Internet Engineering Task Force (IETF)
Request for Comments: 7262
Category: Informational
ISSN: 2070-1721

A. Romanow
Cisco Systems
S. Botzko
Polycom
M. Barnes
MLB@Realtime Communications, LLC

June 2014

Requirements for Telepresence Multistreams

Abstract

This memo discusses the requirements for specifications that enable telepresence interoperability by describing behaviors and protocols for Controlling Multiple Streams for Telepresence (CLUE). In addition, the problem statement and related definitions are also covered herein.

Status of This Memo

This document is not an Internet Standards Track specification; it is published for informational purposes.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Not all documents approved by the IESG are a candidate for any level of Internet Standard; see Section 2 of RFC 5741.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc7262.

Table of Contents

1. Introduction

Telepresence systems greatly improve collaboration. In a telepresence conference (as used herein), the goal is to create an environment that gives the users a feeling of (co-located) presence -- the feeling that a local user is in the same room with other local users and remote parties. Currently, systems from different vendors often do not interoperate because they do the same tasks differently, as discussed in the Problem Statement section below (see Section 4).

The approach taken in this memo is to set requirements for a future specification(s) that, when fulfilled by an implementation of the specification(s), provide for interoperability between IETF protocol-based telepresence systems. It is anticipated that a solution for the requirements set out in this memo likely involves the exchange of adequate information about participating sites; this information that is currently not standardized by the IETF.

The purpose of this document is to describe the requirements for a specification that enables interworking between different SIP-based [RFC3261] telepresence systems, by exchanging and negotiating appropriate information. In the context of the requirements in this

document and related solution documents, this includes both point-to-point SIP sessions as well as SIP-based conferences as described in the SIP conferencing framework [RFC4353] and the SIP-based conference control [RFC4579] specifications. Non-IETF protocol-based systems, such as those based on ITU-T Rec. H.323 [ITU.H323], are out of scope. These requirements are for the specification, they are not requirements on the telepresence systems implementing the solution/protocol that will be specified.

Today, telepresence systems of different vendors can follow radically different architectural approaches while offering a similar user experience. CLUE will not dictate telepresence architectural and implementation choices; however, it will describe a protocol architecture for CLUE and how it relates to other protocols. CLUE enables interoperability between telepresence systems by exchanging information about the systems' characteristics. Systems can use this information to control their behavior to allow for interoperability between those systems.

A telepresence session requires at least one sending and one receiving endpoint. Multiparty telepresence sessions include more than 2 endpoints and centralized infrastructure such as Multipoint Control Units (MCUs) or equivalent. CLUE specifies the syntax, semantics, and control flow of information to enable the best possible user experience at those endpoints.

Sending endpoints, or MCUs, are not mandated to use any of the CLUE specifications that describe their capabilities, attributes, or behavior. Similarly, it is not envisioned that endpoints or MCUs will ever have to take information received into account. However, by making available as much information as possible, and by taking into account as much information as has been received or exchanged, MCUs and endpoints are expected to select operation modes that enable the best possible user experience under their constraints.

The document structure is as follows: definitions are set out, followed by a description of the problem of telepresence interoperability that led to this work. Then the requirements for a specification addressing the current shortcomings are enumerated and discussed.

2. Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

3. Definitions

The following terms are used throughout this document and serve as a reference for other documents.

Audio Mixing: refers to the accumulation of scaled audio signals to produce a single audio stream. See "RTP Topologies" [RFC5117].

Conference: used as defined in "A Framework for Conferencing within the Session Initiation Protocol (SIP)" [RFC4353].

Endpoint: The logical point of final termination through receiving, decoding and rendering, and/or initiation through capturing, encoding, and sending of media streams. An endpoint consists of one or more physical devices that source and sink media streams, and exactly one participant [RFC4353] (which, in turn, includes exactly one SIP user agent). In contrast to an endpoint, an MCU may also send and receive media streams, but it is not the initiator or the final terminator in the sense that media is captured or rendered. Endpoints can be anything from multiscreen/multicamera rooms to handheld devices.

