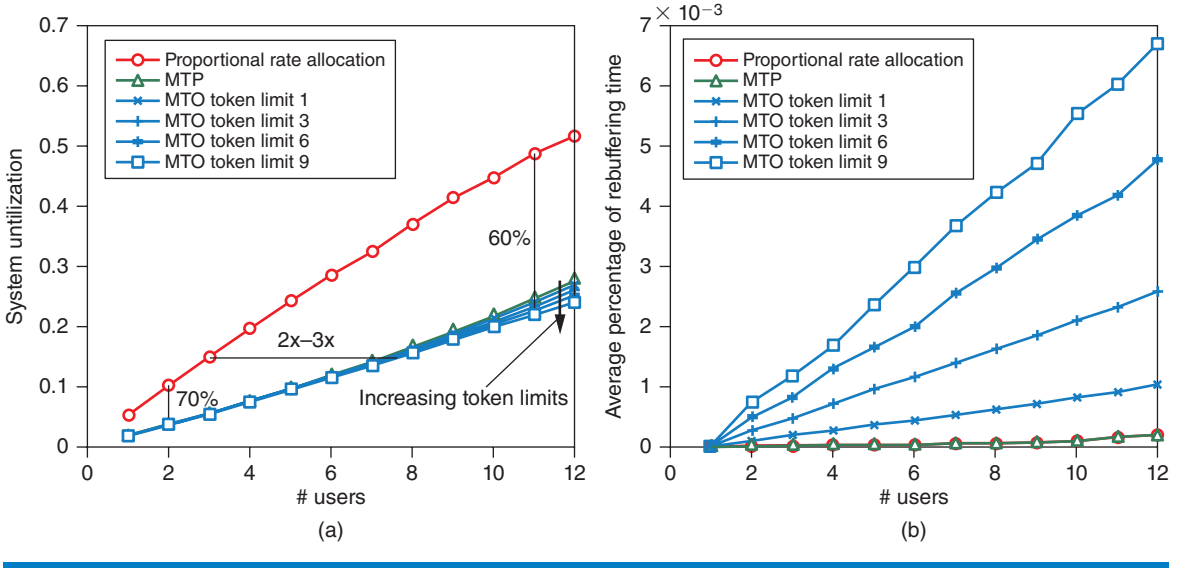


Figure 9: (a) *The system utilization.* The proportional rate allocation has the highest system utilization. MTP and MTO achieve reduced system utilization. MTO schemes do a little better than MTP. An MTO scheme with a higher token limit achieves a lower system utilization. (b) *The average percent rebuffering time.* The proportional rate allocation and MTP have the same percent rebuffering time. MTO schemes result in higher rebuffering time, which increases as the token limit increases
(Source: Lu and de Veciana[19], 2013. Used by permission of the IEEE)

QoE Enhancement over LTE

Quality of experience (QoE) is the key performance evaluation metric for multimedia delivery technologies, accounting for their subjective aspects. Cellular systems, originally focused on voice and subsequently best effort low-rate data traffic, are now transmitting an ever-increasing percentage of video traffic. The main global cellular standards body is 3GPP, with Release 8 and afterwards referred to collectively as its Long Term Evolution (LTE). Providing high QoE to video users is a key objective for LTE system design. A typical LTE system architecture is shown in Figure 10, where a few users are not able to stream high data rate video due to spatio-temporal variations of link conditions and/or because of the geometry itself. One of the other associated major issues is devising appropriate evaluation methodology for quantifying video service capacity of LTE systems.

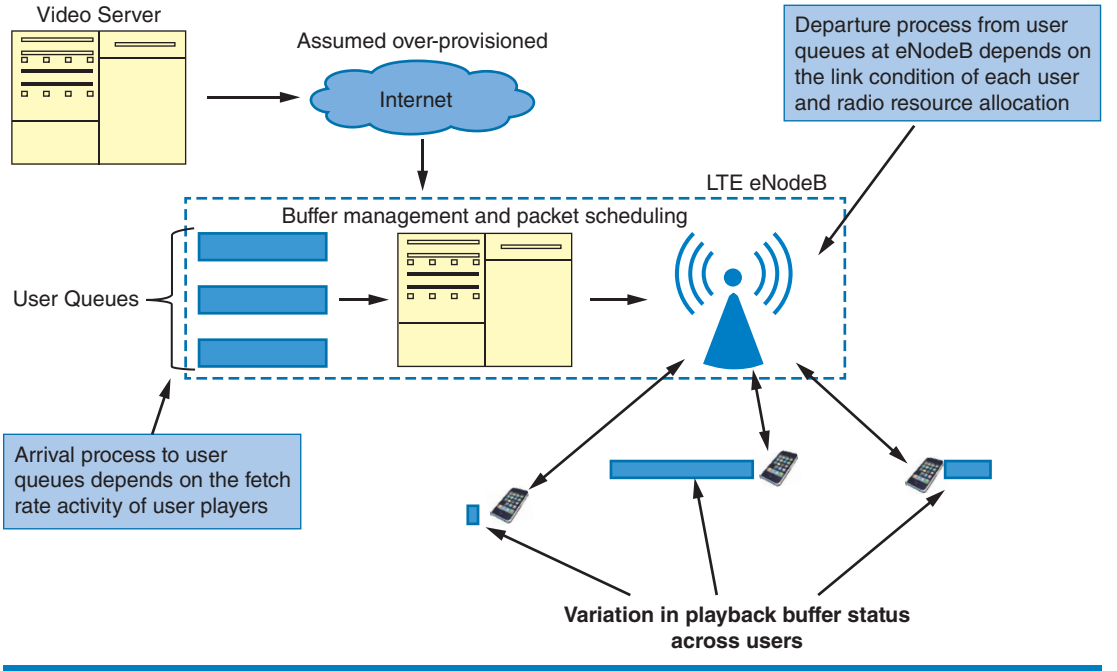


Figure 10: LTE system architecture
(Source: The University of Texas at Austin, 2014)

In Singh et al.[17], we address the above issues by proposing a QoE-based evaluation methodology to assess the LTE system video capacity in terms of the number of unicast video clients that can be simultaneously supported for a given target QoE. We introduce the metric of *rebuffering outage capacity* ($C_{\text{rebuf}}^{\text{out}}$) to quantify the video service capacity, defined as the number of active users that can simultaneously stream video, where users are satisfied A^{cov} percentile of the time, with a user being counted as satisfied if and only if the rebuffering percentage in the user's video streaming session is less than or equal to A^{out}. The presented evaluation methodology incorporates adaptive

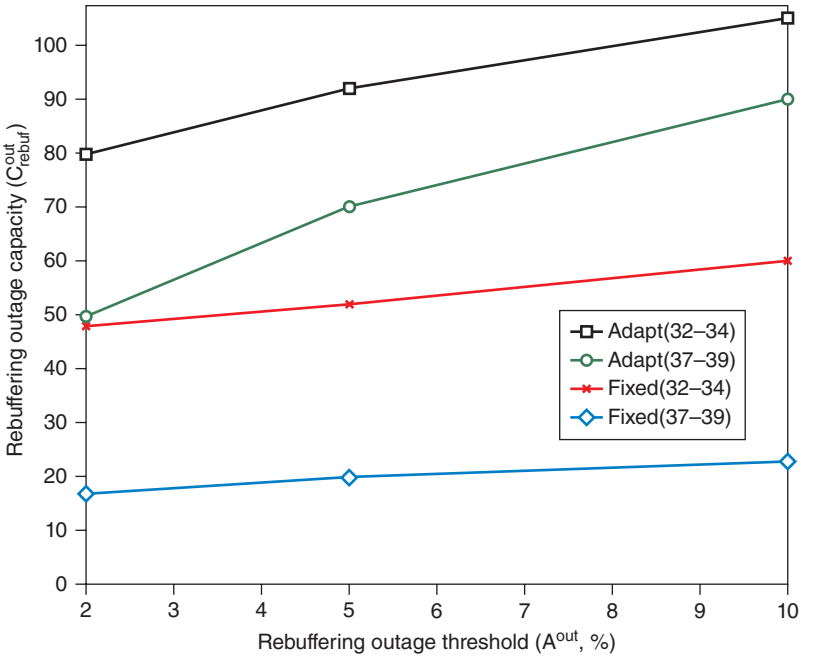


Figure 11: Variation in the rebuffering outage capacity with the rebuffering outage threshold across various configurations (data re-plotted from Singh et al.[17])
(Source: The University of Texas at Austin, 2014)

streaming, a promising technology for video delivery over wireless, and the consequent QoE-capacity tradeoff is presented in Figure 11. The scenarios shown are:

- Fixed (32–34): users fetch a video stream with fixed quality in the range of 32–34 dB PSNR.
- Fixed (37–39): users fetch a video stream with fixed quality in the range of 37–39 dB PSNR.
- Adapt (32–34): users adapt according to link conditions. The representation levels available from the server range from the quality level of 24–26 dB up to the maximum representation level having the corresponding quality in the range of 32–34 dB PSNR.
- Adapt (37–39): users adapt according to link conditions. Configuration is same as Adapt (32–34) with the exception of the maximum available quality being in the range of 37–39 dB PSNR.

"…adaptive streaming proves to be very effective in increasing the rebuffering outage capacity."

As expected, with respect to the defined outage criteria, adaptive streaming proves to be very effective in increasing the rebuffering outage capacity. Also, monotonic increases in C^{out}_{rebuf} with increase in A^{out} are observed. The relative gain from Fixed (37–39) to Adapt (37–39) is much higher than that from Fixed (32–34) to Adapt (32–34). This is because the clients have more video representation levels to switch to in the former. Note that allowing more representation levels for adaptive streaming leads to decrease

in rebuffering outage capacity—compare Adapt (37–39) and Adapt (32–34)—because of the content-agnostic RAN (proportional fair resource allocation is used for computing the results) and greedy client-based implementation of HAS services. This motivates the use of a QoE-aware resource management to work in conjunction with adaptive streaming. To address the above inadequacy, we propose a QoE-aware radio resource management (RRM) framework (PFBF)[17], which allows the network operator to further enhance the video capacity. Once again this approach is a cross-layer solution. Figure 5 exhibits the types of information that would need to be exchanged.

Figure 12 shows the distribution of rebuffering percentage across users for different numbers of users. In the plot, f_{min} is a parameter of the proposed algorithm PFBF, whose higher value implies higher fairness being guaranteed by the network to users in terms of their playback buffer. As compared with the baseline of PF-based scheduling, PFBF allows the network to accommodate more users while keeping the 95th value of rebuffering percent almost the same. For example, fixing A^{cov} = 95 percent and A^{out} = 5 percent, the rebuffering outage capacity for PF, PFBF with f_{min} = 10 and PFBF with f_{min} = 30 is 70, 82, and 87 respectively. Thus, the proposed resource allocation provides a capacity gain in the range of approximately 20 percent, while providing the operator the flexibility to dynamically tune the parameters based on user preferences.

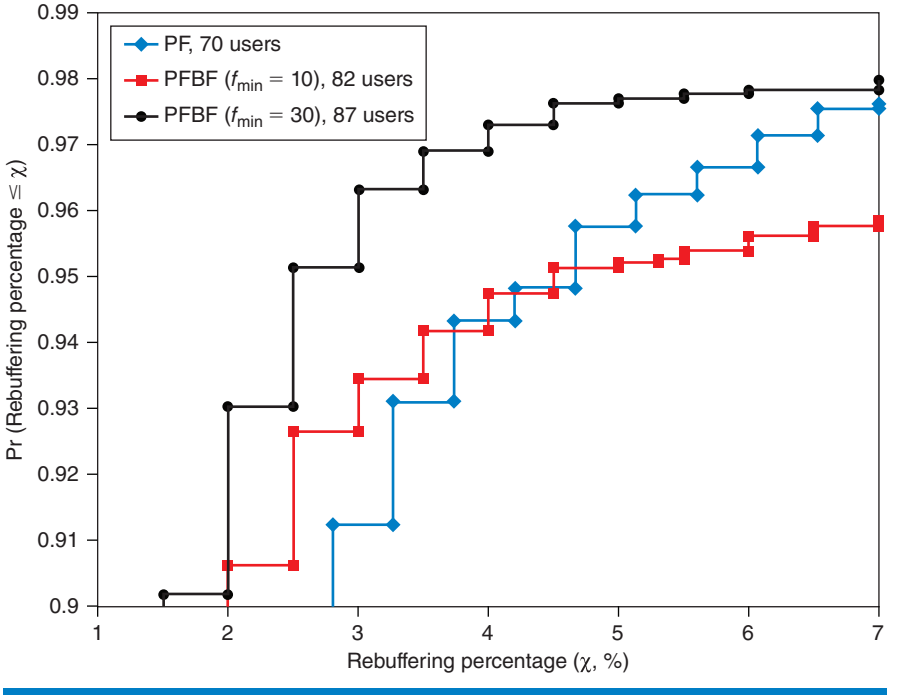


Figure 12: CDF of rebuffering percent across users (data re-plotted from Singh et al.[17])
(Source: The University of Texas at Austin, 2014)

Table 3 summarizes lessons learned on stored video transmission.

Research Agenda	Lesson Learned
QoE Optimized Rate Adaptation	A flexible framework for resource allocation and quality adaptation can be designed to account for multiple QoE factors: average quality, quality variability, rebuffering, cost, and so on. By exploiting tradeoffs amongst heterogeneous videos and users' QoE preferences one can achieve substantial capacity gains (as much as 60%).
Knowledge of future capacity variations	Future wireless networks and applications could be designed to exploit knowledge of anticipated capacity variations based on known mobility patterns. This can enable both better video QoE and more efficient resource utilization.
QoE constrained admission control	Measurement-based admission control, which is sensitive to video quality (for example, 2nd-order eCDF) is a feasible and promising approach to managing users' QoE.
QoE-aware Radio Resource Management for LTE	Rebuffering outage capacity is a suitable capacity metric for LTE systems. A QoE-aware resource allocation in conjunction with adaptive streaming offers a large scope of improvement in video capacity.

Table 3: Summary of Lessons Learned on Stored Video Transmission (Source: The University of Texas at Austin, 2014)

Real-Time Video Transmission

"Real-time video transmission requires maintaining stringent delay bounds to ensure a good user experience."

Real-time video transmission requires maintaining stringent delay bounds to ensure a good user experience. The stringency of the delay bound is further dependent on the specific use case such as video conferencing or live streaming. We address the problem of resource allocation and multiuser scheduling across real-time video users with hybrid delay QoS requirements to maximize a video quality-based utility function in the section "Perceptual Quality Optimized Resource Allocation and Scheduling with Statistical Delay Guarantees."

The delay-sensitive nature of real-time video transmission also motivates the use of unreliable transport protocols, such as user datagram protocol (UDP) for video delivery. This causes the wireless channel impairments, such as losses and delays, to be visible at the APP layer. Consequently, achieving good overall video quality for real-time video requires mitigating channel-induced distortions. Since video quality is the metric of interest from the user perspective, transmission policies should be designed to minimize the impact of losses on video quality. In the section "Perceptual Importance based Video Packet Prioritization," we review a low overhead architecture for real-time video transmission over MIMO channels to mitigate channel-induced video distortions through packet prioritization. In the section "Perceptual Quality

Optimized Unequal Error Protection," we summarize a statistical model using local linear regression to predict a mapping of packet loss fractions in scalable video layers into a measure of video quality degradation, which is used for unequal error protection.

A cross-layer overview of those perceptual optimization techniques is provided in Figure 13, showing the functions of different protocol layers as well as required side information exchange for perceptual quality optimization. A common theme of the proposed cross-layer techniques is adapting the video source and the wireless transmission medium exploiting the video data structure and the wireless channel variability to improve video quality and capacity.

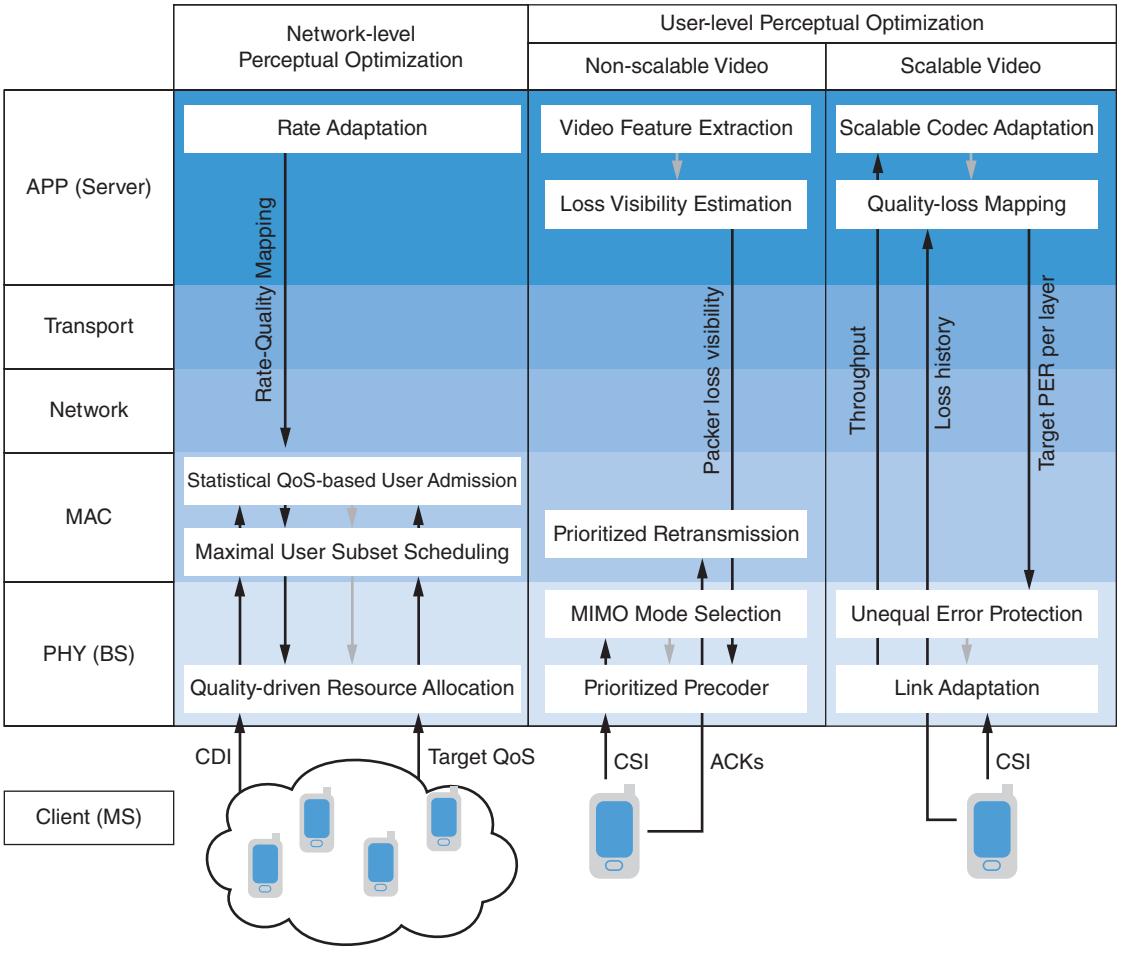


Figure 13: Single-user and multiuser cross-layer approaches to perceptual quality optimization (Source: The University of Texas at Austin, 2014)

Perceptual Quality Optimized Resource Allocation and Scheduling with Statistical Delay Guarantees

Considering a network with multiple video users with possibly different delay requirements sharing a wireless channel resource as shown in Figure 14, we derive the resource allocation and rate adaptation policy that maximizes the

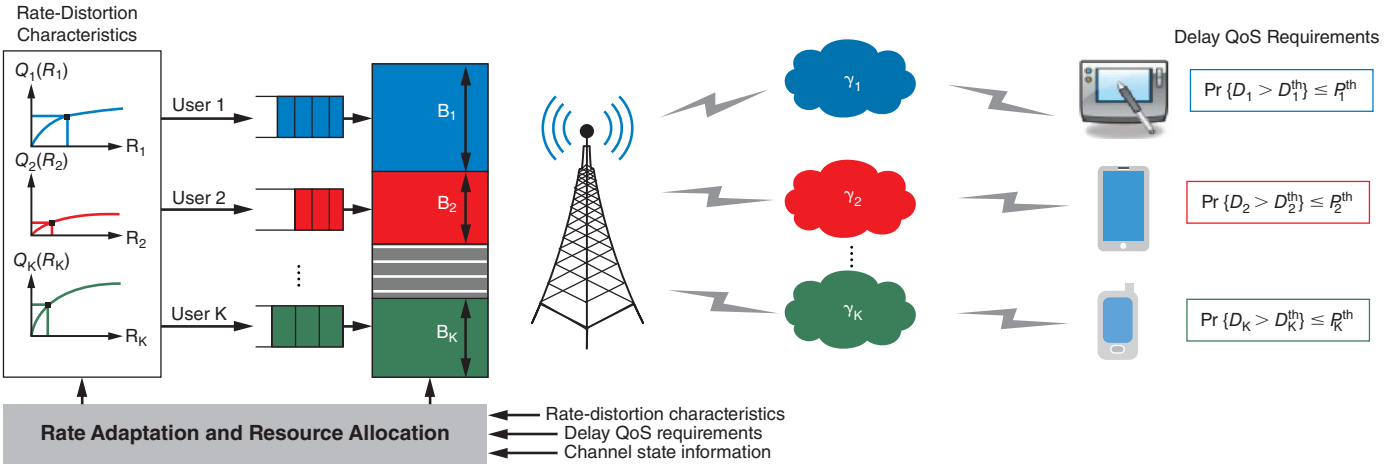


Figure 14: System block diagram for quality-driven resource allocation applied to delay-constrained video transmission (Source: The University of Texas at Austin, 2014)

sum video quality.[22][23] Resource allocation adapts the partitioning of the wireless channel resources across the users, and rate adaptation adapts the video source rate of each user. We show that the optimal operating point per user is such that the rate-distortion slope is the inverse of the supported video source rate per unit bandwidth. The maximum source rate per unit bandwidth is a fundamental measure of the number of video bits per channel use that can be delivered subject to the QoS requirement and we refer to it as the *source spectral efficiency*. We solve the alternative problem of fairness-based resource allocation whereby the objective is to maximize the minimum video quality across users and contrast the solution with the sum quality maximizing policy.

We further derive a policy, termed *maximal user subset scheduling*, to select a subset of users such that all scheduled users can meet their statistical delay requirement. We show that the optimal scheduling policy can be obtained in polynomial time in the number of users and it involves computing the minimum resource allocation required by each user to support their QoS requirement, using it as a sorting criterion, and scheduling the first sorted users such that the sum of their minimum resource requirement does not exceed the total available resources. Under the fairness constraint, a similar solution is obtained with the major difference that the sorting criterion is the video quality corresponding to the minimum rate representation of the video sequence.

We consider a single cell setup where the users are distributed according to a Poisson point process (PPP) in the cell. Half the video users have a delay constraint $D_1^{th} = 2$ sec corresponding to live video users and the other half have a delay constraint $D_2^{th} = 0.3$ sec corresponding to a typical two-way video conferencing user. The target delay bound violation probability is 0.1 for both sets of users and the total system bandwidth is 20 MHz. Figure 15 shows the number of users supported using the proposed maximal user subset

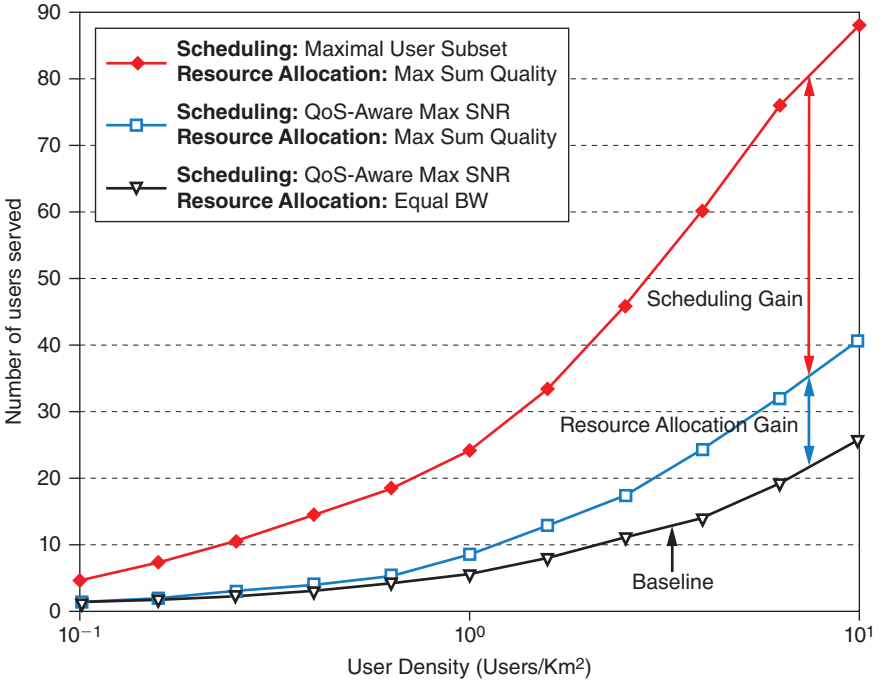


Figure 15: Number of users supported by the proposed scheduling and resource allocation algorithm in comparison to the baselines vs. user density for P_t = 30 dBm and B = 20 MHz
(Source: The University of Texas at Austin, 2014)

scheduling algorithm along with sum quality-maximizing resource allocation. We prove that the maximal user subset scheduling algorithm with sum quality-maximizing resource allocation outperforms any other scheduling/resource allocation combination under the same delay constraints. To distinguish the resource allocation and scheduling gains, we consider two baselines. In both baselines, scheduling is based on QoS-aware Max SNR, that is, the maximum number of the highest SNR users that can be supported such that they can meet their delay constraint is scheduled. In the first baseline, the available bandwidth is divided equally among the scheduled users. In the second baseline, the bandwidth is allocated according to the sum quality-maximizing resource allocation. Thus, the difference between the baselines is the resource allocation gain and the difference between the second baseline and the proposed algorithm is the scheduling gain. We observe that significant gains are achieved in terms of the total T supported in the system. The resource allocation gain corresponds to 1.6x increase in capacity due to the better partitioning of the wireless system resources. The scheduling gain corresponds to 2.2x–3.5x increase in capacity.

Perceptual Importance based Video Packet Prioritization

We review a new low overhead architecture for real-time video transmission to mitigate channel-induced video distortions. Our proposed architecture uses quantized loss visibility scores embedded in the packet header at the expense

of only few extra bits per packet while avoiding a complex cross-layer design.[21][24][25] At the PHY layer, we use the loss visibility values to classify video packets into different priority classes. For a MIMO system, each class of packets is transmitted through a different spatial stream corresponding to a decomposed subchannel of the MIMO channel. The resulting mapping is illustrated in Figure 16, and the physical interpretation of the process is that high priority packets are sent over the more reliable MIMO subchannels.

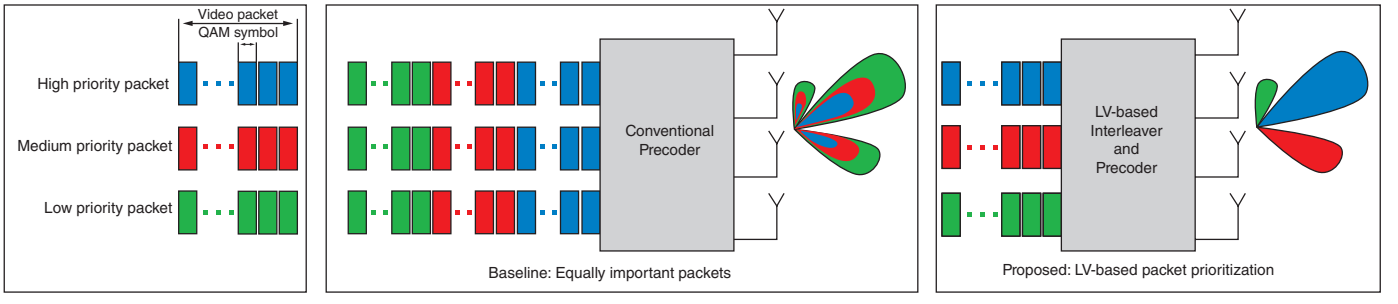


Figure 16: Illustration of the proposed precoder and interleaver design for packet prioritization over MIMO channels (Source: The University of Texas at Austin, 2014)

To optimize the loss-visibility–based transmission policy for high video quality and low latency, we define an optimization metric that generalizes the conventional notion of throughput by weighting each packet in the optimization objective by its loss visibility. Since loss visibility reflects the visual perception of a corresponding packet loss, our optimization metric is a proxy for *the total perceptual value of packets successfully delivered per unit time*. Given the proposed objective function that enables joint optimization of video quality and latency, we derive optimized PHY layer packet prioritization schemes. We derive the optimal packet-stream mapping that maximizes the loss-visibility–weighted throughput objective. The solution can be summarized as follows: (1) The MIMO channel is decomposed into parallel streams, (2) the per-stream transmission rate, that is, modulation order, is chosen to maximize the corresponding throughput per stream, (3) the spatial streams are ordered by their probability of packet error, a function of both the per-stream SNRs and (potentially unequal) modulation orders, and (4) the packets are classified according to a thresholding policy whereby higher priority packets are mapped to high order streams as defined by the ordering in (3). The optimal thresholding policy is such that the load is balanced across streams based on the fraction of packet per priority class, the modulation order per stream, and the retransmission overhead. We show that the solution enables jointly reaping gains in terms of improved video quality and lower latency: a packet prioritization gain results from transmission of more relevant packets over more reliable streams, and an unequal modulation gain results from opportunistically increasing the transmission rate on the stronger streams to enable low latency delivery of high priority packets, as shown in Figure 17.

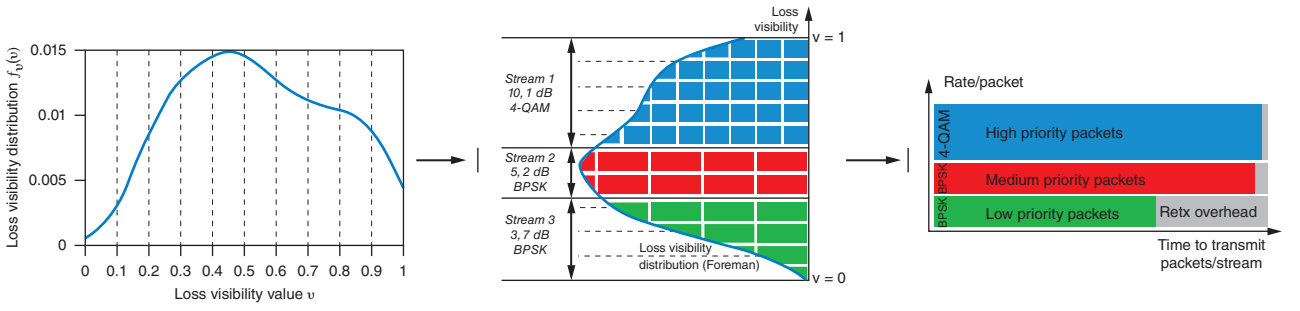


Figure 17: Graphical illustration of loss visibility optimized transmission policy for $\gamma(H) = [10.1; 5.2; 2.7]$ dB and the Forman video sequence; (a) Obtain loss visibility (shown for the Foreman video sequence), (b) Decompose MIMO channel, (c) Determine throughput-maximizing modulation order per stream, (d) Find the optimal thresholding policy. Note that high priority packets achieve both higher rate and reliability
(Source: The University of Texas at Austin, 2014)

Figure 18 demonstrates the video quality gains for a range of antenna configurations for the Foreman video sequence. For a fixed antenna configuration, the gains are maximized when $S = \min(N_t, N_r)$. This is because the large variability in the post-processing SNRs across streams enables more effective packet prioritization. Furthermore, increasing the number of antennas for a fixed number of streams improves video quality but reduces the video quality gain. Gains in excess of 10 dB are achieved over a range of antenna configurations.

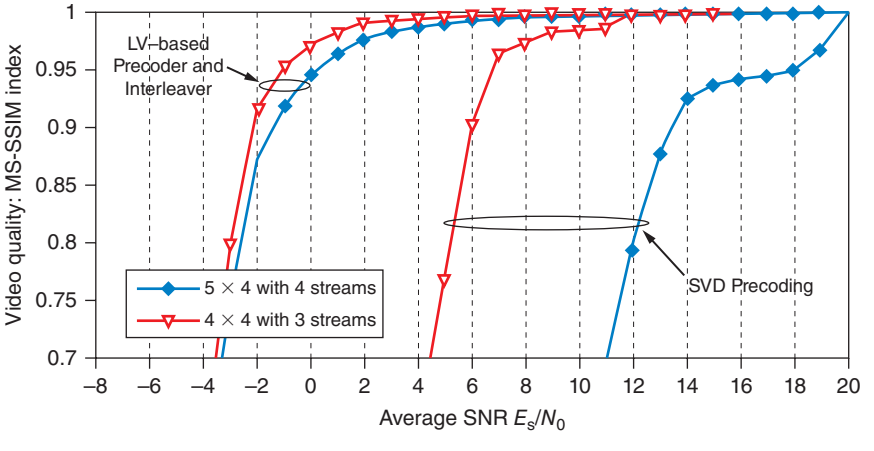


Figure 18: Comparison of the loss-visibility–based packet prioritization vs. nonprioritized MIMO precoding for H.264-encoded Foreman sequence for different antenna configurations over a range of SNRs. The retransmission limit is $r = 4$ and the channel coherence time is 1 GoP
(Source: The University of Texas at Austin, 2014)

Perceptual Quality Optimized Unequal Error Protection

Due to the hierarchical frame structure and inter-layer prediction in scalable video coding, error propagation causes packet losses from different video layers to affect the video quality differently. This motivates selecting potentially

different modulation and coding schemes (MCS) for different packets according to their perceptual relevance to the end-to-end video quality.

"We propose providing unequal error protection (UEP) for SVC-coded content on the basis of a quality-loss mapping technique…"

We propose providing unequal error protection (UEP) for SVC-coded content on the basis of a quality-loss mapping technique that estimates the perceptual quality reduction due to packet losses at each video layer, hence termed perceptually optimized UEP.[20][26][27] Online learning for unequal error protection is motivated by two key insights: on one hand, for real-time video, where a video sequence is not pre-encoded, an offline approach to determining the unequal protection levels is infeasible. On the other hand, an online approach enables adapting to scene changes as well as changes in temporal and spatial characteristics. Since the video natural scene statistics exhibit correlation over short durations of time, the quality-loss mapping $\rho \rightarrow Q^{(k,l)}(\rho)$ varies smoothly over time. Thus, local information is more relevant for adaptation. We emphasize that the notion of locality corresponds to: (1) similarity in quality sensitivity to losses, and (2) proximity in time. Using this intuition, we propose solving the problem using local linear regression over a window of observations. Specifically, the logarithm of the PER is calculated as a linear combination of the previous PER logarithms to minimize a weighted least squares problem with the weights selected according to a smoothing kernel. The packet error rate logarithms are used because the range of PER variations is better captured on a log scale. Thus, the error rate logarithm characterizes the mapping more accurately. Given the target packet error rate that could be supported per video layer, *video-aware link adaptation* can be applied in the physical layer to determine the modulation and coding scheme for each packet such that unequal error protection is achieved.

First, to demonstrate the value of unequal error protection, Figure 19 shows the video quality achieved when a fraction of packets is lost from each

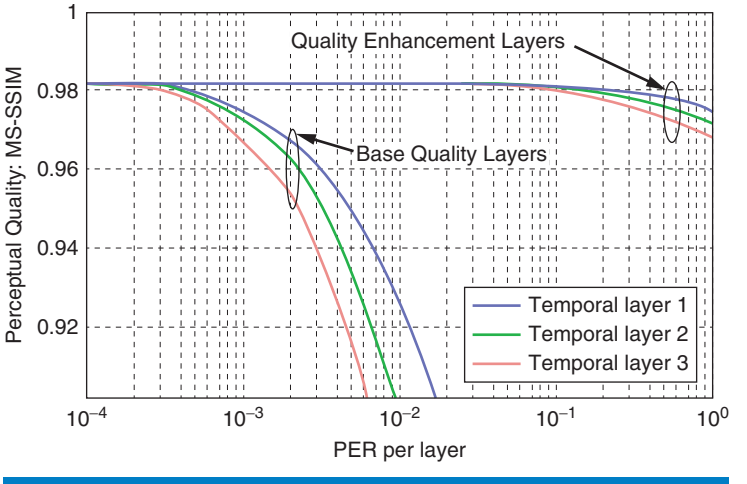


Figure 19: Offline learning of quality-loss mapping of the "riverbed" video sequence
(Source: The University of Texas at Austin, 2014)

temporal-quality layer of the "riverbed" video sequence, averaged over several realizations. Since packet losses in the base quality layer may cause frame losses and error concealment, we notice a significant drop in quality due to losses of packets from the base quality layer. For a given video quality requirement, per-layer PER targets among different quality layers are an order of magnitude apart, which clearly advocates for unequal error protection.

Next, we demonstrate the online learning algorithm on test video sequences from the LIVE video database to ensure its rapid convergence as well as its adaptability to video temporal characteristics. We use a Gaussian kernel with a parameter 0.2. Figure 20 shows a case study of online learning applied to the "pedestrian" video sequence with a learning window of 30 GoPs. Starting with equal error protection for all layers, the corresponding video quality is poor. As the algorithm adapts online, the protection of base quality layers is increased and the protection of quality enhancement layers is relaxed so as to approach the target video quality. The convergence time is very rapid as the target video quality is approached in less than 2 seconds of video playback and is stable throughout playback. Furthermore, we show a baseline corresponding to the base quality layer alone and an upper bound on quality corresponding to lossless transmission. The online algorithm closely approaches that with lossless transmission, thus showing the robustness of the proposed UEP scheme.

"The online algorithm closely approaches that with lossless transmission, thus showing the robustness of the proposed UEP scheme."

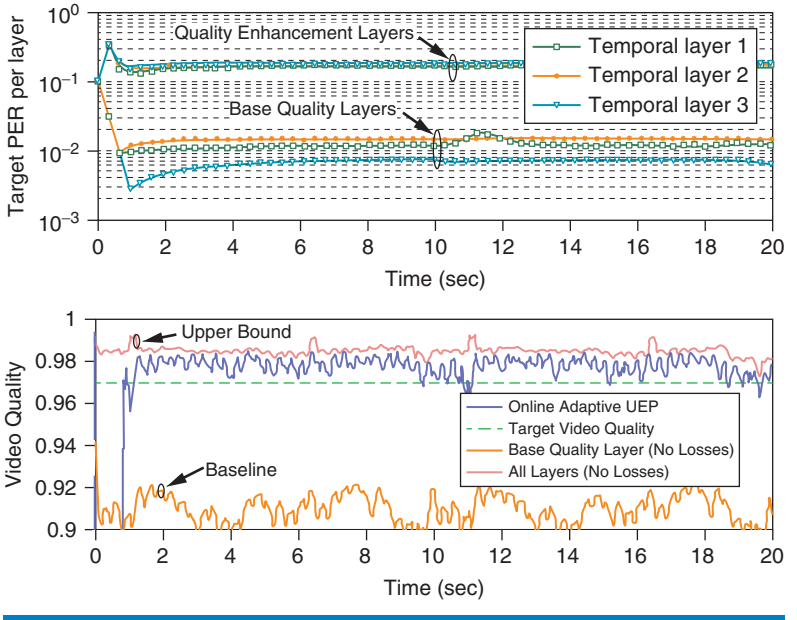


Figure 20: Online learning of quality-loss mapping case study: Top: Unequal protection levels of scalable video layers over time; Bottom: Video quality over time
(Source: A. Abdel Khalek, C. Caramanis, and R. W. Heath, Jr., "Online Learning For Quality-Driven Unequal Protection Of Scalable Video," In Proceedings of *IEEE International Workshop on Machine Learning for Signal Processing* (MLSP), September 2012. Used with permission of the IEEE)

Perceptual Quality Optimized Interference Shaping

As we have noted previously, the rate at which data can be sent over a wireless channel is inherently variable due to variable link conditions. An even bigger source of variation in dense and irregular cellular networks (HetNets), however, is bursty co-channel interference, caused in large part by bursty traffic on lightly loaded small cells. These throughput variations can lead to very large quality variations, especially for real-time video, in which case the throughput variations cannot be smoothed out through buffering, resulting in a degraded quality of experience (QoE) for the user.

"…we propose and analyze a network-level resource management algorithm called interference shaping to smooth out these throughput variations…"

In Singh et al.[28], we propose and analyze a network-level resource management algorithm called *interference shaping* to smooth out these throughput variations, and hence improve the QoE of video users by reducing the variability of interference. Interference shaping operates by decreasing the transmission power, and hence peak rate, of co-channel APs serving bursty best-effort data users. This smooths their transmit power profile and hence the interference caused by them to the video user link, at the cost of a modest rate decrease and latency increase for best-effort users. For video users, QoE is quantified by benchmarking against a metric, which incorporates the strong dependence of the current QoE (which is subjective) on the recent past.

The QoE of data users is evaluated using a framework that quantifies the response of human sensory system to an external stimulus.

The variation of the overall objective quality score H-MS-SSIM and the corresponding QoE (measured as the inverse of DMOS) with the scaling of the peak power (rate) of the interfering base station is shown in Figure 21.

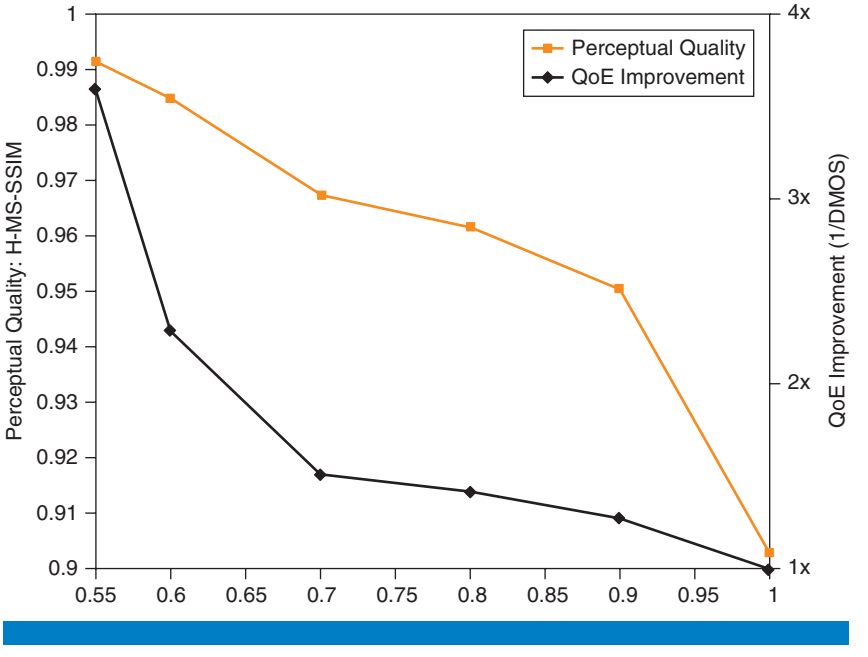


Figure 21: Variation of overall quality (H-MS-SSIM) and QoE with interference shaping in presence of a single dominant interferer
(Source: The University of Texas at Austin, 2014)

The proposed technique increases mean video quality and reduces the quality variations over time, with a net perceptual increase of about 2–3x as compared to the case with no power/rate scaling (scaling factor of 1).

Interference shaping can be applied to both unicast and multicast real-time video streaming with gains proportional to the number of video users sharing the same broadcast and interferers in the latter. The overall lesson from interference shaping is that even without increasing the amount of resources available or increasing the capacity, by simply adapting how the resources are used to account for the presence of video (and its intolerance to bursty SINR), subjective performance can be significantly improved at virtually no cost.

"Interference shaping can be applied to both unicast and multicast real-time video streaming…"

Video Oriented Wireless Network Management

Increasing the network density by deploying many low power/low cost base stations (BSs) and access points (APs) is one of the most promising approaches to boost network capacity and provide the necessary infrastructure that much of the rest of this article's contributions relies upon. Called a *heterogeneous cellular network* (a "HetNet", or HCN) or a small cell network, we envision a diverse and dense deployment of APs differing in transmit powers, radio access technologies (RATs), carrier frequencies, backhaul capacities, and deployment scenarios, resulting in a complex and "organic" network architecture.[45] Due to the much lower power of these small cells as compared to macro BSs, however, a limited number of users appear in their nominal coverage areas, which limits the relief provided to the congested macrocell tier and an overall underutilization of the small cell resources. A summary of lessons learned on real-time video transmission is shown in Table 4.

Research Agenda	Lesson Learned
Multiuser real-time video transmission	Optimizing resource allocation to capture variability in application requirements, delay bounds, and rate-distortion behavior across users enables significant improvements in video quality and capacity.
Single-user real-time video transmission	(1) Using packet importance side information, wireless networks can be redesigned at a low expense to directly minimize the impact of unreliable channel conditions on video quality, (2) Online learning is a key tool for adaptation to temporal video characteristics, scene changes, and enabling unequal error protection.
Interference management	Besides the conventional technique of mitigating interference, smoothing the interference also offers improvement in the QoE of real-time video by helping smooth the quality variations.

Table 4: Summary of Lessons Learned on Real-Time Video Transmission (Source: The University of Texas at Austin, 2014)

Natural questions to ask in such a scenario are:

- Should video users be proactively pushed onto small cells?
- If yes, how much video traffic should be offloaded?
- What is the impact of key system parameters like deployment density or transmit power on the above answers?

The thrust of this section is to answer these kinds of questions both in the context of multi-RAT HetNets and heterogeneous cellular networks (HCNs). The distinction in the two scenarios stems mainly from the interference environment, where networks in different RATs operate on orthogonal bands. As demonstrated below, in contrast to multi-RAT HetNets (such as Wi-Fi offloading), co-channel HetNets require smart interference avoidance in conjunction with offloading in order to leverage the full capacity gains.

"…co-channel HetNets require smart interference avoidance in conjunction with offloading in order to leverage the full capacity gains."

Offloading in Multi-RAT HetNets

Leveraging widely deployed (and very inexpensive) Wi-Fi APs to meet the increasing video traffic demand is an attractive and popular strategy, which further motivates analyzing video traffic across multiple RATs. In this work, offloading across different networks is modeled using the algorithm of association area/cell range expansion, where users are offloaded to smaller cells using an *association bias*. A positive association bias implies an affinity for a small cell by the bias amount, even if it is received at weaker power than the macrocell. Thus, the association bias tunes the aggressiveness of offloading. This association area expansion is demonstrated in Figure 22, where the natural association areas (on left) are expanded by the use of a 15 dB association bias (on right). We propose a tractable and general model to analyze such complex networks[31][32], where the location of APs of each class is assumed to be drawn

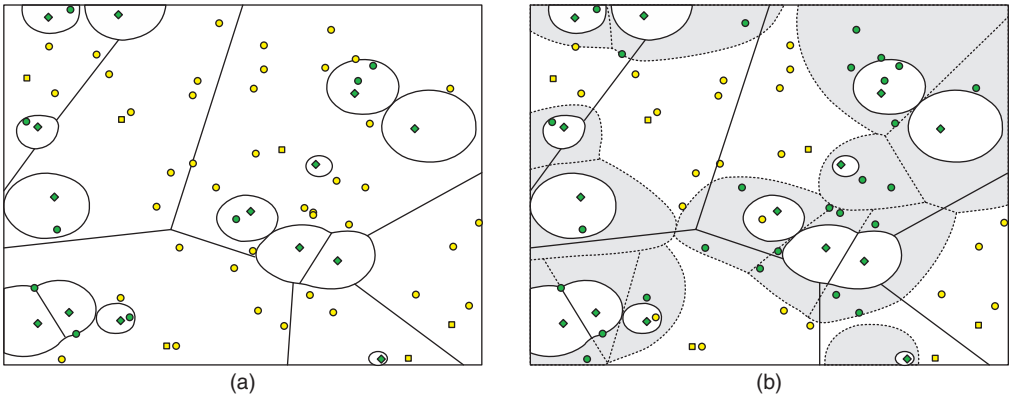


Figure 22: Association regions of a network with two classes of APs. The APs of first class (macrocells) are shown as squares and those of second (small cells) are in diamonds. The users are shown as circles. The natural association regions due to low power of small cells are in (a) and the expanded association regions resulting from the use of bias of 15 dB are shown in (b)
(Source: The University of Texas at Austin, 2014)

from a Poisson point process (PPP). Using the tools from stochastic geometry and Palm calculus[33], the user throughput/rate distribution across the entire network is characterized as a function of key system parameters. The proposed model and analysis is validated by comparing the analytical results with those from a realistic multi-RAT deployment.

Using the developed analysis[31], it is shown that optimal offloading can yield up to 2–3x gains in cell edge and cell median rate, as illustrated in Figure 23. Further, it is shown that the optimal association bias is inversely proportional to the density of the corresponding RAT. This is due to the increasing interference in the orthogonal band. Moreover, the optimal association biases are large (10–14 dB) for the typical density values of small cells—highlighting the aggressive offloading required in orthogonal networks.

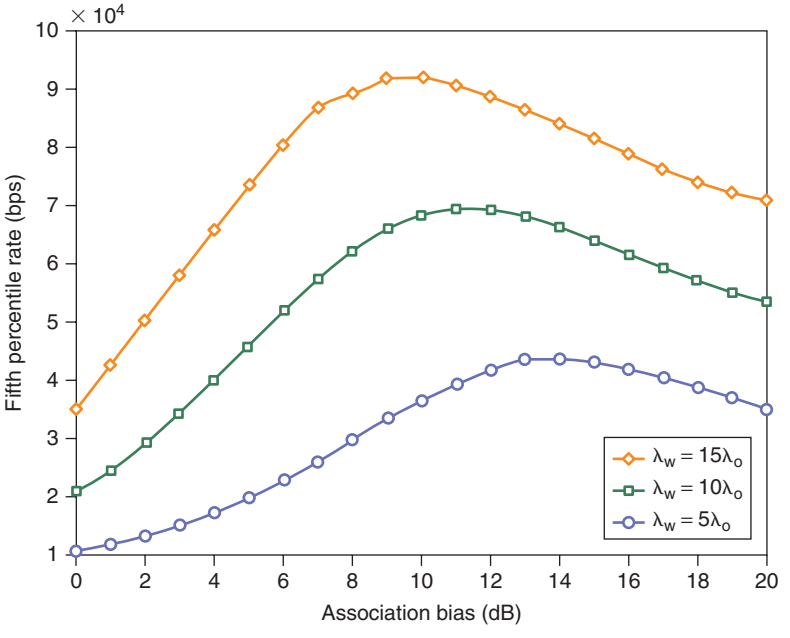


Figure 23: Fifth percentile rate in the network as a function of the association bias for various density ratios of small cell network (orthogonal RAT, index *w*) to macro cell network (index *c*) (Source: The University of Texas at Austin, 2014)

Offloading in Co-Channel HetNets

As established in the previous section, it is desirable to offload mobile users to small cells, which are typically significantly less congested than the macrocells. A key difference in co-channel deployments is that offloaded users not only often have reduced desired power (due to the bias), but also stronger co-channel interference (from the now-interfering macrocell). Therefore, the gains from load balancing are degraded if suitable interference avoidance strategies are not adopted. One such strategy of interference avoidance is resource partitioning, wherein the transmission of the macro tier is periodically muted on a certain fraction (*resource partitioning fraction,* η) of radio resources, that is,

time and/or frequency blocks. The offloaded video users can then be scheduled in these resources by the small cells leading to their protection from co-channel macro tier interference. Perhaps counterintuitively, the entire network throughput significantly benefits from this muting of macrocell resources, despite the macrocells being the traffic bottleneck. As we will see, the gain from offloading with muting is so large that it easily offsets the "waste" of macrocell resources.

The operation of range expansion and muting (resource partitioning) in a two-tier setup is shown in Figure 24. Biasing and muting are strongly coupled: for example, an excessively large association bias can cause the small cells to be overly congested with users of poor SINR, which requires excessive muting by the macrocell to improve the rate of offloaded users. Thus, they must be jointly optimized.

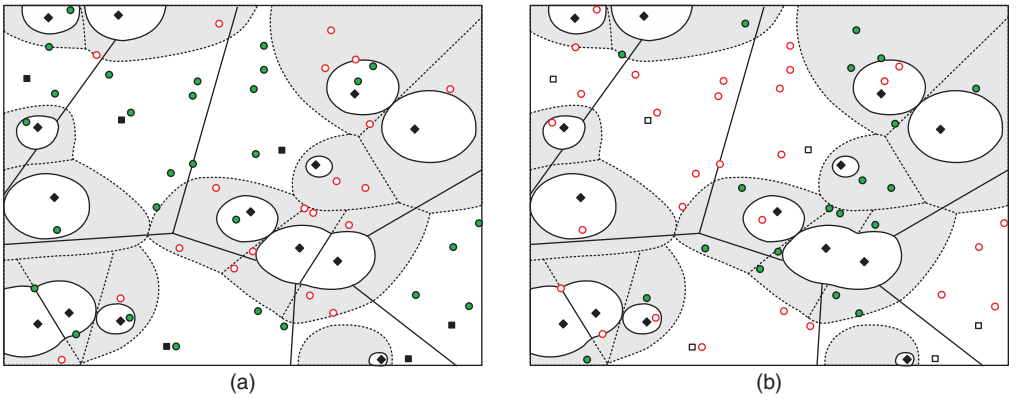


Figure 24: A filled marker is used for a node engaged in active transmission (BS) or reception (user). (a) The macrocells (filled squares) serve the macro users and small cells (filled diamonds) serve the non-range expanded users (filled circles). (b) The macrocells (hollow squares) are muted while the small cells (filled diamonds) serve the range expanded users (filled circles in the shaded region)
(Source: The University of Texas at Austin, 2014)

We propose a tractable and general model[32] to analyze the joint offloading/biasing and resource partitioning/muting in HCNs and characterize the rate distribution across the network as a function of offloading and resource partitioning parameters. Using the developed analysis, it is shown that jointly optimizing resource partitioning and load balancing can lead to 2–3x gain in the cell edge and cell median throughput. Cell median throughput is shown in Figure 25 as a function of association bias and optimal resource partitioning fraction (the fraction of time the macrocell mutes its transmission). Further, it is shown that the optimal partitioning fraction and offloading bias decrease with increasing density of small cells due to increasing interference. Overall, this is a very significant gain in video capacity from a quite simple decentralized strategy with static bias values and partition fractions. Even larger gains should be possible from dynamic adjustment of these values.

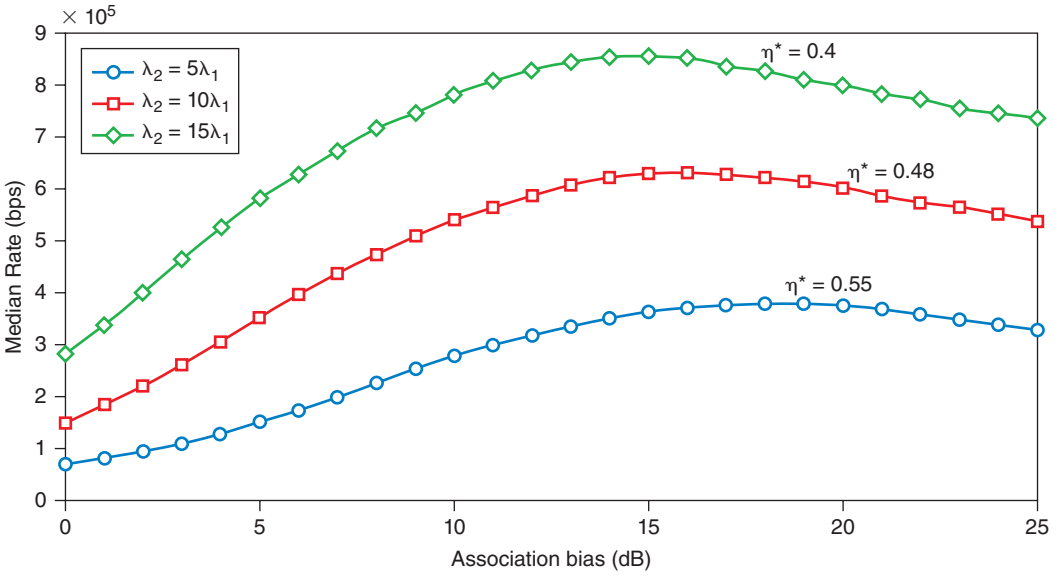


Figure 25: Median rate in the network as a function of the association bias for various density ratios of small cells (index 2) to macrocells (index 1)
(Source: The University of Texas at Austin, 2014)

Conclusion and Lessons Learned

In this article, we summarized the key results and insights into designing next generation wireless networks that can accommodate the anticipated exponential increase in video traffic in a fashion that ensures good QoE of individual video users. Our vision presents a holistic solution to the wireless video capacity problem by capturing the following four key research thrusts: (1) video quality modeling and prediction, (2) rate-adaptive transmission of stored video, (3) cross-layer optimization for real-time video transmission, and (4) load management in heterogeneous wireless networks. In what follows, we summarize the lessons learned in those four research thrusts (see Table 5).

Research Agenda	Lesson Learned
Video traffic management in multi-RAT HetNets	The amount of video traffic offloaded needs to be tuned based on key system parameters like density of small cells and the resource availability at orthogonal RAT.
Video traffic management in HCNs	Offloading to small cells is, in itself, insufficient. Smarter interference avoidance is needed in conjunction.

Table 5: Summary of Lessons Learned on Video Oriented Wireless Network Management
(Source: The University of Texas at Austin, 2014)

The first research thrust pioneered the development of new models and algorithms for full reference, reduced-reference, and no-reference prediction

that achieve high correlation with human judgments of quality recorded in large-scale subjective experiments. The key lesson learned is that natural scene statistics-based models enable accurate quality prediction in the absence of explicit reference information. Further, if only partial information about the video reference can be provided as side information, the performance of video quality predication can be further improved using the novel reduced-reference video quality assessment approaches explained in the article. We further show that it is possible to realize a tradeoff between prediction accuracy and the amount of side information, since our proposed algorithms are flexible enough to allow varying amounts of information available from the reference. In this research thrust, we also created a dynamic system model for capturing the time varying subjective quality on long video sequences. Our experiments show that the short term video quality prediction algorithms are not sufficiently accurate to predict the long term video subjective quality due to effects such as hysteresis. To account for this problem, we have shown that long term subjective quality can be further predicted using a simple dynamic system.

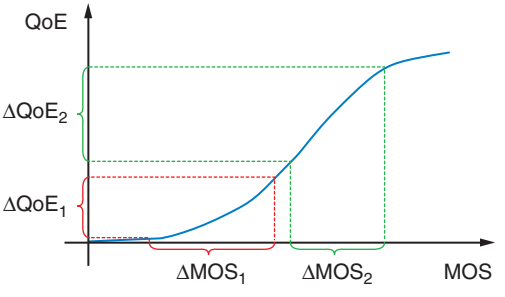


Figure 26: The relationship between MOS and QoE. Because a higher MOS implies a better QoE, QoE is an increasing function of MOS. But a larger difference in MOS (see ΔMOS_1 and ΔMOS_2) does not imply a larger difference in QoE (see ΔQoE_1 and ΔQoE_2) (Source: The University of Texas at Austin, 2014)

The quality metrics developed in this research thrust, along with almost all existing quality metrics, are designed to predict the mean opinion scores (MOSs) obtained from the subjective studies. In most subjective studies reported so far, the subjects only report their quality judgments with respect to each test video.[3][14] Their opinions on the perceptual differences between the test videos are not studied. Hence, the obtained MOSs cannot provide insight into the difference in QoE. In particular, a larger difference in MOSs may not imply a larger difference in QoEs. This is illustrated in Figure 26. Because a large MOS indicates a better QoE, QoE is an increasing function of MOS. The function, however, may be nonlinear and a larger difference in MOS may correspond to a smaller difference in QoE.

The incapability of MOSs to interpret the difference of QoE is especially critical for wireless video transmission. In wireless networks, video users share the limited network resources (such as a resource block in LTE[46]). Most existing resource allocation algorithms are designed to maximize the sum of the MOSs. But these MOS-optimized resource allocation algorithms may not maximize the sum of the QoEs. For example, let's suppose that there are network resources that can either be allocated to user 1 or user 2. If the resources are allocated to user 1, its MOS increases by ΔMOS_1 and its QoE increases by ΔQoE_1. Otherwise, if the resources are allocated to user 2, its MOS increases by ΔMOS_2 and its QoE increases by ΔQoE_1. If ΔMOS_1 is greater than ΔMOS_2, it seems that allocating the resource to user 1 is more beneficial because it maximizes the sum of the MOSs. But, if the mapping from MOS to QoE is a nonlinear mapping (as shown in Figure 26), it is possible that ΔQoE_1 is less than ΔQoE_2. Then allocating the resources to user 1 may not maximize the sum of QoEs. To develop QoE metrics that can reflect the differences in QoE, a more sophisticated design of subjective studies is necessary. For example, in every round of the subjective tests, the subjects could be asked to first view two pairs of videos. Each video pair consists of two distorted versions of the same pristine video. Then, the subjects are asked

to judge which pair of videos has a larger difference in QoE. Based on the feedback of the subjects, a new QoE metric can be obtained to predict the differences in QoE. Although such a subjective study has not been conducted on videos, a similar study on images has been reported in Charrier et al.[47] Using the maximum likelihood difference scaling (MLDS) method[48], a new quality metric was derived to predict the QoE differences of images. Future research should extend the subjective study in Charrier et al.[47] to video QoE assessment. A major limitation of Charrier et al.[47] is that the two image pairs involved in the subjective study correspond to the same pristine image. Thus, the proposed QoE metric can only be applied to the case of image broadcasting in which all users view the same image. In a wireless network, different users may watch different images. In the future, the subjective study needs to be generalized to compare video pairs corresponding to different pristine videos.

