Multi stream signalling

Espen Berger, November 3th 2013 (Revision 7 - WIP)

|  |  |  |
| --- | --- | --- |
| Revision 7 – June 2, 2014 | Espen Berger | Frame marking is a single byte, H264-SVC update, |
| Revision 6 – November 3, 2013 | Espen Berger | Added improved frame marking text, removed lync2013 comments |
| Revision 5 – September 3, 2013 | Espen Berger | RTCP PLI clarification, updated SDP examples |
| Revision 4 – June 17, 2013 | Espen Berger | Clarified MST , added audio level indication RFC6464, updated text on conference control |
| Revision 3 – April 23, 2013 | Espen Berger | Added comment on PLI and FIR usage |
| Revision 2 – March, 2013 | Espen Berger | Initial document |

# Introduction

This document describes Cisco multi stream signalling and simulcast SVC usage.

## Use cases

The multi stream signalling extensions are applicable to these use cases

* Initial call setup
* Negotiation of codec parameters and RTP sessions
* Negotiation of Simulcast SVC capabilities
* Allow middle boxes (e.g. conductor) to detect what an endpoints supports
* Extensibility to support multiple content and camera streams in future products

Some of the use cases will require additional application protocols to control content, e.g. meta-information for multiple content streams.

# Initial call setup

The mandatory call setup mechanism is SIP (RFC3261) with offer/answer (RFC4566) and SDP (RFC3266) for transport and codec parameters.

## Call flow

Regular SIP call flow will be used for call setup. When multi stream is active media will not be sent until the multi stream capability and request signalling is completed.

## Bandwidth

The recommended bandwidth modifiers are described in [IMTC-BANDWIDTH]. In SDP there are modifiers to specify constraints for the whole session and for individual media sessions.

* Session bandwidth b=AS:<value>. The value specified is the total bandwidth in kilobits per second for the session.   
  b=AS:1024
* Media session bandwidth b=TIAS:<value>. The value specified is the total media session bandwidth in bits per second. This means that the sum of all RTP streams in the media session must be below the TIAS value.
* Stream bandwidth a=max-br. This indicates the highest bandwidth in kilobits per second allowed for a single stream within the RTP session.

Examples

b=AS:1024

**m=**video 6002 RTP/AVP 96 97 34 31

b=TIAS:256000

a=rtpmap:96 H264/90000

a=fmtp:96 profile-level-id=428014;max-br=192

## RTP sessions

The multi stream architecture recommend using three RTP sessions, one for audio, one for main video and one for presentation video. Capabilities and stream request are specific to a media session.

A SDP example

m=audio 21000 RTP/AVP 127 …

b=TIAS:64000

a=rtpmap:127 MP4A-LATM/90000

a=sendrecv

m=video 21002 RTP/AVP 122 97 34 110

b=TIAS:1031000

a=rtpmap:122 H264/90000

a=sendrecv

a=content:main

a=label:11

m=video 21004 RTP/AVP 122

b=TIAS:1031000

a=rtpmap:122 H264/90000

a=sendrecv

a=content:slides

a=label:12

The content attribute specify the main role of the RTP session as defined in RFC4796.

The label attributes specify an RTP session identifier as defined in RFC4574.

## Codec negotiation

The supported codecs must be listed in the SDP offer to be available for selection by the remote device.

Multiple codecs should be supported in both endpoints and infrastructure to allow for negotiation of common media codecs and for supporting FEC + video protocols at the same time.

**m=**video 6002 RTP/AVP 96 97

a=rtpmap:96 H264/90000

a=rtpmap:97 H264-SVC/90000

An endpoint capable of receiving multiple codecs should list all the codec options in the SDP answer. It’s recommended that each endpoint and conference server support multiple codecs for improved interoperability.

In multi stream conferencing the preferred codec for on-prem is AAC\_LD and for cloud it’s Opus. In the RTCP SCR request for media the receiver indicate

m=audio 21000 RTP/AVP 127 120

b=TIAS:64000

a=rtpmap:127 MP4A-LATM/90000

a=fmtp:127 profile-level-id=24;object=23;bitrate=64000

a=rtpmap:120 G7221/16000

a=fmtp:120 bitrate=24000

## Codec limits

Individual codec limits can be described in SDP by using the codec specific attributes listed in the corresponding RFC. H.264 specifies the SDP attributes in [RFC6184].

