High Quality Visual Communication based on IMS

Laurits Hamm, Frank Hartung
Ericsson GmbH, Eurolab R&D, Ericsson Allee 1, 52134 Herzogenrath
[laurits.hamm|frank.hartung]@ericsson.com

Abstract

Systems for bidirectional visual communication have been available for a long time, but were rarely used in daily private or business life. However, in the last years the user acceptance of visual communication services has increased significantly. Not only for consumers but especially for businesses and global companies, videotelephony and videoconferencing have become more important, driven by the requirements of reducing travel cost, travel time, and emissions. The 3GPP IP Multimedia Subsystem (IMS) is a standardized framework for deploying multimedia services, such as telephony, messaging and presence. This article discusses advanced visual communication services and how such services and the underlying technical components can be realized. We focus on high quality videoconferencing, using high-definition video and full-band audio.

1 Introduction

1.1 Visual Communication and End User Experience

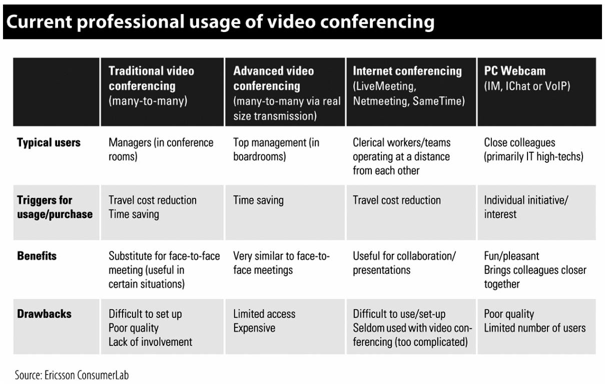
Visual communication is a term for communication services using video and audio transmission. Visual communication systems have been available for a long time. For example, the AT&T Picturephone was commercially available already end of the 1960s. The widest deployment of videotelephony today is in mobile phones: most 3G mobile phones with built-in camera are able to make one-to-one video calls to other mobile phones. However, users never embraced the concept, and only very few video calls are being made. Skype, a free PC-based software for voice and video calls, has become popular, but provides a very limited video resolution and quality.

The situation is different for business use of visual communication. In the business market, videoconferencing and telepresence services are increasingly used. Videoconferencing is a conference call between three or more participants using video and audio. Telepresence is an advanced service using high quality video and audio to achieve a sense of realism and presence between the participants. Telepresence services often use multi-screen video and multi-channel audio with high resolutions and bandwidths. Rooms for telepresence are especially equipped and can allow life-size display of the remote site.

Audio and video quality, resolution, framerate, and bandwidth are not only quantitative improvements. They allow verbal and non-verbal communication between remote sites. Non-verbal communication, such as eye contact, gestures, facial expressions, body language, and voice tone, is an essential part in a real face-to-face conversation. In a standard voice call almost all non-verbal communication is lost and today's mobile phone or PC based videotelephony does not provide a sufficient quality for it. Full-band audio and high-definition (HD) video transmissions allow verbal and non-verbal conversations with a high sense of presence. Multi-screen and multi-channel room installations allow further a spatial orientation of the participants. However, such telepresence systems are often very expensive regarding investment and operation.

1.2 Market Trends and Business Potential

Voice is the dominant service in the telecommunication industry, but video has overtaken in terms of data traffic already and more people start using videotelephony as well. This is also due to increased advertisement and marketing activities around visual communication. Cheap webcams and free internet videotelephony, such as Skype or Apple Facetime, attract people, despite of the low quality. The traditional telecommunication industry has launched fixed and mobile videotelephony years ago but it has not been accepted in a larger scale. The reasons for this are manifold, some of them are interoperability, quality, and price [1]. Standard videotelephony offers only low qualities which has lead to a bad user experience, especially small screens and pixilated images are not appealing. Technology advances, e.g., broadband access and improved codecs, allow higher video qualities nowadays, but there is still a gap between standard videotelephony and high-end videoconferencing and telepresence systems. The later need high bandwidths of several Mbit and connectivity guaranteeing high quality of service (QoS). Operators can profit from this because they have networks already in place to offer and manage high quality services.