The second research thrust proposes stored video streaming techniques that maximize the video QoE taking into account a multitude of factors including average video quality, temporal variability in video quality, rebuffering time and startup delay, as well as cost to the video client. A key lesson learned in this thrust is that achieving good QoE extends beyond optimizing instantaneous video quality. A good adaptive transmission policy should also capture the variability in video quality as well as reduce the risk of rebuffering. Furthermore, capturing heterogeneity in client preferences is critical in achieving good overall QoE. Finally, for practical realization of adaptive video delivery, distributed algorithms that separate the functions of network-driven resource allocation and client-driven quality adaptation are desirable. From experimental observations, we also learned that the constraints on the users' QoE are best captured through the 2nd-order eCDF of video quality over time, which achieves a strong linear correlation with the measured subjective quality of long video sequences. This lends strong support for the eCDF as a good proxy for video QoE. Focusing on the problem of admission control, we show that it is preferable to satisfy the QoE constraints of existing users by selectively blocking newcomers, which enables the overall percentage of users satisfying the QoE constraints among both admitted and blocked users to be improved. Another vector in the research thrust explored the value of knowledge of future capacity variations, for example, through knowledge of mobility patterns. In this setup, we show that we could make capacity variations beneficial to overall QoE by exploiting the storage/buffer on mobile devices and knowledge of future variations. Finally, we define a notion of video service capacity and show that how it captures a tradeoff between rebuffering percentage and users supported as well as use it to demonstrate the value of adaptive streaming. We further show the value of a QoE-aware resource management in conjunction with adaptive streaming in improving video capacity.

"A key lesson learned is that achieving good QoE extends beyond optimizing instantaneous video quality."

The biggest challenge to the adoption and implementation of the proposed stored video streaming techniques is enabling and standardizing information exchanges among network components and layers (see Figure 5). Doing so will require convincing multiple parties that new video streaming algorithms offer substantial advantages, in particular if they require network

"The biggest challenge to the adoption is enabling and standardizing information exchanges among network components and layers"

involvement. The advantages of one protocol over another needs to be evaluated across multiple metrics, such as, for example, QoE, capacity, robustness, and profits. Moreover the potential of these new techniques needs to be evaluated in possibly heterogeneous networks carrying other traffic types as well as video streaming via legacy transport protocols. Doing so requires a systematic methodology to decide on the importance or value of QoE delivered to various heterogeneous users sharing the network—a very difficult task. For the most part the new ideas we have discussed for stored video are receiver driven and have low information overhead. One exception is the set of video delivery mechanisms that exploit future knowledge of capacity variations. Such techniques would require developing infrastructure to predict, for instance, based on historical data and/or wireless coverage maps, or crowdsource estimates for future capacity variations. As discussed in the article, if such information were available, the potential increases in system capacity would be quite high.

For the techniques discussed in this article, there still remain several research directions. For example our work has focused on infrastructure-based cellular networks where resource allocation decisions across contending users can be orchestrated. In future systems one would also expect contention-based Wi-Fi networks and D2D links to play a significant role in video delivery and they should be studied. Also one might consider how the proposed approaches could be migrated into the wireline network, through the use of video QoE management servers. The proposed techniques, based on leveraging future knowledge of capacity variations, assumed these were known accurately, but in practice uncertainty in such estimates would need to be addressed. Finally, although our focus has been protocols that optimize users' QoE, there still is a need to better understand what types of objective metrics are good proxies for users' perceived video quality and tradeoffs.

"For multiuser real-time video transmission, the key finding is that optimizing resource allocation to capture the variability in application requirements, delay bounds, and rate-distortion behavior across users enables significant improvements in video quality and capacity."

The third research thrust focused on the development of cross-layer designs that enable optimizing user-level video perceptual quality as well as providing network-level quality optimizations across users through QoE-driven resource allocation. For multiuser real-time video transmission, the key finding is that optimizing resource allocation to capture the variability in application requirements, delay bounds, and rate-distortion behavior across users enables significant improvements in video quality and capacity. Furthermore, scheduling and user admission policies that capture application-level metrics in addition to channel state information can support significantly more video users under the same resource constraints. For single-user real-time video streams, we show that packet importance, captured with only a few bits of side information, can be used to optimize physical layer techniques, such as beamforming, MIMO precoding, and channel coding, and this enables a significant increase in error resilience. Thus, at a low expense, wireless networks can be redesigned to directly minimize the impact of unreliable channel conditions on video quality. Finally, we demonstrate the value of online learning in dynamically capturing the importance of different video packets over time. We show that packet relevance to QoE can be learned over time by

introducing a notion of locality that captures similarity in quality sensitivity to losses across packets over time. We further demonstrate the power on online learning in adaptation to temporal video characteristics, scene changes, and its value in enabling adaptive unequal error protection. Finally, considering the effect of interference, we show that smoothing out throughput variations through interference shaping can improve the QoE of video users by reducing the variability of interference.

A main research direction remaining open for cross-layer perceptual optimization is nonreference video quality-based perceptual optimization. Video quality metrics that have no access to a reference signal are becoming popular and have been shown to achieve high correlation with subjective quality assessment. The success of these metrics comes due to insights from natural scene statistics and the disruption of those statistics by distortions. In the context of client-driven adaptive streaming, the client can measure and track the video quality without access to the reference and use that to optimize the rate adaptation over time to improve video quality. While the majority of the work presented in this thrust focuses on server-driven adaptation using access to the video reference, similar perceptual quality optimizations can be proposed in a client-driven framework if NR quality assessment is available in the network. For instance, our multiuser real-time video transmission problem can be extended by computing and tracking the rate-quality mapping at the client using NR quality metrics. A quantized version of the quality metric can be fed back to the base station rather than being fed forward from the server. This has two distinct advantages: (1) the client-measured video quality includes the effect of channel distortions in addition to source distortions as opposed to server-measured video quality, which only captures source distortions, and (2) offloading the quality assessment to the client reduces the server load and avoids maintaining a full session state for each client at the server. Furthermore, having the capability to measure the video quality at the receiver improves the online learning aspects of this thrust by enabling the quality-loss mapping function to be computed at the client. It avoids feeding back the index of lost packets to the server to reconstruct the distorted video and compute the reference-based video quality. Instead, the UEP algorithm can be run at the client using the acknowledgment history along with the measured NR video quality. This would enable significant savings in side information exchange and cross-layer overhead as well as reducing the video server load.

"A main research direction remaining open for cross-layer perceptual optimization is nonreference video quality-based perceptual optimization."

The final thrust focuses on load management leveraging stochastic geometric models to develop tractable load balancing strategies capable of enhancing the video capacity of heterogeneous networks. We demonstrate the value of offloading across multiple RATs in improving the video capacity. Controlling the offloading by optimizing an association bias across RATs enables rate improvements. Furthermore, considering co-channel HetNets, interference avoidance strategies become key to maintaining good QoE in the presence of offloading. In this setup, we show that joint offloading and resource partitioning improves the cell throughput.

"We demonstrate the value of offloading across multiple RATs in improving the video capacity."

"…greater strides have to be made by industry on integrating the Wi-Fi standard more tightly with cellular networks."

The proposed offloading and interference avoidance strategies are LTE standard compatible. In fact, the latest release of LTE allows the provision of almost blank subframes (ABSF) for interference avoidance and association bias for offloading in HCNs. The proposed joint interference and offloading, therefore, provides "plug and play" parameters for HCNs. However, to leverage offloading gains in multi-RAT HetNets, greater strides have to be made by industry on integrating the Wi-Fi standard more tightly with cellular networks. As penetration of Wi-Fi networks increases, users should be able to hand off seamlessly from and to cellular networks. As shown in the above research, substantial gains are feasible when video users are offloaded from cellular to Wi-Fi network using appropriate association bias.

Several remaining challenges and open problems relating to the proposed wireless network management policies should be addressed in the future work. Taking into account the mobility of the users while offloading is important. For example, offloading video users with high mobility to small cells may not be beneficial due to the costs associated with handoffs. Incorporating heterogeneous user QoS requirements in the network is another challenge. Users requesting delay-tolerant services can be more aggressively offloaded to Wi-Fi as compared to those requesting delay-stringent services. Incorporating these issues analytically into a comprehensive offloading framework can be quite challenging but can yield valuable insights into the operation of a more realistic network.

Overall, we have developed new insights into the problem of delivering perceptually relevant video to people using wireless communication systems. We developed new perceptual quality metrics and used these metrics to both drive and evaluate adaptation mechanisms for stored video and real-time video, each with different constraints. We used these ideas to provide insight into the how video traffic should be managed in wireless networks using HetNets and multiple RATs. Our approaches set the stage for meeting the requirements brought by video transmission over wireless systems as the endless march for higher quality video content like higher resolution and 3D continues.

References

[1] Wang, Z. and A. C. Bovik, "Mean squared error: Love it or leave it? a new look at signal fidelity measures," *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 98–117, 2009.

[2] Seshadrinathan, K., R. Soundararajan, A. C. Bovik, and L. K. Cormack, "LIVE Video Quality Database," http://live.ece.utexas.edu/research/quality/live video.html, 2012.

[3] Seshadrinathan, K., 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.

[4] Wang, Z., E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers*, vol. 2, Nov. 2003, pp. 1398–1402.

[5] Sheikh, H. R. and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, 2006.

[6] Dobrian, F., V. Sekar, A. Awan, I. Stoica, D. Joseph, A. Ganjam, J. Zhan, and H. Zhang, "Understanding the impact of video quality on user engagement," in *Proceedings of the ACM SIGCOMM 2011 Conference*, 2011, pp. 362–373.

[7] MPEG Requirements Group, "ISO/IEC FCD 23001-6 Part 6: Dynamics adaptive streaming over HTTP (DASH)," http://mpeg.chiariglione.org/working documents/mpeg-b/dash/dash-dis.zip, Jan. 2011.

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

[9] Soundararajan, R. 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, 2013.

[10] Mittal, A., A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, 2012.

[11] Mittal, A., R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, 2013.

[12] Saad, M. and A. C. Bovik, "Blind prediction of natural video quality," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp.1352–1365, 2014.

[13] Chen, C., 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," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 3602–3606.

[14] Chen, C., L. K. Choi, G. de Veciana, C. Caramanis, R. W. Heath Jr., and A. C. Bovik, "Modeling the time-varying subjective quality of http video streams with rate adaptations," *IEEE Trans. Image Process.*, vol. 23, no. 5, 2206–2221, May 2014.

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

[16] Joseph, V. and G. de Veciana, "NOVA: QoE-driven optimization of DASH-based video delivery in networks," in INFOCOM, pp. 82–90, April 27–May 2, 2014.

[17] Singh, S., O. Oyman, A. Papathanassiou, D. Chatterjee, and J. G. Andrews, "Video capacity and QoE enhancements over LTE," in IEEE ICC Workshop on Realizing Advanced Video Optimized Wireless Networks, Jun. 2012.

[18] Chen, C., X. Zhu, G. de Veciana, A. C. Bovik, and R. W. Heath Jr., "Rate adaptation and admission control for video transmission with subjective quality constraints," to appear in the IEEE J. Sel. Topics Signal Process. Previous version available at: http://arxiv.org/abs/1311.6453.

[19] Lu, Z. and G. de Veciana, "Optimizing stored video delivery for mobile networks: The value of knowing the future," in *INFOCOM*, 2013, pp. 2706–2714.

[20] Khalek, A. Abdel, C. Caramanis, and R. Heath Jr., "A cross-layer design for perceptual optimization of H.264/SVC with unequal error protection," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 7, pp. 1157–1171, 2012.

[21] Khalek, A. Abdel, C. Caramanis, and R. Heath Jr., "Loss visibility optimized real-time video transmission over MIMO systems," Available on arXiv http:// arxiv.org/ abs/1301.3174, Oct. 2013.

[22] Khalek, A. Abdel, C. Caramanis, and R. Heath Jr., "Delay-constrained video transmission: Quality-driven resource allocation and scheduling," to appear in the *IEEE Journal of Selected Topics in Signal Processing.* Previous version available on arXiv http:// arxiv.org/abs/1311.5921, Oct. 2013.

[23] Khalek, A. Abdel, C. Caramanis, and R. Heath Jr., "Video quality-maximizing resource allocation and scheduling with statistical delay guarantees," In *Proc. of IEEE GLOBECOM*, 2013.

[24] Khalek, A. Abdel, C. Caramanis, and R. Heath Jr., "Prioritized multimode precoding for joint minimization of source-channel video distortions," In *Proc. of Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, pp. 925–929, 2012.

[25] Khalek, A. Abdel, C. Caramanis, and R. Heath Jr., "Video-aware MIMO precoding with packet prioritization and unequal modulation," In *Proc. of the European Signal Processing Conference (EUSIPCO)*, Aug. 2012.

[26] Khalek, A. Abdel, C. Caramanis, and R. Heath Jr., "Online learning for quality-driven unequal protection of scalable video," In *Proc. of IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Sep. 2012.

[27] Khalek, A. Abdel, C. Caramanis, and R. Heath Jr., "Joint source-channel adaptation for perceptually optimized scalable video transmission," In *Proc. of IEEE Globecom*, Dec. 2011.

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

[29] Chen, C., R. W. Heath Jr., A. C. Bovik, and G. de Veciana, "A markov decision model for adaptive scheduling of stored scalable videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 6, pp. 1081–1095, 2013.

[30] Chen, C., R. W. Heath, A. C. Bovik, and G. de Veciana, "Adaptive policies for real-time video transmission: A markov decision process framework," in *18th IEEE International Conference on Image Processing (ICIP)*, 2011, pp. 2249–2252.

[31] Singh, S., H. S. Dhillon, and J. G. Andrews, "Offloading in heterogeneous networks: Modeling, analysis, and design insights," *IEEE Trans. Wireless Commun*., vol. 12, no. 5, pp. 2484–2497, May 2013.

[32] Singh, S. and J. G. Andrews, "Joint resource partitioning and offloading in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 2, pp. 888–901, Feb. 2014.

[33] Singh, S., F. Baccelli, and J. G. Andrews, "On association cells in random heterogeneous networks," *IEEE Wireless Commun. Lett.*, vol. 3, no. 1, pp. 70–73, Feb. 2014.

[34] Suchow, J. W. and G. A. Alvarez, "Motion silences awareness of visual change," *Current Biol.*, vol. 21, pp. 140–143, January 2011.

[35] Ruderman, D. L., "The statistics of natural images," *Network: Comp. Neural Syst'm*, vol. 5, pp. 517–548, 1994.

[36] Wang, Z., A. C. Bovik, and H. R. Sheikh, "Image quality assessment: From error visibility to structural similarity," 2004, available: https://ece.uwaterloo.ca/z70wang/research/ssim/.

[37] Pinson, M. H. and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcasting*, vol. 10, no. 3, pp. 312–322, September 2004.

[38] Seshadrinathan, K., R. Soundararajan, A. C. Bovik, and L. K. Cormack, "A subjective study to evaluate video quality assessment algorithms," *Proc. of SPIE Human Vision and Electronic Imaging*, January 2010.

[39] Bovik, A., "Automatic prediction of perceptual image and video quality," *Proceedings of the IEEE*, vol. 101, no. 9, pp. 2008–2024, 2013.

[40] Barkowsky, M., B. Eskofier, R. Bitto, J. Bialkowski, and A. Kaup, "Perceptually motivated spatial and temporal integration of pixel based video quality measures," in *Welcome to Mobile Content Quality of Experience*, Mar. 2007, pp. 1–7.

[41] Ninassi, A., O. Le Meur, P. Le Callet, and D. Barba, "Considering temporal variations of spatial visual distortions in video quality assessment," *IEEE J. Sel. Topics Signal Process*., vol. 3, no. 2, pp. 253–265, 2009.

[42] Yang, F., S. Wan, Q. Xie, and H. R. Wu, "No-reference quality assessment for networked video via primary analysis of bit stream," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 11, pp. 1544–1554, 2010.

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

[44] Park, J., K. Seshadrinathan, S. Lee, and A. Bovik, "Video quality pooling adaptive to perceptual distortion severity," *IEEE Trans. Image Process*., vol. 22, no. 2, pp. 610–620, 2013.

[45] Andrews, J. G., "Seven ways that HetNets are a cellular paradigm shift," *IEEE Communications Magazine*, pp. 136–44, March 2013.

[46] Astely, D., E. Dahlman, A. Furuskar, Y. Jading, M. Lindstrom, and S. Parkvall, "LTE: the evolution of mobile broadband," *IEEE Communications Magazine*, vol. 47, no. 4, pp. 44–51, 2009.

[47] Charrier, C., K. Knoblauch, L. Maloney, A. Bovik, and A. Moorthy, "Optimizing multiscale ssim for compression via mlds," *Image Processing, IEEE Transactions on*, vol. 21, no. 12, pp. 4682–4694, 2012.

[48] Maloney, L. T. and J. N. Yang, "Maximum likelihood difference scaling," *Journal of Vision*, no. 8, pp. 573–585, 2003.

Acknowledgments

This research was supported in part by Intel Corp. and Cisco Systems, Inc. under the VAWN program.

Author Biographies

Robert W. Heath Jr. (email: rheath@ece.utexas.edu) is a Cullen Trust Endowed Professor in the Electrical and Communications Engineering Department at the University of Austin, Texas, and Director of the Wireless Networking and Communications Group. He received a BS and an MS in electrical engineering from the University of Virginia and his PhD in electrical engineering from Stanford University. Dr. Heath is also the president and CEO of MIMO Wireless Inc. and chief innovation officer at Kuma Signals LLC. He is co-author of the textbook *Millimeter Wave Wireless Communications*, published by Prentice Hall in 2014. He is a registered Professional Engineer in Texas and a Fellow of the IEEE.

Alan C. Bovik (email: bovik@ece.utexas.edu) holds the Cockrell Family Regents Endowed Chair #4 at The University of Texas at Austin, is a professor in the Department of ECE and the Institute for Neurosciences, and Director of the Laboratory for Image and Video Engineering (LIVE). His papers and books on video processing and computational perception, including the *Handbook of Image and Video Processing* and *Modern Image Quality Assessment* have been cited over 35,000 times. A Thompson Reuters Highly Cited Researcher, he has received nearly every possible career or paper award in the field of video processing, including the 2013 IEEE Signal Processing Society "Society Award."

Gustavo de Veciana (email: gustavo@ece.utexas.edu) is a Cullen Trust Endowed Professor at the University of Texas at Austin. He received his PhD from the University of California, Berkeley in 1993. His research focuses on the analysis and design of wireless and wireline telecommunication networks, architectures and protocols to support sensing and pervasive computing, applied probability, and queueing theory. In 2009 he was designated IEEE Fellow for his contributions to the analysis and design of communication networks.

Constantine Caramanis (email: constantine@utexas.edu) has been on the faculty in the ECE department at the University of Texas at Austin since 2006. He received a PhD in EECS from the Massachusetts Institute of Technology in the Laboratory for Information and Decision Systems (LIDS) and an AB in Mathematics from Harvard University. He received the National Science Foundation CAREER award in 2011. His research interests focus on decision-making in large-scale complex systems, with a focus on learning and computation. Specifically, he is interested in robust and adaptable optimization, high dimensional statistics and machine learning, and applications to large-scale networks, including social networks, wireless networks, transportation networks, and energy networks.

Jeffrey Andrews (email: jandrews@ece.utexas.edu) received the BS in Engineering with High Distinction from Harvey Mudd College, and an MS and PhD in Electrical Engineering from Stanford University. He is the Cullen Trust Endowed Professor (#1) of ECE at the University of Texas at Austin,

Editor-in-Chief of the *IEEE Transactions on Wireless Communications*, and Technical Program Co-Chair of IEEE Globecom 2014. He developed Code Division Multiple Access systems at Qualcomm from 1995 to 1997, and has consulted for entities including Verizon, the WiMAX Forum, Intel, Microsoft, Apple, Samsung, Clearwire, Sprint, and NASA. He is co-author of the books *Fundamentals of WiMAX* (Prentice-Hall, 2007) and *Fundamentals of LTE* (Prentice-Hall, 2010). Dr. Andrews received the National Science Foundation CAREER award in 2007 and has been co-author of ten best paper award recipients. He is an IEEE Fellow and an elected member of the Board of Governors of the IEEE Information Theory Society.

Chao Chen (email: chao.chen@utexas.edu) received his BE and MS in electrical engineering from Tsinghua University in 2006 and 2009, respectively. In 2009, he joined the Wireless Systems Innovation Laboratory (WSIL) and the Laboratory for Image & Video engineering (LIVE) at the University of Texas at Austin, where he earned his PhD in 2013. Since 2014, he has been working in Qualcomm Incorporated in San Diego. His research interests include visual quality assessment, system identification, and network resource allocation.

Michele Saad (email: michele.saad@gmail.com) is a senior engineer and researcher in perceptual image and video quality assessment at Intel. She received her PhD in electrical and computer engineering from the University of Texas at Austin in 2013, BE in computer and communications engineering from the American University of Beirut, Lebanon, in 2007, and MS in electrical and computer engineering from the University of Texas at Austin in 2009. Her research interests include statistical modeling of images and videos, motion perception, design of perceptual image and video quality assessment algorithms, and statistical data analysis and mining and machine learning.

Zheng Lu (email: zhenglu@utexas.edu) is a PhD candidate in the Department of Electrical and Computer Engineering at the University of Texas at Austin. He received his BS in electronic engineering from Tsinghua University in 2009 and his MS in electrical and computer engineering from the University of Texas at Austin in 2011. He interned at Intel Labs in summer, 2013. He is currently working on video delivery in wireless networks and device-to-device communications.

Amin Abdel Khalek (email:akhalek@utexas.edu)received a BE in computer and communication engineering from Notre Dame University in 2008, an ME in electrical and computer engineering from the American University of Beirut in 2010, and a PhD in electrical and computer engineering from the University of Texas at Austin in 2013 with focus on perceptual video optimization. In 2011 and 2012, he interned at Intel Corporation where he was actively involved in developing an end-to-end cross-layer video optimization prototype. In 2013, he interned at Samsung Telecommunications America where he worked on developing precoding algorithms for mmWave antenna systems. In 2014, he became a senior design engineer at Freescale Semiconductor.

Sarabjot Singh (email: sarabjot@utexas.edu)received a B.Tech. in ECE from IIT Guwahati, India, and was awarded the President of India Gold Medal 2010. He is currently a PhD candidate at the University of Texas at Austin, where his research focuses on the modeling, analysis, and design of offloading in wireless heterogeneous networks and self-backhauled millimeter wave networks. His other research interests include RF-localization in indoor networks, scheduling for video streaming in LTE-Advanced, and interference coordination in wireless networks. His paper on multi-RAT offloading received the best paper award at IEEE ICC 2013. His industrial experience includes internships at Alcatel-Lucent Bell Labs in Crawford Hill, New Jersey, at Intel Corporation in Santa Clara, California, and at Qualcomm Inc. in San Diego, California.

CACHING AND CROSS-LAYER DESIGN FOR ENHANCED VIDEO PERFORMANCE

Contributors

Hasti Ahlehagh
University of California, San Diego

Laura Toni
Ecole Polytechnique Federale de Lausanne

Dawei Wang
Adaptive Spectrum & Signal Alignment

Pamela Cosman
University of California, San Diego

Sujit Dey
University of California, San Diego

Laurence Milstein
University of California, San Diego

This article presents a summary of research we have performed on the use of two complementary techniques designed to enhance the capacity and/or the quality of video transmissions over mobile channels. The first technique makes use of a combination of caching and scheduling video sources so as to make them more readily available for wireless transmission to a mobile user. The second technique attempts to more efficiently utilize the wireless channel by taking advantage of cross-layer optimization techniques between the application layer and the physical layer so that, for example, the performance of video transmissions over a given bandwidth is maximized. For both research topics, we quantify the performance gains that can be achieved.

Introduction

Due to the growth in the adoption of smartphones and tablets as well as the increase in popularity of high quality over-the-top (OTT) Internet video, cellular operators have experienced a tremendous rise in data traffic.[1] The growth in video traffic affects many parts of the operator's network. When Internet video is accessed by a mobile device, the video must be fetched from the servers of a content delivery network (CDN) and traverse the mobile carrier core network (CN), radio access network (RAN), and wireless channel to reach the mobile device. While advances in radio technologies and architectures such as LTE, LTE Advanced, small cells, and HetNets will lead to significant increases in wireless channel capacities, they will also exacerbate the capacity challenge and congestion problem in the RAN backhaul. According to a Strategy Analytics report[2], there will be potentially a 16-petabyte shortfall in carriers' CN and RAN backhaul capacity by 2017. Moreover, a serious concern for mobile operators is that the increase in the radio access capacity will be overtaken quickly by the rapid growth in the quality demands and volume of mobile video consumption. In this article, we present the results of our research designed to enhance the performance of video transmission.

The article is divided into two key sections, caching of video and cross-layer design. These two techniques complement one another, in that the former technique applies to the backhaul of a mobile video system and the latter one applies to the links between the mobile units and the base stations. Both of these techniques will be shown to yield sizable gains in performance and capacity.

RAN Caching, Processing, and Scheduling to Enhance Video Capacity and User Experience

The mismatch between end-to-end network capacity and explosive growth in demand for video bandwidth has led to two key negative impacts on user quality of experience (QoE): the initial delay to start a video session, and stalling (buffering) during a video session. Hence, in this work, we address the dual challenge faced by mobile operators—increasing end-to-end mobile video capacity (number of concurrent video sessions) while satisfying expected QoE (acceptable initial delay and probability of stalling)—using alternative techniques beyond increasing backhaul and access capacity.

We address the dual challenge by proposing a mobile video cloud (MVC), shown in Figure 1. In a MVC, video caching and processing is performed in

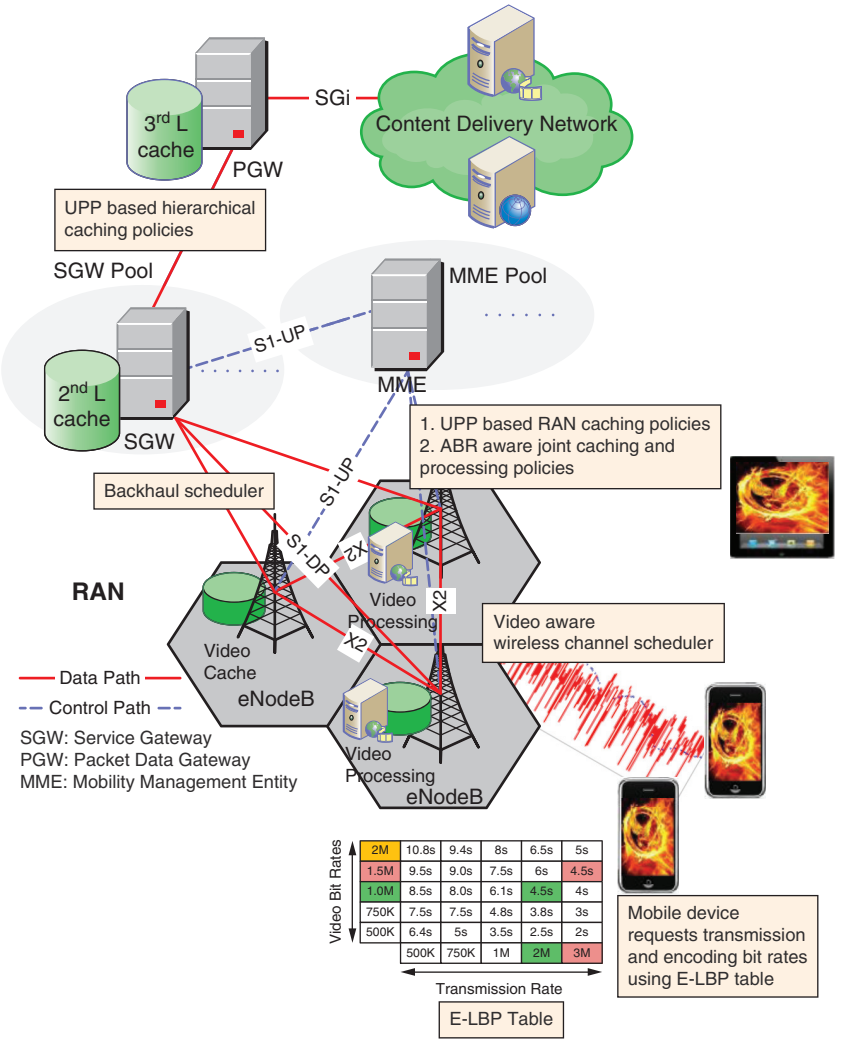


Figure 1: Mobile video cloud for a 4G cellular wireless network, consisting of RAN and hierarchical caching and processing, and video-aware scheduling algorithms (Source: University of California, San Diego, 2014)

a massively distributed way at the base stations (eNodeBs) at the edges of the RAN, and hierarchically in the CN. Getting videos from MVC caches instead of having to fetch videos from Internet CDNs can not only significantly reduce RAN backhaul bandwidth, but can also reduce delay. The latter is also substantiated by our experimentations with 3.5G and 4G mobile networks, which have shown that the probability of achieving low delays increases substantially when content is fetched from mobile CNs as opposed to Internet cloud servers.[3] To ensure effectiveness for the relatively small-sized MVC caches, RAN-aware reactive and proactive caching policies have been developed that utilize user preference profiles (UPPs) of active users in each cell.[3][4][5]

"…wireless channel conditions may prevent a video found in the RAN cache from being delivered to the mobile device in a way that satisfies its QoE requirement."

However, even with the increased bandwidth expected from LTE, wireless channel conditions may prevent a video found in the RAN cache from being delivered to the mobile device in a way that satisfies its QoE requirement. Moreover, between having to fetch videos from Internet CDNs that could not be found in the RAN cache, and videos that may have to be proactively fetched for the RAN caches, the backhaul may get congested. To address this, backhaul and wireless channel scheduling techniques have been developed[3][4] that make novel use of video properties encapsulated by leaky bucket parameters (LBPs)[6] of the videos. Video-aware scheduling using LBPs, together with MVC caching, can maximize the number of concurrent video sessions that can be supported by the end-to-end network while satisfying their delay requirements and minimizing stalling.

To support the growing trend towards adaptive bit rate (ABR) streaming, without having to cache all bit rate versions of all segments of ABR videos, joint RAN caching and processing techniques have been developed that utilize the given backhaul, caching, and processing resources most effectively to maximize video capacity while preserving the low stalling benefit of ABR.[7][8] Video Quality Metric (VQM) is a common perceptually based metric (ranging from zero equals best to one equals worst) for quantifying video quality. Besides initial delay and stalling, we also consider the delivered video bit rate expressed in terms of VQM.

We have developed a statistical Monte Carlo discrete event simulation framework using MATLAB to compare the relative performance of the caching policies proposed in this project, as well as to validate the effectiveness of the proposed scheduling techniques. Using our simulation framework, we have conducted extensive experiments under various user dynamics, cache size, processing capacity, and wireless channel conditions. We will present our key findings, demonstrating the significant capacity and user experience benefits of our proposed MVC architecture and algorithms over conventional CDN techniques and/or ABR streaming alone. Finally, we will present our ongoing research and preliminary results on extending MVC to caching at the mobile devices themselves, working in conjunction with edge caching and processing to enhance the end-to-end capacity and user experience of mobile video delivery.

Next, we briefly describe our research on RAN caches and caching policies, hierarchical caching policies, and scheduling techniques with and without ABR, and present significant results that we obtained using simulation.

Edge Caching and Video-Aware Scheduling

As shown in Figure 1, we proposed video caching at the very edge of the mobile network, with microcaches at base stations (and access nodes of Wi-Fi* hot spots and small cells in HetNets) to address the problem of backhaul congestion and video buffering delay. However, since the proposed approach will lead to thousands of caches, we need to use much smaller sized "microcaches" for RAN caching to keep the overall cost down, as opposed to the large-sized conventional CDN caches. As demonstrated by our research[3][4], conventional video caching policies may not be able to achieve a high cache hit ratio for the RAN microcaches. To address this challenge, and motivated by our empirical studies that show users may have strong preferences for videos belonging to specific video categories (VCs), we developed caching policies that use the video preferences of active users of the RAN cell. We associate a user preference profile (UPP) with each user, which specifies the probabilities that the user will request videos of specific VCs, and can be obtained by tracking the fraction of videos from each VC the user has watched in the past. Using the UPPs of the active users of a cell, we identify the video categories and consequently videos that are most likely and least likely to be requested by the current cell users. We propose two new caching policies: P-UPP (proactive UPP), which proactively caches a subset of the most likely requested videos when there is a change in the active video users in a cell, and R-UPP (reactive UPP), which replaces the least likely requested videos in order to cache the currently requested video in case of a cache miss.[3][4]

"…we developed caching policies that use the video preferences of active users of the RAN cell."

Furthermore, for videos that result in cache misses and need to be fetched from Internet CDNs, we developed a video scheduling approach that allocates the RAN backhaul resources to the video requests so as to reduce video latency and increase video capacity (number of concurrent video sessions that can be scheduled). For all videos that are downloaded either from the RAN microcaches or scheduled successfully through the RAN backhaul, we proposed a video-aware wireless channel scheduler that works with any RAN scheduler as a plugin to increase the number of videos that can be transmitted concurrently through the wireless channel to the requesting mobile devices. Our scheduling techniques use LBP tuples associated with each video to allocate the minimum download rate, R_{min}, needed to meet a desired initial delay without stalling during the video session. Using an optimization formulation, both the RAN backhaul and the wireless channel scheduler allocate the remaining bandwidth among the ongoing video downloads to utilize the spare capacity in order to finish downloads faster and therefore free up bandwidth for future peak demands or to improve QoE beyond desired levels. The backhaul scheduler schedules requested videos according to their minimum rate requirement and transfers the video bits to the mobile device's base station buffer. The wireless channel scheduler assumes an LTE system and allocates both power and bandwidth (subcarriers) to transfer video bits pending in the base station buffer so as to ensure each user's R_{min} according to its LBP. Our proposed power and bandwidth allocation scheme consists of two phases: the first phase is to attempt to assign enough subcarriers to satisfy R_{min} of each user, assuming equal power assignment per subcarrier and starting with

"…we proposed a video-aware wireless channel scheduler that works with any RAN scheduler as a plugin…"

the video request that has the best channel condition. The second phase is to reallocate power, first to ensure R_{min} of each user that was scheduled (allocated subcarriers in the first step), and then using waterfilling to assign the remaining power optimally to the users that were scheduled. Note that due to the heavy traffic load and/or wireless channel conditions, it may not be possible for the scheduler to satisfy a user's R_{min} leading to blocking the user's video request. Moreover, even when a video session has been scheduled, variations in the wireless channel may lead to periods during the session when the video's R_{min} cannot be sustained, which may lead to stalling. Our earlier work[3] provides details of the joint backhaul and wireless channel scheduler.

Our simulation results show that RAN microcaches with the UPP-based caching policies can achieve significantly higher cache hit ratios compared to when no RAN cache is used, and when RAN cache with existing cache policies such as Most Popular Videos (MPV), Least Recently Used (LRU), and Least (Most) Frequently Used (LFU/MFU) are used. As shown in Figure 2(a), for example,

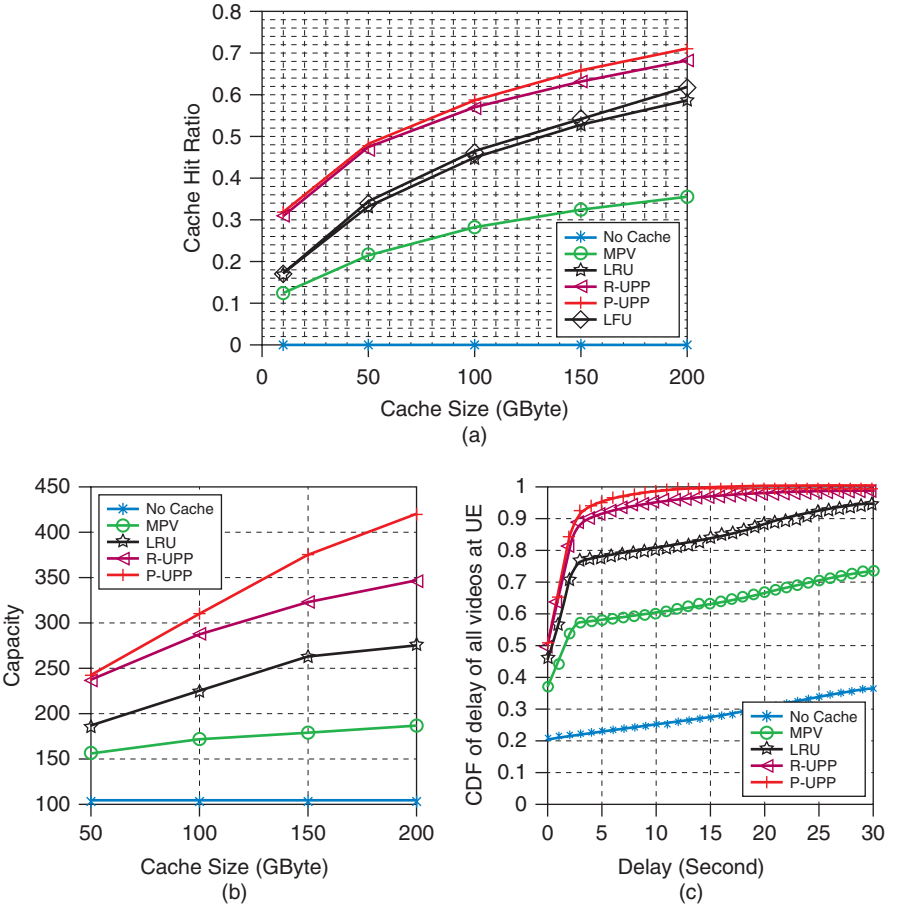


Figure 2: Performance of the caching policies: (a) Cache hit ratio vs. cache size, (b) Capacity vs. cache size, (c) CDF of delay of videos at mobile device (Source: Ahlehagh, H. and S. Dey, "Video aware scheduling and caching in the radio access network," *IEEE Transactions on Networking,* vol. 22, no.5, August 2014. (Copyright IEEE))

when the cache size is 200 GB, P-UPP and R-UPP achieve cache hit ratios of 0.71 and 0.68 respectively, while the LFU, LRU, and MPV policies achieve cache hit ratios of 0.61, 0.58, and 0.35 respectively. As shown in Figure 2(b), together with video-aware backhaul scheduling, UPP-based RAN caching can improve video capacity (number of concurrent video requests served) by up to 300 percent compared to having no RAN caches, and by more than 50 percent compared to RAN caches using LRU. From Figure 2(c), we infer that the probability of achieving an initial delay of 5 seconds or less is about 0.23 with no RAN cache, 0.58 for MPV, 0.77 for LRU, 0.91 for R-UPP, and 0.95 for P-UPP. These results show that using UPP-based RAN caches can greatly improve the probability that video requests can meet initial delay requirements. In networks where the wireless channel bandwidth may be constrained, additional experiments have shown that the application of our video-aware wireless channel scheduler resulted in significantly (up to 250 percent) more video capacity using both Urban Macro (UMa) and Urban Micro (UMi) channel models[3] with very low stalling probability. For instance, the probability of a stall with duration of 10 seconds or higher is almost 0 for P-UPP with our video-aware wireless channel scheduler, while it is around 0.03 for the P-UPP caching policy without our video-aware wireless channel scheduler.

"…UPP-based RAN caching can improve video capacity…"

We also compared the UPP-based policies with conventional caching policies under various simulation conditions defined by varying key parameters such as video popularity ranking (Zipf distribution), user dynamics (that is, mean user inter-arrival and departure rates), and user UPP distribution (how biased the user requests are towards specific video categories). Our simulation results show that even under more challenging conditions such as high Zipf values (implying a fatter tail of the Zipf distribution), higher user dynamics and more uniform UPP distributions than used in the base simulation scenario, the impact on the performance of UPP-based RAN caching is marginal, with continued superior performance compared to no RAN caching, or RAN caching with LFU, LRU, or MPV policies. For example with high user dynamics (user inter-arrival time reduced from 40 seconds to 10 seconds and mean user active time from 2700 seconds to 360 seconds), when the cache size is 200 GB, P-UPP and R-UPP achieve cache hit ratios of 0.68 and 0.65 respectively, marginally lower than the base case, but much higher than the 0.61, 0.58, and 0.35 achieved, respectively, by LFU, LRU, and MPV policies.

Hierarchical Caching

To further improve the capacity and video QoE of the cellular networks, we investigated supplementing the RAN caches with a hierarchical caching scheme, where the gateways in the CN also have video caches. Figure 1 shows our proposed hierarchical caching architecture with caches at SGWs and PGWs of the 4G cellular networks. The hierarchical caching approach can further improve network capacity by enabling multiple cell sites to share caches at higher levels of the hierarchy, thereby improving overall cache hit ratio, without increasing the total cache size used. Hierarchical caching can also help to accommodate mobility; for example, when a user with an active video session moves from one cell to a neighboring cell connected to the same PGW,

"The hierarchical caching approach can further improve network capacity…"

with proper caching of videos in the CN caches, it may become more likely that the video currently being downloaded can be found in the PGW or SGW associated with the new cell.

However, the goals of improving cache hit ratio of users in a given cell and users with mobility across cells can be conflicting, with the amount of video redundancy in the caches of different layers of the hierarchy impacting them differently. A hierarchical caching policy that will result in a higher layer cache including more (versus fewer) videos already existing in the associated lower layer caches will be more effective at increasing cache hit ratio of mobile users (versus static users). Hence, we proposed a hybrid and partially distributed hierarchical caching policy to increase cache hit ratio and provide support for mobility across cells while still aiming to improve the coverage of static users. We also extended our UPP-based caching policies to accommodate the hierarchical caching structure and policy. For instance, when applying R-UPP to hierarchical caching, if the request to the first layer cache (at the base station) results in a cache miss, the request is progressively passed to the next layer in the cache hierarchy until either there is a cache hit or it has reached the root of the tree (the Internet CDN). While the fetched video is traversing the cache hierarchy downward, each cache on the way to the mobile device decides whether to cache the video using the proposed caching algorithm. For more details on hierarchical R-UPP and P-UPP caching policies, refer to our earlier work.[5]

"…we proposed a hybrid and partially distributed hierarchical caching policy to increase cache hit ratio…"

Our simulation results show that using hierarchical caching, with realistic cache sizes, and assuming no wireless channel bandwidth constraints, can enhance network capacity by up to 30 percent compared to caching only in the RAN, with the same total cache size. In the case of mobility, we observe that UPP-based hierarchical policies perform significantly better than RAN-only caches: hierarchical R-UPP results in 47 percent higher capacity than the RAN-only R-UPP. Furthermore, hierarchical R-UPP caching increases capacity by 68 percent compared to hierarchical LRU caching policies. Thus, we infer that significant capacity gains are also observed in cases with user mobility when using UPP-based hierarchical caching.

ABR-aware RAN Caching

ABR streaming has become a popular video delivery technique credited with improving the quality of videos over wireless networks because of its ability to adapt to changing channel conditions. We investigated the opportunities and challenges of combining the advantages of ABR streaming with RAN caching to further increase the video capacity and QoE of wireless networks. Since with ABR, each video is divided into multiple segments that can be requested at different bit rates, a cache hit will require not only the presence of a specific segment but also at the desired bit rate, making ABR-aware RAN caching challenging. To address this without having to cache all bit rate versions of all segments of a video, we add limited processing capacity (Figure 1) to each base station to enable transrating a cached higher rate version to satisfy a request for a lower rate version, thus avoiding the need to cache all bit rates or having to fetch over the backhaul all missing rate versions.

"We investigated the advantages of ABR streaming with RAN caching to further increase the video capacity and QoE…"

Figure 3 explains the overall approach for our joint caching and processing policies. A mobile device requests video 1 with the second highest bit rate, V_{12}. There is an instance of the video in the cache with the desired bit rate for the first second, however, video chunks that correspond to 1–4 seconds are not in the cache and chunks corresponding to 4–5 seconds are cached in the third available rate (lower bit rate than V_{12}). Thus, the video chunks from 1–5 seconds need to be brought in from the backhaul. The remainder of the video chunks exist in the cache with the first (highest) available bit rate (5–8 seconds of V_{11}), so we can use the processing resource to transrate the video bit rate to the desired rate or use the backhaul resource to bring the video chunks. For the last option, either an exact bit rate version can be fetched, or a higher rate version can be fetched and cached, which would allow future requests of lower rate versions to be satisfied using transrating. We developed a joint caching and processing resource allocation algorithm, which, given the available cache size, processing capability, and backhaul bandwidth, selects among available options so as to increase the number of ABR video requests that can be satisfied concurrently.

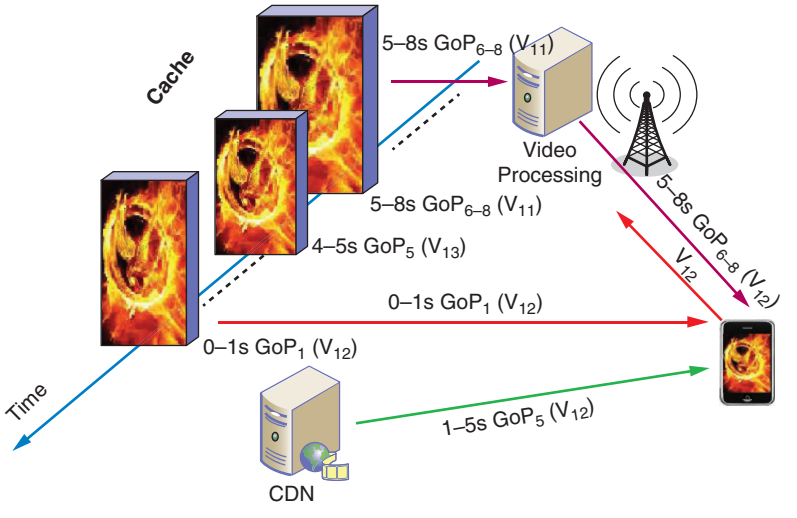


Figure 3: Satisfying ABR video request in proposed ABR-capable RAN caching and processing framework
(Source: Ahlehagh, H. and S. Dey, "Adaptive bit rate capable video caching and scheduling," *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC),* April 2013. (Copyright IEEE))

Our contributions for ABR-aware RAN caching[7][8] are summarized here as follows: (a) addressed the caching challenges imposed by ABR streaming by proposing proactive and reactive ABR-capable caching policies that leverage the processing and cache resources available to increase cache hit ratio, using the resource allocation algorithm above; (b) ABR-aware UPP-based (ABR-P-UPP-P) caching policy that uses a new video bit rate prediction algorithm to proactively cache not only the most likely requested videos by the users of a cell according to their UPPs, but also at the most likely requested rates, depending on the network load and wireless channel conditions; (c) extended

LBP table (E-LBP shown in Figure 1) to consider the additional flexibility of multiple encoding bit rate versions available and developing a new rate adaptation algorithm that runs on the mobile client and uses E-LBP of the requested videos to improve capacity and QoE by adjusting both the video encoding and transmission rates.

Using our MATLAB statistical simulation framework, we conducted extensive experiments under various cache sizes, processing capacities, user distributions, and wireless channel conditions. Table 1 shows the results for the case when the cache size is 150 Gb, transrating capacity is 12 Mbps, users are uniformly distributed across the cell, and the channel exhibits Rayleigh fading with a Doppler frequency of 3 Hz.[8] Table 1 shows that significant gain in end-to-end video capacity of a cellular network can be obtained using ABR-capable RAN caching: up to about 150 percent compared with no RAN caching and no ABR, 66 percent compared to using RAN caching alone, 100 percent compared to using ABR alone, and 23 percent compared to the most effective of the alternate ways we evaluated to enable ABR-aware RAN caching. As shown in Table 1, the above video capacity gains can be achieved while mostly improving the video quality as measured by VQM and stalling probability.

ABR / Caching Policy	Capacity	Probability of Stalling	VQM
1. No ABR, No RAN Cache	99	0.010	1
2. No ABR, RAN Cache [LRU]	148	0.012	1
3. ABR, No RAN Cache	120	0.0002	0.89
4. ABR, RAN Cache [ABR-LRU-P]	208	0.0041	0.88
5. ABR, RAN Cache [Static LRU]	101	0.0114	0.77
6. ABR, RAN Cache [ABR-P-UPP-P]	245	0.0075	0.903
7. ABR, RAN Cache [Highest Rate LRU]	200	0.0072	0.801

Table 1: Video Capacity, Stall Probability, VQM
(Source: University of California, San Diego, 2014)

Cross-Layer Design for Mobile Video Transmission

The research results summarized in this part of the article are typical of a broader set of results that aim to optimize physical/application cross-layer designs for either real-time or archival video, where the channel is doubly selective (that is, exhibits both time and frequency selectivity). 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.

The key goal of the research is to develop improved cross-layer designs for transmission of video waveforms over mobile channels by explicitly accounting for the nonstationary channel statistics inherent in virtually any scenario where there is relative motion between the transmitter and the receiver.[9][10] However, many communications models are inaccurate when the Doppler spread can vary over the duration of a video. For example, the ergodic capacity is probably the most commonly quoted result for the capacity of a rapidly fading channel, but a typical derivation of ergodic capacity assumes that the channel is both stationary and ergodic.[11] However, if the relative velocity between the transmitter and the receiver changes with time, the coherence time changes, and thus the autocorrelation function of the channel gain at any two points depends on the actual instants when the correlation is evaluated. In other words, the channel is not stationary. Even if velocity is constant, relative motion between transmitter and receiver means path loss is a function of time, and most likely the shadowing is a function of time. Thus, unless there is perfect power control, the channel statistics are again nonstationary. Similarly, the coherence bandwidth in most of the literature is taken to be constant. However, since the geometry of the multipath reflections changes over time in the presence of mobility, so does the coherence bandwidth.[9]

"The key goal of the research is to develop improved cross-layer designs for transmission of video waveforms over mobile channels…"

Stated more succinctly, for a stationary channel, the key statistical parameters relevant to communications performance, such as coherence time and coherence bandwidth, are constant, so the system parameters that they influence, such as cyclic prefix length for an OFDM system, or interleaver depth for forward error correction, or pilot spacing for channel estimation, are also constant. For nonstationary channels, the statistical parameters are themselves functions of time, and the appropriate system design becomes more complex.

In particular, much of the literature, while acknowledging that Doppler spread is directly proportional to relative velocity, ends up by quantifying the Doppler spread as a function of the speed of a given vehicle, rather than the relative velocity of the vehicle. This arises most often because scenarios in the literature typically correspond to relative motion between the transmitter and the receiver along a straight line correcting the two, in which case the magnitude of the relative velocity does not vary with time. However, in other cases, the distinction between speed being a scalar and velocity being a vector is ignored, so that a vehicle moving at constant speed is assigned a constant Doppler spread, even though a constant-speed vehicle almost always has a time-varying relative velocity associated with it.

For example, consider the diagram shown in Figure 4, where a moving vehicle is traveling in a vertical direction at a constant speed of s mph, moving from point (x_0,y_0) to point $(x_0,y(t))$. This vehicle communicates with a fixed base station located at the origin, and the magnitude of the relative velocity varies from zero to s. Indeed, this time-varying nature of the relative velocity, and hence of the coherence time, will be true for virtually any path and any speed (constant or otherwise). An obvious exception corresponds to a constant speed vehicle moving along a radial line from the origin, so that if $x_0 = 0$ in Figure 4, the magnitude of the relative velocity is constant.

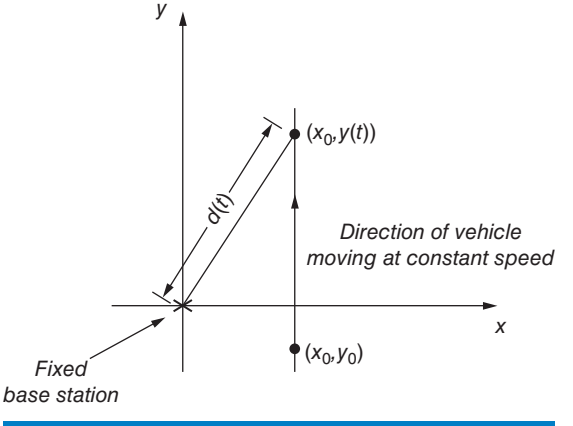


Figure 4: Vehicle traveling at fixed speed, communicating with base station
(Source: University of California, San Diego, 2014)

"…we believe that point optimization should not be the goal; rather the design should aim for robustness…"

Regarding performance, another departure from much of the literature that is basic to this article is the criterion used to design and evaluate system performance. Most analyses use a point optimization as the criterion, meaning for a fixed set of system parameters (including channel parameters), the system performance is optimized on the basis of some objective function, such as minimizing average received video distortion. However, in the presence of arbitrary mobility, we believe that point optimization should not be the goal; rather the design should aim for robustness, meaning the design should yield acceptable performance over a wide range of Doppler spreads and multipath delay spreads. So the second goal of this proposal is to develop understanding and tools for this type of performance evaluation, as well as specific results from optimizing video transmission systems under this approach.

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

Resource Allocation for Video Transmission over Doubly Selective Fading Channels

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 an orthogonal multicarrier waveform 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 (MSEs) across all users, the base station collects the distortion-rate (DR) information for every user. For each bitstream, the most important video information (such as motion vectors and macroblock IDs) 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 5 and 6, the performance of the cross-layer algorithm, described in detail in our earlier work[12][13], is presented. This algorithm is optimized by

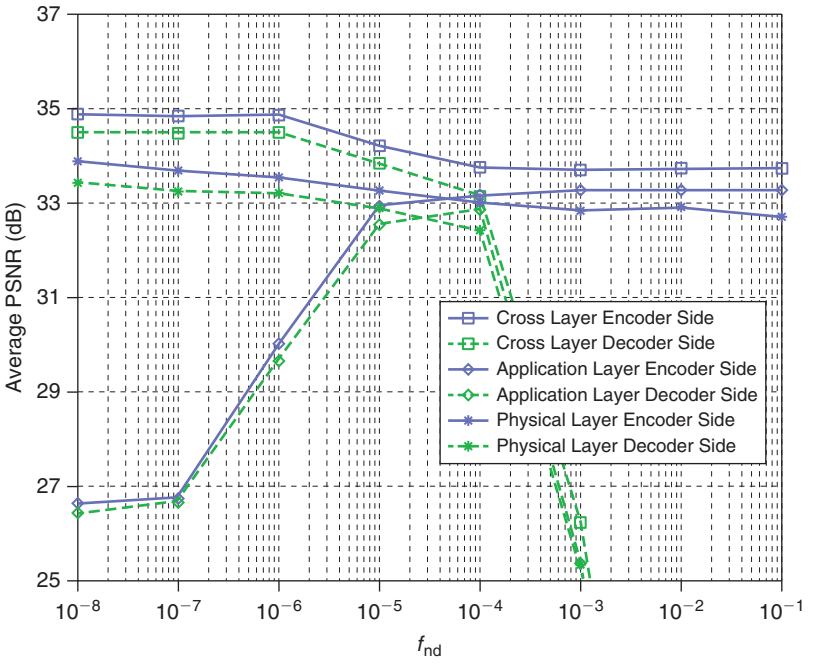


Figure 5: Ls = 100, 16 subcarriers, PSNR versus normalized Doppler spread (Source: University of California, San Diego, 2014)

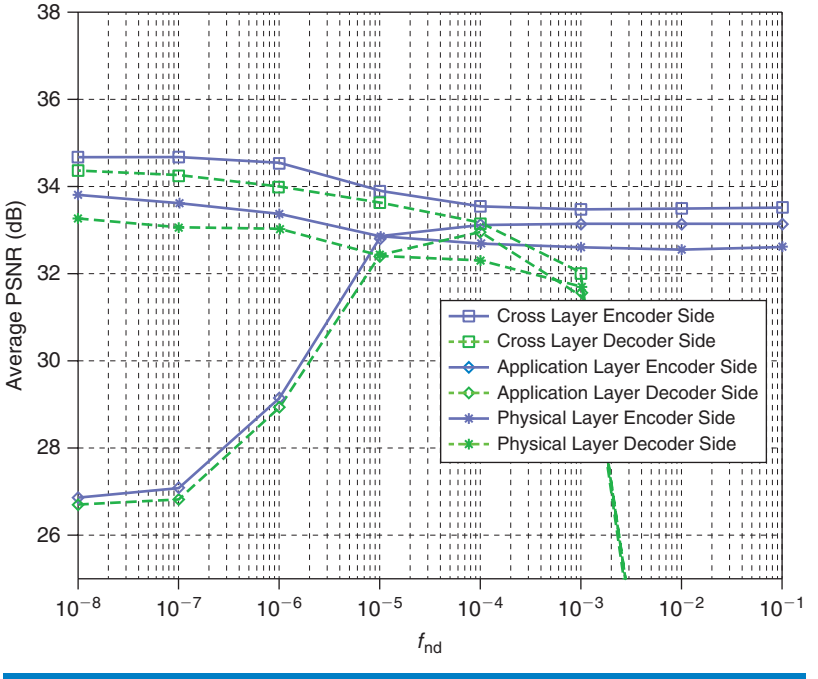


Figure 6: Ls = 25, 16 subcarriers, PSNR versus normalized Doppler spread (Source: University of California, San Diego, 2014)

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 5 and 6, one of which uses just the CSI for the resource allocation (that is, it makes no use of the application-layer information), and the other one of which uses only the DR curves for the resource allocation (that is, it makes no use of the physical-layer information).

Consider Figure 5, which corresponds to three video users competing for bandwidth in increments of OFDM subcarriers, of which there are 16. The ordinate of Figure 5 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.

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."

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 Figure 6, which is the same as Figure 5 except that in Figure 5, the pilot spacing is 100 symbols, and in Figure 6, the pilot spacing is 25 symbols. All three systems can now function properly at Doppler spreads that are about an order of magnitude larger than in the system of Figure 5. Note that if the system has the ability to track the Doppler spread, one can adapt the spacing between pilot tones

by increasing the spacing for large coherence times and decreasing the spacing for small Doppler. In this way, the overhead penalty can be significantly diminished.

Mapping of Video Slices to OFDM Subcarriers for Low-Mobility Users

The research described in this subsection differs in several key ways from that presented in the previous one. The model now under consideration corresponds to single-user transmission at low mobility. Whereas one of the key goals of the preceding subsection was to demonstrate the robust nature of cross-layer design for users moving at arbitrary levels of mobility, in this subsection the key goal is to demonstrate robustness to channel gain for low mobility users.

We considered transmission of nonscalably encoded video sequences over an OFDM system in a slowly varying Rayleigh faded environment. The OFDM waveform consists of N_t subcarriers that 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.

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

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

"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."

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…"

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 earlier work.[14] In Figure 7, the best VQM

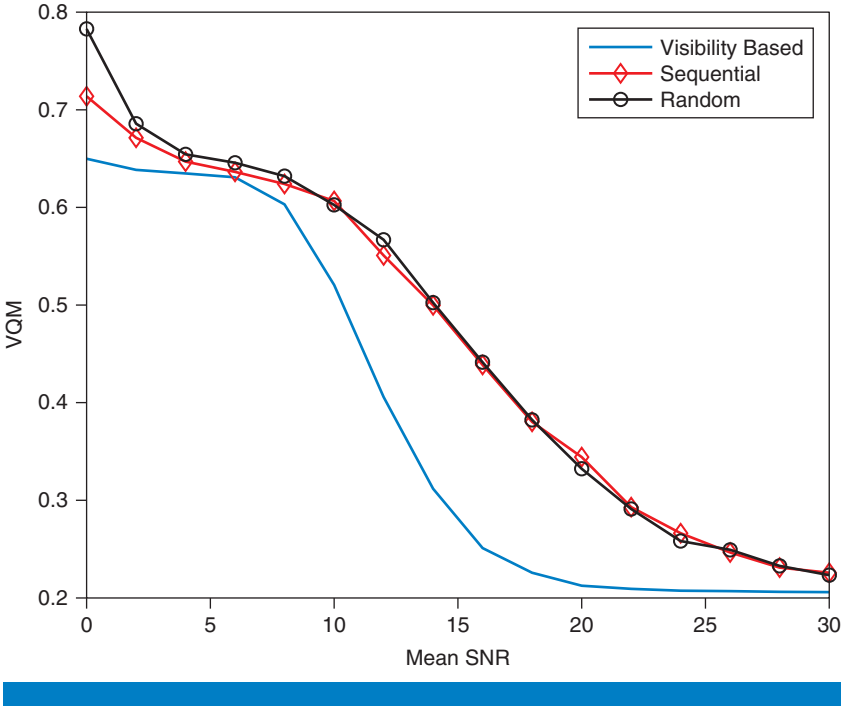


Figure 7: VQM vs. mean SNR for both visibility-based and baseline algorithms for systems with (*N*, *M*) = (32,4)
(Source: L. Toni, P.C. Cosman, and L.B. Milstein, "Channel Coding Optimization Based on Slice Visibility for Transmission of Compressed Video over OFDM Channels," *IEEE Journal on Selected Areas in Communications*, Vol. 30, No. 7, August 2012. (Copyright IEEE))

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, 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 (that is, 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 0.28 in VQM score (for $\gamma = 13$ dB).

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. Figure 8 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 two independent channels are considered, as shown in Figure 9, as expected, due to the limited opportunity for time diversity offered by the channel, all the algorithms lead to almost the same performance.

