QoE-based Transport Optimization for Video Delivery over Next Generation Cellular Networks

Noam Amram¶ Bo Fu† Gerald Kunzmann‡ Telemaco Melia * Daniele Munaretto*§ Sabine Randriamasy* Bessem Sayadi* Joerg Widmer‡ Michele Zorzi*§
¶ LiveU Ltd., Israel
†DOCOMO Euro-Labs, Germany
‡Alcatel-Lucent Bell Labs, Centre Villarceaux, France
*Department of Information Engineering, University of Padova, Italy
§Consortio Ferrara Ricerche, Ferrara, Italy
‡Institute IMDEA Networks, Spain

Abstract—Video streaming is considered one of the most important and challenging application for next generation cellular networks. Current infrastructures are not prepared to deal with the increasing amount of video traffic. The current Internet, and in particular the mobile Internet, was not designed with video requirements in mind and, as a consequence, its architecture is very inefficient for handling video traffic. Enhancements are needed to cater for improved Quality of Experience (QoE) and improved reliability in a mobile network. In this paper we design a novel dynamic transport architecture for next generation mobile networks adapted to video service requirements. Transport optimization of video delivery is achieved through a QoE oriented redesign of networking mechanisms as well as the integration of Content Delivery Networks (CDN) techniques.

I. INTRODUCTION

The MEDIEV AL project aims at defining and implementing network functionalities and associated protocols for an enhanced video delivery over mobile networks. Starting from the current standardization activities in 3GPP and IETF, the project evolves today’s cellular network architecture, integrating sophisticated and intelligent cross layers network functions able to recognize video characteristics, to classify IP traffic accordingly and to adopt operator defined policies for improved video traffic delivery. We design network-level mechanisms that optimize video traffic delivery, taking into account a number of static and dynamic considerations such as wireless access momentary parameters, users subscription profiles, applications requirements in order to enhance traffic delivery accordingly. The approach selected by the MEDIEV AL project is to carefully research the key technology components at each layer of the protocol stack and to enable cross layer optimization for enhanced performance. While cross layer mechanisms are not new by themselves, the key innovation is the ability of parameterize these functions and to trigger algorithms beyond standard operations. Specifically, the paper focuses on three main concepts:

1) efficient IP transport optimized for mobility and integration of content distribution functionality in the mobile network,

2) cross-layer design to adapt the wireless access parameters to the transport network and video traffic requirements,

3) and feedback to the application layer (to support server-side data rate adaptation, transcoding, etc.).

The first key concept focuses on IP transport and leverages the multi-interface characteristics of the mobile devices to redirect IP flows to the most suitable point of attachment. This module closely interacts with the mobility components developed within the MEDIEV AL project. (Those are, however, out of the scope of this paper.) One integral component of our design is the integration of CDN functionality directly within the cellular network architecture. We argue that this is a crucial design choice that will enable large scale content distribution at a reasonable cost and without overloading the operator’s core network.

The second key concept concerns the wireless access, WIFI and LTE, as depicted in Fig. 1. It provides lightweight mechanisms for prioritizing video traffic in the wired part of the network, to enable the wireless access to selectively drop low priority packets belonging to a media stream. This
selective packet dropping is a last resort for traffic adaptation with minimum impact on the user-perceived QoE. While the specific dropping mechanisms are not discussed in the paper, we explain how the framework operates to take optimal decisions.

The final key concept deals with triggering the application layer and content adaptation. The key innovation consists in optimizing network performance as much as possible while minimizing the impact on legacy applications. To this end, the paper describes the required network functions to enable cross layer optimization and provide enhanced QoE. In case the network is not able to grant the requested video traffic parameters, it informs the application layer to take support application-specific countermeasures.

Following an operator-driven approach, we design a transport optimization framework that is in charge of handling resources in the operators network. This framework receives the resource requirements from the video applications (service providers) and performs the following actions to maximize the QoE experienced by video users in the wireless link:

1) resource allocation and negotiation from the core network to the wireless access,
2) optimal handover decisions by interacting with the Mobility entities.

The remainder of the paper is organized as follows. Prior work is reported in Section II, the transport architecture is presented in Section III and we conclude the paper in Section IV.

II. RELATED WORK

Traditional data delivery over the Internet is based on end-to-end approaches between the server and the client. This approach has proven successful in face the wide variety of link conditions prevalent in the Internet, including wireless links (albeit with suboptimal performance): whenever congestion occurs, the congestion detection mechanism enables the source to take appropriate measures to avoid buffer underruns. In the context of video, we have two existing implementations of such an end-to-end approach. One recent approach is to use HTTP to distribute video chunks and use TCP for controlling intermediate congestion [1]. Another older approach is to use an RTP container [2], over UDP where congestion detection is done by exploiting RTCP messages [3] that gather statistics such as transmitted and lost packet counts, jitter, and round-trip delay time. The server-side application may use this information to adapt the video encoding. Among other issues, this second model suffers from the RTCP limited feedback frequency, hence limiting the reaction behavior.