Current professional usage of video conferencing

	Traditional video conferencing (many-to-many)	Advanced video conferencing (many-to-many via real size transmission)	Internet conferencing (LiveMeeting, Netmeeting, SameTime)	PC Webcam (IM, IChat or VoIP)
Typical users	Managers (in conference rooms)	Top management (in boardrooms)	Clerical workers/teams operating at a distance from each other	Close colleagues (primarily IT high-techs)
Triggers for usage/purchase	Travel cost reduction Time saving	Time saving	Travel cost reduction	Individual initiative/ interest
Benefits	Substitute for face-to-face meeting (useful in certain situations)	Very similar to face-to-face meetings	Useful for collaboration/ presentations	Fun/pleasant Brings colleagues closer together
Drawbacks	Difficult to set up Poor quality Lack of involvement	Limited access Expensive	Difficult to use/set-up Seldom used with video conferencing (too complicated)	Poor quality Limited number of users

Source: Ericsson ConsumerLab

Table 1 Usage of video conferencing [1].

Videoconferencing services range from free internet services to high-end professional systems, see **Table 1**. Businesses and global companies are especially interested in visual communication services to increase their efficiency and to reduce costs. Videoconferencing and telepresence are used together with collaboration tools to reduce travel costs and to establish an efficient way-of-working.

However, there exists a gap between cheap but low quality consumer grade visual communication, based on mobile phones and PC applications, and high quality but expensive business videoconferencing and telepresence systems. High quality systems based on off-the-shelf hardware and standard connectivity are not widely available. In this article, we show how high quality videoconferencing services can be realized economically using IMS, standard internet technologies, and off-the-shelf consumer electronics.

2 Visual Communication based on IMS

2.1 IMS – an Interoperable Control Layer for Visual Communication

The IP Multimedia Subsystem (IMS) is a framework for multimedia services in telecommunication networks [2]. 3GPP defined a set of standards to describe the architecture and procedures of IMS [7]. The initial motivation for developing IMS was to offer internet services over mobile networks. It has then grown to also offering fixed services and became an essential part in Next Generation Networks and for fixed and mobile network convergence. One IMS core network can offer services for fixed and mobile networks and offers interworking to other IP and legacy circuit switched networks. 3GPP describes a core network for IMS on top of which services are deployed in the application layer.

IMS is based on IP transport and uses a layered architecture, in which control plane and user plane are separated. The reference architecture defines the logical nodes and interfaces between the nodes [7]. Internet technologies and protocols from the IETF are used in the control and user plane. The most important protocols used in IMS are the Session Initiation Protocol (SIP) for session signaling, Diameter for authentication, authorization and accounting and the Real-time Transport Protocol (RTP) for media transport.

Figure 1 shows an IMS core network and nodes for conferencing services. The P- and S-CSCF (Call and Session Control Function) are SIP servers handling the session control. The P-CSCF is a proxy server and the first point of contact for a terminal in the core network. The main functions of it are to authenticate the user and to verify the received SIP requests. The S-CSCF is the central node in the user plane and does session control. It acts as a SIP registrar and uses a Diameter interface towards the HSS to download the user's service profiles. According to the initial filter criteria defined in the service profile, the S-CSCF triggers the application servers. Different application servers implement different services, and the S-CSCF needs to trigger and forward SIP requests to the correct application servers.

The Home Subscriber Server (HSS) is the central database. It stores the subscriptions of the users and the service profiles. The subscription data is used for authentication and authorization. The service profiles provision the services for a user. Service profiles include operator predefined configurations and user configurations. Hence, the operator can control which services a user is allowed to use and the user can do own configurations.

The application servers (AS) provide the services and different services can be implemented in separate application servers. An application server receives SIP requests over the service control interface (ISC) from the S-CSCF. Different application servers can interact over the ISC interface, and services can be executed after each other by chaining and triggering application servers after each other. For conferencing services we use two application servers, a Conferencing AS and a Booking AS.

The Media Resource Function (MRF) provides a source of media in the network. It can play announcements, mix different media streams, and do transcoding. In a normal SIP call the media plane is established directly between the caller and the callee using RTP without a MRF. In a conference call the MRF plays an important role for mixing the audio and video streams from all participants. The media plane between two users is established using the Session Description Protocol (SDP). The SIP request from the originating caller includes an SDP offer with the supported media types and codecs [14] in the SIP body. The receiving side replies with an SDP answer in the SIP response.