Endpoint Characteristics: include placement of capture and rendering devices, capture/render angle, resolution of cameras and screens, spatial location, and mixing parameters of microphones. Endpoint characteristics are not specific to individual media streams sent by the endpoint.

Layout: How rendered media streams are spatially arranged with respect to each other on a telepresence endpoint with a single screen and a single loudspeaker, and how rendered media streams are arranged with respect to each other on a telepresence endpoint with multiple screens or loudspeakers. Note that audio as well as video are encompassed by the term layout -- in other words, included is the placement of audio streams on loudspeakers as well as video streams on video screens.

Local: Sender and/or receiver physically co-located ("local") in the context of the discussion.

MCU: Multipoint Control Unit (MCU) - a device that connects two or more endpoints together into one single multimedia conference [RFC5117]. An MCU may include a mixer [RFC4353].

Media: Any data that, after suitable encoding, can be conveyed over RTP, including audio, video, or timed text.

Model: a set of assumptions a telepresence system of a given vendor adheres to and expects the remote telepresence system(s) to also adhere to.

Remote: Sender and/or receiver on the other side of the communication channel (depending on context); i.e., not local. A remote can be an endpoint or an MCU.

Render: the process of generating a representation from a media, such as displayed motion video or sound emitted from loudspeakers.

Telepresence: an environment that gives non-co-located users or user groups a feeling of (co-located) presence -- the feeling that a local user is in the same room with other local users and the remote parties. The inclusion of Remote parties is achieved through multimedia communication including at least audio and video signals of high fidelity.

4. Problem Statement

In order to create a "being there" experience characteristic of telepresence, media inputs need to be transported, received, and coordinated between participating systems. Different telepresence systems take diverse approaches in crafting a solution, or they implement similar solutions quite differently.

They use disparate techniques, and they describe, control and negotiate media in dissimilar fashions. Such diversity creates an interoperability problem. The same issues are solved in different ways by different systems, so that they are not directly interoperable. This makes interworking difficult at best and sometimes impossible.

Worse, even if those extensions are based on common standards such as SIP, many telepresence systems use proprietary protocol extensions to solve telepresence-related problems.

Some degree of interworking between systems from different vendors is possible through transcoding and translation. This requires additional devices, which are expensive, are often not entirely automatic, and sometimes introduce unwelcome side effects, such as additional delay or degraded performance. Specialized knowledge is currently required to operate a telepresence conference with endpoints from different vendors, for example to configure transcoding and translating devices. Often such conferences do not start as planned or are interrupted by difficulties that arise.

The general problem that needs to be solved can be described as follows. Today, each endpoint renders the audio and video captures it receives according to an implicitly assumed model that stipulates how to produce a realistic depiction of the remote location. If all endpoints are manufactured by the same vendor, they all share the same implicit model and render the received captures correctly. However, if the devices are from different vendors, the models used for rendering presence can and usually do differ. The result can be that the telepresence systems actually connect, but the user experience will suffer, for example one system assumes that the first video stream is captured from the right camera, whereas the other assumes the first video stream is captured from the left camera.

If Alice and Bob are at different sites, Alice needs to tell Bob about the camera and sound equipment arrangement at her site so that Bob's receiver can create an accurate rendering of her site. Alice and Bob need to agree on what the salient characteristics are as well as how to represent and communicate them. Characteristics may include number, placement, capture/render angle, resolution of cameras and screens, spatial location, and audio mixing parameters of microphones.

The telepresence multistream work seeks to describe the sender situation in a way that allows the receiver to render it realistically even though it may have a different rendering model than the sender.