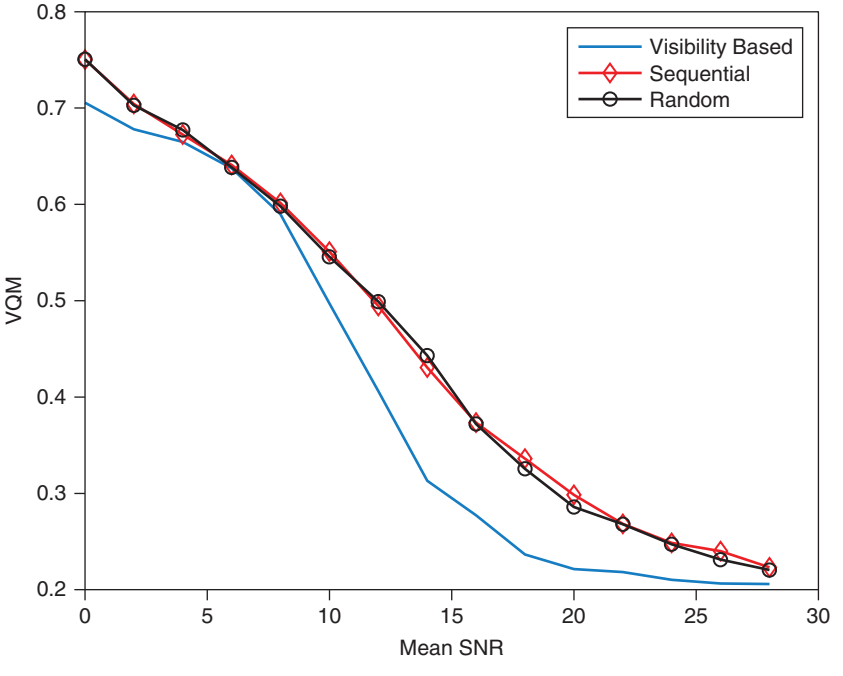


Figure 8: VQM vs. mean SNR for both visibility-based and baseline algorithms for systems with $(N, M) = (8, 16)$ (Source: L. Toni, P.C. Cosman, and L.B. Milstein, "Channel Coding Optimization Based on Slice Visibility for Transmission of Compressed Video over OFDM Channels," *IEEE Journal on Selected Areas in Communications*, Vol. 30, No. 7, August 2012. (Copyright IEEE))

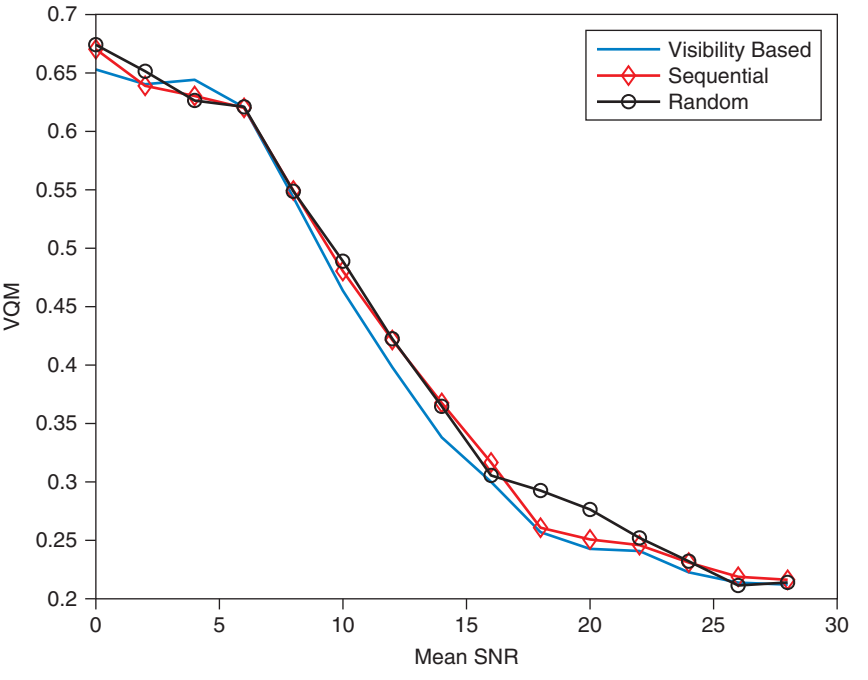


Figure 9: VQM vs. mean SNR for both visibility-based and baseline algorithms for systems with (N, M) = (2, 64)
(Source: L. Toni, P.C. Cosman, and L.B. Milstein, "Channel Coding Optimization Based on Slice Visibility for Transmission of Compressed Video over OFDM Channels," *IEEE Journal on Selected Areas in Communications*, Vol. 30, No. 7, August 2012. (Copyright IEEE))

Conclusions and Future Work

"The caching research demonstrated that very significant gains in end-to-end mobile video capacity can be obtained while meeting key user experience metrics…"

The caching research demonstrated that very significant gains in end-to-end mobile video capacity can be obtained while meeting key user experience metrics of low initial delay and stalling, by caching at the access nodes and gateways of mobile networks, using RAN user-centric caching policies, and coordinated video-aware backhaul and wireless channel scheduling. Our results show that RAN caching with user-preference–based caching policies, together with video-aware backhaul scheduling, can improve video capacity by up to 300 percent compared to having no RAN caches, and by more than 50 percent compared to RAN caches using conventional caching policies like LRU. Furthermore, we showed that using hierarchical caching along with user-centric caching policies can enhance network capacity by up to 30 percent compared to caching only in the RAN given the same cache size. In the case of mobility, we observe that UPP-based hierarchical policies perform significantly better than RAN-only caches: for example, hierarchical R-UPP results in 47 percent higher capacity than the RAN-only R-UPP. When ABR content is available, our research has shown that the addition of limited processing resources at the access nodes, together with the proposed ABR-aware joint caching and processing policies, can significantly increase end-to-end mobile video capacity, while preserving the low stalling benefits of ABR: by up to

150 percent compared with no RAN caching and no ABR, 66 percent compared to using RAN caching alone, and 100 percent compared to using ABR alone. The above video capacity gains can be achieved while mostly improving the video quality as measured by VQM and stalling probability.

We are currently working on expanding the mobile video cloud architecture and algorithms proposed here to include (the caches of) the mobile devices themselves, using base station assistance and coordination with the RAN caches, with the aim of further reducing initial playback delay and stalling, and increasing the video capacity of the wireless channels. We are exploring multiple approaches including (a) caching opportunistically by a mobile device in coordination with RAN caches based on its own UPP, and (b) cooperative caching by multiple neighboring devices based on aggregate UPP and using mobile-to-mobile (M2M) communication with base station assistance to provide requesting videos instead of fetching over cellular links.

"We are currently working on expanding the mobile video cloud architecture and algorithms proposed here to mobile devices…"

We also plan to address an evolving challenge from the increasing deployment of small cells (micro, pico, and femto): while improving access capacity and coverage, they will impose severe burden on backhaul capacity. We plan to explore caching and scheduling opportunities for small cells which can alleviate the related backhaul bandwidth challenge.

Regarding the second part of this article, the essence of the results were summarized in two examples: 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 Figure 5 or Figure 6, 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 Figures 7 through 9, it can be 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.

"…the cross-layer design yields better performance than do either of the two baseline approaches, in some cases by very significant amounts."

Based upon the these observations, our general conclusions are as follows:

1. System designs should be based upon robust performance over a wide operating range, as opposed to optimal performance for a specific operating scenario.
    a. Cross-layer designs typically result in robust performance even if robustness is not explicitly taken into account in the design.
    b. Additional robustness can be achieved by appropriate design, such as adapting system parameters to the nonstationary statistics of the channel.

2. If an adaptive receiver design is not feasible, robustness with respect to Doppler spread can be achieved by designing for the highest anticipated Doppler spread.
3. Meaningful performance gains (such as, for example, in capacity) tend to be very sensitive to specific operating conditions and are the smallest when the system performs very well.

Lastly, with respect to future research, we are further emphasizing the effects of nonstationary channel statistics by designing adaptive systems that track the time-varying coherence time and coherence bandwidth and then adjust key system parameters according. These system parameters include pilot spacing, interleaver depth, buffer size, and pilot power.

References

[1] Erman, Jeffrey, Alexandre Gerber, K. K. Ramakrishnan, Subhabrata Sen, and Oliver Spatscheck, "Over the top video: the gorilla in cellular networks," *Proc. Internet Measurement Conference,* 2012.

[2] Rudd, S. [Online], "Closing the backhaul gap, available: https://www.strategyanalytics.com/default.aspx?mod=reportabstractviewer&a0=8236," Strategy Analytics Report, February 2013.

[3] Ahlehagh, H. and S. Dey, "Video aware scheduling and caching in the radio access network," *IEEE Transactions on Networking*, vol. 22, no. 5, August 2014.

[4] Ahlehagh, H. and S. Dey, "Video caching in radio access network: Impact on delay and capacity," *Proceedings of the 2012 IEEE Wireless Communications and Networking Conference* (WCNC 2012), 2012.

[5] Ahlehagh, H. and S. Dey, "Hierarchical video caching in wireless cloud: Approaches and algorithms," *Proceedings of IEEE International Conference on Communications, Workshop on Realizing Advanced Video Optimized Wireless Networks (ICC ViOpt'12)*, Ottawa, Canada, June 2012.

[6] Ribas-Corbera, Jordi, Philip A. Chou, and Shankar L. Regunathan, "A generalized hypothetical reference decoder for H.264/AVC," *IEEE Transactions on Circuits and Systems*, vol. 13, no. 7, July 2003.

[7] Ahlehagh, H. and S. Dey, "Adaptive bit rate capable video caching and scheduling," *Proceedings of the 2013 IEEE Wireless Communications and Networking Conference (WCNC 2013)*, April 2013.

[8] Ahlehagh, H. and S. Dey, "Enhancing mobile video capacity and quality using rate adaptation, RAN caching and processing," Submitted to *IEEE Transactions on Networking*.

[9] Bernado, L., T. Zemen, A. Paier, G. Matz, J. Karedal, N. Czink, C. Dumard, F. Tufvesson, M. Hagenauer, A. F. Molisch, and C. F. Mecklenbrauker, "Non-WSSUS Vehicular Channel Characterization at 5.2 GHz - Spectral Divergence and Time-Variant Coherence Parameters," *Proceedings of the XXIXth URSI General Assembly in Chicago*, August 2008.

[10] Paier, A., T. Zemen, L. Bernado, G. Matz, J. Karedal, N. Czink, C. Dumard, F. Tufvesson, A. F. Molisch, and C. F. Mecklenbrauker, "Non-WSSUS Vehicular Channel Characterization in Highway and Urban Scenarios at 5.2 GHz Using the Local Scattering Function," *International ITG Workshop on Smart Antennas*, pp. 9–15, February 2008.

[11] Tse, D. and P. Viswanath, *Fundamentals of Wireless Communications*, Cambridge University Press (Cambridge: 2005).

[12] Wang, D., 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.

[13] Wang, D., L. Toni, P. C. Cosman, and L. B. Milstein, "Resource Allocation and Performance Analysis for Multiuser Video Transmission over Doubly Selective Channels," accepted in *IEEE Transactions on Wireless Communications*.

[14] Toni, L., 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.

Acknowledgment

This research was supported by the Intel-Cisco Video Aware Wireless Networks program, and by the National Science Foundation under Grant CCF-0915727.

Author Biographies

Hasti Ahlehagh (hahlehagh@gmail.com) received an master of science degree in electrical engineering from Worcester Polytechnic Institute, Worcester, Massachusetts, in 2004, and is currently pursuing a PhD in electrical and computer engineering at the University of California, San Diego. Before pursuing the PhD, she worked as a senior staff software engineer with the Connected Home Division of Motorola Mobility, Inc., and earlier as a firmware software engineer writing software for 3G wireless networks.

Laura Toni (laura.toni@epfl.ch) received her MS (with honors) in electrical engineering and a PhD 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. During 2007, she was a visiting scholar at the University of California, San Diego, working on video processing over wireless systems. She has been a postdoctoral fellow both at the Italian Institute of Technology and at UCSD, and is currently a postdoctoral fellow at the École Polytechnique Fédérale de Lausanne. Her research interests are in the areas of image and video processing, wireless communications, and underwater communications.

Dawei Wang (dwangnb@gmail.com) received a B.Eng. in electronic engineering (First Class Honors) from the Hong Kong University of Science and Technology (HKUST), Kowloon, Hong Kong SAR, China, in 2008, an MS in 2011 from the University of California, San Diego, and a PhD from UCSD in 2013. He was also a visiting student at the University of Minnesota, Minneapolis. In 2011 he was an intern at Intel Corporation, Hillsboro, Oregon, and is currently employed by Adaptive Spectrum & Signal Alignment, Inc. His research interests are in the areas of communication theory and video processing.

Pamela Cosman (pcosman@eng.ucsd.edu) obtained her BS from CalTech and her PhD from Stanford University, both in electrical engineering. Since 1995, she is with the department of Electrical and Computer Engineering at the University of California, San Diego, where she is currently a professor and Associate Dean for Students of the Jacobs School of Engineering. Her research is in image/video compression and processing, and wireless communications. She directed the Center for Wireless Communications (2006–2008), was an associate editor of the IEEE Communications Letters and Signal Processing Letters, and 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, and a Fellow of the IEEE.

Sujit Dey (sdey@ucsd.edu) received a PhD in computer science from Duke University, in Durham, North Carolina, in 1991. He is a professor with the Department of Electrical and Computer Engineering, University of California, San Diego, with research activities in mobile cloud computing, wireless multimedia, and green computing and communication. He serves as the co-director for the UCSD Center for Wireless Communications. He was the Chief Scientist, Mobile Networks, at Allot Communications from 2012–2013. He founded Ortiva Wireless in 2004, where he served as CTO till its acquisition in 2012. Prior to joining UCSD in 1997, he was a Senior Research Staff Member at the NEC C&C Research Laboratories in Princeton, New Jersey. He is a Fellow of the IEEE.

Laurence Milstein (Milstein@ece.ucsd.edu) received a PhD in electrical engineering from the Polytechnic Institute of Brooklyn in 1968. From 1968 to 1974, he was with the Space and Communications Group of Hughes Aircraft Company, and from 1974 to 1976, he was a member of the Department of Electrical and Systems Engineering, Rensselaer Polytechnic Institute, Troy, New York. Since 1976, he has been with the Department of Electrical and Computer Engineering, University of California at San Diego, where he is the Ericsson Professor of Wireless Communications Access Techniques and former Department Chairman, working in the area of digital communication theory.

FEMTOCACHING AND D2D COMMUNICATIONS: A NEW PARADIGM FOR VIDEO-AWARE WIRELESS NETWORKS

Contributors

Giuseppe Caire
Department of Electrical Engineering, University of Southern California

Andreas F. Molisch
Department of Electrical Engineering, University of Southern California

Video is a main driver for the increased traffic in wireless data networks. This article gives a survey of a novel transmission paradigm that we have developed in the course of the Video-Aware Wireless Networks (VAWN) project. It is based on the following two key properties: (1) video shows a high degree of asynchronous content reuse, and (2) storage is the fastest-increasing quantity in modern hardware. Based on these properties, we suggest caching in helper stations (femtocaching) and/or devices, combined with spectrally efficient short-range communications to deliver video files. For femtocaching, we develop both optimum storage schemes and dynamic streaming policies that optimize video quality. For caching on devices, combined with device-to-device communications, we show that communications within clusters of mobile stations should be used. The cluster size can be adjusted to optimize the tradeoff between frequency reuse and the probability that a device finds a desired file cached by another device in the same cluster. We establish scaling laws that show that under some circumstances the throughput can increase linearly with the number of users, and also analyze the tradeoff between throughput and outage. We finally show that our scheme can be combined with coded multicasting. Simulations demonstrate that network throughput (possibly with outage constraints) can be increased by two orders of magnitude compared to conventional schemes.

Introduction and Motivations

Video transmission is the main driver of the explosive growth in the usage of wireless data transmission. Originally, wireless video mostly implied short video clips (YouTube* or news channels) on the very small screens of smartphones. The recent popularity of tablets and large-screen phones have enabled watching feature-length movies at high resolution on mobile devices, thus greatly increasing the amount of data that have to be transmitted. Numerous market predictions (for example, Cisco[1]) anticipate an increase in the amount of data transmitted each day by almost two orders of magnitude over the next five years. It will then be the dominant source of wireless traffic, by far. These developments, while opening new business models and potentially improving consumer satisfaction, threaten to clog up the already overburdened cellular networks.

Traditional ways of enhancing throughput in cellular systems suffer from significant problems: (1) *increasing the amount of available spectrum* is a drawn-out and costly process, and its effectiveness is limited because the amount of spectrum in the microwave range that can be rededicated to cellular/Wi-Fi*

services is small; (2) *increasing the physical-layer capacity of wireless links* becomes difficult as 4G systems, employing MIMO-OFDM with capacity-achieving codes and interference coordination, are already close to the theoretical limits of what is practically feasible[2][3]; and (3) *decreasing the cell size* is viable but expensive. Deploying small base stations, thus creating pico- and femtocells that enable localized communication and high-density spatial reuse of communication resources, brings the (video) content closer to the users (see for example, Chandrasekhar et al.[4], Madan et al.[5], and references therein). However, one critical bottleneck is the cost of providing *backhaul connectivity* of the small base stations to the cellular operator network.[4] For this reason, new network structures have to be investigated that could provide higher per-user data rates at low cost.

In response to the Call for Proposals for "Video-Aware Wireless Networks"[6], we first proposed in 2010 principles of a new network structure, and from 2012–2014, sponsored by the Intel/Cisco/Verizon VAWN program, we elaborated, quantified, and refined these ideas considerably.[7–21] Our approach 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 *replacing backhaul by caching*. We first consider the use of *femtocaching*, where dedicated "helper nodes" can cache popular files and serve requests from wireless users by enabling localized wireless communication. Such helper nodes are similar to femto-BSs, but they have two key differences: they have *added* a large amount of storage, while they *do not have or need* a high-speed backhaul.

"…we suggest replacing backhaul by caching."

We can achieve an even higher density of caching by using devices themselves as video caches—in other words, using devices as mobile helper stations.[8] Due to the tremendous increase in memory on wireless devices (32–64 gigabytes for tablets, and several hundred gigabytes for laptops), there is ample storage for caching available on wireless devices themselves. The simplest way of using this storage would have each user cache the most popular files (possibly with individual modifications based on the tastes of a particular user). However, this approach incurs inefficiencies due to the fact that many users are interested in similar files, and thus the same videos will be duplicated on a large number of devices. On the other hand, the cache on each device is too small to cache a reasonably large number of files. Thus, it is preferable that the devices "pool" their caching resources, so that different devices cache different files and then exchange them, when the occasion arises, through device-to-device (D2D) communications. It is furthermore advantageous that this exchange process is controlled by the cellular infrastructure, which keeps track of (i) which device has which files in its cache, and (ii) which devices have sufficient channel quality to achieve short-range, spectrally efficient, D2D communications. If a requesting device does not find the file in its neighborhood (or in its own cache), it obtains the file in the traditional manner from the controlling base station.

"We can achieve an even higher density of caching by using devices themselves as video caches…"

Notice that caching is a well-known solution in current content distribution networks (CDNs) over the (wired) Internet.[22] Although caching is a standard

"…our proposed approach consists of caching directly at the wireless edge…"

approach in CDNs, such off-the-shelf solutions do not translate immediately into efficiency gains in the wireless segment since CDNs are implemented in the Internet cloud while, as mentioned above, the bottleneck here is the wireless segment and the limited backhaul connecting dense small cells. 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 approach is radically new and represents a major step forward with respect to conventional CDNs and yields higher spectrum spatial reuse at much lower deployment (CapEX) and operating (OpEX) cost.

This article is a summary of our previously published work (for principles and overviews see[7][8], femtocaching descriptions[9][10][11][12][13][14][15], and device-to-device discussions[16][17][18][19][20][21]). It is organized as follows: "Content Reuse" describes the content reuse of video transmission, "Femtocaching" is dedicated to the principle of femtocaching in helper stations, while "Device-to-Device Communications for Wireless Video" describes the main results of caching on devices combined with D2D communications. The "Conclusions" section gives a summary and outlook for future work.

Content Reuse

"Wireless video distinguishes itself from other wireless content through its strong content reuse."

Wireless video distinguishes itself from other wireless content through its strong content reuse. That is, the same content is seen by a large number of people. This fact was originally used in cellular systems for live streaming systems for exploiting the broadcast nature of the wireless channel.[23][24][25][26][27] Just like in live TV, the number of users tuned in to a particular channel hardly changes the resources required for transmission. However, attempts at practical implementation of such systems (such as, for example, the MediaFLO* system[28]) have failed because users do not want to be bound by predetermined starting times that are inextricably required for such distribution systems.

Rather, the bulk of wireless video traffic is due to asynchronous *video on demand*, where users request video files from some cloud-based server (such as iTunes*, NetFlix*, Hulu*, or Amazon Prime*) at arbitrary times. As indicated in the previous section and expounded upon in the following two sections, the use of caching enables the exploitation of *content overlap*, even in the presence of *asynchronism of requests*. As a matter of fact, time-accumulated viewing statistics show that a few popular videos (YouTube clips, sports highlights, and movies) account for a considerable percentage of video traffic on the Internet. Numerous experimental studies have indicated that Zipf distributions have been established as good models to the measured popularity of video files.[29][30] Under this model, the frequency of the ith popular file, denoted by f_i, is inversely proportional to its rank:

$$f_i = \frac{\frac{1}{i^{\gamma_r}}}{\sum_{j=1}^{m} \frac{1}{i^{\gamma_r}}}, \; 1 \le i \le m. \tag{1}$$

The Zipf exponent γ_r characterizes the distribution by controlling the relative popularity of files. Larger γ_r exponents correspond to higher content reuse; that is, the first few popular files account for the majority of requests. Here, m is the size of the library, which is not the size of all possible files on the Internet, but rather the library of files that are of interest to the set of considered users. Thus, the library size can be a function of the number of considered users n. Let m increase like n^{α}, where $\alpha \geq 0$. The case of constant library size ($\alpha = 0$) occurs, for example, when a provider makes available only a small, regularly rotating set of files to the users. Then $\alpha = 1$, that is, linear increase of m with n, occurs when users have disjoint interests. $\alpha < 1$ corresponds to the case that users share some interests, and thus their file requests overlap. (Though, note that for fixed probability of overlap the total number of requested files m actually increases like $log(n)$.) The case of $\alpha > 1$ can be attributed to the effects of social networking, where the presence of more users spurs an increasing diversity of file requests.

"A further important property of the library is that it changes only on a fairly slow timescale…"

A further important property of the library is that it changes only on a fairly slow timescale (several days or weeks). It can furthermore be shaped by content providers, for example, 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 iTunes). It is thus possible for helper stations and devices to obtain popular content, for example, 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 TB of data storage capacity, enough to store 1000 movies, costs only about USD 100 and could thus be easily added to a helper station. Even on mobile devices, 64 GB of storage could be easily dedicated to caching.

In order to avoid an excessively rosy scenario, we put forward a few disclaimers. The work reported here applies principally to a setting where a content library of relatively large files (such as movies and TV shows) is refreshed relatively slowly (for example, on a dally basis), and where the number of users consuming such a library is significantly larger than the number of items in the library. This may apply to a possible future implementation of a "wireless NetFlix," but it is hardly applicable to a "wireless YouTube" paradigm, where the library items are very short (a few minutes), and the library size is much larger than the typical number of users in a given geographic coverage area. In the latter case, the asynchronous content reuse which caching capitalizes on is hardly existent, and caching the whole YouTube library at the wireless edge would be clearly infeasible. In short, this article reflects a set of results and approaches that are relevant in the case where the caching phase (placement of content in the caches) occurs with a clear time-scale separation with respect to the delivery phase (the process of delivering video packets for streaming to the users), and where the size of the content library is moderate with respect to the user population. Further comments are provided under the "Main Conclusions" heading in the next section.

"…we illustrate our progress on a femtocaching architecture formed by a set of helper nodes…"

Femtocaching

In this section we illustrate our progress on a femtocaching architecture formed by a set of helper nodes (akin to small cell base stations, with large storage capacity and no or very weak backhaul) serving a set of user nodes, which request on-demand video streaming sessions. The two main problems to be addressed for such an architecture are (1) the generalization of dynamic adaptive streaming currently used in wireless video streaming applications-layer protocols such as Microsoft Smooth Streaming (Silverlight)*[31], Apple HTTP Live Streaming*[32] and 3GPP Dynamic Adaptive Streaming over HTTP (DASH)*[33], to the case of a network of multiple users and multiple helpers, and (2) the cache placement problem, that is, how to optimally place the video files into the helper caches. The following two subsections, "Dynamic Adaptive Streaming from Multiple Helpers" and "Optimal Centralized Cache Placement," illustrate some progress on these topics, respectively. "Testbed Experiments" briefly outlines an experimental Wi-Fi–based testbed setup that we have been developing for the sake of demonstration and validation.

Dynamic Adaptive Streaming from Multiple Helpers

Consider a discrete, time-slotted wireless network with user set $\mathcal{U}$ (requesting streaming) and helper set $\mathcal{H}$ (serving such requests). In general, every helper $h \in \mathcal{H}$ has a subset $\mathcal{F}(h)$ of files in the library $\mathcal{F}$ within its cache. Some "infrastructure" nodes (for example, cellular base stations) may be connected directly to a CDN in the core network and have access to the whole library $\mathcal{F}$. Also, because of radio-access technology (RAT) restrictions, not all users and helpers may communicate. Therefore, the set of possible communication links is (h, u).

Each user $u \in \mathcal{U}$ requests a video file $f_u \in \mathcal{F}$, formed by a sequence of chunks, which are independently decodable standalone units[34]. Chunks have a fixed duration T_{chunk} and must be reproduced sequentially at the user end. The streaming process consists of transferring N chunks from the helpers to the requesting users (for simplicity of exposition we assume that all files have the same duration $T_{\text{file}} = NT_{\text{chunk}}$) such that the playback buffer of each user contains the required chunk at the beginning of its playback time. Each file f is encoded at a finite number of different quality levels $m \in \{1, \ldots, N_f\}$. In VBR video coding[35], the quality-rate profile may vary from chunk to chunk in the same file, and across different files. We let $D_f(m, t)$ and $B_f(m, t)$ denote the video quality measure (for example, see Wang et al.[36]) and the number of bits for chunk t of file f at quality level m, respectively. We let $r_{hu}(t)$ denote the source coding rate (bit per chunk) of chunk t requested by user u to helper h. Hence, the streaming scheduler must also allocate the source coding rates $r_{hu}(t)$ satisfying

$$\sum_{h \in \mathcal{H}(u)} r_{hu}(t) = B_{f_u}(m_u(t), t), \quad \forall_u \in \mathcal{U} \tag{2}$$

Notice that this formulation applies also to the case of intrasession network coding or other forms of coding against packet erasures, where the requested data may not be the video-encoded bits but linear combinations thereof, suitably replicated throughout the network for the sake of robustness (see for example Pawar et al.[37]).

We represent the underlying wireless network physical layer through a certain long-term average rate region $\mathcal{R}(t)$. For example, we have considered[14] the region achievable by single-antenna helpers and users operating on the same frequency channel, treating interference as noise, and using intracell orthogonal access. However, our framework generalizes to virtually any arbitrary physical layer. For example, in the more recent work[15] we have considered multiuser MIMO helpers (for example, implemented by 802.11ac wave-2 access points, capable of multiuser spatial multiplexing).

We consider a *Network Utility Maximization* (NUM) formulation[14] in order to systematically design an adaptive dynamic streaming scheduler in the above described femtocaching network. Each helper h has a transmission queue pointing at its served users $\mathcal{U}(h)$, which evolves as

$$Q_{hu}(t+1) = \max\{Q_{hu}(t) - WT_{\text{chunk}} R_{hu}(t), 0\} + r_{hu}(t) \quad (3)$$

In words, at each time t, helper h serves $WT_{\text{chunk}}R_{hu}(t)$ information bits to user u by transmitting over WT_{chunk} physical layer dimensions a rate $R_{hu}(t)$ bits/channel use, where W is the system bandwidth. The input to queue $Q_{hu}(t)$ is formed by the newly requested $r_{hu}(t)$ video-encoded bits. Then, for large files formed by many chunks (in the limit for $N \to \infty$), the NUM problem is given by:

maximize

$$\phi(\bar{D}_u; u \in \mathcal{U}) \quad (4)$$

subject to

$$\lim_{t\to\infty} \frac{1}{t} \sum_{\tau=1}^{t-1} \mathbb{E}[Q_{hu}(\tau)] < \infty \; \forall (h, u) \in \mathcal{E} \quad (5)$$

$$\alpha(\tau) \in A_{\omega(t)} \; \forall \; t, \quad (6)$$

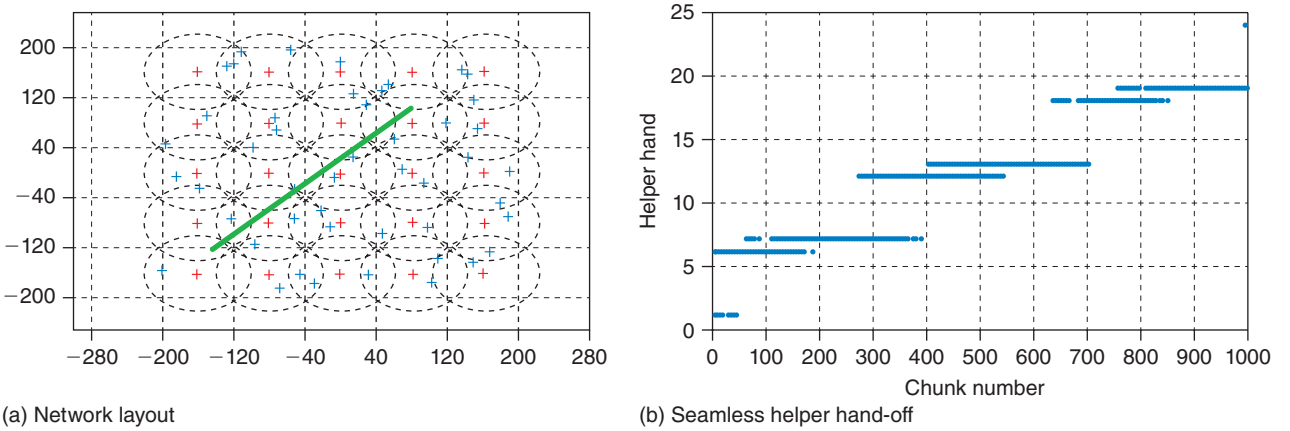
where constraint (5) corresponds to the queues' *strong stability*, $\alpha(t)$ is the decision policy, including the video coding rate requests $\{r_{hu}(t)\}$, the feasible channel coding rate allocation $\{R_{hu}(t)\} \in \mathcal{R}(t)$ and the video quality selection decisions $\{m_u(t)\}$, and $A_{\omega(t)}$ is the set of feasible policies for network state $\omega(t) = \{g_{hu}(t), D_{f_u}(\cdot, t), B_{f_u}(\cdot, t): \forall (h, u) \in \mathcal{E}\}$. We solved problem (4)–(6) through the design of a dynamic policy based on the Lyapunov drift plus penalty (DPP) approach. The resulting policy is decentralized and consists of a distributed multiuser version of a DASH-like protocol, where users make adaptive decisions on which helper to request from and at which quality level each chunk should be requested.

"We consider a Network Utility Maximization (NUM) formulationin order to systematically design an adaptive dynamic streaming scheduler…"

In the more recent work[15] we have considered the case where each user maintains a single *virtual* request queue $Q_u(t)$ and dynamically requests only the chunk at the head of the line. This approach, referred to as "pull" strategy, prevents a user from receiving chunks out of playback order. We have also discussed[14][15] effective prebuffering and rebuffering schemes based on monitoring the maximum chunk delivery delay in a sliding window and setting the buffering time (in multiples of T_{chunk}) sufficiently larger than the maximum observed chunk delay.

"…we have considered the case where each user maintains a single virtual request queue…"

Figure 1 shows a numerical experiment based on a small-cell femtocaching layout, where we let one user move along a linear trajectory at speed 1 m/s (Figure 1(a)). Our dynamic adaptive streaming policy is able to implicitly "discover" new helpers as the user moves across the network, such that chunks are requested from favorable helpers along the streaming session (Figure 1(b)).



(a) Network layout

(b) Seamless helper hand-off

Figure 1: Simulation experiments of the dynamic adaptive streaming policy. (Source: Bethanabhotla et al., 2014.[15] (Copyright IEEE))

Optimal Centralized Cache Placement

In this section we consider the problem of optimal content placement in a femtocaching network where: (1) the network topology is known, (2) the long-term average link rates are known, and (3) the user demand distribution (file popularity) is known. Yet the realization of the users' demands is a random vector, and the caching placement must be done without *a priori* knowledge of the user requests, but only of their statistics.

We have considered the following problem.[12] With the same notation developed before, a femtocaching network is represented by the bipartite graph $\mathcal{G} = (\mathcal{H}, \mathcal{U}, \mathcal{E})$. Here, we are interested in minimizing the average file downloading time with respect to the case where all users download independently from the cellular base station (denoted as helper node 0 in the following). We simplify the physical layer and consider that each link has a fixed average downloading latency indicated by $\omega_{h,u}$, expressed in *time per information bit*. This scenario is consistent with that developed before, considering that the physical layer rate region $R(t)$ is time-invariant in this case, and that the adaptive dynamic scheduling chooses a fixed set of physical layer rates $\mathbf{R} \in \mathcal{R}$, such that $\omega_{h,u} = 1/R_{h,u}$.

The content placement problem treated in this work can be formulated as follows: *for a given file popularity distribution, helper storage capacity and network*

topology, how should the files be placed in the helper caches such that the average sum downloading delay of all users is minimized?

We distinguish between uncoded and coded content placement. In the uncoded case, video-encoded files are cached directly, with possible replication. In the coded case, we consider intrasession coding already mentioned before (for example, using the scheme presented by Pawar et al.[37]).

"We distinguish between uncoded and coded content placement."

Uncoded cache placement: An uncoded cache placement is represented by a bipartite graph $\tilde{\mathcal{G}} = (\mathcal{F}, \mathcal{H}, \tilde{\mathcal{E}})$, such that an edge $(f, h) \in \tilde{\mathcal{E}}$ indicates that a copy of file f is contained in the cache of helper h. We let $\mathbf{X}$ denote the $|\mathcal{F}| \times |\mathcal{H}|$ adjacency matrix of $\tilde{\mathcal{G}}$, such that $x_{f,h} = 1$ if $(f, h) \in \tilde{\mathcal{E}}$ and 0 otherwise. By the cache size constraint, we have that the column weight of X is at most equal to the cache size M (expressed in file units).

The average delay per information bit for user u can be written as:

$$\bar{D}_u = \sum_{j=1}^{|\mathcal{H}(u)|-1} \omega_{(j)_u,u} \sum_{f=1}^{|\mathcal{F}|} \left[\prod_{i=1}^{j=1} (1 - x_{f,(i)_u}) \right] x_{f,(j)_u} P_r(f)$$

$$+ \omega_{0,u} \sum_{f=1}^{|\mathcal{F}|} \left[\prod_{i=1}^{|\mathcal{H}(u)|-1} (1 - x_{f,(i)_u}) \right] P_r(f). \quad (7)$$

where $P_r(t)$ is the request probability distribution, and where the notation $(j)_u$ indicates the helper index in $H(u)$ with the jth smallest delay to user u. The minimization of the sum (over the users) average per-bit downloading delay can be expressed as the integer programming problem:

$$\text{maximize} \sum_{u \in \mathcal{U}} (\omega_{0,u} - \bar{D}_u)$$

$$\text{subject to} \sum_{f \in \mathcal{F}} x_{f,h} \leq M, \quad \forall \quad h$$

$$X \in \{0,1\}^{|\mathcal{F}| \times |\mathcal{H}|}. \quad (8)$$

We showed[12] that (8) is NP-complete. However, it can be formulated as the maximization of a monotone submodular function over matroid constraints, for which a simple greedy strategy achieves at least one half of the optimum value.

Coded content placement: Using intrasession coding, we can obtain a relaxed version of the above problem. In particular, let $\rho = [\rho_{f,h}]$, where $\rho_{f,h}$ denotes the fraction of parity bits of file f contained in the cache of helper h. The delay to download a fraction of parity bits $\rho_{f,h}$ on the link (h, u) is given by $\rho_{f,h}\, \omega_{h,u}$. A file is entirely retrieved when a fraction larger than or equal to 1 of parity bits is downloaded, since by a property of Maximum Distance Separable codes, we have that with a number of parity bits equal to the number of information bits, then the latter can be exactly recovered. The average delay per information bit necessary for user u to download file f, assuming that it can download it from its best j helpers, is given by

"Using intrasession coding, we can obtain a relaxed version of the problem."

$$\bar{D}_u^{f,j} = \sum_{i=1}^{j-1} \rho_{f,(i)_u} \omega_{(i)_u,u} + \left(1 - \sum_{i=1}^{j-1} \rho_{f,(i)_u}\right) \omega_{(j)_u,u}$$

$$= \omega_{(j)_u,u} - \sum_{i=1}^{j-1} \rho_{f,(i)_u} (\omega_{(j)_u,u} - \omega_{(i)_u,u}). \quad (9)$$

Notice that file f can be downloaded by user u from its best j helpers only if $\sum_{i=1}^{j} \rho_{f,(i)_u} \geq 1$. In addition, since the cellular base station contains all files, we always have $\rho_{f,0} = 1$ for all $f \in F$, such that all users can always obtain the requested files by downloading the missing parity bits from the base station.

The delay $\bar{D}_u^f$ incurred by user u because of downloading file f is a piecewise-defined affine function of the elements of the placement matrix ρ, given by

$$\bar{D}_u^f = \begin{cases} \bar{D}_u^{f,1} & \text{if} \quad \rho_{f,(1)_u} \geq 1 \\ \vdots & \vdots \\ \bar{D}_u^{f,j} & \text{if} \quad \sum_{i=1}^{j-1} \rho_{f,(i)_u} < 1, \sum_{i=1}^{j} \rho_{f,(i)_u} \geq 1 \\ \vdots & \vdots \\ \bar{D}_u^{f,|\mathcal{H}(u)|} & \text{if} \quad \sum_{i=1}^{|\mathcal{H}(u)|-1} \rho_{f,(i)_u} < 1, \end{cases} \quad (10)$$

We can show that $\bar{D}_u^f$ is a convex function of ρ. The average delay of user u is given by $\bar{D}_u = \sum_{f=1}^{|\mathcal{F}|} P_r(f) \bar{D}_u^f$. With some further manipulations, the coded placement optimization problem takes on the form:

$$\text{minimize} \sum_{u=1}^{U} \sum_{f=1}^{|\mathcal{F}|} P_r(f) \max_{j \in \{1,2,\ldots,|\mathcal{H}(u)|\}} \{\bar{D}_u^{f,j}\}$$

$$\text{subject to} \sum_{f=1}^{|\mathcal{F}|} \rho_{f,h} \leq M, \quad \forall\, h$$

$$\rho \in [0,1]^{|\mathcal{F}| \times |\mathcal{H}|}, \quad (11)$$

where the optimization is with respect to ρ. In general, the optimum value of delay obtained with the coded optimization is better than the uncoded optimization because any placement matrix with integer entries is a feasible solution to the coded problem. In this sense, the coded optimization is a convex relaxation of the uncoded problem.

"…beyond its theoretical interest, optimal cache placement is unlikely to be useful in practice…"

We conclude this section by mentioning that, beyond its theoretical interest, optimal cache placement is unlikely to be useful in practice since while the user demand distribution $P_r(f)$ may be well estimated and predicted, the network topology is typically time-varying with dynamics comparable with the streaming sessions. Therefore, reconfiguring the caches at this time scale is definitely not practical. However, further computer experiments have also shown that the cache distribution obtained when the mobile stations are in "typical" distances from the helpers also provides good performance for various realizations of random placement of nodes. Furthermore, distributed random caching turns out to be "good enough" as we shall see under the heading "D2D Throughput versus Outage" in the following section. Hence, comparing

optimal placement with random caching yields useful insight into the potential performance gap lost by a decentralized approach. Interestingly, in any reasonable network configuration, it turns out that such a gap is very small.

Testbed Experiments

We have implemented a small Wi-Fi–based testbed to demonstrate the femtocaching idea. In particular, we have implemented a scheme reminiscent of the dynamic adaptive streaming scheme illustrated earlier under the heading "Dynamic Adaptive Streaming from Multiple Helpers," where both helper and user nodes are implemented on an Android* mobile platform (see Figure 2). The helpers create their own hot spot using the tethering mode of Wi-Fi, such that they effectively act as base stations with cached content and no wired (DSL/Ethernet) backhaul.

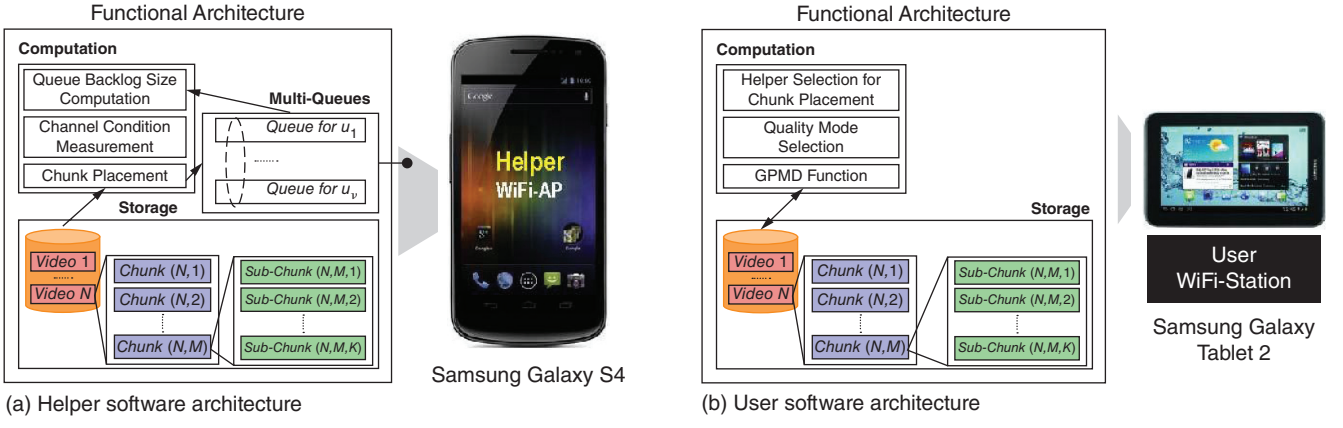


Figure 2: Device software architectures. (Source: Kim et al., 2013.[13] (Copyright ACM))

A practical limitation of Wi-Fi is that users cannot be associated simultaneously with multiple helpers and therefore possibly receive multiple data packets from different helpers. Therefore, we have modified the scheme of the earlier section, implementing a heuristic strategy for user-helper *exclusive* association. In particular, we have implemented the "pull" strategy with a single request queue per user. Then, each user selects the helper node that has the next required chunk according to its own request queue and such that the selected helper has the shortest transmit queue among all helpers having said chunk. The chunk is requested at a quality level that depends on the helper-user Wi-Fi link rate. In this way, we take the helper transmit queue as a proxy for its current load, and this strategy implements a sort of implicit load balancing. It should be noted that this heuristic puts higher priority on receiving the chunk on time rather than obtaining high video-coding quality. In fact, with this strategy, a user always goes for the helper with the smallest transmit queue, even if the obtained link rate is smaller compared to some other helper. The implemented scheme is referred to as a *greedy pull for minimum delay (GPMD) strategy*.

The users can follow two approaches: (1) selecting a possibly different helper at each chunk request or (2) selecting a helper that should transmit a sequence of chunks (thus limiting the number of handoffs between helpers, which in practice contributes to a significant protocol overhead due to the inefficiency of Wi-Fi). In the first case, at each time slot, each user determines which helper should place/stream the next chunk. This is obtained by letting each user send a packet requesting the next video chunk to all neighbor helpers. Each helper replies back with the current queue backlog size. Then, the user selects the helper that has the smallest queue backlog size and sends a chunk download request to said helper. If there are multiple helpers achieving the same smallest queue backlog sizes, the user performs a random selection. The second approach is similar, but helper selection decisions are made less often. For example, in the case of low mobility users, while a user moves across the network, it detects when the current serving helper yields an unacceptably low per-link rate. In this case, it initiates a new request in order to determine a new helper from which to download.

Main Conclusions

"This subsection summarizes a number of lessons learned, system design guidelines, and points for further investigations relative to our work on femtocaching summarized before."

This subsection summarizes a number of lessons learned, system design guidelines, and points for further investigations relative to our work on femtocaching summarized before.

The first observation concerns the validity of the time-scale decomposition that we have implicitly assumed when separating the cache placement phase and the video delivery phase. This assumption is valid for the case of a video library formed by popular content such as movies and TV shows, which is refreshed on a daily or even weekly basis. In this case, we think of a sort of "wireless NetFlix" that pushes into the helpers' caches new content at off-peak times (for example, at night, exploiting the LTE cellular network, without requiring any wired backhaul). In our cache placement problem formulation, we have considered a single popularity distribution. As a matter of fact, the popularity distribution varies significantly with respect to the social group of users, the location, and the time of day. For example, we may imagine that in a train station (for example, Penn Station in New York) the morning commuters are interested in the news and stock market data, while people in a city park in the afternoon are interested in the latest episode of *Modern Family*. Mathematically, by conditioning with respect to a restricted location and time of day, the popularity distribution can have noticeable peaks as compared to averaging over all locations and times. Hence, demands can be more easily predicted and caches better utilized if such "high definition" information is available. Furthermore, the prediction of users' content demands can be further enhanced by taking into account the users' social network interconnections (for example, recommendation systems). All these considerations call for a systematic and coherent study of the problem of content demand prediction in space (with the resolution of the typical area of a small cell corresponding to a helper node) and in time (with the resolution of a few hours). Such prediction can be formulated as a large-dimensional Bayesian inference problem, for which machine learning techniques can be applied. This represents an interesting area for further research.

The second observation is that in our video delivery (streaming/scheduling) from multiple helpers we have assumed that users can request video chunks from multiple helper stations without paying a handover cost. As a matter of fact, if the underlying PHY and MAC wireless network are implemented with off-the-shelf Wi-Fi or Wi-Fi–direct, clients must associate to helper nodes and dynamic association on a per-chunk basis is infeasible because of the slow handover. This calls for a more efficient implementation of the PHY and MAC of the underlying small cell network, allowing for highly dynamic helper-user association.

Finally, a number of improvements to the dynamic streaming and link scheduling algorithms can be considered, as recently done by Bethanabhotla et al.[38] and Joseph and Veciana.[39] These schemes improve upon the basic scheme provided in this survey article, since they attempt a direct control of the playback buffer of the users, ensuring smooth streaming without interruptions, and make use of a single request "virtual queue" that avoids the annoying problem of chunks delivered out-of-order. This may occur in the case where each helper has its own individual transmit queue pointing to a requesting user such that chunks requested from different helpers may be subject to different queuing delays.

Device-to-Device Communications for Wireless Video

We now turn to networks where the devices themselves are caching video files, and transmitting them, upon request, to other devices via high-spectral-efficiency D2D links. For this type of network, we only consider the transmission of video *files*, not streaming, and also neglect the issue of video rate adaptation (these are topics for our future research). In this section, we first outline the principle and mathematical model we consider. We then provide the fundamental scaling laws, both for the sum throughput in the cell (disregarding any fairness considerations), and for the tradeoff between throughput and outage. Under the heading "D2D with Coded Caching," we then describe how D2D communications can be combined with coded multicast. Simulation results described in "Performance in Realistic Settings" illustrate important behaviors.

"We now turn to networks where the devices themselves are caching video files, and transmitting them, upon request, to other devices via high-spectral-efficiency D2D links."

Principle and Mathematical Model

We consider a network where each device can cache a fixed number M video files, and send them—upon request—to other devices nearby. If a device cannot obtain a file through D2D communications, it can obtain it from a macrocellular base station (BS) through conventional cellular transmission. The BS also keeps track of which devices can communicate with each other and which files are cached on each device. Such BS-controlled D2D communication is more efficient (and more acceptable to spectrum owners if the communications occur in a licensed band) than traditional uncoordinated peer-to-peer communications.

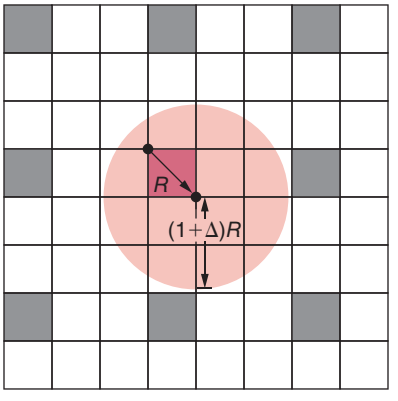


Figure 3: An example of the single-cell layout (Source: Ji et al., 2014.[21] (Copyright IEEE))

Specifically, consider a network that consists of macrocells. For simplicity, each macrocell is assumed to have the shape of a square, with the side of the square having unit length, and with n users per cell. We assume that inter-cell interference from the device BS links is kept to a small level through appropriate cell/frequency planning.[40] There are n users in each cell. We consider either the case that the users are placed on a fixed grid such that the distance of the users is $1/\sqrt{n}$, or we consider a random placement of the nodes in the network, leading to a random geometric graph. Each user makes a request for a file in an i.i.d. (independent identically distributed) manner, according to a given request probability mass function $P_r(f)$.

We furthermore subdivide the cell into equal-sized, disjoint groups of users that we call "clusters" of size (radius) r, with g_c nodes in it. To further simplify the mathematical model, we assume that only nodes that are part of the same cluster can communicate with each other. If a user can find the requested file inside the cluster, we say there is one *potential link* in this cluster; when at least one link is scheduled, we say that the cluster is "active." We use an *interference avoidance* scheme, such that at most one link can be active in each cluster on one time-frequency resource. Intercluster interference is avoided through a frequency reuse strategy as shown in Figure 3. In a simplified protocol model[41] nodes that are within the "reuse distance" cannot communicate at all due to interference (red disk), while nodes/clusters outside the reuse distance are not interfered with at all.

This model is, of course, a major simplification whose assumptions do not hold exactly in practice. Yet, it provides a first approximation to the exact solutions. Furthermore, many of the simplifications do not impact the scaling laws (that is, the functional form of the increase of throughput with number of users), though they do impact the absolute value of the throughput.

We furthermore have to determine which files should be cached by the devices. We consider here two strategies:

- *deterministic* caching, where the BS instructs the devices to cache the most popular files in a disjoint manner; that is, no file should be cached twice in devices belonging to the same cluster. This approach can only be realized if the device stays in the same locations for many hours (the time between refreshing of the cache, which is a rare event, and the time the files are requested by other devices). Performance obtained with the deterministic caching strategy also serves as a useful upper bound for more realistic schemes.
- *random* caching, where each device randomly and independently caches a set of files according to a common probability mass function. In our first papers, we assumed that the caching distribution is also a Zipf distribution, though with a parameter γ_c that is different from γ_r, and which has to be optimized for a particular γ_r and r. Since the Zipf distribution

is characterized by a single parameter, this description gives important intuitive insights into how concentrated the caching distribution should be.

We subsequently found[19] that the optimal caching distribution that maximizes the probability that any user finds its requested file inside its own cluster is given (for a node arrangement on a rectangular grid as described above) by

$$P_c^*(f)=\left[1-\frac{v}{z_f}\right]^+, f = 1, \ldots, m, \quad (12)$$

where $v=\frac{m^*-1}{\sum_{f=1}^{m^*}\frac{1}{z_f}}$, $z_f = P_r(f)^{\frac{1}{M(g_c-1)-1}}$, $m^* = \Theta\left(\min\left\{\frac{M}{\gamma_r}g_c, m\right\}\right)$ and $[\Lambda]^+ = \max[\Lambda, 0]$.

Besides the caching strategy, the main performance factor that can be influenced by the system designer is the cluster size. This is regulated through the transmit power (we assume that it is the same for all users in a cell, but can be optimized as a function of user density, library size, and cache size). Varying the cluster size trades off probability for finding the desired file in the cluster with the frequency reuse. Optimizing cluster size is an important task for the system design.

There are a number of different criteria for optimizing the system parameters. One obvious candidate is the total network throughput. It is maximized by maximizing the number of active clusters. We showed[18] that for deterministic caching, the expected throughput can be computed as

$$E\{T\} = \frac{1}{r^2}\sum_{k=0}^{n}\left(1-\prod_{i=1}^{k}(1-(P_{CVC}(k)-P_r(f_i)))\right)\Pr[K=k]. \quad (13)$$

where $P_{CVC}(k)$ is the probability that the requested file is in the Common Virtual Cache (the union of all caches in the cluster), that is, among the k most popular files. $Pr[K=k]$, the probability that there are k users in a cluster, is deterministic for the rectangular grid arrangement, and

$$\Pr[K=k] = \binom{n}{k}(r^2)^k(1-r^2)^{n-k}, \quad (14)$$

for random node placement.

Scaling Laws for Throughput

We now turn to the scaling laws, which describe the functional behavior of the overall throughput T as a function of the user density. For this analysis, we concentrate on the case that each device make requests according to a Zipf distribution with γ_r, and randomly caches according to a Zipf distribution with parameter γ_c. We note, however, that deterministic and random caching show no fundamental difference in their scaling laws. We use the following notation: given two functions f and g, we say that: $f(n) = O(g(n))$ if there exists a constant c and integer N such that $f(n) \le cg(n)$ for $n > N$; $f(n) = \Omega(g(n))$ if $g(n) = O(f(n))$; $f(n) = \Theta(g(n))$ if $f(n) = O(g(n))$ and $g(n) = O(f(n))$.

"We now turn to the scaling laws, which describe the functional behavior of the overall throughput T as a function of the user density."

We established[16][17] the following lower and upper bounds:

Theorem 1: If the Zipf exponent $\gamma_r > 1$,

i) Upper bound: For any caching policy, $E[T] = O(n)$,

ii) Achievability: Given that $c_1\sqrt{\frac{1}{n}} \le r_{opt}(n) \le c_2\sqrt{\frac{1}{n}}$ (c_1 and c_2 are positive constants that do not depend on n) and using a Zipf caching distribution with exponent $\gamma_c > 1$ then $E[T] = \Theta(n)$.

This theorem shows that if we choose $r_{opt}(n) = \Theta\left(\sqrt{\frac{1}{n}}\right)$ and $\gamma_c > 1$, $E[T]$ can grow linearly with n. For the request distributions that are less concentrated $\gamma_r < 1$, we obtained the following result:

Theorem 2: If $\gamma_r < 1$,

i) Upper bound: For any caching policy, $E[T] = O\left(\frac{n}{m^\eta}\right)$ where $\eta = \frac{1-\gamma_r}{2-\gamma_r}$,

ii) Achievability: If $c_3\sqrt{\frac{m^{\eta+\epsilon}}{n}} \le r_{opt}(n) \le c_4\sqrt{\frac{m^{\eta+\epsilon}}{n}}$ and users cache files randomly and independently according to a Zipf distribution with exponent γ_c, for any exponent $\eta + \varepsilon$, there exists γ_c such that $E[T] = \Theta\left(\frac{n}{m^{\eta+\epsilon}}\right)$ where $0 < \varepsilon < \frac{1}{6}$ and γ_c is a solution to the following equation

$$\frac{(1-\gamma_r)\gamma_c}{1-\gamma_r+\gamma_c} = \eta + \varepsilon.$$

The main conclusion from the scaling law is that for highly concentrated demand distribution, $\gamma_r > 1$, the throughput scales linearly with the number of users. Equivalently, the per-user throughput remains constant as the user density increases; the number of users in a cluster also stays constant. For heavy-tailed demand distributions, the throughput of the system increases only sublinearly, as the clusters have to become larger (in terms of number of nodes in the cluster), to be able to find requested files within the caches of the cluster members.

"…focusing on throughput only is not enough in a D2D network…"

D2D Throughput versus Outage

We noticed[20] that focusing on throughput only is not enough in a D2D network, especially under the protocol "collision" model. In fact, from the network viewpoint, the throughput is maximized by allowing only nearest neighbor communication for the users whose request can be satisfied by a neighbor, and dropping all other users. This, however, yields a very large probability that a user request is not satisfied by the network. A more complete picture is provided by the throughput-outage tradeoff, defined by Ji et al.[20], where we focused on *max-min fairness.* That is, our aim is to maximize the minimum average throughput $\bar{T}_{\min}$ per user, subject to a constraint on the average outage probability p. In other words, the fraction of users that are not served by the system (either because their request is not found in the caches, or because the scheduling policy denies service to these users). The *Throughput-Outage Tradeoff* of a D2D caching network is generally defined as the region of all achievable throughput-outage pairs (T, p). In particular, letting $T^*(p) = \sup\{T: (T, p) \in \mathcal{T}\}$ be the maximum achievable min-throughput per user for given outage probability not larger than p, we have that $T^*(p)$ is the result of the optimization problem:

$$\begin{aligned} &\text{maximize } \bar{T}_{\min} \\ &\text{subject to } p_o \le p, \end{aligned} \tag{15}$$

where p_o indicates the average (over the users) outage probability, and the maximization is with respect to the decentralized cache placement and transmission policies.

Letting $|\mathcal{U}| = n$ (number of users) and $|\mathcal{F}| = m$ (number of files), our main result is that in the regime of n $\to \infty$ and library size m at most linear in n, for any strictly positive p, the optimal throughput of a D2D caching network with one-hop direct communication between sources and destinations scales as $T^*(p) = \Theta\left(\max\{\frac{M}{m}, \frac{1}{n}\}\right)$. The outage probability affects only the multiplicative constant of the leading term, which can be tightly characterized via inner and outer bounds.[20] Interestingly, this scaling law is identical to what can be achieved by *network-coded multicasting* from a single base station[42], and what can be achieved by network-coded D2D delivery (see "D2D with Coded Caching"). Here, this provably optimal (in an information theoretic sense) scaling law is achieved by using simple direct delivery (no intrasession network coding) and simple decentralized cache placement. The related D2D link scheduling is also extremely simple, namely the scheme described earlier under the heading "Principle and Mathematical Model" and also assumed in "Scaling Laws for Throughput". Again, the "magic" of this scheme consists of choosing appropriately the cluster size (see also Golrezaei et al.[11][18]). If the cluster is too small, the spectrum reuse is large but the probability of not finding the desired content in the cluster is also large, resulting in a unacceptably high outage probability. If the cluster size is too large, the content can be found with high probability, but the spectrum spatial reuse is too low. Balancing the tension between spectrum reuse and outage probability, we arrive at the order-optimal result. Interestingly, the scaling $1/n$ corresponds to orthogonal access from a single broadcasting base station with the full library and can be regarded as representative of current conventional systems. In contrast, when $nM \gg m$, the throughput increases linearly with M. In this regime, *offloading the video on demand traffic to the D2D local links by exploiting the storage capacity at each node is a very attractive approach, since storage space is much "cheaper" than scarce resources such as bandwidth or dense base station deployment.*

"…offloading the video on demand traffic to the D2D local links by exploiting the storage capacity at each node is a very attractive approach…"

D2D with Coded Caching

Recently, a *network coded multicasting* scheme exploiting caching at the user nodes was proposed by Maddah-Ali and Niesen.[42][43] In this scheme, the files in the library are divided into blocks (packets) and users cache carefully designed subsets. (In this context, the packetization used for coding may not coincide with the chunk units used by the streaming process.) Then, for a given set of user demands, the base station sends to all users (multicasting) a common codeword formed by a sequence of packets obtained as linear combinations of the original file packets. For the case of arbitrary (adversarial) demands, the scheme of Maddah-Ali and Niesen[42] is shown to be information theoretic near-optimal in the sense that the number of network coded packet transmissions necessary to satisfy any user demand is minimal within a small bounded gap. Both articles by Maddah-Ali and Niesen[42][43] consider

one-hop transmission from the base station only, and it is assumed that all users can receive (at the same rate) one network coded packet per unit time. (In fact, content delivery from a single transmitter (base station) to multiple users with caching is a special case of index coding[44][45][46] where the demands are arbitrary but the "side information" formed by the cached packets is explicitly designed or generated at random with a specific statistical distribution.)

The number of required multicast network coded packets that must be transmitted to satisfy any user demand achieved by the scheme of Maddah-Ali and Niesen[42] is given by

$$N(n, m, M) = n\left(1-\frac{M}{m}\right)\frac{1}{1+\frac{Mn}{m}}.$$

This yields the min per-user throughput $T_{\mathrm{CM}}=\frac{c_0}{N(n, m, N)}$ where C_0 is the common multicasting rate that any user in the system must be able to decode. (For simplicity of exposition, we neglect here the effect of a nonzero outage probability. See Ji et al.[21] for more details.) One can see immediately that for large *m, n,* and finite *M*, the scaling of the symmetric throughput per user is given by $\Theta\left(\max\left\{\frac{M}{m}, \frac{1}{n}\right\}\right)$. As mentioned above, somewhat surprisingly, this is the same scaling behavior of our D2D scheme outlined earlier in "D2D Throughput versus Outage."

At this point, a natural question to ask is whether the gain of D2D local transmission and the gain of network-coded multicasting will be compounded. We considered[20] a network-coded multicasting scheme that involves only local D2D communication (no base station). The caching and delivery scheme is best explained by the example shown in Figure 4. This scheme can be generalized to any *n*, *m*, *M*, and it can be shown that without any spatial reuse (that is, any transmission is heard by all users in the cell, and one transmission per time-frequency slot is allowed), this subpacketization caching and distributed network-coded delivery scheme delivers one packet to all requesting users in $\frac{m}{M}\left(1-\frac{M}{m}\right)$ time slots. This is almost the same as the centralized scheme using the base station by Maddah-Ali and Niesen[42] and achieves the same fundamental scaling law. We also showed that no further scaling law order gain can be obtained if spatial reuse is also exploited. Intuitively, since network coding makes a single (coded) packet useful for as many requesting users as possible, it is better if such a packet is received by all users in the cell, while D2D transmission gets its efficiency from restricting each transmission to a small cluster of nodes. Nevertheless, reducing the transmission range and applying our scheme in clusters, with reuse, yields simpler caching subpacketization and allows transmissions at a higher rate (bit/s/Hz). Hence, the benefits of network-coded multicasting applied to a D2D caching network are reflected in the actual throughput and coding complexity, rather than in the throughput scaling order.

