Technical Disclosure Commons

Defensive Publications Series

January 2020

METHOD FOR FAST BOOTSTRAP AND LOSS RECOVERY FOR CLIENTS CONNECTED TO A SELECTIVE FORWARDING UNIT (SFU)

Jacques Samain

Mohammed Hawari

Andre Surcouf

Follow this and additional works at: https://www.tdcommons.org/dpubs_series

Recommended Citation

Samain, Jacques; Hawari, Mohammed; and Surcouf, Andre, "METHOD FOR FAST BOOTSTRAP AND LOSS RECOVERY FOR CLIENTS CONNECTED TO A SELECTIVE FORWARDING UNIT (SFU)", Technical Disclosure Commons, (January 22, 2020)
https://www.tdcommons.org/dpubs_series/2882

This work is licensed under a Creative Commons Attribution 4.0 License.
This Article is brought to you for free and open access by Technical Disclosure Commons. It has been accepted for inclusion in Defensive Publications Series by an authorized administrator of Technical Disclosure Commons.

METHOD FOR FAST BOOTSTRAP AND LOSS RECOVERY FOR CLIENTS CONNECTED TO A SELECTIVE FORWARDING UNIT (SFU)

AUTHORS:

Jacques Samain
Mohammed Hawari
Andre Surcouf

ABSTRACT

Proposed herein is a technique to reduce the time to obtain an I-frame for video conferencing participants that are unable to decode video stream(s) of the main active speaker(s) by decoding video stream(s) of the main active speaker(s) at a Selective Forwarding Unit (SFU) and re-encoding the video stream(s) using only I-frames. The technique also provides for the ability reduce network traffic by unicasting the I-frames only to participants that may need the I-frames, rather than broadcasting the I-frames to all participants.

DETAILED DESCRIPTION

There are several architectures that can be utilized for video conferencing. For example, a mesh architecture can be utilized in which each participant can establish peer-to-peer connections with each other. In a Multipoint Conferencing Unit (MCU) architecture, a central node gathers media streams from all participants and subsequently encodes and sends a composite stream to all the participants. In a Selective Forwarding Unit (SFU) architecture, a central node gathers media streams from all participants and selectively forwards some of the video streams to the participants.

Due to bandwidth constraints, video conferencing relies on streams that are compressed via codecs, such as VP8. These codecs rely on the transmission of self-sufficient frames, denoted herein as 'I-frames', and partial frames, denoted herein as 'P-frames', in which decoding relies on some context provided by previous frames (e.g., the last I-frame, as well as some previously received P-frames).

When a client joins a video call, the client needs to receive an I-frame for each video stream in order to correctly decode each stream. As I-frames are issued periodically by video encoders and the time at which a client may join a call can vary, some delay may be involved before the client receives an I-frame and, thus, is able to decode a video stream.

Some mechanisms exist to expedite the retrieval of an I-frame, such as a client sending a Full Intra Request (FIR) to a video sender to notify the encoder of the video sender to send an I-frame. However, sending an I-frame increases the bandwidth usage on the uplink of the sender, which can be even more constrained at its downlink (e.g., when connected via cellular or wireless connectivity). Moreover, if a SFU-based architecture is used, such an I-frame will be forwarded to all the clients, resulting in further wasted (downlink) bandwidth.

In addition, if there are losses and the decoder at a client cannot read a frame, a Picture Loss Indication (PLI) will be issued by the client (since sending a FIR in this case is explicitly disallowed, as per Internet Engineering Task Force (IETF) Request For Comments (RFC) 5104). Upon reception of a PLI, the encoder can decide to send a new I-frame to achieve a quick resynchronization at the client. However, sending an I-frame has the same shortcomings as noted above.

This proposal provides a technique that involves video conferencing within an SFU architecture in which the time to transmit the necessary information to a given participant may be reduced for the participant to decode a video. Specifically, the technique may be used either for a participant joining a call and not having the contextual information necessary for decoding (e.g., the last image) or may be used for a participant that has suffered from a packet loss involving the loss of a critical piece of information needed for decoding (e.g., the last image).

The technique of this proposal avoids unnecessary network traffic by keeping the incurred data exchange between the SFU and a given participant. The technique of this proposal can be applied to video conferencing applications that rely on codecs such that the decoding of any frame may only rely on a small number of previous frames.

Typically in VP8, the decoding of any frame may only rely on the previous frame and another frame from the past (e.g., a golden-frame or an alternative reference (altref) frame). At any point in time, the set of frames that may be utilized to decode a current frame may be referred to as a set of 'contextual frames'. With this terminology, the contextual frames of VP8 may include the previously decoded frame and the last golden-frame/altref frame. Note that these frames are not necessarily transmitted as self-sufficient pieces of data (I-frames), but may themselves rely on previously transmitted frames, and so on.

The technique of this proposal provides a new feature at the SFU in which video streams may still be forwarded in a normal manner. For video streams of the main active speaker(s), the streams are also decoded and, at any point in time, a set of contextual frames can be cached and encoded as I-frames, such that they may be self-sufficient and usable by any participant requesting an FIR or a PLI.

Figure 1, below, illustrates example details associated with the technique of this proposal. For the sake of simplicity, only one active speaker is illustrated in Figure 1, however, it is to be understood that this technique can also be utilized for several active speakers.