5. Requirements

Although some aspects of these requirements can be met by existing technology, such as the Session Description Protocol (SDP) [RFC4566], they are stated here to have a complete record of the requirements for CLUE. Determining whether a requirement needs new work or not will be part of the solution development, and is not discussed in this document. Note that the term "solution" is used in these requirements to mean the protocol specifications, including extensions to existing protocols as well as any new protocols, developed to support the use cases. The solution might introduce additional functionality that is not mapped directly to these requirements; e.g., the detailed information carried in the signaling protocol(s). In cases where the requirements are directly relevant to specific use cases as described in [RFC7205], a reference to the use case is provided.

REQ-1: The solution MUST support a description of the spatial arrangement of source video images sent in video streams that enables a satisfactory reproduction at the receiver of the original scene. This applies to each site in a point-to-point or a multipoint meeting and refers to the spatial ordering within a site, not to the ordering of images between sites.

This requirement relates to all the use cases described in [RFC7205].

REQ-1a: The solution MUST support a means of allowing the preservation of the order of images in the captured scene. For example, if John is to Susan’s right in the image capture, John is also to Susan’s right in the rendered image.

REQ-1b: The solution MUST support a means of allowing the preservation of order of images in the scene in two dimensions - horizontal and vertical.

REQ-1c: The solution MUST support a means to identify the relative location, within a scene, of the point of capture of individual video captures in three dimensions.

REQ-1d: The solution MUST support a means to identify the area of coverage, within a scene, of individual video captures in three dimensions.

REQ-2: The solution MUST support a description of the spatial arrangement of captured source audio sent in audio streams that enables a satisfactory reproduction at the receiver in a spatially correct manner. This applies to each site in a point to point or a multipoint meeting and refers to the spatial ordering within a site, not the ordering of channels between sites.

This requirement relates to all the use cases described in [RFC7205], but is particularly important in the Heterogeneous Systems use case.

REQ-2a: The solution MUST support a means of preserving the spatial order of audio in the captured scene. For example, if John sounds as if he is on Susan’s right in the captured audio, John voice is also placed on Susan’s right in the rendered image.

REQ-2b: The solution MUST support a means to identify the number and spatial arrangement of audio channels including monaural, stereophonic (2.0), and 3.0 (left, center, right) audio channels.

REQ-2c: The solution MUST support a means to identify the point of capture of individual audio captures in three dimensions.

REQ-2d: The solution MUST support a means to identify the area of coverage of individual audio captures in three dimensions.

REQ-3: The solution MUST enable individual audio streams to be associated with one or more video image captures, and individual video image captures to be associated with one or more audio captures, for the purpose of rendering proper position.

This requirement relates to all the use cases described in [RFC7205].

REQ-4: The solution MUST enable interoperability between endpoints that have a different number of similar devices. For example, an endpoint may have 1 screen, 1 loudspeaker, 1 camera, 1 mic, and another endpoint may have 3 screens, 2 loudspeakers, 3 cameras and 2 microphones. Or, in a multipoint conference, an endpoint may have 1 screen, another may have 2 screens, and a third may have 3 screens. This includes endpoints where the number of devices of a given type is zero.

This requirement relates to the Point-to-Point Meeting: Symmetric and Multipoint Meeting use cases described in [RFC7205].

REQ-5: The solution MUST support means of enabling interoperability between telepresence endpoints where cameras are of different picture aspect ratios.

REQ-6: The solution MUST provide scaling information that enables rendering of a video image at the actual size of the captured scene.

REQ-7: The solution MUST support means of enabling interoperability between telepresence endpoints where displays are of different resolutions.

REQ-8: The solution MUST support methods for handling different bit rates in the same conference.

REQ-9: The solution MUST support means of enabling interoperability between endpoints that send and receive different numbers of media streams.

This requirement relates to the Heterogeneous Systems and Multipoint Meeting use cases.

REQ-10: The solution MUST ensure that endpoints that support telepresence extensions can establish a session with a SIP endpoint that does not support the telepresence extensions. For example, in the case of a SIP endpoint that supports a single audio and a single video stream, an endpoint that supports the telepresence extensions would setup a session with a single audio and single video stream using existing SIP and SDP mechanisms.