When modifiers are specified in the SDP attributes the values put constraints on every RTP stream with the corresponding RTP payload type.

a=fmtp:122 max-fs=3600

# Simulcasting

Simulcast is defined as the ability to send out the same camera source in multiple qualities. Two features are used;

* Spatial qualities – The same source encoded into multiple spatial resolutions, e.g 360p and 720p.
* Temporal qualities – The same resolution encoded so that multiple temporal qualities (frame rates) can be extracted, e.g. 15/30 or 15/30/60.

Transmission notes

* All qualities should be sent within the same RTP session
* All spatial qualities must have different SSRCs
* Transmission for temporal qualities can support same SSRC for all temporal qualities (SST) or multiple SSRCs for temporal quality (MST).

## Merging of streams

All spatial and temporal qualities from the same media source must have the same CSI (Capture Source ID). The CSRC[0] is used for the CSI value.

## H264-SVC

H264-SVC is the preferred codec for spatial and temporal qualities. UCIF ucconfig mode 1 [UCIF], defines temporal qualities for individual streams.

Negotiation of H.264-SVC must be done in SDP offer/answer

m=video 30428 RTP/AVP 122 123  
a=rtpmap:122 H264-SVC/90000  
a=fmtp:122 packetization-mode=1;mst-mode=NI-T;uc-mode=1  
a=fmtp… // The regular fmtp params for H264 (AVC and SVC)

Additional notes

* An endpoint capable of doing mst-mode must support mst-mode = ‘NI-T’, which is the only mst-mode used in Cisco multi stream
* All endpoint must be capable of receiving and decoding a valid H264-SVC bitstream with all the temporal qualities in the same SSRC (SST-mode) .
* The ‘uc-mode=1’ attribute is used to negotiate UCIF configuration mode 1.
* Cisco will not use multiple spatial qualities within the same SSRC.
* A H264-SVC decoder must support receiving the optional PACSI and prefix nal units, to allow us to be H264-SVC complaint with vendors sending out the full set of RFC6190 nal units.
* A H264-SVC encoder should not send out PACSI NAL units in the base layer. This will allow a re-packetizing MCU to re-packetize from SVC to AVC and forward all packets.

## H265

A simulcast profile of H265 is defined as based on [H265]

* Spatial qualities are sent with unique SSRCs
* Temporal qualities are sent with tx-mode=MSM

# RTP header extensions

RTP header extensions are negotiated with [RFC5285].

## Virtual identifiers

A virtual identifier is used as a hop by hop identifier to map between stream requests and received media streams.

a=extmap:1/sendrecv http://protocols.cisco.com/virtualid

A virtual identifier is an 8-bit value in a RTP header extension. The value can be repeated when a single stream represent multiple virtual identifiers.

0 1 2 3

0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2

+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

| ID=1 |L=0,.N | VID1 | VID2 | VID3 | ++

+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

Comments

* There must be at least one virtual identifier present, that is Length = 0.
* When the length > 0, there are multiple virtual identifiers associated with the RTP packet, which indicates that the same RTP stream has been sent to fulfill multiple stream requests.
* All the virtual identifiers present should match the identifier values requested by the receiver.

## Frame marking

A switching conference server requires frame marking describing the encrypted payload for efficient processing.