Although both models are implemented in current mobile networks, they represent rather limited options to deal with the massive consumption of delay-sensitive, bandwidth intense, and loss-tolerant multimedia applications and the desire for better subjective video quality.

To go further with the improvement of the QoE of a video session, a mobile network-assisted video transport optimization is investigated in the MEDIEVAL project [4] based on two key concepts: the cross layer optimization and the mobile content delivery network.

Cross-layer optimization has been extensively investigated in recent years in order to maximize a predefined utility related to the video content given the time-varying and error-prone network path. Existing cross-layer solutions often span physical and MAC and/or application layers. A typical example of an application layer centric approach is [5] where the heterogeneous characteristics of the video stream are exploited to optimize the packet scheduling under a rate-distortion framework (RaDIO). Nevertheless, RaDIO ignores the capability of the lower layers. In contrast, in [6] video packets are scheduled according to the their various distortion impacts and by considering the lower layer adaptation capabilities. In summary, the transport layer is considered less important for error protection and bandwidth adaptation. Very little work is targeting this direction. In [7] the delay jitter of video transmitted over TCP has been found to be a suitable indicator for triggering IP handovers of mobile terminals. Nevertheless, the transport layer is often viewed only as a pipe and its capabilities to protect the stream and to adapt to bandwidth fluctuation is neglected. In MEDIEVAL, a cross layer optimization strategy will be investigated aiming to exploit all the dependencies between all the protocol layers of the system to improve the performance gain in terms of QoE. Such cross layer algorithms can be formalized as an optimization problem under constraints that can be solved either by a deterministic approach (e.g., Lagrange) or using a stochastic approach that may be more suited to handle this complex system and allow to continuously perform the optimization to adapt to changes in the system (e.g. dynamic programming).

In addition to the cross layer optimization component, the MEDIEVAL project will improve the video transport by re-casting CDN and Peer-to-Peer (P2P) approaches in the framework of an operator-controlled mobile network. In fact, state-of-the-art P2P streaming solutions rarely take a system-wide perspective. Very often a limited set of layers, excluding the transport optimization, is considered. An exception in this sense is represented for example by [8]. The use of network distance for peer selection and the evaluation of content availability at different sources are specific transport layer challenges that MEDIEVAL will explore. CDNs are well investigated and understood in the context of the Internet. However, there is a lack of knowledge about their operation in the context of mobile networks, where mobility support is an important requirement. State-of-the-art content distribution networks usually regard the transport layer (i.e., TCP) as invariant. Moreover, they neglect often the interplay between the transport layer and overlay organization. ALTO-like approaches [9] reduce inter-Autonomous System (AS) traffic by promoting locality but neglect properties of the overlay links other than AS hop distance. Such properties, e.g., delay or loss rate, in connection with the underlying transport protocol, have significant impact on the video quality.

III. ARCHITECTURE

In this section we describe our transport architecture, as shown in Fig. 2. The transport optimization functions provide
optimized video traffic in the mobile operator’s core network through intelligent caching and cross-layer interactions. The main objective is two-fold: (i) reduce the load on the operator’s backbone while (ii) providing a satisfactory QoE to the users.

A. Mobile Content Delivery Network

The traffic increase due to applications involving video is becoming an issue in mobile networks. To reduce the load in the operators mobile core network, MEDIEVAL has opted to implement a “Mobile CDN” (MCDN), where MCDN nodes may cache popular video files.

MCDN deployments need to consider optimization of the number and location of caches as to ensure that the desired performance improvement is achieved at a sufficiently low cost. The particularity of a MCDN compared to a regular CDN is the mobility of terminals requesting content. This creates larger dynamics than in a regular CDN and may require caching content more widely or even relocating content among CDN nodes when a sufficiently large number of nodes move.

MEDIEVAL focuses on the following features of MCDN:

- Content caching management: a Least Frequently Used (LFU) cache algorithm discards the items that are used the least frequently.
- Proactive push of content to caches can be used to maintain a low access delay. This requires that some estimate of future demand is available.
- P2P-inspired content exchange among caches allows to further reduce server load.
- Selection of optimal locations for CDN nodes. Thereby, the physical placement of CDN nodes is determined in advance, based on the physical and functional structure of the mobile operator’s core network.
- Network-aware selection of optimal sources from which to download content (CDN node or external source), under consideration of network layer information and the mobility of users.