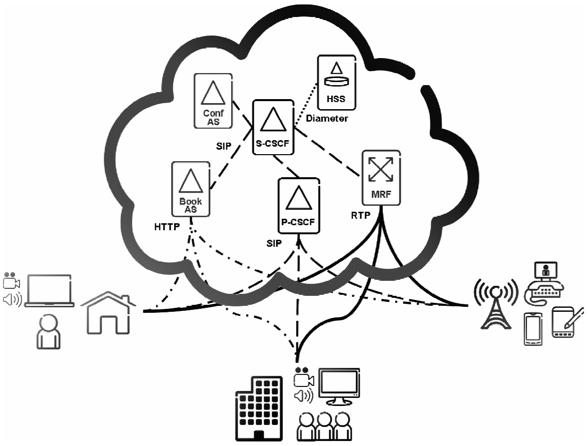


Figure 1 IMS architecture and protocols for conferencing services.

Besides the IMS core network elements **Figure 1** shows two application servers (AS) and one Media Resource Function (MRF). The Booking AS allows the participants to schedule conferences and to manage bookings. The Conference AS executes the conference service for the participants. It acts as a SIP user agent and is the conference focus which terminates the SIP sessions from the participants [9].

The Conference AS interacts with the MRF to requests media resources for new participants. In a videoconference the MRF receives the media streams, audio and video, from the participants. Depending on the clients the media streams can be identical or they can have different characteristics, e.g., codecs or video resolutions. A participant could use a mobile phone, a fixed videophone, or a conference room; and a client could send different numbers of audio and video streams, for example, if it uses multiple cameras or multiple views. The MRF will mix one audio stream for each participant. The audio for one participant is the mix of the audio from the other participants. It is not necessary to mix the audio streams from all participants, but the MRF can choose and mix only relevant streams, e.g., from active speakers. It can further use audio signal processing, e.g., gain control and active speaker detection.

For video the processing is different. The MRF could choose the video from one participant, typically the active speaker, and send it to the other participants. It can send several video streams from different participants or it could mix multiple pictures in one video stream, i.e., split screen or picture-in-picture. The MRF can do transcoding of the audio and video, if clients do not support the same codecs. This is necessary to achieve interoperability. Different available bandwidths due to different access types, multiple video resolutions and different number of audio channels are other reason for transcoding.

Simulcast transmission is another option for video streams. A client using simulcast transmits the same content in different encodings, e.g., in high and low quality at the same time. The advantage is that the complexity of the MRF is reduced and no additional transcoding delay is added to the end-to-end system delay. For real-time communication services end-to-end delays must be optimized, and transcoding of HD video adds a significant delay to the system.

2.2 Videoconferencing Services

Different types of conference services exist, which are distinguished by the conference creation, how participants are added to the conference, and the mixing of media streams [10]. For example, a scheduled dial-in conference is a typical use case for companies: a scheduled conference call is planned ahead of time and booked for a certain start time and duration. Conference participants join the conference by dialing-in to the conference. This conference type is often used by businesses and companies where the participants use especially equipped conference rooms. The system should offer legacy support which means it should be possible for participants to join the conference from a normal circuit switched phone. Another way of creating a conference call is ad-hoc. An ad-hoc conference is not booked in advance. The first conference participant starts the conference call and takes the role of the conference creator. The conference creator invites and adds other participants to the conference.

Depending on the conference type, participants are added to a conference call either by dial-in or by dial-out. To dial-in to a conference the participants use a conference URI, which is distributed beforehand. A dial-out is performed when a user is invited to a conference call by another conference participant, i.e., he is called on his phone or device.

The last criteria distinguishing different conference systems is the media mixing. The media streams send and received by the participants can be mixed and controlled centrally by a media server, the MRF in an IMS system. All media flows are controlled by the media server and it can do transcoding and bitrate adaptation. The contrary architecture for media handling is a distributed media mixing. In this case each participant sends its media streams to all other participants. Obviously, this is difficult for conferences with a larger number of participants, because it creates a high number of media streams and requires high bandwidths. This is especially problematic in the uplink, because many access technologies only offer smaller bandwidths in the uplink than in the downlink, e.g., DSL or HSPA.

2.3 Session Establishment for Scheduled Conferences

We will focus on scheduled dial-in conferences using two application servers and one central media server. Such a conference is also called a tightly coupled conference. The conference creator schedules a conference call via the Booking AS. An interface for booking using HTTP allows integrating the booking service in the user clients, on a booking website or into a calendar application. The Booking AS uses a database to store the booked conferences. Conferences can be booked recurrently, can be modified, and can be deleted.

