ETSI TR 126 980 V17.0.0 (2022-05)

TECHNICAL REPORT

Universal Mobile Telecommunications System (UMTS);
LTE;
5G;
Multimedia telephony over IP Multimedia Subsystem (IMS);
Media handling aspects of multi-stream multiparty conferencing for Multimedia Telephony Service for IMS (MTSI)
(3GPP TR 26.980 version 17.0.0 Release 17)

Reference

RTR/TSGS-0426980vh00

Keywords

5G,LTE,UMTS

ETSI

650 Route des Lucioles
F-06921 Sophia Antipolis Cedex - FRANCE

Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - APE 7112B
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° w061004871

Intellectual Property Rights

Essential patents

IPRs essential or potentially essential to normative deliverables may have been declared to ETSI. The declarations pertaining to these essential IPRs, if any, are publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI Web server (https://ipr.etsi.org/).

Pursuant to the ETSI Directives including the ETSI IPR Policy, no investigation regarding the essentiality of IPRs, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

Trademarks

The present document may include trademarks and/or tradenames which are asserted and/or registered by their owners. ETSI claims no ownership of these except for any which are indicated as being the property of ETSI, and conveys no right to use or reproduce any trademark and/or tradename. Mention of those trademarks in the present document does not constitute an endorsement by ETSI of products, services or organizations associated with those trademarks.

DECT™, **PLUGTESTS™**, **UMTS™** and the ETSI logo are trademarks of ETSI registered for the benefit of its Members. **3GPP™** and **LTE™** are trademarks of ETSI registered for the benefit of its Members and of the 3GPP Organizational Partners. **oneM2M™** logo is a trademark of ETSI registered for the benefit of its Members and of the oneM2M Partners. **GSM**® and the GSM logo are trademarks registered and owned by the GSM Association.

Legal Notice

This Technical Report (TR) has been produced by ETSI 3rd Generation Partnership Project (3GPP).

The present document may refer to technical specifications or reports using their 3GPP identities. These shall be interpreted as being references to the corresponding ETSI deliverables.

The cross reference between 3GPP and ETSI identities can be found under http://webapp.etsi.org/key/queryform.asp.

Modal verbs terminology

In the present document "**should**", "**should not**", "**may**", "**need not**", "**will**", "**will not**", "**can**" and "**cannot**" are to be interpreted as described in clause 3.2 of the ETSI Drafting Rules (Verbal forms for the expression of provisions).

"**must**" and "**must not**" are **NOT** allowed in ETSI deliverables except when used in direct citation.

Contents

Foreword

This Technical Report has been produced by the 3rd Generation Partnership Project (3GPP).

The contents of the present document are subject to continuing work within the TSG and may change following formal TSG approval. Should the TSG modify the contents of the present document, it will be re-released by the TSG with an identifying change of release date and an increase in version number as follows:

Version x.y.z

where:

- x the first digit:
  - 1 presented to TSG for information;
  - 2 presented to TSG for approval;
  - 3 or greater indicates TSG approved document under change control.
- y the second digit is incremented for all changes of substance, i.e. technical enhancements, corrections, updates, etc.
- z the third digit is incremented when editorial only changes have been incorporated in the document.

Introduction

Media handling of Multimedia Telephony Service over IMS, MTSI, are based on 3GPP SA4 TS 26.114 [1]. MTSI media handling is referred to by GSMA in GSMA PRD IR.92 [21] also known as VoLTE and GSMA PRD IR.94 [22] also known as video over LTE. MTSI clients can connect to conferencing IMS communication services. 3GPP conducted a study as part of a work item on Multi-stream Multiparty Conferencing Media-Handling for MTSI. The work objective is to specify an increment to MTSI client media-handling specification TS 26.114 to enable a mass-market multiparty communication service with excellent multiparty user experience and media quality. Such Operator communication service evolution would match proprietary communication services in quality with excellent efficiency and device reach.

The present document captures the study inputs, discussions and findings of the Multi-stream Multiparty Conference Media Handling (MMCMH) work item. It describes a set of use cases with corresponding problem descriptions and potential solutions. The conclusion gives recommendations as to what needs to be normatively specified to achieve the MMCMH work item objectives.