REQ-11: The solution MUST support a mechanism for determining whether or not an endpoint or MCU is capable of telepresence extensions.

REQ-12: The solution MUST support a means to enable more than two endpoints to participate in a teleconference.

This requirement relates to the Multipoint Meeting use case.

REQ-13: The solution MUST support both transcoding and switching approaches for providing multipoint conferences.

REQ-14: The solution MUST support mechanisms to allow media from one source endpoint or/and multiple source endpoints to be sent to a remote endpoint at a particular point in time. Which media is sent at a point in time may be based on local policy.

REQ-15: The solution MUST provide mechanisms to support the following:

* Presentations with different media sources

* Presentations for which the media streams are visible to all endpoints

* Multiple, simultaneous presentation media streams, including presentation media streams that are spatially related to each other.

  The requirement relates to the Presentation use case.

REQ-16: The specification of any new protocols for the solution MUST provide extensibility mechanisms.

REQ-17: The solution MUST support a mechanism for allowing information about media captures to change during a conference.

REQ-18: The solution MUST provide a mechanism for the secure exchange of information about the media captures.

6. Acknowledgements

This document has benefited from all the comments on the CLUE mailing list and a number of discussions. So many people contributed that it is not possible to list them all. However, the comments provided by Roberta Presta, Christian Groves and Paul Coverdale during WGLC were particularly helpful in completing the WG document.

7. Security Considerations

REQ-18 identifies the need to securely transport the information about media captures. It is important to note that session setup for a telepresence session will use SIP for basic session setup and either SIP or the Centralized Conferencing Manipulation Protocol (CCMP) [RFC6503] for a multiparty telepresence session. Information carried in the SIP signaling can be secured by the SIP security mechanisms as defined in [RFC3261]. In the case of conference control using CCMP, the security model and mechanisms as defined in the Centralized Conferencing (XCON) Framework [RFC5239] and CCMP [RFC6503] documents would meet the requirement. Any additional signaling mechanism used to transport the information about media captures needs to define the mechanisms by which the information is secure. The details for the mechanisms needs to be defined and described in the CLUE framework document and related solution document(s).

8. Informative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC3261] Rosenberg, J., Schulzrinne, H., Camarillo, G., Johnston, A., Peterson, J., Sparks, R., Handley, M., and E. Schooler, "SIP: Session Initiation Protocol", RFC 3261, June 2002.

[RFC4353] Rosenberg, J., "A Framework for Conferencing with the Session Initiation Protocol (SIP)", RFC 4353, February 2006.

[RFC4566] Handley, M., Jacobson, V., and C. Perkins, "SDP: Session Description Protocol", RFC 4566, July 2006.

[RFC4579] Johnston, A. and O. Levin, "Session Initiation Protocol (SIP) Call Control - Conferencing for User Agents", BCP 119, RFC 4579, August 2006.

[RFC5117] Westerlund, M. and S. Wenger, "RTP Topologies", RFC 5117, January 2008.

[RFC5239] Barnes, M., Boulton, C., and O. Levin, "A Framework for Centralized Conferencing", RFC 5239, June 2008.

[RFC6503] Barnes, M., Boulton, C., Romano, S., and H. Schulzrinne, "Centralized Conferencing Manipulation Protocol", RFC 6503, March 2012.

[RFC7205] Romanow, A., Botzko, S., Duckworth, M., and R. Even, "Use Cases for Telepresence Multistreams", RFC 7205, April 2014.

[ITU.H323] ITU-T, "Packet-based Multimedia Communications Systems", ITU-T Recommendation H.323, December 2009.

Authors' Addresses

Allyn Romanow
Cisco Systems
San Jose, CA 95134
USA

EMail: allyn@cisco.com

Stephen Botzko
Polycom
Andover, MA 01810
USA

EMail: stephen.botzko@polycom.com

Mary Barnes
MLB@Realtime Communications, LLC

EMail: mary.ietf.barnes@gmail.com