Peer-to-Peer Videoconferencing with H.264 Software Codec for Mobiles

Hans L. Cycon*, Thomas C. Schmidt†, Gabriel Hege†, Matthias Wählisch‡†,
Detlev Marpe§, and Mark Palkow¶
*FHTW Berlin, 10318 Berlin, Germany
†HAW Hamburg, Dept. Informatik, Berliner Tor 7, Germany
‡link-lab, Hönower Str. 35, 10318 Berlin, Germany
§Heinrich-Hertz-Institut, Image Proc. Dept., Einsteinufer 37, 10587 Berlin, Germany
¶daViKo GmbH, Am Borsigturm 40, 13507 Berlin, Germany
Email: {hcycon,hege}@fhtw-berlin.de, {t.schmidt,waehlisch}@ieee.org,
marpe@hhi.fraunhofer.de, palkow@daviko.com

Abstract

A rapidly growing number of carriers offer wireless video services to their customers, taking advantage of high quality video codecs implemented in dedicated hardware of selected mobile devices. In this paper we introduce a video conferencing software*, which seamlessly integrates mobile with stationary users in a provider and device independent fashion. Innovations of this work are twofold. At first we report on a mobile realization of an H.264 video codec and its performance on a standard consumer Smartphone. Operating within the tight bounds of real-time compliance on mobiles, this software is an adapted version of a highly optimized H.264 codec. This DAVC codec, which we introduce along the line, significantly outperforms compatible H.264 realizations and allows for a scalable adaptation of its frame rate. In the second part we present a barrier-resistant peer-to-peer group communication scheme, which scales well for medium-size conferences and accounts for the heterogeneous nature of mobile and stationary participants. An outlook on mobility related group communication issues and future optimizations based on structured communication layers concludes the work.*

Index Terms

Mobile video coding, mobile conferencing, peer-to-peer group communication, distributed SIP conference management

1. Introduction

The idea of augmenting voice calls by video has been around for several decades, but only the flexibility of the Internet generated a noticeable deployment. As compared to audio, video processing places significantly higher demands on end system and network transmission capabilities. The rapid evolution of networks and processors have paved the way for realistic group conferences conducted at standard personal computers, combining about a dozen visual streams of Half-QVGA (240 x 160 pixel @ 15-30 fps) resolution.

Mobile phones and networked consumer portables are now on the spot to deliver sufficient performance for rich multimedia applications and communication, as well. Videoconferencing though, which requires simultaneous decoding and encoding in real-time, poses still a grand challenge to the mobile world. Limited and expensive wireless channels on the one hand, high consumer demands on visual quality on the other, advise applications to take advantage of the latest standard for video coding H.264/AVC [1].

H.264/AVC provides gains in compression efficiency of up to 50 % over a wide range of bit rates and video resolutions compared to previous standards. While H.264/AVC decoding software has been successfully deployed on handhelds, high computational complexity still prevented pure software encoders in current mobile systems. There are however also fast hardware implementations available, which give rise to an increasing offer of device- and operator-bound video services.

In this work we first introduce a pure software solution for real-time video communication on standard

smartphones in section 2. These mobile clients extend a lightweight, feature rich conferencing application developed for an infrastructure compliant use on standard PCs. In the second part we present the underlying peer-to-peer group communication scheme, which performs well for medium-size conferences and accounts for the heterogeneous nature of mobile and stationary participants, cf. section 3. This includes on the one hand SIP [2] standard compliant session signalling with respect to group communication, and on the other hand efficient, serverless media distribution, self-adjusting to the actual network infrastructure support. Extensions to distribute media by source specific multicast in this peer-to-peer model have been previously developed [3], but are currently not focused upon due to limited multicast deployment. Conclusions and an outlook follow in the final section.