The investigated features will be evaluated by answering the following questions:

- How much traffic in the mobile core network can be saved by caching a certain fraction of video files?
- By how much can download times be reduced?
- By how much can mobility induced connection interruptions or buffer underruns be reduced when streaming from a CDN node to a mobile terminal?
- How many more download requests are serviced?

1) CDN architecture: The CDN module provides a mobile CDN solution for video delivery including network based caching. This includes maintaining an efficient and stable overlay topology for the control and management of the CDN nodes, performing load-balancing among the CDN nodes, as well as relaying connections for mobility and caching. We aim at finding optimal locations for deploying the CDN nodes considering costs and the limitations of the core network, optimal distribution of content among the CDN nodes, and optimal node selection, i.e. choosing from the set of CDN nodes that have the desired content the node or subset of nodes that minimize streaming costs while maintaining some form of load balance among the CDN nodes.

Mobile core networks are usually hierarchical, with a central core part as well as branches and leaves in the different regions of a deployment area, for example a country. The specifications of the mobile systems used (e.g. 3G) limit the options for placing CDN nodes in the functional and physical network structure. Thus, an optimal placement of the CDN nodes needs to be determined in advance and would be very difficult to change dynamically. There are many possible options when considering cache placement. Setting up only one CDN node in the core network (close to the ingress point, where the operator’s network is connected to external networks) to cache videos will save bandwidth on the ingress link only. The other extreme is to maintain caches close to the base stations in all the leaves of the network topology. Although such a deployment achieves the highest reduction in load, the deployment cost would be immense and the network management complex. In addition, this approach might suffer from the problem that users within a small region may not access the same video files given the immense number of available videos. For example, a realistic estimate of the number of distinct video files in YouTube is around 143 million videos (as of August 2010) [10]. Therefore, we suggest maintaining a few CDN nodes in major regions.

The size of the storage capacity of the CDN nodes is determined by the tradeoff between storage costs and the popularity of video files. The distribution of the popularity of video files is Zipf-like at the waist with a truncated tail [11], i.e. there are a few extremely popular videos, while most of the videos are accessed only occasionally. As a rule of thumb (the so called Pareto principle) storing around 20% of videos will reduce the data traffic in the network by around 80%. However, this would imply a storage capacity of 272 TB per CDN node for caching YouTube videos only (assuming an average video size of 10 MB [12]). Such cache sizes are not practical and more intelligent caching strategies that work well
with much smaller caches sizes need to be considered.

These caching strategies are implemented in a so-called cache decision module. The decision module requires a continuous monitoring of the current conditions of the system, in particular, the popularity of videos (application monitoring) and the status of the CDN nodes (CDN node control). Based on the popularity and request rate of the videos in the respective region, the decision module determines which data shall be stored at a specific CDN node. Based on the collected data, the decision module also selects optimal sources from the network layers perspective for transmitting the video to the user. In this context, we aim at extending the IETF ALTO protocols [9] to the specifics of mobile core networks and the mobility of users. A first step is to implement a more sophisticated selection algorithm for content location that combines metrics like storage capacity, statistical availability, bandwidth, and latency in a robust way.

The project also considers connection offloading to local wireless networks such as wireless LAN or femto-cells managed by the same operator. In some cases the best strategy is even to jointly stream content via heterogeneous radio networks where some part is served via a cellular macro cell, while other parts are served through a local wireless network. Designing a CDN that spans these very different parts of a mobile operators network (with different backhaul networks and different connections to the mobile operators core network) has the potential for substantial performance improvements but at the same time is highly complex. CDN design in a network composed of femto-cells and/or WLAN routers is much more amenable to distributed designs inspired by peer-to-peer networks [13], rather than the centralized and tightly managed CDNs that are typical today. Resource availability in such local networks is much harder to control and distributed techniques that improve robustness and availability, such as network coding [14], are likely to prove to be important components of the design.

2) Content distribution aware mobility management: Mobile operators network deployments normally consist of a few centralized anchor points in the core network that concentrate the traffic of all users. The MEDIEVAL architecture designs a novel dynamic distributed mobility management architecture for mobile networks (Mobility Functions), where anchors are deployed close to the network edge and are dynamically available to users when needed for mobility. This mobility management is also supported by the MCDN. In the preparation phase of handovers, the Mobility Functions will send a list of feasible handover candidates (e.g. mobile base stations, WLAN routers) to the MCDN decision module. This request is similar to sending a compilation of multiple ALTO requests to the network for the different points of access using different anchor points in the network. The decision module will respond to this request with a rating of the candidates based on its sophisticated selection algorithm for CDN nodes. For example, a handover candidate will receive a lower rating if the corresponding CDN node in its region does not provide a cached copy of the video currently being streamed to the terminal, whereas another candidate may be rated high in case the local CDN node can stream the video file with short delay. These ratings will then be considered by the Mobility Functions (besides other criteria like signal strength) in order to choose an optimal handover candidate. After the successful handover, the final handover decision is reported back to the decision module which can use this information to update the application monitoring and trigger the distribution of content inside the CDN in an optimal way. For example, this may result in a request to a CDN node to start caching a video whose popularity has reached a certain threshold in the corresponding region.