1 Scope

The Technical Report provides a study on the media handling aspects of Multi-stream Multiparty Conferencing for MTSI. The study focuses on enabling

- support to receive multi-stream audio/video at the terminals in a multiparty conferencing,
- support for at least two video contents, e.g. one main and one presentation,
- talker ID provisioning,
- compatibility with MTSI TS 26.114 and GSMA IR.94 (Video over LTE) [22] and GSMA IR.92 (VoLTE) [21], and
- applicability to both mobile and fixed access terminals

High-level use cases along with current limitations and recommendations are documented in the present document.

2 References

The following documents contain provisions which, through reference in this text, constitute provisions of the present document.

- References are either specific (identified by date of publication, edition number, version number, etc.) or non-specific.
- For a specific reference, subsequent revisions do not apply.
- For a non-specific reference, the latest version applies. In the case of a reference to a 3GPP document (including a GSM document), a non-specific reference implicitly refers to the latest version of that document *in the same Release as the present document*.

[1] 3GPP TS 26.114: "IP Multimedia Subsystem (IMS); Multimedia Telephony, Media handling and interaction".

[2] Void.

[3] 3GPP TS 24.147: "Conferencing Using IP Multimedia Core Network; Stage 3".

[4] 3GPP TS 24.605: "CONF Using IP Multimedia Core Network".

[5] 3GPP TS 22.228: "Service Requirements for IP Multimedia Core Network; Stage 1".

[6] 3GPP TS 23.218:"IP Multimedia Session Handling; IP Multimedia Call Model; Stage 2".

[7] 3GPP TS 24.228: "Signalling Flows for IP Multimedia Call Control Based on SIP and SDP; Stage 3".

[8] 3GPP TS 24.229: "IP Multimedia Call Control Protocol Based on SIP and SDP; Stage 3".

[9] 3GPP TR 22.948: "IP Multimedia Subsystem Convergent Multimedia Conferencing".

[10] 3GPP TR 29.847: "SIP Conferencing Models, Flows and Protocols".

[11] GSM Association: "WebRTC Codecs DRAFT v1.3", September 2014, http://www.gsma.com/newsroom/wp-content/uploads/WebRTC-Whitepaper-v13.pdf.

[12] IETF Internet Draft: "Using Simulcast in SDP and RTP Sessions", June 2016, https://datatracker.ietf.org/doc/draft-ietf-mmusic-sdp-simulcast/ (WORK IN PROGRESS).

[13] IETF RFC 4796: "The Session Description Protocol (SDP) Content Attribute", February 2007, https://datatracker.ietf.org/doc/rfc4796/.

[14] GSM Association PRD IR.39: "IMS Profile for High Definition Video Conferencing (HDVC) v5.0", July 2014, http://www.gsma.com/newsroom/wp-content/uploads//IR.39-v5.0.pdf.

[15] IETF RFC 4582: "The Binary Floor Control Protocol (BFCP)", November 2006, https://datatracker.ietf.org/doc/rfc4582/.

NOTE A new version of this RFC is in the process of being published. See [17].

[16] IETF RFC 4583: "Session Description Protocol (SDP) Format for Binary Floor Control Protocol (BFCP) Streams", November 2006, https://datatracker.ietf.org/doc/rfc4583/.

NOTE: A new version of this RFC is in the process of being published. See [18].

[17] IETF Internet Draft, draft-ietf-bfcpbis-rfc4582bis-16: "The Binary Floor Control Protocol (BFCP), November 2015, WORK IN PROGRESS, https://datatracker.ietf.org/doc/draft-ietf-bfcpbis-rfc4582bis/.

[18] IETF Internet Draft, draft-ietf-bfcpbis-rfc4583bis-15: "Session Description Protocol (SDP) Format for Binary Floor Control Protocol (BFCP) Streams", July 2016, WORK IN PROGRESS, https://datatracker.ietf.org/doc/draft-ietf-bfcpbis-rfc4583bis/.

[19] IETF RFC 3550: "RTP: A Transport Protocol for Real-Time Applications", July 2003, https://datatracker.ietf.org/doc/rfc3550/.