2. The daViKo Videoconferencing Software

In this section we give an overview of our reference implementation, a digital audio-visual conferencing system, realised as a serverless multipoint video conferencing software without MCU developed by the authors [4]. It has been designed in a peer-to-peer model as a lightweight Internet conferencing tool aimed at email-like friendliness of use. The system is built upon a fast H.264/MPEG-4 AVC standard conformal video codec implementation [5] called DAVC. By controlling the coding parameters appropriately, the software permits scaling in bit rate from 48 to 1440 kbit/s on the fly.

Audio data is compressed using a 16 kHz speech-optimized variable bit rate codec [6] with extremely short latencies of 40 ms (plus network packet delay). All streams can be transmitted by unicast as well as by multicast protocols. Within the application, audio streams are prioritized over video since user experience is usually more sensitive to losses in audio packets than those of video packets, which both may result from transmission errors or network congestions.

An application-sharing facility is included for collaboration and teleteaching. It enables participants to share or broadcast not only static documents, but also any selected dynamic PC actions like animations including mouse pointer movements. All audio/video - streams including dynamic application sharing actions can be recorded on any site. This system is equally well suited to intranet and wireless video conferencing on a best effort basis, since the audio/video quality can be controlled to adapt the data stream to the available bandwidth.

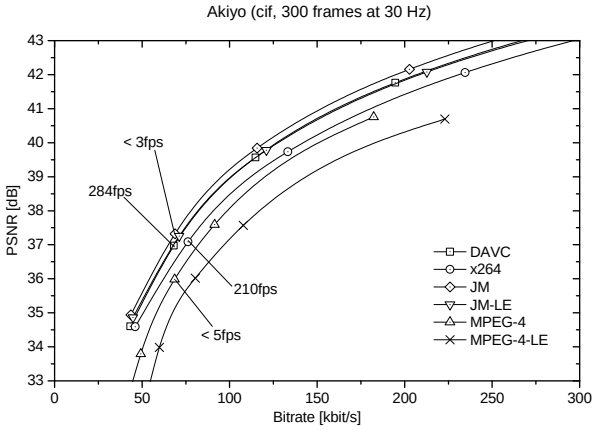


Figure 1. RD plot for the test sequence "Akiyo" in CIF resolution comparing three different H.264/MPEG-4 AVC encoder implementations as well as a RD-optimized MPEG-4 (Part°2) Advanced Simple Profile implementation.

The daViKo conferencing system is available for desktop computers running MS–Windows or Linux and on handhelds with Windows Mobile operating system.

2.1. The DAVC Codec

DAVC, the core of the videoconferencing system, is a fast, highly optimized H.264/MPEG-4 AVC standard implementation. It realizes a Baseline profile, optimized for real-time encoding (as well as real-time decoding) by means of a fast motion estimation strategy including integer-pel diamond search as well as a fast subpel refinement strategy up to $\frac{1}{4}$ pel motion accuracy. Motion estimation includes the choice of several different macroblock (MB) partitions and multiple reference frames, as permitted by the H.264/MPEG-4 AVC standard. For choosing between different MB partitions for motion-compensated (i.e. temporal) prediction and MB-based intra (i.e. spatial) prediction modes, a fast rate-distortion (RD) based mode decision algorithm with early termination conditions has been employed.

In comparison to the well-known open source H.264/MPEG-4 AVC encoder implementation of x264 [7], our DAVC encoder implementation achieves a similar RD performance and a slight increase in encoding speed when using comparable encoder settings. In Figure 1, a typical example of such a comparison between x264 and DAVC is shown. In addition to the RD-performance of those two real-time encoder implementations, this plot also shows the RD behavior

of two non real-time encoder implementations, as given by the H.264/MPEG-4 AVC Joint Model (JM) reference software (with Baseline profile settings) and a MPEG-4 (Part°2) Advanced Simple Profile implementation. The latter two encoders were operated using a high-complexity RD-based mode decision strategy for demonstrating the capabilities of both video coding standards when neglecting any real-time constraints. Figure 1 also contains the number of encoded frames per second (fps) for selected RD points as a measure for maximum encoding speed (e.g., 284 fps for DAVC as compared to 210 fps for x264). Similar results were also achieved for other test sequences.

