SIP Requirements for support of Multimedia and Video

STATUS OF THIS RFC

This document is an Internet-Draft and is in full accordance with all provisions of section 10 of RFC 2026. Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to think of any Internet-Draft as an official standard.

Internet-Drafts are working documents of the Internet Engineering Task Force. An Internet-Draft is required to have at least two weeks of mailing list discussion before being sent to the IETF for discussion. An Internet-Draft must be current in the Internet-Draft Shadow Directories for at least six months but may be updated, replaced, or obsoleted at any time. It is inappropriate to think of any Internet-Draft as an official standard or requirement.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to think of any Internet-Draft as an official standard.

This document examines the way SIP/SDP/RTP/RTCP can be used today to support multimedia services and to deal with some of the presented requirements. In a small number of cases, this document mentions possible directions to enhance SIP in order to add new required functions or to provide the same functionality in a more efficient way.
used in. This information is presented to highlight some possible obstacles and interoperability problems that need to be considered in the way towards the desired networks convergence.

2. Multimedia Application Requirements

2.1 General

Each device has a set of its capabilities in terms of its CPU processing power, additional HW characteristics and the algorithms it supports. During a session's lifetime, its characteristics may be changing dynamically both as a result of network conditions and as part of a broader application (such as interactivity).

All the requirements, presented in this chapter, are required for providing basis reliable sessions: meaning establishing a session of a certain quality, based on capabilities agreement during the session establishment and sustained throughout the duration of the session.

The problem can be described as a lack of expressiveness in three following areas:

- **Capabilities Specification**
- **Resource Reservation**
- **Media Stream Control**

The desired ultimate goal is to:

- Express the capabilities (i.e. supported media, CODEC algorithms, bandwidth, etc.) without a need in configuration
- Signal the total resources, required for a specific session (probably in terms of capabilities)
- Within a session, explicitly open and close media streams, modify the parameters of a certain stream within the boundaries of previously announced capabilities and reserved resources

2.2 Capabilities Specification

2.2.1 Bit Rate

Each CODEC (COder DECoder) algorithm is defined to work at a certain rate or rates measured in bits per second.

SDP [2] has a concept of "Application-Specific Maximum Bandwidth" that can be applied to "m" (media) line and specified in kilobits per second.

The required functionality is to express CODEC capabilities for both ranges of rates and a discrete number of rates. The "discrete number of rates" requirement has a number of purposes:

- Efficient resources allocation
- Multipoint applications where the rates from different sources should be matched
- Interworking with terminals using multiplexing schemes, such as H.320 [12] and H.324 [13], in which only a discrete number of bit rates are available.

The defined capabilities may be used both for resource reservation and the actual control of a media stream.

2.2.2 Advanced CODEC schemes support

Video CODEC algorithms (among them H.263 [9]) have a concept of CIF ‘ Common Intermediate Format Definition with its derivatives: QCIF, 4CIF, etc. This definition of resolution implies the number of pixels and the format of the composed picture.

The challenge of providing quality video services over imperfect networks results in inventing of new coding algorithms with numerous optional operation modes fitting various environments. A famous example is the latest H.263 Recommendation ("H.263+') [10] that specifies a coded representation that can be used for compressing the moving picture component of audio-visual services at low bit rates and has numerous number of standard options described in its Annex.

All of the mentioned characteristics (bit rate, resolution, H.263+ options) are examples of CODECs capabilities. Many CODEC implementations are capable of changing the mode of their operation in real time, as a result of changing conditions, and signaling the new mode within the RTP payload header.

RFP has a definition of a specific RFP Payload header for each CODEC scheme carrying both the configuration and the dynamically changing segmentation information describing the format of the transmitted RTP packet.

In order to use various options dynamically during a session lifetime, without possibility of designing a call, a "capabilities announcement" mechanism should exist, implying modes of operation, supported by each of the sides.

In the future, SDPng work [22] may define a Language, expressive enough to describe different modes of...
operation. SIP may be extended to carry some of the parameters. SIP extensions may be needed to separate the "capabilities announcement" phase from the actual opening of media streams.

2.2.3 Lip Synchronization

One of the basic requirements for a multimedia session is the synchronization between audio and video streams when presenting them to the user.

2.2.3.1 Synch

The different timing of two media streams usually derive from an unequal processing time required for the encoding of the streams by the originating device. This difference is referred to as a skew. Skew is defined as a maximum time that the two media streams are delayed from each other as delivered to the transport network. Skew is usually measured in milliseconds.