[20] IETF RFC 5104: "Codec Control Messages in the RTP Audio-Visual Profile with Feedback (AVPF)", February 2008, https://datatracker.ietf.org/doc/rfc5104/.

[21] GSM Association PRD IR.92, "IR.92 IMS Profile for Voice and SMS", v9.0, April 2015, http://www.gsma.com/newsroom/wp-content/uploads//IR.92-v9.0.pdf.

[22] GSM Association PRD IR.94: "IR.94 IMS Profile for Conversational Video Service", v8.0.1, November 2014, http://www.gsma.com/newsroom/wp-content/uploads//IR.94-v8.01.pdf.

[23] IETF RFC 7728, "RTP Stream Pause and Resume", March 2016.

[24] 3GPP TS 24.173: "IMS Multimedia Telephony Communication Service and Supplementary Services".

[25] IETF RFC 4353: "A Framework for Conferencing with the Session Initiation Protocol (SIP)", February 2006, https://datatracker.ietf.org/doc/rfc4353/.

[26] IETF RFC 3515: "The Session Initiation Protocol (SIP) Refer Method", April 2003, https://datatracker.ietf.org/doc/rfc3515/.

[27] IETF RFC 5939: "Session Description Protocol (SDP) Capability Negotiation", Sept. 2010.

[28] IETF RFC 6184: "RTP Payload Format for H.264 Video", May 2011, https://www.ietf.org/rfc/rfc6184.txt.

[29] IETF RFC 7798, "RTP Payload Format for High Efficiency Video Coding (HEVC)", March 2016.

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

[31] IETF RFC 3261: SIP: "Session Initiation Protocol", June 2002.

[32] 3GPP TR 21.905: "Vocabulary for 3GPP Specifications".

[33] Void.

[34] Void.

[35] Void.

[36] IETF RFC 4575: "A Session Initiation Protocol (SIP) Event Package for Conference State", August 2006.

[37] Void.

[38] Void.

[39] Void.

[40] ITU-T Recommendation H.265 (04/2015): "High efficiency video coding".

[41] 3GPP TS 26.171: "Speech codec speech processing functions; Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; General description".

[42] 3GPP TS 26.441: "Codec for Enhanced Voice Services (EVS); General Overview".

[43] IETF RFC 3264: "An Offer/Answer Model with the Session Description Protocol (SDP)", J. Rosenberg and H. Schulzrinne, July 2002.

[44] IETF Internet Draft, draft-ietf-mmusic-rid-07: "RTP Payload Format Constraints", July 2016 (WORK IN PROGRESS), https://datatracker.ietf.org/doc/draft-ietf-mmusic-rid/.

[45] IETF RFC 7656: "A Taxonomy of Semantics and Mechanisms for Real-Time Transport Protocol (RTP) Sources", J. Lennox, K. Gross, S. Nandakumar, G. Salguiero, and B. Burman, November 2015.

3 Definitions and abbreviations

3.1 Definitions

For the purposes of the present document, the terms and definitions given in TR 21.905 [32] and the following apply. A term defined in the present document takes precedence over the definition of the same term, if any, in TR 21.905 [32].

Mode-set: Used for the AMR and AMR-WB codecs to identify the codec modes that can be used in a session. A mode-set can include one or more codec modes.

MSMTSI client: A multi-stream capable MTSI client supporting multiple streams. An MTSI client may support multiple streams, even of the same media type, without being an MSMTSI client. Such an MTSI client may, for example, add a second video to an ongoing video telephony session as shown in TS 26.114 Annex A.11.

MSMTSI MRF: An MSMTSI client implemented by functionality included in the MRFC and the MRFP.

MSMTSI client in terminal: An MSMTSI client that is implemented in a terminal or UE. The term "MSMTSI client in terminal" is used in the present document when entities such as MRFP, MRFC or media gateways are excluded.

MTSI client: A function in a terminal or in a network entity (e.g. a MRFP) that supports MTSI.

MTSI client in terminal: An MTSI client that is implemented in a terminal or UE. The term "MTSI client in terminal" is used in the present document when entities such as MRFP, MRFC or media gateways are excluded.

