CLUE WG A. Romanow
Internet-Draft Cisco Systems
Intended status: Informational S. Botzko
Expires: July 19, 2013 Polycom
January 15, 2013

Requirements for Telepresence Multi-Streams
draft-ietf-clue-telepresence-requirements-03.txt

Abstract

This memo discusses the requirements for a specification that enables telepresence interoperability, by describing the relationship between multiple RTP streams. In addition, the problem statement and definitions are also covered herein.

Status of this Memo

Copyright Notice

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 the 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.

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; 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. Non IETF protocol based systems, such as those based on ITU-T Rec. H.323, 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.

Telepresence systems of different vendors, today, can follow radically different architectural approaches while offering a similar user experience. It is not the intention of CLUE to dictate telepresence architectural and implementation choices. 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 two 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 must ever take into account information received. 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 to 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 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 [RFC4353], A Framework for Conferencing within the Session Initiation Protocol (SIP).

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 which source and sink media streams, and exactly one [RFC4353] Participant (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 nor 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 single screen/mono audio telepresence endpoint, and how rendered media streams are arranged with respect to each other on a multiple screen/speaker telepresence endpoint. Note that audio as well as video is encompassed by the term layout--in other words, included is the placement of audio streams on speakers as well as video streams on video screens.

Left: to be interpreted as a stage direction, see also [StageDirection(Wikipedia)]

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) also to adhere to.

Remote: Sender and/or receiver on the other side of the communication channel (depending on context); 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.

Right: to be interpreted as stage direction, see also [StageDirection(Wikipedia)]

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

Some degree of interworking between systems from different vendors is possible through transcoding and translation. This requires additional devices, which are expensive, often not entirely automatic, and they 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 sends audio and video captures based upon an implicitly assumed model for rendering a realistic depiction based on this information. If all endpoints are manufactured by the same vendor, they work with the same model and render the information according to the model implicitly assumed by the vendor. However, if the devices are from different vendors, the models they each use for rendering presence can and usually do differ. The result can be that the telepresence systems actually connect, but the user experience suffers, for example because 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 arranement 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 multi-stream work seeks to describe the sender situation in a way that allows the receiver to render it

realistically, though it may have a different rendering model than the sender; and for the receiver to provide information to the sender in order to help the sender create adequate content for interworking.

5. Requirements

Although some aspects of these requirements can be met by existing technology, such as SDP, or H.264, nonetheless we state them here to have a complete record of what the requirements for CLUE are, whether new work is needed or they can be met by existing technology. Figuring this out will be part of the solution development, rather than part of the requirements.

REQMT-1: The solution MUST support a description of the spatial arrangement of source video images sent in video streams which 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.

Use case point to point symmetric, and all other use cases.

REQMT-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.

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

REQMT-1c: The solution MUST support a means to identify the point of capture of individual video captures in three dimensions.

REQMT-1d: The solution MUST support a means to identify the extent of individual video captures in three dimensions.

REQMT-2: The solution MUST support a description of the spatial arrangement of captured source audio sent in audio streams which 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.

Use case point to point symmetric, and all use cases, especially heterogeneous.

REQMT-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 at Susan's right in the captured audio, John voice is also placed at Susan's right in the rendered image.

REQMT-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.

REQMT-2c: The solution MUST NOT preclude the use of binaural audio. [Edt. This is an outstanding issue. Text will be changed when the issue is resolved.]

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

REQMT-2e: The solution MUST support a means to identify the extent of individual audio captures in three dimensions.

REQMT-3: The solution MUST support a mechanism to enable a satisfactory spatial matching between audio and video streams coming from the same endpoints.

Use case is point to point symmetric, and all use cases.

REQMT-3a: 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.

REQMT-3b: The solution MUST enable individual audio streams to be rendered in any desired spatial position.

Edt: Rendering is an open issue. Text will be changed when it is resolved.]

REQMT-4: The solution MUST enable interoperability between endpoints that have a different number of similar devices. For example, one endpoint may have 1 screen, 1 speaker, 1 camera, 1 mic, and another endpoint may have 3 screens, 2 speakers, 3 cameras and 2 mics. Or, in a multi-point conference, one endpoint may have one 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.