Using RTP [1] timestamping services, devices in active multimedia sessions are capable of computing the skew and adjusting the buffers accordingly in order to provide the end-user video and audio display in a synchronized manner.

Additional useful functionality, helping to deal with the lip synchronization problem, is providing the receiving device with a skew metric before the actual media streams are transmitted. Knowing the maximum skew value in advance allows the receiving device, based on the skew metric, to adjust its buffers to the incoming media streams. This allows the buffers to be accurately increasing probability for a high-quality audio and video presentation. Currently this functionality is missing from the SIP/SDP/RTP/RTCP definitions.

G. Levin

2.2.3.2 Association of media streams

A single multimedia session may consist of multiple video or multiple audio streams, addressing multilingual requirements or CODECs multi-rate requirements respectively. In order to implement lip synchronization, an announcement between a certain video stream and its corresponding audio stream is required. Currently this functionality is missing from the SIP/SDP/RTP/RTCP definitions.

2.3 Resource Reservation

An end user may support more than a single CODEC scheme for a specific media type. The initial reason for choosing a specific CODEC (set of a list of supported CODECs) is to match the CODEC scheme supported by the codec used by the originating device. This functionality explicitly chooses a certain combination of audio and video CODECs without exceeding CPU processing power limitations or certain available transport bandwidth.

Using the same system with a defined CPU processing power and a support for testbed CODECs schemes you would be able to select the best CODEC scheme for the specific application. At some time you might want to receive a medical surgery session. In this example, the application running on your computer represents its CODEC's capabilities, CPU constraints and, any IP, the local network limitations. The other side, the equipment preparing the session has its own CODEC's requirements for providing a service with a certain quality. One of the possible ways to separate "resources ability" is by grouping the "capabilities", discussed above. It is an open question, is SIP, as a peer call control protocol, requires actual "resources reservation commands" being required in the other side.

2.3.1 Media Stream control

This section presents requirements to effectively control a particular media stream automatically. These requirements do not refer to manually driven commands such as floor control or camera control.

It is interesting to notice that despite the fact that users commands are not in the scope of a call control protocol (such as SIP), a correlation among them the media streams should be taken into consideration during the design of both protocols. For example, a particular camera (managed by some application protocol) should be synchronized with its originating video stream, managed by SIP.

Video requires broader control than voice. Some examples of video specific control are listed towards the end of this section. It is believed that a convenient means for media streams control is required for both voice and video. We believe that providing an accurate video stream control obviously complicates the problem.

Currently, in SIP/SDP/RTP/RTCP systems the media control commands are divided by SIP/SDP conventions and commands defined for certain CODECs and carried by "contact" RTP control packets, as in [4].

2.4.1 As ability to reference a specific media stream

The first basic requirement is the ability to reference a particular media stream within a session.

This functionality does not currently exist within the SIP. In order to locate a specific thing within a multiple media stream the entire list of media descriptors (i.e. CODECs) has to be examined. Moreover, the remaining side has to perform matching against the old information for each one of the m lines in order to recognize the changes.

This functionality should be fixed for the "new" SIP version, without waiting for the results of SDPng work.

2.4.2 Effective addressing of collision conditions during nonconcurrent operations

In most cases, the change of media stream parameters is to be requested within the same call. Therefore, both of the sides may issue a request for a conflicting change or command simultaneously, generating a so-called "race-condition".

Currently, SIP solves this problem by introducing the "early-after" header field in the bye/bye message. One of
the disadvantages of this approach is that it locks the whole session, when the collision exists in a certain media stream only.

2.4.3 Explicit start and stop of data transmitting in a certain direction

Explicit start and stop of data transmitting in a certain direction is the first basic "control command" out of set of media controls. Its necessity becomes obvious during the design of PSTN interworking. Early establishment of a media path in one direction only, strict billing regulations are just a few of the examples.

Today SIP addresses this functionality by "putting media streams on Hold" by setting "c" destination address to value "0.0.0.0".

2.4.4 Bandwidth changes

Applications involving video are particularly prone to frequent bandwidth changes causing packets lost, error conditions, etc. The first cause for frequent changes is the network changing conditions. Future IP based wireless networks will become a real test bed for SIP services. Similar changing conditions would frequently be caused by the "multimedia nature" of the video sessions. Some examples are presented below.

Today, in any integrated services, "multimedia communication" includes a data session (such as T.120 [16]) bundled together with voice and video sessions. Opening and closure of the data stream may significantly change the desired parameters of the media streams.