"…the benefits of network-coded multicasting applied to a D2D caching network are reflected in the actual throughput and coding complexity…"

Performance in Realistic Settings

We now present some examples based on simulations in the above-described settings. We first consider the case of a single square cell with $n = 500$ users and 1000 cells, using centralized caching and a protocol model. Figure 5

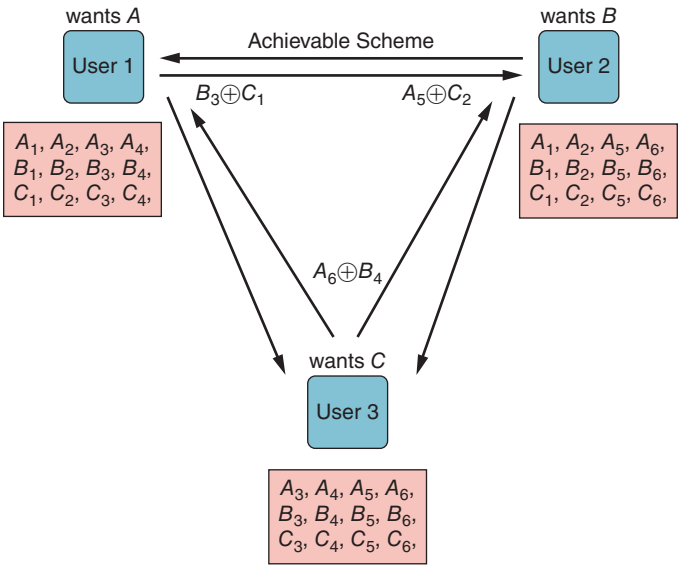


Figure 4: An example of subpacketized caching and network-coded delivery for D2D caching networks. We consider a network with $n = 3$ users, $m = 3$ files (A, B, C) and storage capacity is $M = 2$ files. We divide each file into 6 subpackets (e.g. A is divided into $A_1,\cdots,A_6$.) We let user 1 request A; user 2 requests B and user 3 requests C. The cached subpackets are shown in the rectangles under each user. For the delivery phase, user 1 transmits $B_3 \oplus C_1$; user 2 transmits $A_5 \oplus C_2$ and user 3 transmits $A_6 \oplus B_4$. The normalized number of transmissions is $3 \cdot \frac{1}{6} = \frac{1}{2}$. (Source: Ji et al., 2013.[20] (Copyright IEEE))

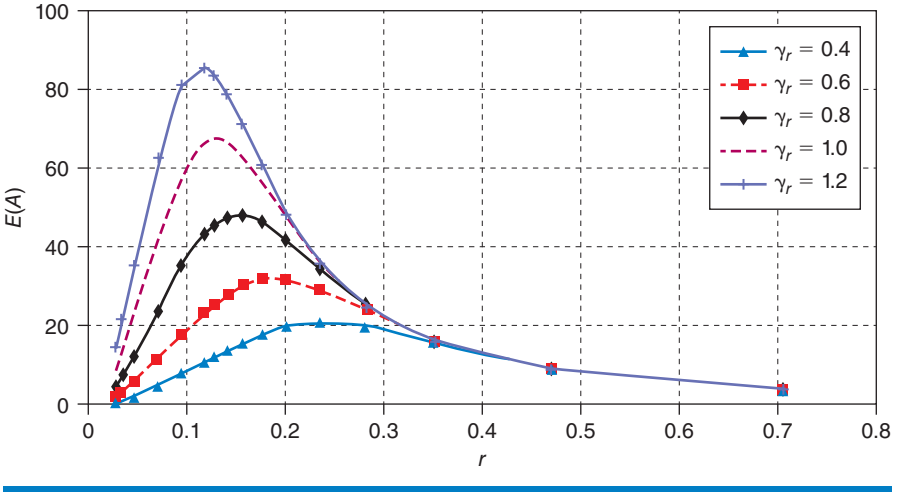


Figure 5: The average number of active clusters versus the collaboration distance for deterministic caching with $n = 500$ and $m = 1000$. (Source: Golrezaei et al., 2012.[18] (Copyright IEEE))

shows the average number of active clusters versus the collaboration distance r (neglecting intercluster interference). We see that the larger γ_r, that is, the more concentrated the request distribution, the smaller the cluster size should be. This is logical, since the probability of finding a desired file even in a small cluster increases with γ_r. Simulations with a more realistic setting (including shadowing and intercluster interference) provided very similar optimal cluster sizes.

We next consider the case of random caching, using a Zipf caching distribution. Figure 6 shows the average number of active clusters versus the collaboration distance r for different values of the exponent of the caching distribution, γ_c. Further simulations showed that increasing γ_r increases the optimum γ_c, since the first few popular files account for the majority of requests and thus to satisfy the users' requests. There is littler need to cache less popular files.

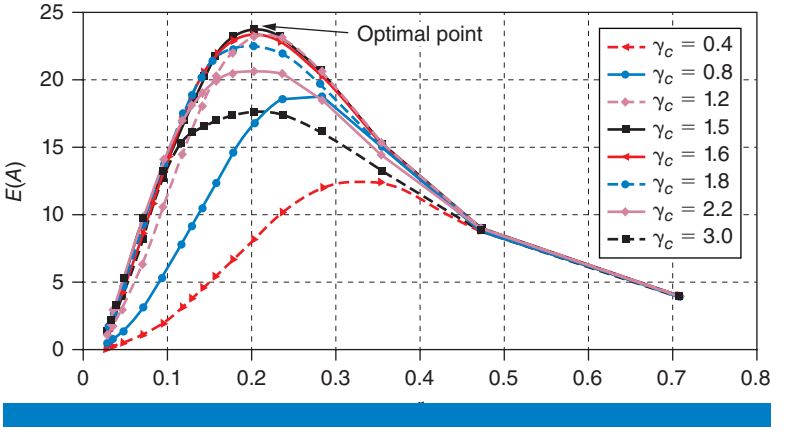


Figure 6: The average number of active clusters versus r for random caching for different γ_c, with $\gamma_r = 0.6$, $n = 500$ and $m = 1000$. (Source: Golrezaei et al., 2012.[18] (Copyright IEEE))

We also performed extensive simulations in more realistic scenarios. We first demonstrate that the throughput of the D2D scheme is markedly (an order of magnitude at low outages) higher than the standard broadcasting from the base station, harmonic broadcasting[47][48][49], and network-coded multicasting.[42] Figure 7 shows an example of such throughput-outage tradeoff performance in a realistic propagation and network topology scenario. We considered indoor office and hotspots environments. Winner models[50] for these wireless propagation channels in these environments were enhanced by models for body shadowing (for more details see Ji et al.[21]), and the capacity for the D2D links was computed based on Shannon's capacity equation. Even in such realistic conditions, the D2D solution provides competitive performance and is significantly simpler to implement than coded multicasting. It is remarkable that while the scaling laws for the coded multicast and D2D schemes are the same, in practical situations the higher capacity of the short-distance links plays a significant role, and a good throughput-outage tradeoff can be achieved even without the use of a BS connection. The good performance of the D2D scheme is tied to the inherent diversity.

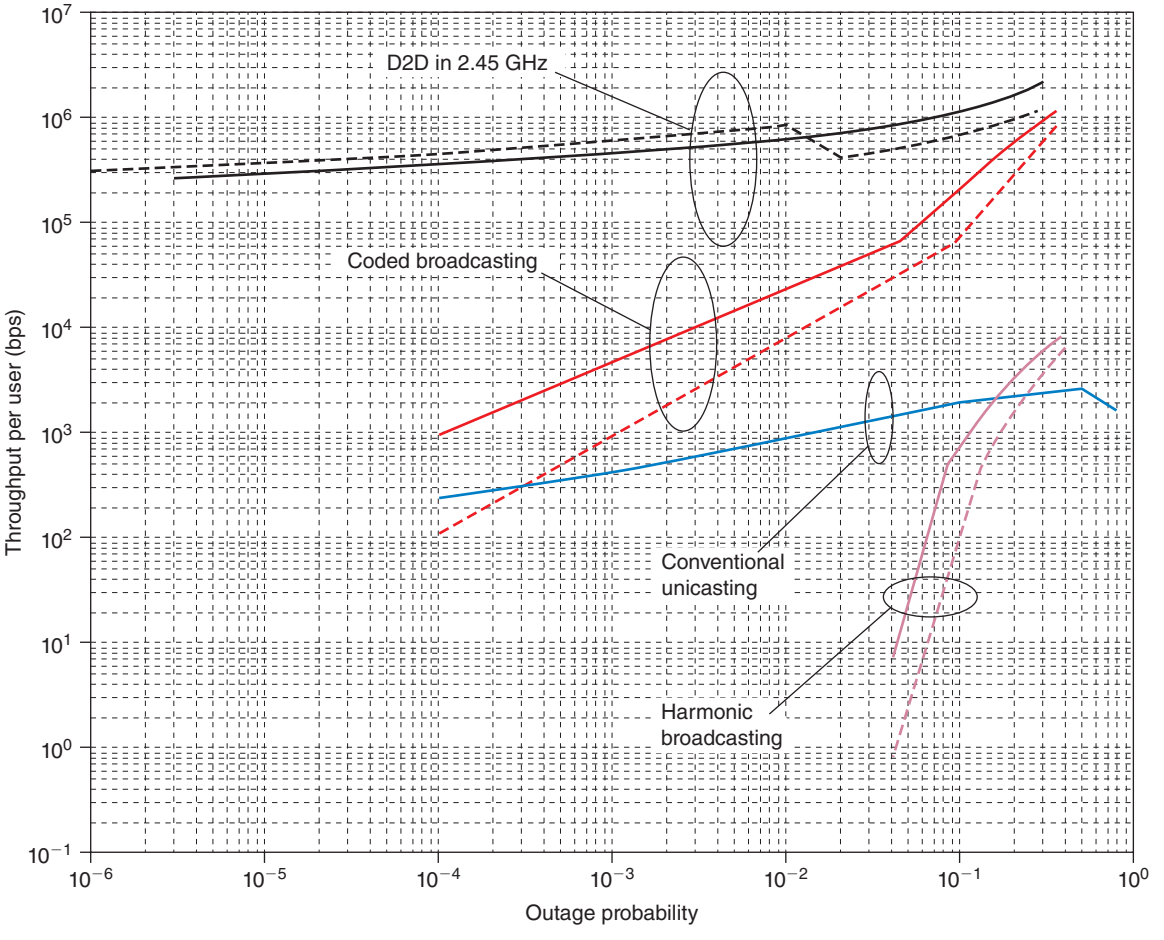


Figure 7: Simulation results for the throughput-outage tradeoff for conventional unicasting, coded multicasting, harmonic broadcasting, and the 2:45 GHz D2D communication scheme under both indoor office and indoor hotspot channel models. Solid lines: indoor office; dashed lines: indoor hotspot. (Source: Ji et al., 2012.[21] (Copyright IEEE))

The theoretical derivations discussed in "D2D Throughput versus Outage" state that the throughput-outage tradeoff does not depend on the number of users or user density as long as n and m are large and $Mn \gg m$. However, the throughput-outage scaling behavior was obtained for the protocol model, where the link capacity for a (feasible) link does not depend on SNR, and thus distance. In practice, the dependence of capacity on distance makes the user density also an important parameter for the system performance. Figure 8 shows that there is a tradeoff between the user density and the throughput, which is caused by two effects: (1) a higher user density allows a smaller cluster size, in turn resulting in shorter links and higher SINR; (2) a small cluster size increases the probability for having LOS interference, which can degrade the performance of the system significantly.

Main Conclusions

From our detailed mathematical investigations, we can draw a number of important conclusions for actual implementation. Firstly, D2D

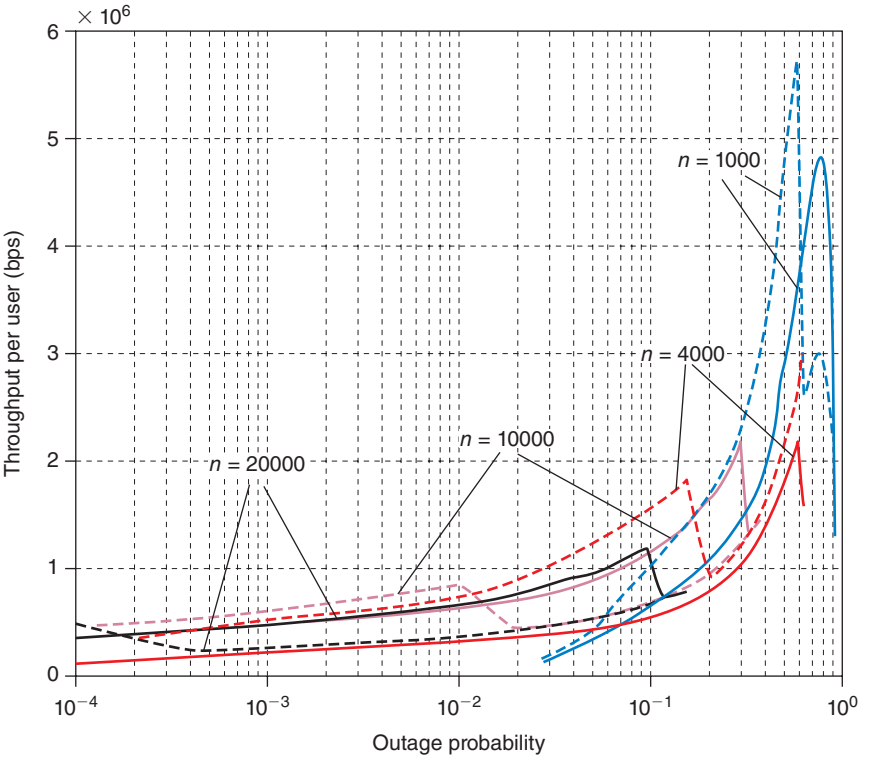


Figure 8: The throughput-outage tradeoff for different user densities. Solid lines: indoor office; dashed lines: indoor hotspot.
(Source: Ji et al., 2012.[21] (Copyright IEEE))

communications with caching on the local devices offers a dramatic increase in the capacity. The actual increase is critically determined by the reuse statistics of the videos. When the library considered by the users has a compact size, then the video throughput per user is independent of the total number of users in a cell. The approach is less useful when the demand distribution is heavy-tailed.

While the mathematical framework allows a detailed analysis, we conjecture that most benefits can be reaped if the "most requested videos" can be stored on 30 devices. From our analysis of upper bounds and achievable schemes, we can conclude that a very simple approach—random caching according to a given probability density function—is scaling-law optimum, and furthermore our numerical simulations showed that the actually achievable numerical values are close to those that can be achieved with an idealized, centralized caching scheme. This tells us not only that a low-complexity implementation is feasible, but also that on-the-fly modifications of caches (for instance, when demand distribution changes) can be done in a simple and straightforward manner. By itself, the D2D scheme might suffer from unacceptably high outage. However, through a suitable combination with transmission from the BS, a very high reliability can be achieved—as a matter of fact, much higher than for a BS alone. Users at a cell edge, who are normally the "problem case," benefit from the possibility of obtaining files from another nearby device. Our simulations

have shown dramatic increase of both reliability and throughput not only with respect to traditional multicast, but also compared with the recently introduced coded multicast, once realistic propagation conditions are taken into account.

A final point that deserves attention is how users can be incentivized to act as caches. In other words, what is the answer to a user asking "why should I use *my* battery to help somebody else get a video more quickly?" In one possible approach, the video file exchange would happen on the basis of reciprocity: only users who provide files (on a time-averaged basis) are also allowed to obtain files. Similar principles have been successfully applied in peer-to-peer file exchanges. Secondly, operators can incentivize users: for example, they might state that files obtained through D2D communications do not count towards the data quota of a particular user (this would be in the interest of operators, since D2D transmissions relieve the pressure on the cellular infrastructure).

Conclusions

"…we have developed a comprehensive framework for caching in wireless networks targeted to on-demand video streaming…"

Supported by the VAWN research initiatives, we have developed a comprehensive framework for caching in wireless networks targeted to on-demand video streaming, which is a killer application for 4G cellular networks, and the root cause of the predicted hundredfold increase in wireless traffic demand in the next five 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.

Besides the problems discussed in this article, a number of interesting problems constitute topics for future research:

- More advanced PHY schemes in femtocaching networks. For example, helper nodes may have multiple antennas and use multiuser MIMO and advanced interference management schemes, beyond the simple Wi-Fi–inspired schemes considered so far.
- More advanced PHY schemes for D2D networks, beyond the simple spectrum reuse and interference avoidance clustering scheme used so far. The optimum scheduling of users is connected to the maximum-independent-set problem, an important (and hard) problem in computer science.
- (Limited) multi-hop in D2D caching networks. Instead of restricting the communication to single-hop, rarer files can be reached through D2D communications without excessive increase in cluster size.
- Transmission schemes that take into account the limitations of existing standards for D2D such as Wi-Fi Direct.
- Analyzing the impact of nonuniform and nonergodic user distributions. Neighbor discovery and channel estimation can be optimized, which is a prerequisite for optimal scheduling of users.

In addition, there is also a large number of interesting problems revolving around the determination and prediction of user request probabilities, the question of incentivizing users to let devices be used for caching purposes, and data authentication and prevention of piracy. While all of these problems seem eminently solvable, as discussions with industry representatives have indicated, further research is needed. We can thus safely say that while our research has shown the feasibility and enormous promise of femtocaching and D2D communications, much remains to be done for a more complete understanding.

References

[1] Cisco, "The Zettabyte Era-Trends and Analysis," 2013.

[2] Dahlman, E., S. Parkvall, and J. Skold, *4G: LTE/LTE-Advanced for Mobile Broadband: LTE/LTE-Advanced for Mobile Broadband*. Academic Press, 2011.

[3] Dohler, M., R. Heath, A. Lozano, C. Papadias, and R. Valenzuela, "Is the phy layer dead?" *Communications Magazine, IEEE*, vol. 49, no. 4, pp. 159–165, 2011.

[4] Chandrasekhar, V., J. Andrews, and A. Gatherer, "Femtocell networks: a survey," *Communications Magazine, IEEE*, vol. 46, no. 9, pp. 59–67, 2008.

[5] Madan, R., J. Borran, A. Sampath, N. Bhushan, A. Khandekar, and T. Ji, "Cell association and interference coordination in heterogeneous lte-a cellular networks," *Selected Areas in Communications, IEEE Journal on*, vol. 28, no. 9, pp. 1479–1489, 2010.

[6] "Video-Aware Wireless Networks" http://software.intel.com/en-us/articles/video-aware-wirelessnetworks.

[7] Golrezaei, N., A. G. Dimakis, A. F. Molisch, and G. Caire, "Wireless video content delivery through distributed caching and peer-to-peer gossiping," in *Signals, Systems and Computers (ASILOMAR), 2011 Conference Record of the Forty Fifth Asilomar Conference on*. IEEE, 2011, pp. 1177–1180.

[8] Golrezaei, N., M. Ji, A. F. Molisch, A. G. Dimakis, and G. Caire, "Device-to-device communications for wireless video delivery," in *Forty Sixth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*. IEEE, 2012, pp. 930–933.

[9] Golrezaei, N., A. F. Molisch, A. G. Dimakis, and G. Caire, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *Communications Magazine, IEEE*, vol. 51, no. 4, pp. 142–149, 2013.

[10] Golrezaei, N., K. Shanmugam, A. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless video content delivery through distributed caching helpers," in *INFOCOM, 2012 Proceedings IEEE*, 2012, pp. 1107–1115.

[11] Golrezaei, N., A. Dimakis, and A. F. Molisch, "Wireless device-to-device communications with distributed caching," in *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*. IEEE, 2012, pp. 2781–2785.

[12] Shanmugam, K., N. Golrezaei, A. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless content delivery through distributed caching helpers," *Information Theory, IEEE Transactions on*, vol. 59, no. 12, pp. 8402–8413, 2013.

[13] Kim, J., F. Meng, P. Chen, H. E. Egilmez, D. Bethanabhotla, A. F. Molisch, M. J. Neely, G. Caire, and A. Ortega, "Adaptive video streaming for device-to-device mobile platforms," in *Proceedings of the 19th annual international conference on Mobile computing & networking*. ACM, 2013, pp. 127–130.

[14] Bethanabhotla, D., G. Caire, and M. Neely, "Joint transmission scheduling and congestion control for adaptive video streaming in small-cell networks," *arXiv preprint arXiv:1304.8083*, 2013.

[15] Bethanabhotla, D., G. Caire, and M. J. Neely, "Adaptive video streaming in MU-MIMO networks," submitted to ISIT 2014, Jan. 2014.

[16] Golrezaei, N., A. G. Dimakis, and A. F. Molisch, "Wireless device-todevice communications with distributed caching," in *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*. IEEE, 2012, pp. 2781–2785.

[17] Golrezaei, N., A. G. Dimakis, and A. F. Molisch, "Device-to-device collaboration through distributed storage," in *Global Communications Conference (GLOBECOM), 2012 IEEE*. IEEE, 2012, pp. 2397–2402.

[18] Golrezaei, N., A. F. Molisch, and A. G. Dimakis, "Base-station assisted device-to-device communications for high-throughput wireless video networks," in *Communications (ICC), 2012 IEEE International Conference on*. IEEE, 2012, pp. 7077–7081.

[19] Ji, M., G. Caire, and A. F. Molisch, "Fundamental limits of distributed caching in D2D wireless networks," in *Information Theory Workshop (ITW), 2013 IEEE*, 2013, pp. 1–5.

[20] Ji, M., G. Caire, and A. F. Molisch, "The throughput-outage tradeoff of wireless one-hop caching networks," *arXiv preprint arXiv:1312.2637*, 2013.

[21] Ji, M., G. Caire, and A. F. Molisch, "Wireless device-to-device caching networks: Basic principles and system performance," *arXiv preprint arXiv:1305.5216*, 2012.

[22] Nygren, E., R. K. Sitaraman, and J. Sun, "The akamai network: a platform for high-performance internet applications," *ACM SIGOPS Operating Systems Review*, vol. 44, no. 3, pp. 2–19, 2010.

[23] Luo, F.-L., *Mobile Multimedia Broadcasting Standards: Technology and Practice*. Springer Verlag, 2008.

[24] Reimers, U., "Digital video broadcasting," *Communications Magazine, IEEE*, vol. 36, no. 6, pp. 104–110, 1998.

[25] Ladebusch, U. and C. Liss, "Terrestrial dvb (dvb-t): A broadcast technology for stationary portable and mobile use," *Proceedings of the IEEE*, vol. 94, no. 1, pp. 183–193, 2006.

[26] Bursalioglu, O., M. Fresia, G. Caire, and H. Poor, "Lossy multicasting over binary symmetric broadcast channels," *Signal Processing, IEEE Transactions on*, vol. 59, no. 8, pp. 3915–3929, 2011.

[27] Li, Y., E. Soljanin, and P. Spasojevi´c, "Three schemes for wireless coded broadcast to heterogeneous users," *Physical Communication*, 2012.

[28] Gao, Q., M. Chari, A. Chen, F. Ling, and K. Walker, "MediaFLO technology: FLO air interface overview," in *Mobile Multimedia Broadcasting Standards*. Springer, 2009, pp. 189–220.

[29] Cha, M., H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system," in *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, ser. IMC '07. ACM, 2007, pp. 1–14.

[30] "http://traces.cs.umass.edu/index.php/network/network."

[31] Zambelli, A., "Iis smooth streaming technical overview," *Microsoft Corporation*, vol. 3, 2009.

[32] Begen, A., T. Akgul, and M. Baugher, "Watching video over the web: Part 1: Streaming protocols," *Internet Computing, IEEE*, vol. 15, no. 2, pp. 54–63, 2011.

[33] Stockhammer, T., "Dynamic adaptive streaming over http–: standards and design principles," in *Proceedings of the second annual ACM conference on Multimedia systems*. ACM, 2011, pp. 133–144.

[34] Sánchez de la Fuente, Y., T. Schierl, C. Hellge, T. Wiegand, D. Hong, D. De Vleeschauwer, W. Van Leekwijck, and Y. Le Louédec, "idash: improved dynamic adaptive streaming over http using scalable

video coding," in *Proceedings of the second annual ACM conference on Multimedia systems*. ACM, 2011, pp. 257–264.

[35] Pancha, P. and M. El Zarki, "Mpeg coding for variable bit rate video transmission," *Communications Magazine, IEEE*, vol. 32, no. 5, pp. 54–66, 1994.

[36] Wang, Z., A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *Image Processing, IEEE Transactions on*, vol. 13, no. 4, pp. 600–612, 2004.

[37] Pawar, S., N. Noorshams, S. El Rouayheb, and K. Ramchandran, "Dress codes for the storage cloud: Simple randomized constructions," in *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*. IEEE, 2011, pp. 2338–2342.

[38] Bethanabhotla, D., G. Caire, and M. J. Neely, "Adaptive video streaming in mu-mimo networks," *CoRR*, vol. abs/1401.6476, 2014.

[39] Joseph, V. and G. de Veciana, "Nova: Qoe-driven optimization of dash-based video delivery in networks," *CoRR*, vol. abs/1307.7210, 2013.

[40] Molisch, A. F., *Wireless Communications*. Second Edition, IEEE Press – John Wiley & Sons, 2011.

[41] Gupta, P. and P. Kumar, "The capacity of wireless networks," *Information Theory, IEEE Transactions on*, vol. 46, no. 2, pp. 388–404, 2000.

[42] Maddah-Ali, M. and U. Niesen, "Fundamental limits of caching," *arXiv preprint arXiv:1209.5807*, 2012.

[43] Maddah-Ali, M. and U. Niesen, "Decentralized caching attains order-optimal memory-rate trade-off," *arXiv preprint arXiv:1301.5848*, 2013.

[44] El Rouayheb, S., A. Sprintson, and C. Georghiades, "On the index coding problem and its relation to network coding and matroid theory," *Information Theory, IEEE Transactions on*, vol. 56, no. 7, pp. 3187–3195, 2010.

[45] Sun, H. and S. A. Jafar, "Index Coding Capacity: How far can one go with only Shannon Inequalities?" *ArXiv e-prints*, Mar. 2013.

[46] Shanmugam, K., A. G. Dimakis, and M. Langberg, "Local Graph Coloring and Index Coding," *ArXiv e-prints*, Jan. 2013.

[47] Juhn, L. and L. Tseng, "Harmonic broadcasting for video-on-demand service," *Broadcasting, IEEE Transactions on*, vol. 43, no. 3, pp. 268–271, 1997.

[48] Pâris, J.-F., S. Carter, and D. Long, "Efficient broadcasting protocols for video on demand," in *Modeling, Analysis and Simulation of Computer and Telecommunication Systems, 1998. Proceedings. Sixth International Symposium on. IEEE,* 1998, pp. 127–132.

[49] Engebretsen, L. and M. Sudan, "Harmonic broadcasting is bandwidth-optimal assuming constant bit rate," *Networks*, vol. 47, no. 3, pp. 172–177, 2006.

[50] WINNER-II, "D1. 1.2, WINNER II channel models," 2007.

Author Biographies

Giuseppe Caire was born in Torino, Italy, in 1965. He received a B.Sc. in Electrical Engineering from Politecnico di Torino (Italy) in 1990, a M.Sc. in Electrical Engineering from Princeton University in 1992, and a PhD from Politecnico di Torino in 1994. He has been assistant professor at the Politecnico di Torino, associate professor at the University of Parma, professor with the Department of Mobile Communications at the Eurecom Institute, Sophia-Antipolis, France, and he is currently a professor of Electrical Engineering at the University of Southern California, Los Angeles and an Alexander von Humboldt Professor at the Technical University of Berlin, Germany. He received the Jack Neubauer Best System Paper Award from the IEEE Vehicular Technology Society in 2003, and the IEEE Communications Society & Information Theory Society Joint Paper Award in 2004 and 2011. Giuseppe Caire has been a Fellow of IEEE since 2005. He has served in the Board of Governors of IEEE Information Theory and was the Society President of the IEEE Information Theory Society in 2011. His main research interests are in the field of communications and information theory. He can be reached at caire@usc.edu.

Andreas F. Molisch is professor of Electrical Engineering and Director of the Communication Sciences Institute (CSI) at the University of Southern California. He previously was at TU Vienna, AT&T (Bell) Labs, Lund University, and Mitsubishi Electric Research Labs. His research interests are novel cellular architectures, wireless propagation channel measurements and modeling, ultrawideband localization and communication, and multi-antenna systems. He has authored 4 books, 16 book chapters, 170 journal papers, and numerous conference papers as well as 70 patents and 60 standards contributions. He is a Fellow of IEEE, AAAS, and IET, and a member of the Austrian Academy of Sciences, as well as recipient of numerous awards. He can be contacted at molisch@usc.edu.

VIDEO DELIVERY OVER WIRELESS NETWORKS: EXPLOITING NETWORK HETEROGENEITY AND CONTENT COMMONALITY

Contributors

Salman Avestimehr
University of Southern California

Tsuhan Chen
Cornell University

Sina Lashgari
Cornell University

Amandianeze Nwana
Cornell University

Md. Saifur Rahman
Cornell University

Sinem Unal
Cornell University

Aaron B. Wagner
Cornell University

We discuss three opportunities to improve the efficiency of video transmission over wireless networks. First, we consider the opportunities that arise from exploiting network heterogeneity, in particular from access to multiple network connections at the users. We consider a new metric of timely throughput that captures the strict per-packet deadline requirement of real-time video traffic, and develop communication protocols that maximize the timely throughput of heterogeneous wireless networks. Next, we consider the gains that can be obtained by multicasting content that is being simultaneously demanded by several users at a single base station. We demonstrate that if multiple users, who are associated with the same Wi-Fi* access point, request the same video at around the same time, then one can substantially reduce the amount of downlink traffic by combining the transmissions instead of using separate unicasts. Finally, we focus on the potential gains from being able to correctly predict what video is going to be watched by users in a network. We postulate that in most networks, the pattern of video requests would resemble that of viral infections in a given population. Using such approximation, we develop new algorithms for predicting future requests of users, and utilizing these predictions for caching.

Introduction

Consumer demand for data services over wireless networks has increased dramatically in recent years, fueled both by the success that Internet streaming has achieved in displacing traditional TV/radio broadcasts and by advances in smartphones, tablets, and netbooks, which have enabled wider viewing patterns that are unconstrained by location or time. These trends are expected to continue as mobile devices evolve and consumers fully embrace Internet video as a primary source of information, communication, and entertainment. This confluence of trends is expected to lead to a 65-fold increase in traffic to mobile devices, the majority of which is expected to be video.

This dramatic increase in demand will severely stress our current wireless infrastructure. The near-term adoption of 4G and near-4G technologies such as LTE, while significantly improving the capacity of wireless networks, is not expected to meet this anticipated consumer demand. While follow-on technologies will certainly yield additional improvements in spectral efficiency, they will probably be inadequate to meet the quickly growing demand for data services and will not be available for several years to come. Meeting this challenge therefore requires finding ways of providing users an equivalent viewing experience for streaming video while placing less demand on the existing network infrastructure, especially during peak periods.

Fortunately, there are many opportunities to improve the efficiency of video transmission over wireless networks while working within the general framework of our existing infrastructure. First, existing mobile devices use only one type of wireless network, such as Wi-Fi or cellular, at a time. By using all available networks simultaneously, a mobile device could, for example, combine the throughput of Wi-Fi networks with the delay guarantees of cellular networks and thereby realize the best aspects of each. Network operators can also exploit such opportunities to balance the traffic between the networks and to survive Internet data explosion in mobile networks. Second, existing networks do not multicast; even if multiple users associated with the same base station view the same video, the network will unicast separate copies to each user for reasons explained below. Removing this inherent redundancy promises to reduce network traffic precisely when it is most needed—when there are many active users. Third, the demand pattern for video content exhibits temporal burstiness, while at the same time, much of the video traffic is predictable from past viewing patterns. Using off-peak periods to speculatively stream videos would reduce peak traffic demands by improved load balancing.

"Fortunately, there are many opportunities to improve the efficiency of video transmission over wireless networks…"

Our goal in this article is to provide an in-depth discussion on the aforementioned three opportunities for video delivery in wireless networks and demonstrate a quantitative analysis of the capacity gains that they can provide. To that end, the article consists of three main sections, which are as follows.

"Our goal in this article is to provide an in-depth discussion on opportunities for video delivery in wireless networks…"

In the first section of this article, we consider the opportunities that arise from exploiting network heterogeneity for video delivery. With the proliferation of a variety of wireless access technologies and the evolution of wireless networks towards heterogeneous architectures, users are often in the connectivity range of several wireless networks. Hence, it is promising that the opportunistic utilization of multiple connections at the users is going to be one of the key solutions to help cope with the phenomenal growth of video demand in mobile devices. We will study optimal strategies that exploit network diversity for delivering real-time video traffic. We will consider a new metric of *timely throughput* that captures the strict per-packet deadline requirement of real-time video traffic, and develop communication protocols that maximize the timely throughput of heterogeneous wireless networks.

In the second section, we consider the gains that can be obtained by multicasting content that is being simultaneously demanded by several users at a single base station. We focus in particular on wireless video delivery in situations that are highly congested, such as sports stadiums, cafés, airports, and train stations. For these venues, we expect a larger amount of overlap among the videos demanded by the users compared with general networks. In a stadium, for instance, spectators all share an interest in a particular sport and even one particular team (or at most two teams). We would thus expect them to be requesting many of the same videos around the same time. The stadium owners could even encourage this by streaming replays or highlights over stadium-installed Wi-Fi access points. At the same time, the inherent

broadcast nature of the wireless medium means that if multiple users, who are associated with the same Wi-Fi access point, request the same video at around the same time, then one can reduce the amount of downlink traffic by combining the transmissions instead of using separate unicasts. This can provide substantial rate benefits exactly where they are needed most, as we shall see.

Finally, in the last section we focus on the potential gains from being able to correctly predict what video is going to be watched by whom at a given time with certain probability. Our work here focuses on the social nature of the requests being made within a particular network, say a campus. We postulate that in most networks, the pattern of video requests would resemble that of viral infections in a given population. Given that an epidemic model is a good approximation of the complex word-of-mouth information propagation, we propose to model the video request patterns as such. Assuming a given epidemic model, we infer the social/influence graph between the users in the network based on historical requests and predict future requests of users by applying the social graph to requests to-date to obtain the probability some user will watch a certain video. We analyze this prediction in the context of caching, and obtain very promising results as shown later in the article.

Exploiting Network Heterogeneity for Video Delivery

"…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."

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 (such as cellular and Wi-Fi), 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 highlights an important problem: how to optimally utilize network heterogeneity for the delivery of video traffic.

In this section, 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 packets. In particular, the packets for such applications have strict per-packet deadlines, 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 by Hou et al.[1][2], which measures the long-term average number of "successful deliveries" (that is, 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 (APs) and M clients, where each client is connected to several out-of-band APs and requests timely traffic. We study the maximum total timely throughput of the network, denoted by C_{T^3}, 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 APs in the network.

This is a very challenging scheduling problem since, at each interval, the number of possible assignments of clients to APs, 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 we 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. We demonstrate the performance of our proposed algorithm via numerical results, and compare it with algorithms that do not take the deadline into account and try to balance the traffic load between different APs.

"…we propose a deterministic relaxation of the problem…"

We also extend the network model to consider fading channels. In addition, we allow each AP to allocate a portion of its available bandwidth to each client during a timeslot. This means that each AP can now access several clients simultaneously. Moreover, we allow for rate adaptation, where according to the time-frequency resource allocated to each client, a certain reward will be obtained.

Model Setup and Problem Formulation

We consider the downlink of a network with M wireless clients, denoted by $Rx_1,\ldots,Rx_M$, and N access points, denoted by $AP_1,\ldots, AP_N$. All APs are connected via reliable links to the backhaul network (see Figure 1).

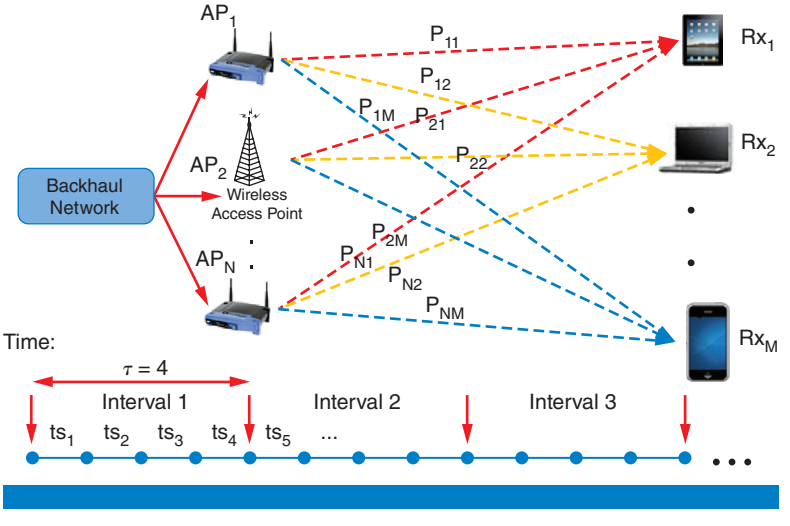


Figure 1: Network Model
(Source: IEEE Transactions on Information Theory, Vol 59, No 2, 2013)

As shown in Figure 1, time is slotted and transmissions occur during timeslots. Furthermore, timeslots are grouped into intervals of length τ, corresponding to the inter-arrival time of the packets (for example, 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 independent and identically distributed (i.i.d.) over time, and have fixed, but possibly different success probabilities. (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 by Lashgari and Avestimehr[3]). 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 timeslot. 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 APs for delivery during that interval. At each timeslot (of that interval), each AP can choose any of the packets that are assigned to it for transmission. At the end of that timeslot the AP will know whether the packet has been successfully delivered. 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 APs…"

The *scheduling policy* determines how the packets are assigned to APs at the beginning of each interval, and which packet each AP transmits at each timeslot. 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 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,

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

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

Main Problem (MP): $C_{T^3} \triangleq \sup_{\eta \in S} T^3(\eta)$,

and the corresponding optimal policy. Characterizing C_{T^3} is challenging, since the dimension of the corresponding optimization problem at each interval (that is, the number of all possible assignments) grows exponentially in 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 for example, Shmoys and Tardos[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 nearly optimal schedule.

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 timeslots expended for this packet to be delivered is a geometric random variable with parameter p_{ij}, with 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 (MP), 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 timeslots 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, that is, 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, for example, Shmoys and Tardos[5]).

"…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."

Relaxed Problem (RP):

$$C_{det} \triangleq \max \sum_{i=1}^{N} \sum_{j=1}^{M} x_{ij}$$

$$s.t. \sum_{j=1}^{M} \frac{x_{ij}}{P_{ij}} \leq \tau$$

$$\sum_{i=1}^{N} x_{ij} \leq 1$$

$$x_{ij} \in \{0,1\}.$$

Main Results

Our main contribution in this article is threefold. First we show that the main problem (MP) can be approximated via its deterministic relaxation, discussed

above. This is stated in the following theorem (the reader is referred to Lashgari and Avestimehr[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 APs (that is, 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 (that is, C_{det}). Moreover, the approximation gap in Theorem 1 is a worst-case bound, and via numerical analysis[3] one can observe that the gap between the main problem (MP) 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.

"…the deterministic relaxation of the main problem can be solved efficiently via a linear-programming (LP) relaxation…"

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 an N-by-M binary matrix where element (i,j) is 1 if and only if the packet requested by Rx_j is packed in bin i. Suppose that the matrix $x^* = [x_{ij}^*]$ is a basic optimal solution to the LP relaxation of RP. Then,

$$C_{det} - \sum_{i=1}^{N}\sum_{j=1}^{M} [x_{ij}^*] \leq N.$$

Note that finding a basic optimal solution to an LP efficiently is straightforward.[6] Hence, Theorem 2 also 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. Later, in the numerical analysis section of this article, we show how our algorithms based on deterministic relaxation of the problem outperform other algorithms that do not consider the deadlines for packets.

So far we considered the same importance for all the flows in the network, and our objective was to maximize T^3. However, it might be the case that in a network some of the flows are more important than the others and should be prioritized accordingly.

Therefore, the objective function should maximize a weighted average of timely throughputs. In particular, weighted total timely throughput of the scheduling policy η, $w - T^3(\eta)$, is defined as

$$w - T^3(\eta) \triangleq \sum_{j=1}^{M} \omega_j R_j(\eta),$$

where ω_j's are arbitrary weights greater than 1, and $R_j(\eta)$ is the timely throughput of Rx_j. Our objective is to find

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

For this extension of the problem we again propose the channel relaxation, which results in a new integer program. This integer program is again a GAP.

The formulation of the relaxed problem is as follows:

Weighted RP:

$$C_{w-det} \triangleq \max \sum_{i=1}^{N}\sum_{j=1}^{M} \omega_j x_{ij}$$

$$s.t. \sum_{j=1}^{M} \frac{x_{ij}}{P_{ij}} \leq \tau$$

$$\sum_{i=1}^{N} x_{ij} \leq 1$$

$$x_{ij} \in \{0,1\}.$$

The following theorem[3] implies that the value of the solution to C_{w-det} is asymptotically the same as the value of the solution to C_{w-T^3} as $C_{w-T^3} \to \infty$ (or equivalently $C_{w-det} \to \infty$).

Theorem 3.

$$-2\omega_{max}\sqrt{N\left(C_{w-det}+\frac{N}{4}\right)} < C_{w-T^3} - C_{w-det} < N\omega_{max},$$

where $\omega_{max} \triangleq \max\{\omega_1,..., \omega_M\}$.

Extensions: Fading Channels and Rate Adaptation

In the section "Model Setup and Problem Formulation" we considered a packet erasure model for channels and assumed that each AP can transmit one packet at a time. We extend the model to consider fading channels in order to better capture the channel physical properties. In addition, we allow each AP to allocate a portion of its available bandwidth to each client during a timeslot. This means that each AP can access several clients simultaneously. Moreover, we allow for rate adaptation, where according to the time-frequency resource allocated to each client, a certain reward will be obtained.

"We extend the model to consider fading channels in order to better capture the channel physical properties."

Consider the network topology and time model described in the section "Model Setup and Problem Formulation." In addition, assume that for $i \in \{1,..., N\}$, AP_i has bandwidth W_i, where $W_i \in \mathbb{N}$, which means at most W_i simultaneous transmissions can occur during a timeslot by AP_i. On the other hand, all the bandwidth of AP_i during a timeslot might be allocated to a certain client.

Define $R_j^{i_1,\ldots,i_N}$ to be the total reward obtained by Rx_j during an interval if it is served $i_1,\ldots,i_N$ times on AP_1, AP_2,…, AP_N, respectively. The amount of this reward is determined by the rate adaptation that is used in the APs. Further, assume that $R_j^{i_1,\ldots,i_N}$ for $j = 1,2,\ldots,M$ is a nonnegative, increasing the function in all dimensions $i_1,\ldots,i_N$.

A scheduling policy η for the system allocates, possibly at random, the bandwidth of each AP to different clients in each timeslot, based on the past history of the system.

Let $q_j(k)$ denote the reward obtained for client Rx_j during interval k under some scheduling policy. The average reward for Rx_j is then defined as

$$q_j = limsup_{k\to\infty} \frac{\sum_{i=1}^{k} q_j(i)}{k}.$$

The objective is to maximize $\Sigma_{j=1}^{M} q_j$, which is the total average reward.

The relaxed problem introduced in the section "Deterministic Relaxation" was in fact a deterministic scheduling problem with binary rewards; that is, either size $\frac{1}{P_{ij}}$ would be allocated to packet of client Rx_j in bin i, which would result in reward one (it will contribute to the objective function by setting $x_{ij} = 1$); or, it would not add to the value of the objective function at all (for $x_{ij} = 0$). Therefore, the value of $\Sigma_{i=1}^{N}\Sigma_{j=1}^{M} x_{ij}$ can be viewed as the total reward resulting from a scheduling policy. Nevertheless, a more practical model for the reward is a function with the input argument being the amount of time frequency allocated to the client. Therefore, the model extension we are considering can also be viewed as a generalization of the deterministic scheduling (RP).

"…the model extension we are considering can also be viewed as a generalization of the deterministic scheduling (RP)."

A similar model has been considered[7] for $N = 1$, where no simultaneous transmissions are allowed, that is, the bandwidth of AP is equal to 1, and intervals for clients are not necessarily equal. They show that for checking if a set of reward requirements is feasible, it is sufficient to look at the average behavior of the system. However, when going from one AP to multiple APs, checking the average behavior is not sufficient, even when multiple simultaneous transmissions are not allowed, and all deadlines are equal.

We focus on maximizing the total average reward, which is the equivalent of C_{T^3} in our original model. To this aim, we first observe that due to the network model being stationary across time, it is sufficient to focus on finding the scheduling that maximizes the total reward over a single interval of length τ and then apply that scheduler to all intervals.[3] Then, we show that the new maximization problem (that is, total reward maximization over one interval) can be solved using dynamic programming. In particular, we define $OPT[m, t_1, \ldots, t_N]$ to be the maximal total reward obtained when only scheduling the first m clients, with the available resource being $t_1, \ldots, t_N$ on AP_1, AP_2,…, AP_N, respectively. Hence, our objective is to find $OPT[M, W_1\tau, \ldots, W_N\tau]$. To this aim, for each $1 \le m \le M$, we iteratively calculate $OPT[m, \ldots]$ based on the values of $OPT[m-1, \ldots]$, which are previously calculated (details of the dynamic programming and analysis of

"…we show that the new maximization problem can be solved using dynamic programming."

its performance can be found in Lashgari and Avestimehr[3]). The following theorem addresses the performance of this dynamic programming algorithm.

Theorem 4. The above Dynamic Programming algorithm solves the problem of finding the maximum total average reward in $O(M\tau^{2N}\Pi_{i=1}^{N} W_i^2)$ time.

Numerical Results

In this section, we numerically demonstrate the performance of our proposed timely-throughput maximization approach, compared to two other algorithms. We consider the network in Figure 2, where there is 1 base station (BS) in the middle, and 4 femto base stations (APs) around it, and a total of 100 receivers randomly located in the coverage area of BS and/or femtocells (APs). Finally, the erasure probabilities of the channels are proportional to the AP-client distances (capped at 1).

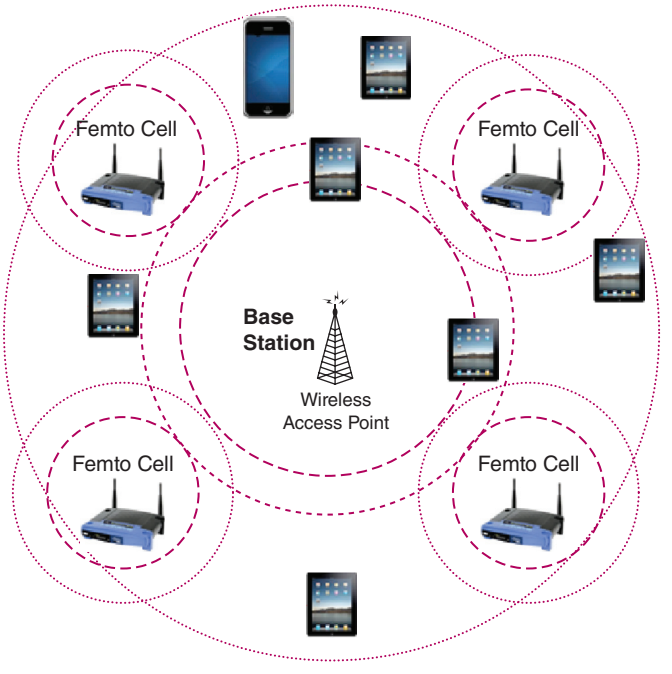


Figure 2: A network with 5 transmitters and 100 clients (Source: Cornell University, 2013)

We consider three different scheduling algorithms. In the first algorithm (the greedy algorithm) each receiver connects to the transmitter that provides it with the strongest channel (that is, the least erasure probability). The second algorithm (the deterministic scheduling algorithm or DS), which is our proposed algorithm, considers a deterministic relaxation of the original stochastic problem (see the section "Deterministic Relaxation"), which is a mixed integer linear program, then solves its LP relaxation, and rounds the solution to obtain a scheduling policy that assigns clients to APs accordingly. Finally, the third algorithm (the load balancing algorithm or LB) ignores packet deadlines, and in order to find a scheduling policy, it assumes that for each unit of time spent by AP_i to serve Rx_j,

a utility equal to p_{ij} is obtained. LB then tries to maximize the total utility of the network subject to some load balancing constraints. In particular, one constraint is that the summation of fractions of times allocated to different clients at the BS must not exceed a certain value L_1. Furthermore, summation of fractions of times allocated to different receivers at an AP must not exceed a certain value L_2. Note that if the coverage radius of a transmitter (for example BS) is larger, it means the transmitter can in general provide a higher channel success probability; therefore it can deliver packets with less time spent. Hence, the BS should intuitively have more capacity (a larger L_1) to accept traffic load than the APs. We consider the case that the ratio of L_1 and L_2 is equal to the ratio of the radius of coverage of BS (radius BS) and the radius of each AP (radius AP). Hence,

$$L_1 = M \times \frac{radius\ BS}{(radius\ BS + 4 \times radius\ AP)}.$$

Similarly,

$$L_2 = M \times \frac{radius\ AP}{(radius\ BS + 4 \times radius\ AP)}.$$

For 20 different realizations of the network we run the network for 200 intervals, where each interval is assumed to contain 30 timeslots. Moreover, we assume intervals are grouped into periods, where each period contains 10 intervals. If a receiver does not receive at least 50 percent of its packets during a period, it will be considered to be in outage during that period.

Figure 3 demonstrates the percentage of intervals during which clients are in outage for the aforementioned three algorithms. As we note, our proposed

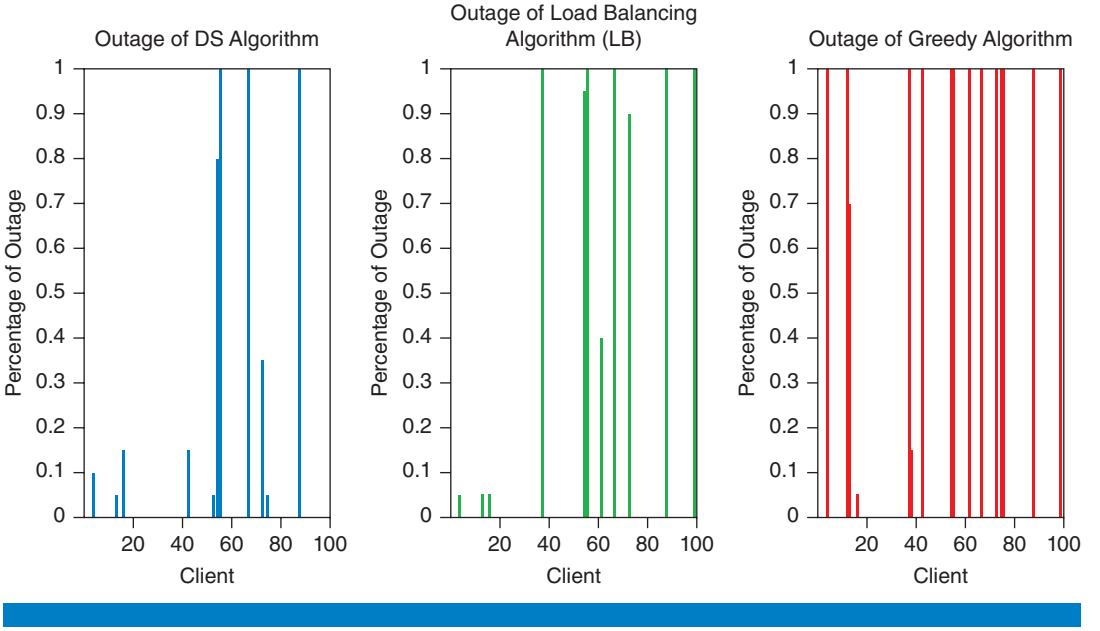


Figure 3: Outage for clients under three scheduling algorithms. Average client outage is as follows: DS (4.7%), LB (7.4%), greedy (13.9%)
(Source: Cornell University, 2013)

timely-throughput maximization approach (DS) utilizes APs most efficiently for traffic delivery and results in a 1.6x gain compared to the load balancing algorithm (LB), and a 3x gain compared to the greedy algorithm in terms of improvement in the percentage of clients in outage.

Summary

In this section, we studied the problem of utilizing network heterogeneity in order to improve the QoS for delivering video traffic to wireless users. In particular, we focused on maximizing timely throughput of heterogeneous wireless networks, in which wireless clients request delivery of time-sensitive traffic and have access to (potentially) multiple access points. We proposed a deterministic relaxation of the stochastic problem to efficiently approximate the optimal solution.

"…we studied the problem of utilizing network heterogeneity in order to improve the QoS for delivering video traffic to wireless users."

In addition, we provided numerical results on the performance of our proposed algorithm and compared it with other algorithms that do not take packet deadlines into account. In particular, we illustrated how our algorithm outperforms (i) the common greedy algorithm, in which each receiver is connected to the transmitter with the strongest signal, by a factor of 3 in terms of percentage of clients in outage, and (ii) the load balancing algorithm, which balances the traffic load across transmitters without taking into account the strict deadlines of the packets, by a factor of 1.6 in terms of percentage of clients in outage. Furthermore, the numerical results also indicated that in terms of total timely throughput our proposed algorithm outperforms the other two algorithms.

Exploiting Content Commonality for Video Delivery

Multicast is a known method for serving common, synchronous content to multiple users, which requires less downlink bandwidth than conventional unicast. Multicast has been widely studied for use in transmitting live events to many users. In the stadium application mentioned in the introduction, however, we expect that many of the users will be demanding the same content but doing so *asynchronously*. For instance, many spectators may wish to watch replays of events that happened earlier in the game or highlights of other games occurring simultaneously. Generally the users will expect to watch these videos *on demand*, that is, asynchronously from other users.

"Multicast is a known method for serving common, synchronous content to multiple users…"

How can one reap some of the benefits of multicasting for on-demand content? Clearly one cannot simply stream a single copy of the video to all of the users, due to their asynchrony. Yet one expects that it should be possible to do better than simply using separate unicasts. Consider, for example, a single Wi-Fi access point, or cellular base station, and a single video streamed from a server that is on a wired network but yet close to the base station. This is rendered in Figure 4. Suppose that clients arrive at the base station randomly and asynchronously and would like to watch the video from the beginning. If two clients enter, the second one somewhat after the first, then we can use the inherent broadcast nature of the wireless medium to reduce the amount of

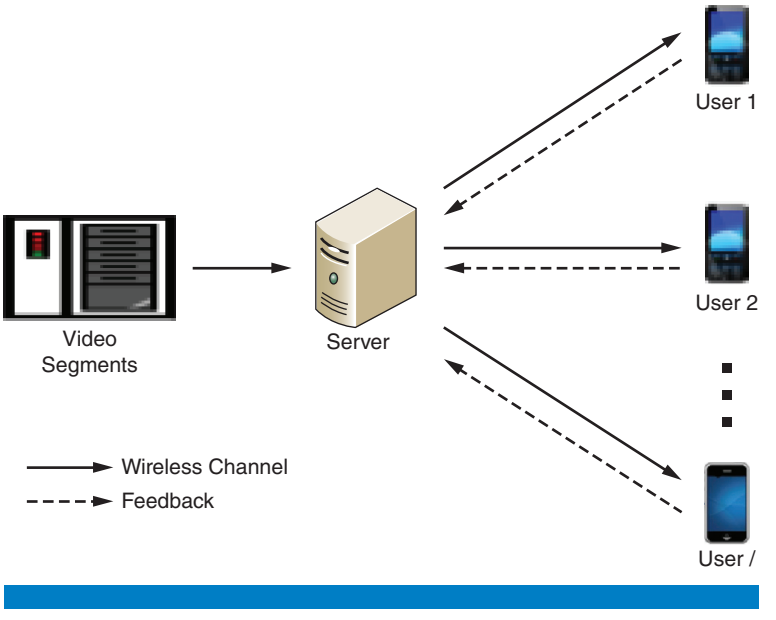


Figure 4: The setup considered for multicasting video-on-demand (Source: Md. S. Rahman and A. B. Wagner[8])

traffic sent on the downlink. We do this by having the second user "listen in" on the packets that are intended for the first user and cache them for later. Our focus in this part will be on the scheduling and coding scheme for this problem and measuring the gains that can be obtained with respect to separate unicast and other schemes in the literature.

Literature Review

The problem of how to reap some of the benefits of multicasting when the clients are asynchronous is not new. The existing technologies that have been proposed can be divided into two groups: "open-loop" schemes in which the transmitter is oblivious to the state of the clients[13][14][15][16][17][18] and "closed-loop" schemes for which the transmitters knows the exact state of each client at each time.[19][20][21][22] Open-loop schemes are necessary when there are a large number of clients, since the amount of feedback required by a fully closed-loop scheme would be prohibitive. For a smaller number of clients, however, open-loop schemes can be very inefficient. Indeed, they assume that clients are constantly arriving and demanding the video, when in reality there might be no clients listening at all. Even when there is a moderate but nonzero number of clients, separate unicast can be more rate efficient.

"Our aim in this work is to create a scheme that uses feedback to handle small numbers of clients efficiently but also scales well…"

Conversely, closed-loop schemes generally perform well when there are small numbers of clients, but they do not scale well as the number of clients grows. Our aim in this work is to bridge the gap between these two types of schemes; that is, to create a scheme that uses feedback to handle small numbers of clients efficiently but also scales well as the number of clients increases and the amount of per-client feedback decreases.

The Model

"We consider several wireless users attached to the same Wi-Fi access point or cellular base station."

We begin with a model that makes simplifying assumptions. We consider several wireless users attached to the same Wi-Fi access point or cellular base station. We assume that all of the users are interested in the same video. The video of interest is divided into N segments. Each segment may be as small as an individual packet, or as large as a significant portion of the entire video. We assume that the ratio of the downlink multicast rate to the video data rate is R. That is, the video could be broadcasted R times in the time it takes to watch it once.

We model the channels from the server to the users as block erasure channels with a low erasure probability. We assume time is divided into slots. During a given slot, a given user's video will advance by one segment if it has either received that segment during a prior transmission of the server or if the server transmits it during the given slot and the user's channel does not erase the transmission during that slot. If neither of these conditions holds, then the user's video *stalls* during that timeslot; that is, it does not advance. Minimizing the amount of stalling will be our initial objective in designing scheduling algorithms.

We do not assume that all I users are demanding video at each time. Instead, we assume that a random fraction of the users are in a *rest* state. Users automatically enter the rest state after they have viewed the final segment of the video. When in the rest state, in each timeslot, a user decides with some fixed probability to begin watching the video from the beginning. This is done independently from slot to slot and from user to user. Users who are currently watching the video decide with some fixed probability to cease watching the video and enter the rest state. We assume that whenever a user enters the rest state it automatically deletes any video segments it has received. As such, one can alternately view a user exiting the rest state as representing the arrival of a new user, with I representing the maximum number of users allowable in the system. We usually assume that the system begins with all users in the rest state. This defines a controlled Markov chain (see Rahman and Wagner[8][9] for a mathematical definition and extensions). We define a cost function over this Markov chain as the number of users whose videos are stalled during a timeslot. That is, the cost during one timeslot is the number of users who are not in the rest state and whose videos do not advance by one segment. Our goal is to minimize the time average of this quantity, normalized by the number of users, over a finite time horizon.

Scheduling Algorithms

In order for the scheduler to make good scheduling decisions, it is useful for it to have feedback from the users, in which they convey that they are currently not in the rest state, where they are in the video, and which segments of the video they currently have and which they need. Let us initially assume that there is perfect feedback: after each timeslot, every client informs the transmitter of its current state, and the transmitter receives this information without error. The system is then a fully observed controlled Markov chain,

"We seek heuristics that perform well yet are efficiently implementable."

and, in principle, the optimum scheduling policy can be described exactly via dynamic programming.[15][16][30] The optimal scheme has extremely high complexity, however. In fact, it is difficult even to simulate for small systems. We thus seek heuristics that perform well yet are efficiently implementable. We shall discuss three.

Algorithm 1: Needed-First and Needed-Most

Needed-first is a simple and intuitive heuristic that performs well in simulation and in experiments although it is not robust in certain respects. The idea is to send the segments that are needed most urgently. To do this, we compute how long each client has until that client's video will stall *assuming no more segments are sent*, that is, how many segments are between that client's current location in the video and the first segment that the client does not have. The segment that is "needed first" is then defined to be the first segment that is needed by the client for which this metric is smallest. Consider the example in Figure 5 in which there are seven segments in the video and two clients. The first client has not viewed any segments and has received all of the odd-numbered segments. The second client has viewed the first two segments and has all of the segments except for the last three. In this case, segment 2 is needed most urgently, because client one will stall after one segment if segment 2 is not received, while client 2 will has two segments left to view before a stall occurs. If the scheduler could only send one segment (that is, $R = 1$) then it would broadcast the second segment. If it could send two segments, then it would send segment 2 and segment 5, since segment 5 is the next-most urgently needed segment after segment 2.