MTSI media gateway (or MTSI MGW): A media gateway that provides interworking between an MTSI client and a non MTSI client, e.g. a CS UE. The term MTSI media gateway is used in a broad sense, as it is outside the scope of the current specification to make the distinction whether certain functionality should be implemented in the MGW or in the MGCF.

Operational mode: Used for the EVS codec to distinguish between EVS Primary mode and EVS AMR-WB IO mode.

Simulcast: Simultaneously sending different encoded representations (simulcast formats) of a single media source (e.g. originating from a single microphone or camera) in different simulcast streams.

Simulcast format: The encoded format used by a single simulcast stream, typically represented by an SDP format and all SDP attributes that apply to that particular SDP format, indicated in RTP by the RTP header payload type field.

Simulcast stream: The RTP stream carrying a single simulcast format in a simulcast.

3.2 Abbreviations

For the purposes of the present document, the abbreviations given in TR 21.905 [32] and the following apply.
An abbreviation defined in the present document takes precedence over the definition of the same abbreviation, if any, in TR 21.905 [32].

AVC Advanced Video Coding
BFCP Binary Floor Control Protocol
CCM Codec Control Messages
LTE Long Term Evolution
MSRP Message Session Relay Protocol
MSMTSI Multi-Stream Multimedia Telephony Service for IMS
MTSI Multimedia Telephony Service for IMS
SDP Session Description Protocol

4 Overview

Clause 5 provides a high-level description of the media handling in current 3GPP conferencing. The rest of the document is organized as follows. clause 6 describes the use cases analysed in this study. Clause 7 provides the conclusion and recommendations for further standardization efforts. Annex A includes some SDP offer/answer examples.

5 Media Handling in Current 3GPP Conferencing

The current 3GPP specifications mentioning conferencing or group communication is mainly focusing on (SIP) signalling aspects, and there is very little on media handling aspects. Those specifications include (list not intended to be exhaustive) [3], [4], [5], [6], [7], [8], [9] and [10].

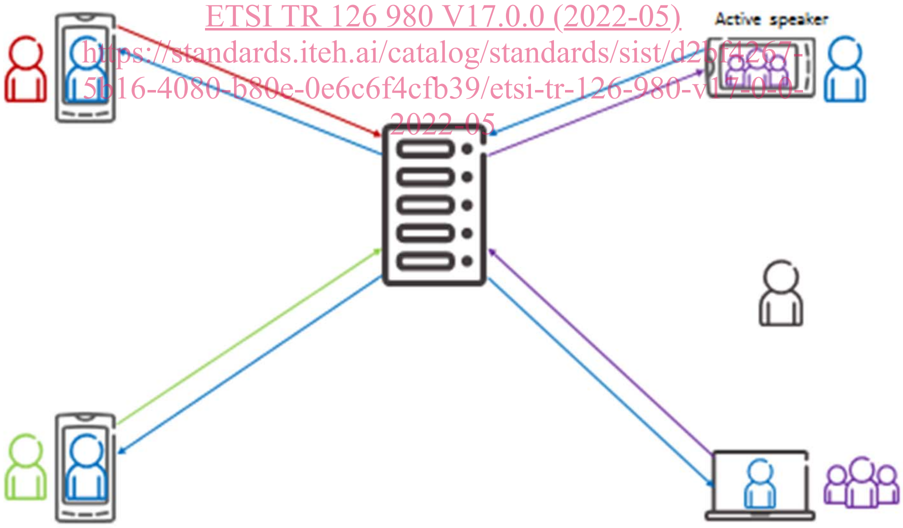

Figure 1. Existing Conference Architecture Example.

This briefly summarizes a few things regarding IMS conferences that are already specified, and that have an impact on media handling:

- A centralized conference with the MRFP as conference focus is assumed [3], where media handling is not explicitly described:
  - MRFP is assumed to be an RTP "mixer" in IETF RFC 3550 [19] sense:

    - One possible implicit assumption is that the conference focus always transcodes (decodes, mixes, and re-encodes) media individually towards every participant.

    - Another possibility is to switch the video RTP stream untouched from one participant to another, and possibly to all other participants.