Note that the DAVC codec also includes some suitable mechanisms to quickly recover from video packet loss.

2.2. Mobile Video Codec

The DAVC codec has been adapted to sustain real-time performance on mobile devices. The mobile codec version operates at reduced complexity for motion compensation with a highly optimized code base for the target platform. This tuning includes the efficient use of the wireless MMX instruction set available at our current target system. Portability is sustained by an ANSI compliant C version, to be augmented incrementally by platform specific injections.

In the DAVC Mobile codec, motion compensation has been restricted to integer-pel block search, which adjusts encoding efforts to the limited processing capacities of the consumer class system. The penalty obtained for coding efficiency remains maintainable though, keeping bandwidth demands of the application within tolerable bounds. Performance values of the mobile encoder are displayed in figure 2 for the Akiyo test sequence in QCIF format.

The application was tested on a 520 MHz Xscale processor in an Asus P735 system. Thereon it can encode the Akiyo test sequence at a rate of 45 fps. In realistic deployment scenarios the application can reliably encode and decode a QCIF video stream in parallel at 10/15 fps, without CPU exhaustion or frame dropping. QCIF @15 fps is the maximal image feed that can be obtained from the front camera in our test equipment and we expect to arrive at realistic real-time performance at this encoding rate after further optimizations. The maximal battery life time at continuous conferencing use, i.e., encoding and decoding of parties permanently in moderate motion complemented by 802.11 WLAN transmission and activated display was measured to slightly exceed two hours.

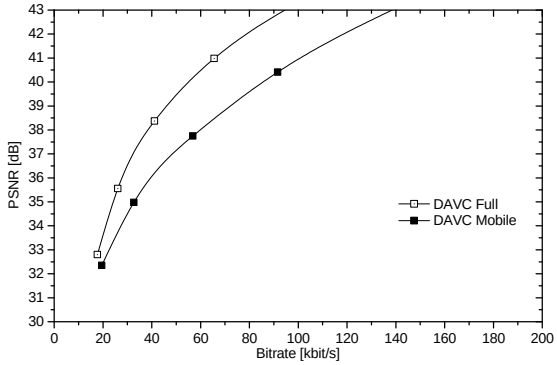


Figure 2. RD plot for test sequence “Akiyo” in QCIF resolution at 10 fps, comparing the DAVC mobile encoder to the DAVC fully optimized implementation. Real-time encoding performance was at 45 fps.

	Framerate	Bandwidth
Desktop	30 fps	190 kbit/s
Desktop	15 fps	120 kbit/s
Smartphone	15 fps	135 kbit/s

Table 1. Performance of DAVC and DAVC Mobile at QCIF resolution

Performance values from an empirical test at vivid camera motion are shown in table 1. Comparison is drawn between the full DAVC codec running on a standard desktop and the mobile DAVC on the handheld. Reduced coding complexity results in an enhanced data rate send by the mobile, but the gross total rate for a bidirectional video exchange at 15 fps complies to 3GPP/UMTS bandwidths constraints. Note that experimental conditions are not fully compatible: The image sequence obtained from the front camera of the mobile is significantly more noisy than our standard USB cameras connected to the desktop, which increases the image complexity and thereby the data rate.

3. Simple, Distributed Point-to-Point Conferencing

Our application aims for simple, flexible, and cost-efficient ad-hoc conferencing functions, which scale appropriately well, but avoid any infrastructure assistance. Such a solution requires group session management and media distribution at peers, which for the sake of standard compliance we realize with group conferencing functions in SIP, cf. [8], [9], [10]. Implemented as pure software on standard personal devices, user agent peers are exposed to severe restrictions in

Figure 3. The mobile video application.

real-world deployments: Often they are located behind NATs and firewalls with network capacities confined to asymmetric DSL or wireless links. Capacity constraints and resilience to node failures require peer-managed ad-hoc conferences to organize in a distributed multi-party model. As a key component, the heterogeneity of clients must be accounted for, whereas the range of scalability is limited to about a dozen parties in videoconferences.