In the IMS architecture, which we described in the previous chapter, the Conference AS is the central managing function. It handles the conference creation and acts as the conference focus. Each conference uses a unique conference URI which the participants use to dial-in to the conference [11]. The conference focus is created when the first participant dials-in to the conference, and the conference URI uniquely identifies the conference focus. The Conference AS manages a SIP dialog per participant and it reserves media resources via the MRF. The SIP protocol together with the Media Server Control Protocol [12] as payload can be used to reserve and update media resources in the MRF. MSCML is transported as a XML SIP body. The protocol allows to manipulate media flows during a conference and to send notifications from the MRF to the Conference AS. This can be used for muting media streams and for active speaker notifications.

The participants can receive information about the conference state by using the conference notification service. The service uses SIP SUBSCRIBE and SIP NOTIFY messages together with a specific event package for conferencing [13]. After a participant has successfully subscribed to the conference events the Conference AS sends notifications every time a conference event occurs. The conference events include information about the conference state, the participants, and the used media. For example, an event will be triggered when a participant joins or leaves the conference.

Figure 2 shows a simplified SIP signaling flow of a participant dialing-in to a conference call and subscribing to conference events. First, the client needs to register in the IMS core network. Then the participant dials-in to the conference sending a SIP INVITE using the conference URI. The Conference AS will reserve media resources at the MRF using SIP INVITE and MSCML. After the media resources are successfully reserved by the MRF the Conference AS sends the SIP 200 OK to the participant and the RTP media plane is established. Finally, the client subscribes to the conference events sending a SIP SUBSCRIBE and receives the initial SIP NOTIFY containing the current conference state.

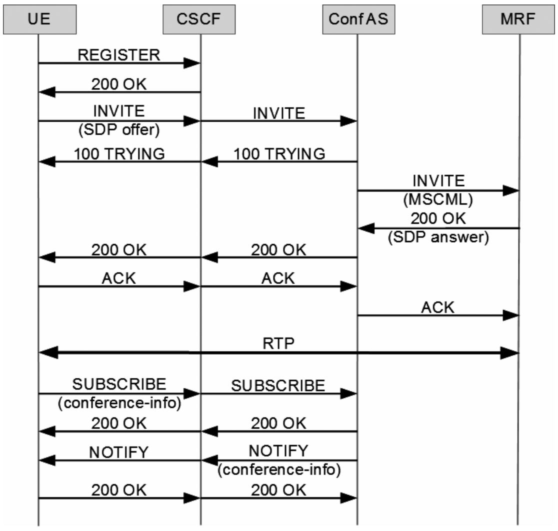


Figure 2 Signaling flow for a user dialing-in to a conference.

2.4 High Quality Audio and Video

The user experience for videoconferencing gains significantly from high quality audio and video, and care must be taken to capture and preserve a high media quality on the recording and the presenting sides. For the audio signal, a full-band codec should be used that is not hampered by the bandwidth limitation of old-fashioned telephone systems. The human voice occupies the frequency range between 50 Hz and 12 kHz, which approximately coincides with the audible frequency range for human adults. This frequency range should be transmitted as transparently as possible, using a full-band (48 kHz) audio codec such as G.719 [5] or at least a wide-band (16 kHz) codec such as AMR-WB [6]. The combination of high quality yet consumer-grade microphones and a full-band stereo audio codec, which operates at a sampling rate of 48 kHz, gives superior audio quality. Users almost have the audio perception of being ‘in the same room’. The bitrate for full-band stereo audio can be up to 128 kbps.

For the video signal, HD video quality is desirable. Consumer grade HD video cameras and webcams offering full HD and HD resolutions (1080p and 720p respectively) are available on the market at affordable price. Many households already have television and computer screens with HD resolutions. Hence, users often do not need to buy new and expensive equipment for high quality visual communication. To encode video an advanced video codec such as H.264 [4] should be used. Today’s codecs offer high compression ratios, however the bitrates for full HD and HD video are still high and can be up to several Mbit per second. Video coding is an area of ongoing research with new and improved codecs under development. Using the Baseline Profile of H.264, average bitrates of 1.5 Mbps and 3 Mbps are realistic for conference scenarios using 720p and 1080p resolutions respectively (with 30 frames per second). Video encoding is a complex task which produces high computational load, and the tradeoff between encoding complexity, bitrate, and encoding delay has to be considered for real-time communication systems.