If congestion in the network (e.g., at a CDN node) is detected, the decision module can also interact with the Mobility Functions to trigger a new evaluation of the mobility anchors of the mobile nodes connected to the overloaded network element. Similar to the communication described above for the handover preparation phase, the Mobility Functions will then send a list of handover candidates (if several candidates exist) for the affected mobile nodes and receive an updated rating of these candidates. The current candidate which was the reason for triggering the new evaluation, will receive a low rating, whereas other handover candidates might have more spare resources and therefore will receive a better rating. It is then up to the Mobility Functions to perform a handover or keep the current point of access.

B. X-Layer Optimizer and Traffic Engineering

Nowadays, QoE is gaining momentum compared to the usual Quality of Service (QoS) as a quality metric. QoE is also called Quality of User Experience and considers the overall experience of the customer when accessing and using a service. In contrast to QoS, QoE is by nature a quite subjective measure, since it represents the individual experience when viewing video contents. The service providers, in order to make their video services commercially successful, should measure the QoE and match it with the level of satisfaction of their customers. For this reason, the optimized resource allocation and traffic engineering techniques in the MEDIEVAL project will be QoE-centric in order to maintain an acceptable perceived quality.

In Fig. 2, we depict two modules in charge of optimizing and adapting the video flows over the network by means of cross-layer operations and traffic engineering actions.

1) Cross-Layer Optimizer: The optimal strategy giving the best video quality is designed by cooperating with the application, the transport, the MAC and physical layers. QoE-based video sensitivity information from the Content Provider (CP) is gathered to deal with the dynamic characteristics of a video source. Thereby, the video sensitivity represents the relationship between the target perceived video quality (for instance, in Fig. 3, measured as VSSIM [15], [16] linearly mapped to mean opinion score (MOS)) and a set of monitored parameters, such as application layer data rate (x-axis in Fig. 3), packet loss, motion activity (dynamic/static videos) or content (simple/complex scenes). The dynamic characteristics of the network, due to congestion either in the core or in the radio access networks, will be assessed by monitoring the users’ channel quality and the available capacity in the streaming path.
Finding the optimal cross-layer strategy will be modeled as a constrained optimization problem where the criterion to maximize is the QoE of the video flow, under constraints related to users’ channel quality and the available bandwidth, the handover candidates obtained from the mobility functions, the fill level of the MAC buffer, and the resilience to losses (Forward Error Correction (FEC), Automatic Repeat reQuest (ARQ)). The goal is to determine the transmission rate, matching the target QoE level requested for a user. The Optimizer is triggered by the mobile network as soon as a severe QoE degradation is observed. To improve it, the Optimizer performs the optimization aforementioned on all video flows being delivered through the congested area, to maintain a level of QoE fairness among the streams involved. Thus, a re-allocation of the channel resources between users is performed such that the average perceived quality (via average MOS for instance) is again maximized. This does not necessarily coincide with maximizing the total throughput, as intuitively observed also from Fig. 3.

The Optimizer takes advantage of the diversity provided by the sensitivity of each video stream, as shown in Fig. 3. Changes in the data rate impact the QoE level in a measure which varies from type of content to another. In the figure, we show the QoE sensitivity of five different videos to changes of date rate when performing open-loop transcoding. The reference video sequences [17] are encoded at a data rate ranging from 400kbps to 480kbps, with QCIF resolution (176 x 144) and at 30 frames per second. The dynamic video Football is most sensitive to variations of data rate while the static video Akiyo is the least sensitive. This feature allows the Optimizer to reduce resources for least sensitive videos and to compensate resources for most sensitive videos when overall network resources are scarce. Besides, different traffic engineering techniques result in a different impact on QoE. Fig. 4 compares different traffic engineering techniques in terms of QoE variation as a function of the available data rate. H.264 transcoding (TransH264), a simple open-loop transcoding, has better performance but is computationally expensive, whereas packet dropping (PckDrop) has overall the worst performance but is convenient for short term adaptations due to the very low complexity. In particular when data rate and available bit-rate are relatively similar (right hand side of the graph), packet dropping provides similar QoE at much lower complexity compared to other adaptation options.