3.1. P2P Adaptive Architecture

A peer-to-peer conferencing system faces the grand challenge to be robust w.r.t. the infrastructure. The role a user agent is able to attain in a distributed scenario needs to be adaptively determined according to constraints of its device and current network attachment. In a simplified scenario, clients may be divided into two groups, distinguished by their ability to act as a SIP conference focus or not. A focus must be globally addressable and have access to necessary processing and network resources.

This elementary adaptation scheme can be based on individual decisions of user agents and gives rise to a hybrid architecture of super peers, chosen from potential focus nodes, and remaining leaf nodes. To decide on its potential role of building a focus, a client at first needs to determine NATs and firewalls. Aside from address evaluation, this is done by a simple probe packet exchange. As the implementation is CPU-type aware, processing restrictions are easily evaluated, as well. However, an a priori judgement on available network bandwidth is not easily obtained. An evaluation of the local link capacity is frequently misleading, as wireless devices may be located behind wired transmitters of lower, asymmetric capacity such as in ADSL. Current experiments to quickly retrieve reasonable estimates of up- and downstream capacity are ongoing on the basis of variable packet size, non-intrusive estimators, cf. [11]. Note that network capacity detection is of vital use for temporal adaptation of the video codecs, as well.

Leaf nodes attach to super peers in subordinate position, whereas potential focus nodes may be assigned to be super peers or leaves. Super peers provide global connectivity among each other and NAT traversal assistance to leaves, while leaf nodes experience super peers in different roles: A leaf nodes sees its next hop super peer as the conference focus, while the remote super peers act as proxies on the path to the leaves behind.[1] This set-up corresponds to the well known architecture of Gnutella 0.6 and successive hybrid unstructured peer-to-peer systems, cf. [12]. Despite its architectural analogy, a routing layer for real-time group applications should follow a different design.

From the perspective of the conference, super peers form a distributed focus. To keep distribution transparent to leaves, each super peer needs to provide full conferencing service functions, e.g., synchronized policing and event notification, and most likely assistance in media mixing and redistribution. Focus nodes consequently require signaling relationships among each other, established on top of an application layer routing of sufficient performance for a simultaneous distribution of media streams.

3.2. Routing Design

The design of a routing will admit critical impact on scalability, application performance, as well as forwarding and maintenance load of the super peers. The three characteristic topologies for routing between N super peers as displayed in figure 4 explore the problem space: On the one extreme, routing on a ring will minimize neighbor states and forwarding load of each peer, but requires $\mathcal{O}(N)$ hops and thus induces large, varying delays. A full mesh, on the other extreme, places the burden of $N-1$ neighbor states to be fed in replicated forwarding, but guarantees a rigid 3-hop forwarding limit and minimal delays. A polygonal mesh of dimension d keeps replication load constant (but dependent on d), while its corresponding path lengths grow as $\mathcal{O}(\sqrt[d]{N})$. Forwarding on a polygonal mesh will require routing intelligence, which is neither needed on a ring nor in a full mesh topology. As

1. This architecture relies on the presence of at least one globally addressable, sufficiently powerful peer. As there are many scenarios, where this is likely to fail, we advise for and offer a permanently deployed 'silent' relay-peer at some unrestricted place.

routing paths in conferencing scenarios are equivalent to the signaling relationships, mesh robustness respectively redundancy of the schemes is equivalent to the number of neighbor states at each peer.