a=extmap:2/sendrecv http://protocols.cisco.com/framemarking

Suggestion for a 1 byte of frame marking extensions

0 1

0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5

+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

| ID=2 | L=0 |D|S|TID |Type |

+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

Frame information

* Discardable (1 bit). The flag must be set to 1 when the frame is disposable. The flag must be true for packets that can be dropped and still provide a decodable media stream.
* Switching point - 1 bit (on|off). The flag must be true for only the first packet in a frame that can be used as switching points, e.g. IDR and GDR frame. A switching point is the first packet where a new receiver can start decoding a video stream without prior frames, e.g an IDR frame from [RFC6184].
* Type (3-bits). An abstract and codec neutral frame type; P=0, IDR=1, GDR=2, LTRF=3.
  + P-frame – A regular video frame referring to one historic frame
  + IDR – a single key frame without references to other frames
  + GDR – a key frame that are transported over multiple regular frames
  + LTRF – Long term reference frame, a frame typically designed to be used for resilience feature, where multiple references frames are kept by both encoder and decoder and reference selection is driven by feedback from receiver.
* T-ID (3 bits). The temporal identifier specifies the temporal dependencies between layers. A higher temporal-id are depending on lower temporal layers. A TID value of 0 does not have any temporal dependencies, and are referred to as the base temporal quality layer. (Both H264-SVC [RFC6190], H265 and other codecs like VP8 uses a 3-bit temporal id field.

Usage comments

* The frame type marking must be present for all packets.
* The switching flag must be set only for the first packet in an IDR frame or GDR sequence.
* The frame marking must match the NAL unit information.
* If len > 0, the first byte must be treated as the base frame marking. Additional bytes should be treated as extensions.

## Audio level

A multi stream conference solution should have a level indication added to each RTP audio packet to be able to calculate active speaker information without decoding audio on conference server.

Audio level indication is specified in RFC6464 and RFC6465.

a=extmap:3/sendrecv urn:ietf:params:rtp-hdrext:ssrc-audio-level

# Video profile

## Lip-sync

Lip-sync will be enabled for all RTP streams with matching RTCP SDES CNAME, see [RFC3550] for details.

## Key frame and picture loss

* RTCP PLI should be used for picture loss indication (RFC4585).
* RTCP FIR should be used for key intra frame request (RFC5104)

A device should use PLI to indicate packet loss and a server should use FIR to request a key frame to be used as switching point.

Both PLI and FIR must use the base layer SSRC.

# Conference control and media streams

Information required for rendering layouts with meta-information like names uses information available from the multi stream protocol and conference control (e.g. XMPP MUC or XCCP).

To link media streams with conference control meta-information the capture source identifiers should be distributed in the conference control protocol.

For conference control protocols based on RFC4575 the capture source identifier should be distributed in the <src-id> element.

<media>

<type>video</type>

<src-id>12345677…</src-id> // the capture source identifier

</media

# Contact header signaling

## Sip.cisco.multistream

The sip.cisco.multistream text must be added into the contact header for enabling multi stream in a SIP call.

sip.cisco.multistream // in contact header

## Number of screens signaling

For allocation of resources in Conductor, endpoints must signal the number of screens planned to be used to allow for proper resource allocation in conductor.

x-cisco-multiple-screen=<number>

x-cisco-multiple-screen=2 // e.g. an MX700

Additional comments

* The attribute should be present in the SIP contact header, when an endpoint is planning to use more than one screen for main video.
* The default value is 1, if attribute is not present.
* The attribute is reused from TIP specification.

# Security

Secure media will in phase 1 be implemented by doing SRTP [RFC3711] hop-by-hop.

# Resilience

Resilience to congestion and packet loss for multi stream will be provided according to MARI Tier 0 fundamentals and MARI Tier 1 mechanisms signaled in the SDP.

# Video capabilities

A multi stream receiver must support decoding of

* All natural aspect ratios used, at least the following; 3:4, 4:3, 16:9, 9:16, 16:10
* All resolutions, due to keeping the original resolution for content

# References

[RFC6184] - <http://tools.ietf.org/html/rfc6184>

[RFC6190] - <http://tools.ietf.org/html/rfc6190>

[IMTC-BANDWIDTH] - SIP Video Profile Best Practices, Date: 6 February 2013. <http://portal.imtc.org/DesktopModules/Inventures_Document/FileDownload.aspx?ContentID=21434>

[multiplex] - http://tools.ietf.org/html/draft-ietf-avtcore-multiplex-guidelines-02

[UCIF] – Unified Communication Specification for H.264/MPEG‐4 Part 10 AVC and SVC Modes V1.0

[H265] - <http://tools.ietf.org/html/draft-ietf-payload-rtp-h265>