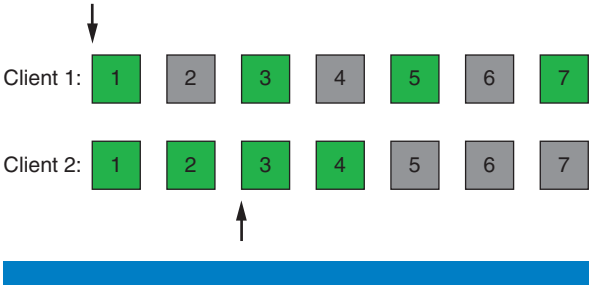


Figure 5: Example for needed-first and needed-most scheduling algorithms
(Source: Cornell University, 2014)

The advantages of this scheme are its simplicity and that it runs very fast in practice. Indeed, experimental tests show that it can schedule and send video data at a rate as high as 400 Mbps over a local host connection, even when scheduling is performed at the packet level. We shall see later that it works well in small, idealized experiments. One disadvantage of this scheme is that it is not robust when relaxing the assumption of perfect feedback—if the transmitter does not know with certainty where the clients are in the video and which segments they need, it is meaningless to talk about which segment is needed most urgently. This scheme is also difficult to extend to the combined unicast/multicast, multiple multicasts, and scalable videos scenarios discussed

later, nor can it be extended to handle other cost measures. Finally the most serious disadvantage is that it can allow one user with a weak channel to dominate the downlink at the expense of a very large number of users who would otherwise have no trouble downloading the video.

Another scheduling algorithm that is also simple and addresses this last issue is to send the segments that are needed by the maximum number of users. We shall call this *needed-most*. The idea is to make most efficient use of the downlink by sending segments that are useful to the maximum number of users. In the above example, the needed-most scheduler would send segment 6, since both clients need it, while all of the others are needed by at most one client. This scheduler is more "democratic" than needed-first in that it does not allow a small number of users to dominate the downlink at the expense of the majority. It does not perform well in small networks, however. For instance, consider a lone client that is partly through the video when a new client arrives. The needed-first scheduler will loop back to the beginning of the video to serve the new client, relying on the first client to use whatever buffer it has accumulated. The needed-most scheduler, on the other hand, will first transmit segments at the end of video that both clients need, leaving the client that just arrived stalled at the beginning for some time.

"The idea is to make most efficient use of the downlink by sending segments that are useful to the maximum number of users."

Figure 6 shows a simulation of the performance of these two scheduling algorithms against pyramid coding, separate unicast, and the approximate dynamic programming scheduler to be discussed shortly. The maximum number of clients I is plotted on the horizontal axis. The fraction of time the average user is stalled is plotted on the vertical axis. The simulation assumes a video with 7 segments, $R = 3$, and a packet loss rate of around 3 percent (the

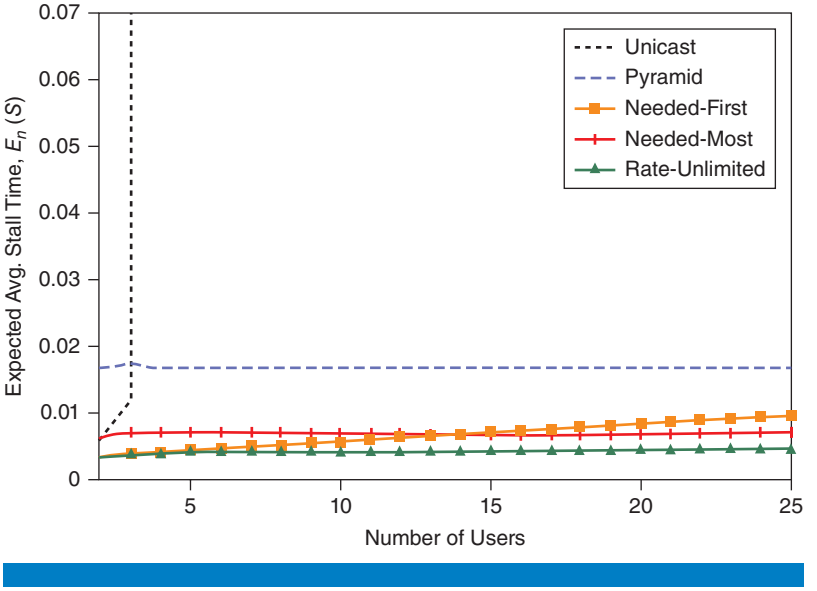


Figure 6: Performance of various scheduling algorithms (Source: Cornell University, 2014)

channel model used in the simulation is actually Markov, not the i.i.d. model described here). Users leave the rest state with probability 0.5, and they enter the rest state with probability 0.01. Since $R = 3$, unicast can handle about three clients, which is clearly seen in the plot. We see that needed-first performs better for a smaller number of clients while needed-most performs better for larger numbers of clients, although both perform much better than pyramid coding or separate unicast.

Next we discuss a more sophisticated algorithm whose performance dominates both needed-most and needed-first. It also has the advantage of being more extensible than either of these.

Algorithm 2: Approximate Dynamic Programming

In principle, the optimum scheduling policy can be described exactly via dynamic programming.[23] The idea is to compute, for each possible set of packets that the transmitter could send during a given timeslot, the expected future cost from the present until the end of the time horizon, assuming the optimal scheduler is used at each point in the future. This characterization has very high complexity, however. Thus we seek approximations to dynamic programming that perform well but can be implemented efficiently. For now, we continue to assume that perfect feedback is available from the clients to the transmitter.

"…we seek approximations to dynamic programming that perform well but can be implemented efficiently."

We shall focus on an approximation of this algorithm that we naturally call *approximate dynamic programming*. The key observation is that the optimal scheduler becomes very simple when there is only one client in the system: keep transmitting the lowest-index segment that the client does not have until the transmission gets through, then proceed to the next lowest-index segment. Our approximate dynamic programming scheduler (ADPS) conducts the following experiment: if only segment i is sent during the current timeslot, then what is the expected future cost from the present until the end of the time horizon, assuming that in all future timeslots each of the users is magically served by its own, exclusive transmitter? In reality of course, all of the users must share a single downlink for all time. For the purposes of this thought experiment, however, we assume that this sharing occurs for only a single timeslot—the one for which we are making a scheduling decision. Then the future cost calculations are performed assuming that the users evolve independently. Since the optimal single-user scheduling algorithm is very simple, this dramatically reduces the complexity of the algorithm. We have also found it helpful to make several other minor simplifications, which are described elsewhere.[8][9]

"…ADPS balances the need to send segments to individual users who are close to stalling against the need to use the downlink efficiently…"

Figure 6 also shows the performance of the ADPS (denoted as "Rate-Unlimited" in the figure) with perfect feedback in comparison with the schemes mentioned in the previous section. We see that it outperforms both needed-first and needed-most. In essence, the ADPS balances the need to send segments to individual users who are close to stalling against the need to use the downlink efficiently by sending segments that help the maximum

number of users possible. We have not proven that this scheme is close to optimal, although in our simulations we have found that its stall-time curve is so close to the horizontal axis that little improvement is possible. Although its performance is attractive, the advantage of ADPS is its extensibility, which we describe next.

Algorithm 3: Extensions of Approximate Dynamic Programming

"The ADPS can be easily extended…"

The ADPS can be easily extended to handle delayed or sporadic feedback, combined multicast/unicast, multiple multicast rates, scalable video, and other cost functions.

Recall that up to now we have assumed that each client, after each timeslot, informs the transmitter of its new state, and that this feedback is received perfectly by the decoder. This is obviously realistic only for systems with few clients, or those for which the segments are very long, or those for which each client is allocated its own, dedicated uplink channel. For many systems, however, including Wi-Fi systems with many clients, none of these conditions occurs and each client must send feedback much less frequently in order to avoid tying up the channel with feedback.

Suppose instead that each client sends feedback once every N segments, where N is on the order of the number of users, so that the aggregate rate of the feedback is remains constant as the number of clients grows. The system then becomes a partially-observed Markov decision process (POMDP), because the true state of the system is never known exactly by the transmitter when it makes it scheduling decisions.

We handle this by forming stochastic state estimates for the clients at the transmitter. That is, for each client, the transmitter maintains a probability distribution over locations in the video and for each client-segment pair, the transmitter maintains a scalar representing the probability that the client has the segment. When the transmitter sends a particular segment, it increases the probability that the client is believed to have the segment, assuming this probability is less than one, to reflect the higher likelihood that it has been received. The amount of increase is determined by the channel statistics. If the channel were perfect, then this probability would be set to one. Likewise, the distribution over locations in the video is updated.[8][9] When feedback is received from a client, all of these distributions briefly become deterministic, because the client's state is known with certainty at that point (that is, all of the segment probabilities will be 0 or 1 and the distribution over locations in the video will become a point mass). Of course, the state will gradually become random again as the system evolves and the transmitter becomes more uncertain about the client's state.

"To make scheduling decisions, we simply generate a random guess of the client's true state according to the recorded distribution."

To make scheduling decisions, we simply generate a random guess of the client's true state according to the recorded distribution. For this "hardened" state, the ADPS can be used to make a scheduling decision. The random realizations of the state are then discarded and regenerated for future scheduling decisions.

Figures 7, 8, and 9 show the simulated performance of the ADPS with partial feedback. In fact, these plots show the performance of ADPS in the extreme situation in which the scheduler learns only when users enter and when they depart. It does not receive any feedback from a client while it is streaming the video. We see that even with this small amount of feedback, the ADPS can provide substantial gains over pyramid scheduling and separate unicast, especially when R is large.

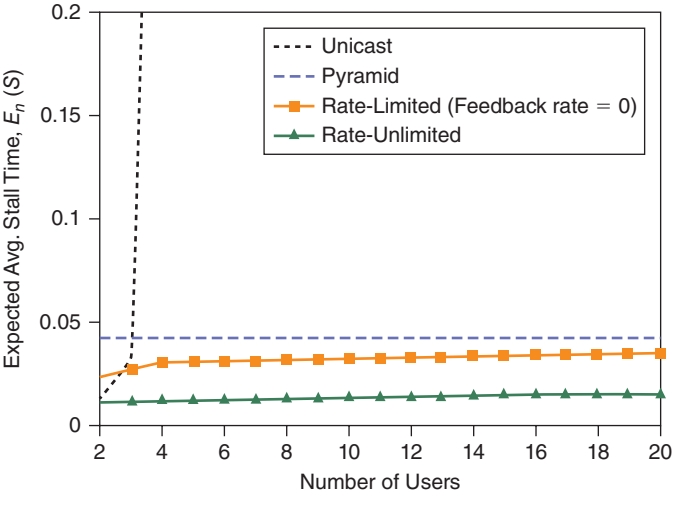


Figure 7: Performance of rate-limited feedback scheduling: $R_{ratio} = 3$ and $N = 7$
(Source: Cornell University, 2014)

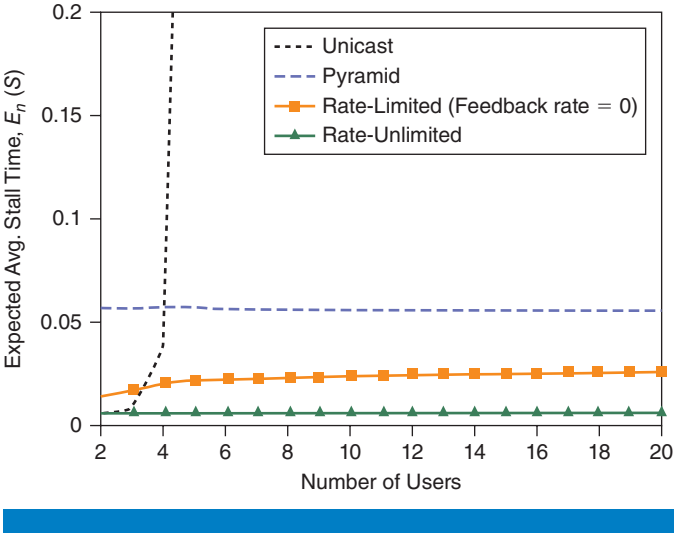


Figure 8: Performance of rate-limited feedback scheduling: $R_{ratio} = 4$ and $N = 15$
(Source: Cornell University, 2014)

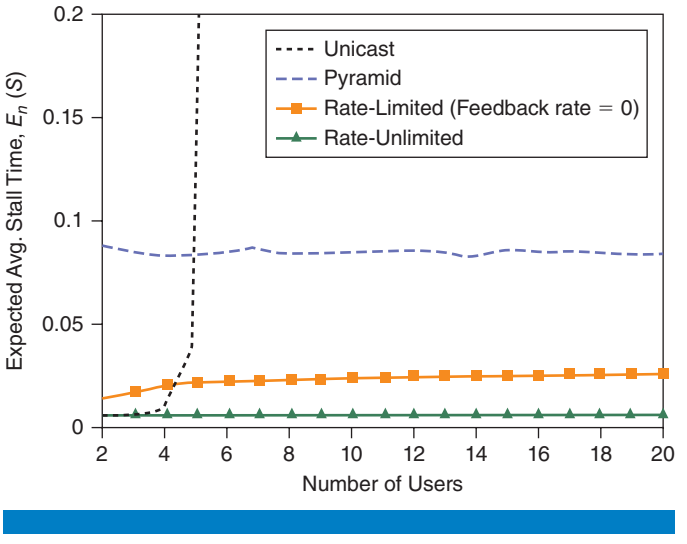


Figure 9: Performance of rate-limited feedback scheduling: $R_{ratio} = 5$ and $N = 31$
(Source: Cornell University, 2014)

The ADPS can also be extended to allow for combined unicasting and multicasting. The ADPS considered so far assumes that all data sent on the downlink is broadcast to all users. Assuming that one desires most transmitted segments to be received, this necessitates transmitting at a rate that can be decoded by the user with the weakest channel. If a segment to be transmitted is only required by a client who has a very strong channel, however, it can be more efficient to unicast the segment to this client at a higher rate. The ADPS can easily be extended to allow for this possibility; one only needs to enlarge the search space of possible actions at each time. The heuristics used to estimate the expected future cost of taking a particular action remain the same. More generally, one could set up several different multicasts at different rates, and the clients would join all of the multicasts whose rate they can support. The transmitter, for each segment that it wishes to send, can decide which multicast to use, based on their rates and coverage. The ADPS can be extended to handle scalable video in a similar way. The scheduler simply has a larger set of options available: in addition to deciding which segment to send, it decides which layer to transmit. Of course, increasing the number of options available to the scheduler inevitably increases its complexity, so the benefit of these extensions must be weighed against a decrease in the speed with which the scheduler can make decisions.

"The ADPS can also be extended to allow for combined unicasting and multicasting."

Although the ADPS algorithm was developed using stall time as the metric, the primary heuristic (considering the coupling between users for one timestep only) will work equally well with other metrics.

Coding

Up to this point, we have not considered the possibility of coding or mixing segments together before transmitting them. Coding is known to provide rate gains in many scenarios and it is natural to consider the gains it can provide here.

It is easy to see that coding can provide some gains. Consider, for example, the prototypical example of *index coding*, depicted in Figure 10. There are two segments and two clients. Each client has one segment and needs the other. How to serve these two clients?

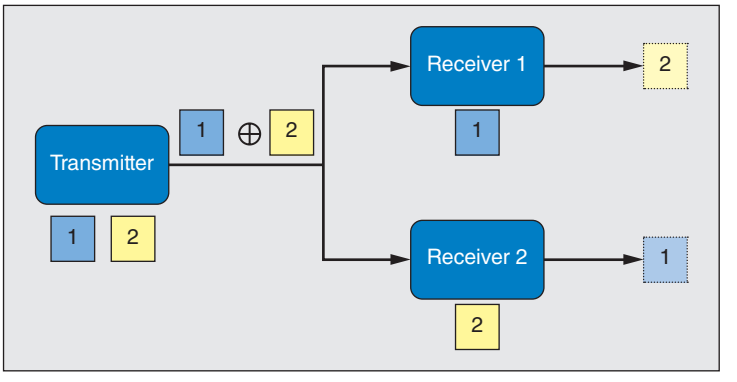


Figure 10: The prototypical example of index coding (Source: Cornell University, 2014)

One could of course send the two segments. If we assume that one segment can be sent every timeslot ($R = 1$), then this requires two timeslots. If the two segments are the same size, however, then the transmitter could transmit, say, their bitwise exclusive OR. This would only require one timeslot, and both clients could recover their desired segment by performing an exclusive OR operation on the received segment with the one they already have.

We have studied how much additional gain this idea provides over scheduling alone by simulating it in combination with the ADPS. Specifically, the scheduler would rank-order the segments in order of transmission priority as before but, rather than simply sending the R segments with the highest priority, we attempt to squeeze a slightly longer list of segments into the same airtime by using coding. Thus if the scheduler ranked the segments in order of decreasing priority as

3, 1, 5, 2, ...

and $R = 2$, then the ADPS without coding would send segments 3 and 1 during the next timeslot. Using coding, however, we might be able to send the segments 3, 1, and 5 in the time that it takes to send two segments without coding. If this is possible we assume that it is done.

We found that this provides negligible gains over scheduling alone, however. Intuitively, this is because the scenario described in the toy example above rarely occurs in practice: it can only happen as a result of a packet drop, which is uncommon. The question then arises: can coding help if one moves beyond the above toy example and does optimal coding? Answering this question is difficult because the optimal scheme for the index coding problem is unknown;

this is a major open problem in information theory and theoretical computer science.

We did, however, find the optimal coding scheme for networks with at most three receivers[10][11][12], using mathematical tools that had not previously been applied to the index coding problem. The ramifications of this result for the intellectual pursuit of solving the index coding problem are discussed elsewhere.[10][11][12] Here we only note that using the optimal coding scheme that emerged from this work again provided only negligible gains over scheduling alone. We conclude that index coding provides little additional performance gain over scheduling.

"We conclude that index coding provides little additional performance gain over scheduling."

Testbed Validation

The ADPS has been implemented on a testbed consisting of a single Wi-Fi access point and several laptops. For the access point, experiments were performed with both an Apple Airport Express* and a Linksys WRT54G* running the DD-WRT open source firmware. Many off-the-shelf consumer-grade Wi-Fi access points fix the multicast rate to be quite low. In our experiments, the Apple Airport Express would not multicast at rates higher than about 10 Mbps, even when all of the clients could support higher rates, although it could unicast about an order of magnitude faster. This made it impossible to fairly compare our scheduler against separate unicast. The main reason for using the Linksys with DD-WRT was that we could manually set the multicast rate to be the same as unicast, 54 Mbps, making for a fairer comparison.

The software for the server and the client was written in C++ and can run entirely in user space. It has been tested successfully under Mac OS X* and Microsoft Windows* under Cygwin. In addition to ADPS, the code implements the needed-first scheduler and pyramid scheduling.[13]

What follows is the result of a typical experiment. The video used was a 67-second HD opening montage of a news program. The video file was approximately 20 MB, so the average bit rate of the video was roughly 2.5 Mbps, although it was encoded using variable bit rate compression. The server was run on a high-performance laptop and up to 15 clients were distributed across four other laptops. The start times of the clients were spread uniformly over a 30-second window. This experiment was run using the WRT54G but with the server and the clients connected to it via wired Ethernet using an Ethernet switch. The reason for using wired Ethernet is that the server multicasts using UDP and does not currently implement flow control. Thus it is important that the Wi-Fi multicast at a nearly constant rate. This is generally the case in wired networks. For wireless networks, we found it to be the case only if the experiment is run in a quiet Wi-Fi environment. We believe that implementing flow control would address this. The needed-first scheduler was used since it runs faster than the ADPS and the added sophistication of the ADPS was not needed for this experiment. We found that the scheduler could easily handle 15 clients without any of the videos stalling,

"…the scheduler could easily handle 15 clients without any of the videos stalling…"

even if it was throttled to send data at 20 Mbps. We did not test it with more clients simply because we ran out of laptops on which to run clients; we found that each laptop could render at most three or four videos at a time.

The lower curve in Figure 11 shows the total amount data sent over the downlink by the scheduler as a function of the number of clients. The upper curve is simply the size of the video file multiplied by the number of clients; this is approximately the amount of data that would be sent on the downlink by separate unicast. We see that already for six clients, the scheduler sends less than half the data required by unicasting. For 15 clients, it is less than one third. While this is already a significant gain, it is worth noting it arguably understates the improvement. The reason is that this plot only shows the downlink traffic. For the scheduler, the amount of feedback traffic is very slight (there is approximately one feedback packet sent for every 60 packets sent on the downlink). For TCP unicast, on the other hand, an ACK will be sent in response to every downlink packet. While these ACK packets will be short, there will be many of them, and collectively they will use up a significant amount of air time.

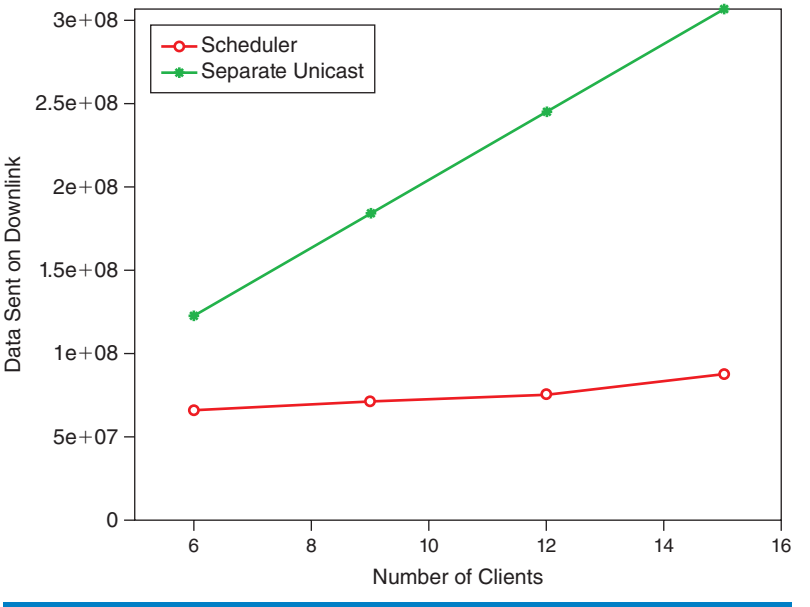


Figure 11: Results from the Wi-Fi testbed (Source: Cornell University, 2014)

Summary and Future Directions

The four main outcomes of this work are (1) a scheduling algorithm, the approximate dynamic programming scheduler (ADPS), which can bridge the divide between open-loop and closed-loop schemes and which is quite extensible, (2) the finding, confirmed both numerically and experimentally, that effective scheduling can provide substantial performance gains compared with the state-of-the-art, and (3) the finding that coding seems to offer little additional benefit beyond what effective scheduling can offer, and (4) a Wi-Fi–based implementation of our algorithms.

Note that the gains described in this part only involve changes to the application layer. Indeed, the gains could be demonstrated using user-space programs on off-the-shelf laptops and Wi-Fi access points. This is important for two reasons. First, it means that it can be combined with other proposals that have been made as part of the VAWN program, so long as those proposals restrict their changes to the lower layers. Second, and more importantly, it means that the technology proposed here could be implemented relatively soon, without large-scale changes to the lower levels of the communication stack. This was an important consideration for the VAWN program.

"…the technology proposed here could be implemented without large-scale changes to the lower levels of the communication stack."

An interesting future direction would be to consider how the performance of the ADPS scales as the number of users grows very large. Although our model is realistic enough that optimality results are difficult to establish for finite number of users, it might be possible to determine how the performance of the optimal scheduler scales with the number of users. This could then be compared with the scaling law exhibited by the ADPS. It would be interesting to consider the effect of coding on the scaling law as well. As the number of clients increases, the transmitter will be less able to retransmit segments that were missed by individual clients. This will result in more coding opportunities.

This work considered only a single video in isolation; we have not explored how to schedule across multiple videos. In this regard, it is worth noting the gains we achieve through multicasting are closely related to the gains achieved via pre-fetch caching that are discussed in the next section. In both cases, we achieve gains by sending data to users in anticipation of them needing it later. In the case of pre-fetch caching, this means unicasting data to users during periods when the downlink would otherwise be idle. For multicasting, it means having lagging users "listen in" on the leading users. Given the similarities between the two approaches, it could be worthwhile to study them jointly as a single problem. This is currently under investigation.

"…the gains we achieve through multicasting are closely related to the gains achieved via pre-fetch caching"…

Acknowledgments

The ADPS scheduling algorithm was developed in collaboration with Md. Saif Rahman. The results on coding were developed in collaboration with Sinem Unal. The Wi-Fi testbed was developed in collaboration with Chris Anthony, C. J. Halabi, Liuyuan Chen, Yan Cao, David Dunn, Aaditya Ramesh, Yixuan Shi, Gil Lee, Krishna Chaitanya Achutuni, Varun Bharadwaj, Xiong Chu, Hsiao-ting Hung, and Karthikeya Kumar Pandit.

Video Traffic Predictions Based on Social Graphs

In large networks as those addressed by VAWN, we face the problem of having many users share the limited network resources while expecting a certain level of quality guarantee. There are many ways of addressing this issue, like limiting the number of users that can access the network, scaling

out the network equipment (which is expensive), or moving the content closer to the users and storing on cheap proxy caches so as to reduce end-to-end network activity.

"…we exploit the widely accepted idea that data in information networks have similar patterns of diffusion to viruses in a population."

As part of VAWN, we exploit the widely accepted idea that data in information networks have similar patterns of diffusion to viruses in a population. We explore using a social graph to predict what video will be watched by whom, and when, based on past requests, which is particularly relevant to VAWN as we strive to improve network efficiency for video. Using this social graph, we evaluate its efficacy on reducing network load by applying it to the caching problem. The idea behind caching is that many of the requests made by the users of the network are for the same objects, so to minimize the end-to-end delay of requests in the network and save limited network resources, the local network should store the items that are likely to be requested again. Crucial to the effectiveness of caching is the ability to predict what will actually be popular and when. The current de-facto standards in industry are crude heuristics that assume the most recent will be most popular or the most viewed in the past will remain the most popular at the given time. By coming up with a model that actually predicts the *who*, *when*, and *with what probability* a video will be requested, we challenge the simplistic assumptions of these industry heuristics, the Least Recently/Frequently Used (LRFU) family of methods.

"We demonstrate the gains that results from implementing socially aware cache policies on media driven by social processes…"

We demonstrate the gains that results from implementing socially aware cache policies on media driven by social processes (for example, YouTube*) can give over the standard LRFU spectrum of policies. The LRFU approaches predict the popularity distribution by computing the combined recency and frequency (CRF) value for each object and caching according to the scores. Computing the CRF is dependent on the weighing parameter γ, which takes on values from 0 to 1 and controls the tradeoff between recency and frequency.[24][25] The two extremes degenerate to the least recently used (LRU) when γ is 1, and least frequently used (LFU) when γ is 0.

We incorporate diffusion in two ways, first by using the inter-arrival time between requests for the same video for prediction, and then by using a virus propagation model over a *latent* social graph to model the spread of the videos between users. We do not have explicit friendship information as found on sites such as Facebook*. Instead, we infer the (directed) edge weights between users in the graph, which are interpreted as the transmission/sharing probability between two users from the network trace of YouTube requests. An example of a social graph is seen in Figure 12.

Related Works

There are several points of difference between our work and prior work. In the work by Wang et al.[26], they observe and track actual friendships within

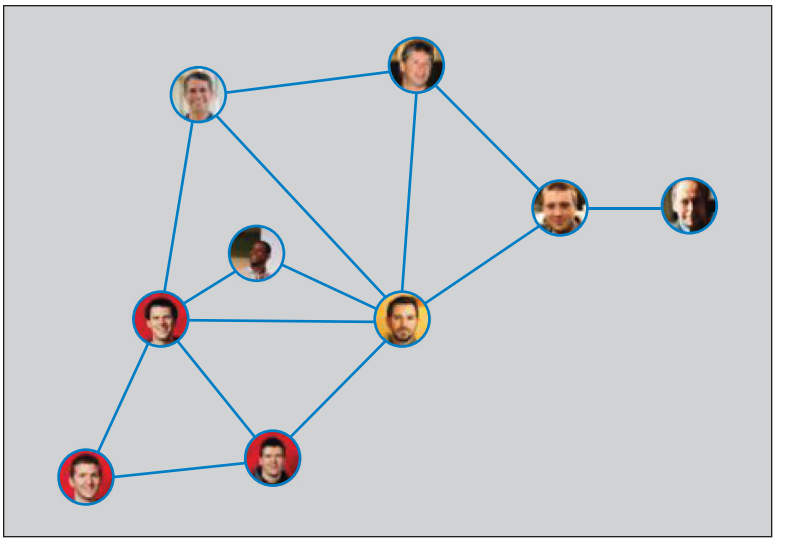


Figure 12: This social graph could represent a network of nine users, where the edges between the users represent the probability they share YouTube videos with each other
(Source: *Symposium on Selected Areas in Communications (GC13 SAC)*, Atlanta, USA, Dec. 2013, pp. 3160–3166)

social networks (Facebook, Twitter*, and so on) and use the patterns of propagation for the use of replicating the content in different sites, whereas in our work, we do not have access to the actual friendships derived from such social networks but instead infer unknown relationships for predicting the popularity distribution of videos in the network under observation. In other works[27][28], they show that the global popularity distribution from the content provider (YouTube) is highly uncorrelated with local popularity on local networks, and this motivates our focus on inferring the local social-influence network.

In recent years, there has been a growing interest in the field of social network analysis and its applications in real-world computational problems. A social network can be described as any network where the realized flow of objects over the links and nodes that make up the network is driven by human action or behavior. Examples include road networks, recommendation networks, citation networks.

"…there has been a growing interest in the field of social network analysis and its applications in real-world computational problems."

In some situations, the human actions are directly observed on application-level networks like Facebook, Twitter, and other social-media websites, where the links between the users are explicitly known. There are also many situations where the human-driven spread of objects in the network is not directly observed over the links. In such cases, to understand the relationship between the users, we must be able to infer from the observed network transactions, the links between these users. One example is the network between a city population and the spread of a virus over

the population. In this case, the spread of an infectious virus over the population are the hidden transactions, while the observable transactions are the various records of infections by clinics, or pharmaceutical drug sales. There have been several works that have addressed the issue of latent social network inference.[29][30][31][32]

"In predicting the popularity distribution, we explore two general approaches, consensus-based approaches and socially based approaches…"

Approach

At a given time t, we observe some user $u \in U(t, w)$ making a request, where $U(t, w)$ is the set of users at time t that made requests in the past w time units. The users are uniquely identified by the order in which they first made requests in the network. Similarly at some time t, we observe a video $v \in V(t, w)$ being requested, where $V(t, w)$ is the set of videos at time t that were requested in the past w time units. The videos are also uniquely identified by the order in which they were first requested. In the rest of this article, for convenience of notation we represent these sets as U and V. We represent these transactions between users and videos as a triplet (u, v, t), and the set of all such triplets in a given interval as the network trace, T. Similar to the single cache performance approximation presented by Li et al.[25], our goal is to predict for some time t, in the future, the popularity distribution of these videos given a history of these triplets before time t. In predicting the popularity distribution, we explore two general approaches, consensus-based approaches dependent on aggregate information in the network, and socially based approaches dependent on explicit information diffusion over nodes in the network. These approaches are evaluated under the framework of caching the k most popular videos. Given the network trace T, we can calculate $\vec{X}(t, k)$ as the k- sparse binary vector of length $|V|$, representing the k videos to be cached at time t, and $X(t) \in \mathrm{N}^{|V|}$, the vector representing the number of views each video has at time t. The hit rate, $H(\vec{X}(t, k) \mid T)$, is then given as follows:

$$H(\hat{X}(t, k) \mid T) = \frac{X(t) \cdot \hat{X}(t,k)}{\| X(t) \|_1}$$

For the purpose of this work, we don't allow partial caching of videos, and we assume all videos have the same size. Caching schemes can be divided into two steps: first, the assignment of scores to the various objects or candidates for the cache, and second, an ordering of these scores to decide which objects will end up in the cache. It is not necessary that these two steps are done separately. Some caching strategies, for example the commonly used least recently used (LRU) scheme, implicitly combine the score assignment and the ordering.

Further details are presented by Nwana et al.[43] We discuss the approaches we explored in the following sections.

Baseline

The de-facto standard for our application is caching according to the CRF values assigned to each video as shown by Li et al.[25] They come up with a model to approximate the performance (hit rate) of the popularity

distribution under the LRFU assumption independent of the actual order the videos arrive, and they demonstrate through trace-based simulations that the approximate model is a good approximation of the actual trace-based simulations using the LRFU scheme (within 5 percent). We compare our approaches to the approximate popularity distribution.

Consensus Approach

Any approach to caching videos from network traces, *T*, that ignores the actual users/watchers of the videos is a consensus approach. Such methods rely on aggregate or average properties derived from the video request patterns, and not particularly how a video diffuses over the nodes in the network. This section discusses some consensus approaches we explored.

"Any approach to caching videos from network traces, that ignores the actual users/watchers of the videos is a consensus approach."

In improving the baseline, we will explore its deficiencies. A deficiency of the baseline is its handling of time and recency. Since underlying the baseline distribution is the LRFU scheme, it is known that the distribution is susceptible to object staleness.[24][33] This is because objects can still have high CRF scores under LRFU even after a long time has elapsed since its last request. We will consider two (orthogonal) ways of incorporating time or the notion of staleness into the baseline.

Inter-arrival Time. In this approach we model the average inter-arrival/inter-viewing times of videos in the network. To do this, for each video we take the average of the intervals between successive views, and then we estimate the parameters of power law distribution fit to these averages. By doing this, if the time that elapsed since we last saw a given video and the time for which we are predicting its popularity is large the probability it is requested is small according to the inter-arrival distribution, thus preventing staleness. We fit to a power law distribution because it has been shown in many real-life information diffusion networks that the propagation time can be modeled by a power law distribution.[34][35][36][37][38][39]

We use the Zipf distribution to represent frequency of requests (because it is a good approximation of the distribution of objects in networks of diffusion[39][40]) and the inter-arrival distribution to represent recency. Using these two distributions, we give a different angle to analyzing recency and frequency with the notion of staleness, which is an Achilles' heel of the LRFU approach. The score for this approach is the product of the recency probability given by the inter-arrival distribution and the frequency probability given by the Zipf distribution. The ordering for this approach is a sort in decreasing probability.

Social Approach

This approach is different from the other ones in that it considers users when predicting the popularity of a video. It predicts for each user the probability of that user watching each video. And given those probabilities, we estimate the number of views for each video by taking the expected number of views that video will get according to the probabilities of users

"Social approach is different from the other ones in that it considers users when predicting the popularity of a video."

watching said videos. Before we can calculate the user-video probabilities, we must first estimate the user-user transmission probabilities. These probabilities are modeled as the edges of a diffusion graph between the users of the network.

Diffusion Model. In classical epidemiology, for a given virus v, we can classify the population into (not necessarily disjoint) sets representing the stage each individual is in the life cycle of that virus. If they have not yet been infected at time t, we say they are in the Susceptible set, $S(t)$. If they have been infected and are infectious, they are in the Infectious set, $I(t)$, and if they have recovered they go into the Recovered set, $R(t)$. In this work, we consider the *S-I* model, where individuals transition from being susceptible to being infectious and remain infectious once infected. The transition from S to I occurs in two stages, first transmission, and then incubation. Before an individual can be said to be infected, the individual must have contracted the virus from a carrier. The difference between the contraction time of the infection and the outbreak of symptoms is the incubation time.

"For this article, the set of users is the population of individuals and the videos are the viruses."

For this article, the set of users is the population of individuals and the videos are the viruses. The probability that a user, u, gets infected by a video, v, at time t is then the same as the probability the individual, $u \in S(t)$, contracted the infection from an already infected individual, $u' \in I(t)$, and the incubation time is the difference between t and the time that u' transmitted the disease to u. In this work we assume that as soon as the user gets infected (watches a video), they immediately transmit the video to all other users not yet infected with some probability. Hence the transmission time from u' to u is the infection time, $\tau^{v}_{u'}$ of u'. Let $\chi_v(u',u,t)$ represent the probability that user u' infects user u with video v at time t. Let Δ represent the incubation period distribution (which is a power-law distribution).

$$\chi_v(u',u,t) = A_{u'u} \cdot \Delta(t - \tau^{v}_{u'})$$

We learn the transmission probabilities (as an adjacency matrix A) under the maximum likelihood framework proposed by Myers et al.[28], and we assume that the incubation times follow a power-law distribution as diffusions in information networks have been shown to follow.[34][35][36][37][38][39] We use the same exponent for the power-law distribution as we do for the inter-arrival approach. The sequence of infections of a given video is called a cascade, and we make an independent cascade assumption, which means each video is transmitted independent of other videos.

"The score for a given video is the expected number of new infections for that video at the given time t."

The score for a given video is the expected number of new infections for that video at the given time t.

$$S_{\text{diffusion}}(v,t) = \sum_{u \in S(t)} \left[1 - \prod_{u' \in I(t)} \left(1 - \chi_v(u',u,t)\right) \right]$$

The ordering is a descending sort of the diffusion scores.

Combined Approach

One deficiency of the purely social approach is that if there is not enough data to create a complete graph based on diffusion, then we are only predicting the views for a small subset of the users in the network, which will lead to underperformance. A remedy for this is for those users that are not part of the diffusion graph, but part of the network, we estimate their probabilities from a consensus approach as previously described.

Yet another apparent drawback of the purely social approach is that not every video watched by users that appears in the diffusion graph is necessarily fully explained through diffusion. Other influences, such as (independent) personal tastes and influence from external sources like news sites and blogs, are bound to play roles in affecting what the user watches as well. We do not attempt to fully model this phenomenon in this work, but we leave it up to a future work.

The score for the combined approach is given by a linear combination of the diffusion approach and the inter-arrival approach. The rank is given by a descending sort of the combined scores.

"The score for the combined approach is given by a linear combination of the diffusion approach and the inter-arrival approach."

Numerical Analysis and Results

Our data is from the University of Massachusetts, Amherst, YouTube network traces described and analyzed by Gill et al.[27] We utilize 120 consecutive hours (from Thu 03/13/2008 19:00 to Tue 03/18/2008 18:10) of YouTube requests from their campus network and we partition this data into a training set over the first sixty hours. The training set contains a total of 79,213 requests made by 7,260 unique users over 58,345 unique videos. The testing set contains a total of 96,568 requests made by 6,383 unique users over 72,528 unique videos. In the dataset, there are a total of 10,349 unique users and 120,973 unique videos. For our experiments, we make our cache update periods in units of length, 1 hour. For each of these periods, we predict the cache content and as our performance metric, we look at the average hit rate over all the time periods in our testing set for different cache sizes.[43]

Baseline. As explained earlier, we rank each video in decreasing order according to its approximate popularity under the LRFU scheme it got during that time. We employ a window size of $w = 28$ for our baseline, because empirically on our training set we see that this window size gives the best average hit rate.

Inter-arrival Time. For this approach we have the constraint that the cascades used in learning the parameters for the inter-arrival distribution must be of at least length five, that is, at least five requests for the video. This is to avoid the noise added by shorter cascades. We then calculate the scores as described in the section "Approach," using the output of our baseline method as the input for the Zipfian distribution[39][40]. We compare the result of this approach to our baseline (Figure 13) and we get on average about an 11.5-percent improvement in hit rate.

"…we get on average about an 11.5-percent improvement in hit rate."

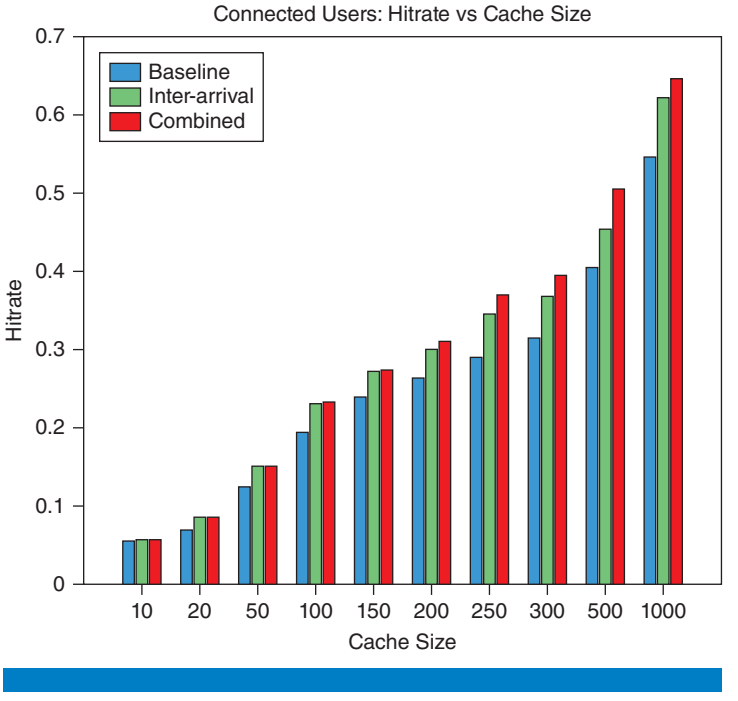


Figure 13: The comparison between the combined and inter-arrival on connected users. Starting from medium cache sizes, the combined approach outperforms the inter-arrival approach (Source: *Symposium on Selected Areas in Communications (GC13 SAC),* Atlanta, USA, Dec. 2013, pp. 3160–3166)

Social. To learn the incoming transmission probabilities on the training set, we must also learn power-law distribution parameters. We learn these parameters exactly under the same assumptions as the inter-arrival time approach. An added constraint in choosing parameters were that there are at least three unique users making the requests that are of that cascade. This is an attempt to remove noise by increasing the probability that one of those views was as a result of sharing between users.

On our testing set, we relearn the adjacency matrix every 10 hours, and limit ourselves to only inferring from the past $w = 60$ hours of history each time we learn a new adjacency matrix. We use the algorithm described by Myers et al.[29] with a sparsity of 300.

After these parameters and transmission probabilities are learned, we proceed to calculate the future scores of the video for each of the periods in our testing set. For each hour, we consider as prediction history all the requests that were made in the last $w = 16$ hours. We calculate the score using the function described in the section "Social Approach," with the appropriate adjacency matrix.

Because of the constraints on valid cascades, we find that on average only about 40 percent of the users in the network are used for inference, which

implies only 40 percent of the nodes in the graph are in a connected component and the others are just isolated nodes. This ultimately leads to underperformance (Figure 14) since for about 60 percent of the users we can make no estimation of the probability it views a given video. So in order to gauge the usefulness of this approach, we also run it on an augmented dataset (Figure 15) where only users in the largest connected component are in the dataset.

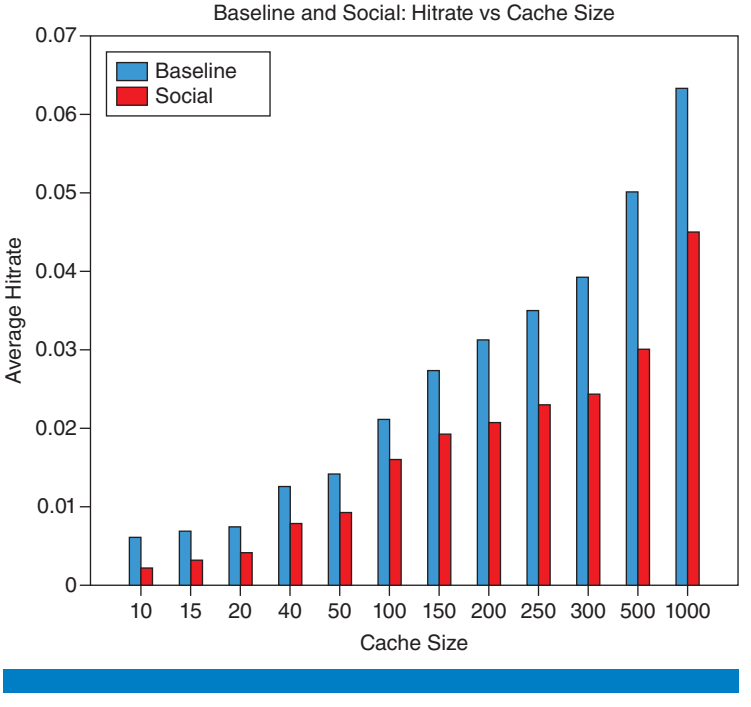


Figure 14: The comparison between the social approach and baseline on all users. As was mentioned in the section "Combined Approach," the social approach suffers from the fact that only 40% of all users have nonzero request probabilities
(Source: *Symposium on Selected Areas in Communications (GC13 SAC)*, Atlanta, USA, Dec. 2013, pp. 3160–3166)

Combined. As previously noted in the section "Combined Approach," it is unlikely that even in a connected graph, the volume of videos watched will be completely accounted for by diffusions over the graph. Myers et al.[41] showed that only about 71 percent of information volume on Twitter* could be accounted for by network diffusions. To that end, on our data we performed an experiment to figure the best weighting (ranging from 0 percent to 100 percent in steps of 10) between the social scores and the inter-arrival scores, and, corroborating the conclusion of Myers et al.[41], it came out to be 70 percent from social and 30 percent from inter-arrival. We also analyze the performance of the combined approach under the full data, and under just the connected component.

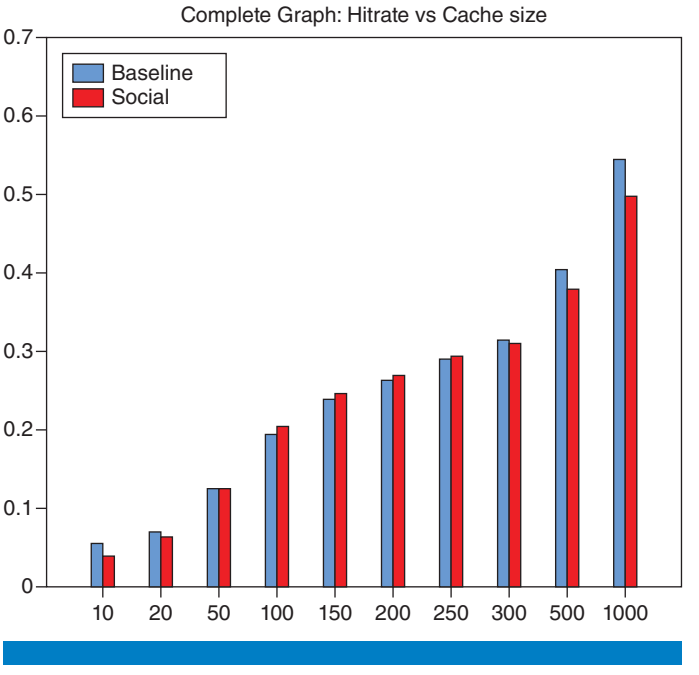


Figure 15: The comparison between the social approach and baseline when on connected users. We can see from this that the social popularity distribution, which implicitly takes into account recency and frequency through diffusion, is within 5% of the approximate popularity distribution based on LRFU if indeed all the users are governed by diffusion
(Source: *Symposium on Selected Areas in Communications (GC13 SAC)*, Atlanta, USA, Dec. 2013, pp. 3160–3166)

Under the full dataset, we see an improvement of 13.2 percent from the baseline to the combined (Figure 16), and under just the connected components, our improvement rises to 21.1 percent (Figure 13).

We also compare the performance of the combined to the inter-arrival approach. From our experiments we see that the combination of social and inter-arrival gives an improvement of 1.6 percent over inter-arrival on this full dataset (Figure 16), and 5.5 percent when only users from the connected component are used (Figure 13).

"…the combination of the social approach with a consensus approach (inter-arrival) generally outperforms any of the individual approaches."

From our results, we see that for the social approach to give considerable gain, it is imperative that the underlying structure be a connected graph, and not a graph with mostly isolated vertices. We have also seen that the combination of the social approach with a consensus approach (inter-arrival) generally outperforms any of the individual approaches. This is because on this dataset, and as we suspect on most social networks, neither diffusion nor consensus can fully explain the request patterns of the users in the network. Users tend to have their own preferences and are also

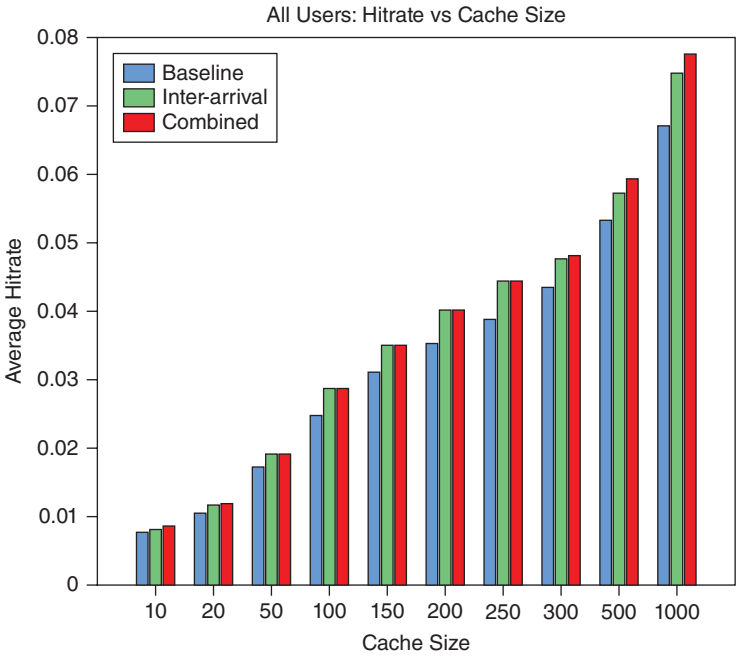


Figure 16: The comparison between the combined and inter-arrival on all users. The combined approach outperforms the inter-arrival approach especially on the larger cache sizes, but still caches more relevant videos even on smaller cache sizes
(Source: *Symposium on Selected Areas in Communications (GC13 SAC)*, Atlanta, USA, Dec. 2013, pp. 3160–3166)

influenced by different external media like the news or blogs, and such behavior has been studied by Myers et al.[41], where they show that only about 71 percent of information volume on Twitter can be attributed to network diffusion.

Method A	Method B	% improvement
Baseline	Viralness (small cache)	6.2
Baseline	Inter-Arrival	11.6
Baseline	Combined	13.2

Table 1: Summary of Results on All Users
(Source: *Symposium on Selected Areas in Communications (GC13 SAC)*, Atlanta, USA, Dec. 2013, pp. 3160–3166)

Method A	Method B	% improvement
Baseline	Inter-Arrival	15.6
Baseline	Combined	21.1

Table 2: Summary of Results on Connected Users
(Source: *Symposium on Selected Areas in Communications (GC13 SAC)*, Atlanta, USA, Dec. 2013, pp. 3160–3166)

"…our model captures the idea of videos being spread like diseases over a network of users…"

Summary and Future Directions

To achieve the goals as defined in VAWN, in this work we have shown that by leveraging social cues inferred from the requests made by users in the network, through an estimated latent social graph over these users, we can better predict the popularity distribution of the videos being requested by the users. Our preferred method of combining the distributions resulting from the social approach and the inter-arrival (staleness) approach is shown to outperform other approaches. This is because our model captures the idea of videos being spread like diseases over a network of users where some users are more likely to infect (and be infected by) other users, which the other approaches do not. These considerations result in a 14-percent improvement over the baseline.

One of the roadblocks we faced in this work is the inadequate amount of data both in terms of recency and volume, so an immediate follow-up to this work is to gather more recent data on a longer scale from different network sites and verify our findings from this work. We also aim to address the issue of the robustness of the algorithms to varying of the parameters and the performance when the uniform video size assumption is lifted. Based on the inferred social graph, we would also like to investigate community-based approaches to prediction and potentially find the users most influential to widespread propagation of information. As mentioned in the introduction, this work has wider implications in the domain of network management where the knowledge of future user requests is important for efficient management of systems and resources, so another future direction is to quantify how much the social driven approach helps scheduling and coding in multicast scenarios.

Concluding Remarks

In this article, we provided an in-depth discussion on three opportunities, created by network heterogeneity and content commonality, for video delivery in wireless networks, and demonstrated a quantitative analysis of the capacity gains that they can provide.

In the first part of the article, we considered the opportunities that arise from exploiting network heterogeneity for video delivery. We studied optimal strategies that exploit network diversity for delivering real-time video traffic. We considered a new metric of *timely throughput* that captures the strict per-packet deadline requirement of real-time video traffic, and developed communication protocols that maximize the timely throughput of heterogeneous wireless networks. Via numerical analysis we demonstrated the gain from our algorithms compared with other load-balancing algorithms that do not take packet deadlines into account; in particular we showed a reduction of 1.6x in client's outage in a network with 5 access points and 100 users. Our results promise 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.

"…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."

In the second part of the article, we considered the gains the can be obtained by multicasting content that is being simultaneously demanded by several users at a single base station, in particular in situations that are highly congested, such as sports stadiums, cafés, airports, and train stations. We provided several scheduling algorithms for asynchronous multicast in such networks. We demonstrated, both numerically and experimentally (via a Wi-Fi–based testbed), that effective scheduling can provide substantial performance gains compared with the state of the art.

Finally, in the last section we focused on the potential gains from being able to correctly predict what video is going to be watched by whom at a given time with certain probability. By leveraging social cues inferred from the requests made by users in the network, we provided a prediction algorithm for the popularity distribution of the videos being requested by the users. We showed that our prediction method, which is based on combining the distributions resulting from the social approach and the inter-arrival (staleness) approach, outperforms other existing approaches by a 14-percent improvement in prediction error.

References

[1] Hou, I. H., V. Borkar, and P. R. Kumar, "A theory of QoS for wireless," In *Proc. of IEEE INFOCOM*, 2009.

[2] Hou, I. H., A. Truong, S. Chakraborty, and P. R. Kumar, "Optimality of periodwise static priority policies in real-time communications," In *Proc. of CDC*, 2011.

[3] Lashgari, S. and A. S. Avestimehr, "Timely Throughput of Heterogeneous Wireless Networks: Fundamental Limits and Algorithms," to appear in *IEEE Transactions on Information Theory*, available at http://arxiv.org/abs/1201.5173, 2012.

[4] Lashgari, S. and A. S. Avestimehr, "Approximating the Timely Throughput of Heterogeneous Wireless Networks," In *Proc. of IEEE ISIT*, 2012.

[5] Shmoys, D. B. and E. Tardos, "An approximation algorithm for the generalized assignment problem," *Mathematical Programming*, 62:461474, 1993.

[6] Jain, K., "A factor 2 approximation algorithm for the generalized Steiner network problem," *Combinatorica*, Springer, 2001.

[7] Hou, I. H. and P. R. Kumar, "Scheduling periodic real-time tasks with heterogeneous reward requirements," *IEEE RTSS*, 2011.

[8] Rahman, Md. S. and A. B. Wagner, "Multicasting for Wireless Video-On-Demand," 51th Annual Allerton Conference on Communications, Control, and Computing, University of Illinois, Urbana-Champaign, Oct. 2013, to appear.

[9] Rahman, Md. S. and A. B. Wagner, "Multicasting for Wireless Video-On-Demand," in preparation.

[10] Unal, S. and A. B. Wagner, "General Index Coding with Side Information: Three Decoder Case," IEEE Int. Symp. Inf. Theor. Proc., 2013, to appear.

[11] Unal, S. and A. B. Wagner, "A Rate-Distortion Approach to Index Coding," Proc. ITA, 2014, to appear.

[12] Unal, S. and A. B. Wagner, "A Rate-Distortion Approach to Index Coding," in preparation.

[13] Viswanathan, S. and T. Imielinski, "Metropolitan area video-on-demand service using pyramid broadcasting," *Multimedia Systems*, vol. 4, pp. 197–208, 1996.

[14] Aggarwal, C. C., J. L. Wolf, and P. S. Yu, "A permutation-based pyramid broadcasting scheme for video-on-demand systems," *Proc. of the International Conference on Multimedia Computing and Systems*, pp. 118–26, 1996.

[15] Hua, K. A. and S. Sheu, "Skyscraper Broadcasting: A new broadcasting scheme for metropolitan video-on-demand systems," *Computer Communication Review*, pp. 89–100, 1997.

[16] Juhn, L.-S. and L.-M. Tseng, "Harmonic broadcasting for video-on-demand service," *IEEE Trans. Broadcasting*, vol. 43, no. 3, pp. 268–271, Sept. 1997.

[17] Janakiraman, R., M. Waldvogel, and L. Xu, "Fuzzycast: efficient video-on-demand over multicast," in *Proc. IEEE INFOCOM*, 2002, pp. 920–929.

[18] Janakiraman, R., M. Waldvogel, W. Deng, and L. Xu,"Achieving scalable and efficient video-on-demand over multicast," IBM Research Report RZ-3495, Dec. 2002.

[19] Dan, A., D. Sitaram, and P. Shahabuddin, "Scheduling policies for a non-demand video server with batching," *Proc. ACM Multimedia*, pp. 15–23, 1994.

[20] Hua, K. A., Y. Cai, and S. Sheu, "Patching: a multicast technique for true video-on-demand services," *Proc. ACM Multimedia*, pp. 191–200, 1998.

[21] Gao, L. and D. Towsley, "Supplying instantaneous video-on-demand services using controlled multicast," *Proc. IEEE Int. Conf. on Multimedia Computing and Systems*, pp. 117–21, vol. 2, 1999.

[22] Aggarwal, V., R. Calderbank, V. Gopalakrishnan, R. Jana, K. K. Ramakrishnan, and F. Yu, "The effectiveness of intelligent

scheduling for multicast video-on-demand," in *Proc. 17th ACM Int. Conf. Multimedia*, pp. 421–430, 2009.

[23] Kumar, P. R. and P. Varaiya, Stochastic Systems: Estimation, Identification, and Adaptive control. Prentice-Hall Information and System Sciences Series, Englewood Cliffs, NJ: Prentice Hall, 1986.

[24] Lee, D., J. Choi, J. hunKim, S. H. Noh, S. L. Min, Y. Cho, and C. S. Kim, "Lrfu: A spectrum of policies that subsumes the least recently used and least frequently used policies," in *Proceedings of the 1999 ACM SIGMETRICS Conference on Measurement and Modeling of Computer Systems*, 2001, pp. 134–143.

[25] Li, Z., G. Simon, and A. Gravey, "Caching Policies for In-Network Caching," in *ICCCN 2012: IEEE International Conference on Computer Communication Networks*, IEEE, Ed., 2012.

[26] Wang, Z., L. Sun, X. Chen, W. Zhu, J. Liu, M. Chen, and S. Yang, "Propagation-based social-aware replication for social video contents," in *Proceedings of the 20th ACM international conference on Multimedia*, ser. MM '12, New York, NY, USA: ACM, 2012, pp. 29–38 [Online] Available: http://doi.acm.org/10.1145/2393347.2393359.

[27] Zink, M., K. Suh, Y. Gu, and J. Kurose, "Characteristics of youtube network traffic at a campus network - measurements, models, and implications," *Comput. Netw.*, vol. 53, no. 4, pp. 501–514, Mar. 2009 [Online] Available: http://dx.doi.org/10.1016/j.comnet.2008.09.022.

[28] Gill, P., M. Arlitt, Z. Li, and A. Mahanti, "Youtube traffic characterization: a view from the edge," in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, ser. IMC '07. New York, NY, USA: ACM, 2007, pp. 15–28 [Online] Available: http://doi.acm.org/10.1145/1298306.1298310.

[29] Myers, S. A. and J. Leskovec, "On the convexity of latent social network inference," in *NIPS*, 2010, pp. 1741–1749 [Online] Available: http://books.nips.cc/papers/files/nips23/NIPS2010_1257.pdf.

[30] Liben-Nowell, D. and J. Kleinberg, "The link prediction problem for social networks," in *Proceedings of the twelfth international conference on Information and knowledge management*, ser. CIKM '03, New York, NY, USA: ACM, 2003, pp. 556–559 [Online] Available: http://doi.acm.org/10.1145/956863.956972.

[31] De Choudhury, M., W. A. Mason, J. M. Hofman, and D. J. Watts, "Inferring relevant social networks from interpersonal communication," in *Proceedings of the 19th international conference on World wide web*, ser. WWW '10, New York, NY, USA: ACM, 2010, pp. 301–310 [Online] Available: http://doi.acm.org/10.1145/1772690.1772722.

[32] Gomez-Rodriguez, M., J. Leskovec, and A. Krause, "Inferring networks of diffusion and influence," *CoRR*, vol. abs/1006.0234, 2010.

[33] Oneil, E. J., P. E. Oneill, and G. Weikum, "The lru-k page replacement algorithm for database disk buffering," in *Proc. ACM SIGMOD International Conference on Management of Data*, Washington, D.C, 1993, pp. 297–306.

[34] Mitzenmacher, M., "A brief history of generative models for power law and lognormal distributions," *Internet Mathematics*, vol. 1, pp. 226–251.

[35] Newman, M. E. J., "Clustering and preferential attachment in growing networks," *Phys. Rev. E*, 2001.

[36] Barabsi, A.-L. and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999 [Online] Available: http://www.sciencemag.org/content/286/5439/509.abstract.

[37] Mihaljev, T., L. de Arcangelis, and H. J. Herrmann, "Inter-arrival times of message propagation on directed networks," *CoRR*, vol. abs/1011.0630, 2010.

[38] Capocci, A., A. Baldassarri, V. D. P. Servedio, and V. Loreto, "Statistical properties of inter-arrival times distribution in social tagging systems," *CoRR*, vol. abs/1210.2752, 2012.

[39] "Zipf's law," http://en.wikipedia.org/wiki/Zipf's law, accessed: 03/11/2013.

[40] Yu, H., D. Zheng, B. Y. Zhao, and W. Zheng, "Understanding user behavior in large-scale video-on-demand systems," in *Proceedings of the 1st ACM SIGOPS/EuroSys European Conference on Computer Systems 2006*, ser. EuroSys '06, New York, NY, USA: ACM, 2006, pp. 333–344 [Online] Available: http://doi.acm.org/10.1145/1217935.1217968.

[41] Myers, S. A., C. Zhu, and J. Leskovec, "Information diffusion and external influence in networks," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD '12, New York, NY, USA:

ACM, 2012, pp. 33–41 [Online] Available: http://doi.acm.org/10.1145/2339530.2339540.