- It is not described which video stream the MRFP should distribute to the different participants:

  - One possible and reasonable assumption is that the video from some "active speaker" is distributed to other participants, which would require some "active speaker" decision in the MRFP that in turn could be based on speech activity analysis of the audio streams from every participant.

  - If "active speaker" is distributed, it is common on the market to not distribute media from that "active speaker" to itself, but rather the previous "active speaker" (as depicted in Figure 1).

  - Another possible assumption is that all, or at least most, participant videos are re-sized, composed, and transcoded into a checkerboard layout.

  - A third possible assumption is that some type of floor control, e.g. based on Binary Floor Control Protocol (BFCP) [3] and [15], is used, where the usage details in that case are so far left unspecified.

  - When MRFP is not transcoding, when changing from forwarding one participant's video to another participant's video, and since encoded video typically makes use of temporal redundancy, this change can only be made at a point in the video stream that does not depend on any previous part of that video stream – a so called "intra" picture. When deciding to make a change of forwarded video, the MRFP can trigger the UE to send such intra picture by issuing an RTCP CCM FIR command to the UE, as described in RFC 5104, and make the actual switch only when that intra picture arrives to the MRFP. Timing, reliability and bandwidth aspects of FIR transmission are described in RFC 5104. A MTSI UE is already required to support and react on FIR. When changing to a new source and if the new source is inactive (on hold or not established) then that stream has to be activated before the MRFP can switch to it.

- SIP conference call control includes three allowed options [3] and [8]:

  - Each participating UE calls in to conference (SIP INVITE).

  - The originating UE calls in to conference and requests it to call out to other participants (SIP INVITE with recipient list).

  - A UE has an ongoing point-to-point or three-party call that is moved into a conference (SIP REFER).

- MRFC always includes "isFocus" tag in its SIP signalling [3] (regardless if it is a SIP request or response), which lets the UE know that it is signalling with a conference and not another UE.

- The conference may optionally make use of explicit floor control through Binary Floor Control Protocol (BFCP) [3] and [15]:

  - The use of a floor control protocol allows explicitly, and even manually, controlling which participant's video is distributed to others by the MRFP.

  - The use of this "application" media stream is negotiated through SDP [16].

  - TCP transport of BFCP is assumed, possibly because this was until fairly recently the only specified transport in IETF, but many BFCP implementations on the market instead use UDP in a straightforward way, and there is well progressed work in IETF to describe this in an update to the BFCP RFC [17].

The MMCMH work item objectives include enabling multi-stream audio/video support at the terminals. In addition, as specified in the MMCMH WI objectives, the conference focus and the terminals may receive stereo streams for further processing and rendering. The clauses below present the multi-stream audio and video use cases, where the terminals receive and decode the multiple streams of audio/video and thumbnails, and render them at the terminal that is potentially transcoder-free. Conferencing using IP multimedia core network and with conference focus mixing (e.g. with MRFP) are addressed in 3GPP TS 24.147 and IETF RFC 4353.

6 Use cases

6.1 Overview

This clause contains multimedia group communication use cases that enables multi-stream video and audio support at the terminals.

6.2 Use case A: Transcoding Free Continuous Presence

When calling in to a group video call, the user is able to see video from more than a single one of the other participants in the call, which is commonly referred to as "continuous presence". This is typically desirable in a group communication for a user to be able to see the reactions of more than a single participant.

In contrast, when receiving video from one participant at a time, several different approaches to choose that single participant are possible, subject to implementation in the conference focus. It may be that the active speaker is chosen, based on conference focus analysis of some (unspecified) voice activity measure of all participants. It may be based on a chair person's explicit and manual control of the conference focus, typically requiring a floor control protocol, such as for example BFCP, [15]. It can also be based on other approaches, such as for example an automatic, timed round-robin among all participants.

The participant layout or the number of participants simultaneously visible in a continuous presence layout is neither specified nor specifically restricted in this use case. An implementation will however always be limited, either by the receiving UE capability, or by group call network resources. Examples of receiving UE capability limitations are available display size, and decoding resources. Examples of network resource limitations are network bandwidth and conference focus processing capacity. Different UE implementations may typically have different amounts of limitation, and a solution could possibly accommodate that by letting the conference focus adapt the layout to individual UE.