Our problem focuses on moderately sized peer-to-peer conferences of simple and robust nature. Therefore a favorable routing scheme is easily identified. The full mesh topology outperforms alternative schemes in forwarding efficiency and robustness, while scaling well up to a hundred nodes, provided a significant fraction of unrestricted, high-performance super peers is available. In addition, this scenario is bound to low complexity, since no routing intelligence beyond standard SIP logic of next hop proxying is required. We thus use a full mesh topologies here as the favorable approach to mid-size multi-party conversations.

3.3. SIP Representation

To explore the corresponding conference scenario in detail, consider an ad-hoc join. A client submits an INVITE to any party. It thereby needs to indicate its potential roles in some way. As a corresponding client protocol extension has not been specified yet, cf. [13], we use a proprietary payload here.

The callee may be a conference focus or leaf node. In the first case, it will be aware of the overall leaf node distribution from the conference event states and will transfer the newly joining party to the least occupied super peer by a REFER, e.g.,

```
REFER sip:lucy@psychic.org SIP/2.0
...
CSeq: 9380 REFER
Refer-To: <sip:hypnotic-talks@\\
              vain-focus.circles.com>
Content-Length: 0
```

In the second situation, the contacted leaf node will issue a re-INVITE to attach the new conference member to its focus, which in turn may refer the caller to another focus for the sake of load balancing, e.g.,

```
INVITE sip:lucy@psychic.org SIP/2.0
...
CSeq: 1199 INVITE
Contact: <sip:hypnotic-talks@\\
              my-focus.circles.com>
Content-Type: application/sdp
```

Having indicated its ability to serve as a super peer, the newly arrived party may be selected to join the group of focus nodes. This decision is taken by its current super peer and realized via a 3rd party invite issued to the group of all established focus nodes. The elected super peer will thereby establish point-to-point signaling relationships with all correspondents, leading to an immediate formation of the full mesh conference topology. Focus election and leaf node distribution is conducted in an individual step-by-step way, following an eager strategy. Its implementation is driven by maximizing resistance to the overall environment and redundancy of super peers.

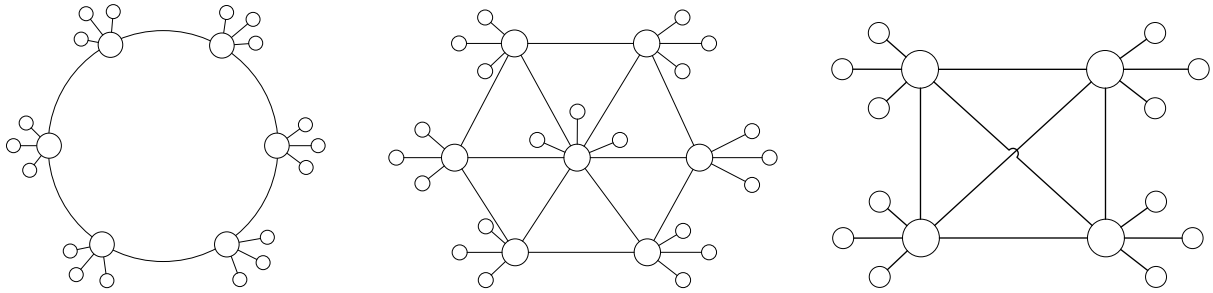
Note that media negotiations have been part of the initial arrival steps for each party. Media distribution will naturally follow the paths of the established routing topology, where super peers act as two-sided media controllers: They can combine media streams arriving from their attached leaf nodes before peering them within the focus mesh, but withhold media arriving from neighboring super peers, which are not needed at clients in order to assure a lean transmission to lightweight or mobile leaves.

4. Conclusions & Outlook

We have presented a peer-to-peer software for high-quality videoconferencing on mobiles, admitting utmost flexibility with respect to end systems, operators and network provisioning. To the best of our knowledge, this is the first software implementation of an H.264 video encoder that operates in real-time on mobile phones. An adaptive, fully distributed conference management scheme with SIP has been developed as part of the multi-party scenario. This hybrid peer-to-peer model accounts for client capabilities as well as network attachment, and does scale well beyond standard use.