[42] "Knitro optimization software," http://www.ziena.com/matlabknitro.html, accessed: 07/23/2013.

[43] Nwana, A. O., S. Avestimehr, and T. Chen, "A latent social approach to YouTube popularity prediction," in *Globecom 2013 – Symposium on Selected Areas in Communications (GC13 SAC)*, Atlanta, USA, Dec. 2013, pp. 3160–3166.

Author Biographies

Salman Avestimehr received a PhD in electrical engineering and computer science from the University of California, Berkeley in 2008. He was a postdoctoral scholar at the Center for the Mathematics of Information (CMI) at the California Institute of Technology, Pasadena, in 2008. He was also an assistant professor at the ECE school of Cornell University from 2009 to 2013. Dr. Avestimehr has received a number of awards, including the Communications Society and Information Theory Society Joint Paper Award, the Presidential Early Career Award for Scientists and Engineers (PECASE), the Young Investigator Program (YIP) award from the U. S. Air Force Office of Scientific Research, and the National Science Foundation CAREER award. He is currently an associate editor for the *IEEE Transactions on Information Theory*. He can be contacted via email at avestimehr@ee.usc.edu.

Tsuhan Chen has been with the School of Electrical and Computer Engineering, Cornell University, Ithaca, New York, since January 2009, where he is the David E. Burr Professor of Engineering, and served as the Director of the School of Electrical and Computer Engineering from 2009 to 2013. From October 1997 to December 2008, he was a professor with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania, where he served as the Associate Department Head from 2007 to 2008. He received an MS and PhD in electrical engineering from the California Institute of Technology, Pasadena, California, in 1990 and 1993, respectively.

Tsuhan served as the Editor-in-Chief for *IEEE Transactions on Multimedia* in 2002–2004. He received the Benjamin Richard Teare Teaching Award in 2006, the Eta Kappa Nu Award for Outstanding Faculty Teaching in 2007, and the Michael Tien Teaching Award in 2014. He served as Vice President 2012–2013, and as President 2013–2014, of the ECE Department Head Association. He is a Fellow of IEEE. He can be contacted via email at tsuhan@cornell.edu.

Sina Lashgari received his BS in Electrical Engineering from Sharif University of Technology in 2010. He joined Cornell University in 2010, where he is pursuing his PhD 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. Sina spent the summer of 2013 in Qualcomm Flarion Technologies, where he worked on developing new algorithms for interference management in device-to-device communication networks. In 2010 Sina received an Irwin D. Jacobs Fellowship. He can be contacted via email at sl2232@cornell.edu.

Amandianeze (Amandy) Nwana received two bachelor of science degrees in Electrical and Computer Engineering and Computer Science both from Carnegie Mellon University in 2010, and his MS in Electrical and Computer Engineering from Carnegie Mellon University in 2011. He joined Cornell University in 2011, where he is currently pursuing his PhD. in the School of Electrical and Computer Engineering. His research interests are centered around social and information networks, on how to extract and leverage the rich and complex information implicit in these kinds of networks to solve both traditional and novel estimation, prediction, and detection tasks. Amandy was recognized as a finalist in the Qualcomm Innovation Fellowship in 2013 with his joint proposal on "Advanced Transmission and Prediction Techniques for High-Density Homogeneous Networks." In 2011 Amandy received a three-year Cornell Sloan Fellowship. He can be contacted via email at aon3@cornell.edu.

Md. Saifur Rahman received his MS and PhD in Electrical and Computer Engineering from Cornell University in 2011 and 2012, respectively, and received his B.Tech. in Instrumentation Engineering from Indian Institute of Technology Kharagpur in 2006. He worked as a post-doctoral associate in the School of Electrical and Computer Engineering at Cornell University from January to October 2012. Since November 2012, he has been with Samsung Research America - Dallas, where he is currently a senior standards engineer working on next generation communication standards. His areas of interest include wireless communication, optimization, and compression. He spent the summer of 2011 interning at Samsung Telecommunications America. He also worked at Samsung Research India – Bangalore during 2006 and 2007. He was a recipient of an Irwin D. Jacobs Fellowship in 2008. He was nominated for the ECE director's PhD thesis award for his doctoral work. He can be reached at md.rahman@samsung.com.

Sinem Unal received her BS in Electrical and Electronics Engineering with a minor in Mechatronics Engineering from Middle East Technical University, Turkey in 2011. She is pursuing her PhD at Cornell University in the School of Electrical and Computer Engineering. Her research interests include coding and scheduling for wireless networks and index coding using information theoretic tools. She was a summer intern at Bell Labs in 2014, where she worked on the coded caching problem. She received an Irwin D. Jacobs Fellowship in 2011. She can be contacted via e-mail at su62@cornell.edu.

Aaron B. Wagner received a BS in Electrical Engineering from the University of Michigan, Ann Arbor, in 1999 and an MS and PhD in Electrical Engineering and Computer Sciences from the University of California,

Berkeley, in 2002 and 2005, respectively. During the 2005-2006 academic year, he was a postdoctoral research associate in the Coordinated Science Laboratory at the University of Illinois at Urbana-Champaign and a visiting assistant professor in the School of Electrical and Computer Engineering at Cornell University. Since 2006, he has been with the School of Electrical and Computer Engineering at Cornell, where he is currently an associate professor. He has received the NSF CAREER award, the David J. Sakrison Memorial Prize from the U.C. Berkeley EECS Department, the Bernard Friedman Memorial Prize in Applied Mathematics from the U.C. Berkeley Department of Mathematics, and teaching awards at the department, college, and university level at Cornell. He is currently serving as associate editor for *IEEE Transactions on Information Theory.* He can be contacted via email at wagner@ece.cornell.edu.

DELIVERING ENHANCED 3D VIDEO

Contributors

Yury Gitman
Lomonosov Moscow State University

Can Bal
University of California, San Diego

Mikhail Erofeev
Lomonosov Moscow State University

Ankit Jain
University of California, San Diego

Sergey Matyunin
Lomonosov Moscow State University

Kyoung-Rok Lee
University of California, San Diego

Alexander Voronov
Lomonosov Moscow State University

Jason Juang
University of California, San Diego

Dmitriy Vatolin
Lomonosov Moscow State University

Truong Nguyen
University of California, San Diego

Autostereoscopic displays are expected to gain higher popularity in comparison with devices that require a viewer to wear special glasses to see 3D. Nonetheless, the industry might not be ready to deliver high quality 3D to viewer. In this article we examine each stage of the 3D content lifecycle from its creation to display in the user's home. We present various algorithms for solving existing problems in each of these steps.

To avoid the creation of low quality 3D, we propose a set of methods for automatic quality assessment. To enable easy multiview content creation, a disparity estimation method is proposed. We discuss two methods of efficient 3D compression to save bandwidth in wireless channels. We propose a system for 3D display quality assessment to ensure that 3D video quality is not affected by the display device. Finally, we describe a system for carrying out subjective comparisons that will assist in further improvement of the above-mentioned methods.

Introduction

The challenge of video transport in future wireless networks includes the increasing prevalence of new data types like HD and 3D. While HD (high-definition) video increases data size by expanding resolution, 3D (stereoscopic) video increases it through the use of dual video streams, one intended for each eye of the viewer, or even more than ten video streams for multiview autostereoscopic displays, enabling the watching of 3D from several points of view without special glasses. Raw 3D video data is thus at least twice the size of raw 2D video, although much work has been done to develop compression standards (for example, MVC extensions to H.264/AVC) that improve data size in practice.

A viewer's perceptual system relies on several cues to extract information about the relative distance between objects within 3D space. These can be divided into two groups: monocular cues and binocular cues. While 2D video systems use only monocular cues (such as motion parallax, perspective, and occlusions), 3D video systems also utilize binocular cues to give the viewer an even stronger sensation of scene depth. For example, *binocular parallax* refers to the difference between left and right images, something the human visual system automatically translates into perceived depth. *Eye convergence* refers to the manner in which human eyes point inward at one another as an object gets closer to the viewer.

While conventional approaches to 3D video formatting like Side-By-Side and Top-and-Bottom incorporate separate left and right images into a single video frame or temporally interlace left and right images sequentially within the data stream (frame sequential), newer approaches leverage the notion of a *depth map*.

Depth maps provide information on the relative distances of surfaces within a scene from a viewpoint and, when combined with 2D or multiview video data, can be used to reconstruct a second view frame at the decoder using depth-image-based rendering (DIBR) techniques. Depth maps may be used to adjust the disparity between views on different displays and to generate multiple views in autostereoscopic displays. They may also be associated with compression strategies for transmitting 3D video in capacity-constrained wireless networks.

We believe the challenge of transporting 3D video over wireless networks of the future requires an end-to-end approach that considers all phases of the transport pipeline: content creation, delivery, and display. Content creation determines the encoding and formatting of data, matching data to user display requirements and creating opportunities for data compression. Delivery considerations include the role of data compression and robustness against errors that occur due to wireless channel effects. Display considerations focus on device characteristics, as well as complexity and performance of rendering and playback. How to define and measure quality in 3D video playback for a particular device and 3D encoding scheme is a key challenge.

"We believe the challenge of transporting 3D video over wireless networks of the future requires an end-to-end approach that considers all phases of the transport pipeline…"

This article reviews our research contributions to each part of this end-to-end pipeline within the VAWN research program. The next section, "Content Creation: Disparity Estimation and Processing, and Quality Metrics," discusses our work on content creation, including disparity estimation, which is a building block for depth map construction, the handling of common errors in depth map construction, and our development of VQMT3D, a tool for 3D video quality assessment. In the section "Delivery," we review our contributions to 3D data delivery, including multiview video compression based on depth map propagation, multiview video plus depth coding with depth-based prediction mode, and mixed resolution stereoscopic coding. In the section "Display," we discuss our work on 3D displays including autostereoscopic displays, tools for subjective testing, and automatic device testing. In the section "Conclusion," we close with recommendations to the industry community.

Content Creation: Disparity Estimation and Processing, and Quality Metrics

There are two common approaches to creating 3D content (besides rendering): capturing with a stereo camera system or converting from a 2D video. Capturing with a stereo camera rig seems to be the most natural way to create stereo video, but typically, captured video suffers from various mismatches in camera settings and inaccurate spatial alignment. This implies the need for 2D-to-3D conversion instead of capturing for some scenes, even for feature films. Converting from a 2D video to 3D has its own difficulties.

The process of 2D-to-3D conversion first requires the content creator to determine the distance of each pixel from the camera for each one of the frames of the entire 2D video. These distance values associated with each pixel

are then stored as a separate 2D video called the *depth map*. While conversion can potentially allow the production of 3D content without any impairments, the quality of the converted video strongly depends on the quality of the depth maps. The labor-intensive nature of depth map creation typically leads to low-quality conversion. In this section, we will also discuss methods for visual quality estimation for both captured and converted content.

It is also possible to capture 3D content with more than two cameras. Such a multiview video capture process is even more challenging because the complexity of camera alignment increases with the number of views. This is the main reason for limited availability of multiview video content which requires autostereoscopic displays. Thus multiview content creators commonly prefer an intermediate approach; that is, capturing stereo video with subsequent conversion. In the case of stereo-to-multiview conversion, the depth map associated with each of the views can be estimated without manual labor. For this, the pixels of the left-view image are matched with pixels of the right-view image in order to estimate a shift map, referred to as the *disparity map*. Given the disparity map, the depth map can easily be calculated, which enables synthesis of additional views.

In this section, we present methods for 3D content creation, beginning with our disparity estimation algorithm.

Disparity Estimation

Disparity estimation is an integral problem associated with the delivery of 3D content (2D + depth, multiview + depth format), and it plays a direct role in both compression efficiency and the need for wireless network capacity to deliver 3D content to mobile devices. In a two-camera imaging system, disparity is defined as the vector difference between the object point in each image relative to the focal point. It is this disparity that allows for depth estimation of objects in the scene via triangulation of the object point in each image. Figure 1 shows

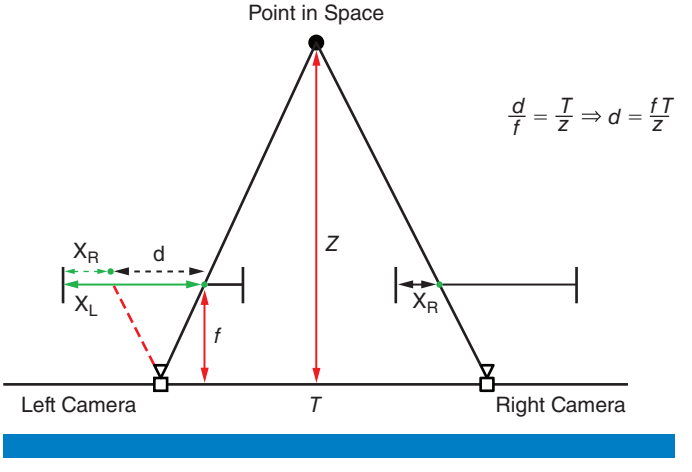


Figure 1: Relating depth to disparity
(Source: Ramsin Khoshabeh, PhD thesis: "Bringing Glasses-free Multiview 3D into the Operating Room," UCSD, 2012)

the inverse relationship between depth z and disparity d, as identified next to the figure.

The proposed disparity estimation algorithm consists of five main components: similarity measure, support weight, disparity computation, occlusion filling, and *total variation* (TV) refinement. The block diagram of the overall system is shown in Figure 2.

For similarity measure, we propose a three-mode census transform with a noise buffer to be more tolerant of image noise in flat areas and a cross-square census to increase the reliability of census measure. We suggest the effective combination of three cost measures formulated as

$$C_0(q, q_d) = 3 - \exp\left(-\frac{\Delta H_{qqd}}{\gamma_H}\right) - \exp\left(-\frac{\Delta I_{qqd}}{\gamma_I}\right) - \exp\left(-\frac{\Delta G_{qqd}}{\gamma_G}\right) \quad (1)$$

where ΔH_{qqd}, ΔI_{qqd}, and ΔG_{qqd} stand for census, color, and gradient respectively.

For support weight, the adaptive support weight is based on the strength of grouping by similarity and proximity. We suggest the following conditional adaptive support weight:

$$w(c, q) = \begin{cases} g_{\gamma_s}(\Delta S_{cq}) & \text{if } \Delta S_{cq} < E \\ g_{\gamma_s}(\Delta S_{cq})\, g_{\gamma_p}(\Delta p_{cq}) & \text{otherwise} \end{cases} \quad (2)$$

where r_s and r_p are empirical similarity parameters, ΔS_{cq} is the RGB color difference between the center pixel and the neighboring pixel, Δp_{cq} is the spatial distance between pixel c and pixel q, and E is a color difference threshold determining the similar color between two pixels.

For disparity computation, once the support weights are calculated, the aggregated cost is computed by aggregating the raw similarity measures, scaled by the support weights in the window. The aggregated matching cost between pixel c and pixel c_d is given in the weighted mean as

$$A(c, c_d) = \frac{\Sigma q \in W_c, q_d \in W_{cd}\, w(c, q) w(c_d, q_d) c_0(q, q_d)}{\Sigma q \in W_c, q_d \in W_{cd}\, w(c, q) w(c_d, q_d)} \quad (3)$$

where W_c and W_{cd} represent the left and right support windows, respectively, and the function $w(c_d, q_d)$ is the support weight of pixel q_d in the right window. After the aggregated matching costs have been computed within the disparity range, the disparity map is obtained by determining the disparity d_p of each pixel p through the Winner-Takes-All (WTA) algorithm.

For occlusion filling, first a left-right consistency check is performed to detect unreliable pixels. The unreliable pixels are defined as the ones that have nonmatching disparities on the left and right images.

Then, the reliable pixels within a cross-based neighborhood vote for the candidate disparity value at (x, y). The pixels with missing disparity values are then filled with the majority votes of the reliable pixels in the voting region.

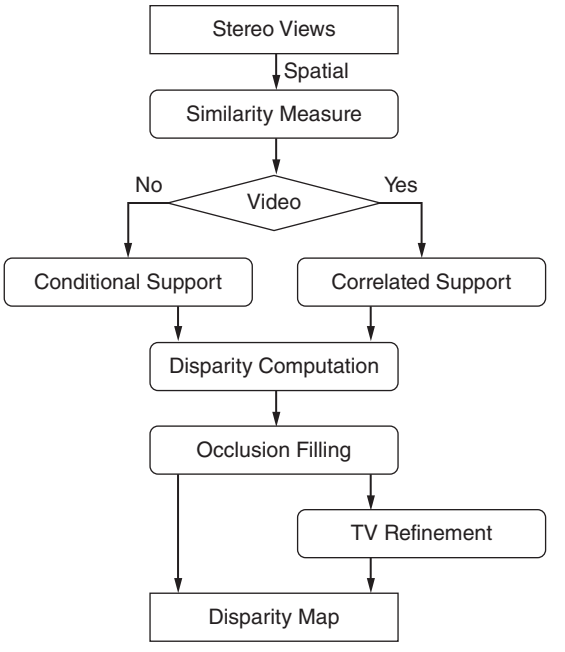


Figure 2: Block diagram of the proposed disparity estimation approach
(Source: Zucheul Lee, PhD thesis: "Disparity estimation and enhancement for stereo panoramic and multi-array image/video," UCSD, 2014)

The final step in the algorithm is the TV refinement. The block diagram illustrated in Figure 3 captures how this refinement process works at a high level. TV refinement involves solving the following minimization problem:

$$\underset{\mathbf{f}}{\text{minimum}}\ \mu\|\mathbf{f}-\mathbf{g}\|_1+\|\mathbf{Df}\|_2 \tag{4}$$

where $g = vec(g(x, y, t))$ and $f = vec(f(x, y, t))$ are the initial disparity and the optimization variables, respectively. The operator **D** is the spatiotemporal gradient operator that returns the horizontal, vertical, and temporal forward finite difference of **f**.

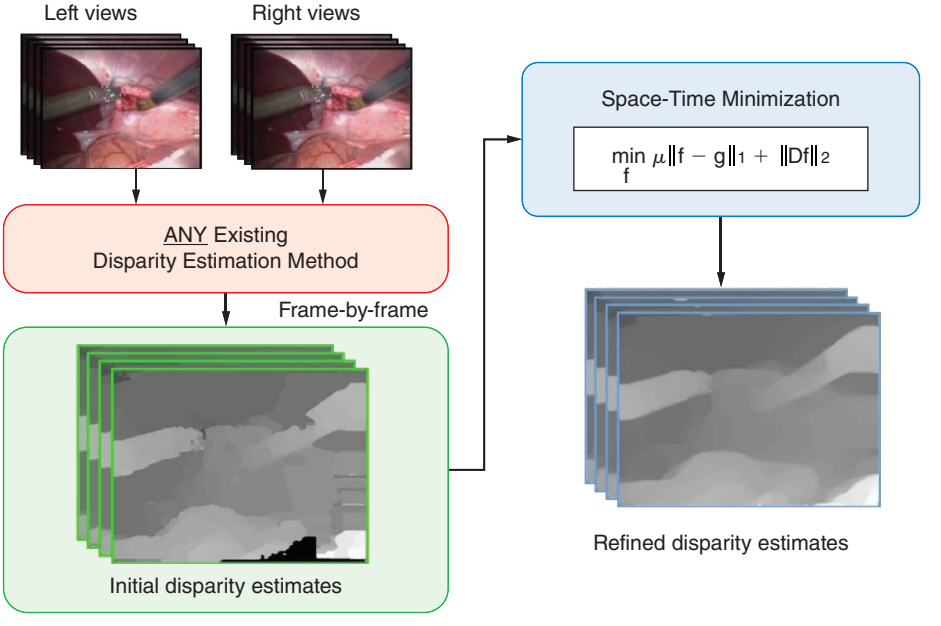


Figure 3: Space time minimization overview
(Source: Ramsin Khoshabeh, PhD thesis: "Bringing Glasses-free Multiview 3D into the Operating Room," UCSD, 2012)

An augmented Lagrangian method solves the above minimization problem by the following steps (at the *K*th iteration):

$$\mathbf{f}_{k=1}=\ \underset{\mathbf{f}}{\text{argmin}}\ \frac{\rho_o}{2}\ \|\mathbf{r}_k-\mathbf{f}+\mathbf{g}\|^2+\rho_r\|\mathbf{u}_k-\mathbf{Df}\|^2+\mathbf{z}_k^T\mathbf{f}+\mathbf{y}_k^T\mathbf{Df},$$

$$\mathbf{v}_{k+1}=\mathbf{Df}_{k+1}+\frac{1}{\rho_r}\mathbf{y}_k,$$

$$\mathbf{u}_{k+1}=\max\left\{\mathbf{v}_{k+1}-\frac{1}{\rho_r},0\right\}\cdot\frac{\mathbf{v}_{k+1}}{\|\mathbf{v}_{k+1}\|2},$$

$$\mathbf{r}_{k+1}=\max\left\{\left|\mathbf{f}_{k+1}-\mathbf{g}+\frac{1}{\rho_o}\mathbf{z}_k\right|-\frac{\mu}{\rho_o},0\right\}\cdot sign\left(\mathbf{f}_{k+1}-\mathbf{g}+\frac{1}{\rho_o}\mathbf{z}_k\right),$$

$$\mathbf{y}_{k+1}=\mathbf{y}_k-\rho_r(\mathbf{u}_{k+1}-\mathbf{Df}_{k+1})$$

$$\mathbf{z}_{k+1}=\mathbf{z}_k-\rho_o(\mathbf{r}_{k+1}-\mathbf{f}_{k+1}+\mathbf{g})$$

In the iterative method described above, the first problem (known as the f-subproblem) can be solved using Fast-Fourier Transform for fast computation.

The performance evaluation makes us of the Middlebury datasets with ground truth disparity maps provided by the Middlebury online benchmark.[1][2] Table 1 summarizes the quantitative results taken from the Middlebury database methods. Our method achieves excellent results, ranking 13th out of about 130 methods and it is the best performing local method.

Methods	Rank	Avg Err(%)	Tsukuba	Venus	Teddy	Cones
Proposed	13	5.12	2.10	0.12	5.46	2.12
PatchMatch	15	4.59	2.09	0.21	2.99	2.47
CostFilter	20	5.55	1.51	0.20	6.16	2.71
InfoPermeable	21	5.51	1.06	0.32	5.60	2.65
GeoSup	28	5.80	1.45	0.14	6.88	2.94
AdaptDisCalib	37	6.10	1.19	0.23	7.80	3.62
SegmentSupport	53	6.44	1.25	0.25	8.43	3.77
AdaptWeight	67	6.67	1.38	0.71	7.88	3.97

Table 1: Local method performance evaluation on the Middlebury datasets. (Source: Zucheul Lee, PhD thesis, "Disparity estimation and enhancement for stereo panoramic and multi-array image/video," UCSD, 2014)

To assess the performance of the proposed method quantitatively on stereo videos, we use five synthetic stereo videos with ground truth disparity from the University of Cambridge.[3] We compare three methods without occlusion filling to compare their performance. The LASW method ranks 67th and the Cost-filter, which is one of the best performing local methods, ranks 20th on the Middlebury benchmark. Table 2 shows the average percentage of bad pixels over all frames and illustrates that the proposed method has the best performance.

Video	# of frames	LASW	CostFilter	Proposed method
Tunnel	99	1.435%	2.157%	0.997%
Book	40	5.933%	4.919%	3.601%
Temple	99	10.15%	10.70%	10.36%
Street	99	9.978%	7.305%	7.246%
Tanks	99	5.714%	4.826%	4.811%

Table 2: Performance comparison of methods on five stereo videos (Source: Zucheul Lee, PhD thesis, "Disparity estimation and enhancement for stereo panoramic and multi-array image/video," UCSD, 2014)

"…the depth map is critical for efficiently displaying and transmitting 3D content…"

Depth Upsampling and Processing

As mentioned earlier, the depth map is critical for efficiently displaying and transmitting 3D content, especially for multiview displays where multiple views can be generated at the display using the depth map rather than transmitting each view separately. The more accurate the depth map, the more efficiently the content can be compressed. This reduces the capacity needed when sending over a wireless network and allows a higher quality rendering to be achieved. A depth map is typically estimated using stereo vision systems and the disparity estimation procedure; however recent advances in capture technology also allow capturing the depth maps directly using real-time depth sensors. Regardless of whether the depth maps are estimated or directly captured, depth maps contains errors. These errors can be roughly categorized into two broad categories:

- Errors in transition areas: Inadequate calibration, occlusion areas, or motion artifacts often lead to wrong depth values at object boundaries when aligned with color images.
- Random noise on geometrically flat or smooth surfaces: Properties of the object surface, lighting conditions, or systematic errors may generate noise on the surface.

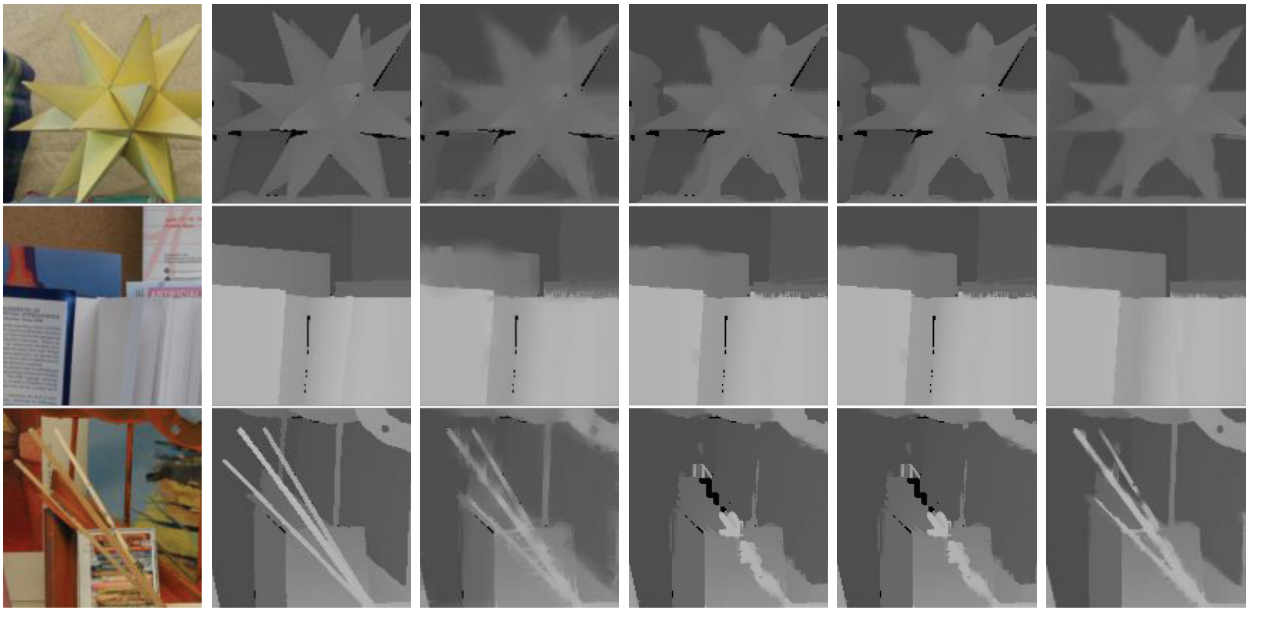
In our work, we investigate a method that can fix both of these errors. Our method takes a color image I and a corresponding lower resolution depth map D as inputs. The process consists of upsampling, sample selection, sample refinement, and robust multilateral filtering of the depth map. Before the refinement step, we begin by measuring the depth reliability and finding the unreliable regions. In this sample selection stage, for every pixel in the unreliable region, we collect depth samples from reliable regions and select the best sample that yields the highest fidelity. Then these samples are refined by sharing their information with their neighbors' selected samples. Finally, a robust multilateral filter is applied to reduce noise in smooth areas, while preserving sharpness along the edges.

The proposed method has been implemented with GPU programming and tested on a computer with an Intel® Core™ i7 2.93-GHz CPU and an NVIDIA GeForce GTX 460* graphics card. Our implementation can produce an output of 26 fps on average for a 640×480 input video.

For a quantitative comparison with the state-of-the-art methods presented by Garcia et al.[4], we utilize the *Moebius*, *Books*, and *Art* scenes from the Middlebury dataset.[5] In this dataset, the disparity maps are of the same resolution as the color images. Hence, we generated the input disparity maps by downsampling the ground truth by a factor of 3x, 5x, and 9x.

Garcia et al.[4] use the structural similarity (SSIM) measure to compare the evaluated methods; however it is not appropriate for evaluation in this context. SSIM cannot yield meaningful results for the regions with unknown depth values and Middlebury's ground truth disparity maps contain such regions. Instead, for a fair comparison, we calculate the average percentage of bad pixels

with an error threshold of 1 for all known regions. In this measure, pixels with a disparity error greater than the threshold are regarded as bad pixels. This is the same scoring scheme employed in the Middlebury evaluation. Figure 4 and Table 3 show that our method performs better than all the competing methods.



(a) Color image (b) Ground truth (c) JBU (d) PWAS (e) AMF (f) Proposed

Figure 4: Visual comparison on the Middlebury dataset. The upsampling methods include: (c) JBU[6], (d) PWAS[7], (e) AMF[4], (f) proposed method
(Source: Kyoung-Rok Lee, PhD thesis, "Accurate, efficient, and robust 3D reconstruction of static and dynamic objects," UCSD, 2014)

Dataset		JBU[6]	PWAS[7]	AMF[4]	Proposed
	3x	7.43	4.68	4.5	**3.62**
Moebius	5x	12.22	7.49	7.37	**4.87**
	9x	21.02	12.86	12.75	**9.02**
	3x	5.4	3.59	3.48	**2.38**
Books	5x	9.11	6.39	6.28	**3.58**
	9x	15.85	12.39	12.24	**7.11**
	3x	15.15	7.05	6.79	**5.07**
Art	5x	23.46	10.35	9.86	**6.91**
	9x	38.41	16.87	16.87	**11.7**

Table 3: Quantitative comparisons (average percentage of bad pixels)
(Source: Kyoung-Rok Lee, PhD thesis, "Accurate, efficient, and robust 3D reconstruction of static and dynamic objects," UCSD, 2014)

In addition, we further evaluate our refinement method by applying the algorithm to all the disparity maps generated by the methods submitted to the Middlebury stereo evaluation. Figure 5 shows the improvement in terms of the percentage of bad pixels. Note that the proposed method improves the

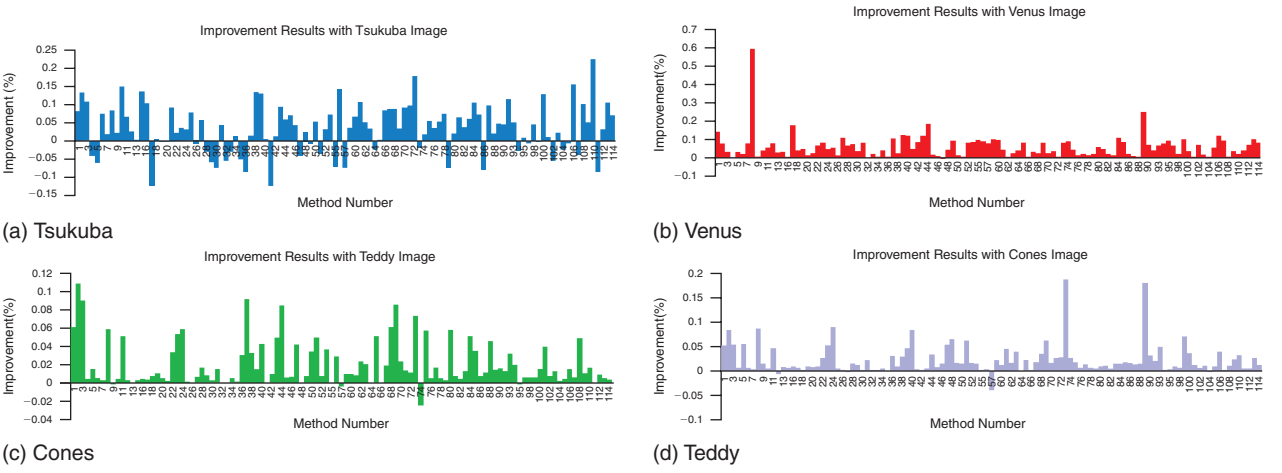


Figure 5: Percentage improvement in terms of number of bad pixels after applying the proposed algorithm to all the 109 methods on the Middlebury stereo evaluation[2]
(Source: Kyoung-Rok Lee, PhD thesis, "Accurate, efficient, and robust 3D reconstruction of static and dynamic objects," UCSD, 2014)

disparity estimates for the majority of the methods. One limitation of the proposed algorithm is that its performance drops when the input images are small and complex, or when the initial disparity estimates contain significant errors.

VQMT3D: Video Quality Measuring Tool for 3D

The quality of experience for any video viewing is important, especially for 3D and multiview content. As such, it's important to understand all the different aspects of 3D that affect the end user quality of experience. As such when it comes to evaluating the rate vs. distortion/quality tradeoff for delivering the content, the communications rate, pre- or postprocessing, and end user quality can all be jointly optimized. Therefore, significant effort in this project went towards understanding 3D video quality impairments and towards creating tools that can help identify these issues. A cinematographer of a 2D film should consider many factors simultaneously, such as scene composition, color camera settings, light, focus position, depth of field, and amount of zoom. A cinematographer of a stereoscopic film additionally pays attention to depth budget distribution and various cameras cross-settings, such as geometry, color, blurriness, and time synchronization.

We have checked the quality of 30 well-known films (besides converted films) and found more than 1000 frames with inter-view mismatches or excessive parallax.[8] In summary, we have concluded that manual quality control leaves a lot of artifacts even in released versions, resulting in less consistent and thus less redundant video-content, which is harder to compress. Therefore process automatization is a necessary and important requirement in stereoscopic video creation.

We started a project to automatically detect all common problems (Video Quality Measurement Tool 3D). Within the project, we developed a distributed system that produces per-frame charts of each metric and automatically extracts problem frames. Finally, the quality-estimating report is generated. We have already published four reports and collected feedback. More than 20 stereographers provided valuable feedback on found artifacts, and their comments were included in our reports.

Some films are captured in 2D and then converted to S3D at the postprocessing stage to avoid various mismatches during S3D capturing. However this approach produces its own specific artifacts, with no reliable classification. Currently, we propose algorithms to detect the cardboard effect and edge-sharpness mismatch artifacts. In future work, we plan to investigate other artifacts.

We believe that our work will motivate the development of stereoscopic video quality standards. In relation to VAWN program goals, that means that network traffic will be decreased since low-quality stereoscopic video contains inter-view mismatches requiring extra bits from encoder.

"We believe that our work will motivate the development of stereoscopic video quality standards."

Quality Issues in Stereo Video Capturing

Usually, the binocular impairments belong to one of the following types:

- geometry mismatch
- color mismatch
- sharpness mismatch
- time asynchrony

We designed a set of methods to automatically estimate each of them excluding time asynchrony. Another problem with stereoscopic videos is excessive parallax that causes accomodation-vergence conflict. We have a method to detect potentially problematic frames with respect to excessive parallax.

Excessive parallax. Many authors[9][10][11][12][13][14] consider the accommodation-vergence problem caused by excessive parallax to be the primary cause of visual discomfort and fatigue associated with viewing 3D. The problem, in essence, is that under natural viewing conditions, vergence and accommodation stimuli are equal to each other, whereas in stereo films, they can be significantly different due to excessive horizontal parallax. Normally, even in natural viewing, there exists a certain degree of tolerable mismatch between accommodation and vergence, since neither mechanism is ideal and both have limited accuracy and sensitivity. A large depth of focus makes accurate accommodation useless because retinal image quality is constant when changing accommodation by (0.2–0.3) D, an exact tuning to the object distance appears to be excessive, and accommodation mechanisms perform only minimal adjustment. From a functional point of view, high accuracy is not very important. On the contrary, a viewer must see clearly both the object of interest and its surroundings. Some quantitative data on this topic is provided by Daum.[15]

We designed an algorithm to detect excessive-parallax scenes.[16] It can be briefly summarized by the following steps:

1. Rough estimation of the inter-view disparity map
2. Disparity map filtering
3. Creation of a distribution graph that shows the number of pixels according to the disparity in the timeline. Beforehand, we ask the user to input the display system characteristics where the maximum artifact-free parallax level can be obtained.
4. Detection of the problem scenes.

An example of parallax monitoring graph is shown in Figure 6. We then mark frames that have a parallax exceeding a certain value.

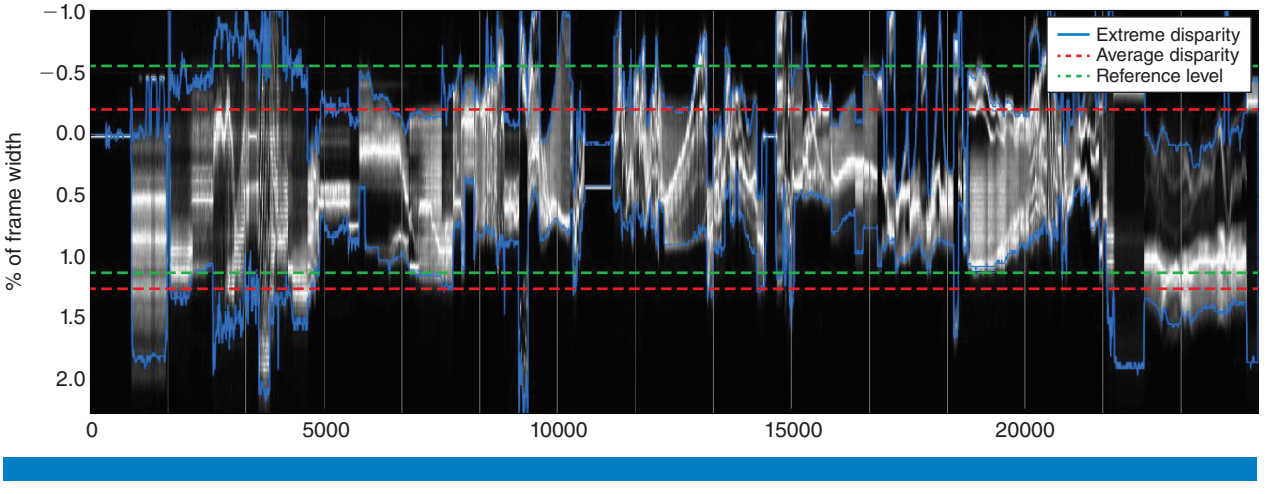


Figure 6: Example of a parallax control chart for the movie *Pina*. The parallax clearly differs from scene to scene and a large parallax is achieved only in a small portion of scenes.
(Source: Lomonosov Moscow State University, 2013)

Geometry distortions. When stereo images are displayed on a screen frontoparallel to the viewer so that horizontal lines on the screen are parallel to the line joining the eyes, all the conjugated points corresponding to virtual objects in the spatial scene should have only horizontal displacement, not vertical. Indeed, in the real world, any point in space, along with the two eyes of the viewer, defines the epipolar plane that intersects the screen in a horizontal line. Therefore, the rays directed from the eyes to a single object should intersect the screen at points located on horizontal lines; that is, at points displaced horizontally on the screen. Thus, to simulate a real scene, a correct stereo pair should have no vertical disparities, and on the screen, the left and right images must be presented without vertical disparities.

At the same time, it is necessary to take into account that in natural vision, vertical disparities are usually present. For example, in cases where the distances from the left and right eyes to the object in view are different, the two retinal images will be of different sizes in both horizontal and vertical dimensions. Psychophysical experiments reveal that the vertical-disparity gradients

significantly affect perception of 3D form, depth, and size.[17][18][19] Regarding quantitative data, Stevenson and Schor in 1997[20] found that matching stereo images is not restricted to epipolar lines and that people are able to estimate depth accurately even in cases of vertical disparities up to $45'$ (retinal angular minute). Currently we estimate vertical parallax and tilt.[16][21]

Color mismatch. It is well known that under natural viewing conditions involving a real scene, the binocular visual mechanisms use not only geometric disparities (cues based on the relative positions of objects in depth) but also differences in luminance, contrast, and color between the left and right images, as well as asynchrony in their temporal changes.[22] In particular, local differences in luminance are characteristic of images containing lustrous surfaces and transparent objects. In this way, color and luminance differences can cause false perception. Our system of binocular vision, however, is sufficiently robust to handle color differences between views. Recently, new data has surfaced indicating that depth perception can survive significant interocular differences in luminance levels up to 60 percent.[23][24]

"Our system of binocular vision, however, is sufficiently robust to handle color differences between views."

Our metric of color differences is described elsewhere.[16][21][25]

Sharpness mismatch. The effect of blur can significantly affect binocular perception. Specialists in binocular vision have long been aware that when image blur is asymmetric, the quality of the binocular percept is determined mainly by the sharper of the two images. In particular, this conclusion was mentioned by Stelmach et al.[26] This conclusion is true only for a certain range of stimulus parameters, however. A more recent study[27] has shown that the perceived quality of an asymmetrically degraded image pair is roughly the average of both perceived qualities when one of the two views is degraded by very strong JPEG compression. Kooi and Toet[28] also note that fusing a good image with a blurred image requires a few more seconds than fusing the original image pair. Unfortunately, insufficient data is available (concerning various questions that need to be clarified) to perform a comprehensive analysis of the issue.

A detailed description of our sharpness-mismatch detection algorithm is presented elsewhere.[21][25] The metric is designed to detect differences in high frequencies caused by focus mismatches and also by inaccurate postprocessing, differences in motion blur, and asymmetric compression.

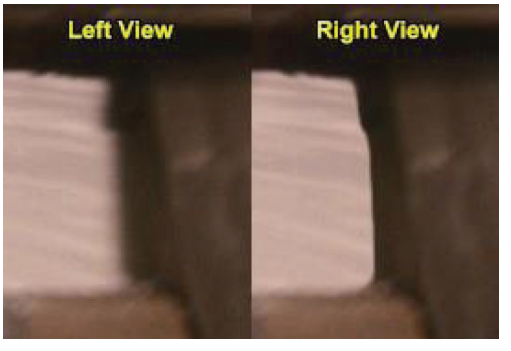
Quality Issues in Stereo Video Conversion

Converted videos are potentially better for VAWN goals since there exists a ground-truth sequence of depth maps, thus they can be easily represented in 2D+Z or MVD formats (we will discuss the advantages of these formats for compression in the section "Delivery"). However, the process of conversion requires significant manual work (including depth drawing), thus the price/quality ratio is low, resulting in audience distrust. Our research on the quality of captured stereoscopic video motivates us to study the quality of 2D-to-3D conversion. We believe that our work will improve quality of stereo creation by 2D-to-3D conversion.

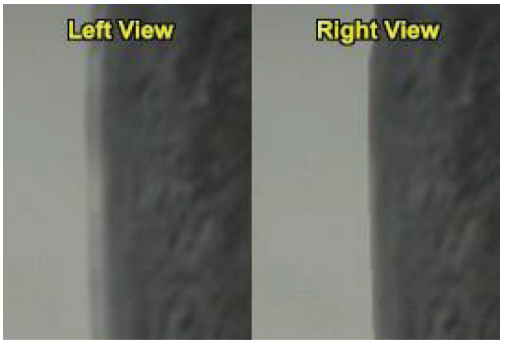
"Our research on the quality of captured stereoscopic video motivates us to study the quality of 2D-to-3D conversion."

Edge-sharpness mismatch. The term edge-sharpness mismatch (ESM) describes defective stereo pairs with specific asymmetric impairments. It refers to any inconsistencies in object edges between the stereoscopic views (edge-sharpness variation, edge doubling, and so on). Under the viewing conditions in a real environment, such situations rarely occur. In the case of 2D-3D conversion, however, the likelihood of ESM occurrence can be rather high. During the 2D-3D conversion workflow, ESM can be caused (besides an inaccurate depth map) by:

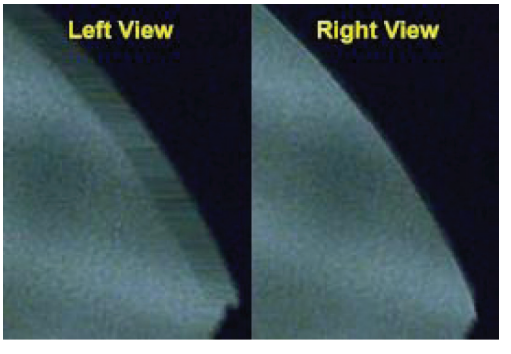
- Use of a "rubber sheet" occlusion-filling technique, defined as warping the pixels surrounding the occlusion regions to avoid the explicit occlusion-filling step (Figure 7a)
- Lack of proper alpha-channel treatment (Figure 7b)
- Simple occlusion-filling techniques where background or foreground pixels are stretched across the entire occluded region (Figure 7c)



(a) Result of "rubber sheet" occlusion filling



(b) Result of the absence of proper alpha channel



(c) Result of a very simple occlusion-filling technique based on stretching foreground and background pixels

Figure 7: Examples of edge-sharpness mismatch. Each example is a magnified fragment of a stereoscopic picture
(Source: Lomonosov Moscow State University, 2014)

Our proposed approach[29] is based on edge detection and matching:

1. Disparity map construction
2. Estimation of edge map for each view
3. Matching edge pixels using disparity map

4. Per-pixel edge sharpness mismatch estimation
5. Rejecting the results when the background changes significantly

Cardboard effect. Cardboard effect is a term referring to an unnatural flattening of objects in perceived visual images (the objects look like pieces of cardboard placed parallel to the screen). Long before video professionals gained widespread experience with stereoscopic movies, stereo photographers observed the cardboard effect. These photographers analyzed its causes and tried to formulate shooting conditions that minimize it.[30] In the case of 2D-3D conversion, the cardboard effect refers to mismatch between the perceived frontal size of the object (the visual angle occupied by the object in the visual field) and its perceived depth (thickness).

The proposed algorithm[29] for flat foreground objects detection consists of the following steps:

1. Disparity map construction
2. Mean-shift segmentation of the obtained map
3. Calculation of the median disparity value for each segment
4. Calculation of the variance around this value

Delivery

The main problem underlying the VAWN program is the rising amount of video content being transferred over wireless networks while network bandwidth remains the same. The increasing popularity of 3D video makes the problem even worse because of the additional data required for 3D. Hence the delivery of 3D video with the least amount of bitrate increase is desired. Although some communication networks cannot be extended for various reasons (such as cost and physical limitations), a 3D-specific video codec can help minimize this additional bitrate. However, most existing 3D video compression methods yield a significant bitrate increase over their 2D counterpart to achieve the same perceptual quality.

The current standard for 3D video compression is the multiview coding (MVC) extension of H.264/AVC. Bitrates generated by MVC are linearly proportional to the number of encoded views[31], and thus not scalable for use with autostereoscopic displays. Moreover, with multiview video representation, the number and location of the views are restricted to the captured data. In this work we pay attention to promising encoding techniques based on depth map storage that address the limitations of MVC. Specifically, two approaches to compact depth map storage are presented based on existing MVD and 2D+Z formats (Figure 8).

"In this work we pay attention to promising encoding techniques based on depth map storage…"

Multiview + depth format (MVD) assumes depth map storage and transmission for one view or several views. The presence of a depth map significantly decreases the complexity of intermediate view generation. This format is also compatible with display systems with two views. The size of additional data is even larger here than for pure multiview formats. However,

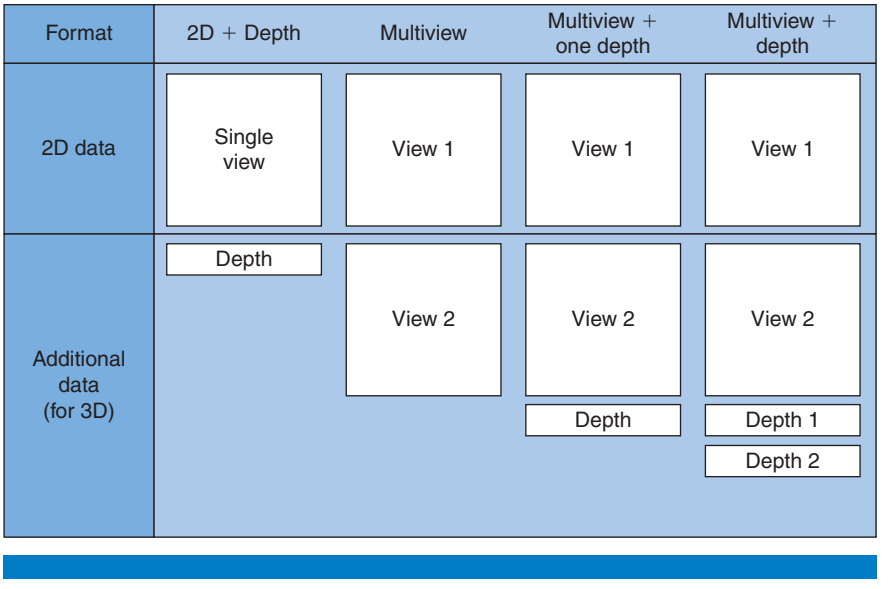


Figure 8: Comparison of 3D video formats in terms of additional data required for extending 2D video to 3D
(Source: Lomonosov Moscow State University, 2013)

the additional depth information enables new coding tools that exploit the correlation between the views of a 3D video better than existing tools. With depth-based 3D video, it is possible to synthesize virtual views and use them as a means of prediction within the codec. Currently, the Joint Collaborative Team on 3D Video Coding Extension Development (JCT-3V) is working on the standardization of a depth-based 3D video representation and its coding methods.

The 2D+depth format is supported by the 2D-to-3D video conversion industry, Dolby 3D format[32], and several TV set manufacturers. This format provides a 3D video experience with minimal additional data. The depth map can be compressed very effectively so almost all available bandwidth is used for 2D streaming. Consequently, this format is the most promising one in terms of compatibility with 2D devices and minimal additional data. This section presents three main options for reducing the overall bitrate needed to deliver 3D content: (1) 2D+depth compression, (2) MVC+depth compression, and (3) mixed resolution coding.

Multiview Video Compression Based on Depth Map Propagation

Let's consider the additional data that should be added to a 2D video stream to provide a 3D video experience. To minimize this data, we use the most effective 3D video format: 2D+depth.

The naïve solution for 2D+Z compression is using conventional 2D video codecs for both 2D and depth maps. This approach doesn't take advantage of the strong correlation between the 2D view and depth map. The depth map also has a structure different from the video structure, and since it has

no texture, conventional video compression approaches are not efficient. There are several methods for utilizing correlation in 2D video and depth maps. For example, Choi et al.[33] use a 2D image to increase the frame rate and resolution of the depth map obtained using a depth sensor. Modified cross-bilateral filtering is used to increase spatial resolution. Frame rate is doubled using temporal interpolation based on motion vectors estimated from 2D video. Depth map restoration from sparse key frames requires a more complex interpolation procedure because of the larger difference between distant frames.

De Silva et al.[34] perform joint compression of video and depth maps using motion vectors estimated from the source video. The input scene is segmented and extracted background is transmitted independently. Depth maps are considered only for foreground objects. Additionally, motion vectors are estimated and transmitted with the compressed video stream.

Our proposed approach is based on sparse representation of depth maps. The general pipeline is shown in Figure 9, where only downsampled key frames for a depth map are compressed. In the simplest case, the encoder for a depth map only selects every *n*th frame, downsamples it, and compresses using any conventional image or video codec, yielding a very low bitrate for the depth map.

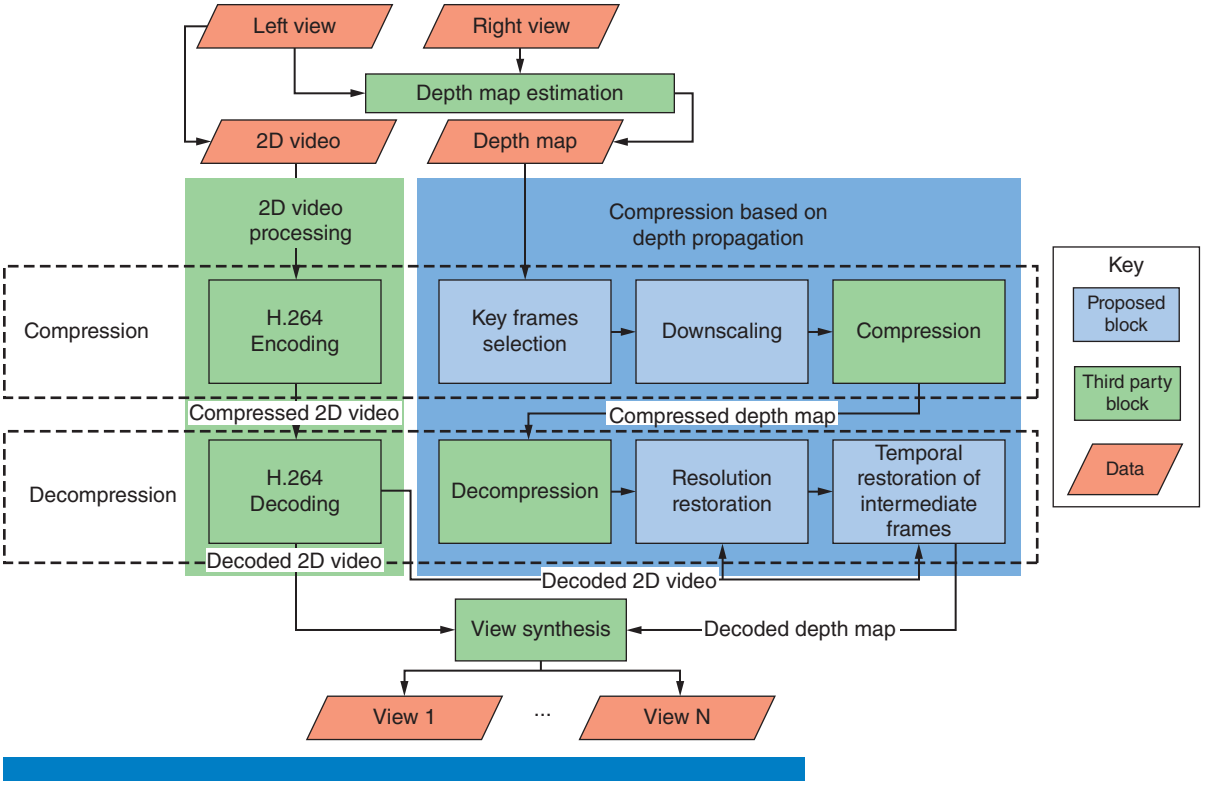


Figure 9: Depth propagation-based compression scheme for 3D video (Source: Lomonosov Moscow State University, 2013)

The decoder decompresses the 2D video stream and then uses it for depth map decompression. First the decoder restores spatial resolution of the depth map for key frames using appropriate decoded high-resolution 2D frames. This can be done using numerous depth map upscaling methods.[4][6][35] The method is used by YUVsoft Depth Upscale[35], where rough edges in a low-resolution depth map are enhanced using a high-resolution 2D color frame.

Then the decoder restores missing depth maps for the rest of the frames. We use block-based motion estimation to get information about object motion from the 2D video. The depth map is assumed to be correlated to 2D video motion flow and the depth map can be interpolated using key frames and motion information. Appropriate nonlinear edge-aware postprocessing is used to conceal motion compensation artifacts. Such a propagation procedure is held for each video interval between key frames. Depth is propagated from the first and the last depth key frame of the interval. Then the two depth maps are merged according to the degree of confidence in motion information, based on Simonyan et al.[36]

The final decoding step depends on the target device, where the required views are generated using a dedicated algorithm.

Quantitative Results

The performance of depth-propagation–based compression is compared with depth map compression using x264 codec. 2D video compression is the same for both cases, so the bitrate of 2D video is not considered. We consider only the bitrate of additional data (that is, depth map bitrate).

"Due to the complexity of stereoscopic perception there is no generally accepted method for quality measurement."

Due to the complexity of stereoscopic perception there is no generally accepted method for quality measurement. Research in this direction is still in progress.[37][38] The most common approach is using a conventional 2D quality metric (such as PSNR). Direct measurements of the depth map difference before and after compression are not relevant because the depth map influences synthesized view quality in a nontrivial way. Therefore 2D quality metrics are applied to synthesized views, for example by Lee et al.[39] Both methods were used for quality evaluation of the propagation-based approach.

In our experiments, we used constant intervals between key frames. We tested intervals of 10, 20, 40, and 100 frames. Depth key frames were downscaled using factors 1, 2, and 4. For key frame compression we used the JPEG 2000 image codec. Full-resolution key frames are restored using YUVsoft Depth Upscale.[35] The full depth map is restored from key frames using YUVsoft Depth Propagation.[40] We took the best settings of the proposed propagation scheme in terms of quality and bitrate.

The proposed technique demonstrates very promising results (Figures 10 and 11). It outperforms depth map coding using x264 up to 15 times in bitrate while preserving the same quality. The highest gain is achieved on the lowest bitrates.

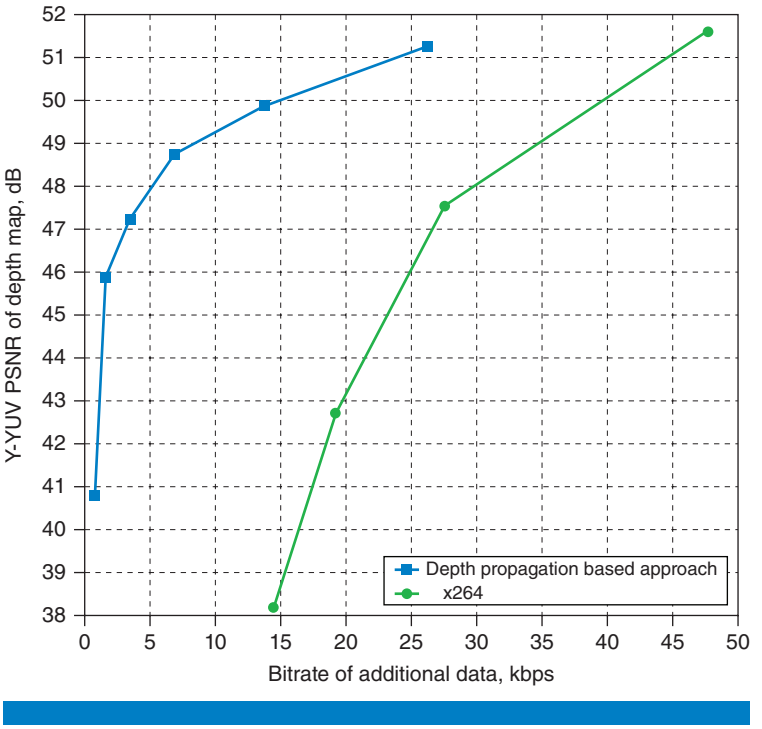


Figure 10: Comparison of depth map compression using depth propagation based approach and x264. Only the bitrate of additional data is considered.
(Source: Lomonosov Moscow State University, 2013)

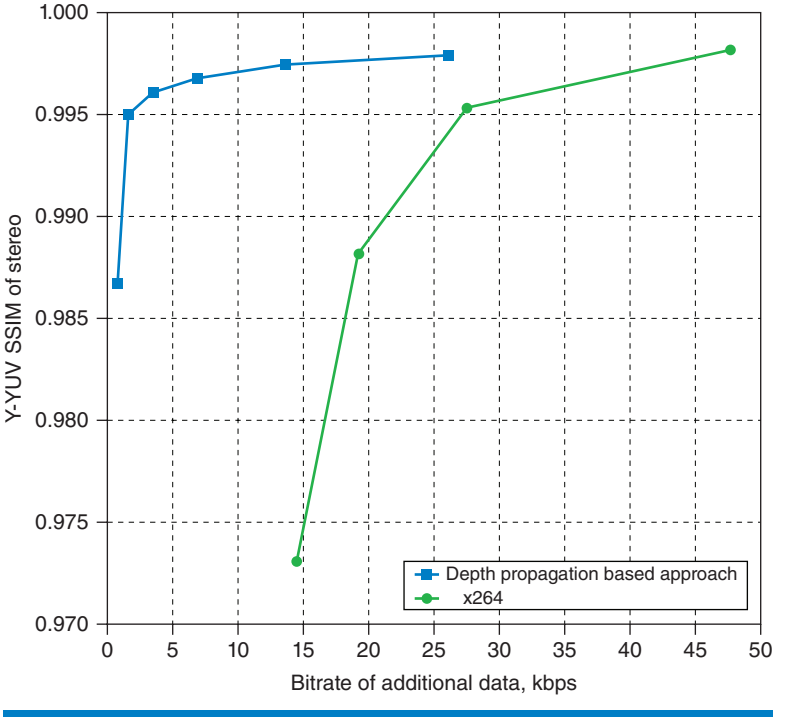


Figure 11: Comparison of reconstructed stereo for depth propagation based compression and x264. Only the bitrate of additional data is considered.
(Source: Lomonosov Moscow State University, 2013)

While the depth-propagation–based approach demonstrates good potential compression, there are a number of ways it can be improved. One of the improvements is adaptive key frames selection. Static scenes require a lower bitrate, so the algorithm can use sparser key frames. Dynamic scenes must be encoded using dense key frames to achieve acceptable quality. Starting from a sparse arrangement, we add key frames to maximize the target metric. An example of per-frame SSIM of stereo for a part of the *Basketball* sequence is presented in Figure 12. Adaptive key frames selection demonstrates even more gain in comparison with the basic approach with equidistant key frames, shown in Figure 13.