Use case is asymmetric point to point and multipoint.

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

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

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

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

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

Use case heterogeneous and multipoint.

REQMT-10: The solution MUST make it possible for endpoints without support for telepresence extensions to participate in a telepresence session with those that do.

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

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

Use case multipoint.

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

REQMT-14: The solution MUST support mechanisms to make possible for either or both site switching or segment switching. [Edt: This needs rewording. Deferred until layout discussion is resolved.]

REQMT-15: The solution MUST support mechanisms for presentations in such a way that:

- Presentations can have different sources

- Presentations can be seen by all

- There can be variation in placement, number and size of presentations

REQMT-16: The solution MUST include extensibility mechanisms.

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

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

6. Acknowledgements

This draft has benefitted from all the comments on the mailing list and a number of discussions. So many people contributed that it is not possible to list them all.

7. IANA Considerations

There are no IANA considerations associated with this specification.

8. Security Considerations

Requirement Paragraph 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 CCMP for a multi-party 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 XCON Framework [RFC5239] and CCMP [RFC6503] documents would meet the requirement. Any additional signaling mechanism used to transport the information about media captures would need to define the mechanisms by the which the information is secure. These mechanisms would need to be defined and described in the CLUE framework document and related solution document(s).

9. 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.

[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.

[StageDirection(Wikipedia)]
Wikipedia, "Blocking (stage), available from http://en.wikipedia.org/wiki/Stage_direction#Stage_directions", May 2011, <http://en.wikipedia.org/wiki/Stage_direction#Stage_directions>.

Appendix A. Open issues

OPEN-1 Binaural Audio [REQMT-2C] The need to support of binaural audio is unresolved, and the "MUST NOT preclude" language in this requirement is problematic. The authors believe this requirement needs to be either changed or withdrawn, depending on how the issue is resolved.

OPEN-2 Reference to Rendering [REQMT-3b] This is the only requirement which refers to rendering. It may also be empty, since receivers can rendering audio captures as they wish. This is deferred until broader discussion on rendering requirements is concluded.

OPEN-3 Conference modes [REQMT-14] This wording of this requirement is problematic in part because the conference modes (site switching and segment switching) are not defined. It at least needs rewording. This is deferred until broader discussion on layout is concluded.

OPEN-4 Need to capture requirement that attributes can change at any time during the call.

OPEN-5 Need to add requirement for three dimensions in the right place

OPEN-6 Multi-view, is there a requirement needed?

Appendix B. Changes From Earlier Versions

Note to the RFC-Editor: please remove this section prior to publication as an RFC.

B.1. Changes from draft -02

Updated IANA section - i.e., no IANA registrations required.

Added security requirement Paragraph 18.

Added some initial text to the security section.

B.2. Changes from draft -01

Cleaned up the Problem Statement section, re-worded.

Added Requirement Paragraph 17 in response to WG Issue #4 to make a requirement for dynamically changing information. Approved by WG

Added requirements #1.c and #1.d. Approved by WG

Added requirements #2.d and #2.e. Approved by WG

B.3. Changes From Draft -00

- Requirement #2, The solution MUST support a means to identify monaural, stereophonic (2.0), and 3.0 (left, center, right) audio channels.

  changed to

  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.

- Added back references to the Use case document.

  * Requirement #1 Use case point to point symmetric, and all other use cases.

  * Requirement #2 Use case point to point symmetric, and all use cases, especially heterogeneous.

  * Requirement #3 Use case point to point symmetric, and all use cases.

  * Requirement #4 Use case is asymmetric point to point, and multipoint.

  * Requirement #9 Use case heterogeneous and multipoint.

  * Requirement #12 Use case multipoint.

Authors' Addresses

Allyn Romanow
Cisco Systems
San Jose, CA 95134
USA

Email: allyn@cisco.com

Stephen Botzko
Polycom
Andover, MA 01810
US

Email: stephen.botzko@polycom.com