Two factors raise the need for signaling feedback especially for video. First the high compression ratio of video encoding makes its transmission especially vulnerable to packet loss. A packet loss of only some percent can destroy the decoding on the receiving side completely. The second reason is the temporal prediction between pictures used in video encoding. Due to the prediction it is only possible to start decoding a video stream at certain points, where a frame without temporal prediction (so called I-frame) is transmitted. Feedback can be sent in the media plane directly using RTCP based feedback [15, 16]. RTCP based feedback for H.264 includes negative acknowledge (nack), picture loss indication (pli) and full-intra request (fir) messages.

```
1.   v=0
2.   o=Exa 34135268 0 IN IP4 192.168.123.101
3.   s=Example of G.719 and H.264
4.   c=IN IP4 192.168.123.101
5.   t=0 0
6.   b=AS:1650
7.   a=tcap:1 RTP/AVPF
8.   m=audio 49152 RTP/AVP 97 98
9.   b=AS:150
10.  b=RS:0
11.  b=RR:2000
12.  a=rtpmap:97 G719/48000/2
13.  a=rtpmap:98 AMR-WB/16000/1
14.  a=fmtp:98 mode-change-capability=2; max-
       red=220; octet-align=1
15.  a=ptime:20
16.  a=maxptime:240
17.  a=sendrecv
18.  m=video 49154 RTP/AVP 100 101
19.  b=AS:1500
20.  b=RS:0
21.  b=RR:2500
22.  a=rtpmap:100 H264/90000
23.  a=fmtp:100 profile-level-id=42C01F;sprop-
       parameter-sets=Z0LAH5ZkAoAt0IAAAAMAgA
24.  a=imageattr:100 send [x=1280,y=720]
25.  a=rtpmap:101 H264/90000
26.  a=fmtp:101 profile-level-id=42e00c;sprop-
       parameter-sets=J0LgDZWgUH6Af1A=
27.  a=imageattr:101 send [x=320,y=240]
28.  a=rtcp-fb:* trr-int 5000
29.  a=rtcp-fb:* nack
30.  a=rtcp-fb:* nack pli
31.  a=rtcp-fb:* ccm tmmbr
32.  a=rtcp-fb:* ccm fir
```

Figure 3 Example of an SDP offer including full-band audio, HD video and RTCP based feedback.

Transport of audio and video streams uses the RTP/RTCP transport protocol. When audio and video streams are being set up, an SDP-based negotiation takes place in which information about the used codecs and transport is exchanged [8]. **Figure 3** shows an example SDP offer for audio and video with different codecs, here G.719, AMR-WB, and H.264. Audio and video use two separate media lines and for different codecs use different payload types (see Figure 3 line 8 and line 18). A bandwidth of 1.65 Mbps is reserved for the session including audio and video (see Figure 3 line 6). For H.264 video the client supports RTCP based feedback (see Figure 3 line 28-32).

3 Ericsson Visual Communication Prototype

Ericsson has developed a prototype for visual communication. **Figure 4** shows an experimental set-up of a meeting room with the Ericsson Visual Communication system. The hardware is high-end yet consumer grade commercial off-the-shelf hardware: a standard desktop PC, HD LCD television screens, microphones, and loudspeakers. The prototype offers high quality videoconferencing services based on IMS [3] using full-band stereo audio and HD video. The video is encoded in parallel in three versions: HD (720p), half-resolution (360p), and thumbnail. All three versions are sent uplink to the media server, which acts as a switchboard and forwards the suitable video stream to the other participants. The video of the active speaker is distributed to the other endpoints in HD quality, while the non-active participants are distributed in a lower resolution and presented smaller on the user interface. RTCP based feedback is used between the clients and the media server. When a new participant enters the conference, a full-intra request is sent for the video streams such that the respective encoders are forced to produce random access points (I-frame), which the new participant can use to hook into the ongoing video streams. If packet losses occur on the link between the media server and a client, packets are re-transmitted. If re-transmission using RTCP based nack feedback is not enough, pli or fir is used to recover the video streams.