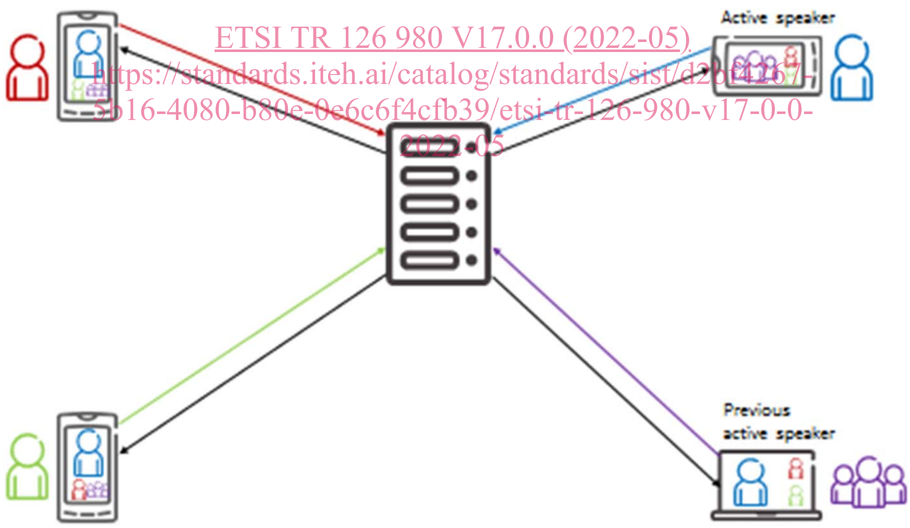

Figure 2. Continuous presence example.

If the chosen layout is able to fit all participants in the group communication, no further action has to be taken. However, when the number of group participants is larger than the number of participants in the chosen layout, those participants have to be selected somehow, just as for the single video layout described above. The selection of participants to include in such continuous presence layout can be based on the same principles as for the single video case. For example, if a voice activity approach is used, the N currently most active speakers can be chosen. In that case, it is often desirable that the current active speaker is highlighted in some way, for example by using a larger video image size.

Typically, a separate composed video layout has to be created for each receiver, since it is often not desirable to show the receiving user as part of such composition. Specifically, the currently active speaker should also be shown something else than itself in active speaker position, for example the previously active speaker. If it is desirable to show a self-view video, this is much more efficient to solve locally in the sending UE, since that video then neither has to occupy any composition resources in the conference focus nor any downlink bandwidth. A video layout is possible to re-use in group communications where the total number of participants is at least two more than the number of participants included in the video layout.

The receiving user should ideally be able to impact the received layout, but it may also be decided by some policy implemented in the UE application, in the conference focus, or some combination.

6.2.1 Problem Description

Assuming that a UE can receive only a single video stream, creation of the composed "continuous presence" picture has to happen elsewhere, typically in the conference focus media handling part. Such composition requires decoding of video from all of the participants to be composed, re-sizing them to fit the layout, composing the layout in the decoded pixel domain, and re-encoding the resulting video. This transcoding operation introduces increased end-to-end delay and decreases video quality, similar to what is described in [11] (although that document focuses on transcoding between different video codecs). It also requires a significant amount of transcoding and video composition resources in the conference focus, per group video communication participant.

To summarize, the problem with this approach to continuous presence is threefold:

1. Increased end-to-end delay
2. Decreased video quality
3. Increased amount of resources in the conference focus

The increased bandwidth with respect to multi-stream vs. transcoding is considered in a separate use case in clause 6.4.

6.2.2 Proposed Solution

The suggested solution is instead using local video composition of decoded video in the receiving UE, meaning that it receives and independently decodes all of the video streams to be used for composition. The conference focus is then neither composing any continuous presence image nor transcoding it, but just forwarding video streams from the sending participants to appropriate receivers.

The composition can be part of the normal video display process and does not introduce any noticeable extra video delay. In addition, the video composition process is also under full control of the receiving UE, and leaves significant freedom to the local graphical user interface (GUI) to layout the different videos, and can easily (but optionally) allow the user of the receiving UE to impact such layout, without or with minimal changes to the conference focus or received video streams.

