Thesis Summary

Name: Boris Grozev
Thesis title: Efficient and Scalable Video Conferences With Selective Forwarding Units and WebRTC

Video conferencing is becoming more and more accessible to people as the hardware, networks and software systems improves. The WebRTC technology brings a state of the art multimedia stack and real-time communication capabilities to browsers, which makes conferencing available to anyone with a web browser or a mobile device. This creates a great opportunity for the development of rich multimedia web applications, and video conferencing applications in particular. The topology that such applications use, i.e. how the endpoints are connected to one another, is very important for the user experience and for the cost of the service. There are a few different architectures which have been used, and in recent years centralized architectures using Selective Forwarding Units (SFUs) have gained popularity. They have some important advantages, such as low end-to-end delay and low CPU use, but there are also some challenges and problems unique to SFUs which need to be solved in order to provide the best possible service. Perhaps the most important one being managing the bandwidth required for a conference.

In this thesis we examine different ways to improve SFU based conference to make them more efficient and scalable. It is based in part on the research in the publications listed below.

In the first part (see the plan below) we concentrate on managing the required bandwidth. We first present the idea of LastN[1], in which we automatically order the endpoints in the conference by their speech activity and only forward a subset of streams. We then present an experimental evaluation of simulcast[2], which is a mode in which endpoints encode and transmit multiple versions of their stream. We investigate the effects on the senders, servers and receivers in terms of CPU use, bandwidth, and image quality. We find that these two features greatly reduce the bandwidth required for a given conference.

Next we look at the bandwidth required for the conferencing service as a whole, and detail the implementation of a system which uses peer-to-peer connections for one-to-one calls[5] (reducing the load on the infrastructure) and switches to a server based connection seamlessly when the conference size grows to more than two. We examine the different sets of features that are better used in each mode and we perform a measurement study to quantify the impact on the infrastructure cost as well as the connection quality.
In the rest of the chapters of part 1 we present other improvements and optimizations for SFU based conferences. We look at server performance optimizations, improving the initial bitrate of streams, and we propose a novel way to multiplex sessions using a single IP address/port pair.

In Part II we move on from conferences using a single SFU to a distributed model. We present Considerations for deploying a geographically distributed video conferencing system[4], in which we explore the problems, pitfalls and potential gains from distributing a conference across multiple media servers located in different geographical regions. We find that a lot depends on the type of conferencing service, the user base and the ways that users use the system. We analyze those for one particular production service and propose ways to improve it.

We also detail the implementation of a system for cascaded SFUs, and look at some experimental results.

Part III introduces another topic: end-to-end encrypted conferences. We present PERC double media encryption for WebRTC 1.0 sender simulcast[3] in which we propose a way to extend the PERC[8] system for end-to-end encrypted conferences to support simulcast. Specifically we extend the additional PERC headers in RTP packets with fields to contain the original SSRC and timestamp, allowing a media distributor (a.k.a a Selective Forwarding Unit) to modify those fields. We also propose ways to adapt PERC to support other mechanisms used in WebRTC today such as bandwidth probing using padding packets and packet retransmissions using RTX. We followed up this publication with two internet drafts[6,7], but they did not gain traction in the working group.
Thesis plan

This is an abbreviated table of contents of the thesis.

1. Introduction and State of The Art
   Part I: Managing bitrate in SFU-based conferences, and other optimizations
2. Last N: Relevance-Based Selectivity
3. Simulcast for real-time conferencing
4. Switching between SFU and full mesh
5. Performance optimizations for SFUs
6. Rapid optimization of initial media stream bitrate
7. Multiplexing sessions in using Interactive Connectivity Establishment
   Part II: Conferences with cascaded SFUs
8. Problems and goals
9. The Octo protocol
10. On-demand signaling and selective internal forwarding
11. Selective Internal Forwarding
    Part III: End-to-end encrypted conferences with an SFU
12. Approaches for end-to-end conferencing and the PERC effort
13. Using PERC with simulcast and WebRTC
List of publications


[4] Considerations for deploying a geographically distributed video conferencing system, Boris Grozev, George Politis, Emil Ivov, Thomas Noël, 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 8-10 Jan. 2018


References