The same effect may exist in multimedia applications, where instead of using G.728 derivatives, video streams are presented to the user in separate windows. This mode of operation may be advantageous to the user, who has a better control over the session and can use network resources in a more effective way.

Additional example of an "application condition" in a multi-conferencing service, where adding of a participant may result in a change of transmitted stream parameters or in a reconsideration of conference capabilities.

2.4.5 Video/CODEC Specific Commands

Various video specific techniques have been used in today's networks in order to cope with the conditions mentioned above with minimal service degradation and as seamlessly as possible to the users. Below are some of the examples:

H.261 and H.263 video CODECs have a notion of picture's building blocks: "full picture", GOB and MacroBlock (MB). The decoder would have an ability to recognize synchronization degradation and explicitly request from an encoder for a "full picture", a whole GOB or a whole MacroBlock.

In SIP/SDP/RTP/RTCP systems, the only analogous functionality is defined in RFC-2032 "RTP Payload Format for H.261 Video Streams" [4] that defines a "Full INTRA-frame Request" (FIR) to be carried in RTCP "reverse" control packet. This technique is definitely an exception in normal RTP design and therefore does not work in all cases. No corresponding functionality has been defined for H.263 CODEC in [5] and [6].

A simple example of a video specific command is a request to "freeze a picture" in this case originated from the encoder towards the decoder. In case the encoder is aware of oncoming massive changes in the transmitted picture, it would request the decoding side to stop presenting the changes, until a new stable image is encoded and transmitted.

Another inherent example of a video specific command is a request to change the tradeoff between temporal and spatial resolutions, i.e. the tradeoff between the rate of the samples and the resolution of the picture. This request would be originated by the decoder towards the encoder (if the encoder has the capability to dynamically change the tradeoff).

2.4.6 Transmission of media stream commands

Today the media stream commands are transmitted by RTCP. "FIR" command defined for H.261 and described above is the example for this technique. The benefit of this approach is that the commands flow the same path as the media stream itself and therefore are synchronized in time.

On the other hand, the RTCP approach has a number of following drawbacks:
- RTCP is an unreliable transport channel
- RTCP mechanisms were originally designed to work in primarily multicast environment. This may result in an incoherent sequence of commands issued in both directions
- Difficulty in RTCP potential use of capabilities defined by SIP (e.g. [5] and [6]).

3. Interoperability with existing video systems

Today several protocols are defined to support multimedia system both for Circuit Switched and Packet networks. Part of them such as H.323[11] and H.324M[12] are deployed. Others (such as H.320M) are intended to be used in future networks.
It is important to be aware of the architecture of these systems and the interesting challenges they introduce. Below is the list of these specifications.

3.1 H.320

3.2 H.324 and H.324M

3.3 H.323

4. Conclusion
This draft is a first attempt to present and summarize issues needed for video services support in SIP/SDP/RTP/RTCP systems. Fast of the solutions may be defined in a short period of time. More advanced features or complicated problems will be resolved in the future by extending the existing frameworks.

It is important to be aware of the current limitations or open issues in the standard, based on the imperativeness of the requirements, SIP allows for extensions adding functionality in a standard interoperable manner.

An alternative possible approach might be the definition of conventions for certain SIP based multimedia systems.

5. Security Considerations
This document does not introduce new security requirements to existing SIP/SDP/RTP/RTCP systems.

6. References


7. Acknowledgements

Sasha Ruditsky, Yair Mirasla, Eli Zorar, Itamar Gilad andDanny Levin participated in earlier discussions on this topic.

8. Authors' Addresses

Orit Levin
RADVision Inc.
575 Corporate Drive Suite 420
Mahwah, NJ 07430
Phone: +1 201 529 4300
Email: orit@radvision.com

Full Copyright Statement

Copyright © The Internet Society (2000). All Rights Reserved.

This document and translations of it may be copied and furnished to others, and derivative works that comment on or otherwise explain it or assist in its implementation may be prepared, copied, published and distributed, in whole or in part, without restriction of any kind, provided that the above copyright notice and this paragraph are included on all such copies and derivative works. However, this document itself may not be modified in any way, such as by removing the copyright notice or references to the Internet Society, except as needed for the purpose of developing Internet standards in which case the procedures for copyrights defined in the Internet Standards process must be followed, or as required to translate it into languages other than English.

The limited permissions granted above are perpetual and will not be revoked by the Internet Society or its successors or assigns.

This document and the information contained herein is provided on an "AS IS" basis and THE INTERNET SOCIETY AND THE INTERNET ENGINEERING TASK FORCE DISCLAIMS ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.