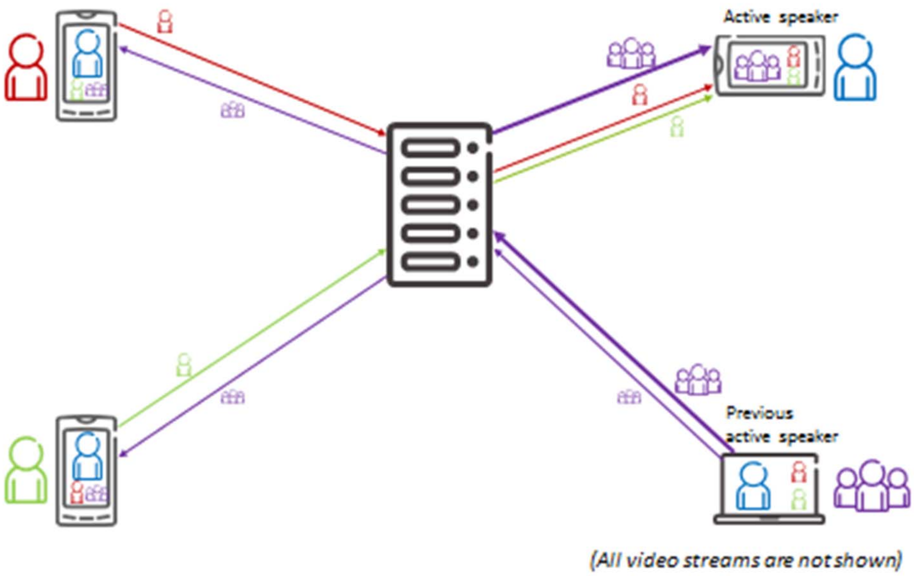

Figure 3. Transcoding-free multi-stream continuous presence example.

If the receiving UE is able to express capability for maximum number of received video streams and also their corresponding "sizes", it is fairly easy to use that information to construct meaningful local video layouts. For example, assuming that the conference focus uses voice activity detection for the group communication participants, and further assuming it is able to send videos for the N most active speakers, where the current active speaker is provided in normal resolution while the rest (N-1) is provided in low resolution ("thumbnails"). A conference focus having this information per connected UE can easily choose which, and which number of participant videos to forward to a specific UE.

To support that the active speaker is shown in normal size on receiving UE while thumbnails are smaller, the conference focus needs that active speaker to send a normal size video, while the others being shown as thumbnails may send smaller sized videos. To accommodate the previous active speaker to be shown in normal size as active speaker to the current active speaker (instead of itself), the previous active speaker may have to send both a normal sized video (to be forwarded by conference focus to current active speaker), and a small sized video (to be forwarded as thumbnail to other participants). This way of sending multiple simultaneous representations of the same content is called "simulcast" [12].

This can be accomplished in SDP by letting the active speaker be described by the already present video m-line, and add a set of additional video m-lines (number can be decided by UE capability) to describe the additional (possibly thumbnail) videos. The advantages with this approach are that the number of and details for each additional thumbnail can be negotiated (and also rejected) independently. This approach is also fully in line with existing SDP semantics, where additional m-lines describe media that are sent in addition to and simultaneously with other m-lines (like, for example, the current audio and video m-lines).

When it comes to simulcast (see above), this is slightly different, and there is ongoing work in IETF on this issue. Different (and simultaneous) representations of the same video source should be described by a single m-line (see details in [12]). It is of course possible to express capability for and negotiate the use of simulcast.

Note that the way to implement multi-stream in this scenario does not require any specific video codec type. Any video codec type can be used, as long as the sending and receiving UE use compatible video codecs, described and negotiated by SDP, for example the mandatory TS 26.114 video codec H.264.

6.3 Use case B: Screen Sharing

In a group video communication, it is sometimes desirable for a user to show something else than the video from the camera to the other participants, like a document, image or something else that can be shown on the user's local screen.

6.3.1 Problem Description

The basic problem is that there is no commonly accepted interchange format to transfer screen content between peers, although several proprietary formats do exist. It is in principle possible to send screen content as regular video, if