With respect to the reliability of video flows, i.e. resilience to losses, tuning the FEC across the protocol stack could be sufficient, in some specific scenarios, to solve the problem of QoE degradation. It is well known that video applications (unlike data application) can cope with bit errors. For this reason, allocating more redundancy to the application layer (AL-FEC) compared to the physical one (PHY-FEC) improves the QoE. The problem will be tackled in a general form in the MEDIEVAL project by taking into account FEC, ARQ and HARQ mechanisms. Furthermore, as soon as multiple Points of Access (PoA) are available, the Optimizer might decide to divert the video flows and re-allocate network resources in order to minimize the impact on QoE perceived by the other users at the PoA.

For practical scenarios, even after computing the best solution, our transport framework might not be capable of maintaining the QoE level requested at the user side. In such a case, a signal is sent back to the video service layer to perform content adaptation at the source, for instance, by changing the encoding parameters of the video flow from the source.

![Fig. 3. Utility functions for different video sequences obtained with open-loop transcoding, mapping the video rate to the perceived video quality.](image)

![Fig. 4. Utility functions for a dynamic video (Football) and a static video (Mother and Daughter), mapping the video rate to the perceived video quality when applying transcoding (TransH264) and packet dropping (PktDrop).](image)
low priorities by the CP. In this context we propose the use of scalable video encoders, such as H.264/SVC [18], designed to address the issues of serving heterogeneous user terminals. To improve the received video quality, packets with different priority levels are generated. According to the channel state and the fill level of the MAC buffers, packets from layers with low priority can be dropped in order to ensure the correct transmission of packets with higher priority, thus guaranteeing a minimum video quality.

Assuming the delivery of videos encoded with H.264/SVC [18], when a congestion in the mobile network is detected, the operator may decide to drop the least important video packets in the flow in case of small losses, while whole video layers may be dropped whenever the bandwidth reduction persists. The decision of where to drop packets/layers depends on the geographical location of the congestion. It is preferable to perform packet/layer dropping at the congested base station only, while the same action may also be performed at core network node responsible for serving a certain congested area (wider range than the previous one, i.e. including more base stations). Another possible action is to re-schedule the network and channel resources among users. This also includes the possibility of offloading the mobile operator’s cellular radio access network by switching a video stream to an available WLAN network within the radio range.

IV. Conclusions

In this paper we have built an intelligent video aware mobile network architecture which addresses the expected growth in demand for video content delivery in the coming years. This novel architecture takes QoE as its centric paradigm and is aiming to improve the joint perceived QoE of the different users by handling congestion through cross layers techniques. By exploiting the well developed CDN concept into future mobile networks, we have defined the mobile CDN (MCDN) in order to improve the utilization of network resources over the last mile. Techniques for P2P traffic optimization are also considered through advancing of the ALTO service to take mobile aspects into account. In our cross layers approach, through a tight collaboration of the video service layer with the underlying wireless access networks, traffic offloading from a congested network segment is offered. Moreover, traffic engineering mechanisms will be implemented taking advantages of H.264 SVC partitions to allow smart packets dropping at the underlying networks with minimal and smoother QoE degradation, in comparison to random dropping. We have shown with preliminary results that QoE varies in a non-linear manner as a function of the application data-rate for a number of video sequences, hence allowing the network to perform joint resource optimization such as allocating more resources to video flows where their QoE curves, as in Fig. 3, shows a potential large increase in the perceived QoE on the expense of flows with minimal QoE degradation, thus improving the overall joint QoE. Mechanisms for traffic shaping at the network are diverse and include bandwidth limitations, per-flow mobility, packet and layer filtering, content adaptation and more. The exact relation and use of traffic shaping mechanisms including their inter-relations are left for future work.

The immediate next steps in our research plan focus on the implementation and the evaluation of our transport framework as a collection of modules and algorithms with the aim of improving the QoE by adapting the video to the wireless constraints, selecting the traffic engineering policies for IP routing in the access and core network and adapting the data rate at the application layer. The P2P and the MCDN coexistence will be further analyzed and a set of practical scenarios will be selected to evaluate the enhancements introduced by the EU MEDIEVAL project on video delivery mechanisms for next generation cellular networks.

V. Acknowledgement

This work is supported by the EU MEDIEVAL research project. MEDIEVAL (MultiMEDia transport for mobile Video AppLications) is a medium-scale focused research project (STREP) of the 7th Framework Programme of the European Commission, addressing the core of the strategic objective “The Network of the Future”.

REFERENCES