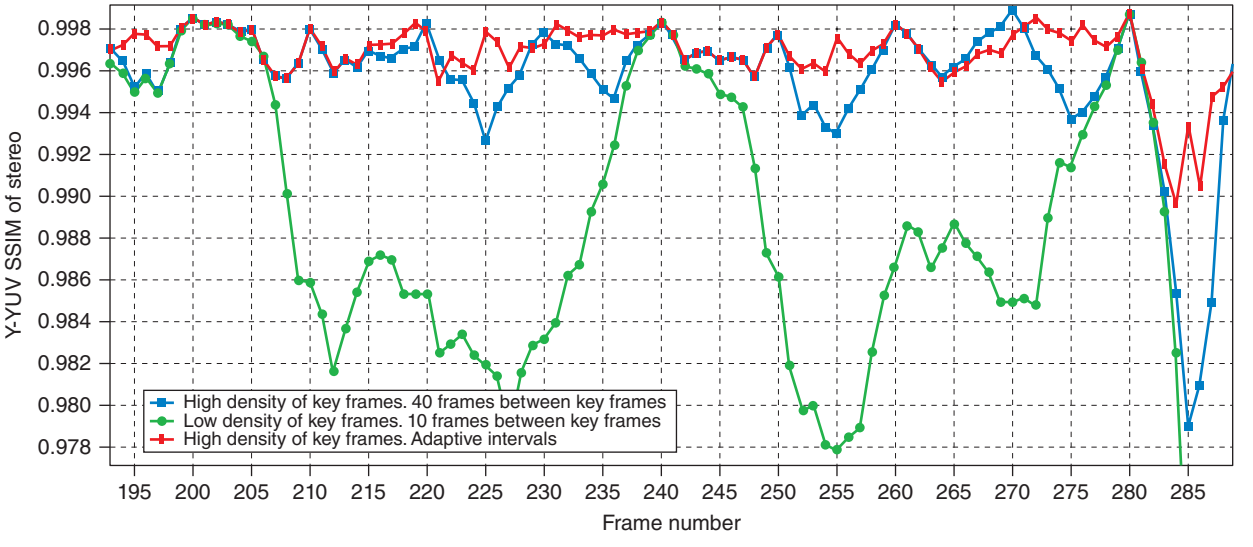


Figure 12: Per-frame SSIM scores of proposed depth-propagation-based compression with different key strategies for choosing frames
(Source: Lomonosov Moscow State University, 2013)

Advantages and Disadvantages

The proposed technique demonstrates good compression performance, where a quality 3D experience is delivered with minimal additional bandwidth. Compression efficiency is achieved at the expense of high computation cost of decompression. Depth map propagation on the decoder requires a rather large buffer of decoded 2D frames. The number of frames depends on the compression rate, which can reach tens of frames. Consequently, a small delay at the decoder is inevitable. The depth-propagation–based approach inherits all the advantages and disadvantages of the 2D+Z video format. This format and MVD support an arbitrary number of displayed views and arbitrary parallax tuning in a reasonable range. This capability is important due to the variety of existing 3D display sizes and formats. Each of them

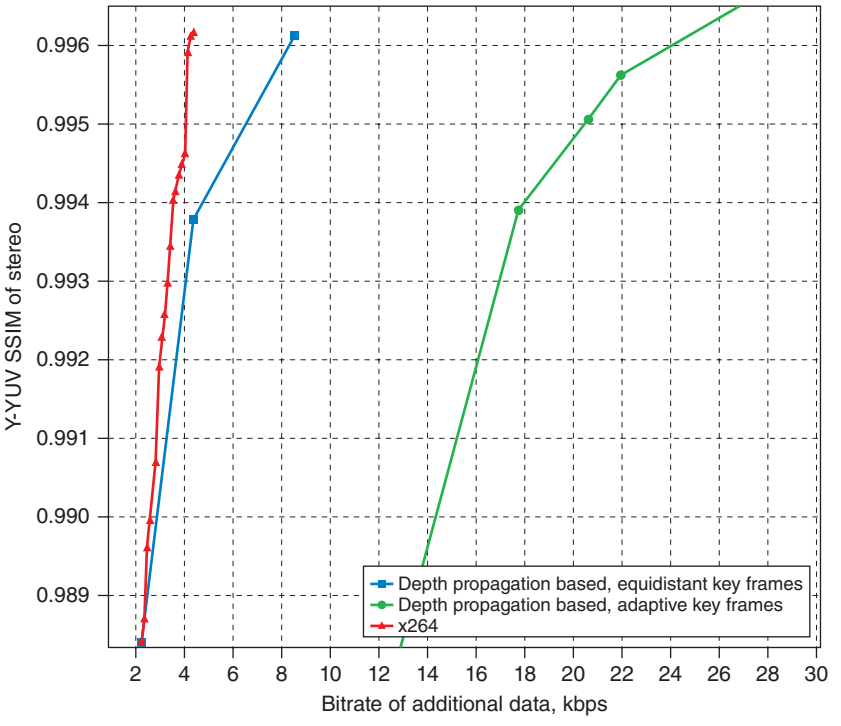


Figure 13: Comparison of reconstructed stereo for depth-propagation–based compression and x264 on a sequence from the *Pirates of the Caribbean* trailer. Only the bitrate of additional data is considered (Source: Lomonosov Moscow State University, 2013)

requires appropriate stereo parameters. The 2D+Z format is also supported as a native format in a variety of displays. The 2D+Z format is impossible to use for correct processing of semitransparent objects. However, this drawback can be concealed by additional data for semitransparent region representation. The quality of the final image strongly depends on the quality of stereo occlusion processing. The inpainting algorithm for these areas is critical, although the inpainting area is typically small. On small screens, parallax must be high, but the overall size of the display is low. On the large screens, parallax must be small, so the area to fill is not very big. This allows one to successfully apply inpainting algorithms.

Multiview Video Plus Depth Coding with Depth-based Prediction Mode

An alternative to the 2D+depth compression technique discussed in the previous section is to utilize depth information as a complement to existing multiview coding (MVC). This section describes how this can be done and quantifies the resulting improvements in quality and bitrate needed to deliver the content.

Depth-based Prediction Mode (DBPM)

DBPM allows the use of a synthesized reference picture for prediction without any high level syntax changes to the MVC and requires only simple macroblock-level syntax changes to the standard. It can be used concurrently with existing prediction modes of MVC without introducing a significant overhead due to changes in syntax.[41][42]

In our codec, the base (first) view video and its associated depth map are encoded with H.264/AVC individually. Then while encoding a frame of an additional view, a virtual view is rendered from the base view data at the corresponding time instance. The virtual view is rendered simply by the well-known Depth Image Based Rendering (DIBR)[43] algorithm by projecting the base view pixels onto the current viewpoint, without any hole filling at the disocclusion regions. Once the proposed mode is signaled to the decoder, the decoder refers to this reference virtual view and copies the collocated macroblock for prediction. If there is additional residual information in the bitstream it also adds the residual information. An illustration of DBPM is provided in red in Figure 14a along with the DIBR operation, which is depicted in blue. A block diagram of the proposed codec is provided in Figure 14b.

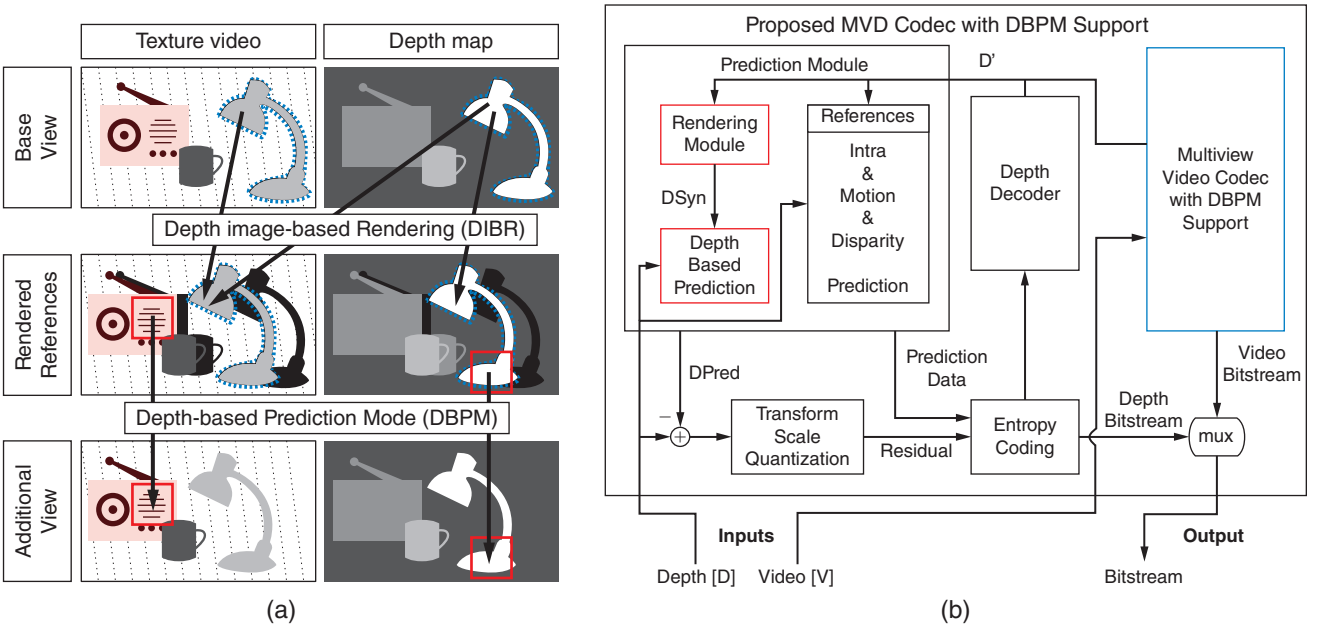


Figure 14: (a) Illustration of the depth-based prediction mode (b) Block diagram of the proposed MVD codec with DBPM support

(Source: C. Bal and T. Q. Nguyen, "Multiview Video Plus Depth Coding With Depth- Based Prediction Mode," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 6, pp. 995–1005, Jun. 2014)

Rate-Distortion Analysis

We compare the coding performance of the proposed codec with MVC using rate-distortion (RD) curves and Bjontegaard Delta Rate (BD-Rate).[44] RD curves measure the coding performance of a codec in terms of a quality metric (such as PSNR) and the corresponding bitrate levels. BD rate values measure the percent bitrate savings between two RD curves, where negative values signify gain. We analyze the rate-distortion performance of DBPM in two different contexts. First, we provide results for the proposed MVD codec. The proposed codec encodes the texture videos and the depth maps, both with DBPM support. In comparison, we use MVC to encode texture and depth channels disjointly. Second, to isolate the contribution of the DBPM support for depth maps, we analyze its prediction performance in the context of depth map coding.

The BD rate results for coding MVD data using the proposed codec versus MVC are reported in Table 4. These results show that the proposed codec can achieve up to 9.2 percent bitrate savings with DBPM support, and as expected, the gains vary depending on the depth map quality. For example, *GTFly* and *UndoDancer* are among the sequences with the largest gain since they are computer-generated sequences with ground truth depth maps. In comparison, the depth maps of the *Newspaper1* sequence are noisy and consist of both temporal inconsistencies and spatial errors. This leads to inefficient coding of the depth maps and geometric distortions in the rendered references for DBPM. Thus, *Newspaper1* is among the sequences that benefited the least from DBPM support.

Depth QP	PoznanHall2	PoznanStreet	UndoDancer	GTFly	Kendo	Balloons	Newspaper1
26	–5.26	–3.81	–7.06	–9.19	–2.03	–2.57	–2.55
31	–5.35	–4.04	–5.52	–9.03	–2.53	–3.16	–3.17
36	–4.95	–3.75	–3.93	–8.51	–2.84	–3.33	–3.52
41	–4.49	–3.09	–2.94	–7.50	–2.88	–3.53	–3.64

Table 4: BD rate (%) for coding MVD data, 3 views—measured against MVC (Source: C. Bal and T. Q. Nguyen, "Multiview Video Plus Depth Coding With Depth- Based Prediction Mode," *IEEE Trans. Circuits Syst. Video Technol.,* vol. 24, no. 6, pp. 995–1005, Jun. 2014)

BD rate allows measurement of bitrate savings in a concise manner, yet it fails to associate these savings with the absolute number of bits saved. Hence, in addition to the BD rate results, we also provide the RD curves for the *GTFly*

sequence, which yields the most gain for the proposed codec. Figure 15 shows that the proposed codec can deliver the same quality of MVD data with up to 900 kbps less bitrate than MVC.

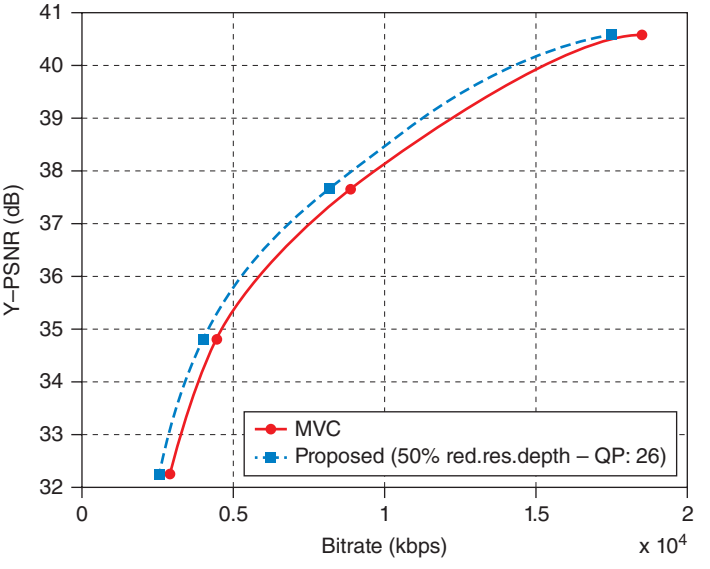


Figure 15: RD curves for *GTFly*, 3 views, coding MVD data (Source: C. Bal and T. Q. Nguyen, "Multiview Video Plus Depth Coding With Depth- Based Prediction Mode," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 6, pp. 995–1005, Jun. 2014)

We also provide BD rate results for depth map coding in Table 5. Since depth maps consist of piece-wise smooth regions, DBPM faces stronger competition against existing MVC prediction modes. Looking at the results in Table 5, DBPM proves to be successful with up to 9.9 percent bitrate savings when the depth maps are accurate. On the other hand, for depth maps with limited accuracy, the encoder chooses DBPM infrequently and the DBPM-enabled codec starts to yield slightly worse performance than MVC. These bitrate losses are limited to around or less than 1 percent, and they are due to the syntax overhead introduced by DBPM.

PoznanHall2	PoznanStreet	UndoDancer	GTFly	Kendo	Balloons	Newspaper1
1.03	–1.69	–2.98	–7.56	0.08	0.32	0.75

Table 5: BD rate (%) for coding depth maps, 3 views—measured against MVC (Source: C. Bal and T. Q. Nguyen, "Multiview Video Plus Depth Coding With Depth- Based Prediction Mode," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 6, pp. 995–1005, Jun. 2014)

Mixed Resolution Stereoscopic Coding

A final technique to reduce the required over-the-air bitrate is to blur one eye's view compared to the other, allowing for a lower overall bitrate. Mixed resolution stereoscopic coding (MRSC) relies on the perceptual phenomenon of binocular suppression, where if one eye's view of the world is blurry while the other eye's view is sharp, then the fused 3D percept of the scene will appear almost as sharp as the high resolution view and will be faithfully represented in depth.[45] Mixed resolution coding implements this idea by transmitting a stereo pair comprised of one full resolution image and one lower resolution image.

"Mixed resolution stereoscopic coding (MRSC) relies on the perceptual phenomenon of binocular suppression…"

One concern for MRSC is that one eye continually receives a low resolution or blurry input. In the following subsection, we investigate the perceptual response for two methods of MRSC.[46] The first method, single-eye blur, is to blur all frames of the video corresponding to one of the eyes. The second method, alternating-eye blur, is to blur alternate frames of each view, such that there is one blurry and one sharp frame at each time instance, and the view that is blurred alternates with each frame.

There are applications, such as high quality viewing or decoder-side processing, for which a full-resolution stereo pair is beneficial. A super-resolution algorithm for single-eye blur MRSC videos is presented by Jain and Nguyen.[47]

Quality Experiment

In order to compare the perceived quality of the two processing methods, we asked subjects which of the pair of videos, each processed according to one of the two methods, they preferred. We used four high quality stereoscopic video clips, four blur levels corresponding to a diameter of 2, 4, 8, or 16 pixels (1.1 arcmin to 9.0 arcmin) of a disk kernel, and three frame rates: 30 Hz, 60 Hz, and 120 Hz. Each of the 48 unique test conditions was repeated four times for a total of 192 trials, which were tested in random order. Stimuli were presented on a pair of CRT displays viewed through a mirror stereoscope at a distance of 6.4 feet (6° horizontal per eye). Twenty-three subjects participated.

Figure 16 shows the proportion of trials where subjects preferred single-eye blur over alternating blur, as a function of blur diameter, for the three different frame rates. We did not see any consistent difference in preferences for blur type between the four different source videos, so we have combined data across that factor.

For a refresh rate of 30 Hz it is clear that single-eye-blur is preferred. For 60 Hz and 120 Hz, there is no real evidence of a preference below blur diameters of 8 pixels.

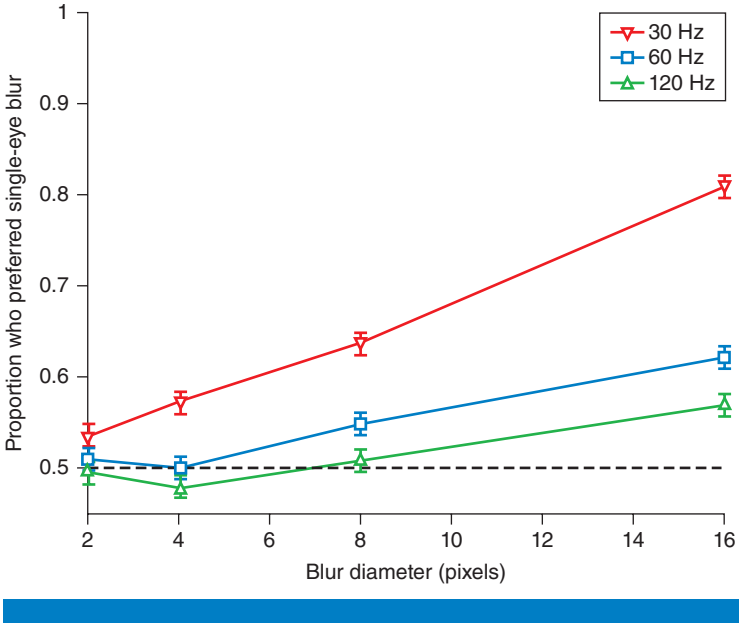


Figure 16: Proportion of trials in which single-eye blur was preferred over alternating-eye blur
(Source: Ankit K. Jain, PhD thesis, "Perceived Blur in Stereoscopic Video: Experiments and Applications," UCSD, 2014)

Fatigue Experiment

In this experiment, we compiled video clips into a single 5-minute source video and looped it twice to make a 10-minute exposure. We showed subjects this video processed according to each of the two processing methods and had them continuously rate their visual comfort level using a slider with scores ranging from 1 (very uncomfortable) to 5 (very comfortable) using the system presented by Jain et al.[48] Twenty-two subjects participated.

The mean scores across subjects over the duration of the test videos are shown in Figure 17. For both test methods, the mean scores generally range from 3 to 4 (fair to good comfort). The alternating blur method is rated higher, with an overall mean of 3.86 compared to 3.74 for the single-eye blur.

The dashed lines in Figure 17 indicate the ends of each of the four clips, and the dotted lines indicate a scene change. In addition to the general preference for alternating blur over time, there is some dependence on the type of content as certain clips are less straining to watch than others. For instance, the "Looney Tunes" clip from about 2:05–4:25 and again from 7:05–9:25 reflects the largest difference in scores between the two methods, with the alternating blur being preferred. The animation contains high contrast, flat textures, and sharp edges. All of these features produce

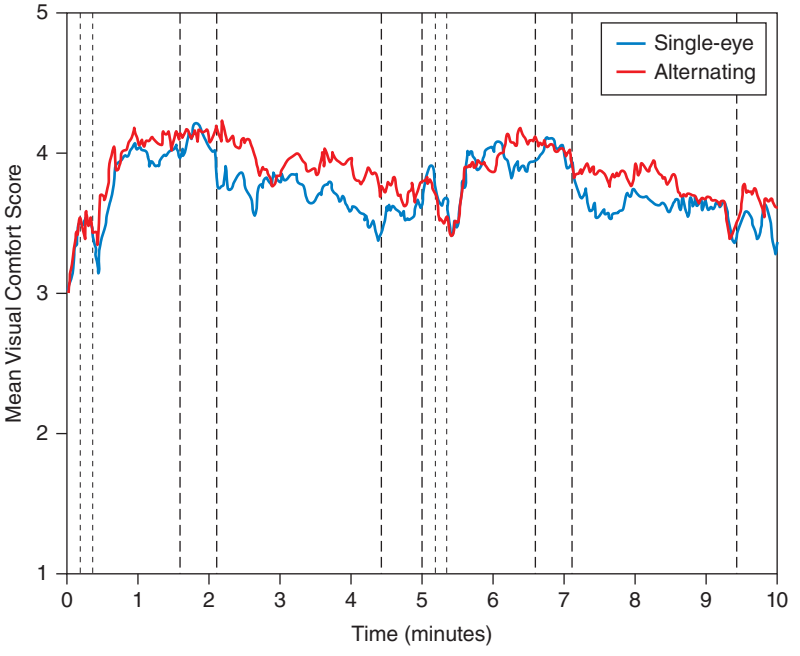


Figure 17: Mean scores across subjects over duration of video for each processing method. Dashed lines separate each video clip in the sequence
(Source: Ankit K. Jain, PhD thesis, "Perceived Blur in Stereoscopic Video: Experiments and Applications," UCSD, 2014)

artifacts under asymmetric blur, which are quite salient in the single-eye blur case.

The mean scores are relatively high for both methods considering the level of blur applied to the sequence. A 20-pixel diameter was chosen for the disk filter, which corresponds to a downsampling ratio of 8.33 in each dimension or about 69.44 overall.

Display

Usually, the quality of compression is measured by a rate-distortion ratio, thus there exist two ways to decrease traffic over networks:

- to decrease the bitrate for the same distortion level,
- to decrease the level of distortions for the same bitrate.

This section is mainly about the second approach, but viewing the distortion and quality from the end user perspective. For 3D viewing, quantifying quality is still an unknown and difficult task. Without a way to quantify quality for different bit rates and distortions, it is difficult if not impossible to optimize the end-to-end delivery system, which was an objective of the VAWN research. The final viewing quality is determined by the quality of the video itself and

the display equipment. Thus, it is important to study the quality of the whole stereoscopic video life cycle and the problems of automatic stereoscopic-display adjustment and choice of the proper content. The first part of this section will cover the unique processing requirements needed to render content for autostereoscopic displays, which relies heavily on an accurate depth map as discussed earlier. Then a tool will be presented that helps to capture subjective quality scores efficiently, which can hopefully be used more broadly with the industry in order to better quantify 3D video quality for different content, bitrate, and distortions. Finally, another tool is presented to help automatically compare the quality of the end display, since displays have a significant impact on the end user quality. It is hoped that these tools will help move the industry towards a deeper understanding of the end user perception of 3D video quality.

Autostereoscopic Display

Various autostereoscopic display technologies have surfaced in recent years. In general, they work by projecting images of a scene into the space in front of them to create two or more spatially separated perspectives. Then, based upon where an individual stands in reference to the display, each eye will perceive a different viewing angle, which leads to a disparity between what each eye sees. This disparity is translated into depth perception by the human visual system.

Although we typically use an 8-view display for a richer 3D effect, for simplicity Figure 18 illustrates a display with just two views using a sheet of

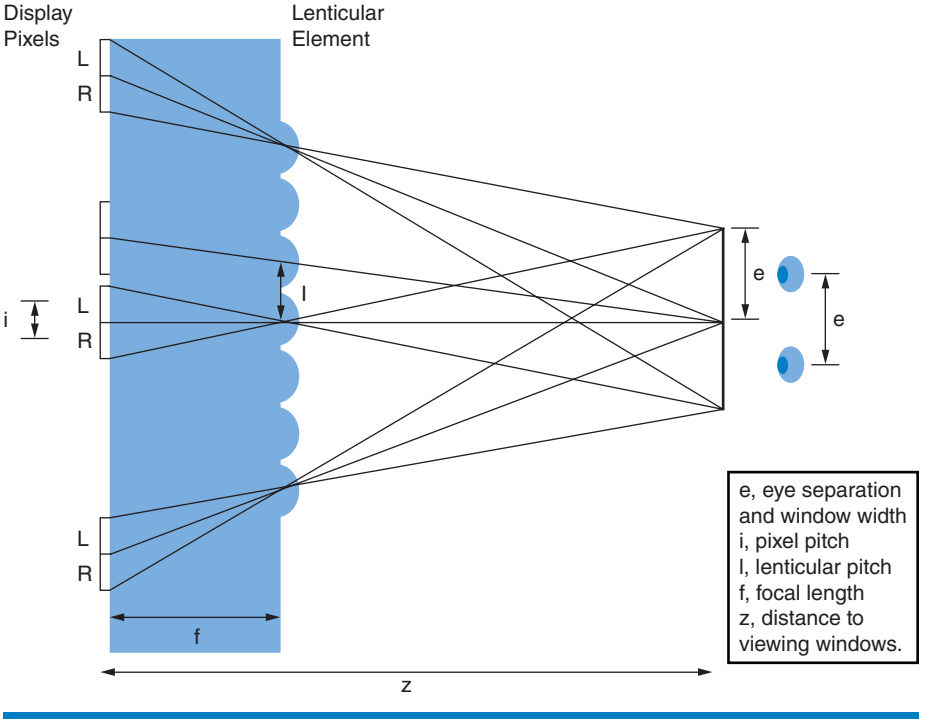


Figure 18: A top-down view of a 2-view lenticular display (Source: Ramsin Khoshabeh, PhD thesis: "Bringing Glasses-free Multiview 3D into the Operating Room," UCSD, 2012)

lenses called a lenticular sheet. The curved lenses angle the light emitted from a traditional LCD in such a way that the image incident on the left eye is slightly different from that of the right eye, creating a 3D sensation. As can be readily seen from Figure 18, with just a 2-view display, only one viewer can perceive 3D from a fixed location. Using an 8-view display allows for eight viewing zones with eight different perspectives, also known as sweet spots. This effectively allows all onlookers to see 3D from multiple vantage points. A typical problem with autostereoscopic displays is that moving between these sweet spots can produce an experience that resembles double vision when, for example, the eyes are seeing half of two separate views.

While autostereoscopic displays offer users the ability to see 3D without having to wear any specialized glasses, they require multiple viewpoints of the scene in order to display content. Typically, they require five, eight, or nine stereoscopically aligned views in order to properly display a 3D effect. In our case, this means that we would need to construct a camera system with at least eight cameras. In addition, there are strict requirements that the cameras must be as nearly identical to each other as possible, and extremely and fixed in orientation for an accurate 3D representation. Therefore, it is impractical to capture data with such camera systems. Instead, we limit ourselves to capturing the data with just two cameras. With that, it becomes a matter of providing a robust stereo-to-multiview conversion solution to take in the camera input and visualize it on an autostereoscopic display. Figure 19 shows the result of our work rendering the remaining six stereoscopically aligned views given a pair of images as input.

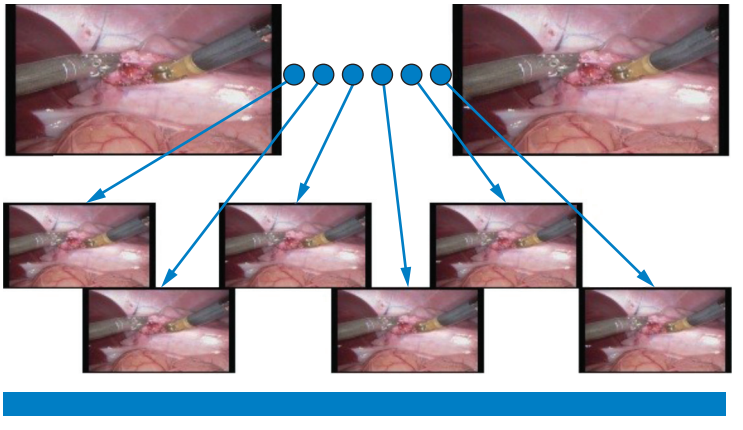


Figure 19: Rendering of six interpolated views from a stereo pair (Source: Ramsin Khoshabeh, PhD thesis: "Bringing Glasses-free Multiview 3D into the Operating Room," UCSD, 2012)

Tally: A Subjective Testing Tool

Many labs around the world conduct research that relies heavily on perceptual experiments with video. Commonly, data is collected by asking subjects to write their responses to stimuli using pen and paper, and then manually entering this data into computerized spreadsheets. Not only is this process extremely slow,

it is also prone to error. Some researchers have written custom software to automate this process, but it is not generally applicable, is not made widely and freely available, does not work for both 2D and 3D content, and does not permit testing multiple subjects simultaneously.

We developed Tally[48], a subjective testing tool, as a web-based system to solve these problems. Additionally, Tally's web-based design allows data to be accessible from anywhere, allows many people to use the same system with their individual history and data securely saved and privately accessible, allows experiments to be repeated with identical parameters and methods, and allows remote collaboration between labs through a sharing feature.

Our system consists of three major pieces: the desktop application, the web front end, and the server back end. The basic workflow of a subjective experiment is depicted in Figure 20. Prior to the experiment, the researcher uses the web front end to create a test and run it. Then, subjects log into the web front end (website) and select the appropriate test to begin. Once they are ready, the server tells the desktop application which video to play. The desktop application receives the command and plays the video to the display device.

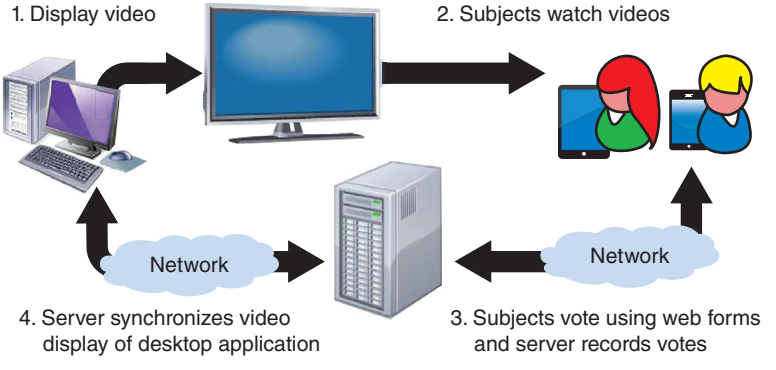


Figure 20: Workflow of the subjective testing tool (Source: Ankit K. Jain, PhD thesis, "Perceived Blur in Stereoscopic Video: Experiments and Applications," UCSD, 2014)

Note that only the file name is sent across the network; the actual videos are stored locally on the machine connected to the display. Subjects then vote on the video using any web-enabled device such as a smartphone, tablet, laptop, or desktop, and their scores are transmitted to the server and recorded. Once the video is done playing, the server tells the desktop which video to play next, and the process repeats until all test videos have been shown. After the test, the server automatically aggregates the data and makes it available for download in several different formats.

Tally is free, open source, cross-platform, and very customizable. Any web-enabled device can be used to vote, most any video player can be used to play the videos, and any display device can be used to show the videos. We natively

support most of the standard test methods of the ITU[49], but also allow for custom test methods to be added. Tally, along with full documentation and installation instructions, is available for download at the project website (http://github.com/canbal/Tally).

Automatic Device Testing

Finally, the display device itself has a significant impact on the end user perceived quality of the 3D video. Therefore, it's important to understand how to evaluate the quality of the display and how different displays compare. Eventually, this information could be used to optimize the compression and bitrates needed to deliver a good quality of experience that takes into account the specific characteristics and quality of the display. The problem of viewing-device fair comparison existed long before the market of 3D viewing devices started growing and some solutions were proposed.[50] Nowadays the problem is urgent again. The 3D viewing devices are much more diverse than 2D ones and the space of their characteristics has more dimensions. Figure 21 shows a partial classification of existing device types.

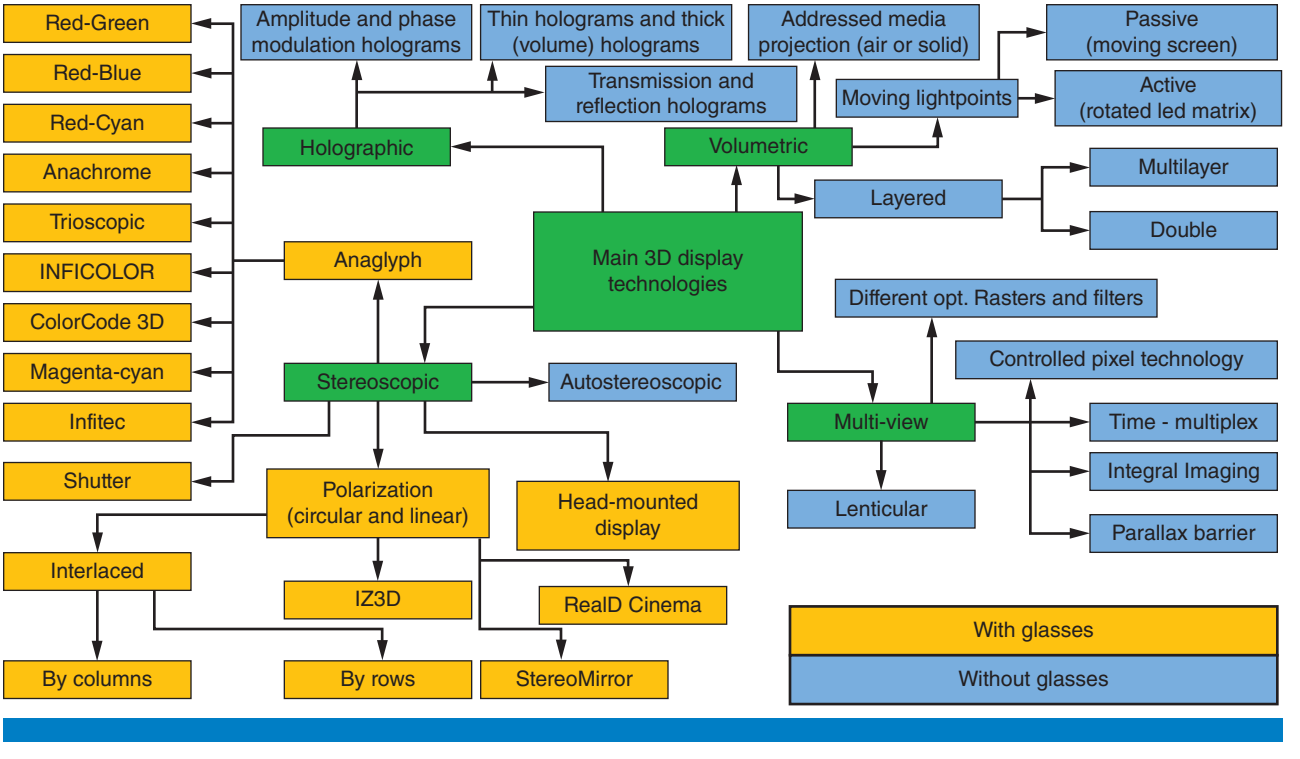


Figure 21: Partial classification of existing stereoscopic devices (Source: Lomonosov Moscow State University, 2012)

Absence of a fair comparison methodology leads to unfair competition between manufacturers and undermines user confidence in the whole market. Creation of easy-to-use software that performs a complete estimation of 3D viewing device characteristics and a database with a detailed description of each device is needed.

Content creators are also interested in understanding end-user display devices because they could provide device-specific content and require device-specific settings for that content.

Proposed Pipeline

Our proposed pipeline is based on the well-known concept of special *patterns* that enable estimation of each individual device characteristic. However, understanding *patterns* is not an easy-to-use approach, especially when the required result is a quantitative one. Consequently, we propose a system requiring from the user a series of device shots and locating automatically the relative position of the shot using special QR codes placed behind the screen, robustly estimating device characteristics at each spatial position in front of the device (normally, manufacturers declare the best value). The result of the measurement is continuous maps of each characteristic spatial distribution interpolated from points of shots. We briefly illustrate how the system works in the Figure 22. Some common problems for specific technologies are:

- Crosstalk
- Geometric distortions (stereoscopic two-projector systems)
- Time asynchrony (stereoscopic two-projector systems)
- Brightness decrease

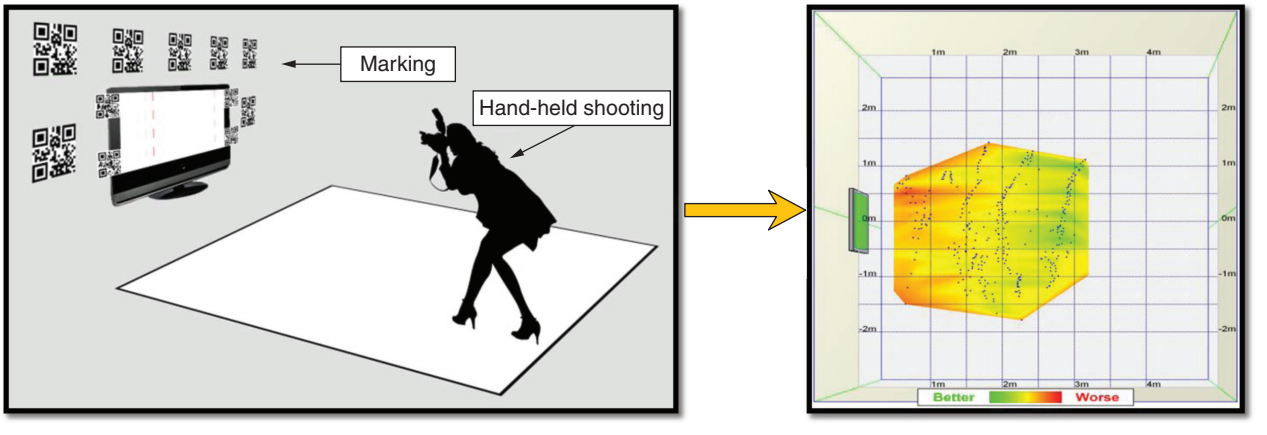


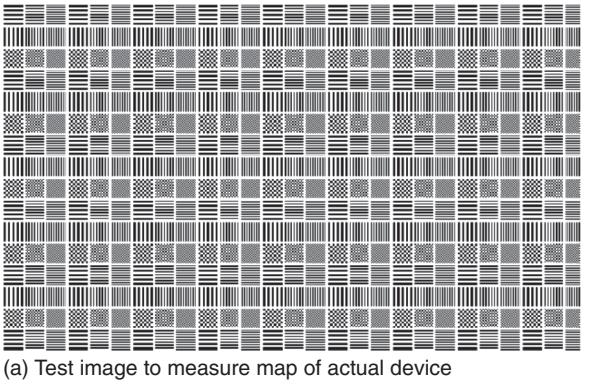
Figure 22: Brief illustration of our proposed pipeline. A user performs a series of shots from different positions in front of the device. The system determines each shot position relative to the device using QR codes placed behind the device and estimates a set of available characteristics. Finally, sparsely estimated characteristics are interpolated into a continuous map
(Source: Lomonosov Moscow State University, 2013)

Moreover, some integral characteristics are important to improve viewing quality:

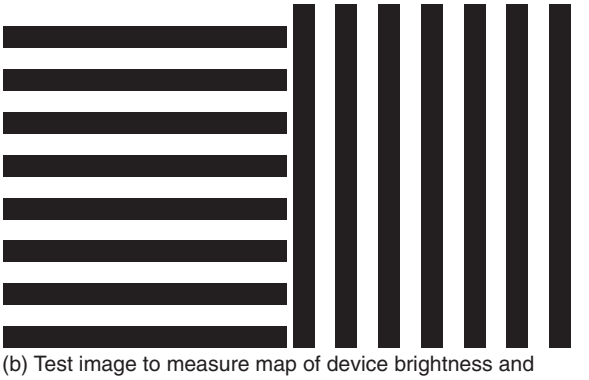
- Optimal observing distance
- Width of view zones (autostereoscopic displays)
- Actual resolution
- Actual number of views (autostereoscopic displays)

Estimated Characteristics

Brightness and crosstalk. Viewing quality of a 3D device can not be expressed by one single value. Actually, viewing quality significantly changes with respect to the observer's position. Mainly, it is determined by changes in brightness and amount of crosstalk (view mixes). To simplify the testing process, we use the pattern from Figure 23b, enabling us to measure brightness and crosstalk maps for each view simultaneously.



(a) Test image to measure map of actual device resolutions



(b) Test image to measure map of device brightness and crosstalk

Figure 23: Examples of test images to measure characteristics of a 3D viewing device (Source: Lomonosov Moscow State University, 2012)

Resolution. Most of the S3D viewing devices assume decreases in horizontal resolution up to the number-of-view-zone times. The exact number depends on the amount of crosstalk. Normally, the resolution reduction is spatially uniform, thus we measure only one number (not a map). We show the test image we used to estimate actual resolution in Figure 23a.

Conclusion

Although most of the wireless optimization techniques studied as part of the VAWN research program were focused on the delivery of 2D content, it's important to also understand how these same concepts could be applied to 3D eventually. As such, this research was intended to increase the fundamental understanding of the various factors that impact 3D content delivery, from content creation to compression to end user quality prediction. The studies mentioned in the previous sections lead us to several conclusions, which we would like to communicate to the industry community:

- 3D content creation requires more quality control than 2D creation does. The VQMT3D project revealed that most 3D films contain numerous impairments that can potentially cause eyestrain and headaches. Recent waining interest in 3D can be explained by it's unacceptable quality.

To avoid further decrease in interest, the industry needs to introduce strict quality standards to the 3D creation stage. Additionally, high quality 3D can be better compressed than video with a large amount of stereoscopic impairments.

- Autostereoscopic multiview devices are expected to gain popularity over displays that require users to wear glasses. Capturing content for the autostereoscopic displays with camera arrays is currently impractical. The content for such devices should be generated using depth maps, which can either be estimated from data captured by stereo cameras or captured by real-time depth sensors.
- 3D representations employing depth maps look promising due to their scalability (that is, support for various autostereoscopic displays) and low amount of additional data in comparison with 2D streams. Unfortunately, as autostereoscopic multiview devices gain popularity, due to its additional bitrate requirements the MVC standard will not suffice to deliver 3D videos to multiview displays over current modern wireless networks. Hence it is important to invest effort into developing 3D codecs specific to depth-based 3D video representations.
- It is important to study the quality of the entire end-to-end 3D system and not focus only on the rate distortion ratio of the codec. No one will watch a video with significant artifacts introduced at the creation stage despite the absence of distortions introduced during delivery and display stages. Any quality gain at the delivery stage can easily be negated at the display stage due to a low quality display. Therefore a quality standard for 3D displays should be created. Also an update to the conventional ITU recommended methodologies for subjective experiments is required. The software tools developed for measuring 2D video quality can no longer be used for 3D. Our open source software, Tally, is a good candidate solution to this problem both for industrial and scientific communities.

In this article, we presented several issues arising in the 3D-video life cycle (content creation, delivery, processing, and display) and proposed several concepts to mitigate these issues. A considerable amount of work still remains to be done. Results reported on depth-map–based compression schemes must be validated by conducting extensive subjective tests.

We would like to highlight the work that should be done to better understand how various factors influence 3D visual quality. In the content creation section, we have proposed several methods to detect artifacts that, according to medical experts' opinions, can potentially cause eyestrain. Measurement of the correlation between proposed metrics values and actual eyestrain is still an open problem. Actually, determining the value of the viewer's eyestrain is a challenge on its own, although some promising work exists in this field.[51] We hope that the research studies presented in this article will help to bring high-quality 3D video to every home while avoiding wireless network overload.

References

[1] Scharstein, D. and R. Szeliski, "High-accuracy stereo depth maps using structured light," in *Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2003, pp. I–195–I–202.

[2] Scharstein, D. and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, no. 1–3, pp. 7–42, 2002.

[3] Richardt, C., D. Orr, I. Davies, A. Criminisi, and N. A. Dodgson, "Real-time spatiotemporal stereo matching using the dual-cross-bilateral grid," in *European Conference on Computer Vision (ECCV)*, vol. 6313, 2010, pp. 510–523.

[4] Garcia, F., D. Aouada, B. Mirbach, T. Solignac, and B. Ottersten, "A new multi-lateral filter for real-time depth enhancement," in *Advanced Video and Signal-Based Surveillance*, 2011, pp. 42–47.

[5] Hirschmuller, H. and D. Scharstein, "Evaluation of cost functions for stereo matching," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, 2007, pp. 1–8.

[6] Kopf, J., M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," in *ACM Transactions on Graphics (TOG)*, vol. 26, no. 3, 2007.

[7] Garcia, F., B. Mirbach, B. Ottersten, F. Grandidier, and A. Cuesta, "Pixel weighted average strategy for depth sensor data fusion," in *International Conference on Image Processing (ICIP)*, 2010, pp. 2805–2808.

[8] "Video Quality Measuring Tool 3D Project," http://www.compression.ru/video/vqmt3d.

[9] Rushton, S. K. and P. M. Riddell, "Developing visual systems and exposure to virtual reality and stereo displays: Some concerns and speculations about the demands on accommodation and vergence," *Applied Ergonomics*, vol. 30, no. 1, pp. 69–78, 1999.

[10] Hoffman, D. M., A. R. Girshick, K. Akeley, and M. S. Banks, "Vergence-accommodation conflicts hinder visual performance and cause visual fatigue," *Journal of Vision*, vol. 8, pp. 1–30, 2008.

[11] Ukai, K. and P. A. Howarth, "Visual fatigue caused by viewing stereoscopic motion images: Background, theories, and observations," *Displays*, vol. 29, no. 2, pp. 106–116, 2008.

[12] Howarth, P. A., "Potential hazards of viewing 3-D stereoscopic television, cinema and computer games: A review," *Ophthalmic and Physiological Optics*, vol. 31, no. 2, pp. 111–122, 2011.

[13] Tam, W. J., F. Speranza, S. Yano, K. Shimono, and H. Ono, "Stereoscopic 3D-TV: Visual comfort," *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 335–346, 2011.

[14] Banks, M. S., J. C. A. Read, R. S. Allison, and J. W. Simmon, "Stereoscopy and the human visual system," *Motion Imaging Journal*, vol. 121, no. 4, pp. 24–43, 2012.

[15] Daum, R. M., "Clinical management of binocular vision: Heterophoric, accommodative and eye movement disorders," in *Optometry & Vision Science*, vol. 71, 1994, p. 414.

[16] Voronov, A., A. Borisov, and D. Vatolin, "System for automatic detection of distorted scenes in stereo video," in *International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, 2012, pp. 138–143.

[17] Ogle, K. N., "Induced size effect: I. a new phenomenon in binocular space perception associated with the relative sizes of the images of the two eyes," *Archives of Ophthalmology*, vol. 20, no. 4, pp. 604–623, 1938.

[18] Gåding, J., J. Porrill, J. E. W. Mayhew, and F. J. P., "Stereopsis, vertical disparity and relief transformations," *Vision Research*, vol. 35, pp. 703–722, 1994.

[19] Allison, R. S., B. J. Rogers, and M. F. Bradshaw, "Geometric and induced effects in binocular stereopsis and motion parallax," *Vision Research*, vol. 43, no. 17, pp. 1879–1893, 2003.

[20] Stevenson, S. B. and C. M. Schor, "Human stereo matching is not restricted to epipolar lines," *Vision Research*, vol. 37, no. 19, pp. 2717–2723, 1997.

[21] Voronov, A., D. Vatolin, D. Sumin, V. Napadovsky, and A. Borisov, "Methodology for stereoscopic motion-picture quality assessment," in *Stereoscopic Displays and Applications*, vol. 8648, 2013.

[22] Anderson, B. L., "A theory of illusory lightness and transparency in monocular and binocular images: The role of contour junctions," in *Perception*, vol. 26, 1997, pp. 419–454.

[23] Boydstun, A., J. Rogers, L. Tripp, and R. Patterson, "Stereoscopic depth perception survives significant interocular luminance differences," in *Journal of the Society for Information Display*, vol. 17, 2012, pp. 467–471.

[24] Patterson, R., A. Boydstun, J. Rogers, and L. Tripp, "Stereoscopic depth perception and interocular luminance differences," in *Digest of Technical Papers*, vol. 40, 2009, pp. 815–818.

[25] Voronov, A., D. Vatolin, D. Sumin, V. Napadovskiy, and A. Borisov, "Towards automatic stereo-video quality assessment and detection of color and sharpness mismatch," in *International Conference on 3D Imaging (IC3D)*, 2012, pp. 1–6.

[26] Stelmach, L., W. Tam, D. Meegan, A. Vincent, and P. Corriveau, "Human perception of mismatched stereoscopic 3D inputs," in *International Conference on Image Processing (ICIP)*, vol. 1, 2000, pp. 5–8.

[27] Seuntiens, P., L. Meesters, and W. Ijsselsteijn, "Perceived quality of compressed stereoscopic images: Effects of symmetric and asymmetric JPEG coding and camera separation," in *ACM Transactions on Applied Perception*, vol. 3, 2006, pp. 95–109.

[28] Kooi, F. and A. Toet, "Visual comfort of binocular and 3D displays," in *Displays*, vol. 25, 2004, pp. 99–108.

[29] Aleksei, B., B. Aleksandr, D. Vatolin, and M. Erofeev, "Automatic detection of artifacts in converted s3d video," in *Stereoscopic Displays and Applications*, 2014.

[30] Ivanov, B. T. and L. A. L., "Stereoscopic photography," *The Art Magazine*, 1959.

[31] Müller, K., P. Merkle, and T. Wiegand, "3-D video representation using depth maps," *Proceedings of IEEE*, vol. 99, no. 4, pp. 643–656, 2011.

[32] "Dolby gets support of the foundry, Cameron—Pace group for glasses-free 3D," http://goo.gl/BgyOjL.

[33] Choi, J., D. Min, and K. Sohn, "2d-plus-depth based resolution and frame-rate up-conversion technique for depth video," *IEEE Transactions on Consumer Electronics*, vol. 56, pp. 2489–2497, 2010.

[34] De Silva, D. V. S. X., W. A. C. Fernando, and S. L. P. Yasakethu, "Object based coding of the depth maps for 3d video coding," *IEEE Transactions on Consumer Electronics*, vol. 55, no. 3, pp. 1699–1706, 2009.

[35] "YUVsoft depth upscale," http://goo.gl/mrqEFI.

[36] Simonyan, K., S. Grishin, and D. Vatolin, "Confidence measure for block-based motion vector field," in *Graphicon*, 2008, pp. 110–113.

[37] Hewage, C. T. E. R. and M. G. Martini, "Reduced-reference quality evaluation for compressed depth maps associated with colour plus depth 3d video," in *ICIP,* IEEE, 2010, pp. 4017–4020.

[38] Kim, W.-S., A. Ortega, P. Lai, D. Tian, and C. Gomila, "Depth map distortion analysis for view rendering and depth coding," in *Image Processing (ICIP), 2009 16th IEEE International Conference on*, 2009, pp. 721–724.

[39] Lee, C., J.-I. Jung, and Y. .-S. Ho, "Inter-view depth pre-processing for 3d video coding," in *ISO/IEC JTC1/SC29/WG11, m22669*, 2011, pp. 1–7.

[40] "YUVsoft depth propagation," http://goo.gl/IDdcFe.

[41] Bal, C. and T. Q. Nguyen, "Depth-based prediction mode for 3D video coding," in *International Conference on Image Processing (ICIP)*, 2013.

[42] Bal, C. and T. Q. Nguyen, "Multiview video plus depth coding with depth-based prediction mode," *Circuits and Systems for Video Technology*, 2013.

[43] Fehn, C., "Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3D-TV," *Stereoscopic Displays and Virtual Reality Systems*, vol. 5291, pp. 93–104, 2004.

[44] Bjontegaard, G., "Calculation of average PSNR differences between RD-curves," *VCEG-M33*, Mar. 2001.

[45] Julesz, B., *Foundations of Cyclopean Perception* (Univ. Chicago Press, 1971).

[46] Jain, A. K., A. E. Robinson, and T. Q. Nguyen, "Comparing perceived quality and fatigue for two methods of mixed resolution stereoscopic coding," *Circuits and Systems for Video Technology*, 2013.

[47] Jain, A. K. and T. Q. Nguyen, "Video super-resolution for mixed resolution stereo," in *International Conference on Image Processing (ICIP)*, 2013.

[48] Jain, A. K., C. Bal, and T. Q. Nguyen, "Tally: A web-based subjective testing tool," in *International Workshop on Quality of Multimedia Experience (QoMEX)*, 2013, pp. 128–129.

[49] ITU-R, "Recommendation ITU-R BT.500-13: Methodology for the subjective assessment of the quality of television pictures," International Telecommunication Union, Tech. Rep., 2012.

[50] "The Lagom LCD monitor test pages," http://www.lagom.nl/lcd-test/.

[51] Chen, W., "Multidimensional characterization of quality of experience of stereoscopic 3d tv," PhD dissertation, Université de Nantes, 2012.

Author Biographies

Yury Gitman (ygitman@graphics.cs.msu.ru) graduated from the Computer Science department of Lomonosov Moscow State University in 2014. His main research interests are still image inpainting, visual attention modeling, edge aware filtering, and video matting. His recent work include semiautomatic visual attention models (http://compression.ru/video/savam) and the objective video matting benchmark (http://videomatting.com).

Can Bal (Cbal@ucsd.edu) is currently a PhD candidate in the Electrical and Computer Engineering Department at the University of California, San Diego. He received his BS and MS in Electrical and Electronics Engineering from Bilkent University, Turkey in 2007 and 2009 respectively. His research interests are in the field of 3D video compression and include depth-based 3D video coding, virtual view synthesis algorithms, and analysis of perceptual quality for 3D video.

Mikhail Erofeev (merofeev@graphics.cs.msu.ru) graduated from the Computer Science department of Lomonosov Moscow State University in 2013. Currently he is a PhD student in MSU's Graphics and Media Lab. His research interests are video and image matting, machine learning, 3D video generation and visual saliency modelling. Mikhail is one of the major contributors to the video matting methods benchmark (http://videomatting.com).

Ankit K. Jain (ankitkj@ucsd.edu) received a BS from Stanford University, Stanford, California, in 2005, and MS and PhD from the University of California, San Diego, in 2010 and 2014, respectively, all in Electrical Engineering. From 2005 to 2008, he was with the Embedded Digital Systems Group, MIT Lincoln Laboratory, Lexington, Massachusetts. He is currently a member of the technical staff at Pelican Imaging Corporation, Santa Clara, California. His research interests include image processing, 3-D video processing, computer vision, and human binocular vision.

Sergey Matyunin (smatyunin@graphics.cs.msu.ru) is a PhD student in the Graphics and Media Lab at Lomonosov Moscow State University. He received his specialist degree at MSU in 2011. The topics of his research are digital image processing, 3D video compression, and 2D-to-3D conversion.

Kyoung-Rok Lee (krl006@ucsd.edu) received a B.Eng. degree in Computer Engineering from the Kyungpook National University, Daegu, Korea, in 2008, an MS in Computer Science from the University of California, San Diego, in 2010, and a PhD in Electrical Engineering from the University of California, San Diego in 2014. His research interests are in video processing and computer vision including depth refinement, Simultaneous Localization And Mapping (SLAM), and 3D reconstruction.

Alexander Voronov (avoronov@graphics.cs.msu.ru) graduated from the Computer Science department of Lomonosov Moscow State University in 2011. As a member of the MSU Graphics and Media lab in 2008–2013, he participated in research on S3D-video generation and post-processing.

During 2012–2013 he was the head of the VQMT3D research project, which is dedicated to S3D video objective quality estimation and artifact analysis (http://compression.ru/video/vqmt3d/). Since 2013 Alexander has worked at Intel as a software engineer in the area of video compression.

Jason Juang (jajuang@ucsd.edu) is a PhD student in the Video Processing Lab in the Department of Electrical Computer Engineering, University of California, San Diego, led by Prof. Truong Nguyen. He received his MS at University of California, San Diego, and his BS at National Taiwan University. He is specializing in the field of computer vision and graphics, specifically in stereo vision. He did an internship at Qualcomm R&D and has worked previously at Tetravue.

Dmitriy Vatolin (dmitriy@graphics.cs.msu.ru) graduated from the Computer Science department of Lomonosov Moscow State University in 1996 and got his PhD in 1999. The theme of his PhD thesis was "Optimization methods of fractal image compression." Dmitriy has been teaching a computer graphics course at MSU since 1997. He wrote the book *Algorithms of Image Compression* in 1999 and coauthored *Methods of Data Compression* in 2003. Dr. Vatolin supervised collaborative research projects with Intel and Samsung. Three PhD students have graduated under his supervision. He created one of the largest websites devoted to data compression (http://compression.ru). He teaches courses on methods of 3D and 2D video and image processing and compression.

Truong Q. Nguyen (tqn001@eng.ucsd.edu) [F'05] is currently a professor in the ECE Department of the University of California, San Diego. His current research interests are 3D video processing and communications and their efficient implementation. He is the coauthor (with Prof. Gilbert Strang) of a popular textbook, *Wavelets & Filter Banks,* Wellesley-Cambridge Press, 1997. He has over 400 publications. Prof. Nguyen received the IEEE Transaction in Signal Processing Paper Award (1992). He received the NSF CAREER Award in 1995. He served as associate editor for *IEEE Transaction on Signal Processing*, *Signal Processing Letters*, *IEEE Transaction on Circuits & Systems*, and *IEEE Transaction on Image Processing*.

IMPROVING VIDEO PERFORMANCE WITH EDGE SERVERS IN THE FOG COMPUTING ARCHITECTURE

Contributors

Xiaoqing Zhu
Cisco Systems

Douglas S. Chan
Cisco Systems

Hao Hu
Cisco Systems

Mythili S. Prabhu
Cisco Systems

Elango Ganesan
Cisco Systems

Flavio Bonomi
IoXWorks Inc.

This article introduces a novel computing paradigm—fog computing—that extends the current practice of cloud computing to the network edges closer to the clients. The main advantages of fog computing include low latency, location awareness, support for real-time analytics, and support for mobility. We further showcase how fog computing may transform many video applications and services, ranging from intelligent caching for on-demand video delivery and proxy-assisted rate adaptation for live video streaming to enhanced interactivity in virtual desktop infrastructure (VDI) systems and real-time video analytics for surveillance cameras. We believe that the rich interactions between the fog and the cloud will shape the next wave of innovations in the information technology (IT) industry.

Introduction

The proliferation of smartphones and tablets, along with advances in broadband mobile networking technologies, has fueled the rapid growth of mobile media traffic. According to forecasts from Cisco Visual Networking Index, two thirds of the world's mobile data will be video traffic by 2015.[1] Such a trend imposes daunting technical challenges for existing network infrastructures and in the meantime opens up unprecedented opportunities for novel architectures and solutions.

Video delivery over the Internet today still abides by the conventional client-server model. That is, a client's request for video content is first routed out of the client's local network gateway and traverses the Internet to servers where the requested video data is hosted. Subsequently, packets of the requested video traverses the Internet back towards the local network where the client is located. In wireless networks, the access node is typically the bottleneck where congestion is likely to occur. In order to track the wireless link status, an end-to-end feedback mechanism is employed. However, end-to-end delay can be large in a wireless environment, hence sender-based rate adaptation can suffer from obsolete feedback information from receivers.

Industry has already made several advances that build upon this traditional paradigm to alleviate the problem caused by long end-to-end delays. For instance, Akamai Technologies, one of the world's largest Internet content delivery networks (CDNs), has deployed over 137,000 servers in 87 countries within over 1,150 networks.[2] This allows requests for popular web contents (including video clips such as those found on YouTube*) to be redirected to a local proxy and served from there with lower latency. In addition, recent standardization efforts on MPEG-DASH advocate a pull-based mode for video rate adaptation—video

rate selections are driven by clients riding over HTTP/TCP connections, so as to relieve the servers from the computational burden of sophisticated congestion control schemes and to leverage the prevalence of CDN deployments.