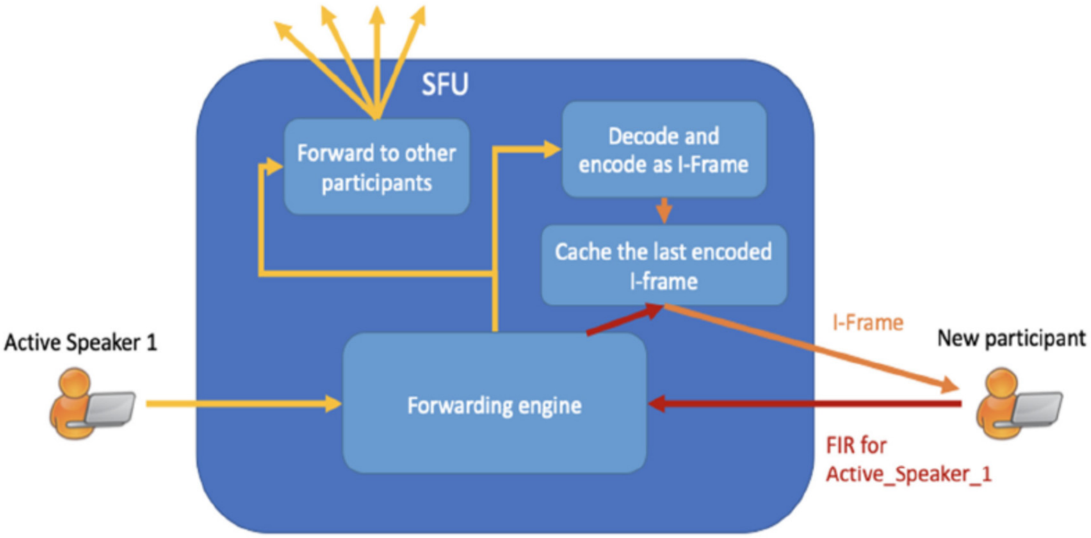
SFU
Forward to other participants
Decode and encode as I-Frame
Cache the last encoded I-frame
I-Frame
Active Speaker 1
New participant
Forwarding engine
FIR for Active_Speaker_1

Figure 1

For the example illustrated in Figure 1, the active speaker sends its video stream to the SFU as a stream of I- and P-frames. The SFU forwards the streams to the other participants and also decodes the stream and re-encodes the contextual frames as a stream of I-frames. Thus, the SFU always has an up-to-date I-frame cached for the video stream of the active speaker.

When a new participant joins the call, the SFU can start sending the new participant the video stream of the active speaker. However, if the first frame that the new participant receives is a P-frame, the new participant will issue an FIR for the video stream of the active speaker. Rather than propagating the FIR to the sender, the SFU is capable of directly sending the last encoded contextual frames to the new participant, so that the new participant is able to decode further P-frames from the active speaker.

Although the example of Figure 1 is focused on a participant joining a call, the technique of this proposal can also be used to recover from picture loss events at a participant. For example, if a participant sends a PLI for a video stream of an active speaker, the SFU can also intercept the PLI and send contextual I-frames to the participant on behalf of the active speaker.

Several factors may be considered for implementing this technique. For example, for Real-time Transport Protocol (RTP) implementations, SFU re-encoded contextual I-frames should have the same RTP sequence numbers as the P-frames originating from a sender, since participants rely on RTP sequence number to detect loss. Considering rate adaptation, when an active speaker's encoder changes its settings, the encoder at the SFU is to reflect these changes as well. For security considerations, since the SFU decodes and encodes video streams from active speakers, the participants may be configured to trust the SFU.

Further, for considerations involving a change of an active speaker, when the active speaker changes, the SFU may spawn a new decoder/encoder and can delete the decoder/encoder for a participant that is no longer an active speaker. For resource allocation considerations, the technique may be enhanced such that it may be applied only to the main active speakers (e.g., the first three in the list) to limit impacts of this technique on central processing unit (CPU) load/utilization.

The technique of this proposal may provide several advantages. For example, the time to video can be reduced for a participant. As an FIR from a participant can be intercepted at the SFU, which can directly send the contextual I-frames to the participant, one round-trip time (RTT) can be saved between the active speaker and the SFU. Additional savings may be realized based on the time that would have otherwise been involved for the encoder of the active speaker to encode an I-frame.

Further, the technique of this proposal avoids unnecessary network traffic. As noted for current deployments, an FIR from a given participant is communicated to an active speaker and a subsequent I-frame is forwarded to all participants. As I-frames are typically bigger in size than P-frames, the network load for current deployments may be momentarily increased for all participants based on transmission of the I-frame to all participants even though it may have been needed by only one of the participants. Such network traffic may be avoided by utilizing the technique of this proposal.

In still another example, if there is a loss between an active speaker and the SFU, the decoder at the SFU may identify the loss with the active speaker quicker than participants may identify the loss such that the SFU can act on the loss (e.g., retransmit a frame, send a PLI to the sender, etc.).

In summary, this proposal provides a technique to reduce the time to obtain an I-frame for video conferencing participants that are unable to decode video stream(s) of the main active speaker(s) by decoding the video stream(s) of the main active speaker(s) at an SFU and re-encoding the video stream(s) using only I-frames. Thus, the technique provides for opportunistically reconstructing I-frames at a video bridge and sending such frames only to the clients that may need an I-frame. By unicasting the I-frames only to participants that may need the I-frames, rather than broadcasting the I-frames to all participants, the technique also provides for the ability reduce network traffic.