In future work we will concentrate on further optimization and generalization of the video coding software to make it available for a wider variety of platforms. Network adaptation and capacity evaluation will require further work to arrive at estimates that reliably serve the needs in real world environments, as well.

Additional research will target at benefits possibly inherited from key-based routing. As common application layer multicast schemes, which rely on dedicated shared or source specific trees, are significantly sensitive to client departure and of insufficient performance in medium size groups, and as conference routing actually can be seen as an application layer broadcasting problem, new and highly optimized structured broadcast algorithms are desirable. The bidirectional shared tree approach introduced in [14] may be a promising point to start at.



(a) Ring (b) Polygon (c) Full mesh

Figure 4. Peer-to-peer routing toplogies on the overlay

Acknowledgement

This work is supported by the German Bundesministerium für Bildung und Forschung within the project *Moviecast* (http://moviecast.realmv6.org).

References

[1] ITU-T Recommendation H.264 & ISO/IEC 14496-10 AVC, "Advanced Video Coding for Generic Audiovisual Services," ITU, Tech. Rep., 2005, draft Version 3.

[2] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, and E. Schooler, "SIP: Session Initiation Protocol," IETF, RFC 3261, June 2002.

[3] T. C. Schmidt, M. Wählisch, H. L. Cycon, and M. Palkow, "Scalable Mobile Multimedia Group Conferencing based on SIP initiated SSM," in *Proc. of 4th European Conference on Universal Multiservice Networks – ECUMN'2007*. Washington, DC, USA:IEEE Computer Society Press, February 2007, pp. 200–209.

[4] M. Palkow, "The daViKo homepage," 2008, http://www.daviko.com.

[5] J. Ostermann, J. Bormans, P. List, D. Marpe, N. Narroschke, F. Pereira, T. Stockhammer, and T. Wedi, "Video Coding with H.264/AVC: Tools, Performance and Complexity," *IEEE Circuits and Systems Magazine*, vol. 4, no. 1, pp. 7–28, April 2004.

[6] "The Speex projectpage," http://www.speex.org, 2007.

[7] "VideoLan: x264 - a free h264/avc encoder," http://www.videolan.org/developers/x264.html, 2007.

[8] A. Johnston and O. Levin, "Session Initiation Protocol (SIP) Call Control - Conferencing for User Agents," IETF, RFC 4579, August 2006.

[9] R. Mahy, R. Sparks, J. Rosenberg, D. Petrie, and A. Johnston, "A Call Control and Multi-party usage framework for the Session Initiation Protocol (SIP)," IETF, Internet Draft - work in progress 9, November 2007.

[10] T. C. Schmidt and M. Wählisch, "Group Conference Management with SIP," in *SIP Handbook: Services, Technologies, and Security*, S. Ahson and M. Ilyas, Eds. Boca Raton, FL, USA: CRC Press, November 2008, to appear, on invitation.

[11] R. Prasad, C. Dovrolis, M. Murray, and kc claffy, "Bandwidth Estimation: Metrics, Measurement Techniques, and Tools," *IEEE Network*, vol. 17, no. 6, pp. 27–35, November–December 2003.

[12] R. Steinmetz and K. Wehrle, Eds., *Peer–to–Peer Systems and Applications*, ser. LNCS. Berlin Heidelberg: Springer–Verlag, 2005, vol. 3485.

[13] D. Bryan, P. Matthews, E. Shim, and D. Willis, "Concepts and Terminology for Peer to Peer SIP," IETF, Internet Draft - work in progress 01, November 2007.

[14] M. Wählisch and T. C. Schmidt, "Between Underlay and Overlay: On Deployable, Efficient, Mobility-agnostic Group Communication Services," *Internet Research*, vol. 17, no. 5, pp. 519–534, November 2007.