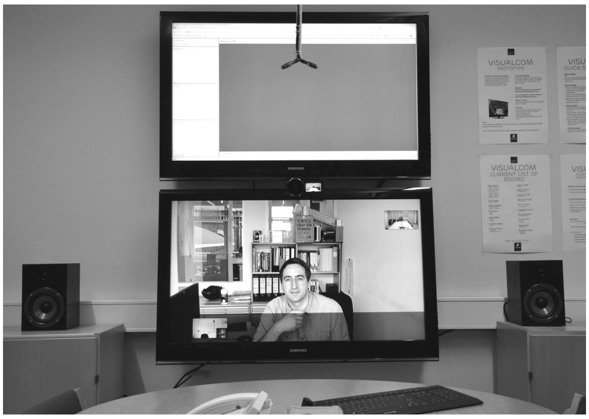


Figure 4 Conference room equipped with the Ericsson Visual Communication prototype.

4 Outlook

Currently, there is a strong trend to reduce travel. This is not only driven by cost awareness, but also by consideration of sustainability, natural resources and the goal to reduce CO2 emissions. Face-to-face meetings are a basic need, but not all meetings need to be face-to-face. In line with the growing need, there is a very fast deployment evolution of visual communication. We strongly believe that high quality visual communication will very soon become ubiquitous, almost as ubiquitous as voice communication is today.

This places a number of challenges. Visual communication will need to handle a high heterogeneity in terms of devices, accesses, and networking technologies. Devices could be mobile phones with small screens, fixed phones with larger screens, and conference rooms with multiple HD screens. Multi-channel and multi-screen installations and even 3D video could be used. Asymmetric access technologies like DSL, HSPA, or LTE pose special challenges. Interoperability between devices from different vendors, which is not given today, will be a prerequisite. Also, new functionality will be incorporated. Face recognition can be used to link people to information and context. Advanced computer graphics and scene composition will allow placing physically separated people in one virtual room.

5 Conclusion

High quality visual communication services, such as high-definition videoconferencing and telepresence, can be realized using the IMS core network and hence using standardized functions, procedures, and interfaces. We explained the basic IMS architecture and the most important functions needed to deploy high quality videoconferencing. We discussed videoconferencing services in general and scheduled dial-in conferencing especially.

High quality visual communication services use HD video and full-band audio. A high-end media plane and high user experience are realized using multiple media streams and state-of-the-art codecs as G.719 and H.264. The media plane uses either a central media server with transcoding and bitrate adaptation or, simulcast transmission without central transcoding.

High quality visual communication based on IMS bridges the current gap between cheap low quality PC based video telephony and expensive high-end telepresence systems. Using standard consumer hardware, this has the potential of enabling mass-market high quality visual communication.

Most likely, in a few years from now, the times when high quality visual communication was not a part of everyday life will be hard to imagine.

6 Literature

[1] Axelsson, Michael: Visual communication goes mass market. Ericsson Business Review: no. 01, 2011

[2] Camarillo, Gonzalo: The 3G IP Multimedia Subsystem. John Wiley & Sons, 2nd Edition 2006

[3] Mecklin, Opsenica, Rissanen, Valderas: ImsInnovation - Experiences of an IMS Testbed. ONIT 2009

[4] ITU-T H.264: Advanced video coding for generic audiovisual services

[5] ITU-T G.719: Low-complexity, full-band audio coding for high-quality, conversational applications

[6] ITU-T G.722.2: Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB)

[7] 3GPP TS 23.228: IP Multimedia Subsystem; Stage 2

[8] 3GPP TS 26.114: IP Multimedia Subsystem (IMS); Multimedia Telephony; Media handling and interaction

[9] 3GPP TS 24.147: Conferencing using the IP Multimedia (IM) Core Network (CN) subsystem; Stage 3

[10] IETF RFC 4353: A Framework for Conferencing with the Session Initiation Protocol (SIP)

[11] IETF RFC 4579: Session Initiation Protocol (SIP) Call Control - Conferencing for User Agents

[12] IETF RFC 5022: Media Server Control Markup Language (MSCML) and Protocol

[13] IETF RFC 4575: A Session Initiation Protocol (SIP) Event Package for Conference State

[14] IETF RFC 3551: RTP Profile for Audio and Video Conferences with Minimal Control

[15] IETF RFC 4585: Extended RTP Profile for Real-time Transport Control Protocol (RTCP)-Based Feedback (RTP/AVPF)

[16] IETF RFC 5104: Codec Control Messages in the RTP Audio-Visual Profile with Feedback (AVPF)