Nevertheless, the current video delivery model and aforementioned state-of-the-art techniques rely on some form of direct interaction between video server and client. In other words, there are presently no concepts of automatically and dynamically optimizing video delivery based on information about a client that is already known or can only be accurately measured at devices local to or near the client's network. Examples of such information include: the local network conditions (including the wireless channel) or traffic statistics that are monitored by a client's local network switch node or gateway, a client's feedback control signals that could become stale after traveling beyond a few hops from the local network, and cached contents in an edge server that can be used for re-composing the requested video without a new version being fetched from the remote data center. As such, current performance optimization techniques for video applications and services do not yet leverage advantageous properties of what we at Cisco Systems call the *fog computing architecture.*

"…there are presently no concepts of automatically and dynamically optimizing video delivery based on information about a client…"

As proposed by Bonomi[3] and Bonomi et al.[4], the emerging fog cmputing architecture is a highly virtualized platform that provides compute, storage, and networking services between end devices and traditional cloud computing data centers—typically, but not exclusively—located at the edge of network. Figure 1

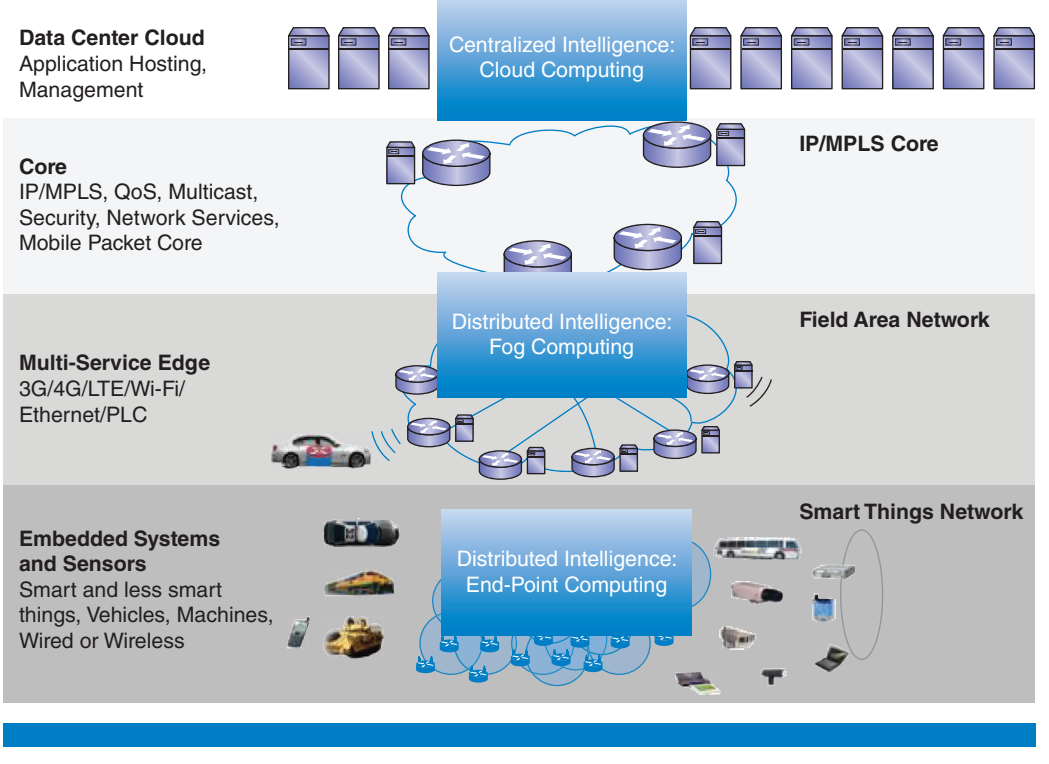


Figure 1: An overview of fog computing architecture (Source: Cisco Systems, 2014)

"…fog is a "distributed" form of cloud that stays close to the ground."

presents this idealized information and computing architecture and illustrates an implementation of fog computing. The namesake derives from the notion that fog is a "distributed" form of cloud that stays close to the ground. It needs to be mentioned that the notion of fog computing echoes similar concepts as advocated by others in the industry, such as "Intelligent Edge" architecture from Intel[5] and "Cloudlet" from Microsoft.[6]

In the next section, we introduce our vision of the fog computing paradigm, delineate its characteristics, and outline emerging platforms that support fog services. We then describe a series of video applications that can potentially benefit from a fog computing architecture, ranging from intelligent caching and adaptive streaming for video content delivery to interactive virtual desktop infrastructure (VDI) and real-time video analytics. Whenever applicable, we also point to research outcomes from the Video Aware Wireless Networks (VAWN) program that can potentially benefit from, or be integrated into, this new fog computing paradigm. Finally, we conclude the article by speculating on how fog computing may enable novel video applications and services.

The Fog Computing Architecture and Vision

In the cloud computing paradigm, web applications and information processing are centralized at data centers deployed in a limited number of locations at the core of the Internet. While this model has numerous technical and economic advantages, not all applications and services are suitable to be migrated to the cloud, nor can they be efficiently supported by the interplay of network endpoints and centralized data centers. For instance, latency-sensitive applications require nodes in the vicinity to meet their delay requirements, and they therefore demand a tight control of the locations of the compute and storage elements. Furthermore, an emerging wave of Internet deployments—most notably the Internet of Things (IoT)—requires mobility support and geo-distribution *in addition to* location awareness and low latency. These attributes call for a significant extension of the cloud infrastructure to the edge of the network—a paradigm we call fog computing[3][4], or, briefly, fog.

"We envision fog computing to be a highly distributed instantiation of cloud computing…"

We envision fog computing to be a highly distributed instantiation of cloud computing that provides computing, storage, and network services from the edge of the network. Figure 1 illustrates the concept of fog in a high-level end-to-end architecture view: fog's resources are located between endpoints and remote data centers (the cloud) and facilitate the operations of both. Our vision is that the fog will enable applications and services to exist and efficiently utilize resources across this entire stack. The goal is to add to the existing value of cloud computing with advantages such as reduced latency, improved power efficiencies, enhanced security, and significantly higher scalability for mobile and distributed services.

Rather than cannibalizing cloud computing, fog computing enables a new breed of applications and services that leverage the fruitful interplay between

the two computing paradigms. While the fog and the cloud share a common set of building blocks—computing, storage, and networking resources—the notion of "edge of the network" implies a number of characteristics that make the fog a significant extension of the cloud. To further contrast their differences, we list below several of fog's salient characteristics along with motivating application examples:

"…fog and the cloud share a common set of building blocks—computing, storage, and networking resources…"

- *Edge location, location awareness, and low latency*. Wireless access points or cellular mobile gateways are prime examples of a fog network node. The origins of the fog can be traced to early proposals to support endpoints with rich services at the edge of the network, including applications with low latency requirements (such as gaming, video streaming, and augmented reality).
- *Support for online analytic and real-time interactions*. The fog is positioned to play a significant role in the ingestion and processing of data close to the source. Important fog applications involve real-time interactions rather than batch processing.
- *Geographical distribution*. In sharp contrast to the more centralized cloud, the fog is well suited for applications and services that involve widely distributed deployments. The fog, for instance, will play an active role in supporting high quality video streaming to moving vehicles via proxies and access points positioned at the network edge and close to the users.
- *Support for mobility*. It is essential for many fog applications to communicate directly with mobile devices, therefore to support mobility techniques, such as LISP[7], that decouple host identity from location identity and provide a distributed directory system.

"It is essential for many fog applications to communicate directly with mobile devices…"

- *Scalability.* The fog plays an essential part in scaling up Internet services by several orders of magnitude. The smart grid is a salient application example of the inherently distributed systems that require distributed computing and storage resources for a very large number of nodes. A second example is large-scale sensor networks for monitoring the environment, due to their wide geographical distribution.
- *Heterogeneity*. Fog nodes come in different form factors, and are built upon heterogeneous resource platforms. They also tend to be deployed in a wide variety of environments.
- *Interoperability and federation*. Seamless support of certain services (video streaming is a good example) requires the cooperation of different providers. Hence, fog components must be able to interoperate, and services must be federated across domains.

In this article, we focus our discussions on how fog computing may transform video applications and services. Whenever applicable, we also point to research outcomes from the VAWN program that can potentially benefit from or fit into this new paradigm. We refer interested readers to Bonomi[4], J. Zhu et al.[8], and X. Zhu et al.[9] for discussions on other use cases of fog computing.

"…we explore how to reap the performance benefits from hosting video contents even closer to the network edge."

Novel Wireless Network Architectures for Video Caching Services

In this section, we identify and discuss additional values that a fog computing model can bring to video caching services. In particular, we explore how to reap the performance benefits from hosting video contents even closer to the network edge.

Whenever a user requests a piece of video content, the video data is delivered across the Internet, and, most likely, via content delivery networks (CDNs). However, it is inefficient to repeatedly field requests of the same video from different users belonging to the same local wireless access network. Such a practice wastes backhaul network bandwidth, incurs unnecessary latency for the users, and strains potentially scarce wireless resources. One way to remedy this is to design better caching algorithms at the network edge based on video content popularity (see Breslau et al.[10] and references therein).

As with web file requests, studies have found that online video requests also follow a Zipf distribution.[11] Furthermore, the popularity of online video contents changes relatively slowly, typically over several days or weeks. This implies that in a wireless environment, there is a high likelihood for users in the same vicinity to request the same video around the same time. In other words, since users of the same vicinity will most probably be associated to the same access point (AP) or base station, the AP is expected to field requests for the same video stream many times. Based on these findings, several projects within the VAWN program have propose novel caching designs—by extending and tailoring existing CDN frameworks to mobile architectures—so as to not only reduce video service delay and jitter but also to reduce Internet and wireless bandwidth consumption.[12][13][14][31][33][34]

Predictive Caching for Video Delivery to Roaming Mobile Clients

One characteristic of a fog computing service is the leveraging of information specific to and readily available at the wireless access local network. For example, a mobile device's wireless network has detailed information of the wireless channel between the device and its surrounding base stations. Such information can in turn be used to derive many attributes of the device, such as its potential roaming path, that can be utilized to improve user experience. We will describe one such service in the context of seamless video delivery to roaming mobile clients on Cisco's 802.11 wireless LANs.[15]

In the Cisco 802.11 wireless LAN infrastructure, one central agent, known as the Wireless LAN Controller (WLC), manages APs belonging to the same deployment. APs are configured to communicate with a WLC using a secure CAPWAP tunnel. (Note: CAPWAP, which stands for Control and Provisioning of Wireless Access Points, is the protocol specified by IETF RFC 5415.) While the WLC can co-locate on the same network with the APs, they can also be deployed without the presence of a WLC in the local network. Such APs are called Hybrid Remote Edge Access Point (H-REAP) devices. As illustrated in Figure 2, they communicate with their local network's router (using the CAPWAP

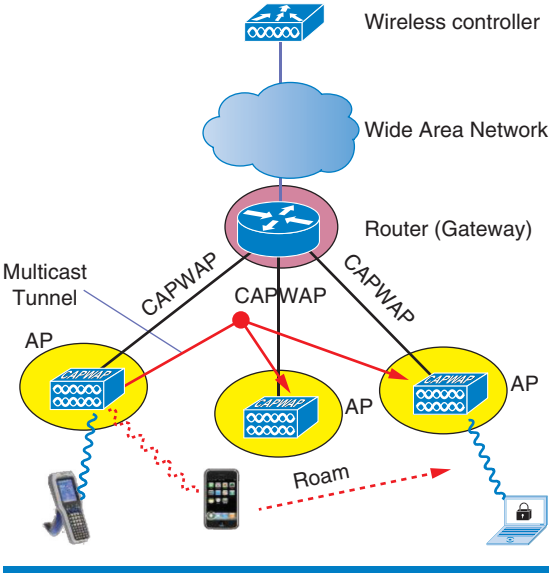


Figure 2: Dynamic multicast to predictively cache video content for roaming clients (Source: Cisco Systems, 2014)

protocol), which, in turn, communicates with the WLC across the WAN. In the context of fog computing, the APs represent the edge network nodes.

This H-REAP configuration saves the need for deploying a controller specific to a branch or remote office network, while still allowing client traffic to be tunneled securely to and from the controller at the central office. However, tunneling of such centrally switched traffic through the WAN incurs additional latency. Our internal measurement tests indicate that depending on status of the WAN link, delay of packet delivery to a client can range from a few milliseconds to 2 seconds.

"…delay it can be detrimental to the quality of experience (QoE) of time-sensitive traffic…"

Although this delay may be acceptable for certain use cases (such as web browsing and email), it can be detrimental to the quality of experience (QoE) of time-sensitive traffic (like video content delivery). This unfavorable situation is further exacerbated when mobile clients roam between APs. Consider when a branch-office wireless user is watching a centrally switched video and roams from one H-REAP AP to another. Once the authentication and association steps with the new AP are completed, this client will need to renegotiate with the video server via the central office to download the video content from where it was left off. These initial link setup steps can easily require up to a few seconds.[16] If the user's video playout buffer—which typically holds several seconds of video—is near empty, the user will experience an interruption in the video playback when roaming occurs. Clearly, such effects are undesirable.

As motivated previously, one way to mitigate this is to employ the AP's cache for improving the delay of roaming clients in H-REAP environments, described as follows.

Consider a wireless client that has started a video stream playback. Video traffic is first fetched via the Internet and delivered to the controller. Then the controller forwards the video to the AP over the WAN link, and the AP in turn transmits the video to the client. The client maintains a playout buffer to absorb delay jitters in video transmission. Typically, the buffer contains a few seconds of video for ensuring smooth playback at the receiver.

Now the wireless client begins to roam and moves away from its associated AP, a situation that can be detected from the decrease in the client's received signal strength indicator (RSSI) at the AP. Having detected this client's impending roaming activity, the following steps are executed:

1. The controller predicts the list of APs that the client will most likely roam towards. This can be implemented in many ways. For example, the controller can check to see which of these neighboring APs have received probe requests at RSSI from the alleged client. Similar neighbor information can be obtained via provisions in the 802.11k or Cisco Compatible Extensions (CCX) standards. One can even incorporate 802.11v's BSS Transition Management framework to direct roaming clients to specific APs.
2. To minimize service interruption, the roaming client's profile is pushed to this list of APs. The client profile includes the client's association information (which helps to speed up link setup with the new AP) and

video application flow contexts. The latter enables the destination APs to start accepting and buffering the video stream data for active flows on the client that is still associated with its original AP.

3. While continuing to send video data to the client through the currently associated AP, the controller now also begins to send video traffic of the client to the caches on the aforementioned candidate APs. Instead of sending individual flows of the client traffic to each of these APs through their respective CAPWAP unicast tunnel, the client's video traffic is broadcasted to all these APs using a CAPWAP multicast tunnel. When the APs receive the video data, they will only retain the latest *X* seconds of the video stream in their caches and discard older data, where *X* is tunable parameter. It suffices for them to store only the latest *X* seconds since the client is still receiving the video stream continuously from its current AP. A minimal value of *X* helps to minimize the storage of predictive roaming video data.
4. Now consider that the client has finally moved close enough to one of the predicted APs and begins associating with it. After link setup, this new AP can begin sending the cached video from the point where the previous AP has left off. The new AP can recognize the starting point to send to the clients with the assistance of a synchronized application context from the controller; for example, the final RTP sequence number that clients have received and acknowledged or the CAPWAP sequence number of the video packet can be forwarded to the new AP.
5. Once roaming is completed, the controller reverts back to the usual unicast tunnel and stops multicasting the video stream to all of the predicted APs.

"In summary, because the client no longer needs to make new video requests the user can enjoy seamless video playback despite roaming to different APs."

In summary, because the client no longer needs to make new video requests through the controller at the new AP—a process that can incur greater delay than what can be accommodated by the client's video buffer—the user can enjoy seamless video playback despite roaming to different APs.

One can further improve the efficiency of this system by observing that the caching data and control signaling can be sent directly between APs on the same network instead of going through the controller via the WAN connection. As Figure 2 illustrates, the multicast tunnel and control signals can be sourced and destined between a set of H-REAP APs. (The prediction algorithm can run on the client's currently associated AP by utilizing the standards mentioned in Step 1 above.) Therefore, neither data nor control path traffic would be subject to the WAN bottleneck.

Distributed Video Caching and Storage with Coding

In practice, a cache has finite storage. Thus, to efficiently manage the cache, a video object that has not been requested for some time must inevitability be deleted. However, once the video object enters the local network that contains the APs, transporting it between the APs has a significantly low cost. Consequently, our system can achieve better network and caching efficiencies by:

1. Prolonging the duration that a recently requested video object is accessible within the local network;

2. Allowing nodes belonging to the same local network to share their recently requested video objects;
3. Achieving the above conditions without transporting across the local network the video object in its entirety.

It turns out that these behaviors can be accomplished by designing a distributed caching system that applies the theory of network coding for distributed storage, which recently appeared in the literature (see Dimakis et al.[17][18] and the references therein). While various coding techniques for distributed storage have been employed in practice previously, the application of network coding to these systems is only just being considered at this time. And more importantly, network coding also introduces new coding constructions that can significantly reduce the local network bandwidth requirements, as compared with the current approach of using Reed-Solomon or other existing codes.

Intelligent Proxy for Live Video Streaming

The previous section explained how extra storage resources in fog nodes can be leveraged to improve delivery of on-demand video via intelligent caching at edge nodes. In this section, we describe how computational resources in fog nodes can be harnessed to enhance the performance of live video streaming. In live streaming, multiple clients are interested in watching the same content (such as a music concert or a sporting event) at the same time. In such a case, it is wasteful to stream multiple copies of the same content to users in the local network. Instead, only the highest-quality version of the video stream needs to be transferred to the fog node at the edge of the network. The fog node will, in turn, transcode the stream down to various quality versions that match the link rates and device capabilities (such as display size) of individual clients.

"We have built a proof-of-concept demonstration to showcase the advantages of fog-assisted media content adaptation for live streaming."

We have built a proof-of-concept demonstration to showcase the advantages of fog-assisted media content adaptation for live streaming. Figure 3 shows the system setup. The required computational modules—in the form of either real-time video transcoding processes or proxy servers—are hosted as Linux

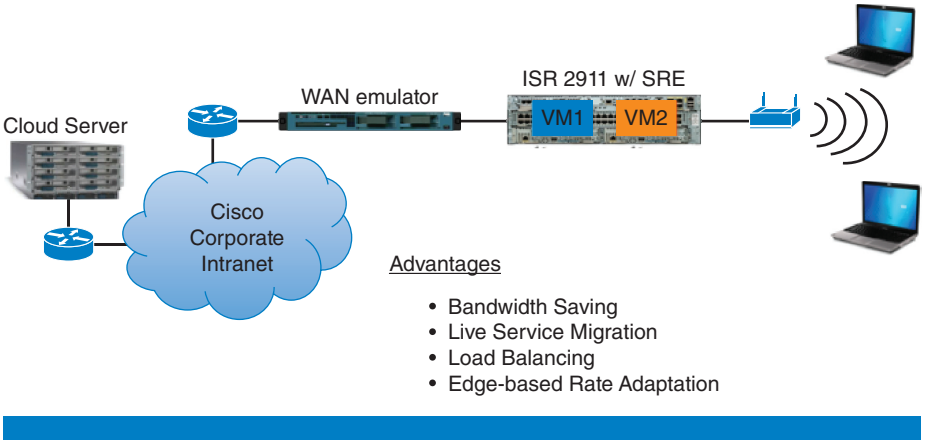


Figure 3: Proof-of-concept demo for fog-assisted adaptive video streaming (Source: Cisco Systems, 2014)

virtual machines (VMs) on the Service-Ready Engine (SRE) blade, as part of Cisco's Integrated Service Router (ISR) edge router.[19]

For our prototype, we host open-source VLC media players[20] on the VMs for transcoding and rate adaptation. The video source stream is hosted on a separate cloud server. The fog proxy retrieves the video stream from the cloud and forwards it to a local multicast address. Multiple instances of VLC are invoked to transcode this incoming stream. The clients subscribe to these ports on the ISR. To ensure we are able to repeat and analyze test results, we create a somewhat more controlled environment. We emulate the link from the cloud server to the fog proxy using a WAN emulator, with tunable link rate, delay, and packet loss rate.

The demo scenarios include the following:

- *Backhaul bandwidth saving.* Instead of serving multiple versions of the same video content for clients with heterogeneous devices and wireless channel conditions, the system can save backhaul bandwidth by streaming the highest-quality version of the stream to the edge node and employing its real-time transcoding capabilities to tailor the video bit rate and spatial resolution to individual clients. In addition to bandwidth saving, with this architecture we are able to alleviate or even completely eliminate the negative impact of bandwidth reduction in the backhaul on the user experience. Transient reduction in bandwidth between the edge and the original content server does not trigger the content server to reduce the streaming quality because we reduce the simultaneous requests for the same content drastically.
- *Dynamic load balancing and service migration.* As new clients become active in the system, each invokes a new real-time transcoding module. The system can dynamically allocate these new processes to a pool of available virtual machines (VMs) on the SRE blade, so as to balance out the computational load per VM. As the number of clients changes over time, the system is capable of migrating the transcoding processes from one VM to another without interrupting the video viewing experience at the client.
- *Low-latency rate adaptation*. In addition to hosting real-time video transcoding modules on edge nodes, the system can naturally shift the rate adaptation decision modules to the edge server. The rate adaptation module can react to abrupt bandwidth changes in clients in a more agile manner, bypassing the long round-trip delays seen by the remote video server.
- *Proxy-assisted weighted bandwidth sharing.* In the presence of multiple video streams, hosting rate adaptation proxy at the fog proxy can provide the additional benefit of weighted-fair bandwidth sharing among these clients. The system can be configured to accommodate a shared bottleneck in the backhaul link or within the wireless access region. The weight of individual clients can be update dynamically, depending on several factors: video content complexity, client's service priority, form factor of the device, wireless channel conditions, and so on. Our demo system supports weighted bandwidth sharing for both UDP/RTP-based and TCP/HTTP-based video streams.

"…we are able to alleviate or even completely eliminate the negative impact of bandwidth reduction in the backhaul…"

"The rate adaptation module can react to abrupt bandwidth changes in clients in a more agile manner…"

- *Proxy-assisted adaptive streaming for legacy and thin clients.* The fog proxy can act on behalf of a legacy or thin client, so as to interact with the video server in the cloud via sophisticated or proprietary rate-adaptation protocols, such as HTML5 and WebRTC. Provisioning an intelligent proxy/server at the edge has several significant benefits: seamless transcoding across different protocols, better optimization decisions; offloading computational complexity from the mobile client for longer battery life, and enabling new transport features without modifying either content server or client.

A similar architecture can also be applied to optimizing wireless video delivery for live streaming or video-on-demand (VoD) services. In the following, we showcase the benefit of a proxy-based solution for adapting the scalable video streams at the edge of a wireless network, right where congestion over the wireless links occurs. This allows the rate adaptation module to constantly monitor the bottleneck buffer level, which, in turn, reflects variations in the throughput and delay of wireless links for all receivers.

"…we showcase the benefit of a proxy-based solution…"

In this work, we adopt the latest H.264/SVC standard[21] for lightweight in-network rate adaptation. By combined usage of temporal scalability and amplitude scalability, a wide bit-rate range (with a factor of more than 10) is allowed. The resulting scalable video stream can be decoded at different frame rates (FR) and quantization step sizes (QS). We further leverage the parametric models from prior work[22] to explicitly account for the impact of FR and QS on rate and subjective video quality of the encoded scalable stream. These models enable our system to choose the best combination of FR and QS, along with the corresponding temporal and amplitude layers, given a rate constraint for each stream. The adaptation of both FR and QS supports video delivery over a wide range of rates that result from a wide range of channel conditions in wireless networks.

The goal of the video adaptation module at a proxy node is to maximize the overall viewing experience of all traversing streams. The problem can be broken down into two steps: i) to allocate the video rate for each stream based on their respective rate-quality relations, wireless link throughputs, and the common bottleneck buffer level; and ii) to extract video packets belonging to the appropriate temporal and amplitude layers from each scalable video stream based on the allocated rate. Given the optimal rate-quality tradeoff derived from the original rate and quality models, the first subproblem of multi-stream-rate allocation is solved by maximizing the weighted sum of user video qualities under a total network utilization constraint. We propose an iterative solution, whereby the per-stream rate is calculated based on periodic updates of bottleneck buffer level and relative link throughputs. The second subproblem can be solved offline, by preordering the video temporal and amplitude layers based on the parametric rate and quality models, so that each additional layer offers maximum quality improvement per rate increment.

"The goal of the video adaptation module at a proxy node is to maximize the overall viewing experience of all traversing streams."

Figure 4 shows the system diagram of the proposed rate adaptation proxy at the edge node. The link buffer monitor periodically checks the bottleneck queue length. It is also responsible for estimating the link throughput for each receiver. In our system, the packets' inter-departure time at the interface

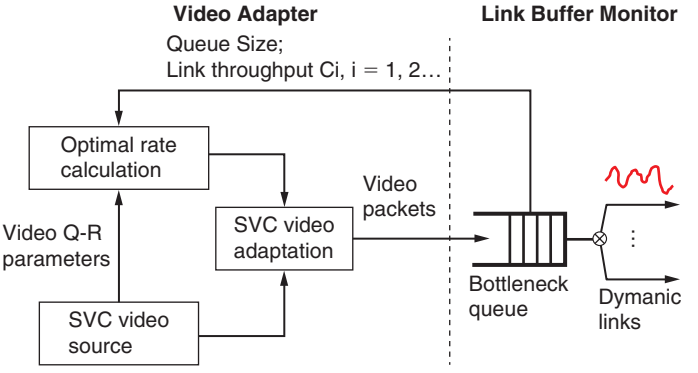
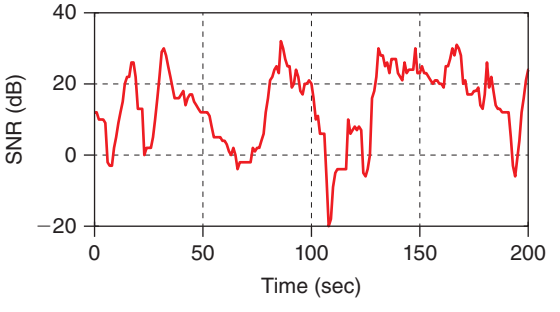
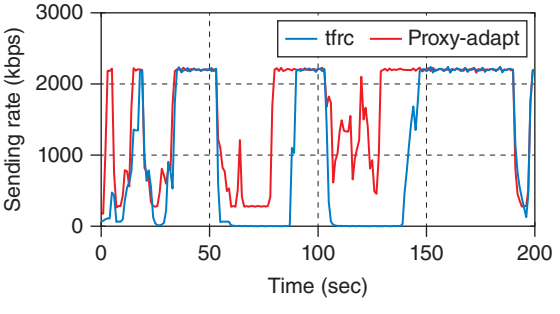


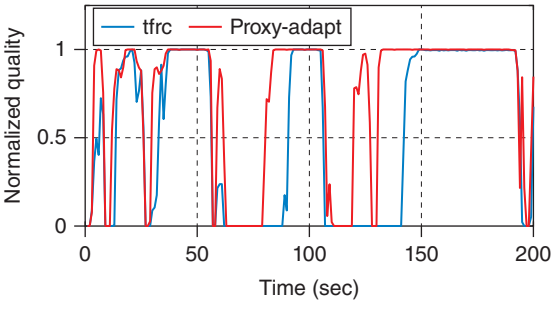
Figure 4: Main components in proxy-based video rate adaptation (Source: Cisco Systems, 2014)



(a) Wireless SNR traces from real-world measurement. The trace was collected while driving around Mountain View, California at an average speed of 20 mph.



(b) Comparison of sending rate



(c) Comparison of normalized video quality

Figure 5: Proxy-based adaptation vs. TFRC over a single dynamic link (Source: Cisco Systems, 2014)

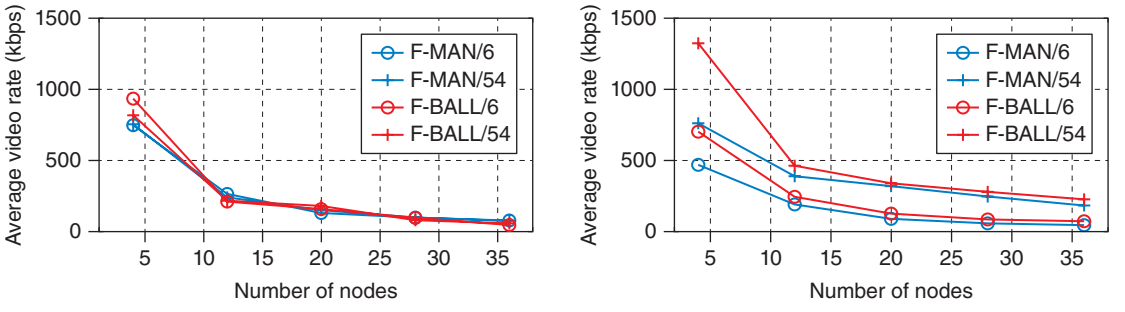
queue is inspected for deriving the instantaneous throughput of the link that transports the packet under consideration (via dividing the packet length by the inter-departure time for that packet). Then, the link throughput C_i for user i can be estimated by averaging over a number of packets.

The optimal rate allocation module will calculate the new stream rate based on the feedback from the link buffer monitor and the video rate-quality parameters embedded in the SVC stream. Then, the SVC stream is adapted to the new rate by simply sending video packets up to the target rate—assuming the stream is preordered in a quality-optimized manner. Ideally, the preordering procedure should be carried out at the video encoder so that the packets arriving at the proxy are already in optimal orders. In practice, it is also possible to order the SVC layers at the proxy, if the video server is agnostic of the rate-quality mode in use and does not know how to preorder the packets.

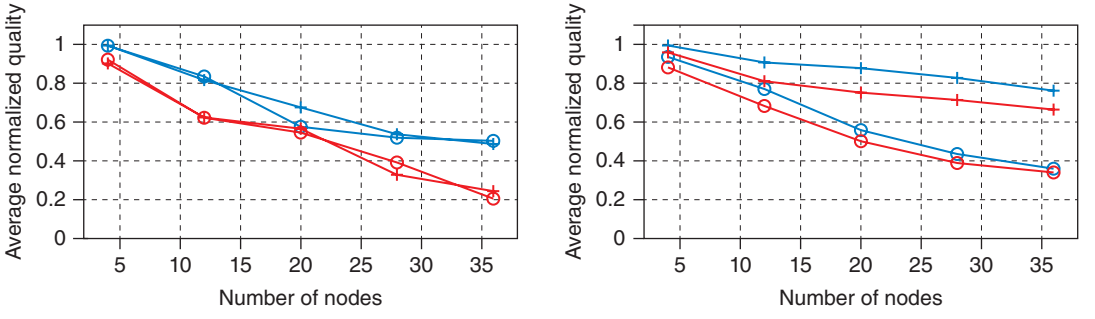
Some of our results based on extensive ns-2[23] simulations are highlighted below. We used two representative video sequences, FOREMAN and FOOTBALL, in CIF resolution (352x288 pixels) and a frame-rate of 30 frames-per-second (fps). They are encoded with JSVM version 9.12[24] into SVC streams with 5 CGS layers and 5 temporal layers. All video packets are preordered offline in a quality-optimized manner.

We first compare the proxy-based adaptation scheme against TFRC for a single video stream over a time-varying wireless link. As can be observed from Figure 5, when the channel condition is good and stable, for example, around time 50, 90, and 150, TFRC achieves good performance. However, it recovers slowly from poor channel conditions. In contrast, the proxy-based adaptation scheme can adjust the sending rate quickly, thereby significantly improving the video playback quality over TFRC. The average normalized qualities over time for proxy-based adaptation and TFRC are 0.66 and 0.47 respectively.

Next, we consider the scenario where *multiple video users share an AP.* The number of concurrent users ranges from 4 to 36; half of them are watching FOREMAN and the other half are watching FOOTBALL. In both groups



(a) Average video rate for each category of users: TFRC vs. proxy-based



(b) Average normalized quality for each category of users: TFRC vs. proxy-based

Figure 6: Normalized quality and video rate when multiple streams share a single AP (Source: Cisco Systems, 2014)

of receivers, half have a high PHY link rate of 54 Mbps and the other half have a low link PHY rate of 6 Mbps. Figure 6 shows the resulting video rate and normalized quality per user groups, averaged over 60 seconds after convergence. With TFRC, all users receive similar rates regardless of their video characteristics and link status. This leads to head-of-line blocking by the slow-link users, especially when the system load is high. As a result, the more complex video of FOOTBALL is delivered at a lower quality. In contrast, users receive different rates from the proxy-based scheme, depending on their respective video content characteristics and link qualities. This leads to more balanced video qualities across users, as well as higher aggregate video quality.

Interested readers can find a more detailed description of the scheme, together with more extensive evaluation of results in the article by H. Hu et al.[25] Within the VAWN program, the research theme of carrying out more intelligent and video-aware resource allocation at the edge of the network has also been explored extensively.[26][27][28][32]

Performance Enhancement for Interactive Applications

In addition to improving user experience for video-on-demand (VoD) and live streaming services, the presence of a fog proxy can also significantly benefit other interactive applications. This section illustrates this point using Virtual Desktop Infrastructure (VDI) as an application example.

One key challenge with VDI is to display the remote desktop on thin clients, especially when users are working with graphics-centric applications—for example, browsing web pages and PDFs, PowerPoint, Flash—the rendering of images and videos to the end user can be painfully sluggish to the point of hindering user productivity.

The culprit of this issue is the high latency and low bandwidth of WAN or mobile links that connect the server to the end users, and optimizing VDI performance over constrained network links is known to be a hard problem. In VDI, graphic components are rendered together with the desktop image at the VDI server, to be sent across the WAN to the user. When there are no rasterized graphics, the remote desktop components can typically be transported to the end user at a relatively low data rate (for example, via vector graphical commands). However, if graphics are involved, like when the user plays a YouTube video, the VDI server requests the video from its own location, combines the video stream with the remote desktop composite image, and sends it to the user; repeating each time a new video frame is played. The resulting data rate for continuously updating the remote display can easily overload the WAN link. Consequently, the video may end up being rendered in a very choppy manner.

To mitigate this problem, we propose a solution to separate the graphical and nongraphical elements of the remote desktop, as shown in Figure 7. Our proposed new architecture introduces a local server at the access edge near the client to offload some of the computing responsibilities from the cloud (that

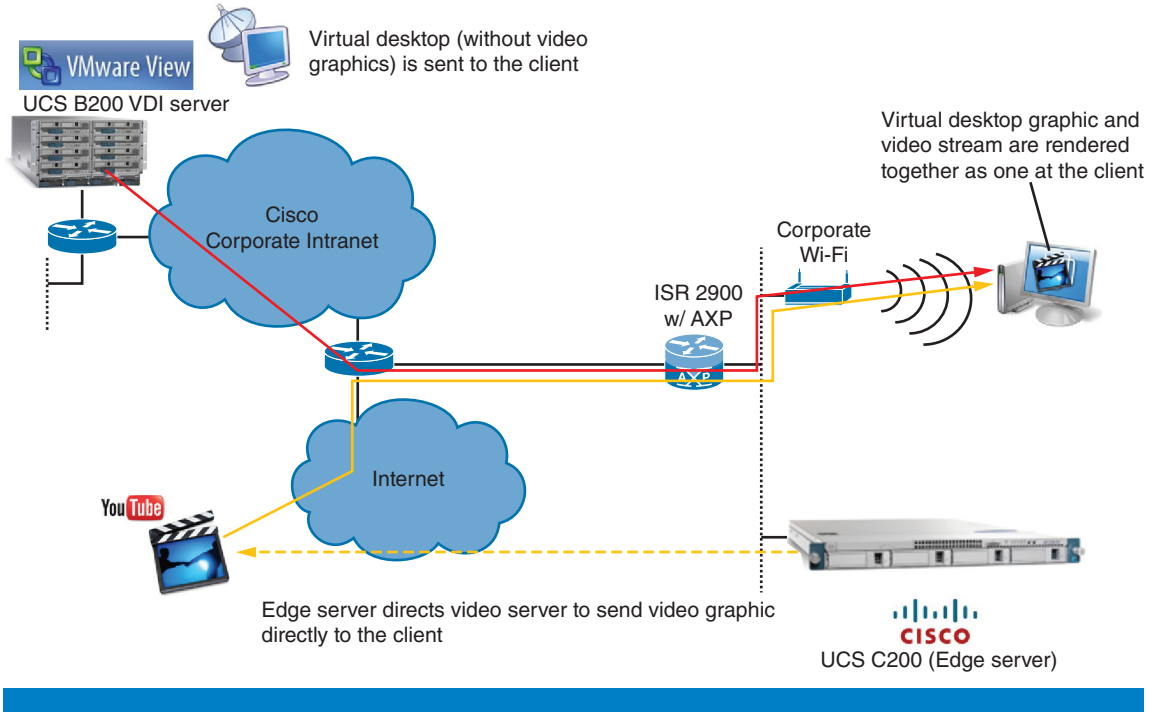


Figure 7: Architecture of our proposed VDI system (Source: Cisco Systems, 2014)

is, rendering at the remote VDI server across the WAN). This new architecture fits well into our fog computing paradigm, which is gradually being adopted by the industry.

In our proposed architecture, the VDI server in the cloud will only send nongraphical elements across the WAN, whereas the local fog server handles the graphical elements. For instance, the desktop image sent by the VDI server only draws out the window frame of the video stream, whereas the video stream itself is "filled" in by the local fog server. When the remote user watches YouTube, the requested video stream is fetched and rendered into video images by the local UCS C200 server. This is then sent to the end user, which also receives the desktop image without the video portion from the remote VDI server across the WAN. The video is ultimately combined with the desktop image for the user to view.

Note that in the conventional VDI setting, the graphics requested (like the YouTube video stream) traverses the WAN twice, first from the graphics' server to the VDI server and then from the VDI server to the end user. Our proposed fog-based architecture has successfully eradicated such inefficiency. It is possible to exploit the local server's knowledge of the end users' network conditions and further enhance our proposed solution by using this knowledge to adapt the graphical elements' data rates. We illustrate it in Figure 8 and an example workflow. More detailed information can be found in the U. S. patent application by Chan et al.[29]

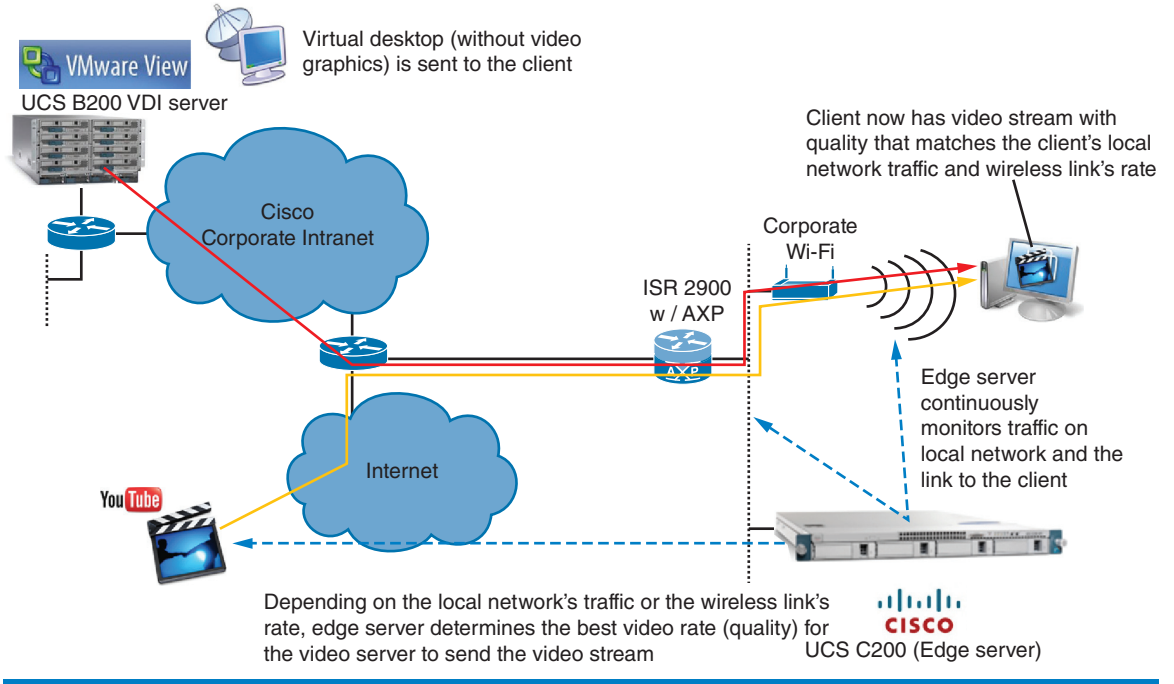


Figure 8: Process flow of our proposed VDI system (Source: Cisco Systems, 2014)

1. When the end user requests a video stream on the remote desktop, the edge server intercepts this action as the request is passed to the VDI server and extracts the link of the video stream.
2. The edge server communicates with the VDI server that the video stream will be handled by the local server.
3. Using the knowledge of the end user's link and the local network's traffic conditions, the edge server determines a data rate for the video stream suitable for this end user. Typically, the edge server should select the highest data rate deemed sustainable on the link such that the highest quality of the video stream will be delivered.
4. As in the new architecture, the edge server then requests the video stream from a video source and ensures it is delivered to the end user at the determined suitable data rate.
5. After the video stream has started, the edge server continues to monitor the network's traffic condition, in particular the link pertinent to the end user. If the conditions change, the edge server determines a new suitable data rate accordingly and manages the video stream such that the end user obtains it at the new rate.

The new IT trend of "Bring Your Own Device" (BYOD) in the enterprise is rapidly and dramatically changing our workspace. We envision that VDI will soon become a core component in future IT architectures that embrace BYOD by making the workspace more secure and more efficient. The proposed VDI system (as illustrated in Figure 8) is capable of adapting content locally to boost the quality and/or executing additional plug-ins to enhance security, and therefore it is well suited for enterprise IT deployment.

Real-Time Video Analytics in Physical Security Cameras

"Use of compute and storage resources just in time from fog nodes to perform video analytics fits well with the idea of taking computing closer to data."

Physical security cameras are widely deployed across many industries and businesses. As sensors in the Internet of Things, video cameras place a high demand on bandwidth and computing resources, and are capable of detecting and collecting the status of many other environmental parameters. Use of compute and storage resources just in time from fog nodes to perform video analytics on contents captured by physical security cameras fits well with the idea of taking computing closer to data.

Traditionally, physical security cameras were analog cameras connected to digital video recorders (DVRs) with propriety software and hardware. Recently, there has been a rise in the use of IP connected cameras and network video recorders (NVRs), which provide many benefits in terms of deployment, flexibility, and general operations.

Video streams from cameras in a branch location are rarely directly streamed to a central data center, since the bandwidth requirement for transporting video from multiple cameras is higher than what is normally available in terms

of branch connectivity. Hence NVRs are often located on premises whereby the cameras are connected via a high speed LAN. Another recent trend is to completely eliminate the NVR and record the video in the cameras themselves, utilizing onboard storage.

In both cases, there is a need to archive "important video" back to the central site. Typically, important video can be identified through the use of motion detection algorithms coupled with operation schedules. However, motion detection tends to suffer from false positives due to lighting and scene changes. Use of video analytics is emerging as a popular and attractive alternative for identifying important video for the purpose of alerting the operator, recording, and long term archiving. Such video analytics can be performed in real time on the live video or post facto on recorded video.

Another important usage model pertaining to video analytics is to convert video into usable data at the branch and transmit only the data, typically an order of magnitude smaller, to the cloud and/or headquarters. Once video is converted into usable data, operations such as searching and statistical trend extraction can be performed at the cloud/headquarters level across multiple cameras. As an example, Figure 9 shows the architecture for metadata management and APIs provided for upper-layer applications in Cisco's video surveillance manager.

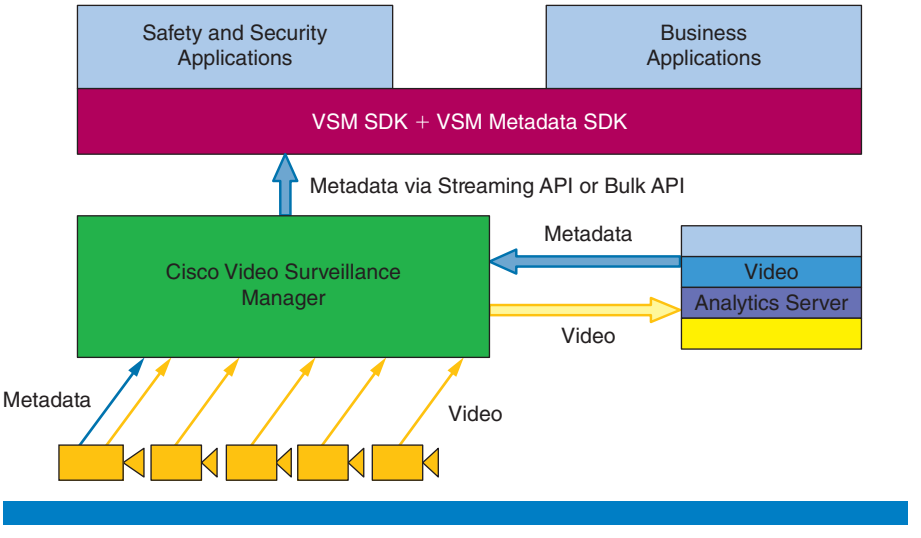


Figure 9: Metadata management and APIs in Cisco video surveillance manager (Source: Cisco Systems, 2014)

In summary, the use of fog computing achieves the following important objectives that are especially valuable to video as sensor data:

- Bandwidth scalability
- Compute scalability (memory and compute power)
- Storage scalability
- Relevance scalability (branch vs. HQ data relevance)
- Application scalability

These points are further illustrated by the following two use cases. Figure 10 provides a more comprehensive summary of potential use cases for utilizing video analytics in IoT.

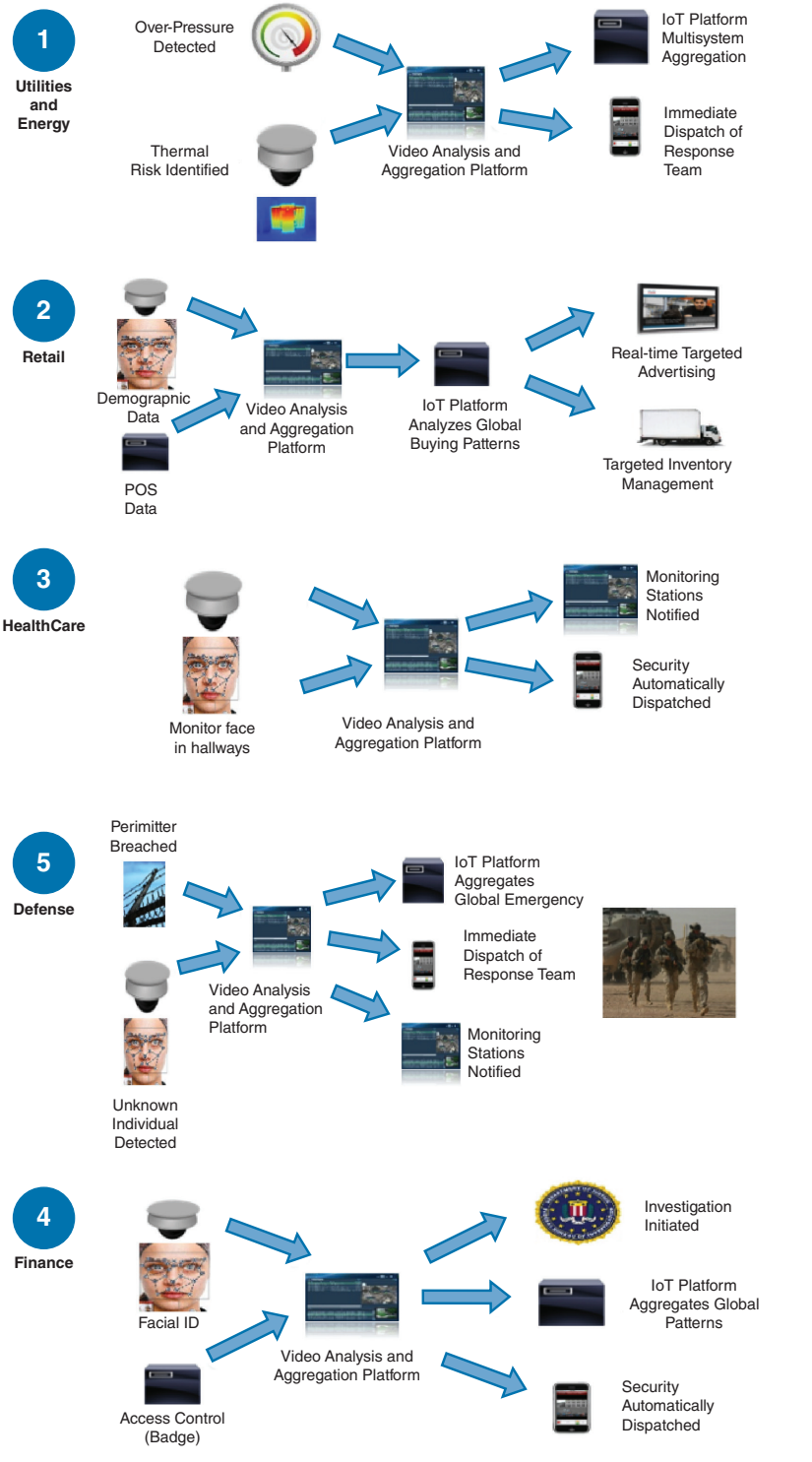


Figure 10: Example use cases for utilizing video analytics in IoT (Source: Cisco Systems, 2014)

Use Case A: Combining an Employee Badge with Soft Demographic Data in Facility Security

The primary means of providing physical access control is through the use of badges issued by the facility management. However, this provides weak security because anyone who gains access to a badge also gains physical access to the buildings and/or locations; the photograph on a badge is hardly ever cross-examined by security personnel. Introducing biometrics for strengthening badge security is expensive in terms of both capital expense (CAPEX) and operational expense (OPEX), therefore it has not seen wide adoption except for really critical areas such as vaults.

Currently, Cisco is working with a large financial firm to develop a novel alternative solution that correlates the following to enhance badge security in a nonintrusive manner:

- Badge credentials
- MAC addresses with location information from mobile devices (from the Cisco Mobile Services Engine)
- Soft biometrics (for example, gender, age, height, and race) from Cisco video analytics platforms

The system will build upon machine learning algorithms so that the correlations are automatically learned over an initial deployment period.

Use Case B: Urban Parking

Cisco partner Streetline[30] offers many creative parking solutions for parking lots and large urban areas. Currently, sensors on the ground are the primary means by which parking occupancy is detected. Embedding and connecting sensors in a large area, particularly when unplanned, is expensive. We are considering utilizing existing physical security cameras on the premises to detect parking occupancy in a cost-effective manner. To reduce the cost even further, the fog computing platform can be utilized to offer just-in-time video analytics based on schedules and proximity of consumers.

"…the fog computing platform can be utilized to offer just-in-time video analytics based on schedules and proximity of consumers."

We believe that converting video into data and providing a platform with simple-to-use SDKs will enable a large number of business applications to benefit from video analytics. Judicial use of the fog computing platform to provide just-in-time computation to reduce the size of video into usable data for aggregate data processing in a central/cloud location further saves costs and improves operational efficiency to video analytics applications.

Discussions and Outlook

As computational and storage resources become increasingly affordable and prevalent, the emerging fog computing architecture opens up new avenues for video applications. This article has surveyed a few application examples that may reap the benefits of fog: intelligent caching at the edge for video content delivery, proxy-based transcoding and rate adaptation for live video streaming,

"…intelligence at the edge of the network echoes strongly several of the research topics explored by the VAWN program…"

improving interactivity of virtual desktop infrastructure (VDI) by edge-based rendering, and improving operational efficiency by extracting video analytics from physical security cameras directly at branch locations. This recurring theme of deploying and exploiting intelligence at the edge of the network echoes strongly several of the research topics explored by the VAWN program, in particular, femtocaching[31], QoE-optimized resource allocation and rate adaptation in cellular networks[32], edge-caching and video-aware scheduling at the mobile video cloud[33], and intelligent scheduling for wireless video-on-demand multicasting.[34]

In addition to improving existing applications, fog computing also boasts the potential to enable new service and applications. For instance, the presence of fog nodes may make it finally practical to mine user-generated media content (for example, photos and videos captured by smartphones and tablets) in a location-aware manner. Instead of uploading all captured media data into a remote cloud for data mining and analytics extraction or carrying out feature extractions locally at each mobile client, the fog node at the edge of the network may perform feature extractions and correlation across local data first before invoking high-level global analytics from cloud servers. This will relieve the computational burden from mobile clients, reduce the communication burden on wireless backhaul networks, and improve response time in extracting any location-based analytics.

At Cisco, we are continuing to explore other potential use cases of fog computing, how it may benefit the broad spectrum of rich media applications and services, and how it may shape the future of the information technology (IT) industry.

Acknowledgements

We would like to thank Chuck Byers, Ivy H. Liu, Rodolfo Milito, Rong Pan, Francis T.M. Pang, Jiang Zhu, and many other colleagues at Cisco who collaborated with us on the various topics described in this article.

References, Industry Publications, and Presentations

[1] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2012-2017, http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.html.

[2] http://www.akamai.com/html/about/facts_figures.html.

[3] Bonomi, F., "Cloud and Fog Computing: Trade-Offs and Applications," in *Proc. International Symposium on Computing Architecture* (ISCA), EON Workshop, June 2011.

[4] Bonomi, F., R. Milito, J. Zhu, and S. Addepalli, "Fog Computing and Its Role in the Internet of Things," in *Proc. ACM SIGCOMM Mobile Cloud Computing* (MCC'12), August, 2012.

[5] Intel Corporation, "Intel Architecture at the Edge for Greater Flexibility and Scalability," Intel Solutions Brief, 2011.

[6] Mahadev, S., P. Bahl, R. Caceres, and N. Davies, "The Case for VM-Based Cloudlets in Mobile Computing," *IEEE Pervasive Computing*, vol.8, no.4, pp.14–23, Oct.–Dec. 2009.

[7] LISP (Locator/ID separation Protocol) open-source implementation: http://www.lispmob.org.

[8] Zhu, J., D. S. Chan, M. Prabhu, P. Natarajan, H. Hu, and F. Bonomi, "Improving Web Sites Performance Using Edge Servers in Fog Computing Architecture," in *Proc. IEEE International Symposium on Mobile Cloud, Computing and Service Engineering* (MobileCloud'13), March, 2013.

[9] Zhu, X., J. Zhu, R. Pan, M. S. Prabhu, and F. Bonomi, "Cloud-Assisted Streaming for Low-Latency Applications," in *Proc. IEEE International Conference on Computing, Networking and Communications* (ICNC'12). pp. 949–953, January 2012.

[10] Breslau, L., P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web Caching and Zipf-like Distributions: Evidence and Implications," in *Proc. INFOCOM*, pp. 126–134, March 1999.

[11] Cha, M., H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "Analyzing the Video Popularity Characteristics of Large-scale User Generated Content Systems," *IEEE/ACM Trans. Networking*, vol. 17, no. 5, pp. 1357–1370, October 2009.

[12] Ahlehagh, H. and S. Dey, "Video caching in Radio Access Network: Impact on delay and capacity," in *Proc. IEEE Wireless Communications and Networking Conference* (WCNC'12), pp. 2276–2281, 2012.

[13] Ahlehagh, H. and S. Dey, "Hierarchical video caching in wireless cloud: Approaches and algorithms," in *Proc. IEEE International Conference on Communications* (ICC'12), pp. 7082–7087, June 2012.

[14] Shanmugam, K., N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "FemtoCaching: Wireless Content Delivery Through Distributed Caching Helpers," *IEEE Trans. Information Theory*, vol. 59, no. 12, pp. 8402–8413, December 2013.

[15] Liu, I. H., D. S. Chan, and F. T. M. Pang, "Predictive Caching and Tunneling for Time-Sensitive Data Delivery to Roaming Client," US Patent Application No. 13/626,983, 2012.

[16] Emmelmann, M., "Fast Initial Link Set-Up PAR," IEEE 802. 11–10/1152r1, Sep. 2010.

[17] Dimakis, A. G., P. B. Godfrey, M. J. Wainwright, and K. Ramchandran, "Network Coding for Distributed Storage Systems," *IEEE Trans. Information Theory*, Aug. 2010.

[18] Dimakis, A. G., K. Ramchandran, Y. Wu, and C. Suh, "A Survey on Network Codes for Distributed Storage," *Proc. IEEE*. vol. 99, no. 3, pp. 476–489, Mar. 2011.

[19] Cisco Service-Ready Engine (SRE) Modules, http://www.cisco.com/c/en/us/products/interfaces-modules/services-ready-engine-sre-modules/index.html.

[20] VLC media player: http://www.videolan.org/vlc/index.html.

[21] ITU-T Recommendation H.264-ISO/IEC 14496-10(AVC), Advanced Video Coding for Generic Audiovisual Services, Amendment 3: Scalable Video Coding, ITU-T and ISO/IEC JTC 1, 2005.

[22] Wang, Y., Z. Ma, and Y.-F. Ou, "Modeling rate and perceptual quality of scalable video as functions of quantization and frame rate and its application in scalable video adaptation," in *Proc. Packet Video Workshop*, May 2009, pp. 1–9.

[23] NS-2 Network Simulator: http://www.isi.edu/nsnam/ns.

[24] JSVM SVC Reference Software, http://ip.hhi.de/imagecom_G1/savce/downloads/.

[25] Hu, H., X. Zhu, Y. Wang, R. Pan, J. Zhu, and F. Bonomi, "Proxy-Based Multi-Stream Scalable Video Adaptation over Wireless Networks Using Subjective Quality and Rate Models," *IEEE Trans. Multimedia,* vol. 15, no. 7, pp. 1638–1652, November 2013.

[26] Khalek, A. A., C. Caramanis, and R. Heath, "Video Quality-Maximizing Resource Allocation and Scheduling with Statistical Delay Guarantees," in *Proc. IEEE Global Communications Conference* (GLOBECOM'13), December 2013.

[27] Joseph, V. and G. de Veciana, "NOVA: QoE-driven Optimization of DASH-based Video Delivery in Networks," in *Proc. IEEE INFOCOM 2014*, April 2014.

[28] Chen, C., X. Zhu, G. de Veciana, A. C. Bovik, and R. W. Heath, "Rate Adaptation and Admission Control for Video Transmission with Subjective Quality Constraints," *IEEE Journal on Selected Areas of Signal Processing*, in print.

[29] Chan, D. S., J. Zhu, and H. Hu, "Rate-Adapted Delivery of Virtual Desktop Elements by Edge Servers in Fog Computing Environments," *US Patent Application No. 14/056,051,* 2013.

[30] Streetline Inc., http://www.streetline.com.

[31] Caire, G. and A. F. Molisch, "Femto-caching and D2D communications: a new paradigm for Video-Aware Wireless Networks," *Intel Technology Journal, Special Issue on Video-Aware Wireless Networks,* 2014.

[32] Heath, Jr., R. W., A. C. Bovik, G. de Veciana, C. Caramanis, J. Andrews, C. Chen, M. Saad, Z. Lu, A. A. Khalek, and S. Singh, "Perceptual Optimization of Large Scale Wireless Video Networks," *Intel Technology Journal, Special Issue on Video-Aware Wireless Networks,* 2014.

[33] Ahlehagh, H., L. Toni, D. Wang, P. Cosman, S. Dey, and L. Milstein, "Caching and Cross-Layer Design for Enhanced Video Performance," *Intel Technology Journal, Special Issue on Video-Aware Wireless Networks,* 2014.

[34] Avestimehr, S., T. Chen, S. Lashgari, A. Nwana, S. Rahman, S. Unal, and A. Wagner, "Video Delivery over Wireless Networks: Exploiting Network Heterogeneity and Content Commonality," *Intel Technology Journal, Special Issue on Video-Aware Wireless Networks,* 2014.

Author Biographies

Xiaoqing Zhu (xiaoqzhu@cisco.com) is a technical leader in the Chief Technology and Architecture Office (CTAO) at Cisco Systems Inc. She received the B.Eng. degree in Electronics Engineering from Tsinghua University, Beijing, China. She earned both an MS and a PhD in Electrical Engineering from Stanford University. Prior to joining Cisco, Dr. Zhu interned at IBM Almaden Research Center in 2003 and at Sharp Labs of America in 2006. Her research interests span multimedia applications, networking, and wireless communications. She received the best student paper award at ACM Multimedia 2007. She also won the best presentation award at IEEE Packet Video Workshop in 2013.

Douglas S. Chan (douglas.chan@ieee.org) is a Visiting Research Scholar in Electrical Engineering and Computer Science at the University of California, Berkeley. His work is affiliated with the Swarm Lab and Berkeley Wireless Research Center. Prior to that (2006–2014), Doug was a Technical Leader at Cisco Systems where he worked in R&D on wireless LAN products and their standardizations. Later at Cisco, Doug was a member of Cisco Advanced Architecture and Research (Enterprise Networking Labs) where he investigated

emerging paradigms like fog computing, data analytics, and the Internet of Things. Doug received his M.Eng. and PhD in Electrical Engineering from Cornell University. Doug is a Senior Member of IEEE and has received recognitions from the IEEE Standards Association for his contributions to the 802.11n and 802.11y wireless LAN standards.

Hao Hu (hoohawk@gmail.com) received his BS from Nankai University and MS from Tianjin University in 2005 and 2007 respectively, and a PhD from the Polytechnic Institute of New York University in January 2012. He has been with the Advanced Architecture and Research group and ENG Labs at Cisco Systems, San Jose, California. He interned at the Corporate Research, Thomson Inc., New Jersey in 2008 and Cisco Systems, California in 2011. His research interests include multimedia networking, distributed systems, and data analytics.

Mythili S. Prabhu (mythili.sprabhu@gmail.com) is a senior performance engineer at Akamai Technologies. She received her Bachelor of Engineering degree at Visvesvaraya Technological University. She earned her Master of Science in Electrical Engineering at the University of Southern California. Prior to joining Akamai, she worked in the Advanced Architecture and Research group at Cisco. Mythili's research interests include media streaming, transport layer optimizations, and content delivery.

Elango Ganesan (eganesan@cisco.com) is an expert in large-scale video processing and video management, with a strong background in data center and cloud orchestration technologies. He is currently focusing on enabling the Internet of Things at Cisco. Prior to joining Cisco in 2003, Elango developed a multi-service platform offering for SSL, IPSec, firewall, and load-balancing at Nexsi (acquired by Juniper Networks), and earlier spent six years at Intel where he was responsible for multiple core microprocessor innovations. He holds numerous patents across his areas of expertise, and has a PhD in Computer Engineering.

Flavio Bonomi (fgbonomi@gmail.com) is the founder and Chief Technology Officer at IoXWorks, Inc., which provides Internet of Things consulting and advisory services, including the incubation of new startups. Previously, Flavio was a Cisco Fellow, Vice President, and Head of the Advanced Architecture and Research Organization. He co-led (with JP Vasseur) Cisco's Internet of Things initiative during which he shaped a number of research and innovation efforts relating to mobility, security, communications, distributed computing, and data management. Before joining Cisco in 1999, Flavio was at AT&T Bell Labs from 1985 to 1995 where he worked on architectures and research, mostly related to the evolution of ATM. Later, Flavio was Principal Architect at two Silicon Valley startups, ZeitNet and Stratum One. Flavio received his MS and PhD in Electrical Engineering from Cornell University in 1981 and 1985, respectively.